Decoding Employee Attrition: A Unified Approach with XAI and AHP

Gabriel Marín Díaz[†] *and José Javier Galán Hernández*
Faculty of Statistics, Complutense University
Madrid, 28040

[†]*E-mail: gmarin03@ucm.es*
www.ucm.es

In the face of escalating employee attrition challenges, organizations are increasingly relying on artificial intelligence (AI) to predict and address turnover. This paper explores the application of explainable AI (XAI) to identify potential employee turnover, analyzing its impact on organizational productivity and stability. The second section focuses on AI techniques that leverage historical data to forecast attrition, enabling proactive interventions. The third part introduces XAI to enhance model transparency, providing HR professionals with deeper insights to develop targeted retention strategies aligned with individual employee needs. Integrating the Analytic Hierarchy Process (AHP) model becomes imperative to assign weights to criteria identified by AI as significant. This incorporation aims to introduce the human factor into decision-making.

Keywords: XAI; Artificial Intelligence; AHP; Human Resources; Talent Attraction.

1. Introduction

The contemporary labor market, particularly within Human Resources (HR), is undergoing rapid evolution, marked by the pervasive challenge of 'involuntary turnover'—employee departures contrary to employer intentions [1]. This phenomenon, accelerated by the pandemic and the 'Great Resignation' since 2021, poses multidimensional challenges for enterprises globally, resulting in tangible and intangible costs.

Recent findings from Randstad Research's "Labour Turnover Report in Spain" [2] reveal a significant 38.5% increase in turnover rates for 38.5% of Spanish enterprises in the past twelve months. While 77% attribute turnover to employees pursuing opportunities elsewhere, secondary factors include workforce concerns

arising from recent crises, aspirational salary expectations, and a demand for increased flexibility.

Addressing this, AI emerges as a promising solution, particularly through Explainable AI (XAI) methodologies [3]. XAI not only sheds light on predictive models but also identifies critical factors contributing to predictions [4]. This knowledge empowers proactive implementation of targeted retention strategies, aligning with employee needs [5].

Incorporating AHP into the methodology allows for the assignment of weights to the characteristics of the learning model, considering the context of interpretability [6]. This strategic weighting enhances the overall effectiveness of the model in addressing the nuances of employee turnover.

The proposed approach facilitates actionable insights, allowing organizations to understand, intervene, and reshape strategies to mitigate turnover risks.

By emphasizing the interpretability of the XAI model, decision-makers can pinpoint pivotal factors driving attrition within their context, leading to targeted interventions such as fostering a supportive culture and addressing compensation concerns. The methodology also enhances recruitment by identifying attributes associated with employee longevity, guiding efforts towards successful and enduring employment relationships.

Advocating for a paradigm shift, this article envisions organizations proactively implementing strategies for enhanced employee retention and optimized recruitment processes. The subsequent sections detail the development and application of the XAI model, reviewing current XAI methodologies, discussing the methodological framework, applying the XAI model in a business context, integrating AHP to assign weights to learning model characteristics for enhanced interpretability, and presenting discussions, conclusions, and future work..

2. Related Work

Studies related to employee attrition using machine learning techniques are scarce, with a total of 26 studies identified based on the following search variables in Web of Science: TS = (ATTRITION) AND TS = (EMPLOYEE) AND TS = (MACHINE LEARNING), Figure 1. These studies do not address algorithm interpretability; rather, they focus on incorporating predictive models to facilitate early detection of employee attrition. However, it is noteworthy that interpretability is not discussed or employed in these studies.

An important research frontier is unfolding within corporate settings, encompassing the integration of predictive models and interpretability into decision-making processes. Moreover, the inclusion of a pivotal determinant in

decision-making, the human factor (AHP), provides a distinctive opportunity for exploration and advancement in this domain.

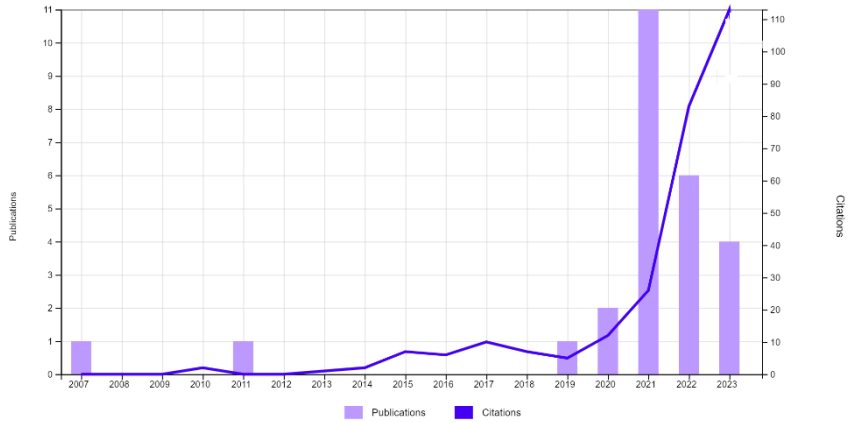

Fig. 1. Publications and citations (26). TS = (ATTRITION) AND TS = (EMPLOYEE) AND TS = (MACHINE LEARNING).

The predominant approach at present emphasizes the adoption of model-agnostic interpretability tools. The strategy involves separating the interpretation method from the employed model, facilitating the automation of interpretability. The use of agnostic methods allows for the replacement of both the learning model and the interpretation method, resulting in significantly scalable capabilities [7].

3. Methodology

The methodology we are going to present is based on the Knowledge Discovery in Databases (KDD) framework [8], which employs principles to extract actionable insights from data, with a particular focus on the variables determining employee departures. Subsequently, aligning hiring and promotion strategies with these identified variables allows for a targeted and informed approach to human resource management. The incorporation of XAI and AHP emphasizes the commitment to transparency and understanding in decision-making processes.
Explainable AI contributes to the interpretability of models, ensuring that decisions driven by artificial intelligence are accessible and understandable to human decision-makers. Simultaneously, AHP, as a decision-making model, provides a systematic and mathematical foundation for evaluating and prioritizing

variables, incorporating both psychological and mathematical principles into the decision-making process. The fusion of these methodologies aims to create a comprehensive and intelligible framework for informed and strategic decision-making in the realm of human resource management, Figure 2.

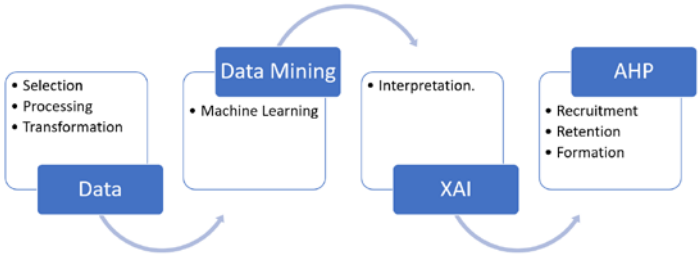

Fig. 2. KDD - Employee attrition analysis and HR recruitment, retention and training plans.

3.1. *Data Object*

The initial phase in the Knowledge Discovery in Databases (KDD) process is the "Data" stage [9]. This stage is primarily focused on comprehending the domain and establishing the fundamental groundwork for the entire data analysis undertaking.

Data processing encompasses the tasks of cleaning, transforming, and organizing raw data into a format suitable for subsequent analysis [10]. This critical step addresses challenges such as handling missing values, identifying outliers, and mitigating noise within the dataset.

This process entails transforming raw data into an appropriate format that enhances its quality and facilitates effective analysis [11]. Techniques employed in this stage include normalization, aggregation, discretization, and attribute construction.

3.2. *Data Mining*

In this employee attrition case study, we'll follow a structured workflow, applying various classification algorithms like Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, or Gradient Boosting [3]. Each algorithm will predict an employee's likelihood of leaving based on relevant features. After predictions are generated, we'll evaluate each employee's attrition tendency, measuring model accuracy and constructing AUC/ROC curves for assessing predictive ability. Additionally, a detailed confusion matrix will be

created for each algorithm, offering insights through metrics like precision, recall, specificity, and F1 score.

3.3. *XAI*

Throughout time, algorithms have primarily concentrated on exploratory data analysis to support decision-making processes. Although acquiring and analyzing data is often adequate for well-informed decisions, there are situations where validating decisions based on patterns becomes essential. In this regard, AI holds particular importance. Historically, machine learning algorithms have placed a premium on attaining high precision in results, progressively gravitating towards black-box algorithms. As a result, the interpretability of these decision-making algorithms has been restricted, leading to the emergence of interpretable algorithms [7], Figure 3. These interpretable models are explored through various methods outlined below.

1. Inherent Interpretable Models [12]: Models such as Linear Regression, Decision Trees, and Logistic Regression feature transparent structures and explicit relationships between features and predictions, making them inherently interpretable.
2. Rule-based Interpretable Models [13]: This category aims to simplify intricate patterns into clear-cut rules, promoting transparency in decision-making. Models like decision trees and rule-based systems utilize simple, understandable rules, establishing direct connections between input features and predictions.
3. Local Interpretable Models [13]: Focusing on explaining predictions at the instance level, these models offer a detailed understanding of how the model arrived at a specific prediction for a particular data point. Techniques like LIME fall into this category, providing insights into the model's behavior around a specific data point.
4. Model-Agnostic Approaches [7]: Model-agnostic methods, including SHAP, LIME, and Permutation Feature Importance, offer interpretability for a wide range of models without relying on their internal architecture, enhancing transparency across diverse machine learning models.
5. Feature Importance Techniques [14]: Techniques like ELI5 and Partial Dependence Plot (PDP) highlight the significance of each feature in the model's predictions, offering a straightforward understanding of feature contributions.

6. Surrogate Models [15]: Simpler and more interpretable models trained to approximate the predictions of complex models, surrogate models serve as a bridge between highly complex, less interpretable models and the need for comprehensible insights into their functioning.
7. Visualizations and Plots [16]: Techniques like partial dependence plots, SHAP summary plots, and feature contribution plots offer intuitive graphical representations of the model's behavior and feature influences.

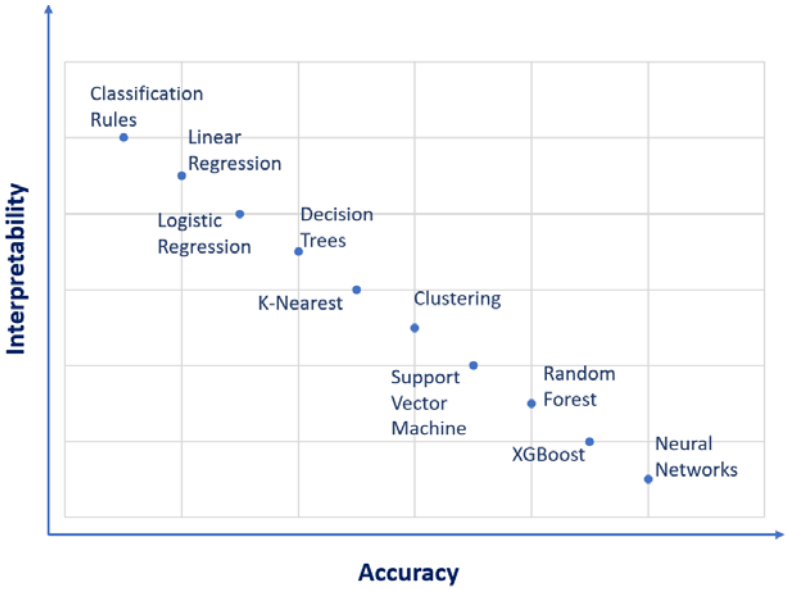

Fig. 3. Accuracy vs Interpretability.

3.4. *AHP*

The Analytic Hierarchy Process (AHP) is a decision-making methodology that can significantly contribute to the personnel selection process by enhancing interpretability. AHP involves structuring complex decision problems into a hierarchical framework, allowing decision-makers to systematically evaluate and prioritize various factors. This structured approach aids in understanding the relationships and importance of different criteria in the decision-making process. In the context of personnel selection, AHP can be employed by first defining the overarching goal, which, in this case, is selecting the most suitable candidate for a position. The decision criteria, such as skills, experience, cultural fit, and potential for growth, are then identified and organized hierarchically, to

summarize, the criteria that the Interpretability model has deemed predominant in the employee attrition process will be crucial in the company's recruitment, retention, and personnel training processes. Subsequently, pairwise comparisons are conducted to determine the relative importance of each criterion in relation to the goal. These comparisons are facilitated by decision-makers, providing a qualitative assessment of the significance of one criterion over another, Figure 4. The pairwise comparisons generate a set of numerical values that represent the relative weights of each criterion. A mathematical process is then applied to derive a consistent set of weights that reflect the decision-makers' preferences. The final outcome is a prioritized list of criteria based on their importance in achieving the overall goal.

The AHP offers a valuable contribution to the personnel selection process, leveraging the significant criteria identified through AI interpretability models. AHP provides a structured and coherent framework, introducing a human-centric approach to decision-making.

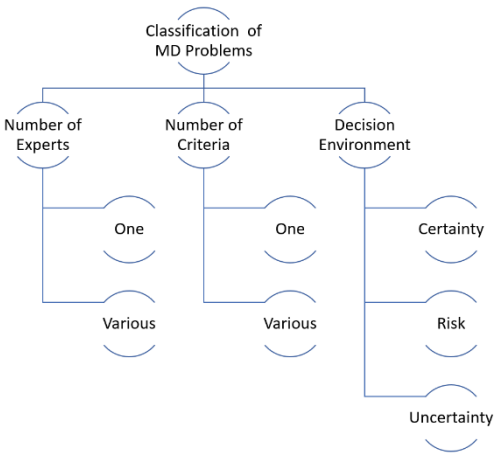

Fig. 4. Classification of MD Problems.

4. Conclusions

The methodology employed, leveraging eXplainable Artificial Intelligence (XAI) to identify critical employee attrition criteria and subsequently applying Analytic Hierarchy Process (AHP) decision theory, holds promising implications for diverse organizational processes.

The utilization of XAI in discerning key attrition criteria not only enhances transparency in decision-making but also provides valuable insights applicable to various facets of human resource management. The identified criteria, proven to be significant in predicting employee turnover, can be effectively translated and incorporated into distinct operational domains such as personnel selection, retention strategies, and employee training programs.

Furthermore, the detailed process elucidated in this study is not confined to a specific industry or sector; rather, it serves as a versatile framework that can be extended to any business context. By systematically integrating XAI and AHP, along with fuzzy linguistic models, organizations gain a comprehensive understanding of the factors influencing workforce dynamics. This holistic approach not only aids in the formulation of targeted and effective strategies for personnel-related processes but also promotes adaptability, ensuring its applicability across diverse business environments.

References

[1] S. Rughoobur-Seetah, "The Unprecedented Lockdown: The consequences of job loss," *Zagreb Int. Rev. Econ. Bus.*, vol. 24, no. 2, pp. 1–23, 2021, doi: 10.2478/zireb-2021-0008.

[2] Randstad Research, "Informe de rotación laboral en España," *Randstad Res.*, 2022.

[3] G. Marín Díaz, J. J. Galán, and R. A. Carrasco, "XAI for Churn Prediction in B2B Models: A Use Case in an Enterprise Software Company," *Mathematics*, vol. 10, no. 20, 2022, doi: 10.3390/math10203896.

[4] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, pp. 1–45, 2021, doi: 10.3390/e23010018.

[5] D. Mishra, "Review of literature on factors influencing attrition and retention," *Int. J. Organ. Behav. Manag. Perspect.*, vol. 2, no. 3, pp. 435–445, 2013.

[6] O. S. Vaidya and S. Kumar, "Analytic hierarchy process: An overview of applications," *Eur. J. Oper. Res.*, vol. 169, no. 1, pp. 1–29, 2006, doi: 10.1016/j.ejor.2004.04.028.

[7] C. Molnar, "Interpretable Machine Learning. A Guide for Making Black Box Models Explainable.," *Book*, p. 247, 2019, [Online]. Available: https://christophm.github.io/interpretable-ml-book.

[8] U. Shafique and H. Qaiser, "A Comparative Study of Data Mining Process Models (KDD , CRISP-DM and SEMMA)," *Int. J. Innov. Sci. Res.*, vol. 12, no. 1, pp. 217–222, 2014, [Online]. Available: http://www.ijisr.issr-journals.org/.

[9] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, “Leakage in data mining: Formulation, detection, and avoidance,” *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 4, pp. 1–21, 2012, doi: 10.1145/2382577.2382579.

[10] F. Kamiran and T. Calders, *Data preprocessing techniques for classification without discrimination*, vol. 33, no. 1. 2012.

[11] S. A. Alasadi and W. S. Bhaya, “Review of Data Preprocessing Techniques,” *Journal of Engineering and Applie Sciences*, vol. 12, no. 16. pp. 4102–4107, 2017.

[12] Z. C. Lipton, “The mythos of model interpretability,” *Commun. ACM*, vol. 61, no. 10, pp. 35–43, 2018, doi: 10.1145/3233231.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’ Explaining the Predictions of Any Classifier,” *NAACL-HLT 2016 - 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Demonstr. Sess.*, pp. 97–101, 2016, doi: 10.18653/v1/n16-3020.

[14] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, “Accurate intelligible models with pairwise interactions,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. Part F1288, pp. 623–631, 2013, doi: 10.1145/2487575.2487579.

[15] L. Breiman, “Stacked regressions,” *Mach. Learn.*, vol. 24, no. 1, pp. 49–64, 1996, doi: 10.1023/A:1018046112532.

[16] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Section 2, pp. 4766–4775, 2017.