PhD–FDEF–2022–15
The Faculty of Law, Economics and Finance

DISSERTATION

Defence held on 23/09/2022 in Esch–sur–Alzette
to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN SCIENCES ÉCONOMIQUES

by

NICCOLO' GENTILE
Born on 28 June 1992 in Barletta (Italy)

ESSAYS ON THE ECONOMICS OF WELLBEING AND MACHINE LEARNING

Dissertation defence committee

Dr. Conchita D'Ambrosio, dissertation supervisor
Professor, Université du Luxembourg

Dr. Alexandre Tkatchenko, chairman
Professor, Université du Luxembourg

Dr. Gautam Tripathi, vice–chairman
Professor, Université du Luxembourg

Dr. Andrew Clark, expert in advisory capacity
Professor, Paris School of Economics

Dr. Michela Gianna Bia, member
Research Fellow, Luxembourg Institute for Socio–Economic Research

Dr. Chiara Binelli, member
Professor, Università di Bologna

To my family, my friends, and my supervisors.

Table of Contents

Abstracts

In Chapter 1, we apply Machine Learning (ML) methods to predict and interpret life satisfaction using data from the UK British Cohort Study. We discuss the application of first Penalized Linear Models and then of one non–linear method, Random Forests. We present two key model–agnostic interpretative tools for the latter method: Permutation Importance and Shapley Values. With a parsimonious set of explanatory variables, neither Penalized Linear Models nor Random Forests produce major improvements over the standard Non–penalized Linear Model. However, once we consider a richer set of controls, these methods do produce a non–negligible improvement in predictive accuracy. Although marital status and emotional health continue to be the most important predictors of life satisfaction, as in the existing literature, gender becomes insignificant in the non–linear analysis.

In Chapter 2, we further assess the potential of ML to help us better understand wellbeing. To do so, we analyze wellbeing data on over a million respondents from Germany, the UK, and the United States. In terms of predictive power, ML approaches do perform better than traditional models. Although the size of the improvement is small in absolute terms, it turns out to be substantial when compared to that of key variables like health. We moreover find that drastically expanding the set of explanatory variables doubles the predictive power of both OLS and the ML approaches on unseen data. The variables identified as important by our ML algorithms – *i.e.*, material conditions, health, and meaningful social relations – are similar to those that have already been identified in the literature. In that sense, our data–driven ML results validate the findings from conventional approaches.

In Chapter 3, we predict and analyze the determinants of health. There is a change in the target compared to the previous two chapters: we now focus on objective health outcomes. In particular, ML methods are applied to predict health outcomes in the German Socio–Economic Panel, under two specifications: pooling data across multiple years, and applying the Mundlak transformation on the same pooled data. The dependent variable of interest is Number of doctor visits in the last three months. We discuss the application of ML Regression and Clustering techniques, and after

presenting the different nature of the independent variables, and the rationale behind the choice of the considered ML algorithms, we present the findings, using accuracy scores suited to compare all models. The analysis of the distribution of the variables in the clusters created by the algorithm, along with novel model–agnostic interpretative tools (Shapley Values), allows us to better interpret the results. We find that ML algorithms – Random Forest in our case – lead to large improvements in predictive accuracy, especially in clusters. Self–rated measures of health, gender and disability status represent the most important drivers in healthcare utilization, in line with the existing literature.

Co–author Statement

The last chapter:

“Healthcare utilization and its evolution during the years: building a predictive and interpretable model”

is a single–author paper. In the second one:

“Human Wellbeing and Machine Learning”

the first authorship is jointly shared between me, Dr. Ekaterina Oparina, and Dr. Caspar Kaiser. This reflects our contributions. In order to decide the order in the listing of our names, we further used the AEA randomization tool – confirmation code available at: https://www.aeaweb.org/journals/policies/random–authororder/search?RandomAuthorsSearch%5Bsearch%5D=oJsh_ZMZJwhH.
By randomness, the list happened to be Oparina, Kaiser, Gentile. The other authors include Prof. Alexandre Tkatchenko, Prof. Andrew Clark, Prof. Jan–Emmanuel De Neve, and Prof. Conchita D’Ambrosio.

In the first one:

“What Makes a Satisfying Life? Prediction and Interpretation with Machine–Learning Algorithms”

I am the only first author, reflecting my contributions. The other co–authors are Prof. Andrew Clark, Prof. Conchita D’Ambrosio, and Prof. Alexandre Tkatchenko.

General Introduction

Human wellbeing and life satisfaction: normal and crisis times

The study of the determinants of physical and psychological wellbeing is at the core of research in Economics. For instance, Richardson *et al.* (2014), analyzing the association between the notions of subjective wellbeing and utility, conclude that there exists a strong correlation between subjective wellbeing and the psycho–social components of multi attribute utility. As of today, there is a vast and comprehensive literature regarding self–assessed wellbeing, which is more in detail presented and described in the Introductions of Chapter 1 and Chapter 2. Overall, as discussed in Clark and Lepinteur (2022), it has been shown that despite its subjective nature, questions relating to wellbeing and life satisfaction are associated with brain activity (Urry *et al.* 2004), likelihood of marital breakup (Guven *et al.* 2012), quitting your job (Clark 2001), productivity (Oswald *et al.* 2015), and voting preferences (Liberini *et al.* 2017, Ward 2020). In recent times, the theme of subjective wellbeing has attracted increasing interest, as a consequence of the common psychological struggles associated with the COVID–19 pandemic. For instance, D'Ambrosio *et al.* (2021) show, using data from the COME–HERE surveys, an increase in life satisfaction from March 2021 to July 2021 in Luxembourg, a period of loosening of the containment measures.

Considering 643 people, when asked about reporting their degree of life satisfaction from 0 to 10 (0 being "completely dissatisfied"), replies from 8 to 10 were observed, respectively, 19%, 8.9% and 9.9% of the times in March 2021, and 22.2%, 11.8% and 15.7% of the times in June 2021. Using the same data, Clark and Lepinteur (2022) show a strong negative correlation between life satisfaction and the *stringency index*, a measure consisting of nine indicators representing different containment measures: the average life satisfaction reported in their sample is 6.34, well below the historical 7–to–8 averages for OECD countries in normal times. In a similar study, Dymecka *et al.* (2021) observed a negative correlation between the fear of COVID–19 and the sense of coherence, health–related hardiness and life satisfaction.

From subjective to objective health – healthcare utilization

When it comes to objective health, the literature on the application of Machine Learning algorithms is increasing by the day. For instance, Toh and Brody (2021) define three areas in which Machine Learning applications are leading to particularly interesting improvements in healthcare, including medical imaging, natural language processing of medical documents and analysis of genetic data. The aim of the research in these areas is to increase diagnostic accuracies, detection of anomalies and improve genetic–rooted predictions. One specific example is Roth *et al.* (2016), using Convolutional Neural Networks (CNNs) to predict colonic polyps, sclerotic spline metastases and enlarged lymph nodes from CT scans image. Similarly, Dou *et al.* (2016) use the same algorithm to detect cerebral microbleeds from susceptibility weighted MRI scans. As can be noticed, all these applications in the healthcare domain are specific to one diagnosis or physical condition. The aim of this work is instead to predict and interpret the determinants of *healthcare utilization*, considered traditionally in the literature via measures like number of doctor visits or nights spent in hospital in a given period. In these cases, the demand for healthcare utilization is estimated starting from a broader set of individual characteristics rather than unstructured data like images or sounds. A thorough review of the existing literature in health economics, and healthcare utilization in particular, is provided in the Introduction of Chapter 3.

The importance of predicting for policymaking

Two of the key characteristics of all the quantitative studies about wellbeing and healthcare utilization are that estimations are performed mostly considering parametric inflexible methods, and second that only parsimonious models – in terms of included independent variables – are considered. One key consequence of these two practices is that we are capable of predicting only a small fraction of the variability of healthcare and wellbeing, as measured via the R–squared. The key reason behind these choices is that, in research, a higher attention is traditionally paid to model that lead to readily interpretable results – as for instance a Linear Regression and the estimated coefficients – at the cost of low predictive accuracy. In terms of policy applications, the possibility of intuitively establishing a quantitative relationship between the dependent and the independent variables – and, with due attention, also causal – is particularly appealing. However,

there are instances in which predictive accuracy becomes even more important in the context of policymaking. An example in this direction is the work of Kleinberg *et al.* (2015). They used a Machine Learning algorithm (Least Absolute Shrinkage and Selection Operator – *LASSO*) to better predict the rate of survival of elderly potential recipients of joints replacement surgery. In a simulation study, they concluded that the obtained increase in predictive accuracy could lead to the possibility of avoiding more than 10,000 surgeries to individuals who would die within 1 to 12 months after the surgery itself, and that therefore would never experience its benefits (and only bore the physical and economic costs).

Another example of particular relevance of accurate predictions for policymaking is Bansak *et al.* (2018). In their work, they start noticing that in Switzerland refugees are assigned across the cantons randomly, only taking into account a proportionality criterion. Using as measure of successful integration the probability of being employed within three years, they fit a Machine Learning model to better redistribute the refugees. They find that with their method, the probability of employment within three years increased from the current 15% to 26%. They also observed similar results (from 25% to 50%) in the US.

In our context, being able to better predict wellbeing and healthcare utilization has important policy implications. Subjective wellbeing has a strong correlation with depression (Gigantesco *et al.*, 2019, Lagnado *et al.*, 2017, World Health Organization latest guidelines), and can therefore lead to informed policy measures (Dolan *et al.*, 2012). Being able to promptly identify subjects at risk of depression can lead to timely life–saving interventions. In a similar fashion, predicting who is more in need of healthcare can lead to a more efficient healthcare market by reducing moral hazard. The effect of moral hazard on the health insurance market is for instance described in Breyer *et al.* (2004). In this case, accurate predictions also on the lower end of the distribution (low need for healthcare) are important, since the possibility to redistribute resources to whom is more in need of healthcare (and can't afford it) is directly related with avoiding allocating resources to individuals who either don't need healthcare or can already afford it. A highly inflexible algorithm like the Linear Regression may perform particularly poorly in predicting the tails of the distributions. Moreover, the estimated coefficients represent average marginal effects, and not specific for each individual. In order to address this issue, in Chapter 1 and Chapter 3, to

interpret the findings produced by the Machine Learning algorithms, we use *Shapley Values*: differently from the coefficients of a Linear Regression, the Shapley Values allow to interpret the marginal contribution of a given independent variable not on average, but at the individual level. That is, the Shapley Value of income in predicting life satisfaction tells us its marginal effect at each level of income: including or not income among the predictors changes life satisfaction in a different way for people earning, say, 100.000 GBP a year than how it does for people earning, say, 10.000 GBP. This is an information we would not immediately infer simply considering the Linear Regression coefficients.

In Chapter 3, moreover, given the sufficiently large sample size, we also clustered the data based on the independent variables, trying to find homogenous groups of individuals in which the prediction task would become easier. However, differently from the traditional literature in Econometrics (Wooldridge, 2003), we did not ex–ante choose on which variables to cluster, but rather we considered an *Unsupervised Learning* algorithm called *K–Means–Clustering*, hence letting the machine choose by itself how to better organize the data. Throughout all the chapters, all the predictive tasks were performed considering *Machine Learning* algorithms, benchmarked in their predictive power against the Linear Regression.

Artificial Intelligence and Machine Learning: an introduction

Almost everyday, across the world–leading newspapers and news channels, there is at least one article or report about Artificial Intelligence (AI) and its applications, whether talking about some new discovery, a new regulatory debate or issues caused by AI powered systems. AI is a discipline with a long history. Despite its recent spotlight in the news, AI is widely considered to be born in 1956, during the *Dartmouth Summer Research Project on Artificial Intelligence* workshop. In particular, John McCarty, the computer and cognitive scientist who coined the term, defined AI as "the science and engineering of making intelligent machines" (McCarthy *et al.*, 1956). In this context, the quintessential expression of intelligence, both in animals (including humans) and machines, is *learning*. For this reason, *Machine Learning* (ML) is commonly described as a branch of AI. More precisely, ML is defined as "the field of study that gives computers the ability to learn without being explicitly programmed", as per definition of Samuel (1959). Similarly, according to Mitchell (1997), "A computer program is said to learn from experience E with respect

to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E".

In this work, we focus our attention on *Statistical Learning*: building on the previous definition, we define the experience E as the *data* we are considering, the task T as the *prediction of a dependent variable*, and the performance measure P as the *degree of accuracy* in predicting our dependent variable.

The scientific literature around Machine and Statistical Learning is rich, ranging from high level learning material to frontier research. A thorough description of Statistical Learning Theory is beyond the scope of this work, and we remind the reader to cornerstone works in the discipline like *"The Elements of Statistical Learning",* Hastie *et al.* (2009), and *"Machine Learning and Pattern Recognition"*, Bishop (2006). Here, the focus will be in describing the basic principles of the discipline, focusing in particular on the challenges posed by the minimization of the *Expected Generalization Error* and the role of *interpretability*, since particularly relevant across the chapters.

Machine Learning: predicting and interpreting

Machine Learning algorithms can be divided in multiple subgroups, depending on the specific task they aim to solve. In this work, the focus is on two categories: *Supervised* and *Unsupervised Learning* algorithms.

- *Supervised Learning (SL)*: In SL, the goal is to learn a pattern (an approximating function) between a set of *independent variables* and a *dependent variable* we are interested in predicting. We do so by fitting our algorithms on a group of individuals known as the *training set*, hence obtaining an approximation function (*i.e.*, learning a pattern between the independent variables and the dependent one). Then, we want to use the same learned pattern to predict the dependent variable over a new group of individuals, the *test set*, and then evaluate the accuracy of the model by comparing the predicted and the actual values of the dependent variable on this group. If the dependent variable is a discrete number or a class – *i.e.*, spam vs. non spam email – we talk about *Classification*, whereas if it is a continuous variable – income, height, weight – we talk about *Regression*. In the literature,

the independent variables are also known as *features* or *inputs*, whereas the dependent variable is also known as *target* or *outcome*. Examples of Classification and Regression methods are, for instance, *K–Nearest–Neighbors Classifiers* and *Regression Trees*. It is called "supervised" since we are supervising the learning process by specifying to the computer (the machine in our case) which outcome we are interested in predicting.

- *Unsupervised Learning (UL):* In UL, the goal is to process the independent variables to come up with a summary of the data. If we aim at grouping variables we talk about *Dimensionality Reduction*, while if we aim at grouping people into groups based on their independent variables we talk about *Clustering*. Examples of Dimensionality Reduction techniques and Clustering are, for instance, *Principal Component Analysis* and *K–Means–Clustering*. In this case, it is called "unsupervised" since we are not telling the machine which specific outcome needs to be learned.

We said that the ultimate aim of (Supervised) Machine Learning algorithms is to predict in the most accurate possible manner over test set individuals (out–of–sample). One of the key challenges in this process is balancing the necessity for flexibility, hence the capacity of modelling complex data–generating–processes, and its opposite, namely the necessity for an algorithm to remain stable with respect to noisy observations or outliers.

In this work, when it comes to Supervised Learning, the focus is uniquely on Regression tasks. Here, the maximization of the predictive accuracy is conceptually equivalent to the minimization of a *loss function* measuring the distance between the predicted and true values of the dependent variable. As loss function, in this work we consider the *Residual Sum of Squares*. Once again, for a formal description of the *Bias–Variance Tradeoff* and its decomposition, a thorough description is available in "*The Elements of Statistical Learning*", Hastie *et al.* (2009), in particular Chapter 7, "*Model Assessment and Validation*". An excellent explanation can also be found in Mehta *et al.* (2019). What is relevant for this work is to notice that the Expected Generalization Error of an algorithm is characterized by *three* components, two of which balance each other:

- *The Bias:* Following Metha *et al.* (2019, p.13), the bias "measures the deviation of the expectation value of our estimator (*i.e.*, the asymptotic value of our estimator in the

infinite data limit) from the true value". In Hastie *et al.* (2009, p.223) it is instead defined as "the amount by which our estimates differ from the true mean".

Broadly speaking, a low–bias algorithm is one capable of producing a very flexible fitting curve, hence getting closer in predicting all the values in the training set. On the contrary, a high–bias algorithm is a very inflexible one, producing more similar estimates across all individuals.

- *The Variance:* in Mehta *et al.* (2019, p.13), the variance is defined as measuring "how much our estimator fluctuates due to finite–sample effects", and similarly in Hastie *et al.* (2009, p.223) as "the expected square deviation of a prediction in a given point from its mean".

  Hence, a low–variance algorithm will lead to a fitted curve that will not change strongly if an outlier or noisy observation appears in the training set, whereas a high–variance algorithm will behave in the opposite manner.
- *The Irreducible Error:* in Hastie *et al.* (2009, p.223) it is defined as "the variance of the target around its true mean", whereas in Mehta *et al.* (2019) is left undefined, and simply called *Noise*. This error component is called "Irreducible" since it is independent w.r.to the fitted algorithm, meaning that it remains the same whatever algorithm we are considering. This is because this component of error is usually caused by either *Measurement Errors* in the dependent variable or by the omission of relevant independent variables in the model (*Omitted Variables*).

It is intuitive to see how Bias and Variance are in competition: if a fitted model doesn't change its shape significantly in response to a new observation (low variance), it will inevitably be incapable of getting closer in predicting the outcome across the entire dataset (high bias), and vice versa. Two examples exemplifying the bias – variance tradeoff are *Linear Regressions* and *Neural Networks*. A Linear Regression is based on the parametric assumption of linearity in the parameters, hence producing as fitted curve a hyperplane. At each new individual observed within the learning process, at most what will change is the slope of the fitted hyperplane: it is therefore a low – variance, high – bias algorithm. On the contrary, Neural Networks are instead an example of *universal approximators* – as based on the Universal Approximation Theorem (Csáji 2001,

Hornik *et al.*, 1989) – meaning that there exists at least one kind of network capable of approximating any kind of function. This intuitively suggests low bias and high variance. If our goal was to fit uniquely the training set, a low bias – high variance algorithm would always be our best choice. However, since our goal is to predict on the test set (out–of–sample), this is not necessarily the case. When the training set is small, a low bias – high variance algorithm may end up modeling as real pattern fluctuations that are instead random, associated with the aforementioned issues of omitted variables and measurement errors. Such error would instead not be committed by a low variance – high bias algorithm which, in its incapability of modeling strong real fluctuations, would also (correctly) ignore random ones.

On the other hand, on larger training samples, the role of random fluctuations will eventually converge to averaging 0, meaning that low bias – high variance algorithm should perform better also when used to predict on the test set. As such, there is the question of what is a sufficiently large training sample in the context of predicting wellbeing and healthcare utilization.

In this work, therefore, beside the economic questions associated with subjective wellbeing and healthcare utilization, it is addressed also the more technical question of how bias and variance behave w.r.to the sample sizes in our contexts, and if the nature of the dependent variable can have an impact. More specifically:

- In Chapter 1, the goal is to predict self–assessed life satisfaction using a relatively small sample composed of 8,867 individuals from the British Cohort Study (BCS) in 2004: 80% of them were used to train the models. We consider two datasets, one including the 8 variables also considered in Layard *et al.* (2014) – with only exception of physical health (in this work considered objective) – and a one with 21 variables (96 including the transformation of categorical variables in dummies). The technical question associated with this chapter is therefore *whether ML algorithms can lead to increases in predictive accuracy also on a relatively small training set, or if instead in this case is better to stick to a low – variance model like a (Penalized) Linear Regression*. The key result of this chapter is that, in terms of predictive accuracy, ML algorithms and Linear Regression perform similarly, with the second therefore being preferable (Occam's Razor Principle).

However, a non–negligible improvement is observed, considering both high and low bias algorithms, with the inclusion of more variables.

- In Chapter 2, the goal is predicting wellbeing intended as both life satisfaction and positive and negative affects, but using larger datasets. The considered data are from the American Gallup Daily Poll, the UK household longitudinal study (UKHLS) and the German Socio–Economic Panel (SOEP). In the Gallup dataset, across the 2010–2018 years, we have samples ranging from 115,192 observations to 351,875; in the UKHLS samples ranging from 29,605 observations to 40,679; and in the SOEP from 26,089 observations to 32,333. Also in this case, we considered two variations of each dataset, one with less variables – covering the standard demographic, economic and health individual characteristics – and a richer one, including up to 450 variables in SOEP and in UKHLS, and 67 in Gallup. In this chapter, therefore, *building on the findings from the first chapter, the potential of low bias algorithms is further investigated on larger datasets.*
- In Chapter 3, the focus is on a more objective dependent variable, healthcare utilization.[1] In this case, only the SOEP dataset is used, in a specification including 19 independent variables and 208,903 individuals. Moreover, we also considered Unsupervised Learning clustering algorithms to automatically divide the individuals in groups in which the predictive task would become easier. Hence, the technical question in this chapter is about the Irreducible Error: *is predicting more objective outcomes easier than predicting subjective ones? Are objective variables less subject to measurement errors and less influenced by omitted predictors?* The question is addressed considering a large dataset and clustering the data.

Finally, as economists, simply predicting is not sufficient. While in the previous section of the General Introduction we have seen cases in which relevant policy decisions relied uniquely on accurate predictions, being able to address variable–specific effects remains crucial. And one of

[1] Arguments have been made that healthcare utilization – measured as Number of doctor visits in the last three months – may in turn also be a subjective variable. The argument would be that people may not remember the exact number, and just give an approximation. While this is true, defining objectivity and subjectivity on a scale, said variable can still be considered more objective than self-assessed life satisfaction. The remaining quota of lack of memory in reporting represents a perfect example of the aforementioned measurement error.

the main critiques moved to ML algorithms is that they are “black–boxes”, producing accurate predictions – also generalizable – but being silent on the specific role of each independent variable. In recent times, the development of *model–agnostic interpretative tools* is allowing researchers to untangle the relationship between the independent and dependent variables also in more complex models. Interpretative tools are employed in all chapters, including eventual ad hoc modifications.

Chapter 1

What Makes a Satisfying Life? Prediction and Interpretation with Machine Learning Algorithms

1.1 Introduction

One of the major domains of Social Science is the understanding of individual well–being, with the aim of predicting what makes a successful life. This success in well–being terms can be defined either objectively or subjectively: the former focuses on measures such as income or consumption, where those with more economic resources are considered to be better–off, while the latter relies on individuals' own evaluations of how well their life is going. We here consider this second type of measure, commonly called subjective well–being.

The prediction of subjective well–being starts with the analysis of its associations with a set of key observable characteristics, which can be at either the individual or a more–aggregated level (see Clark, 2018, for a survey). We will here focus on individual–level characteristics. One of the central individual variables is income, both in absolute terms and expressed relative to others (Clark and Oswald, 1996, and Luttmer, 2005, are two analyses including relative income), and another (conditional on income) is unemployment (Winkelmann and Winkelmann, 1998, and Clark and Oswald, 1994). With respect to other non–pecuniary characteristics, the married are more satisfied than the non–married (see Stutzer and Frey, 2006, for a discussion of selection into marital status), and the correlations with both physical and mental health are typically positive (Dolan *et al*., 2008), with Layard *et al*. (2014) and Clark *et al*. (2018) finding the correlation with emotional health to be larger. The association between subjective well–being and education is on the contrary more ambiguous (see Chapter 3 of Clark *et al*., 2018). Women are often found to be more satisfied with their lives (Helliwell *et al*., 2016) but at the same time report more stress (Kahneman and Deaton, 2010). While there is a vibrant literature on subjective well–being and

age, this will not be relevant in the analysis we carry out here, which is based on one wave of a birth–cohort dataset (in which all respondents are therefore the same age).

The vast majority of these findings regarding the individual correlates of well–being come from parametric models. These models are, however, more useful in terms of explaining rather than predicting the dependent variable, at a potential cost in terms of predictive accuracy. The related statistical and methodological arguments will be presented below. At the same time, the growing computing power of current machines (including computers) has recently made Machine Learning (henceforth ML) widely available. Broadly, ML looks for a pattern (in general, non–linear) that maps a set of explanatory variables to the dependent variable of interest in a training set of data, and then focuses on generalizations, *i.e.* on obtaining good predictions of the dependent variable on data from outside of this training set.

Our aim here is to see whether two key ML algorithms – Penalized Linear Models and Random Forests – can provide more–accurate predictions of subjective well–being than does the more–traditional linear model (which we will henceforth call non–penalized linear regression). The model we analyze is that in Layard *et al.* (2014), the aim of which (as indicated in their article title) is the prediction of life satisfaction; this thus provides a natural starting point for our analysis.

The greater predictive accuracy of ML models comes at the cost of being less–easily interpretable than non–penalized linear regressions. Following Kim *et al*. (2016), interpretability refers to "the degree to which a human can consistently predict the model's result". We will below apply model–agnostic methods to our results in order to render the predictions from Random Forests more interpretable.

The remainder of the chapter is organized as follows. Section 2 describes the British Cohort Study data that we use in our empirical applications. The results are then presented in Section 3, and interpreted in Section 4. Last, Section 5 concludes.

1.2 Data

We use the same dataset as in Layard *et al.* (2014), the British Cohort Study (BCS). This is a birth–cohort study, covering all individuals in the UK who were born in the second week of March

1970 (cls.ucl.ac.uk/cls–studies/1970–british–cohort–study/). Since the birth wave of the survey in 1970, there have been ten other waves ("sweeps") at ages 5, 10, 16, 26, 30, 34, 38, 42, 46 and 51. Layard *et al.* (2014) focus on the life satisfaction that respondents report at age 34.
Of the 17,000 initial births recorded, 8,867 individuals provided information at age 34 on all of the variables that we will use in the analysis, as listed below.

We initially consider only the eight adult age–34 explanatory variables that appear in Layard *et al.* (2014): these are our explanatory variables, which we use to predict *Life Satisfaction*, our dependent variable. The only variable that we treat differently from them is health. Our health measure comes from the BCS analysis in Clark and Lepinteur (2019), and is the number of conditions from which the individual suffers; that in Layard *et al.* (2014) is instead self–assessed health at age 26 measured on a scale of 1 to 4 (from "Bad" to "Excellent"). We prefer an objective health measure for common–method variance reasons (even if the subjective health measure in Layard *et al.*, 2014, is lagged by two waves).

Our eight initial explanatory variables are the following:

- *Ln***(***income***)** at age 34. Household equivalent disposable income using the OECD equivalence scale, expressed in Pounds.
- *Educational Achievement* at age 34. This is a single variable with six distinct cardinal values, obtained from a regression of male log full–time earnings on having a family, childhood emotion and conduct, and five education dummies. The resulting values are 0.750 (PhD or Master), 0.486 (Degree), 0.237 (A–level), 0.188 (GCSE), 0.043 (CSE), and 0 (No qualifications; this was the omitted category in the regression).
- *Employment* at age 34. A dummy variable for not being unemployed at the time of the interview.
- *Has a Partner* at age 34. This is a single variable with four distinct cardinal values, obtained from a regression of life satisfaction on three family dummies and a number of life–success variables. The resulting estimated coefficients on the family dummies are 0.685 (Married and cohabiting with children), 0.530 (Married/cohabiting without children), –0.004 (Single with children), and 0 (the omitted category: Single without children).

- *Good Conduct* between ages 16–34: One unit of "crime" here is being found guilty by a criminal court or formally cautioned at a police station. Good Conduct is the maximum observed number of crimes between ages 16 and 34 years in the BCS sample (25 crimes) minus the individual's own number of crimes.
- *Physical Health* at age 26. This is a cardinal variable for the number of health conditions from which the individual suffers, from a list of 15 (see Appendix B.[2] We multiply this figure by –1, so that higher values refer to better physical health.
- *Mental Health* at age 26. This is the sum of the respondent's replies at the age–26 BCS wave to 24 questions covering aspects such as worry and irritation, and physical symptoms like poor appetite and headache. The total number of conditions, multiplied by –1, is our index of mental health. 665 individuals had missing values for mental health at age 26; for these individuals we take their value at age 30 instead.
- *Gender*. 1 if female, 0 for male.

The dependent variable is *Life satisfaction* at age 34. This comes from the following question: "Here is a scale from 0–10. On it "0" means that you are completely dissatisfied and "10" means that you are completely satisfied. Please tick the box with the number above it which shows how dissatisfied or satisfied you are about the way your life has turned out so far."

Our expanded analysis of life satisfaction adds 16 additional explanatory variables reflecting life at age 34: *Number of people in the household, Number of natural children of the Cohort Member in the household, Number of non–natural children of the Cohort Member in the household, Number of rooms in the household, Type of accommodation, BMI, Alcohol units per week, Cohort Member's main activity, Highest academic qualification, Disability status, Whether the mother is alive, Whether the father is alive, Marital status, Weekly smoking habits, Tenure status*, and *Whether health limits everyday activities*. These new explanatory variables are likely highly correlated with some of the eight original explanatory variables: we will discuss this issue below when presenting the results. The descriptive statistics of all our variables appear in

[2] We retain the two-wave lag (*i.e.* using age-26 values) in order to be consistent with Layard *et al*. (2014). Information on some, but not all, of the conditions used to construct the Physical Health index are also available at age 34.

Appendix Table A, which also contains the coding details for all the variables, including Type of accommodation, Alcohol units per week, and Cohort Member's main activity.

The treatment of missing values depends on the nature of the variable. Missing values for categorical variables are not imputed. The rationale here is that the missing values are not at random, and potentially contain additional information about the individual. We instead consider the missing categories (there may be more than one for a given variable) as separate values to be used in the empirical analysis. Of the 16 new explanatory variables proposed above, the only categorical variable with significant missing information is Alcohol units per week (which is measured in categories), with 1,683 missing values. These correspond to individuals who reported never drinking or only on special occasions (these individuals are assigned a missing value code of –1 in the BCS questionnaire). The next most–frequent occurrences of missing values are for BMI and Whether the father is alive, with much smaller numbers of 246 and 121.

In the Linear Regression models, we create dummies for each value of the following categorical variables: Type of accommodation, BMI, Alcohol units per week, Cohort Member's main activity, Highest academic qualification, Disability status, Whether the mother is alive and Whether the father is alive (both of these are categorical, as they distinguish between the living parent being in or outside of the household), Marital status, Weekly smoking habits, Tenure status, and Whether health limits everyday activities.

The numerical variables Number of people in the household, Number of natural children of the Cohort Member in the household, Number of non–natural children of the Cohort Member in the household, and Number of rooms in the household have, respectively, 25, 25, 25, and 53 missing values, also labelled via negative numbers. We impute the negative missing values for these variables by the mean of the observed value. Nonetheless, there are only few observations that have missing values for these numerical variables in the dataset (between 0.3% and 0.6% of the observations), and our findings are unaffected if we instead simply drop the observations with missing values for these numerical variables. In the Random Forest analysis, these missing negative values were left as they appear in the data, as here the different explanatory variables' values only serve to define the sample splits, with the actual numerical values not affecting the calculation of the value of the dependent variable.

1.3 Machine Learning Algorithms: Presentation and Results

The choice of the ML technique to be used depends on the *interpretability−predictive accuracy* trade–off (see James *et al.*, 2013, for a discussion). In general, the most internally–interpretable algorithms are the least flexible: these less–flexible algorithms provide straightforward intuitions about the relationship between each of the explanatory variables and the dependent variable. If we wish the model to be *interpretable*, we may then prioritize less–flexible models. If, on the contrary, we are more concerned about accurate *predictions*, we may sacrifice interpretation in favor of more–flexible complex models. Accurate predictions may be at a premium, for example, in contexts in which we already have strong theoretical arguments regarding the explanatory variables–dependent variable relationship, and want to establish the best–possible predictive map. Linear Regression and Deep Feedforward Neural Networks can be considered as two polar examples in this trade–off continuum.

Nonetheless, the interpretability–predictive accuracy trade–off is not a strict dichotomy. As we will see below, model–agnostic interpretative tools also allow for inference in more–flexible methods. Equally, inflexible methods can produce similar (or even better) performance than more–flexible ones (for example, if the joint distribution of the explanatory variables and the dependent variable is relatively simple to model).

We will start our analysis of subjective well–being in the BCS data in the following sub–section by considering linear models. In order to estimate all the algorithms, we used *scikit–learn*, scientific library in Python (Pedregosa *et. al.*, 2011), and *glmnet* in R (Friedman *et al.*, 2010).

1.3.1 Non–Penalized and Penalized Linear Regressions

The standard linear non–penalized regression is our benchmark. This is a special case of an *Elastic Net Regression*, the general form of which is (see Zou and Hastie, 2005):

$$\min_{\boldsymbol{\beta}\in R^k} \sum_{i=1}^{n}(y_i - \boldsymbol{x_i'\beta})^2 \quad + \quad \lambda\left[\frac{1-\alpha}{2}\sum_{j=1}^{k}\beta_j^2 + \alpha\sum_{j=1}^{k}|\beta_j|\right] \qquad (1)$$

where λ and α are *hyperparameters*, *i.e.* parameters that are used to regulate the learning process, whose value has to be determined before the estimation of the β's. Penalizing by the sum of

squares of the betas produces coefficient shrinkage, balancing the bias and variance of the estimates. It does not however yield a parsimonious model as all variables are retained – none of the coefficients are shrunk to 0. Automatic variable selection instead comes from penalizing the sum of the absolute values of the betas (Zou and Hastie, 2005). The values of λ and α can either be input manually (*ex ante*) or discovered via cross–validation (*tuning*: see Section 3.1.2 below). We first consider five different values of α, (0, 1, 0.25, 0.50 and 0.75), and in each case use 5–fold cross–validation on the training set (which will cover 80% of the individuals) to find the optimal value of λ.

The linear non–penalized regression empirical loss function (*i.e.* that of OLS) is given by Equation (1) with $\lambda = 0$. When $\lambda \neq 0$, a value of $\alpha = 0$ corresponds to the *Ridge Regression* minimization problem, and $\lambda \neq 0$ and $\alpha = 1$ to the *LASSO Regression* minimization problem (where *LASSO* stands for *Least Absolute Shrinkage and Selection Operator*).

In linear regression, the goal is to estimate the unknown mapping under the assumption that the dependent variable is linear in the parameters, by minimizing the squared distance between the predicted and observed values.

We analyze these four cases ($\lambda = 0$, and $\lambda \neq 0$ with α either 0, 1, or in the interval) in turn, discussing the rationale for each case and the ensuing results.

1.3.2 Linear Regression – Non–Penalized

The standard linear regression model corresponds to $\lambda = 0$. Defining $X \in \mathbb{R}^{n \times k}$ as the matrix whose element $x_{i,j}$ is the value of the j^{th} explanatory variable for the i^{th} individual, the (non–penalized) linear regression minimization problem is usually presented as:

$$min_{\boldsymbol{\beta} \in \mathbb{R}^k} \sum_{i=1}^{n} (y_i - \boldsymbol{x_i'\beta})^2 \tag{2}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of values of the continuous dependent variable for each of the n individuals in the sample. The underlying assumed mapping is linear in the parameters:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(\mathbf{0},\ \sigma^2 I_n)\ . \tag{3}$$

Additional standard requirements are the conditional mean independence of the error term with respect to the explanatory variables (formally, $E(\boldsymbol{\varepsilon}|X) = \mathbf{0}$), no perfect multicollinearity, so that no one column in X can be expressed as a linear combination of the others (or more simply

that *rank*(X) = $k < n$) and that the error term $\boldsymbol{\varepsilon}$ is distributed as in (3). The latter can be relaxed by allowing for *heteroscedasticity* (where the variance of the error term's distribution is individual–dependent), which often appears as a robustness check. Under these conditions, it is well–known (the Gauss−Markov Theorem) that the Least Squares estimator solving (2)

$$\widehat{\boldsymbol{\beta}}_{OLS} = (X'X)^{-1}X'\boldsymbol{y} \tag{4}$$

is the Best Linear Unbiased Estimator (*BLUE*), in that it has the lowest variance of all the unbiased linear estimators.

Given its additive structure, the Linear Regression is arguably the most–interpretable model, as $\hat{\beta}_{OLS,j}$ is the predicted change in y_i following a unit change in $x_{i,j}$, for all individuals i and keeping all other explanatory variables $\boldsymbol{x}_{i,-j}$ constant. If the variables are standardized, a similar interpretation holds in terms of the correlation between standard deviations, and the square of each estimated coefficient $\hat{\beta}_{OLS,Stand,j}$ shows how much the explanatory variable $\boldsymbol{x}_j$ contributes to the dependent variable's variance, ignoring its covariance with the other explanatory variables (Layard *et al.*, 2014). Nonetheless, Linear Regression is inflexible due to the stringent parametric linearity assumption and the other requirements noted above.

We will compare the performance of our models using the Test Mean Squared Error (MSE), considering a random split where 80% of the individuals appear in the training set (S) and the remaining 20% are in the test set (T). In general, S and T have no individuals in common and come from the same data–generating process. We train our algorithms on the set S to learn the mapping $\hat{f}: \mathbb{R}^{\mathrm{k}} \rightarrow \mathbb{R}$. We then assess the empirical quality of this mapping via the following statistic:

$$MSE_{test} = \frac{1}{n(T)} \sum_{i:(\boldsymbol{x_i}, y_i) \in T} \left(\hat{f}(\boldsymbol{x_i}) - y_i\right)^2 \tag{5}$$

where *n(T)* represents the cardinality of the test set T. For instance, in the case of linear regression:

$$\hat{f}(\boldsymbol{x_i}) = \boldsymbol{x_i}'\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS,train}} \tag{6}$$

for all the individuals i in T, with $\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS,train}}$ having been learned from the training set. We add the subscript 'train' to the estimated coefficients to stress that these come from the training set, but are evaluated in terms of their ability to map the explanatory variables onto the dependent variable using the data from the test set.

We now present the Test MSEs for predicting life satisfaction, as well as the Training MSEs, defined as in (5) but over the elements in the Training Set *S*. All non–dummy explanatory variables are standardized (standardization is a normalization and does not affect the quality of the fit). *Original* refers to the model including only the eight adult explanatory variables from Layard *et al.* (2014), and *Extended* to the 21 – explanatory variable model (which become 96 once the dummies are created from the categorical variables) corresponding to five of the eight original explanatory variables in Layard *et al.* (2014) and the 16 new explanatory variables. Three of the eight original explanatory variables are dropped (or rather expanded) in the Extended model. The Original explanatory variable "Has a partner" is now redundant, as the newly–added variables include both respondent marital status and the number of natural and non–natural children. Equally, educational achievement is replaced by the highest academic qualification, and the original employed dummy is now one of the categories of the newly–added respondent main–activity variable. In order to avoid potential multicollinearity issues, we omit the most–populous category for each categorical explanatory variable, and drop entirely all categories covering fewer than 15 individuals: these dropped categories are listed in Appendix D.[3] As a result, the number of explanatory variables falls from 96 to 72.

All models were fitted 100 times with 100 different randomly–drawn train–test splits (in all of which 80% of observations were assigned to the training set). Table 1 lists the average Mean Squared Errors from these 100 different splits, with their associated standard deviations in parentheses.

Table 1. The Performance of the Linear Regression

	Training MSE	**Test MSE**
Original	2.78 (0.03)	2.79 (0.11)
Extended	2.57 (0.02)	2.65 (0.09)

Notes. These figures show the average performance of linear regressions in predicting life satisfaction in 100 different train–test splits, with 80% of the sample in the training set and the error calculated on the remaining 20% test–set individuals. Standard deviations are in parentheses.

[3] Without this exclusion, there are 18 perfectly multicollinear cases (out of the 100). This occurs with sparse categorical dummy cells, when all of the 1's are randomly-allocated to the test set (producing a column of 0's in the training set).

Adding the 16 new explanatory variables – for a total of 72 plus the constant – in the Extended model improves the Test Set performance, with a reduction in the MSE of 5.3% (from a figure of 2.79 to 2.65). Moreover, while in the Original dataset the training and testing accuracy figures are almost identical, in the Extended case the Training MSE is 3% lower than the Test MSE (2.65 *vs.* 2.57).

The procedure to avoid multicollinearity does nonetheless involve a potentially substantial loss of information. Considering, for instance, Marital Status, we of course have to drop one category in order to estimate the coefficients on the other categories: here we drop the most–populous category ("Married", with 4,817 observations); we in addition drop "Widowed" (12 observations) and "Other missing" (3 observations), for which we therefore do not estimate a coefficient. However, these small groups may still be of policy interest – especially the Widowed, whose life satisfaction (as we will see with Random Forests) is particularly low. As such, Machine Learning algorithms that are capable of dealing with multicollinear datasets can be of use, as they allow us to model the relationship with the dependent variable for individuals in these more sparsely–populated categories. To this end, in what follows we consider Penalized Linear Regressions that allow for the inclusion of all of the response categories, for a total of 96 explanatory variables.

1.3.3 Multicollinearity and Ridge Regression

The Ridge Regression estimator (Hoerl and Kennard, 1970) corresponds to the minimization of (1) with $\alpha = 0$:

$$\min_{\boldsymbol{\beta} \in R^k} \sum_{i=1}^{n} (y_i - \boldsymbol{x_i'\beta})^2 + \frac{\lambda}{2} \sum_{j=1}^{k} \beta_j^2 \tag{7}$$

where λ is a tuning parameter. It can be shown that the Ridge Regression estimator from (7) is:

$$\widehat{\boldsymbol{\beta}}_{Ridge} = (X'X + \lambda I_k)^{-1} X'\boldsymbol{y}. \tag{8}$$

The Ridge estimator can be calculated even under perfect multicollinearity, as $\lambda > 0$. In the case of harmful, but not perfect, multicollinearity, it can be seen that the presence of λ reduces the absolute values of the estimates. The larger is the chosen λ (via hyperparameter tuning or ex–

ante choice), the greater is the coefficient shrinkage – although the coefficients never become zero.

The variance of the Ridge estimator is:

$$Var(\widehat{\boldsymbol{\beta}}_{Ridge} \mid X) = \sigma^2(X'X + \lambda I_k)^{-1}X'X(X'X + \lambda I_k)^{-1} < \sigma^2(X'X)^{-1} = Var(\widehat{\boldsymbol{\beta}}_{OLS} \mid X). \quad (9)$$

This variance is smaller than that from OLS for every $\lambda > 0$. However, $E(\widehat{\boldsymbol{\beta}}_{Ridge} \mid X) \neq \beta$ due to shrinkage, so that the coefficients are *biased* under the linearity assumption, whereas $E(\widehat{\boldsymbol{\beta}}_{OLS} \mid X) = \beta$. The broad idea behind the use of the Ridge estimator is that by introducing some bias into the estimates, we can reduce the variance up to a point at which the associated MSE is lower than that from OLS.

The Ridge Regression results appear in Table 2. The optimal λ^* here is chosen from a grid of 100 values via *5–fold cross−validation*[4] on the training set solving (7). The λ^* producing the smallest average cross–validated MSE is then introduced into (7), producing the Ridge estimator in (8). Last, the fitted model is used to assess the quality of the fit on the data in the test set, measured via the Test Set MSE as in (5). The procedure is again applied with 100 different random train–test splits, and the results refer to the average performance and associated standard deviations. We also list the mean and standard deviation of λ^*. Note that standardization is required here for all explanatory variables, including the dummies, given the presence of the penalization term.

Table 2. The Performance of the Ridge Regression

	Training MSE	**Test MSE**	**λ∗**
Original	2.78 (0.03)	2.79 (0.11)	0.06 (0.001)
Extended	2.57 (0.02)	2.65 (0.10)	0.35 (0.08)

Notes: This table lists the mean performance and optimal $\lambda*$ of the Ridge regression predicting life satisfaction over 100 different train–test splits, each with 80% of the sample in the training set and the error calculated on the remaining 20% of individuals in the test set. The $\lambda*$ obtained for each split comes from a 5–fold cross–validation on the training set. Standard deviations appear in parentheses.

[4] The training set is split into *k* equally-sized blocks for *k*-fold cross-validation. One of these *k* blocks is used for validation, and the model is fitted on the remaining *k*-1 blocks. This process is repeated *k* times until each of the *k* blocks has been used for validation. The cross-validated score for a given hyperparameter value is the average validation score (the MSE in our case) over the *k* folds (we here use 5 folds).

Prediction in the test set using the Ridge estimator on the Extended dataset now always produces a reasonable Test Error, even without dropping any dummy variable.

In the Original dataset, the Ridge estimator's lower variance does not suffice to offset the loss in accuracy: the Test MSE of the Ridge estimator is 2.79. This reflects the absence of multicollinearity in the Original model. On, the contrary, the estimated average value of λ^* in the Extended model is almost six times that in the Original model: this reflects the multicollinearity discussed above. The standard deviation of λ^* is small as compared to its mean, so that the optimal values found across the 100 train–test splits were very similar to each other.

In terms of performance, the Ridge estimator produces a Test MSE that is 5.3% lower in the Extended (2.65) than that in the Original model (2.79). The new explanatory variables provide more–detailed information on the socioeconomic determinants of individual well–being, including marital status and wealth (approximated by housing–tenure status and the number of rooms in the household). In the Original model, these latter were limited to the explanatory variables of Has a Partner and Log Income. In order to estimate a coefficient for each of the categories of each categorical explanatory variable, we have however had to introduce bias into the estimates, in that the Ridge coefficients are biased estimates of the true β. A better way of describing the determinants of subjective well–being more thoroughly appears in the discussion of the Shapley Values in the Random Forest in Section 4 below.

1.3.4 Variable Selection and LASSO Regression

An alternative to the Ridge is the *LASSO* regression (Tibshirani, 1996). The empirical loss function here comes from setting $\alpha = 1$ in (1):

$$\min_{\boldsymbol{\beta}\in R^k} \sum_{i=1}^{n}(y_i - \boldsymbol{x_i'\beta})^2 + \lambda \sum_{j=1}^{k} |\beta_j|. \qquad (10)$$

The LASSO minimization problem in (10) may have multiple solutions, although they always produce the same predicted values, so that the Test MSE remains a valid measure of the quality of fit (Tibshirani, 2013). Outside of some particular cases, no closed–form expression for the LASSO estimator exists. There are a number of numerical methods solving (10), including *Coordinate Descent*, the method used in the *glmnet* package of R. Additional details on

Coordinate Descent and other solution techniques can be found in Friedman *et al.* (2010) and Van Wieringen (2020).

The key characteristic of the LASSO penalization is that it induces *variable selection*: even with no particularly large values of λ, one or more of the $\hat{\beta}_{Lasso,j}$ may be shrunk to 0; this is only the case for the Ridge estimator when the estimated coefficients were already zero in the OLS estimation without penalization. The difference between the two approaches reflects the shapes of the constraints imposed on the estimates by the two penalizations. A more detailed explanation can be found in Hastie *et al.* (2009, Ch.3).

The optimal λ^* values were obtained using the same procedure as described above for the Ridge estimator. The results appear in Table 3, which also lists the number of *non–zero* coefficients associated with the optimal cross–validated λ^*. The figures refer to standardized values and show the means and standard deviations over 100 different random train–test splits.

Table 3. The Performance of the LASSO Regression

	Training MSE	**Test MSE**	**$\lambda*$**	**Non–zero coefficients**
Original	2.78 (0.03)	2.79 (0.11)	0.002 (0.002)	9 [out of 9] (0.20)
Extended	2.58 (0.02)	2.64 (0.09)	0.02 (0.004)	51 [out of 97] (5.60)

Notes: These figures show the average performance, optimal λ^* and number of non–zero coefficient figures in a LASSO regression predicting life satisfaction over 100 different train–test splits, each with 80% of the sample in the training set and errors calculated over the remaining 20% of individuals in the test set. λ^* is obtained from 5–fold–cross–validations on the training set. Standard deviations appear in parentheses.

The predictive performance of the LASSO regression is comparable to that of the Ridge regression, and the same conclusions regarding bias and variance, overfitting and underfitting as in the OLS and Ridge case apply.

In the Original model, shrinkage to 0 was confined to one explanatory variable out of 9 (8 plus the constant) in four cases out of the 100 train–test splits. On the contrary, in the Extended model an average of 46 coefficients (out of 97) were shrunk to 0.

As for the Ridge estimator, the LASSO estimator solves the numerical multicollinearity issues found for the OLS estimator in the Extended model when we did not drop the dummy associated with the most populous category and all categories with fewer than 15 individuals. The 16 new

explanatory variables (with their 97 associated categories) yield a greater predictive accuracy of 5.7% in testing.

While the performances of the Ridge and LASSO estimators are then comparable, the latter has the advantage of automated explanatory–variable selection via the shrinkage to zero. This may help reduce model complexity, further reducing its variance and making it easier to interpret. We nonetheless may still wish to obtain estimates for all of the coefficients, after explanatory–variable selection has been carried out *ex ante*. In general, Tibshirani (1996) concludes that with $n > k$ (*i.e.*, more observations than independent variables) the Ridge estimator outperforms the LASSO estimator. Furthermore, if two explanatory variables are collinear, the LASSO estimator does not shrink both of the associated $\hat{\beta}_{Lasso,j}$ coefficients, but rather only one of them. As such, LASSO does not have the desirable *Grouping Effect*, where two highly–correlated explanatory variables should attract similar estimated coefficients (and identical coefficients in absolute value if the two are perfectly correlated: see Zou and Hastie, 2005).

The *Elastic Net*, first developed by Zou and Hastie (2005), is considered to overcome the weaknesses of the LASSO estimator, but retains its attractive explanatory variable–selection property.

1.3.5 Between Ridge and LASSO: The Elastic Net

The general Elastic Net minimization problem in Zou and Hastie (2005) was set out in Equation (1) above, of which OLS, Ridge and LASSO are special cases. In general, the estimator that solves this problem is

$$min_{\boldsymbol{\beta}\in R^k} \sum_{i=1}^{n}(y_i - \boldsymbol{x_i'\beta})^2 + \frac{\lambda_2}{2}\sum_{j=1}^{k}\beta_j^2 + \lambda_1 \sum_{j=1}^{k}|\beta_j| \qquad (11)$$

where $\alpha \in (0,1)$ in Equation (1) is the ratio of λ_1 over $\lambda_1 + \lambda_2$, and thus shows the relative weights given to the two types of penalization.

Were we to optimize over pairs of (λ_1, λ_2), we may find the same cross–validated log–likelihood for two different pairs and thus not be able to distinguish between them: the same log–likelihood can come from a very sparse model in which more coefficients are shrunk to 0 ($\lambda_1 >> \lambda_2$) or one that is not sparse ($\lambda_2 >> \lambda_1$). We thus instead optimize over α, rephrasing the Elastic Net minimization problem in (11) as that in (1). The introduction of α allows us to tune the model

over the pairs (λ, α). We here consider three possible values for α, 0.25, 0.50 and 0.75, hence either giving 3/4 of the weight to one of the two forms of penalization or weighting them equally. The results are listed in Table 4.

Table 4. The Performance of the Elastic Net Regression

	$\alpha = 0.25$				$\alpha = 0.50$				$\alpha = 0.75$			
	Train MSE	**Test MSE**	$\lambda*$	$\widehat{\beta_J} \neq 0$	**Train MSE**	**Test MSE**	$\lambda*$	$\widehat{\beta_J} \neq 0$	**Train MSE**	**Test MSE**	$\lambda*$	$\widehat{\beta_J} \neq 0$
Original	2.78 (0.03)	2.79 (0.11)	0.007 (0.003)	9 (0.14)	2.78 (0.03)	2.79 (0.11)	0.004 (0.003)	9 (0.20)	2.78 (0.03)	2.79 (0.11)	0.003 (0.002)	9 (0.17)
Extended	2.58 (0.02)	2.64 (0.09)	0.07 (0.01)	54 (5.35)	2.58 (0.02)	2.64 (0.09)	0.04 (0.01)	52 (5.37)	2.58 (0.02)	2.64 (0.09)	0.03 (0.01)	51 (6.01)

Notes: These figures show the average performance, optimal λ^* and number of non–zero coefficient figures in three elastic–net regressions predicting life satisfaction. 100 different train–test splits are carried out, each with 80% of the sample in the training set and the error calculated on the remaining 20% of individuals in the test set. The values of α are *ex–ante* fixed and reflect the relative weights on the two penalization terms. λ^* was obtained via 5–fold cross–validations on the training set. Standard deviations appear in parentheses.

As can be seen in Table 4, the three variants of the Elastic Net we consider do not yield much improvement in terms of predictive performance over the Ridge or LASSO regressions. From the $\widehat{\beta_J} \neq 0$ columns, there is shrinkage for over 40 explanatory variables in all three Elastic–Net estimations.

Our main conclusion from considering penalized and non–penalized linear regressions is then that there is no reason to believe that the linear non–penalized regression overfits the Original data and, given the reliability of the estimates in the training data with no evidence of harmful multicollinearity, it is probably preferable to avoid introducing bias. Conversely, in the Extended dataset, the 16 additional explanatory variables improve the Test Set performance with a reduction in the MSE of 5.3%. Moreover, while in the Original dataset the training and testing accuracy were almost identical, in the Extended model we observe a Training MSE that is 2.3% lower than the Test MSE.

We next introduced penalization, and retained all of the dummies in the analysis. We do not observe any additional improvement here: the Test MSE for the Ridge estimator is 5.3% lower in the Extended (2.65) than in the Original model (2.79), as was the case for the non–penalized regression. In general, fitting a multicollinear linear regression can be of interest in any case, as

we may wish to assess the marginal effects of some explanatory variables while adding other (possibly correlated) controls. Moreover, the addition of (relevant) multicollinear explanatory variables can in theory still lead to improved test accuracy, and hence a fuller model to interpret (although this is not the case in the data that we analyze here).

In what follows, we move beyond linear estimation to the next algorithm in the interpretability–complexity trade–off: *Regression Trees* and their *ensemble*, the *Random Forest*. For the latter, we will explore two *Model–Agnostic Interpretable Algorithms – Permutation Importance* and *Shapley Values* – that will help us to interpret the results.

1.3.6 Regression Trees and Random Forest: Stratifying the Explanatory Variable Space

Classification and Regression Trees have a considerable history. The Regression Trees we now turn to were presented in Breiman (1984). The overall idea is to divide the explanatory variable space into *J* distinct and disjoint sets, the *terminal nodes* or *leaves* of the tree. The dependent variable value for each individual in a leaf is the mean of the dependent variable of all the individuals who are in the same leaf. Individuals fall into a leaf by moving along one of the branches of the tree, depending on values of their explanatory variables.

The subsequent splits along the *branches* of the tree define the *internal nodes* obtained by *recursive binary splitting*. Starting from the top of the tree – at which point every individual belongs to the same set (so that this is a *top−down* approach) – a *greedy* procedure is implemented, where the preferred split is that which is the best at that specific point, independent of any subsequent steps.

These procedures tend to overfit the training set, producing deep trees with too–long branches, and so produce estimators with high variance and low bias. There will be only few training individuals in each of the final leaves, and a poorly–defined outcome variable, $\hat{y}_{t_{k,train}}$. For this reason, *Random Forests (ensembles* of trees) are preferred, along with regularization criteria for each tree.

Random Forests are constructed via *bootstrap aggregation*, which can be either *non–parametric* or *parametric*. In the former case, no assumptions are made regarding the data–

generating process, and new observations are constructed by sampling with reintroduction from the training set. On the contrary, in the latter we assume a well–defined parametric model for the data–generating process.

As we are looking for evidence *against* the linearity (parametric) assumption, we consider nonparametric bootstrapping, which is the general practice in the applied Random Forest literature. *Bootstrap Aggregation* or *bagging* consists in averaging the prediction of B fitted models, each labelled b, over the S^b different bootstrapped samples, with the aim of reducing the variance of the final estimator.

The entire Random Forest, and each Regression Tree in it, has the same expected value, and hence the same bias. As Tibshirani (2013, p.596) notes, "Increasing the number of trees does not cause the Random Forest sequence to overfit".

A key element in the lower variance is the number m of explanatory variables used for the split at each internal node of each tree, $m \leq k$, where k is the total number of explanatory variables. The correlation between two generic trees in the forest rises with m, although the bias falls with m.

One of the most interesting features of Random Forests is the possibility of *leaving categorical and ordinal explanatory variables as they are, without creating dummies.* We now present the Random Forest results for both the Original and Extended models. The average predictive performance continues to be calculated over 100 Random Forests with 100 different random train–test splits. In each of these, 400 trees were constructed with non–parametrically bootstrapped data. The procedure differs from that in the Penalized Linear Regressions, where we looked for the optimal λ^* *in each train–test split*. Conversely, the optimal structure of the trees in the forest was established, via 5–fold cross–validation, on a single train–test split (the first) using 4000 trees. The penalizations used were the number of explanatory variables at each split, the maximum depth of the branches, and the minimum number of training individuals per leaf. The Shapley Values, describing the marginal effects of the different explanatory variables at an individual level, are instead calculated considering only the Random Forest in train–test split 1. The results are presented in Table 5.

1.3.7 Random Forest: Results

Cross–validation was used as the optimizing strategy, so as to be consistent with the linear regressions. Table 5 (fourth column) shows that the algorithm always prefers a random subset of the explanatory variables over including them all – in order to avoid overfitting – and over considering one variable only – which would have been too restrictive. More precisely, the algorithm considers a subset composed of only the (rounded) square root of the number of all of the variables, which is a rule–of–thumb value to trade–off between overfitting and underfitting. Regarding the maximum depth – intended as number of internal splits – of each branch of each tree, longer trees are unsurprisingly required in the Extended dataset of 21 explanatory variables, given the potential for more–complex relationships.

Table 5. The Performance and the Optimal Hyperparameters of the Random Forest

	Average Training MSE	Average Test MSE	Number of trees	Number of considered explanatory variables per split	Maximum depth of branches	Minimum individuals per leaf
Original	2.67 (0.03)	2.79 (0.10)	400	round($\sqrt{8}$) = 3	8	15
Extended	2.19 (0.02)	2.66 (0.10)	400	round($\sqrt{21}$) = 5	13	8

Notes: These figures show the average performance of 100 Random Forests over 100 different train–test splits in predicting life satisfaction. The optimal number of explanatory variables to be considered at each split of each tree, the maximum depth of each branch of each tree, and the minimum number of training individuals to be left in each leaf of each tree were *ex–ante* obtained via 5–fold–cross–validation on the first train–test split 1.

Table 6. The Performance of the Random Forest Compared to Linear Regression

	Lin. Reg. MSE Train	Lin. Reg. MSE Test	R.F. MSE Train	R.F. MSE Test	R.F. Improvement in Training Set	R.F. Improvement in Test Set
Original	2.78 (0.03)	2.79 (0.11)	2.67 (0.03)	2.79 (0.10)	4.12%	0%
Extended	2.57 (0.02)	2.65 (0.09)	2.19 (0.02)	2.66 (0.10)	17.35%	–0.38%

Table 6 compares the performance of Linear Regressions and Random Forests, in both training and testing. We first note a considerable improvement in training set accuracy over the linear regressions of 4.1% and 17.4% in the Original and Extended specifications respectively,

while accuracy does not change much in testing. Comparing across both algorithms and specifications, the Extended–model Test MSE in the Random Forest (2.66) is a 4.9% improvement over the Original–model MSE in Unpenalized Linear Regression (2.79).

We now present *Permutation Importance* and the *Shapley Values* calculated for the Random Forest, and a comparison of the latter to the Linear Regression results. As well as discussing the Random Forest's predictive accuracy, these will allow us to understand how the different explanatory variables affect life satisfaction.

1.4 Interpreting the Findings: Opening the Black Box

The interpretation of the ML results requires additional calculations beyond fitting, as opposed, for instance, to the interpretation of the explanatory variable coefficients in linear regressions. Model–agnostic tools are used to this end.

The choice of the best model–agnostic interpretability approach depends on a number of factors, including the complexity cost of the algorithm, and whether we are interested in *sparse* or *full* interpretations, or extracting new, derived predictive algorithms from the fitted model (see Molnar, 2019, for details). We will here consider *Permutation Importance* and *Shapley Values*, applied to the results from the Random Forest. We first focus on the *Shapley Values*, as they are interpretable in terms of both their importance – defined via their absolute mean for each explanatory variable – and their marginal effects, and provide a clearer image of the fitted model. *Permutation Importance* instead tells us which explanatory variables, once randomized, most increase the MSE. Last, *Learning Curves* allow us to understand the overall complexity of the underlying data–generating process.

1.4.1 Shapley Values and TreeSHAP

The Shapley Value is a solution concept from co–operative game theory introduced by Shapley (1951) and formalized in Shapley (1953). The underlying idea is that the way in which a certain sum obtained by a group of players is split depends on how much each member contributes to the outcome.

Applied to Machine Learning, the *game* is the predictive task and the *players* are the different explanatory variables that work together to produce the *gain*, namely the difference between the prediction for a given individual and the "average prediction in the sample" (Molnar, 2019, Chapter 5.9). The *Shapley Value* of an explanatory variable is "the average of all marginal contributions across all possible coalitions of explanatory variables" (Molnar, 2019, Chapter 5.9). Shapley Values are calculated at the individual level. If we have k explanatory variables and we are interested in calculating the Shapley Value for one of them, say variable j, we will consider all the possible 2^{k-1} coalitions of the remaining $k-1$ explanatory variables.

In each of these 2^{k-1} coalitions, we calculate the difference between the predicted value *with* and *without* the value of the j^{th} explanatory variable for individual i, $x_{i,j}$. This reveals the *marginal contribution* of the explanatory variable j in predicting the dependent variable. The values of the explanatory variables that do not appear in a coalition are *eliminated*, by randomly replacing individual *i's* value of that explanatory variable with that of another individual. The Shapley Value for explanatory variable j for individual i is then the weighted average of its marginal contributions across all of the 2^{k-1} coalitions, with the weights depending (in a U–shaped way) on the number of explanatory variables included in the coalitions.

Formally, define $\boldsymbol{x}_i$ as the vector of explanatory variables for individual i, and $\{x_{i,1}, \dots, x_{i,k}\}$ as the set of all of the values of the k explanatory variables considered for i. Let S be the coalitions of players considered in a given step – that is, the coalition of explanatory variables used in the model – and $f: 2^{k-1} \rightarrow \mathbb{R}$ a *value function*. The Shapley Value of the explanatory variable j for individual i is formally defined as:

$$\phi\ \left(x_{i,j}\right) = \sum_{S \subseteq \{x_{i,1},\dots,x_{i,k}\} \setminus \{x_{i,j}\}} \frac{n(S)!(k-n(S)-1)!}{k!} \left[f_{\boldsymbol{x}_i}\left(S \cup \{x_{i,j}\}\right) - f_{\boldsymbol{x}_i}(S)\right]. \qquad (12)$$

The value taken by explanatory variable j for individual i then contributes $\phi\ \left(x_{i,j}\right)$ "to the prediction of this particular instance compared to the average prediction for the dataset" (Molnar, 2019, Chapter 5.9).

It is immediate to see that the calculation of Shapley Values is costly, as we calculate values for 2^{k-1} coalitions *for every individual* in the sample and *for every explanatory variable*. A number of ways of addressing this issue have been proposed, including Monte Carlo sampling by Štrumbelj *et al.* (2014).

We here consider the *TreeSHAP* algorithm of Lundberg *et al.* (2018), where the value function is the expected value of the prediction conditional on the explanatory variables in the coalition *S*: $f_{x_i}(S) = E\,[f(x_i) \mid S]$. The direct estimation of $f_{x_i}(S)$ would have computational complexity of $O(BL2^k)$, where *B* is the number of trees in the forest, *L* the maximum number of final leaves in any tree, and *k* the number of explanatory variables. The *TreeSHAP* algorithm greatly reduces the computational complexity to $O(BLD^2)$, where *D* is the maximum depth of any tree.

The key measure that can be derived from Shapley Values is the *Shapley Feature Importance*, that is, the mean absolute value of the Shapley Values for variable *j* calculated over all of the *i* individuals in the training set:

$$I_{Shap}(X_j) = \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} |\phi(x_{i,j})|. \tag{13}$$

We calculate $I_{Shap}(X_j)$ for each explanatory variable in each of the 100 train–test splits, and average these to produce *Average Mean Absolute Shapley Values* with their associated standard deviations. Formally, labelling the different train–test splits as *train(1), train(2),…,train(100)*, this average value is given by:

$$Avg\left[I_{Shap}(X_j)\right] = \frac{1}{100} \sum_{t=1}^{100} \frac{1}{n(train(t))} \sum_{i=1}^{n(train(t))} |\phi_t(x_{i,j})|, \tag{14}$$

where $n(train(t)) = 7093$ (*i.e.*, 80% of the sample size of 8,867) in all the 100 splits, and $\phi_t(x_{i,j})$ represents the Shapley Value of explanatory variable *j* for training individual *i* in the t^{th} training set. The results appear in Figure 1 and Table 7 for the Original model, and Figure 3 and Table 8 for the Extended model.

1.4.1.1 Average Mean Absolute Shapley Values: Original Model

The Average Mean Absolute Shapley Values are depicted in Figure 1: the most important explanatory variable is the composite variable "Has a Partner". This changes the absolute predicted value of life satisfaction by on average 0.36 over the 100 train–test splits; the second most important explanatory variable is Emotional Health, with an average effect of 0.19.

Figure 1: Average Mean Absolute Shapley Values in the Original Model

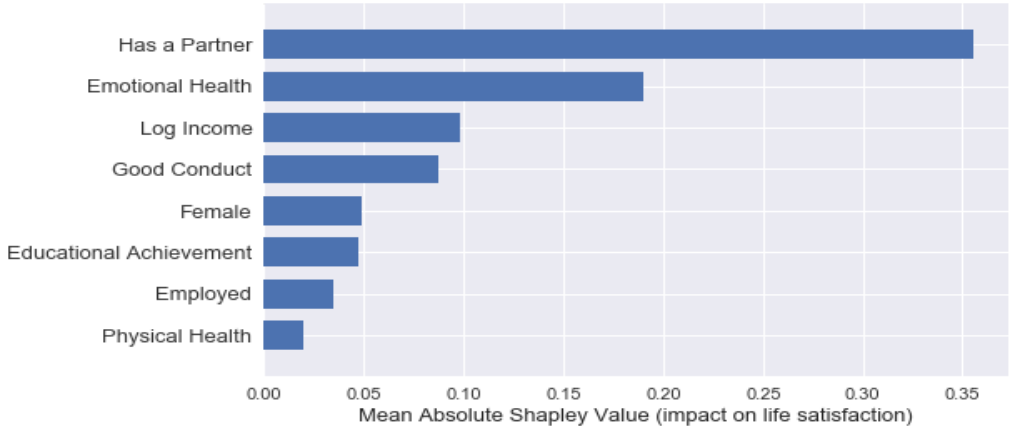


Table 7. Average Mean Absolute Shapley Values in the Original Model

Explanatory Variables	Average MASV	SD MASV
Has a Partner	0.36	0.01
Emotional Health	0.19	0.01
Log Income	0.10	0.01
Good Conduct	0.09	0.01
Female	0.05	0.01
Educational Achievement	0.05	0.01
Employed	0.03	0.00
Physical Health	0.02	0.00

Notes: This table shows the Average Mean Absolute Shapley Value (MASV) for each explanatory variable calculated over the same 100 different train–test splits considered in the Random Forests. Original model. Standard deviations are in parentheses.

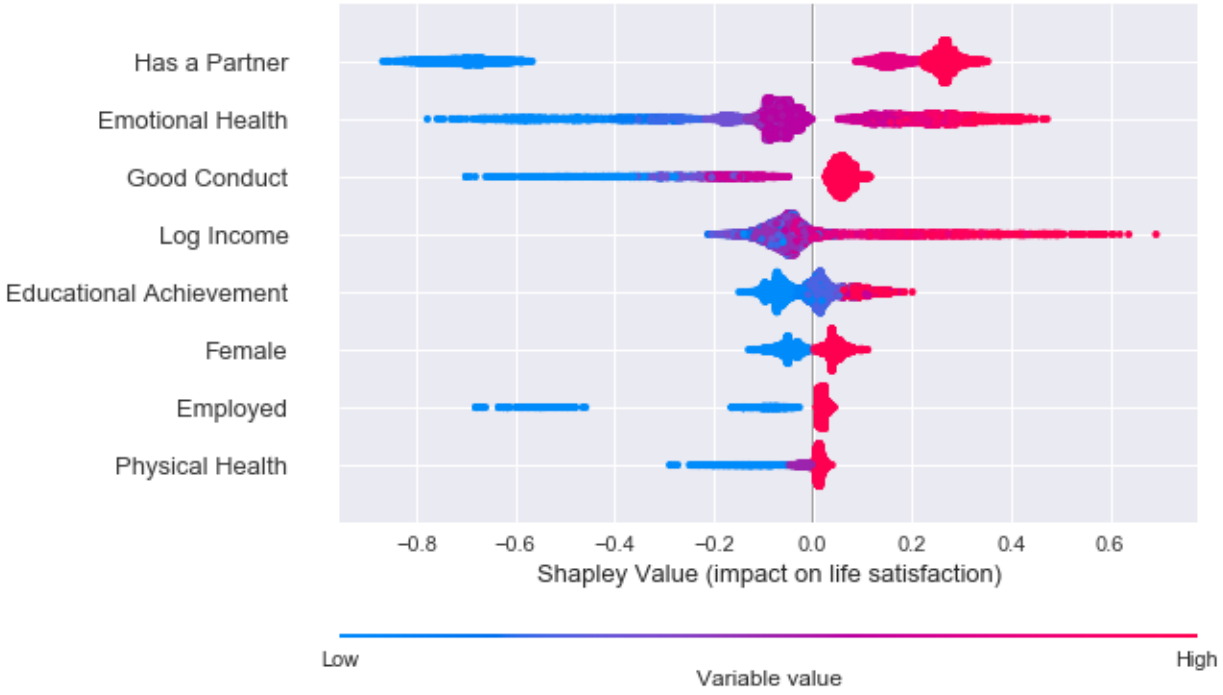
The Shapley Values can also tell us in which direction the explanatory variables affect the findings. The values presented below refer *to one Random Forest only (that calculated on train–test split 1)*. Nonetheless, given that the performance of this Random Forest and the average over all 100 forests are similar, the results there are generalizable. The rankings of the Average MASVs in Table 7 (calculated over the 100 Random Forests) and those in Figure 2 below are also similar.

The dots depicted in Figure 2 are the Shapley Values by individual by explanatory variable, the $\phi \quad (x_{i,j})$ in Equation (12), with the explanatory variables on the vertical axis and the Shapley Values on the horizontal axis. The explanatory variables are ranked from the most Shapley Important (Has a Partner) to the least (Physical Health), as shown in Figure 1.

The colors of the dots reveal whether the explanatory variable for that individual has a high or low value, ranked by color intensity ranging from red (high) to blue (low). Overlapping dots create ‘clouds’ that help to illustrate the distribution of the Shapley Values.

The patterns in Figure 2 allow a more–detailed understanding of the average absolute values plotted in Figure 1. Consider, for instance, Has a Partner, which is the most important explanatory variable: the two highest values of this variable, from Section 2, are 0.685 and 0.530, for being married with and without children respectively. The associated Shapley Values for these two highest values of Has a Partner in the first row of Figure 2 are represented by the red and purple dots, respectively. As can be seen, the Shapley Values of the Has a Partner variable are mostly clustered in the [0.1, 0.35] or [–0.9, –0.5] intervals: being Married with or without children increases life satisfaction, on average, by 0.1 to 0.35 points relative to the “average prediction for the dataset” (Molnar, 2019). Conversely, the two lowest values that the Has a Partner variable takes, 0 and –0.004 (for being single with and without children respectively), correspond to the blue Shapley Values and are associated with lower life satisfaction of 0.5 to 0.9 points.

Figure 2: Shapley Values by individual by explanatory variable – Original Model



Notes: The dots in each line represent the Shapley Values (as shown on the horizontal axis) for each individual for the variable indicated. The redder dots refer to higher values of the explanatory variable in question, and the bluer dots to lower values. Shapley Values at the individual level are calculated from the Random Forest fitted on training–test split 1.

The results are even more interesting for Emotional Health. As this explanatory variable is more continuous, the Shapley Values are distributed more uniformly. Having a high value of Emotional Health increases life satisfaction by 0.1 to 0.45 points. There is also a long left tail: predicted life satisfaction can be up to 0.8 points lower for the individuals with the lowest values of emotional health.

Criminality (Good Conduct) is the third–most important variable. The highest value here is for those who reported no crimes. As is evident from the figure, having no criminal record has only a small impact on predicted life satisfaction; instead, having committed crimes can sharply reduce satisfaction by up to 0.7 points. The logic here is that while no criminal record is normal (and so does not make the individual much more satisfied with life), having reported crimes is associated with sharply lower satisfaction. The same pattern is found for being employed and good physical health: being employed and not having health problems do not have positive effects on life satisfaction, but the lack of them (being unemployed or having health problems) has a sizeable negative effect. Health problems having such a large effect may reflect the relatively young age (34) of our sample.

Last, low income does not strongly negatively affect life satisfaction (the majority of the blue–dot Shapley Values are close to zero), but there is a large positive impact of higher income, of up to 0.7 points.

This ranking of explanatory variables is important for policy. Population life satisfaction can then be improved by focusing on the individuals in the left tails of the Shapley Values. Here Emotional Health, Family situation, Unemployment and Criminality appear central, as the explanatory variables associated with the largest drops in life satisfaction.

1.4.1.2 Average Mean Absolute Shapley Values: Extended Model

Figure 3 and Table 8 show the results for the Extended model. Marital Status and Emotional Health behave similarly to Has a Partner and Emotional Health in the Original model. The

individual Shapley Values for this extended set of variables, analogous to those for the Original model in Figure 2, appear in Figure 4. Many of these variables seem to have a systematic relationship with life satisfaction, as revealed by the separate clusters of dots according to the variable's different values (and the color of the individual dots).

Figure 3: Average Mean Absolute Shapley Values in the Extended Model

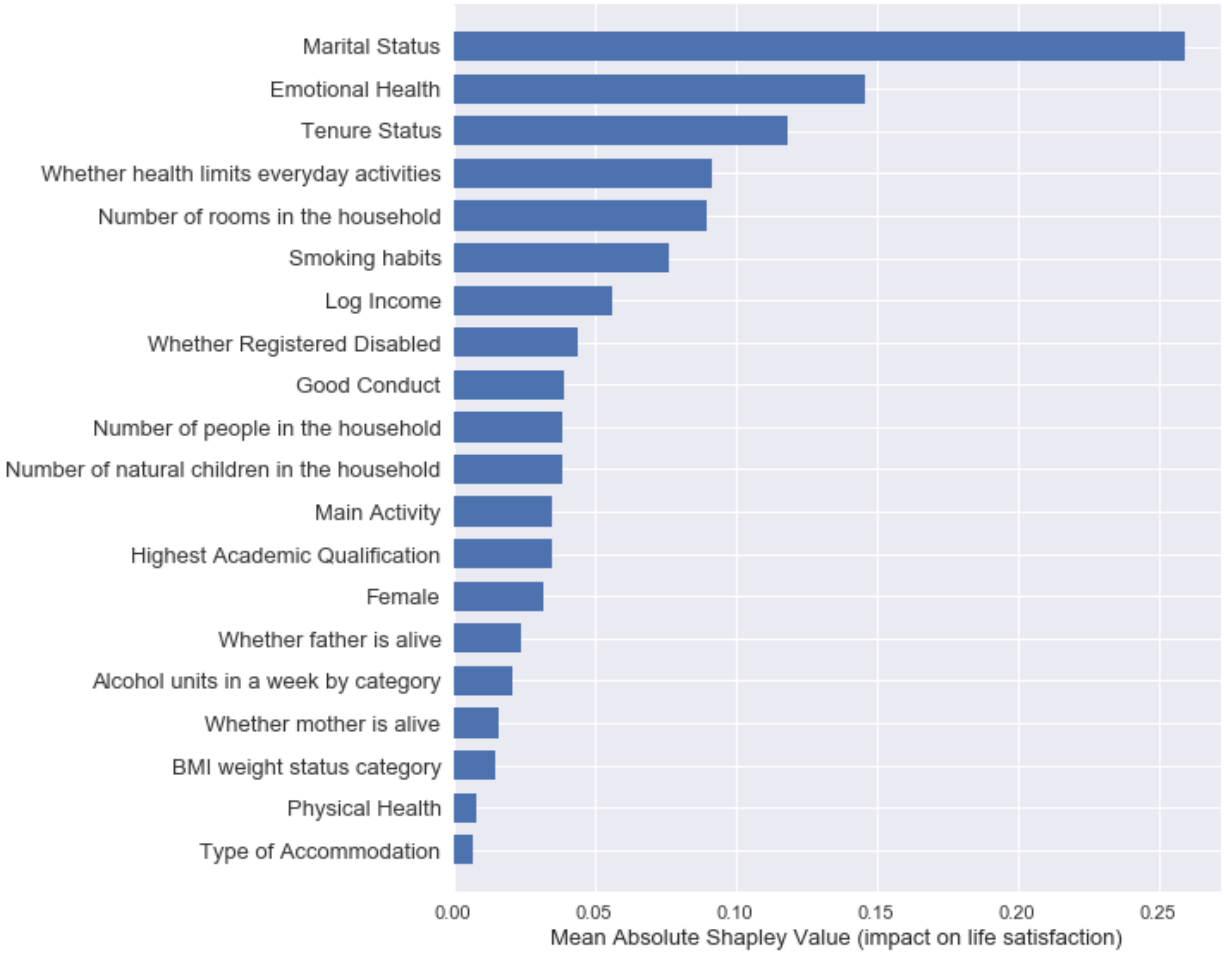
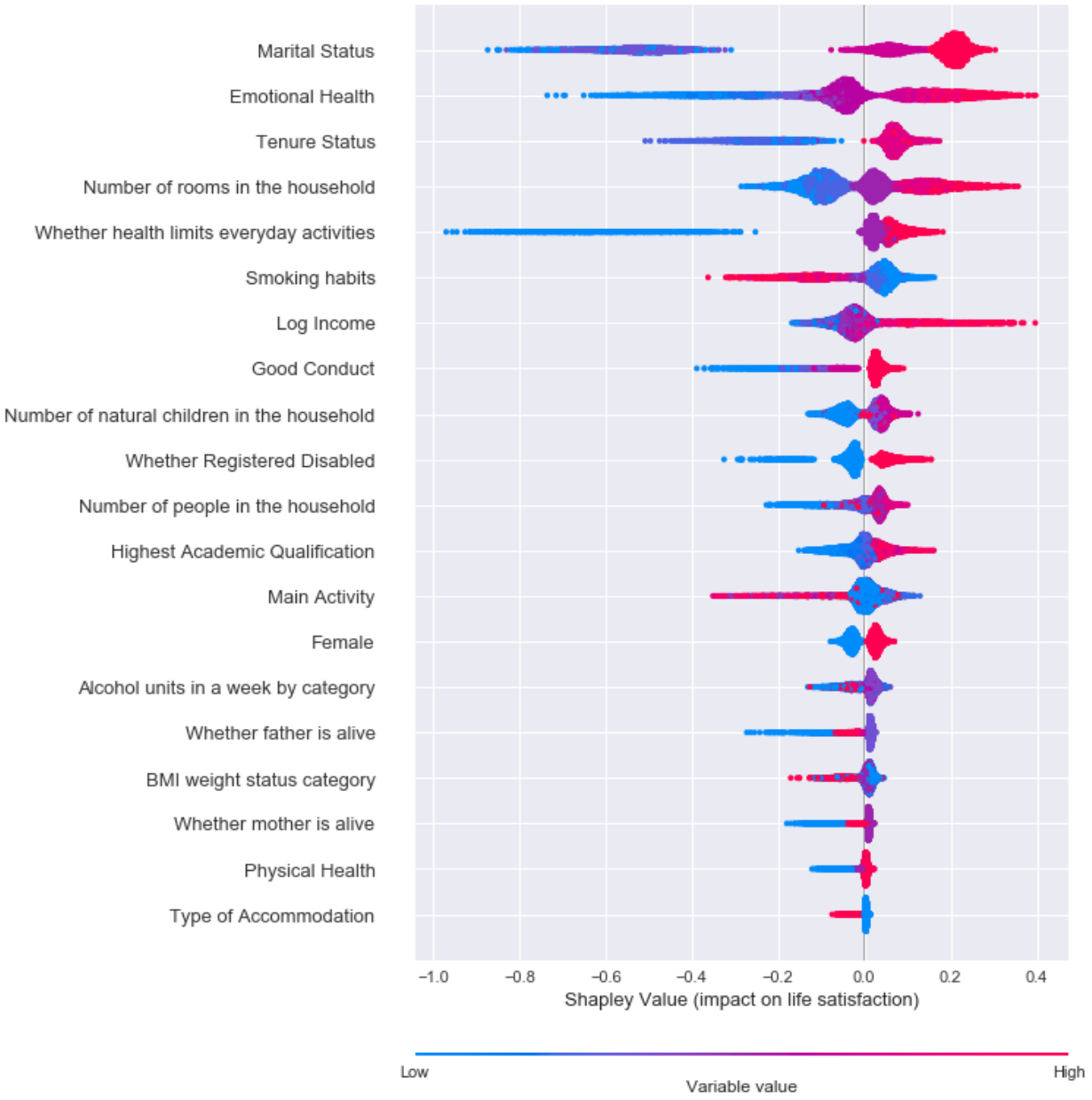


Table 8. Average Mean Absolute Shapley Values in the Extended dataset

Explanatory variable	Average MASV	SD MASV
Marital Status	0.26	0.008
Emotional Health	0.15	0.009
Tenure status	0.12	0.009
Number of rooms in the household	0.09	0.008
Whether health limits everyday activities	0.09	0.006
Smoking habits	0.08	0.007
Log Income	0.06	0.004
Number of natural children in the household	0.04	0.005
Number of people in the household	0.04	0.003
Good Conduct	0.04	0.004
Whether registered disabled	0.04	0.006
Main Activity	0.04	0.003
Highest academic qualification	0.03	0.005
Female	0.03	0.004
Whether father is alive	0.02	0.003
Alcohol units in a week by category	0.02	0.003
Whether mother is alive	0.02	0.003
BMI category	0.01	0.002
Type of accommodation	0.01	0.002
Physical Health	0.01	0.001
Number of non–natural children in the household	0.01	0.001

Notes: This table shows the Average Mean Absolute Shapley Values in the Extended dataset. Standard deviations appear in the right–hand column.

Figure 4: Shapley Values by Individual by Explanatory Variable – Extended Dataset



Notes: The dots in each line represent the Shapley Values (as shown on the horizontal axis) for each individual for the variable indicated. The redder dots refer to higher values of the explanatory variable in question, and the bluer dots to lower values. Shapley Values at the individual level are calculated from the Random Forest fitted on training–test split 1.

Marital Status in the Extended model is a different variable from Has a Partner in the Original model, as it now does not include the presence of children (children appear in a separate variable),

and takes on more values than simply Single or Married, now including Separated, Divorced, and Widowed (which are assigned the values of 3, 2 and 1 respectively, the lowest values for this variable). There is wide variation in the marginal effects for marital status, where the highest values (representing Married and Cohabiting, with values of 6 and 5) have a positive impact of up to 0.3 life–satisfaction points, but Single, Separated, Divorced or Widowed have large negative effects of 0.3 to 0.9 points.

Health limiting everyday activity has the largest negative impact on predicted life satisfaction, of up to 1 point, and behaves in the same way as Disability and Criminality (Good Conduct). Physical health, which is towards the bottom of Figure 4, has almost no effect on life satisfaction. We might wonder whether this reflects the inclusion of both disability and health limitations in the Extended Model. However, dropping these latter two continues to produce only very small Shapley Values (as illustrated in Figures 1 and 2, where this is the only physical–health variable). Our age–34 respondents report only few of the 15 health conditions in Appendix B: over–three quarters have none, and only 5% report two or more.

The impact of Emotional Health is again more–continuously distributed, with a large effect as illustrated in Figure 3. Some of the other explanatory variables are of more marginal importance, including gender, education, number of children, number of people in the household, and the type of accommodation. The first two of these were equally relatively unimportant in the Original Model.

1.4.2 Comparing Mean Absolute Shapley Values to the Linear Regression Coefficients

The MASV associated with an explanatory variable is its average absolute marginal effect on the predicted dependent variable. This measure is intuitively comparable to the coefficients from linear regression, which also reflect the marginal effect of a unitary change in the explanatory variable on the dependent variable. We here compare the two, taking only the Random Forest with 4000 trees fitted on training set 1. Insignificant coefficients (p–values > 0.05) are reported as 0. We start with the Original model.

1.4.2.1 Shapley Values and Regression Coefficients: Original model

It is intuitive to compare the MASVs, which reflect the mean absolute marginal impact of each explanatory variable, to the absolute linear regression coefficients. The results appear in Table 9, where the variables are ranked by MASV. The ranking in the two columns is identical for the continuous variables (which are all standardized). The comparison between the two columns is more difficult to carry out for Employed and Female, as these two coefficients are not standardized. The estimated coefficients are therefore larger than they would have been had the variables been standardized. On the other hand, standardization has no impact in Random Forests.

Table 9. Random Forest Mean Absolute Shapley Values and Absolute Linear Regression Coefficients – Original Model

Explanatory variable	MASV	|Coefficients|
Has a Partner	0.355	0.470
Emotional Health	0.177	0.293
Good Conduct	0.096	0.134
Log Income	0.092	0.117
Ed. Achievement	0.051	0.078
Female	0.047	0.216
Employed	0.034	0.988
Physical Health	0.021	0.000

Notes: This table compares Mean Absolute Shapley Values calculated from the optimized Random Forest to the Absolute Linear Regression Coefficients. All variables are standardized but the Employed and Female dummies in the Original model.

1.4.2.2 Shapley Values and Regression Coefficients: Extended Model

The comparison in the Extended Model is less straightforward. While the Shapley Values in this case can be interpreted in the same way as for the Original Model, this is not the case for the Ridge Regression Coefficients, as in the Extended Model we have added multiple (ordinal) multiclass categorical explanatory variables that are divided into dummies. We thus require a unique measure for these explanatory variables that is comparable to the MASVs from all of the coefficients on the associated dummies. We here choose the absolute weighted mean coefficient over all of the associated dummies, with the weights being the fraction of individuals in each of

the explanatory–variable categories. In this case, since the coefficients are from a Ridge regression, they also are standardized.

Formally, suppose that the explanatory variable X_j is a multiclass categorical variable with *k* categories, split into *k* dummies for the Ridge Regression. Let $\chi_{j,l}$ be the proportion of individuals in the training set in the l^{th} category of explanatory variable *j*:

$$\chi_{j,l} = \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} I\left(x_{i,j} = l\right) \qquad (15)$$

where $I(x_{i,j} = l)$ is the indicator function with value 1 if individual *i* belongs to the *l*th category of the *j*th explanatory variable, and 0 otherwise. Then, given the $\hat{\beta}_{j,1}, \ldots., \hat{\beta}_{j,k}$ estimated Ridge Regression coefficients, the Derived Coefficient is:

$$\hat{\beta}_j = \sum_{l=1}^{k} \chi_{j,l} \left|\hat{\beta}_{j,l}\right| \ . \qquad (16)$$

We only carry out this calculation for the multiclass categorical explanatory variables (which are indicated by underlined coefficients in the final column below). Numerical discrete variables (whether binary, such as Female, or with multiple values, like Number of People in the Household), and the variables that are treated as numerical continuous (Log Income and Emotional Health) enter the Ridge Regression as they are, and the absolute coefficient in the table below is entered directly from the regression output.

Table 10. Random Forest Mean Absolute Shapley Values and Absolute Ridge Regression Coefficients – Extended Model

Explanatory variable	**MASV**	**|Coefficients|**
Marital Status	0.260	0.292
Emotional Health	0.138	0.190
Tenure Status	0.113	0.105
Number of rooms in the household	0.100	0.110
Whether health limits everyday activities	0.089	0.104
Smoking Habits	0.066	0.076
Log Income	0.055	0.067
Good Conduct	0.047	0.060
Number of natural children in the household	0.043	0.001
Registered disabled	0.041	0.050
Number of people in the household	0.039	0.022
Main Activity	0.031	0.090

Highest academic qualification	0.031	0.048
Female	0.029	0.136
Alcohol units in a week by category	0.024	0.054
Father is alive	0.022	0.090
BMI category	0.018	0.034
Mother is alive	0.017	0.047
Physical Health	0.008	0.009
Type of Accommodation	0.007	0.009
Number of non–natural children in the household	0.006	0.029

Notes: This table compares Mean Absolute Shapley Values calculated from the optimized Random Forest to the Absolute Ridge Regression Coefficients. The results are from the Extended model. The underlined coefficients in the final column are calculated using Equation (16).

In Table 10, the values of the multiclass categorical explanatory variables (calculated via Equation (16)) are underlined. The two most important variables in both columns are marital status and emotional health. The most–notable difference between the two columns of Table 10 is the estimated effect of Female (which is standardized in Ridge Regression): here the MASV is more than four times smaller than the associated Ridge coefficient. In the Ridge Regression, Female is the third most–important explanatory variable. But in terms of MASVs it is only the 14th most–important explanatory variable. However, the estimated Ridge Regression coefficients should perhaps be taken with a grain of salt, as there is some risk that they overestimate the expected marginal impact of the explanatory variables on the dependent variable, given the assumed linearity of the dependent variable in the parameters and, under the linearity assumption, their bias. We conclude this section by discussing Permutation Importance, to assess the impact of each explanatory variable in determining the model's predictive accuracy.

1.4.3 Permutation Importance

The idea of Permutation Importance is simple. Once we have randomized, via shuffling, one of the explanatory variables in the test set, say the *j*th, its Permutation Importance is defined as the difference between the scoring metric that we consider (in our case, the MSE) calculated from the actual X_j and its shuffled version, X_{j^*} , keeping all of the other variables unshuffled at their original values. This operation is performed multiple times, and Permutation Importance is then calculated as the average difference in the scoring metric across the multiple repetitions. While

this operation can be carried out for both the test and training sets (see Breiman, 2001), we here consider only the Test Set, as this represents a diagnostic measure of predictive accuracy. The results refer to the Random Forest on train–test split 1.

Table 11. Random Forest Permutation Importance – Original Model

Explanatory variable	Weight (Standard Deviation)
Has a Partner	0.113 (0.011)
Emotional Health	0.043 (0.006)
Log Income	0.024 (0.005)
Good Conduct	0.013 (0.004)
Employed	0.009 (0.002)
Female	0.004 (0.002)
Physical Health	0.004 (0.002)
Educational Achievement	0.002 (0.001)

Notes: This table shows Permutation Importance calculated on the Test Set of the Original Model considering the best–performing Random Forest, measuring the fall in predictive accuracy across 100 shuffles of each explanatory variable. The figures in parentheses are standard deviations.

Table 12. Random Forest Permutation Importance –Extended Model

Explanatory variable	Weight (Standard Deviation)
Marital Status	0.064 (0.007)
Whether health limits everyday activities	0.028 (0.005)
Emotional Health	0.027 (0.004)
Log Income	0.011 (0.003)
Main activity	0.010 (0.002)
Tenure Status	0.010 (0.003)
Smoking habits	0.007 (0.002)
Number of rooms in the household	0.006 (0.003)
Good Conduct	0.004 (0.002)
Number of natural children in the household	0.003 (0.001)
Whether Registered Disabled	0.003 (0.002)
Number of people in the household	0.003 (0.002)
Whether father is alive	0.002 (0.001)
Female	0.002 (0.001)
Whether mother is alive	0.001 (0.001)
Highest Academic Qualification	0.001 (0.001)
Alcohol units in a week by category	0.001 (0.001)
Type of Accommodation	0.000 (0.000)

Number of non–natural children in the household	0.000 (0.000)
Physical Health	0.000 (0.000)
BMI weight status category	0.000 (0.001)

Notes: This table shows Permutation Importance calculated on the Test Set of the Extended Model considering the best–performing Random Forest, measuring the fall in predictive accuracy across 100 shuffles of each explanatory variable. The figures in parentheses are standard deviations.

The first intuitive finding from Tables 11 and 12 is that, in the Original Model with 8 explanatory variables, the average marginal impact of randomizing explanatory variables on predictive accuracy is greater than in the richer Extended Model with 21 explanatory variables. It is also clear that Permutation Importance is not monotonic with respect to the cardinality of the explanatory variable. Take, for example, Has a Partner and Log Income in the Original model. The former takes on only 4 different values, while the latter is continuous. Hence, when randomizing (shuffling) the former, the probability that an individual's shuffled value is the same as their original value is higher, which in turn should mechanically reduce its Permutation Importance. Nonetheless, the Permutation Importance of Has a Partner is almost five times higher than that of Log Income: Permutation Importance then does capture the actual importance of an explanatory variable in predicting life satisfaction, rather than simply modeling the noisy characteristics of the explanatory variable itself, such as its cardinality.

1.5 Discussion

In this work we have constructed a predictive model for life satisfaction using data from the British Cohort Study (BCS). We evaluate the predictive performance of our models relative to the benchmark OLS regression in Layard *et al.* (2014). We first use only the eight original adult variables that appeared there (with a different version of self–assessed physical health, as updated in Clark and Lepinteur, 2019), and then turn to an Extended model that has 21 explanatory variables: 5 of the original 8, plus 16 new variables (some of which are more–detailed versions of the other 3 of the original 8). Splitting these categorical variables up into their separate values produces 96 dummy variables.

We found no evidence of improvement in model fit using more–advanced ML methods. In the Extended model, we first have to penalize the linear models due to numerical problems

including multicollinearity, or exclude from the analysis some of the least–populated categories. The Extended Model with the 16 new explanatory variables allows us to improve the predictive accuracy, in testing, by 5.3% in terms of a lower Average Test MSE figure.

The best–optimized Random Forest produced no improvement over the Penalized Linear Regressions on the test set in the Extended Model.

Last, to help interpret the importance of the different explanatory variables in the prediction of life satisfaction, we considered two model–agnostic interpretability tools applied to the Random Forest: Permutation Importance and Shapley Values. The latter allows the comparison of the machine–learning results to the estimated coefficients from Penalized Linear Regressions.

Shapley Values assess the marginal impact of the (significant) different explanatory variables at the individual level. In other words, Shapley Values do not pick up the *average* effect of a one–unit change in the explanatory variable (as for the coefficients of a linear regression model) but the marginal impact of *every single value of that explanatory variable*. Another advantage of using a Machine Learning algorithm like Random Forest, where the explanatory variables do not need to be split in dummies (as long as they are ordinal), is that we can take into account the categories that we dropped in the Linear Unpenalized Regression. The comparison of the Random Forest Shapley Values to the estimated Ridge Regression coefficients suggests that some caution should be exercised regarding coefficient size in the latter. This in particular applies to gender: in the Extended dataset, there is a significant difference between the Female MASV and the linear regression coefficient, with the latter being nine times larger than the former. This is in line with Oparina and Srisuma (2022) who, in non–parametric estimation of the measurement error in reported life satisfaction, find a negative relation between female and *latent* life satisfaction (*i.e.*, the true value of the variable), but a positive coefficient for *reported* life satisfaction.

Our work here has considered the subjective judgment of life satisfaction, but we believe that the prediction of objective variables will also benefit from non–linear machine–learning analyses.

Regarding the most important predictors of life satisfaction, our comprehensive analysis confirms that Marital Status as well as Emotional and Physical Health (in terms of limitations to everyday activities) are always the most important explanatory variables, in line with the findings from the existing literature.

Appendix

Appendix 1

Explanatory Variables	Mean	SD	Min	Max
Log Income	9.28	0.598	6.23	12.4
Educational Achievement	0.20	0.251	0	0.75
Employed	0.98	0.130	0	1
Has a Partner	0.48	0.285	0.00	0.66
Good Conduct	24.50	1.699	0	25
Female	0.52	0.500	0	1
Marital Status – Other missing	0.00	0.018	0	1
Marital Status – Married	0.54	0.498	0	1
Marital Status – Cohabiting	0.21	0.404	0	1
Marital Status – Single (never married)	0.19	0.394	0	1
Marital Status – Separated	0.02	0.149	0	1
Marital Status – Divorced	0.03	0.182	0	1
Marital Status – Widowed	0.00	0.037	0	1
Type of Accommodation – Not Applicable	0.01	0.076	0	1
Type of Accommodation – A house or bungalow	0.88	0.326	0	1
Type of Accommodation – Flat or Maisonette	0.11	0.310	0	1
Type of Accommodation – Studio flat	0.00	0.044	0	1
Type of Accommodation – A room / rooms	0.00	0.041	0	1
Type of Accommodation – Something else	0.00	0.057	0	1
Tenure Status – Refusal	0.00	0.065	0	1
Tenure Status – Do not Know	0.00	0.015	0	1
Tenure Status – Own (outright)	0.05	0.221	0	1
Tenure Status – Own – buying with help of a mortgage/loan	0.69	0.462	0	1
Tenure Status – Pay part rent and part mortgage (shared/equity ownership)	0.01	0.067	0	1
Tenure Status – Rent it	0.19	0.393	0	1
Tenure Status – Live here rent–free	0.04	0.185	0	1
Tenure Status – Squatting	0.00	0.015	0	1
Tenure Status – Other	0.02	0.147	0	1
Main Activity – Do not know	0.00	0.015	0	1
Main Activity – Full–time paid employee	0.58	0.494	0	1
Main Activity – Part–time paid employee (under 30 hours a week)	0.16	0.365	0	1
Main Activity – Full–time self–employed	0.08	0.273	0	1
Main Activity – Part–time self–employed	0.02	0.127	0	1

Main Activity – Unemployed and seeking work	0.02	0.137	0	1
Main Activity – Full–time education	0.01	0.092	0	1
Main Activity – On a government scheme for employment training	0.00	0.028	0	1
Main Activity – Temporarily sick/disabled	0.00	0.042	0	1
Main Activity – Permanently sick/disabled	0.02	0.153	0	1
Main Activity – Looking after home/family	0.10	0.302	0	1
Main Activity – Other	0.01	0.108	0	1
Highest Academic Qualification – Do not know	0.00	0.034	0	1
Highest Academic Qualification – None	0.09	0.286	0	1
Highest Academic Qualification – CSE	0.15	0.359	0	1
Highest Academic Qualification – GCSE	0.09	0.289	0	1
Highest Academic Qualification – GCE O Level	0.24	0.428	0	1
Highest Academic Qualification – A/S Level	0.02	0.128	0	1
Highest Academic Qualification – Scottish School Certificate, Higher School Certificate	0.02	0.145	0	1
Highest Academic Qualification – GCE A Level (or S Level)	0.05	0.225	0	1
Highest Academic Qualification – Nursing or other para–medical qualification	0.02	0.128	0	1
Highest Academic Qualification – Other teaching qualification	0.01	0.086	0	1
Highest Academic Qualification – Diploma of Higher Education	0.08	0.267	0	1
Highest Academic Qualification – Other degree level qualification such as graduate membership	0.05	0.217	0	1
Highest Academic Qualification – Degree (e.g. BA, BSc)	0.12	0.325	0	1
Highest Academic Qualification – PGCE–Post–graduate Certificate of Education	0.02	0.135	0	1
Highest Academic Qualification – Higher degree (e.g. PhD, MSc)	0.04	0.205	0	1
Whether Registered Disabled – Do not know	0.00	0.030	0	1
Whether Registered Disabled – Yes	0.02	0.132	0	1
Whether Registered Disabled – No but long–term disability	0.63	0.482	0	1
Whether Registered Disabled – No and no long–term disability	0.35	0.477	0	1
Whether health limits everyday activities – Yes	0.07	0.258	0	1
Whether health limits everyday activities – No but health problems since last interview	0.51	0.500	0	1
Whether health limits everyday activities – No and no health problems since last interview	0.42	0.494	0	1
BMI weight status category – Insufficient data	0.03	0.164	0	1
BMI weight status category – Underweight (< 18.5)	0.01	0.119	0	1
BMI weight status category – Normal (18.5–24.9)	0.47	0.499	0	1
BMI weight status category – Overweight (25–29.9)	0.33	0.470	0	1
BMI weight status category – Obese (30 and above)	0.16	0.368	0	1
Smoking habits – Other missing	0.00	0.011	0	1
Smoking habits – Never smoked	0.45	0.498	0	1
Smoking habits – Ex smoker	0.24	0.425	0	1
Smoking habits – Occasional smoker	0.07	0.246	0	1
Smoking habits – Up to 10 a day	0.09	0.290	0	1
Smoking habits – 11 to 20 a day	0.13	0.337	0	1
Smoking habits – More than 20 a day	0.02	0.144	0	1

Smoking habits – Daily but frequency not stated	0.00	0.026	0	1
Alcohol units in a week by category – Never drinks or only on special occasions	0.19	0.392	0	1
Alcohol units in a week by category – None reported	0.08	0.266	0	1
Alcohol units in a week by category – 1 to 14	0.48	0.500	0	1
Alcohol units in a week by category – 15 to 21	0.10	0.305	0	1
Alcohol units in a week by category – 22 to 39	0.10	0.294	0	1
Alcohol units in a week by category – More than 39	0.05	0.226	0	1
Whether mother is alive – Do not know	0.00	0.032	0	1
Whether mother is alive – Missing	0.00	0.055	0	1
Whether mother is alive – Yes in household	0.07	0.254	0	1
Whether mother is alive – Yes	0.86	0.346	0	1
Whether mother is alive – No	0.03	0.158	0	1
Whether mother is alive – No reported dead last sweep	0.04	0.197	0	1
Whether father is alive – Do not know	0.01	0.105	0	1
Whether father is alive – Missing	0.00	0.051	0	1
Whether father is alive – Yes in household	0.05	0.220	0	1
Whether father is alive – Yes	0.79	0.410	0	1
Whether father is alive – No	0.05	0.218	0	1
Whether father is alive – No reported dead last sweep	0.10	0.300	0	1
Number of people in the household	3.11	1.274	1	10
Number of natural children in the household	1.09	1.090	0	8
Number of non–natural children in the household	0.07	0.357	0	4
Number of rooms in the household	4.70	1.531	1	12
Physical Health	0.30	0.610	0	4
Emotional Health	0.83	0.119	0	1

Appendix 2

Physical Health
Please tick all that apply. Have you suffered from any of these…
Hay Fever
Asthma
Bronchitis
Wheezing when you have a cold flu
Skin problems
Fit, convulsions, epilepsy
Persistent joint of back pain
Diabetes
Persistent trouble with teeth, gums or mouth

Cancer
Stomach or other digestive problems
Bladder or kidney problems
Hearing difficulties
Frequent problems with periods or other gynecological problems
Other health problem

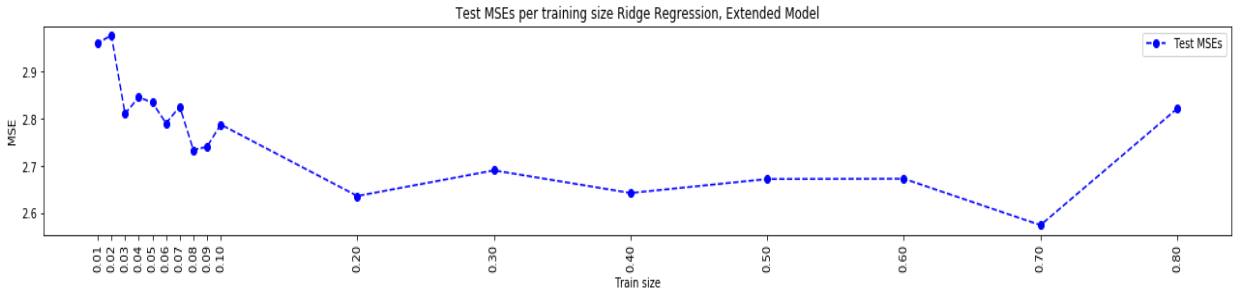
Appendix 3

Learning Curves refer to the behavior of the MSE calculated on the Test Set as function of the size of the training set, based on the idea that more–complex data–generating processes (DGP) may require larger training sets. The understanding of the necessary size of the training set required to correctly learn the DGP is useful for a number of reasons. First, should we be interested in carrying out new analyses on the same data, we can save time by fitting the new algorithms only to the required amount of training data. Second, this can help us to better understand the complexity of the underlying DGP. And last, it can provide guidance for the training set size required for the analysis of similar, but not identical, data. In the Extended model, we limit our discussion to the Ridge Regression, and for the Original model we present the Unpenalized Learning Curves. In both the Original and Extended models, all of the five different Penalized Linear Regressions considered have similar learning behavior. We also plot the curves from Random Forests for both models, trained on non–standardized values.

Figure 5: Learning Curve of Linear Regression on the Original Data

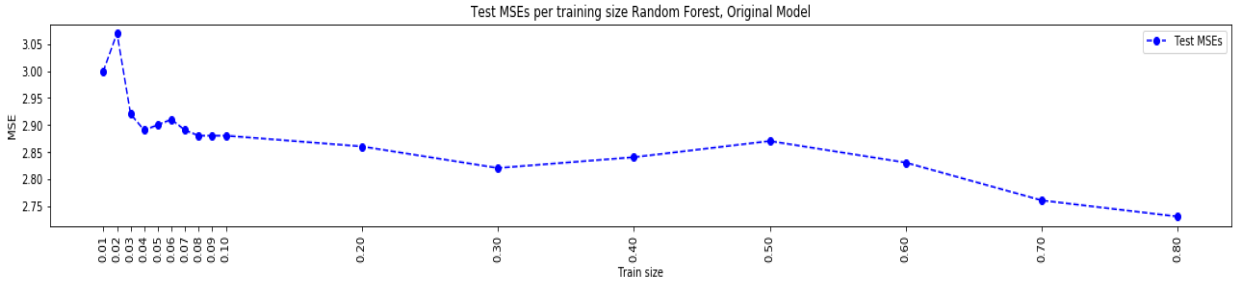
In the Original model, we start with the Unpenalized Linear Regression. Here the DGP is already fully learned with only 2% of individuals in the training set. This is consistent with our finding that an Unpenalized Linear Regression is the best choice for these data, and that the linearity assumption holds: the correct DGP is learned very quickly. Here, the MSEs converge to the bias only.

Figure 6: Learning Curve of the Ridge Regression on the Extended Data



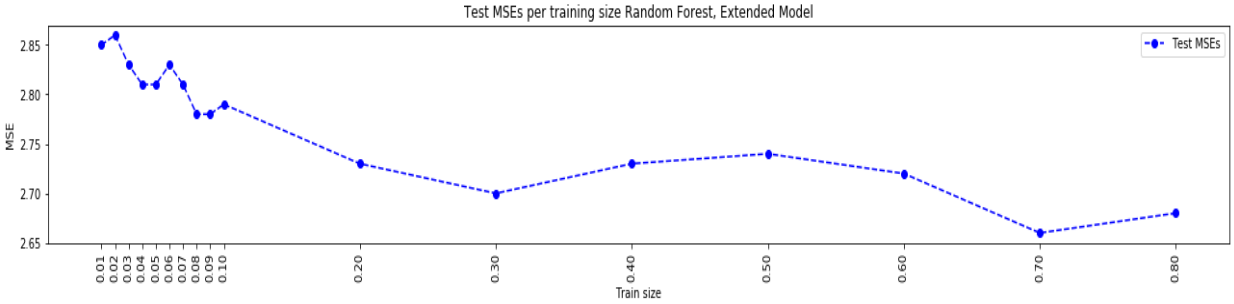
The behavior in the Ridge Regression on the Extended dataset is similar to that in the Unpenalized Linear Regression in the Original dataset. In this case, the Test MSE also stabilizes for training sets including more than 20% of individuals.

Figure 7: Learning Curve of the Random Forest on the Original Data



For the Learning Curves in the Random Forest, in the Original model the DGP is learned confidently with 3% of observations in the training set.

Figure 8: Learning Curve of the Random Forest on the Extended Data



The DGP is also confidently learned with 10% of individuals in the training set in the Extended model, with the Test MSE thereafter remaining constant.

Appendix 4

Categories with at most 15 individuals	**Number of individuals**
Type of Accommodation – A room / rooms	15
Main Activity – Don't Know	2
Main Activity – On a government scheme for employment training	7
Main Activity – Wholly Retired	1
Whether registered disabled – Don't Know	8
Highest academic qualification – Don't Know	10
Marital Status – Widowed	12
Marital Status – Other missing	3
Whether mother is alive – Don't Know	9
Smoking habits – Daily but frequency not stated	6
Smoking habits – Other missing	1
Tenure Status – Squatting	2
Tenure Status – Don't Know	2

Chapter 2

Human Wellbeing and Machine Learning

2.1 Introduction

Over the last 40 years, researchers from various fields have established an immense literature on the correlates and determinants of subjective wellbeing (Clark 2018, Diener *et al.* 2018, Nikolova and Graham, 2020). In parallel, international organisations (OECD 2020) and national governments (ONS 2021) have turned to subjective wellbeing data as a key tool for policy analysis. However, despite the widespread use of wellbeing scores, our current ability to predict wellbeing is limited. Conventional linear models, where variables are selected based on intuition or theory, explain little individual–level variation. Typically, models of individual wellbeing produce an R–squared of no more than 15%.

In response, we here evaluate whether Machine Learning (ML) algorithms can improve our capacity to understand wellbeing.

We answer two research questions:

- RQ1: Are ML algorithms significantly better at predicting wellbeing than conventional linear models? What is the upper limit on our ability to predict wellbeing based on survey data?
- RQ2: Are the variables that are identified by ML algorithms as important in predicting wellbeing the same as those in the conventional literature?

To answer these questions, we use Random Forests (Breiman 2001, Hastie *et al.* 2009), Gradient

Boosting (Friedman 2001, Natekin and Knoll 2013), and Penalized Regressions (Tibshirani 1996) as examples of ML algorithms. Random forests and Gradient boosting are tree–based algorithms that have been shown to perform well with tabular data (Shwartz–Ziv and Armon 2022).[5]

Penalized Regressions are a convenient tool for analyses that involve large number of covariates, like ours (Tibshirani 1996). Generally, these techniques can identify more–complex models of wellbeing than traditional linear models, potentially improving predictive performance. Unlike standard regression techniques, these algorithms allow for the inclusion of an arbitrary number of variables, and, in the case of our tree–based methods, can identify nonlinearities and interactions between variables.

Earlier works on wellbeing and Machine Learning focused on relatively small country – and year–specific samples (Margolis *et al.* 2021), or particular drivers of wellbeing, such as age (Kaiser *et al.* 2022).

We carry out our empirical analysis using three of the largest currently–available datasets that include wellbeing information: the German Socio–Economic Panel (SOEP), the UK Household Longitudinal Study (UKHLS), and the American Gallup Daily Poll. The SOEP has data on about 30,000 unique respondents and 400 distinct variables; the UKHLS surveys around 40,000 individuals in each wave and has over 500 distinct variables; and each year of the Gallup data has information on around 200,000 respondents with approximately 60 distinct variables. We can thus study the extent to which utilizing more information about individual respondents improves the predictive power of wellbeing models.

Regarding RQ1, we find that ML algorithms predict somewhat better than standard linear models. The size of this improvement is small in absolute terms, but substantial when compared to the predictive power of key variables, such as health. Increasing the number of variables in the model from a standard set (we call this the "Restricted Set") to all available data (the "Extended Set") has a far larger effect on predictive model performance.

Predictive accuracy, judged by the R–squared on unseen data, roughly doubles for both OLS and ML methods. Independently of the type of algorithm, an R–squared of 0.30 appears to be the

[5] 1In contrast, other ML algorithms, such as neural networks, tend to perform poorly on tabular data, which is why we do not consider them here (Borisov *et al.* (2022)). In preliminary analyses we did indeed find that feed-forward neural networks yielded performances that were no better than OLS.

feasible maximum given the available data.[6] For RQ2, our data–driven ML results validate the findings of the conventional literature. We find that variables related to respondents' social connections, health and material conditions are consistently among the most important in predicting wellbeing.

Variable importance is assessed using Permutation Importances (Breiman 2001, Kuh *et al.* 2002) and by computing pseudo partial effects for all algorithms, including OLS. In general, there is substantial correlation in variable importance rankings across algorithms (ρ = 0.58 to ρ = 0.83), so that ML approaches and OLS are largely in agreement in terms of what matters most for wellbeing.

2.2 Materials and Methods

2.2.1 Data

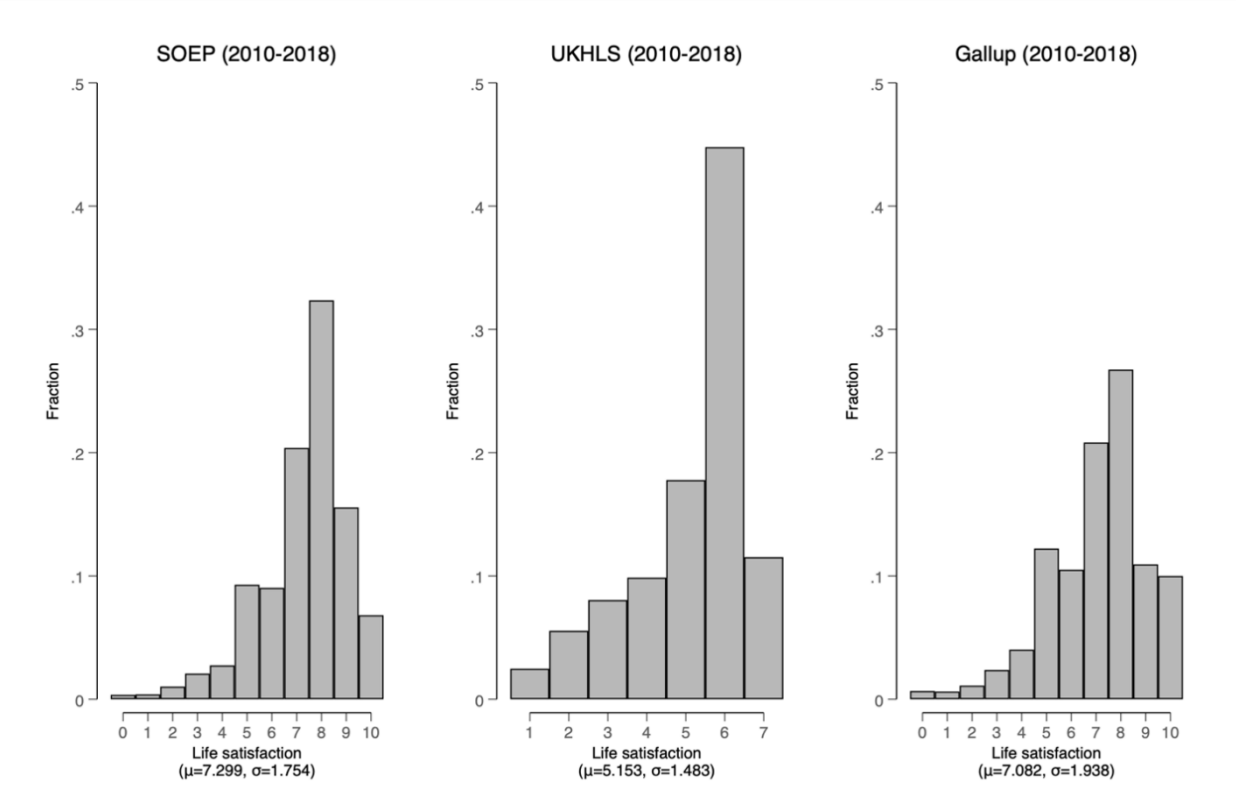
We analyze data from three nationally–representative surveys over the 2010 to 2018 period: the German Socio–Economic Panel (SOEP), the UK Longitudinal Household Survey (UKHLS), and the US Gallup Daily Poll (Gallup).

The Gallup data covers the US adult population, with daily cross–sectional telephone–based surveys of 500 (1000 until 2012) respondents. After removing incomplete data, this yields an annual sample ranging from N=115,192 (in 2018) to N=351,875 (in 2011). Wellbeing is measured by the Cantril Ladder of Life (Cantril 1965), which asks: "Please imagine a ladder with steps numbered from zero at the bottom to ten at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?" Answers are recorded on a scale from 0 to 10, with equal steps between response options.[7]

[6] Our estimations on the extended set of variables, which include all of the variables apart from other measures of subjective wellbeing, produce R-squared figures of between 0.25 and 0.32 across the different datasets.

[7] There has been controversy about whether such data can support inferences about underlying wellbeing (Bond and Lang (2019), Chen *et al.* (2019), Kaiser and Vendrik (2020), Schröder and Yitzhaki (2017)). We here remain agnostic about this issue. We instead rather ask which algorithms and models best predict the answers to wellbeing questions, without making any further claims about how these answers relate to respondents' underlying feelings.

Figure 1: Histograms of life satisfaction for SOEP, UKHLS and Gallup data.



The SOEP is representative of the German adult population, with interviews conducted in person. To allow for a direct comparison with the Gallup data, we here consider the survey period between 2010 and 2018. In each year, between N= 26,089 and N= 32,333 observations are available. Life

satisfaction is measured on a scale from 0 to 10, from the question: "We would like to ask you about your satisfaction with your life in general, please answer according to the following scale: 0 means completely dissatisfied and 10 means completely satisfied: How satisfied are you with your life, all things considered?"

The UKHLS is representative of the UK adult population. Interviews are conducted in person. We again confine our analysis to the same 2010–2018 period (corresponding to Waves 2 to 10). The number of available annual observations is between N=29,605 to N=40,679. Life satisfaction is measured on a 1 to 7 scale. Respondents are asked: "How dissatisfied or satisfied are you with your life overall?"

Descriptive statistics and histograms of each wellbeing measure appear in Figure 1. The wellbeing distributions are very similar across datasets. As is typically found in high–income countries, wellbeing is strongly left–skewed.

2.2.2 Algorithms

We model wellbeing using four kinds of algorithms. First, as our baseline and corresponding to the workhorse of a great deal of research on subjective wellbeing, we apply Ordinary Least Squares (OLS) to solve Linear Regressions. OLS estimates are the solution to the problem:

$$argmin_b \sum_{i=1}^{N} (x_i' b - s_i)^2 \quad (1)$$

Here, x_i is a vector of explanatory variables and *b* the vector of coefficients. The wellbeing of respondent *i* is denoted by s_i. Let $\hat{b}$ be the solution to Equation 1. Then, the predicted wellbeing level on the respondent *i* is $\widehat{s_i} = x_i'\hat{b}$. When using OLS, the researcher implicitly assumes that reported wellbeing is a linear combination of the chosen set of explanatory variables *x*. If these assumptions are an appropriate description of the true data–generating process, OLS will provide accurate predictions of individual wellbeing. In applications with a large number of covariates, the performance of OLS may degrade due to overfitting or multicollinearity between included explanatory variables.

The second algorithm we consider, the Least Absolute Shrinkage and Selection Operator (LASSO), tackles this issue by adding a penalty for the sum of the magnitudes of the estimated coefficients. In particular, LASSO estimates are the solution to:

$$argmin_b \sum_{i=1}^{N} (x_i'b - s_i)^2 + \lambda \sum_{k=1}^{K} |b_k| \quad (2)$$

Here, λ is a hyperparameter, the preferred value of which is found using a grid search. LASSO and OLS are equivalent for $\lambda = 0$. Although LASSO may improve predictions by reducing the risk of overfitting, the algorithm continues to assume an additive functional form. Nevertheless, one helpful property of LASSO is that it shrinks coefficients on the variables with low explanatory power to zero. In some specifications, we thus use LASSO as a device for variable selection.

The third and fourth algorithms we consider – Random Forests (RF) and Gradient Boosting (GB) – are based on Regression Trees (Breiman 1984). Regression Trees are generated via a recursive binary splitting algorithm. The algorithm splits the sample along values of covariates and predicts the outcome in each subsample, or *node*, as the mean outcome within each node. More formally, at each step *k*, the data *D* is split into two nodes, $D_{L,k}$ and $D_{R,k}$. The location of the split within the data is determined by some variable x_j and some threshold $\tau_{k,j}$. The nodes $D_{L,k}$ and $D_{R,k}$ are defined as (see Hastie *et al.* 2009):

$$D_{L,k} = \{x \mid x_j < \tau_{k,j}\};\ D_{R,k} = \{x \mid x_j \geq \tau_{k,j}\} \ (3).$$

The predicted values are the mean value of *s* within each node, *i.e.* $\hat{s}D_{m,k} = N^{-1}{}_{D_{m,k}} \sum_{i:X_i \in D_{m,k}} s_i$, for *m* in *{L,R}*, where $N_{D_{m,k}}$ is the number of respondents in each node. The splitting variable x_j and the threshold $\tau_{k,j}$ are determined by minimizing the following residual sum of squares:

$$min_{j,\tau_{k,j}} \sum_{i:X_i \in D_{L,k}} \left(s_i - \hat{s}D_{L,k}\right)^2 + \sum_{i:X_i \in D_{R,k}} \left(s_i - \hat{s}D_{R,k}\right)^2 \quad (4)$$

Finally, the nodes $D_{L,k}$ and $D_{R,k}$ are in turn used as inputs for the next step. This procedure is repeated until some final number of *leaves* is found. By construction, every split reduces the

in–sample mean squared error (MSE).[8] Hence, if the size of the tree is not limited, the algorithm will overfit the data. Limiting the maximum tree size can ameliorate this issue by reducing the variance of the predictions. However, this comes at the cost of increasing the bias of the resulting estimates (Hastie *et al.* 2009). Alternatively, the variance in the predictions can be reduced by aggregating the predictions from multiple trees. Random Forests and Gradient Boosting are both examples of this strategy.

Specifically, Random Forests, the third algorithm we consider, rely on averaging across a large number of trees (which we set to 1,000 for all the three datasets)[9]. Each individual tree has low bias but high variance. When the correlation between the trees is low, averaging across the predictions of multiple trees reduces the variance of the predictions without introducing additional bias. To carry out this procedure, each individual tree is grown on a nonparametrically bootstrapped sample of the original data. The correlation between trees is further reduced by considering only a random subset of all covariates at each split. The size of this subset, Nvars, is a hyperparameter that we select based on a grid search.

The fourth algorithm, Gradient Boosting, proceeds by sequentially fitting regression trees on the residuals of the predictions of the previous collection of trees[10]. Intuitively, each subsequent tree attempts to explain the variance that was not explained by the previous trees. We begin with the predictions $\hat{s}_{T_1}$ of a first tree T_1 and calculate the residual $\hat{s}_{T_1} - s_i = e_{T_1}$. A second tree T_2 is then fitted on these residuals to obtain predicted residuals $\hat{e}_{T_1}$.The updated overall predictions are then given by $\hat{s}_{T_1} + \hat{e}_{T_1} = \hat{s}_{T_2}$. A third tree is subsequently trained on the residuals $\hat{s}_{T_2} - s_i = e_{T_2}$. This process is repeated *Ntrees* times, producing increasingly accurate predictions of *s*. Since gradient boosted collections of trees overfit with large *Ntrees*, we select this hyperparameter via

[8] Mean squared error measures the average of the squares of the errors – the average squared difference between the predicted and reported levels of wellbeing.

[9] The performance of the random forest is non-decreasing in the number of trees. In our application, increasing the number of trees to 2,000 for UKHLS and Gallup and to 10,000 for SOEP yields qualitatively similar results. The final number of trees was chosen to render the optimisation less computationally-expensive.

[10] This construction of the trees, when the residuals from the previous tree are used to build the following tree, is specific to a case when the partitioning of the tree is chosen to minimise the sum of squared residuals in each node. The construction differs when other objective functions are used. See Friedman (2001) and Hastie *et al.* (2009) for the general case. We here use a standard implementation of gradient boosting. In preliminary tests, we also evaluated the performance of extreme gradient boosting (XGBoost; Chen and Guestrin 2016) in the Gallup and SOEP datasets. The use of XGBoost only yielded negligible improvements compared to standard gradient boosting, which is why we here focus on the latter.

a grid search. Moreover, to further reduce overfitting, the size of the update at each step is reduced by adding a penalty $0 < \lambda \leq 1$ and predictions are updated with the rule $\hat{s}_{T_k} + \lambda \hat{e}_{T_k} = \hat{s}_{T_{k+1}}$. The penalty λ is also selected via grid–search[11]. As is customary, the algorithms are trained on the training set, which here contains 80% of the sample. Each algorithm's performance is then estimated on the test set, which contains the remaining 20% of observations. Optimal hyperparameters are chosen via 4–fold cross validation. Optimal hyperparameters for all the datasets can be found in Appendix Table A1. Each of these algorithms are implemented using the *scikit–learn* library in Python (Pedregosa *et al.* 2011). To evaluate the stability of our results across time, where feasible, we train each algorithm on each survey–wave combination separately.

2.2.3 Explanatory variables

We evaluate each algorithm's performance for two different sets of explanatory variables.

As noted above, we first consider a restricted set of variables that are observed in all three of the datasets, which cover basic demographics as well as economic and health variables. We specifically include: sex, age, age–squared, ethnicity, religiosity, number of household members, number of children in the household, marital status, log household income (equivalised using the modified OECD scale), general health status, disability status, body mass index, labour–force status, working hours, home ownership, area of residence, and interview month. A more detailed description of these variables is provided in Appendix Table A2. These variables are typical in the conventional literature on subjective wellbeing. This restricted set of variables will then allow us to assess the performance of ML algorithms relative to OLS in a standard estimation setting.

We also evaluate each algorithm on a much larger extended sets of explanatory variables. Here, we only use the 2013 Wave of Gallup and SOEP, and Wave 3 of the UKHLS (which covers 2011–2012)[12]. Our dataset includes all of the available variables, apart from direct measures of subjective wellbeing (such as domain satisfaction, happiness, or subjective health) or mental

[11] The maximum size of each tree in the gradient-boosting algorithm is significantly smaller than in the case of random forests. Consequently, the individual trees in such an ensemble are called weak learners (Freund 1995, Freund and Schapire 1999).

[12] These waves/years were chosen as they include personality traits in the SOEP and UKHLS.

health. We also exclude variables with more than 50% missing values. The resulting Gallup dataset contains 67 variables, and around 450 variables are retained in the SOEP and UKHLS. Missing values for continuous variables are assigned the observed means, while missing values for categorical variables are assigned a new category[13]. We convert categorical variables into a set of dummies, one for each category. The full list of variables in this extended set appears in the supplementary material. The large number of variables in the extended set produces significant computational burden. At the same time, it is evident that some portion of these variables will have no predictive power for wellbeing. We therefore use LASSO as a device to select the explanatory variables (Tibshirani 1996, Ahrens *et al.* 2020)[14]. We have carried out the estimations on both the full–extended set and the post–LASSO extended set. Typically, both approaches perform similarly. For simplicity, we only show results for the approach that performed better in each individual case.

2.2.4 Assessing Variable importance

To answer our second research question, we need to assess how important each explanatory variable is in enabling our algorithms to predict wellbeing. We do so in two ways.

We first use Permutation Importances (PIs) to measure the degree to which each algorithm relies on a given variable in making its predictions (Molnar 2019)[15].

PIs are calculated by randomly shuffling a given variable's observed values across individuals in the test data and evaluating the extent to which the predictive performance (in terms of R–squared) of a given algorithm falls when permuting the variable's values. This operation is carried out 10 times. The reported PI is the average change in the R–squared across these 10 iterations. The greater the average fall in the R–squared, the more important is the variable.

To understand the direction of our variables' effects we also report *pseudo partial effects* (PPEs). These are calculated by taking the difference in predicted wellbeing after setting each explanatory

[13] Processing categorical variables and removing perfectly collinear variables respectively yields 210, 542, and 957 effective explanatory variables in the Gallup, SOEP and UKHLS datasets.

[14] Using LASSO on the restricted set of variables produced a similar performance to OLS, with optimal $\lambda = 0$.

[15] Shapley values are an alternative option to assess feature importances. We did not compute Shapley Values because of their substantial computational complexity (Lundberg *et al.* 2018; Yang 2021), and since our pseudo marginal effects already allow us to identify the direction of variables' effects.

variable to a given set of values. Specifically, for continuous and ordinal variables we set the variable to the third and first quartile of their distributions and calculate the mean difference in predicted wellbeing. For binary variables (including dummies for all of the categorical variables), we predict wellbeing when setting each individual's value to either 0 or 1.

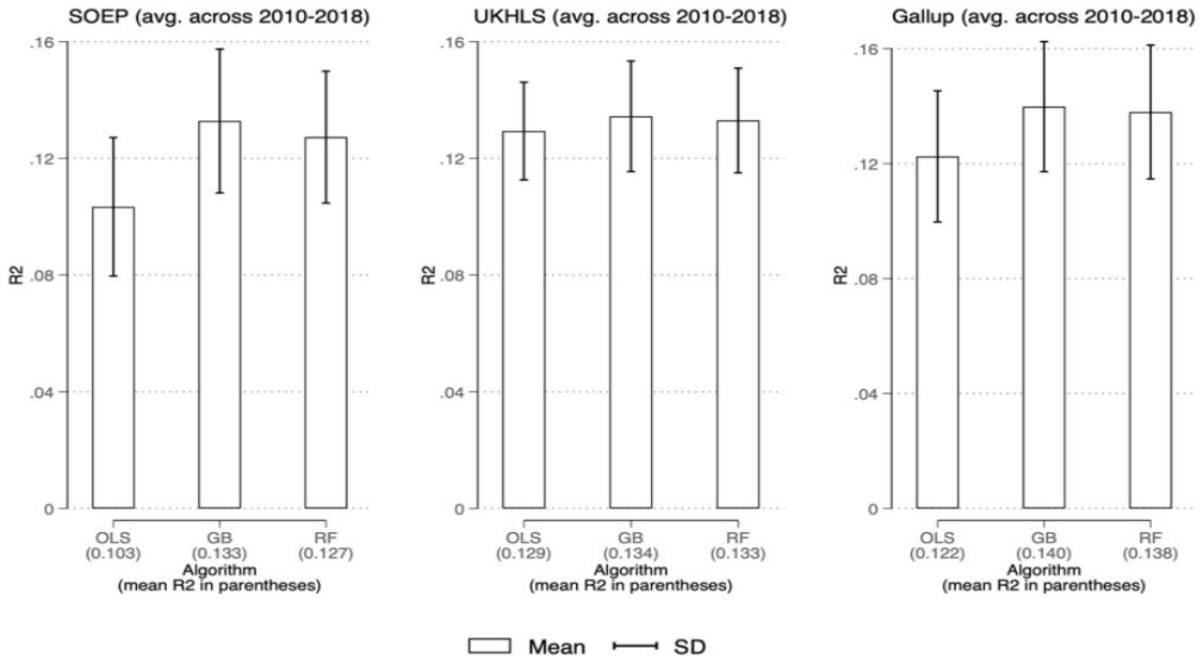
A key advantage of PIs and PPEs is that they can be used with any kind of algorithm, allowing us to compare the way in which each algorithm makes use of the available data.

2.3 Results

2.3.1 Model performance

We begin with RQ1, *i.e.* whether ML algorithms significantly outperform OLS in predicting wellbeing. As noted, OLS is the standard approach followed in the conventional literature.

Figure 2: R–squared figures from OLS, GB and RF using the restricted set of variables. The R-squareds are computed using the unseen testing data



2.3.1.1 The Restricted Set of explanatory variables

We start with the analysis based on the restricted set of covariates, which includes the variables that are typical in many conventional wellbeing estimations. Figure 2 depicts the performance of each algorithm on the test–set portion of each dataset. We use R–squared as our primary evaluation metric in order to facilitate the comparison with previous analyses.

In Figure 2 each algorithm is trained separately for each year between 2010 and 2018. The values refer to the average R–squared across these years and their standard deviations. The R–squareds are very similar across datasets, ranging from 0.10 (SOEP) to 0.14 (Gallup). Gradient Boosting (GB) and Random Forests (RF) yield larger R–squared values than OLS in each case. Specifically, Random Forests yield absolute increases in R–squared of 0.024 (SOEP), 0.004 (UKHLS) and 0.016 (Gallup); the respective improvements from using Gradient Boosting are slightly larger, with respective R–squared gains of 0.030, 0.005, and 0.018[16].

ML algorithms thus do outperform Linear Regressions, and Gradient Boosting always outperforms Random Forests.

These gain figures considered on their own are hard to interpret. To illustrate the substantive size of these improvements, we compare them to the change in predictive performance when omitting information on respondent's health status – a key wellbeing predictor – in our baseline OLS regressions.

[16]These gains are calculated from the test set, which was not used for training the algorithm. In the training set, *i.e.* the data that is observed by each algorithm, the improvement from performance of the ML algorithms over OLS is larger (see Appendix Figure A1). The predictive capacity of the ML algorithms applied to the test set does not then seem to be constrained by underfitting. Of course, performance in the training set is not per se indicative of the quality of an algorithm. A decision tree with as many leaves as training individuals would yield an MSE of 0. However, this model would perform extremely poorly when used to assess unseen test data.

Table 1. An illustration of the size of the improvements from using ML

	OLS, full	OLS, no health	GB	GB gain as % of loss from removing health
Panel A: Restricted set of variables				
SOEP	0.103	0.075 (Δ=0.028)	0.133 (Δ=0.030)	107%
UKHLS	0.129	0.095 (Δ=0.034)	0.134 (Δ=0.005)	15%
Gallup	0.122	0.093 (Δ=0.029)	0.140 (Δ=0.018)	62%
Panel B: Extended set of variables				
SOEP	0.284	0.240 (Δ=0.043)	0.318 (Δ=0.035)	81%
UKHLS	0.215	0.197 (Δ=0.018)	0.243 (Δ=0.028)	155%
Gallup	0.270	0.240 (Δ=0.031)	0.280 (Δ=0.018)	58%

Notes: The figures refer to the R–squared values from the test–set.

Panel A of Table 1 lists the changes in the test–set R–squared of the OLS regression when omitting this information and compares this figure to the gain from using Gradient Boosting. As benchmarked against the gain from adding health information, the prediction–improvement figure from Gradient Boosting (as our best ML algorithm) lies between 15% and 107%. When evaluated in this way, the gains from using ML do look substantial.

2.3.1.2 The Extended Set of explanatory variables

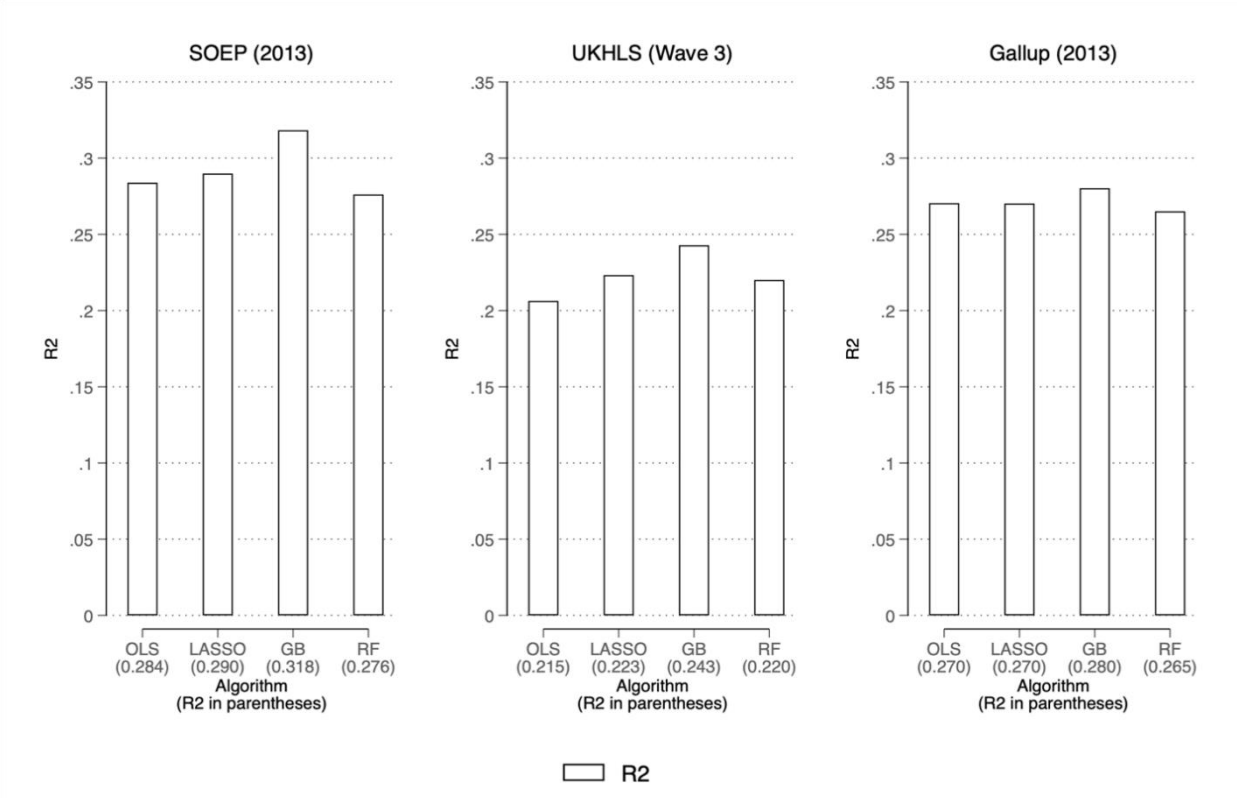
Adding further explanatory variables should increase our ability to predict wellbeing. Given the greater flexibility of the ML algorithms, we should expect these to benefit more from additional variables than OLS. To test this, we estimate all of our models on the extended sets of variables. As explained in Section 2.3, these extended sets include all of the variables available in the 2013 waves of the SOEP and Gallup, and Wave 3 of the UKHLS.

Figure 3 depicts our main results[17]. The R–squared figure approximately doubles using the extended set for all algorithms, including OLS. The OLS R–squared is now 0.28 for the SOEP, 0.21 in the UKHLS and 0.27 for Gallup. As such, standard economic specifications do not fully exploit the predictive information available in typical large–scale survey data.[18]

[17] The results for the training set can be found in Appendix Figure A2.

[18] All of these R-squared estimates are obtained using the test set. Hence, these improvements cannot be attributed

Figure 3: R–squared figures from OLS, LASSO, GB and RF using the extended set of variables. The R–squareds are computed using the unseen testing data



Gradient Boosting remains the best–performing algorithm and clearly predicts better than OLS.

to a mechanical increase in the share of explained variance due to adding more variables to the model.

The absolute gain in the R–squared from Gradient Boosting over OLS is now 0.034, 0.028 and 0.010 for the SOEP, UKHLS and Gallup respectively. Random Forests now tend to perform poorly, underperforming OLS for SOEP and Gallup. This has also been observed in other empirical applications where covariates were measured with error (Reis *et al.* 2018).

We again interpret the size of the gains from Gradient Boosting by comparing them to those from the inclusion of respondents' health information when using OLS.[19] The results in Panel B of Table 1 illustrate that these gains are again substantial, being approximately equivalent to the role of health in predicting wellbeing.

We thus conclude that tree–based ML algorithms can provide improvements in predictive performance over conventional methods. These gains are moderate in absolute terms, but are meaningful when compared to the predictive power of health. However, we also note that these gains are obtained with algorithms that take up to 100 times more time to estimate.[20]

The use of ML algorithms thus involves a trade–off between computational burden and predictive performance.

There are multiple reasons that can explain why nonlinear ML methods do not yield a substantial improvement in predicting human wellbeing compared to Linear Regression. First, most independent variables in the datasets we have used are binary or categorical. Such datasets cannot exhibit nonlinearity except by interaction terms between variables. Therefore, if a large number of the variables present only take binary values the ways that improvements can occur with non–linear models are limited. It is possible that non–linear relationships do exist but the variables concerned have a small contribution to the outcome. This is particularly likely if there are many variables contributing to the outcome, as is the case in our extended set of independent variables. Additionally, it may be that the non–linearity is present only at the extremes of the distribution where only few points exist.

As well as improvements in performance, ML may also indicate new, and potentially–overlooked, variables that are key in explaining subjective wellbeing. The next section explores this idea.

[19] In these extended specifications, there are multiple variables related to health in each dataset. We remove 21, 19 and 12 health-related variables in the Gallup, the SOEP and UKHLS respectively.

[20] This figure is based on a comparison between OLS and RF on the Gallup data with the extended dataset.

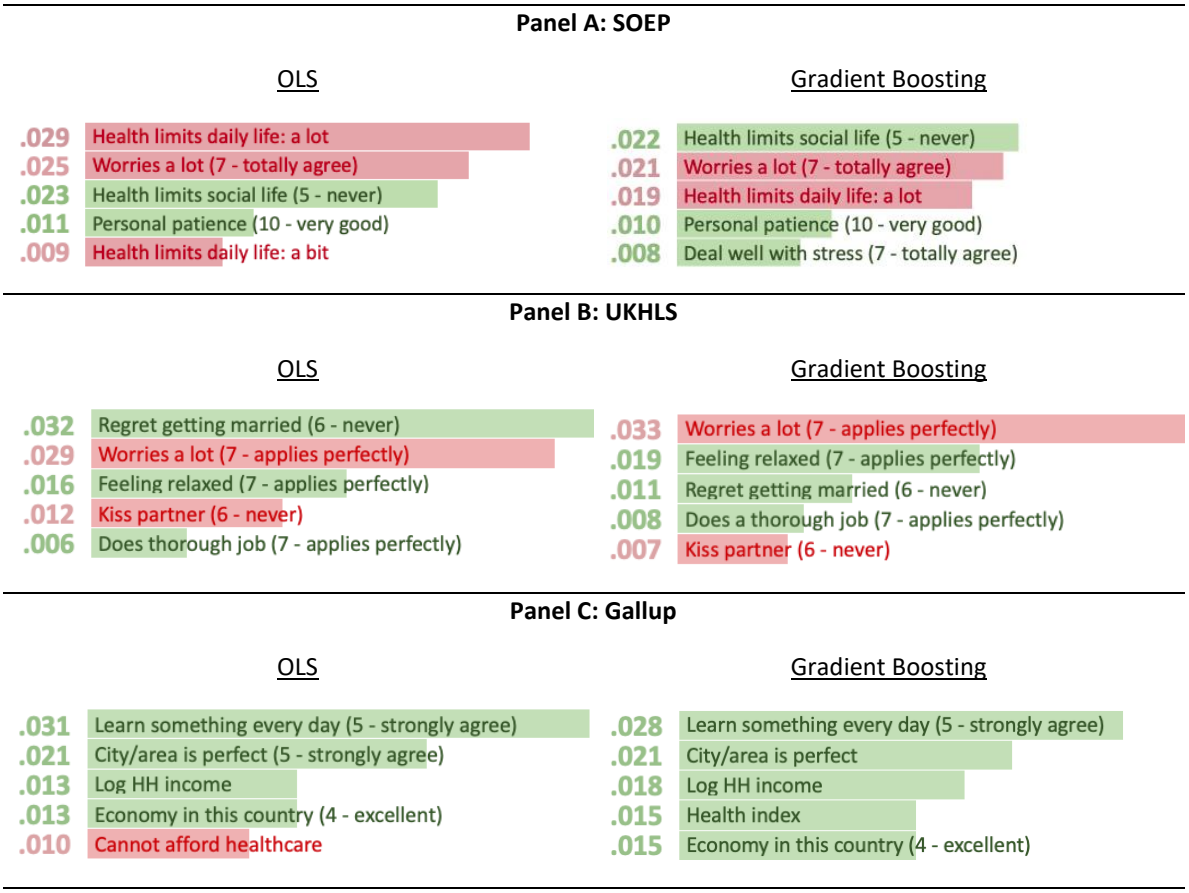
2.3.2 Variable importance

In this section we ask whether the variables that ML identifies as important in predicting life satisfaction correspond to those emphasised in the conventional literature. We do so by estimating variable importances, as discussed in Section 2.4. Our ML–based findings turn out to fit well with the results in previous analyses.

We start by estimating variable importances in the extended dataset, which provides more possibilities for the identification of important variables that do not appear in conventional wellbeing models. Figure 4 lists the five most–important variables identified in OLS and GB, which is the best performing ML algorithm, in each dataset[21]. The bars and numerical values refer to Permutation Importance, *i.e.* the drop in the model's R–squared when the values of the variable are randomly permuted across respondents. The variables that are negatively associated with average wellbeing are in red, and those with a positive association in green. In all three countries, individual health and interpersonal relationships are among the most–important predictors. As expected, respondents whose health limits their activities are on average less satisfied, while people with fulfilling relationships are typically more satisfied with their lives. The directions of the estimated effects are in line with those in the previous conventional work. ML algorithms and OLS thus generally agree on the direction and approximate size of the most–important variables (see Appendix Table A3 for the effect–size estimates).

[21] We present the Top-10 most-important variables for OLS, RF and GB in the three datasets in Appendix Table A3.

Figure 4: Permutation importance and pseudo partial effects of OLS and GB on the extended set of variables, 5 most–important variables



Notes: The bars and numerical values represent permutation importance and are coloured red for variables with negative pseudo partial effects and green otherwise. For Likert–scale variables, the highest category is reported.

As a more systematic measure of the degree of agreement between ML and OLS, we calculated the correlations of the ranks (in terms of their permutation importance) of each variable across algorithms and datasets. The results appear in Table 2.

Table 2. Correlations between the Permutation Importance ranks in different algorithms

	OLS vs. GB	OLS vs. RF	GB vs. RF
SOEP	0.70	0.58	0.79
UKHLS	0.75	0.67	0.86
Gallup	0.86	0.69	0.82

Notes: The correlation figures refer to the Top–100 variables (using the OLS ranking). These are Spearman rank correlations.

There is strong agreement between GB and RF in all three datasets, with the rank correlation figure never falling below 0.79. The correlations with the OLS ranking are somewhat lower, with a minimum value of 0.58 (OLS *vs.* RF in SOEP). Nevertheless, we can strongly reject ($p < 0.001$) the null hypothesis that the rankings are uncorrelated, supporting our conclusion that the OLS and ML algorithms are in broad agreement.

Apart from the conventional variables used in wellbeing analysis, such as health and interpersonal relationships, the algorithms also identify personality traits as important predictors in the UKHLS and SOEP. Personality traits, unfortunately, do not appear in the Gallup survey. In the UK data, measures associated with (the absence of) neuroticism (*i.e.* worrying, and feeling relaxed) appear in the Top–3. In German data, worrying a lot, being able to deal with stress, and patience are among the most–important variables in all empirical approaches. This is in line with previous research underlining the potential advantages of including personality traits in wellbeing regressions (Ferrer–i–Carbonell and Frijters 2004, Proto and Zhang 2021).

Beyond these similarities, there are some cross–country differences. The most striking refer to the importance of financial factors. These are important in the US (e.g., HH income and being able to pay for healthcare) but not in the other countries. To see whether this is a genuine finding or a consequence of differences in variable availability across countries, we carry out the same analysis using the restricted set (for which we have a common set of variables). When we do so, the cross–country differences in the importance of income largely disappear. As shown in Appendix Table A4, the most–important variables identified in these harmonized datasets are very similar across the three countries. They include health, income, marital and employment status, as well as home–ownership – which is a proxy for wealth – and age. Sex and ethnicity are only important in the US. Education is among the most important factors in the US and Germany, but not in the UK.

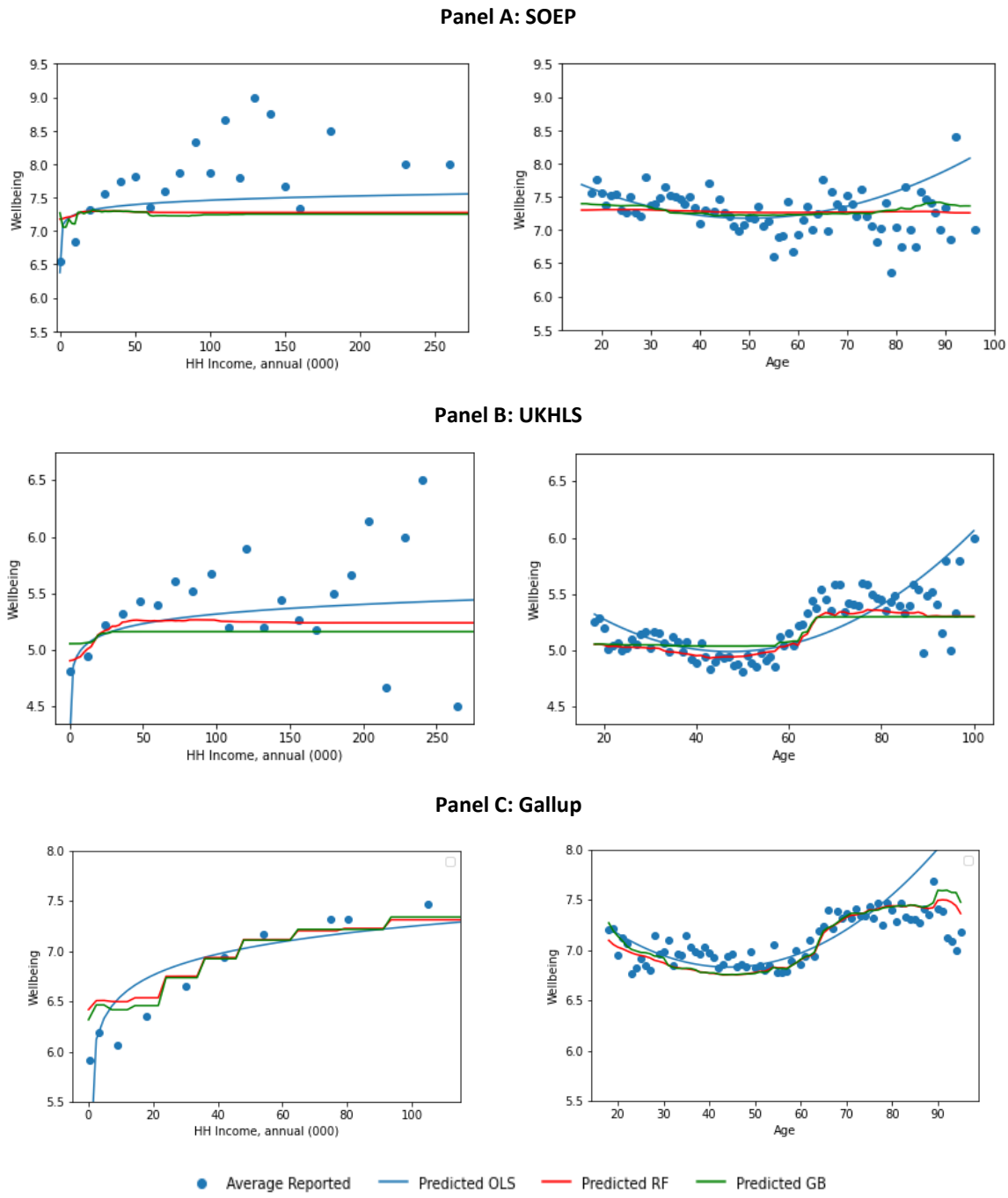
2.3.3 Additional analyses and robustness tests

2.3.3.1 Wellbeing by age and income

The preceding section concluded that the kinds of variables that ML finds to be important – and the estimated direction of their association with wellbeing – are largely in line with the results in the conventional literature. We here present a detailed analysis of two variables that have attracted a great deal of interest in the conventional literature: age and income. In OLS estimation, the functional forms associated with these two variables are imposed by the analyst, while they are instead freely estimated in our tree–based ML algorithms.

The results appear in Figure 5 and Appendix Figure A3.

Figure 5: The mean effects of age and household income on wellbeing, restricted set of variables



Notes: In the UKHLS and the SOEP, income is measured continuously, and we calculated equivalised annual income. This latter is trimmed at a figure of 250 thousand in local currency. This restriction retains over 99.9% of the income distribution in each country. Income in Gallup is collected in bands, and

household size data was not collected in 2013. We therefore analyze non–adjusted banded household income data in the US.

In the OLS estimation, illustrated in blue, we assume a quadratic form for age, and a log–linear functional form for income, which are very common functional forms in this literature. The relationships for RF are in red, and those for GB in green.

For low to medium incomes, both ML algorithms track the assumed log–normal functional form remarkably closely, in line with the conventional literature. However, once we reach relatively–high equivalised annual income figures, above 50,000 EUR in the SOEP or 40,000 GBP in the UKHLS, the ML algorithms suggest that wellbeing no longer increases with income. We cannot confirm this finding in the US, as income in Gallup appears in bands with the highest band being 100,000 USD or above. In 2013, 100,000 USD were approximately equivalent to 70,000 GBP or 78,000 EUR[22]. In addition, the Gallup 2013 wave did not collect data on household size. As a result, household income in Gallup is not directly comparable to the adjusted equivalent incomes in SOEP and UKHLS. Given these caveats, we do not find evidence of satiation in the US data. Our ML findings are therefore in line with previous work on wellbeing using US data (Kahneman and Deaton 2010, Killingsworth 2021).

With respect to the relationship between age and wellbeing, our ML estimations replicate the well–known approximate U–shape up to age 70 (*e.g.*, Cheng *et al.* 2017), which is more pronounced in the US. However, unlike the smooth U–shape assumed in the OLS approach, we find a much more pronounced "kink" at around age 65 for each dataset and ML–algorithm. We suspect that this kink reflects the gains in wellbeing following on from retirement (Gorry *et al.* 2018, Wetzel *et al.* 2016).

2.3.3.2 Positive and negative affect

We also evaluate the performance of Gradient Boosting and Random Forests on measures of positive and negative affect. The results show that our findings are not specific to the use of life evaluations as the measure of subjective wellbeing, but generalize to affect (or mood). In the 2013 Gallup data, positive affect is measured by the average figure from dummy variables indicating

[22] https://data.oecd.org/conversion/purchasing-power-parities-ppp.htm

whether the respondent felt happiness or joy, or smiled during the previous day. Negative affect is calculated analogously from dummies indicating pain, worry, sadness, and anger. In the German SOEP, positive affect is the self–reported frequency of being happy in the last 4 weeks (on a 1 to 5 scale), and negative affect as the average of the self–reported frequency from three questions about being angry, sad, or worried in the last four weeks (all measured on a 1 to 5 scale). The UKHLS dataset does not contain comparable affect data and is not used in this part of analysis.

The results for the Gallup data appear in the top panels of Figure A4 and Table A5. It is notable that negative affect is easier to predict than positive affect. This finding holds across algorithms, with R–squared figures ranging from 0.423 and 0.464 for negative affect, and between 0.261 and 0.296 for positive affect. Random Forests and Gradient Boosting outperform both OLS and LASSO. As was the case for life evaluations, Gradient Boosting again performs the best, with gains in R–squared over OLS of 0.041 for negative affect and 0.036 for positive affect. Regarding variable importances, Table A5 shows that good health is even more important for predicting positive and negative affect in the Gallup data than it was for life evaluation. Moreover, in line with previous work (*e.g.*, Kahneman and Deaton 2010), variables relating to material conditions – like income – do not feature in the list of the set of most–important variables when modelling affect. As shown in Table A6 and the bottom panels of Figure A4, the results are qualitatively similar in the German data: Gradient Boosting again performs best, and positive affect is harder to predict than negative affect.

2.3.3.3 Panel data

Our main findings regarding the ML estimation of wellbeing are also robust to exploiting the panel dimension of the German SOEP and the UKHLS. As there is no standard procedure for the introduction of individual fixed effects in the ML algorithms that we use, we implement an approach similar to the Mundlak correction for linear models (Mundlak 1978, Wooldridge 2010). We pool all years of the UKHLS and SOEP data, demean all covariates at the individual level, and include both an individual's average value over time of each covariate as well as their year–

specific deviations from their individual mean. The level of wellbeing is the dependent variable, as was the case in the analysis above.

The relative predictive performance of the OLS and ML in the pooled dataset is similar to the findings for individual years. In the UKHLS, the OLS R–squared is 0.140. The use of RF produces a small improvement, with the R–squared increasing to 0.143. Gradient Boosting provides a further improvement, yielding an R–squared of 0.150. In the German SOEP, the OLS R–Squared is 0.122, with once again both the Random Forest and Gradient Boosting leading to better R–Squared figures of, respectively, 0.150 and 0.156. As shown in Tables A7 and A8, the most important variables predicting the level of wellbeing are almost exclusively the average values of the individual covariates. One exception in both the UKHLS and SOEP is the Health limits activities variable. As such, deviations in individual health status (from their average value) seem to be important for the level of individual wellbeing.

2.4 Discussion

We draw three main conclusions from our analysis above.

First, tree–based ML approaches do indeed perform better at predicting wellbeing than more conventional linear models. Although the gains in R–squared we obtain are modest in absolute terms, they are comparable with – and sometimes exceed – the extent to which information on respondents' health can improve wellbeing predictions. Comparing the algorithms we consider, Gradient Boosting consistently outperforms Random Forests.

Second, when we use all of the non–wellbeing variables that are available in each dataset as predictors, we more than double the explained variation in wellbeing for all of the estimation procedures that we analyze. This extended set of variables produces R–squared figures of around 0.3. These values look to be the maximum achievable with the current survey data.

Third, almost all of the variables that turn out to be important in the specifications using of all the available data relate to health, economic conditions, personality traits, and personal relationships. This purely data–driven process thus picks out the same core determinants of wellbeing as have been identified in the conventional literature. In that sense, ML approaches validate the previous human–guided search for the determinants of wellbeing. This looks to be good news for the field.

We see two directions for future research.

The first is to further explore the capabilities of ML models. We have focused our analysis here on tree–based methods, which are powerful algorithms that perform well in multiple contexts. However, given the specificities of wellbeing data, we might find further improvements by using other algorithms (*e.g.*, Kernel Ridge, Vovk 2013), or by using a combination of theory–based modelling and algorithmic approaches. Another potential approach is using a combination of unsupervised and supervised learning. For example, it might be possible to separate the whole dataset into overlapping clusters of individuals chosen based on subsets of independent variables. Then, the predictive performance of non–linear ML models could be substantially higher when applied to such clusters, as compared to using one global model for the whole dataset as done in our work. Moreover, we have currently only focused on identifying variables that are key for the successful prediction of wellbeing. A natural next step is to extend the use of ML–based algorithms to investigate the variables that are most important for wellbeing in a causal sense (Wager and Athey 2018).

Second, our analyses focused on rich Western countries. As such, it remains an open question whether our findings would also hold in a more global setting, *e.g.* in countries where material needs are much more acute. Insofar as there may be greater scope for improving wellbeing in low– and middle–income countries (Helliwell *et al*. 2022), applying ML approaches in this setting may be particularly valuable going forward.

Appendix

Appendix 1

Table A1. Optimal hyperparameters used in the extended specifications (post–LASSO extended specification in parentheses).

	Panel A: Random Forest		
	SOEP	Gallup	UKHLS
MaxDepth	96 (70)	70 (70)	30 (20)
Nvars	225 (65)	80 (80)	400 (130)
Ntrees	1000 (1000)	1000 (1000)	1000 (1000)
MinLeaf	1 (1)	5 (5)	15 (5)
	Panel B: Gradient Boosting		
	SOEP	Gallup	UKHLS
MaxDepth	8 (8)	3 (3)	5 (7)
Nvars	75 (30)	40 (40)	100 (30)
Ntrees	6000 (2000)	16000 (16000)	2000 (2000)
MinLeaf	1 (1)	1 (1)	1 (1)
Learning rate (λ)	0.005 (0.01)	0.0063 (0.0063)	0.01 (0.01)

*Notes***:** Hyperparameters are identified via a grid search by minimizing the average MSE across 4 folds of cross–validation. MaxDepth is the maximum depth of each branch of each tree. Nvars is the maximum number of randomly–picked variables used to perform splits within each tree. MinLeaf is the minimum number of training individuals that must be in each leaf of a given tree (fixed to 1 for gradient boosting). Ntrees is the number of trees fitted (fixed to 1,000 for random forests). The learning rate (λ) is the rate at which predictions are updated (only applicable to gradient boosting).

Appendix 2

Table A2. List of variables in the restricted set.

Variable	SOEP	UKHLS	Gallup
Age	16 – 105	18 – 103	18 – 99
Area of residence	16 distinct values	12 regions	51 distinct values
BMI	11.10 – 84.50	11.80 – 74.20	7.19 – 152.56

Disability status	Binary	Binary	n.a.
Education	18 – 7 (years of education)	6 distinct values	6 distinct values
Employment status	Binary	12 distinct values	4 distinct values
Equivalised Log HH income	0 – 13.88	–0.80 – 12.52	3.40 – 9.90
Ethnicity/Migration background	3 distinct values (migration background)	18 distinct values (ethnicity)	5 distinct values (ethnicity)
Health	0 – 396 (doctor visits in prev. year)	Health limits activities (3 distinct values)	Binary (self–assessed health problems)
Housing status	4 distinct values	6 distinct values	n.a.
Marital status	5 distinct values	10 distinct values	6 distinct values
Month of interview	12 distinct values	24 distinct values	12 distinct values
Number of children in HH	0 – 11	0 – 9	0 – 15
Number of people in HH	1 – 16	1 – 16	1 – 99
Religion	10 distinct values	Binary	8 distinct values
Sex	Binary	Binary	Binary
Working hours	0 – 6669	0 – 180	4 distinct values

Notes: For continuous variables, the range is reported. For SOEP, possible values for the categorical variables are: *Area of residence*: Each of the 16 Bundesländer. *Ethnicity/Migration background*: No migration background, Direct migration background, Indirect migration background. *House ownership status*: Main Tenant, Sub–Tenant, Owner, Nursing Home/ Retirement Community. *Marital status*: Married, Single, Widowed, Separated, Divorced. *Religion*: Catholic, Protestant, Christian Orthodox, Other Christian, Muslim, Muslim (Shiite), Muslim (Sunnite), Muslim (Alevite), Other, No religion. For UKHLS, possible values for the categorical variables are: *Area of residence*: North East, North West, Yorkshire and the Humber, East Midlands, West Midlands, East of England, London, South East, South West, Wales, Scotland, Northern Ireland. *Education:* Degree, Other higher degree, A–level etc, GCSE etc., Other qualification, No qualification. *Employment status:* Self–employed, Paid employment(ft/pt), Unemployed, Retired, On maternity leave, Family care or home, Full–time student, LT sick or disabled, Govt training scheme, Unpaid, family business, On apprenticeship, Doing something else. *Ethnicity*: British/English/Scottish/Welsh/Northern irish, Irish, Gypsy or Irish traveller, Any other white background, White and black caribbean, White and black african, White and asian, Any other mixed background, Indian, Pakistani, Bangladeshi, Chinese, Any other asian background, Caribbean, African, Any other black background, Arab, Any other ethnic group. *Health*, health limits moderate activities: Yes, a lot; Yes, a little; No, not at all. *House ownership status*: Owned outright, Owned/being bought on mortgage, Shared ownership (part–owned part–rented), Rented, Rent free, Other. *Marital status*: Single and never married/in civil partnership, Married, In a registered same–sex civil partnership, Separated but legally married, Divorced, Widowed, Separated from civil partner, A former civil partner, A surviving civil partner, Living as couple. For Gallup, possible values for the categorical variables are: *Area of residence*: 51 States. *Education*: Less than high school, High school, Technical/Vocational school, Some college, College graduate, Post–graduate. *Employment status*: Employed, Self–employed, Employed and self–employed, not employed. *Ethnicity*: White, Other, Black, Asian, Hispanic. *Marital status*: Single, Married, Separated, Divorced, Widowed, Living with partner (not married). *Religion*: Protestant, Catholic, Jewish, Muslim,

Mormon, Other Christian, Other, No religion. *Working hours*: 30 or more hours per week, 15 to 29 hours per week, 5 to 14 hours per week, less than 5 hours per week.

Appendix 3

Table A3. Permutation Importance (PI) and Pseudo Partial Effects (PPE) in OLS, GB and RF on the Extended Set of variables: the 10 most–important variables.

	OLS			Random forest			Gradient boosting		
	Variable	PI	PPE	Variable	PI	PPE	Variable	PI	PPE
Panel A: SOEP									
1	Health limits daily life: a lot	.029	–.780	Health limits social life	.032	.154	Health limits social life	.022	.172
2	Worry a lot	.025	–.146	Health limits daily life: a lot	.028	–.742	Worry a lot	.021	–.100
3	Health limits social life	.023	.187	Worry a lot	.020	–.113	Health limits daily life: a lot	.019	–.628
4	Personal patience	.011	.129	HH income	.018	.202	Personal patience	.010	.174
5	Health limits daily life: a bit	.009	–.266	Deal well with stress	.015	.160	Deal well with stress	.008	.128
6	Partner in HH	.008	.222	Personal patience	.008	.106	Health limits daily life: a bit	.006	–.220
7	No monthly savings	.008	–.186	No annual holiday trip	.007	–.114	Partner in HH	.006	.152
8	Deal well with stress	.006	.080	No monthly savings	.007	–.110	Risk tolerance	.006	.036
9	House needs repair	.005	–.126	Not unemployed	.006	.303	HH income	.006	.152
10	Hours of sleep on workday	.004	.077	Unemployment benefit	.005	–000	Number of doctor visits	.006	–.086
Panel B: UKHLS									
1	Regret getting married	.032	.418	Worries a lot (Big 5)	.030	–.146	Worries a lot (Big 5)	.033	–.188

2	Worries a lot (Big 5)	.029	–.274	Feeling relaxed (Big 5)	.027	.238	Feeling relaxed (Big 5)	.019	.212
3	Feeling relaxed (Big 5)	.016	.240	Health limits kind of work	.009	.040	Regret getting married	.011	.209
4	Kiss partner	.012	–.218	Belong to neighbourhood	.009	–.179	Does a thorough job (Big5)	.008	.069
5	Does thorough job (Big 5)	.006	.112	Age squared	.009	.007	Kiss partner	.007	–.110
6	Share interests w. partner	.006	–.161	Regret getting married	.009	.137	Age squared	.007	.002
7	Belong to neighbourhood	.005	–.107	Health limits work amount	.008	.032	Health limits kind of work	.007	.053
8	Sociable (Big 5)	.005	.094	Does thorough job (Big 5)	.007	.053	Health limits work amount	.006	.049
9	Health limits work amount	.005	.070	Consider divorce (never)	.006	.106	Belong to neighbourhood	.006	–.162
10	Long term sick or disabled	.005	–.420	Sociable (Big 5)	.006	.081	Sociable (Big 5)	.006	.126
Panel C: Gallup									
1	Learn something every day	.031	.43	Learn something every day	.033	.34	Learn something every day	.028	.35
2	City/area is perfect	.021	.32	City/area is perfect	.026	.42	City/area is perfect	.021	.39
3	Log HH income	.013	.15	Log HH income	.021	.30	Log HH income	.018	.26
4	Economy in this country	.013	.21	Cannot afford healthcare	.021	–.54	Health index	.015	.16
5	Cannot afford healthcare	.010	–.38	Economy in this country	.015	.21	Economy in this country	.015	.22
6	Health limits activities	.010	–.04	Physical health index	.013	.15	Cannot afford healthcare	.013	–.40

7	Health encouragement	.010	.12	Health limits activities	.010	–.03	Health encouragement	.008	.17
8	Physical health index	.010	.14	Health encouragement	.010	.17	Health limits activities	.008	–.01
9	Female	.008	.24	Female	.005	.13	Age and age–squared	.005	.03
10	Ever diag. w depression	.008	–.28	Ever diag. w. depression	.005	–.16	Female	.005	.25

Notes: The following variables are shown. SOEP: Dummies: *Health limits daily life a lot*, *Health limits daily life a bit*, *Partner in HH*, *No monthly savings*, *Not unemployed*, *No emergency reserves*, and *No annual holiday trip*. Likert scales: *Limited socially due to health* (1 – always to 5 – never), *Worries a lot* and *Deals well with stress* (1 – not at all to 7 – totally agree), *Personal patience* (0 – very bad to 10 – very good), *House needs repair* (1 – in good condition, 3 – needs major renovation). Continuous: *Log HH income*, *Hours of sleep*, *Number of Doctor visits*, *Risk Tolerance and Unemployment Benefit.*

UKHLS: Dummies: *Health not limiting activities.* Likert scales: *Pain interferes with work* (1 – not at all to 5 – extremely), *Regret getting married*, *Share interests w. partner*, *Consider divorce* and *Kiss partner* (1 – all the time, 6 – never), *Health limits work amount* and *Health limits kind of work* (1 – all of the time, 5 – none of the time); Big 5 traits, including *Worries a lot*, *Feeling relaxed*, *Does thorough job*, *Is sociable* (1 – does not apply to 7 – applies perfectly), *Belong to neighbourhood* (1 – strongly agree – 5 strongly disagree). Continuous: *Age squared*. Gallup: Dummies: *Cannot afford healthcare*, *Female*, *Ever diagnosed with depression*. Likert scales: *Learn something every day*, *City/area is perfect* and receives *Health encouragement* (1 – strongly disagree, 5 – strongly agree), *Economy in this country* (1 – poor to 4 – Excellent), *Health limits activities in the last month (0 to 30 days)*. Continuous: *Age*, *age squared*, *Log HH income*, *Physical health index*.

Table A4. Permutation Importance (PI) and Pseudo Partial Effect (PPE) in OLS, GB and RF on the Restricted Set of variables: the 10 most–important variables.

	OLS			Random forest			Gradient boosting		
	Variable name	PI	PPE	Variable name	PI	PPE	Variable name	PI	PPE
Panel A: SOEP									
1	Age and age–squared	0.10	–1.70	Adjusted Income	0.13	0.27	Adjusted Income	0.14	0.46
2	Adjusted Income	0.10	0.26	Age and age–squared	0.12	–0.14	Age and age–squared	0.13	–0.18
3	Number of doctor visits	0.08	–0.14	Number of doctor visits	0.11	–0.28	Number of doctor visits	0.12	–0.63

4	Marital Status – Single	0.07	–0.40	Disability Status	0.04	–0.40	Disability Status	0.03	–0.45
5	N of children in HH	0.06	0.30	N of children in HH	0.03	0.07	Working hours	0.02	–0.29
6	Disability Status	0.04	–0.52	N of people in HH	0.03	0.02	N of years of education	0.02	0.17
7	N of people in HH	0.03	–0.17	N of years of education	0.02	0.07	N of children in the HH	0.02	0.08
8	N of years of education	0.03	0.11	House Ownership: Owner	0.02	0.12	N of people in HH	0.02	–0.16
9	Marital Status – Divorced	0.02	–0.38	Working hours	0.01	0.04	Marital Status – Single	0.02	–0.19
10	Marital Status – Separated	0.02	–0.74	BMI	0.01	–0.02	Marital Status – Separated	0.01	–0.53
Panel B: UKHLS									
1	Health limits activities: a lot	.024	–.670	Age	.040	.052	LT sick or disabled (empl.)	.018	–.587
2	Single	.020	–.336	HH income	.015	.161	Age	.015	.052
3	LT sick or disabled (empl.)	.017	–.797	Health limits activities: a lot	.014	–.377	Health limits activities: a lot	.012	–.377
4	Age	.018	.015	Not disabled (health)	.014	.215	Not disabled (health)	.010	.215
5	Health limits activities: a bit	.014	–.327	Health limits activities: a bit	.012	–.226	Renting house	.007	–.106
6	Not disabled (health)	.011	.240	LT sick or disabled (empl.)	.011	–.587	Health limits activities: a bit	.007	–.226
7	Retired	.010	.235	Unemployed	.006	–.193	HH income	.006	.161
8	Renting house	.008	–.208	Renting house	.005	–.106	Unemployed	.006	–.193
9	Unemployed	.008	–.343	Single	.005	–.136	Retired	.005	.099
10	HH income	.008	.083	Retired	.003	.099	Single	.003	–.136
Panel C: Gallup									
1	Health limits activities	.064	.84	HH income	.062	.48	HH income	.067	.48
2	HH income	.049	.30	Health limits activities	.057	.69	Health limits activities	.054	.71

3	Post–graduate education	.026	.58	Age and age–squared	.046	.43	Age and age–squared	.041	.44
4	Married	.013	.33	Married	.013	.26	Married	.013	.27
5	College Graduate	.010	.37	Female	.010	.23	Female	.013	.29
6	Female	.010	.29	Post–graduate education	.008	.43	Post–graduate education	.008	.34
7	Age and age–squared	.008	.24	Body Mass Index	.005	.29	Body Mass Index	.005	–.12
8	Hispanic	.003	.28	Working Hours Missing	.005	–.12	Hispanic	.003	.15
9	Atheist	.003	–.19	Hispanic	.003	.06	Black	.003	.10
10	High school graduate	.003	.17	Asian	.003	.02	Working Hours Missing	.003	–.06

Notes: The total list of variables in the Restricted Set appears in Table A1.

Table A5. Permutation Importance (PI) and Pseudo Partial Effect (PPE) in OLS, GB and RF for positive and negative affect: the top 10 most–important variables (using 2013 Gallup data with the Extended Set of variables).

	OLS			Random forest			Gradient boosting		
	Variable	PI	PPE	Variable	PI	PPE	Variable	PI	PPE
Panel A: Positive affect									
1	Age	0.14	–0.26	Physical health index	0.07	0.42	Physical health index	0.16	0.62
2	Age squared	0.09	–0.26	Learn something every day	0.06	0.43	Learn something every day	0.05	0.49
3	Physical health index	0.09	0.66	Not treated with respect	0.03	–1.39	Not treated with respect	0.03	–1.13
4	Learn something every day	0.05	0.82	Health encouragement	0.02	0.13	Health encouragement	0.02	0.14
5	Not treated with respect	0.03	–1.52	Diagnosed w. depression	0.01	0.27	BMI	0.01	0.02
6	Health encouragement	0.02	0.23	City/area is perfect	0.00	0.17	Diagnosed w. depression	0.01	0.34

7	In workforce	0.01	0.44	Health limits activities	0.00	–0.01	Has any health problems	0.01	–0.26
8	Diagnosed w. depression	0.01	0.52	BMI	0.00	0.09	City/area is perfect	0.00	0.17
9	Not working	0.00	–0.32	Age squared	0.00	–0.11	Health limits activities	0.00	0.21
10	Tuesday	0.00	–0.33	Age	0.00	–0.11	Female	0.00	0.20
Panel B: Negative affect									
1	Physical health index	0.26	–0.11	Physical health index	0.31	–0.15	Physical health index	0.50	–0.18
2	Not treated with respect	0.03	0.16	Not treated with respect	0.04	0.17	BMI	0.04	–0.02
3	Diagnosed w. depression	0.02	–0.09	BMI	0.03	–0.01	Not treated with respect	0.03	0.15
4	Age squared	0.01	–0.03	Diagnosed w. depression	0.02	–0.07	Has any health problems	0.02	0.06
5	BMI	0.01	–0.03	Health limits activities	0.01	–0.02	Diagnosed w. depression	0.02	–0.07
6	Has any health problems	0.01	0.04	Has any health problems	0.01	0.02	Health limits activities	0.02	–0.06
7	Cannot afford healthcare	0.01	–0.05	Cannot afford healthcare	0.01	–0.04	Had a cold yesterday	0.01	0.07
8	Wednesday	0.00	0.05	City/area is perfect	0.00	–0.02	Cannot afford healthcare	0.01	–0.04
9	Neck or backpain	0.00	–0.03	Neck or backpain	0.00	–0.02	Headache yesterday	0.00	0.02
10	Time Zone E	0.00	0.03	Age	0.00	–0.04	City/area is perfect	0.00	–0.02

Notes: The following variables are shown.: Dummies: Health limits daily life a lot, Health limits daily life a bit, Partner in HH, No monthly savings, Not unemployed, No emergency reserves, Last word in financial decisions–NA, Psychiatric problems, Female, and No annual holiday trip. Likert scales: Limited socially due to health (1 – always to 5 – never), Worries a lot, Importance: To help others (1 – Very Important to 4 – Not important), Deals well with stress (1 – not at all to 7 – totally agree), Personal patience (0 – very bad to 10 – very good), House needs repair (1 – in good condition, 3 – needs major renovation), Attend cinema/concerts (1 – Daily to 4 – Infrequent), Am Sociable (1 to 7), Visit neighbours/friends (1 – Daily to 5 – Never), Use of social networks (1 – Daily to 5 – Never), Health affects tiring tasks (1 – A lot to 3 – Not at all), and Physical pain last 4 weeks (1 – Always to 5 – Never). Continuous: Log HH income, Hours

of sleep, Number of doctor visits, Risk tolerance, Unemployment benefit, Excursions/short trips, Number of close friends, Hours of childcare per day, Annual pension.

Table A6. Permutation Importance (PI) of OLS, GB and RF for levels of wellbeing of the 10 most–important variables (using pooled UKHLS data with the Restricted Set of variables). For each covariate, the models include the average value and the annual deviation from that average.

OLS		Random forest		Gradient boosting	
Variable name	PI	Variable name	PI	Variable name	PI
Health limits activities: a lot (avg.)	.041	Age (avg.)	.025	Age (avg.)	.026
Not disabled (health) (avg.)	.020	Not disabled (health) (avg.)	.020	Not disabled (health) (avg.)	.022
Married (avg.)	.019	Health limits activities: a lot (avg.)	.018	Health limits activities: a lot (avg.)	.021
Health limits activities: a bit (avg.)	.017	Health limits activities: a bit (avg.)	.014	Health limits activities: a bit (avg.)	.014
LT sick or disabled (empl.) (avg.)	.015	LT sick or disabled (empl.) (avg.)	.011	HH income (avg.)	.012
Age (avg.)	.013	HH income (avg.)	.009	LT sick or disabled (empl.) (avg.)	.012
Retired (avg.)	.012	Married (avg.)	.006	Married (avg.)	.009
HH income (avg.)	.010	Retired (avg.)	.005	Retired (avg.)	.006
Unemployed (avg.)	.007	Unemployed (avg.)	.004	Unemployed (avg.)	.005
Rents the house/flat	.005	Health limits activities: a bit	.003	Health limits activities: a lot	.004

Note: All covariates apart from month, ethnicity and sex are split into individual means and deviation from the mean. Individual averages are denoted by *(avg.)*; variables without additional notes are the deviations from the individual means.

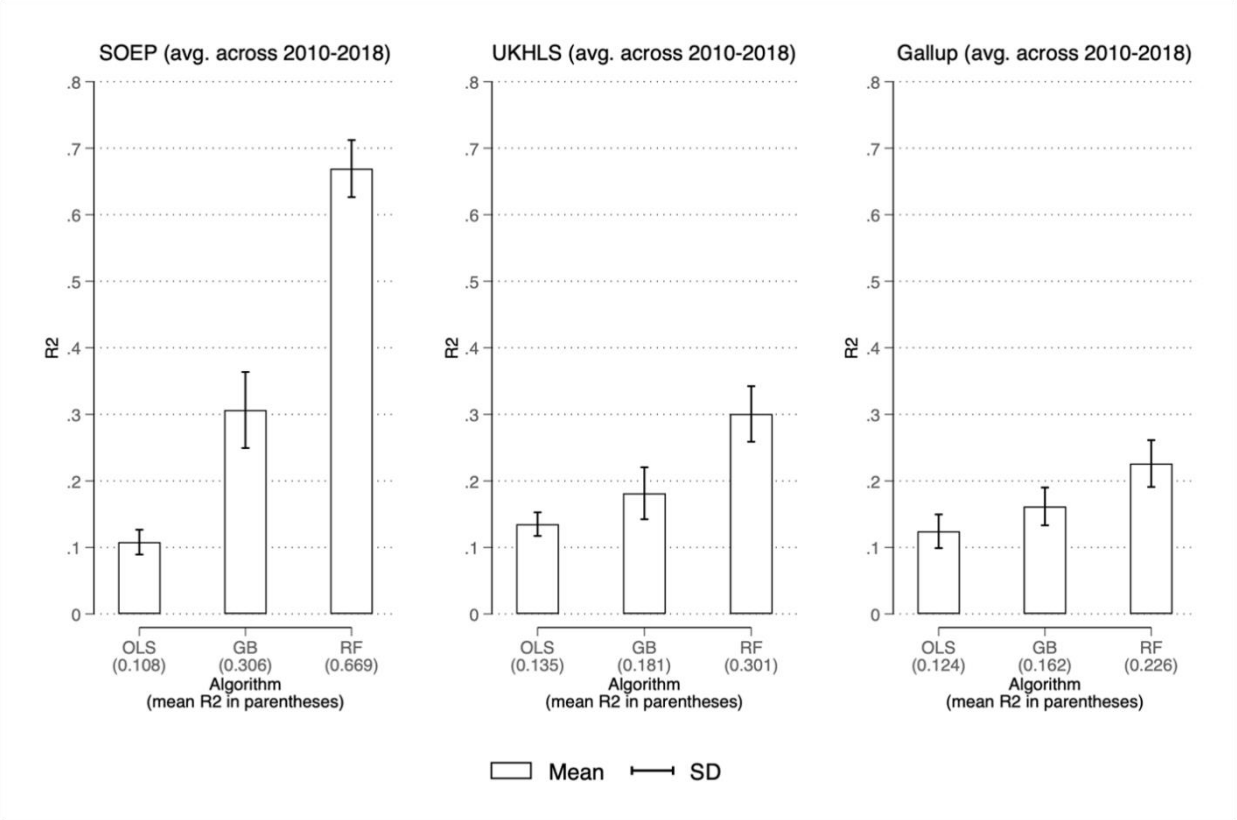
Table A7. Permutation Importance (PI) of OLS, GB and RF for deviations from the average wellbeing and individual level of wellbeing of the 10 most–important variables (using pooled SOEP data with the Restricted Set of variables).

OLS		Random forest		Gradient boosting	
Variable name	PI	Variable name	PI	Variable name	PI
Age (avg.)	.082	Age (avg.)	.126	Age (avg.)	.124
Number of doctor visits (avg.)	.039	Adjusted Income (avg.)	.059	Adjusted Income (avg.)	.049
Adjusted Income (avg.)	.039	Number of doctor visits (avg.)	.041	Number of doctor visits (avg.)	.042
N. of children in the hh (avg.)	.025	Not disabled (health) (avg.)	.021	Not disabled (health) (avg.)	.016
Not disabled (health) (avg.)	.016	N. of people in hh (avg)	.014	Age	.010
Single (avg.)	.016	N. of children in hh (avg.)	.011	N. of people in hh (avg.)	.009
Divorced (avg.)	.007	House Owner	.009	N. of children in hh (avg.)	.008
N. of people in hh (avg.)	.006	Age	.008	Number of doctor visits	.007
Number of doctor visits	.005	Number of doctor visits	.005	Single	.006
House Owner	.005	Number of years of education	.005	House Owner	.006

Notes: All covariates apart from month, ethnicity and sex are split into individual means and deviation from the mean. Individual averages are denoted by *(avg.)*; variables without additional notes are the deviations from the individual means. For each covariate, the models include the average value and the annual deviation from that average.

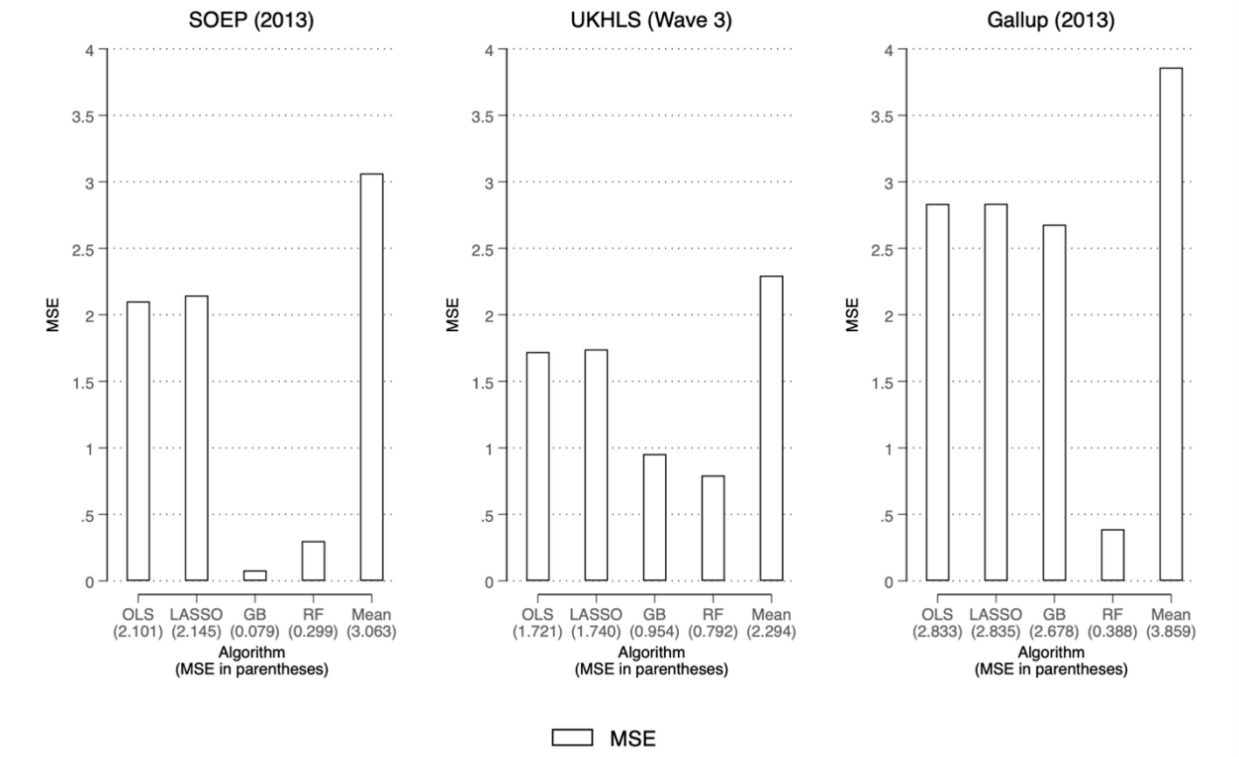
Appendix 4

Figure A1: The R–squared from OLS, GB and RF on the Restricted Set of variables on the training set



Notes: The R–squareds are calculated from the training data and are not representative of out–of–sample performance.

Figure A2: The R–squareds from OLS, LASSO, GB, RF, and mean on the Extended Set of variables on the training set

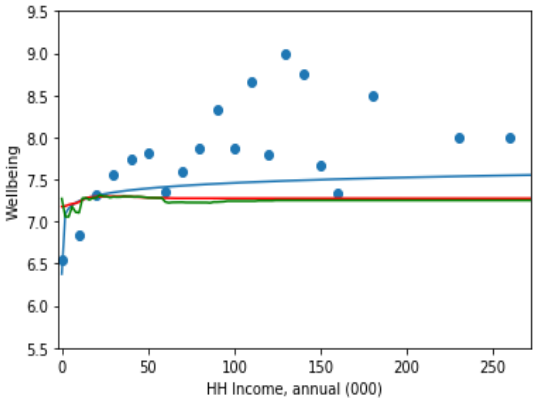


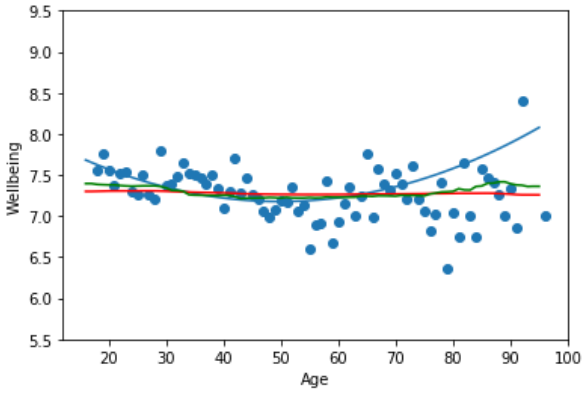
Notes: The R–squareds are calculated from the training data and are not representative of out–of–sample performance.

Appendix 5

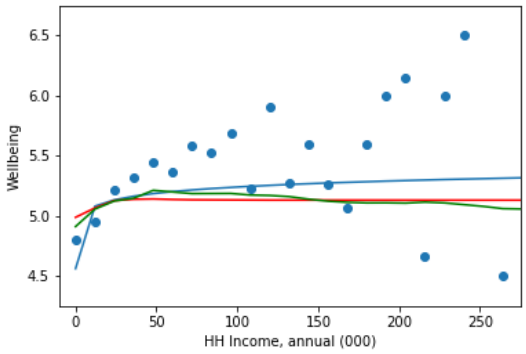
Figure A3. Mean effects of age and household income on wellbeing in the Extended Set of variables.

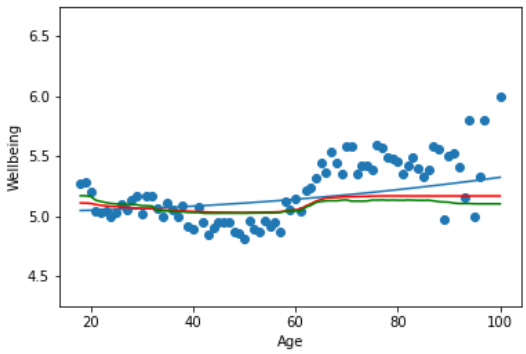
Panel A: SOEP



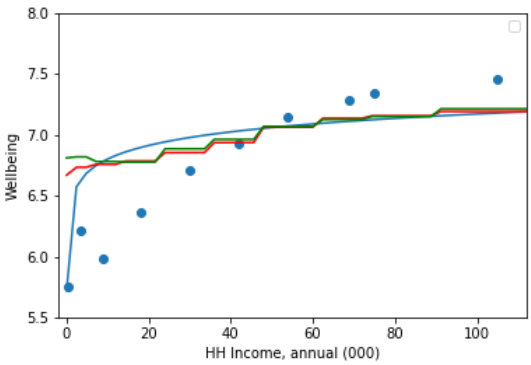


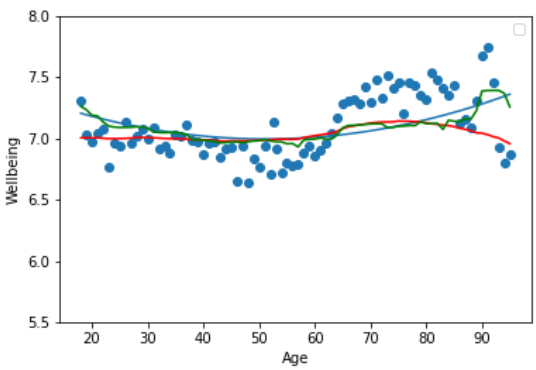
Panel B: UKHLS





Panel C: Gallup

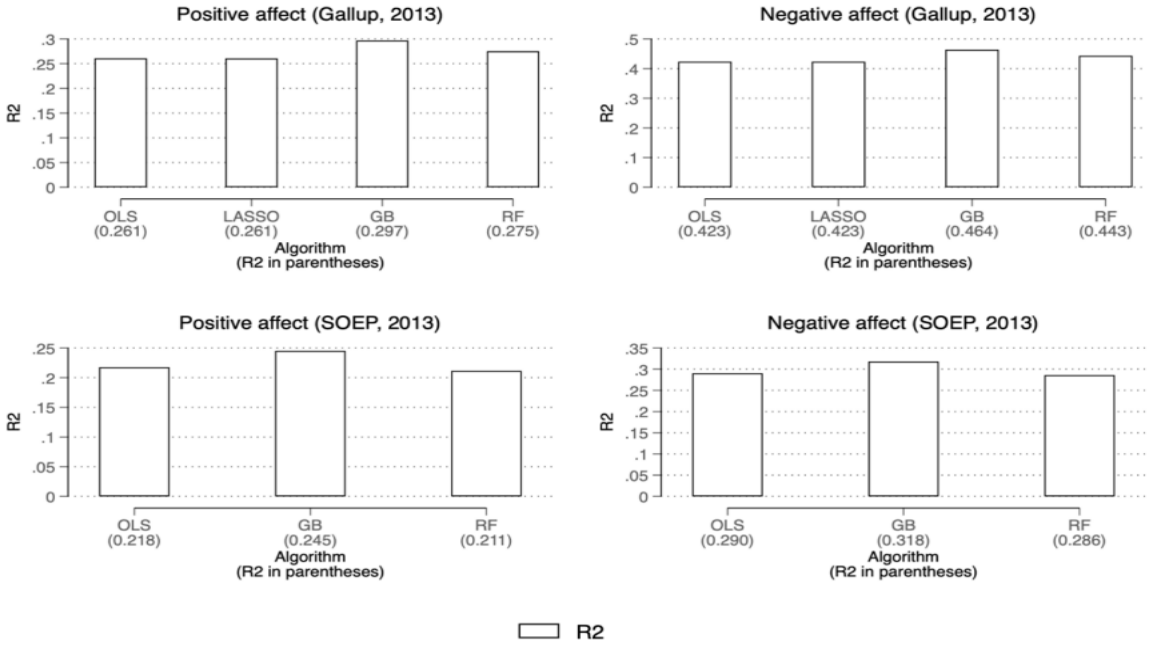




Notes: For the UKHLS and the SOEP annual income is constrained to be less than or equal to a figure of 250 000 in the local currency. This covers over 99.9% of the income distribution in both countries. In SOEP and UKHLS, incomes are recorded as a continuous variable and equivalence–scale adjusted HH income is used for the analysis. Income data in Gallup is collected in income bands, and household size data was not collected in 2013. We here thus use non–adjusted HH income data.

Appendix 6

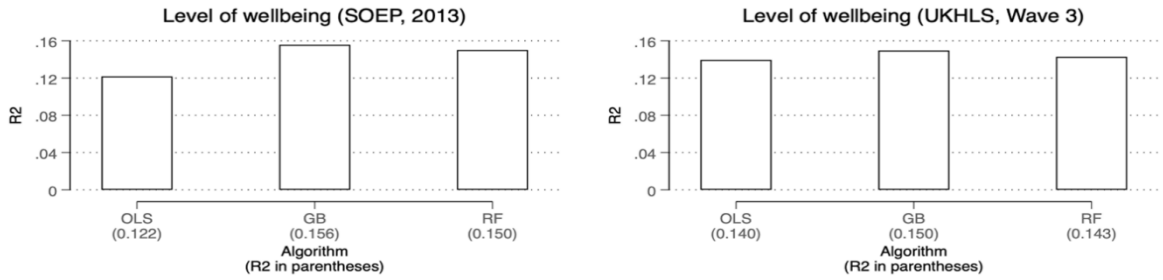
Figure A4: The R–squared from OLS, LASSO, GB and RF when positive and negative affect using 2013 Gallup and 2013 SOEP data with the Extended Set of variables



Notes: The R–squareds are calculated from unseen 'testing data'.

Appendix 7

Figure A5: The R–squared from OLS, LASSO, GB and RF when modelling the level of wellbeing with Mundlak terms using 2013 SOEP and Wave 3 UKHLS data with the Restricted Set of variables



Notes: The R–squareds are calculated from unseen 'testing data'.

Chapter 3

Healthcare utilization and its evolution during the years: building a predictive and interpretable model

3.1 Introduction

The field of study now called health economics is considered to be born following Arrow's 1963 seminal paper *"Uncertainty and the Welfare Economics of Medical Care"* – Arrow (1963). As one of the founding fathers of the study of the implications of uncertainty and imperfect information, in his paper Arrow studied their role in the market of medical care. The conclusion of his work is that these markets fail to meet the conditions necessary to reach the socially desirable equilibrium. Fundamental issue in such failure is the role of moral hazard, both in doctors and patients. Andersen (1968) describes the role of variables that predispose and enable utilization. In particular, variables that predispose utilization include gender, age and household composition, while variables enabling utilization include income and education. Grossman (1972) is among the first to both build a theoretical and empirical model to predict demand for medical services. His key findings are that the demand for medical services increases with age and wage rate, while it is negative in education as long as it leads to the creation of health capital. Seminal paper in health economics is also Manning *et al.* (1987): in their work, Manning and colleagues ran a randomized experiment to empirically address the endogeneity of the demand for medical care. Their key finding is that, counterintuitively, in postwar US the diffusion of health insurance causally explained only a small fraction of the increase in the demand for medical care, in the order of one–tenth. Their possible explanation is that instead the increase in demand is mainly due to the development of new treatments and technologies that allow to treat patients that otherwise would have simply died (they cite for instance the case of renal dialysis and transplantation, developed in 1950).

Overall, beside direct prediction of specific conditions, the current research in health outcomes and economics has focused on (self–reported) health status, mortality, and healthcare utilization. In this work, we focus on the topic of healthcare utilization. From a health economics perspective, modelling health care utilization provides better insights into the inequalities (differences in access/usage of health care services) and inequities (differences in access/usage of health care services, which are not necessarily driven by corresponding – justifiable – "needs") that may arise from socio–economic heterogeneity with respect to access and usage of health care systems. For the prediction and explanation of healthcare utilization, we focus on the application of *Supervised* and *Unsupervised Machine Learning* techniques.

With *Supervised Learning*, we refer to the set of algorithms aimed at predicting a dependent variable *y* as function of a set of independent variables *X*. Since the relationship is not specified by the researcher, it is *learned* by the algorithms based on the provided data. In the case of a Linear Regression, for instance, the aim is to learn the best value of the coefficients. Instead, in the case of a Regression Tree – Breiman (1984) – the aim is to learn the best set of sequential splits in the variables' space, in order to produce a prediction in each final node (Regression Trees will be described in Section 3). We consider these two Supervised Learning algorithms in this work.

With *Unsupervised Learning*, instead, we refer to the set of algorithms aimed at finding patterns in the data, without predicting a particular outcome. With these algorithms, we either focus on reducing the dimensionality of the variables' space (for instance, considering methods like *Principal Component Analysis*), or we aim at finding relevant *clusters* of individuals in the data (for instance, with methods like *K–Means*). Principal Component Analysis was used to create two Physiological and Psychological Health scales, already available in the SOEP dataset. For an in–depth description of Supervised and Unsupervised Machine Learning techniques, we remind the reader to *"The Elements of Statistical Learning"*, Hastie *et al.* (2009).

Also, we focus on finding clusters of individuals, using the aforementioned K–Means. An alternative possibility would consist in clustering as in standard Econometric approach. For instance, Abadie *et al.* (2022) estimate a log–linear regression of earnings on an indicator function for some college. Their finding is that the clustered standard errors are 20 times larger, robust

standard error. For a broader discussion about clustering in Econometrics, a key reference is Wooldridge (2003), *"Cluster–sample methods in applied Econometrics"*.

In this work, we have instead decided to cluster in an unsupervised manner, *i.e.*, letting the algorithm choose at which level to perform the clustering. This is a novel approach, potentially leading to clustering on variables we would have otherwise not think of. A detailed analysis of what are the characteristics of each of the found clusters is available in Appendix 6.

In each of them, then, we investigate whether Random Forest can indeed yield increases in predictive accuracy over the Linear Regression, meaning that eventual nonlinearities become easier to model in these subgroups. Throughout the chapter, we will refer to the analysis in clusters as at the "*local"* level (as opposed to the analysis on the entire dataset, at the *"global"* level).

To summarize, the aim of this chapter is to reply to the following *three* research questions:

- RQ1: Can Machine Learning algorithms allow us to predict objective health outcomes more efficiently than traditional linear models?
- RQ2: Are Unsupervised Learning algorithms identifying clusters in the data we would have not thought of otherwise, and useful for predictions?
- RQ3: Which variables are most important in predicting healthcare utilization?

Associated with RQ1 there is also the question of whether objective variables are easier to predict than subjective ones. Since no large absolute improvements were observed considering ML to predict subjective well–being, considering an objective measure of health as dependent variable, and observing ML algorithms outperforming linear methods, may be an indication that objective measures are indeed better suited for this kind of algorithms. Explanation for this would be a smaller *Irreducible Error* – Hastie *et al.* (2009) – *i.e.* the variance of the error term, encompassing the role of unobserved predictors and measurement errors.

The rest of the chapter is structured as follows: in Section 2, we describe the German Socio–Economic Panel, going in detail of the considered dependent and independent variables and the associated *feature engineering* process. In Section 3, we describe the two considered specifications (*Pooled* and *Transformed Pooled*), as well as all the implemented Machine Learning algorithms. In Section 4, we present the predictive analytics' results, and in Section 5 the analysis of the determinants of healthcare utilization. Section 6 concludes.

3.2 Data

We make use of the German Socio–Economic Panel (SOEP) longitudinal survey data. This dataset contains information on approximately 11,000 private households between 1984 (1990 for former GDR) and 2018. This panel dataset contains a lot of information on socio–economic variables of individuals and households, as well as health information (use of healthcare). This makes the longitudinal survey data very useful to test our models.

There exist a rich literature using the SOEP focusing on health outcomes. Moor *et al.* (2018) found consistent educational inequalities in self–rated health and health–related quality of life across the period 1994–2014. Leopold (2019) investigates the hypothesis of increasing educational differences in health in age, comparing the results considering both subjective, semi–subjective and objective measures of health. Her conclusion is that the hypothesis holds for men considering subjective and semi–subjective measures, but not the objective ones. Opposite findings held for women. This work represents an interesting indication that indeed the findings in empirical health economic research may be sensitive to the nature of the considered dependent variable. Schmitz (2011), in his Ph.D dissertation, uses the SOEP across all the chapters to study the role of inefficiencies in the German healthcare sector. Before moving onto describing the different algorithms, their rationale and the results, we present and discuss the considered dependent and independent variables.

3.2.1 Dependent variables

Aim of this work is to predict and explain the determinants of Healthcare Utilization.
Investigating the SOEP dataset, we came up with two different measures capturing it:

- *Number of doctor visits.* A first dependent variable is the frequency of doctor visits. This variable has often been used to model the use of healthcare services. For instance, Ygzaw *et al.* (2020) considered the number of visits to a physician as dependent variable – on Norwegian data – to assess whether (and how) it correlates with health–related researches on Internet (finding a positive association).

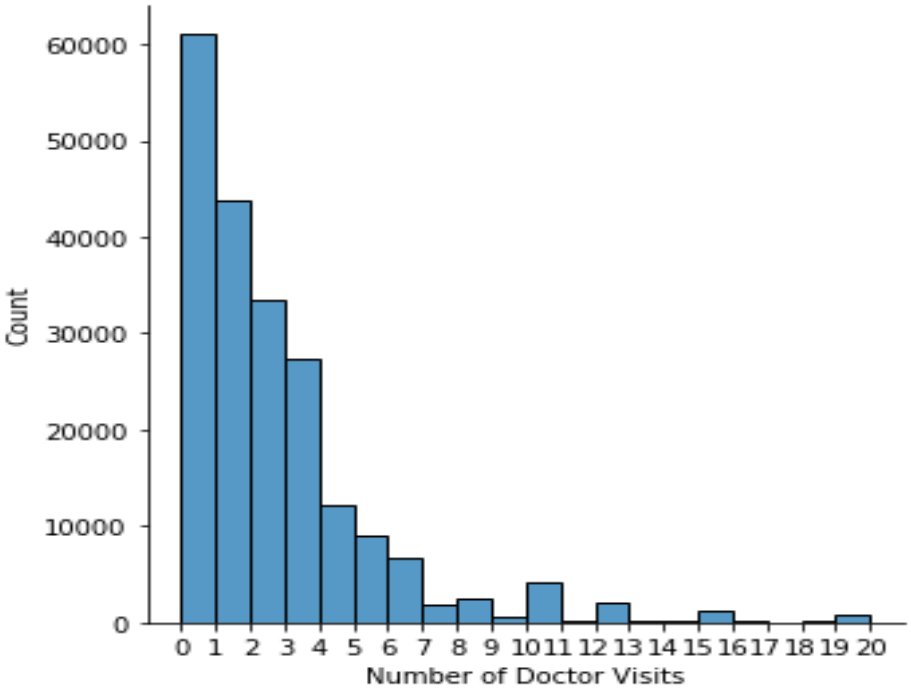
  In the SOEP in particular, Number of doctor visits was considered by the

aforementioned Schmitz (2011) precisely as a measure of healthcare utilization. The SOEP data contains the Number of doctor visits of each individual in the past 3 months. Yet, it has not been consistently asked before 1995 (no records in 1990 and 1993 and more than 50% missing data before 1988 and 1994). However, it has well over 99% completion rate since 1995 (except for 2013 and 2016–2018). Therefore, we decided to use data starting from 2000 onwards.

- *Number of nights spent in hospital*. An alternative dependent variable could be the number of nights spent in hospital during the last year.

It needs to be noted, though, that the Number of nights spent in hospital elicits a higher degree of healthcare need (you only spend a night in hospital for relatively more severe injuries and health problems). In this sense, we expect a much higher degree of zero–observations.
Indeed, we observe that about 80% of the full dataset (682,447 observations) are zero (no overnight hospital stays). For this reason, we decided to solely focus on Number of doctor visits in the last three months as dependent variable. In this case, 99.5% of the observations had less than 20 visits. We anyway included also the outliers (including a maximum of 99 observations). The rationale behind this choice is that being able to predict who is more in need of healthcare services is as important (if not more) than doing it for the general population.

Fig. 1 Distribution of Number of Doctor Visits in the last three months



Notes: Distribution capped at maximum 19 visits (only 0.5% of the individuals have gone to the doctor in the last three months at least 20 times).

3.2.2 Independent variables

Several characteristics are considered to be important predictors of healthcare usage. These characteristics can either reside on the personal and the family level. We categorize all variables in two important groups: *need–based* and *non–need–based* variables. Need–based variables are characteristics which predict valid/justifiable use of healthcare services; for example, self–rated health scores, objective health measures (physical as well as psychological), unhealthy lifestyle, habits and behaviors such as smoking and alcohol consumption, as well as conditions like obesity. Non–need–based features are predictors to the access/opportunity to use healthcare services, beyond need–based motives. These variables are related to socio–economic status. For example, these include marital status, educational level, employment status and disposable income, as well as having a paid health insurance. In all models, we control for age, which can be argued to be a need–based predictor (old age often leads to deterioration of health status and more frequent use

of health care). However, it can also be argued to be non–need–based: older people can be argued to have more time (higher opportunity), be wealthier (higher spending–power), and/or use health care for non–illness–related reasons. We also control for gender, for which can be argued that females are, on average, expected to more frequently visit doctors. Owens (2008) for instance finds this discrepancy to be particularly large for women between 45 and 64, due to the development of chronic conditions associated with the menopause. From the SOEP databases, we extract the following independent variables:

3.2.2.1 Need–based Independent Variables

- *Self–Rated Health status***.** This variable is rated on a 5–point Likert scale, ranging from Very Good[1] to Bad[5]. We reverse the order of the categories to express health, rather than the lack of it. This variable contained about 15% missing values in the full dataset. We imputed the data using a *flexible time–trend approach* (more details about this approach are provided in Appendix 4).
- *Disability status*. This variable expresses the degree of hindrance/legal handicap which reduces the individual's ability to work/be employed. This variable ranges between 0% and 100%. This variable is not asked every year for each individual. Hence, we need to impute about 10% of the values of this variable. We use a logical imputation approach: we assume that disability status does not decrease easily. Thus, we make sure that this variable does not decline over time for each individual. We compute the lag (one year earlier) and lead (one year in the future) of each data point (missing values included). We impute the missing data point using the arithmetic mean of both boundary observations. In this way, we implicitly assume that the disability percentage gradually increases over time until it reaches the future value.
- *Smoking habits*. This variable is originally measured by the (current) number of cigarettes, pipes and cigars smoked on a daily basis: it is included in the SOEP questionnaire on a biannual basis, starting in 2002. We winsorize the original variable at a 99% level (*i.e.*, 36 items per day). In this way, we avoid unrealistically high numbers (values up to 236 were observed). This variable is however rich of missing information (about 75% of the total dataset and 61% of the dataset for when this variable would be deemed available). As

such, we adopt *feature engineering* approach which circumvents missing information and still retains some information on the smoking behavior of individuals, based on whether the individuals have ever smoked. Hence, we have two variables associated with the smoking habits:

"Moving Average Smoke": in it, missing values were imputed across years, considering the moving average of smoked cigarettes in those years in which smoking occurred to impute missing values.

"*Whether Ever Smoked":* binary variable for whether the individual has ever smoked.

- *Body Mass Index*. This variable is computed as weight in kilogram divided by the square of height in meter. Weight and height data are included in the SOEP questionnaire on a biannual basis, starting in 2002. We winsorize the resulting BMI values at the 99.9% level (54.4) to remove extreme observations. Due to the SOEP questionnaire design, we face a lot of missing observations in the data (64% of the data since 2002 is missing). Given the reasonable stability of height over time (for adult individuals), the only change of variation is weight. As such, imputation of BMI would be equivalent to changes in weight. Also in this case, to impute missing values, we considered the flexible time–trend approach considered for Self–Rated Health.
- *Physical and Psychological health scales*. This is a set of specific, *objective and subjective* health–related scales. The measures are based on Norm–based Scoring (NBS) using the SF–12 (Short–Form (12 items)) Health Survey of Ware *et al.* (1996).

  The SOEP survey included these measures biannually from 2002 onwards.

  There are 8 dimensions of health outcomes, namely: Physical Fitness, Role–Physical, Bodily Pain, General Health, Vitality, Social Functioning, Role–Emotional, and Mental Health – Andersen *et al.* (2007). Two composite scales were made using the 8 subscales above and made available in the SOEP. Principal component analysis with varimax rotation is used, which results in two orthogonal (uncorrelated) composite variable: *Psychological Scale* and *Physical Scale.* Also in this case, missing values were imputed using the flexible time–trend approach based on Generalized Additive Models described in Appendix 4.

3.2.2.2 Non–Need based Independent Variables

- *Type of health insurance*. This categorical covariate records the type of health insurance to which the individual is subjected. This categorical covariate is categorized as:
    - 0: no insurance (which was a valid response–option)
    - 1: compulsory/statutory insurance
    - 2: private insurance (exclusively)
    - 3: compulsory/statutory insurance augmented with private insurance.

  We decided to not transform this variable in dummies, given its ordinal natural. Having no insurance guarantees less coverage than having a statutory one, in turn leading to less coverage than a private one, and this one to less coverage than a combination of the two. Missing values were imputed using the mode, compulsory/statutory insurance.

- *Marital status.* Categorical covariate which records the family situation of each individual. We recategorize the categorical variable into a binary indicator, discerning whether the individual is alone or not:
    - Alone [1] if "Married, but separated", "Single", "Divorced", "Widowed", "Partner abroad", "Legally cohabiting, but not living together".
    - Else [0]: "Married" or "Legally cohabiting and living together".

  The choice of recategorize Marital Status in a binary for loneliness is in contrast with for instance Gentile *et al.* (2022) – first chapter – where instead the possibility to consider all the different Marital Statuses allowed, via Shapley Values, to investigate the impact of all the different statuses on the self–assessed life satisfaction. In this case, however, under the assumption that being Not Alone may represent a non–need based condition positively associated with opportunity of access to healthcare, the specific marital status becomes irrelevant. That is, we assume that whether you are married or simply cohabiting, there is someone that else may take care of your home and eventual kids while you're visiting a doctor. On the contrary, whether you are separated, single, divorced or widowed, such possibility would not be there, hence making all these different statuses

equivalent given our aim. Missing values were imputed as 0 (Married or cohabiting), the most frequent value.

- *Employment status.* This categorical covariate records the employment status of the individuals as close as possible.

  – FT: full–time employed

  – PT: part–time employed

  – VTraining: vocational training (e.g., education, unpaid internship, …)

  – NO: unemployed

  – Pension: retired

  Missing values in Employment Status (less than 10 individuals) were dropped as part of dropping missing information of Number of doctor visits and Household income.

- *Household income (per capita).* This variable includes the Household's Post–Government Income after taxes and government transfers of all individuals in the household. This measure includes: labor income, income from assets, retirement income, unemployment benefits, and alimony, minus taxes. As such, this measure reflects disposable income for the household (before rent and fixed costs). We winsorize the household income at a 99.5% level (150,568 euro) to truncate extreme values in the distribution. This truncation will downplay the possible outlying impact of very few observations earning more than this amount. We also compute household income per capita, by dividing the household income by the number of household members (irrespective of their role/status in the household – *e.g.*, including children). We winsorize this variable separately at a 99.5% level (55,865 EUR). As this variable is expected to be the most important non–need based characteristic, missing values were simply dropped.

- *Educational level.* This categorical covariate records the largest educational level ever obtained by the individual. We use two variables, due to the change in the ISCED–nomenclature over time (in 2011 the coding was changed). The combination of both levels ensures that we obtain the most fine–grained information as possible. The categories were decided as follows:

- Lower: “in school”, “inadequate schooling”, “general elementary schooling”
- “SE”: (middle) vocational and abitur + post–secondary (but non tertiary)
- “BA”: higher vocational studies (bachelor–level)
- “MA”: higher education (master level or higher)

We observed several individuals who had a drop in the educational level over time, which is logically impossible. Hence, we correct these observations by replacing drops with the higher value of both. Missing observations were imputed only if the lag (one year before) and the lead (one year after) have the same value.

3.2.2.3 Controls

- *Gender*. Gender of the individual, with 0 being female and 1 male.
- *Age.*

Putting all together, our final *pooled* dataset consisted therefore of 208,903 individuals and 19 independent variables, representing an unbalanced panel of 11 years. More precisely, we decided to consider only the years from 2004 to 2014 (both included). First reason for this choice is that health variables like BMI and Smoking Habits are recorded starting from 2002 and only in even years: we however omitted 2002 itself as needed for interpolation when imputing them. Moreover, considering these years we also have that all the individuals have been interviewed at least three times, and Household Income and Number of doctor visits are never missing. Finally, we also dropped the individuals for which Self–Rated Health, BMI and Smoking Habits were still missing (since impossible to use the considered flexible time–trend approach).

On the *Transformed Pooled* (described below), we instead end up with 37 variables, and the same individuals: given the considered statistical modelling choices, the lack of balance in the panel is not an issue.

By definition, Number of doctor visits (in the last three months) is a *count variable*, meaning that it can only assume non–negative integer values. In this case, considering a Linear Regression may result in two issues: first, we may have nonsensical predictions, *i.e.* negative values. Second, the homoscedasticity assumption would be violated.

There are multiple ways to account for the two above. One way, is to consider a log transformation of the dependent variable. In our case, however, this is not an optimal solution given the abundancy of 0s in the dataset (29.20% of the observations). Better, Generalized Linear Models (GLM) are considered. Zuur *et al*. (2009) describes the procedure of building a GLM in two steps, namely making an assumption about the distribution of the dependent variable and choosing a link function between its expected value and a linear combination of the independent variables. A Linear Regression is a specific case of GLM when the dependent variable is assumed to be distributed normally and the link function is simply the identity function (hence the dependent variable's expected value being simply the fitted linear combination of the independent variables). However, despite these limitations, in the main text we focus on comparing Machine Learning algorithms with Linear Regression. The issue of properly treating Number of doctor visits as a count variable, and why overall in our case using a Linear Regression is not leading to significant issues, is extensively addressed in Appendix 3. All the considered robustness checks proved the appropriateness of considering the Linear Regression as benchmark: its main advantage is that it also leads to the estimation of easy–to–interpret coefficients, which will be discussed at length – and compared with the Shapley Values from the Random Forests – in Section 5.

Before presenting the results on the considered data, we therefore introduce the two considered specifications. Then, we describe the algorithms: Linear Regression, Random Forest, and K–Means–Clustering. Poisson Regression and Negative Binomial are directly described in Appendix 3, together with the results.

3.3 Statistical Modelling

We started pooling together data across all the 11 years, and all the analyses were performed on this pooled dataset, both with and without additional transformations. Formally, in the linear case, we started estimating:

$$y_{i,t} = X'_{i,t}\beta + \varepsilon_{i,t} \quad (1)$$

where $y_{i,t}$ is the value of the dependent variable at time *t* for individual *i*, whereas $X_{i,t}$ is the 19×1 vector including the values of the *19* independent variables for individual *i* at time *t*.

Moreover, we also followed Mundlak (1978) strategy, *i.e.* including both *group–mean and group–demeaned variables* for each individual. Formally, defining:

$$\overline{x}_{i,j} = \frac{1}{T}\sum_{t=1}^{T} x_{i,j,t} \quad (2)$$

as the group–mean value of variable *j* for individual *i* over the *T =11* time periods, the Mundlak Estimator solves:

$$y_{i,t} = \overline{X}'_i\gamma + \left(X_{i,t} - \overline{X}_i\right)'\delta + \varepsilon_{i,t} \quad (3)$$

where $\overline{X}_i$ is the 19×1 vector including the values of $\overline{x}_{i,j}$ of the *19* independent variables, and $\left(X_{i,t} - \overline{X}_i\right)$ the 19×1 vector including the values of the differences, at time *t*, from $\overline{x}_{i,j}$. In the Machine Learning case, the two equations to be estimated become:

$$y_{i,t} = f(X_{i,t}) + \varepsilon_{i,t} \quad (4)$$

and

$$y_{i,t} = f\left(\overline{X}_i, X_{i,t} - \overline{X}_i\right) + \varepsilon_{i,t} \quad (5)$$

since in Regression Trees – Breiman (1984) – we make no assumptions over the functional form with respect to the parameters.

For simplicity, for the reminder of this work, we will refer to the dataset including group–mean and within–group–demeaned variables as to the *"Transformed Pooled"* dataset (as opposed to the *"Pooled"* one).

3.3.1. Supervised Learning – predicting with Linear Regression and Random Forest

In the previous paragraph, we have already introduced the Linear model, under both specifications (equation 1 and 3, respectively, for the Pooled and Transformed Pooled). Both equations are solved by the estimates for the parameters minimizing the sum of squared distances of the

predicted values from the true ones of the dependent variable.

In the case of equation 4 and 5, instead, we make no specific assumption with respect to the relationship of the dependent variables and the parameters.

In this work, we solve them considering an ensemble of *Regression Trees*, called *Random Forest.*

For a formal and detailed discussion of Regression Trees and Random Forest, we remind the reader to Hastie *et al.* (2009). Here we just provide a description of their inner working and an illustrative example.

In a Regression Tree, we start with all the individuals belonging to the same group. Then, the algorithm starts considering one of the independent variables, and a threshold within it, so that the individuals are split in two subgroups. In each of the two groups, in turn, the Residual Sum of Squares (RSS) is computed, where the predicted value of the dependent variable is the average across all the individuals in that same subgroup. This operation is done also for the other independent variables, and for each of them for multiple thresholds (if not binary, where you have only one threshold).

The combination of variable–threshold that leads to the lowest sum of RSSes across the two subgroups is finally considered to perform the split.

As an example, suppose that we are trying to predict Number of doctor visits using only Gender (binary in our data) and Income. The algorithm will start considering as partition variable Gender, hence leaving all men in one subgroup and women in the other. The RSS in both subgroups is computed and summed up (leading to, say, 1.5). It will then consider Income, with threshold point, say, 30,000 euro, meaning that all the individuals earning less than/equal to 30,000 will be in one subgroup, and all higher earners in the other subgroup (irrespectively of their Gender). Once again, the RSS in both groups is computed and summed up, leading to, say, 1.6. Finally, again in Income, another threshold point will be considered, say 15,000, and the previous operation will be done, leading to a sum of RSSes, for instance, of 1.55.

Hence, in this example, the algorithm will finally consider Gender as variable to perform the split, on its only available cutoff point.

This operation is done multiple times, up until when a maximum depth (number of splits) ex–ante fixed by the researcher will be reached. Such amount is found via cross–validation. A too generous

maximum depth may lead to trees overly capable of predicting in the *training set* – the in–sample observations used to build the tree/compute the coefficients – but doing poorly on the *test set* – the out–of–sample observations considered to validate the models. Conversely, a value too strict for the max depth may lead to bad performances on both the sets (*underfitting*).

The maximum depth of each branch is an example of a *regularization criterion*: it is considered to address the risk of *overfitting* – the problem of overperformance on the training set and underperformance on the test set. Moreover, we also impose to the algorithm to consider only a random subset of the independent variables at each point to perform the split (instead of them all), another regularization criterion. In the above example, for instance, we may impose that either Gender or Income should be considered to perform the split, choosing randomly which of the two. Finally, to further smooth the predictions, we consider multiple independent and identically distributed Regression Trees, built on nonparametrically bootstrapped samples of the training set: such ensemble is called a *Random Forest* (Breiman 2001). In this case, instead of considering the prediction of one single tree, we consider as final prediction the average of all the predictions of the trees in the forest. The more uncorrelated the trees are, the more *ensembling* trees improves the prediction. In our case, we consider 1000 trees.

More details about the other two regularization criteria – maximum depth of each branch of each tree and maximum number of independent variables to be considered at each split of each tree – for all the cases are presented in Appendix 2.

3.3.2 Unsupervised Learning – clustering with K–Means–Clustering

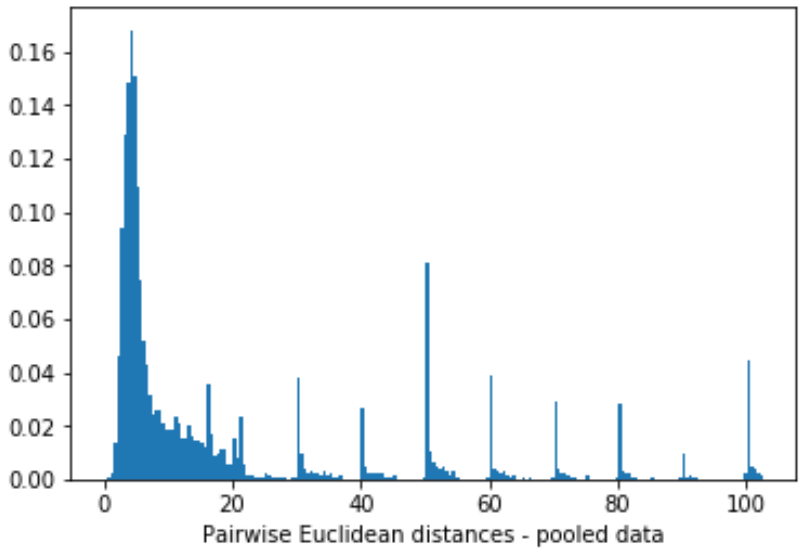
We already mentioned the use of Principal Component Analysis (PCA) for the creation of the Physical and Psychological health scales. As we mentioned initially, Unsupervised algorithms can also be used to identify clusters of individuals: to this aim, we considered *K–Means–Clustering*. K–Means–Clustering is an algorithm first introduced by MacQueen (1967), where *K* refers to the final number of clusters. It is also called "Lloyd's algorithm", referring to one of the key papers about it, Lloyd (1982). We use the K–Means implementation provided by the Python library *scikit–learn* – Pedregosa *et al.* (2011). Suppose we have *n* training individuals and that we have ex–ante chosen to have *k* clusters. The algorithm works as follows:

1. First, randomly pick *k* individuals in the training set, representing one–individual–only clusters.
2. Then, assign to each of these one–individual–only clusters all the remaining *n* – *k* individuals. The assignment is based on *closeness*, where closeness is defined based on the Euclidean distance.
3. In each of these *k* clusters, compute the *centroid*, namely the vector with the averages of the *p* variables computed over all the individuals in the cluster.
4. Re–assign as such the individuals to the clusters whose centroid is the closest (redoing step 2 and step 3).

When do we stop the process (steps 2 – 3 – 4)? At the end of step 3, in each cluster we compute the "Frobenius norm of the difference in the cluster centers of two consecutive iterations" (*scikit-learn* documentation). We iterate steps 3 and 4 up until when either this quantity is no larger than an ex-ante fixed tolerance (0.0001) for improvement, or at most after 300 iterations. The algorithm will eventually converge to a local minimum, but its value will strongly depend on the *k* initial random individuals chosen to create the clusters. Hence the whole procedure (1 – 4) is repeated 10 times. The repetition out of these 10 in which the sum across the clusters of all the *inertias* – sum of squared distances of each individual in each cluster from the centroid – is the smallest is the finally considered one.

Using K–Means allows us to further explore the data and interpret the results. In order to evaluate if indeed clustering – whether using K–Means or manually – is a promising way to proceed, we also compute all the $n(n-1)/2$ Euclidean distances across all the individuals in the dataset and plot them. All independent variables are standardized.

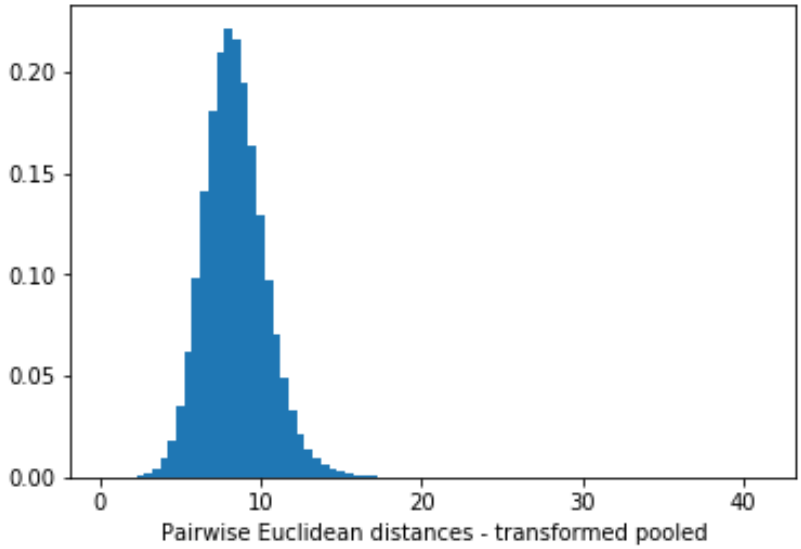
Figure 2: Histogram of the Pairwise Euclidean distances, Pooled dataset



Notes: Pairwise distances computed across 110,000 of the individuals in the pooled data (not the whole dataset for working memory limitations).

Excluding the 0 distances between each individual and oneself in the same year, on a subsample of 110,000 individuals – representing 52.56% of the overall dataset – we observe a minimum possible Euclidean distance of 0.04, and a maximum one of 102.54. As can be noticed in the above graph, there are indeed indications of the presence of multiple clusters in the data, in terms of subgroups of individuals who have the same Euclidean distance from each other. This is an indication that clustering may indeed lead to the identification of subgroups in which predicting algorithms can perform better – as compared to the global level, and in which eventual nonlinearities can be more easily modeled using ML algorithms.

Figure 3: Histogram of the Pairwise Euclidean distances, Transformed Pooled dataset



Notes: Pairwise distances computed across 110,000 of the individuals in the pooled data (not the whole dataset for working memory limitations).

Conversely, manually inspecting the pairwise Euclidean dataset in the Transformed Pooled, we do not see an immediate, intuitive presence of clusters. The minimum distance observed, on a subsample of 110,000 individuals, is 0.23, and the maximum is 41.36.

It is nonetheless still possible that the aforementioned K–Means–Clustering algorithm will be capable of finding patterns not immediately visible.

Now that both the specifications, the different algorithms, and the clustering procedure have been defined, we move on presenting the results.

3.4 Results

We start presenting the results at the *global* level, both for the Pooled and Transformed Pooled, by comparing the Test R2 of both Linear Regressions and Random Forests.

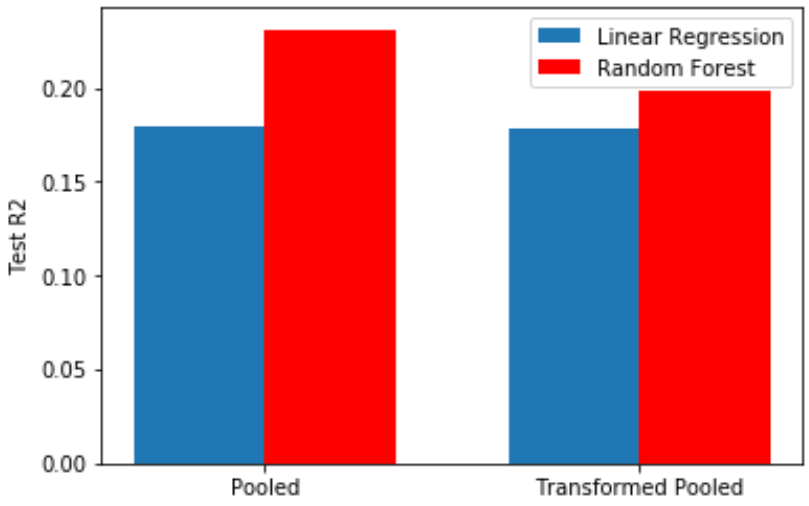
We then proceed discussing the results at the *local* level, namely in clusters, both on the Pooled

dataset and on the Transformed Pooled dataset. In each of the clusters, we compare the performances of Linear Regressions and Random Forests.

Finally, we open the "black box" by considering Linear Regressions' coefficients and Random Forests' Shapley Values, as well as *ablation studies*.

3.4.1 Global level analysis: Pooled and Transform Pooled

Figure 4: Test R2 of Linear Regression and Random Forest



Notes: Test R2 of Linear Regression (blue) and Random Forest (red) on Pooled and Transformed Pooled data, global level. Test Set includes 20% of the individuals (41,781).

As can be seen from Fig.4, we indeed observe Random Forest outperforming Linear Regression under both the considered specifications. The models were trained on a random subset of 80% of individuals (training set), and validated on the remaining 20% (test set). In particular, under the Pooled specification, the Test R2 produced by the Linear Regression and the Random Forest were, respectively, 0.1800 and 0.2312, yielding a relative improvement of the latter over the former of 28.44%. Under the Transformed Pooled specification, we also notice an improvement using Random Forest, although smaller: the Test R2 associated with Linear Regression and Random Forest were, respectively, 0.1782 and 0.1990, implying a relative improvement of 11.68%. These findings do indeed suggest the presence of nonlinearities in the data, especially in the Pooled specifications. When instead we take into account the panel dimension of the dataset applying the Mundlak correction (Transformed Pooled), there are still evidences of nonlinearities

in the data–generating process, but either less than in the Pooled case, or being more difficult to model.

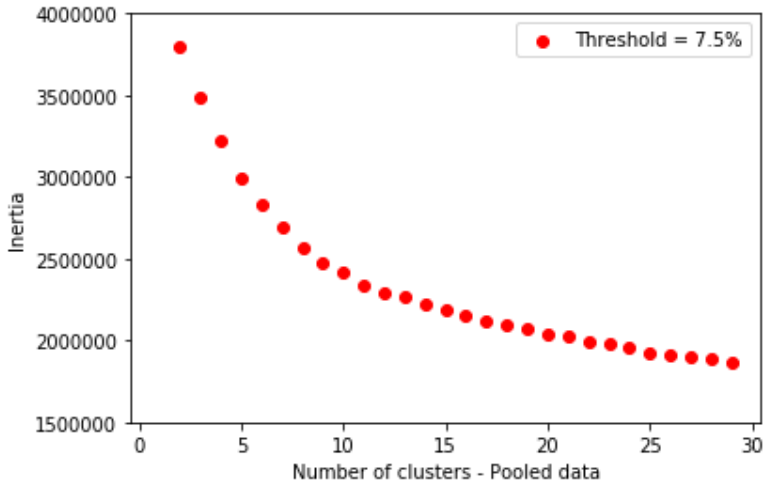
3.4.2 Local level analysis: clusters on Pooled and Transformed Pooled data

Before exploring in detail the predictive analytics results in the clusters, we briefly describe the results of the clustering procedure itself under the two specifications.

3.4.2.1 Clustering on the Pooled data

On the Pooled data, the considered threshold for the reduction in inertias was 7.5%, leading to a total of 5 clusters. The choice of such values is associated with the so called *Elbow Method*, based on plotting the number of clusters (horizontal axis) vs. the sum of inertias (vertical axis) over all the clusters. The sum of inertias is monotonically non–increasing in the number of clusters. At a certain point, the marginal decrease in inertias associated with a unitary increase in clusters doesn't reduce the sum of inertias sufficiently (less than the threshold), which therefore leads to an "elbow" in the curve.

Figure 5: "Elbow Method" for clustering on Pooled dataset



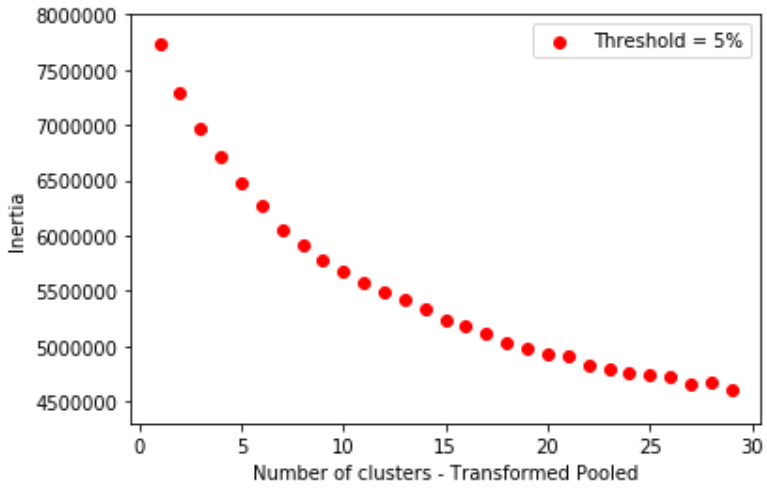
As can be seen in Figure 5, for the Pooled data, the "elbow" appears at five clusters: moving to six clusters reduced the sum of inertias from 2,987,103 to 2,833,446, a reduction of 5.42%, less than 7.5%. Moving instead from four to five reduced it from 3,221,765 to 2,987,104, a reduction of 7.86%.

The first interesting finding associated with this is that K–Means adopts – as optimal number of clusters – a value smaller than the number of years, and in general a number of clusters smaller than the number of peaks observed in Figure 2. The algorithm is therefore leading to results non immediately intuitive. Investigating the clusters, we indeed find that the year of observation is not relevant, since the clusters are not year driven.

3.4.2.2 Clustering on the Transformed Pooled data

Differently from the Pooled data, in the Transformed Pooled specification we considered a 5% threshold, leading to 3 clusters.

Figure 6: "Elbow Method" for clustering on Pooled dataset



When clustering on the Transformed Pooled, the "elbow" was less evident than in the Pooled case. In this case, the optimal number of clusters was only 3: moving from two to three clusters reduced the sum of inertias from 7,294,618 to 6,974,122 – a 4.60% reduction – whereas moving from three to four reduced it from 6,974,122 to 6,716,059 – a 3.84% reduction – hence not respecting the 5% threshold and leading the algorithm to stop.

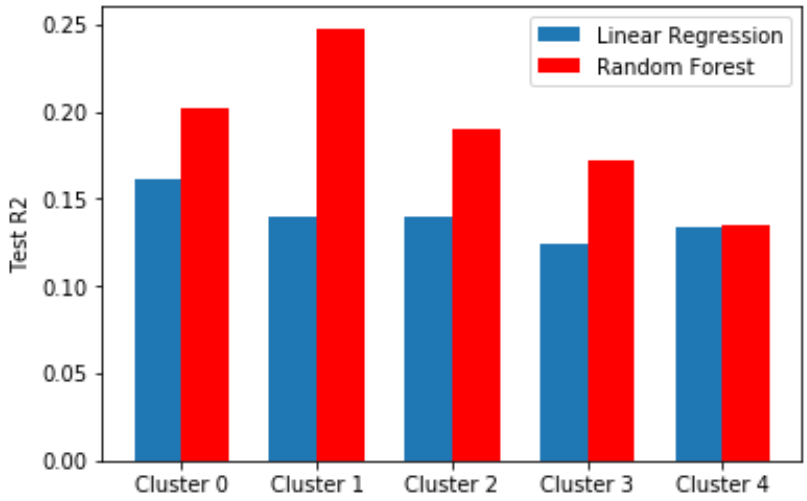
Overall, no particular correlations emerge within variables in the clusters derived under both specifications. The only correlations greater than 0.50 (or smaller than –0.50) are between the

dummies derived from the categorical variables, or between whether ever smoked and the moving average of smoked cigarettes. Similarly, strong correlations were observed between self–assessed health and the objective physical health scales. What instead is more of interest is the resulting distribution of Number of doctor visits (Appendix 1) – explaining the results in the following section – keeping in mind that Number of doctor visits was *not* included in the clustering process.

3.4.2.3 Results on clusters on Pooled data

Under both specifications, we observed the Random Forest leading to large improvements in the Test R2 with respect to the OLS. We start detailing the results for the clusters on the Pooled data.

Figure 7: Test R2 of Linear Regression and Random Forest on clusters from Pooled dataset



In each of the clusters, like at the global level, the dataset was split in training set (80% of the individuals) and test set (including the remaining 20%). The Random Forest was tuned separately in each cluster, hence leading to different architectures in each of them (the hyperparameters are presented in Appendix 2). Table 1 summarizes the results of Figure 7.

Table 1: Test R2 of Linear Regression and Random Forest in the five clusters derived from the Pooled dataset.

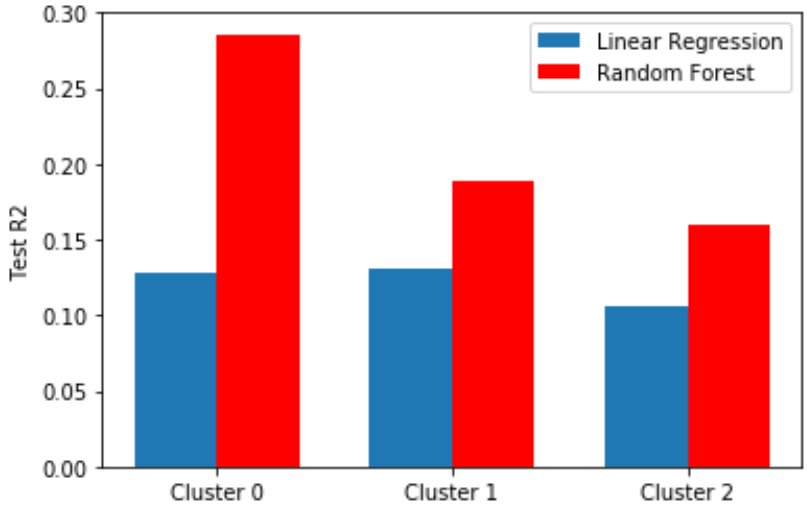
Cluster	Size (% whole dataset)	Test R2 LinReg	Test R2 RF	Test R2 RF / Test R2 LinReg (%)
Cluster 0	37057 (17.74%)	0.1611	0.2019	25.31%
Cluster 1	51799 (24.79%)	0.1393	0.2479	77.94%
Cluster 2	57886 (27.71%)	0.1399	0.1897	35.96%
Cluster 3	50573 (24.21%)	0.1237	0.1725	39.47%
Cluster 4	11588 (5.55%)	0.1333	0.1345	0.90%

As can be noticed, in all the clusters except the last – the smallest – Random Forest led to large improvement in predictive accuracy over the Linear Regression. This result in clusters confirms the result found at the global level, where indeed Random Forest was improving the Test R2 of 0.0512, implying a relative improvement of 28.44%. Considering the weighted average – by cluster size – of the Test R2 of Linear Regression and Random Forest, the improvement given by the latter is 42.98%, with the two measures being respectively 0.1392 and 0.1991 respectively. As can be seen in Appendix 1, interestingly enough, Cluster 1 is also the cluster with the highest level of nonzero visits (the only cluster where the mode is three and not zero). This is an indication that the Random Forest is better suited to capture nonzero values. In Appendix 2, where the optimal hyperparameters of the Random Forest are presented, can also be noticed that when predicting in Cluster 1 the optimal trees had indeed the longest branches, indicating the presence of strong nonlinearities.

3.4.2.4 Results on clusters on Transformed Pooled data

The previously observed pattern in clusters from the Pooled data appear also more significantly on the clusters from the Transformed Pooled.

Figure 8: Test R2 of Linear Regression and Random Forest on clusters from Transformed Pooled dataset



Also in this case a further 80 – 20 train–test split was performed in each cluster, and the Random Forest was optimized every time. The results of Figure 8 are summarized in Table 2.

Table 2: Test R2 of Linear Regression and Random Forest in the three clusters derived from the Transformed Pooled dataset.

Cluster	Size (% whole dataset)	Test R2 LinReg	Test R2 RF	Test R2 RF / Test R2 LinReg (%)
Cluster 0	58965 (28.22%)	0.1278	0.2859	123.69%
Cluster 1	81421 (38.98%)	0.1303	0.1889	44.95%
Cluster 2	68517 (32.80%)	0.1062	0.1602	50.86%

In the Transformed Pooled case, the in–cluster improvements due to Random Forest (over Linear Regression) are larger than in the Pooled case. Moreover, the cluster in which the improvement

is the largest this time is the smallest (in terms of individuals). These results show that when taking into account the panel dimension there are indeed strong nonlinearities in the data– generating process, which however are mostly modeled when the individuals are split in clusters. To compare, at the global level, we had observed that Random Forest was improving over the Linear Regression of only 0.018, accounting for a relative improvement of 11.68% – even less than the global level analysis on the Pooled data. Considering the weighted average across clusters as before, in this case the weighted Test R2 of Linear Regression and Random Forest are, respectively, 0.1217 and 0.2069, for an improvement of 69.99%.

Also in this case, as shown in Appendix 1, Cluster 0 is the only one where three is the mode (not zero). This is once again an indication that the Random Forest is particularly capable of predicting nonzero values. And indeed, the trees of the optimal forest in this cluster were also the longest.

Given the observed increases in predictive accuracy – under both specifications, and both at the global and local level (in particular) – it becomes crucial to understand *what* are the variables that are driving the most the increase in predictive accuracy. To do so, we present the *Shapley Values* – Shapley (1953) – of the Random Forest under the two specifications at the global level, and at the local level in the most populous clusters (for the other clusters, they are presented in Appendix 5).

3.5 Interpreting the results: what predicts Number of doctor visits

Shapley Values were first introduced in cooperative game theory – Shapley (1951) and Shapley (1953) – as a concept to fairly distribute the gains from a cooperative game.

Applied to Machine Learning, Shapley Values are quickly becoming the main interpretative tool used by data scientists to interpret the outcomes of Machine Learning algorithms otherwise uninterpretable.

The main shortcoming of this algorithm is its computational complexity: the number of computations performed to obtain them grows exponentially in the number of variables.

Recently, Lundberg *et al.* (2018), have developed a methodology called *TreeSHAP* – specific for tree–based methods – that renders the Shapley Values easier to compute, by reducing the complexity from exponential in the number of independent variables to quadratic in the maximum depth of each branch of each tree in the Random Forest.

For a detailed explanation of Shapley Values we remind the reader to Shapley (1953), Lundberg *et al.* (2018), and Molnar (2019). According to Molnar (2019), the *Shapley Value* of an explanatory variable is "the average of all marginal contributions across all possible coalitions of explanatory variables" (Molnar 2019, Chapter 5.9). Here, we focus on describing them considering an example from our case.

Suppose that in our Pooled specification instead of having 19 independent variables we had only three, namely – for the sake of the example – Gender, Income and Physical Scale (Physcale). Without loss of generality, imagine we are interested in the Shapley Value of the variable Income for a given individual *i*. After having trained our Random Forest, its computation would consist of the following steps:

1. Keep both Gender and Physcale as they are, and compute the difference in predicted Number of doctor visits keeping or not the variable Income. With "keeping or not" a variable, we mean shuffle randomly it for individual *i*.
2. Keep only Gender (hence shuffle randomly Physcale), and compute the difference in predicted Number of doctor visits keeping or not the variable Income.
3. Keep only Physcale (hence shuffle randomly Gender), and compute the difference in predicted Number of doctor visits keeping or not the variable Income.
4. Exclude both Physcale and Gender (hence shuffle randomly both Gender and Physcale), and compute the difference in predicted Number of doctor visits keeping or not the variable Income.

The Shapley Value for the variable Income for individual *i* will be the weighted average of the four differences in predicted Number of doctor visits – across the four previous cases – including or not Income, with weights depending on the size of the coalitions of the other variables. In particular, in the previous case the weights would be 1/3 for the case with both Gender and Physcale and for the case with none of them (points 1 and 4), whereas they would be 1/6 for both the case with only Gender and only Physcale (points 2 and 3).

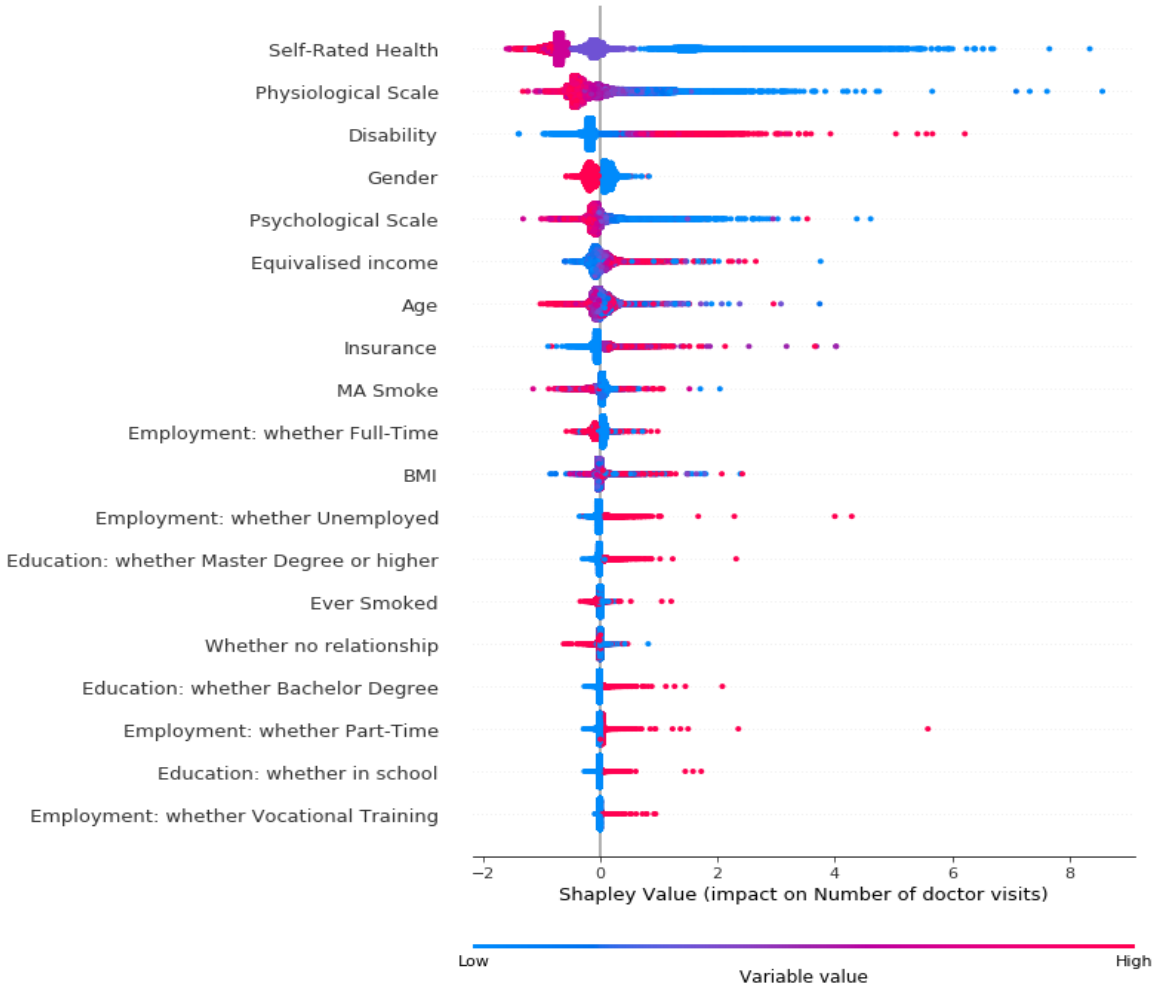
This example should make clear how Shapley Values can be hard to compute, since if we have *p* independent variables, to compute the Shapley Value of one of them, for *any* individual, we have to consider 2^{p-1}possible *coalitions* of the other variables. On the other hand, they support the

interpretation of contributions in terms of marginal effects, extremely useful especially for its comparability with the coefficients from the Linear Regression. For this reason, we compute them considering only building Random Forests with 100 trees instead of 1000: in terms of predictive accuracy, no major changes were observed. To interpret the Linear Regressions, and make a comparison across the two algorithms, we simply consider the estimated coefficients.

3.5.1 Interpreting the results: Shapley Values in the Pooled data

We start presenting the results at an individual level in the Pooled dataset. The Shapley Values are computed on the Test Set.

Figure 9. Shapley Values at an individual level, Pooled (test) data



In Figure 9, we can see the distribution of the Shapley Values across the variables, with the most important on top. That is, Self–Rated Health is the variable whose absolute mean of Shapley Values is the largest.

To understand how to read the above graph, let's consider Self–Rated health: each dot represents one individual in the Test Set. Given our reversing of the order of the categories, the variable is organized to express health, hence a value of 1 means "Bad", whereas 5 means "Very Good". Blue dots are associated with individual reporting 1 ("Bad"), with instead red dots are associated with individual reporting 5 ("Very Good"). The shades of color – moving from blue to red – are associated with the other three intermediate levels.

On the horizontal axis, there is reported the Shapley Value associated with each dot (individual). For instance, consider the extreme value on the right in correspondence of Self–Rated Health: for this person, including her reported Self–Rated Health ("Bad" in this case) led to a predicted Number of doctor visits *larger* by 8 visits on average – with respect to when her Self–Rated Health was not included – across all the possible coalitions of the other variables.

Shapley Values are interesting because they allow us to provide marginal interpretation at the individual level, or more in general at different points of the distribution.

Looking again at Self–Rated Health, we find that including or not this information is more relevant for people who rate it poorly than for individuals who rate it well.

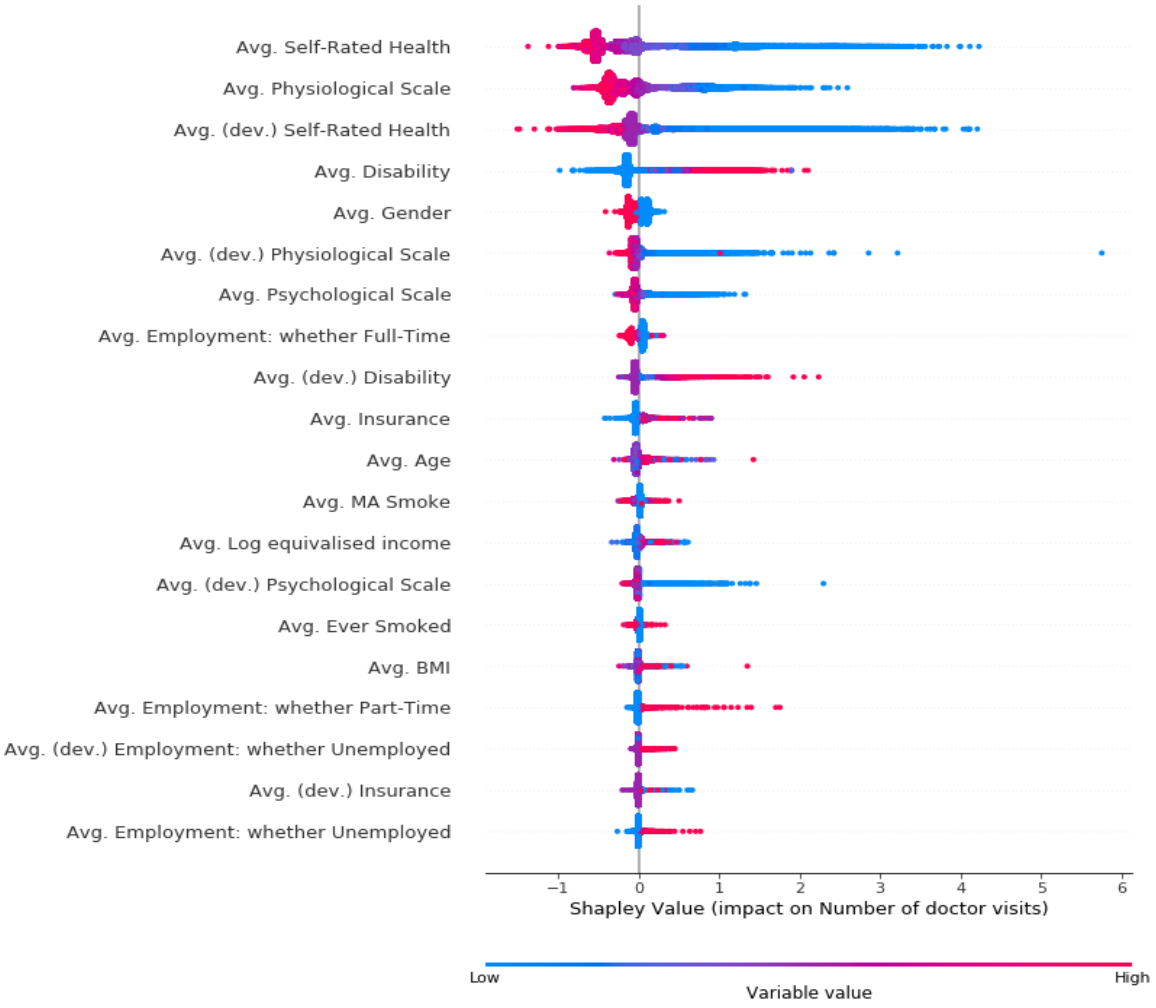
For individuals who rate it poorly, we observe marginal changes in the predicted amount – up to 8 visits – as previously described. Conversely, for those who reported it as "Very Good", we observe that including this information or not led to an increase in predicted Number of doctor visits of at most 2 visits. That is: including Self–Rated health for individuals who feel "Very good" lead to a predicted Number of doctor visits *smaller* by 2 visits at most on average – with respect to when Self–Rated Health was not included – across all the possible coalitions of the other variables.

Gender appears to be the fourth most important variables, with indeed women (blue dots, since 0 in the dummy Gender) visiting the doctor more often. While in terms of Mean Absolute Shapley Values Gender it is the fourth most important, we anyway notice no outlier (extreme values either on the right or left), meaning that for no one in the dataset including or not their gender changed

the average predicted Number of doctor visits in an extreme manner. Finally, we also note that having high income is associated with higher Number of doctor visits (excluding this information for high earners can reduce the average predicted Number of doctor visits up to 3), whereas having lower incomes is not associated with strong changes. This reflects the aforementioned "Non–need–based" nature of Income as predictor, and the public nature of the health system in Germany. The other variables associated with higher marginal positive impacts on Number of doctor visits are Higher education levels, age, better insurance coverage and being Unemployed.

3.5.2 Interpreting the results: Shapley Values in the Transformed Pooled data

Figure 10. Shapley Values of the 20 most important variables at an individual level, Transformed Pooled (test) data



In the Transformed Pooled, we have both the group–mean and group–mean–deviations variables. The former ones are identified with the denomination "Avg.", whereas the latter ones via "Avg. (dev.)" When analyzing the Shapley Values in the Transformed Pooled, we note a similar behavior to the Pooled ones. We notice that the Group–Mean and Deviation from Group–Mean Self–Rated health are the first and third most important variables, with the latter associated also with slightly larger negative Shapley Values than the former (left tail: *i.e.*, *decrease* in predicted Number of doctor visits when the variable *is* into the coalitions). This means that people stop going to the doctor as soon as they feel better, independently from how many times they had gone in the previous years.

Also interesting, we note that the Deviation from Group–Mean Disability Status (9^{th} most important variable) is mostly associated with positive Shapley Values (longer right tail): this suggests that deviations from the mean are usually positive, indicating an increase in the percentage of disability throughout years, in turn indicating a larger demand for healthcare. The Shapley Values at the local level – in the clusters – showed patterns similar to their global counterparts, under both specifications. For this reason, since what is particularly relevant at the cluster level is the possibility to model nonlinearities, we instead focus only on comparing the Mean Absolute Shapley Values and the Absolute Linear Regression Coefficients. This allows to understand with respect to which variables there were nonlinearities in the clusters, that instead via Machine Learning we were able to model.

3.5.3 Interpreting the results: MASVs and Coefficients in cluster from Pooled

We start presenting the Absolute Coefficients (ACs) and the Mean Absolute Shapley Values (MASVs) from cluster 2, the most populous from the Pooled dataset – the results for all the other clusters are presented in Appendix 5. We focus our attention only on the top ten most important variables as per MASV. All the reported coefficients were significant at 0.001 level: coefficients whose associated p–values from t–test is greater than 0.001 are directly reported as 0.

Table 3: Comparison of Absolute Coefficients (ACs) from Linear Regression vs. Mean Absolute Shapley Values (MASVs) on Cluster 2 obtained from Pooled data.

Variable	|Coefficient|	MASV	Ranking as per |Coefficient|	Ranking as per MASV
Self–Rated Health	0.7766	0.4992	1	1
Physio. Scale	0.5015	0.2774	3	2
Gender	0.2538	0.1939	4	3
Disability	0.5878	0.1230	2	4
Psych. Scale	0.1552	0.1008	7	5
Eq. income	0.09	0.0813	10	6
Insurance	0.1031	0.0526	8	7
Age	0.1725	0.0431	6	8
BMI	0	0.0318	11–18	9
MA Smoke	0.181	0.0225	5	10

Notes: Coefficients are computed on the Training Set, whereas the MASVs on the Test Set.

The first strong difference between the two measures is that the MASVs are always smaller than the Absolute Coefficients, with only exception of (the non–significant) BMI. This is related with the MASVs being derived from the Random Forest rather than the Linear Regression, hence already incorporating interaction effects. The Spearman–Rank correlation between the two measures is 0.68.

Particularly interesting is to notice how the AC of Disability is almost five times larger than its associated MASV, meaning that the percentage of disability is less relevant under the (more accurate) nonlinear specification. In terms of (the more comparable) ranking indeed, we notice that while it has the second largest AC, it only has the fourth largest MASV.

Similarly, we notice that, in this cluster, while the AC associated with Moving Average smoked cigarettes is the fifth largest, the MASV is only the tenth, and around nine times smaller. A possible conclusion that can therefore be drawn is that the nonlinearities captured at the cluster level using Random Forest are associated with a better modeling of the relationship between Number of doctor visits and the percentage of Disability, the Gender and the Moving Average Number of Smoked Cigarettes.

Finally, it's also interesting to compare the MASVs ranking in the Cluster with the ranking on the entire Pooled Test Set (vertical axis of Fig. 9).

On the entire Pooled Test Set, we notice a ranking similar to the MASVs in the cluster – with few exceptions, including the reversion of ranking between Disability and Gender.

3.5.4 Interpreting the results: MASVs and Coefficients in cluster from Transformed Pooled

We now present the same findings on cluster 1 from the Transformed Pooled.

Table 4: Comparison of (ACs) from Linear Regression vs. Mean Absolute Shapley Values (MASVs) on Cluster 1 obtained from Transformed Pooled data.

Variable	|Coefficient|	MASV	Ranking as per |Coefficient|	Ranking as per MASV
Avg. (dev.) Self–Rated Health	0.4958	0.3333	3	1
Avg. Self–Rated Health	0.5772	0.2967	2	2
Avg. Physiological Scale	0.3879	0.1859	4	3
Avg. Gender	0.1975	0.1073	8	4
Avg. (dev.) Physiological Scale	0.2719	0.1018	6	5
Avg. Disability	0.5964	0.0918	1	6
Avg. Eq. income	0.0856	0.0871	14	7
Avg. Insurance	0.1106	0.0812	12	8
Avg. Employment: whether FT	0.3463	0.0728	5	9
Avg. Psychological Scale	0.1097	0.0460	13	10

Notes: Coefficients are computed on the Training Set, whereas the MASVs on the Test Set.

In this case, the first noticeable difference is in Avg. Disability: while it has the largest AC, its MASV is only the sixth, with a MASV that is more than six times smaller than its associated AC. Also at the global level, we indeed noticed that Avg. Disability is the fifth most important variable in terms of MASVs (vertical axis of figure 10). The degree of Spearmen – Rank correlation among MASVs and ACs this time is lower, being 0.64.

This result – decreased relevance of Disability in the nonlinear specification – is also in line with what previously described in the Pooled specification. Here, the change in ranking is even larger. Interestingly, we notice that the Gender becomes more relevant in the nonlinear specification, as well as the Average Income, Average Insurance and Average Psychological Scale.

Variables that in the Random Forest lost importance but were among the top ten according to the Linear Regression included Avg. (dev.) Disability, Avg. Age, Avg. (dev.) Psychological Scale, and Avg. Employment: whether Vocational Training.

The Random Forest considered as more important of the above four the Avg. (dev.) Physiological Scale, the Avg. Equivalised income, the Avg. Insurance, and the Avg. Psychological Scale. The two versions of Self–Rated Health – Avg. and Avg. (de.) – were among the top three variables according to both the Linear Regression and the Random Forest.

The key conclusions that can be drawn comparing the ACs and MASVs in the clusters – under both specifications – is that the variables that play the largest role in predicting Healthcare Utilization are (the degree of) Disability, the SF–12 derived Physiological Scale and the Self–Rated Health. These variables also explain the difference in predictive accuracy between the Linear Regression and the Random Forest. For this reason, we further explore how the performance of the models degrades when these variables are deleted from the model. This operation, in the Machine Learning literature – and in general in Artificial Intelligence – is part of the so called *ablation studies*.

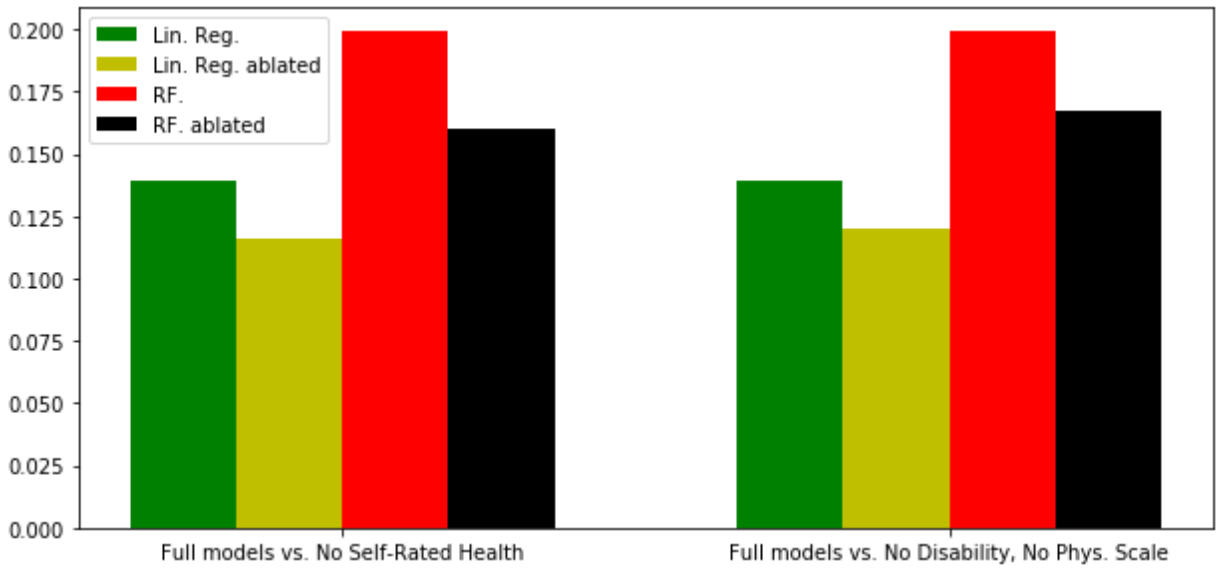
3.5.5 Interpreting the results: ablation of Disability, Physiological Scale and Self–Rated Health in clusters

In order to properly assess the role played by Disability, Physiological Scale and Self–Rated Health, we decided to proceed with the following comparisons:

1) Comparing the performance of the full model (all variables included) vs. model ablating *only* Self–Rated Health.
2) Comparing the performance of the full model (all variables included) vs. model ablating *both* Disability and Physiological Scale.

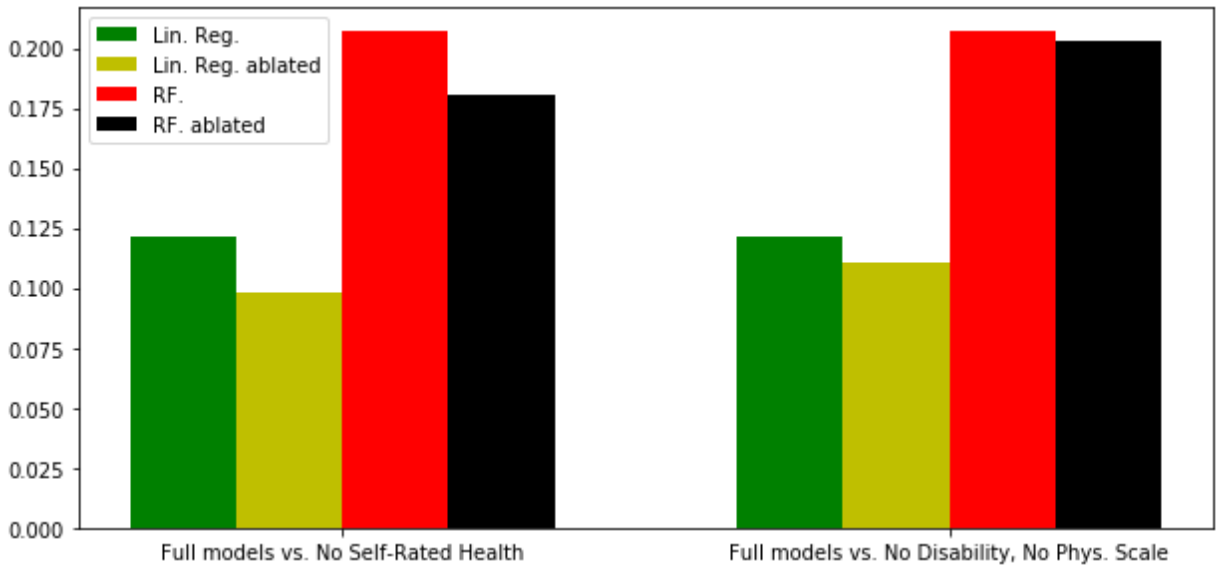
Self–Rated Health is always the most important variable, hence the choice of ablating it alone. In particular, in both 1) and 2), we computed the Weighted Average Test R2 across the clusters (weights being their relative size) and compared it across four cases: Linear Regression full model vs. ablated model, Random Forest full model vs. ablated mode. The results for the clusters from the Pooled specification are in Figure 11.

Figure 11. Weighted Average Test R2 in the clusters from Pooled specification, ablation study



The four bars on the left represents – respectively – the Weighted Average Test R2 (henceforth: *WATR*) of the full Linear Regression, of the Linear Regression without Self–Rated Health, of the full Random Forest, and of the Random Forest ablating Self–Rated Health. The four bars on the right represent the same quantities, but obtained ablating both Disability and Physiological Scale. The first interesting result to notice is that ablating Self–Rated Health alone and Disability and Physiological Scale together leads to a similar degradation in the average accuracy. The WATR of the Linear Regression and Random Forest ablating only Self–Rated Health are, respectively, 0.1160 and 0.1604, whereas the same measures ablating Disability and Physiological Scale together are 0.1199 and 0.1675 (light green and black bars, on both sides). This once again confirms the overall dominance of subjective over objective health in determining healthcare utilization (in the entire dataset, Self–Rated Health has a 0.7 positive correlation with Physiological Scale and a –0.36 negative one with Disability). In both cases, we observe a strong degradation w.r.to the full models, being the WATR of the Linear Regression and Random Forest of the full model (deep green and red bars, on both sides), respectively, 0.1399 and 0.1991.

Figure 12. Weighted Average Test R2 in the clusters from Transformed Pooled specification, ablation study



In this case, in abating each variable, we dropped both its group–mean and deviation from group–mean versions. We notice again a similar pattern to the clusters from the pooled specification: ablating Self–Rated Health alone led to a major degradation in the WATR than jointly ablating Disability and Physiological Scale. The Linear Regression and Random Forest's WATR ablating Self–Rated Health are, respectively, 0.0979, and 0.1803. Instead, ablating Disability and Physiological Scale, they are, respectively, 0.1104 and 0.2031 (light green and black bars, both sides). Moreover, we notice that ablating Disability and Physiological Scale, in general, degrades the performance of the two algorithms only marginally: the WATR of the full Linear Regression and Random Forest are, respectively, 0.1217 and 0.2069 (deep green and red bars, both sides). This is an indication of the overall stability, over time, of the degree of disability and objective health – at least in the considered time span.

3.6 Discussion

In this work, we aimed at replying three research questions.

The first question was whether Machine Learning techniques would help us in better predicting healthcare utilization – intended here as Number of doctor visits in the last three months – with

respect to traditional Linear Regression models. To do so, we considered a parsimonious set of variables, distinguished in need–based and not–need–based predictors. The need–based predictors included measures of subjective and objective health, as well as measures of psychological health. The non–need based predictors included variables like income, marital and employment status, describing the possibility to access healthcare. Age and gender controlled. In terms of data, we considered two specifications: in one, we pooled all individuals across 11 years, leading to an unbalanced panel of 208,903 individuals. In the other, in order to harness the presence of possible time effects, we considered a Mundlak transformation on these same data, hence including as predictors both the individual average of each variable across the years, as well as the deviation in each year from it.

As main Machine Learning algorithm we considered Random Forests, representing an excellent balance of computational complexity, flexibility, and variance in the predictions. As customary, we split the data in training and test set: on the former, we estimated the OLS coefficients/obtained the structure of the trees in the forest, and on the latter we computed the R2, our main evaluation metric. We found that under both specifications Random Forest consistently outperformed the Linear Regression.

Under the Pooled specification, using Random Forest led to an increase in predictive accuracy of 0.0512, implying a relative improvement of 28.44% (from 0.1800 to 0.2312).

The same result emerged under the Transformed Pooled specification, although smaller in magnitude: in this case, the Random Forest outperformed the Linear Regression by 0.018, implying a relative improvement of 11.68% (from 0.1990 and 0.1782).

We can therefore conclude that on the whole dataset, under both specifications, there are nonlinearities that Machine Learning algorithms are indeed capable of capturing.

This led us to further investigate the data moving from the *global* level (entire dataset) to the *local* level (clusters). We started considering the pairwise Euclidean distances to check if there were clusters immediately detectable, and then proceeded considering K–Means–Clustering to let the algorithm automatically identify them in an unsupervised manner.

Considering ad hoc stopping criteria, the algorithm identified five main clusters under the Pooled specifications and three under the Transformed Pooled one.

Within clusters, under both specifications, we found that the Random Forest led to improvements over the Linear Regression in predicting healthcare utilization even more significant than those at the global level. In the five clusters from the Pooled dataset, we noticed relative improvements in the predictive accuracy ranging from an increase in Test R2 of 77.94% (from 0.1393 to 0.2479) to 25.31% (from 0.1611 to 0.2019), with only the smallest cluster showing a similar performance between the two algorithms.

In particular, the improvement appeared to be the largest in the only cluster where the mode is not zero but three, implying that the Random Forest is particularly suited to capture the nonlinearities in predicting higher degrees of healthcare utilization.

When considering the clusters under the Transformed Pooled specification, the degree of improvement in predicting healthcare utilization associated with Random Forest is even larger. We observed the Random Forest leading to a 123.69% improvement in one of the clusters (Test R2 from 0.1278 to 0.2859), and an improvement of around 45% and 51% in the other two. Once again, the increase is the largest in the cluster where the mode is not zero but three. We can therefore conclude that, at the local level, the nonlinearities present in the data–generating process can be modeled even more accurately, despite the decrease in the training size with respect to the global level analysis. Moreover, the Random Forest outperformed the Linear Regression in particular where zero is not the mode, indicating that the nonlinearities in the data–generating process are particularly strong at higher degrees of healthcare utilization. Finally, we wanted to understand better which variables are the most important in predicting healthcare utilization.

To do so, we harnessed the internal interpretability of the Linear Regression by considering its estimated coefficients when significant at the 0.001 level. For the Random Forest, instead, we computed the Shapley Values, considering them both at the individual level – to assess positive and negative marginal effects – as well as terms of mean absolute values, hence comparable with the absolute coefficients.

At the global level, we focused only on the presenting the Shapley Values at the individual level, whereas within the clusters we compared the Absolute Coefficients vs. the Mean Absolute Shapley Values. The reason behind this choice is that the improvement in predictive accuracy is

more significant in the clusters, making therefore more interesting here to understand which variables drove the most the predictions between the two algorithms.

Conversely, at the global level, it becomes more interesting to understand in which *direction* the variables affect the Number of doctor visits (the observed patterns, in this sense, were similar at the local level).

Under both the Pooled and Transformed Pooled specification, we consistently found Self–Rated Health to be the most important variable, immediately followed by the more objective Physiological scale.

In particular, under both specifications, we noticed long blue tails associated with it. This means that low levels of Self–Rated Health explain higher degrees of healthcare utilization more than how high levels of Self–Rated Health explain low degrees of it. A similar pattern can also be noticed for the percentage of Disability: high levels of disability predict high levels of healthcare utilization more than how low levels of disability predict low levels of healthcare utilization. Considering instead the comparison of the Absolute Coefficients and Mean Absolute Shapley Values in the clusters allowed us to understand where the differences in predictive accuracy between the two algorithms (Linear Regression and Random Forest) were emerging. In the largest cluster derived from the Pooled specification, we observed that the degree of Disability and the Moving Average smoked cigarettes are more relevant in the Linear Regression than in the Random Forest, where instead Gender and Income become more important.

In the largest cluster from the Transformed Pooled specification, instead, we noticed that the deviation from the intertemporal average Self–Rated Health is the most important variable according to the Random Forest, but only the third according to the Linear Regression.

Similarly to the local analysis on the Pooled dataset, we again found that the degree of Disability's importance is highly overestimated by the Linear Regression – largest Absolute Coefficient – while it is only the sixth most important variable according to the Random Forest. And once again, at the local level, Gender assumed a more important role in the Random Forest rather than in the Linear Regression.

The key conclusions that can be drawn are that, in the major clusters under both specifications, most of the increased predictive accuracy given by the Random Forest is associated with a better

modeling of the relationship between the Number of doctor visits, Gender (women needing more healthcare than men) and Disability, especially for people needing more healthcare.

We see four main possible developments over the current work.

First, the improvements yielded by the Random Forest over the Linear Regression under both specifications open up the possibility to consider algorithms that are even more flexible. Further developments may include considering algorithms like *Gradient Boosting*, *Extreme Gradient Boosting*, *Kernel Ridge Regression*, and *Neural Networks*.

Second, it could be interesting to consider a broader set of independent variables, describing in a more granular manner the characteristics we already included. In this work, we did not consider this possibility since, by considering a parsimonious set of predictors, we already observed substantial improvements in predictive accuracy using Random Forest, worth exploring more in detail.

Third, the clustering procedure could be performed differently, either considering different algorithms or trying to cluster the data manually. In particular considering the first way, it would be interesting to observe how different algorithms organize the data differently, and which variables drive the clustering process across the different methods.

Finally, other health–related variable may be considered as target. For instance, the initially mentioned Number of nights spent in the hospital in the last year could be considered, as long as the dataset were to be crafted to consider only people with nonzero number of nights. Similarly, other variables not associated with healthcare utilization may be considered, as for instance mortality risk, or visits to a specialist.

Appendix

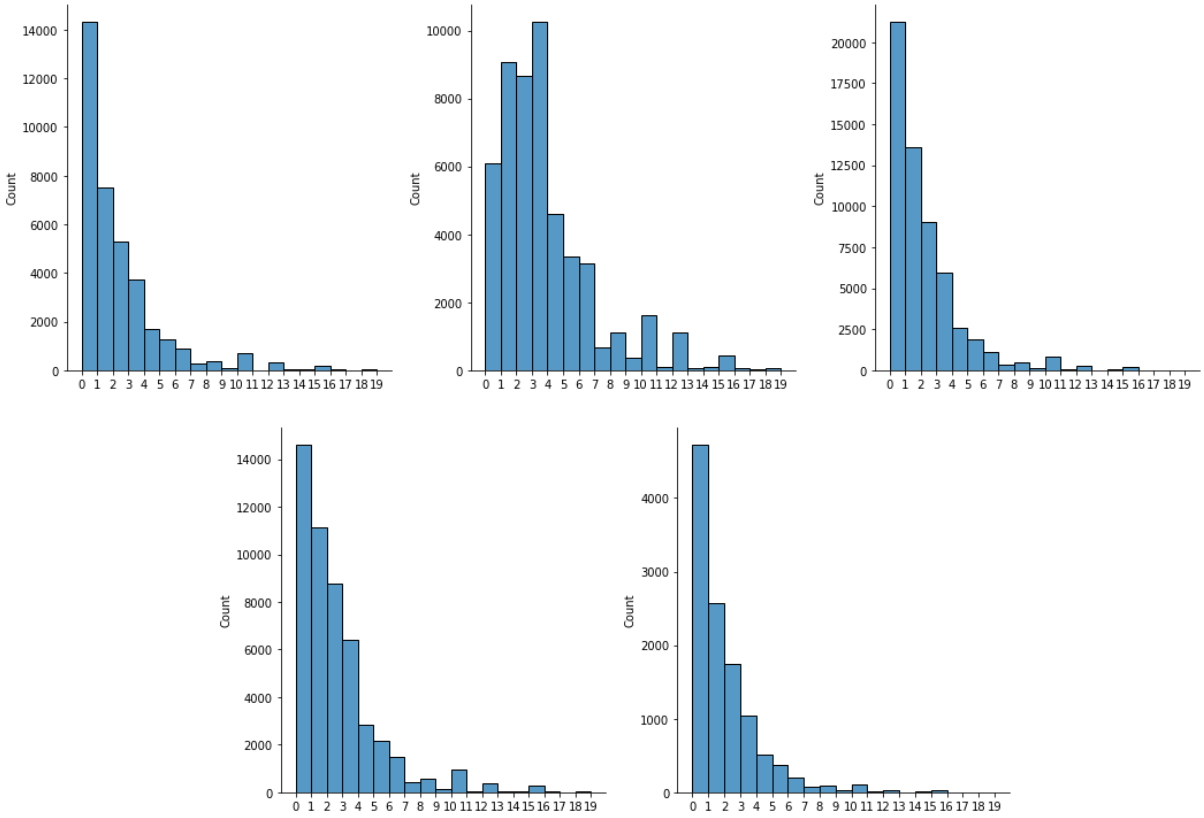
The Appendices are structured as follows:

In Appendix 1, we show the distribution of Number of doctor visits in the clusters created under both specifications. It is interesting to see how it is distributed since the clusters are created only considering the independent variables. In Appendix 2, we present the optimal hyperparameters found fitting the Random Forest both at the global and local level, under both specifications. In Appendix 3, we check whether treating Number of doctor visits as a count variable – hence

predicting it using a Poisson and Negative Binomial Regression – leads to significant changes with respect to the Linear Regression. In Appendix 4, we describe in greater detail the considered imputation techniques. In Appendix 5, we present the ACs and MASVs for all the remaining clusters, and in Appendix 6 we analyze in–depth how the clusters are built.

Appendix 1

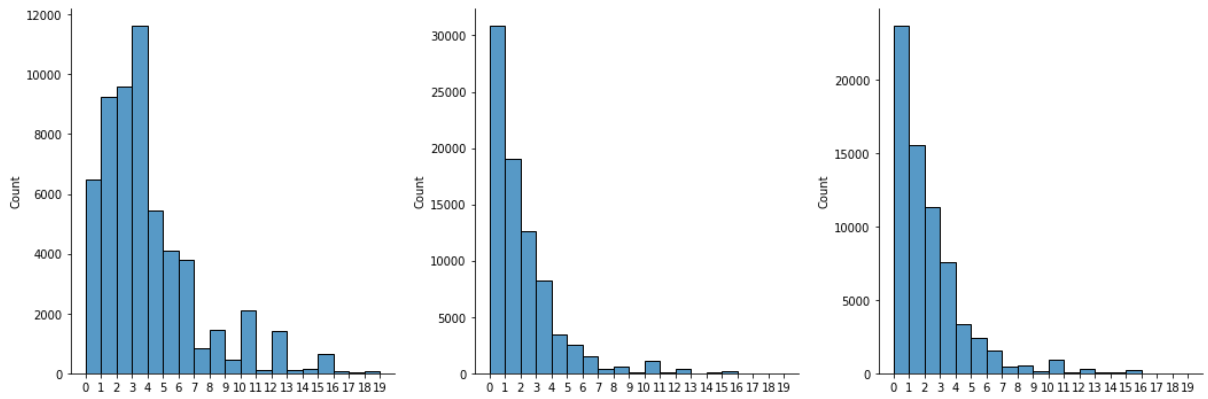
Fig.11 Distribution of Number of doctor visits in the last three months in the five clusters derived from the Pooled dataset.



Notes: Top left is Cluster 0, bottom right is Cluster 4. Values truncated at less than 20 visits for readability.

As can be noticed, in all the five clusters Number of doctor visits is centered around 0, with a long right tail. Only exception is Cluster 1, where the mode is three, and in general the distribution presents more nonzero individuals.

Fig.11 Distribution of Number of doctor visits in the last three months in the five clusters derived from the Transformed Pooled dataset.



Notes: Left is Cluster 0, right is Cluster 2. Values truncated at less than 20 visits for readability.

Once again, we observe that in one of the considered clusters the mode is not zero, but rather three. And similarly to the previous case, this is also the cluster where using the Random Forest over the Linear Regression lead to the largest improvements, confirming ML's better capability at predicting higher degrees of healthcare utilization.

Appendix 2

Table 5: Optimal hyperparameters of the Random Forest in the Pooled and Transformed pooled specifications, entire dataset.

Specification	Max. depth each branch	N. of considered variables to split	Number of trees in the forest
Pooled	23	7	1000
Transformed Pooled	11	9	1000

Notes: The algorithms are trained via 4–fold–cross–validation on 80% of the individuals (training set).

Interestingly, in the Pooled specifications the algorithm considers longer trees than in the Transformed–Pooled one, despite the former having almost half the predictors (19 and 37, respectively). In the entire datasets, we also observed the Random Forest outperforming the Linear Regression more in the Pooled specification than in the Transformed Pooled.

Table 6: Optimal hyperparameters of the Random Forest in the clusters from both the Pooled and Transformed pooled specifications

Specification	Max. depth each branch	N. of considered variables to split	Number of trees in the forest
Cluster 0 – Pooled	12	3	1000
Cluster 1 – Pooled	25	2	1000
Cluster 2 – Pooled	11	6	1000
Cluster 3 – Pooled	21	2	1000
Cluster 4 – Pooled	9	2	1000
Cluster 0 – Transf. Pooled	35	9	1000
Cluster 1 – Transf. Pooled	23	9	1000
Cluster 2 – Transf. Pooled	23	9	1000

Notes: The algorithms are trained via 4–fold–cross–validation on 80% of the individuals (training set).

Under both specifications, we notice that the optimal trees are the deepest in the two clusters where the improvement of the Random Forest was the largest over the Linear Regression, and in which there was a higher concentration of nonzero doctor visits. Compared to the forests at the global level, we notice that in particular in the three clusters derived from the Transformed Pooled, the trees are optimized to model nonlinearities more accurately.

Appendix 3

We here address the potential bias deriving from treating Number of doctor visits as a continuous numeric variable rather than a count variable. We do so presenting the results of two GLMs: Poisson Regression and Negative Binomial Regression.

We mentioned in the text that a GLM is built first by making an assumption about the distribution of the dependent variable, and then by choosing a link function between its expected value and the linear combination of predictors and parameters – Zuur *et al.* (2009). In the case of a Poisson

Regression, Number of doctor visits is assumed to be distributed according to a Poisson Distribution:

$$P(Y = y) = \frac{e^{-\lambda}\,\lambda^{y}}{y!} \quad (6)$$

hence endowed with the *equidispersion property*, meaning that $E(Y) = Var(Y) = \lambda$. The considered link function is the exponential one:

$$E(Y_i|\boldsymbol{x_i}) = e^{\boldsymbol{x_i'\beta}} \quad (7)$$

One key drawback of the Poisson distribution is indeed its equidispersion property. If we observe that $E(Y_i|\boldsymbol{x_i}) < Var(Y_i|\boldsymbol{x_i})$ we talk about *overdispersion*, and *underdispersion* vice versa. In order to address this problem, we also considered a Negative Binomial modeling:

$$P(Y = y) = \frac{\Gamma(\theta + y)}{y!\,\Gamma(\theta)}\left(\frac{\theta}{\theta+\lambda}\right)^{\theta}\left(1 - \frac{\theta}{\theta+\lambda}\right)^{y} \quad (8)$$

with, in this case, $Var(Y) = \lambda + \alpha\lambda^2$ and $\alpha \equiv 1/\theta$. It can be proved that for $\theta \to inf$ we revert back to the Poisson Regression. For this reason, α is referred in the literature as "dispersion", "shape", "aggregation", "heterogeneity" or "clustering" coefficient.

In order to fit and predict with a Negative Binomial Regression, we first need to estimate α. We follow Cameron and Trivedi (1990)'s procedure. The idea is to test for overdispersion comparing:

$$H_0: Var(Y) = \lambda$$
$$H_1: Var(Y) = \lambda + \alpha\lambda^2$$

by first estimating $\widehat{\lambda_\iota} = e^{\boldsymbol{x_i'\hat{\beta}}}$ via Poisson modeling, and then estimate α via the following auxiliary Linear Regression *without intercept*:

$$\frac{\left(y_i - \widehat{\lambda_\iota}\right)^2 - y_i}{\widehat{\lambda_\iota}} = \alpha\widehat{\lambda_\iota} + \varepsilon_i \quad (9)$$

The significance of $\hat{\alpha}$ allows to reject or not H_0. If the found p–value is significant, the value of $\hat{\alpha}$ is then plugged into the Negative Binomial in (8) and the regression is run. In our data, we found that $\hat{\alpha}$ was always significant at the 0.001 level, both at the local and global level under the two specifications. In Table 7, we present the results of both the Poisson Regression and Negative

Binomial in predicting on the Test Set. For readability, we also report the already presented values for the Linear Regression and Random Forest.

Table 7: Comparison of the Test R2 from the Linear Regression and the Poisson Regression under both specifications, both at the local and global level.

Specification	Test R2 Linear Reg.	Test R2 Poisson Reg.	Test R2 Neg. Bin. Reg.	Test R2 Rand. Forest
Pooled – entire dataset	0.1800	0.1858	0.1767	0.1990
Transf. Pooled – entire dataset	0.1782	0.1866	0.1821	0.2312
Cluster 0 – Pooled	0.1611	0.1953	0.1927	0.2019
Cluster 1 – Pooled	0.1393	0.1369	0.1319	0.2479
Cluster 2 – Pooled	0.1399	0.1553	0.1525	0.1897
Cluster 3 – Pooled	0.1237	0.1349	0.1353	0.1725
Cluster 4 – Pooled	0.1333	0.1342	0.1352	0.1345
Cluster 0 – Transf. Pooled	0.1278	0.1309	0.1274	0.2859
Cluster 1 – Transf. Pooled	0.1303	0.1412	0.1395	0.1889
Cluster 2 – Transf. Pooled	0.1062	0.1169	0.1153	0.1602

As can be noticed from table 7, the Poisson Regression and the Negative Binomial tend to perform similarly to the Linear Regression, with only noticeable exception the Cluster 0 from the Pooled specification. In any case, both continue to perform poorly when compared to the Random Forest. The resilience of the Linear Regression when compared to the more appropriate counting methods is explained by the little tendency in predicting negative values.

Table 8: Percentage of negative predictions of negative Number of doctor visits by OLS

Specification	%. of neg. preds.	Avg. of neg. preds.
Pooled – entire dataset	2.11	–0.2417
Transf. Pooled – entire dataset	2.23	–0.3000
Cluster 0 – Pooled	4.41	–0.4111
Cluster 1 – Pooled	0.53	–0.3365
Cluster 2 – Pooled	2.17	–0.1259
Cluster 3 – Pooled	0.98	–0.2011
Cluster 4 – Pooled	0.48	–0.0945
Cluster 0 – Transf. Pooled	0.42	–0.5419
Cluster 1 – Transf. Pooled	2.24	–0.1925
Cluster 2 – Transf. Pooled	1.27	–0.1884

As can be seen, only a small fraction of the predictions via OLS was negative, and their average still close to 0 under all circumstances.

Appendix 4

When describing the different variables we noted that in Self–Rated Physical Health, BMI and both the Physiological and Psychological scales, missing values have been imputed using a flexible time–trend approach. In this section, we describe it in greater detail. The basic is to interpolate the target variable using a time–trend. However, a linear time trend would in many cases be unrealistic (as it assumes unidirectional and monotonous evolution of time). We use the following interpolation algorithm:

1. Identify individuals with missing data points between 2000 and 2014.

2. Identify individuals with at least 3 available observations on the target variable (this is required to build a time–trend model. Note that this also excludes individuals who entered the SOEP panel after 2011).

3. Fit a *Generalized Additive Model (GAM*) for each person, predicting the target variable using a smoothing spline of time as predictor.

The model uses a piecewise cubic model of time (*i.e.*, a cubic–spline basis) to model a flexible time–trend. The software uses a built–in generalized cross–validation approach to determine the number of "pieces" (*i.e.*, knots) and the penalization term. The latter penalization is used to penalize the spline function specification at the knots (the boundaries of each "piece") and, thus, creates a continuous and smooth function of time. The R–package mgcv is used to estimate the interpolation model for each individual with missing data (*i.e.*, each person is allowed to have his/her idiosyncratic time trajectory).

4. Impute the missing data points using the GAM's prediction for each individual.

5. If necessary, apply a range restriction on the imputed values

For instance, avoid predicting negative values for a strictly positive variable and respect the theoretical range of the measurement scale. For this step, we use either (1) the theoretical bounds of the measurement scale (e.g., 1 and 5 for a 5–point Likert scale) or (2) the minimum and maximum value of the original scale, to winsorize the imputed values.

Appendix 5

Here, we present the comparison of MASVs and ACs also for all the other clusters – under both specifications. We remind that the values for Cluster 2 from the Pooled specification and Cluster 1 from the Transformed Pooled specification are already presented in the main text (Tables 3 and 4 respectively).

Table 9: Comparison of (ACs) from Linear Regression vs. Mean Absolute Shapley Values (MASVs) on Cluster 0 obtained from Pooled data (top ten by MASVs only)

Variable	|Coefficient|	MASV	Ranking as per |Coefficient|	Ranking as per MASV
Self–Rated Health	0.9073	0.5974	1	1
Physi. Scale	0.5398	0.3531	3	2
Gender	0.2393	0.2017	5	3
Disability	0.5712	0.2013	2	4
Psych. Scale	0.2237	0.1670	6	5
Eq. income	0.1738	0.0949	7	6
Empl.: Full–Time	0.1040	0.0873	9	7
Empl.: Unempl.	0.0882	0.0695	9	8
BMI	0.0536	0.0691	9	9
MA Smoke	0.0622	0.0651	9	10

Notes: Coefficients are computed on the Training Set, whereas the MASVs on the Test Set. The Spearman Rank correlation considering the top ten variables by MASV and the associated ranking by |Coef| is 0.92.

Table 10: Comparison of (ACs) from Linear Regression vs. Mean Absolute Shapley Values (MASVs) on Cluster 1 obtained from Pooled data (top ten by MASVs only)

Variable	|Coefficient|	MASV	Ranking as per |Coefficient|	Ranking as per MASV
Self–Rated Health	1.0061	0.6979	1	1
Physi. Scale	0.6046	0.5473	2	2
Disability	0.3176	0.4232	7	3
Psych. Scale	0.2637	0.2626	8	4
Insurance	0.2006	0.1204	9	5

Age	0.3192	0.1092	6	6
Eq. income	0.1774	0.1089	10	7
BMI	0	0.0831	13	8
MA Smoke	0.3439	0.0669	5	9
Whether no relat.	0.1244	0.0632	12	10

Notes: Coefficients are computed on the Training Set, whereas the MASVs on the Test Set. The Spearman Rank correlation considering the top ten variables by MASV and the associated ranking by |Coef| is 0.71.

Table 11: Comparison of (ACs) from Linear Regression vs. Mean Absolute Shapley Values (MASVs) on Cluster 3 obtained from Pooled data (top ten by MASVs only)

Variable	**|Coefficient|**	**MASV**	**Ranking as per |Coefficient|**	**Ranking as per MASV**
Self–Rated Health	0.8512	0.4640	1	1
Phys. Scale	0.5802	0.3554	2	2
Psych. Scale	0.2347	0.1547	5	3
Disability	0.5308	0.1200	3	4
Eq.income	0.167	0.1016	7	5
Age	0.3252	0.0884	4	6
Insurance	0.1216	0.0864	8	7
Gender	0.1923	0.0605	6	8
Empl: Unempl.	0	0.0571	11	9
BMI	0	0.0551	11	10

Notes: Coefficients are computed on the Training Set, whereas the MASVs on the Test Set. The Spearman Rank correlation considering the top ten variables by MASV and the associated ranking by |Coef| is 0.89.

Table 12: Comparison of (ACs) from Linear Regression vs. Mean Absolute Shapley Values (MASVs) on Cluster 4 obtained from Pooled data (top ten by MASVs only)

Variable	**|Coefficient|**	**MASV**	**Ranking as per |Coefficient|**	**Ranking as per MASV**
Self–Rated Health	0.5982	0.3156	2	1
Phys. Scale	0.7012	0.1883	1	2
Gender	0.2126	0.1438	4	3
Psych. Scale	0.2842	0.1429	3	4
Disability	0	0.0458	7	5
MA Smoke	0.1905	0.0433	5	6
Eq. income	0	0.0420	7	7
BMI	0	0.0416	7	8
Whether no relat.	0	0.0330	7	9
Age	0	0.0321	7	10

Notes: Coefficients are computed on the Training Set, whereas the MASVs on the Test Set. The Spearman Rank correlation considering the top ten variables by MASV and the associated ranking by |Coef| is 0.87.

Table 13: Comparison of (ACs) from Linear Regression vs. Mean Absolute Shapley Values (MASVs) on Cluster 0 obtained from Transformed Pooled data (top ten by MASVs only)

Variable	|Coefficient|	MASV	Ranking as per |Coefficient|	Ranking as per MASV
Avg. Self–Rated Health	0.9436	0.6724	1	1
Avg. (dev.) Self–Rated Health	0.672	0.5128	2	2
Avg. Phys. Scale	0.509	0.4217	3	3
Avg. Disability	0.3227	0.3857	4	4
Avg. Insurance	0.3157	0.2006	5	5
Avg. Psych. Scale	0.2695	0.1938	7	6
Avg. (dev.) Phys. Scale	0.2650	0.1302	8	7
Avg. (dev.) Disability	0.0707	0.1159	14	8
Avg. Eq. income	0.1839	0.1061	10	9
Avg. Age	0.3056	0.0786	6	10

Notes: Coefficients are computed on the Training Set, whereas the MASVs on the Test Set. The Spearman Rank correlation considering the top ten variables by MASV and the associated ranking by |Coef| is 0.79.

Table 14: Comparison of (ACs) from Linear Regression vs. Mean Absolute Shapley Values (MASVs) on Cluster 2 obtained from Transformed Pooled data (top ten by MASVs only)

Variable	|Coefficient|	MASV	Ranking as per |Coefficient|	Ranking as per MASV
Avg. (dev.) Self–Rated Health	0.4287	0.3968	2	1
Avg. Self–Rated Health	0.6143	0.3360	1	2
Avg. Empl: Full–Time	0.2454	0.2242	8	3
Avg. Gender	0.2503	0.2026	7	4
Avg. Phys. Scale	0.3905	0.1772	3	5
Avg. (dev.) Phys. Scale	0.269	0.1228	6	6
Avg. Empl.: whether Part–Time	0.1749	0.1027	10	7
Avg. Insurance	0.1512	0.0904	12	8
Avg. Eq. income	0.1041	0.0822	17	9
Avg. Age	0.3411	0.0813	5	10

Notes: Coefficients are computed on the Training Set, whereas the MASVs on the Test Set. The Spearman Rank correlation considering the top ten variables by MASV and the associated ranking by |Coef| is 0.64.

Appendix 6

In Appendix 1, we have observed the behavior of Number of doctor visits in each cluster. Here, instead, we investigate in detail which individuals are there in each cluster based on their independent variables. Given the observed differences in predictive accuracies between clusters, and between the analysis at the local and global level, it is important to understand what the peculiarities of each cluster are. To perform the comparison, we observe, for each variable, the difference between the mean in the cluster and at the global level, including the absolute deviation both in absolute and relative terms. For what it concerns Education and Employment, we do not present the statistics for the two reference categories Secondary Education and Pension, since they were not considered in running the algorithms (and can be derived marginally from the other dummies). For what it concerns the Clusters from the Transformed Pooled data, we present only the statistics regarding the mean of the individual group–mean variables, since the mean of the individual deviations from group–mean variables is always 0 (can be proved formally, not an empirical fact of our data).

Table 15: Comparison of Mean between Cluster 0 from Pooled specification and Entire Pooled

Variable	Mean in Entire (1)	Mean in Cluster (2)	|(2) – (1)|	(2) / (1) %
Self–Rated Health	3.37	3.33	0.04	98.81%
Disability	7.65	4.5	3.15	58.82%
Ever Smoked	0.28	1	0.72	357.14%
MA Smoke	3.23	12.44	9.21	385.14%
BMI	26.11	26.11	0	100.00%
Insurance	1.44	1.35	0.09	93.75%
Whether no relationship	0.38	0.48	0.1	126.32%
Eq. income	15631.57	14441.1	1190.47	92.38%
Gender	0.47	0.6	0.13	127.66%
Age	49.41	43.83	5.58	88.71%
Psychological Scale	50.17	49.04	1.13	97.75%
Physiological Scale	49.3	50.07	0.77	101.56%
Education: whether Bachelor	0.13	0.07	0.05	53.85%
Education: Master or higher	0.17	0.08	0.09	47.06%
Education: in school	0.08	0.08	0	100.00%
Employment: Full–Time	0.39	0.63	0.24	161.54%

Employment: Unemployed	0.13	0.2	0.08	153.85%
Employment: Part–Time	0.18	0.12	0.06	66.67%
Employment: Vocational Training	0.06	0	0.06	0.00%

The key peculiarity in cluster 0, with respect to the entire Pooled dataset, is that K–means clustered all people that have smoked at least once (against the 28% of the dataset), with therefore more than triple number of smoked cigarettes (3.85 times more cigarettes on average). Moreover, in Cluster 0, there are no people in Vocational Training – unsurprisingly considering that already in the entire dataset they only represent the 6%. We also observe that there are more Full–Time workers than Unemployed people.

Table 16: Comparison of Mean between Cluster 1 from Pooled specification and Entire Pooled

Variable	Mean in Entire (1)	Mean in Cluster (2)	|(2) – (1)|	(2) / (1) %
Self–Rated Health	3.37	2.81	0.56	83.38%
Disability	7.65	21.78	14.13	284.71%
Ever Smoked	0.28	0.14	0.14	50.00%
MA Smoke	3.23	1.26	1.97	39.01%
BMI	26.11	27.31	1.2	104.60%
Insurance	1.44	1.38	0.05	95.83%
Whether no relationship	0.38	0.32	0.06	84.21%
Eq. income	15631.57	15604.31	27.27	99.83%
Gender	0.47	0.47	0	100.00%
Age	49.41	70.67	21.27	143.03%
Psychological Scale	50.17	51.05	0.88	101.75%
Physiological Scale	49.3	41.09	8.21	83.35%
Education: whether Bachelor	0.13	0.13	0	100.00%
Education: Master or higher	0.17	0.14	0.03	82.35%
Education: in school	0.08	0.1	0.02	125.00%
Employment: Full–Time	0.39	0.01	0.38	2.56%
Employment: Unemployed	0.13	0.02	0.11	15.38%
Employment: Part–Time	0.18	0.01	0.17	5.56%
Employment: Vocational Training	0.06	0	0.06	0.00%

For what it concerns Cluster 1, the key difference is in the average degree of disability, that in the cluster is, on average, around 2.85 times larger. This also explains why this is the only cluster

derived from the Pooled specification where the mode of Number of doctor visits is 3 and not 0, since instead all the other variables are similar in the entire dataset.

Table 17: Comparison of Mean between Cluster 2 from Pooled specification and Entire Pooled

Variable	Mean in Entire (1)	Mean in Cluster (2)	|(2) – (1)|	(2) / (1) %
Self–Rated Health	3.37	3.64	0.27	108.01%
Disability	7.65	2.42	5.23	31.63%
Ever Smoked	0.28	0.08	0.2	28.57%
MA Smoke	3.23	0.89	2.35	27.55%
BMI	26.11	26.17	0.06	100.23%
Insurance	1.44	1.57	0.14	109.03%
Whether no relationship	0.38	0.34	0.04	89.47%
Eq. income	15631.57	19177.96	3546.39	122.69%
Gender	0.47	0.67	0.2	142.55%
Age	49.41	44.35	5.06	89.76%
Psychological Scale	50.17	50.68	0.51	101.02%
Physiological Scale	49.3	52.92	3.62	107.34%
Education: whether Bachelor	0.13	0.19	0.06	146.15%
Education: Master or higher	0.17	0.3	0.14	176.47%
Education: in school	0.08	0.02	0.06	25.00%
Employment: Full–Time	0.39	0.99	0.6	253.85%
Employment: Unemployed	0.13	0	0.13	0.00%
Employment: Part–Time	0.18	0	0.18	0.00%
Employment: Vocational Training	0.06	0	0.06	0.00%

For what concerns Cluster 2, the only peculiarity is that almost anyone in this cluster is employed Full–Time (99% vs. the entire dataset's average of 39%). We also observe a larger proportion of men (67% vs. the 47% of the entire dataset) and of Master or higher graduated (30% vs. 17% in the entire dataset).

Table 18: Comparison of Mean between Cluster 3 from Pooled specification and Entire Pooled

Variable	Mean in Entire (1)	Mean in Cluster (2)	|(2) – (1)|	(2) / (1) %
Self–Rated Health	3.37	3.53	0.16	104.75%
Disability	7.65	2.65	5.01	34.64%
Ever Smoked	0.28	0.14	0.14	50.00%
MA Smoke	3.23	1.4	1.83	43.34%
BMI	26.11	25.45	0.65	97.47%
Insurance	1.44	1.43	0.01	99.31%
Whether no relationship	0.38	0.29	0.09	76.32%
Eq. income	15631.57	13220.25	2411.33	84.57%
Gender	0.47	0.13	0.33	27.66%
Age	49.41	43.57	5.84	88.18%
Psychological Scale	50.17	49.61	0.56	98.88%
Physiological Scale	49.3	51.44	2.14	104.34%
Education: whether Bachelor	0.13	0.12	0.01	92.31%
Education: Master or higher	0.17	0.15	0.02	88.24%
Education: in school	0.08	0.06	0.02	75.00%
Employment: Full–Time	0.39	0	0.39	0.00%
Employment: Unemployed	0.13	0.36	0.23	276.92%
Employment: Part–Time	0.18	0.64	0.46	355.56%
Employment: Vocational Training	0.06	0	0.06	0.00%

In cluster 3, on the contrary, we observe lower levels of disability and people who have ever smoked (and consequently of smoked cigarettes), as well as a higher percentage of women. Interestingly, in this cluster we have only people unemployed or working part time.

Table 19: Comparison of Mean between Cluster 4 from Pooled specification and Entire Pooled

Variable	Mean in Entire (1)	Mean in Cluster (2)	|(2) – (1)|	(2) / (1) %
Self–Rated Health	3.37	3.95	0.58	117.21%
Disability	7.65	2.56	5.09	33.46%
Ever Smoked	0.28	0.23	0.05	82.14%
MA Smoke	3.23	2.33	0.9	72.14%
BMI	26.11	23.28	2.83	89.16%

Insurance	1.44	1.29	0.14	89.58%
Whether no relationship	0.38	0.94	0.55	247.37%
Eq. income	15631.57	12368.66	3262.91	79.13%
Gender	0.47	0.48	0.02	102.13%
Age	49.41	22.94	26.47	46.43%
Psychological Scale	50.17	49.73	0.44	99.12%
Physiological Scale	49.3	56.1	6.8	113.79%
Education: whether Bachelor	0.13	0.02	0.11	15.38%
Education: Master or higher	0.17	0.02	0.14	11.76%
Education: in school	0.08	0.36	0.28	450.00%
Employment: Full–Time	0.39	0	0.39	0.00%
Employment: Unemployed	0.13	0	0.13	0.00%
Employment: Part–Time	0.18	0	0.18	0.00%
Employment: Vocational Training	0.06	1	0.94	1666.67%

The dominant characteristics of the individuals in Cluster 4 is the average age, less than half than in the entire datasets. This explains also the lower average degree of disability (around one third than in the whole dataset), the more–than–double proportion of singles, and the presence of students and trainees.

Table 20: Comparison of Mean between Cluster 0 from Transformed Pooled specification and Transformed Pooled

Variable	Mean in Entire (1)	Mean in Cluster (2)	|(2) – (1)|	(2) / (1) %
Avg. Self–Rated Health	3.37	2.61	0.76	77.38%
Avg. Disability	7.65	23.03	15.38	300.96%
Avg. Ever Smoked	0.28	0.20	0.09	69.58%
Avg. MA Smoke	3.23	2.25	0.98	69.69%
Avg. BMI	26.11	28.02	1.91	107.32%
Avg. Insurance	1.44	1.29	0.15	89.82%
Avg. Whether no relationship	0.38	0.32	0.06	84.77%
Avg. Eq. income	15631.57	13934.67	1696.90	89.14%
Avg. Gender	0.47	0.42	0.05	90.29%
Avg. Age	49.41	66.45	17.04	134.49%
Avg. Psychological Scale	50.17	49.36	0.82	98.38%
Avg. Physiological Scale	49.30	39.25	10.05	79.62%
Avg. Education: Bachelor	0.13	0.10	0.03	78.56%
Avg. Education: Master or higher	0.17	0.09	0.07	56.07%

Avg. Education: in school	0.08	0.11	0.04	148.17%
Avg. Employment: Full–Time	0.39	0.09	0.30	22.37%
Avg. Employment: Unemployed	0.13	0.12	0.01	90.92%
Avg. Employment: Part–Time	0.18	0.08	0.10	45.60%
Avg. Employment: Voc. Training	0.06	0.00	0.05	5.71%

The key characteristic of the individuals in Cluster 0 from the Transformed Pooled specification is that, on average across the years, they had a more than tripled average degree of disability. As per Figure 11 in Appendix 1, this is indeed the only cluster obtained from the Transformed Pooled specification where the mode of Number of doctor visits was 3 and not 0. The ratio of the cluster mean and entire sample mean for the average of vocational training is not 0 since, in the cluster, the mean of the average time in vocational training is actually 0.003.

Table 21: Comparison of Mean between Cluster 1 from Transformed Pooled specification and Transformed Pooled

Variable	Mean in Entire (1)	Mean in Cluster (2)	|(2) – (1)|	(2) / (1) %
Avg. Self–Rated Health	3.37	3.64	0.27	108.00%
Avg. Disability	7.65	1.94	5.71	25.35%
Avg. Ever Smoked	0.28	0.33	0.05	118.41%
Avg. MA Smoke	3.23	3.89	0.66	120.43%
Avg. BMI	26.11	26.21	0.10	100.37%
Avg. Insurance	1.44	1.60	0.17	111.70%
Avg. Whether no relationship	0.38	0.33	0.05	85.75%
Avg. Eq. income	15631.57	19512.94	3881.37	124.83%
Avg. Gender	0.47	0.73	0.27	156.99%
Avg. Age	49.41	47.29	2.11	95.72%
Avg. Psychological Scale	50.17	51.52	1.35	102.70%
Avg. Physiological Scale	49.30	52.76	3.47	107.03%
Avg. Education: Bachelor	0.13	0.18	0.05	140.86%
Avg. Education: Master or higher	0.17	0.28	0.11	164.91%
Avg. Education: in school	0.08	0.02	0.06	21.91%
Avg. Employment: Full–Time	0.39	0.82	0.43	211.25%
Avg. Employment: Unemployed	0.13	0.03	0.10	22.08%
Avg. Employment: Part–Time	0.18	0.04	0.14	21.15%
Avg. Employment: Voc. Training	0.06	0.01	0.05	13.60%

In cluster 1 from the Transformed Pooled data, the only key characteristic we observe is a higher presence of highly educated men with full–time jobs.

Table 22: Comparison of Mean between Cluster 2 from Transformed Pooled specification and Transformed Pooled

Variable	Mean in Entire (1)	Mean in Cluster (2)	|(2) – (1)|	(2) / (1) %
Avg. Self–Rated Health	3.37	3.70	0.34	109.96%
Avg. Disability	7.65	1.21	6.45	15.76%
Avg. Ever Smoked	0.28	0.29	0.01	104.30%
Avg. MA Smoke	3.23	3.29	0.06	101.80%
Avg. BMI	26.11	24.35	1.76	93.25%
Avg. Insurance	1.44	1.36	0.07	94.86%
Avg. Whether no relationship	0.38	0.50	0.11	130.04%
Avg. Eq. income	15631.57	12479.55	3152.02	79.84%
Avg. Gender	0.47	0.19	0.28	40.63%
Avg. Age	49.41	37.25	12.15	75.40%
Avg. Psychological Scale	50.17	49.26	0.91	98.19%
Avg. Physiological Scale	49.30	53.83	4.53	109.19%
Avg. Education: Bachelor	0.13	0.09	0.04	69.89%
Avg. Education: Master or higher	0.17	0.10	0.07	60.67%
Avg. Education: in school	0.08	0.12	0.04	151.34%
Avg. Employment: Full–Time	0.39	0.13	0.25	34.60%
Avg. Employment: Unemployed	0.13	0.26	0.13	200.41%
Avg. Employment: Part–Time	0.18	0.43	0.25	240.52%
Avg. Employment: Voc. Training	0.06	0.16	0.10	283.82%

On the contrary, in Cluster 2 from the Transformed Pooled we observe a higher frequency of women with no full–time jobs and lower levels of education.

Conclusions

In this work, the key research question has been to study the determinants and improve the accuracy in predicting wellbeing and healthcare utilization using novel Machine Learning techniques.

In Chapter 1, the question of predicting and interpreting wellbeing – here declined as self–assessed life satisfaction – has been addressed using the same dataset (only exception: physical health) considered by Layard *et al.* (2014) in their seminal paper, and then considering a larger set of variables. The employed Machine Learning algorithms were Random Forest and Penalized Regressions, although experiments with other algorithms have also been carried out.

In terms of predictive accuracy, we did not observe any specific improvement considering the restricted set of variables, and a non–negligible one considering the extended set. This improvement is however mostly related with the presence of more variables than with the use of different algorithms. Nonetheless, in terms of interpretation of the findings, the application of Shapley Values on the Random Forest has allowed to us gain novel insights, producing marginal effects at all levels of the independent variables' distributions. In particular, we found out, using Machine Learning, that the role of gender may be overestimated in linear estimations.

In Chapter 2, we have addressed the question of predicting and interpreting wellbeing – declined both as self–assessed life satisfaction as well as positive and negative affects. We considered three different datasets: the American Gallup Daily Poll, the UK Household Longitudinal Study and the German Socio–Economic Panel – with sample sizes ranging from 30,000 individuals to more than 300,000 in one single year, and richer models with up to 450 independent variables. In this case, the abundance of data made predictions also across different Machine Learning algorithms significantly different. Indeed, an increase in predictive accuracy can be observed specifically across Machine Learning algorithms, and while not particularly large in absolute terms, it becomes non–negligible when compared to the changes in predictive accuracy associated with the ablation of physical health variables. Similar results were observed also exploiting the panel dimension of the dataset, using a Mundlak–like transformed set of variables. In terms of interpretability, the focus is put more on Permutation Importances – confirming the standard

findings in the literature, with interesting differences across countries – and on the study of the relationship between wellbeing and income and age. In this case, it is interesting to notice that also (the nonparametric) Machine Learning algorithms confirm the U–Shaped hypothesis of wellbeing w.r.to age and its concavity in income.

Finally, in Chapter 3, the attention is on predicting and interpreting the determinants of healthcare utilization, representative of a more objective facet of individuals' health. To study it, the analysis was done considering the SOEP data, on a rich pooled dataset built across 11 years, including more than 200,000 individuals. The analysis was performed considering two specifications, one of simply pooled data and another one with Mundlak–like transformed variables. On top of this, we further explored the possibility of moving the analysis at a *local* level, hence considering clusters of individuals automatically identified by the algorithm (therefore free from the researcher's assumptions). In this case, strong increases in predictive accuracy – using Machine Learning algorithms – can be observed, both considering the pooled dataset as well as the Mundlak–like transformed variables (with the gains being larger under the former specification at the global level, and under the latter at the local level). The increases in accuracies ranges from 50% larger R–squared to more–than–doubled ones when the analysis is performed in the clusters. The description of how the clusters are composed, along with the ablation studies of independent variables like Self–Rated Health, Disability and Physiological Scale, confirmed that the increase in accuracy is mostly related to the use of different algorithms, better capable of modeling relationship between healthcare utilization (and higher degrees of it in particular) and variables like Gender, Disability, and Physiological Scale.

Overall, the fundamental take of this work is that Machine Learning algorithms *can* lead to novel discoveries in economic and social sciences. On smaller datasets, mostly via different insights in terms of interpretability, whereas on larger datasets also in terms of predictive accuracy. Moreover, they can be particularly useful when the considered dependent variable is (more) objective, and the analysis is moved to a smaller group of individuals, automatically identified. In this case, predictive accuracies can increase significantly, and the insights can vary as well. Regarding the technical questions made in the General Introduction, we can conclude that on small datasets, considering a subjective dependent variable, low bias – high variance algorithms

may be unnecessary. On larger datasets, including more predictors, still focusing on a subjective dependent variable, the results suggest that indeed Machine Learning algorithms can lead to increases in accuracy, but to observe improvements large also in absolute terms more individuals may be needed. Finally, the results of the last chapter suggest that a more objective variable may be less sensitive to measurement errors and omitted variables, given its higher degree of predictability using Machine Learning algorithms, also without clustering. However, strong improvements in accuracies, and differing results in terms of interpretability, emerge in particular considering the analysis in clusters. Therefore, social scientists and economists interested in the application of Machine Learning algorithms should first consider how interested they are in increasing predictive accuracy – which we have argued in the General Introduction to be fundamental in the case of wellbeing and healthcare utilization – the availability of data, and the eventuality of clustering.

One key point common across the chapters is that with more data, low bias – high variance algorithms tend to increase their performance, and that high bias – low variance algorithm at least do not worsen it. This suggest therefore that different kinds of data – *big data* – may lead to particularly interesting new findings in the study of both wellbeing and healthcare utilization. Companies like Meta (Facebook–Instagram), Amazon, TikTok, Twitter, and Apple are already well–known to use data from users to better predict which content people may enjoy, which product may be interesting in buying, or even regularly recording health variables like heart rate (*e.g.*, Apple Watch). In this case, we talk about "big data" not only because of the size of the dataset, but also because of their *dynamism.* In other words, new data are constantly produced and used to update the models. Similarly, researchers in finance are starting to advocate for the application of Machine Learning algorithms to *nowcast* phenomena (stock prices, inflation) – rather than *forecast*, since in the long run "black swans" rendering past predictions useless may occur (Lipton and Lopez De Prado, 2020).

In the context of social sciences, and wellbeing and healthcare utilization in particular, datasets like those of the aforementioned companies may help find patterns in behaviors associated with lower/higher degrees of life satisfaction, and open up the possibility to better investigate (also in real time) how the effect of social media consumption – and all its associated network effects –

can lead to increased/decreased chronic degrees of life satisfaction. Exploring internet research, buying behavior and contents of interest on social media may also help predict the expected healthcare necessities. In general, future datasets in social sciences – on which Machine Learning algorithms may be exploited in all their potential – should allow for more dynamic, and possibly less biased regular estimations of degrees of wellbeing and health.

Bibliography

Abadie, A., Athey, S., Imbens, G.W., and Wooldridge, J. (2022) "When Should You Adjust Standard Errors for Clustering?", published online at https://economics.mit.edu/files/13927

Ahrens, A., Hansen, C. B., and Schaffer, M. E. (2020), "lassopack: Model selection and prediction with regularized regression in Stata", *Stata Journal*, 20, 176–235.

Andersen, H.H., Mühlbacher, A., Nübling, M., Schupp, J., and Wagner, G.G. (2007), "Computation of Standard Values for Physical and Mental Health Scores Using the SOEP version of SF–12v2", *Schmollers Jahrbuch*, 127, 171–182.

Andersen, R. (1968), *Behavioral model of families' use of health services*, University of Chicago: Center for Health Administration Studies.

Arrow, K.J. (1963), "Uncertainty and the Welfare Economics of Medical Care", *American Economic Review*, 53, 941–973.

Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Duncan, L., and Weinstein, J., (2018), "Improving refugee integration through data–driven algorithmic assignment", *Science*, 359, 325–329.

Bishop, C. M., (2006), *Pattern Recognition and Machine Learning*, New York: Springer.

Bond, T. N., and Lang, K. (2019), "The Sad Truth about Happiness Scales", *Journal of Political Economy*, 127, 1629–1640.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2022), "Deep Neural Networks and Tabular Data: A Survey", *Preprint at ArXiv:2110.01889*

Breiman, L. (2001), "Random Forests", *Machine Learning*, 45, 5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984), *Classification and Regression Trees*, New York: Chapman and Hall, Wadsworth.

Breyer, F., Zweifel, P., and Kifmann, M., (2004), *Gesundeisökonomik,* Springer, Berlin Heidelberg New York, fifth edition.

Cameron, A.C., Trivedi, P.K., (1990), "Regression–based tests for overdispersion in the Poisson model", *Journal of Econometrics*, 46, 347–364.

Cantril, H. (1965), *The pattern of human concerns*. New Brunswick: Rutgers University Press.

Chen, L.–Y., Oparina, E., Powdthavee, N., and Srisuma, S. (2019), "Have Econometric Analyses of Happiness Data Been Futile? A Simple Truth About Happiness Scales", *Preprint ArXiv:1902.07696*.

Chen, T., and Guestrin, C. (2016), "Xgboost: A scalable tree boosting system", *Proceedings of the 22nd ACM SIKGDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Cheng, T. C., Powdthavee, N., and Oswald, A. J. (2017), "Longitudinal Evidence for a Midlife Nadir in Human Well-Being: Results from Four Data Sets", *The Economic Journal*, 127, 126–142.

Clark, A. E., (2018), "Four Decades of the Economics of Happiness: Where Next?", *Review of Income and Wealth*, 64, 245–269.

Clark, A. E., (2001), "What Really Matters in a Job? Hedonic Measurement Using Quit Data," *Labour Economics,* 8, 223–42.

Clark, A.E., and Lepinteur, A. (2019), "The Causes and Consequences of Early–adult Unemployment: Evidence from Cohort Data", *Journal of Economic Behavior & Organization*, 166, 107–124.

Clark, A.E., and Oswald, A.J. (1994), "Unhappiness and Unemployment", *Economic Journal*, 104, 648–59.

Clark, A.E., and Oswald, A.J. (1996), "Satisfaction and Comparison Income," *Journal of Public Economics,* 61, 359–381.

Clark, A.E., Flèche, S., Layard, R., Powdthavee, N., and Ward, G. (2018), *The Origins of Happiness: The Science of Well–being over the Life Course*, Princeton University Press, Princeton NJ.

Clark, A.E., Lepinteur, A, (2022), "Pandemic Policy and Life Satisfaction in Europe*",* *Review of Income and Wealth 2022*, 68, 393–408.

Csáji, B.C., (2001) *Approximation with Artificial Neural Networks*; MSc Thesis, Faculty of Sciences, Eötvös Loránd University, Hungary.

D'Ambrosio, C., Greiff, S., Ratti, L., and Vögele, C., (2021), "Pandemic Life in Luxembourg 2021", PANDEMIC Research News – August 2021.

Diener, E., Lucas, R. E., and Oishi, S., (2018), "Advances and open questions in the science of subjective well–being", *Collabra Psychology*, 4.

Dolan, P. and Metcalfe, R., (2012) "Measuring subjective wellbeing: recommendations on measures for use by national governments", *Journal of Social Policy*, 4, 409–427.

Dolan, P., Peasgood, T., and White, M. (2008), "Do We Really Know What makes us happy? A Review of the Economic Literature on the Factors Associated with Subjective Well–being" *Journal of Economic Psychology*, 29, 94–122.

Dou, Q., Chen, H., Yu, L., Zhao, L., Qin, J., Wang, D., Ct Mok, V., Shi, L., and Heng, P. A., (2016), "Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks," *IEEE Trans. Med. Imag.*, 35, 1182–1195.

Dymecka, J., Gerymski, R., Machnik–Czerwik, A., Derbis, R., and Bidzan, M., (2021), "Fear of COVID–19 and Life Satisfaction: The Role of the Health–Related Hardiness and Sense of Coherence", *Frontieriers in psychiatry, 12.*

Ferrer-i-Carbonell, A., and Frijters, P. (2004), "How Important is Methodology for the Estimates of the Determinants of Happiness?", *Economic Journal*, 114, 641–659.

Freund, Y. (1995), "Boosting a Weak Learning Algorithm by Majority", *Information and Computation,* 121, 256–285.

Freund, Y., and Schapire, R. E. (1999), "A Short Introduction to Boosting", *Journal of Japanese Society for Artificial Intelligence,* 14, 771–780.

Friedman, J. H. (2001), "Greedy function approximation: A gradient boosting machine", *Annals of Statistics*, 29, 1189–1232.

Friedman, J., Hastie, T., and Tibshirani, R., (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent", *Journal of Statistical Software*, 33, 1–22.

Gigantesco, A., Fagnani, C., Toccaceli, V., Stazi, M.A., Lucidi, F., Violani, C., and Picardi, A., (2019), "The Relationship Between Satisfaction With Life and Depression Symptoms by Gender", *Frontiers in psychiatry*, 10.

Gorry, A., Gorry, D., and Slavov, S. N. (2018), "Does retirement improve health and life satisfaction?", *Health Economics*, 27, 2067–2086.

Grossman M. (1972), "On the concept of health capital and the demand for health*", Journal of Political Economy*, 80, 223–255.

Guven, C., Senik, C.c and Stichnoth, H., (2012), "You Can't be Happier than Your Wife. Happiness Gaps and Divorce", *Journal of Economic Behavior & Organization*, 82, 110–30.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer.

Helliwell, J.F., Huang, H., and Wang, S. (2016), "The Distribution of World Happiness", *World Happiness Report*.

Hoerl, A.E., and Kennard, R.W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems", *Technometrics*, 12, 55–67.

Hornik, K., Stinchcombe, M., White, H., (1989), "Multilayer Feedforward Networks are Universal Approximators", *Pergamon Press*, 2, 359–366.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning: with Applications in R*, New York: Springer.

Kahneman, D., and Deaton, A. (2010), "High Income Improves Evaluation of Life but Not Emotional Well–being." *Proceedings of the National Academy of Sciences*, 107, 16489–16493.

Kaiser, C., and Vendrik, M. C. (2020), "How threatening are transformations of happiness scales to subjective wellbeing research?", *SocArxiv Preprint*. https://osf.io/gzt7a/

Kaiser, M., Otterbach, S., and Sousa–Poza, A. (2022), "Using machine learning to uncover the relation between age and life satisfaction", *Scientific Reports*, 12, 5263.

Killingsworth, M. A. (2021), "Experienced well–being rises with income, even above $75,000 per year", *Proceedings of the National Academy of Sciences*, 118.

Kim, B., Khanna, R., and Koyejo, O.O. (2016), "Examples are Not Enough, Learn to Criticize! Criticism for Interpretability". *Advances in Neural Information Processing Systems*, 29, 2280–2288.

Kleinberg, J., Ludwig, J. Mullainathan, S., and Obermeyer, Z., (2015), "Prediction Policy Problems", *American Economic Review*, 105, 491–495.

Kuh, D., Hardy, R., Langenberg, C., Richards, M., and Wadsworth, M. E. J. (2002), "Mortality in adults aged 26–54 years related to socioeconomic conditions in childhood and adulthood: Post war birth cohort study", *BMJ*, 325, 1076–1080.

Lagnado, A.M., Gilchrist, K., Cvancarova Smastuen, M., and Memon, A., (2017), "Is subjective wellbeing associated with depression? A cross–sectional survey in southeast England", *10th European Public Health Conference: Parallel sessions.*

Layard, R., Clark, A.E., Cornaglia, F., Powdthavee, N., and Vernoit, J. (2014), "What Predicts a Successful Life? A Life–Course Model of Well–Being", *Economic Journal*, 24, 720–738.

Leopold, L. (2019), "Health Measurement and Health Inequality Over the Life Course: A Comparison of Self–rated Health, SF–12, and Grip Strength", *Demography*, 56, 763–784.

Liberini, F., Redoano, M., and Proto, E., (2017) "Happy Voters", *Journal of Public Economics*, 146, 41–57.

Lipton, A. and López de Prado, M. "Three Quant Lessons from COVID–19" Available at SSRN: https://ssrn.com/abstract=3580185

Lloyd, S.P. (1982), "Least squares quantization in PCM", *IEEE Transactions on Information Theory*, 28, 129–137.

Lundberg, S.M., and Lee, S.I. (2017), "A Unified Approach to Interpreting Model Predictions", Proceedings of the *31st Conference on Neural Information Processing Systems*, 4768–4777.

Lundberg, S.M., Erion, G.G., and Lee, S. (2019), "Consistent Individualized Explanatory Variable Attribution for Tree Ensembles", *Preprint at arXiv:1802.03888.*

Luttmer, E. (2005), “Neighbors as Negatives: Relative Earnings and Well–Being,” *Quarterly Journal of Economics*, 120, 963–1002.

MacQueen, J. B. (1967), "Some Methods for classification and Analysis of Multivariate Observations”, *Proceedings of 5–th Berkeley Symposium on Mathematical Statistics and Probability"*, 1, 281–297.

Manning, W.G., Newhouse, J.P., Duan, N., Keeler, E.B., and Leibowitz, A. (1987), “Health insurance and the demand for health care: evidence from a randomized experiment”, *American Economic Review*, 77, 251–277.

Margolis, S., Elder, J., Hughes, B., and Lyubomirsky, S. (2021), “What Are the Most Important Predictors of Subjective Well–Being? Insights From Machine Learning and Linear Regression Approaches on the MIDUS Datasets”, *PsyArXiv*

McCarthy, J., Minsky, M.L., Rochester, N., and Shannon, C.E., (1956), “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence”.

Mehta, P., Wang, C.H., Day, A.G.R., Richardson, C., Bukov, M., Fisher, C.K., and Schwab D.J (2019), “A high–bias, low–variance introduction to Machine Learning for physicists”. Physics Report, 810, 1–124.

Mitchell, T., (1997), *Machine Learning*, McGraw–Hill, New York.

Molnar, C. (2019), *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*.

Moor, I., Günther, S., Knöchelmann, A., Hoebel, J., Pförtner, T.K., Lampert, T., and Richter M. (2018) “Educational inequalities in subjective health in Germany from 1994 to 2014: a trend analysis using the German Socio–Economic Panel study (GSOEP)”. *BMJ Open*, 8.

Mundlak, Y. (1978), “On the pooling of time series and cross section data”, *Econometrica*, 46, 69–85.

Natekin, A., and Knoll, A. (2013), “Gradient boosting machines, a tutorial”, *Frontiers in Neurorobotics*, 7.

Nikolova, M., and Graham, C. (2020), “The Economics of Happiness”, In K. F. Zimmermann (Ed.), *Handbook of Labor, Human Resources and Population Economics*, 1–33, Springer International Publishing.

OECD. (2020), *How's Life? 2020: Measuring Well–being*, OECD.

ONS. (2021), *Well–being* – Office for National Statistics.

Oparina, E., and Srisuma, S. (2022), "Analyzing Subjective Well–Being Data with Misclassification", *Journal of Business & Economic Statistics*, 40, 730–743.

Oswald, A. J., Proto, E., and Sgroi, D., (2015), "Happiness and Productivity", *Journal of Labor Economics*, 33, 789–822.

Owens, G.M. (2007), "Gender differences in health care expenditures, resource utilization, and quality of care". *J. Manag. Care Pharm*, 14, 2–6.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011), "Scikit–learn: Machine Learning in Python", *Journal of Machine Learning Research*, 12, 2825–2830.

Proto, E., and Zhang, A. (2021), "COVID–19 and mental health of individuals with different personalities", *Proceedings of the National Academy of Sciences*, 118.

Reis, I., Baron, D., and Shahaf, S. (2018), "Probabilistic random forest: A machine learning algorithm for noisy data sets", *The Astronomical Journal*, 157.

Richardson, J., K.Y., Chen, G., Khan, M.A., and Iezzi A. (2014), "Subjective Wellbeing versus Utility: Incommensurable or mismeasured constructs", working paper 04–14, *Centre for Health Economics, Monash University*.

Roth, H.R., Lu, L., Liu, J., Yao, J., Seff, A., Cherry, K., Kim, L., and Summers, R.M., (2016), "Improving computer–aided detection using convolutional neural networks and random view aggregation," *IEEE Trans. Med. Imag.*, 35, 1170–1181.

Samuel, A. L., (1959), "Some studies in machine learning using the game of checkers", *IBM Journal of Research and Development*, 3, 210–229.

Schröder, C., and Yitzhaki, S. (2017), "Revisiting the evidence for cardinal treatment of ordinal variables", *European Economic Review*, 92, 337–358.

Shapley, L.S. (1953), "A Value for n–person Games", *Contributions to the Theory of Games, Princeton University Press, Princeton*, 2, 307–317.

Shwartz–Ziv, R., and Armon, A. (2022), "Tabular data: Deep learning is not all you need", *Information Fusion,* 81, 84–90.

Štrumbelj, E., and Kononenko, I. (2014), "Explaining Prediction Models and Individual Predictions with Explanatory Variable Contributions", *Knowledge and Information Systems*, 41, 647–665.

Stutzer, A., and Frey, B.S. (2006), "Does Marriage Make People Happy, Or Do Happy People Get Married?", *Journal of Socio–Economics*, 35, 326–347.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society B,* 58, 267–288.

Tibshirani, R.J. (2013), "The Lasso Problem and Uniqueness", *Electronic Journal of Statistics*, 7, 1456–1490.

Toh, C., Brody, J.P., (2021), "Applications of Machine Learning in Healthcare", in Kheng, T.Y., (ed.), *Smart Manufacturing – When Artificial Intelligence Meets the Internet of Things*, IntechOpen, London.

Urry, H. L., Nitschke, J.B., Dolski, I., Jackson, D.C., Dalton, K.M., Mueller, C.J., Rosenkranz, M.A., Ryff, C.D., Singer, B.H., and Davidson, R.J., (2004), "Making a Life Worth Living: Neural Correlates of Well–being", *Psychological Science*, 15, 367–72.

Vovk, V. (2013), "Kernel ridge regression", In B. Schölkopf, Z. Luo, and V. Vovk (Eds.), *Empirical inference: Festschrift in honor of Vladimir N. Vapnik*, 105–116. Springer Berlin Heidelberg.

Wager, S., and Athey, S. (2018), "Estimation and inference of heterogeneous treatment effects using random forests", *Journal of the American Statistical Association*, 113, 1228–1242.

Ward, G., (2020), "Happiness and Voting: Evidence from Four Decades of Elections in Europe", *American Journal of Political Science*, 64, 504–518.

Ware, J. Jr., Kosinski, M., and Keller, S.D. (1996), "A 12–Item Short–Form Health Survey: construction of scales and preliminary tests of reliability and validity". *Med Care*, 34, 220–233.

Wetzel, M., Huxhold, O., and Tesch–Römer, C. (2016), "Transition into Retirement Affects Life Satisfaction: Short– and Long–Term Development Depends on Last Labor Market Status and Education", *Social Indicators Research*, 125, 991–1009.

Winkelmann, L., and Winkelmann, R. (1998), "Why are the Unemployed so Unhappy? Evidence from Panel Data," *Economica*, 65, 1–15.

Wooldridge, J. (2010), *Econometric analysis of cross section and panel data*. MIT press.

Wooldridge, J., (2003), "Cluster–Sample Methods in Applied Econometrics", *American Economic Review*, 93, 133–138.

Yang, J. (2021), "Fast treeshap: Accelerating shap value computation for trees", Preprint at: *arXiv2109:09847*.

Yigzaw, K.Y., Wynn, R., Marco–Ruiz, L., Budrionis, A., Oyeyemi, S.O., Fagerlund, A.J., and Bellika, J.G. (2020), "The Association Between Health Information Seeking on the Internet and Physician Visits (The Seventh Tromsø Study – Part 4): Population–Based Questionnaire Study", *Journal of Medical Internet Resesearch*, 22.

Zou, H., and Hastie, T. (2005) "Regularization and variable selection via the elastic net", *Journal of the Royal Statistical Society B*, 67, 301–320.

Zuur, A., Ieno, E.N., Walker, N., Savelielev A.A., and Smith, G.M., (2009), *Mixed effects models and extensions in ecology with R*, NY: Springer.