**IBM-HR ANALYTICS EMPLOYEE ATTRITION AND PERFORMANCE**

**Abstract**

Employees are the backbone of the organization. Organization's performance is heavily based on the quality of the employees. Here we apply machine learning techniques to analyse the employee information for improving his/her position in the organization. Compensation and job performance information from revenue rates and personnel characteristics to payroll and service history, never before have HR executives had such liberated right to use to individual details. In this work, we are applying random forest classification, which facilitates employee classification based on their monthly income and informal way to execute analytics on data.

**Some business questions:**

1. What factors are contributing more to employee attrition?
2. What type of measures should the company take in order to retain their employees?
3. What business value does the model bring?
4. Will the model save lots of money?
5. Which business unit faces the attrition problem?

**OBJECTIVES**

* Visualizing the performance of employees.
* Analysing the factors causing attrition.

1. **INTRODUCTION**
2. **ABOUT PROJECT**

Attrition and performance analysis of employees working in a company. The dataset included in our project contains attributes in respective to age, gender, job role, martial status, department of employee like more. There are many factors that is taken into consideration hence for the HR they can visualize each and every aspect of employee to find their performance and attrition.

**THE PROJECT CONSISTS OF THE FOLLOWING MODULES:**

* DATA VISULIZATION AND ANALYSIS ON EMPLOYEE’S DATA.
* BUILDING MODEL TO FIND FEATURES CAUSING ATTRITION.

1. **REVIEW OF LITERATURE**

The several areas in which organisations can adopt technologies that will support decision-making: artificial intelligence is one of the innovative technologies that is widely used to assist organisations in business strategies, organisational aspects and people management. Good amount of research has been done while considering the factors for the employee attrition prediction. We can analyse all the factors influencing employee attrition so that HR and project managers use them accordingly retaining valuable employee.

1. **SYSTEM ANALYSIS**

The primary objective of our study is to analyse the factors causing employee attrition so that the company can use it to retain valuable employee. The dataset we chosen contains attributes like age, gender, department, overtime, and so on to analyse the employee’s performance and attrition. The model that we develop should reduce the difficulties for HR or company to find the employee’s attrition.

1. **FEASIBILTY STUDY**

The online dataset consists of records of Age, Attrition, BusinessTravel, DailyRate,

Department, DistanceFromHome, Education, EducationField, EmployeeCount,

EmployeeNumber, EnvironmentSatisfaction, Gender, HourlyRate, JobInvolvement,

JobLevel, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, MonthlyRate,

NumCompaniesWorked, Over18, OverTime, PercentSalaryHike, PerformanceRating,

RelationshipSatisfaction, StandardHours, StockOptionLevel, TotalWorkingYears,

TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole,

YearsSinceLastPromotion, YearsWithCurrManager.

1. **SYSTEM SPECIFICATION**

**5.1. Hardware Requirements:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | * Processor | | | |  |  | Intel Core i3 (7thGen) | | |
|  | * Clock Speed | | | |  |  | 2.8 GHz | | |
|  | * RAM | | | |  |  | 2 GB RAM (4 GB PREFERABLE) | | |
|  | * Hard Disk | | | |  |  | 1 TB HDD | | |
|  | * Monitor | | | |  |  | 15.6 Inches | | |
|  | * Mouse | | | |  |  | HP S500 Wireless Mouse | | |
|  | * Keyboard | | | |  |  | Full-size island-style keyboard with number Keypad | | |
|  |  |  | | | |  |  | |

**5.2. Software Requirements:**

* Operating System : Windows 10
* Front-End Tool : Jupyter Notebook

1. **PROJECT DESIGN**
   1. **Project description**

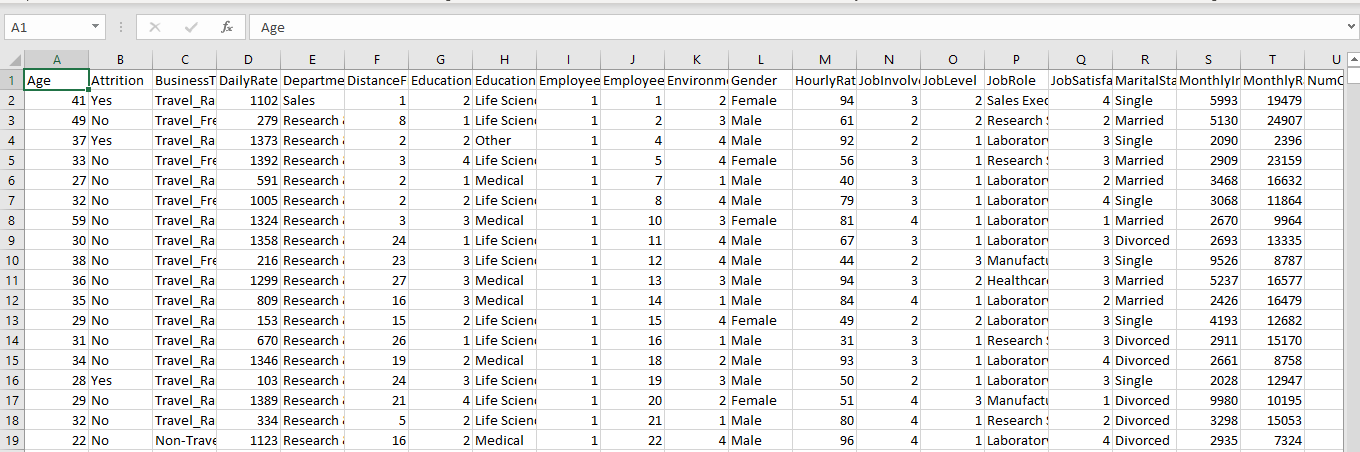
To analyse the factors causing employee attrition so that the company can use it to retain valuable employee. The dataset we chosen contains attributes like age, gender, department, overtime, martial status, education field and so on to analyse the employee’s performance and attrition.

* 1. **Data flow diagram**

Result

Logistic Regression and Random forest classifier

* 1. **Database design**

****

**Table name:** WA\_Fn-UseC\_-HR-Employee-Attrition

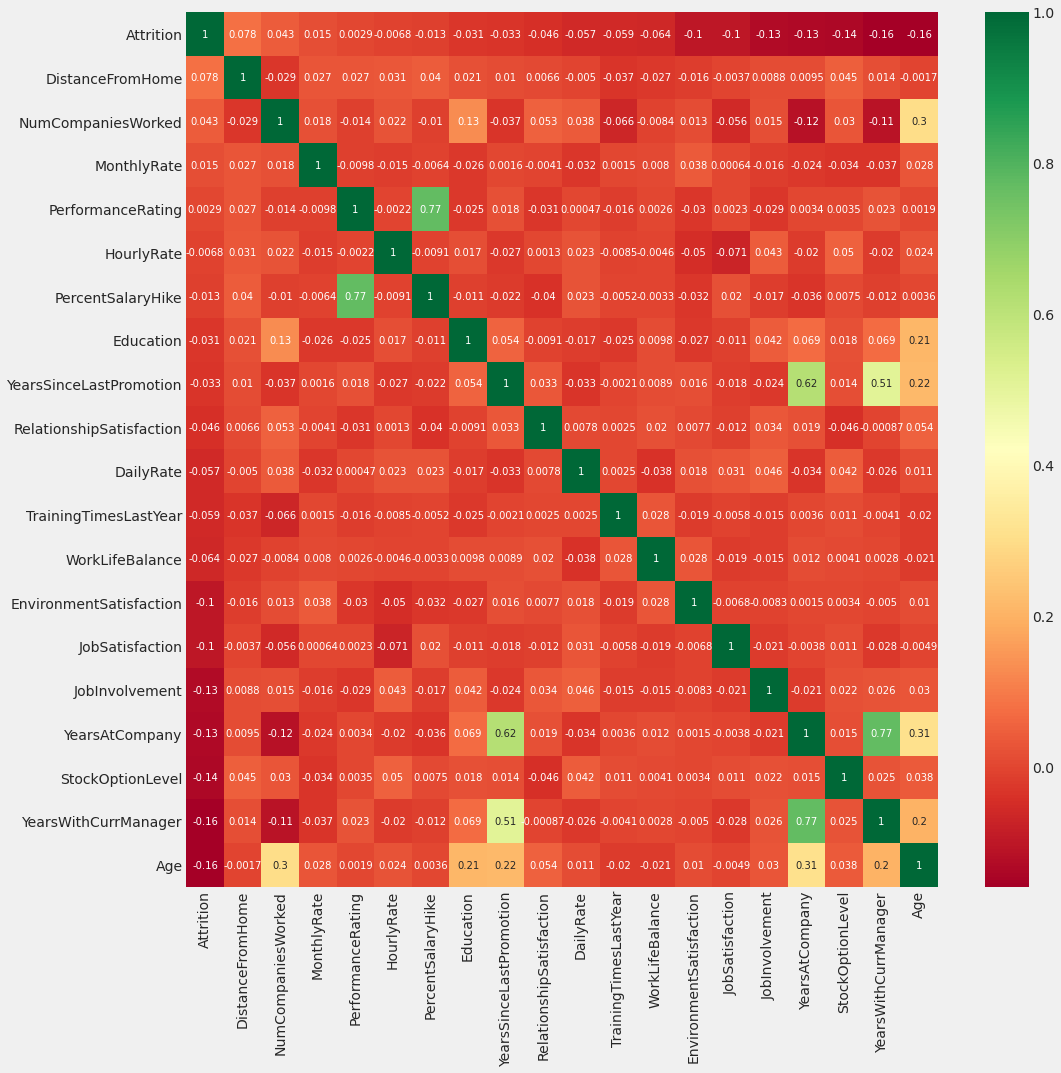
|  |  |  |  |
| --- | --- | --- | --- |
| **S.NO** | **FIELD NAME** | **DATA TYPE** | **DESCRIPTION** |
| 1. | Age | Int64 | Age of the employee |
| 2. | Attrition | Object | Attrition or not? |
| 3. | BusinessTravel | Object | Does travel for business? |
| 4. | DailyRate | Int64 | Daily rate of employee |
| 5. | Department | Object | Department of employee |
| 6. | DistanceFromHome | Int64 | Distance from home |
| 7. | Education | Int64 | How many education degree |
| 8. | EducationField | Object | Field of education |
| 9. | EmployeeCount | Int64 | Employee count |
| 10. | EmployeeNumber | Int64 | Employee number |
| 11. | EnvironmentSatisfaction | Int64 | Whether satisfied by environment |
| 12. | Gender | Object | Gender |
| 13. | HourlyRate | Int64 | Hourly rate |
| 14. | JobInvolvement | Int64 | Involvement in job |
| 15. | JobLevel | Int64 | Job level |
| 16. | JobRole | Object | Job role |
| 17. | JobSatisfaction | Int64 | Job satisfaction |
| 18. | MaritalStatus | Object | Martial status of employee |
| 19. | MonthlyIncome | Int64 | Income |
| 20. | MonthlyRate | Int64 | Monthly rate |
| 21. | NumCompaniesWorked | Int64 | Number of companies worked for? |
| 22. | Over18 | Object | Age over 18? |
| 23. | OverTime | Object | Works overtime? |
| 24. | PercentSalaryHike | Int64 | Salary hike in percentage |
| 25. | PerformanceRating | Int64 | Performance rating |
| 26. | RelationshipSatisfaction | Int64 | Satisfaction in relationship |
| 27. | StandardHours | Int64 | Standard hours in working |
| 28. | StockOptionLevel | Int64 | Stock option level |
| 29. | TotalWorkingYears | Int64 | Total working years |
| 30. | TrainingTimesLastYear | Int64 | Training time in last year |
| 31. | WorkLifeBalance | Int64 | Balance in work and life |
| 32. | YearsAtCompany | Int64 | Working years in this company |
| 33. | YearsInCurrentRole | Int64 | Years working in current role |
| 34. | YearsSinceLastPromotion | Int64 | Years since last promotion |
| 35. | YearsWithCurrManager | Int64 | Years with current manager |

1. **SYSTEM TESTING AND IMPLEMENTATION**
   1. **Testing Fundamentals**

* **Correlation plot:** This will be helpful in finding the columns which are highly correlated with the attrition as in our dataset there is a separate column indicating whether the employee is in attrition or not. It is also helpful in finding correlation factor for some important attributes like job level, etc…
* **Logistic regression:** It is a model used for classification, we use in our data to classify attrition employee and normal employee.
* **Random Forest Classifier:** It is ensemble technique used to improve the accuracy of the model.
  1. **Implementation**
* Technology has got us covered, Machine Learning powered to classify the attrition of employees.
* Using this we can predict employee’s attrition by just giving past data to our training model.
* We can use this also to find valuable employees in the company to improve the performance of the company.

1. **OUTPUT SCREEN LAYOUTS AND SAMPLE CODING**

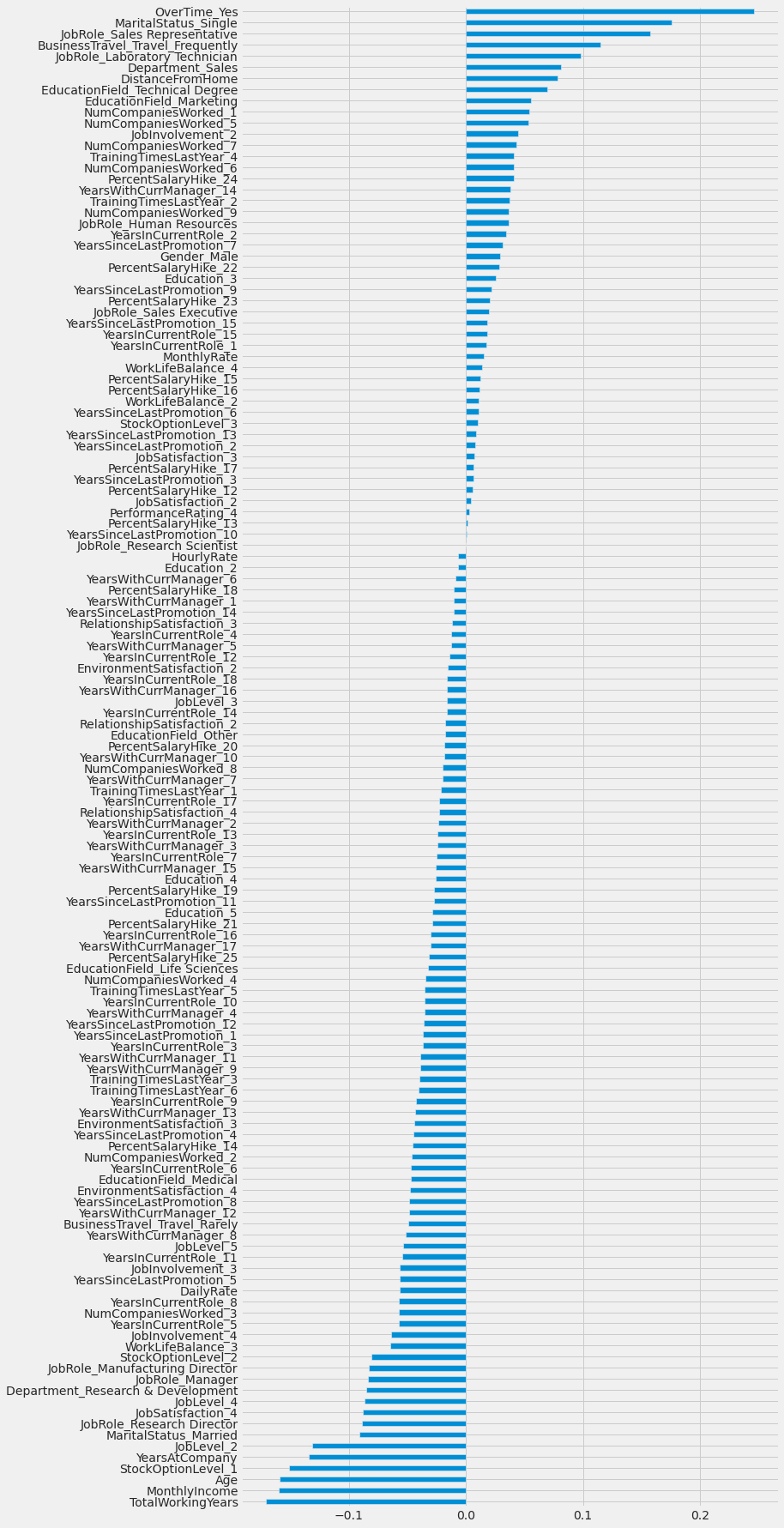
**Correlation plot for analysis of factors causing attrition**

****

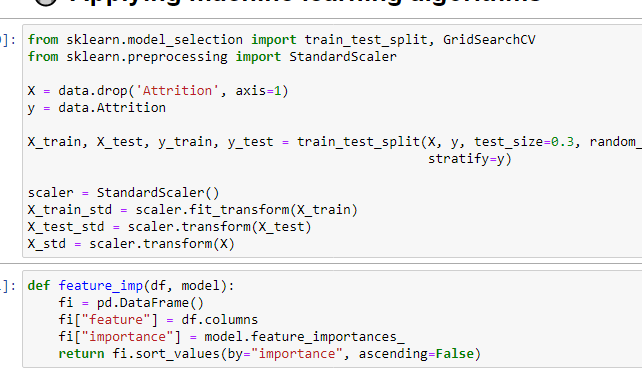
**Data pre-processing**

****

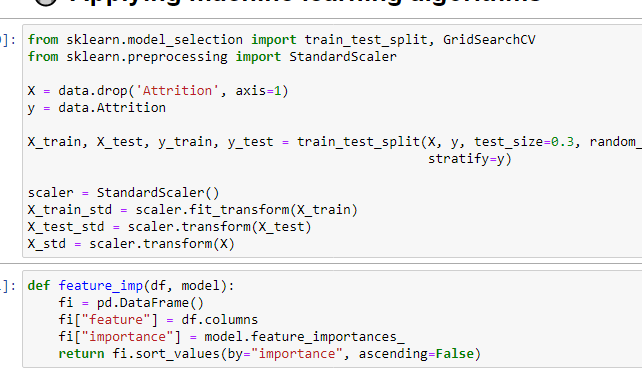
**Histplot of correlation with attrition column**

****

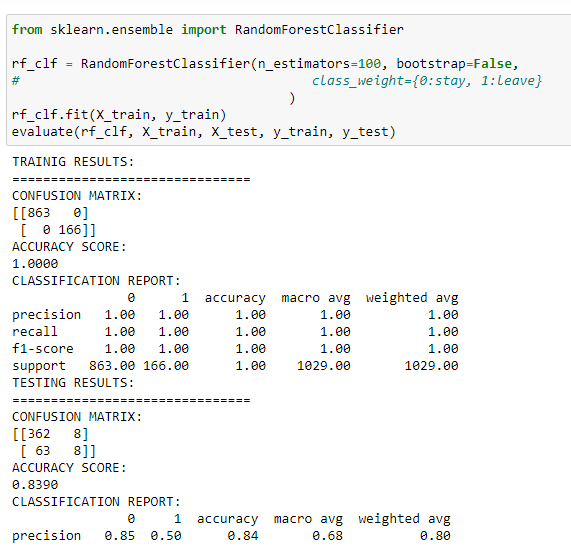
**Selecting important features**

****

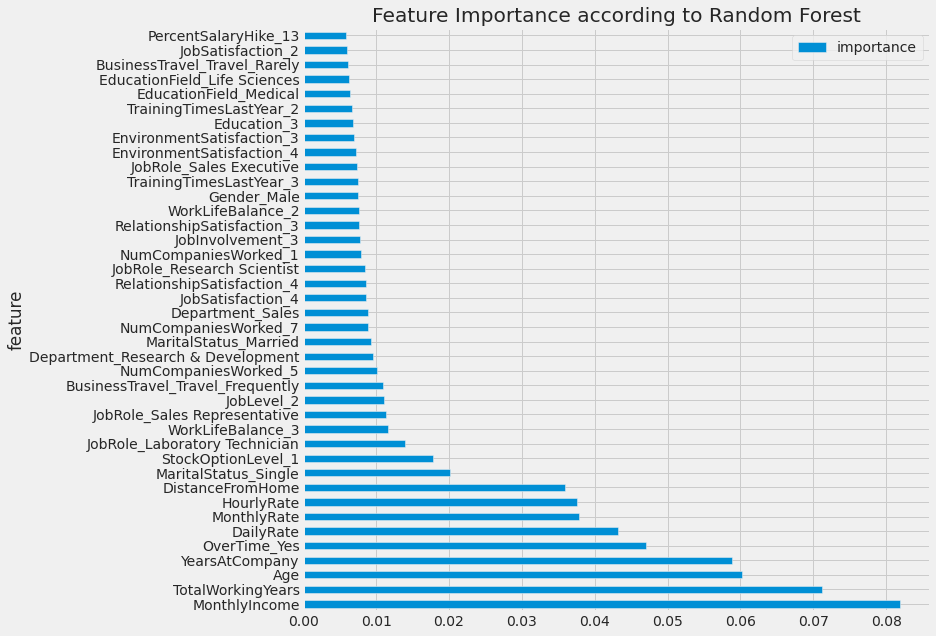
**Logistic Regression**

****

**Random Forest Classifier**

****

**Feature importance according to random forest**

****

**DESCRIPTION:**

Now a days most of the employees quit or leave the job because of various factors and some stay even with low performance. Our primary goal is to predict the employee’s attrition and also for measuring the employee’s performance.

**SAMPLE CODING**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

df = pd.read\_csv("/kaggle/input/ibm-hr-analytics-attrition-dataset/WA\_Fn-UseC\_-HR-Employee-Attrition.csv")

df.head()

df.info()

df.describe()

for column in df.columns:

print(f"{column}: Number of unique values {df[column].nunique()}")

print("==========================================================")

object\_col = []

for column in df.columns:

if df[column].dtype == object and len(df[column].unique()) <= 30:

object\_col.append(column)

print(f"{column} : {df[column].unique()}")

print(df[column].value\_counts())

print("====================================")

object\_col.remove('Attrition')

from sklearn.preprocessing import LabelEncoder

label = LabelEncoder()

df["Attrition"] = label.fit\_transform(df.Attrition)

col = df.corr().nlargest(20, "Attrition").Attrition.index

plt.figure(figsize=(15, 15))

sns.heatmap(df[col].corr(), annot=True, cmap="RdYlGn", annot\_kws={"size":10})

# Transform categorical data into dummies

dummy\_col = [column for column in df.drop('Attrition', axis=1).columns if df[column].nunique() < 20]

data = pd.get\_dummies(df, columns=dummy\_col, drop\_first=True, dtype='uint8')

data.info()

from sklearn.model\_selection import train\_test\_split, GridSearchCV

from sklearn.preprocessing import StandardScaler

X = data.drop('Attrition', axis=1)

y = data.Attrition

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42,

stratify=y)

scaler = StandardScaler()

X\_train\_std = scaler.fit\_transform(X\_train)

X\_test\_std = scaler.transform(X\_test)

X\_std = scaler.transform(X)

from sklearn.linear\_model import LogisticRegression

lr\_clf = LogisticRegression(solver='liblinear', penalty='l1')

lr\_clf.fit(X\_train\_std, y\_train)

evaluate(lr\_clf, X\_train\_std, X\_test\_std, y\_train, y\_test)

from sklearn.ensemble import RandomForestClassifier

rf\_clf = RandomForestClassifier(n\_estimators=100, bootstrap=False,

# class\_weight={0:stay, 1:leave}

)

rf\_clf.fit(X\_train, y\_train)

evaluate(rf\_clf, X\_train, X\_test, y\_train, y\_test)

1. **RESULT AND DISCUSSION**

Thus based on the project there is some solution for HR and company to find employee attrition and measure employees performance. The project depends on data because the attrition is calculated from employee’s data so the data must be original and more training data can increase the model accuracy also.

1. **CONCLUSION AND FURTHER WORK**

This project incorporates all the requirements. In this paper, it deeply study the employee’s profile data to measure the performance. It calculates the similarity of employee’s attrition and logistic regression is used to classify them. Random forest classifier is used to improve the model performance. In future an application development on this will be useful in tracking performance and attrition of employees.

**REFERENCES:**

<https://esource.dbs.ie/handle/10788/3497>

<https://www.kaggle.com/faressayah/ibm-hr-analytics-employee-attrition-performance>

<https://blog.perceptyx.com/employee-attrition-analytics>

<https://incentius.com/blog-posts/analyzing-attrition-performance/>

<http://inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/January2018FBL/IBM_Attrition_VSS.html>