Human Resource analytics

Assignment 2

Cindrella KC 40429497

Dhanush MS 40412492

Mrunmayee Bapat 40420299

Manogna B R 40426970

Rohan Mahesh Patil 40395741

Word count : 3122

Table of Content

Table of Figures

Predicting and Mitigating Employee Attrition in the Pharmaceutical Industry

1.0 Abstract

This report investigates the factors contributing to high employee attrition within a pharmaceutical firm, utilizing predictive analytics. The objectives include identifying root causes and developing proactive HR strategies. Leveraging HRIS data and technologies like KNIME, the study aims to provide the Board with informed decision-making strategies. The comprehensive methodology employs the CRISP-DM framework, exploring data, conducting feature engineering, and employing models like decision trees. The report emphasizes feature improvement, construction, and scaling to enhance model performance. The accuracy was highest for the gradient boosting model at 86.136%, followed by the random forest method at 85%, and the decision tree method at 74.545%. The final model evaluation considers machine learning metrics and interpretability, showcasing the potential for advanced analytics in mitigating attrition challenges and fostering organizational stability.

2.0 Scope and Overview

2.1 Company Overview

Employee attrition is a major issue for the pharmaceutical company, which tracks HR metrics via monthly Excel reports. This persistent issue hinders operations, efficiency, and financial success. They want to use advanced analytics techniques to overcome their reporting system's limitations. The goals are to better comprehend the data and to address high attrition. The company's Human Resource Information System (HRIS) dataset includes attributes, performance measurements, and attrition. Despite having this data, the firm struggles to understand the complex causes of high turnover. To unravel this mystery, complex analytics tools like Knime and machine learning classification models like Gradient Boosting, Random Forest Classifier, and Decision Trees must be used instead of traditional reporting methods. This will help analysts identify causes and create effective attrition-reduction measures.

2.2 Issues relating to high attrition.

The company's 19.15% attrition rate requires a thorough investigation into its causes. According to Jain and Nayyar (2018), attrition is the loss of employees owing to voluntary resignations, death, retirement, and other factors. According to Mozaffari et al. (2022), employee attrition encompasses voluntary and involuntary departures. To predict future attrition patterns, use predictive analytics like KNIME. The goal is to analyse HRIS data to identify key factors that cause employee turnover and present the Board with actionable insights. This predictive modelling approach actively manages attrition and promotes a more stable and engaged workforce to inform strategic decisions.

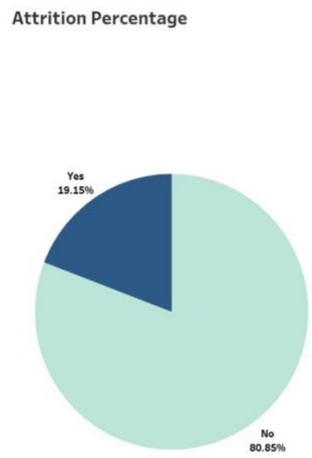


Fig 2.2: Attrition scenario in percentage

Many firms worry about staff turnover due to its complex and far-reaching implications. High attrition costs money, reduces expertise, and lowers employee engagement. Mozaffari et al. (2022) found that high personnel turnover hurts company efficiency. This supports previous research showing that turnover hurts productivity and organisational goals (Han, 2020).

Staff turnover causes many issues that are serious. Knowledge depletion, workflow disruptions, financial costs, and HR stress are issues. The effects also affect team dynamics, employee happiness, client relationships, and service quality, affecting the organization's effectiveness. When hiring replacements, the company pays for recruitment, training, and the complicated interview procedure (Alduayj & Rajpoot, 2018).

3.0 Literature review

Business leaders must deal with competent and important personnel leaving. This issue highlights the need of using machine learning to reduce talent erosion in companies. Machine learning is crucial to strategic decision-making as companies compete for top personnel (Sarah S. Alduayj, 2018). A blended study methodology shows that qualitative and quantitative methodologies can identify employee churn and staff loss issues. Machine learning improves the model's accuracy in identifying staff attrition concerns (Fatemeh Mozaffari, 2022). The study built predictive models using categorization. It assigns each dataset item to a pre-defined class or group. This classification method uses decision trees, linear programming, neural networks, and statistical analyses (Alao et al., 2013).

This study used machine learning models to predict employee turnover based on employee attributes to help management quickly identify and retain at-risk talent. Training on imbalanced data with quadratic SVM yielded the best results (0.50 F1 score), balancing classes with the ADASYN method improved model performance (F1 scores between 0.91 and 0.93 for cubic SVM, random forest, and KNN), and manual under sampling for class balance led to slightly lower performance but still significant F1 scores. This approach shows how machine learning can reduce staff turnover (Alduayj et al., 2018).

For employee attrition prediction, use logistic regression, decision trees, random forests, SVMs, KNN, and Naive Bayes. These models consider demographics (age, gender), job-related factors (satisfaction, tenure), work environment elements (balance, relationships), career development (growth opportunities), personal factors (commute, family obligations), and engagement indicators (feedback, sentiment). These variables predict attrition and reveal historical patterns and relationships. The models use these characteristics to predict employee departure based on dataset patterns. These factors' selection and use affect the model's attrition trend forecasting accuracy, helping businesses develop proactive retention strategies. RS Shankar et al. (2018).

In Marjorie Laura Kane-Sellers' 2007 dissertation, she examined the factors affecting voluntary employee turnover in the professional sales force of a Fortune 500 North American industrial manufacturing corporation. The study examined Voluntary Turnover (VTO) to better understand HRD initiatives that could improve employee retention. The 14-year dataset showed staff dynamics over time. The initial database has 21,271 employee clock-numbered observations. The study used these broad observations to understand voluntary turnover dynamics and inform HRD initiatives to improve employee retention.

4.0 Methodology

Prior to constructing any machine learning model, several preliminary activities must be undertaken. Data storage, data preprocessing, feature engineering, model construction, model enhancement, and ultimately deploying the constructed model for usage, known as model deployment. Wirth, Rudiger, and Jochen Hipp devised a systematic guide called Cross Industry Standard Process for Data Mining (CRISP-DM) to outline the common processes involved in this process (Hotz, 2023). The CRISP-DM methodology provides a clear and structured framework for solving business problems that includes the process of generating models (Refer to the diagram). We will be implementing the CRISP-DM technique to address our business problem. From this point on, we have already dealt with the Business Understanding process. Now, we will proceed to cover each of the other sections.

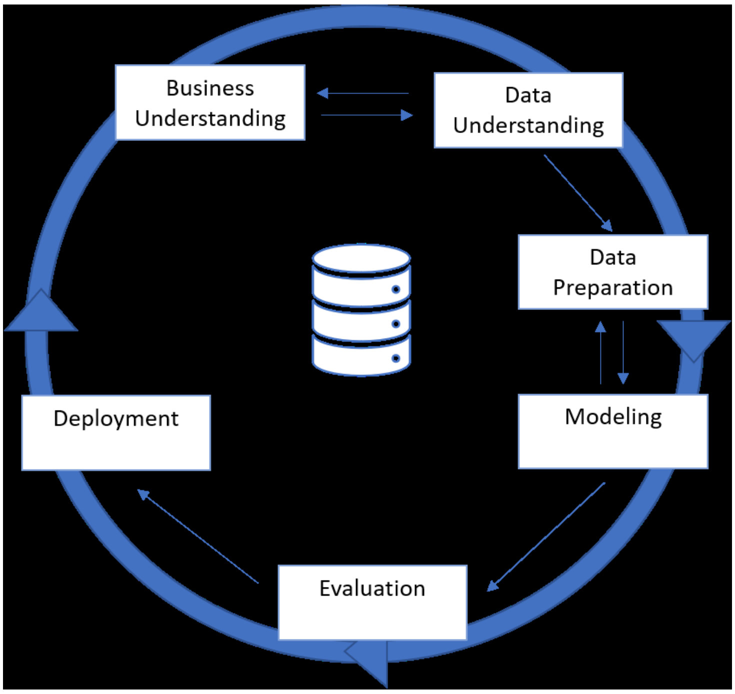


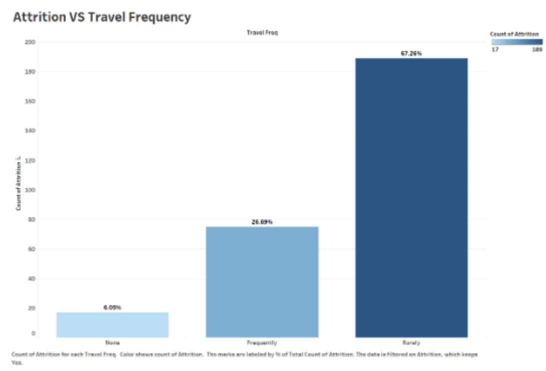
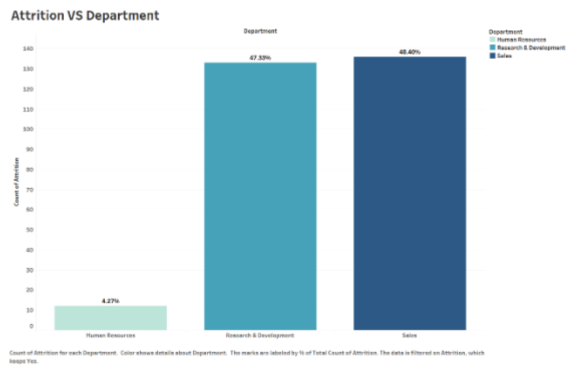
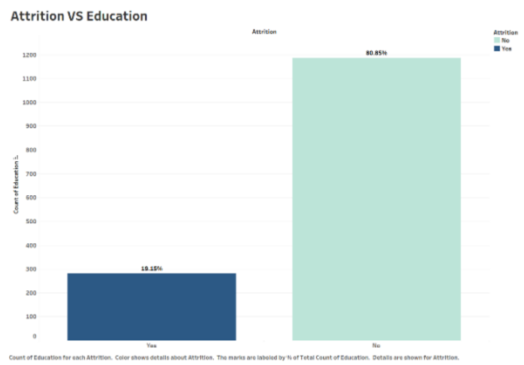
Fig 4: CRISP – DM methodology

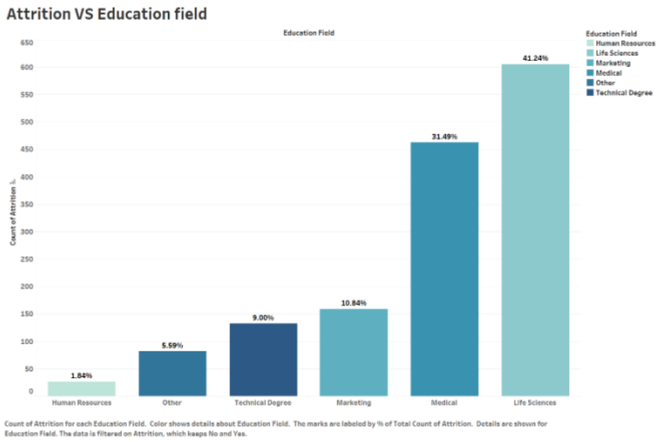
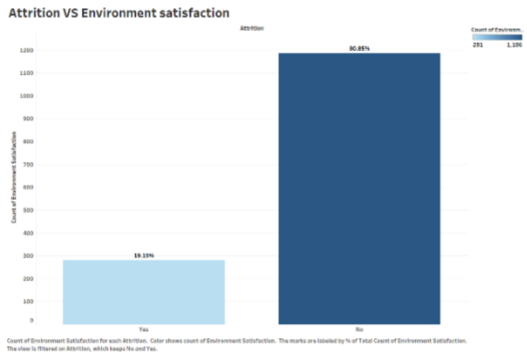
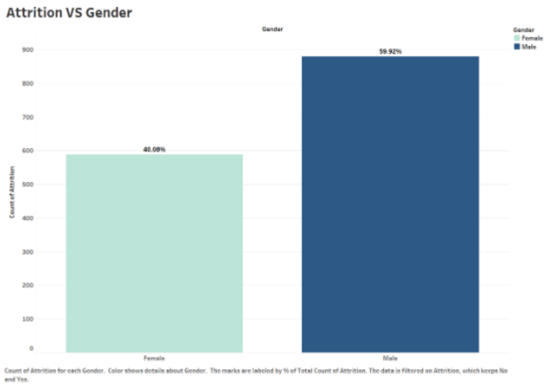
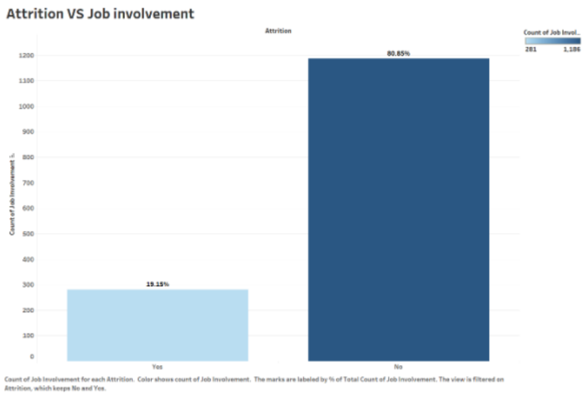
4.1 Exploratory Data Analysis

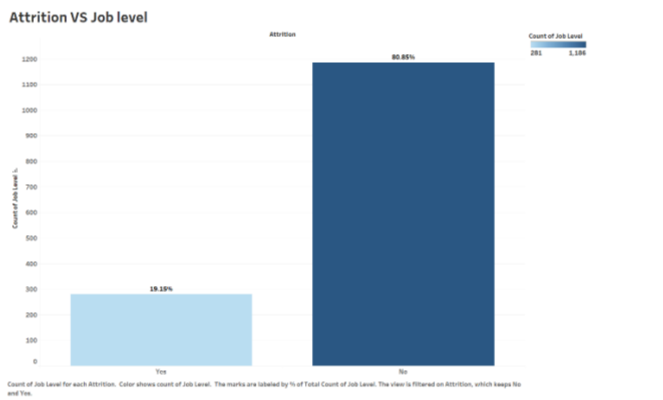
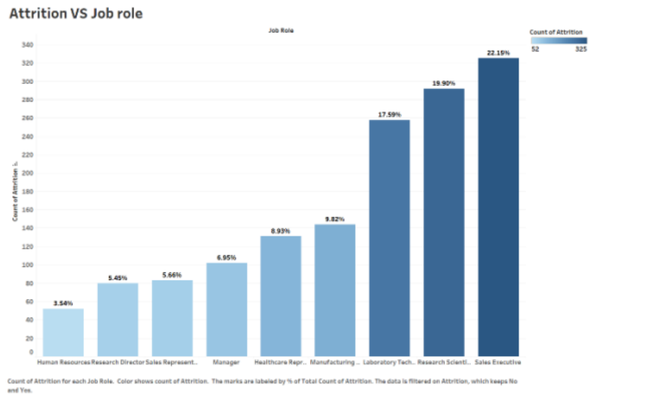
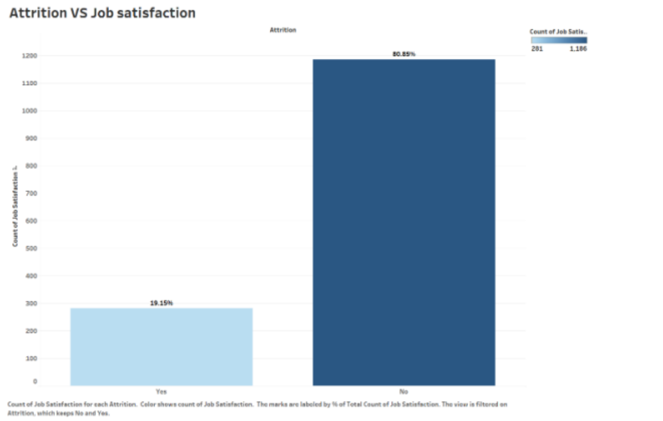
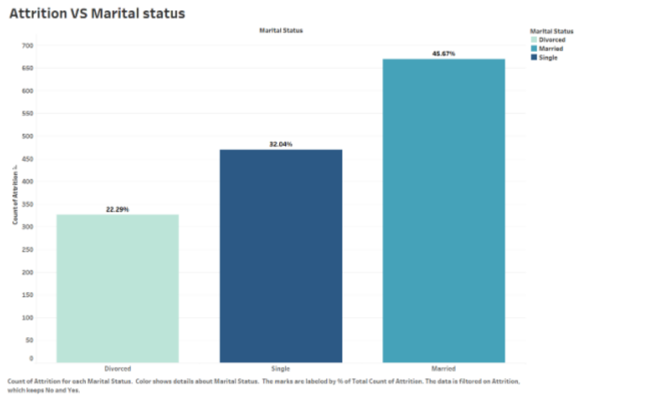
Exploratory Data Analysis begins by examining raw data for trends and patterns (James, 2023). The dataset has 1467 items in 35 variables. These include 25 numeric/integer, 9 text, and 1 identifier and one date variable. There are no logic variables or factors. Uniform two variables have zero variance. 97.14% of variables have complete cases, while 2.86% have <50% missing data. There are no variables with 50%–90% or 90%+ missing data.

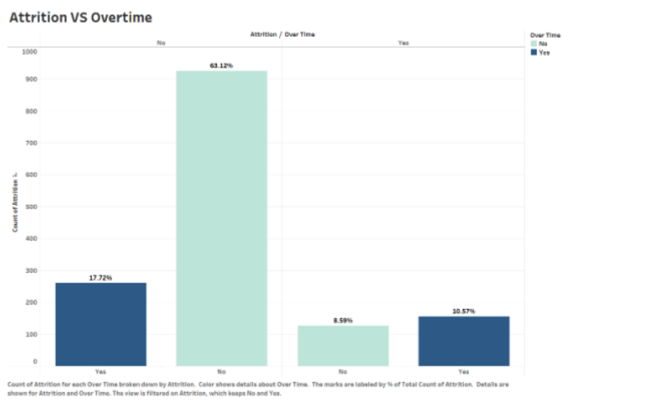
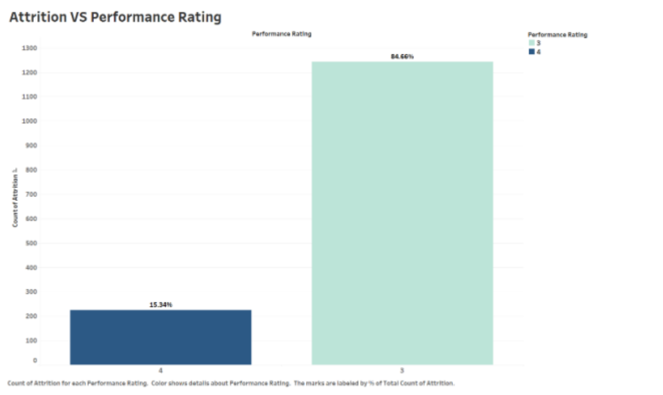
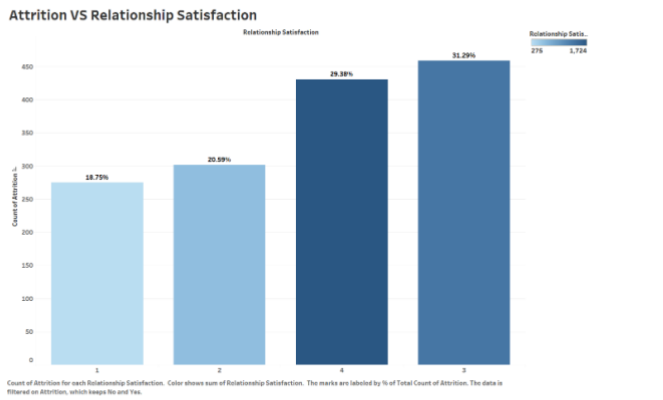
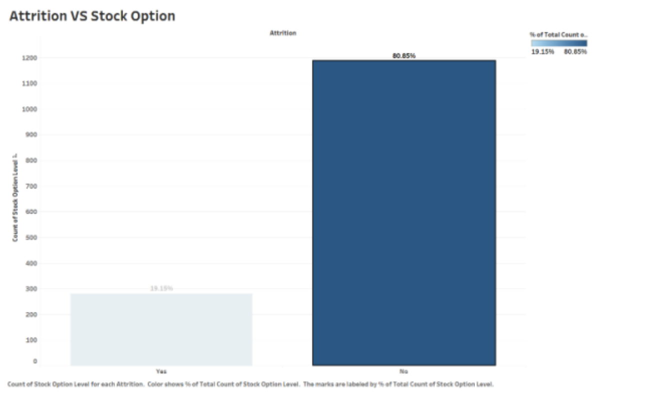
For numerous dataset variables, the table shows numerical statistics per group (Attrition: Yes/No). Each variable's descriptive statistics include min, max, mean, median, SD, CV, IQR, skewness, kurtosis, lower bound of the 25th percentile (LB.25%), upper bound of the 75th percentile (UB.75%), and outlier count.

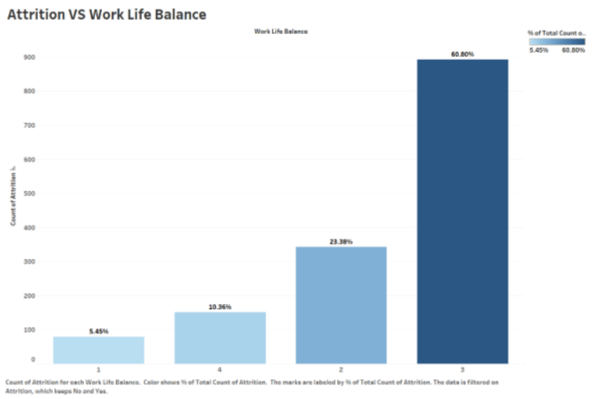
Categorical:

SL NO	Data variable name	Distinct values	Visualization
1	**ATTRITION**	Yes and No	
2	**TRAVEL FREQUENCY**	Rarely, Frequently, None	
3	**DEPARTMENT**	HR, Sales, and R & D	
4	**EDUCATION**	Below Collège, College, Bachelor, Master, Doctor	

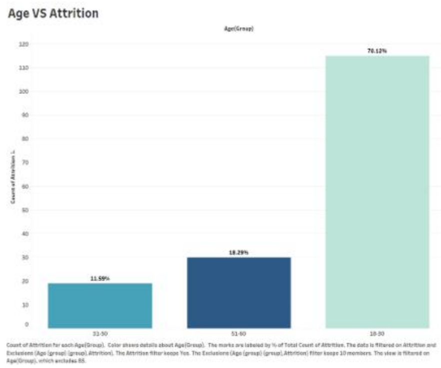
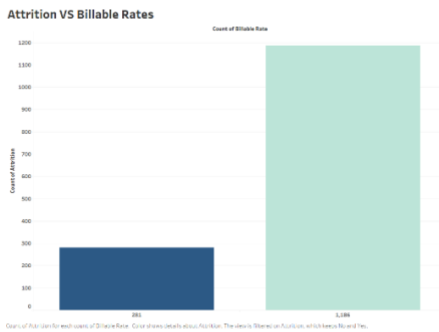
5	**EDUCATION FIELD**	HR, Life Science, Marketing, Medical Science, Others, Technical	
6	**ENVIRONMENT SATISFACTION**	Low, Medium, High, Very High	
7	**GENDER**	Male and Female	
8	**JOB INVOLVEMENT**	Low, Medium, High, very high	

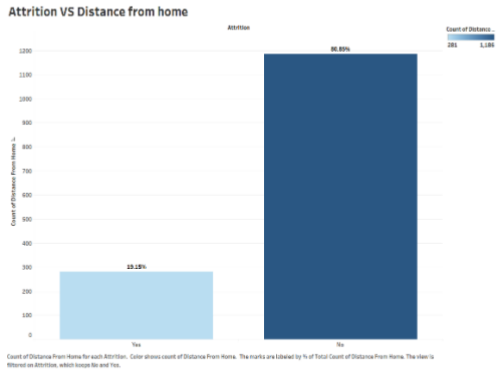
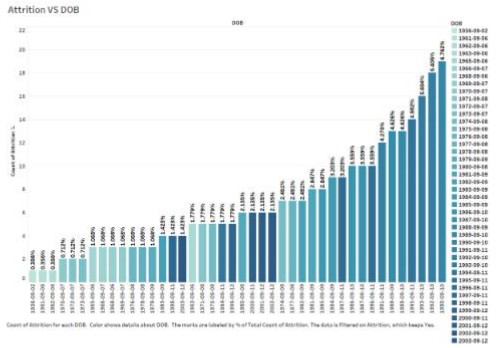
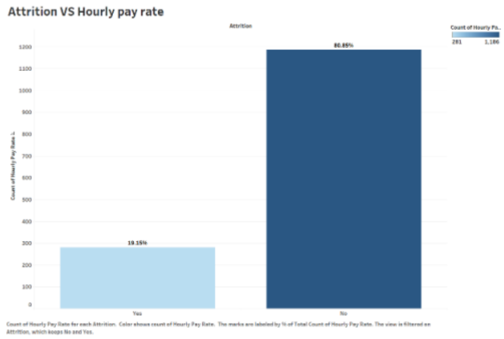
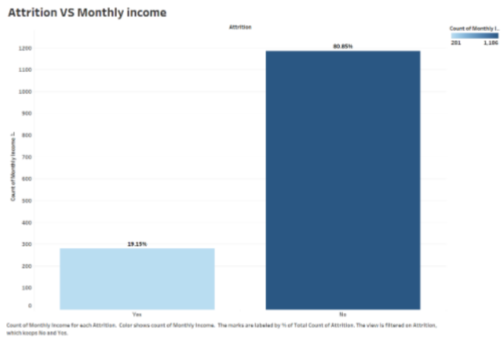
9	**JOB LEVEL**	Junior to Senior	
10	**JOB ROLE**	Sales Executive, Research Scientist, Laboratory Technician, Manufacturing Director, Healthcare, Representative, Manager, Sales Representative, Research Director, Human Resources	
11	**JOB SATISFACTION**	Low, Medium, High, Very high	
12	**MARITAL STATUS**	Single, Married and Divorced	

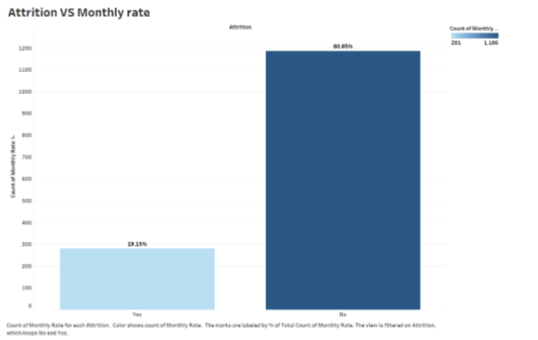
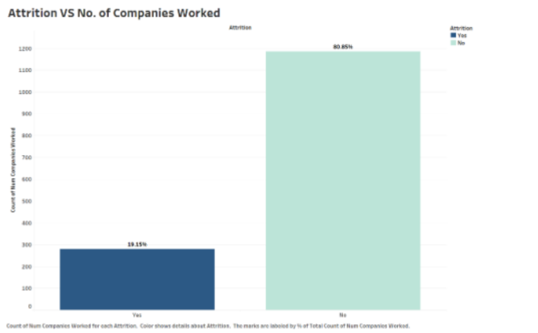
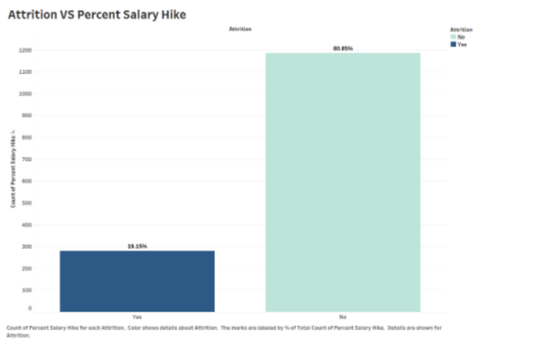
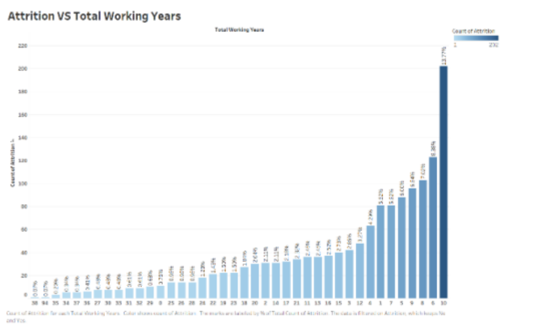
13	**OVERTIME**	Yes and No	
14	**PERFORMANCE RATING**	Low, Good, Excellent, Outstanding	
15	**RELATIONS SATISFACTION**	Low, Medium, High, Very High	
16	**STOCK OPTION LEVEL**	None, Low, Medium, High	

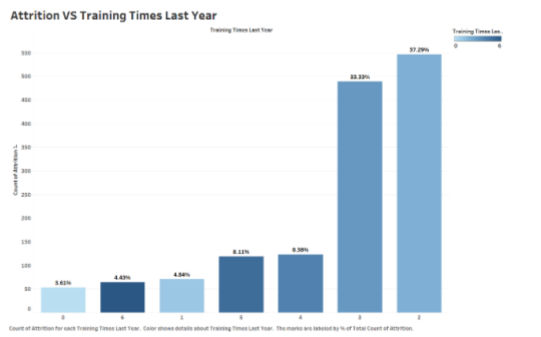
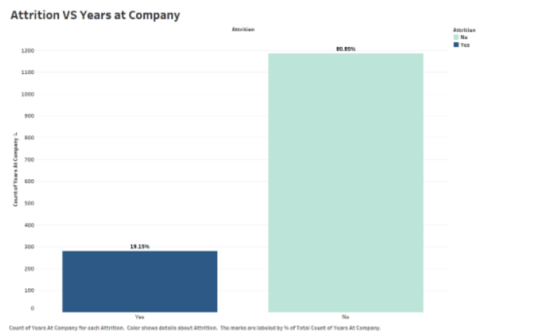
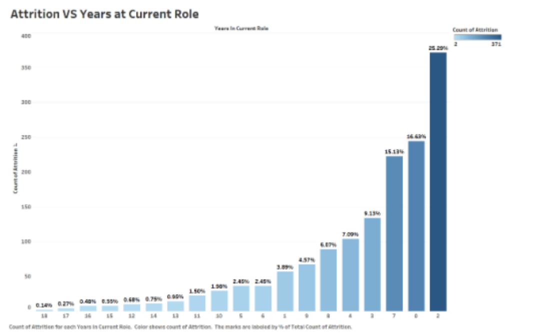
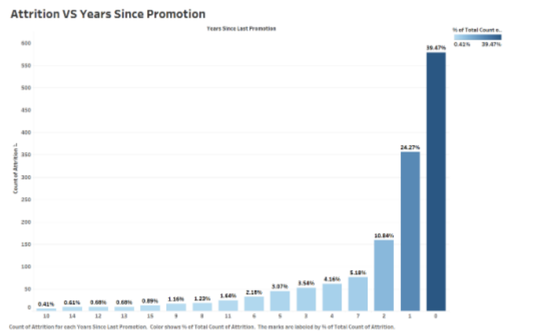
17	**WORK LIFE BALANCE**	Bad, Good, Better, Best	

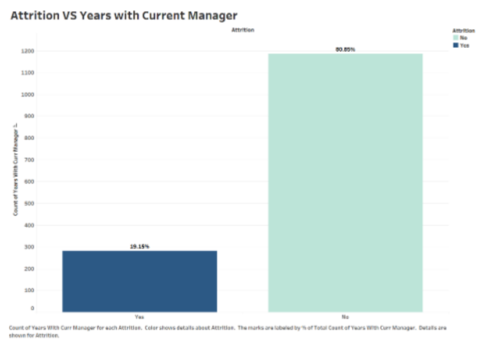
Numerical:

SL NO	Data variable Name	Minimum	Mean	Median	Maximum	Visualization
1	AGE	7	36.97	36	85	
2	BILLABLE RATE	102	802.9	802	1499	

3	DISTANCE FROM HOME(MILES)	1	9.2002	7	29	
4	DOB	1936-09-02	-	-	2003-09-12	
5	HOURLY RATE	30	65.87	66	100	
6	MONTHLY INCOME ($)	1009	6505	4908	19999	

7	MONTHLY RATE	2094	14323	14255	26999	
10	NUMCOMPANIES WORKED	0	2.695	2	9	
11	PECENT SALARY HIKE	11	15.21	14	25	
12	TOTAL WORKING YEARS	0	11.32	110	94	

13	TRANING TIMES LAST YEAR	0	2.801	3	6	
14	YEARS AT COMPANY	0	7.01	5	40	
15	YEARS IN CURRENT ROLE	0	4.32	3	18	
16	YEARS SINCE LAST PROMOTION	0	2.192	1	15	

17	YEARS WITH CURRENT MANAGER	0	4.123	3	17	

4.2 Feature Engineering

The technique of altering and changing data into a format that optimally depicts the underlying problem that an ML Algorithm is attempting to predict while mitigating inherent complexities and biases within data is known as feature engineering. Feature Engineering is next step in our process of model building right after EDA(See the figure 4.2). Feature Engineering is an umbrella term which has five steps in it Feature Improvement, Feature Construction, Feature Selection, Feature extraction and finally Feature learning (Ozdemir, 2022).

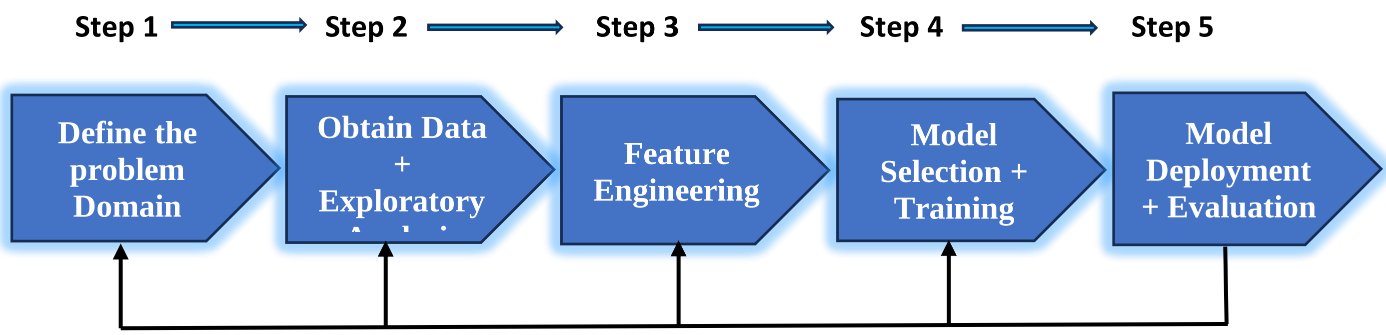


Fig 4.2: Flow chart of Feature Engineering

How do we know that the Feature Engineering techniques employed improved the performance of the model? There are 4 approaches as mentioned by (Ozdemir, 2022), Machine Learning Metrics, Interpretability, Fairness and Bias and ML complexity and speed. We will use Machine Learning Metrics (Accuracy and Cohen Kappa) and Interpretability for comparison. We chose Decision tree algorithm as it has in-built feature selection technique.

The Base Model Performance (4.2.1) with overall accuracy of 71.16% and Cohen's Kappa (k) 0.108.

Scorer View

Confusion Matrix

	No (Predicted)	Yes (Predicted)	
No (Actual)	281	68	80.52%
Yes (Actual)	56	25	30.86%
	83.38%	26.88%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
71.16%	28.84%	0.108	306	124

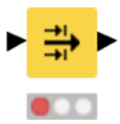
Row ID	True ...	False ...	Tru...	False ...	Recall	Precision	Sensitivity	Specificity	F-me...
No	281	56	25	68	0.805	0.834	0.805	0.309	0.819
Yes	25	68	281	56	0.309	0.269	0.309	0.805	0.287

Fig 4.2.1: Model Statistics before Feature engineering

4.2.1 Feature Improvement

Feature Engineering involves enhancing structured features through various transformations, such as handling missing values, standardisation, and normalisation of numerical and categorical data (Ozdemir, 2022). Our dataset has negligible missing values, omitting the need for imputation.

Table below shows Feature Improvement clearly delineates all the steps performed under the Feature Improvement step. Kindly refer to the table.

Data Variable	Data Type	DQ Issues	Description	Node Used
Age	Numerical Data	7,77,85	Data points for individuals aged 7, 77, and 85 are considered outliers because standard employment ages range from 18 to 60 years, rendering employment outside this range, particularly at age 7, highly improbable.	**Row Filter**  Filtering the Age Outlier

Department	Categorical Data	HR, Human Resources, R & D, Resource and Development	The same categories are listed twice and should be unified under one label, either as HR and R&D or Human Resource and Resource and Development.	**String Manipulation** Improving Department Column
Total Working Years	Numerical Data	94	This data point is likely an outlier, as having 94 years of work experience is extremely unlikely.	**Row Filter** Filtering the Age Outlier
Attrition	Categorical Data	NA	The Attrition is highly imbalanced, so to make it balanced SMOTE resampling technique was used	**SMOTE** Attrition Resample

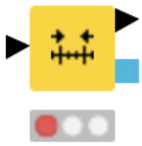
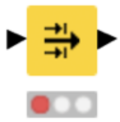
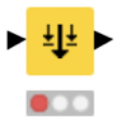
4.2.2 Feature Construction

There are primarily two methods of feature transformation: log transformation and Box-Cox transformation(Zheng & Casari, 2018 and Ozdemir, 2022 and). Nevertheless, both of these transformations exclusively operate on data that is strictly positive (Ozdemir, 2022). We employ the Yeo-Johnson transformation to handle negative values (Ozdemir, 2022) .We will utilise the Box-Cox transformation, which is an extension of the log-transformation.

As seen in the EDA of the numerical variable "YearsSinceLastPromotion" (Skewness = 1.979) and "MonthlyIncome" (Skewness = 1.366) exhibit notable skewness, indicating the need for modifications. Furthermore, the variables "YearsAtCompany" (with a skewness value of 1.761) and "TotalWorkingYears" (with a skewness value of 1.705) demonstrate a moderate level of skewness, indicating a potential requirement for transformation.

4.2.3 Feature Scaling

There are two primary techniques for feature scaling: Min-Max Standardisation and z-score normalisation(Ozdemir, 2022). Feature scaling is conducted to standardise all the feature variables to a uniform scale (Ozdemir, 2022). However (Huyen, 2022 and Ozdemir, 2022) argues that Decision Trees, Random Forests, and Gradient Boosting are considered insensitive to feature scaling, which is why standardisation is considered unnecessary for these algorithms.

Data Variable	Data Type	DQ Issues	Description	Node Used
YearsSinceLast Promotion, YearsAtCompany, MonthlyIncome	Numerical	Skewness	Due to skewness it might affect the accuracy of the models performance	**Normalizer**  **Normalization of Variables**
Total Working Years	Numerical	94	Total Working Years	**Row Filter**  Filtering the Age Outlier
ID and DOB	Numerial and Date	NA	ID and DOB are irrelevant columns for the predictive analysis	**Column Filter**  Filtering ID and DOB Column

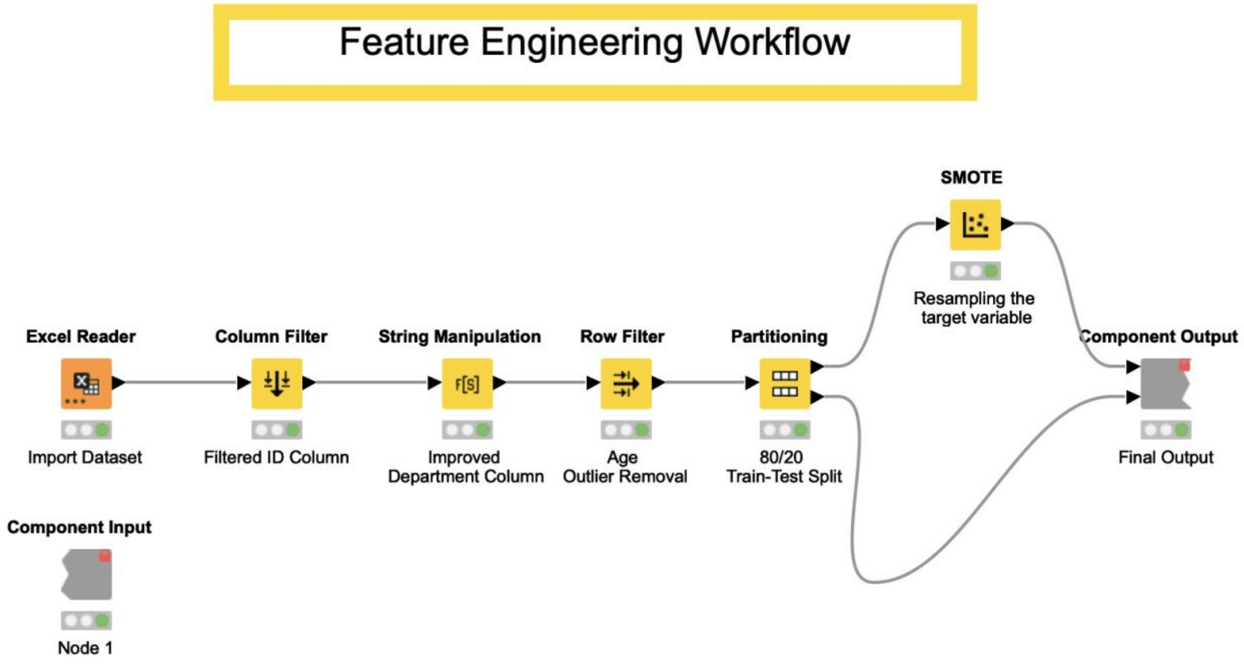


Fig 4.2.3: Feature engineering workflow

The Entire workflow of feature engineering performed in KNIME is shown in the Figure().

4.3 Model Building

Machine learning, a subset of artificial intelligence, streamlines analytical model creation by automating data analysis. It relies on computers' ability to derive insights from data, minimizing human intervention (Sarmento et al., 2021).

The process involves integrating the Feature Engineering pipeline into a KNIME meta node and selecting three distinct classification models: Decision Trees, Random Forest, and Gradient Boosted. Data segmentation via the Partitioning Node (80-20 split) precedes model training with Learner Nodes. Predictions for "Attrition" utilize Predictor Nodes.

Uniform seed "40395741" is applied. Accuracy is assessed by the Scorer Node, while the ROC Curve Node visualizes model performance. All modeling occurs within KNIME's native nodes, detailed in a comprehensive table.

1. Decision Tress

Decision trees, efficient for classification and prediction, offer a hierarchical, rule-explaining model. Nodes represent categories or decisions based on attribute-values, guiding instance categorization. Beginning at the root, traversal along branches leads to a leaf node, finalizing instance categorization (Tahir et al., 2006).

The KNIME Decision Tree Learner configuration optimally sets 'Attrition' as the class column, using the Gini index to ensure effective split quality. It features Reduced Error Pruning to enhance model generalization, and an average split point for fair decision boundaries. The capacity to handle 10,000 records for viewing and the use of 8 threads demonstrates the platform's robust data processing capabilities, geared towards achieving high model accuracy.

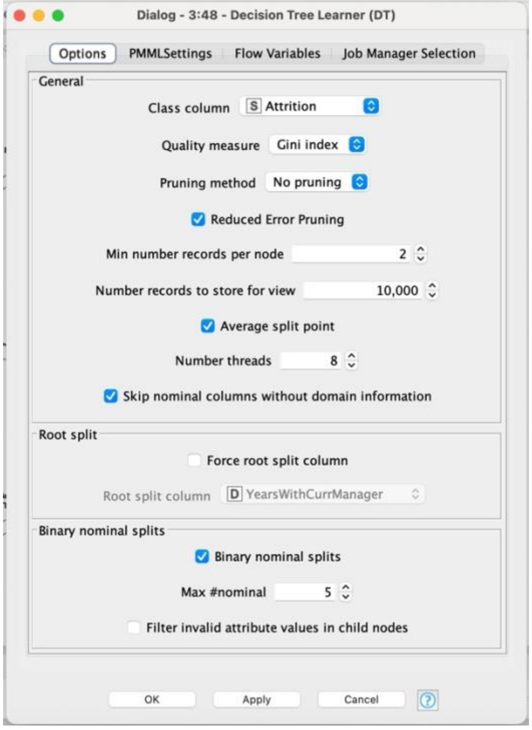


Fig 4.3.1: KNIME Decision tree Parameters

The table displays the nodes utilised for constructing the model and their respective purposes:

Node name in Knime	Node pictorial Representation	Description
Decision Tree Learner		The Decision Tree Learner in KNIME constructs a tree-based predictive model from input data, utilizing feature attributes to make decisions and predict outcomes for classification or regression tasks.
Decision Tree Predictor		The Decision Tree Predictor in KNIME applies a trained decision tree model to new data, making predictions based on the learned patterns and structures within the tree, allowing for classification or regression tasks.
Scorer		The Scorer node in KNIME evaluates the performance of machine learning models by comparing predicted values against actual values, providing metrics like accuracy, precision, recall, and F1-score for classification tasks or error metrics like RMSE for regression tasks.

2. Random Forest

Random Forest, a supervised learning method, classifies and predicts data. It differs from decision trees in locating root nodes for feature splitting. Effective with missing values, it requires numerous trees to prevent overfitting, enhancing prediction accuracy (Rony et al., 2021; Hassan et al., 2021).

In KNIME, Random Forest Learner predicts 'Attrition' using key attributes like 'Age' and 'TravelFreq' etc . Gini Index guides split decisions for fair node distribution. Limited tree depth (10) and minimum node size (1) capture intricate data patterns, avoiding overfitting for a comprehensive analysis.

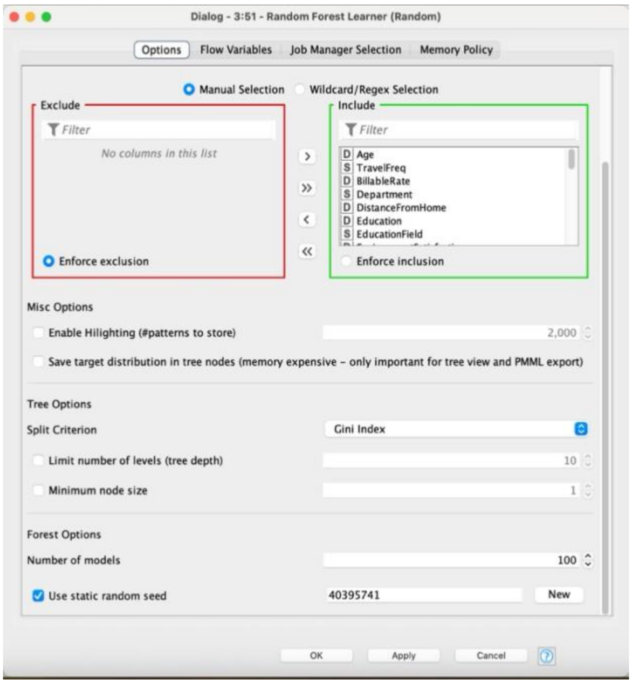
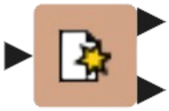


Fig 4.3.2: KNIME Random Forest Parameters

The table displays the nodes utilised for constructing the model and their respective purposes:

Node name in Knime	Node pictorial Representation	Description
Random forest Leaner Learner		Generates an ensemble of decision trees using random subsets of data, combining their predictions to create a robust model for classification or regression tasks.
Random Forest predictor		Takes a trained Random Forest model and uses it to predict outcomes on new data within KNIME, leveraging the collective intelligence of multiple decision trees to make accurate predictions.

Scorer		The Scorer node in KNIME evaluates the performance of machine learning models by comparing predicted values against actual values, providing metrics like accuracy, precision, recall, and F1-score for classification tasks or error metrics like RMSE for regression tasks.

3. Gradient Boosted Trees

Gradient boosting, a machine learning approach for regression and classification, amalgamates weak prediction models, often decision trees (Breiman, L., 2020). It incrementally constructs models, extending boosting's capabilities by optimizing differentiable loss functions. Inspired by Breiman, it views boosting as an optimization technique for relevant cost functions.

The configuration for the Gradient Boosted Trees model in KNIME is specifically customised, focusing on predicting 'Attrition' using carefully selected key attributes to improve the model's relevance. The restriction on the maximum number of levels in the trees is in accordance with a more precise modelling technique. The configuration of 100 trees with a learning rate of 0.1 demonstrates a purposeful and advanced approach to learning from the data.

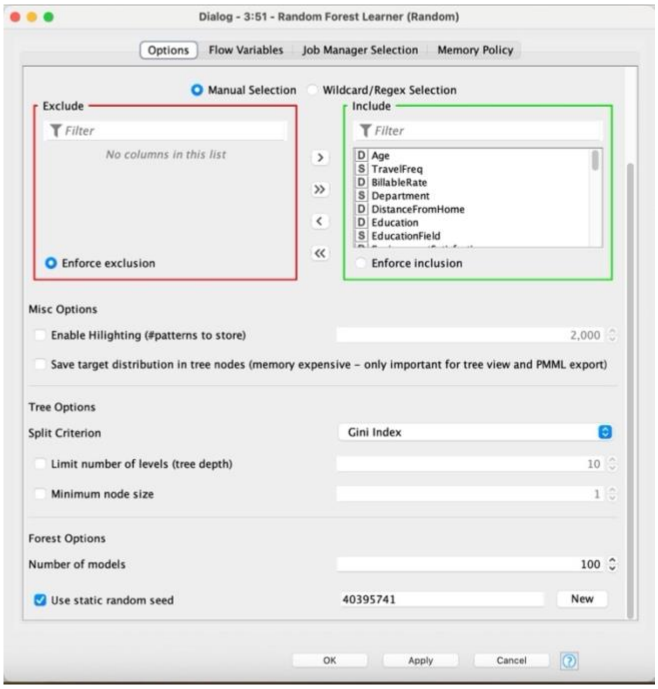
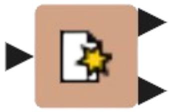


Fig 4.3.3: KNIME Gradient Boosted Trees

The table displays the nodes utilized for constructing the model and their respective purposes:

Node name in Knime	Node pictorial Representation	Description
Gradient Boosted Trees Learner		Constructs a predictive model by sequentially adding decision trees to refine predictions and enhance accuracy, specifically designed for classification or regression tasks.

Gradient Boosted Trees predictor		Applies a pre-trained Gradient Boosted Trees model to new data in KNIME, providing highly accurate predictions by leveraging the sequentially improved ensemble of decision trees.
Scorer		The Scorer node in KNIME evaluates the performance of machine learning models by comparing predicted values against actual values, providing metrics like accuracy, precision, recall, and F1-score for classification tasks or error metrics like RMSE for regression tasks.

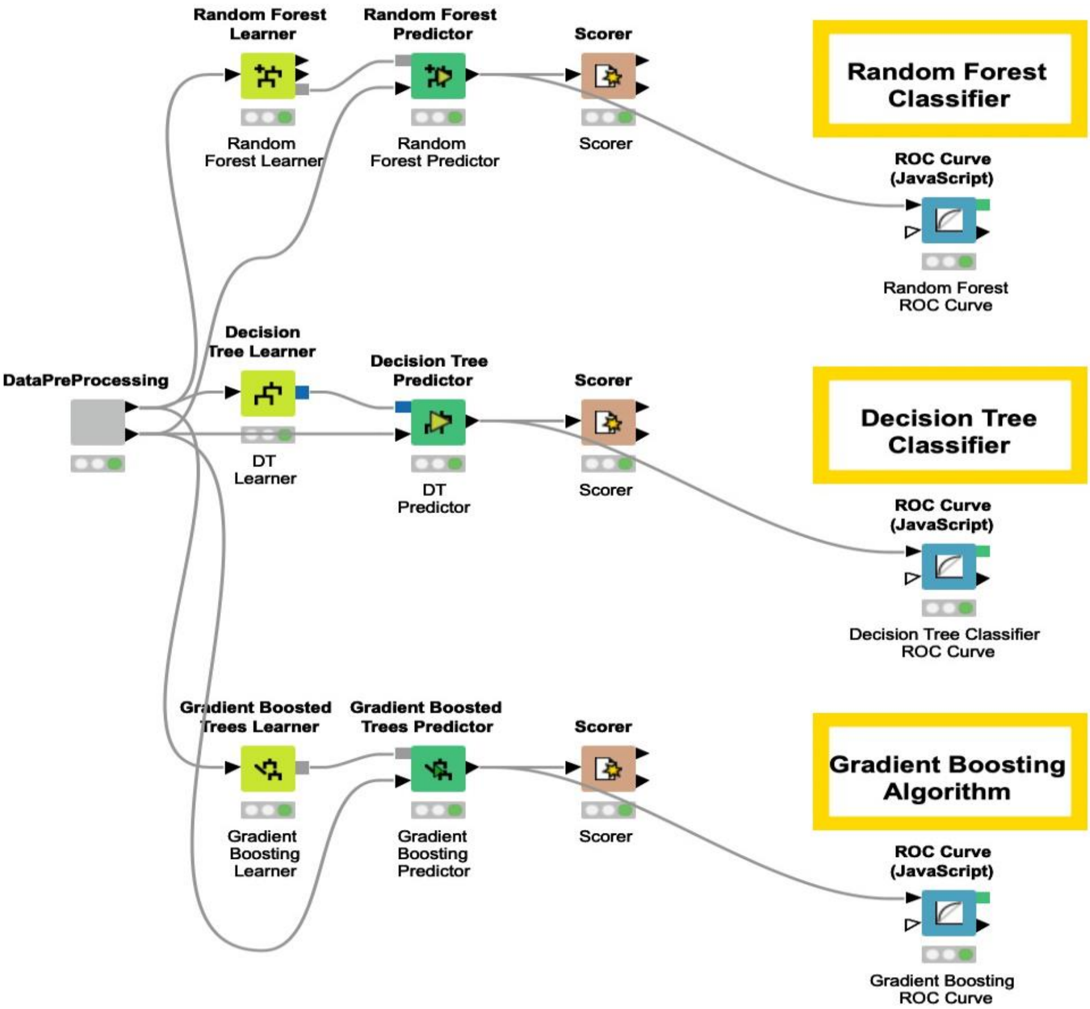


Fig 4.3.4: Model Building workflow in KNIME

4.4 Interpreting Decision Tree

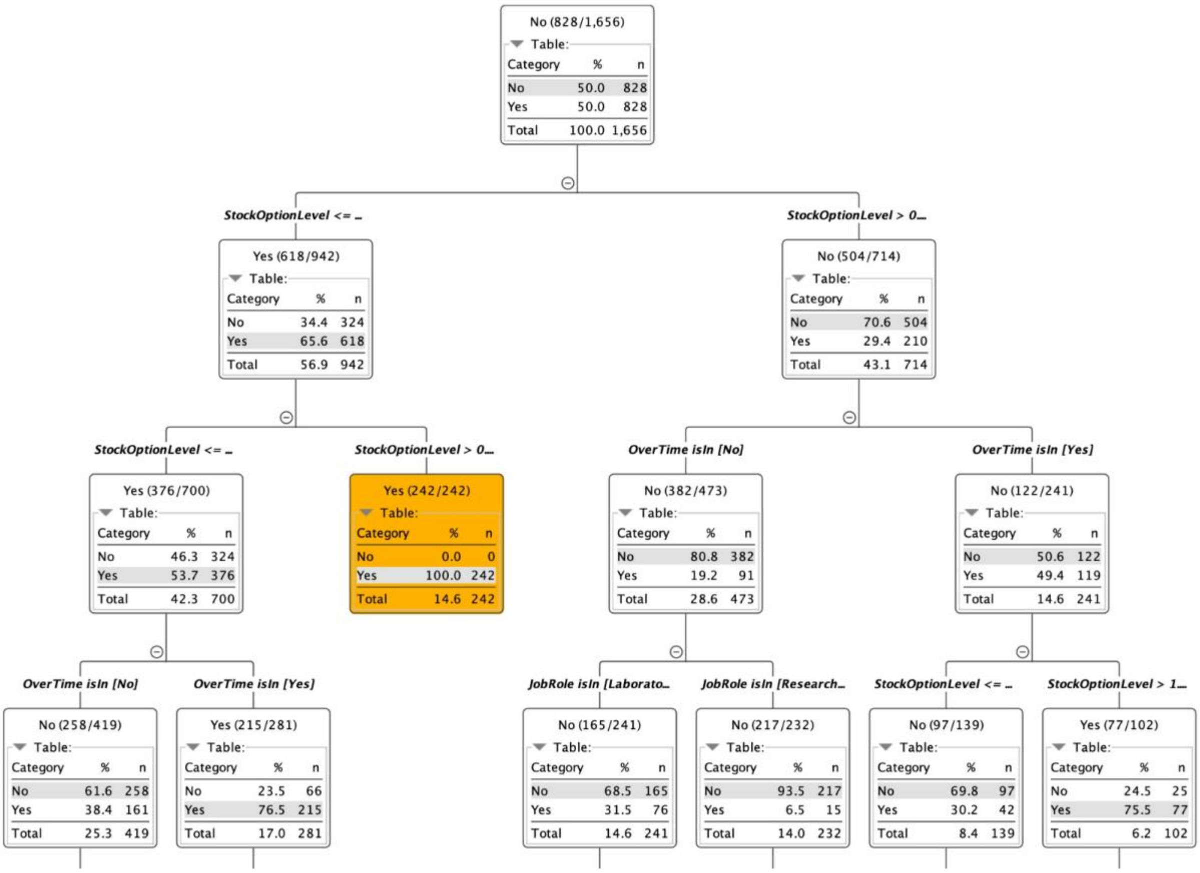


Fig 4.4: Decision Tree

Our study, done using the KNIME Analytics Platform, utilised a decision tree model that employed a binary split technique to classify prospective staff attrition. This model utilised key features, namely 'StockOptionLevel' and 'Overtime', which played significant roles in the decision-making process of the tree. The branches and nodes of the tree revealed a distinct pattern of how these variables influenced attrition outcomes, enabling us to further explore the predictive features of employee turnover.

The decision tree graphically displayed the systematic divisions, each intended to progressively improve the forecast precision. For instance, the original division based on the variable 'StockOptionLevel' was subsequently narrowed down by considering the 'Overtime' status and 'JobRole'. Each branch in the decision tree represents a distinct probability of attrition. The level of granularity provided a thorough and practical understanding, allowing HR departments to precisely identify key elements that contribute to employee turnover.

By utilising a confusion matrix to assess the model's performance, we obtained a commendable overall accuracy of 74.545%. The model accurately predicted 328 cases out of 440, demonstrating its reliability for developing and implementing HR policies. The accuracy statistic, bolstered by a Cohen's Kappa score, confirms the model's ability to accurately classify attrition, showcasing its practical usefulness in real-world HR situations.

Furthermore, the confusion matrix emphasised the model's effectiveness, especially in accurately forecasting the "No" attrition category with a substantial percentage of real negatives. The model's prediction's favourable element highlights its efficacy and dependability. Essentially, the decision tree model has demonstrated its worth as an invaluable analytical tool, delivering a remarkable level of precision in forecasting staff attrition. The successful implementation of machine learning techniques in this study demonstrates the potential for deriving valuable insights that aid in strategic HR decision-making and proactive retention efforts.

4.5 Model Evaluation

Model's Accuracy and result

Model	Correctly Classified	Incorrectly Classified	Accuracy (%)	Error (%)	Cohen's Kappa (%)
Random Forest	374	66	85.0	15.0	0.419
Decision Tree	328	112	74.545	25.455	0.251
Gradient Boosting	379	61	86.136	13.864	0.509

The Gradient Boosting model demonstrates exceptional prediction performance, with the best accuracy rate of 86.136% and the strongest Cohen's Kappa score of 0.509. These results indicate a robust classification ability with significant reliability. The Random Forest model exhibits impressive predictive accuracy, achieving an 85% score, along with a respectable Cohen's Kappa score of 0.419, which indicates its usefulness. Nevertheless, the Decision Tree model, although still valuable, exhibits a relatively lower level of accuracy (74.545%) and a moderate Cohen's Kappa score (0.251), indicating that it might potentially be enhanced with additional optimisation. The combination of these metrics demonstrates the superior performance of ensemble approaches compared to individual predictors in intricate classification tasks.

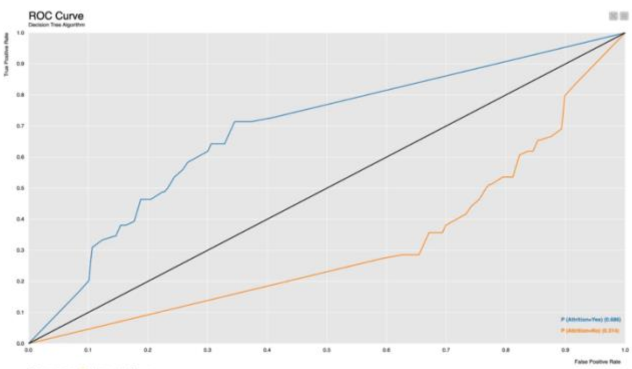


Fig 4.5.1: ROC curve decision Tree

The ROC curve shows a decision tree algorithm's moderate performance for the "Yes" class (AUC: 0.686) and poor performance for the "No" class (AUC: 0.314).

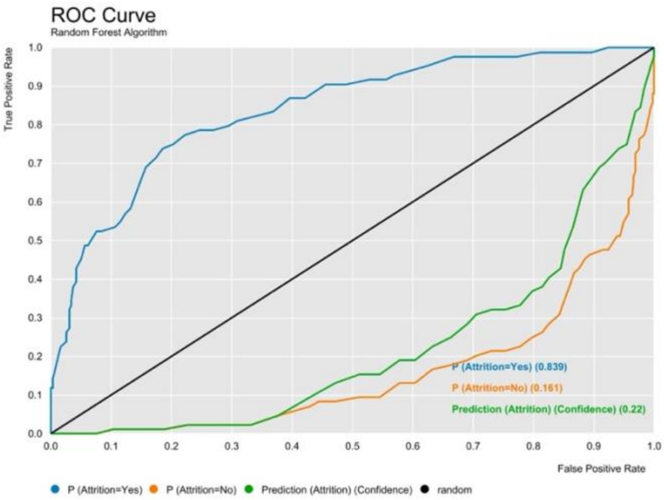


Fig 4.5.2: ROC curve of Random Forest algorithm

The ROC curve for the Random Forest Algorithm shows good performance for the "Yes" class (AUC: 0.839), but the "No" class has a low AUC of 0.161, with overall higher confidence in predictions (0.22) compared to random chance.

The ROC curve for the Gradient Boosting model indicates a strong ability to predict the "Yes" class (AUC: 0.842) and poor performance on the "No" class (AUC: 0.158), with a confidence level for predictions marked at 0.23.

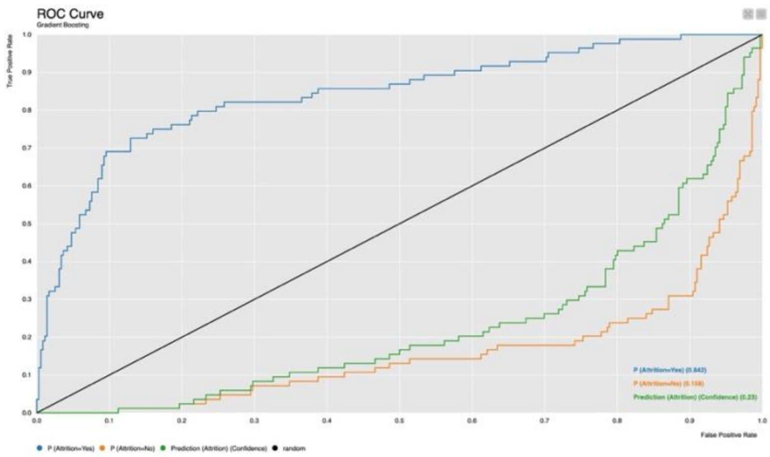


Fig 4.5.3: ROC curve of Gradient boosting model

Comparison of the model curves

The ROC curve analysis across the three machine learning models—Decision Tree, Random Forest, and Gradient Boosting—reveals distinct performance characteristics. The Decision Tree algorithm shows moderate and balanced predictive performance for both "Yes" and "No" classes, with its simplicity and interpretability but with a potential for overfitting. In contrast, the ensemble models, Random Forest, and Gradient Boosting, display strong predictive abilities for the "Yes" class, with AUCs over 0.8, but fall short in predicting the "No" class, with AUCs around 0.16. These ensemble methods, which are more complex and less interpretable than single trees, mitigate overfitting through aggregation and sequential correction, respectively, suggesting a trade-off between model complexity and performance. Also, it is noteworthy to observe the feature engineering has had an impact on increasing the accuracy of decision tree from 71.17 % to 74.545%.

5.0 Results and Discussion

5.1 Key Insights

1. The CRISP-DM technique emphasizes data mining's importance in solving business challenges methodically.

2. Normalization and absent value management are examples of feature engineering strategies that improve model performance.

3."SMOTE for Class Imbalance" fixes dataset class imbalance concerns to improve model performance.

4. Predictive Model Construction: The KNIME Analytics Platform was used to do rigorous predictive analysis with an 80%-to-20% training-to-testing ratio.

5. Gradient Boosting Model: Had the highest accuracy rate of 86.136%, showing strong employee attrition prediction.

6. The decision tree model's balance was 85% accurate and interpretable, but not as precise as gradient boosting.

"Underperformance of the Random Forest Model": The accuracy of 74.545% suggests this dataset is tough to manage.

7.Model Evaluation Rigor: Accuracy and Cohen's Kappa statistics were used to analyze performance, providing Decision Tree scores of 0.419, Random Forest scores of 0.251, and Gradient Boosting scores of 0.509.

8. The study indicates that variables such as 'StockOptionLevel' and 'Overtime' have a significant impact on employee attrition. It was observed that employees with a lower number of stock options and mandatory overtime have a higher propensity to resign. These elements, along with additional ones, were crucial in the construction of predictive models, highlighting their substantial effect on the turnover rates of staff in the pharmaceutical sector.

5.2 Future Recommendations

Maximizing Decision Tree Model Use: Given its notable accuracy and interpretability, the Decision Tree model's insights could be pivotal in HR analytics. Future strategies should focus on using this model to identify key factors driving attrition and develop targeted interventions. Enhancing the model with more complex data could further refine these insights, aiding strategic HR decisions. (Benjamín et al., 2022)

Focus on Feature Engineering and Data Quality: Given the importance of feature engineering in model performance, future efforts should concentrate on refining these techniques.

Additionally, ensuring high data quality, including addressing class imbalances and missing data, will be crucial. (Benjamín et al., 2022)

Expanding Model Evaluation Metrics: While the report used accuracy and Cohen's Kappa statistic, incorporating additional evaluation metrics like Precision, Recall, and F1-Score can provide a more comprehensive understanding of model performance, especially in scenarios of class imbalance. (Benjamín et al., 2022)

Integration into HR Decision-Making Processes: The insights gained from these models should be integrated into the HR decision-making processes to proactively address factors contributing to attrition, thus enhancing employee retention strategies. (Benjamín et al., 2022)

5.3 Limitations

1.Data and decision tree limitations:

The data that is currently available may limit the model's performance. The accuracy of the model might be increased by compiling more thorough data on employee experiences. Decision trees include drawbacks, such as the possibility of overfitting in situations with insufficient training data and difficulties managing continuous variables (Zorman et al., 1997)

2.Complexity of attrition:

A variety of internal and external factors can have an impact on the complex phenomenon of employee attrition. The intricacy of these influences could restrict the predicted accuracy of the model. The tendency to oversimplify intricate relationships may make thorough research and accurate predictions more difficult. (Zorman et al., 1997)

3.Interpretability Challenges:

Even though the decision tree model works well, interpretability issues may arise because of its opaqueness when elucidating intricate interactions between features.

4.Model Validation:

The main areas of emphasis for the current evaluation are Cohen's Kappa and accuracy. To guarantee the model's dependability, robust validation approaches should be incorporated into subsequent versions.

6.0 References

- Alduayj, S.S. and Rajpoot, K. (2018) 'Predicting employee attrition using machine learning', 2018 International Conference on Innovations in Information Technology (IIT) [Preprint]. doi:10.1109/innovations.2018.8605976 Available at https://ieeexplore.ieee.org/abstract/document/8605976 Accessed on 27th Dec 2023].

- Benjamín, M.-Q. et al. (2022) 'A predictive model implemented in Knime based on learning analytics for timely decision making in virtual learning environments', International Journal of Information and Education Technology, 12(2), pp. 91–99. doi:10.18178/ijiet.2022.12.2.1591 Available at https://www.researchgate.net/profile/Syed-Muzamil Basha/publication/334605232_Comparative_Study_on_Performance_of_Document_Classification_Using_Supervised_Machine_Learning_Algorithms_KNIME/links/5d35969f4585153e5916bdae/Comparative-Study-on-Performance-of-Document-Classification-Using-Supervised-Machine-Learning_Algorithms-KNIME.pdf [Accessed on 3rd Jan 2023].

- Fafalios, S., Charonyktakis, P., & Tsamardinos, I. (2020). "Gradient Boosting Trees." Gnosis Data Analysis PC. Available at: https://www.gnosisda.gr/wp-content/uploads/2020/07/Gradient_Boosting_Implementation.pdf [Accessed on : 4th January , 2024]

- Hassan, Md.M. et al. (2021) 'Diabetes prediction in healthcare at early stage using Machine Learning Approach', 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) [Preprint]. doi:Available at : https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9579869 [Accessed on 4th January 2024)].

- Huyen, C. (2022) Designing machine learning systems: An iterative process for production-ready applications. Sebastopol, CA: O'Reilly Media.

- James, G. et al. (2023) An introduction to statistical learning with applications in Python. Cham: Springer International Publishing.

- Jain, R. and Nayyar, A. (2018) 'Predicting employee attrition using XGBoost machine learning approach', 2018 International Conference on System Modeling & Advancement in Research Trends (SMART) [Preprint]. doi: Available at:

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8746940 [Accessed on: 4th November 2023].

- Kuhn, M. and Johnson, K. (2019) Applied predictive modeling. New York: Springer.

- Lama, D. and Mishra, S. (2017) A decision making model for human resource management in organizations using data mining and Predictive Analytics, Academia.edu. Available at: https://www.academia.edu/26628757/A_Decision_Making_Model_for_Human_Resource_Management_in_Organizations_using_Data_Mining_and_Predictive_Analytics (Accessed: 27 December 2023).

- Marín Díaz, G., Galán Hernández, J.J. and Galdón Salvador, J.L. (2023) 'Analyzing employee attrition using explainable AI for strategic HR decision-making', Mathematics, 11(22), p. 4677. doi:10.3390/math11224677 Available at https://core.ac.uk/download/pdf/234697248.pdf [Acceesed on 27th Dec 2023].

- Mozaffari, F. et al. (2022) Employee attrition prediction in a pharmaceutical company using both machine learning approach and qualitative data, Benchmarking: An International Journal. Available at: https://www.emerald.com/insight/content/doi/10.1108/BIJ-11-2021-0664/full/html (Accessed: 27 December 2023).

- Ozdemir, S. (2022a) Feature engineering bookcamp. Shelter Island, NY: Manning Publications Co.

- Patway, M.J. et al. (2021) 'Bank deposit prediction using ensemble learning', Artificial Intelligence Evolution, pp. 42–51. doi:10.37256/aie.222021880 Available at https://ojs.wiserpub.com/index.php/AIE/article/view/880/591 [Accessed on 7th Jan 2023].

- Raza, A. et al. (2022) 'Predicting employee attrition using machine learning approaches', Applied Sciences, 12(13), p. 6424. doi:10.3390/app12136424 Available at https://www.mdpi.com/2076-3417/12/13/6424 [Accessed on 27th Dec 2023].

- Rony, M.A. et al. (2021) 'Identifying long-term deposit customers: A machine learning approach', 2021 2nd International Informatics and Software Engineering Conference (IISEC) [Preprint]. doi:Available at : https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9672452 [Accessed on 3rd jan 2024)].

- Sarmento, A.G. et al. (2021) ‘Applying big data analytics in DDOS forensics: Challenges and opportunities’, Cybersecurity, Privacy and Freedom Protection in the Connected World, pp. 235–252. doi:Available at : https://link.springer.com/chapter/10.1007/978-3-030-68534-8_15 [Accessed on 6th January 2024]].

- Shankar, R.S. et al. (2018) ‘Prediction of employee attrition using Datamining’, 2018 ieee international conference on system, computation, automation and networking (icscan) [Preprint]. doi:10.1109/icscan.2018.8541242 Available at https://www.citethisforme.com/cite/sources/journalautociteconfirm [Accessed on 27th Dec 2023].

- Tahir, N.M. et al. (2006) ‘Feature selection for classification using Decision Tree’, 2006 4th Student Conference on Research and Development [Preprint]. doi:10.1109/scored.2006.4339317 Available at https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4339317 [Accessed on 2nd Jan 2023].

- Zorman, M. et al. (1997) The limitations of decision trees and automatic learning in Real World Medical Decision making, Journal of medical systems. Available at: https://pubmed.ncbi.nlm.nih.gov/9555627/ (Accessed:07 January 2024).

- Zheng, A. and Casari, A. (2018) Feature Engineering for Machine Learning: Principles and techniques for Data scientists. Beijing: O’Reilly.

6.0 Appendix

Meeting 1: December 27, 2023 (All members present)
The primary objective of the initial meeting was to foster acquaintance among team members. Identified and outlined the activities essential for accomplishing the assignment.

Meeting 2: December 29, 2023 (All members present)
Focused on formulating the comprehensive process for model building. Emphasized understanding the methodology section with careful attention to detail.

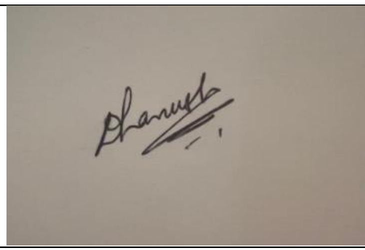
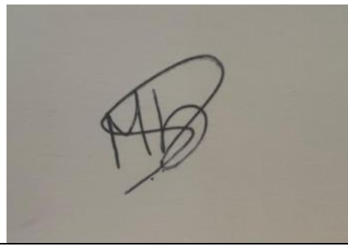
Meeting 3: December 30, 2023 (All members present)
Central focus on familiarizing the team with the KNIME software and determining the requisite nodes for executing various model building tasks, including EDA, Data Preprocessing, and Feature Engineering.

Meeting 4: January 2, 2024 (All members present)
Concentrated on assigning written work for the report and delineating tasks concerning model building on KNIME software. Tasks executed by team members:

- **Dhanush**: Model Building, Scope and Overview, Built Models using KNIME.
- **Rohan**: Methodology, Model Building using KNIME, Results and Discussion.
- **Manogna**: Literature Review, EDA, Model Building, Models on KNIME.
- **Mrunmayee**: Abstract, Literature Review, Results and Discussion, EDA.
- **Cindrella**: Literature Review, Scope and Overview, EDA, Results and Discussion.

Signed Declaration:

We all hereby declare that we have coordinated and completed the group assignment for Human Resource Analytics submitted to Queens' University Belfast, the tasks that were completed and submitted are as mentioned in the activity report above.

Name	Signature
Cindrella KC - 40429497	
Dhanush MS - 40412492	
Mrunmayee Bapat - 40420299	
Manogna B R - 40426970	
Rohan Mahesh Patil - 40395741