Group Coursework Submission Form

Specialist Masters Programme

Please list all names of group members: 1. Verma, Shubham: 230021430 2. Rawat, Nikhil: 230044335 3. Amarnani, Pooja: 230022030 4. Sista, Sai Srimanth - 230045961	**GROUP NUMBER:**	**15**

MSc in: BUSINESS ANALYTICS

Module Code: SMM636

Module Title: Machine Learning

Lecturer: Dr Rui, Zhu	**Submission Date: 04/Mar/2024**

Declaration:

By submitting this work, we declare that this work is entirely our own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work we acknowledge that we have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. We also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days of lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked online):

Deduction for Late Submission:

Final Mark: %

Task of the Dataset

"Enhancing Workforce Stability: A Predictive Approach to Employee Attrition"

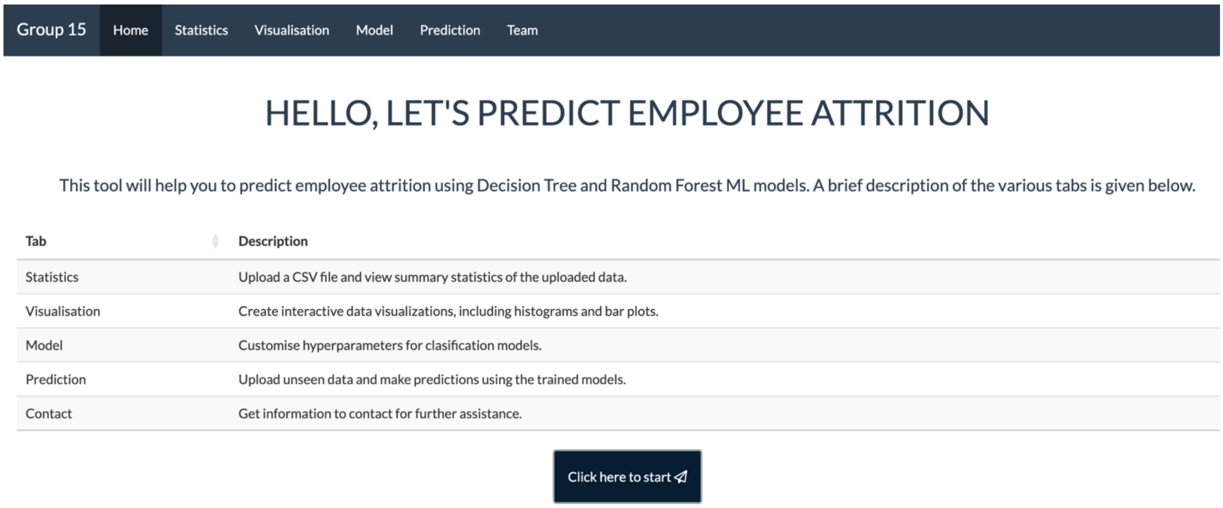
Employee attrition presents a critical challenge for businesses, leading to financial losses and operational instability. Addressing this, we aim to offer HR a robust predictive tool, utilizing classification models on employee data to anticipate departures and facilitate timely HR interventions for improved organizational efficiency.

Application Design & Functionality

Design

The user interface for our R Shiny application is structured into tabs for an intuitive user experience. Starting with a welcoming "Home" tab, users are introduced to our employee attrition prediction tool using Decision Tree and Random Forest models.
Subsequent tabs include:

- Statistics: Easily upload a CSV file, handpick variables, and gain insights through summary statistics.
- Visualisation: Interactively explore data by dynamically selecting variables. Generate histogram and bar charts based on variable type, with adjustable histogram bins.
- Model: Presents a model summary, offering users the flexibility to choose between Random Forest (RF) and Decision Tree (DT) models. Customise hyperparameters such as minimum number of trees, number of variables at each split for RF and minimum sample for split, and complexity for DT with results being displayed dynamically.
- Prediction: Select a model and predict employee attrition on your dataset and export results in a table format for further analysis.
- Team: Find contact information of the development for user support.



HELLO, LET'S PREDICT EMPLOYEE ATTRITION

This tool will help you to predict employee attrition using Decision Tree and Random Forest ML models. A brief description of the various tabs is given below.

Tab	Description
Statistics	Upload a CSV file and view summary statistics of the uploaded data.
Visualisation	Create interactive data visualizations, including histograms and bar plots.
Model	Customise hyperparameters for clasification models.
Prediction	Upload unseen data and make predictions using the trained models.
Contact	Get information to contact for further assistance.

Figure 1. Landing page of the application with description of various tabs, complemented by an action button to navigate to the subsequent tabs.

Functionality

Dataset

In this project, we have used the Employee Attrition dataset of an organisation. The dataset has 487 observations and 34 features. The dataset has both continuous and categorical variables such as age, gender, working years, salary hike, job satisfaction etc. The target variable, labelled "Attrition," is binary indicating whether an employee will leave the organisation or not.

Data pre-processing

Several data cleaning steps were undertaken, including the examination of null values and outliers, feature selection, encoding of categorical features, and analysis of the value counts of the target variable to assess data balance.

Following pre-processing, no null values were detected, redundant features were eliminated based on correlation analysis with the target variable, and the dataset exhibited a balanced distribution, with approximately 51% of observations indicating attrition.

Model Training and Testing:

Utilising the processed dataset derived from the previous step, featuring 487 rows and 30 refined features, we conducted a split, allocating 80% of the data for training and the remaining 20% for testing. Decision tree and random forest models were trained using the training dataset. The application gives the user a further ability to tune the hyperparameter of the classification models. Specifically, for the DT model, users can optimise the minimum number of samples for split and the complexity parameter. In the case of the RF model, users can fine-tune parameters such as the number of variables at each split and the minimum number of trees. This approach ensures a personalised modelling experience, empowering users to enhance the predictive capabilities of both models based on their specific requirements.

Tuning hyperparameters is a crucial step in model development, as it helps strike a balance between underfitting and overfitting. Underfitting occurs when a model is too simple and fails to capture the underlying patterns in the data, whereas, overfitting happens when a model is too complex, essentially memorising the training data but performing poorly on new, unseen data. By allowing users to adjust hyperparameters, the application enables a personalized modelling experience, empowering users to enhance the predictive capabilities of both models based on their specific requirements while mitigating the risks of underfitting or overfitting.

Following the training process, predictions are made on the testing set. The application provides a detailed performance evaluation, presenting a confusion matrix along with metrics such as Precision, Recall, F1 score, and Accuracy.

Predictions on User-Uploaded Data:

Users can upload new, unlabelled employee data via the application and use the pre-trained model from the previous step to generate predictions. A table of these predictions can be exported for further review if needed. Figure 2 presents a detailed flowchart, guiding users through the entire process, from application launch to prediction.

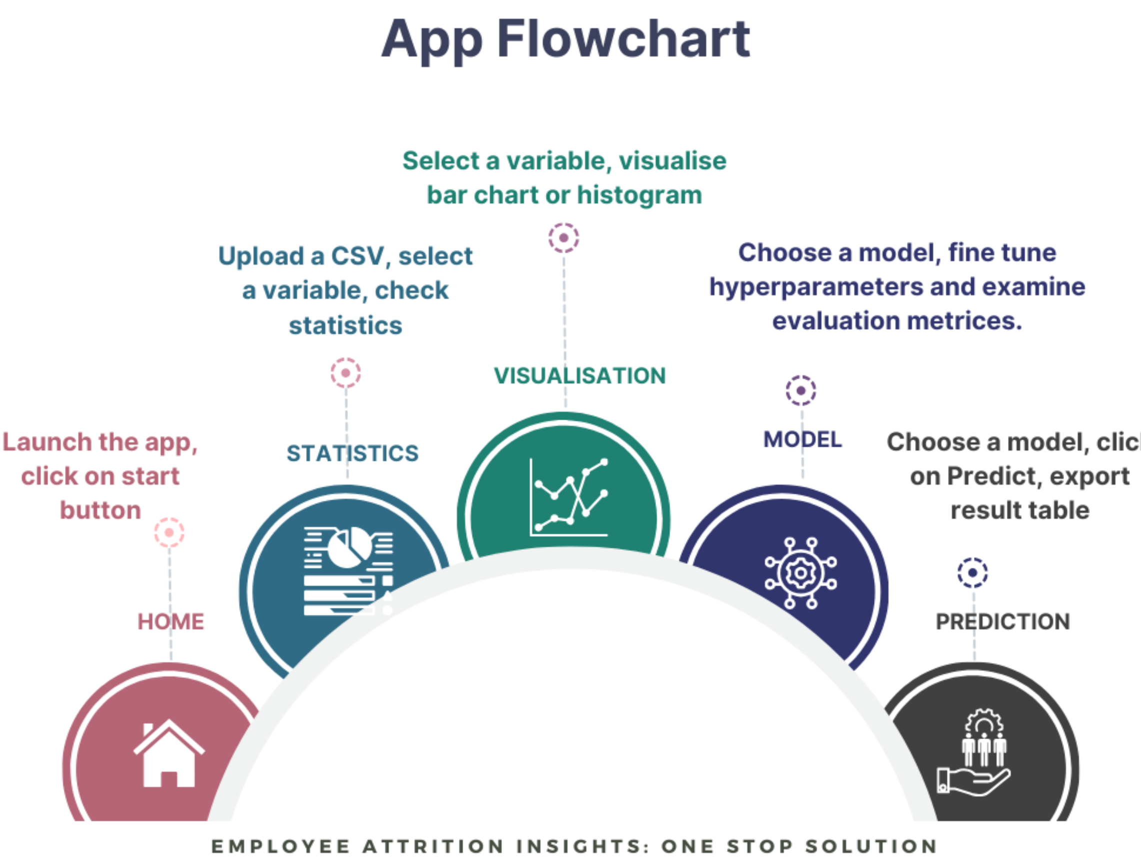

App Flowchart
Select a variable, visualise bar chart or histogram
Upload a CSV, select a variable, check statistics
Choose a model, fine tune hyperparameters and examine evaluation metrices.
VISUALISATION
Launch the app, click on start button
STATISTICS
MODEL
Choose a model, click on Predict, export result table
HOME
PREDICTION
EMPLOYEE ATTRITION INSIGHTS: ONE STOP SOLUTION

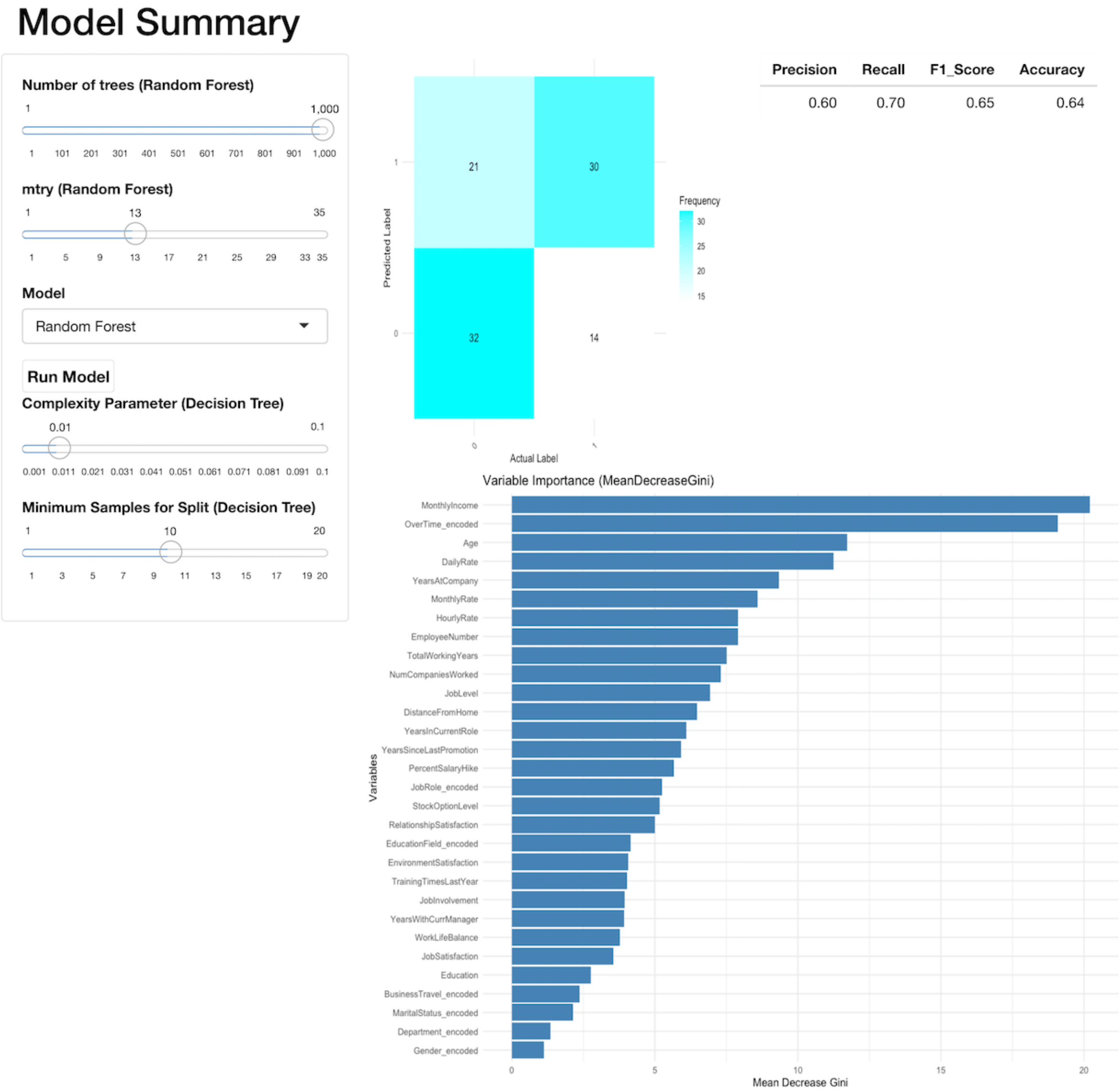
Figure 2. Navigating the Employee Attrition Insights App: From Data Upload to Predictive Analysis

Analysis of the Results

For analysis, both Random Forest and Decision Tree classifiers were employed to predict employee attrition. RF, with its ensemble approach, tends to offer higher accuracy due to multiple decision trees reducing variance. The DT provides easy-to-follow logical paths, making its predictions highly interpretable. Both classifiers were evaluated based on precision, recall, and F1 scores, with Random Forest performing slightly better overall. For Random Forest, 'mtry' and the 'number of trees' influence the robustness and accuracy of the model, with a higher number of trees generally providing more stable predictions. The Decision Tree's 'complexity parameter' and 'minimum samples for split' affect the depth and overfitting potential of the model. Key variables impacting attrition, such as 'MonthlyIncome' and 'Overtime', were identified by both models.

The analysis of the two classifiers for predicting employee attrition reveals distinct performance metrics, for example:

- In Figure 3, the Random Forest classifier, with a slightly lower accuracy of 0.64, has a precision of 0.60 and recall of 0.70, indicating it is slightly less precise but comparable in detecting actual attrition instances.



Model Summary

Number of trees (Random Forest): 1,000

mtry (Random Forest): 13

Model: Random Forest

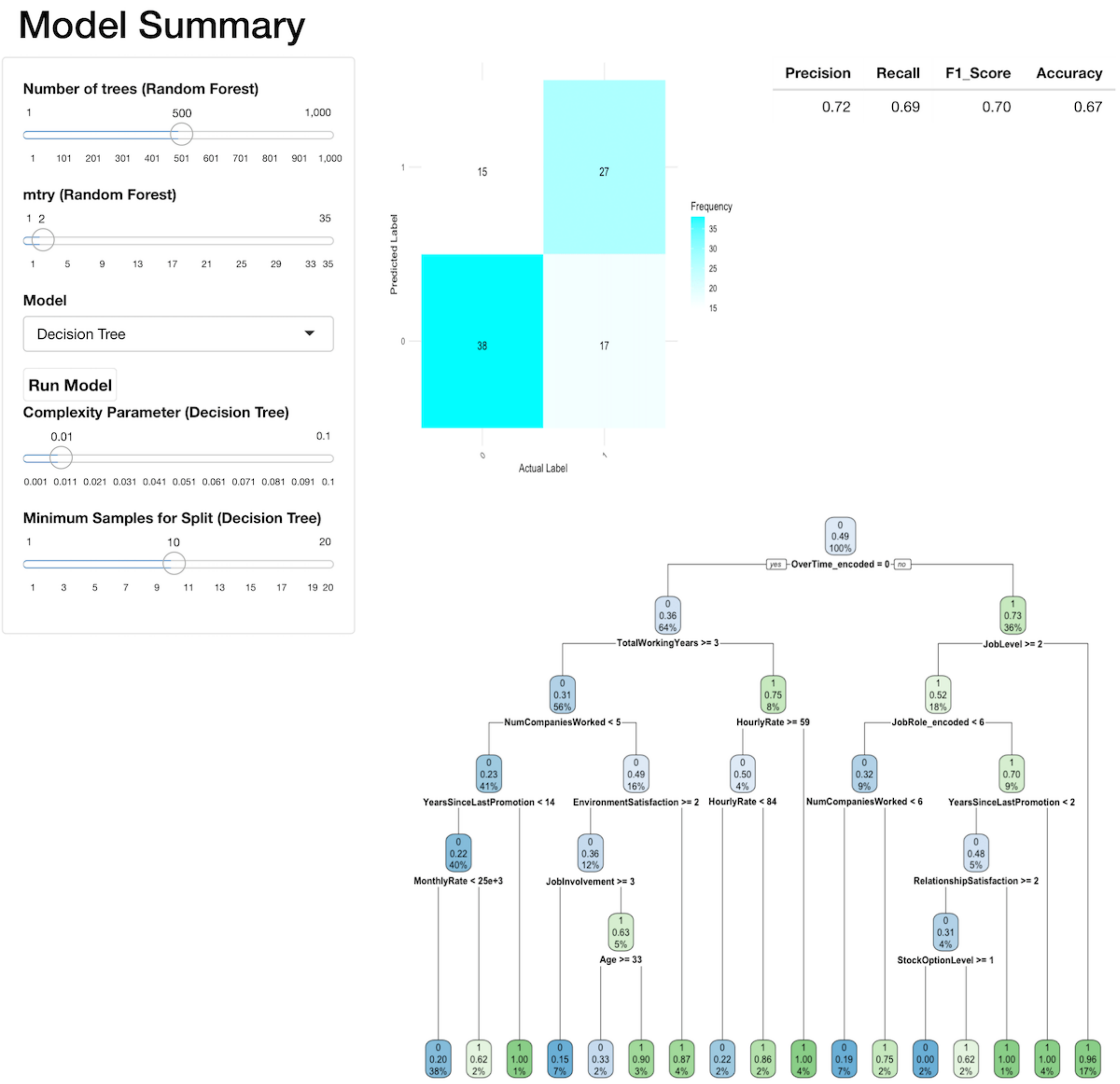
Complexity Parameter (Decision Tree): 0.01

Minimum Samples for Split (Decision Tree): 10

Precision	Recall	F1_Score	Accuracy
0.60	0.70	0.65	0.64

Figure 3: Random Forest

- In Figure 4, the Decision Tree classifier shows an accuracy of 0.67 with a precision of 0.72, indicating a reliable prediction when an employee is at risk of leaving. The recall rate of 0.69 suggests the model's robustness in identifying true attrition cases.



Model Summary

Number of trees (Random Forest): 500

mtry (Random Forest): 2

Model: Decision Tree

Run Model

Complexity Parameter (Decision Tree): 0.01

Minimum Samples for Split (Decision Tree): 10

Precision	Recall	F1_Score	Accuracy
0.72	0.69	0.70	0.67

Figure 4: Decision Tree

The F1 scores are 0.70 for Random Forest and 0.65 for Decision Tree, reflecting a balanced precision-recall trade-off. Both models highlight 'MonthlyIncome' and 'OverTime' as top features influencing attrition. The Decision Tree's structure showcases the decision rules used, providing transparent criteria that led to its predictions, which can be crucial for interpretability in HR decision-making.

Conclusion:

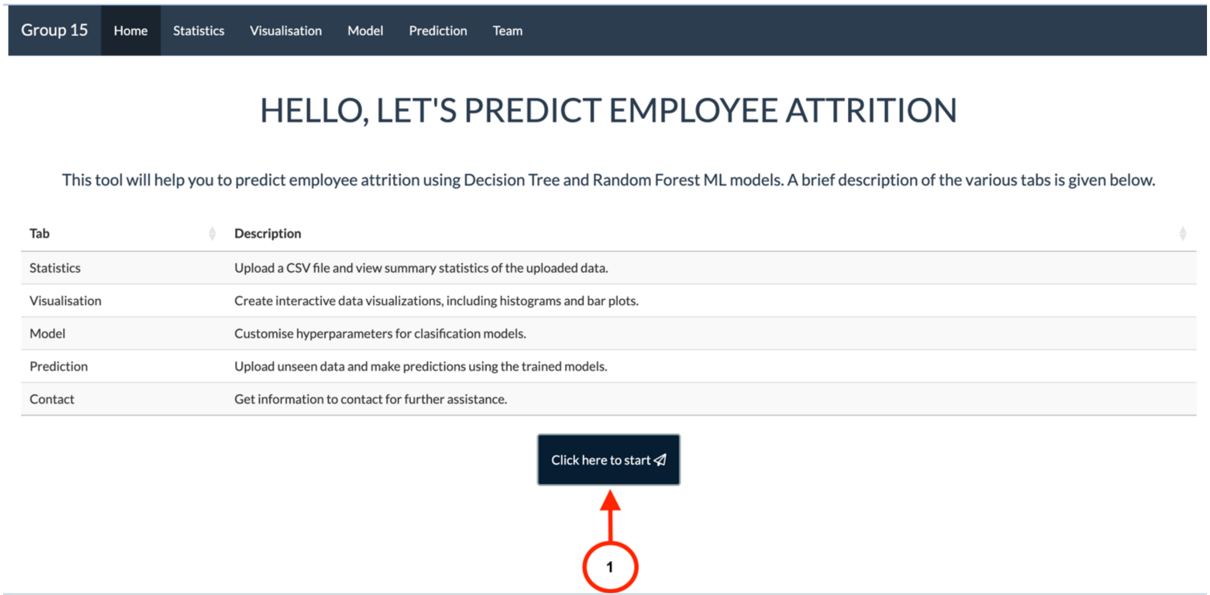
This app can significantly aid HR departments in predicting employee attrition, which is crucial for managing workforce stability and planning. By utilizing machine learning models, the app analyses employee data to identify patterns and factors contributing to attrition. This predictive capability allows HR to intervene proactively, implement retention strategies, and make informed decisions regarding recruitment, training, and employee engagement. It offers insights into the variables affecting attrition, helping HR to tailor their strategies to individual needs and organizational goals, ultimately reducing turnover and its associated costs.

Number of words: 1050

Appendix

Please refer to the below manual for using the application.

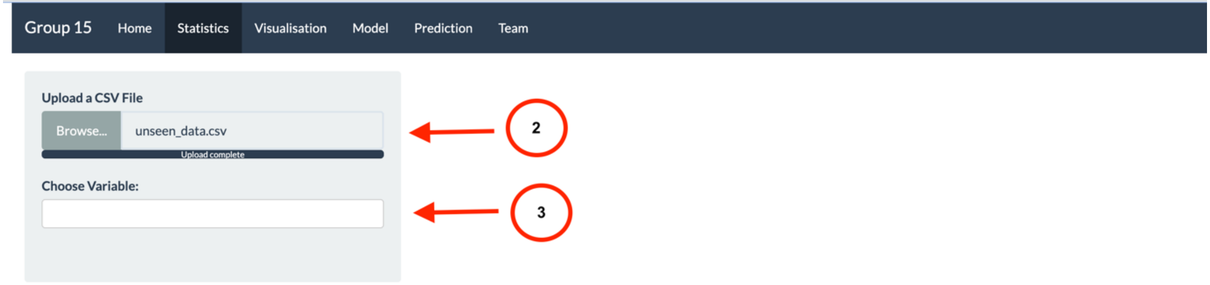
Step 1: Begin by clicking the designated button to start the journey.
Step 2: Browse and select the data file named "unseen_data.csv" from your computer.
Step 3: Choose the variables you wish to analyse to view their statistics.
Step 4: Navigate to the visualization tab and select variables to observe their trends.
Step 5: Adjust parameters according to your visualization preferences.
Step 6: In the Model tab, select from two available model types.
Step 7: Execute the chosen model.
Step 8: Fine-tune parameters specific to the selected model.
Step 9: In the prediction tab, select the desired model type.
Step 10: Click on the "predict" button to generate the report.
Step 11: Utilize the "export" function to download the file.
Step 12: Interpret predictions on the extreme right side: 0 signifies "No" while 1 signifies "Yes".



HELLO, LET'S PREDICT EMPLOYEE ATTRITION

This tool will help you to predict employee attrition using Decision Tree and Random Forest ML models. A brief description of the various tabs is given below.

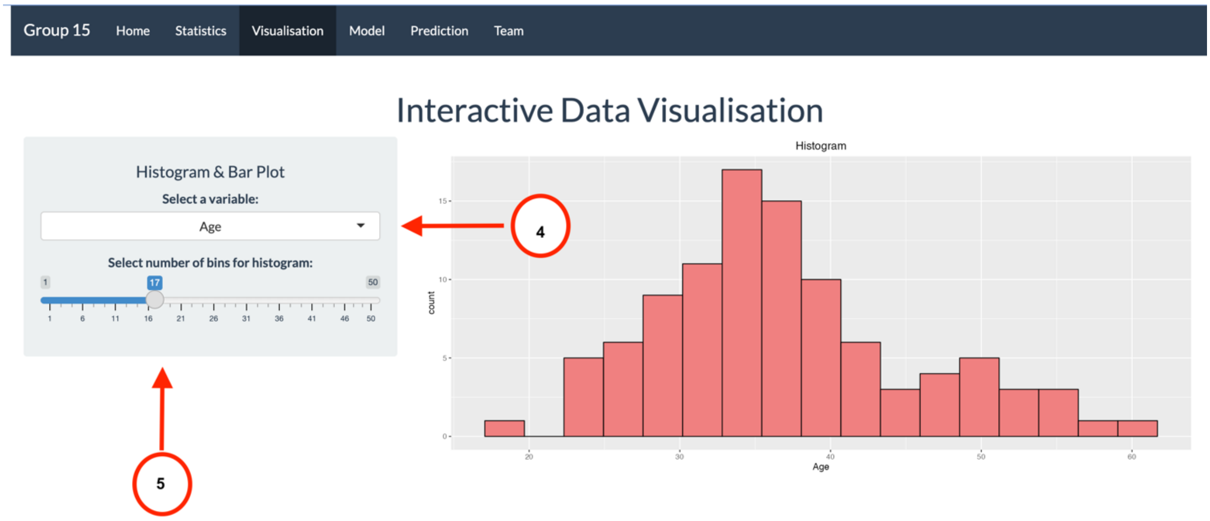
Tab	Description
Statistics	Upload a CSV file and view summary statistics of the uploaded data.
Visualisation	Create interactive data visualizations, including histograms and bar plots.
Model	Customise hyperparameters for clasification models.
Prediction	Upload unseen data and make predictions using the trained models.
Contact	Get information to contact for further assistance.



Upload a CSV File

unseen_data.csv

Choose Variable:

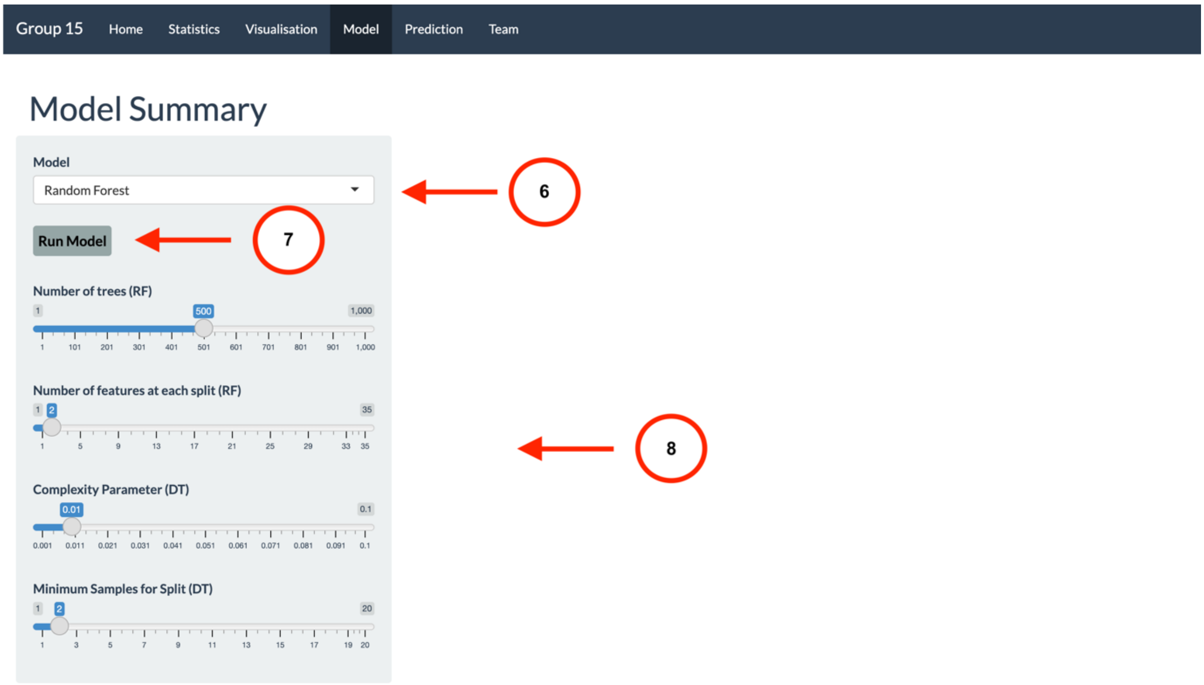


Interactive Data Visualisation

Histogram & Bar Plot

Select a variable: Age

Select number of bins for histogram: 17



Model Summary

Model: Random Forest

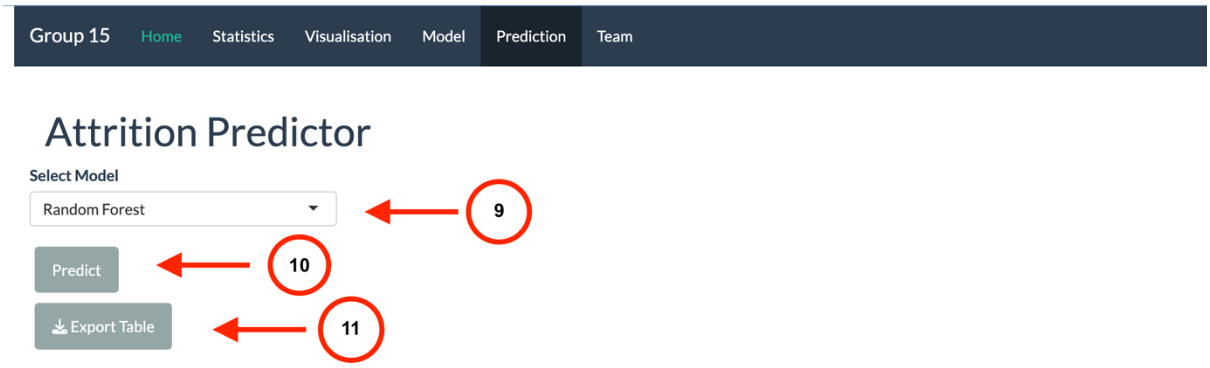
Run Model

Number of trees (RF): 500 (1 – 1,000)

Number of features at each split (RF): 2 (1 – 35)

Complexity Parameter (DT): 0.01 (0.001 – 0.1)

Minimum Samples for Split (DT): 2 (1 – 20)

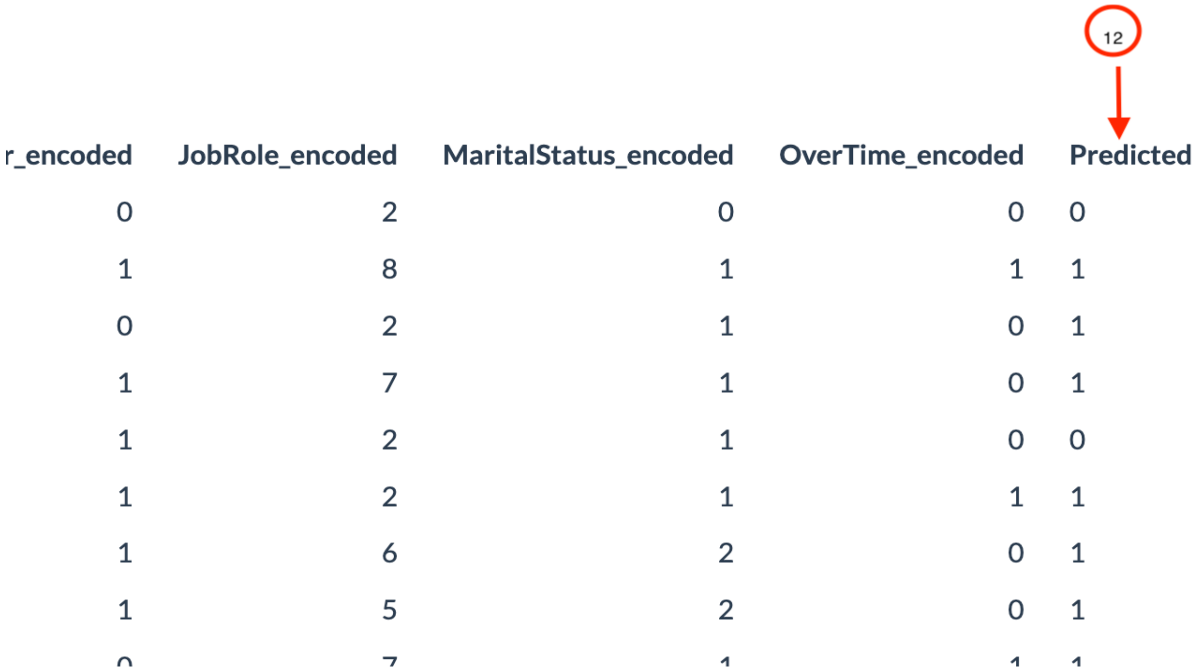


Attrition Predictor

Select Model: Random Forest

Predict

Export Table



r_encoded	JobRole_encoded	MaritalStatus_encoded	OverTime_encoded	Predicted
0	2	0	0	0
1	8	1	1	1
0	2	1	0	1
1	7	1	0	1
1	2	1	0	0
1	2	1	1	1
1	6	2	0	1
1	5	2	0	1
0	7	1	1	1