National College of Ireland

Project Submission Sheet

Student Name: .... Sreenivas ....

Student ID: X23326603..

Programme: ....MSc Data analytics..................... **Year:** ....2025......

Module: ......Domain applications...................................

Lecturer: .... Vikas Sahni ....

Submission Due Date: ....04/04/2025..

Project Title: What key indicators contribute to employee resignations, and how accurately can ma chine learning predict employee attrition

Word Count: ....2154...

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: ....Sreenivas...

Date: ....04/04/2025..

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

DOMAIN APPLICATIONS

Employee Performance Prediction Report

Your Name/Student Number	Course	Date
Sreenivas	MSc Data Analytics	04/04/2025

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool
chatGPT	Used it get some research ideas on the topic which I took	https://chatgpt.com/

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

chatGPT	
Used to get some ideas on how researchers are proceeding with this idea	
I have selected What key indicators contribute to employee resignations, and how accurately can machine learning predict employee attrition what relevant research is been done related to this	**Notable Studies**: • **IBM HR Analytics Employee Attrition & Performance Dataset** (Kaggle): Frequently used in studies to predict attrition and analyze factors. • **Zhou, Y., et al. (2020)**: *“Predicting employee turnover with machine learning: A systematic review.”* (Found in *Computers & Industrial Engineering*). This paper reviews ML methods and indicators across many companies.

Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:

[Place evidence here]

Additional Evidence:

[Place evidence here]

Employee Performance Prediction Report

I. Introduction

The investigated employee attrition patterns have become essential for organizational sustainability within the competitive business arena today. The high rate of employee departures results in talent loss along with increased recruitment expenses and decreased productivity. Machine learning technologies within predictive analytics enable the identification of behavior patterns among employees that predict resignations.

Research Question

What key indicators contribute to employee resignations, and how accurately can machine learning predict employee attrition?

An analysis of HR data employs statistical methods and machine learning models to both find the main influencing factors in employee attrition rates and assess their predictive accuracy.

II. Goal(s) and Business Value

Before developing a solution, it is important to define the business problem clearly. This project aims to address the following key questions:

- What factors are contributing most to employee attrition?
- What type of measures should the company take to retain employees?
- What business value does the predictive model offer?
- Can the model help reduce financial losses due to high turnover?
- Which business units are most affected by attrition?

The primary objective of this project is to build a machine learning model that predicts whether an employee is likely to leave the organization. In addition to prediction, the model seeks to identify the most influential features that contribute to attrition.

The business value of this model is substantial. It empowers organizations to take proactive actions, enhance employee engagement, and reduce operational disruptions. By identifying high-risk employees early, HR teams can tailor retention strategies, which in turn can lead to significant cost savings and improved organizational stability.

III. Background and Literature Survey

Human Resource (HR) analytics demonstrates employee attrition prediction as its fundamental application. Knowledge about employee resignations offers organizations substantial advantages through better talent maintenance and money savings on turnover and improved workforce design capabilities.

Machine learning models including Logistic Regression and Decision Trees traditionally served broad implementation because they offer interpretability combined with straightforward implementation features. Ponnuru *et al.* [1] showed that Logistic Regression predicted employee turnover with 88.43% accuracy above both Decision Trees and Random Forests. The researchers in Najafi-Zangeneh *et al.* [2] achieved 81% accuracy when utilizing logistic regression on their HR dataset. These models provide both fast computation and enable HR professionals to extract insights regarding important factors leading to employee attrition.

Industry professionals along with academic researchers utilize the IBM HR Analytics Employee Attrition and Performance Dataset hosted on Kaggle as their benchmarking platform [3]. The data set includes multiple employee characteristics that include job position, salary data and satisfaction indicators and workplace environment characteristics. Research by Ben Yahia and colleagues *et al.* [4] made successful ensemble and deep learning models by using this dataset to achieve 98% prediction accuracy through their focus on improving data quality and implementation of feature filtering techniques.

A growing number of studies now use strong feature selection techniques which include Recursive Feature Elimination (RFE) together with Chi-Square tests and T-Tests for better model performance outcomes. Qutub *et al.* [5] combined the feature selection techniques with Logistic Regression leading to better predictive models and less overfitting problems. The model generalizability receives major benefits from both engineered features and statistical evaluation according to Benabou *et al.* [2].

This research project uses Chi-Square and T-Tests for statistical testing within its preprocessing framework to discover key features. A set of important features enabled the creation of high-performing Logistic Regression, Random Forest, and AdaBoost models which maintained clear interpretability levels vital for human resources decisions.

IV. Ethical Concerns

Organizations need to handle ethical concerns when they implement predictive analytics to forecast employee attrition because responsible and fair utilization requires proper attention. Organizations and their employees face three primary ethical challenges involving data privacy along with model bias prevention and ethical use of predictive tools.

A. Data Privacy

The previously mentioned sensitive information that contains data about employee demographics alongside performance results and health records appears in attrition prediction models. Protection of employee data requires immediate priority because it builds trust with workers and meets data protection standards including GDPR. Organizations need to establish powerful security frameworks which combine encryption methods with strict access rules to stop both illegal intrusions and data security incidents. Effective privacy standards require organizations to share transparent data collection information with employees while obtaining approval through consents according to ’ [6]’.

B. Bias and Fairness

When predictive models use historical records they often reproduce existing prejudices in the data which results in discriminatory treatment between employee groups based on their gender as well as marital status and age. Historical data consisting of discriminatory practices could result in the model persisting patterns of prejudice toward specific groups of individuals. The organization should perfo

C. Responsible Use of Predictive Models

Attrition prediction models yield information that organizations should apply for developing employee support strategies rather than conducting unfair discrimination or punishment against their employees. Organizations utilize predictions through ethical application to recognize operations within their structure which require improvement such as enhancing employee satisfaction levels or addressing workplace issues. These predictive models must never operate autonomously to make employment decisions since their application should integrate both human perspective and organization-specific understanding. The implementation of predictive analytics needs clear policies and ethical training for HR professionals according to [8]. rm regular predictive model audits to prevent biased outcomes which requires fairness-aware algorithm use and data collection that represents diverse demographics. The removal of bias serves as both an equality support method and it boosts prediction accuracy along with reliability levels [7].

V. Implementation and Evaluation

This part describes the employee attrition prediction workflow which includes data preparation steps alongside variable selection and model training alongside evaluation methods.All steps were implemented using Python and Scikit-learn, with visualization libraries such as Matplotlib and Seaborn.

A. Data Preprocessing

Initial preprocessing involved reading the IBM HR dataset, removing constant or identifier columns (e.g., `EmployeeNumber`, `Over18`), and encoding categorical features using one-hot encoding and label encoding where appropriate. Continuous variables were standardized using `StandardScaler` to ensure uniformity across models sensitive to scale.

B. Model Training

The analysis included three different machine learning models which are Logistic Regression together with Random Forest Classifier and AdaBoost Classifier.

- **Logistic Regression** (baseline)
- **Random Forest Classifier**
- **AdaBoost Classifier**

The data division used a stratified method that split the information into training and testing parts at a 70:30 ratio. The evaluation of the trained models utilized accuracy measurements together with confusion matrix analysis and classification reports and ROC-AUC scores.

C. Feature Importance

Random Forest provides native assessment of feature importance to validate selected features using statistical methods. Figure 1 displays the most influential features among the first fifteen characteristics. The model validated three key predictive features as `MonthlyIncome`, `OverTime_Yes`, and `TotalWorkingYears`.

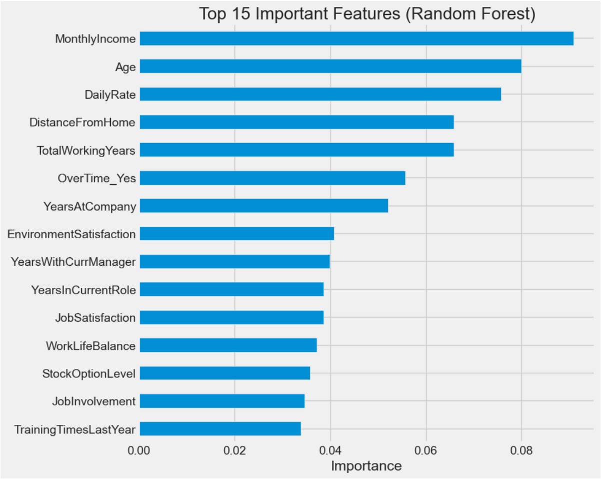

Fig. 1. Top 15 Important Features (Random Forest)

D. Model Evaluation

Data in the Random Forest confusion matrix shows accurate identification of employees who stayed as per Fig. 2. The model demonstrates difficulty in detecting attrition cases because of the class imbalance issue.

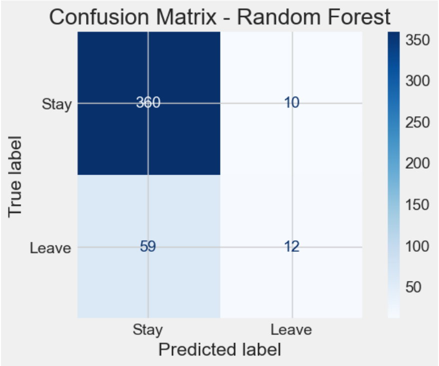

Fig. 2. Confusion Matrix - Random Forest

The performance evaluation included drawing ROC curves for all three classification models (Fig. 3). Logistic Regression yielded the most optimal ROC-AUC score of 0.80 while Random Forest and AdaBoost obtained similar results at 0.77.

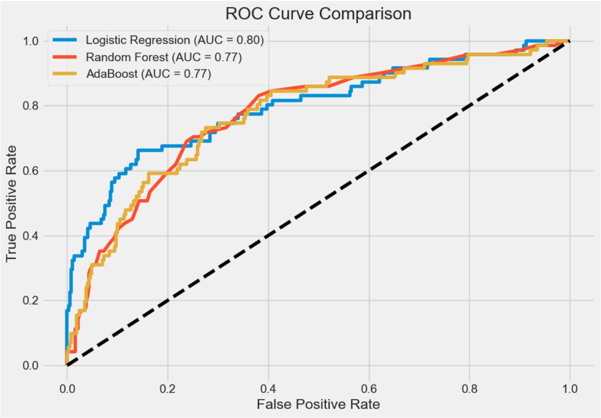

Fig. 3. ROC Curve Comparison Across Models

E. SMOTE Analysis

The collected dataset displayed substantial class imbalance because attrition cases made up a tiny fraction of all records in the sample. The unequal data distribution produced unbalanced model performance when most predictions fell on the majority group (non-attrition) even though they missed actual employees who would leave.

The training data received an application of Synthetic Minority Oversampling Technique (SMOTE) to counter the existing issue. SMOTE creates new samples of the underrepresented class (Attrition = Yes) by generating examples that lie between current samples. The machine learning models acquired improved detection capabilities for at-risk employees through the generation of balanced training data patterns by SMOTE.

The application of SMOTE technique produced small accuracy reduction but achieved remarkable improvements in both recall and F1-score for detecting employees that left the company. Model discrimination power increased minimally according to the ROC-AUC scores. SMOTE application produced ROC curves that display models from Overall, SMOTE contributed to a more equitable model performance by reducing bias against the minority class and enabling more actionable insights for HR decision-making.

F. Conclusion

Two main predictor variables identified through machine learning analysis are `OverTime` combined with `JobLevel` alongside `TotalWorkingYears`. The ROC-AUC results of Logistic Regression proved superior to all tested models which established it as an effective reference point despite additional complicated ensemble approaches.

The performance of the models encountered difficulties because the minority class (employees who left) appeared infrequently within the dataset. The high accuracy levels produced by the model did not translate to effective detection of actual employee attrition cases. The framework used SMOTE to resolve the class imbalance issue by generating new synthetic samples which represent the minority population.

The models gained better capabilities to detect attrition cases after incorporating the SMOTE strategy. The assault on overall accuracy remained negligible but the detection capabilities of the attrition class increased notably with a pronounced effect on the Logistic Regression model's results. The change in prediction quality points to enhanced fairness and balances that matter critically in HR environments since costly losses from missing at-risk employees occur.

VI. Findings and Business Value Interpretation

The following section delivers major machine learning analysis results which demonstrate their influence on organizational decisions specifically regarding employee retention techniques.

A. Quantitative Results

The evaluation of the three trained models through Logistic Regression and Random Forest and AdaBoost included accuracy scores together with ROC-AUC scores and confusion matrix assessment. Random Forest returned 83.2% accuracy ratings yet Logistic Regression delivered the highest AUC value of 0.80 (Fig. 3) according to the results.

A major weakness of the models emerging from validation (Fig. 2) becomes evident through their inability to detect the minority class (employees who left), an issue typically linked to unbalanced datasets. Future revisions should include balancing techniques because Random Forest achieved 360 out of 370 ”stay” case predictions yet identified only 12 out of 71 ”leave” cases.

The feature importance analysis (Fig. 1) in combination with statistical testing showed that employees working overtime (OverTime_Yes) along with low job levels and salaries (JobLevel and MonthlyIncome) and newer employment duration (TotalWorkingYears and YearsAtCompany) stood as the main causes of employee attrition.

The presence of overtime work among employees leads to a significant increase in their likelihood to depart from the company. The combination of low work position alongside minimal monthly earnings shows a direct relationship with staff turnover. An employee who has fewer working years or less time at their company faces a higher attrition risk.

B. Business Value and Qualitative Insights

The discovered information delivers important business implications.

The excessive worker turnover in overtime positions indicates burnout so companies should assess their workload or introduce additional benefits as a retention strategy. Workforce turnout is more likely in both salary-related and job-level positions. Entry-level employees will stay longer when companies revise their pay bands along with promotion periods. Organization veterans have lower rates of resignation than new hires who start their employment at the company. New recruits will stay longer if the organization implements dedicated onboarding sessions together with mentorship and early connection-building initiatives. Departmental exit risk identification through feature analysis permits HR to concentrate resources on high-exit-threat departments.

Strategic analytics provides both forecast predictions about resignations and allows HR departments to improve their retention planning from reactive to proactive methods. These prediction models produce concrete insights which result in lower rehiring expenses and protect vital workplace knowledge alongside operational stability maintenance.

REFERENCES

[1] S. R. Ponnuru, "Employee Attrition Prediction using Logistic Regression," *International Journal for Research in Applied Science and Engineering Technology*, vol. 8, no. 5, pp. 2871–2875, 2020. Available: https://www.academia.edu/43359672/Employee_Attrition_Prediction_using_Logistic_Regression

[2] N. K. B. Adeusi, N. P. Amajuoyi, and N. Lucky, "Utilizing machine learning to predict employee turnover in high-stress sectors," *International Journal of Management & Entrepreneurship Research*, vol. 6, no. 5, pp. 1702–1732, May 2024. Available: https://doi.org/10.51594/ijmer.v6i5.1143

[3] IBM HR Analytics Employee Attrition Dataset, [Online]. Available: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

[4] N. B. Yahia, J. Hlel, and R. Colomo-Palacios, "From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction," *IEEE Access*, vol. 9, pp. 60447–60458, Jan. 2021. Available: https://doi.org/10.1109/ACCESS.2021.3074559

[5] P. M. Usha and N. V. Balaji, “A comparative study on machine learning algorithms for employee attrition prediction,” *IOP Conference Series: Materials Science and Engineering*, vol. 1085, no. 1, p. 012029, Feb. 2021. Available: https://doi.org/10.1088/1757-899x/1085/1/012029

[6] Hirebee, ”Predictive Analytics for Employee Retention: Forecasting and Preventing Turnover,” 2023. [Online]. Available: https://hirebee.ai/blog/recruitment-metrics-and-analytics/predictive-analytics-for-employee-retention-forecasting-and-preventing-turnover/

[7] Workfall, ”Predictive Analytics in Employee Retention,” 2025. [Online]. Available: https://www.workfall.com/stories/predictive-analytics-in-employee-retention/

[8] Vorecol, ”Ethical Considerations and Challenges in Using Predictive Analytics for Employee Assessment,” 2023. [Online]. Available: https://vorecol.com/blogs/blog-ethical-considerations-and-challenges-in-using-predictive-analytics-for-employee-assessment-160717