ML MODELS TO ADDRESS ATTRITION

HR Analytics Assignment 2 - Group 28

Group Members:

Rishita Yogesh Kasat - 40408876

Muhammad Muneeb Ullah Ansari - 40426685

Anjana Vinukumar Krishna - 40423135

Ahmed Ali - 40426107

Shayan Qayyum - 40425337

Word Count: 3300

Table of Contents

Section 1: Introduction

Employee attrition, the gradual reduction in an organization's workforce, casts a long shadow on competitiveness. Whether driven by retirement, resignation, termination, or restructuring, this exodus carries significant challenges. Institutional knowledge walks out the door, morale plummets, and workloads swell for those left behind. The ripple effects extend beyond internal disruptions, potentially raising red flags for prospective recruits and damaging an organization's reputation. In 2021 alone, the U.S. Bureau of Labor Statistics reported a staggering 57.3% attrition rate, underscoring the urgency of addressing this issue. (BLS, 2022)

Yet, many organizations still view human resources as a cost center rather than a strategic investment, overlooking the detrimental impact of losing valuable talent. High attrition rates signal employee dissatisfaction and unstable labor forces, both critical roadblocks to long-term growth and competitive advantage. The costs are multifaceted, encompassing not only recruitment and training expenses but also disruptions in productivity, workflow, and overall work environment.

Understanding the underlying factors driving employee departures is crucial. While individual decisions play a role, organizational policies and actions significantly influence them (Dalton & Mesch, 1990). This is where machine learning algorithms emerge as a powerful tool. Their ability to predict employee attrition can inform proactive engagement strategies and data-driven decision-making, ultimately boosting organizational performance.

This study delves into the complex web of factors contributing to employee attrition, including training opportunities, performance appraisal processes, employee attitudes, and delegation practices. We leverage the predictive power of Decision Trees, Random Forests, and Binary Logistic Regression algorithms, known for their improved accuracy, to shed light on the drivers of talent churn. By identifying these factors, organizations can implement targeted interventions, such as tailoring recruitment efforts to attract individuals with higher retention potential.

The following sections delve deeper into this crucial topic: Section 2 explores existing research on employee attrition, Section 3 outlines our methodology and initial findings, Section 4 analyzes the improved model accuracy and its implications, and Section 5 concludes the study with insights and actionable recommendations.

Section 2: Literature Review

Machine learning applications extend across diverse areas like malware detection (Gaurav et al., 2023), phishing website identification (Almomani et al., 2022), and healthcare security (Wassan et al., 2022). However, its potential in predicting employee attrition has garnered significant attention in recent years. Pioneering research in this domain includes Liu's (2014) case study analyzing Chilean labor market data (112 responses) and Nagadevara and Srinivasan's (2007) investigation of demographic influences and absenteeism on attrition, achieving a model accuracy of 79.58% through improved logistic regression (Nagadevara & Srinivasan, 2007). Notably, Rombaut and Guerry (2018) focused solely on work-related factors in their study of employee turnover.

Several studies leverage logistic regression for prediction. Ponnuru et al. (2020) employ it to forecast employee turnover, while Najafi-Zangeneh et al. (2021) utilize it for analyzing attrition reasons, albeit without focusing on model accuracy improvement (Najafi-Zangeneh et al., 2021). Fallucchi et al. (2020) and Qutub et al. (2021) utilize various models but refrain from addressing accuracy enhancement (Fallucchi et al., 2020; Qutub et al., 2021). Bhartiya et al. (2019) adopt a structured approach, employing multiple algorithms (SVM, Decision Tree, KNN, Random Forest, Naive Bayes). Their Random Forest model achieved the highest accuracy (83.3%) (Bhartiya et al., 2019). Similarly, Joseph et al. (2021) attained an 86% accuracy score with their emotional evaluation approach to predict attrition (Joseph et al., 2021).

Decision trees also show promise. Alao and Adeyemo (2013) analyzed data from over 300 Nigerian employees (1978-2006) and achieved a 74% accuracy rate, identifying salary and tenure as key influencers (Alao & Adeyemo, 2013). Pushing the boundaries further, Al-Darraji et al. (2021) achieved a remarkable 94% accuracy using deep learning techniques (Al-Darraji et al., 2021). Lazzari et al. (2022) conducted a large-scale European survey, employing logistic regression and LightGBM to study turnover reasons (Lazzari et al., 2022). Numerous researchers have explored factors influencing and predicting employee attrition, including Subhashini and Gopinath (2020), Vasa and Masrani (2019), and Zhao et al. (2018) (Subhashini & Gopinath, 2020; Vasa & Masrani, 2019; Zhao et al., 2018).

Section 3: Methodology

The dataset exhibited minimal occurrences of missing values, necessitating limited data cleaning. Notably, the primary adjustment involved standardizing department names to ensure consistency, specifically reconciling "HR" with "Human Resources" and unifying "R & D" with "Research & Development." This rectification was prompted by an anomaly detected in the statistics node, manifesting as a bar plot for the department variable with five bars instead of the expected three as per the data dictionary. The erroneous labels were harmonized with the most prevalent responses, namely "Human Resources" and "Research & Development," utilizing the string replacer node.

Furthermore, scrutiny revealed outliers in two columns, namely age and total working years. The identification process involved the deployment of a boxplot node within the workflow. Regarding age, three outliers were observed, surpassing the prescribed range of 18-60. Adhering to legal working and retirement age parameters, as well as the conventional outlier criteria (1stQ – 1.5IQR and 3rdQ + 1.5IQR), values beyond the stipulated range were excluded from the dataset.

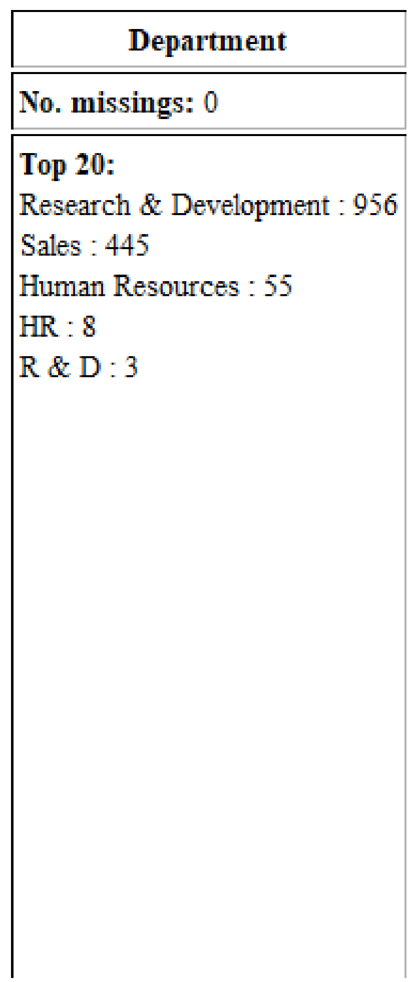

Figure 1: Statistics view of Department

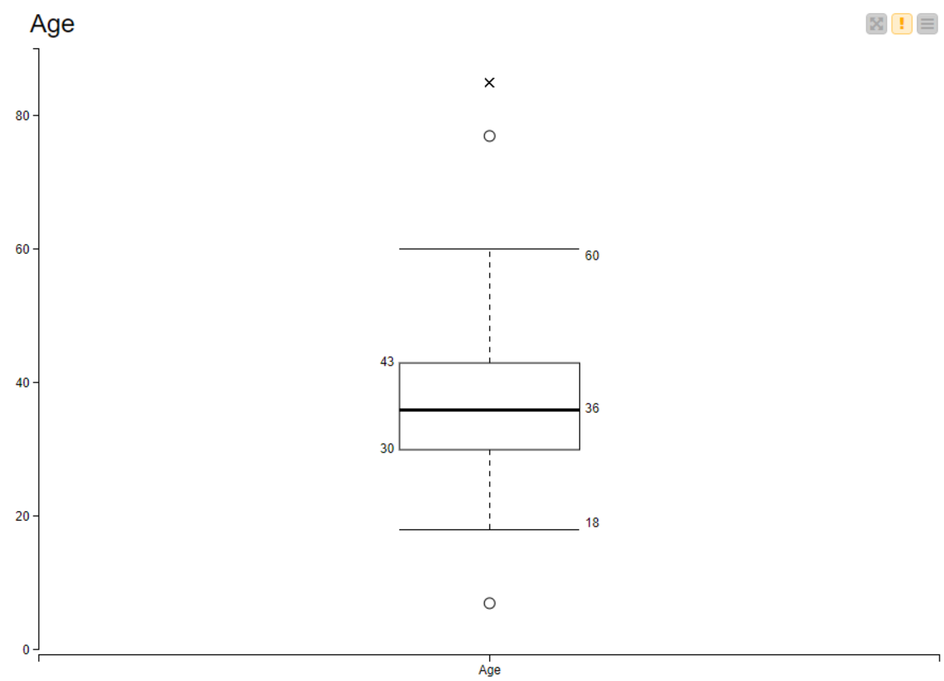

Figure 2: Boxplot of Age

Concerning total working years, eight values exceeded the 3rdQ + 1.5IQR range, with seven of them forming a distinct cluster. Despite their statistical classification as potential outliers, given their concentrated nature, these values were deemed non-anomalous and retained. Notably, the dataset's substantial size, comprising 1467 observations, facilitated the omission of seven clustered values without significant impact. However, an extreme outlier at 94 was removed using the row filter node.

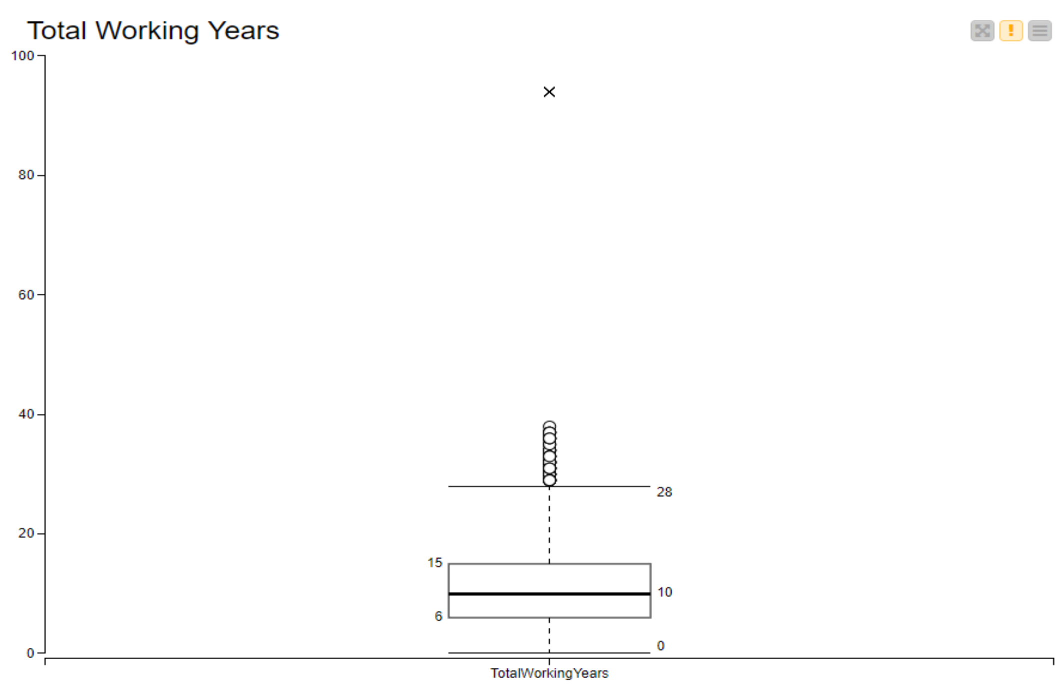

Figure 3: Boxplot of Total Working Years

Upon implementing these adjustments, an Excel writer node was utilized to generate a refined dataset, subsequently employed in the development of prediction models. Initial steps involved data partitioning, effectuated through a partitioning node, with an 80-20 split based on random draws. It is imperative to highlight the utilization of an equal size sampling node, considering the substantial imbalance in attrition labels, with "no" comprising 1184 instances compared to 279 instances of "yes." Additionally, a column filter node was applied to streamline the dataset by excluding variables with potential negligible impact on attrition.

This was evident from the first part of the analytics project where insights were derived relating to trends in attrition.

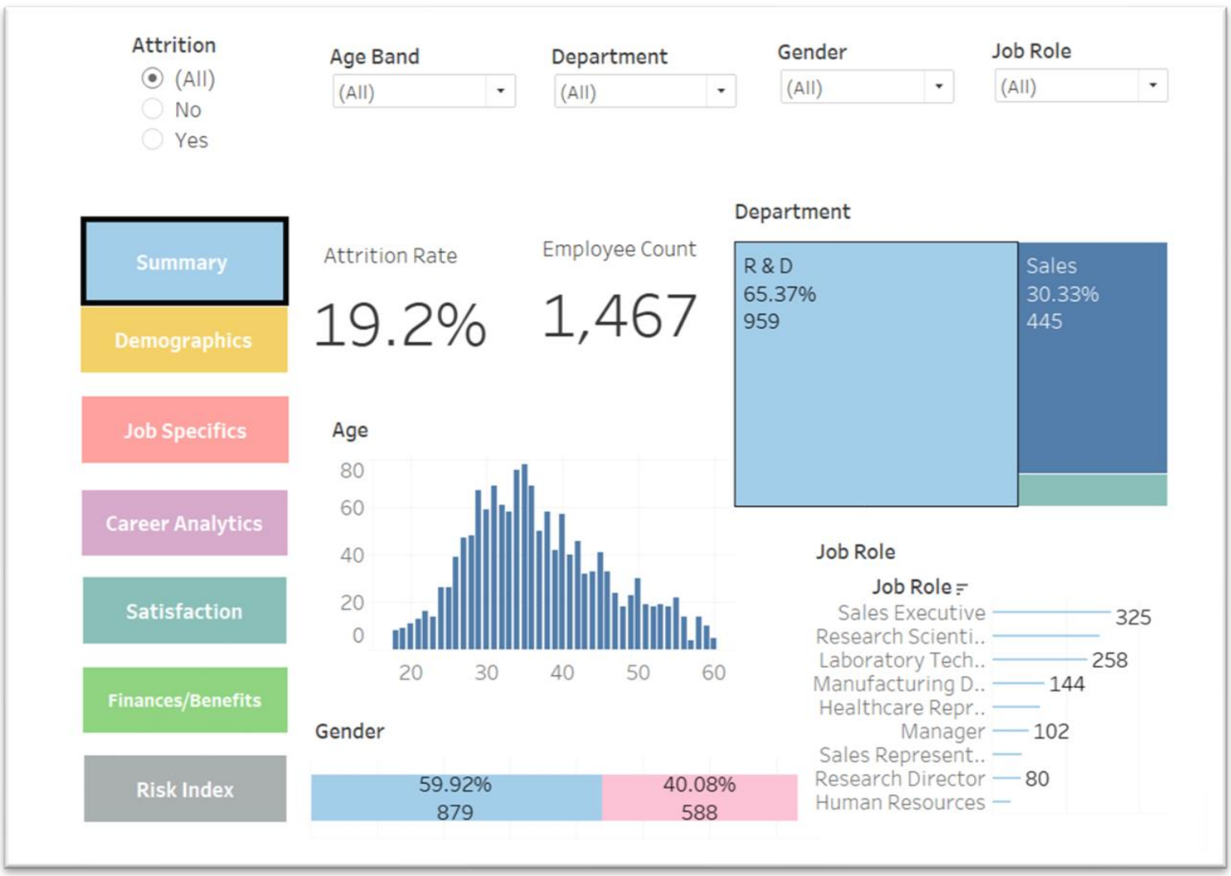

Figure 4: Attrition Dashboard

Removing such variables might cause a higher prediction accuracy of the model. Consequently, certain variables were removed which are as follows.

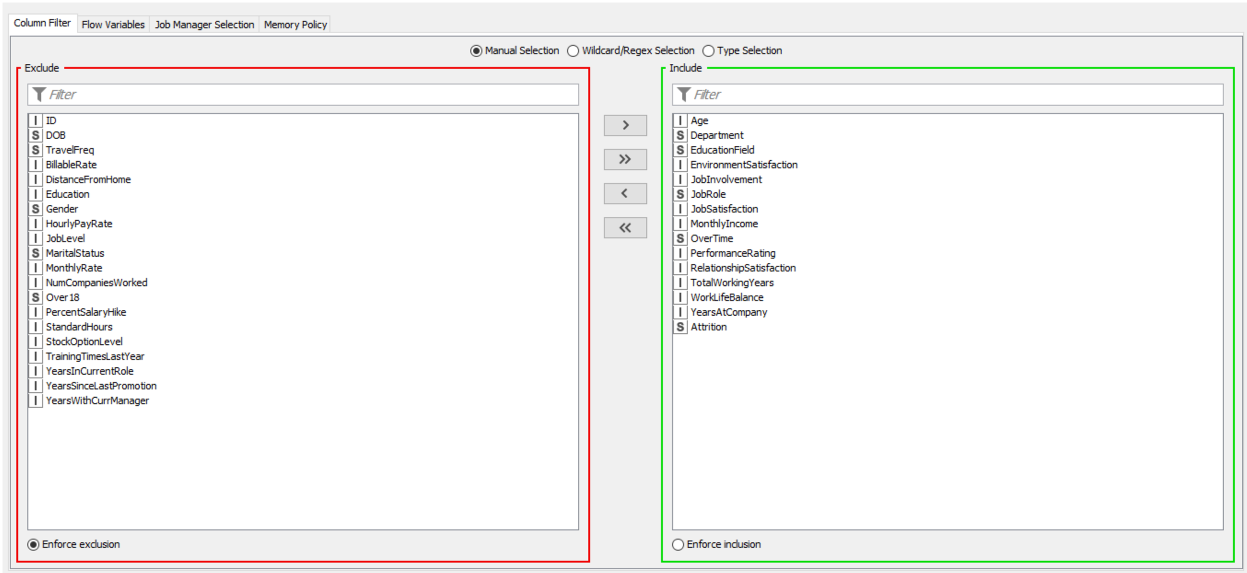

Figure 5: Columns excluded from model

We incorporated pruning methods, encompassing both minimal description length (MDL) and reduced error pruning, within our decision tree model. These methods, classified as post-pruning techniques, involve allowing the tree to grow initially and subsequently pruning nodes individually. Although pre-pruning methods, which prevent excessive branching during the initial growth phase, are alternative options, post-pruning techniques are generally preferred. This preference is rooted in the goal of mitigating overfitting and preventing the development of a biased model.

In the context of Predictive Model Markup Language (PMML) settings, given the absence of missing values in our dataset, the specification of a strategy to handle missing values is unnecessary. However, we must address the "No True Child" problem, which arises when a node is reached where a decision cannot be made for the current input. In such instances, we opt for the pragmatic approach of returning the last prediction, which corresponds to the majority class of the preceding node within the same branch.

Section 4: Model Building

4.1 Decision Tree

After initiating the decision tree learner node, it becomes evident that the initial split is exclusively based on job role, with no further nodes or branching beyond that point. The node purity is weak for each root node, except managing director, health care representative, research director, and manager. However

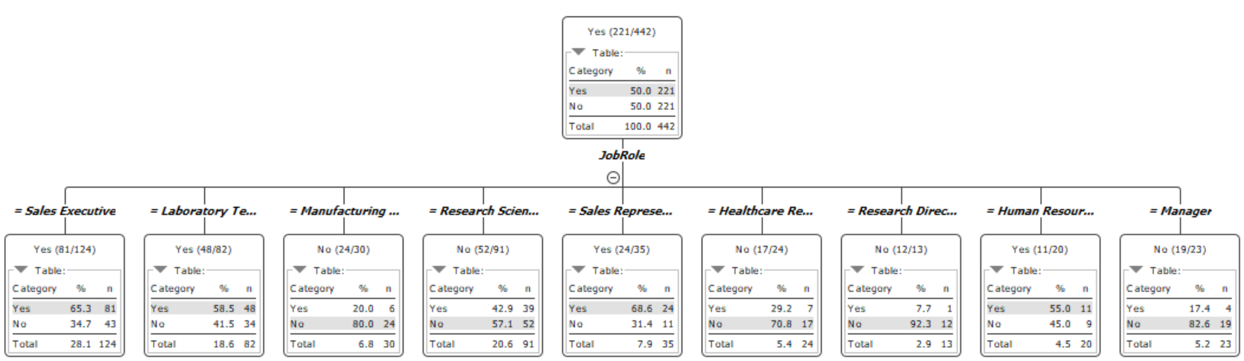

Figure 6: Iteration 1 of Model

basic attrition count analysis by job roles reveals that certain roles already exhibit a lower attrition count i.e. HR has none. Consequently, a high node purity in these instances does not signify statistical significance. For nodes demonstrating poorer node purity, the outcomes are inconclusive and do not provide conclusive indications.

To address this challenge, a potential solution involves activating the binary nominal splits option within the decision tree learner node, leading to splits exclusively in binary variables. To maintain a balanced approach and mitigate the risk of overly shallow trees, we established a constraint by setting the maximum number of root splits to 10. This decision was made along with the existing activation of post-pruning methods. It is noteworthy that opting for a lower number of splits could invoke pre-pruning methods, potentially causing the tree to exhibit elevated variance due to its shallowness. The implementation of these adjustments leads to the generation of the following decision tree:

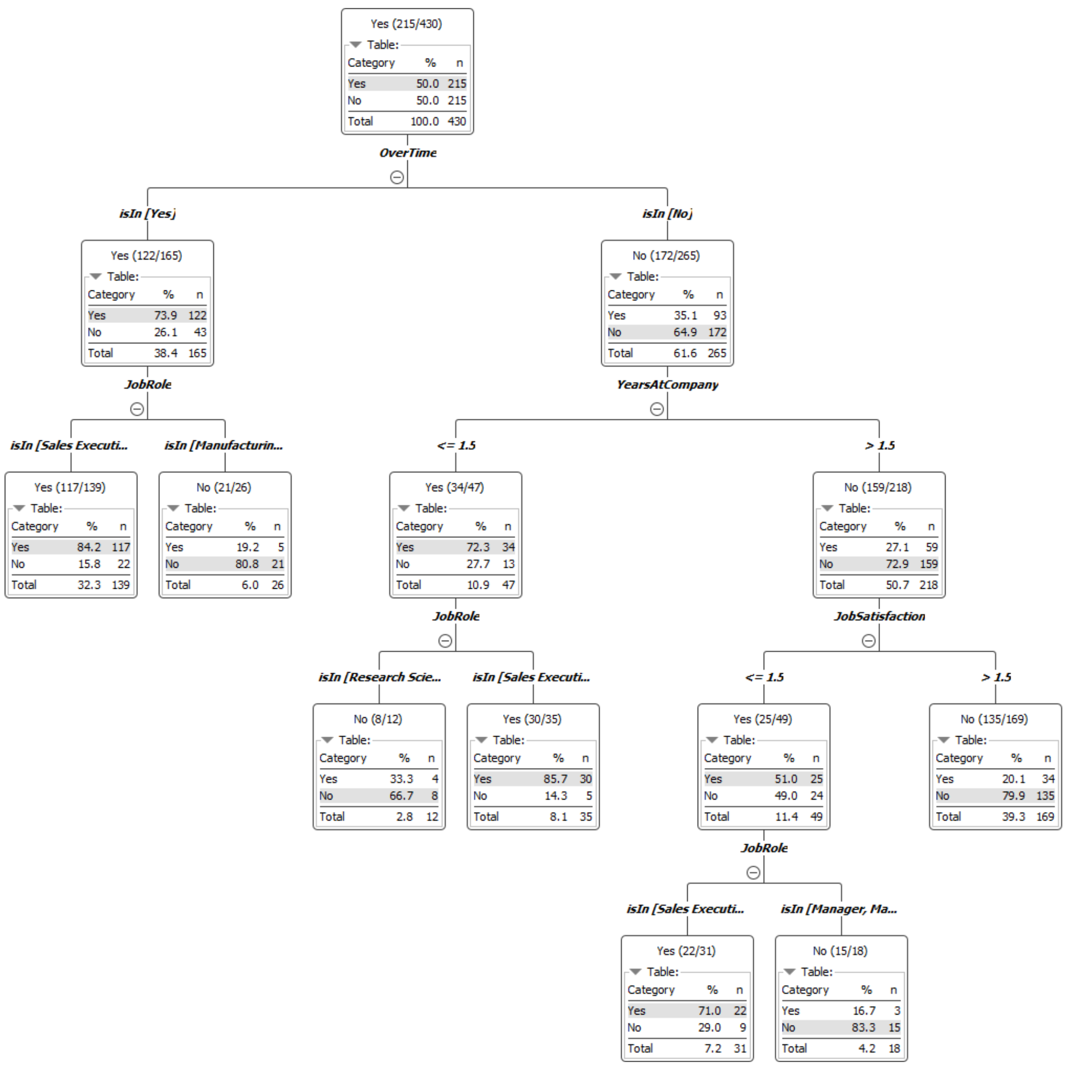

Figure 7: Iteration 2 of Model

This development aligns with the findings from the preceding segment of the project, wherein heightened attrition rates were observed in specific departments and job roles, primarily attributed to unfavorable employee perceptions of working conditions. While the foundational node purity appears commendable, decisions based solely on job role and department fail to yield comprehensive insights. Such splits are insufficient for encapsulating the intricate factors contributing to attrition, limiting the attribution of attrition values to job roles and departments alone. Consequently, the current manifestation of the model exhibits a pronounced bias.

To remedy this bias, a more refined approach involves additional filtering of columns from the dataset. Specifically, the columns "Department" and "JobRole" were excluded, resulting in the generation of the subsequent model. This refined model aims to offer a more nuanced understanding of the factors influencing attrition by eliminating the confounding effects associated with job roles and departments.

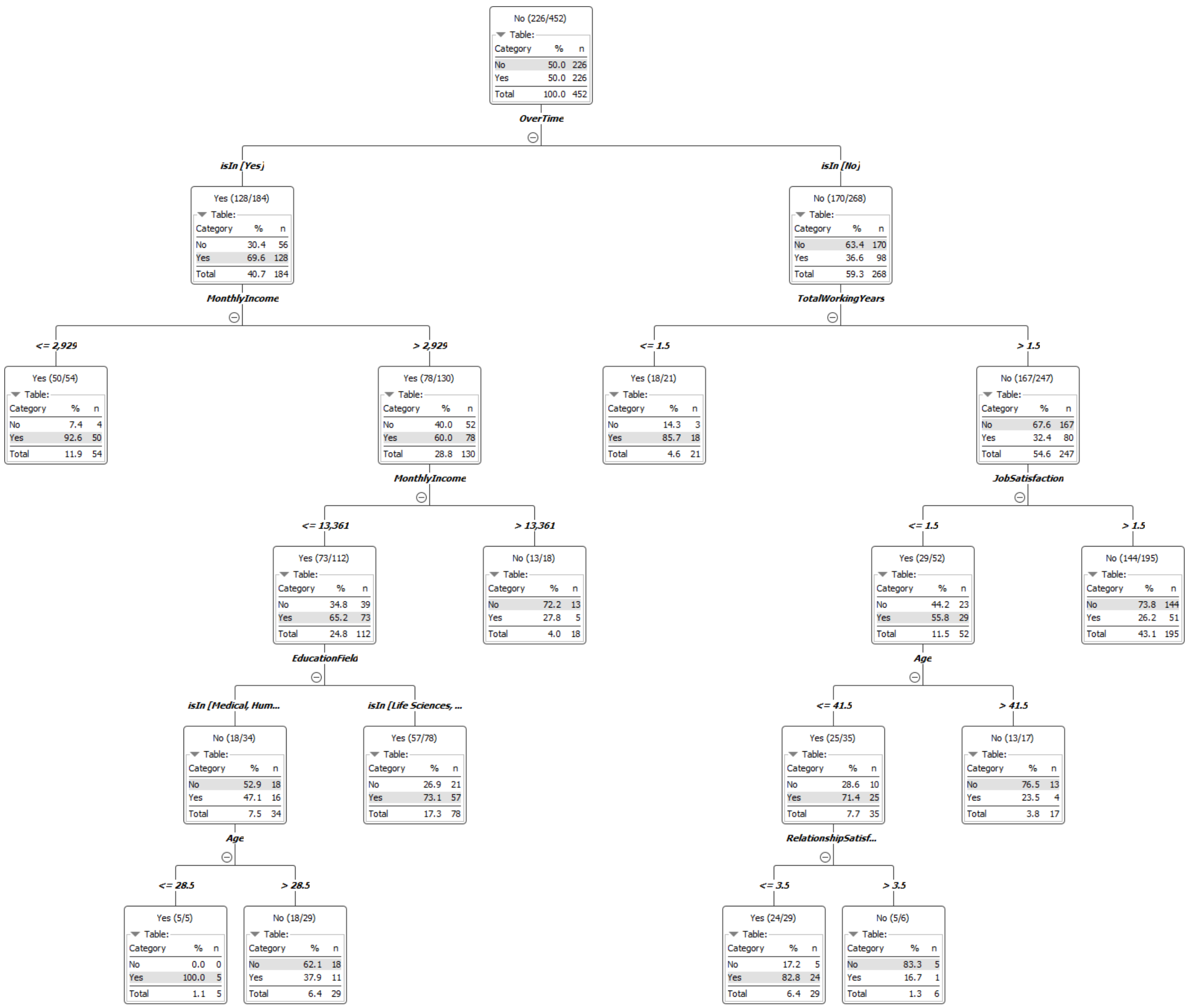

Figure 8: Final Iteration of Model

The current configuration of nodes provides a comprehensive understanding of factors potentially associated with attrition. While the educational field, at first glance, might not appear to yield conclusive insights, its significance becomes apparent in the context of the company being a pharmaceutical entity. The educational field of an employee serves as an indicative measure of their compatibility with the organization, influencing their likelihood of intending to depart.

Within the ensemble of 10 root nodes, notable observations include 2 nodes with purities surpassing 90%, 3 nodes exceeding 80% purity, 4 nodes surpassing 70% purity, and 1 node with purity surpassing 60%.

These purity levels collectively underscore the robust construction of the model, suggesting its efficacy. Consequently, the model is deemed sufficiently well-constructed, warranting further evaluation.

In the configuration of the decision tree predictor node, an additional step involves appending columns featuring normalized class distributions. This augmentation facilitates the depiction of prediction probabilities. Subsequently, this information proves instrumental in the creation of an ROC (Receiver Operating Characteristic) curve during the model evaluation process.

4.2 Accuracy Testing

Subsequently, we will employ the scorer node to assess the accuracy of the model. The ensuing presentation comprises the confusion matrix and associated accuracy statistics for the model.

Attrition \ ...	Yes	No
Yes	41	13
No	95	144

Correct classified: 185 | Wrong classified: 108

Accuracy: 63.14% | Error: 36.86%

Cohen's kappa (κ): 0.228%

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
Yes	41	95	144	13	0.759	0.301	0.759	0.603	0.432	?	?
No	144	13	41	95	0.603	0.917	0.603	0.759	0.727	?	?
Overall	?	?	?	?	?	?	?	?	?	0.631	0.228

Figure 9: Accuracy Test for Decision Tree

The overall accuracy of our model stands at 0.631, a figure that may initially seem modest. This implies that, of all predictions made, our model achieved accuracy in 63% of instances. However, a more nuanced understanding can be gleaned by scrutinizing the true positives and false negatives.

Examining the positive accuracy metrics, the Recall (Sensitivity or True Positive Rate) is 0.759, Precision is 0.301, and the Type II error (False Negatives) amounts to 13 observations. These metrics indicate that our model correctly predicted approximately 76% of instances where employees left, but only 30% of the times when it predicted an employee's departure. This suggests that our model frequently predicts employee departures, but with a relatively lower accuracy rate. Shifting focus to negative accuracy, the Specificity (correctly identified instances of the negative class) is 0.603, signifying that our model made accurate predictions in 60% of instances where employees did not leave.

In conclusion, our model exhibits greater proficiency in predicting employee departures than non-departures, aligning with the primary objective of the model. However, it is noteworthy that a substantial portion of positive predictions are incorrect. From a business perspective, it is advantageous to anticipate potential departures, even if some predictions prove to be inaccurate.

Analyzing Cohen's Kappa, a value of 0.228 indicates fair agreement between predicted and actual classifications beyond chance, albeit falling short of an ideal scenario. This suggests there is still room for improvement in the model, potentially through exploring alternative models beyond Decision Trees.

Following is the ROC curve for the model.

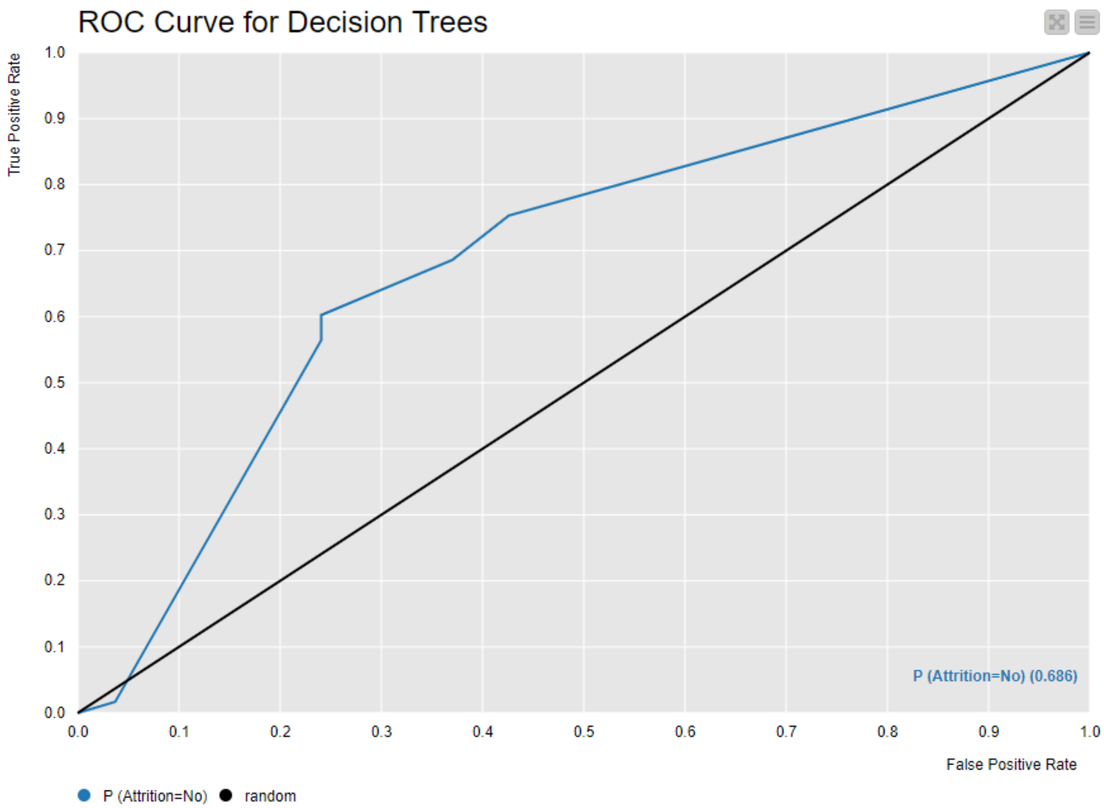

Figure 10: ROC Curve for Decision Tree

The ROC curve serves as an evaluative tool by utilizing the false positive rate and true positive rate, which are derived from applying diverse threshold values to the positive class probabilities predicted by the model. This graphical representation elucidates the model's trade-off between the false positive rate and true positive rate. Ideally, the curve should manifest a True Positive Rate of 1.0 and a False Positive Rate of 0.0. In our model, the curve indicates that, at a 0.8 True Positive Rate, the corresponding False Positive Rate stands at 0.5. This alignment suggests a decent performance with identifiable room for enhancement. Moreover, this finding is consistent with the observed tendency of our model to more frequently predict "Yes" for attrition rather than "No."

Given the identified scope for improvement, our next step involves the creation of alternative models utilizing different algorithms while maintaining constant parameters in the analysis. This strategic approach aims to explore alternative methodologies and potentially enhance the predictive capabilities of the model.

4.3 Other ML Techniques

Following are the results of a Random Forest Model.

Attrition \ ...	Yes	No
Yes	40	14
No	67	172

Correct classified: 212 Wrong classified: 81

Accuracy: 72.355% Error: 27.645%

Cohen's kappa (κ): 0.334%

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
Yes	40	67	172	14	0.741	0.374	0.741	0.72	0.497	?	?
No	172	14	40	67	0.72	0.925	0.72	0.741	0.809	?	?
Overall	?	?	?	?	?	?	?	?	?	0.724	0.334

Figure 11: Accuracy Test for Random Forest

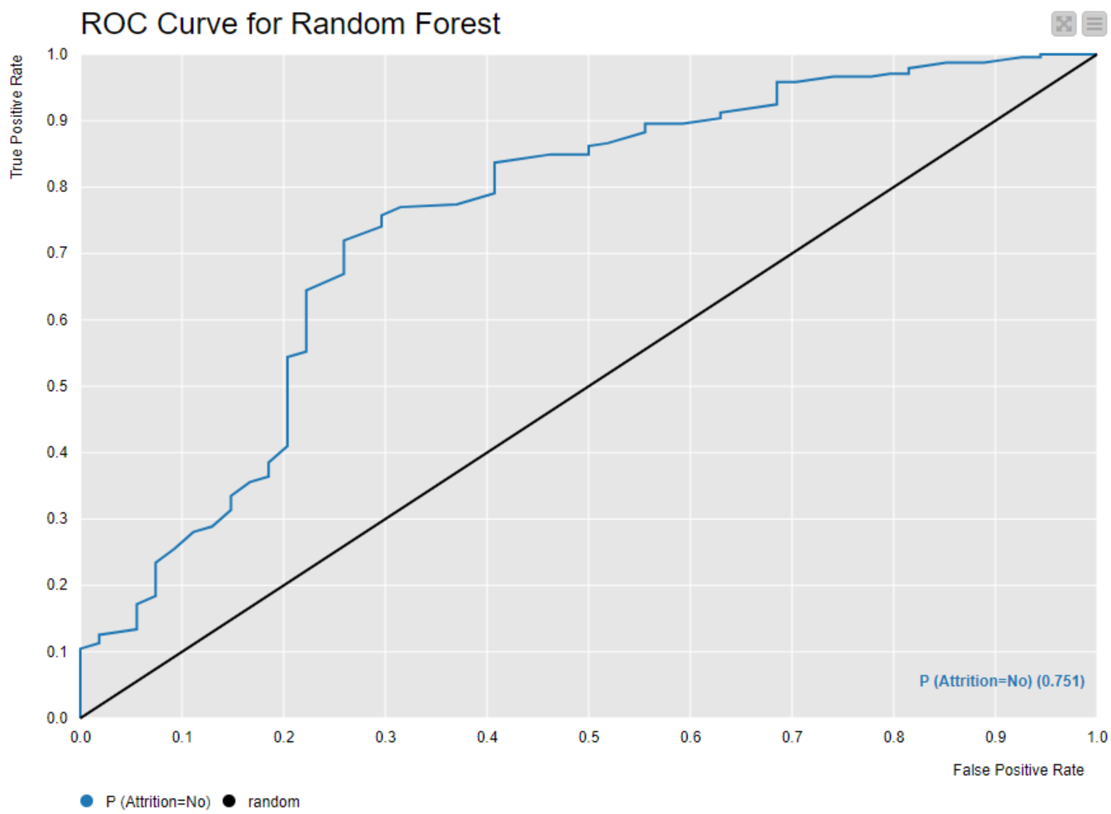

Figure 12: ROC Curve for Random Forest

The following are the results for Tree Ensemble.

Attrition \ ...	Yes	No
Yes	39	15
No	66	173

Correct classified: 212 Wrong classified: 81

Accuracy: 72.355% Error: 27.645%

Cohen's kappa (κ): 0.327%

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
Yes	39	66	173	15	0.722	0.371	0.722	0.724	0.491	?	?
No	173	15	39	66	0.724	0.92	0.724	0.722	0.81	?	?
Overall	?	?	?	?	?	?	?	?	?	0.724	0.327

Figure 13: Accuracy Testing for Tree Ensemble

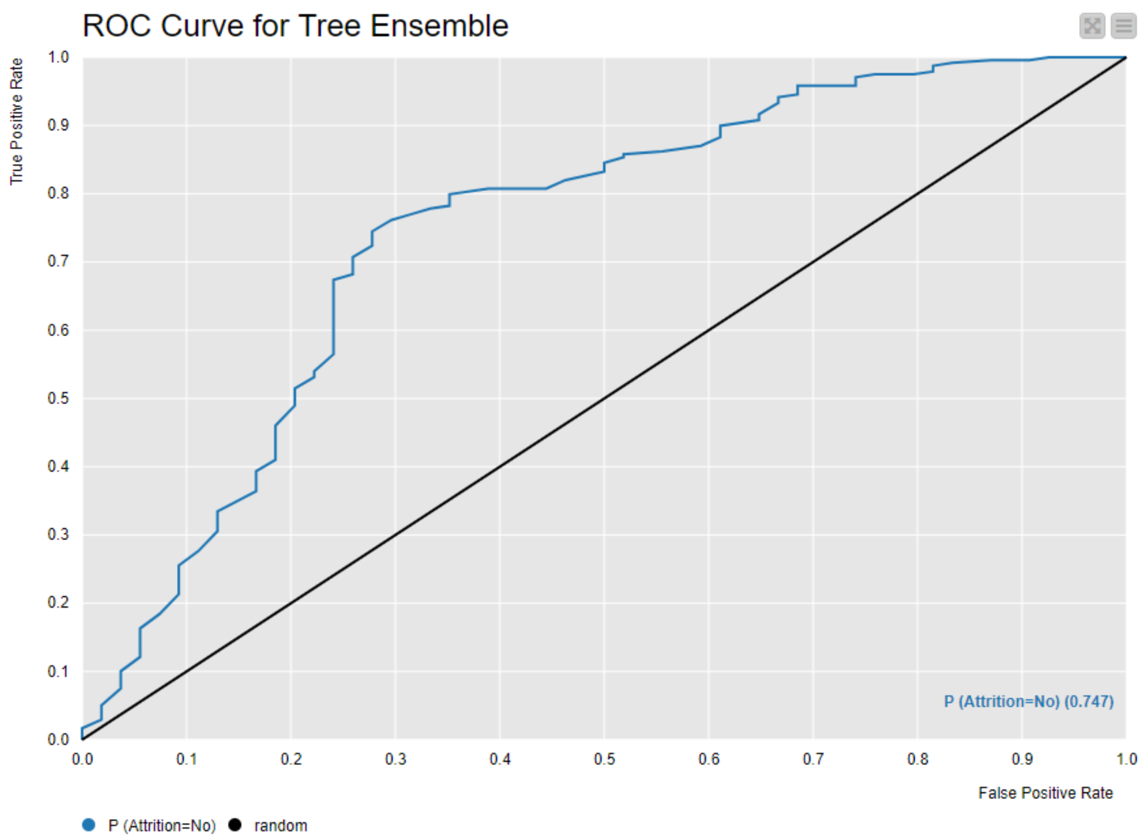

Figure 14: ROC Curve for Tree Ensemble

The following are the results for Gradient Boosted Trees.

Row ID	Yes	No
Yes	38	16
No	75	164

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
Yes	38	75	164	16	0.704	0.336	0.704	0.686	0.455	?	?
No	164	16	38	75	0.686	0.911	0.686	0.704	0.783	?	?
Overall	?	?	?	?	?	?	?	?	?	0.689	0.274

Figure 15: Accuracy Statistics for Gradient Boosted Trees

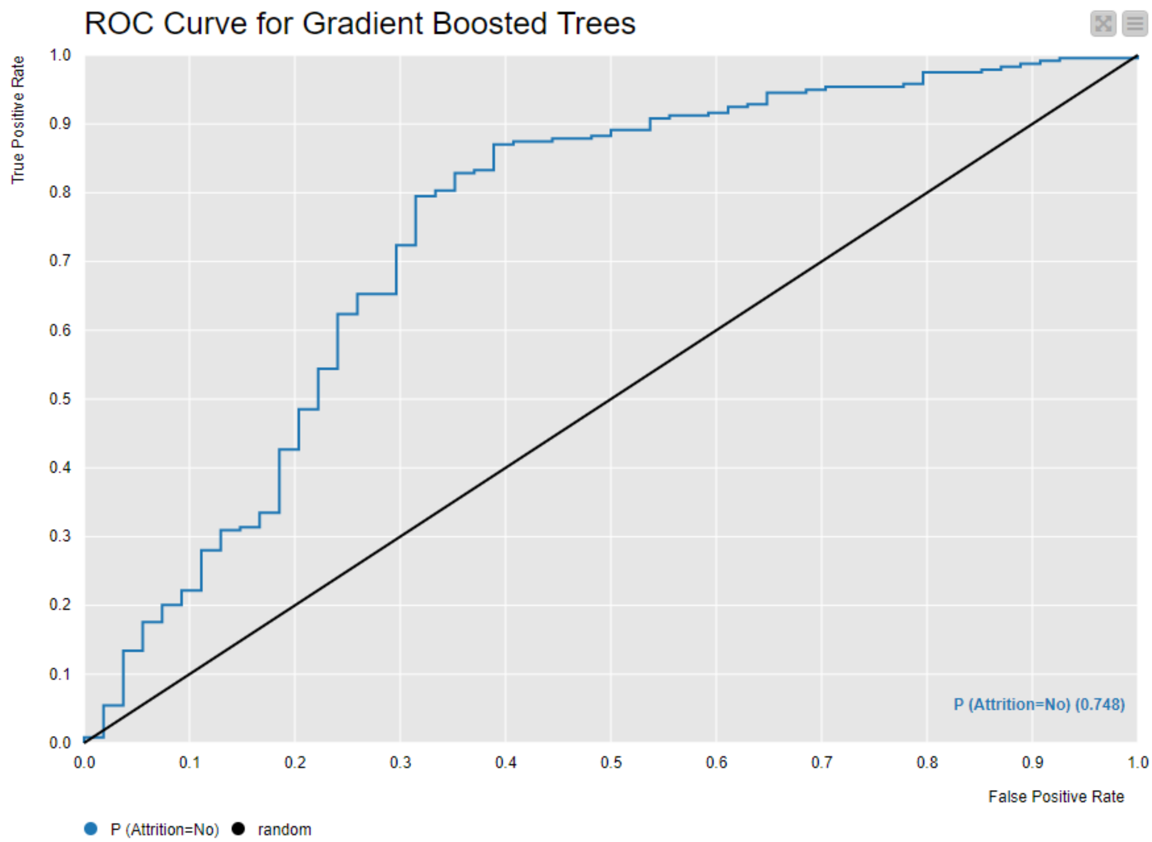

Figure 16: ROC Curve for Gradient Boosted Trees

Section 5: Findings and Discussion

The following table combines all the important statistics for each of the four models to compare.

	Recall	Precision	Specificity	Cohen's Kappa	True Positive Rate at 0.5 False Positive Rate
Decision Trees	0.759	0.301	0.603	0.228	0.8
Random Forest	0.741	0.374	0.72	0.334	0.85
Tree Ensemble	0.722	0.371	0.724	0.327	0.84
Gradient Boosted Trees	0.704	0.336	0.686	0.274	0.9

Figure 17: Comparative Figures for all models

The Decision Trees model emerges as the most proficient in predicting attrition, attributed to its highest recall, demonstrating its capability to identify a substantial portion of actual positive instances. However, the model's efficacy is tempered by significant drawbacks, including the lowest precision and a notable trade-off between true positive and false positive rates. This trade-off signifies a discernible tendency to misclassify many observations as "Yes" for attrition, compromising the model's precision. Additionally, the Decision Trees model exhibits the lowest specificity, indicating a limited ability to accurately identify instances of employees who do not leave. Moreover, the Cohen's Kappa metric reveals the model's inadequacy in achieving substantial agreement between predicted and actual classifications beyond what would be expected by chance.

In contrast, the Random Forest model emerges as a compelling alternative with nuanced strengths. While it falls slightly short of Decision Trees in terms of recall, it excels in precision, specificity, and Cohen's Kappa. The Random Forest model not only maintains high precision, signifying a lower rate of false positives, but also demonstrates the highest specificity, indicating a superior ability to accurately identify instances of employees who do not leave. This nuanced trade-off positions Random Forest as a favorable choice, particularly when considering its balanced performance in minimizing misclassifications.

Furthermore, the Random Forest model exhibits the second-highest true positive rate for a 0.5 false positive rate, underscoring its ability to achieve a favorable balance between true positives and false positives. While Random Forest may not surpass Decision Trees in predicting employee departures, its notable advantage lies in mitigating the misclassification of observations as "Yes" for attrition. This feature enhances the model's overall predictive accuracy, particularly in the context of predicting "No" for attrition.

The enhanced reliability of Random Forest results from a higher agreement between predicted and actual classifications, as denoted by Cohen's Kappa. This metric underscores the model's capacity to achieve a level of agreement beyond what would be anticipated by random chance. In summary, the Random Forest model emerges as a promising alternative, offering a balanced and reliable predictive performance across various metrics.

Insights

In an organizational context, the insights derived from the analysis of machine learning (ML) techniques, specifically Decision Trees and Random Forest, provide invaluable guidance for employers seeking to understand and address employee attrition. (Joseph et.al., 2021) The proficiency of the Decision Trees model lies in its capability to identify a substantial portion of employees at risk of attrition, offering a targeted approach to workforce management. However, caution is advised regarding precision, as the model exhibits a trade-off with false positives, indicating occasional misclassifications. Employers are encouraged to investigate further before making decisive actions based solely on Decision Trees predictions. (LI et al., 2009) To refine intervention strategies, employers can leverage additional employee insights, such as feedback, performance reviews, and engagement surveys, in conjunction with Decision Trees predictions. This holistic approach enhances the accuracy of identifying underlying issues contributing to attrition, facilitating the development of tailored intervention plans for specific departments or job roles flagged by the model.

Considering the limitations of Decision Trees, employers are urged to explore Random Forest as an alternative ML model. The Random Forest model, with its strengths in precision, specificity, and overall reliability, provides a more balanced approach to attrition prediction, mitigating the misclassification of employees likely to leave. (EDUCBA, 2023) Incorporating Random Forest into the HR analytics toolkit contributes to a more nuanced understanding of attrition factors. For strategic workforce planning, employers should focus on high-attrition roles identified by the models, allocating resources strategically to areas prone to attrition. Balancing predictive accuracy by evaluating the trade-off between true positive and false positive rates is crucial to maintaining a balanced and efficient workforce management strategy.

Acknowledging the room for improvement in both Decision Trees and Random Forest models, employers are encouraged to invest in continuous model refinement and explore alternative algorithms to achieve higher predictive accuracy. Establishing a feedback loop that incorporates real-world outcomes into the model training process ensures the adaptability of models to the dynamic nature of employee behaviors and organizational changes. To enhance employee engagement, employers can use ML insights as a starting point for addressing underlying issues contributing to attrition. Implementing changes in policies, work culture, or communication strategies based on identified factors can foster a positive work environment and support proactive retention measures. (Zutavern, 2022)

When evaluating the success of attrition prediction models, employers are advised to consider multiple metrics beyond accuracy, including precision, recall, specificity, and Cohen's Kappa. Striving for cohesion between model predictions and real-world outcomes ensures a comprehensive assessment of the models' performance. By incorporating these insights into organizational strategies, employers can harness the power of ML techniques to proactively address employee attrition, foster a positive work environment, and optimize workforce management practices. (Alsubaie et.al., 2024)

Conclusion

The report detailed the data preprocessing steps, placing emphasis on outlier handling and the standardization of department names. Decision tree models, enriched with post-pruning techniques, were deployed, and alternative models, including Random Forests and Gradient Boosted Trees, were explored. While the decision tree model showcased commendable recall, its vulnerability to a trade-off between true positives and false positives unveiled certain limitations. In stark contrast, the Random Forest model emerged as a promising alternative, exhibiting a balanced performance in precision, specificity, and Cohen's Kappa. Its capacity to mitigate misclassifications, especially in the prediction of "No" for attrition, positions it as a robust choice for organizations aspiring to implement an effective attrition prediction model.

In summation, this study underscores the strategic significance of embracing machine learning models for attrition prediction. While the decision tree model demonstrated proficiency, the Random Forest model surfaced as a balanced and reliable alternative, furnishing actionable insights for organizations to fortify their talent retention strategies. As attrition persists as a formidable challenge to workforce stability, the adoption of data-driven methodologies becomes imperative for organizations aiming to sustain a competitive advantage in the dynamic landscape of talent management.

Appendix:

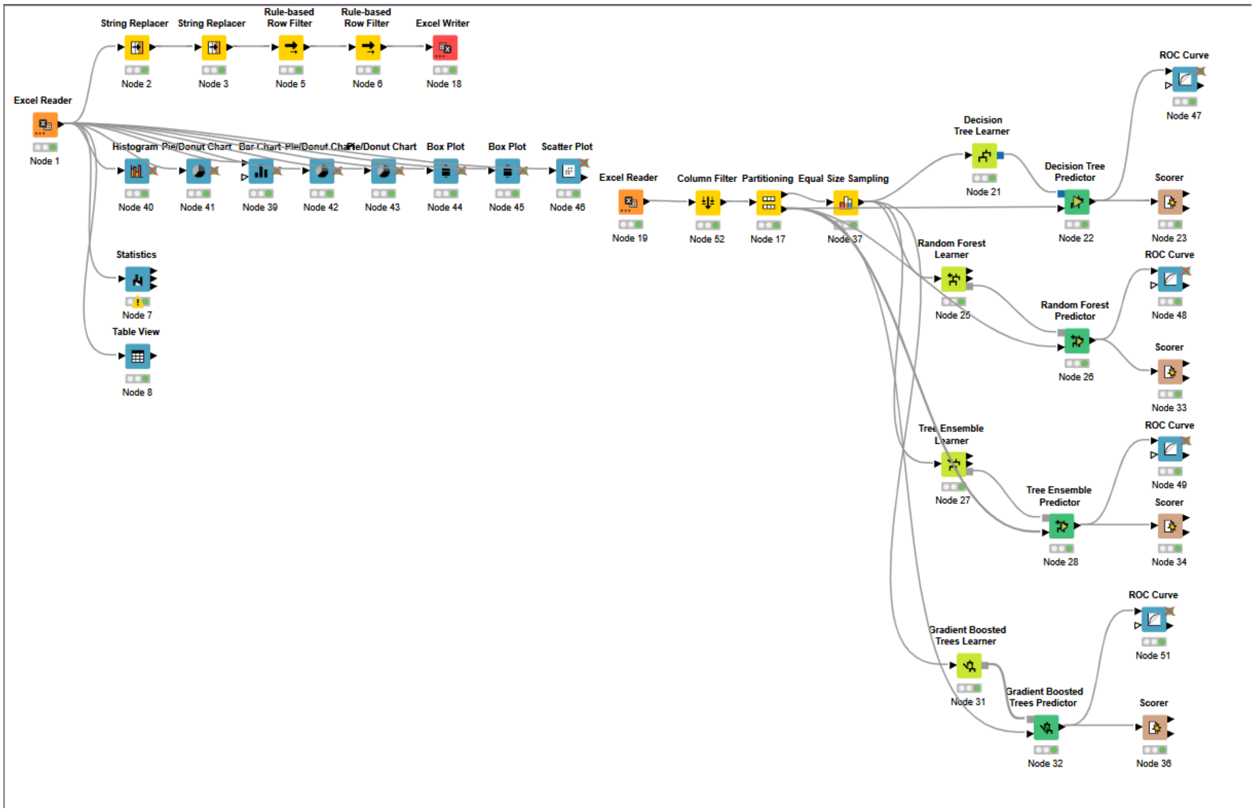

Figure 18: KNIME Workflow

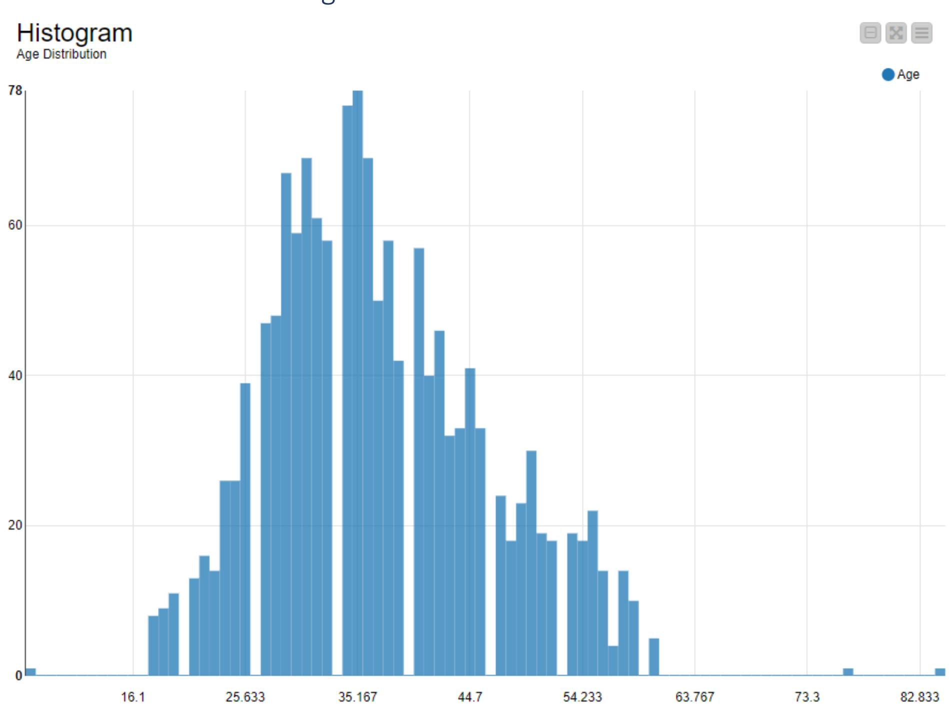

Figure 19: Histogram for Age

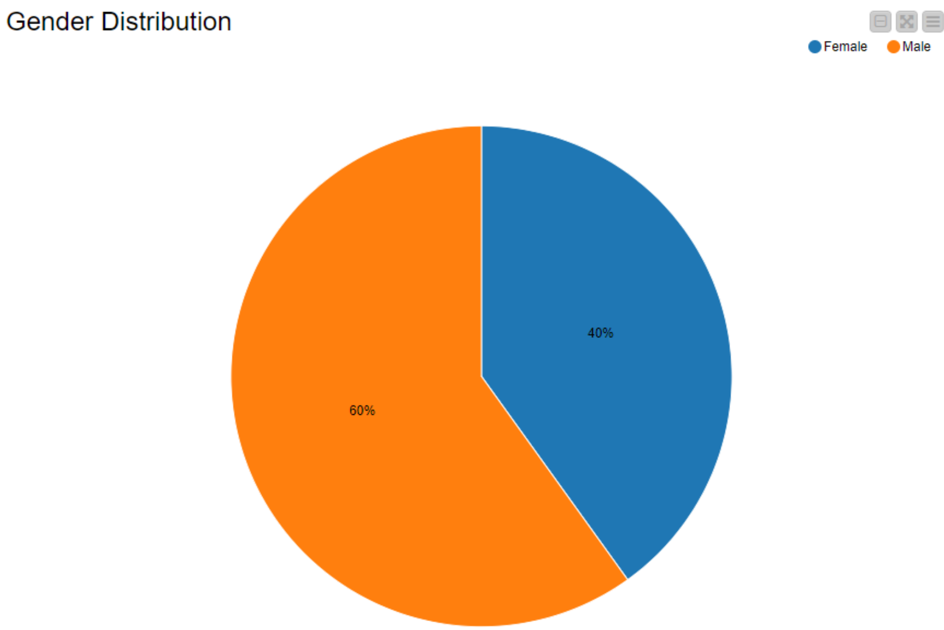

Figure 20: Pie Chart of Gender distribution

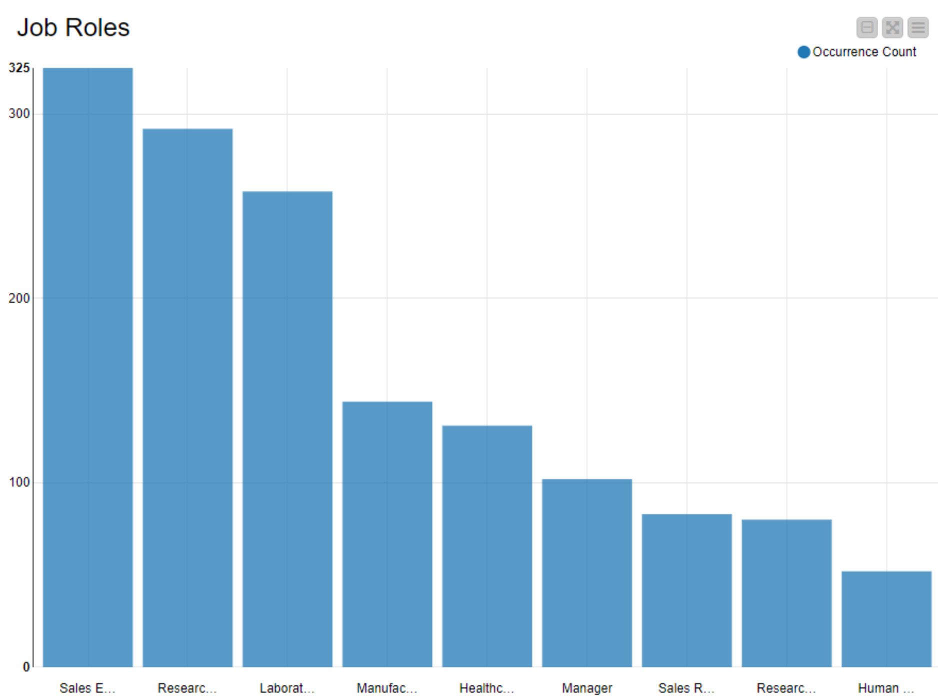

Figure 21: Bar chart for Job Roles

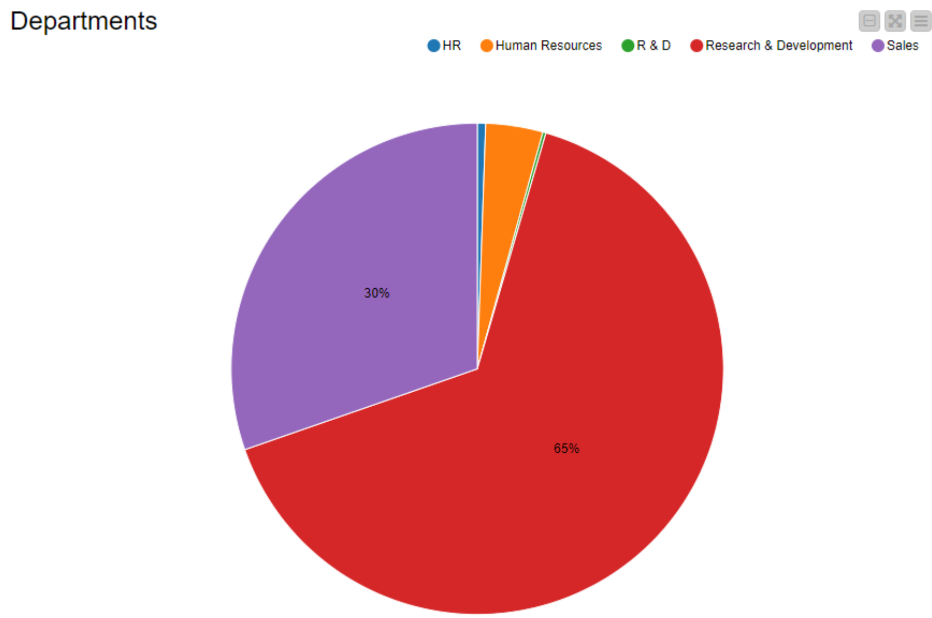

Figure 22: Pie chart for Department distribution

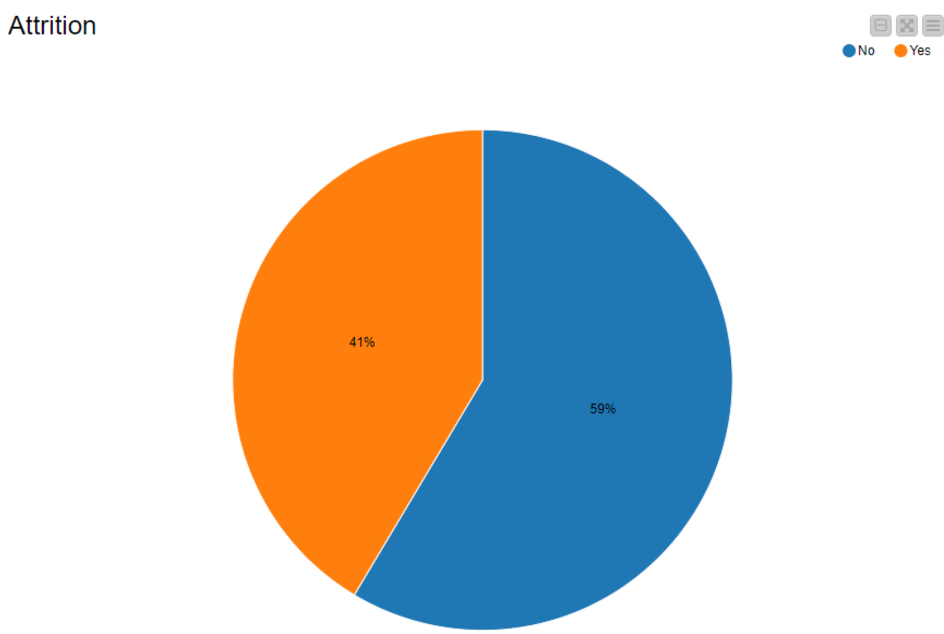

Figure 23: Pie chart for Attrition distribution

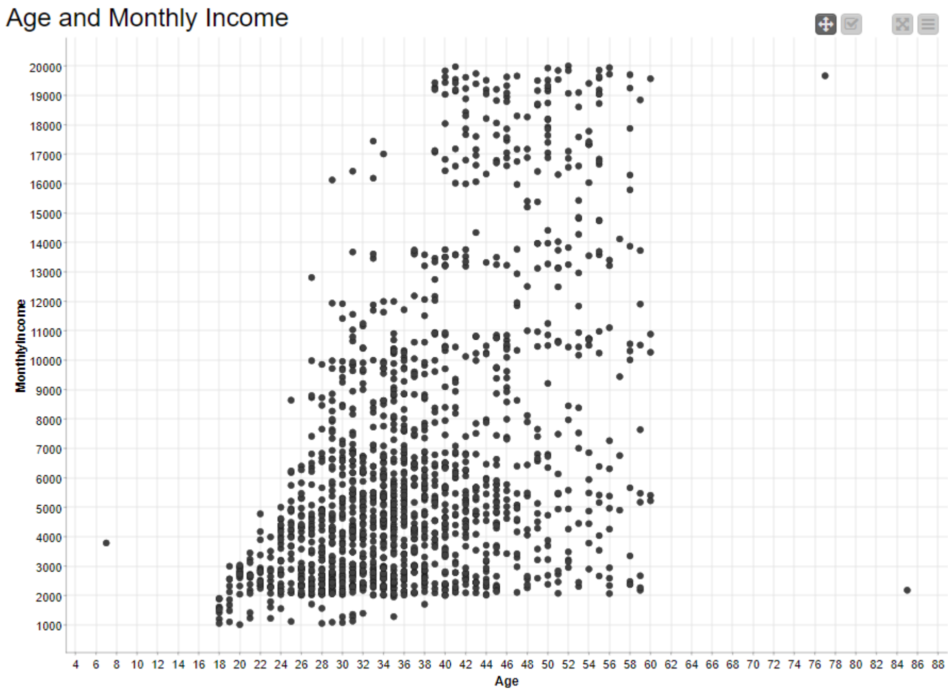

Figure 24: Scatter plot for Age and Monthly Income

# Declaration of Equal Contribution in University Group Project

We, the undersigned members of **Queen's University Belfast**, Queen's Business School, enrolled in the **Masters in Business Analytics (T)** program, working on the group project titled "**ML Models to Address Attrition**" by *Group 28* for the course "**MGT7222: Human Resources Analytics**", hereby declare that all members of the group have contributed equally to the successful completion of the project. Each member has actively participated in the planning, research, analysis, and presentation phases, ensuring a fair distribution of responsibilities and work.

Group Members:

Name	Student Number
Muhammad Muneeb Ullah Ansari	40426685
Shayan Qayyum	40425337
Anjana Vinukumar Krishna	40423135
Rishita Yogesh Kast	40408876
Ahmed Ali	40426107

We hereby acknowledge and affirm that:

- The workload was distributed equitably among all members.
- Each member actively participated in group discussions and decision-making processes.
- All members contributed to the creation and editing of project documents, reports, and presentations.
- We have communicated effectively throughout the project, sharing information and updates regularly.

In witness whereof, we have affixed our signatures below:

Signature	Name	Date
	Muhammad Muneeb Ullah Ansari	7th January, 2023
Shayan Qayyum	Shayan Qayyum	7th January, 2023
	Anjana Vinukumar Krishna	7th January, 2023
	Rishita Yogesh Kasat	7th January, 2023
	Ahmed Ali	7th January, 2023

This document serves as a formal acknowledgment of our shared commitment and contributions to the group project.

Queen's University Belfast,
Queen's Business School,
7th January, 2023.

Figure 25: Declaration

References

1. Al-Darraji, S. *et al.* (2021) 'Employee attrition prediction using deep neural networks,' *Computers*, 10(11), p. 141.
2. Alao, D., & Adeyemo, A. (2013) *ANALYZING EMPLOYEE ATTRITION USING DECISION TREE ALGORITHMS*. https://iiste.org/Journals/index.php/CIS/article/view/10148.
3. Almomani, A. *et al.* (2022) 'Phishing website detection with semantic features based on machine learning classifiers,' *International Journal on Semantic Web and Information Systems*, 18(1), pp. 1–24.
4. Alsubaie, F. and Aldoukhi, M. (2024) ‘Using machine learning algorithms with improved accuracy to analyze and predict employee attrition’, Decision Science Letters, 13(1), pp. 1–18.
5. Bhartiya, N., Jannu, S., Shukla, P., & Chapaneri, R. (2019). "Employee Attrition Prediction Using Classification Models." *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, Bombay, India, 2019, pp. 1-6.
6. Branham, L. (2005) 'Planning to become an employer of choice,' *Journal of Organizational Excellence*, 24(3), pp. 57–68.
7. Breiman, L. (1996). "Out-of-Bag Estimation." Statistics Department: University of California Berkeley.
8. Ceriani, L. and Verme, P. (2011) 'The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini,' *The Journal of Economic Inequality*, 10(3), pp. 421–443.
9. Dalton, D. R., & Mesch, D. J. (1990). "The Impact of Flexible Scheduling on Employee Attendance and Turnover." Administrative Science Quarterly, 370–387.
10. Desboulets, L. (2018) 'A review on variable selection in regression analysis,' *Econometrics*, 6(4), p. 45.
11. EDUCBA (2023) Decision tree limitations: Learn the limitations of decision trees, EDUCBA. Available at: https://www.educba.com/decision-tree-limitations.
12. Fallucchi, F. *et al.* (2020) 'Predicting employee attrition using machine learning techniques,' *Computers*, 9(4), p. 86.
14. Gaurav, A., Gupta, B. B., & Panigrahi, P. K. (2023). "A comprehensive survey on machine learning approaches for malware detection in IoT-based enterprise information systems." Enterprise Information Systems, 17(3).
15. Joseph, R., Udupa, S., Jangale, S., Kotkar, K., & Pawar, P. (2021). "Employee Attrition using Machine Learning and Depression Analysis." Proceedings - 5th International Conference on Intelligent Computing and Control Systems, ICICCS 2021.
16. Lazzari, M., Álvarez, J.M.S. and Ruggieri, S. (2022) 'Predicting and explaining employee turnover intention,' *International Journal of Data Science and Analytics*, 14(3), pp. 279–292.
16. Liu, J. L. (2014). "Main Causes of Voluntary Employee Turnover: A Study of Factors and Their Relationship with Expectations and Preferences." University Of Chile.
17. LI, F., LI, Y. and WANG, C. (2009) ‘Uncertain Data Decision Tree Classification algorithm’, Journal of Computer Applications, 29(11), pp. 3092–3095.
18. Nagadevara, V., & Srinivasan, V. (2007). "Early Prediction of Employee Attrition in Software Companies-Application of Data Mining Techniques." The 10th International Conference of the Society of Global Business and Economic Development.
19. Najafi-Zangeneh, S. *et al.* (2021) 'An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection,' *Mathematics*, 9(11), p. 1226.

20. Ponnuru, S., Merugumala, G., Padigala, S., Vanga, R., & Kantapalli, B. (2020). "Employee Attrition Prediction using Logistic Regression." International Journal for Research in Applied Science and Engineering Technology, 8(5), 2871–2875.
21. Qutub, A., Al-Mehmadi, A., Al-Hssan, M., Aljohani, R., & Alghamdi, H. S. (2021). "Prediction of Employee Attrition Using Machine Learning and Ensemble Methods." *International Journal of Machine Learning and Computing, 11(2), 110–114.
22. Rombaut, E. and Guerry, M.-A. (2018) 'Predicting voluntary turnover through human resources database analysis,' *Management Research Review*, 41(1), pp. 96–112.
23. IBM HR Analytics Employee Attrition & Performance (2017). https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset.
24. Subhashini, M., R, Gopinath. (2020) 'EMPLOYEE ATTRITION PREDICTION IN INDUSTRY USING MACHINE LEARNING TECHNIQUES,' *ResearchGate.*
25. Vasa, J. et al. (2019) "Foreseeing Employee Attritions using Diverse Data Mining Strategies," International Journal of Recent Technology and Engineering (IJRTE). Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP.
26. Wassan, S. *et al.* (2022) 'Gradient boosting for health IoT federated learning,' *Sustainability*, 14(24), p. 16842.
27. U.S. Department of Labour (2022). Census of Fatal Occupational Injuries Summary New York, NY.
28. Zhao, Y. *et al.* (2018) 'Employee Turnover Prediction with Machine Learning: A Reliable Approach,' in *Advances in intelligent systems and computing*, pp. 737–758.
29. Zutavern, S. (2022) 'Making sense of the new work environment: How employers can support newcomers; onboarding', AERA 2022.