A MINI- PROJECT REPORT ON

***“Employee Attrition Rate Prediction”***

SUBMITTED BY

**Riteshkumar Singh 18IT1068**

**Rupal Sonje 18IT1086**

**Soham Salkar 18IT1050**

Supervisor

**Mrs. Soumi Banerjee**



## Department of Information Technology

Dr. D. Y. Patil Group’s

**Ramrao Adik Institute of Technology**

Nerul, Navi Mumbai (Affiliated to University of Mumbai**)**

(2022)



**Ramrao Adik Institute of Technology**

## (Affiliated to the University of Mumbai)

#### Dr. D. Y. Patil Vidyanagar,Sector 7, Nerul, Navi Mumbai 400706.

**CERTIFICATE**

## This is to certify that, the Mini Project titled

**“Employee Attrition Rate Prediction”**

### is a bonafide work done by

**Riteshkumar Singh 18IT1068**

**Rupal Sonje 18IT1086**

**Soham Salkar 18IT1050**

**Project Guide (Mrs. Soumi Banerjee)**

**Head of Department (Dr. Ashish Jadhav)**

**Mini Project Approval**

This is to certify that the Mini Project entitled “**Employee Attrition Rate Prediction**” is a bonafide work done by Riteshkumar Singh Roll no.18IT1068, Rupal Sonje Roll no.18IT1086, Soham Salkar Roll no. 18IT1050 under the guidance of Mrs. Soumi Banerjee. This work has been approved as R programming Mini Project for Final year Information Technology.

Examiners : 1...............................

2...............................

Supervisors : 1...............................

2...............................

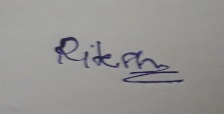
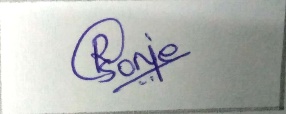
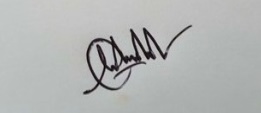
Date :

Place :

**DECLARATION**

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

**Name and Roll No. of Students Signature**

1. Riteshkumar Singh (18IT1068) 
2. Rupal Sonje (18IT1086) 
3. Soham Salkar (18IT1050) 

Date:

Place:

**ACKNOWLEDGEMENT**

The project “Employee Attrition Rate” is creative work of many minds. A proper synchronization between individual is must for any project to be completed successfully. One cannot imagine the power of the force that guides us all and neither can we succeed without acknowledging it.

We would like to express our gratitude to Principal **Dr. Mukesh D. Patil** and **Dr. Ashish Jadhav,** our Head of the department, Information Technology Engineering for encouraging and inspiring us to carry out the project in the department lab.

We would also like to thank our Guide **Mrs. Soumi Banerjee**, Department of the Information Technology Engineering for her expert guidance, encouragement and valuable suggestions at every step.

We also would like to thank all the staff members Department of the Information Technology Engineering for providing us with the required facilities and support towards the completion of the project.

Last but not the least we are thankful to our parents and friends for their constant Inspiration, encouragement and well wishes by which we have made a challenging project.

Riteshkumar Singh 18IT1068

Rupal Sonje18IT1086

Soham Salkar 18IT1050

**TABLE OF CONTENTS**

##### Sr. No. Topic

ABSTRACT

* 1. INTRODUCTION
  2. DATASET
  3. PROBLEM STATEMENT
  4. METHODOLOGY
  5. RESULT
  6. CONCLUSION
  7. REFERENCES

**ABSTRACT**

Employees are the significant resources of any association. In any case, in the event that they quit job unexpectedly, it might bring about immense expense to any organization. Since new hiring will consume money and time as well as the newly hired employees set aside some effort to make the particular organization productive. Subsequently in this paper we attempt to develop a model which will predict employee attrition rate dependent on HR analytics dataset. Several factors lead to employee attrition. Such factors are analyzed to reveal their intercorrelation and to demonstrate the dominant ones. Our work was tested using the imbalanced dataset of IBM analytics, which contains 35 features for 1470 employees. To get realistic results, we derived a balanced version from the original one. Finally, cross-validation is implemented to evaluate our work precisely. This model can be utilized by the Human Resource branches of the organizations to shape proficient methodologies to hold the important representatives before they start searching for new employments like by giving a hike in their compensation.

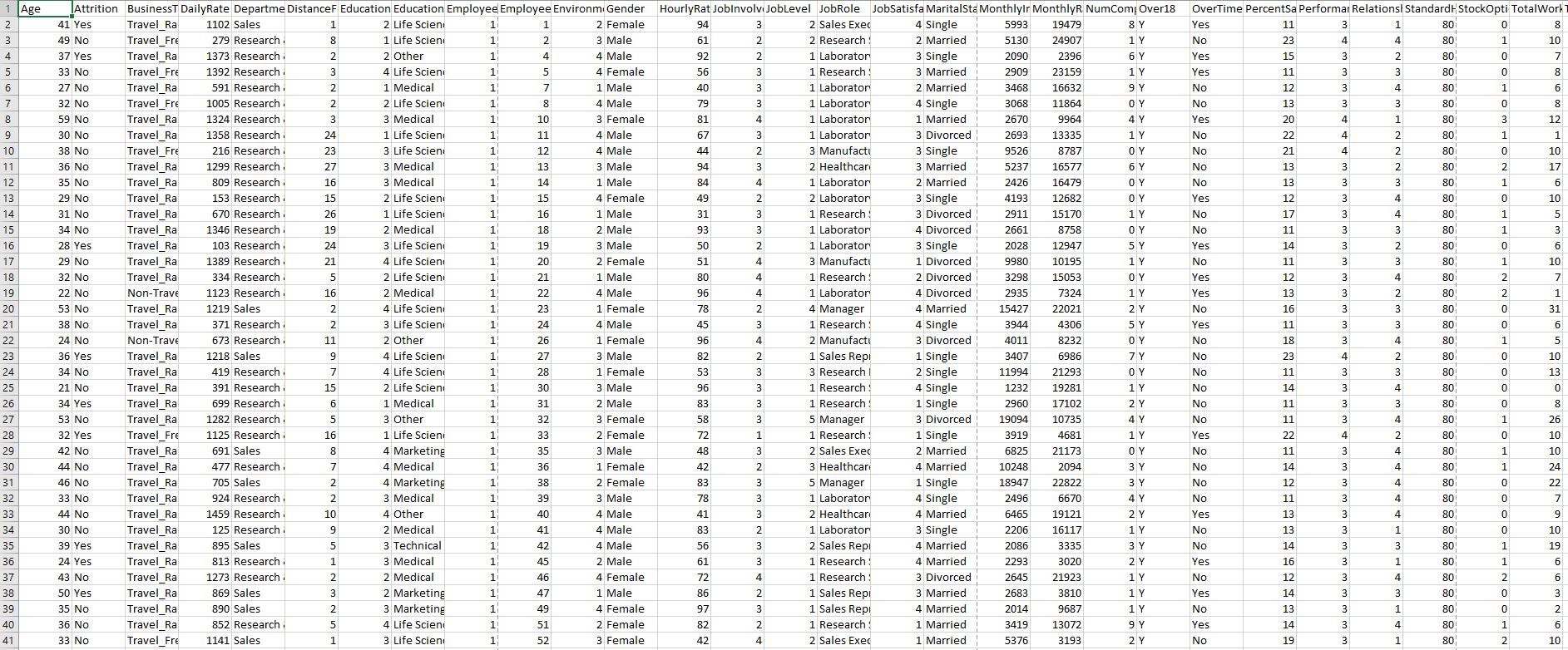
**Introduction**

Employee Attrition is a reduction in manpower in any organization where employees resign. Employees are the valuable assets of any organization. It’s necessary to know whether the employees are dissatisfied or are there any other reasons for leaving the respective job. These days for better opportunities, employees are eager to jump from one organization to other. But if they leave jobs unexpectedly, it may cause huge loss for organization. New hiring will consume money and time, and also the freshly hired employees take time to make the respective organization profitable. Retention of skilled and hardworking employees is one of the most critical challenges faced by many organizations. Hence, by improving employee satisfaction and providing a desirable working environment, we can certainly reduce this problem significantly. When an Employee leaves an organization, the reasons are determined by a variety of factors, some of the reason of leaving the organization could be better-paying job outside, a bad relationship with boss, pursuing higher studies, relocating due to family reasons, fired from organization, job Dissatisfaction, salary not as per expectation, poor relationship with team members, poor working environment, lack of opportunity for career development, overtime, workload etc. Using the historical data kept in human resources (HR) departments, analysts can build and prepare a machine learning model that can predict the workers who are leaving the organization. Such models are prepared to look at the connection between the features of both active and terminated workers. This model is applicable for employees who have completed their probation period. If the employee has recently joined the organization, then it is difficult to predict their dissatisfaction factors as they are not a confirmed employee before their allocated probation period. Predicting employee attrition before it occurs can help the administration to prevent it or at least reduce its effect.

Here we will be using IBM HR analytics dataset from Kaggle website to train and test the machine learning model. Here we will be using Logistic Regression algorithm as a method of predicting employee attrition. This is done by using dataset and treating the problem as a classification task.

**DataSet**

* The IBM HR Analytics Employee Attrition & Performance is a fictional data set created by IBM data scientists.
* This dataset includes 35 features for 1470 samples of two classes (current and former employees). These samples are not balanced; there are 237 positive samples (former employee) and 1233 negative samples (current employee).
* The categorical values are converted to numeric values in order to make the classification algorithm more effectual. For example, categorical attribute ‘Business Travel’ contains three values such as Travel-Rarely, TravelFrequently, Non-Travel. Hence it is converted to 1, 2 and 3 respectively.



## **PROBLEM STATEMENT**

Employees are the most important part of an organization. Successful employees meet deadlines, make sales, and build the brand through positive customer interactions. Employee attrition is a major cost to an organization and predicting such attritions is the most important requirement of the Human Resource department in many organizations. Employee resignations are a reality for any business. However, if the situation isn't handled properly, key staff members' departures can lead to a downturn in productivity. The organization may have to employ new people and train them on the tool that is being used, which is time consuming. Most organizations are interested in knowing which of their employees are at the risk of leaving. Organizations try to prevent employee attrition by using machine learning algorithms to predict the risk of an employee leaving, and then take pro-active steps for preventing such an incident.

## **METHODOLOGY**

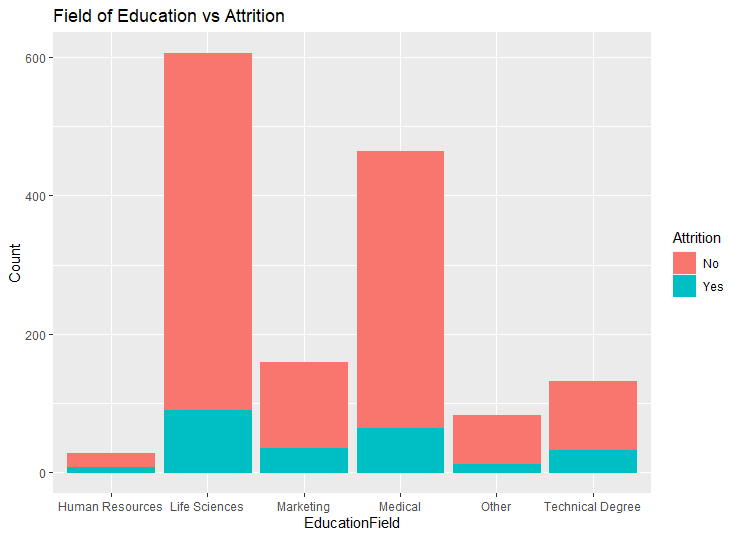
1. **Data Preprocessing**

Preprocessing operation is a crucial step in machine learning, which significantly improves the model performance. If suppose any feature value in the record contain any null value or undefined or irrelevant value then separate that entire record from the original dataset and place that record into training dataset, else if the record contain perfect data with all features then place that into test dataset. Test dataset contain all important features to predict employee attrition or employee attrition and training dataset contain irrelevant data.

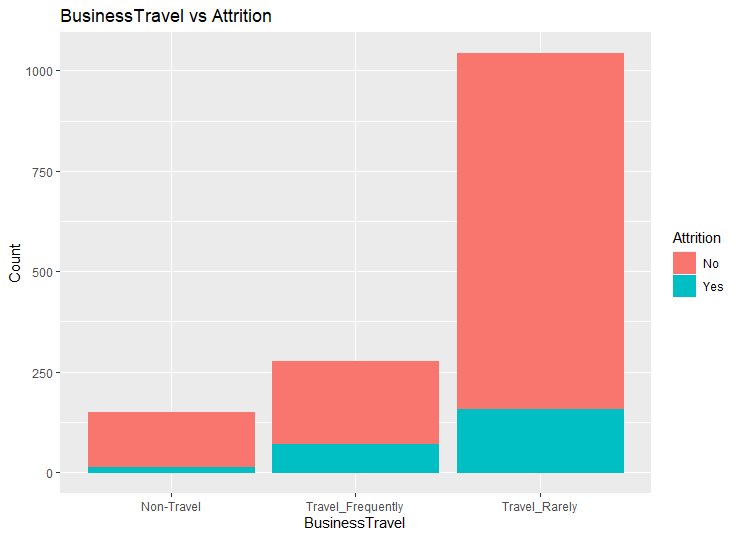
1. **Data Analysis and Visualization**

Visualisation is an important tool for insight generation, but it is rare that you get the data in exactly the right form you need. Often you’ll need to create some new variables or summaries, or maybe you just want to rename the variables or reorder the observations in order to make the data a little easier to work with. Functions from dplyr, ggplot2 packages of tidyverse mostly do the work of data transformation.

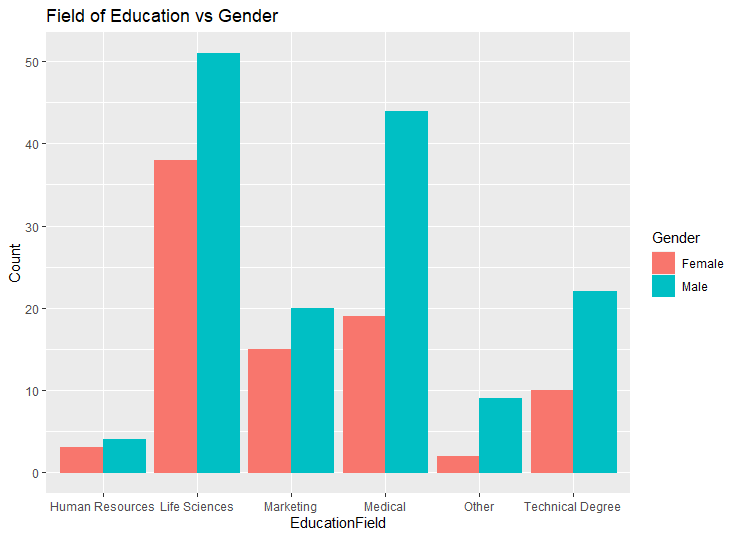
Here are several graphs generated by the system with respect to attrition:



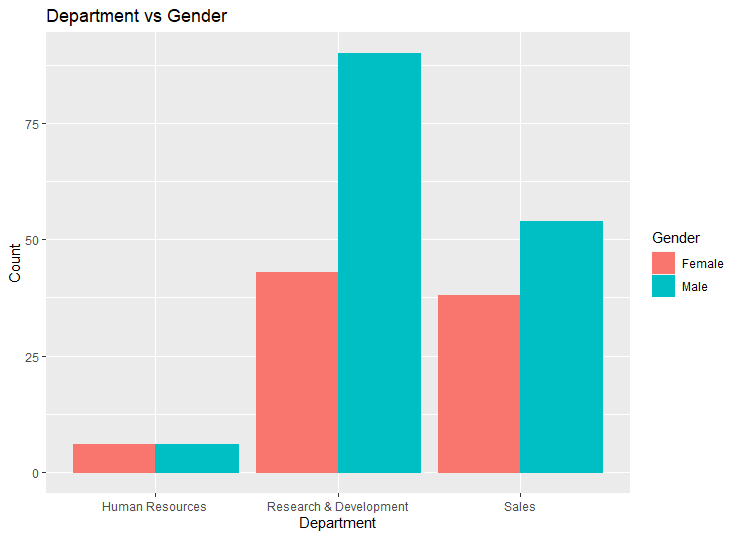
The bar graph represents the comparison among the Field of Education with respect to Attrition.



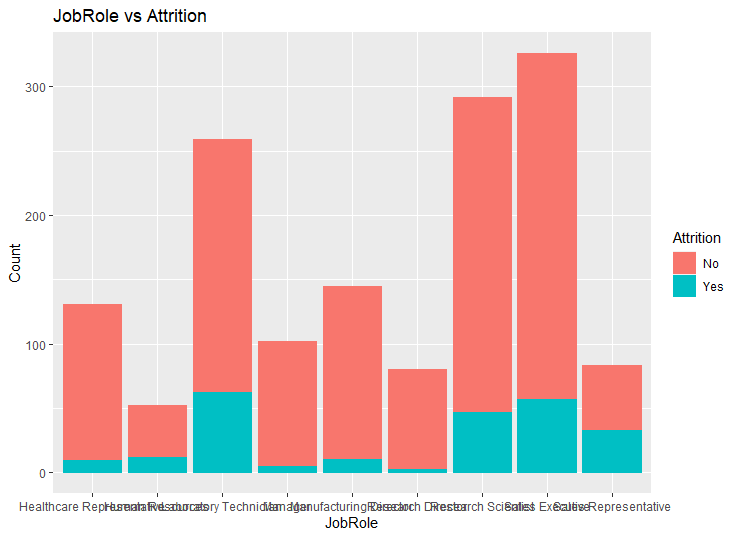
The bar graph the bar graph of comparison among the Business Travel with respect to Attrition. Attrition rate of Travel Rarely is 14.96%, Travel Frequently is 24.91% and Non-Travel is 08.00%.



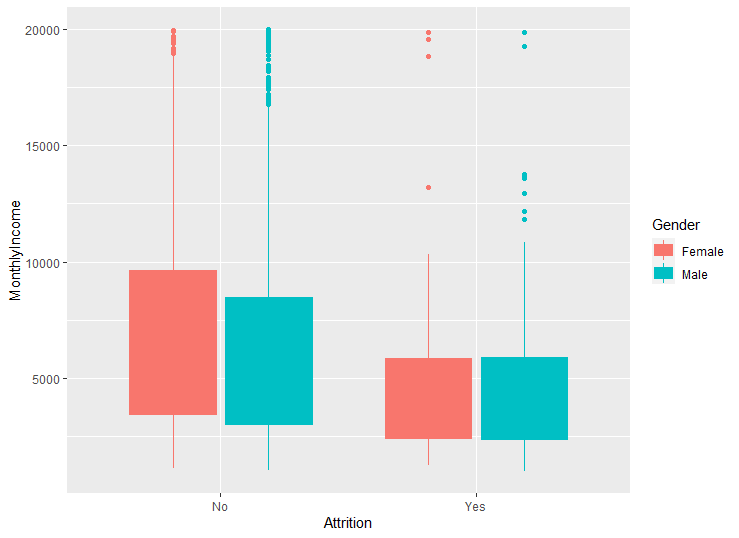
The bar graph represents the comparison of Education Field with respect to Gender.



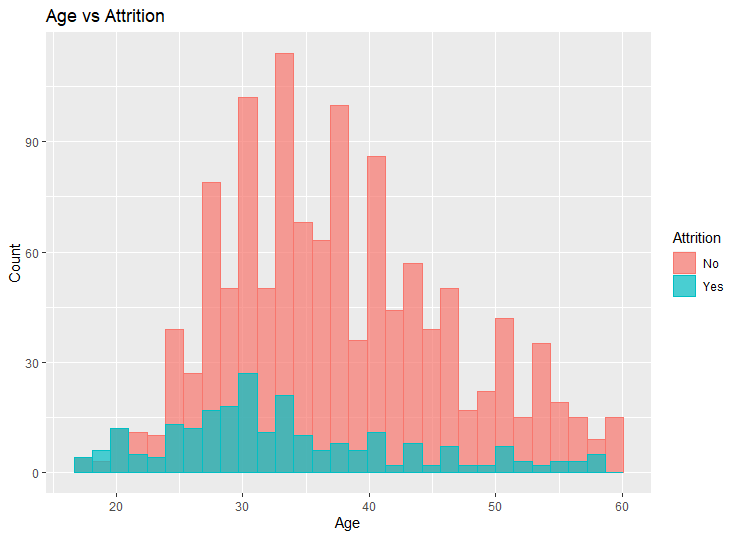
The bar graph represents the comparison of Department of employee with respect to Gender.



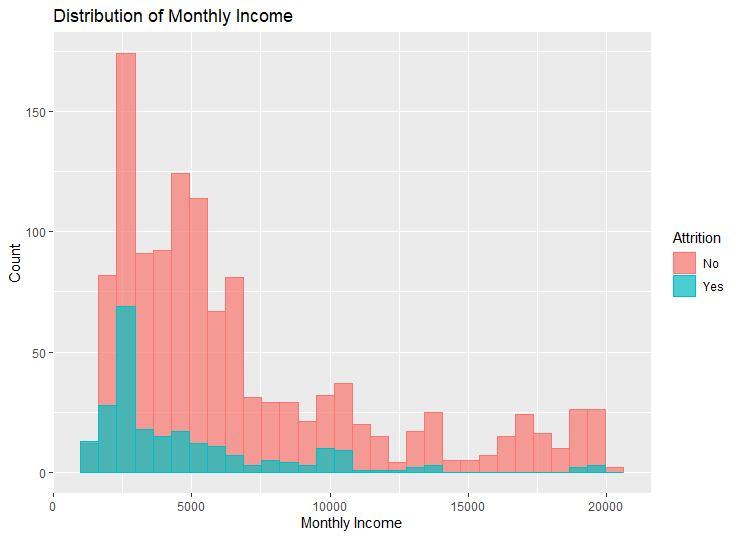
The bar graph represents the comparison of Job Role of employee with respect to Gender.



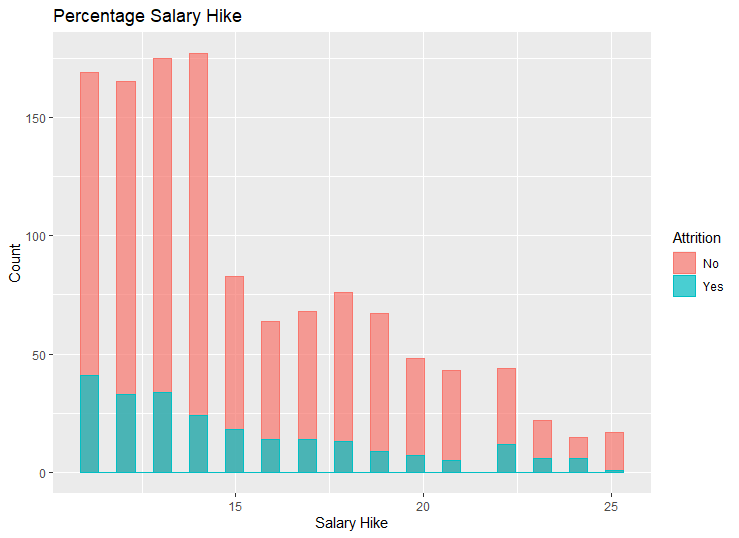
The box plot graph represents the comparison of Monthly Income of employee with respect to Attrition.



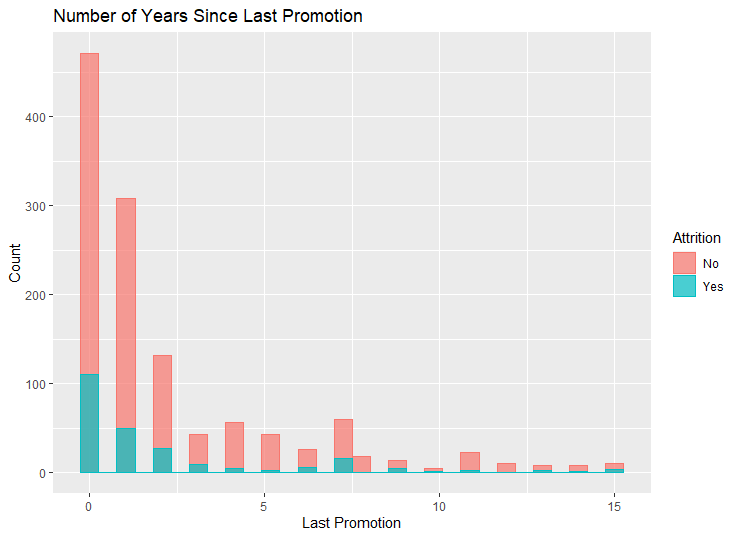
The histogram represents the comparison of age of employee with respect to attrition.



The histogram represents the comparison of distribution of monthly income with respect to attrition.



The histogram represents the comparison of distribution of percentage salary hike with respect to attrition.



The histogram represents the comparison of distribution of years since last promotion with respect to attrition

1. **Test Dataset and Training Dataset**

Separating data into test datasets and training datasets is an important part of evaluating data mining models. By this separation of total data set into two data sets we can minimize the effects of data inconsistency and better understand the characteristics of the model. The dataset was split 70-30 into training and test sets. The models were trained using their optimal configurations on the training dataset. The trained model was then used to predict on the 30% test set.

1. **Model Building**

Machine Learning is most important technology towards data analysis for quality prediction and evaluation. There are various algorithms in machine learning which are used to predict the appropriate class of new or unseen data. In our system we used different machine learning algorithms to find out the reasons for employee attrition. For this dataset we will be using logistic regression algorithm for model building. Logistic regression is a regression model that fits the values to the logistic function. It is useful when the dependent variable is categorical. The general form of the model is 𝑃(𝑌|𝑋,𝑊)= 1⁄(1+e^(-(w0+Σwixi)) Logistic regression is often used with regularization techniques to prevent overfitting.

**RESULT**

###Step 1: Read data

attrition.df <- read.csv("C:\\Users\\nikhi\\Downloads\\kaggle\\HR-Employee-Attrition.csv", stringsAsFactors = TRUE)

#preprocessing the data

#all relevant variables are categorical

#base level for attrition is "No" . changing it to "Yes"

attrition.df$Attrition<- as.numeric(attrition.df$Attrition == "Yes")

attrition.df$Attrition <- factor(attrition.df$Attrition)

### Step 3: Partition data

set.seed(132)

# select variables

selected.var <- c(29,19,16,23,1,4,15,6,32,13,2)

selected.df <- attrition.df[, selected.var]

str(selected.df)

# partition the data

train.index <- sample(1:nrow(attrition.df), nrow(attrition.df)\*0.7)

train.df <- selected.df[train.index, ]

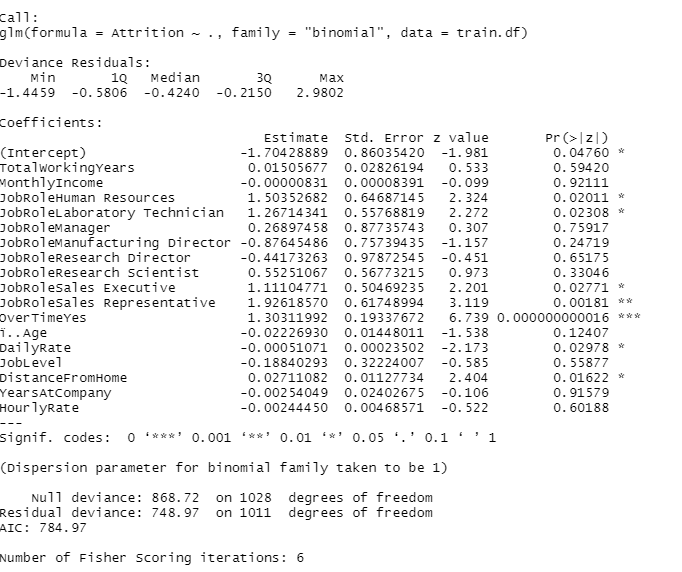
valid.df <- selected.df[-train.index, ]

### Step 4: Fit a logistic regression model

logit.reg <- glm(Attrition ~ ., data = train.df, family = "binomial")

options(scipen = 999)

summary(logit.reg)



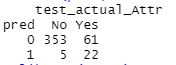
### Step 5: Generate outcome by comparing predicted probability with the cutoff probability

logit.reg.pred <- predict(logit.reg, valid.df, type = "response")

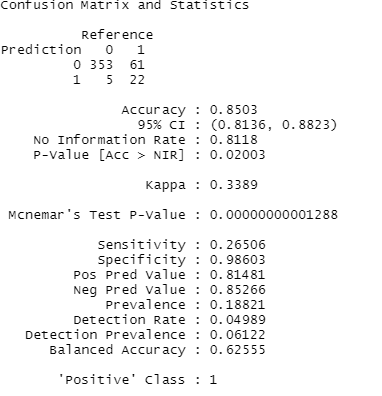
pred <- ifelse(logit.reg.pred > 0.4, 1, 0)

test\_actual\_Attr <- factor(ifelse(valid.df$Attrition==1,"Yes","No"))

table(pred,test\_actual\_Attr)



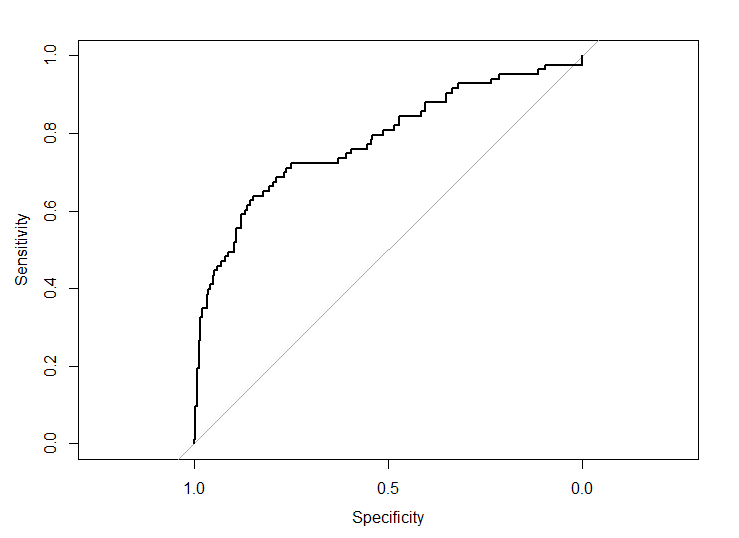
confusionMatrix(factor(pred), factor(valid.df$Attrition), positive = "1")



###Step 6: Generate ROC curve

r <- roc(valid.df$Attrition, logit.reg.pred)

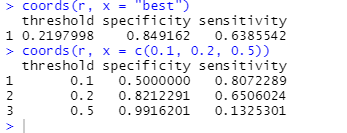
plot.roc(r)



# find the best threshold, with a default best method "youden"

coords(r, x = "best")

coords(r, x = c(0.1, 0.2, 0.5))



**Conclusion**

The proposed work can assist the human resources department in providing the necessary information about the potential decision of an employee to leave the organization. Depending on employee signals, our method predicts whether there is a potential risk of employee attrition. We have analyzed the employee’s dataset to obtain the most features that encourage the employee to leave the organization. Additionally, the correlations among various features are also presented. Our findings, in this regards, shows that overtime hours, job level, and monthly income are the most effective features that influence the employee decision. Using the dataset offered by IBM analytics is still a challenging task due to its imbalanced nature. This leads us to create a synthetic version of this dataset to build a stable classifier that can support realistic prediction. Thorough experiments have been conducted to measure the effectiveness of our method in terms of accuracy and precision. The proposed method has shown a high performance compared to state-of-the-art techniques that used the same dataset. The accuracy, using the imbalanced and synthetic balanced datasets was 85.06% respectively.

**REFERENCES**

[1] Cotton, J.L. and Tuttle, J.M., 1986. “Employee turnover: A metaanalysis and review with implications for research” Academy of management review, pp.55-70.

[2] Liu, D., Mitchell, T.R., Lee, T.W., Holtom, B.C. and Hinkin, T.R., 2012. “When employees are out of step with coworkers: How job satisfaction trajectory and dispersion influence individual-and unit-level voluntary turnover”. Academy of Management Journal, pp.1360-1380.

[3] Heckert, T.M. and Farabee, A.M., 2006. “Turnover intentions of the faculty at a teaching-focused university”. Psychological reports, pp.39-45.

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