**PROJECT REPORT**

1. **INTRODUCTION**
   1. **TITLE:** A study on employee attrition using supervised machine learning.
   2. **OBJECTIVES OF RESEARCH:**

* Describe the input and output of the classification model for the given secondary data set.
* Tackle both binary and multiclass classification cases.
* Develop various classification models for the given secondary data set.
* Identify the optimal model based on it’s accuracy and other metrics.
* Implement these techniques in python.
  1. **PROBLEM STATEMENT:**

“Attrition is said to be the gradual reduction in the number of employees through retirement, resignation or death. It can also be said as Employee Turnover or Employee Defection. “Attrition is a critical issue and pretty high in the industry these days. It’s the major problem which highlights in all the organizations. Whenever a well-trained and well adapted employee leaves the organization, it creates a vacuum. So, the organization loses key skills, knowledge and business relationships. Modern managers and personal administrators are greatly interested in reducing attrition in organizations, in such a way that it will contribute to the maximum effectiveness, growth and progress of the organization

Retaining employees is a critical ongoing effort. One of the biggest challenges in having managers in the place that understands it is their responsibility to create and sustain an environment that fosters retention. Staff requires reinforcement, direction and recognition to grow and remain satisfied in their positions.

The main objective of this study is to know the reasons, why attrition occurs, to identify the factors which make employees dissatisfy, to know the satisfactory level of employees towards their job and working conditions.

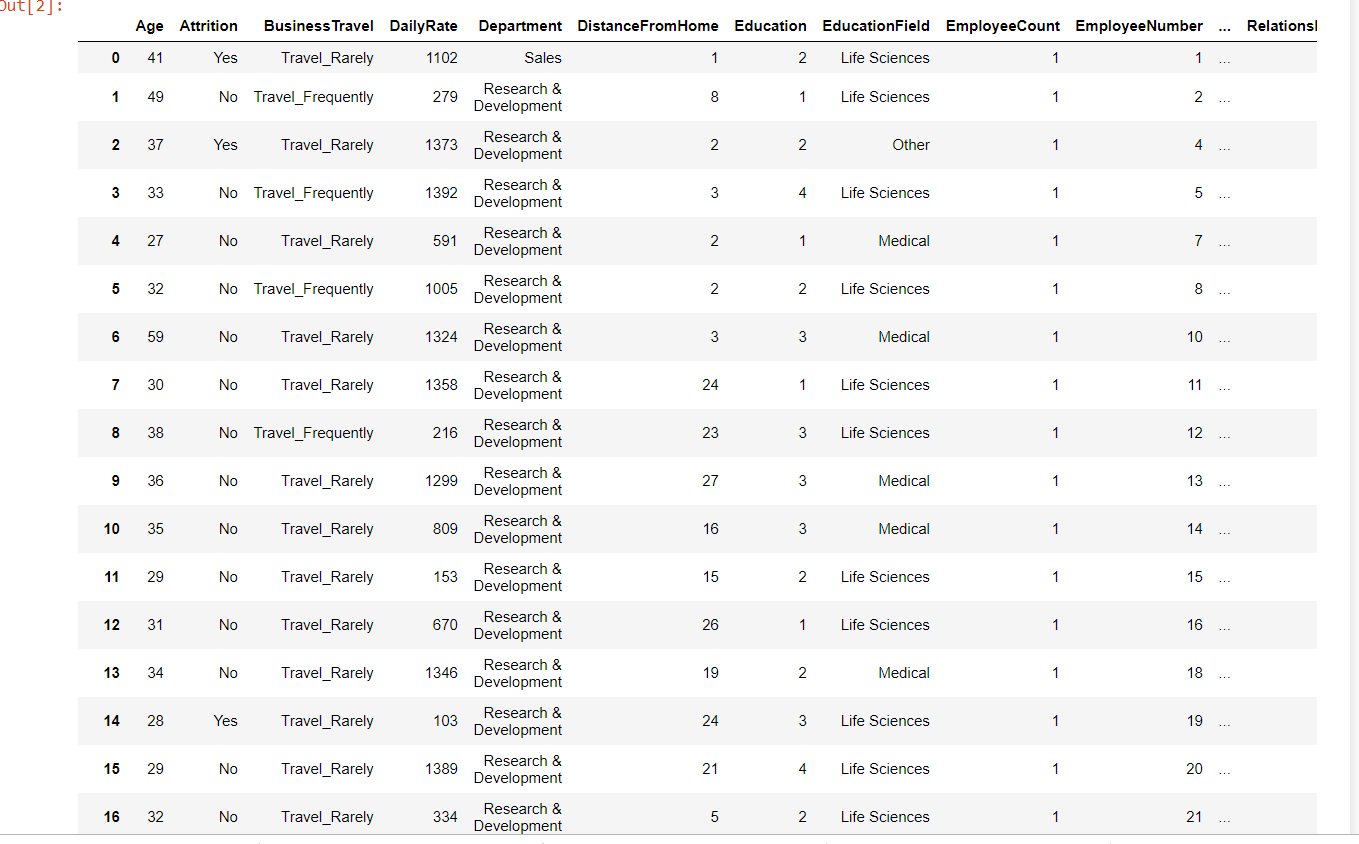
1. **LITERATURE REVIEW**:

Despite several studies carried out on employee retention, the strategic human resource researchers are still investigating the causal mechanisms between HR practices and firm’s performance mostly related to voluntary turnover as a critical component (Shaw, Gupta and Demery, 2005) as employee retention plays a vital role in bridging the gap between the macro strategies and micro behaviour in Organizations. This is because it ensures stability and connects the experiences of individuals in Organizations on a continuous basis to the critical measures of success factors in the Organization. The decision of leaving the Organization is not easy for an individual employee as well as significant energy is spent on finding new jobs, adjusting to new situations, giving up known routines and interpersonal connection and is so stressful (Boswell, Boudreau and Techy, 2005). Therefore, if timely and proper measures are taken by the Organizations, some of the voluntary turnover in the Organization can be prevented. The reasons for employee turnover may vary from external environmental factors such as economy that influence the business that in turn affects the employment levels (Pettman 1975; Mobley, 1982, Schervish, 1983; Terborg and Lee, 1984) to Organizational variables such as type of industry, occupational category, Organization size, payment, supervisory level, location, selection process, work environment, work assignments, benefits, promotions and (Mobley, 1982; Arthur, 2001). The other factors that influence employee turnover in Organizations include the individual work variables like demographic variables, integrative variables like job satisfaction, pay, promotion and working condition (Pettman, 1975; Mobley 1982; Arthur 2001) and the individual nonworking variables such as family related varibles (Pettman, 1975; Mobley, 1982;). Any of the above factors could be the reasons, but the decision process to leave or stay in the Organization is to be periodically examined to understand the specific reasons that prompted them to take such a step and the Organizations should be mainly concerned about voluntary turnover and not involuntary turnover as it is within their control. Also, it is found that employees who perform better and are intelligent enough have more external employment opportunities available compared to average or poor performance employees and thus they are more likely to leave (Trevor, 2001). High rates of voluntary turnover of such employees are often found to be harmful or disruptive to firm’s performance (Glebbeck & Bax, 2004). Therefore, the acquisition, development and retention of talent form the basis for developing competitive advantage in many industries and countries (Pfeffer, 1994, 2005). Organizations failing to retain high performers will be left with an understaffed, less qualified workforce that ultimately hinders their ability to remain competitive (Rappaport, Bancroft, & Okum, 2003). Three studies incorporated attitudinal and/or behavioural changes over time to better predict turnover. Sturman and Trevor (2001) found that quitters’ performance over time did not significantly change while stays’ performance slope was positive. Demographic factors cannot be ignored as age, tenure, level of education, level of income, job category, gender have influenced employee retention and have been found to have stable relationship with turnover intention.

1. **DATA COLLECTION:**

Secondary data shared by speakers and mentors in the form of dataset.

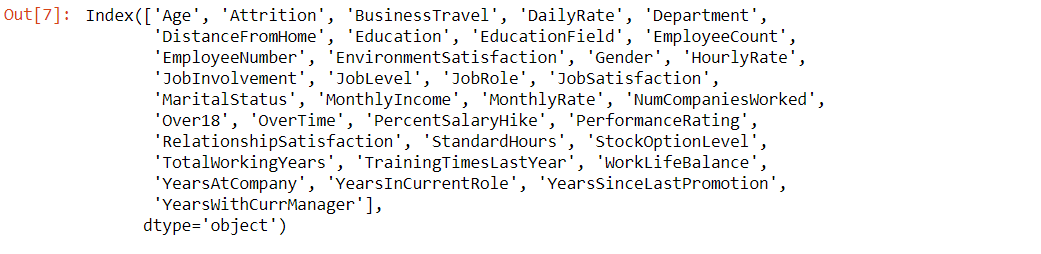
1. **DATA ANALYSIS AND INTERPRETATION:**
   1. **IDENTIFYING DEPENDENT AND INDEPENDENT PARAMETERS:**

****

The data set is read using pandas and is analyzed as shown above.

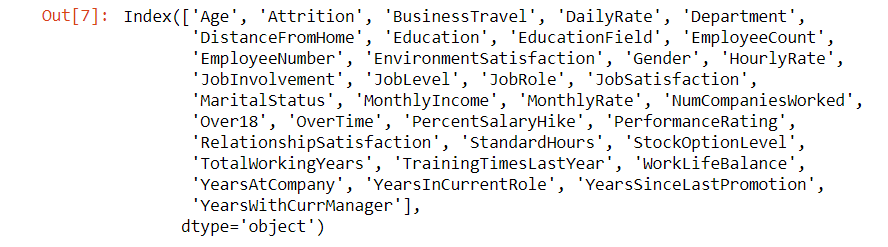


We observe that there are 1470 rows and 35 columns.

By analyzing all the columns, we see that the column “Attrition” refers to the dependent parament while all the other parameters are independent.

* 1. **EXPLORATORY DATA ANALYSIS:**

**Observing the columns:**



By observing the columns, we can come to a conclusion that “EmployeeNumber” and “EmployeeCount” do not affect the dependent parameter.

By further analyzing univariate boxplots, we can drop column “StandardHours” as all the employees have got the same value.

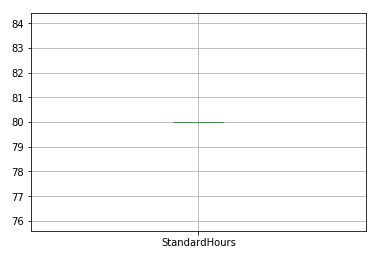


FIG 1

By examining the column “Over18” we find that all the employees are above 18 so even this column can be dropped.



**Understanding the reasons for positive attrition:**

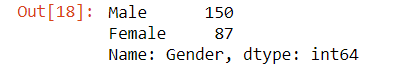
For this, we consider the subset of the given dataset such that it consists only the data of the employees who have attrited.



We observe that there are 237 employees who have attrited.

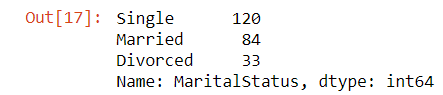
**Affect of categorical parameters on positive attrition:**

Let us first see how the gender affects the attrition.



We observe that males are more likely to attrite than females.

Based on marital status:

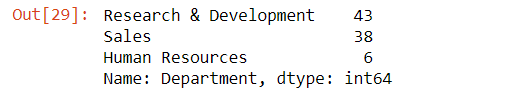


From this we can say that singles are more likely to attrite while attrited employees with divorced marital status are very few.

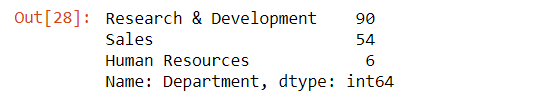
Furthermore, we classify the subdatset into smaller sets by considering males and females separately and inspect the data based on various other parameters.

The following are some noteworthy observations:

Female Employees and Department:

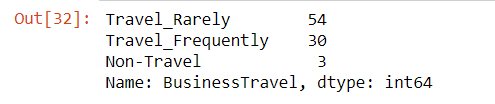


Male Employees and Department:

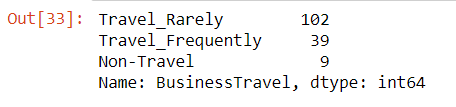


From this we can observe that in both the cases employees belonging to Research & Development department attrited the most.

Female Employees and BusinessTravel:



Male Employees and BusinessTravel:



Thus, employees who travelled rarely attrited more, while, the employees who did not travel attrited less.

**VISUALISATION:**

**PercentSalaryHike:**

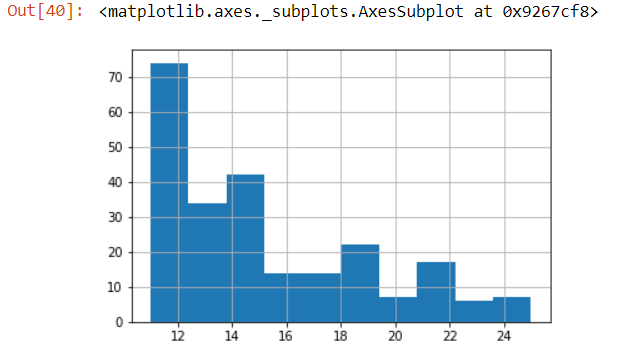
****

FIG A

We observe that Employees with less Salary Hike attrited the most.

**DistanceFromHome:**

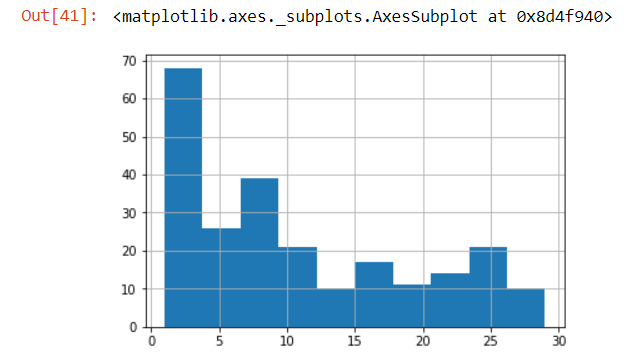


FIG B

This is an interesting observation which states that employees who lived close by attrited the most.

**JobLevel:**

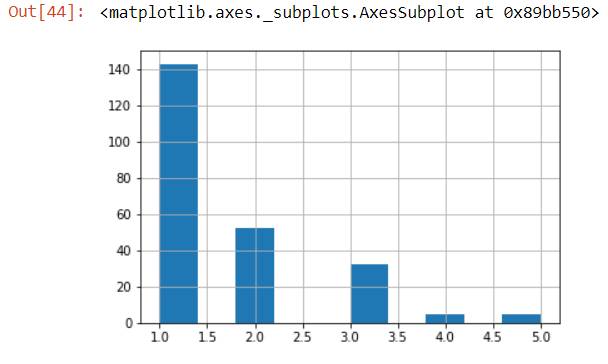
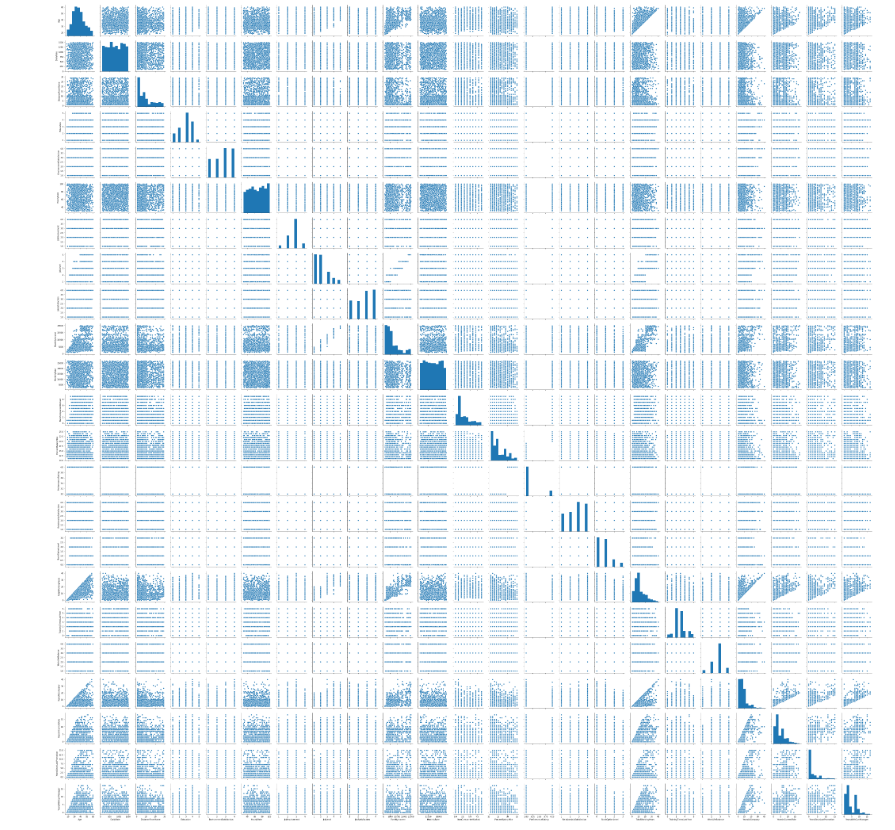
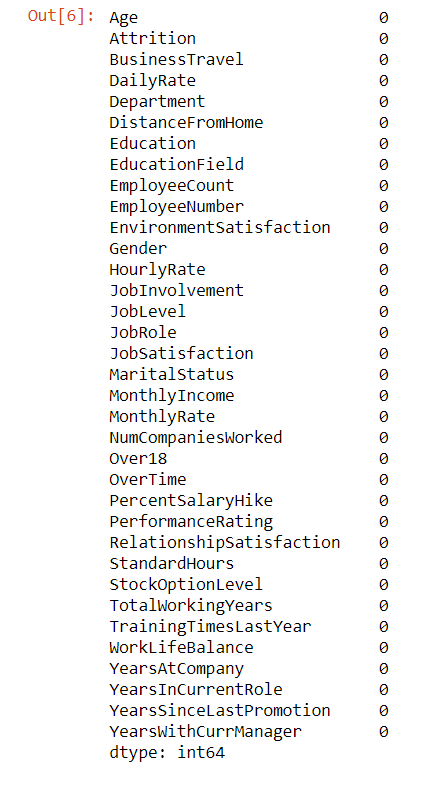
****

FIG C

We see that as the Job level increased the number of attrited employees decreased.

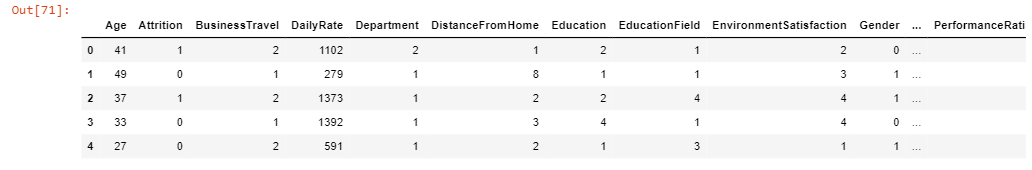
**Pairplot:** 

**4.3.DATA CLEANING:** Missing values: 

We observe that the given data set is free from missing values, hence, data imputation is not necessary.

**RECODING:**

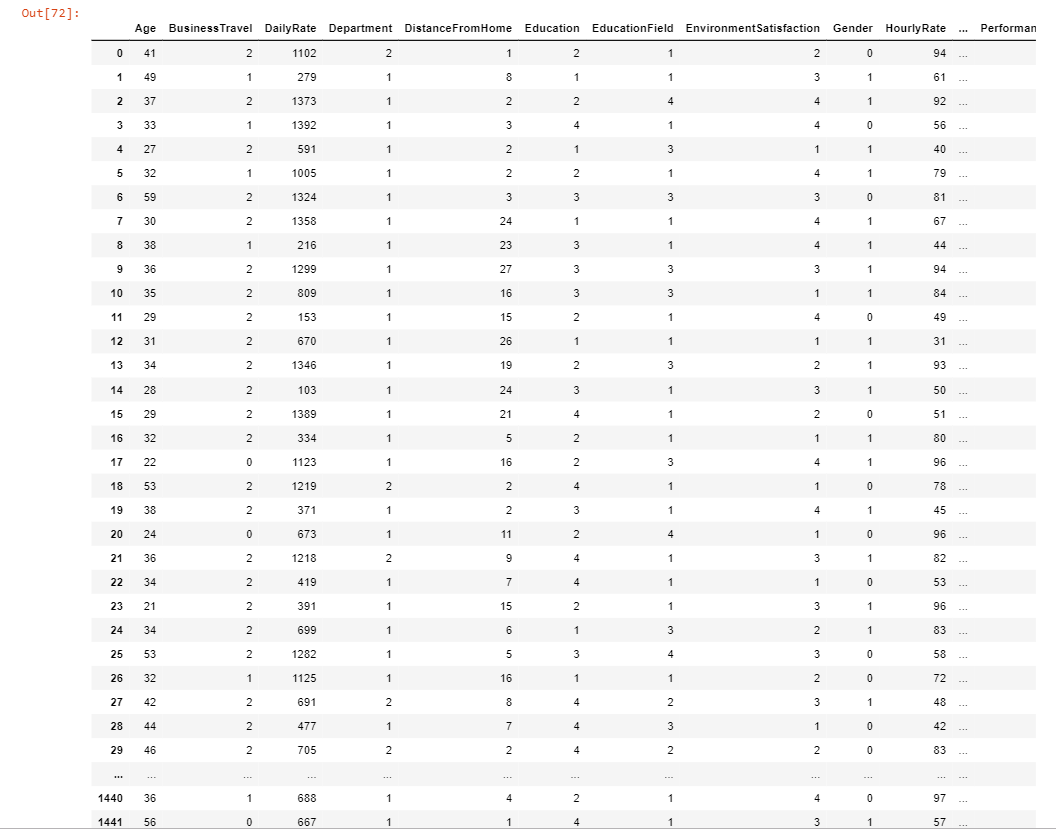
The categorical data present in the dataset are recoded in order model it properly.

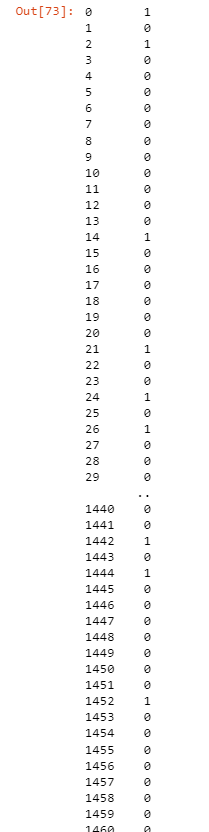


Now, all the categorical columns are recoded as shown above.

**FEATURE ENGINEERING:**

The dependent parameter attrition is dropped and stored in another variable so as to differentiate between independent and dependent parameters.





Attrition column stored in another variable is as shown above.

**Splitting the data into test and train data:**

Train data:

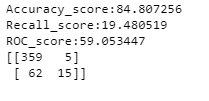


Test data:



**MODELLING (Key Metrics):**

**LINEAR REGRESSION:**

****

**ROC curve:**

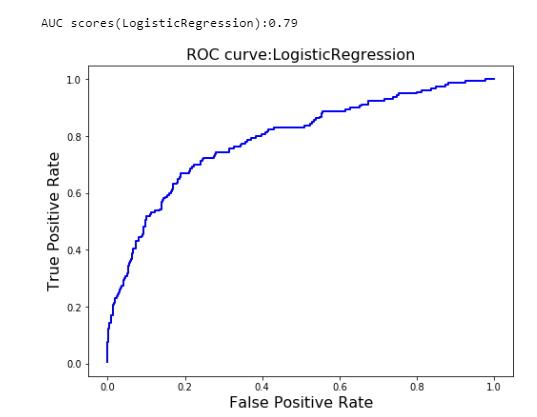
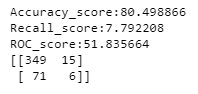
****

FIG I

**KNN:**

****

**ROC curve:**

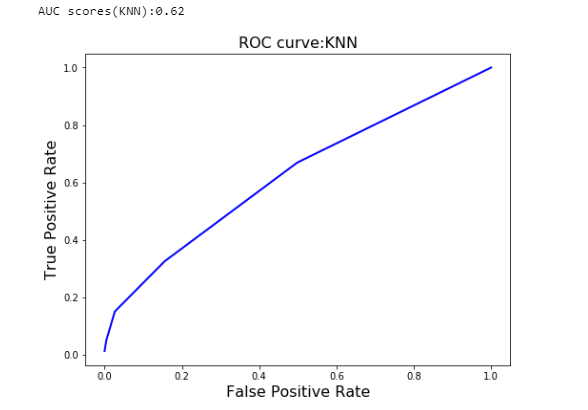
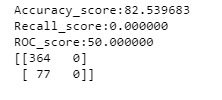
****

FIG II

**SVM:**

****

**ROC curve:**

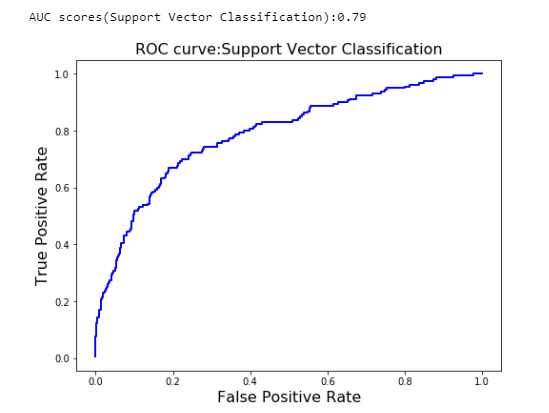
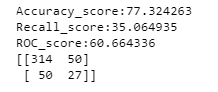
****

FIG III

**DECISION TREE:**

****

**ROC curve:**

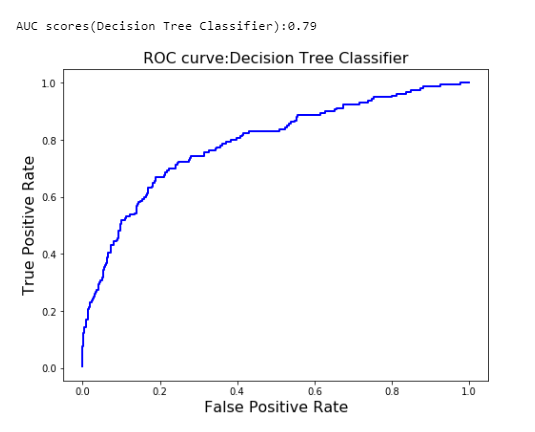
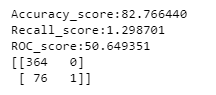
****

FIG IV

**RANDOM FOREST:**

****

**ROC curve:**

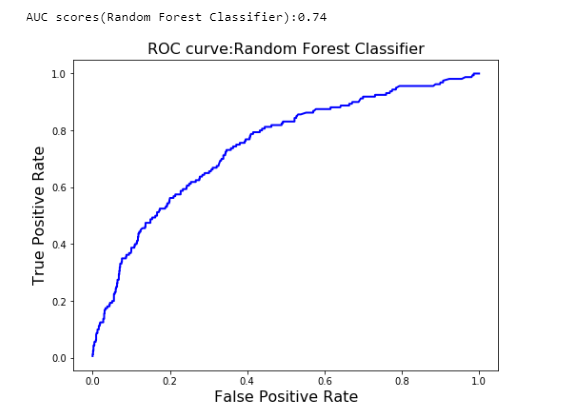
****

FIG V

1. **FINDINGS AND SUGGESTIONS:**

* Males are more likely to attrite than females.
* Singles are more likely to attrite while attrited employees with divorced marital status are very few.
* Employees belonging to Research & Development department attrited the most.
* Employees who travelled rarely attrited more, while, the employees who did not travel attrited less.
* Employees with less Salary Hike attrited the most.
* An interesting observation states that employees who lived close by attrited the most.
* As the Job level increased the number of attrited employees decreased.

After applying various models of classification, we find that the accuracy of all the models is quite toothsome study other key metrics as well.

By examining other key metrics, we observe that Decision Tree has got an excellent confusion matrix which has greater true positive and true negative values, while the false positive and false negative values are less, which is optimal.

Also, the recall score of Decision tree is more which indicates that it is a good model which is free from TYPE I and TYPE II error.

The AUC score of Decision tree is also closer to 1, which is ideal.

1. **CONCLUSION**:

* The input and output of the classification model are described.
* Both the binary and multiclass classification cases are tackled.
* Various classification models for the given secondary dataset are developed.
* Decision Tree model is identified to be the optimal model based on the key metrics.
* The above techniques are implemented in python.

1. **REFERENCES:**

<https://medium.com/datadriveninvestor/choosing-the-best-algorithm-for-your-classification-model-7c632c78f38f>

<https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>

1. **DOCUMENTED BY :**

**K NAGA SRI NAVYA**

Mail id:[kotanavya456@gmail.com](mailto:kotanavya456@gmail.com)



**THANK YOU**