**A close up of a sign

Description automatically generated**

Pace University

New York

Presented by,

Snehal Rode(U01835898)

**Employee attrition analysis using predictive techniques**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Index** | **Content** | **Page Number** |
| 1. | Introduction | 1 |
| 2. | Problem Statement | 1 |
| 3. | Dataset Preparation and Transformation | 4 |
| 4. | Approach to solve problem | 9 |
| 5. | Result Analysis | 13 |
| 6. | Conclusions | 17 |
|  | Appendix |  |

**ACKNOWLEDGEMENTS**

Foremost, I would like to express my sincere gratitude to my advisor Prof. David Benjamin for the continuous support of my study and giving me right direction, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of study through the necessary videos and important links and writing of this thesis. I could not have imagined having a better advisor and mentor for my Data mining study.

Thanks sir for lovely videos and amazing assignments and projects that made my interest in this field even more!

Regards,

Snehal Rode [U01835898]

1. Introduction

Around 4500 people are employed by the one of the huge corporation at any given time. However, about 20% of its workforce departs each year. Each of them must be filled from the skill pool of candidates on the job market. The management thinks that this degree of attrition—employees quitting either on their own volition or as a result of being fired—is detrimental to the business for the following reasons:

1. The projects of the former employees are delayed, which makes it challenging to fulfill deadlines and damages the

company's reputation with customers and partners.

1. In order to attract new talent, a sizable department must be maintained.
2. The majority of the time, new hires need to be taught for the position and/or given time to adjust to the organization.

So to avoid above situations, if management gets the right prior prediction then they can able to find the ways to retain the employee similarly they will get an idea of changing the workplace enviornment.

# Problem Statement

Classify the employees those are at higher attrition risk( meaning they are leaving the job) based on various categories and combination of attributes. Additionally, also find out important factors that are influencing employess to leave organization using appropriate feature selection. The results thus obtained will be used by the management to understand what changes they should make to their workplace, in order to get most of their employees to stay.

**Dataset Description**

**Dataset Name:** HR Analytics Case Study

**URL:** https://www.kaggle.com/datasets/vjchoudhary7/hr-analytics-case-study

There are a total of 3 data files used in this study. The following are the key attributes of each of them.

(i) employee\_survey\_data.csv

|  |  |
| --- | --- |
| **Attribute** | **Explanation** |
| EmployeeID | Unique Employee ID |
| EnvironmentSatisfaction | Employee's Work Environment Satisfaction Level  (1: 'Low' , 2: 'Medium' , 3: 'High' , 4: 'Very High') |
| JobSatisfaction | Employee's Job Satisfaction Level  (1: 'Low' , 2: 'Medium' , 3: 'High' , 4: 'Very High') |
| WorkLifeBalance | Employee's Work Life Balance Rating Level  (1: 'Low', 2: 'Medium', 3: 'High' , 4: 'Very High') |

(ii) manager\_survey\_data.csv

|  |  |
| --- | --- |
| **Attribute** | **Explanation** |
| EmployeeID | Unique Employee ID |
| JobInvolvement | Employee's Job Involvement Level  (1: 'Low', 2: 'Medium', 3: 'High', 4: 'Very High') |
| PerformanceRating | Employee's performance rating last year |

(iii) general\_data.csv

|  |  |
| --- | --- |
| **Attribute** | **Explanation** |
| Age | Age of the employee |
| Attrition | Whether the employee has left the organisation |
| BusinessTravel | How frequent the employee travelled for business in the last year |
| Department | Employee's department |
| DistanceFromHome | Distance between Office and Employee's home (in km) |
| Education | Employee's level of education  (1:  'Below College' , 2:  'College' , 3:  'Bachelor's Degree' , 4:  'Masters Degree', 5:  'Doctorate') |
| EducationField | Employee's field of education |
| EmployeeCount | Employee count |
| EmployeeID | Unique Employee ID |
| Gender | Employee's gender |
| JobLevel | Employee's job level on a scale of 1 to 5 |
| JobRole | Employee's role title |
| MaritalStatus | Employee's marital status |
| MonthlyIncome | Employee's monthly income (in Rupees per month) |
| NumCompaniesWorked | Total number of companies the employee has worked for |
| Over18 | Whether the employee is above 18 years of age |
| PercentSalaryHike | Employee's salary hike last year (in percentage points) |
| StandardHours | Employee's standard working hours (duration) |
| StockOptionLevel | Employee's stock option level |
| TotalWorkingYears | Employee's total number of working years (entire life) |
| TrainingTimesLastYear | Number of times employee attended training last year |
| YearsAtCompany | Employee's total number of working years (in the company) |
| YearsSinceLastPromotion | Employee's number of years since last promotion |
| YearsWithCurrManager | Employee's number of years working under current manager |

# Data Preparation and transformation.

To embark with first I have downloaded data from Kaggle, after that I get to know that this dataset is the combination of 3 different dataset. Therefore for analysis first we need to merge the data into one dataset and for that I have performed following steps:-

1. There are 3 files, general\_data, manager\_survey\_data, employee\_survey\_data, on these 3 files I have performed a merge action command to get a combined version. After merging 3 files following is the combined attribute list:-

Table

Description automatically generated

Step 2) Convert csv file into arff format (weka s/w support arff format)

I have loaded a csv file into weka tool and then saved that file using save option which got saved in .arff format.

Step 3) Performing some preprocessing tasks like handling missing values (using weka filters: Replace with Missing values, Replace Missing Value), exploring data as fig1.

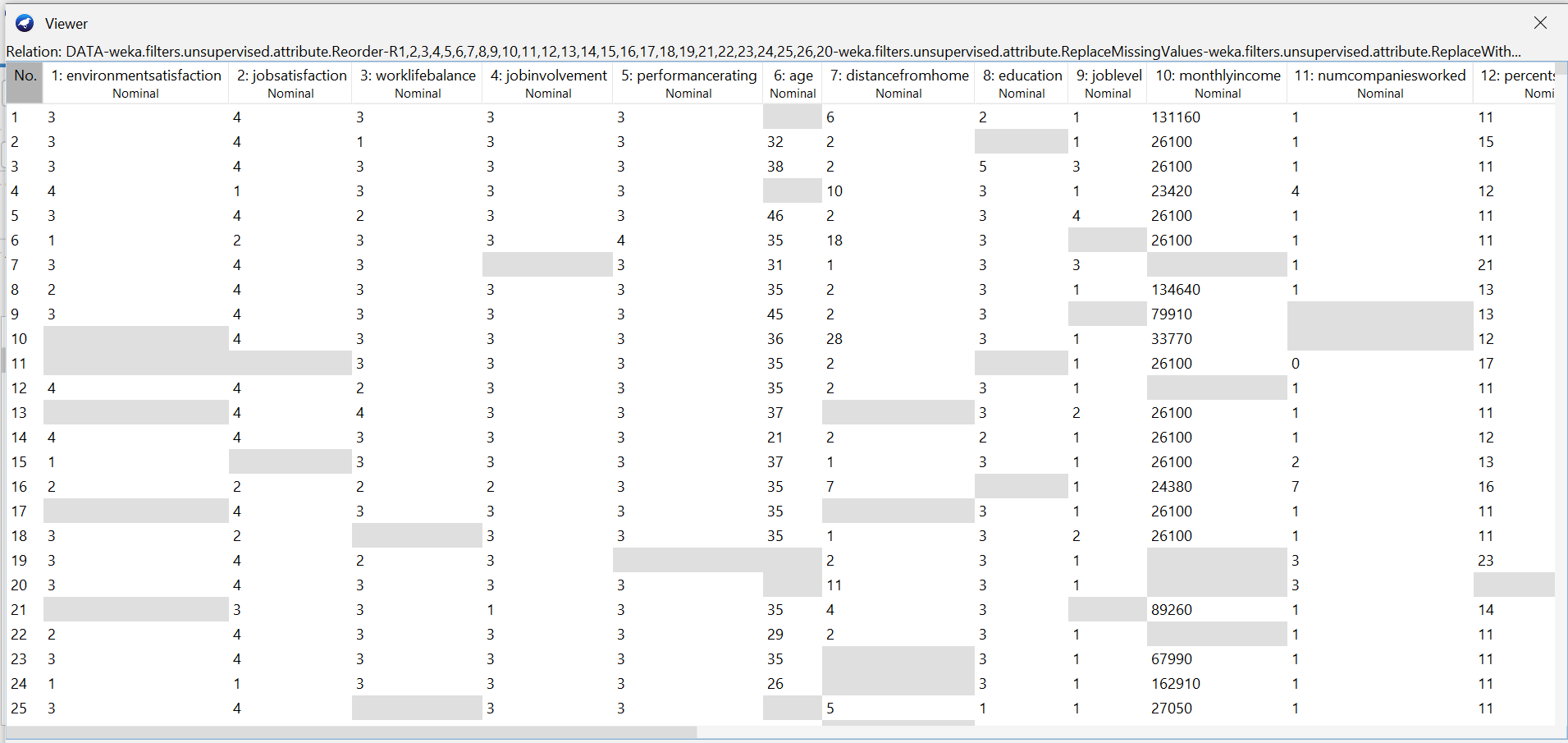


*Figure-1:  Attribute distributions Red = No Attrition, Blue = Yes Attrition*

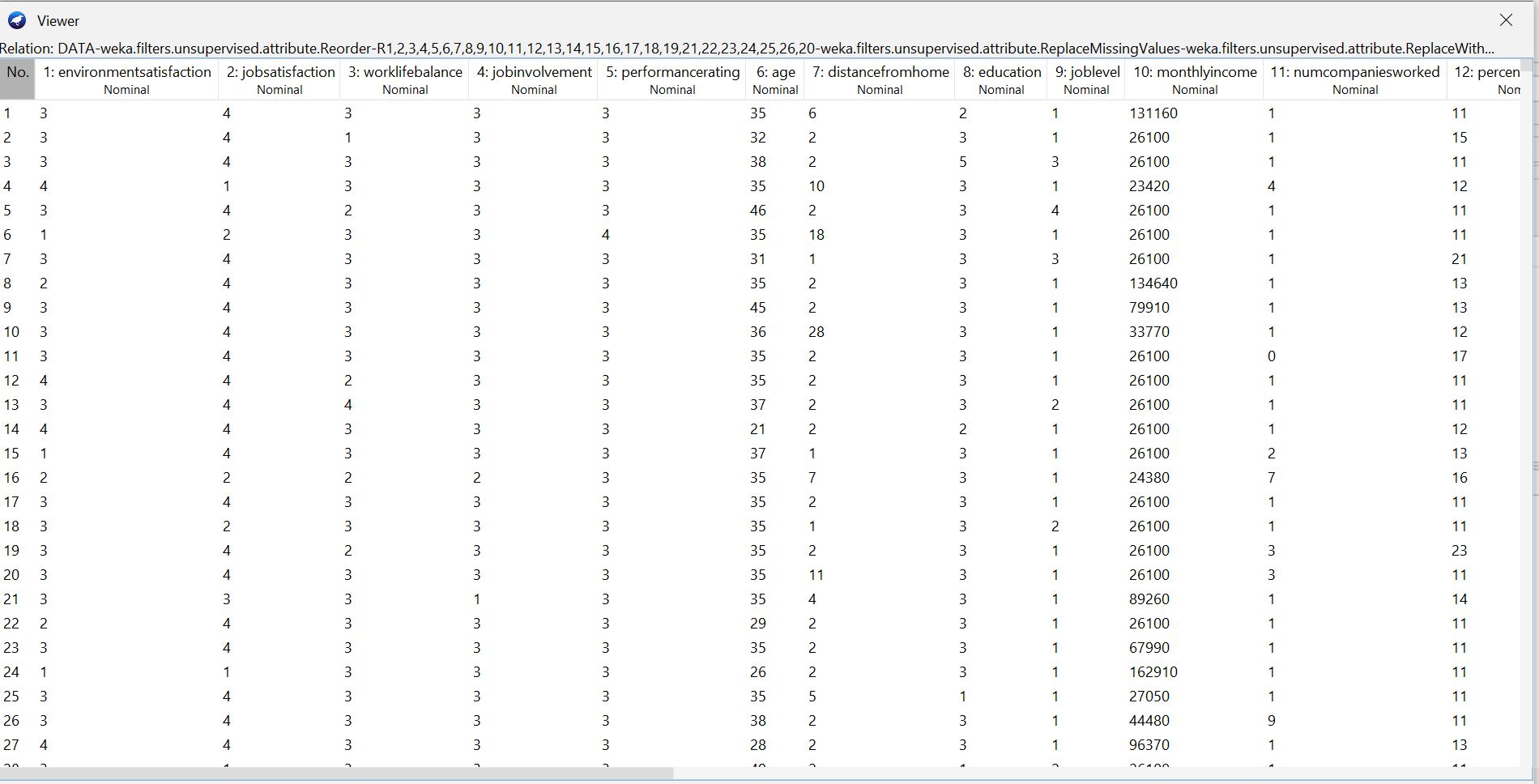
To understand/explore our dataset with respect to our class label attrition, Figure 1 shows clearly that many people are leaving the company as per the attribute/features. Main aim is to find the cause of this to happen. And also see by given set of features of new employee to predict that they will stay or leave the company.

I thought to replace the null values using mean instead of removing those rows so I checked in the graphical view into viewer After seeing from the image below, we can see that there is missing data. Now how to resolve this issue. So in weka we have two great filters for use to resolve this issue which is firstly we will import the data then select the “RemoveWithMissingValue” that resulted in the figure shown below. After getting that data further we use “RemoveMissingValue” that gives me the following result as shown in the figure.

Replace with Missing values

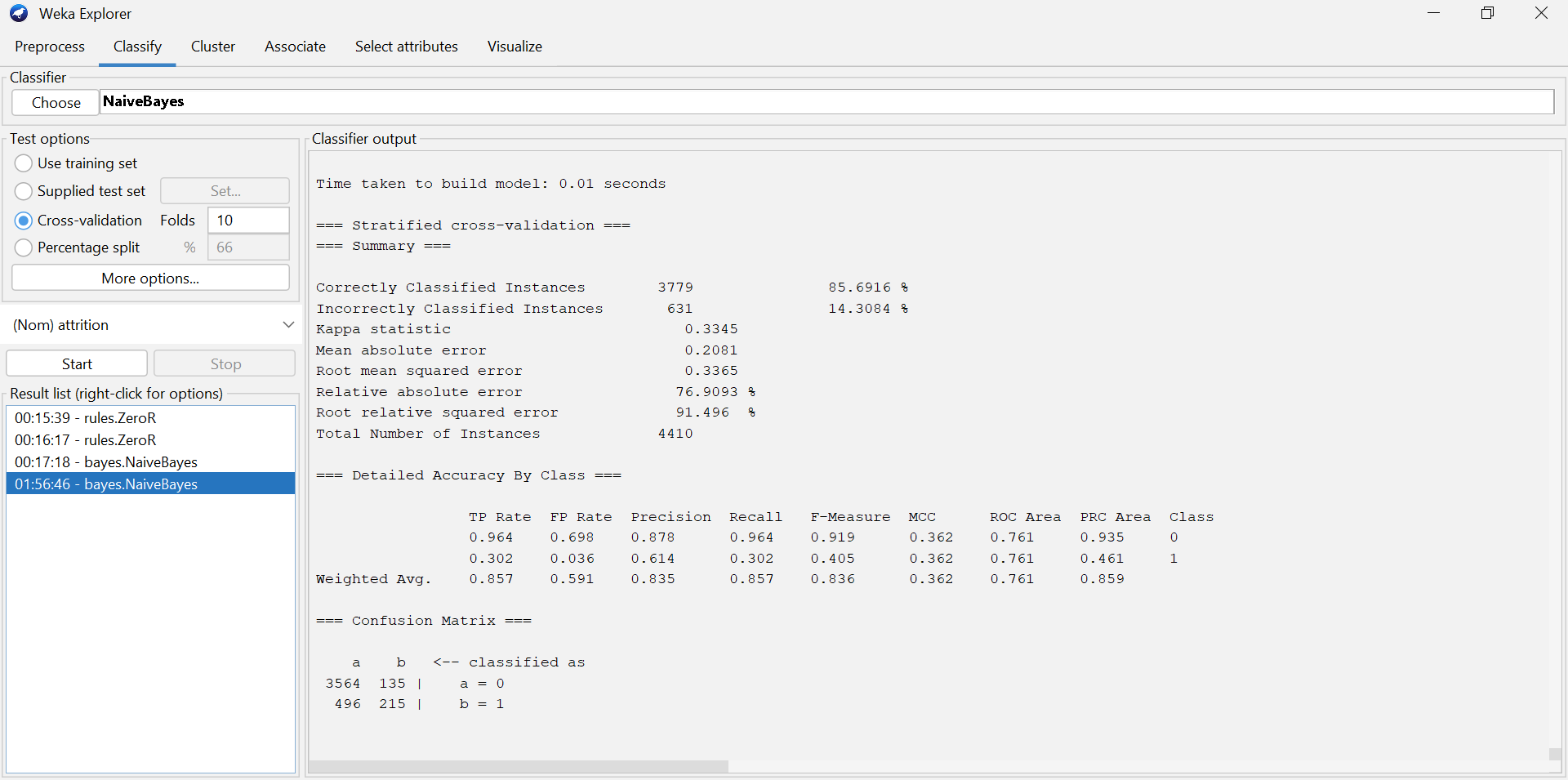


Replace Missing Value



1. **Approach to solve the problem (Algorithm summary)**

**Naïve Bayes:** The naive Bayes Algorithm is one of the popular classification machine learning algorithms that helps to classify the data based upon the conditional probability values computation. It implements the Bayes theorem for the computation and used class levels represented as feature values or vectors of predictors for classification. Naive Bayes Algorithm is a fast algorithm for classification problems. This algorithm is a good fit for real-time prediction, multi-class prediction, recommendation system, text classification, and sentiment analysis use cases. Naive Bayes Algorithm can be built using Gaussian, Multinomial and Bernoulli distribution. This algorithm is scalable and easy to implement for a large data set.



**SMO**: Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line.

A picture containing text, screenshot, receipt

Description automatically generated

**Random Forest**: It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. It uses bagging and features randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. It is a simple but popular algorithm that follows a top-down approach. Each node in the decision tree represents an attribute, and the leaf represents the outcome. Branches that link nodes to leaves are the decisions or the rules for prediction. Finally, the root node is the attribute that best describes the training dataset. Thus, the overall process is diagrammed into a tree-like structure.

A screenshot of a computer

Description automatically generated with low confidence

**Comparing Results:-**

|  |  |  |  |
| --- | --- | --- | --- |
| **class 0 class 1** | **Naïve Bayes** | **SVM** | **Random Forest** |
| **TP Rate:** | 0.964 | 0.978 | 1.000 |
| 0.302 | 0.848 | 0.909 |
| **FP Rate:** | 0.698 | 0.152 | 0.091 |
| 0.036 | 0.02 | 0.000 |
|  |  |  |  |
| **Accuracy** | 85% | 95.54% | 98.41% |
| **Precision** | 0.878 | 0.968 | 0.981 |
| 0.614 | 0.890 | 1.000 |
| **Recall** | 0.964 | 0.978 | 1.000 |
| 0.302 | 0.848 | 0.909 |
| **F1 Score** | 0.919 | 0.973 | 0.990 |
| 0.405 | 0.869 | 0.952 |
| **ROC** | 0.761 | 0.913 | 0.995 |
| 0.761 | 0.913 | 0.995 |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Naïve Bayes** | **SVM** | **Random Forest** |
| **Mean Absolute Error** | 0.21 | 0.12 | 0.22 |
| **Root Mean Squared Error** | 0.34 | 0.34 | 0.32 |
| **Relative Absolute Error** | 76.29 | 43.40 | 81.52 |
| **Paired T-Tester** | 0.96 | 0.99 | 1.00 |
| **False Positive Rate** | 0.70 | 0.69 | 0.90 |
| **IR Precision** | 0.88 | 0.88 | 0.85 |
| **IR Recall** | 0.96 | 0.99 | 1.00 |

As the majority of data is about class 0, which represents an employee staying with the firm, it is noted that the classification of class 1, which represents an employee who will quit the company, presents a challenge owing to the problem of class imbalance. The models are thus seen to be overfit and more likely to predict "No" attrition. All of the models' observed outcomes have good accuracy, high F1 scores, and high ROC scores, but their strong performance is more effective at recognizing the positive cases. The class imbalance issue can be solved in a number of methods, including oversampling and undersampling.

The different models, which are Random Forest, Support Vector Machine and Naive Bayes are compared based on the following metrics:

***Most accurate*:** All the models have fairly good accuracy, however the best accuracy was observed by using Random Forest.

***Best ROC Area under curve*:** ROC depicts the comparison of True Positives vs False positives, higher the ROC score, the better the model performs. In this category, Random Forest again scored the highest within all the models.

Best ‘Yes’ True Positive Rate: Perhaps, the most important metric for our models is their ability to identify True Positives. Random Forest was again observed to perform the best. Random Forest had a TP Rate of 0.909 meaning 90% of the ‘Yes’ tuples were correctly classified. It is therefore observed that Random Forest outperforms the others by having the highest ‘Yes’ TP rate at 90% while maintaining a low FP rate. The overall accuracy is also the highest for Random Forest along with a high ROC score.

According on the metrics mentioned above, it appears that the model is far more effective in identifying employees in class 0, or those who are likely to depart the company.

1. **Conclusion**

Finding out how different classifiers predict results differently and with greater accuracy was interesting. Comparing Random Forest to other classifier methods, it outperforms them. It was intriguing to observe how well the data fit the various learning layers that contribute to outcome prediction. We attempt to extract attributes from the best classifier algorithm (Random Forest), and it is obvious that these qualities—job engagement, monthly pay, years at company, years with manager, and business transition—make sense as clear indications of employee attrition in the organization.

**Appendix**

1. 3 different algorithm output are as mentioned above.
2. Different attributes used for above algorithm

Graphical user interface, application

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedGraphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

Feature Selection

Graphical user interface, application, Word

Description automatically generated