AN EFFICIENT SUPERVISED MACHINE LEARNING MODEL USED FOR CLASSIFICATION AND PREDICTION OF EMPLOYEE ATTRITION

Subhash Chandra Jat
Professor-CSE
Rajasthan College of Engineering for Women
Jaipur, India
subhashcjat@gmail.com

Neha Singh
Computer Science & Engineering
Rajasthan College of Engineering for Women
Jaipur, India
kneha0182@gmail.com

Abstract*—A company's employees are its most important asset. It is crucial to know who could quit the company due to the high expense of professional training, the loyalty that has grown over the years, and the delicate nature of some organizational roles. Employee attrition (EA) can occur for several causes. Employees leave a company either voluntarily or involuntarily; this phenomenon is known as attrition. Researchers must investigate the use of machine learning (ML) in corporate organizations due to the increasing fascination with the topic among corporate decision-makers and executives. The use of ML models to investigate employee turnover (ET) is the focus of this study. Using data analysis and ML approaches, this research primarily aims to explain a methodical flow for forecasting attrition. Acquiring data, preprocessing it, visualizing it, engineering features, balancing it, and finally classifying it using a ML approach are the stages involved. Automated and precise prediction of EA is achieved in this work through the use of supervised ML models. The article details the use of k-best for feature engineering and ML classification techniques such as Random Forests (RFs) and decision trees (DTs) to forecast the likelihood of ET for each given new hire. Using the Kaggle dataset titled "IBM HR analytics employee Attrition Performance," this article trains and evaluates a RF model with cross-validation in a Python environment. The ultimate goal of effectively identifying attrition is to help any company enhancevarious retention tactics on important workers and raise the pleasure of those employees. Classification predictions were evaluated using the following 5 efficiency indicators: f1-score, Accuracy, Confusion Matrix, Precision, Recall, & the ROC Curve. The findings showed that the RF classifier had the best performance, with an accuracy of 96%. Businesses' capacity to avoid ET might be significantly affected by the results of this study.

Keywords-Employee Attrition, IBM HR analytics, Machine Learning, Random Forest, Decision Tree, Feaureenginerring, Random oversamplling.

I. Introduction

EA is the procedure by which workers leave an organization following causes, such as the involuntary resignation of employees. The phrase EA describes the normal process by which employees depart an organisation. There is a wide range of potential causes that might lead to ET[1]. There is a greater rate of employee turnover than there are new hires at the organization. The organization suffers a loss as an outcome of the openings that stay unfilled after the person departs the organisation where they were employed. The rate of staff turnover is a useful indicator for determining the degree of development that an organisation has made. It is clear from the high attrition rate (AR) that personnel are departing their positions regularly. The high AR leads to the loss of organizational benefits, which is the outcome of the situation. The rate of employee turnover needs to be managed in order to ensure that the organisation continues to make development[2].

The attrition process can be better understood by examining several forms of EA. Whether or whether a worker willingly quits is a measure of the attrition type. An example of involuntary attrition would be a company firing an employee. An example of external attrition would be an workers who leaves one organization to join another. Internal attrition occurs when an employee moves to a new position inside the same organisation as a result of a raise[3][4]. The percentage of workers that quit within a certain time period is called the EA rate. Taking stock of the AR will help us zero in on the root problems and find solutions that will keep valuable employees from leaving. The average AR is obtained by dividing the sum of all employee departures by the average staff count for a given time period. Using the AR, we may determine the company's performance within a certain time frame[5][6].

The capacity for machines to learn from past data and generate predictions is known as ML, and it is a subfield of AI. In today's data science landscape, ML is an essential tool. ML approaches aim to outperform humans in terms of accuracy [7]. The decision-making process makes use of the ML models. ML is an automated process. Machines learn from the cleaned-up data and use it to make conclusions about fresh data. The fundamental purpose of ML algorithms is to understand data by discovering patterns within it.

ML is finding more and more uses in modern technology every day. ML has several practical uses across many different fields. Common real-world issues, such as picture identification [8], ML has several applications in fields such as online commerce, transportation forecasting, voice recognition, text categorization, social analysis, healthcare, agriculture, and the stock market[9][12]. The purpose of using ML techniques is to forecast staff turnover [10][11][13]. Our suggested research's primary contributions to the field of EA prediction are as follows:

- This framework design and implement machine learning based supervised techniques based on IBM HR analytics EAPerformance’dataset from Kaggle with multiple attributed for employee attrition prediction**.**

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

- In order to extract useful knowledge from the dataset, the Employee Exploratory Data Analysis (EEDA) was utilised. The elements that influence the rate of EA were investigated.
- To achieve statistical equality in the dataset, resampling using the Random oversampling technique was employed. Due to the balanced data and higher predictive accuracy scores, the model forecasting difficulty is reduced;
- A comparison of the four methods used to determine of evaluating performance accuracy score using K-Fold cross-validation;
- An innovative ML approach called Random Forest Classifier (RFC) that is tree-based and used to forecast ET;
- We examined the performance validation using several assessment methodologies, including a confusion matrix and ROC curve analysis, on our suggested strategy.
- The optimum performance fit analysis approach was determined by conducting a comparison study of the deployed ML models based on their accuracy score values.
- to acquire the greatest possible accuracy scores when compared to ML approaches and modern research, we optimized the suggested RFC methodology as an invention.

This research is structured as follows for the parts that follow: **Section 2** delves into the literature that is relevant to our inquiry. the study issue description and methodological analysis; **Section 3** discusses the recommended techniques within the context of EA. In **Section 4**, we cover the outcomes and assessments of the research project we had in mind. Our research study's findings and plans for the future form **Section 5.**

II. Literature Review

In this part, we review the literature that is relevant to our investigation. Summaries of previously utilized techniques and research findings for forecasting EA provide the basis of the relevant literature. The literature review focused on the most current, applicable, currenttechniques.

The reasons for ET could be investigated through an investigation, and a learning model should be developed to forecast ET [3]. The purpose of the study was to apply ML approaches to forecast ET and examine the organisational elements that contributed to EA. A comparative analysis was conducted using the four ML methods. A 93% success rate in predicting ET was attained by the suggested enhanced Extra Trees Classifier (ETC) method. According to current state-of-the-art investigations, the suggested method performed better.

In[14],employs a total of three types of ML to forecast ET using the 35-feature IBM Watson dataset. In terms of efficiency measurements, the findings show that the Logistic Regression (LR) ML approach generate the best findings for the dataset, with an accuracy of 87% and the best recall rate (0.36).

In[15], used an ML model to forecast employee turnover and conduct a depression study based on this data. They achieved an accuracy of 86.0% for forecasting AR by applying techniques like Support Vector Machine (SVM), Decision Tree Classifier (DTC), & RFC to this dataset after completing preprocessing procedures..

manyMLapproaches [16], Our study included several statistical methods, such as LR and the RF classifier, to determine the main factors impacting ET. With an accuracy of 87% vs. 85%, the RFC model was the clear winner. In comparison to the LR model, the RFC model achieved better results in terms of recall, precision, or F1 score.

This research employs [17], prediction of an employee's departure using LR. Features that They assess include but are not limited to, job satisfaction, gender, years of employment, average monthly hours worked, and average education level. Additional approaches to this problem-solving include DTs, RF, KNN, and SVM. A LR accuracy of 88.0952% was achieved by dividing the dataset in half and using 80% of it for training and 20% for testing the method.

Methods for Predicting EA via Data Analysis & are the Main Emphasis of This Work [18]. Our findings suggest that RFC outperformed the others with an accuracy of 83.3%, but Naive Bayes NBs) and SVM scored higher for True Positive (TP) classifications and showed larger Area Under Curve (AUC) values.

This study uses ML algorithms to examine ET [6]. Three primary studies were carried out to forecast ET utilising synthetic data generated by IBM Watson. In the initial trial, the initial class-imbalanced dataset was trained using the following ML approaches: KNN, SVM, and RF using various kernel functions. Retraining the new dataset using the aforementioned ML models was the subject of the second study, which aimed to address class imbalance utilising the ADASYN technique. To achieve class parity in the third trial, data was manually undersampled. Consequently, the best performance (0.93 F1-score) was obtained by training an ADASYN-balanced dataset using KNN (K = 3). Feature selection and RF allowed us to reach an F1-score of 0.909 via 12 features out of a total of 29.

In this paper[19], In order to forecast the likelihood of ET, they use a number of ML models applied to the IBM HR Analytics Data set available on Kaggle. Using the AUC metric, researchers also evaluate the precision of these models. With RF, They identify the most important factors influencing EA, and the Xg boost categorization helps us determine which sorts of workers are most prone to leaving. They also go over several strategies that businesses may take to maintain employee interest.

A. Reseacrh gaps

Although the literature on forecasting employee attrition using machine learning approaches has provided helpful insights, there are still significant research gaps that need to be addressed. To

start, different studies seem to use different features and approaches, which causes the model performance measures to vary. Results may be more easily compared if relevant feature selection was standardized and a uniform strategy was used. Furthermore, additional research and comparisons of sophisticated machine learning algorithms are required, taking into account aspects like hybrid models, deep learning, and ensemble approaches. Organizational decision-makers rely on models' interpretability and explainability, although these aspects have received less attention in the current research. A more consistent and broadly applicable model for forecasting employee turnover might be developed if these gaps are filled up by future studies.

III. Research Methodology

In this section provide the problem statement or prposedmethodlogy for this frarmework. Section provide the each techniques of proposed system with deep discusison.

IV. Problem Statement

Employees are the backbone of every company's productivity and power. In today's cutthroat business environment, it is extremely difficult for organisations to hold on to their regular employees. One of the most significant challenges faced by HR analytics companies is ET. Businesses put a lot of money into staff training because they want to reap benefits from it down the road. The loss of a valuable employee is a missed opportunity for the business. Outstanding businesses seldom have access to competent people. Owners of businesses have the challenge of retaining skilled workers. Companies might lose a lot of money when employees leave since their knowledge and productivity are so valuable. This is why we provide a ML model in this study that can anticipate ET using a variety of predictive analytics methods.

V. Methodology

This paper proposes a method for predicting EA through the use of ML. our study process, describe in detailed. For the implementation of an effective system for predicting employee attrition use Python programming language and jupyter notebook. The research methodology is divided into many steps and phases. The first stage in using the IBM HR Analytics EA & Performance dataset for research purposes is to gather the necessary data. The Kaggle website is used to gather the dataset. Once the data has been collected, preprocess the data for data cleaning and make it more valuable. In a preprocessing step, to obtain useful insights use exploratory data analysis. When developing models and making predictions, it is helpful to extract features using feature engineering techniques that have been used to DT and SelectKBest. This is done via feature correlation. The dataset was shown to be unbalanced through analysis. Use random oversampling to even out the dataset. The next step in creating a model is to prepare a preprocessed dataset. Then, split the dataset into two types training and testing. The ratio of the split dataset is 85:15. For model building, apply machine learning techniques like a random forest classifier. We used a fully-tuned ML model. Employ performance metricx such as ROC curve, recall, accuracy, precision, and f1-score to assess a ML model. Once the information on the workers has been

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

provided, a more generic version of the suggested model may be prepared to forecast ET. In Figure 1 we can see the research study's methodology. Whole flowchart description provide in this methodology while each phases discirbed in below sebcetions.

1) Data collection

Data collection is a very initial and common step. Gather the IBM HR Analytics dataset on ET and performance for this research. The dataset in question was retrieved from the Kaggle website. Examine key questions like "Compare average monthly income by education & turnover" or "Show me a breakdown of distance from home by job role and attrition" to uncover the elements that lead to ET. The data scientists at IBM developed this hypothetical dataset. With 35 columns and 1470 rows, this dataset is rather large.

2) Data preprocessing

After data collection, move forward to the preprocessing step. Data pre-processing is ancrusialfirst phase in the knowledge discovery process. Data reduction and transformation are two of the many procedures needed. Improving the quality of raw data is crucial for ensuring that learning techniques work efficiently and accurately. As a result, selecting appropriate learning techniques and carrying out appropriate data pretreatment stages allow for accurate analysis of the acquired data [20]. The data in the present research was cleaned and transformed using many pretreatment procedures. The following are the measures to take:

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

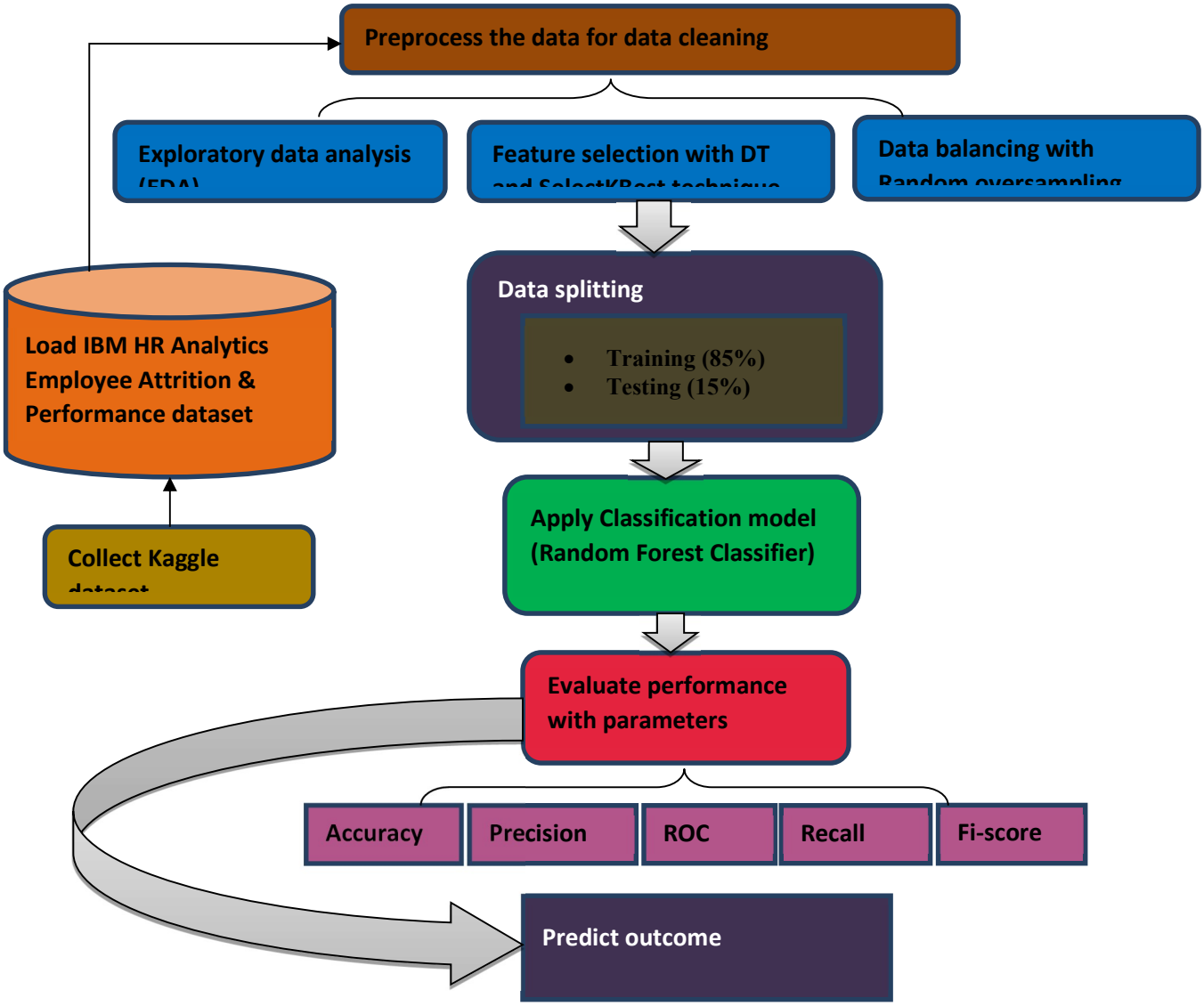


Figure 1: Proposed Flowchart for employee attrition

a) Feature Selection (SelectKBest technique)

One of the most important steps in extracting the necessary data items from a dataset using existing knowledge is feature selection (FS). There are a lot of characteristics in the dataset that are not relevant here, but we extracted the ones that help with performance monitoring and decision-making and placed them aside. Classification accuracy improves when the dataset contains only useful, highly predictable variables.Classification performance is therefore improved by having just important characteristics and decreasing the amount of superfluous attributes. The SelectKBestmethodutilize the k highest score to select features for feature extraction. This approach may be used with either classification or regression data by modifying the'score_func' option. Among the many steps involved in getting a big dataset ready for training

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

is picking the right features. It shortens the training period by allowing us to remove irrelevant data.

b) Feature Importance (Decision Tree)

The value of each characteristic in relation to your output variable is shown by its score; a greater score indicates that the feature is more significant. We will be utilizing the DTC to extract the top features for the dataset, which is an inherent class with Tree Based Classifier. Feature significance is one of such classes.

Medical diagnosis and credit risk assessment are just two examples of the many fields that have found use for DT, which are extremely strong techniques that can suit complicated datasets. For better human comprehension, DT learning approximates a goal function using a "if-then" rule tree [21]. In this way, it develops a DT for each subset of a dataset sequentially, beginning with the largest subset at the very top (the "root" node). The two types of nodes in the finished DT—decision nodes and leaf nodes—can process numerical and categorical input respectively[22]. A DT example employing the most highly linked attributes to the objective (the attrition) is shown in Fig. 4.2. It has been demonstrated,

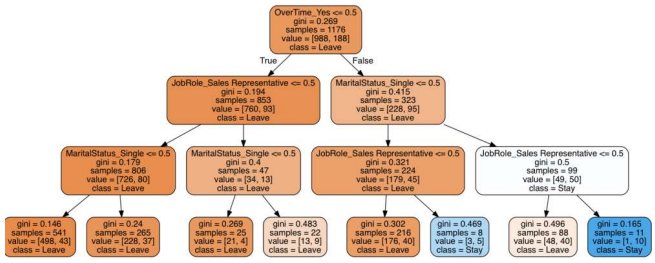


Figure 2. A decision tree generated by using extracted features from IBM HR dataset[23].

c) Data balancing (Random Oversampling)

One issue with classification challenges is imbalanced data categorization. When the classes in the training set are not evenly distributed, we say that the data is unbalanced. Particularly for minority groups, this can cause models to have poor predictive accuracy[24]. Random oversampling methods can be used to tackle this problem of class imbalance. The practice of random oversampling involves augmenting the training data with more instances of specific minority classes. We can pick several at random and replace them in the minority class instead of recreating each duplicate. It is possible to execute oversampling many times. The fact that it is one of the oldest techniques still in use speaks much about its reliability[25].

3) Data splitting

In every implementation, data splitting is a crucial responsibility. Data splitting refers to the procedure of dividing a dataset into multiple sections. Separate the dataset into a training set

and a testing set for the study. The dataset is split between training data making up 85% and testing data making up 15%.

4) Classification techniques

Select ML approaches for the aim of constructing models. In ML, computers take in data as input and use statistical analysis to learn new ideas, such as how to categorize and forecast data. For this research, the RFC will be used to forecast ET.

1) Random forest classifier

One kind of supervised classification system is the RF technique. By merging several DTs, the RF idea may be used to construct DTs. DTCs and other single-tree classifiers are susceptible to problems like noisy or outlier data, which can impact the accuracy of the classifier. In contrast, RFC include randomization, making them extremely resistant to noise and outliers [26]. Both data randomness and function randomness are produced by this classifier. Both of these types of unpredictability are distinct entities. Considering that it is utilized to combine numerous DTs, this classifier possesses a lot of hyperparameters. The DT is a tool that assists in making decisions. In the realm of ML with supervision, the RF method is widely considered to be among the most powerful methods for producing classifications and regressions. In order to train data, RF makes use of numerous DT. After each tree casts a vote for a classification label for a particular dataset, the RF model determines which category received the most votes from the DTs[27]. It displays the potential outcomes using a graph that resembles a tree. The DT will generate a set of rules when fed a training dataset containing features and objectives. Predictions may be made using these Rules.

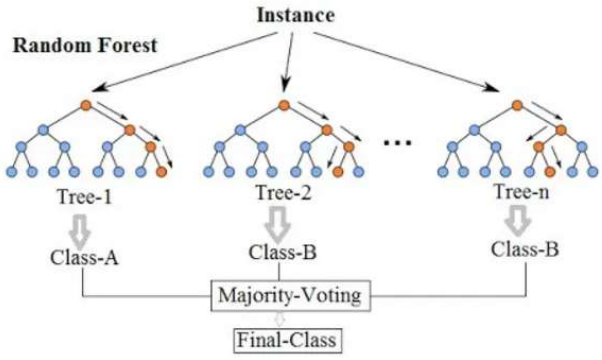


Figure 3: Random forest workflow

To determine the optimal course of action, RFs construct DTs using randomly choose data samples, collect forecasts from each tree, and finally cast a vote. It is also possible to correctly assess the value of the characteristic. It may be expressed using this formula [28]:

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \dots\dots\dots\dots (1)$$

In this equation, the symbol nij represents the importance of node j, the symbol wj represents the weighted number of samples that arrived at node j, the symbol Cj represents the impurity value of node j, the symbol left(j) represents the child node from the left split on node j, and the symbol right(j) represents the child node from the right split on node j.

5) Model building

The scikit-learn package in Python is used to carry out a number of essential phases in the process of constructing a RF model. Importing necessary libraries is the first step. These libraries include the RandomForestClassifier from the sklearn.ensemble package and a variety of evaluation metrics from the sklearn.metrics package. The Random Forest Classifier is then initialised with certain parameters such as 'n_estimators' (which indicates the number of trees in the forest, which is set to 10 in this instance) and 'random_state' (which ensures repeatability). In the next step, the model is trained by applying the fit technique to the training data, which consists of X_train and y_train. Through the use of the cross_val_score function, a 10-fold cross-validation is carried out in order to undertake an accurate assessment of the model's performance. The training data is divided into ten folds by this function, which then trains and evaluates the model on each fold in an iterative manner. Finally, the function outputs a list of accuracy scores for additional analysis. The building and assessment of a Random Forest model for predictive tasks is made possible by this all-encompassing method.

6) Model Evaluation

For the purpose of model evaluation, use some parameter for check efficiency of our model. Performance metrics were utilized in order to conduct an analysis of how well the model performed on the testing set. many different parameters, including F1 score, accuracy, precision, recall, and ROC score, are using to predicting employee attrition.

VI. Proposed Algorithm

Give a rundown of the suggested algorithms that were utilized to execute the EA prediction in the present research.

Proposed Algorithm: predicting employee attrition

Step 1: Install python simulation tool and jupyter notebook

- Import python libraries includingseaborn, Pandas, NumPy, Matplotlib. pyplot, etc.

Step 2: Data Collection

- Firstly, collect the IBM HR Analytics EA& Performance dataset.
- Dataset collect from Kaggle online resource.

Step 3: Data Preprocessing

- EDA (Exploratory Data Analysis)

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

- SelectKBest technique for Feature Selection

Step 5: data balancing

- Random Oversampling

Step 4: Data Splitting

- Training (85%)
- Testing (15%)

Step 5: classification model

- Apply machine learning model for employee attrition prediction.
- Random Forest Classifier

Step 6: Model Training

- Train the model for implementation of the employee attrition.

Step 7: Model Evaluation

- Accuracy
- Precision
- Recall
- F1-Score

Step 8: predict outcome

VII. RESULTS & DISCUSSIONS

various research are conducted to train the RF model utilizing the suggested ML methodology. The dataset used for these studies is the Kaggle dataset for IBM HR analytics EA performance. To provide a consistent and impartial assessment of the suggested RF model, a uniform methodology was employed for both the training and testing phases. A GPU (more especially, an NVIDIA Tesla T4 GPU) and fourteen gigabytes of DDR4 RAM are required to execute the suggested methods. The solution makes use of many Python-based libraries, such as matplotlib, SK-Learn, pandas, numpy, seaborn, etc. Ten K-fold cross validation was employed to train the suggested model. The five metrics employed for assessment in the present research were F1-score, precision, recall, accuracy, & ROC. All implemented results on train and test dataset provide below with using performance measures, also provide the dataset description and visualization results with EDA.

VIII. Dataset Description

Table 1 shows the characteristics extracted from the Kaggle-obtained IBM HR Analytics EA & Performance dataset. To find out what causes ET, you may ask questions like "Compare average monthly income by education and attrition" or "Show me the breakdown of distance from home by job role and attrition." The data scientists at IBM developed this hypothetical dataset. There are

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

a total of 1470 rows and 35 columns in the dataset. The EA & Performance dataset from IBM HR Analytics has the following important features:

Table 1: Dataset parameters

No.	Education	Environment Satisfaction	Job Involvement	Job Satisfaction	Performance Rating	Relationship Satisfaction	Work Life Balance
1.	'Below College'	'Low'	'Low'	'Low'	'Low'	'Low'	'Bad'
2.	'College'	'Medium'	'Medium'	'Medium'	'Good'	'Medium'	'Good'
3.	'Bachelor'	'High'	'High'	'High'	'Excellent'	'High'	'Better'
4.	'Master'	'Very High'	'Very High'	'Very High'	'Outstanding'	'Very High'	'Best'
5.	'Doctor'	-	-	-	-	-	-

The inclusion of these attributes to the dataset provides extra context, which enables a more nuanced view of the factors that influence employee performance and turnover. Some examples of elements that may play a significant influence in moulding the entire work experience include education level, job happiness, and the ability to maintain a healthy work-life balance. These factors can also have an effect on employee attrition rates and job performance.

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

IX. EDA (Exploratory Data Analysis)

A technique to data summarization known as EDA involves identifying the data's primary properties and then using appropriate representations to display them. EDA provides a concise overview of the data collection, including its size, kinds, missing data, and columns and rows. Find and fix missing data, incorrect data types, and incorrect values; eliminate inaccurate data. Bar graphs, histograms, or box plots are the visual representations of data distribution provided by EDA. Find the relationships (correlations) between the variables and show them on a heat map. Dataset information with an in-depth analysis for ET is presented in the following graphic representations:

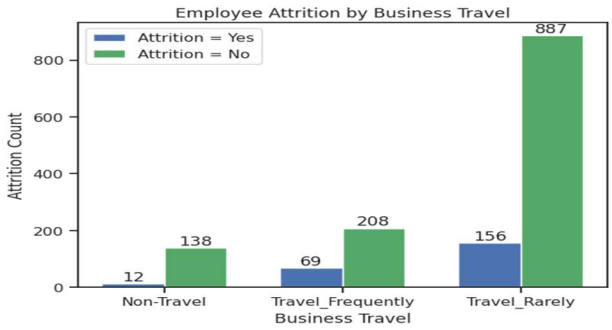


Figure 4: Bar graph of Employee Attrition by Business Travel

Figure 4 shows the plot a bar graph between Attrition Count(Y-axis) and Business Travel (x-axis). In figure there are three business travel such as non-travel, travel frequency and travel rarely. The last travel rarely gets highest no attrition with 887 counts and lower attrition count is 12 non-travel.

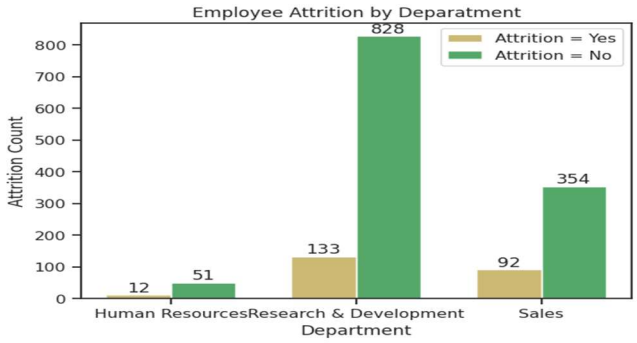


Figure 5: Bar graph of Employee Attrition by Department

Figure 5 shows the plot a bar graph between Attrition Count(Y-axis) and Department(x-axis) shows the attrition count between different department. The highest attrition of no research development with of 828 count, and lower attrition of human resource with of 12 counts.

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

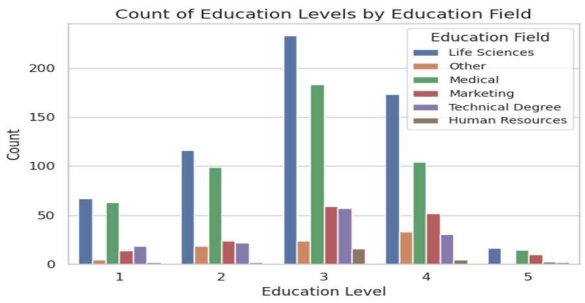


Figure 6: Bar graph of Employee Attrition by Department

Figure 6shows the education Level counts bar graph with different education levels and fields. There are five education level the highest level of life science with 250 approx. counts. And lower count plot of human resources with 5thlavel.

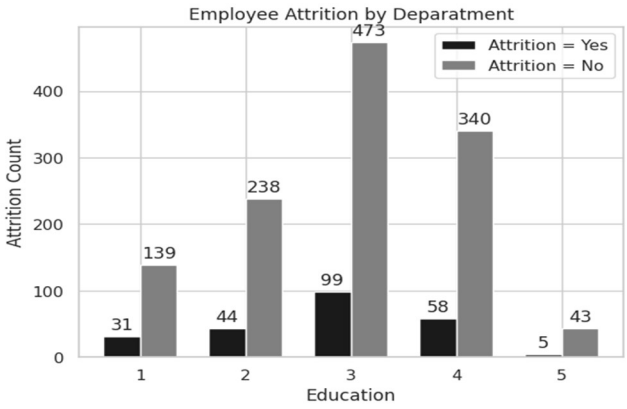


Figure 7: Histogram of Employee Attrition by Department

In figure 7 shows the histogram of attrition count vs education attribute on different education field where, education mapping = {1: 'Below College', 2: 'College', 3: 'Bachelor', 4: 'Master', 5: 'Doctor'}. The highest histogram of No attrition with count values 473, respectively.

Figure 8: Scatter plot of Monthly Income vs Total Working Years.

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

The above figure 8shows the Scatter plot of Monthly Income vs Total Working Years. In figure x-axis shows the total working years and monthly income shows in y-axis.

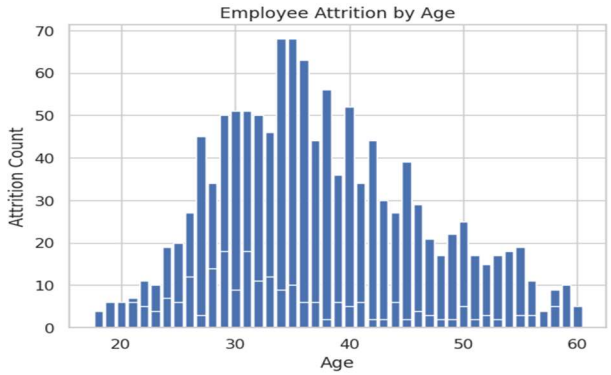


Figure 9: Employee Attrition by age shows attrition count of different ages.

Figure 9 shows the employee Attrition by age shows attrition count of different ages. In figure x-axis shows the age with 10 to 60 and y-axis shows the attrition counts. The highest attrition by age of 35 approx.

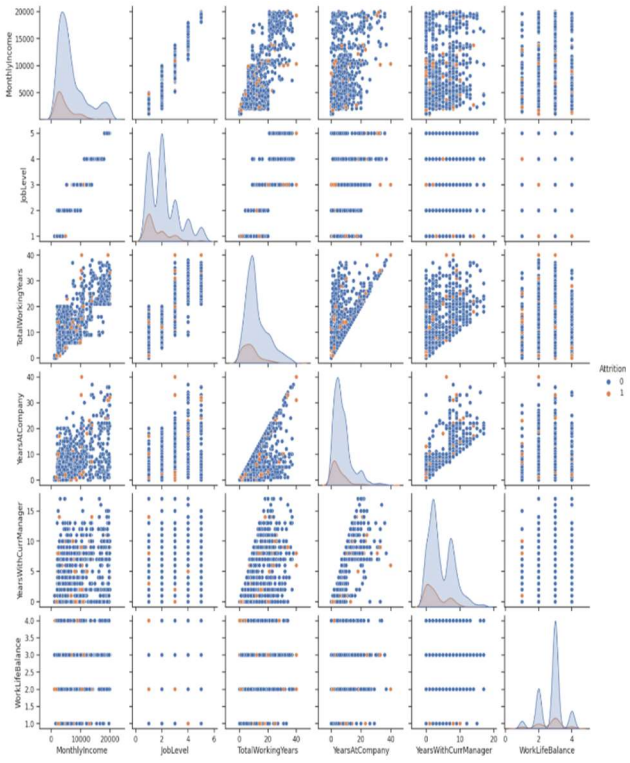


Figure 10: Pair plot of different column.

A matrix of scatterplots that illustrates the correlations between various pairs of variables is known as a pair plot. This is shown in figure 10, which can be found above. Its purpose is to facilitate the visualisation of potential correlations or trends.

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

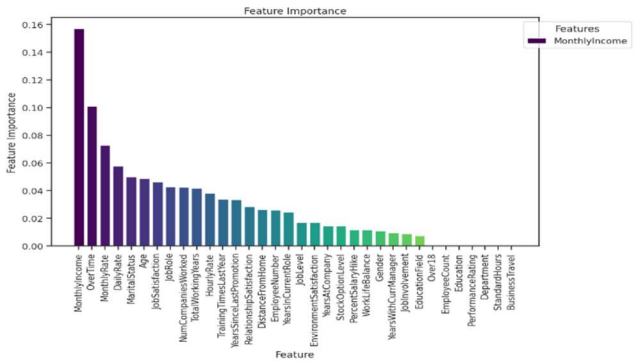


Figure 11: Feature Importance graph of different features in dataset for feature selection through Decision Tree algorithm

The above figure 11present the feature importance graph of different features in dataset for feature selection through decision tree algorithm. Through the use of the Decision Tree technique, the following image shows a graph that illustrates the importance of different features included inside the dataset. When it comes to anticipating employee attrition, it is helpful to have an awareness of what features are the most prevalent.

```
Index(['JobLevel', 'MaritalStatus', 'MonthlyIncome', 'OverTime',
       'TotalWorkingYears', 'YearsInCurrentRole'],
      dtype='object')
```

Figure 12: Mine selected features with SelectKBest feature selection

A different approach to feature selection than the Decision Tree method is shown by this Figure 12, which advocates the use of the SelectKBest technique. It is possible that this method produced better results in terms of picking relevant characteristics, since it was said that the accuracy was enhanced.

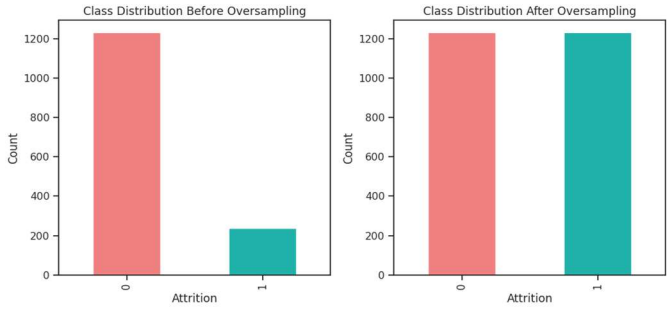


Figure 13: Balanced Data by using Random Oversampling

The above figure 13 shows the Balanced Data by using Random Oversampling in form of bar graphs. First bar graph shows data imbalanced and second graph shows the balanced dataset. The effect that random oversampling has on the balance of dataset is probably shown by this picture.

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

Specifically, in context of forecasting attrition, random oversampling is a strategy that is used to alleviate class imbalance.

X. Performance Measures

Metrics for assessment provide light on how well a model performs. One important characteristic is that assessment measures can distinguish between various model outcomes. Measures for evaluating the proposed approach for EA in the present research include ROC metrics, recall scores, accuracy scores, precision scores, and confusion matrices.

1) Confusion Matrix

N is the number of classes being predicted, and the resulting matrix is N X N. We will be using a confusion matrix similar to Table 2 for this topic. The efficacy of an algorithm can be better understood using either the table or matrix structure. You may think of a confusion matrix as a table that displays the expected value (TN) alongside the actual value (TP).

Table 2: Representation of cells in confusion matrix

	Predicted:0	Predicted:1
Actual:0	TN	FP
Actual:1	FN	TP

The first component, referred to as TP, involves the identification of values that are acknowledged as accurate and indeed, they were accurate. The second category is referred to as a false positive (FP), which arises when inaccurate data points are mistakenly classified as true. False negative (FN) is a term used to describe a scenario in which a verifiable truth is mistakenly categorized as a bad outcome. The fourth category is referred to as a TN when the numerical number is actually negative. The following performance measures shows below:

2) Accuracy Score

As a percentage, accuracy measures how many forecasts were precisely correct of all of them. Eq. (2) was used to determine accuracy.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{Fp} + \mathrm{TN} + \mathrm{FN}} \dots\dots\dots\dots\dots\dots (2)$$

3) Precision Score

Precision is the proportion of positive cases that were identified correctly. The accuracy of a prediction model is demonstrated by its precision score. We used equation (3) to get the precision.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \dots\dots\dots\dots\dots\dots (3)$$

4) Recall Score

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

"Recall" means the proportion of true positives that are correctly identified. Equation (4) was used to compute recall.

$$Recall = \frac{\text{TP}}{TP + FN} \dots\dots\dots\dots\dots\dots (4)$$

5) F1 Score

The F1-Score is the harmonic mean of the recall and accuracy scores for a classification issue. The formula that was used to determine F1 is (5).

$$F1 - Score = \frac{2(Precision * Recall)}{\text{Precision} + \text{Recall}} \dots\dots\dots\dots\dots\dots (5)$$

The ROC curve plots the ratio of true positives (Sensitivity) to FPs (100 - Specificity) as a function of the distinguishing threshold. The receiver operating characteristic (ROC) curve shows the relationship between decision thresholds, sensitivity, or specificity.

XI. Experimental Analysis of proposed model with train and test dataset

The analysis presented in the following section pertains to the outcomes derived from the conducted experiments for employee attrition. The results of the investigation are displayed using the Python programming language. The Kaggle website was utilised to acquire the IBM HR Analytics EA& Performance dataset utilized in this study. The forthcoming section will undertake a thorough analysis of the research findings for employee attrition. The outcomes of the simulations generated by the random forest method with cross validation under consideration are encouraging for employee attrition. Furthermore, the proposed model undergoes evaluation through the utilisation of performance metrics including f1-score, accuracy, precision, recall, and PROC. This section outlines the results obtained from the classification procedure, which will be further expounded upon in the following discourse.

Table 3: Train and Test Parameter Performance of Random Forest (RF) for employee attrition

Parameters	Training performance	Testing performance
Accuracy	100	96
Precision	100	96
Recall	100	96
F1-score	100	96
ROC	100	100

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

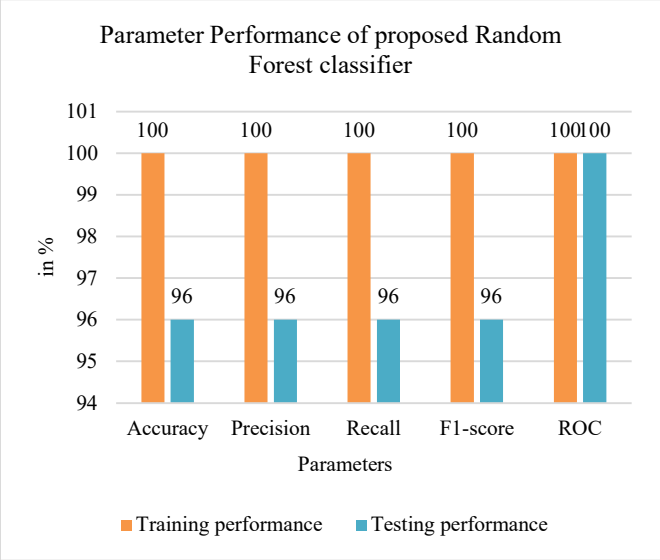


Figure 14: Bar graph of train and test performance of RF model with classification parameters

Table 3 shows Random Forest (RF) model training and testing performance across categorization parameters for employee attrition. This model performs in several parameters, including recall, accuracy, precision, F1-score, and ROC. Training for EA yields perfect results for the RF model in terms of recall, accuracy, precision, F1-score, and ROC. A 96% F1-score, recall, accuracy, and precision indicate that the model generalises effectively to new data on EA. The 100% ROC score indicates a high true positive rate and low false positive rate for employee attrition. Figure 14 shows the RF model's consistency and efficacy in classification tasks via a bar graph of train and test performance for employee attrition.

```
Accuracy: 0.9594594594594594
Classification Report: on test data
              precision    recall  f1-score   support

           0       0.97      0.95      0.96       184
           1       0.95      0.97      0.96       186

    accuracy                           0.96       370
   macro avg       0.96      0.96      0.96       370
weighted avg       0.96      0.96      0.96       370
```

Figure 15: Classification report on testing data of proposed random forest classifier

The figure 15 shows the Classification report on testing data of proposed random forest classifier with IBM-HR-Analytics-Attrition-Dataset that contain two classes yes or no. The no class model obtains 97% precision, recall 95%, and f1-score 96% with support 184, while proposed model obtain on class yes precision of 95%, recall 97% and f1-score 96% with support 186, respectively. The proposed model obtains overall classification results on train dataset performance in all parameters 96% with support 370, respectively.

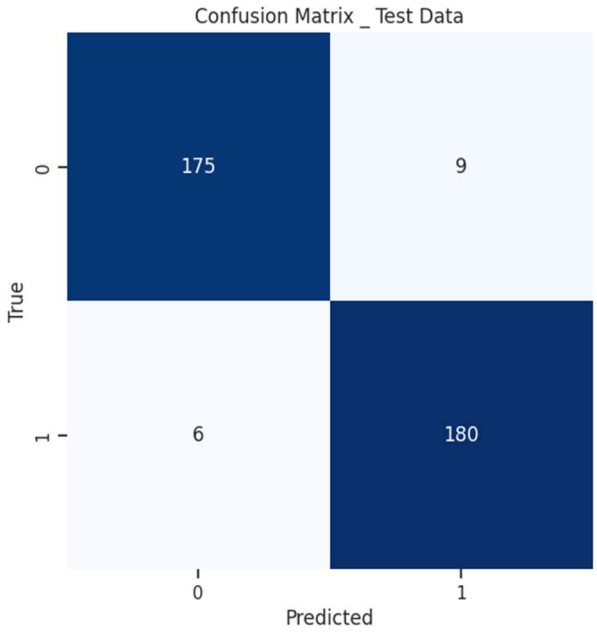


Figure 16: Confusion matrix on testing data of proposed random forest classifier

The proposed model testing confusion metrix shows in figure 16 for employee attrition. In this matrix 0,0 = True negative, 0,1 = False negative, 1,0 = False positive, and 1,1 = true positive. The proposed model predicted true prediction instance of 175, true negative instance of 180, while false positive instance of 9 and false negative instance of 6, respectively.

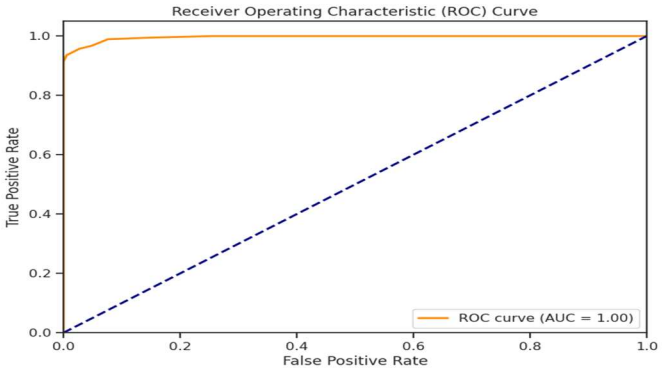


Figure 17: ROC curve on testing data of proposed random forest classifier

The figure 17shows the ROC curve on testing data of proposed random forest classifier for employee attrition. In figure x-axis shows the FPR value of ROC and y-axis shows the TPE value of ROC. This is a graph that illustrates how well a system for classification performs across all of the categorization thresholds. Proposed Random forest classifier obtain 100% ROC curve performance on input dataset for employee attrition.

```
Accuracy on Training Data: 0.9985687022900763
Classification Report on Training Data:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1049
           1       1.00      1.00      1.00      1047

    accuracy                           1.00      2096
   macro avg       1.00      1.00      1.00      2096
weighted avg       1.00      1.00      1.00      2096
```

Figure 18: Classification report on train data of proposed random forest classifier

The figure 18 shows the Classification report on train data of proposed random forest classifier. The proposed model obtains overall classification results on train dataset performance in all parameters 100% with support 2096, respectively, for employee attrition.

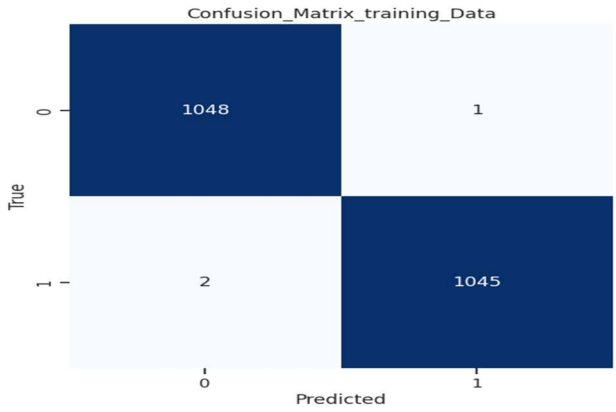


Figure 19: Confusion metrix on train data of proposed random forest classifier

The proposed model train confusion metrix shows in figure 19, for employee attrition. In this matrix 0,0 = TP, 0,1 = FN, 1,0 = FP, and 1,1 = TP. The proposed model predicted true prediction instance of 1045, true negative instance of 1045, while false positive instance of 1 and false negative instance of 2, respectively.

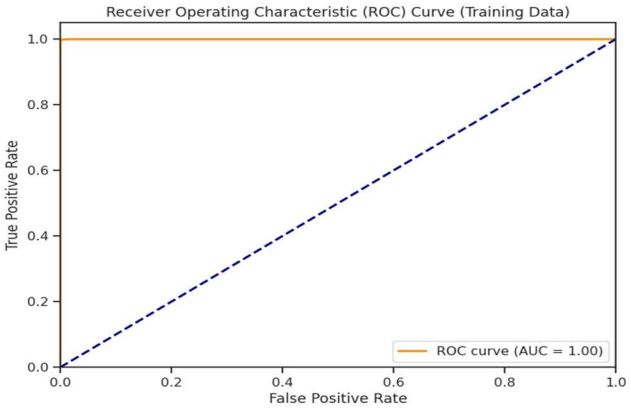


Figure 20: ROC curve on train data of proposed random forest classifier

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

TPR versus FPR are shown against one another at various categorization levels in a ROC curve. There will be more TPs and FPs as a consequence of a decrease in the classification threshold since more objects will be classed as positive. Figure 20 shows a typical ROC curve for ET, which was generated with a model performance of 100%.

XII. Comparison Between Base and Proposed Models and Discussion

Table 4compares Base Models (Extra Tree, Logistic Regression, Decision Tree) with a Proposed model employing Random Forest (RF) across key categorization parameters to handle Employee Attrition. The Extra Tree (ET) model has the best accuracy among basic models at 91%, followed by the DTand LR models at 81% and 73%, respectively. The suggested Random Forest model surpasses all basic models with 96% accuracy, precision, recall, and a 100% F1-score. It seems that the Random Forest model is better at forecasting and treating staff attrition than the basic models. Ensemble approaches like Random Forest may improve employee attrition prediction performance, providing a solid solution for workforce management organizations seeking accurate and dependable forecasts.

Table 4: Base and Proposed comparison with classification parameters for Employee Attrition

Parameters	Base Models			Proposed
	ET	LR	DT	RF
Accuracy	91	73	81	96
Precision	90	73	80	96
Recall	91	72	84	96
F1-score	91	73	82	100

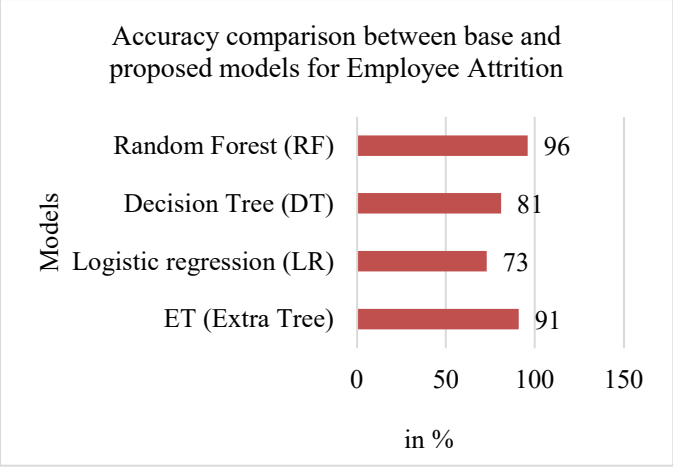


Figure 21: Bar graph of accuracy performance for base and proposed models' comparison

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

Within the context of Employee Attrition, this figure 21, presents a comprehensive comparison of the accuracy performance of several models. The Extra Tree (ET) model obtains an accuracy of 91%, the Logistic Regression (LR) model scores 73%, the Decision Tree (DT) model reaches 81%, and the RF model that was developed performs substantially better than all of these models with an accuracy of 96%. The bar graph provides a clear and compelling illustration of the significant improvement that the RF model brings to the table in terms of properly forecasting employee turnover in comparison to the base models.

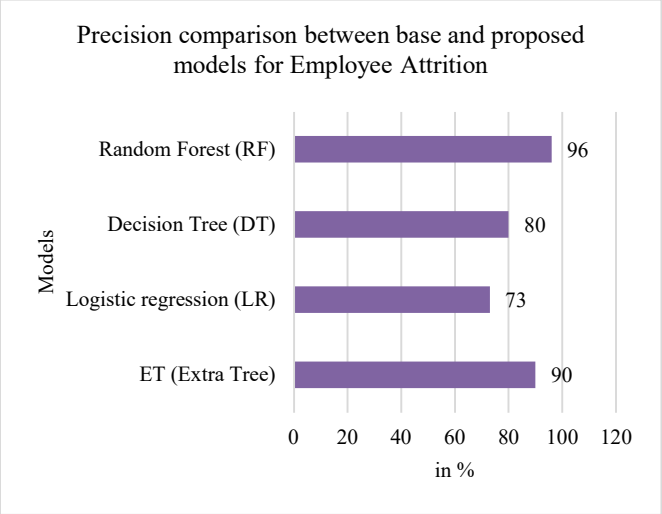
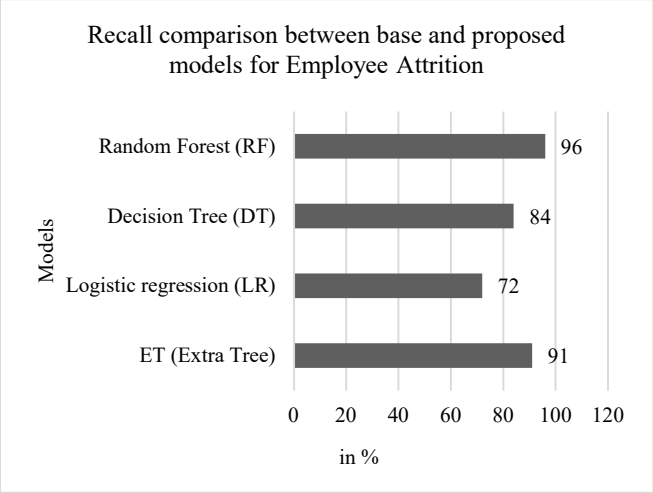


Figure 22: Bar graph of precision performance for base and proposed models' comparison

Looking at the performance of precision, the Extra Tree (ET) model has a precision of 90%, the Logistic Regression (LR) model has a precision of 73%, the Decision Tree (DT) model has a precision of 80%, and the Random Forest (RF) model that was suggested is exceptional with a precision that reaches 96%. The Figure 22provides a visual representation of the RF model's superiority in significantly reducing the number of false positives, which makes it especially successful in detecting workers who are really at danger of leaving their positions.



Vol. 21, No. 1, (2024)
ISSN: 1005-0930

Figure 23: Bar graph of recall performance for base and proposed models' comparison

The base ET model obtains a recall performance of 91%, the LR model achieves 72%, the DT model reaches 84%, and the RF model that was suggested retains a high recall score of 96%. The dependability of the RF model in reducing the number of false negatives is shown by this Figure 23, which illustrates the model's capacity to detect a significant percentage of workers who are really at danger of leaving their positions.

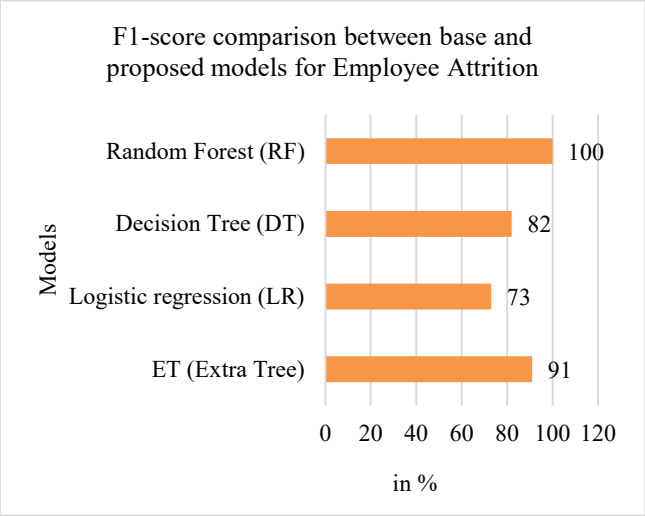


Figure 24: Bar graph of f1-score performance for base and proposed models' comparison

The F1-score graph reveals that the Extra Tree (ET) model received a perfect score of 91%, the Logistic Regression (LR) model received 73%, the DT model received 82%, and the RF model received an astounding 100%, shows in Figure 24. This number highlights the remarkable balance between accuracy and recall that the RF model has, which makes it an appealing option for businesses that are looking for a complete and accurate predictive model for controlling employee turnover.

In the discussion and evaluation of the findings focus around the Random Forest (RF) model that was created for forecasting employee attrition. We compare this model to some of the most fundamental ones, including DT, LR, and ET. Employing the EA & Performance dataset from IBM HR Analytics, the EDA provides comprehensive understanding of several factors impacting ET. These insights are displayed via bar graphs, scatter plots, and histograms. In the part that focuses on performance measurements, the evaluation metrics that were used are described in depth. Some examples of these measures involve the F1-score, confusion matrix, precision, recall, accuracy, and ROC metrics. Exhibiting remarkable performance in both the training and testing phases, the suggested RF model achieves a perfect score of one hundred percent across all metrics during the training phase and maintains high scores of ninety-six percent throughout the testing phase. The powerful prediction capabilities of the model are further validated by classification reports, confusion matrices, and ROC curves with further validation. The comparison analysis

proved that the RF model outperformed the basic models in every metric: recall, accuracy, precision, or F1-score. The RF model is positioned as a powerful instrument for workforce management as a result of this comprehensive analysis. It provides organisations with a dependable solution for forecasting and resolving employee's attrition.

XIII. CONCLUSION AND FUTURE WORK

In conclusion, the Random Forest (RF) model that was suggested demonstrates a balanced and accurate approach, demonstrating solid performance in forecasting staff attrition. This research highlights the better capabilities of the model via the provision of a detailed analysis as well as graphic representations. By utilizing the 1470-row and 35-column IBM HR Analytics EA & Performance dataset, a substantial understanding of the factors impacting ET has been achieved. With its performance of one hundred percent across a variety of measures during training and its maintained high accuracy of ninety-six percent during testing, the RF model appears as a viable tool for organizations that are looking to solve difficulties related to workforce management and improve retention tactics.

The outcomes of our study assist organizations in overcoming the problem of staff turnover. In light of the limits of the research and the route that we want to take in the future, we will use deep learning methods in order to forecast staff turnover. Furthermore, we will improve the dataset feature space in order to produce findings that are more accurate by using deep learning approaches.In terms of future development, there are a number of different paths that may be taken to improve and expand. It is possible that the flexibility and predictive capacity of the model might be improved by including more datasets from a variety of sources. This would enable the model to capture a wider range of characteristics that would influence employee turnover.

References

[1] S. Krishna and S. Sidharth, "Workforce analytics: Predicting employee attrition using machine learning approach," in *Building Resilient Organizations*, 2022. doi: 10.4324/9781003313663-16.

[2] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, "Predicting Employee Attrition Using Machine Learning Approaches," *Appl. Sci.*, 2022, doi: 10.3390/app12136424.

[3] S. George, K. A. Lakshmi, and K. T. Thomas, "Predicting Employee Attrition Using Machine Learning Algorithms," in *Proceedings - 2022 4th International Conference on Advances in Computing, Communication Control and Networking, ICAC3N 2022*, 2022. doi: 10.1109/ICAC3N56670.2022.10074131.

[4] P. Gangrade and N. Vijayvargiya, "Analysis of Employee Attrition Rate in Indian IT Industry and Prediction using Machine Learning Approach," *Int. J. Innov. Res. Sci. Eng. Technol. | An ISO*, 2021.

[5] P. K. Jain, M. Jain, and R. Pamula, "Explaining and predicting employees' attrition: a machine learning approach," *SN Appl. Sci.*, 2020, doi: 10.1007/s42452-020-2519-4.

[6] S. S. Alduayj and K. Rajpoot, "Predicting Employee Attrition using Machine Learning," in *Proceedings of the 2018 13th International Conference on Innovations in Information Technology, IIT 2018*, 2018. doi: 10.1109/INNOVATIONS.2018.8605976.

[7] S. Y. Bansal, B. Kaur, and J. R. Saini, "A Novel Optimized Approach for Machine Learning Techniques for Predicting Employee Attrition," in *2022 International Conference on Smart Generation Computing, Communication and Networking, SMART GENCON 2022*, 2022. doi: 10.1109/SMARTGENCON56628.2022.10084237.

[8] X. Jia *et al.*, "Mapping soil pollution by using drone image recognition and machine learning at an arsenic-contaminated agricultural field," *Environ. Pollut.*, 2021, doi: 10.1016/j.envpol.2020.116281.

[9] N. Reshma Ramchandra and C. Rajabhushanam, "Machine learning algorithms performance evaluation in traffic flow prediction," in *Materials Today: Proceedings*, 2021. doi: 10.1016/j.matpr.2021.07.087.

[10] N. Aljedani, R. Alotaibi, and M. Taileb, "HMATC: Hierarchical multi-label Arabic text classification model using machine learning," *Egypt. Informatics J.*, 2021, doi: 10.1016/j.eij.2020.08.004.

[11] I. J. Tsai, W. C. Shen, C. L. Lee, H. D. Wang, and C. Y. Lin, "Machine Learning in Prediction of Bladder Cancer on Clinical Laboratory Data," *Diagnostics*, 2022, doi: 10.3390/diagnostics12010203.

[12] L. S. Ganthi, Y. Nallapaneni, D. Perumalsamy, and K. Mahalingam, "Employee Attrition Prediction Using Machine Learning Algorithms," in *Lecture Notes in Networks and Systems*, 2022. doi: 10.1007/978-981-16-5120-5_44.

[13] S. Aggarwal, M. Singh, S. Chauhan, M. Sharma, and D. Jain, "Employee Attrition Prediction Using Machine Learning Comparative Study," in *Smart Innovation, Systems and Technologies*, 2022. doi: 10.1007/978-981-16-6482-3_45.

[14] M. Maharana, R. Rani, A. Dev, and A. Sharma, "Automated Early Prediction of Employee Attrition in Industry Using Machine Learning Algorithms," in *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO 2022*, 2022. doi: 10.1109/ICRITO56286.2022.9965017.

[15] R. Joseph, S. Udupa, S. Jangale, K. Kotkar, and P. Pawar, "Employee attrition using machine learning and depression analysis," in *Proceedings - 5th International Conference on Intelligent Computing and Control Systems, ICICCS 2021*, 2021. doi: 10.1109/ICICCS51141.2021.9432259.

[16] S. Sharma and K. Sharma, “Analyzing Employee’s Attrition and Turnover at Organization Using Machine learning Technique,” in *2023 3rd International Conference on Intelligent Technologies, CONIT 2023*, 2023. doi: 10.1109/CONIT59222.2023.10205676.

[17] S. Gupta, G. Bhardwaj, M. Arora, R. Rani, P. Bansal, and R. Kumar, “Employee Attrition Prediction in Industries using Machine Learning Algorithms,” in *Proceedings of the 17th INDIACom; 2023 10th International Conference on Computing for Sustainable Global Development, INDIACom 2023*, 2023.

[18] N. Bhartiya, S. Jannu, P. Shukla, and R. Chapaneri, “Employee Attrition Prediction Using Classification Models,” in *2019 IEEE 5th International Conference for Convergence in Technology, I2CT 2019*, 2019. doi: 10.1109/I2CT45611.2019.9033784.

[19] N. Darapaneni *et al.*, “A Detailed Analysis of AI Models for Predicting Employee Attrition Risk,” in *IEEE Region 10 Humanitarian Technology Conference, R10-HTC*, 2022. doi: 10.1109/R10-HTC54060.2022.9929893.

[20] Q. Li *et al.*, “Using fine-tuned conditional probabilities for data transformation of nominal attributes,” *Pattern Recognit. Lett.*, 2019, doi: 10.1016/j.patrec.2019.08.024.

[21] J. R. Quinlan, “Induction of decision trees,” *Mach. Learn.*, 1986, doi: 10.1007/bf00116251.

[22] C. Sehra, “Decision Trees Explained Easily,” *Medium*, 2018.

[23] A. Qutub, A. Al-Mehmadi, M. Al-Hssan, R. Aljohani, and H. S. Alghamdi, “Prediction of Employee Attrition Using Machine Learning and Ensemble Methods,” *Int. J. Mach. Learn. Comput.*, vol. 11, no. 2, pp. 110–114, 2021, doi: 10.18178/ijmlc.2021.11.2.1022.

[24] Y. Sun, A. K. C. Wong, and M. S. Kamel, “Classification of imbalanced data: A review,” *Int. J. Pattern Recognit. Artif. Intell.*, 2009, doi: 10.1142/S0218001409007326.

[25] B. Das, N. C. Krishnan, and D. J. Cook, “RACOG and wRACOG: Two probabilistic oversampling techniques,” *IEEE Trans. Knowl. Data Eng.*, 2015, doi: 10.1109/TKDE.2014.2324567.

[26] A. Ravi, A. R. Khettry, and S. Yelandur Sethumadhavachar, “Amazon reviews as corpus for sentiment analysis using machine learning,” in *Communications in Computer and Information Science*, 2019. doi: 10.1007/978-981-13-9939-8_36.

[27] V. Nagadevara, V. Srinivasan, and R. Valk, “Establishing a link between employee turnover and withdrawal behaviours: application of data mining techniques,” *Res. Pract. Hum. Resour. Manag.*, 2008.

[28] A. P. Rodrigues *et al.*, “Real-Time Twitter Spam Detection and Sentiment Analysis using Machine Learning and Deep Learning Techniques,” *Comput. Intell. Neurosci.*, 2022, doi: 10.1155/2022/5211949.

Vol. 21, No. 1, (2024)
ISSN: 1005-0930