mathematics

*Article*

# Unsupervised Modelling of E-Customers' Profiles: Multiple Correspondence Analysis with Hierarchical Clustering of Principal Components and Machine Learning Classifiers

**Vijoleta Vrhovac, Marko Orošnjak *, Kristina Ristić, Nemanja Sremčev, Mitar Jocanović, Jelena Spajić and Nebojša Brkljač**

Department of Industrial Engineering and Engineering Management, Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia; violeta.vrhovac@uns.ac.rs (V.V.); kristina.ristic@uns.ac.rs (K.R.); nextesla@uns.ac.rs (N.S.); mitarj@uns.ac.rs (M.J.); stankovicj@uns.ac.rs (J.S.); n.brkljac@uns.ac.rs (N.B.)

* Correspondence: orosnjak@uns.ac.rs

**Abstract:** The rapid growth of e-commerce has transformed customer behaviors, demanding deeper insights into how demographic factors shape online user preferences. This study performed a threefold analysis to understand the impact of these changes. Firstly, this study investigated how demographic factors (e.g., age, gender, education) influence e-customer preferences in Serbia. From a sample of $n = 906$ respondents, conditional dependencies between demographics and user preferences were tested. From a hypothetical framework of 24 tested hypotheses, this study successfully rejected 8/24 (with $p < 0.05$), suggesting a high association between demographics with purchase frequency and reasons for quitting the purchase. However, although the reported test statistics suggested an association, understanding how interactions between categories shape e-customer profiles was still required. Therefore, the second part of this study considers an MCA-HCPC (Multiple Correspondence Analysis with Hierarchical Clustering on Principal Components) to identify user profiles. The analysis revealed three main clusters: (1) young, female, unemployed e-customers driven mainly by customer reviews; (2) retirees and older adults with infrequent purchases, hesitant to buy without experiencing the product in person; and (3) employed, highly educated, male, middle-aged adults who prioritize fast and accurate delivery over price. In the third stage, the clusters are used as labels for Machine Learning (ML) classification tasks. Particularly, Gradient Boosting Machine (GBM), Decision Tree (DT), k-Nearest Neighbors (kNN), Gaussian Naïve Bayes (GNB), Random Forest (RF), and Support Vector Machine (SVM) were used. The results suggested that GBM, RF, and SVM had high classification performance in identifying user profiles. Lastly, after performing Permutation Feature Importance (PFI), the findings suggested that age, work status, education, and income are the main determinants of shaping e-customer profiles and developing marketing strategies.

**Keywords:** e-commerce; customer profiles; demographics; user preferences; multiple correspondence analysis; hierarchical clustering; machine learning

**MSC:** 62H25; 62H30; 62H17; 62F03

check for updates

**Citation:** Vrhovac, V.; Orošnjak, M.; Ristić, K.; Sremčev, N.; Jocanović, M.; Spajić, J.; Brkljač, N. Unsupervised Modelling of E-Customers' Profiles: Multiple Correspondence Analysis with Hierarchical Clustering of Principal Components and Machine Learning Classifiers. *Mathematics* **2024**, *12*, 3794. https://doi.org/10.3390/math12233794

Academic Editors: Wen Zhang, Xiaofeng Xu, Jun Wu and Kaijian He

Received: 4 November 2024
Revised: 26 November 2024
Accepted: 27 November 2024
Published: 30 November 2024



## 1. Introduction

E-commerce is developing faster than ever, with an increase in the number of transactions [1,2]. This global trend allows retailers to expand their reach with as many customers as possible [3]. Additionally, the COVID-19 pandemic has further accelerated consumer habits, with many consumers switching to online shopping [4,5], posing new challenges for e-tailers [6–8]. With the rise in online consumers, e-tailers are forced to improve customer experience throughout the purchasing process [9,10]. Similar trends are visible in the

Serbian market, where e-shopping is becoming pronounced, emphasizing the importance of e-commerce in modern business.

According to Ranđelović [11], despite the numerous benefits of e-commerce, consumers in Serbia still need to accept this form of commerce. One of the reasons for the slow adoption of e-commerce is consumer distrust, especially regarding the security of transactions [12]. Still, in 2023, the e-commerce market in Serbia grew by 34.5%, reaching a value of 955.7 million dollars, and it is predicted that by 2027, the market will grow to 1.65 billion dollars. The growth in e-commerce users is estimated to reach 4.36 million customers by 2027, comprising 62.5% of Internet users in Serbia [13]. Regardless of these positive trends, Serbia faces challenges such as the preference for cash-on-delivery payments and a continued lack of trust in the security of online transactions [14].

Demographic data, such as gender, age, income, and education, are vital in analyzing consumer preferences and behavior [15,16]. These variables enable market segmentation and the adaptation of marketing strategies to the specific needs of different consumers [17]. For example, income indicates purchasing power and can help target marketing campaigns to different socio-economic segments [18]. Education and age influence how consumers use technology and how they behave when shopping online [19]. Understanding these demographic factors allows e-tailers to identify target groups and optimize customer experience [20]. In Serbia, where e-commerce is growing, insight into demographic variables can help overcome obstacles and improve consumer acceptance of e-commerce. This provides valuable information for adjusting e-tailers' market strategies and enhancing user experience.

Considering the research problem's contextual settings, this study's primary idea is to switch from a traditional statistical exploratory analysis to an ML approach in identifying the user profiles of e-customers. Firstly, an analysis of how demographic factors influence the online behavior of e-customers in Serbia is performed. Specifically, how demographics affect user preferences is of interest here, particularly purchase frequency (PURFREQ), the most important property when buying for the first time (MIPB1T), the most important property before repeating the purchase (MIPBREP), and reasons for quitting the online purchase (RFQ), using a hypothetical framework. However, given that statistical testing highlights significant relationships between categories, it needs to provide an understanding of how these categories interact. As a response, Multiple Correspondence Analysis with Hierarchical Clustering on Principal Components (MCA-HCPC) is performed as an unsupervised model for labelling user profiles. Secondly, after obtaining class categories, i.e., labels, an ML classification using a dataset from user profiles that includes demographics and user preferences is conducted. Lastly, the most important features of the best-performing algorithms are extracted using Mean Dropout Loss (MDL).

The rest of this study is structured as follows. Section 2 provides an in-depth description of the survey used for this study. In addition, rigorous data analysis procedures are performed, starting from a priori sample determination, study selection, and the proposition of hypotheses in this study. The third section provides information about descriptives, including demographics and user behavior variables. Next, the most significant and non-significant results are provided to ensure the transparency and replicability of the findings. Also, given that existing association tests fail to indicate exact interactions between categories, an MCA is performed to provide a deeper understanding of the association between variables, including all of the variables in this study, and the variables are reduced to only those reported as statistically significant. This study provides a discussion section that explains the findings obtained, and finally, it gives concluding remarks, limitations, and implications of this study in the last section.

## 2. Literature Review

When analyzing user profiles in e-commerce and the influence of demographics, limited studies include supervised or unsupervised learning models. Previous studies primarily rely on user descriptive profiles and inferential statistics for hypothesis testing [1,15,21]. In contrast, others extend the analysis with multivariate regression [22] and factor models [23–25], whereas item answers are used as continuous values assuming a normal distribution. In addition, a sample of studies that use advanced mathematical modelling of e-customer profiles is identified. For example, Hristoski I. et al. [26], using customer behavior model graphs (CBMGs), identify 12 typologies of online shoppers. The CBMGs work as an $N \times N$ transitional probability matrix that denotes relative frequencies of invoking specific e-commerce functions. These frequencies are then used to map a unique transitional probability matrix of users with similar behavior.

Next, Swarnakar P. et al. [27] use Logistic Regression (LR) and Artificial Neural Networks (ANN) to predict online purchase behaviors to help e-retailers develop suitable strategies. They first perform a statistical analysis of demographics and user preferences to identify relevant factors affecting online behavior before using LR and ANN to predict user behavior. Chen X. et al. [28] used deep learning algorithms to predict user demographics based on the multisource user characteristics and preferences data in a vice versa approach.

On the other hand, some have tried identifying user profiles through cluster modelling. For instance, Bellini P. et al. [29] perform k-means clustering to identify user profiles based on demographics and preferences. The results show that after identifying different clusters of user profiles, a stimulus was generated for the customers, which increased buyers' purchases by 3.48%. Kuruba, Y. et al. [30] Customer segmentation was successfully performed using distributed clustering models. Lastly, a study by Hørlück J. et al. [31] fails to identify user behavior clusters using an unsupervised Gaussian Mixture Model with Hierarchical Probabilistic Clustering. It neither falsifies nor supports the hypothesis of identifying clear user profiles based on buying behavior.

Furthermore, understanding that demographics play an important role in shaping user profiles, a brief overview of demographic variables affecting user preferences is analyzed. The literature reports that e-customers between the ages of 18 and 50 are the most active [32], with young adults being prevalent [33,34], mainly driven by convenience and interactivity [35]. Older adults value product selection and post-consumption for repurchases [36], while younger adults do. The studies on gender suggest an inconclusive contribution [37,38], reporting both significant and non-significant findings. However, some suggest that women are "shopping for fun" while males emphasize "quick shopping" behaviors [39]. Additionally, Naseri and Elliott [40] indicate that education significantly influences the adoption of online shopping [41], while also income status [42] plays an important role, where both income and education tend to have a higher probability of online shopping [43,44]. Lastly, employment (work status) determines purchase behaviors [45–47]. A more comprehensive list of demographics that may interest the reader is described elsewhere [48,49].

Major challenges still exist. Although proposed typologies, clusters or classes of user profiles can be successfully mapped into different user profiles, the biggest issue is that most clusters and user profiles overlap and are not mutually exclusive. Also, from previous studies, there is a lack of evidence and methods describing the actual contribution of each variable (features) in describing contribution or shared variance. In this study, unsupervised modelling of user profiles for obtaining cluster labels is performed. Machine Learning (ML) predicts user profiles and identifies the most important features.

## 3. Materials and Methods

### 3.1. The Multistage Model of the Data Workflow Framework

The data workflow framework consists of three stages (Figure 1). The first stage explains a priori sample determination, survey development and data collection. Also, this stage explains the research hypothesis framework for reducing a raw dataset to a dataset comprising only statistically significant associations. Lastly, the first stage explains an in-depth statistical analysis regarding the statistically significant association between demographics and user preferences.

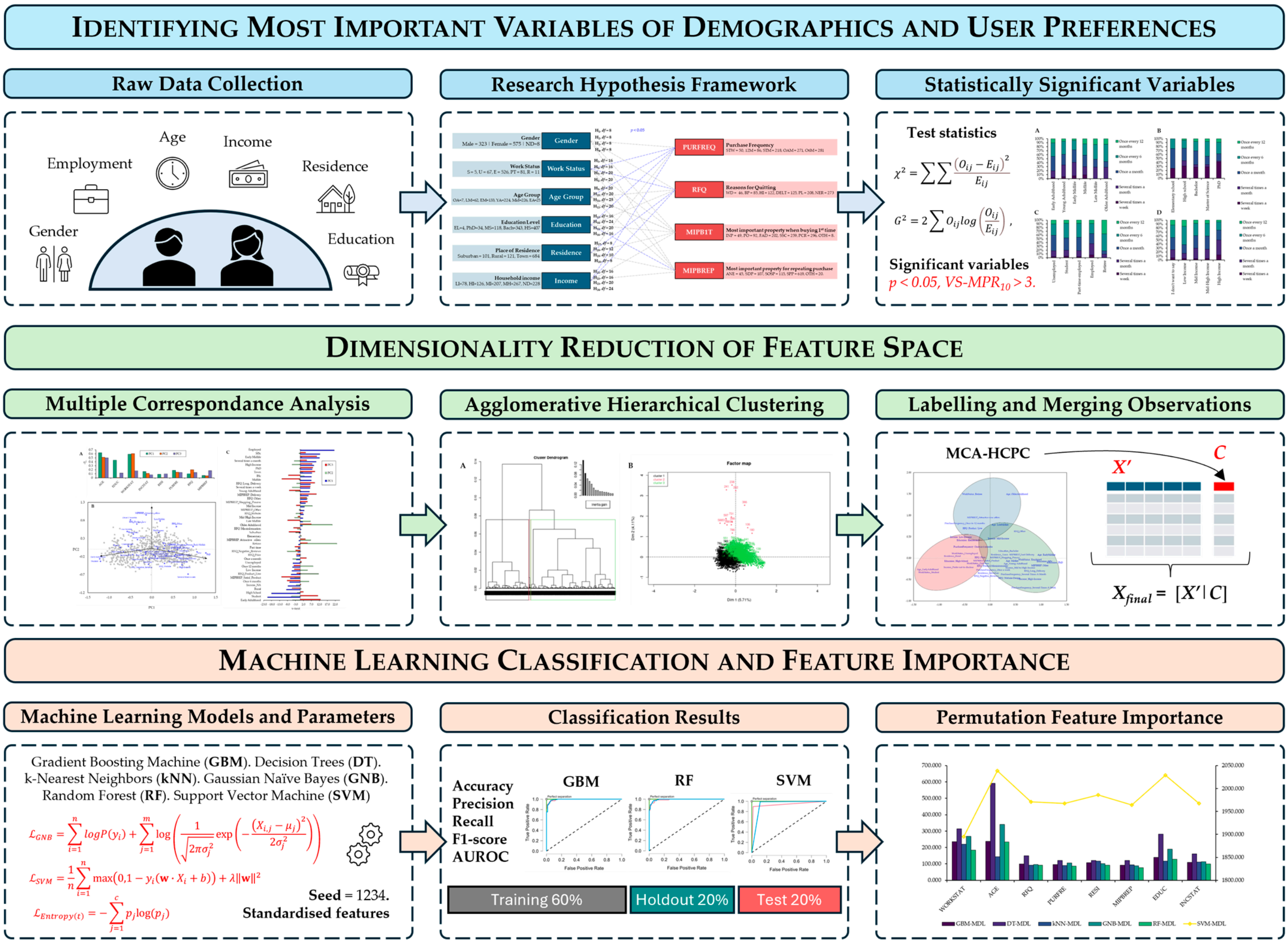



**Figure 1.** The data workflow framework.

The second stage explains the Multiple Correspondence Analysis (MCA) procedure for reducing the dimensionality of a dataset into two-dimensional space. Next, the study extracts relevant features by including only the top two Principal Components that explain the most variation, i.e., inertia. Also, the Agglomerative Hierarchical Clustering (AHC) is performed on MCA's Principal Components to obtain clusters of user preferences. The class label vector is merged with the raw dataset after identifying class labels, i.e., user profiles.

The third segment proposes and explains the utilized ML classifiers, i.e., ML algorithms, setting the parameters and loss functions used for the analysis. The second part of the third stage describes the main classification metrics used for evaluating the model performance and ML classification performance metrics—accuracy, precision, recall, F1-score and area under receiver operating characteristic (AUROC). Finally, the last part of the stage includes the extraction of the most relevant features using Permutation Feature Importance (PFI) [50] that builds upon the Mean Dropout Loss (MDL) metric [51].

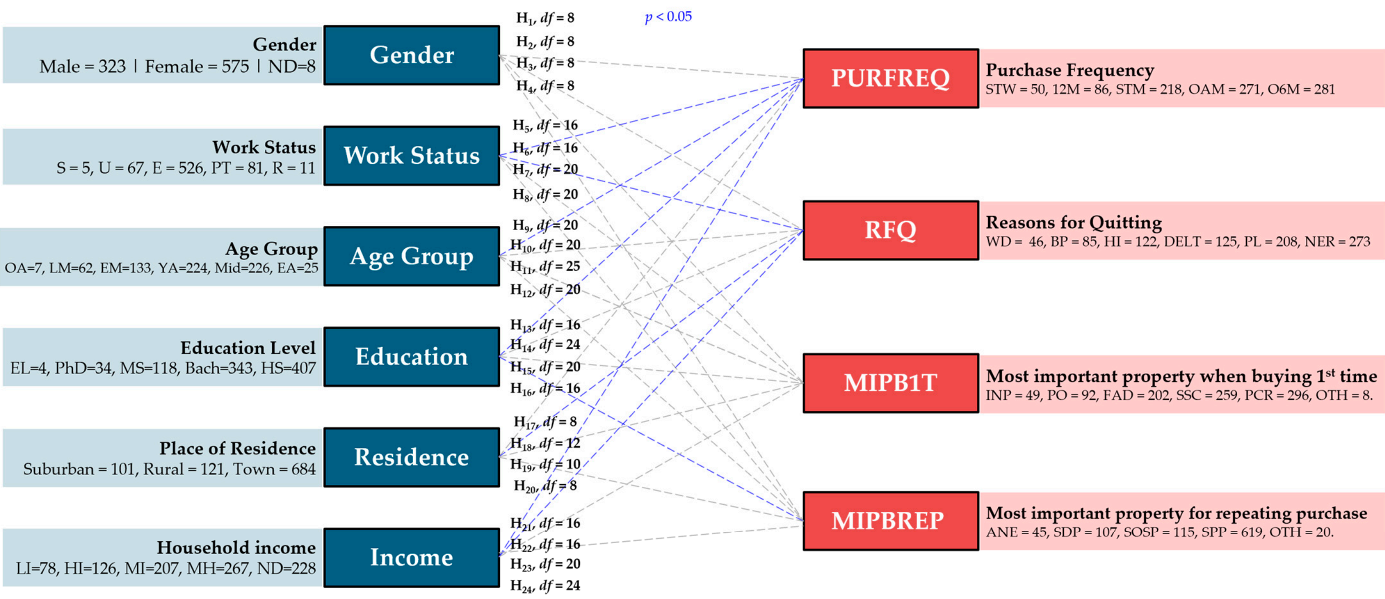
### 3.2. Data Collection and Sample Size

The research was realized during the four weeks of November–December 2022 using two strategies for securing a representative sample. The survey was created on the Google Forms platform and distributed exclusively online. The target group was adult citizens of the Republic of Serbia who had experience ordering goods and durable products via the Internet and delivering them to a specific location. The survey included respondents with experience delivering goods such as electrical appliances, clothing, footwear, furniture, tools, small home and yard use items, sports equipment, and similar items. Before participating in the survey, respondents gave their consent to participate. They were informed about the research objectives and that their answers would be anonymous and analyzed in groups. They were informed that they could stop participating at any time.

A priori sample size was determined as follows. Firstly, the sample size is determined based on the minimum required sample size for performing $\chi^2$ test statistics. To do so, a G*Power (v.3.1.9.6) is used for calculating sample size per parameters: Effect size $w = 0.3$ (medium effect), $\alpha = 0.05$, power ($1 - \beta$ error probability) = 0.80 and $df = 36$ (determined as the most of 7 categories per variables $j$ and $k$, such that $df = (j-1)(k-1)$). The output statistic shows non-centrality parameter $\lambda = 26.46$, with ${\chi^2}_{crit} = 50.998$, actual power ($1 - \beta$ error probability) = 0.801, and minimum sample size of $n = 294$. Secondly, the sample size is calculated per Hamburg [52]—a commonly used sample size calculator can be found online at www.calculator.net (accessed on 1 November 2024). The minimum required representative sample is $n = 385$. Lastly, given that the sample size from previous similar studies [39,41] was higher than needed, data collection at $n = 906$ respondents was stopped.

### 3.3. The Research Hypothesis Framework

The research hypothesis framework (Figure 2) is designed to test the conditional dependencies, i.e., the existence of an association between different demographic properties of e-consumers and user preferences (and behavior). To do so, variables primarily reported in previous studies [32,53] regarding demographic properties are used—gender, work status, age group, education, place of residence, and income. Next, a test of association considering PURFREQ (purchase frequency), MIPB1T (most important property when buying the first time), MIPBREP (most important property before repeating the purchase), and RFQ (reasons for quitting the online purchase) is performed.




**NOTE:** ND = Prefer not to disclose; S = Student; U = Unemployed; E = Employed; PT = Part Time Employed; R = Retired; OA = Older Adulthood; LM = Late Midlife; EM = Early Midlife; YA = Young Adulthood; Mid = Midlife; EA = Early Adulthood; EL = Elementary School; PhD = Having PhD degree; MS = Master of Science; Bach = Bachelor degree; HS = High School degree; LI = Low Income; HI = High Income; MI = Mid Income; MH = Mid to High Income; STW = Several Times a Week; 12M = Once every 12 months or less; STM = Several Times a Month; OAM = Once a Month; O6M = Once every 6 months; WD = Due to website design; BP = Due to the need for better price; HI = Due to inappropriate and hidden information about the product; DELT = Due to long delivery time; PL = Because I want to see the product live; NER = Because of the negative reviews; INP = To offer the option of in-store pickup; PO = To offer the payment option I prefer; FAD = Fast and accurate delivery; SSC = To have a secure shopping certificate; PCR = To have positive customer reviews; OTH = Other; ANE = Attractive new offers; SDP = Satisfaction with the delivery of products; SOSF = Satisfaction with online shopping process; SPP = Satisfaction with purchased products.

**Figure 2.** The research hypothetical framework.

For hypothesis testing, Pearson's $\chi^2$ test statistic is chosen, described as follows:

$$\chi^2 = \sum\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad (1)$$

such that $Oij$ is observed frequencies, $E_{ij}$ is expected frequencies computed as $R_i \times C_j/N$, where $R_i$ and $C_j$ represent row and column marginals, while $N$ is the total observations. However, the $G^2$ likelihood ratio is reported using the following equation:

$$G^2 = 2\sum O_{ij} log\left(\frac{O_{ij}}{E_{ij}}\right), \quad (2)$$

as a more robust method for evaluating the goodness-of-fit and association between variables in contingency tables, given that the $\chi^2$ statistic is only approximated by the $\chi^2$ distribution and worsens when expected frequencies are relatively small, which is a common controversy using $\chi^2$ test statistic, $G^2$ is included because it provides more robust measurements for large dimensional tables. It is commonly discussed that $G^2$ advantage over traditional $\chi^2$ is that $G^2$ for large contingency tables can be neatly decomposed into smaller components [54], which cannot be achieved by the $\chi^2$ test. Even so, as the sample size increases, statistics tend to converge. The degrees of freedom $df$ is determined by $df = (r - 1)(c - 1)$, where $r$ and $c$ represent classes of row and column profiles, respectively.

Furthermore, the diagnosticity of $p$ values using the VS–MPR (Vovk–Sellke Maximum $p$-Ratio) calibration score is performed. The score is computed as VS–MPR = $-e \times p \times \ln(p)$ and is commonly referred to as a lower bound to BF (favoring $H_0$ over $H_1$) [55]. The reason for including VS–MPR is that a large body of research [56–58] is calling into question the traditional threshold ($p < 0.05$) for deciding whether a model is statistically significant, i.e., rejects the null hypothesis. The following labelling intervals are used as evidence in favor of the alternative over the null hypothesis: VS–MPR$_{10}$ = 1–3 anecdotal, VS–MPR$_{10}$ = 3–10 substantial, VS–MPR$_{10}$ = 10–30 strong, VS–MPR$_{10}$ = 30–100 very strong, and finally VS–MPR$_{10}$ > 100 as decisive evidence [59].

### *3.4. Multiple Correspondence Analysis Hierarchical Clustering of Principal Components*

To understand MCA's description, let the initial raw dataset be a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times m}$, where $n$ is the number of observations and $m$ is the number of variables that explain demographics and user preferences. To reduce the dimensionality of dataset $\boldsymbol{X}$, hypothesis testing is performed (see Equations (2) and (3)) to keep only statistically significant variables ($p < 0.05$, VS–MPR$_{10}$ > 3). After removing non-significant variables, the reduced $\mathrm{X}' \subset \mathrm{X}$ dataset comprises only significant variables. Hence, the $\boldsymbol{X'}$ is subjected to MCA to further reduce dimensionality by transforming raw data into Principal Components (PCs).

The first step in MCA is to convert a raw categorical dataset $\boldsymbol{X'}$ into indicator matrix $\boldsymbol{Z} \in \mathbb{R}^{n\times q}$, where $q$ is the number of class categories from the retained variable set. Next, from the defined indicator matrix $\boldsymbol{Z}$, centring is performed by subtracting row, and column means to obtain the centered matrix $\boldsymbol{Z}_C$:

$$\boldsymbol{Z}_{C,\,ij} = \boldsymbol{Z_{ij}} - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z_{ij}} - \frac{1}{q}\sum_{i=1}^{n}\boldsymbol{Z_{ij}} + \frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{q}\boldsymbol{Z_{ij}}, \quad (3)$$

which adjusts for marginal distributions, ensuring a standardized matrix for PC extraction. Using Singular Value Decomposition (SVD), $\boldsymbol{Z_C}$ is decomposed as follows:

$$\boldsymbol{Z_C} = \boldsymbol{U\Sigma V^T}, \quad (4)$$

where $\boldsymbol{U} \in \mathbb{R}^{n\times k}$ are singular vectors in the reduced space, $\boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}$ is a diagonal matrix with singular values of each PC, and $\boldsymbol{V} \in \mathbb{R}^{q\times k}$ is the singular vector matrix representing variable contributions of each PC. The top $k$ components are selected based on the first two PCs that explain the most variance as PCs $\in \mathbb{R}^{n \times k}$.

Finally, after obtaining reduced dataset PCs, an HCPC is performed to identify clusters, i.e., label user profiles in a reduced space. For performing distance computation and obtaining distance matrix $\boldsymbol{D}$, an Euclidian distance $\boldsymbol{D}(i,j)$ between each pair of observations in PCs is performed:

$$\boldsymbol{D}(i,j) = \left\| PCs_i - PCs_j \right\|^2, \quad (5)$$

which yields distance matrix $\boldsymbol{D} \in \mathbb{R}^{n \times n}$, capturing similarities between respondents. Ward's method is selected to apply hierarchical agglomerative clustering on PCs to merge clusters and minimize within-cluster variance. The results are represented via a dendrogram. Linkage $L$ selects the clusters, assigning labels to observations $C = \{c_1, c_2, \ldots, c_n\}$. As a last step, cluster $\boldsymbol{C}$ labels are merged with dataset $X'$, resulting in the final dataset $\boldsymbol{X_{final}} = [\boldsymbol{X'} \mid \boldsymbol{C}]$, which is then subjected to classification by ML algorithms.

### 3.5. Machine Learning Classifiers

Machine Learning classification is performed as follows. Let dataset $\boldsymbol{X_{final}} \in \mathbb{R}^{n \times (m+1)}$ be the final dataset, where $n$ is the number of observations and $m$ represents the features (demographics and user preferences), while the last column, denoted as $y \in \{1, 2, \ldots, c\}$, represents the cluster labels assigned via MCA-HCPC, such that $c$ represents a total number of selected clusters.

The splitting of the dataset into train (with holdout)/test dataset is performed, such that $\boldsymbol{X_{train}} \in \mathbb{R}^{n_{train} \times m}$ be training matrix with $y_{train} \in \{1, 2, \ldots, c\}^{n_{train}}$ labels. Similarly, let $\boldsymbol{X_{test}} \in \mathbb{R}^{n_{test} \times m}$ with $y_{test} \in \{1, 2, \ldots, c\}^{n_{test}}$ be the test dataset. Specifically, the models are trained with train/validation/test per 60%/20%/20% split from the $\boldsymbol{X_{final}}$ dataset. We aim to train the ML classifier $\mathcal{M}$: $\mathbf{X_{train}} \rightarrow y$ by optimizing loss function $\mathcal{L}$. For proposed ML classifiers—Gradient Boosting Machine (GBM), Decision Tree (DT), k-Nearest Neighbors (kNN), Gaussian Naïve Bayes (GNB), Random Forest (RF), and Support Vector Machine (SVM)—the loss functions are defined for GBM:

$$\mathcal{L}_{GBM} = -\sum_{i=1}^{n} \sum_{j=1}^{c} y_{i,j} \log\left(\hat{y}_{i,j}\right), \quad (6)$$

where in multi-class problem $y_{i,j}$ is a binary indicator (1 if label $j$, otherwise 0), and $\hat{y}_{i,j}$ is the predicted probability for class $j$. For DT and RF, a Gini impurity is estimated as:

$$\mathcal{L}_{Gini(t)} = 1 - \sum_{j=1}^{c} p_j^2, \quad (7)$$

and Entropy as:

$$\mathcal{L}_{Entropy(t)} = -\sum_{j=1}^{c} p_j \log\left(p_j\right), \quad (8)$$

where algorithms select the feature that maximizes the reduction in Gini (or Entropy). For Gaussian Naïve Bayes, the classifier calculates the posterior probability for each class:

$$\mathcal{L}_{GNB} = \sum_{i=1}^{n} logP(y_i) + \sum_{j=1}^{m} \log\left(\frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{\left(X_{i,j} - \mu_j\right)^2}{2\sigma_j^2}\right)\right) \quad (9)$$

where $P(y_i)$ is the prior probability of class $y_i$, while $\mu$ is the mean and $\sigma^2$ is the variance of feature $j$ conditioned on $y_i$. The SVM loss function is determined per Hinge loss:

$$\mathcal{L}_{SVM} = \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i(\mathbf{w} \cdot X_i + b)) + \lambda \|\mathbf{w}\|^2, \quad (10)$$

where $y_i \in \{-1, 1\}$ represents the true class label, $\mathbf{w}$ is the weight vector, while $X_i$ is the feature vector, $b$ is the bias, and $\lambda$ is the regularization parameter controlling the margin. Given the classification problem, the selection of the highest-performing algorithm is conducted per classification evaluation metrics of accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{11}$$

where $TP$ = true positive, $TN$ = true negative, $FP$ = false positive, and $FN$ = false negative. The recall is calculated as:

$$Recall = \frac{TP}{TP + FN}\ . \tag{12}$$

The precision is estimated as follows:

$$Precision = \frac{TP}{TP + FP}\ , \tag{13}$$

and finally *F1-score* is calculated as:

$$F1 - score = \frac{2\ TP}{2\ TP + FP + FN}\ , \tag{14}$$

for estimating the classification performance of ML algorithms. Additionally, an AUROC is considered by adopting it to multi-class problems.

To ensure transparency and replicability of the results, the following parameters (and hyperparameters) are set for the ML models. For GBM, the following parameters are set: Shrinkage = 0.1, with 1.0 interaction depth with a minimum number of observations per node = 10. The training used per tree is 50%. The number of trees is optimized such that the maximum number is 100. For the DT algorithm, the minimum number of observations for split is 20, with a minimum number of observations in terminal = 7, with maximum interaction depth = 30. The DT tree complexity is optimized with a maximum complexity penalty = 1.0. The kNN algorithm settings are set as follows: weights = Rectangular with Euclidian distance metric used. The number of nearest neighbors is optimized such that the maximum allowed nearest neighbors = 10. For the RF algorithm, the training data used per tree are 50%, while features are split automatically with an optimized number of trees set to a maximum of 100 trees. The SVM algorithm is set with the following parameters: Weights = Linear, Tolerance of termination criterion = 0.001, with function $\xi$ = 0.01. The costs of constraints violation are optimized with maximum violation cost = 5. There were no smoothing parameters set for GNB algorithms. However, all algorithms contain scaled features, and the seed is 1234. Lastly, to ensure the reliability of the analysis, i.e., to avoid overfitting and underfitting, a 10-fold cross-validation is performed of the training (with validation) dataset. To ensure objectiveness and replicability in testing the results of individual classifiers, the parameters (and hyperparameters) used can be input and replicated in JASP v0.19.1.

## 4. Results

### *4.1. Descriptive Statistics*

The characteristics (Figure 3) show the following. This study comprises 906 participants, ranging from 21 to 77 years old (median = 34.90, SD = 12.67). Most were female (63%, $n$ = 575), while <1% of participants selected "preferred not to disclose" ($n$ = 8). Most of the respondents had completed high school (44.92%, $n$ = 407), followed by a Bachelor's degree (37.86%, $n$ = 343) and a Master's degree (13.02%, $n$ = 118). The respondents were primarily situated in towns (75.5%, $n$ = 684), followed by rural (13.36%, $n$ = 121) and suburban areas (11.15%, $n$ = 101). The employment characteristics show that 58% ($n$ = 526) of respondents are employed. The income status is classified according to monthly earnings of RSD (Republic Serbia Dinar) in categories as low income (<50.000 RSD), mid income (50.000–100.000 RSD), mid-high income (100.000–200.000 RSD), high income (>200.000 RSD). The results show that most respondents had mid-high income (29.47%, $n$ = 267), 25.17% ($n$ = 228) did not want to say their income, followed by mid income (22.85%, $n$ = 207), high income (13.91%, $n$ = 126), and low income (8.61%, $n$ = 78). Lastly, the age distribution shows average AVGAge = 35.914 with SDAge = 12.671, ranging from 21 to 77 years old. The age distribution is then coded as follows: early adulthood (18–24 years), young adulthood (25–34 years), early midlife (35–44 years), midlife (45–54 years), late midlife (55–64 years), older adulthood (>65 years). As per coded categories, the descriptives show that early adulthood (28.04%, $n$ = 254) is the most dominant category, followed by midlife (24.94%, $n$ = 226) and young adulthood (24.72%, $n$ = 224), early midlife (14.68%, $n$ = 133), late midlife (6.84%, $n$ = 62), and older adulthood (0.77%, $n$ = 7).

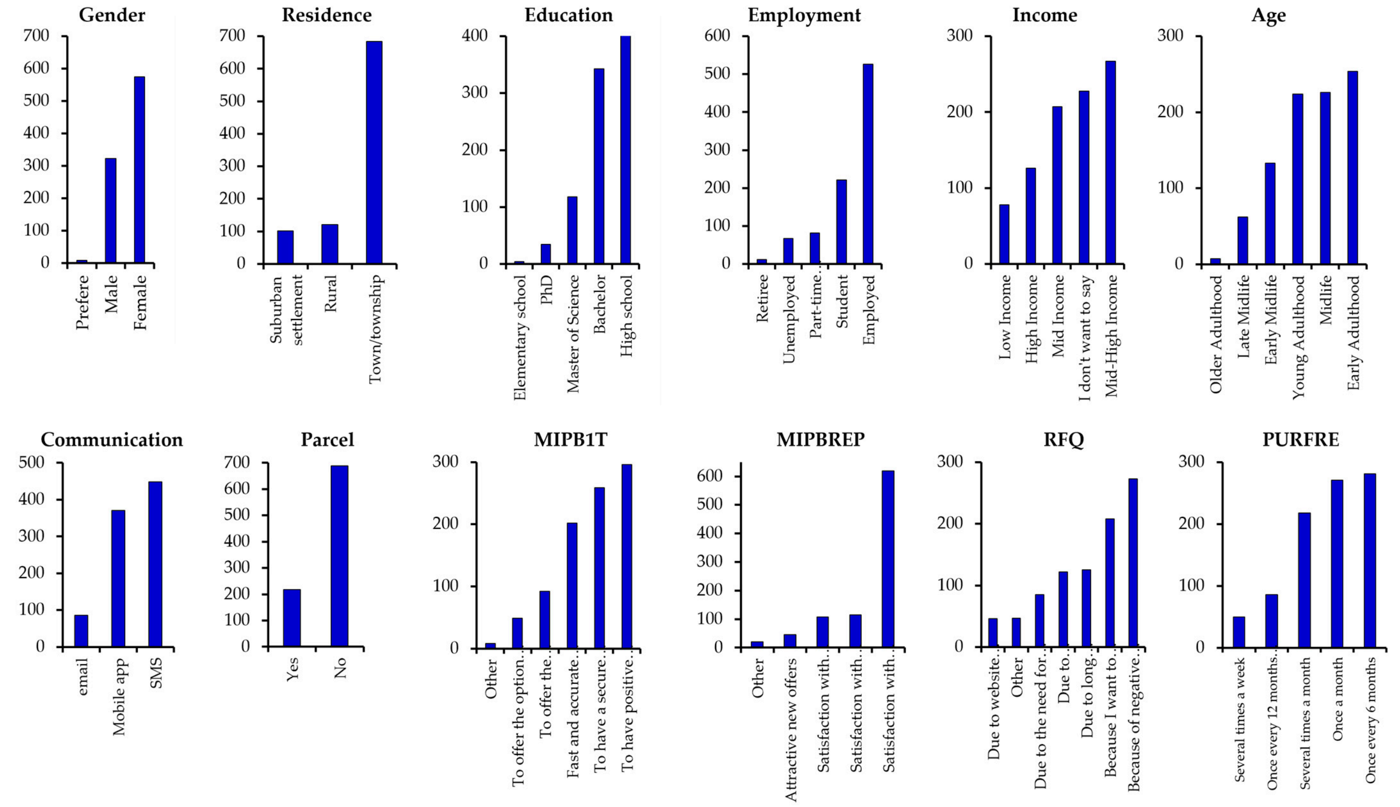



**Figure 3.** Descriptive statistics of demographic data (**top row**) and user preferences (**bottom row**).

Supplementary items such as "What communication channel would you prefer to have with the supplier?" labelled as Communication, and "Used parcel locker?" show that most users prefer SMS (49.56%, $n$ = 449) and Mobile App (40.95%, $n$ = 371) and uses parcel locker (24.06%, $n$ = 218), respectively. The main preferences, such as MIPB1T, show that most users consider positive reviews (32.67%, $n$ = 296) when buying the first time, followed by "To have a secure shopping certificate" (28.59%, $n$ = 259), and "Fast and secure delivery" (22.30%, $n$ = 202). The MIPBREP property suggests that "Satisfaction with products purchased" (68.32%, $n$ = 619) is the most dominant reason for repeating the purchase, followed by "Satisfaction with the online shopping (12.69%, $n$ = 115), "Satisfaction with the delivery of purchased products" (11.81%, $n$ = 107), and others. The RFQ shows that negative reviews (30.13%, $n$ = 273), followed by "Because I want to see the product live" (22.96%, $n$ = 208) and "Long delivery time" (13.80%, $n$ = 125) are the most common reasons why users quiet their purchase. Lastly, the PURFRE show that most users buy every six months (31.02%, $n$ = 281), followed by "Once a month" (29.91%, $n$ = 271).

Contingency tables are constructed before performing the statistical inferential statistics and $\chi^2$ tests to investigate the association between variables. However, due to the limited length and extensive number of tables (24 tables) for each model comparison, the complete analysis with tables is provided in Supplementary Materials in cases the reader is interested in following the full analysis. In the following, the final results of each model comparison using both $\chi^2$ and $G^2$ test statistics are provided, including the effect size (for $\chi^2$ test) and VS–MPR ratio.

### *4.2. Hypothesis Testing*

After performing hypotheses testing as a proposed framework, 8/24 hypotheses regarding the conditional dependencies were rejected, i.e., the association between e-customer demographics and user preferences (Table 1). Namely, only cases where both $\chi^2$ and $G^2$ suggest statistically significant association ($p < 0.05$) were considered and where there is at least moderate, i.e., substantial evidence in favor of the alternative hypothesis (VS–MPR$_{10}$ > 3). In such instances, results will be more robust and reliable. Additionally, non-significant test results are reported in the appendices (Table A1).

**Table 1.** Statistical analysis reported per $\chi^2$ test.

| Variables | Test | Value | *df* | *p* | Cramer's *V* | VS–MPR |
|---|---|---|---|---|---|---|
| AGE—PURFRE | $\chi^2$ test statistic | 50.519 | 20 | 0.001 | 0.118 | 229.606 |
| | $G^2$ likelihood ratio | 54.838 | 20 | 0.001 | | 843.603 |
| EDU—PURFRE | $\chi^2$ test statistic | 38.498 | 16 | 0.001 | 0.103 | 43.022 |
| | $G^2$ likelihood ratio | 39.004 | 16 | 0.001 | | 49.631 |
| RESI—RFQ | $\chi^2$ test statistic | 24.599 | 12 | 0.017 | 0.117 | 5.349 |
| | $G^2$ likelihood ratio | 25.710 | 12 | 0.012 | | 7.025 |
| EDU—MIPBREP | $\chi^2$ test statistic | 33.834 | 16 | 0.006 | 0.097 | 12.456 |
| | $G^2$ likelihood ratio | 30.371 | 16 | 0.016 | | 5.516 |
| INCSTAT—PURFRE | $\chi^2$ test statistic | 52.421 | 16 | 0.001 | 0.120 | 3392.513 |
| | $G^2$ likelihood ratio | 49.376 | 16 | 0.001 | | 1221.498 |
| INCSTAT—RFQ | $\chi^2$ test statistic | 41.019 | 24 | 0.017 | 0.106 | 5.413 |
| | $G^2$ likelihood ratio | 41.087 | 24 | 0.016 | | 5.484 |
| WORKSTAT—PURFRE | $\chi^2$ test statistic | 51.892 | 16 | 0.001 | 0.120 | 2834.207 |
| | $G^2$ likelihood ratio | 50.487 | 16 | 0.001 | | 1766.807 |
| WORKSTAT—RFQ | $\chi^2$ test statistic | 50.697 | 24 | 0.001 | 0.118 | 47.147 |
| | $G^2$ likelihood ratio | 54.134 | 24 | 0.001 | | 115.292 |

Surprisingly, no evidence suggested that demographic properties are conditionally dependent, i.e., associated with MIPB1T ($p > 0.05$). There is, however, anecdotal evidence (VS–MPR$_{10}$ = 2.386) suggesting that age ($\chi^2 = 37.455$, $p = 0.052$) tends to be associated with MIPB1T but fails to reject the null ($p > 0.05$). Hence, the null hypothesis that indicates a user preference describing the "Most important property when buying from a webshop for the first time?" failed to be rejected.

The results investigating the association of demographics to the PURFRE variable suggests that age ($\chi^2 = 50.519$, $G^2 = 54.838$, $p = 0.001$), education (*03C7*$^2$ = 38.498, $G^2 = 39.004$, $p = 0.001$), income ($\chi^2 = 52.421$, $G^2 = 49.376$, $p = 0.001$), and work status ($\chi^2 = 51.892$, $G^2 = 50.487$, $p = 0.001$), ranging from very strong (VS–MPR$_{10}$ > 30) in education, to decisive evidence (VS–MPR$_{10}$ > 100) of age, income and work status in associations with PURFRE. Next, the analysis of associations with RFQ suggests that work status provides very strong evidence ($\chi^2 = 50.697$, $G^2 = 54.134$, $p = 0.001$, VS–MPR$_{10}$-$\chi^2$ = 47.147), followed by substantial evidence regarding residence ($\chi^2 = 24.599$, $G^2 = 25.710$, $p = 0.017$, VS–MPR$_{10}$-$\chi^2$ = 5.349) and income ($\chi^2 = 41.019$, $G^2 = 41.087$, $p = 0.017$, VS–MPR$_{10}$-$\chi^2$ = 5.413). Lastly, the test considering MIPBREP suggests only the existence of association to education with substantial ($G^2 = 30.371$, $p = 0.016$, VS–MPR$_{10}$ = 5.516) to strong ($\chi^2 = 33.84$, $p = 0.006$, VS–MPR$_{10}$ = 12.456) evidence in favor of the alternative.

The results suggest statistical dependency between demographics and user preferences (excluding MIPB1T). However, more needs to be understood about how specific underlying classes associate with each other. Hence, MCA uses the $\chi^2$ distance better to understand the relationship between investigated demographics and user preferences.

### *4.3. Multiple Correspondence Analysis*

The MCA analysis suggests an improved understanding of the association between class categories. Namely, the explained inertia is 11.797% in the first two components and 16.1% in the first three (Figure 4A). Next, the $\eta^2$ correlation coefficient (Figure 4B), representing the degree of association between variables and Principal Components, suggests that all three components explain the AGE variable well. The second component mainly captures AGE ($\eta^2$ = 0.516) and WORKSTAT ($\eta^2$ = 0.607), while user preferences include PURFRE ($\eta^2$ = 0.141) and RFQ ($\eta^2$ = 0.201). Lastly, AGE ($\eta^2$ = 0.499), EDU ($\eta^2$ = 0.119) and WORKSTAT ($\eta^2$ = 0.175) consider demographic variables explained by the PC3, with PURFRE ($\eta^2$ = 0.135), RFQ ($\eta^2$ = 0.134) and MIPBREP ($\eta^2$ = 0.181) user preferences captured. Most of the information is well explained by the first two components, with the MIPBREP variable slightly higher $\eta^2$. Hence, the association between variables is discussed within the first two PCs.

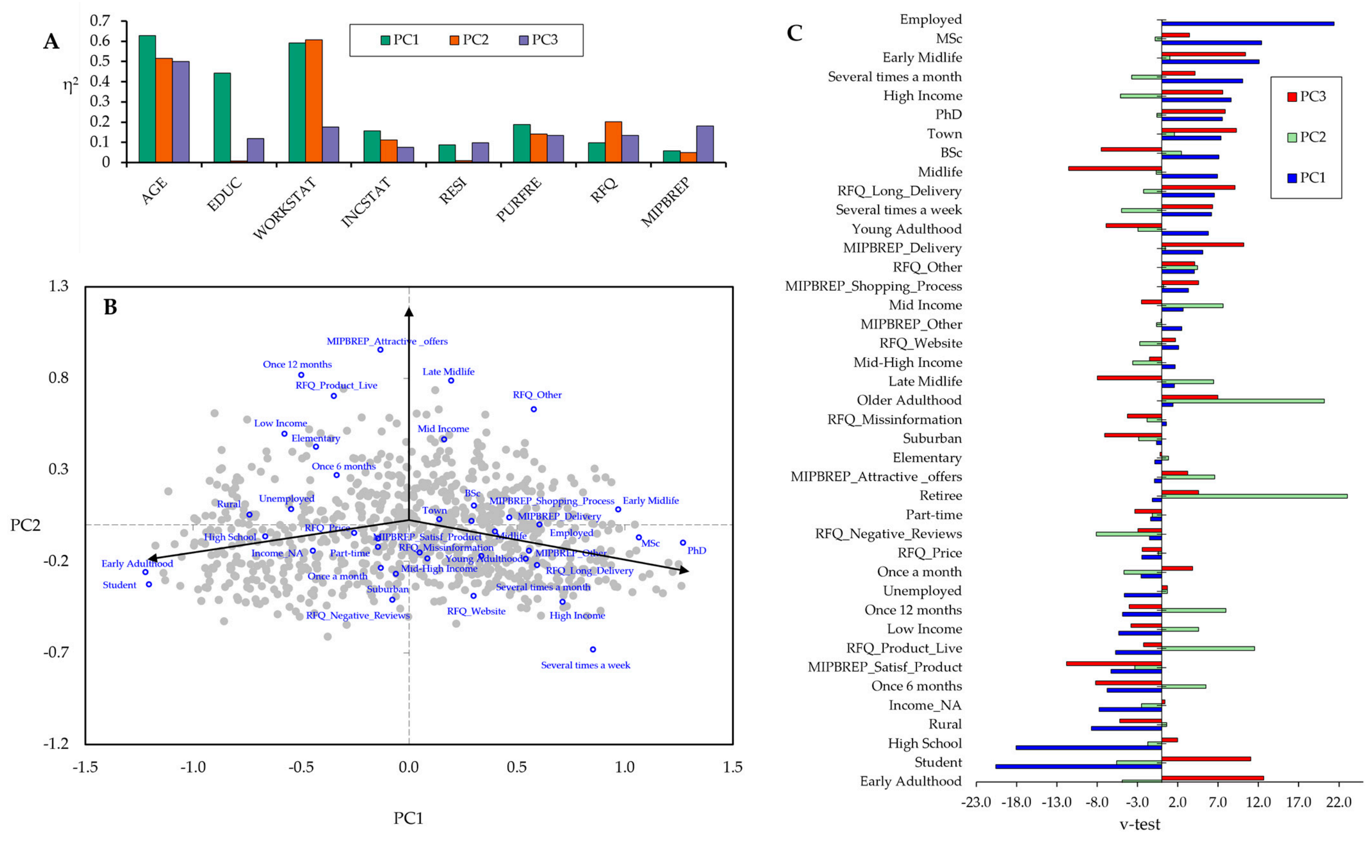


**Figure 4.** MCA analysis including (**A**) $\eta^2$ coefficient of categories concerning PCs; (**B**) MCA biplot of respondents (grey color) and class categories of categorical variables; (**C**) v-test score of class categories ($z > 1.96$, $z < -1.96$).

The MCA biplot (Figure 4C) identifies three latent projections describing demographics and user preferences. Namely, the vertical mainly suggests late midlife to older adults, low to mid income, and with purchasing frequency from 6 to 12 months, who prefer person-to-person purchase, i.e., seeing the product live, which is mainly described by the PC2. Observing the negative side of PC1(−) and PC2(−), i.e., the left bottom quadrant, mainly describes students, high school education, early adults, unemployed and part-time employed respondents, rural areas, and with purchasing frequency from once a month to once in six months. This can also be supplemented by the v-test (Figure 4C) score, as it mainly quantifies the class category distance from the average. Hence, PC1-PC3 can offer similar information on class categories in factor plots.

Lastly, the bottom right quadrant, captured mainly by the PC1(+) inertia, describes the respondents as early midlife and employed mostly with an MSc degree, with town residence and high income. These respondents are characterized by user preferences describing purchasing power from several times a week (CTR = 1.78, $\cos^2$ = 0.04, v-test = 6.20) to several times a month (CTR = 3.77, $\cos^2$ = 0.11, v-test = 10.6), while most crucial property before repeating the purchase (MIPBREP) being the satisfaction with the delivery (CTR = 1.13, $\cos^2$ = 0.029, v-test = 5.12) and satisfaction with the online shopping process (CTR = 0.474, $\cos^2$ = 0.012, v-test = 3.33) while the main reason for quitting the purchase is long delivery (CTR = 1.80, $\cos^2$ = 0.05, v-test = 6.52). The full list of coordinates for PCs, including cos2, CTR, and v-test, is provided in Table A4.

### 4.4. Classification Results

The training (and holdout set) classification accuracy (Table 2) suggests that the highest classification accuracy is obtained via SVM (0.950), GBM (0.939) and RF (0.928). The SVM suggests that classification metrics are mostly consistent across precision, recall, F1-score and AUROC. However, the GNB, instead of RF, show high precision, recall and F1-scores, while AUROC results suggest that GBM (0.994) and RF (0.994) are the highest.

**Table 2.** Machine learning classification metrics (* represents results of 10-fold cross-validation).

| Algorithm | Accuracy | Precision | Recall | F1-Score | AUROC |
|---|---|---|---|---|---|
| **GBM** | **0.959 (* 0.956)** | 0.943 (* 0.938) | 0.939 (* 0.934) | 0.936 (* 0.930) | **0.994 (* 0.994)** |
| **DT** | 0.917 (* 0.876) | 0.902 (* 0.878) | 0.917 (* 0.898) | 0.909 (* 0.863) | 0.894 (* 0.882) |
| **kNN** | 0.884 (* 0.930) | 0.870 (* 0.881) | 0.884 (* 0.895) | 0.876 (* 0.887) | 0.796 (* 0.802) |
| **GNB** | 0.939 (* 0.911) | 0.942 (* 0.914) | 0.939 (* 0.930) | 0.936 (* 0.928) | 0.843 (* 0.836) |
| **RF** | 0.928 (* 0.928) | 0.914 (* 0.914) | 0.928 (* 0.928) | 0.920 (* 0.920) | **0.994 (* 0.994)** |
| **SVM** | 0.950 (* 0.936) | **0.954 (* 0.944)** | **0.950 (* 0.942)** | **0.949 (* 0.940)** | 0.902 (* 0.886) |

From a cluster of user profiles, the labelling shows a significant dataset imbalance; therefore, accuracy results can also be misleading, although precision and recall metrics are useful when false positives and false negatives are problematic, respectively. F1-score and AUROC are particularly useful since they address the case of the imbalanced dataset. Therefore, the results from the holdout set show that GBM (0.959), SVM (0.950), GNB (0.939) and RF (0.928) are the highest-performing classifiers. The results after the 10-fold cross-validation report are similar: GBM (0.956), SVM (0.936), RF (0.928) and GNB (0.911).

The classification results (Figure 5A) show that GBM (0.994), RF (0.994), and SVM (0.902, 10FCV = 0.886) are the highest-performing classifiers. Although the interest here is in PFI rankings, MDL scores of all ML classifiers are provided. Overall, one can conclude that age plays a crucial role in ML classifiers, followed by work status, education status, and income status. Still, the highest-performing ML classifiers suggest that age (GBM, RF, and SVM) is the most important feature, followed by work status (GBM, RF), education (GBM, RF, SVM) and residence (SVM). Hence, ensemble learners (GBM, RF) offer similar results, while SVM suggests similar results in decision boundaries.

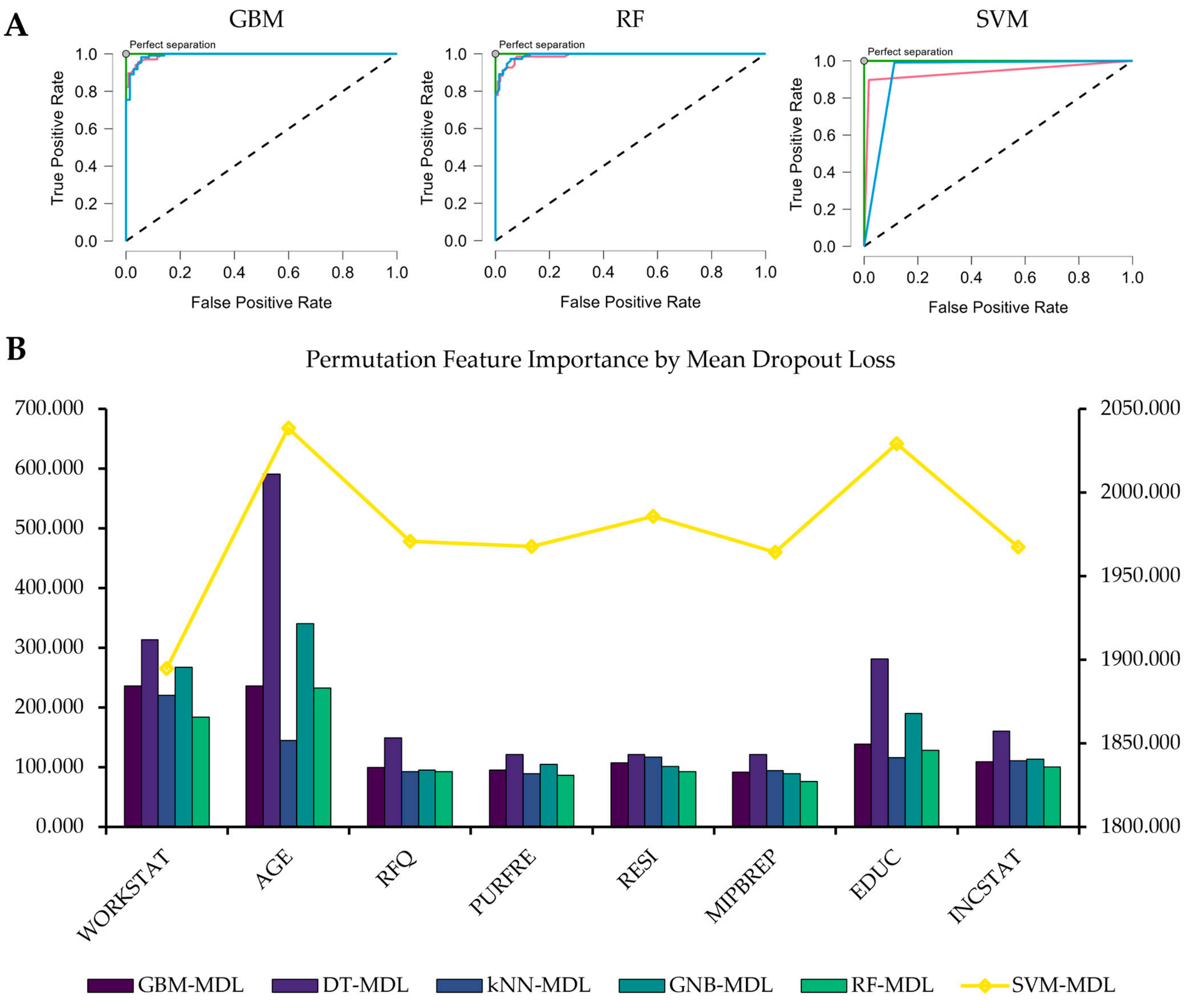



**Figure 5.** Machine Learning Classification of (**A**) Receiver Operating Characteristic Curve representing Cluster 1 (red), Cluster 2 (green) and Cluster 3 (blue), and (**B**) Permutation Feature Importance estimated by Mean Dropout Loss.

## 5. Discussion

### *5.1. Hypothesis Testing Results*

The following conclusions are obtained. Age indicates a significant association with purchase frequency ($p$ = 0.001, $V$ = 0.118). From a general trend (Figure 6A), early (18–24 years) and young adults (25–34 years) dominate in purchasing from at least once a month to once every six months, presumably due to a combination of comfort with digital technology. However, there is a significant increase in purchasing of early midlife (35–44 years) and midlife (45–54 years) several times a month and even several times a week, which was quite surprising since most of the prior literature report the dominance in frequent purchases of young adults (21–30 years). This may be attributed to the cause that parents tend to shift priorities and behavior (e.g., career, family, investments).

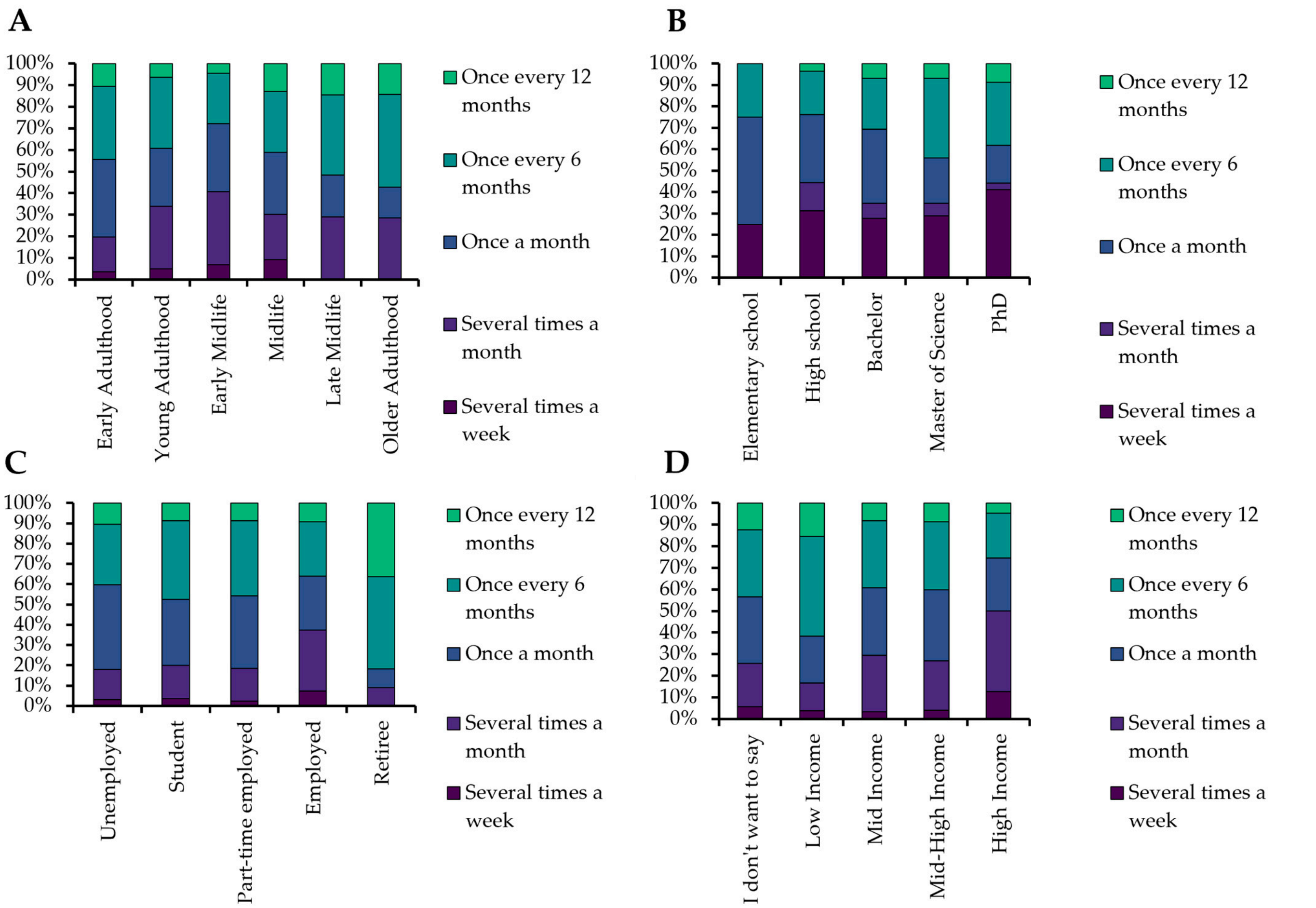



**Figure 6.** The purchase frequencies with corresponding demographics are (**A**) age, (**B**) education, (**C**) work status, and (**D**) income.

Analyzing the association between education and purchase frequency (Figure 6B), the evidence suggests that higher educational attainment corresponds to a more selective behavior. Namely, there is an increase in purchases from several times a month to once a month, while there is an increase in users with small purchases from once in six months to once in twelve months. Master's and PhD degree holders tend to shop less frequently but with more deliberate timing and thoughtful behavior in their purchases. Bachelor's degree holders are slightly more active in online shopping than high school graduates, probably due to higher income levels and better comfort with digital content. The discussion on elementary school education is inconclusive since only four subjects participated in the questionnaire.

The work status (Figure 6C) exhibits similar behavior patterns of unemployed respondents, students and part-time employees. The evidence shows a significant increase in purchase frequency among employed participants. On average, there is a 25.7% increase in purchases several times per month and a 16.5% increase in purchase frequency once a month. In contrast, 6.8% and 4.9% drop in purchases "once a month" and "once every six months", respectively. Lastly, there is a significant drop from 12–25% in purchasing from several times a week to once a month, and a substantial increase in the purchases of "once every six months" (6.9%) and "once every 12 months" (37%) of the retirees.

The results comparing income status and purchase frequency (Figure 6D) show the highest effect ($\chi^2 = 52.421$, $V = 0.120$, VS–MPR$_{10}$ = 3392.513, $G^2 = 49.376$, VS–MPR$_{10}$ = 1221.5) among tested variables, suggesting extreme evidence (VS–MPR$_{10}$ > 100). There is, on average, a 29.97% increase in purchase frequency "several times a week" comparing High with Low to mid-high income respondents. Also, there is a 14.02% increase in purchasing "several times a month" compared to Low to Mid-High income respondents. There is a significant drop of 2.91%, 9.7% and 12.13% in purchase frequency "once a month", "once every six months", and "once every twelve months", respectively.

The association between residence and RFQ (Figure 7A) suggests that suburbans' main reason for quitting is to see the product live (20% higher). At the same time, rural areas state that the main reasons for quitting are inappropriate or hidden information (8–16% higher) and complicated searches on the website (4–20% higher). The prevalent factor across all groups, particularly in rural areas, is the desire to see the product in person, while inappropriate/hidden information is the main reason for quitting the purchase.

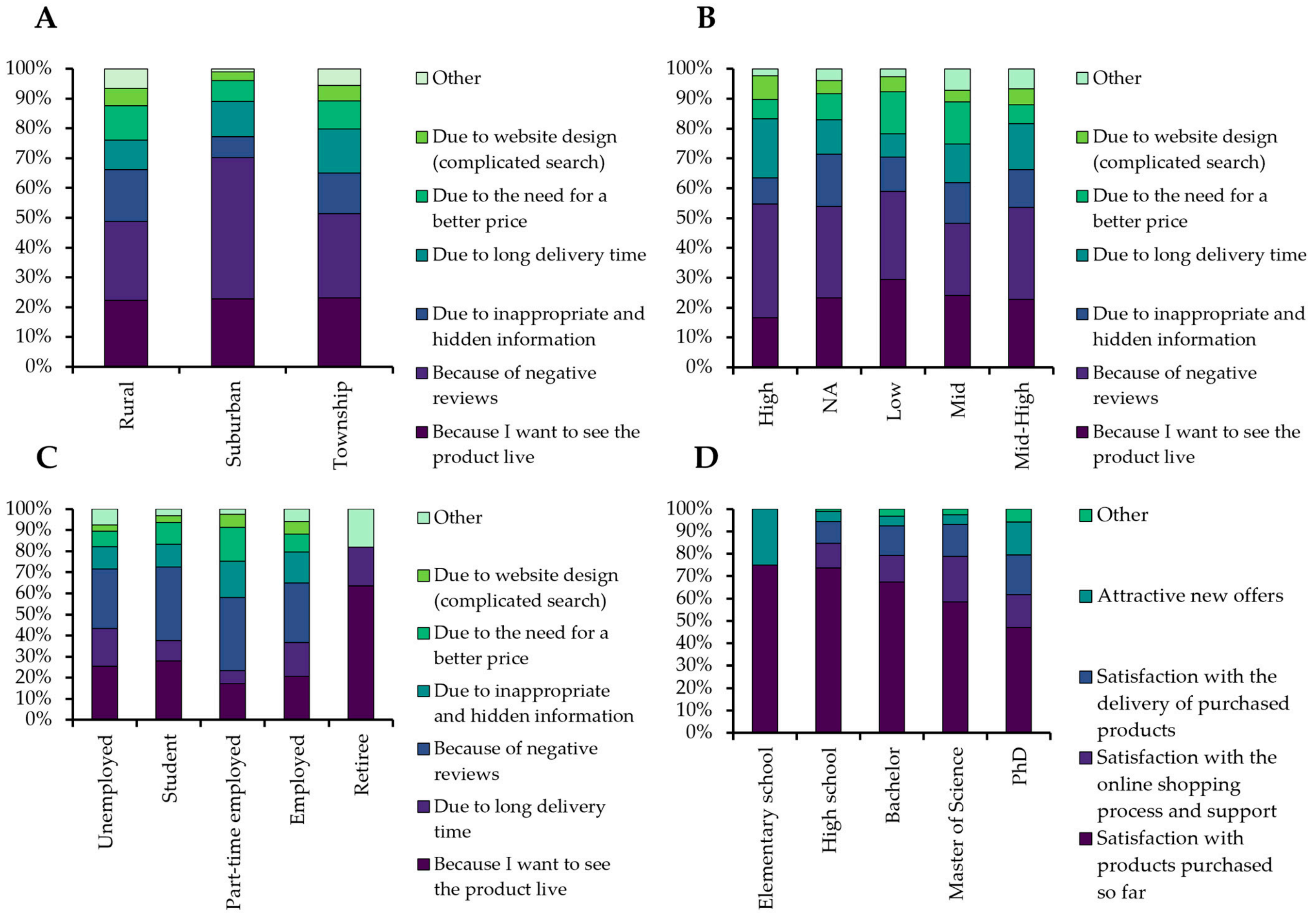



**Figure 7.** The frequencies of reasons for quitting (RFQ) variable and corresponding demographics (**A**) residence, (**B**) income, (**C**) work status. The frequencies of MIPBREP (most important property before repeating the purchase) and demographic (**D**) income.

Regarding the dependency between income and RFQ (Figure 7B), the evidence suggests that negative reviews (30%) are the most common reasons for quitting. However, higher-income respondents cite long delivery times (29.5%) and website design (29.9%) as the most common reasons for quitting, suggesting that convenience and user experience are more critical. In comparison, the need for a better price is a minor concern (12.8%) for these users. The low-income respondents cite the need for a better price (28.4%) and reflect a more price-sensitive group, while also the need to see the product live (25.3%). Paradoxically, the respondents who preferred not to disclose their income cite that the highest dissatisfaction rate is due to inappropriate or hidden information (27.4%) about the product as a potential concern over transparency and trustworthiness. Lower-income users prioritize affordability and assurance, while higher-income respondents prioritize efficiency and convenience.

The analysis of work status and RFQ (Figure 7C) shows that unemployed respondents and students demonstrate more risk-averse attitudes. In contrast, unemployed respondents cite negative reviews and long delivery times as primary concerns for quitting. At the same time, students also cite negative reviews in addition to a need for a better price, which is also the main reason (32.5%) emphasized in responses of part-time employees. The employed e-consumers cite hidden information and website design as the primary reasons, aligning somewhat with part-time workers. Still, although this suggests that individuals generally have more disposable income, they expect a high standard of service.

Lastly, the results regarding the association between education and MIPBREP (Figure 7D) show that as education levels increase, consumers tend to prioritize the broader shopping experience, such as satisfaction with delivery and customer services, rather than product—or price-related issues. This underlines the need for e-tailers to provide more comprehensive and personalized services to accommodate the expectations of more educated consumer groups.

### *5.2. Multiple Correspondence Analysis with Hierarchical Clustering on PCs*

For the MCA, mainly using significant variables, three potential clusters may be identified along the axes (Figure 4C). Hierarchical Clustering on Principal Components (HCPC) is performed to maintain objectivity in the identification of user profiles. The HCPC method uses Euclidian distance for clustering of points, while Ward's linkage is used for cluster selection. The complete analysis is performed in R studio *FactoMineR* (v2.11).

The dendrogram (Figure 8A) suggests that three clusters are selected, while the interpretation can also be suitable for selecting up to six clusters based on the inertia gain. However, for the simplicity of interpretation, three clusters are selected for the analysis (Figure 8B). The data behind clusters are provided in Appendix B. To understand the interpretation of tables (e.g., see Table A5), let us go through the features "Cla/Mod", "Mod/Cla", "Global", "*p*-value", and "v-test". For instance, Cluster 1 (black) shows that age = early adulthood ("Cla/Mod" = 92.913, v-test = 24.465) suggest that 92.9% belongs to Cluster 1, while 78.67% of age proportion ("Mod/Cla") in Cluster 1 is early adulthood. The "Global" shows the overall proportion of a particular class in a complete dataset. Note that only classes (with associated categories) that are statistically significant (v-test = ±1.96, *p*-value < 0.05) are represented.

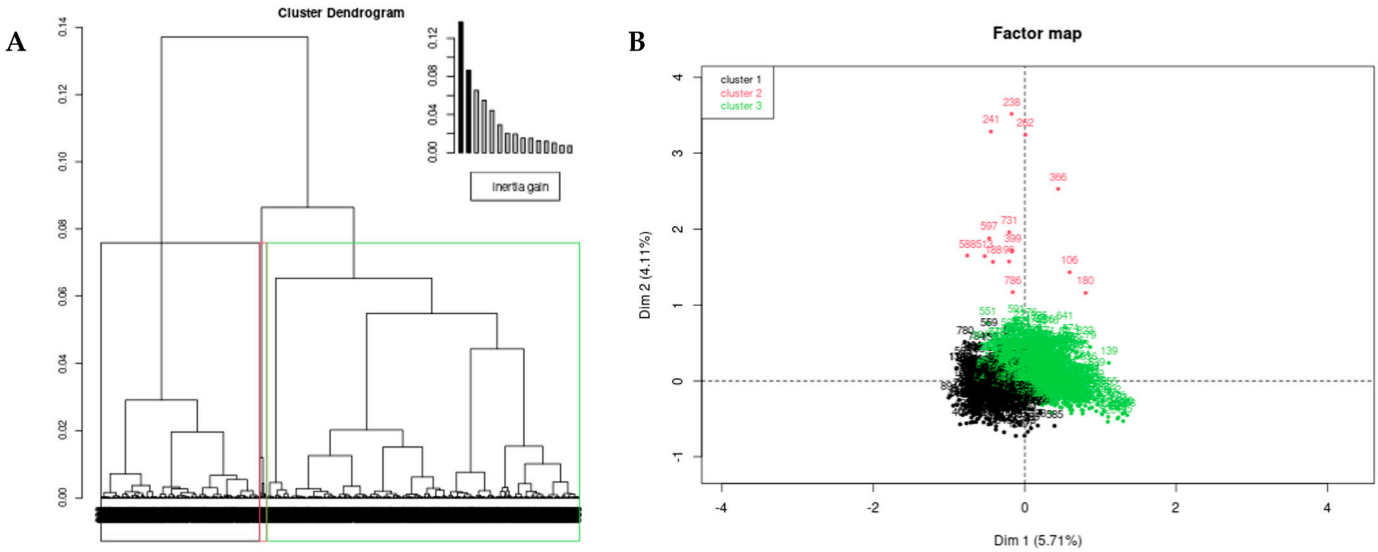



**Figure 8.** Agglomerative Hierarchical Clustering of observations represented via (**A**) a dendrogram with observations ($x$-axis) and distance measured ($y$-axis); and (**B**) identified clusters based on the first two Principal Components.

Cluster 1 (black), therefore, suggests that respondents are mainly in age = early adulthood, Students (with high school diploma), WORKSTAT = Unemployed, from RESI = Rural areas, and Gender = Females (and persons that do not want to disclose gender identity) being over-represented. A particular group of respondents show an association with PURFRE = Once a month, MIPB1T = To have positive customer reviews, RFQ = Because of negative reviews, and RFQ = I want to see the product live. This cluster suggests that women are dominant e-consumers, mainly driven by positive reviews for repeating the purchase (and vice versa, quitting the because of negative reviews), prefer seeing the product live with purchasing via mobile apps. The purchase frequency is characterized by moderate purchase frequency, suggesting at least once a month.

Cluster 2 (red) mainly describes WORKSTAT = Retiree (Cla/Mod = 100%) and age = older adulthood (Cla/Mod = 100%), with minimal purchase frequency of once every 12 months, that is mainly driven by attractive new offers when wanting to repeat the purchase but are also quitting if they cannot see the product in person. Ultimately, although with limited respondents (Global = 1.214%), this suggests that these e-consumers are retirees and infrequent cautious shoppers, and the key to repeating the purchase of these respondents is customer service and trust.

Cluster 3 (green) is comprised of diverse demographics. Namely, the cluster suggests that e-customers are mainly WORKSTAT = Employed (Cla/Mod = 95.82%), AGE = early midlife—midlife—late midlife (99.11%, 94.74%, and 95.16% respectively), EDUC = BSc—MSc—PhD (74.64%, 91.52%, and 97.1% respectively), and mostly dominated by males (73.1%) with township residence (67.54%). The user preferences of these e-consumers, frequent consumers (several times a week to several times a month), show that fast and accurate delivery is essential when buying the first time, while inappropriate and hidden information and long delivery are the main reasons for quitting the purchase. Overall, this cluster resembles early to late midlife, highly educated respondents, mostly male frequent shoppers who prioritize convenience and service efficiency.

### *5.3. Validity of Findings from Classifiers and Feature Importance*

The ML classifiers offer several critical insights into the effectiveness of distinguishing different e-customer profiles based on demographics and user preferences. Namely, the SVM, GBM and RF algorithms demonstrate the highest accuracy, achieving robust performance across proposed metrics. The SVM achieved the highest overall accuracy, suggesting high efficiency in handling multi-class problems with imbalanced datasets. This can be attributed to the SVM capacity to capture class separability by optimized hyperplanes. The Hinge loss optimization enables SVM to minimize misclassifications near decision boundaries, which is particularly useful in the case of MCA-HCPC user profile labelling. Similarly, ensemble methods, i.e., GBM and RF algorithms, performed consistently well, with GBM outperforming in the F1-score. This can be attributed to GBM's iterative boosting process prioritizing correcting previously falsely classified labels by fine-tuning. On the other hand, TF's robust performance is seen from the use of multiple Decision Trees, ultimately reducing the risk of overfitting and enhancing generalizability.

Although performing adequately, the GNB classifier offered less competitive outcomes in multi-class accuracy and AUROC than the SVM and ensemble methods. Such results may be due to GNB's feature independence assumption, which results in oversimplified data relations in complex user profiles. Lastly, the KNN classifier showed lower performance, presumably due to sensitivity to feature scaling and variations across classes, which were prevalent in this categorical dataset.

After obtaining important ML performance features, the interpretation of e-customer profiles suggests the following. Age is consistently ranked as the highest and most influential, so one can assume that younger adults are typically of the most interest to e-tailers. Hence, it suggests a correlation between age technology adoption and comfort with online shopping. Younger adults usually display more purchases and are drawn by user-friendly and interactive platforms. On the other hand, older adults often favor quality assurance and transparency. Ultimately, age-based preferences can reflect generational differences in digital engagement, where young people are more amenable to risk and immediacy, while the elderly prioritize security and familiarity.

Another critical factor is work status, which affects purchasing power and shopping references. Employed individuals tend to have more disposable income and thus prefer efficient and value-driven experiences. This suggests that fast and reliable delivery services are more important than price. Conversely, students and unemployed e-customers exhibit patterns in shopping behavior that reflect more price-sensitive decisions that place greater emphasis on reviews and promotional offers.

Unlike the work status, which can be considered an indirect indicator of budgetary flexibility and consumption priorities, the importance of the education feature may suggest two things: (a) technology savviness and (b) service quality expectations. What is meant by this? From the perspective of the sample demographics, higher education is associated with comfort in navigating through digital platforms, leading to higher purchases and trust in e-commerce. Similarly, educated individuals also exhibit higher standards for customer service, preferring e-tailers with strong reputations consistent with reliability, customer support, and transparency. The feature offers additional insights into the relationship between cognitive engagement and customer loyalty factors, as educated e-customers might scrutinize product information and reviews more closely.

Income also acts as one of the main determinants of purchasing power and often correlates with shopping frequency. The findings also suggest that higher-income respondents prioritize convenience, service quality and delivery over price, making them more inclined to purchase "premium" products or services. E-tailers may utilize such implications in providing platforms that offer enhanced customer experiences. On the other hand, low-income e-consumers typically emphasize affordability, promotional deals and discounts. Finally, while not the most dominant feature, the residence also differentiates user profiles, particularly for urban and rural e-customers. Specifically, rural users may face logistical constraints regarding reliability and transparency of delivery, which makes them more inclined to quiet if logistical support appears uncertain.

ML classifiers' feature importance findings show major implications for tailored e-commerce strategies. As the most influential feature, age stresses the generational divide in digital engagement, distinguishing younger and older adults as users seeking convenience and interaction and users prioritizing transparency and security, respectively. Simultaneously, work status and income further contrast purchasing power and preferences, where employed and high-income users prioritize efficient service over cost. At the same time, students and lower-income customers remain price-sensitive often guided by reviews and promotional offers. As a cognitive component, education plays a role in service standards and critical engagement regarding product details. Overall, the proposed ML models' findings offer meaningful insights and strategies for e-tailers in shaping user-profiles and developing marketing strategies.

## 6. Conclusions

### *6.1. Concluding Remarks*

This study performs a threefold analysis. Firstly, this study determines statistically significant variables by investigating the association between customer demographics and user behavioral preferences in the Republic of Serbia on a survey of $n = 906$ respondents. The findings show an 8/24 significant association, with extreme evidence considering the association between variables: age and purchase frequency, income status and purchase frequency, and work status and purchase frequency. Interestingly, there is no significant association reported between demographics and items of "Most important property when purchasing for the first time" ($p < 0.05$). Still, all reported tests suggest a small effect size, suggesting that more variables explain user preferences outside of used demographics.

Secondly, given that tests do not expose particular user profiles, an MCA-HCPC (Multiple Correspondence Analysis Hierarchical Clustering on Principal Components) identifies three main clusters. The first cluster mainly comprises early adult students (unemployed), primarily females from rural areas. These profiles are characterized by moderate purchase frequency (at least once a month) driven by positive customer reviews and the desire to see the product in person. The second cluster describes retirees who exhibit infrequent (once in 12 months) purchases driven by attractive offers but are hesitant if they cannot inspect the product. The third cluster includes a more diverse group but primarily describes males in early to late midlife stages who are employed and prioritize fast and accurate delivery.

Finally, merged datasets with $C$ cluster labels are subjected to the following ML algorithms: Support Vector Machine, Gradient Boosting Machine, Decision Trees, Random Forest, k-Nearest Neighbors, and Gaussian Naïve Bayes. The results suggest >90% classification accuracy, where GNB, RF, and SVM offer the highest classification metrics considering accuracy, precision, recall, F1-score, and AUROC. Additionally, from the obtained classifiers, important features were extracted by feature importance relying on Mean Dropout Loss. The evidence suggests that age, work status, education, income, and residence are the main determinants for shaping user profiles and developing e-tailers' marketing strategies.

### *6.2. Limitations of This Study*

Regarding the demographics, there is a low representation of the elderly and residents from rural areas, such as a high percentage of e-customers who did not want to disclose their income. This introduced a slight increase in heterogeneity in the sample, i.e., reduced confidence in the relationship between income status and user preferences, which may impact the application of the research results. For instance, the heterogeneity impacted the Machine Learning classification with uneven representations of demographics (e.g., low elderly population and missing income disclosures), introducing potential bias in training. To address this, samples are stratified during splitting to ensure a proportional representation of demographic variables. Also, to avoid possible bias in reporting the results, performance metrics sensitive to imbalanced datasets, such as F1-score and AUROC, are used to ensure robustness in reporting results despite sample heterogeneity.

The Multiple Correspondence Analysis suggests low inertia in the first three components. This may affect the interpretation of association among class categories. This certainly does not downplay findings, given that most of the captured (inertia) variance corresponds well to investigated test statistics. However, a higher percentage of captured inertia would undoubtedly increase the confidence in understanding user profiles.

### *6.3. Implications*

Regarding e-customer preferences, website design plays a minimal role for most respondents, showing that the user interface of online stores plays less of a barrier than logistical and trust-related issues. Thus, improved logistics (e.g., fast and delivery), mainly for township areas, are a significant concern, likely due to infrastructure issues. Also, product transparency (e.g., hidden or inappropriate information) plays a vital role in quitting the purchase, in addition to negative reviews. Thus, platforms could benefit from more proactive review management, encouraging satisfied customers to leave positive reviews and addressing negative feedback promptly. Future research will focus more on the behavior of specific subgroups, which could provide a better understanding of e-customer preferences and behavioral patterns.

**Supplementary Materials:** The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/math12233794/s1.

**Author Contributions:** Conceptualization, V.V.; methodology, M.O. and V.V.; validation, V.V.; formal analysis, M.O.; investigation, V.V.; resources, V.V. and K.R.; data curation, V.V.; writing—original draft preparation, V.V., M.J., J.S., N.S., K.R. and M.O.; writing—review and editing, V.V., M.J., J.S., N.S. and M.O.; visualization, M.O.; supervision, V.V., M.J., J.S., N.B. and N.S.; project administration, V.V., N.S., M.J., N.B. and J.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-65/2024-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project "Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad" (No. 01-3394/1).

**Data Availability Statement:** Data are available in Supplementary Materials.

**Conflicts of Interest:** The authors declare no conflicts of interest.

**Author Statement:** The large language models are used here to ensure the manuscript's grammatical and scientific writing quality. Specifically, Grammarly removes grammar and spelling errors and language corrections, while OpenAI's GPT-4 tool is used in refining sentences. The LLM and AI tools used here are solely used as accelerators for enhancing the writing process, assisting in language and spelling checks and improving writing accuracy. These tools are not used to generate new ideas, insights or sources of intellectual content within the manuscript. All of the ideas, comments, discussions, drawings, illustrations and processing of images (and graphs) originated and are solely the work of the authors of this manuscript.

## Appendix A. Non-Statistically Significant Findings

**Table A1.** Chi-square test statistics of (un)conditionally dependent relationships (non-significant results).

| Variables | Chi-Squared Tests | Value | *df* | *p* | Cramer's *V* | VS–MPR |
|---|---|---|---|---|---|---|
| RESI—PURFRE | $\chi^2$ test statistic | 13.502 | 8 | 0.096 | 0.086 | 1.638 |
| | $G^2$ Likelihood ratio | 14.657 | 8 | 0.066 | | 2.047 |
| AGE—RFQ | $\chi^2$ test statistic | 43.165 | 30 | 0.057 | 0.098 | 2.262 |
| | $G^2$ Likelihood ratio | 42.747 | 30 | 0.062 | | 2.141 |
| EDU—RFQ | $\chi^2$ test statistic | 21.600 | 24 | 0.603 | 0.077 | 1.000 |
| | $G^2$ Likelihood ratio | 21.506 | 24 | 0.609 | | 1.000 |
| AGE—MIPB1T | $\chi^2$ test statistic | 37.455 | 25 | 0.052 | 0.091 | 2.386 |
| | $G^2$ Likelihood ratio | 36.759 | 25 | 0.061 | | 2.160 |
| EDU—MIPB1T | $\chi^2$ test statistic | 24.007 | 20 | 0.242 | 0.081 | 1.071 |
| | $G^2$ Likelihood ratio | 19.474 | 20 | 0.491 | | 1.000 |
| RESI—MIPB1T | $\chi^2$ test statistic | 10.860 | 10 | 0.369 | 0.077 | 1.000 |
| | $G^2$ Likelihood ratio | 13.190 | 10 | 0.213 | | 1.116 |
| AGE—MIPBREP | $\chi^2$ test statistic | 20.786 | 20 | 0.410 | 0.076 | 1.000 |
| | $G^2$ Likelihood ratio | 19.676 | 20 | 0.418 | | 1.000 |
| RESI—MIPBREP | $\chi^2$ test statistic | 12.540 | 8 | 0.129 | 0.083 | 1.394 |
| | $G^2$ Likelihood ratio | 14.275 | 8 | 0.117 | | 1.896 |
| INCSTAT—MIPBREP | $\chi^2$ test statistic | 6.601 | 16 | 0.980 | 0.043 | 1.000 |
| | $G^2$ Likelihood ratio | 6.959 | 16 | 0.974 | | 1.000 |
| INCSTAT—MIPB1T | $\chi^2$ test statistic | 22.425 | 20 | 0.318 | 0.079 | 1.010 |
| | $G^2$ Likelihood ratio | 23.378 | 20 | 0.271 | | 1.040 |
| WORKSTAT—MIPBREP | $\chi^2$ test statistic | 23.385 | 16 | 0.104 | 0.080 | 1.564 |
| | $G^2$ Likelihood ratio | 18.538 | 16 | 0.293 | | 1.023 |
| WORKSTAT—MIPB1T | $\chi^2$ test statistic | 22.403 | 20 | 0.319 | 0.319 | 1.009 |
| | $G^2$ Likelihood ratio | 23.386 | 20 | 0.270 | | 1.040 |
| GENDER—PURFRE | $\chi^2$ test statistic | 6.853 | 8 | 0.553 | 0.061 | 1.000 |
| | $G^2$ Likelihood ratio | 7.311 | 8 | 0.503 | | 1.000 |
| GENDER—MIPBREP | $\chi^2$ test statistic | 14.223 | 8 | 0.074 | 0.089 | 1.914 |
| | $G^2$ Likelihood ratio | 13.556 | 8 | 0.094 | | 1.654 |
| GENDER—MIPB1T | $\chi^2$ test statistic | 6.432 | 10 | 0.778 | 0.060 | 1.000 |
| | $G^2$ Likelihood ratio | 7.038 | 10 | 0.722 | | 1.000 |
| GENDER—RFQ | $\chi^2$ test statistic | 20.983 | 12 | 0.051 | 0.108 | 2.436 |
| | $G^2$ Likelihood ratio | 22.061 | 12 | 0.037 | | 3.025 |

**Table A2.** Categories (PC1, PC2, PC3), including contributions, $\cos^2$ similarity and v-test—significant categories.

| Categories | PC1 | CTR | $\cos^2$ | v-Test | PC2 | CTR | $\cos^2$ | v-Test | PC3 | CTR | $\cos^2$ | v-Test |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Early Adulthood | −1.22 | 18.57 | 0.58 | −22.93 | −0.26 | 1.15 | 0.03 | −4.88 | 0.68 | 9.01 | 0.18 | 12.67 |
| Student | −1.20 | 15.70 | 0.47 | −20.57 | −0.33 | 1.58 | 0.03 | −5.58 | 0.65 | 7.20 | 0.14 | 11.05 |
| High School | −0.67 | 8.83 | 0.36 | −18.07 | −0.06 | 0.11 | 0.00 | −1.73 | 0.07 | 0.17 | 0.00 | 1.99 |
| Rural | −0.74 | 3.24 | 0.08 | −8.73 | 0.05 | 0.02 | 0.00 | 0.64 | −0.44 | 1.84 | 0.03 | −5.21 |
| Income_NA | −0.45 | 2.21 | 0.07 | −7.76 | −0.14 | 0.31 | 0.01 | −2.49 | 0.02 | 0.01 | 0.00 | 0.40 |
| Once 6 months | −0.34 | 1.55 | 0.05 | −6.76 | 0.27 | 1.39 | 0.03 | 5.47 | −0.41 | 3.63 | 0.08 | −8.22 |
| MIPBREP_Satisf_Product | −0.14 | 0.62 | 0.04 | −6.30 | −0.08 | 0.24 | 0.01 | −3.33 | −0.27 | 3.45 | 0.15 | −11.82 |
| RFQ_Product_Live | −0.35 | 1.24 | 0.04 | −5.72 | 0.70 | 6.90 | 0.15 | 11.54 | −0.14 | 0.30 | 0.01 | −2.23 |
| Low Income | −0.58 | 1.27 | 0.03 | −5.32 | 0.50 | 1.30 | 0.02 | 4.59 | −0.41 | 1.03 | 0.02 | −3.80 |
| Once 12 months | −0.50 | 1.05 | 0.03 | −4.85 | 0.82 | 3.87 | 0.07 | 7.97 | −0.41 | 1.14 | 0.02 | −4.01 |
| Unemployed | −0.55 | 0.98 | 0.02 | −4.64 | 0.09 | 0.03 | 0.00 | 0.73 | 0.08 | 0.04 | 0.00 | 0.72 |
| Once a month | −0.13 | 0.22 | 0.01 | −2.55 | −0.24 | 1.02 | 0.02 | −4.66 | 0.20 | 0.81 | 0.02 | 3.86 |
| RFQ_Price | −0.25 | 0.27 | 0.01 | −2.45 | −0.05 | 0.01 | 0.00 | −0.44 | −0.25 | 0.41 | 0.01 | −2.41 |
| RFQ_Negative_Reviews | −0.08 | 0.08 | 0.00 | −1.51 | −0.41 | 3.10 | 0.07 | −8.11 | −0.15 | 0.47 | 0.01 | −2.93 |
| Part-time | −0.14 | 0.08 | 0.00 | −1.36 | −0.12 | 0.08 | 0.00 | −1.14 | −0.35 | 0.79 | 0.01 | −3.33 |
| Retiree | −0.35 | 0.07 | 0.00 | −1.16 | 6.91 | 35.30 | 0.59 | 23.04 | 1.38 | 1.62 | 0.02 | 4.59 |
| MIPBREP_Attractive _offers | −0.13 | 0.04 | 0.00 | −0.91 | 0.96 | 2.76 | 0.05 | 6.57 | 0.47 | 0.78 | 0.01 | 3.25 |
| Elementary | −0.43 | 0.04 | 0.00 | −0.86 | 0.43 | 0.05 | 0.00 | 0.85 | −0.10 | 0.00 | 0.00 | −0.20 |
| Suburban | −0.06 | 0.02 | 0.00 | −0.65 | −0.27 | 0.49 | 0.01 | −2.86 | −0.66 | 3.45 | 0.06 | −7.05 |
| RFQ_Missinformation | 0.05 | 0.01 | 0.00 | 0.58 | −0.15 | 0.19 | 0.00 | −1.81 | −0.36 | 1.21 | 0.02 | −4.24 |
| Older Adulthood | 0.53 | 0.10 | 0.00 | 1.40 | 7.60 | 27.15 | 0.45 | 20.16 | 2.62 | 3.76 | 0.05 | 6.96 |
| Late midlife | 0.20 | 0.12 | 0.00 | 1.59 | 0.79 | 2.59 | 0.05 | 6.43 | −0.98 | 4.64 | 0.07 | −7.99 |
| Mid-High Income | 0.09 | 0.10 | 0.00 | 1.67 | −0.19 | 0.62 | 0.01 | −3.60 | −0.08 | 0.12 | 0.00 | −1.50 |
| RFQ_Website | 0.30 | 0.20 | 0.01 | 2.09 | −0.39 | 0.47 | 0.01 | −2.71 | 0.25 | 0.22 | 0.00 | 1.71 |
| MIPBREP_Other | 0.56 | 0.30 | 0.01 | 2.51 | −0.14 | 0.03 | 0.00 | −0.65 | −0.02 | 0.00 | 0.00 | −0.08 |
| Mid Income | 0.16 | 0.27 | 0.01 | 2.67 | 0.47 | 3.01 | 0.06 | 7.62 | −0.15 | 0.37 | 0.01 | −2.49 |
| MIPBREP_Shopping_Process | 0.29 | 0.47 | 0.01 | 3.33 | 0.02 | 0.00 | 0.00 | 0.23 | 0.40 | 1.42 | 0.02 | 4.56 |
| RFQ_Other | 0.58 | 0.77 | 0.02 | 4.07 | 0.63 | 1.26 | 0.02 | 4.43 | 0.58 | 1.24 | 0.02 | 4.10 |
| MIPBREP_Delivery | 0.47 | 1.13 | 0.03 | 5.12 | 0.04 | 0.01 | 0.00 | 0.43 | 0.93 | 7.14 | 0.12 | 10.19 |
| Young Adulthood | 0.34 | 1.24 | 0.04 | 5.79 | −0.17 | 0.44 | 0.01 | −2.96 | −0.40 | 2.78 | 0.05 | −6.88 |
| Several times a week | 0.85 | 1.78 | 0.04 | 6.20 | −0.68 | 1.57 | 0.03 | −4.96 | 0.87 | 2.93 | 0.04 | 6.31 |
| RFQ_Long_Delivery | 0.54 | 1.80 | 0.05 | 6.52 | −0.19 | 0.29 | 0.01 | −2.24 | 0.76 | 5.59 | 0.09 | 9.11 |
| midlife | 0.40 | 1.76 | 0.05 | 6.92 | −0.04 | 0.02 | 0.00 | −0.66 | −0.67 | 7.81 | 0.15 | −11.55 |
| BSc | 0.30 | 1.53 | 0.06 | 7.09 | 0.10 | 0.25 | 0.01 | 2.44 | −0.32 | 2.73 | 0.06 | −7.50 |
| Town | 0.14 | 0.65 | 0.06 | 7.37 | 0.03 | 0.04 | 0.00 | 1.58 | 0.18 | 1.65 | 0.10 | 9.28 |
| PhD | 1.27 | 2.69 | 0.06 | 7.55 | −0.10 | 0.02 | 0.00 | −0.58 | 1.33 | 4.66 | 0.07 | 7.88 |
| High Income | 0.71 | 3.13 | 0.08 | 8.60 | −0.42 | 1.50 | 0.03 | −5.09 | 0.63 | 3.85 | 0.06 | 7.57 |
| Several times a month | 0.59 | 3.77 | 0.11 | 10.06 | −0.22 | 0.72 | 0.02 | −3.74 | 0.24 | 1.01 | 0.02 | 4.14 |
| Early midlife | 0.97 | 6.14 | 0.16 | 12.11 | 0.08 | 0.06 | 0.00 | 1.03 | 0.83 | 7.21 | 0.12 | 10.41 |
| MSc | 1.07 | 6.56 | 0.17 | 12.40 | −0.07 | 0.04 | 0.00 | −0.80 | 0.30 | 0.81 | 0.01 | 3.45 |
| Employed | 0.61 | 9.43 | 0.51 | 21.40 | 0.00 | 0.00 | 0.00 | 0.02 | −0.26 | 2.70 | 0.09 | −9.09 |

## Appendix B. Agglomerative Hierarchical Clustering

**Table A3.** Significant variables identified in the first cluster.

| Cluster 1 Variables | Cla/Mod | Mod/Cla | Global | *p* Value | v-Test |
|---|---|---|---|---|---|
| AGE = Early Adulthood | 92.913 | 78.667 | 28.035 | 0.000 | 24.465 |
| WORKSTAT = Student | 97.738 | 72.000 | 24.393 | 0.000 | 24.261 |
| EDUC = High school | 50.860 | 69.000 | 44.923 | 0.000 | 10.295 |
| WORKSTAT = Unemployed | 55.224 | 12.333 | 7.395 | 0.000 | 3.846 |
| RESI = Rural | 48.760 | 19.667 | 13.355 | 0.000 | 3.823 |
| MIPB1T = To have positive customer reviews | 41.554 | 41.000 | 32.671 | 0.000 | 3.722 |
| PURFRE = Once a month | 41.697 | 37.667 | 29.912 | 0.000 | 3.545 |
| GEND = Female | 37.043 | 71.000 | 63.466 | 0.001 | 3.333 |
| INCSTAT = I don’t want to say | 42.105 | 32.000 | 25.166 | 0.001 | 3.288 |
| PPL = No | 35.756 | 82.000 | 75.938 | 0.002 | 3.040 |
| RFQ = RFQ_Because of negative reviews | 39.194 | 35.667 | 30.132 | 0.011 | 2.531 |

**Table A3.** *Cont.*

| Cluster 1 Variables | Cla/Mod | Mod/Cla | Global | p Value | v-Test |
|---|---|---|---|---|---|
| CCP = Mobile app | 37.736 | 46.667 | 40.949 | 0.014 | 2.451 |
| GEND = Prefer not to disclose | 75.000 | 2.000 | 0.883 | 0.021 | 2.309 |
| RFQ = RFQ_Because I want to see the product live | 39.423 | 27.333 | 22.958 | 0.029 | 2.179 |
| RFQ = Due to long delivery time | 24.800 | 10.333 | 13.797 | 0.032 | −2.151 |
| MIPB1T = Fast and accurate delivery | 26.733 | 18.000 | 22.296 | 0.028 | −2.201 |
| CCP = SMS | 29.621 | 44.333 | 49.558 | 0.027 | −2.209 |
| PURFRE = Several times a week | 18.000 | 3.000 | 5.519 | 0.016 | −2.401 |
| RFQ = RFQ_Other | 17.021 | 2.667 | 5.188 | 0.013 | −2.483 |
| RESI = Town/township | 30.848 | 70.333 | 75.497 | 0.012 | −2.513 |
| WORKSTAT = Retiree | 0.000 | 0.000 | 1.214 | 0.012 | −2.523 |
| RFQ = Due to inappropriate and hidden information | 22.951 | 9.333 | 13.466 | 0.009 | −2.609 |
| INCSTAT = Mid Income | 25.604 | 17.667 | 22.848 | 0.008 | −2.640 |
| AGE = Young Adulthood | 25.893 | 19.333 | 24.724 | 0.008 | −2.671 |
| PPL = Yes | 24.771 | 18.000 | 24.062 | 0.002 | −3.040 |
| GEND = Male | 25.077 | 27.000 | 35.651 | 0.000 | −3.858 |
| PURFRE = Several times a month | 22.018 | 16.000 | 24.062 | 0.000 | −4.075 |
| EDUC = Bachelor | 23.615 | 27.000 | 37.859 | 0.000 | −4.791 |
| EDUC = PhD | 0.000 | 0.000 | 3.753 | 0.000 | −4.926 |
| AGE = Late midlife | 0.000 | 0.000 | 6.843 | 0.000 | −6.907 |
| EDUC = Master of Science | 7.627 | 3.000 | 13.024 | 0.000 | −6.929 |
| AGE = Early midlife | 3.759 | 1.667 | 14.680 | 0.000 | −8.851 |
| AGE = midlife | 0.442 | 0.333 | 24.945 | 0.000 | −14.177 |
| WORKSTAT = Employed | 3.802 | 6.667 | 58.057 | 0.000 | −23.228 |

**Table A4.** Significant variables identified in the second cluster.

| Cluster 2 | Cla/Mod | Mod/Cla | Global | p Value | v-Test |
|---|---|---|---|---|---|
| WORKSTAT = Retiree | 100.000 | 78.571 | 1.214 | 0.000 | 9.891 |
| AGE = Older Adulthood | 100.000 | 50.000 | 0.773 | 0.000 | 7.577 |
| RFQ = I want to see the product live | 4.327 | 64.286 | 22.958 | 0.001 | 3.248 |
| CCP = SMS | 2.450 | 78.571 | 49.558 | 0.031 | 2.153 |
| MIPBREP = Attractive new offers | 6.667 | 21.429 | 4.967 | 0.032 | 2.143 |
| PURFRE = Once every 12 months | 4.651 | 28.571 | 9.492 | 0.043 | 2.023 |
| WORKSTAT = Student | 0.000 | 0.000 | 24.393 | 0.019 | −2.340 |
| AGE = Young Adulthood | 0.000 | 0.000 | 24.724 | 0.018 | −2.363 |
| INCSTAT = Mid-High Income | 0.000 | 0.000 | 29.470 | 0.007 | −2.686 |
| WORKSTAT = Employed | 0.380 | 14.286 | 58.057 | 0.001 | −3.281 |

**Table A5.** Significant variables identified in the third cluster.

| Cluster 3 | Cla/Mod | Mod/Cla | Global | p Value | v-Test |
|---|---|---|---|---|---|
| WORKSTAT = Employed | 95.817 | 85.135 | 58.057 | 0.000 | 23.920 |
| AGE = midlife | 99.115 | 37.838 | 24.945 | 0.000 | 14.315 |
| AGE = Early midlife | 94.737 | 21.284 | 14.680 | 0.000 | 8.620 |
| EDUC = Master of Science | 91.525 | 18.243 | 13.024 | 0.000 | 6.993 |
| AGE = Late midlife | 95.161 | 9.966 | 6.843 | 0.000 | 5.734 |
| EDUC = Bachelor | 74.636 | 43.243 | 37.859 | 0.000 | 4.628 |
| EDUC = PhD | 97.059 | 5.574 | 3.753 | 0.000 | 4.470 |
| PURFRE = Several times a month | 77.064 | 28.378 | 24.062 | 0.000 | 4.255 |
| GEND = Male | 73.065 | 39.865 | 35.651 | 0.000 | 3.661 |
| PPL = Yes | 74.312 | 27.365 | 24.062 | 0.001 | 3.233 |
| AGE = Young Adulthood | 74.107 | 28.041 | 24.724 | 0.001 | 3.214 |
| RFQ = Due to inappropriate | 77.049 | 15.878 | 13.466 | 0.003 | 2.980 |

**Table A5.** *Cont.*

| Cluster 3 | Cla/Mod | Mod/Cla | Global | *p* Value | v-Test |
|---|---|---|---|---|---|
| PURFRE = Several times a week | 82.000 | 6.926 | 5.519 | 0.009 | 2.624 |
| RESI = Town/township | 67.544 | 78.041 | 75.497 | 0.016 | 2.419 |
| MIPB1T = Fast and accurate | 71.782 | 24.493 | 22.296 | 0.028 | 2.195 |
| INCSTAT = Mid Income | 71.498 | 25.000 | 22.848 | 0.033 | 2.130 |
| RFQ = Due to long delivery | 73.600 | 15.541 | 13.797 | 0.035 | 2.109 |
| RFQ = Other | 78.723 | 6.250 | 5.188 | 0.044 | 2.010 |
| RFQ = Because of negative | 60.440 | 27.872 | 30.132 | 0.043 | −2.022 |
| CCP = Mobile app | 61.456 | 38.514 | 40.949 | 0.041 | −2.039 |
| PURFRE = Once every 6 months | 60.498 | 28.716 | 31.015 | 0.041 | −2.041 |
| INCSTAT = Low Income | 53.846 | 7.095 | 8.609 | 0.029 | −2.184 |
| GEND = Prefer not to disclose | 25.000 | 0.338 | 0.883 | 0.027 | −2.215 |
| RFQ = RFQ_Because I want to see | 56.250 | 19.764 | 22.958 | 0.002 | −3.097 |
| GEND = Female | 61.565 | 59.797 | 63.466 | 0.002 | −3.163 |
| INCSTAT = I don't want to say | 56.579 | 21.791 | 25.166 | 0.002 | −3.174 |
| PURFRE = Once a month | 57.565 | 26.351 | 29.912 | 0.001 | −3.182 |
| PPL = No | 62.500 | 72.635 | 75.938 | 0.001 | −3.233 |
| AGE = Older Adulthood | 0.000 | 0.000 | 0.773 | 0.001 | −3.443 |
| MIPB1T = MIPB1T_To have positive customer reviews | 57.095 | 28.547 | 32.671 | 0.000 | −3.600 |
| WORKSTAT = Unemployed | 43.284 | 4.899 | 7.395 | 0.000 | −3.812 |
| RESI = Rural | 49.587 | 10.135 | 13.355 | 0.000 | −3.818 |
| WORKSTAT = Retiree | 0.000 | 0.000 | 1.214 | 0.000 | −4.473 |
| EDUC = High school | 47.666 | 32.770 | 44.923 | 0.000 | −10.136 |
| WORKSTAT = Student | 2.262 | 0.845 | 24.393 | 0.000 | −23.557 |

## References


1. Svobodová, Z.; Rajchlová, J. Strategic Behavior of E-Commerce Businesses in Online Industry of Electronics from a Customer Perspective. *Adm. Sci.* **2020**, *10*, 78. [CrossRef]
2. Kim, J.; Yum, K. Enhancing Continuous Usage Intention in E-Commerce Marketplace Platforms: The Effects of Service Quality, Customer Satisfaction, and Trust. *Appl. Sci.* **2024**, *14*, 7617. [CrossRef]
3. Vakulenko, Y.; Shams, P.; Hellström, D.; Hjort, K. Online Retail Experience and Customer Satisfaction: The Mediating Role of Last Mile Delivery. *Int. Rev. Retail. Distrib. Consum. Res.* **2019**, *29*, 306–320. [CrossRef]
4. Buldeo Rai, H.; Mommens, K.; Verlinde, S.; Macharis, C. How Does Consumers' Omnichannel Shopping Behaviour Translate into Travel and Transport Impacts? Case-Study of a Footwear Retailer in Belgium. *Sustainability* **2019**, *11*, 2534. [CrossRef]
5. Zennaro, I.; Finco, S.; Calzavara, M.; Persona, A. Implementing E-Commerce from Logistic Perspective: Literature Review and Methodological Framework. *Sustainability* **2022**, *14*, 911. [CrossRef]
6. Guo, J.; Liu, X.; Jo, J. Dynamic Joint Construction and Optimal Operation Strategy of Multi-Period Reverse Logistics Network: A Case Study of Shanghai Apparel E-Commerce Enterprises. *J. Intell. Manuf.* **2017**, *28*, 819–831. [CrossRef]
7. Biclesanu, I.; Anagnoste, S.; Branga, O.; Savastano, M. Digital Entrepreneurship: Public Perception of Barriers, Drivers, and Future. *Adm. Sci.* **2021**, *11*, 125. [CrossRef]
8. Voicu, M.-C.; Sîrghi, N.; Toth, D.M.-M. Consumers' Experience and Satisfaction Using Augmented Reality Apps in E-Shopping: New Empirical Evidence. *Appl. Sci.* **2023**, *13*, 9596. [CrossRef]
9. Olsson, J.; Hellström, D.; Vakulenko, Y. Customer Experience Dimensions in Last-Mile Delivery: An Empirical Study on Unattended Home Delivery. *Int. J. Phys. Distrib. Logist. Manag.* **2023**, *53*, 184–205. [CrossRef]
10. Vrhovac, V.; Vasić, S.; Milisavljević, S.; Dudić, B.; Štarchoň, P.; Žižakov, M. Measuring E-Commerce User Experience in the Last-Mile Delivery. *Mathematics* **2023**, *11*, 1482. [CrossRef]
11. Ranđelović, D. Internet Prodaja u Republici Srbiji. *Pravo-Teor. i Praksa* **2017**, *34*, 13–24.
12. Vasic, N.; Kilibarda, M.; Kaurin, T. The Influence of Online Shopping Determinants on Customer Satisfaction in the Serbian Market. *J. Theor. Appl. Electron. Commer. Res.* **2019**, *14*, 70–89. [CrossRef]
13. TRADE Serbia—Country Commercial Guide. Available online: https://www.trade.gov/country-commercial-guides/serbia-ecommerce (accessed on 10 September 2024).
14. Kaurin, T.; Kilibarda, M.; Fakultet Beograd, S. Analiza Determinanti Elektronske Trgovine Na Tržištu Srbije. In Proceedings of the 17th International Symposium INFOTEH-JAHORINA, Jahorina, East Sarajevo, Republic of Srpska, 21–23 March 2018.
15. Martinović, M.; Barać, R.; Maljak, H. Exploring Croatian Consumer Adoption of Subscription-Based E-Commerce for Business Innovation. *Adm. Sci.* **2024**, *14*, 149. [CrossRef]

16. Borsiak-Dańska, B.; Grzelak, M.; Roszko-Wójtowicz, E. The Development of Innovation and Infrastructure in the European Countries Fostering the Growth of the E-Commerce Sector. In *Financial Stability, Economic Growth and Sustainable Development*; Routledge: London, UK, 2024; pp. 251–278.
17. Agarwal, V.; Govindan, K.; Darbari, J.D.; Jha, P.C. An Optimization Model for Sustainable Solutions towards Implementation of Reverse Logistics under Collaborative Framework. *Int. J. Syst. Assur. Eng. Manag.* **2016**, *7*, 480–487. [CrossRef]
18. Kotler, P.; Keller, K. *Marketing Management*, 15th ed.; Pearson: London, UK, 2015; ISBN 0133856461.
19. Morris, M.G.; Venkatesh, V. Age Differences in Technology Adoption Decisions: Implications for a Changing Work Force. *Pers. Psychol.* **2000**, *53*, 375–403. [CrossRef]
20. Zavali, M.; Lacka, E.; de Smedt, J. Shopping Hard or Hardly Shopping: Revealing Consumer Segments Using Clickstream Data. *IEEE Trans. Eng. Manag.* **2023**, *70*, 1353–1364. [CrossRef]
21. Vieira, J.; Frade, R.; Ascenso, R.; Prates, I.; Martinho, F. Generation Z and Key-Factors on E-Commerce: A Study on the Portuguese Tourism Sector. *Adm. Sci.* **2020**, *10*, 103. [CrossRef]
22. Barutcu, S. E-Customer Satisfaction in the E-Tailing Industry: An Empirical Survey for Turkish E-Customers. *Ege Acad. Rev.* **2010**, *10*, 15–35. [CrossRef]
23. Ghosal, I. A Demographic Study of Buying Spontaneity on E-Shoppers: Preference Kolkata (West Bengal). *J. Technol. Manag. Grow. Econ.* **2015**, *6*, 65–75. [CrossRef]
24. Ansari, Z.A.; Qadri, F.A. Role of E-Retailer's Image in Online Consumer Behaviour—Empirical Findings from E-Customers' Perspective in Saudi Arabia. *Int. Bus. Res.* **2018**, *11*, 57. [CrossRef]
25. Tkalčič, M.; Chen, L. Personality and Recommender Systems. In *Recommender Systems Handbook*; Springer US: New York, NY, USA, 2022; pp. 757–787.
26. Hristoski, I.; Kostoska, O. On the Taxonomies and Typologies of E-Customers in B2C e-Commerce. *Balk. Near East. J. Social Sci.* **2018**, *4*, 130–148.
27. Swarnakar, P.; Kumar, A.; Kumar, S. Why Generation Y Prefers Online Shopping: A Study of Young Customers of India. *Int. J. Bus. Forecast. Mark. Intell.* **2016**, *2*, 215. [CrossRef]
28. Chen, X.; Guo, Y.; Xu, H.; Yan, H.; Lin, L. User Demographic Prediction Based on the Fusion of Mobile and Survey Data. *IEEE Access* **2022**, *10*, 111507–111527. [CrossRef]
29. Bellini, P.; Palesi, L.A.I.; Nesi, P.; Pantaleo, G. Multi Clustering Recommendation System for Fashion Retail. *Multimed. Tools Appl.* **2023**, *82*, 9989–10016. [CrossRef] [PubMed]
30. Kuruba Manjunath, Y.S.; Kashef, R.F. Distributed Clustering Using Multi-Tier Hierarchical Overlay Super-Peer Peer-to-Peer Network Architecture for Efficient Customer Segmentation. *Electron. Commer. Res. Appl.* **2021**, *47*, 101040. [CrossRef]
31. Hørlück, J.; Christiansen, T.; Hansen, L.K.; Larsen, J. Are All E-Customers Alike? In Proceedings of the 1st Nordic Workshop on Electronic Commerce, Halmstad, Sweden, 28–29 May 2001.
32. Ansari, S.; Farooqi, R. Moderating Effect Of Demographic Variables on Attitude towards Online Shopping: An Empirical Study Using PROCESS. *IOSR J. Bus. Manag.* **2017**, *19*, 47–54.
33. Noorshella, C.N.; Abdullah, A.M.; Nursalihah, A.R. Examining the Key Factors Affecting E-Service Quality of Small Online Apparel Businesses in Malaysia. *Sage Open* **2015**, *5*, 215824401557655. [CrossRef]
34. Kalia, P.; Kaur, N.; Singh, T. A Review of Factors Affecting Online Buying Behavior. *Apeejay J. Manag. Technol.* **2016**, *11*, 58–73. [CrossRef]
35. Bhat, S.A.; Darzi, M.A. Exploring the Influence of Consumer Demographics on Online Purchase Benefits. *FIIB Bus. Rev.* **2019**, *8*, 303–316. [CrossRef]
36. Hettich, D.; Hattula, S.; Bornemann, T. Consumer Decision-Making of Older People: A 45-Year Review. *Gerontologist* **2018**, *58*, e349–e368. [CrossRef]
37. Pascual-Miguel, F.J.; Agudo-Peregrina, Á.F.; Chaparro-Peláez, J. Influences of Gender and Product Type on Online Purchasing. *J. Bus. Res.* **2015**, *68*, 1550–1556. [CrossRef]
38. Garczarek-Bąk, U.; Szymkowiak, A. Gender Differences in E-Commerce. *Handel Wewnętrzny* **2018**, *4*, 250–261.
39. Hansen, T.; Møller Jensen, J. Shopping Orientation and Online Clothing Purchases: The Role of Gender and Purchase Situation. *Eur. J. Mark.* **2009**, *43*, 1154–1170. [CrossRef]
40. Naseri, M.B.; Elliott, G. Role of Demographics, Social Connectedness and Prior Internet Experience in Adoption of Online Shopping: Applications for Direct Marketing. *J. Target. Meas. Anal. Mark.* **2011**, *19*, 69–84. [CrossRef]
41. Kalia, P. Does Demographics Affect Purchase Frequency in Online Retail? *Int. J. Online Mark.* **2017**, *7*, 42–56. [CrossRef]
42. Luo, X.; Niu, C. *E-Commerce Participation and Household Income Growth in Taobao Villages (English)*; Policy Research Working Paper, no. WPS 8811; World Bank Group: Washington, DC, USA, 2019.
43. Agudo-Peregrina, Á.F.; Hernández-García, Á.; Acquila-Natale, E. The Effect of Income Level on E-Commerce Adoption. In *Encyclopedia of E-Commerce Development, Implementation, and Management*; IGI Global: Hershey, PA, USA, 2016; pp. 2239–2255.
44. Paun, C.; Ivascu, C.; Olteteanu, A.; Dantis, D. The Main Drivers of E-Commerce Adoption: A Global Panel Data Analysis. *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 2198–2217. [CrossRef]
45. Çebi Karaaslan, K. Determinants of Online Shopping Attitudes of Households in Turkey. *J. Model. Manag.* **2022**, *17*, 119–133. [CrossRef]

46. Karaaslan, K.Ç. Analysis of the Factors Affecting Credit Card Use and Online Shopping Attitudes of Households in Turkey with the Bivariate Probit Model. *Int. J. Electron. Financ.* **2022**, *11*, 189. [CrossRef]
47. Imran, M.; Asif, M.; Sajjad, W. Impact of Employment Status on Online Shopping Preferences: A Case Study of Women in Rawalpindi. *J. Bus. Insight Innov.* **2022**, *1*, 19–28.
48. Singh, K.; Basu, R. Online Consumer Shopping Behaviour: A Review and Research Agenda. *Int. J. Consum. Stud.* **2023**, *47*, 815–851. [CrossRef]
49. Wang, S.; Cheah, J.; Lim, X. Online Shopping Cart Abandonment: A Review and Research Agenda. *Int. J. Consum. Stud.* **2023**, *47*, 453–473. [CrossRef]
50. Molnar, C.; König, G.; Herbinger, J.; Freiesleben, T.; Dandl, S.; Scholbeck, C.A.; Casalicchio, G.; Grosse-Wentrup, M.; Bischl, B. General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*; Springer: Cham, Switzerland, 2022; pp. 39–68.
51. Orošnjak, M.; Beker, I.; Brkljač, N.; Vrhovac, V. Predictors of Successful Maintenance Practices in Companies Using Fluid Power Systems: A Model-Agnostic Interpretation. *Appl. Sci.* **2024**, *14*, 5921. [CrossRef]
52. Hamburg, M. *Basic Statistics: A Modern Approach*, 3rd ed.; Harcourt Brace Jovanovich: San Diego, CA, USA, 1985.
53. Burkolter, D.; Kluge, A. Online Consumer Behavior and Its Relationship with Socio-Demographics, Shopping Orientations, Need for Emotion, and Fashion Leadership. *J. Bus. Media Psychol.* **2011**, *2*, 20–28.
54. Howell, D.C. *Chi-Square Test—Analysis of Contingency Tables*; University of Vermont: Burlington, VT, USA, 2011.
55. Altman, N.; Krzywinski, M. Interpreting P Values. *Nat. Methods* **2017**, *14*, 213–214. [CrossRef]
56. Halsey, L.G.; Curran-Everett, D.; Vowler, S.L.; Drummond, G.B. The Fickle P Value Generates Irreproducible Results. *Nat. Methods* **2015**, *12*, 179–185. [CrossRef]
57. Altman, N.; Krzywinski, M. P Values and the Search for Significance. *Nat. Methods* **2017**, *14*, 3–4. [CrossRef]
58. Wagenmakers, E.-J. A Practical Solution to the Pervasive Problems of *p* Values. *Psychon. Bull. Rev.* **2007**, *14*, 779–804. [CrossRef]
59. Garcia, A.M.R.-R.; Puga, J.L. Deciding on Null Hypotheses Using P-Values or Bayesian Alternatives: A Simulation Study. *Psicothema* **2018**, *30*, 110–115. [CrossRef] [PubMed]