National College of Ireland

Predicting Employee Attrition with a Comprehensive Machine Learning Approach: Utilizing Ensemble Methods and Hyperparameter Optimization

MSc Research Project
Data Analytics

Sreejith Sridhar
Student ID: X22242376

School of Computing
National College of Ireland

Supervisor: Prof. Christian Horn

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sreejith Sridhar		
Student ID:	X22242376		
Programme:	MSc. Data Analytics	**Year:**	2024
Module:	MSc. Research Project		
Supervisor:	Prof. Christian Horn		
Submission Due Date:	12/08/2024		
Project Title:	Predicting Employee Attrition with a Comprehensive Machine Learning Approach: Utilizing Ensemble Methods and Hyperparameter Optimization		
Word Count:	7458	**Page Count:**	21

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 bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Sreejith Sridhar
Date:	12TH August 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

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

Table of Contents

Predicting Employee Attrition with a Comprehensive Machine Learning Approach: Utilizing Ensemble Methods and Hyperparameter Optimization

Sreejith Sridhar
X22242376

Abstract

Employee attrition poses significant challenges for organizations, leading to high recruitment costs, the loss of key personnel, and diminished employee morale. Traditional attrition prediction approaches like surveys and interviews lack the data and timeliness needed for effective intervention. This work develops an effective employee turnover prediction model using machine learning to overcome these constraints. The study seeks to discover attrition factors and create a reliable forecasting model. The CRISP-DM structure guides the research from business needs to data preparation. The dataset was retrieved from Kaggle and thoroughly pre-processed to remove duplicates, encode category variables, and normalise numerical characteristics. The Research tested Logistic Regression, Decision Tree, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Random Forest, and AdaBoost. Ensemble approaches, especially the Voting Classifier, improved prediction accuracy. GridSearchCV hyperparameter adjustment improved model performance. The optimised Voting Classifier surpassed others in accuracy, precision, recall, F1-score, and ROC AUC, scoring Accuracy of 85.03% and ROC score of 0.845. Overtime, experience, stock option level, and distance from home greatly affect attrition, according to feature importance analysis. Using Streamlit, this research provides an interactive tool for analysing attrition risk and early intervention alternatives. To improve attrition prediction, advanced feature engineering, time-series analysis, and ensemble model interpretability will be investigated in future work.

Keywords: Employee Attrition, Machine Learning, Prediction Model, CRISP-DM, Feature Importance, Voting Classifier, GridSearchCV, Hyperparameter, Ensemble.

1 Introduction

Employee attrition, commonly known as turnover, poses a significant difficulty for organizations since it necessitates the substitution of current personnel with new recruits. This issue results in higher recruitment expenses, diminished organizational expertise, and reduced employee satisfaction, which has a substantial influence on organizational performance (Margherita, 2022). Conventional approaches like surveys and departure interviews sometimes suffer from inaccuracy and delays, creating challenges for Human Resources departments in implementing preventative measures (George et al., 2022). Hence, it is crucial to create prediction technologies that can precisely anticipate staff loss, enabling organizations to deploy efficient retention measures.

The objective of this study is to utilize machine learning approaches to overcome the restrictions of conventional attrition prediction methodologies. Conventional methods, such as administrative assessments and historical examination, are frequently influenced by personal opinions and do not effectively identify complicated patterns in employee behaviour (George et al., 2022; Krishna et al., 2023). Recent progress in machine learning enables the analysis of extensive datasets, the identification of difficult relationships, and the provision of immediate forecasts. Consequently, it offers a more reliable approach to forecasting employee turnover (Al-Alawi & Ghanem, 2024). While there exists substantial literature on the development of several individual models, very limited work has been proposed on how to enhance the models' performance by efficiently using multiple models in parallel. To fill this gap in the existing literature, this research seeks to explore whether, and how, different machine learning techniques and particularly ensembles, are effective in predicting employee turnover. In addition, the objective is to conduct a thorough examination of the significance of different features to identify the primary elements that contribute to attrition. This aspect has not been extensively investigated in previous research studies.

The primary participants of this initiative include both academic scholars and industrial professionals. This research enhances the current knowledge by conducting a comparative analysis of several machine learning models and their efficacy in predicting employee attrition. From a business point of view, HR departments may employ the results and created models to effectively execute proactive retention tactics. This will result in lower turnover rates, reduced recruiting costs, and the preservation of organizational expertise and morale.

The main research question motivating this project is: **"How can hyperparameter-optimized machine learning models, including ensemble methods like Voting Classifiers, effectively predict employee attrition and identify the most influential features contributing to employee turnover?"**

The main goals of this study are to assess the efficacy of different machine learning models, namely Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, AdaBoost, and ensemble methods, in accurately predicting employee attrition. Additionally, the study aims to identify the crucial factors that influence employee attrition through thorough feature engineering and selection techniques. Furthermore, the study aims to enhance model performance by employing techniques like GridSearchCV and cross-validation. Lastly, the study aims to create an interactive API for HR departments that utilizes the most effective model to predict employee attrition.

This study aligns to the CRISP-DM model, guaranteeing a methodical approach to data mining. The approach encompasses the following procedures: Business Understanding is identifying the issue of staff attrition and its influence on organizational success; Data Understanding and Preparation, this process involves using a dataset obtained from Kaggle[1], carrying out data cleaning, and doing exploratory data analysis (EDA) to get insights into data distributions and correlations. Modelling involves the development and training of several machine learning models, including ensemble approaches. These models are then optimized using GridSearchCV. Evaluation involves the valuation of model performance by

[1] https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/data

utilizes criteria such as accuracy, precision, recall, F1-score, and ROC AUC. Deployment involves the implementation of the most effective model in an interactive interface using Streamlit for the purpose of real-time attrition prediction.

This report contains further five parts. Section 2 of the report presents a comprehensive assessment of the previous research conducted in the areas of data mining and employee attrition. The research discusses the methodologies and implementations in Sections 3 and 4, and Evaluation in Section 5. Section 6 of the study discusses potential future work, thereby concluding the report.

2 Related Work

2.1 Introduction to Employee Attrition

Employment turnover or more commonly known as employee attrition is a movement of employees whereby existing employees in an organization are replaced by new employees. Hiring expenses had risen, knowledge of previous candidates is lost, morale among workers reduced significantly. High turnover rates have significant adverse effects on the organization performance (Margherita, 2022). Keeping a steady and effective staff requires anticipating and controlling turnover. This project gives HR departments a proactive tool by using machine learning to predict probability of staff turnover.

2.2 Importance of Predicting Employee Attrition

To combat employee turnover, Human resources (HR) departments are increasingly turning to machine learning and sophisticated analytics. Proactive measurements are challenging because traditional techniques, like surveys and exit interviews, sometimes lack accuracy and timeliness. Huge datasets may be analyzed by machine learning to identify trends that indicate which employees are more likely to leave. Al-Alawi and Ghanem (2024) emphasise feature engineering such as employee demographics, job satisfaction, and performance metrics to boost accuracy and indicate the advancements made in the use of machine learning algorithms for attrition prediction. Also, they show how crucial it is to use various assessment criteria, including F1-score, accuracy, precision, and recall. I used these insights to my project by designing appropriate features and using these thorough metrics to assess my models to guarantee effective performance.

2.3 Strategic Implications of HR Analytics

Strategic HR management requires a thorough understanding of employee attrition through HR analytics. Margherita (2022) offers a comprehensive analysis of the state of HR analytics, classifying research into enablers of HR analytics, applications, and value creation. With its deeper insights into human behaviour and organisational performance, this framework facilitates understanding how artificial intelligence and advanced analytics may change HR operations. Important obstacles are also highlighted by Margherita's work, including data protection, and integrating HR analytics with more comprehensive corporate plans.

Leveraging the full potential of predictive analytics in HR requires addressing these issues. Organisations may anticipate attrition and create focused interventions to keep prized people by matching predictive models with strategic HR goals. Within the framework of my project, I made certain that my predictive models are in line with strategic HR objectives by prioritising elements that are both pertinent and implementable. This coordination enables HR departments to not only forecast attrition but also develop tactics that can successfully prevent it.

2.4 Integrating Predictive Analytics into HR Strategy

The integration of Al-Alawi and Ghanem's (2024) analysis with Margherita's (2022) strategic framework offers a complete approach for effectively addressing employee turnover. Machine learning models may be created and optimised to accurately forecast attrition, while the overall strategic framework guarantees that these models are in line with the aims and ethical standards of the organisation. Therefore, by adopting this approach that covers all the aspects of Human Resources Management, the roles of the human resource experts involved can easily transform from the predominantly tactical bench to being a strategic force that handles labor stability and optimizes performance in an organization (George et al., 2022; Krishna et al., 2023). The best models include Logistic Regression, Random Forest, Decision Trees, Support Vector Machines (SVM), and K-nearest Neighbors (KNN) which I included in my work. Grid search and cross-validation approaches were employed to optimise each model's performance and verify that it aligned with organisational policies for proactive attrition control.

2.5 Traditional Methods of Predicting Attrition

Employers have always used both qualitative and quantitative techniques to forecast staff loss. Though helpful, these approaches frequently fall short of the accuracy and forecasting abilities achieved by modern machine learning strategies.

Employee surveys and exit interviews: While exit interviews offer detailed insights into the reasons behind workers departures, employee surveys collect data on employee happiness, engagement, and reasons for leaving. These approaches, while useful, are more reactive than proactive in that they identify attrition causes after the fact. Managerial Judgements: To forecast possible attrition, managers rely on their interactions and observations of their workforce. But because of its subjectivity and reliance on the manager prejudices and expertise, this method produces inconsistent and erroneous forecasts (George et al., 2022).

Historical Analysis: To forecast future turnover, historical analysis looks at historical attrition trends. This approach is helpful, but it assumes that historical trends will hold true, which may not always be the case because organisational dynamics are changing. Statistical Modelling: Conventional statistical techniques that find correlations between variables and attrition include regression analysis. But complicated interactions and non-linear connections, which are prevalent in attrition data, are a challenge for these models (Krishna et al., 2023).

The weakness of conventional employee attrition prediction techniques is highlighted by recent research. George et al. (2022) draw attention to their imprecision, pointing out that biases and insufficient data frequently cause these approaches to miss the complex and multidimensional character of attrition. Krishna et al. (2023) draw attention to their poor predictive capacity, which is complicated by their reliance on preset factors and linear connections. Furthermore, real-time analysis and managing huge, complicated datasets are challenges for standard methods, requiring stronger approaches like the ensemble learning approach employed in this research.

2.6 Machine Learning Prediction

Based on the above limitations, modern research is more and more on the application of machine learning approaches to predict employee turnover. This study agrees with George et al. (2022) and Krishna et al. (2023) that the machine learning model particularly, ensemble techniques are more efficient in predicted accuracy and robustness in comparison to the traditional approach. The algorithms of machine learning possess the ability to analyze large data, discover complex dependencies, and provide real time forecasts and recommendations. It is descriptive in the aspect of attrition management, which enables organizations to identify the workers who are at risk of leaving and control the attempts that can retain them.

Halibas et al. (2019) noted that analysing the variables involved and creating new variables can substantially help determine customer turnover, which may also help the staff turnover prediction. They used different ML algorithms and address the problem of unequal distribution of classes by applying the SMOTE technique. Wardhani and Lhaksmana (2022) used logistic regression and a feature selection technique, recursive feature elimination in their study. The independent input variables' interactions and large-scale data processing were effectively managed by Isha et al. (2024) through XGBoost. Moreover, in their reasoning, the authors also emphasized the usage of hyperparameters tuning and cross validation for the selection of features.

Chaurasia et al. (2023) described in detail that artificial neural networks (ANN) can learn non-linear interactions and perform better than conventional methods. In another study, Xiahou and Harada (2022) developed an architecture that innovatively fused K-Means clustering with Support Vector Machines (SVM) for predicting customers' churn rate and staff turnover rate. Random Forest with an emphasis on the role of feature importance scores was applied to collect the results, Pratt, Boudhane and Cakula (2021). Konar, Das, and Das (2023) incorporated genetic algorithm with XGBoost with the help of SMOTE and ADASYN for handling the class imbalance problem and optimizing the hyperparameters focusing on the significance of feature importance scores.

Qutub et al. (2021) showed that Logistic Regression, Random Forest, and Gradient Boosting improve employee attrition prediction. Wu (2022) used LSTM networks and Transformer models to forecast attrition in real time, enabling proactive staff retention. Ensemble approaches for telecom churn prediction by Pandithurai and Sriman (2023) reveal comparable staff attrition possibilities. For attrition prediction, Baghla and Gupta (2022) recommended tweaking model parameters for best performance. Chaurasia et al. (2023) showed ANNs can catch complicated patterns, improving attrition forecasts.

Research today highlights the application of machine learning for analyzing and predicting employee turnover because the models created are far more accurate and reliable compared to the prior models. Advanced machine learning algorithms such as the ensemble method could deal with large data sets, look for intricate patterns, and offer constant insights and forecasts, which can improve attrition control. Improving accuracy, feature selection, hyperparameters' optimization, and how to work with an imbalanced set of classes are essential for better results. Relevant predictive models that have been successfully employed include logistic regression, random forests, artificial neural networks, and new hybrid techniques, which have demonstrated potential for employee attrition forecasting, providing vital insights to organizations on how to retain their vulnerable employees. The following table presents the methods employed by other authors to forecast the rates of employee turnover.

Table 1: Methodologies Used for Employee Attrition Prediction Studies

Authors & Year	Methodologies Used
Al-Alawi and Ghanem (2024)	Neural Network, LR, The Hybrid Model (SOM And BPN)
George, Lakshmi, and Thomas (2022)	Extra Tree Classifier, Random Forest, Gradient Boosting, ADA Boost
Krishna, Dwivedi, and Murti (2023	Random Forest, K-NN, Naive Bayes
Wardhani and Lhaksmana (2022)	Logistic Regression
Isha et al. (2024)	XGBoost
Chaurasia et al. (2023)	Artificial Neural Networks
Xiahou and Harada (2022)	K-Means, Support Vector Machine (SVM)
Pratt, Boudhane, and Cakula (2021)	Random Forest Algorithm
Konar, Das, and Das (2023)	Genetic Algorithm, XGB Classifier, Parameter Optimization
Qutub et al. (2021)	Ensemble Methods
Chen Wu (2022)	Sentiment Analysis
N. B. Yahia, J. Hlel and R. Colomo-Palacios, (2021)	Ensemble Methods and Deep Learning Models

2.7 Feature Engineering, Selection and Evaluation Metrics

Exploratory Data Analysis and Initial Feature Engineering: Halibas et al. (2019) applied EDA and feature engineering in telecom customer attrition modelling. They showed that accurate prediction requires detailed EDA and feature engineering. EDA, pattern identification, and feature creation impact attrition were comparable in this project. Mohamad et al. (2021) utilised categorisation and feature selection to predict employee attrition. They found important characteristics using Pearson's Chi-square and correlation analysis. Feature selection reduced noise and focused on the most important characteristics, improving the model's predictive performance.

Table 2: Factors that Lead to Employee Attrition

Author	Features
S. George, K. A. Lakshmi, and K. T. Thomas (2022)	Age, Job Satisfaction, Work Environment, Compensation
S. Krishna, R. Dwivedi, and A. Murti (2023)	Age, Monthly Income, Overtime, Years at Company, Total Working Years, Environment Satisfaction, Marital Status
Isha, N. Thapliyal, S. Solanki, N. K. Pandey, and S. Papola, 2024	Age, Job satisfaction, daily rate, and Employee number
A. Chaurasia, S. Kadam, K. Bhagat, S. Gauda, and P. Shingane, 2023	Employee tenure, performance ratings, salary
Pratt.M, Boudhane.M, and Cakula, (2021)	Monthly Income, Age, Daily Rate, Total Working Years and Monthly Rate
M. R. Mohamad, F. Hanum Nasaruddin, S. Hamid, S. Bukhari, and M. T. Ijab (2021)	Business Travel, Department, Education Field, Environment Satisfaction, Job Related Attributes, Distance from Home

2.8 Current Trends and Future Directions

Advanced, ethical, and comprehensive employee attrition prediction methods have emerged. Subhashini and Gopinath (2020) emphasise the use of numerous machine learning algorithms in the business to improve forecast accuracy and dependability. Optimising model performance requires robust feature selection and preprocessing. Yahia, Hlel, and Colomo-Palacios (2021) highlight the move from big data to deep data, using deep learning to gain insights from complicated datasets. Understanding complex employee behaviour patterns enhances attrition risk identification.

When using AI, ethical issues are crucial, placing a strong emphasis on privacy, transparency, and bias reduction (Tiwari, 2023). Establishing moral standards and policies guarantees that workers are treated fairly. AI-based attrition prediction systems, like the ones Agarwal et al. (2023) showed, provide HR departments with useful information and suggestions. Future initiatives include resolving ethical issues, creating integrated systems that include predictive and prescriptive analytics for efficient retention methods, and further improving model accuracy through advanced analytics and deep learning. These developments will aid businesses in better managing their human resources, which will improve employee performance and stability.

2.9 Implementation and Model Development

Using the outcomes of this research I built an interactive model to predict the employee attrition rate using Streamlit which also displays the probability of attrition rate. In the training and assessment of these models, Accuracy, precision, recall, F1-score were applied with the help of the Support Vector Machines, Random Forest, Logistic Regression, Decision Trees, KNNs, and Voting Classifiers. Thus, it is aligning with Annamalai et al. (2023), which

enabled the proactive management of workforce stability, and performance improvement through enhanced development speed, transparency and scalability.

3 Methodology

This study's approach is based on the CRISP-DM framework. The six stages of this methodical technique are Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. Following the CRISP-DM framework guarantees that the project is in line with business goals and makes use of trustworthy data for analysis.

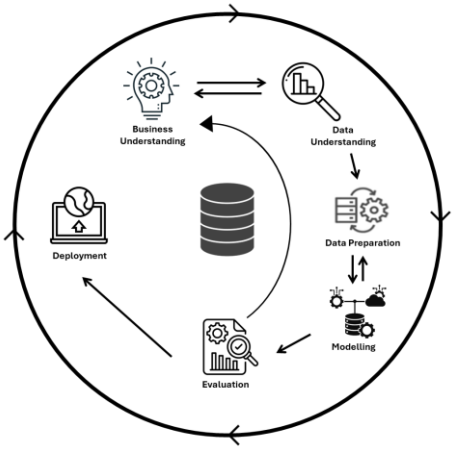

Figure 1: CRISP-DM Methodology

3.1 Business Information

The purpose of this research is to establish a model, using a machine learning approach, to predict an employee's propensity to leave the company based on specific attributes. It is applied in predicting the possibility of staff turnover and used to incorporate strategies that would help in increasing staff retention at the same time minimizing on the costs incurred in high turnover and improving the organization's effectiveness. Staff turnover is a very critical area of concern because it affects the knowledge level and rhythm within an organization. Therefore, it is crucial to identify the main reasons that contribute to attrition and precisely predict the probability of most likely instances.

3.2 Data Information and Analysis

The dataset utilized in this research was initially generated by IBM data scientists as a fictitious dataset. Obtained a copy from Kaggle[2]. The dataset has a range of attributes pertaining to workers, such as demographic details, job positions, remuneration, and other pertinent elements that impact attrition. The dataset has 1470 entries and 35 characteristics, including important variables such as Age, BusinessTravel, Department, DistanceFromHome, Education, EnvironmentSatisfaction, Gender, JobRole, JobSatisfaction, MaritalStatus MonthlyIncome, OverTime, and WorkLifeBalance.

[2] https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/data

3.3 Data Preprocessing

The original dataset was thoroughly scrutinised for any instances of missing values and duplicate entries to maintain the integrity of the data. The dataset was checked for missing values and no instances were detected. Additionally, duplicate entries were eliminated to ensure the integrity of the dataset. Columns with unique numbers for each employee, such as EmployeeNumber, were eliminated. In addition, columns that had just one unique value, such as EmployeeCount, Over18, and StandardHours, were eliminated as they do not offer any valuable information for the model.

Exploratory data analysis (EDA) was performed to gain an understanding of several characteristics of the data including the shape of the distributions and relations between attributes. Many arithmetic and graphic methods were applied for condensation of the data and for the search of the patterns and semi patterns, and for the search of the outliers. The essential sections consisting of Descriptive Statistics including measures such as mean, median, and standard deviation, were calculated for numerical aspects to gain insight into their distribution. Box plots and bar charts were employed to visually represent the distribution of characteristics and the correlation between categorical factors and attrition. The visualisations provide valuable insights into the primary elements that influence employee churn.

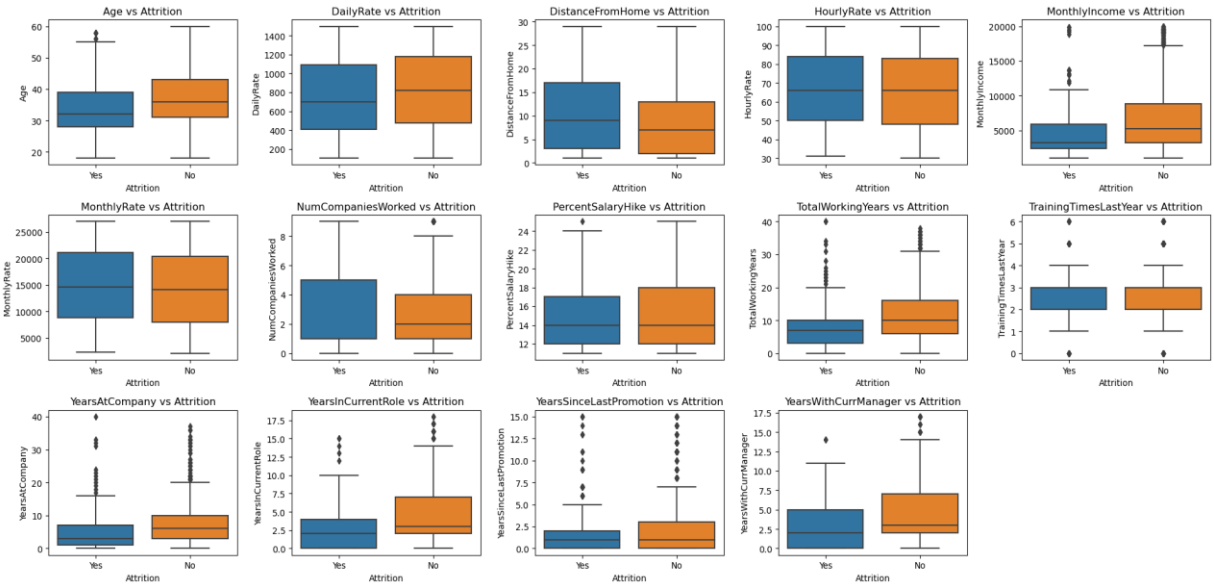

Figure 2: Distribution Of Numerical Features Against Attrition

The above box plots show the spread of several numerical attributes against employee attrition; those who left compared with those who remained. Several indicators indicate that attrition tends to occur among the younger employees earning lower monthly remuneration, residing farther away from the workplace, and have lesser total working years, years of service with the company, and years with the current immediate supervisor or manager. Also, the employees who departed can be seen to have worked previously at more firms as compared to others. These findings stress multiple factors that impact the level of employee turnover, showing directions to the HR departments, enabling them to detect the employees

that can potentially leave or create proper strategies for their retention stressing the methodological approach of the current study toward attrition rate prediction.

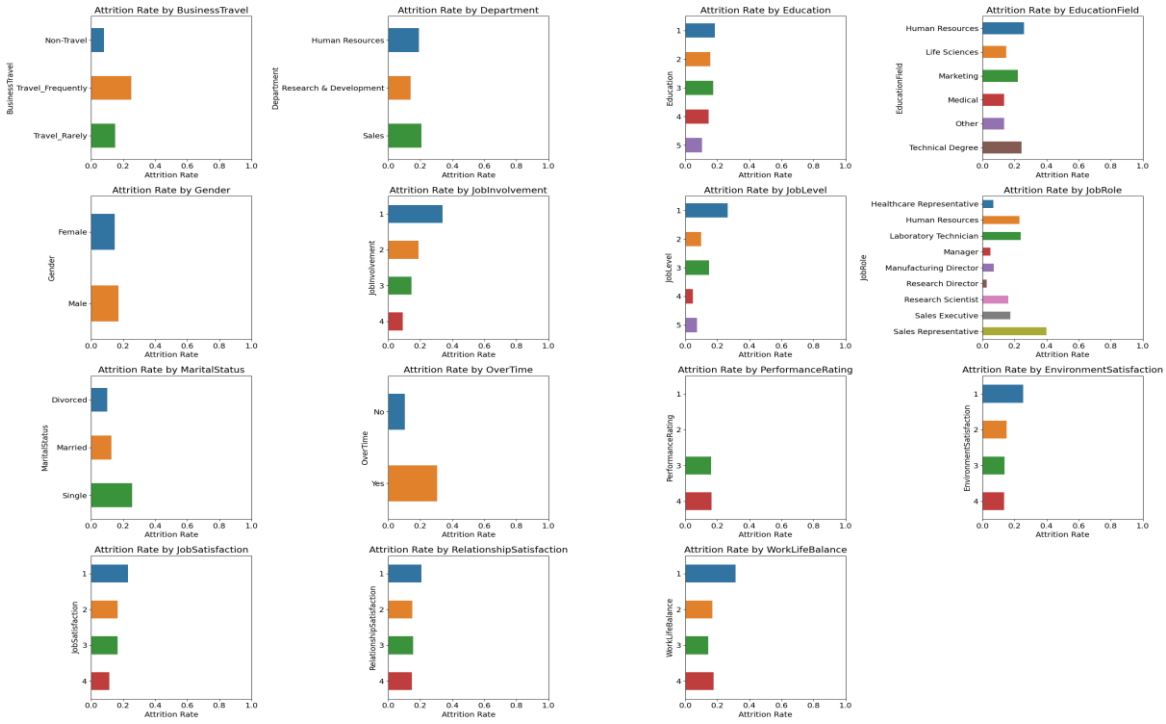

Figure 3: Distribution of Categorical Features Against Attrition

The series of bar plots show the attrition rate of the employees taking into consideration all categorical features and presenting the comparisons between those who left the company and those who did not. Analyzing the visualizations part of the report, it is possible to highlight that the Sales and Human Resources departments, the workers with the low level of job satisfaction, and the workers who often travel for business, have a higher attrition rate. On the same note, employees are working extra hours. These plots highlight the significance of knowing the effects of position attributes, satisfaction levels, and work circumstances on quitting behavior, which can give essential records for HR authorities to work on influential retention approaches to enhance organizational stability.

3.4 Feature Engineering

Categorical variables were encoded to convert non-numeric data into a format that is appropriate for machine learning techniques. The process of label encoding was employed to transform categorical data into numerical representation, while one-hot encoding was utilised to generate binary columns for each category. This phase guaranteed that the models were able to accurately read the categorical data. Normalisation was utilised to standardise numerical characteristics, ensuring that the data is scaled within a certain range. This process improves the effectiveness of machine learning algorithms. The MinMaxScaler method was employed to convert the following numeric features Age, DailyRate, DistanceFromHome, HourlyRate, MonthlyIncome, MonthlyRate, PercentSalaryHike, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsWithCurrManager, and

YearsSinceLastPromotion. A thorough methodology was employed to preprocess, explore, and optimize features, guaranteeing an effective dataset for modelling. The techniques described in this methodology offer a strong basis for creating predictive models that can effectively assess the probability of employee turnover using specific criteria.

4 Model Implementation

In the CRISP-DM model, the model implementation phase encompasses the enhancement of the machine learning models to predict the likelihood of employees' turnover. The first is the choice of algorithms, with the second process being the training of plenty of models with the help of the prepared data. Methods such as GridSearchCV are used in hyperparameter tuning to increase the efficiency of the model. Logistic regression, Decision Tree, Random Forest, SVM, KNN, and the best two Ensemble techniques are used to evaluate the most suitable model for the experiment. The chosen model is subsequently verified with unfamiliar test data to guarantee its prediction precision and dependability.

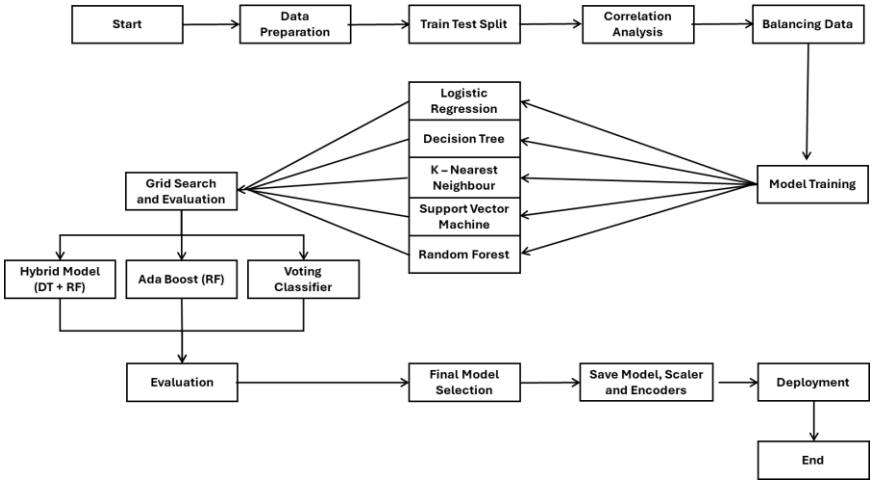

Figure 4: Model Architecture

4.1 Balancing Training and Testing Data

The dataset was partitioned into training validations, and testing sets to assess the models' performance. The training set was utilized to train the models, but the testing set was exclusively allocated for assessing the models' performance. This division meant that the models were evaluated on data that had not been previously seen, therefore offering an accurate assessment of their ability to make predictions. A correlation analysis was done to determine the correlations between the various characteristics prior to moving further with the model training. The correlation coefficients between numerical features were visualized by creating a correlation heatmap. To prevent multicollinearity, which might have a detrimental effect on model performance, features with strong correlation coefficients were found. It was discovered that JobLevel and TotalWorkingYears had a strong correlation. To avoid duplication and possible overfitting, JobLevel was eliminated from the dataset.

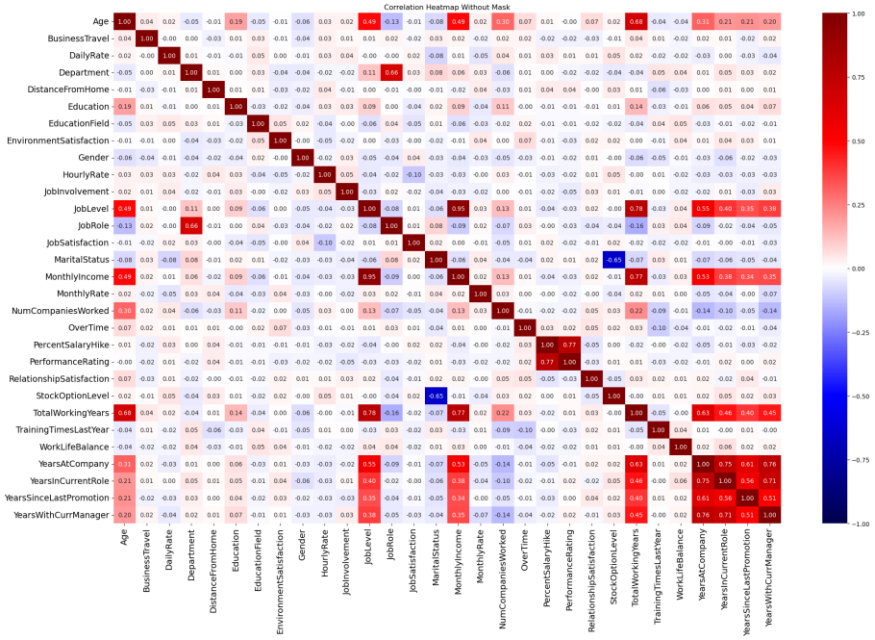

Figure 5: Correlation Heat Map

The data showed an imbalance between the two classes; non-attrition had far more cases than attrition instances. Concerning this problem, Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE performs by oversampling the minority class which means it brings examples of this class close and equilibrates the data in a way that does not promote models with bias for the major class.

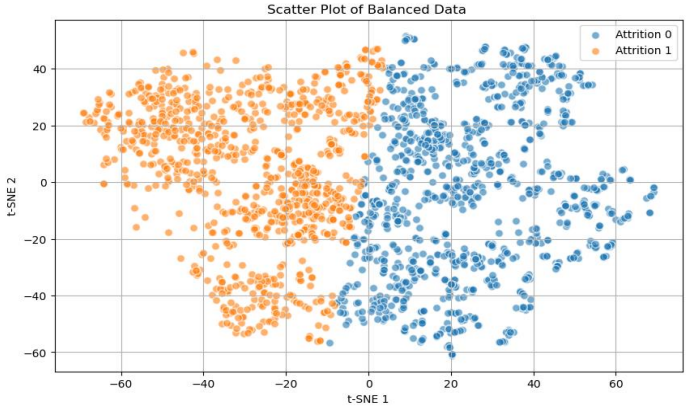

Figure 6: Scatter plot of Attrition

To see the distribution of balanced data in a two-dimensional space, a t-SNE (t-Distributed Stochastic Neighbor Embedding) transformation was also applied to the training set. The balanced dataset's separation between the two classes of Attrition is highlighted by the t-SNE graphic in conjunction with KMeans clustering. Understanding the uniqueness of the techniques and the efficacy of the balancing approach is made easier with the use of this visualization. After balancing, the scatter plot below shows a distinct distinction between the attrition groups, demonstrating the efficacy of the SMOTE approach and supporting visual validation of the data preparation processes.

4.2 Models and Hyperparameter Tuning

Multiple machine learning models were created and fine-tuned to forecast employee turnover rates. Every model received hyperparameter tuning to optimize its performance. Several machine learning models were created and optimized for predicting the employee's attrition, with each model undergoing hyperparameter tuning to improve performance. The Logistic Regression model achieved an accuracy of 0.8605 after optimization, whereas the Decision Tree Classifier increased its accuracy to 0.6667. The accuracy of the K-Nearest Neighbors models significantly improved, reaching an accuracy of 0.5952. The Support Vector Machine (SVM) model maintained a high accuracy of 0.8673 after adjustment. Also, the hyperparameters of Random Forest model were optimized which improved its performance and the accuracy achieved was 0.8571. These improvements show that it is necessary to tune hyperparameters concerning the increase of the model accuracy in the prediction of employee attrition.

Voting Classifier- A Voting Classifier with soft voting was developed to exploit the advantages of many classifiers. This ensemble technique consolidated the predictions of the top-performing models (Decision Tree, Random Forest, SVM, KNN, and Logistic Regression) to enhance the overall accuracy. Soft voting is a technique where the projected probabilities of each classifier are averaged. This approach frequently results in greater performance compared to using a single model.

Voting Classifier with Best Grid- A final Voting Classifier was constructed utilizing the optimal hyperparameters acquired via grid search for each individual model. The table below displays the models that are included in this ensemble approach.

Table 3: Parameters used for each model

Model	Hyperparameters
Logistic Regression	C=0.1, penalty='l2', solver='newton-cg', random_state=42
Decision Tree	max_depth=2, min_samples_leaf=5, random_state=42
K-Nearest Neighbors	metric='manhattan', n_neighbors=6
Support Vector Machine	C=10, gamma=1, kernel='linear', probability=True
Random Forest	max_depth=340, max_features='log2', n_estimators=2000

5 Evaluation and Results

These include the accuracy, precision, recall, and F1-score were among the evaluation techniques used in conducting the analysis of the prediction of models of employee attrition. These measures offer a thorough grasp of how effectively the models forecast situations that result in attrition as well as those that do not. In addition, the models' discriminative ability was evaluated using the Area Under the ROC Curve (AUC) and the Receiver Operating Characteristic (ROC) curve.

5.1 Ensemble Methods

Hybrid Model (Decision Tree + Random Forest)- To combine the advantages of Random Forest and Decision Tree classifiers, a hybrid model was created. Attrition probabilities were predicted using the Decision Tree classifier, and the Random Forest classifier further assessed cases with a high confidence of "No Attrition". The goal of this method was to combine the robustness of Random Forests with the accessibility of Decision Trees. When compared to the Decision Tree model alone, the hybrid model showed an accuracy increase of 0.7823, demonstrating the value of combining models to improve predictive performance. But for a realistic implementation, the hybrid model's complexity could be a challenge.

The AdaBoost model, utilizing Random Forest as the underlying estimator, exhibited considerable enhancement in prediction accuracy, with an accuracy score of 0.8537. The iterative nature of AdaBoost enables it to prioritize misclassified examples, leading to improved model capacity and reduced bias. Using AdaBoost with the Random Forest base estimator resulted in an improved balance between accuracy and recall, which is essential for handling the class imbalance in the attrition dataset. The result is consistent with the findings of Qutub et al. (2021), who likewise emphasized the efficacy of ensemble approaches in enhancing accuracy of predictions.

Voting Classifier- The Voting Classifier aggregated the predictions of many models, including Decision Tree, Random Forest, SVM, KNN, and Logistic Regression, using soft voting. The purpose of this ensemble technique was to enhance the total accuracy by focusing on the capabilities of each individual classifier. The Voting Classifier attained an accuracy score of 0.8469, suggesting that combining predictions from many models might result in a more resilient and consistent predictive model for employee attrition.

Voting Classifier with best Grid- To improve the Voting Classifier, the optimal hyperparameters found via grid search for each individual model were utilized. The optimized ensemble technique achieved a little higher accuracy score of 0.8503, highlighting the significance of hyperparameter adjustment in enhancing model performance. The Voting Classifier with the Best Grid achieved superior performance compared to the regular Voting Classifier, highlighting the importance of optimizing each model in the ensemble. This technique is consistent with the research conducted by Pandithurai and Sriman (2023), which also showed that ensemble approaches are beneficial in enhancing predicted accuracy for comparable classification issues.

5.2 Results

The provided table presents a concise overview of the classification outcomes for several models used on the employee attrition dataset. It specifically emphasizes important performance measures like accuracy, precision, recall, and F1-score. The Voting Classifier with soft voting and the Grid Search optimized Support Vector Machine (SVM) were identified as the most successful models. The Voting Classifier demonstrated an accuracy of 85.03%, effectively achieving a balance between precision (91%) and recall (55%). The use of this ensemble strategy, which combines several models, yielded a resilient and dependable predictive capacity, positioning it as one of the most proficient models. The Grid Search

optimized SVM model attained an accuracy of 86.73%, with precision of 89% and recall of 40%. This model also shows a robust capability to forecast employee turnover, but with a lower recall rate in comparison to the Voting Classifier. The results demonstrate that both the Voting Classifier and the SVM are very proficient at predicting employee attrition.

Table 4: Overview of Classification Results for Different Models

Model	Accuracy	Precision	Recall	F1-Score	TP	TN	FP	FN
Logistic Regression	0.8605	0.90	0.43	0.49	20	233	14	27
Decision Tree	0.6667	0.91	0.64	0.38	30	166	81	17
K-Nearest Neighbors	0.5952	0.89	0.60	0.32	28	147	100	19
Support Vector Machine	0.8673	0.89	0.40	0.49	19	236	11	28
Random Forest	0.8571	0.88	0.28	0.38	13	239	8	34
Voting Classifier (Best Grid)	0.8503	0.91	0.55	0.54	26	224	23	21
Hybrid Model (DT + RF)	0.7823	0.92	0.62	0.48	29	201	46	18
AdaBoost	0.8537	0.87	0.23	0.34	11	240	7	36

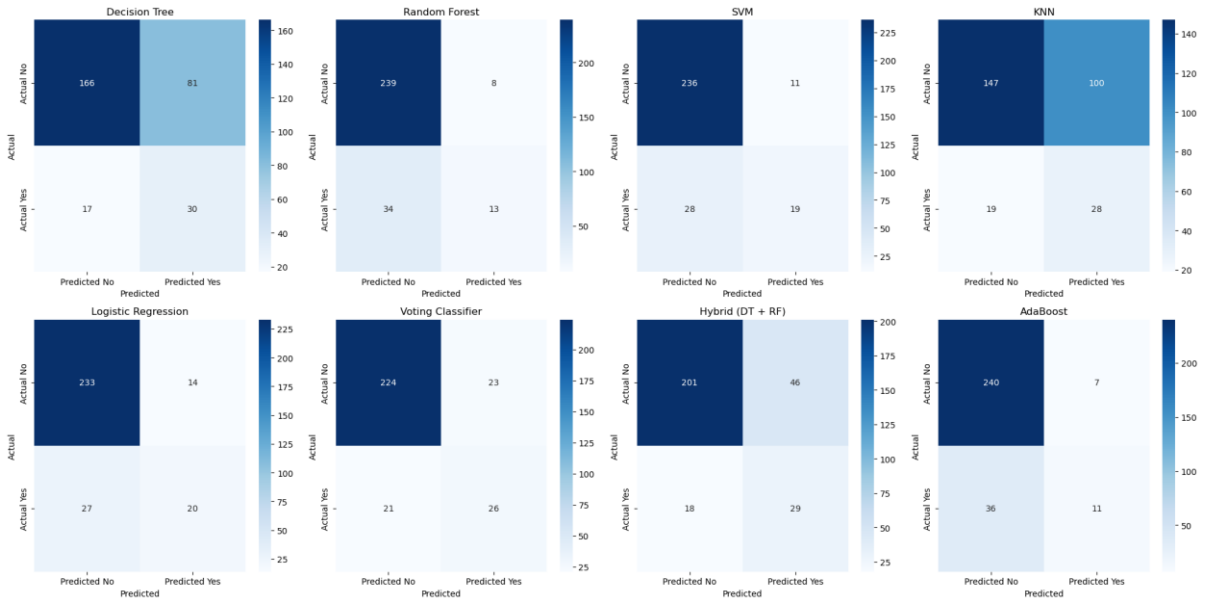

Figure 7: Confusion Matrix Plot for All Models

The ROC curve seen below depicts the effectiveness of several classifiers in forecasting employee attrition. The ROC curve is a graphical method that presents the sensitivity, namely the true positive rate, against the specificity, which is the 1-specificity or false positive rate, based on the adopted thresholds. Particularly, the AUC is a detailed measure of the classification accuracy reflecting the model's effectiveness with respect to any criterion. A greater AUC value indicates superior model performance.

The ROC curve clearly demonstrates that the Support Vector Machine (SVM) classifier earned the maximum AUC value of 0.865, suggesting its excellent ability to differentiate between attrition and non-attrition instances. The Voting Classifier, which utilizes soft voting to integrate different models, achieved a high performance with an AUC of 0.845. The second in line was the Logistic Regression Classifier with an AUC of 0.849. The accuracy of the Random Forest Classifier was 0.830, while the AUC was given by AdaBoost Classifier with the value of 0.837. This shows that both models have good predictive power as the value of 0.837 is high. The Decision Tree and Random Forest algorithms were employed in the Hybrid Model, and it was observed that it had good accuracy with an AUC of 0.794. This model takes full advantage of the features of Decision Tree, it also contains some features of Random Forests.

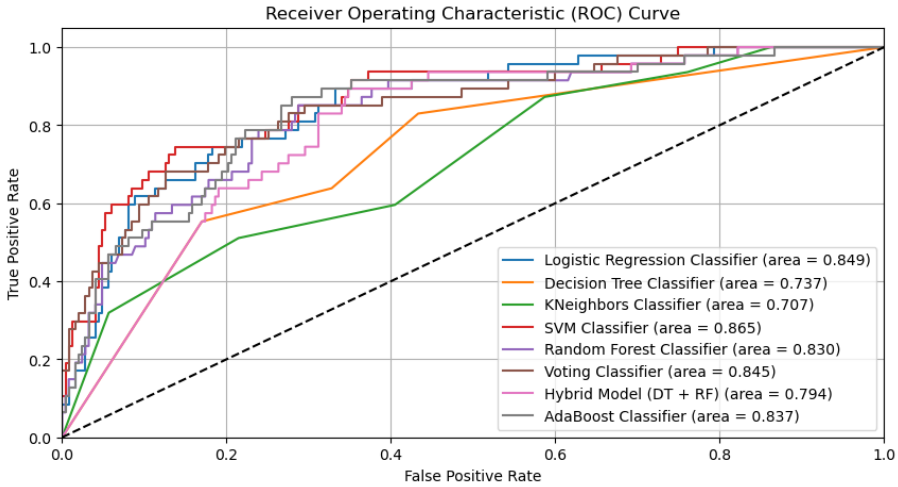

Figure 8: ROC Curve Results

The Decision Tree Classifier and the K-Neighbors Classifier had the lowest AUC of 0.737 and 0.707, respectively, which indicates that the performance of these models is less trustworthy as compared to the other models with lower values on the scale. In summary, the ROC curve demonstrates that ensemble approaches and hyperparameter-tuned models are highly helpful in enhancing predicting accuracy for employee attrition. The Support Vector Machine (SVM) and Voting Classifier have been identified as the most effective approaches among the options considered.

5.3 Best Model – Voting Classifier Best Grid

After conducting a thorough assessment of several models, such as Logistic Regression, Decision Tree, KNN, SVM, Random Forest, AdaBoost, and various ensemble methods, the Voting Classifier (Best Grid) was identified as the most effective model for predicting employee attrition. The model obtained a ROC AUC score of 0.845, demonstrating a robust capability to distinguish between employees who are likely to depart and those who are likely to remain. The performance of the model was well-balanced, as indicated by its accuracy of 0.8503, precision of 0.53, and recall of 0.55 for the attrition class. The Voting Classifier employs an ensemble technique that effectively exploits the capabilities of its constituent models, including Decision Tree, Random Forest, SVM, KNN, and Logistic Regression. This

leads to more consistent and trustworthy predictions in comparison to using individual models alone. The durability and adaptability of this technology make it well-suited for a wide range of complicated datasets, allowing for practical implementation in real-world situations.

The individual models of the Voting Classifier were carefully optimized using GridSearchCV to get the best possible performance for the features of the attrition dataset. When comparing the SVM with the Voting Classifier, the SVM had a slightly higher individual ROC AUC score of 0.865. However, the Voting Classifier showed a more balanced performance across all measures. Although the SVM showed high precision and recall, the Voting Classifier proved to be more efficient in predicting attrition instances. The Voting Classifier's ensemble technique helps reduce the danger of overfitting and model-specific biases, providing a more robust and generalized result. Moreover, the Voting Classifier's capability to integrate the features of various algorithms leads to a reduction in both variance and bias, resulting in a more dependable option for predicting employee's turnover. The Voting Classifier, with its superior predictive skills and practical use, appears as the best approach for effectively detecting employees who are likely to leave. This makes it a valuable tool for human resource management.

5.4 Feature Importance Analysis

An examination of feature importance in the Voting Classifier was quite advantageous in as much as it helped me identify the key attributes involved in employee churn. The model has identified the following characteristics as the most significant: OverTime, Years in current Role, level of Stock Option and Distance from Home. One of the most influential components was OverTime that brought about the vital role of workload and work-life balance for defining the attrition rates. Opposite to the findings of the previous studies carried out by S. George et al. (2022) and S. Krishna et al. (2023) which provided conclusions entirely based only on Age and Monthly Income. The fact that OverTime and Years in Current Role proved to hold significant major significance for attrition means that these previously neglected factors need to be included in further research.

Table 5: Top 5 Important features responsible for Attrition

Rank	Features Observed	Details
1	Overtime_0	Employee does not work overtime.
2	YearsInCurrentRole	No. of years the employee has been in their current role.
3	StockOptionLevel_1	The employee's Stock level (Level 1).
4	DistanceFromHome	Distance between the employee's home and workplace.
5	YearsWithCurrManager	No. of years the employee has worked with their current manager.

The very positive results of the Voting Classifier strengthen the analysis, and it appears that the ensemble methods are of paramount relevance when it comes to achieving a

comprehensive understanding of the attrition process. It also highlights that organizations must continue to support the work-life balance, where they try to provide financial rewards for employees as the best ways of retaining them.

6 Conclusion and Future Work

In this research, I followed a structured procedure to model and forecast the employee attrition rate by utilizing several machine learning techniques according to the CRISP-DM framework. In this task, I carefully dealt with different difficulties that could arise during the pre-processing of the data set which include the issue of class imbalance and multicollinearity. The algorithms Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, AdaBoost, and ensemble classifiers namely Voting Classifier were tested for model accuracy. As for the Voting Classifier, the model with the optimal parameters obtained from grid search turned out to be the best one with an accuracy of 0.8503 and a ROC AUC of 0.845. Compared with the separate classifiers, this model was much more balanced and quite accurate since the individual classifiers' advantages were united. The ensemble approach offered the advantage of diversification, which greatly enhanced its applicability for use in real life human resource management for accurate determination of employees at high risk of turnover. Most notably, only the Voting Classifier predicted attrition cases more accurately than SVM while showcasing a strong comprehensive way of prediction.

Despite the success demonstrated by the ensemble methods and specifically the Voting Classifier, the following are some ideas on how to improve the given model and make it more efficient for predicting employee attrition. Possible research further could be related to the more complex feature engineering methods, extracting more of the likely informative features from the data. Adding time-series analysis could help identify trends and patterns over time, which could help create a more robust prediction model. It is, however, crucial to note that although the Voting Classifier is a very effective ensemble method there are issues of interpretability because of the complexity of the algorithms; this can be an area that needs improvement in future, where research should be directed towards methods of developing better methods of interpreting these models and provide some insights into factors that have been used to make the attrition predictions. Furthermore, when the model predicts a high probability of an employee departing, an ethical dilemma occurs. In these situations, HR must take this forecast carefully and use it as a foundation for assistance and engagement rather than the final decision ,to guarantee that workers receive impartial and equitable treatment. The application of these models in real time can provide the departments of Human Resources with timely alerts and practical information which can ensure proactive action when it is required to retain the key employees. Moreover, it may be beneficial to consider utilizing deep learning models like neural networks and comparing their performance to these forecasting techniques.

To simplify the actual use, I have developed a Streamlit API that incorporates the Voting Classifier Best Grid architecture. This API facilitates immediate forecasts and user-friendly interactions, allowing HR managers to input employee variables such as age, distance from

home, and other relevant variables to obtain rapid evaluations of attrition risk. Validating the efficacy and robustness of the created models may be achieved by applying them to datasets from other sectors. An examination of several industries would aid in understanding the wide effectiveness of predictive models. To summarize, our research has established a strong basis for forecasting employee attrition through the utilization of machine learning models. The encouraging outcomes from the Voting Classifier motivate more investigation and improvement to enhance prediction skills, offering important instruments for strategic human resource management.

Acknowledgements

I like to convey my sincere thanks to my supervisor, Prof. Christian Horn, for his important guidance, insightful comments, and unwavering support during this project. His support and guidance have been vital in the successful completion of this project. I express my gratitude to the School of Computing at the National College of Ireland for supplying the essential resources and a supportive atmosphere for my study. Finally, I would like to thank Kaggle for providing open datasets that were crucial for the success of my project.

References

Al-Alawi, A.I. and Ghanem, Y.A., 2024, January. Predicting Employee Attrition Using Machine Learning: A Systematic Literature Review. In *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)* (pp. 526-530).

Margherita, A., 2022. Human resources analytics: A systematization of research topics and directions for future research. *Human Resource Management Review*, *32*(2), p.100795.

George, S., Lakshmi, K.A. and Thomas, K.T., 2022, December. Predicting Employee Attrition Using Machine Learning Algorithms. In *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)* (pp. 700-705).

Krishna, S., Dwivedi, R. and Murti, A., 2023, November. Predictive Analytics for Employee Attrition: A Comparative Study of Machine Learning Algorithms. In *2023 Third International Conference on Digital Data Processing (DDP)* (pp. 180-187).

Halibas, A.S., Matthew, A.C., Pillai, I.G., Reazol, J.H., Delvo, E.G. and Reazol, L.B., 2019, January. Determining the intervening effects of exploratory data analysis and feature engineering in telecoms customer churn modelling. In *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)* (pp. 1-7).

Wardhani, F.H. and Lhaksmana, K.M., 2022. Predicting Employee Attrition Using Logistic Regression With Feature Selection. *Sinkron: jurnal dan penelitian teknik informatika*, *6*(4), pp.2214-2222.

Thapliyal, N., Solanki, S., Pandey, N.K. and Papola, S., 2024, May. Employee Attrition Analysis Using XGBoost. In *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)* (pp. 1-6).

Chaurasia, A., Kadam, S., Bhagat, K., Gauda, S. and Shingane, P., 2023, May. Employee Attrition Prediction using Artificial Neural Networks. In *2023 4th International Conference for Emerging Technology (INCET)* (pp. 1-6).

Xiahou, X. and Harada, Y., 2022. B2C E-commerce customer churn prediction based on K-means and SVM. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(2), pp.458-475.

Pratt, M., Boudhane, M. and Cakula, S., 2021. Employee attrition estimation using random forest algorithm. *Baltic Journal of Modern Computing*, *9*(1), pp.49-66.

Konar, K., Das, S. and Das, S., 2023, January. Employee attrition prediction for imbalanced data using genetic algorithm-based parameter optimization of XGB Classifier. In *2023 International Conference on Computer, Electrical & Communication Engineering (ICCECE)* (pp. 1-6).

Qutub, A., Al-Mehmadi, A., Al-Hssan, M., Aljohani, R. and Alghamdi, H.S., 2021. Prediction of employee attrition using machine learning and ensemble methods. *Int. J. Mach. Learn. Comput*, *11*(2), pp.110-114.

Wu, C., 2022, November. Real Time Attrition Prediction Mechanism Based on Deep Learning. In *2022 2nd International Signal Processing, Communications and Engineering Management Conference (ISPCEM)* (pp. 128-131).

Pandithurai, O. and Sriman, B., 2023. Telecom Churn Prediction Using Voting Classifier Ensemble Method and Supervised Machine Learning Techniques. In *ITM Web of Conferences* (Vol. 56, p. 05012). EDP Sciences.

Baghla, S. and Gupta, G., 2022. Performance evaluation of various classification techniques for customer churn prediction in e-commerce. *Microprocessors and Microsystems*, *94*, p.104680.

Subhashini, M. and Gopinath, R., 2020. Employee attrition prediction in industry using machine learning techniques. *International Journal of Advanced Research in Engineering and Technology*, *11*(12), pp.3329-3341.

Mohamad, M.R., Nasaruddin, F.H., Hamid, S., Bukhari, S. and Ijab, M.T., 2021, November. Predicting employees' turnover in IT industry using classification method with feature selection. In *2021 International conference on computer science and engineering (IC2SE)* (Vol. 1, pp. 1-7).

Tiwari, R., 2023. Ethical and societal implications of AI and machine learning. *International Journal of Scientific Research in Engineering and Management*, *7*(01).

Yahia, N.B., Hlel, J. and Colomo-Palacios, R., 2021. From big data to deep data to support people analytics for employee attrition prediction. *Ieee Access*, *9*, pp.60447-60458.

Agarwal, S., Bhardwaj, C., Gatkamani, G., Gururaj, R., Darapaneni, N. and Paduri, A.R., 2023, June. AI Based Employee Attrition Prediction Tool. In *International Conference on Multi-disciplinary Trends in Artificial Intelligence* (pp. 580-588). Cham: Springer Nature Switzerland.

Brockett, N., Clarke, C., Berlingerio, M. and Dutta, S., 2019, December. A system for analysis and remediation of attrition. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 2016-2019).

Annamalai, R., Deena, S., Venkatakrishnan, R., Shankar, H., Harshitha, Y.S., Harini, K. and Reddy, M.N., 2023, December. Automating Machine Learning Model Development: An OperationalML Approach with PyCARET and Streamlit. In *2023 Innovations in Power and Advanced Computing Technologies (i-PACT)* (pp. 1-6).