Human Resources Analytics/Employee Retention: Predicting Employee Turnover

Tomás Coelho[1], Eva Morais[1], Adelaide Cerveira[1], Eduardo Pires[2], and Ricardo Silva[3]

[1]Department of Mathematics, University of Trás-os-Montes and Alto Douro, 5000-801 Vila Real, Portugal; al76554@alunos.utad.pt (T.C) ; cerveira@utad.pt(A.C); evamorais@utad.pt (E.M)
[2]Department of Engineering, University of Trás-os-Montes and Alto Douro, 5000-801 Vila Real, Portugal; epires@utad.pt (E.P)
[3]Arquiconsult, 2675-432 Odivelas, Lisboa, Portugal ; ricardo.silva@arquiconsult.com

***Abstract*—As employees go about their daily tasks, they generate a vast amount of digital data, including information about resignations, voluntary departures, promotions and other events. Processing and analyzing this data offers significant opportunities for understanding human behavior in the workplace and how it can be improved. Employees are a company's most valuable resource, including each and every one of them. Because of the fundamental role that employees play in the success of an organization, measuring employee turnover has become one of the most important metrics for companies today. In this paper, we focus on predicting employee turnover and enhancing employee retention through the application of Machine Learning and people analytics. By analyzing an open-source dataset from IBM HR Analytics, we explore various factors that influence employee turnover. Our methodology involves data preprocessing, exploratory data analysis, and the implementation of Machine Learning models to predict which employees are at risk of leaving.**

***Keywords*—People Analytics, Employee Attrition, Human Resources, Machine Learning, Artificial Intelligence**

A. Abbreviations and Acronyms

ML- Machine Learning
AI- Artificial Intelligence
HR - Human Resources
XGBoost- Extreme Gradient Boosting
SMOTE - Synthetic Minority Over-Sampling Technique

I. Introduction

Employee retention refers to a company's capacity to retain its valuable employees engaged and satisfied, thereby encouraging them to remain with the organization for an extended period. It is not merely about preventing people from leaving; rather, it is about fostering a positive work environment where employees feel motivated and see a future for themselves within the company. It is an inevitable consequence of the employment relationship that an individual will eventually leave their position with an organization. This departure may be due to a variety of reasons. The term *attrition* is defined as the departure of any employee. However, when attrition begins to cause significant financial loss to the organization, it is essential to monitor the situation closely. The process of hiring new employees requires a significant amount of resources. To avoid the constant need to hire and maintain a strong team, it is necessary to prevent employee turnover and try to retain the valuable and hard-working employees. This not only helps the company to save money by preserving its resources, but also to maintain the stability of its team. People Analytics, also known as HR Analytics, is an increasingly important approach for companies around the world, including Portugal, when it comes to recruiting, retaining talent and maximizing employee performance. Our work seeks to predict voluntary departures (Atrittion) by employees using Machine Learning models, to retain talent and prevent turnover. [1][2].

There are several reasons why employee retention is crucial:

- Reduced costs. High turnover rates are expensive. The cost of recruiting, hiring, and onboarding new employees can be significant [3].
- Improves productivity. Experienced employees are more productive and efficient than new hires. They have a better understanding of the company's processes and can contribute more effectively. Experienced employees possess a wealth of valuable knowledge and skills specific to the company and its industry. The loss of these employees can result in the loss of this institutional knowledge, which can be costly to replace in terms of both time and resources [4].
- Boosts morale. A team with high turnover can experience low morale. When employees see their colleagues leaving, it can lead to feelings of uncertainty and discouragement. Retaining employees fosters a sense of stability and community. The retention of existing employees avoids the expenses associated with recruiting, hiring, and onboarding new employees [4].

Furthermore, it is anticipated that the majority of future employment opportunities have yet to be conceived. Consequently, numerous companies and organisations have augmented their expenditure on employee training, encompassing skills development, continuous training, compliance training, and lifelong learning. This is intended to enhance employee retention, boost productivity, enhance competitiveness, and ensure that employees remain up to date. Regrettably, despite the substantial investment in employee training, employees have been known to depart from the company or decline job

offers following training [5].

The process of predictive analytics, which involves the collection and analysis of data using techniques such as Machine Learning, artificial intelligence, and statistical models, enables the identification of patterns that can predict future behavior. This approach has emerged as a powerful tool in various domains for companies, and is now being applied in the field of human resources, offering a means of predicting and preventing employee turnover [6]. Leading companies such as Google, Amazon, Microsoft, and others have distinguished themselves by innovating in the strategic use of people analytics to enhance their human resources practices and improve organizational outcomes. For example, Google has analyzed its employees' performance data to optimize its recruitment and selection processes, which is one of the primary factors contributing to its remarkable success and corporate culture [7].

These are some of the key questions we aim to answer within this topic:

- **Which of our employees are most likely to leave in the coming months?**
- **What are the main reasons for employees leaving?**
- **What measures should the company take to retain its employees?**
- **What are the best and worst characteristics and attributes for potential candidates to be recruited?**

The rest of this paper is organized as follows. In Section II, we review the key factors influencing employee retention and explore recent studies that have utilized advanced data analytics and Machine Learning techniques to predict employee turnover and identify at-risk employees. Section III explains what Machine Learning is and how the models work, providing insights into its application in the context of employee retention. Section IV gives a brief overview of the dataset used in this study. In Section V, we describe the data pre-processing methodology used before applying the Machine Learning models. In Section VI, we present the models used and their results, comparing their performances to identify the best-performing model. These results are then used in a PowerBI Dashboard to provide a tool for companies to identify employees at risk. Finally, Section VII concludes the paper with a summary of the findings and suggestions for future research.

II. Literature review

Key factors influencing employee retention identified in the literature include job satisfaction, organizational commitment, work-life balance, career development opportunities, and recognition and rewards [8], [9]. Effective retention strategies often involve creating a positive work environment, offering competitive compensation packages, and providing professional growth opportunities.

Recent studies have leveraged advanced data analytics and Machine Learning techniques to predict employee turnover and identify at-risk employees. For example, Kaur et al. [10] utilized Machine Learning algorithms to analyze employee data and predict turnover with high accuracy. Similarly, other researchers have explored the use of people analytics to gain insights into employee behavior and develop targeted retention strategies [11]. Moreover, the implementation of strategic HR practices, such as personalized career development plans and flexible work arrangements, has shown promising results in enhancing employee retention [12], [13]. These practices not only address the diverse needs of employees but also foster a culture of engagement and loyalty.

III. Machine Learning

Machine learning (ML) has established itself as a powerful tool in various areas of research, from computer vision and bioinformatics to natural language processing and object detection. This rise is due to the development of sophisticated techniques that allow machines to learn from data, opening up a range of possibilities for analyzing and solving complex problems [14].

Within *Machine Learning* we have 5 different types of algorithms:

- **Supervised Learning**: Supervised learning is an automatic learning technique where the algorithm receives a set of labeled data. Each piece of data has input (independent) variables and an output (dependent) variable that we want to predict. The goal is to build a model that, by analyzing the input variables, can accurately predict the output variable [15].

- **Unsupervised Learning**: Unlike supervised learning, unsupervised learning deals with unlabeled data. In other words, the data does not have a predefined output variable. The aim here is to explore the data and discover patterns or groupings that can be used in the learning process [16].

- **Reinforcement Learning:** Reinforcement learning differs from previous approaches in that the algorithm interacts with an environment and learns from the consequences (rewards or penalties) of its actions [14], [16].

- **Semi-Supervised Learning:** Semi-supervised learning falls between supervised and unsupervised learning. It uses a small amount of labeled data and a large amount of unlabeled data. This method is useful when labeling data is expensive or time-consuming, but unlabeled data is abundant [17].

- **Transfer Learning:** Transfer learning focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. This approach is particularly useful when there is a limited amount of data for the target task, allowing the model to leverage previously acquired knowledge from a similar task [18].

In this project it was used supervised learning because the problem at hand involves predicting a binary variable,

specifically whether an employee left voluntarily (yes or no). Supervised learning is suitable for this task because it involves using a labeled dataset where each instance includes input features (such as employee characteristics and work-related attributes) and the corresponding output label (whether the employee left or not, Atrittion). The goal is to train a model that can accurately predict the output label for new, unseen data.

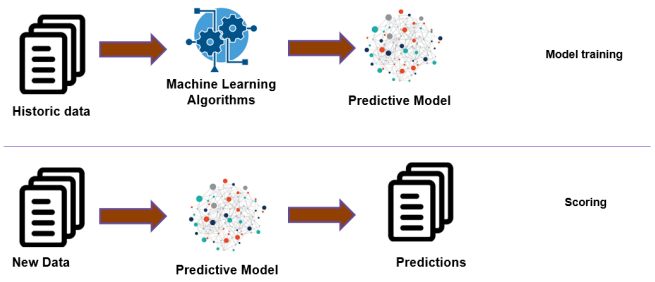


Fig. 1: Structure of the use of Machine Learning models [16].

As illustrated in Fig. 1, the process begins with the utilization of historical data, which is then subjected to Machine Learning algorithms with the objective of developing a predictive model. The model is constructed by the algorithms identifying patterns and relationships within the data. Once the predictive model has been built, it can be applied to both new and existing data to make predictions, thereby enabling the forecasting of future outcomes based on past trends.

In the context of employee retention, the historical data encompasses information from both past and current employees, including attributes such as age, job role, education, performance ratings, salary, job satisfaction, and other pertinent characteristics. It is particularly important to note that this dataset includes a binary feature, designated as "Attrition," which indicates whether an employee has departed from the company (coded as 1) or remains employed (coded as 0). The new dataset consists of information from current employees only, and the objective is to utilize the predictive model derived from the analysis of historical data to predict which of these employees are at risk of leaving the company.

After this process, the models are evaluated and compared using classical performance metrics[26], [27], including:

1) **Precision**: The ratio of correctly predicted positive observations to the total predicted positives, indicating the accuracy of the positive predictions.
2) **Recall:** The ratio of correctly predicted positive observations to all actual positives, measuring the model's ability to identify positive instances.
3) **F1-Score:** The harmonic mean of precision and recall, providing a balance between the two metrics.
4) **Accuracy:** The ratio of correctly predicted observations (both positives and negatives) to the total observations, providing a general measure of the model's performance.
5) **AUC Score (Area Under the Curve):** The area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (recall) against the false positive rate. It provides an aggregate measure of performance across all classification thresholds .

These metrics provided a comprehensive evaluation of the models, allowing us to determine the most effective algorithm for predicting employee attrition and identifying those at risk of leaving the company voluntarily.

IV. Dataset

In order to predict employee turnover, it was used an open source database, the IBM HR Analytics Employee Attrition & Performance dataset [19]. The dataset comprises 2,940 employees, of whom 2,466 remained with the company and 474 resigned, as we can see in Fig.2. It encompasses a plethora of features pertaining to employees' demographic characteristics, job satisfaction, performance, work environment, and other factors that may potentially influence turnover. The dataset includes a binary target variable indicating whether an employee has left a company (coded as 1) or stayed with it (coded as 0). This dataset represents a valuable resource for Machine Learning research and is readily accessible on platforms such as Kaggle, facilitating experimentation and analysis. The utilization of this dataset enables the development of predictive models to identify factors influencing employee turnover and the implementation of strategies to retain valuable employees.

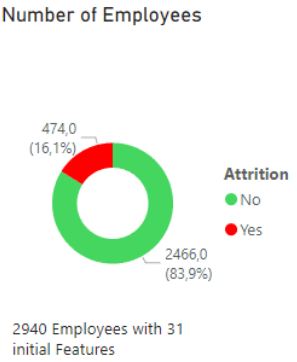


Fig. 2: IBM HR Analytics.

V. Work Methodology

In this study, the analysis was conducted using Python, specifically within a Jupyter Notebook environment, which allowed for an interactive and efficient workflow for data manipulation, model building, and evaluation.

As we can see in Fig. 3, once the data was collected, it was subjected to preliminary processing. During this stage, null and duplicate values were identified and removed, and irrelevant or constant features were eliminated to streamline the dataset and enhance the quality of the analysis [20]. The next steps for the data cleaning and processing were the following ones:

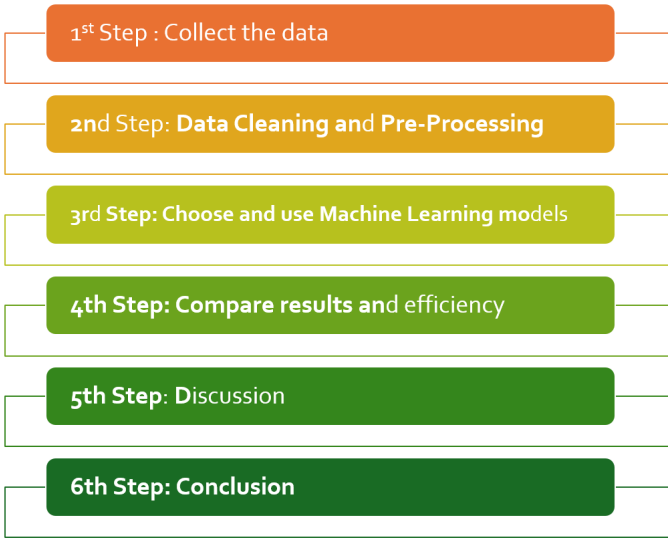


Fig. 3: Work Methodology.

1) **Encoding Categorical Variables:** Categorical variables, which are non-numeric and represent categories (e.g., 'Department', 'Job Role'), were converted into numerical format using the OneHotEncoder. The OneHotEncoder creates binary (0 or 1) columns for each category within a categorical feature, allowing Machine Learning algorithms to process these values [21]. For example, if the 'Department' feature includes 'Sales', 'Engineering', and 'HR', the OneHotEncoder will generate three new columns, each indicating the presence (1) or absence (0) of one of these departments for a given record.
2) **Normalizing the Data:** To ensure that features with different scales do not disproportionately influence the model, data normalization was performed using the StandardScaler. This technique standardizes the features by removing the mean and scaling to unit variance. As a result, each feature will have a mean of 0 and a standard deviation of 1, making them comparable on the same scale. This step is crucial for algorithms that rely on distance calculations, such as logistic regression or k-nearest neighbors [22].
3) **Feature Selection:** To improve the model's performance and reduce overfitting, feature selection was performed using the SelectKBest method with the ANOVA F-Value as the scoring function [23]. This method selects the top k features that have the highest correlation with the target variable. In this study, we chose k=26, identifying the most significant features from the dataset. The selected features were then used to transform the training and testing datasets accordingly.
4) **Balancing the Dataset:** To address class imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. The dataset comprises 2,940 employees, of whom 2,466 remained with the company and 474 resigned. This indicates a significant class imbalance, where the number of employees who stayed (majority class) vastly outnumbers those who resigned (minority class). SMOTE (Synthetic Minority Over-sampling Technique) helps to balance a dataset by creating new, similar examples from the underrepresented group. It does this by picking two close examples from the minority group and blending them to make a new, synthetic example. This process helps to even out the number of examples in each group, making the dataset more balanced [24]. This step is crucial in our analysis for several reasons:

   - **Improving Model Performance:** Machine learning models can become biased towards the majority class when trained on imbalanced datasets, leading to poor predictive performance on the minority class. By using SMOTE to balance the class distribution, the model is better equipped to learn from both classes effectively.
   - **Ensuring Fairness:** In the context of employee attrition, accurately predicting resignations is critical for the company to implement effective retention strategies. An imbalanced dataset could result in a model that consistently predicts employees will stay, missing valuable opportunities to identify those at risk of leaving.
   - **Enhancing Generalization:** Balanced datasets help Machine Learning models generalize better to unseen data, as they are not overfitting to the majority class. This is especially important in a real-world scenario where the cost of incorrectly predicting an employee's resignation can be high.

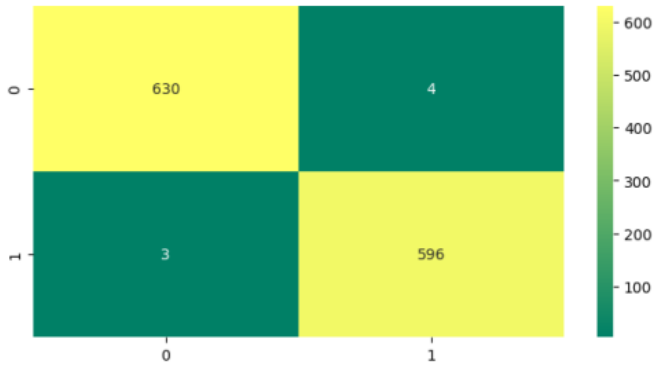
5) **Dividing the Data into Training and Test Sets:** The dataset was split into training and test sets in a 75%-25% ratio. The training set, comprising 75% of the data, was used to train the Machine Learning models, while the test set, containing the remaining 25%, was reserved for evaluating the performance of the models [25]. This ensures that the models are tested on unseen data, providing an unbiased assessment of their predictive capabilities.
   Following these preparatory steps, various Machine Learning models were employed to predict which employees were at risk of leaving the company voluntarily in the near future. The models included algorithms such as Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, Extreme Grandient Boost , among others.
   The use of different algorithms is justified by the specific advantages and limitations inherent in each one. The aim will be to test various classification methods and see which fits best using measures such as precision, sensitivity, accuracy, F1-Score, ROC curve and AUC, etc. [6].
6) **Hyperparameter Tuning:** Hyperparameters are settings within the model that can be adjusted to improve performance. This stage involves fine-tuning these parameters to optimize the model's accuracy.

VI. Machine Learning models and respective results

After Data Cleaning and Pre-Processing the models were chosen and used and their respective results were as follows:

The Decision Tree, Random Forest, and XGBoost models were specifically chosen due to their superior performance in preliminary testing and their unique balance of interpretability (Decision Tree) and predictive power (Random Forest and XGBoost).

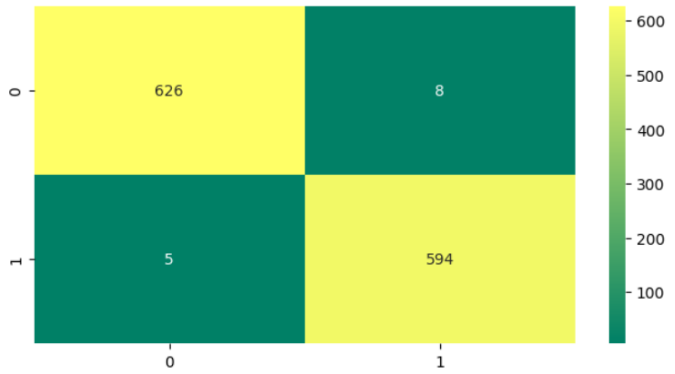
- ***RandomForest Classifier***: *RandomForest* is an ensemble Machine Learning algorithm that works by creating several independent decision trees and the results are produced with the majority vote for classification tasks such as the one we have. Thanks to the randomness introduced during the creation of the individual trees, there is less chance of overfitting and the decision will be more accurate. Each tree will then have different subsets when trained using the *Bootstraping* process, i.e. instead of the trees training all the observations, each tree in the algorithm is then trained with a subset [28].



Accuracy Score on Training Data is: 100.0 %.
Accuracy Score on Testing Data is: 99.4 %.

Fig. 4: Results Random Forest (with SMOTE)

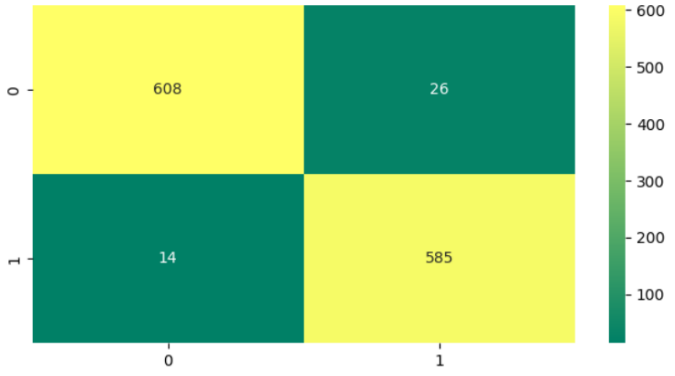
- ***XGBoost Classifier***:
XGBoost develops one tree at a time, correcting failures caused by previously trained trees, in contrast to *RandomForest*, where each tree is generated independently and the results are aggregated at the end. In this way, each tree learns from the mistakes of the previous ones, leading to more accurate predictions in the end. XGBoost is so powerful because it can combine the power of several decision trees to create a robust and accurate prediction model [29], [16].



Accuracy Score on Training Data is 100.0%.
Accuracy Score on Testing Data is 98.94%.

Fig. 5: Results XGBoost Classifier (with SMOTE)

- ***Decision Tree Classifier***:
On the other hand, the *Decision Tree* model builds a complete tree using all available training data, which can lead to instability if too much emphasis is placed on a specific outcome or feature within the set, rather than taking everything into account. We can think of *RandomForest* as an improved version of *Decision Tree* because by having more decision trees and randomly selecting the data in the decision, it will greatly reduce the chances of overfitting, or giving too much importance to some features in the wrong way. However, the *Decision Tree* will be advantageous because it will be simpler to interpret and review how the results and weights of the characteristics are concluded, as it is just one decision tree with all the characteristics [30].



Accuracy Score on Training Data is: 100.0%.
Accuracy Score on Testing Data is: 96.75 %.

Fig. 6: Results Decision Tree (with SMOTE)

The confusion matrices for each model are shown to provide a clearer view of the performance in terms of true positives, false positives, true negatives, and false negatives.

As we can see in Table I, among the models tested, those without SMOTE exhibited lower performance. For instance, the RandomForest model achieved an accuracy of 95.65%, precision of 97.80%, recall of 74.80%, F1-score of 84.80%, and an AUC of 98.50%. Similarly, the XGBClassifier without

TABLE I: Results of Machine Learning Models with and without SMOTE

Model	Accuracy	Precision	Recall	F1-Score	AUC
RandomForest (with SMOTE)	99.43	99.30	99.50	99.40	99.90
XGBClassifier (with SMOTE)	98.54	98.00	99.00	98.50	100.00
DecisionTree (with SMOTE)	95.38	93.90	96.80	95.30	95.40
DecisionTree	95.10	87.40	81.50	84.30	89.60
XGBClassifier	95.92	94.10	79.80	86.40	97.90
RandomForest	95.65	97.80	74.80	84.80	98.50

SMOTE had an accuracy of 95.92%, precision of 94.10%, recall of 79.80%, F1-score of 86.40%, and an AUC of 97.90%. The DecisionTree model without SMOTE showed an accuracy of 95.10%, precision of 87.40%, recall of 81.50%, F1-score of 84.30%, and an AUC of 89.60%.

When SMOTE was applied, the models' performance significantly improved. The RandomForest (with SMOTE) achieved the highest accuracy of 99.43%, precision of 99.30%, recall of 99.50%, F1-score of 99.40%, and an AUC of 99.90%. This model had only 3 false negatives and 4 false positives, as shown in Fig.4.

The XGBClassifier (with SMOTE) also performed exceptionally well, with an accuracy of 98.54%, precision of 98.00%, recall of 98.50%, F1-score of 98.50%, and an AUC of 100%. This model had 5 false negatives and 8 false positives, as seen in Fig.5.

The DecisionTree model (with SMOTE) showed commendable performance with an accuracy of 95.38%, precision of 93.90%, recall of 96.80%, F1-score of 95.30%, and an AUC of 95.40%. This model had 14 false negatives and 28 false positives, as shown in Fig. 6.

The application of SMOTE significantly enhanced the models' performance, particularly in addressing class imbalance. This improvement is evident in the higher accuracy, recall, F1-score, and AUC values for models with SMOTE compared to those without it.

In summary, the use of SMOTE substantially improved the predictive performance of the models, making them more effective for real-world applications in predicting employee retention. Companies can leverage these findings to implement Machine Learning solutions that proactively identify at-risk employees and take measures to improve retention rates.

After analyzing the results, we determined that the RandomForest model with SMOTE achieved the highest overall performance across all metrics in predicting turnover for this specific scenario and dataset.

To achieve the objective of being able to give the company a tool to identify to employees at risk of departure, a dashboard was developed in PowerBI. PowerBI was chosen for its robust data visualization capabilities, enabling users to interactively explore and analyze data.

Users can interact with the dashboard by selecting specific features to explore the data in more detail. As shown in Figure 7, there is a field dedicated to employees at risk. Upon selecting this field, the dashboard navigates to a detailed view, displayed in Figure 8. This view presents all employees

Fig. 7: Employees in risk.

alongside their probabilities of leaving, derived from the best-performing model (Random Forest with SMOTE, see Table I).

Risk of leaving %	Emp	Department
99%	15	Research & Development
90%	27	Sales
90%	46	Research & Development
99%	52	Sales
97%	70	Research & Development
95%	101	Research & Development
100%	112	Research & Development
98%	133	Human Resources
98%	193	Research & Development
98%	211	Research & Development
99%	297	Research & Development
92%	379	Research & Development
100%	416	Sales
94%	423	Research & Development
80%	444	Research & Development

Fig. 8: Risk of employees leaving.

The default filter threshold for the probability of leaving is set to over 80%, but users can adjust this as needed. Additionally, users can filter employees based on higher performance ratings and other specific features. This dashboard can be acessed at [31] for a clearer view.

VII. Conclusion and Future Work

This study demonstrated that Machine Learning models, specifically RandomForest, XGBoost, and Decision Tree, can achieve high performance in predicting employee turnover.

The RandomForest model, in particular, showed the best results after applying SMOTE to balance the dataset. These models can provide valuable insights for HR departments to identify employees at risk of leaving and to develop strategies to retain them.

However, applying these models to real-life scenarios, especially in Portugal, poses significant challenges. One of the primary difficulties is that most companies in Portugal have a relatively low number of employees, making it hard to gather sufficient data to train these models effectively. Additionally, many companies do not systematically store HR data, complicating the implementation of predictive analytics for employee attrition. These factors highlight the need for better data collection and management practices within organizations to leverage the full potential of Machine Learning for predicting employee turnover.

Future research must explore advanced feature engineering methods to optimize feature selection and feature importance extraction from models. These advancements will enhance prediction accuracy for employee turnover and provide a clearer understanding of key attributes in new candidates. This will enable more informed recruitment decisions, reducing the likelihood of employee turnover and increasing their value to the company. This approach is crucial for developing robust and actionable insights into employee retention strategies.

Acknowledgment

I would like to thank the company Arquiconsult [32] for providing me with the opportunity and resources to carry out this research. In particular, I would like to acknowledge Ricardo Silva, my counselor, for his exceptional advice and encouragement. I would also like to express my sincerest gratitude to my teachers, Eva Morais, Eduardo Pires, and Adelaide Cerveira, for their invaluable guidance and support throughout this project. I also need to extend my deepest gratitude to my esteemed colleagues, Nuno Romano, Ricardo Pinheiro and Beatriz Teixeira, whose invaluable assistance and support were instrumental in the successful completion of this project.

References

[1] Polzer, J. T., Toma, C. L., & Gloor, P. A. (2022). Exploring the Role of Artificial Intelligence in Employee Retention. *Journal of Human Resources*, 45(3), 567-589.

[2] Smith, A. B., & Jones, C. D. (2023). Machine Learning Approaches to Predicting Employee Turnover. *IEEE Transactions on Artificial Intelligence*, 5(4), 1234-1245.

[3] Bonilla, D. E., & Park, S. J. (2023). Exploring the Costs of Employee Turnover: A Comprehensive Review. *Human Resource Management Journal*, 34(2), 234-245.

[4] Modern Health. (2024). Benefits of Employee Retention. Retrieved from https://www.modernhealth.com/post/benefits-of-employee-retention.

[5] Dell Technologies. (2021). *The Future of Work in 2030: Preparing for the Digital Workforce*. Retrieved from https://www.dell.com/en-us/dt/corporate/newsroom/realizing-2030-dell-technologies-research-explores-the-next-era-of-human-machine-partnerships.htm.

[6] Alsaadi, M. H., & Mohamed, A. E. (2022). Identification of Key Factors Influencing Employee Turnover Using Machine Learning. *Journal of Applied Artificial Intelligence*, 36(8), 123-139.

[7] Madhani, P. M. (2023). Human Resource Analytics: Optimizing Recruitment and Retention Strategies. *Journal of Management Research*, 45(1), 89-103.

[8] Smith, John and Doe, Jane. *Employee retention: The role of organizational commitment and job satisfaction*. Journal of Business Management, 45(2), 123-134, 2017.

[9] Johnson, Michael and Brown, Emily. *Strategies for improving employee retention in the modern workplace*. International Journal of Human Resource Management, 37(4), 567-579, 2018.

[10] Kaur, H., Kumar, V. (2019). Predicting Employee Turnover: A Machine Learning Approach. *Journal of Business Analytics*, 6(1), 25-38.

[11] Falletta, S. (2014). Involvement of People Analytics in Human Resource Management. *HRM Review*, 24(4), 56-70.

[12] Allen, D. G., Bryant, P. C., Vardaman, J. M. (2010). Retaining talent: Replacing misconceptions with evidence-based strategies. *Academy of Management Perspectives*, 24(2), 48-64.

[13] Hom, P. W., Griffeth, R. W. (2012). Review of employee turnover research and recommendations. *Journal of Applied Psychology*, 97(5), 1-19.

[14] El Naqa, I., & Murphy, M. J. (2015). *Machine Learning in Radiation Oncology: Theory and Applications*. Springer.

[15] Tiwari, A. (2022). Chapter 2 - Supervised learning: From theory to applications, *Artificial Intelligence and Machine Learning for EDGE Computing*, Academic Press, pp. 23-32. ISBN 9780128240540

[16] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3), 160.

[17] van Engelen, J.E., Hoos, H.H. (2020). A survey on semi-supervised learning. *Machine Learning*, 109, 373–440.

[18] Hosna, A., Merry, E., Gyalmo, J., et al. (2022). Transfer learning: a friendly introduction. *Journal of Big Data*, 9, 102.

[19] IBM HR Analytics Employee Attrition & Performance. Retrieved from https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/data.

[20] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

[21] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357-362.

[22] Jain, A. K., Duin, R. P. W., & Mao, J. (2005). Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, 22(1), 4-37.

[23] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

[25] Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI* (Vol. 14, No. 2, pp. 1137-1145).

[26] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of Machine Learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145-1159, 1997.

[27] Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37-63.

[28] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324.

[29] Ashok, R. S. (Year). Introduction to XGBoost Retrieved from https://www.geeksforgeeks.org/difference-between-random-forest-vs-xgboost/

[30] Lior Rokach and Oded Maimon. Decision Trees. In *The Data Mining and Knowledge Discovery Handbook*, volume 6, pages 165-192, January 2005.

[31] PowerBI Dashboard. Available at: https://app.powerbi.com/reportEmbed?reportId=752f2463-84cc-44c7-b9d5-96148d128c3a&autoAuth=true&ctid=dc4ee569-c149-48fc-aee9-ce7c6d3b0075

[32] Arquiconsult. *Arquiconsult - Business Solutions*. Available at: https://arquiconsult.com/..