DOI: 10.1002/ksa.12389

EDITORIAL

Knee Surgery, Sports Traumatology, Arthroscopy WILEY

The artificial intelligence advantage: Supercharging exploratory data analysis

Abstract

Explorative data analysis (EDA) is a critical step in scientific projects, aiming to uncover valuable insights and patterns within data. Traditionally, EDA involves manual inspection, visualization, and various statistical methods. The advent of artificial intelligence (AI) and machine learning (ML) has the potential to improve EDA, offering more sophisticated approaches that enhance its efficacy. This review explores how AI and ML algorithms can improve feature engineering and selection during EDA, leading to more robust predictive models and data-driven decisions. Tree-based models, regularized regression, and clustering algorithms were identified as key techniques. These methods automate feature importance ranking, handle complex interactions, perform feature selection, reveal hidden groupings, and detect anomalies. Real-world applications include risk prediction in total hip arthroplasty and subgroup identification in scoliosis patients. Recent advances in explainable AI and EDA automation show potential for further improvement. The integration of AI and ML into EDA accelerates tasks and uncovers sophisticated insights. However, effective utilization requires a deep understanding of the algorithms, their assumptions, and limitations, along with domain knowledge for proper interpretation. As data continues to grow, AI will play an increasingly pivotal role in EDA when combined with human expertise, driving more informed, data-driven decision-making across various scientific domains.

Level of Evidence: Level V - Expert opinion.

KEYWORDS

artificial intelligence, exploratory data analysis, feature engineering, machine learning, orthopedic research

INTRODUCTION

Explorative data analysis (EDA) is a crucial step in any scientific project, aimed at uncovering valuable insights and patterns within the data. Traditional EDA techniques often involve manual inspection, visualization, and traditional statistical methods to understand the data's characteristics, identify outliers, and explore relationships between variables. However, the advent of artificial intelligence (AI) and machine learning (ML) has opened up new avenues for enhancing EDA, offering more sophisticated approaches that can improve the efficacy with which EDA is conducted.

AI and ML algorithms can play a significant role in feature engineering and selection, two critical components of EDA. Even AI and ML techniques that have been around for relatively long, like tree-based models, regularized regression, and clustering can automatically identify the most important features, uncover complex relationships between variables, and reveal hidden groupings or patterns within the data (Table 1). It is worth noting that some of these methods can be considered to be at the intersection of traditional statistical methods and modern ML [15]. These methods can streamline the feature engineering process, provide valuable insights for interaction terms (representing the joint effect of two or more features on the outcome), and can reveal the importance of undervalued variables, both in preparation for traditional statistics or further ML applications [9, 10].

Unsupervised learning algorithms like clustering can be leveraged to discover intrinsic groupings, detect anomalies, and summarize data during EDA. These techniques can uncover hidden segments, outliers, or trends that may not be immediately apparent, enabling more in-depth analysis and understanding of the data.

By integrating AI and ML into the EDA process, data scientists can not only accelerate certain tasks but also

Abbreviations: AI, artificial intelligence; EDA, explorative data analysis; GBM, gradient boosting machines; ML, machine learning; THA, total hip arthroplasty; XAI, eXplainable AI.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2024 The Author(s). *Knee Surgery, Sports Traumatology, Arthroscopy* published by John Wiley & Sons Ltd on behalf of European Society of Sports Traumatology, Knee Surgery and Arthroscopy.

TABLE 1 Machine learning techniques in explorative data analysis.

Technique	Description	Application in explorative data analysis
Tree-based models	Class of algorithms using a tree-like model of decisions. Includes decision trees, random forests, and gradient boosting machines (GBMs).	– Identifying important features: Ranks feature importance. – Detecting interactions: Captures complex feature interactions. – Visualization: Provides transparent decision-making.
Regularized regression	Adds a penalty to the regression model to prevent overfitting. Includes ridge regression, lasso regression, and elastic net.	– Handling multicollinearity: Shrinks coefficients. – Feature selection: Identifies and removes irrelevant features. – Model interpretability: Produces simpler models.
Clustering	Unsupervised learning technique grouping similar data points. Includes K-means, hierarchical clustering, and DBSCAN.	– Discovering patterns: Identifies natural groupings. – Anomaly detection: Detects outliers. – Cohort segmentation: Groups patients/study participants into similar segments.

uncover more sophisticated insights and patterns, ultimately leading to more robust and accurate predictive models or data-driven decisions.

AI/ML IMPROVES FEATURE ENGINEERING AND SELECTION DURING EDA

Algorithms like tree-based models, explainable boosted machines, or regularized regression can identify the most important features and their relationships with the target variable [1, 8].

These algorithms have built-in mechanisms for feature importance ranking, which can guide feature engineering efforts. For instance, tree-based models compute feature importance scores based on metrics that highlight the features that contribute most to the model's predictive power [8]. Regularized regression methods, like Lasso perform automatic feature selection by driving the coefficients of irrelevant features to zero, effectively eliminating them from the model [12]. For example, here are two common approaches to using Lasso for feature selection:

1. Lasso as the Final Model: Fit a Lasso regression model and use the nonzero coefficients to identify the selected features. This approach performs feature selection and model fitting simultaneously [12].
2. Two-Stage Approach: First, fit a Lasso model with a fixed penalty to identify a subset of relevant features. Then, fit a separate model (e.g., ordinary least squares, ridge regression) using only the selected features [6].

The two-stage approach, also known as the 'Relaxed Lasso', can be beneficial when the Lasso's shrinkage effect is too severe, leading to biased estimates for the selected features. By separating feature selection and model fitting, the second stage can estimate coefficients without the Lasso's shrinkage penalty [6].

It is important to note that while Lasso can handle multicollinearity better than ordinary least squares, it may arbitrarily drop one of a group of highly correlated features. In such cases, alternative methods like elastic net or manual feature selection may be preferable [14].

Venäläinen et al. employed Lasso regression to develop risk prediction models for common adverse outcomes after primary total hip arthroplasty (THA) [13]. Lasso was applied to the training cohort, which consisted of two-thirds of the data from the Finnish Arthroplasty Register, to identify subsets of variables that were most predictive of each outcome. By shrinking less important feature coefficients to zero, Lasso helped create parsimonious models that included only the most relevant predictors.

Furthermore, these algorithms can uncover complex, non-linear relationships between features and the target variable, providing insights for creating interaction terms or higher order polynomial features during feature engineering [4]. The hierarchical structure of tree-based models can also reveal feature combinations and decision rules that are predictive of the target, informing the creation of new, engineered features.

The ability to assist with ranking and selecting relevant features is a powerful aspect of these algorithms, streamlining the feature engineering process and reducing the risk of including irrelevant noise in the final model.

Clustering to find hidden groups and similarities

Unsupervised learning algorithms like clustering can be used to explore the inherent structure and groupings within the data during EDA [3]. This can reveal

14337347, 2024, 11, Downloaded from https://esskajournals.onlinelibrary.wiley.com/doi/10.1002/ksa.12389 by Cochrane Luxembourg, Wiley Online Library on [05/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

patterns, subgroups, or segments that may not be immediately apparent.

Clustering algorithms work by grouping similar data points together into clusters based on their proximity or similarity in the feature space [3]. The goal is to maximize the similarity within clusters while maximizing the dissimilarity between different clusters. This allows analysts to uncover hidden relationships, trends, or outliers that can provide valuable insights into the data.

Some key advantages of using clustering for EDA include:

1. *Discovering intrinsic groupings*: Clustering can automatically identify inherent groupings or segments within the data, which can be useful for customer segmentation, market analysis or identifying subpopulations in scientific studies.
2. *Detecting anomalies*: Outliers or anomalies that deviate significantly from the main clusters can be detected, potentially revealing interesting or unexpected cases.
3. *Data summarization*: By grouping similar data points together, clustering can provide a concise summary or representation of the data, facilitating understanding and further analysis.
4. *Feature exploration*: Clustering can help explore the relationships between features and their influence on the formation of clusters, potentially revealing important patterns or correlations.

Common clustering algorithms like K-means [5], hierarchical clustering, DBSCAN, and Gaussian mixture models are widely used for EDA across various domains, including customer analytics, image analysis, bioinformatics, and more. However, it is important to note that the choice of algorithm and its configuration can significantly impact the results, and domain knowledge is often required to interpret and validate the clusters.

Thong et al. used an unsupervised clustering method to group together the encoded representations of 3D spine reconstructions generated by a stacked auto-encoder, a type of ML algorithm [11]. This data-driven approach revealed 11 distinct subgroups amongst 915 surgical adolescent idiopathic scoliosis patients, demonstrating that even within the established Lenke classification types, there are subgroups characterized by specific combinations of curve location, kyphosis, and lordosis.

Recent AI/ML advances that can enhance EDA

While tree-based models, regularized regression, and clustering have clear potential to enhance EDA, more recent AI and ML advances can provide additional benefits. In the area of eXplainable AI (XAI), several techniques have been proposed that can help find smaller and more understandable models with fewer features which can be important for further analysis. As one example, a recent technique can produce families of ML models with high predictive power and then study the feature importance for the family as a whole [2]. This avoids potential problems of the methods mentioned above that calculate feature importance based on a single, preferred model. By considering the so-called Rashomon set of all good models the feature importance scores are more stable and less sensitive to minor variations in model training.

There are even approaches to automate larger parts of the EDA process by leveraging modern AI and ML approaches [7]. Three main types of solutions have been studied: EDA recommender system, user interestingness prediction, and full EDA automation. While the former two augments the scientist by recommending which aspects of the data they can explore further, the latter uses sequence learning and generative AI methods based on Deep Learning to generate whole EDA reports with the steps, analyses, visualizations, and conclusions of the AI 'user' which the scientist can then review, use, and build on [7].

CONCLUSION

The integration of AI and ML into EDA offers a powerful set of tools for enhancing insights and understanding data. By leveraging the aforementioned techniques, data scientists can improve feature selection, uncover complex relationships between variables, and reveal hidden groupings or patterns. These AI-driven approaches streamline feature engineering and provide valuable guidance for creating new engineered features. However, effective utilization requires a deep understanding of the algorithms, their assumptions and limitations, along with domain knowledge for proper interpretation. As data continues to grow, AI will play an increasingly pivotal role in EDA and drive more informed, data-driven decision-making when combined with human expertise.

AUTHOR CONTRIBUTIONS

All listed authors have contributed substantially to this work. Felix C. Oettl performed literature research. Felix C. Oettl, Jacob F. Oeding and Christophe Ley performed primary manuscript preparation. Editing and final manuscript preparation were performed by Robert Feldt, Michael T. Hirschmann, Kristian Samuelsson, and Felix C. Oettl. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

The authors have no funding to report.

14337347, 2024, 11, Downloaded from https://esskajournals.onlinelibrary.wiley.com/doi/10.1002/ksa.12389 by Cochrane Luxembourg, Wiley Online Library on [05/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

ETHICS STATEMENT

Not applicable.

Felix C. Oettl[1,2]
Jacob F. Oeding[3,4,5]
Robert Feldt[6]
Christophe Ley[7]
Michael T. Hirschmann[8]
Kristian Samuelsson[3,4,9]
ESSKA Artificial Intelligence Working Group

[1]*Hospital for Special Surgery, New York, New York, USA*
[2]*Department of Orthopedic Surgery, Balgrist University Hospital, University of Zürich, Zurich, Switzerland*
[3]*Department of Orthopaedics, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden*
[4]*Sahlgrenska Sports Medicine Center, Göteborg, Sweden*
[5]*Mayo Clinic Alix School of Medicine, Mayo Clinic, Rochester, Minnesota, USA*
[6]*Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden*
[7]*Department of Mathematics, University of Luxembourg, Esch-sur-Alzette, Luxembourg*
[8]*Department of Orthopedic Surgery and Traumatology, Kantonspital Baselland, Liestal, Switzerland*
[9]*Department of Orthopaedics, Sahlgrenska University Hospital, Mölndal, Sweden*

Correspondence
Kristian Samuelsson, University of Gothenburg, Göteborgsvägen 31, 431 80 Mölndal, Sweden.
Email: kristian.samuelsson@gu.se

ORCID

Felix C. Oettl http://orcid.org/0000-0001-9721-685X
Jacob F. Oeding http://orcid.org/0000-0002-4562-4373
Robert Feldt http://orcid.org/0000-0002-5179-4205
Christophe Ley http://orcid.org/0000-0003-2751-8902
Michael T. Hirschmann http://orcid.org/0000-0002-4014-424X
Kristian Samuelsson http://orcid.org/0000-0001-5383-3370

REFERENCES

1. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. & Elhadad, N. (2015) Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission, *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. Sydney, NSW, Australia.
2. Donnelly, J., Katta, S., Rudin, C. & Browne, E.P. (2024) The Rashomon importance distribution: getting RID of unstable, single model-based variable importance. *ArXiv*. [Preprint]
3. Eckhardt, C.M., Madjarova, S.J., Williams, R.J., Ollivier, M., Karlsson, J., Pareek, A. et al. (2023) Unsupervised machine learning methods and emerging applications in healthcare. *Knee Surgery, Sports Traumatology, Arthroscopy*, 31, 376–381. Available from: https://doi.org/10.1007/s00167-022-07233-7
4. Liu, Y., Li, Y., Yang, W. & Hu, J. (2023) Exploring nonlinear effects of built environment on jogging behavior using random forest. *Applied Geography*, 156, 102990. Available from: https://doi.org/10.1016/j.apgeog.2023.102990
5. Lloyd, S. (1982) Least squares quantization in PCM | IEEE Journals & Magazine | IEEE Xplore. *IEEE Transactions on Information Theory*, 28(2), 129–137.
6. Meinshausen, N. (2007) Relaxed lasso. *Computational Statistics & Data Analysis*, 52, 374–393.
7. Milo, T. & Somech, A. (2020) Automating exploratory data analysis via machine learning: an overview, *Proceedings of the 2020 ACM SIGMOD international conference on management of data*. Portland, OR, USA.
8. Nembrini, S., König, I.R. & Wright, M.N. (2018) The revival of the Gini importance? *Bioinformatics*, 34, 3711–3718. Available from: https://doi.org/10.1093/bioinformatics/bty373
9. Pruneski, J.A., Pareek, A., Kunze, K.N., Martin, R.K., Karlsson, J., Oeding, J.F. et al. (2023) Supervised machine learning and associated algorithms: applications in orthopedic surgery. *Knee Surgery, Sports Traumatology, Arthroscopy*, 31, 1196–1202. Available from: https://doi.org/10.1007/s00167-022-07181-2
10. Pruneski, J.A., Williams, R.J., Nwachukwu, B.U., Ramkumar, P.N., Kiapour, A.M., Martin, R.K. et al. (2022) The development and deployment of machine learning models. *Knee Surgery, Sports Traumatology, Arthroscopy*, 30, 3917–3923. Available from: https://doi.org/10.1007/s00167-022-07155-4
11. Thong, W., Parent, S., Wu, J., Aubin, C.E., Labelle, H. & Kadoury, S. (2016) Three-dimensional morphology study of surgical adolescent idiopathic scoliosis patient from encoded geometric models. *European Spine Journal*, 25, 3104–3113. Available from: https://doi.org/10.1007/s00586-016-4426-3
12. Tibshirani, R. (1997) The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16, 385–395. Available from: https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3
13. Venäläinen, M.S., Panula, V.J., Klén, R., Haapakoski, J.J., Eskelinen, A.P. & Manninen, M.J. et al. (2021) Preoperative risk prediction models for short-term revision and death after total hip arthroplasty: data from the Finnish arthroplasty register. *JBJS Open Access*, 6, e20.00091. Available from:. https://doi.org/10.2106/JBJS.OA.20.00091
14. Zou, H. & Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67, 301–320. Available from: https://doi.org/10.1111/j.1467-9868.2005.00503.x
15. Zsidai, B., Kaarre, J., Narup, E., Hamrin Senorski, E., Pareek, A. & Grassi, A. et al. (2024) A practical guide to the implementation of artificial intelligence in orthopaedic research-Part 2: a technical introduction. *Journal of Experimental Orthopaedics*, 11, e12025. Available from: https://doi.org/10.1002/jeo2.12025

14337347, 2024, 11, Downloaded from https://esskajournals.onlinelibrary.wiley.com/doi/10.1002/ksa.12389 by Cochrane Luxembourg, Wiley Online Library on [05/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License