# Human Resource Manager's Prediction

## Introduction

The purpose of this project is to predict the employee who leaves the company. It will help the company to make the right decision on the employee willing to leave.

For a company hiring and retaining the employees is a complex task that require time and skills,

With only several people on board, every new hire is crucial for the team's success.

Most of all, small business owners spend around 40 percent of their work on tasks that do not generate income, such as hiring.

The average cost to hire an employee is minimum of 20000/- per head in India, with around 52 days to fill an empty position.

## Problem

With taking all the work in the process of hiring and the details of employees with their interests. Predicting the employee who is willing to leave the company and get to know the precautions need to be taken by the company in order to retain the employee based on the drawbacks.

## Data Description

Data Acquired from  Kaggle

Available at https://github.com/karthikband1/Human-Resourse-Manager-Prediction

Dataset consists of the following,

Age, Education, Job Satisfaction, Job Involvement, Gender, Marital Status, Hourly Rate and much more.

## Libraries Used To Develop The Project

**Pandas**: For creating and manipulating data frames.

**Numpy**: For numerical calculations.

**Models Used**: Logistic Regression and Random Forest.

**Scikit Learn**: For importing Logistic Regression, Random Forest, accuracy score, confusion matrix, classification report.

**One Hot Encoder**: For converting the categorical values to the numerical values.

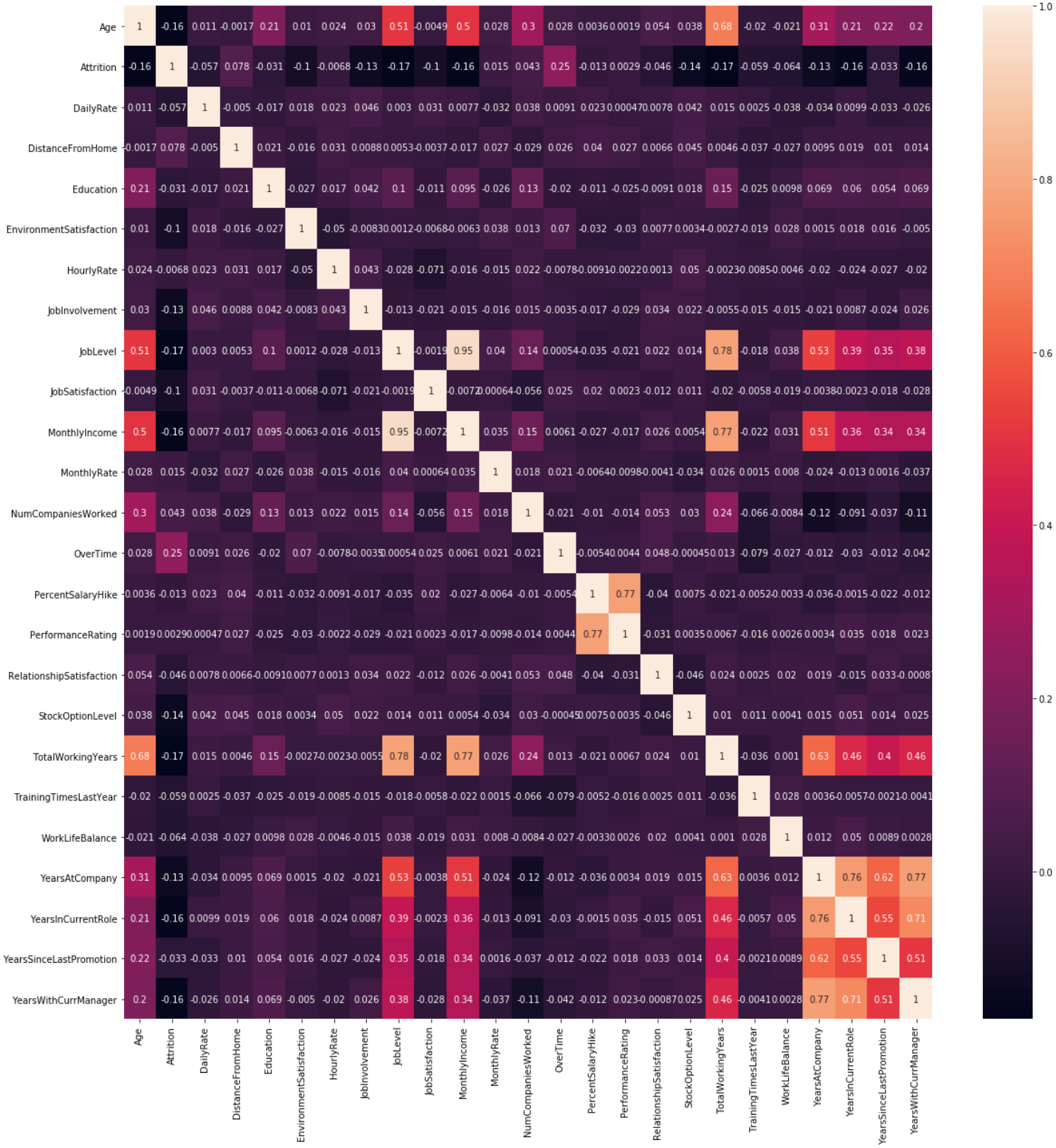
**Min Max Scaler**: Used for the normalization of the data to fit within a range.

**Seaborn**: Plotting Module.

**Matplotlib**: Plotting Module.

# Discussion Section

## Correlation



From the heatmap we have observed that the highest correlations are 0.78,0.77,0.95,0.68

0.78 - correlation between "Job Level" and "Total Working Years"

0.77 - correlation between "Monthly Income" and "Total Working Years"

0.95 - correlation between the both "Monthly Income, Job Level" and "Job Level, Monthly Income"

0.68 - correlation between the both "Age, Total Working Years"

## Left And Stayed employees

**Legend** 1-Left  0-Stayed

1.Comparison on the basis of age we come to know that no people left the company at the age of 54 and 60.The highest Left from the company is in the ages ranging from 18 to 22,atmost 50% to 75% left the company, Ages from 34 to 38 the stayed percentage of employees is much higher compared to the left one's.

2.Comparison on the basis of Job Level ,It's quite simple from the graph that the higher the job level the lower the left percentage(taking the percentage of the stayed in count).Maybe its due to the concern of money the employees get, keeping in note that the companies offer salary based on their job level.

3.The married employees are more willing to stay the company comparing to the rest married>divorce>single. Based on the Total Working Years people are more tending to leave the company at the start but shows a hike at 10 years of experience work in the company and also at the age of 40 most of the people are tending to leave the company.

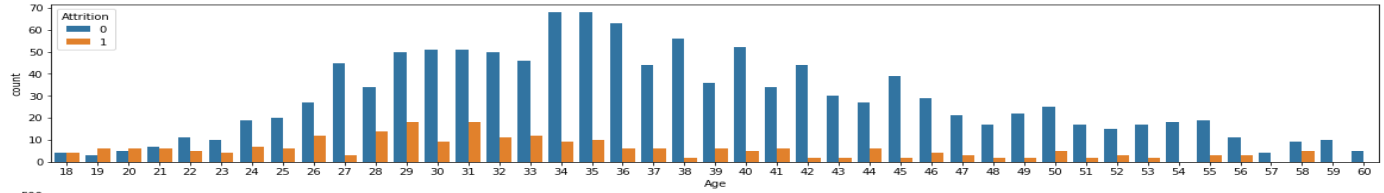
4.On the gender basis the left percentage of boys and girls is same taking the strength of the genders inconsideration, As the  Job Involvement increases the percentage of left people decreases.

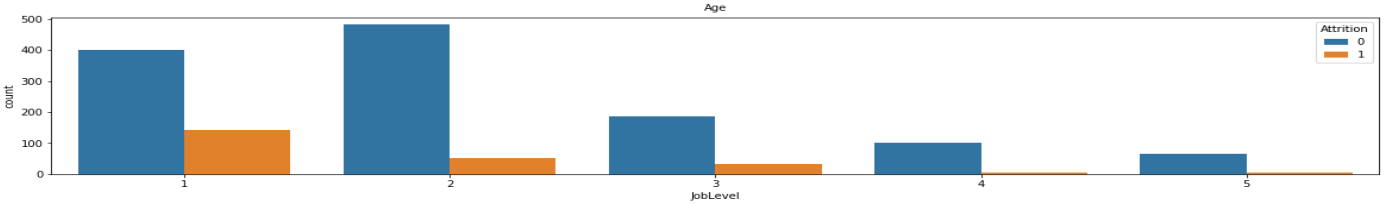
5.Most of the people shows interest at a small distance from their home, as we can see the shorter the distance the more the people stay.

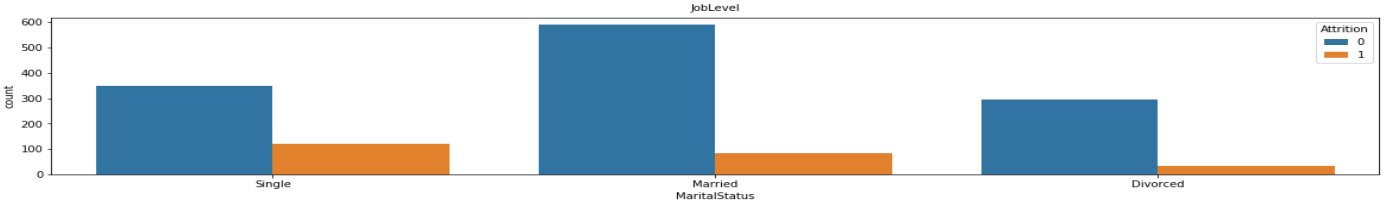
6.Its clearly shown that as the employees get promotion are staying and the at the long year of promotion we can see that there is less percentage in the both Left and Stayed, it shows that most of the employees leaving the companies after certain amount of time.

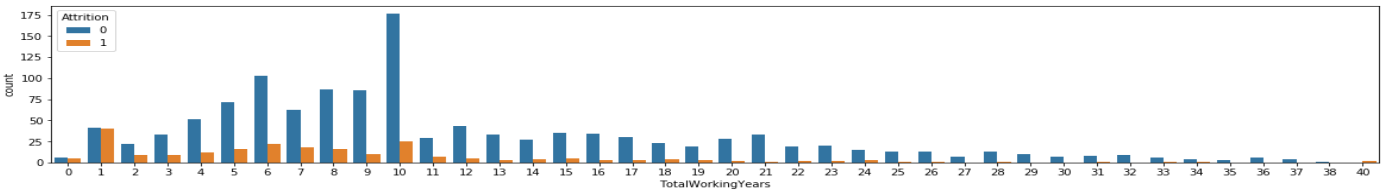
7.From the current manager comparison we come to know that as the pay increases the left percentage decreases

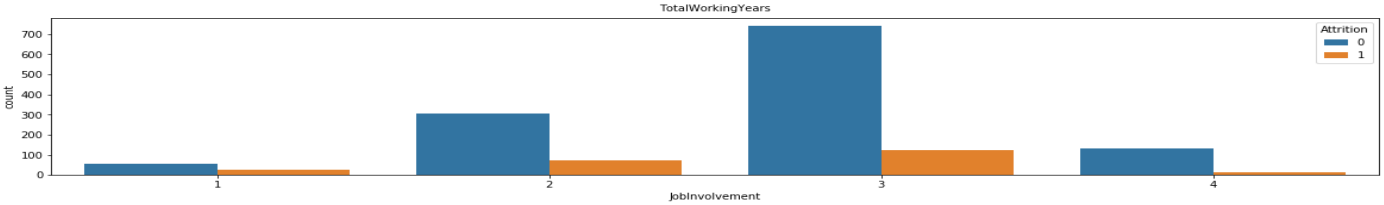
8.Comparison with hourly pay, at the pay of 30 there is no percentage of people left as it is the starting stage or we can say as they are the freshers, and at the pay of 42 there is more percentage in staying, even though the pay is high from 87 people starting to leave the company ,maybe the age and other factors place role here.

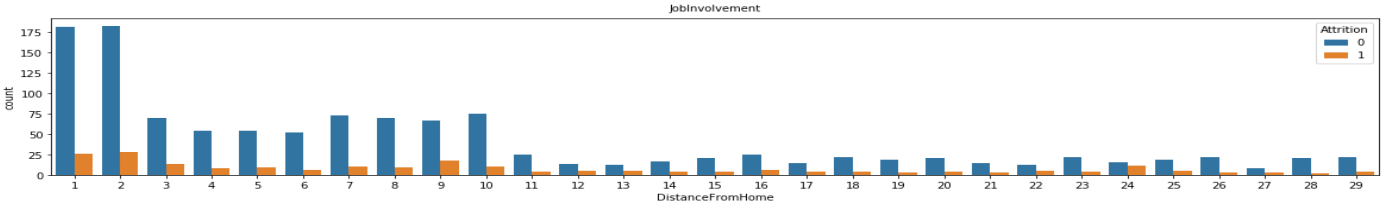

Attrition
0
1
count
Age

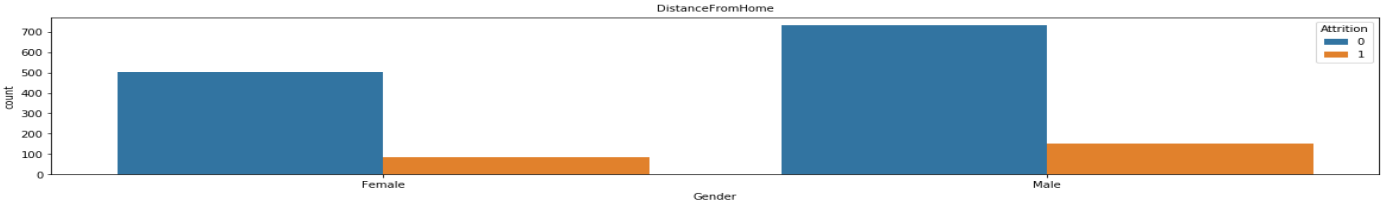

Attrition
0
1
count
JobLevel

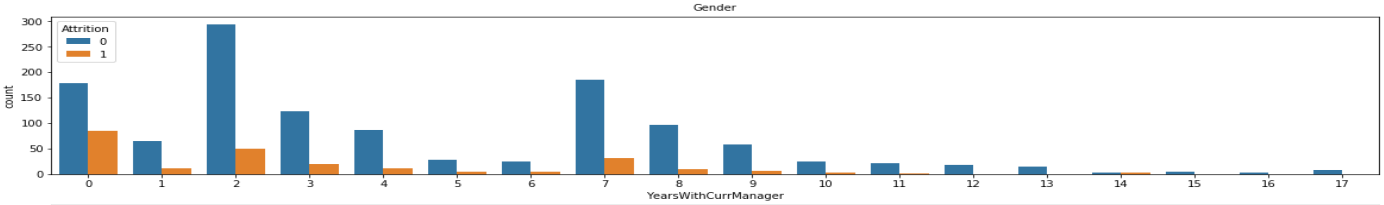

Attrition
0
1
count
Single
Married
Divorced
MaritalStatus

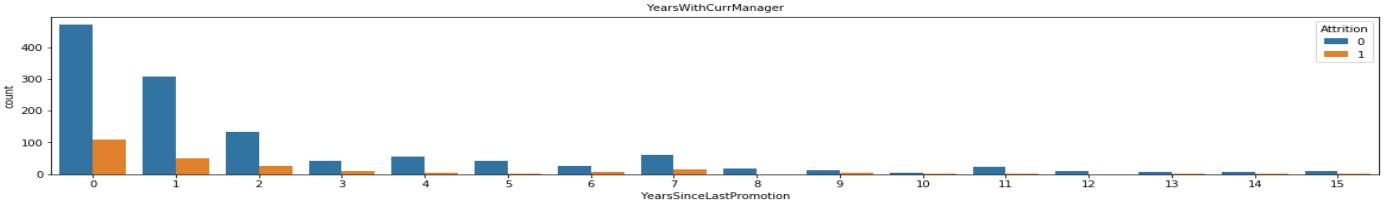

Attrition
0
1
count
TotalWorkingYears

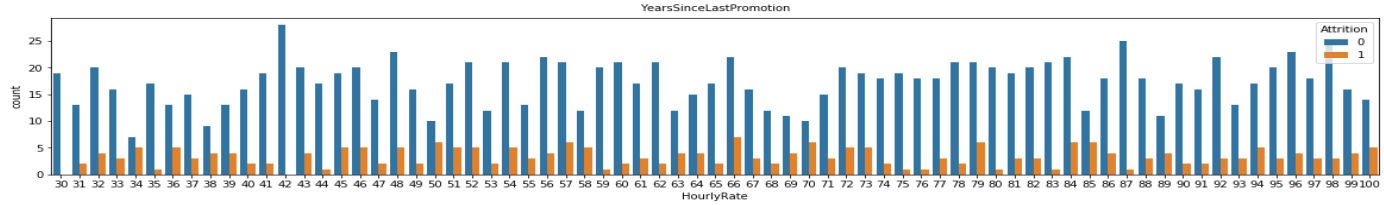

Attrition
0
1
count
JobInvolvement


Attrition
0
1
count
DistanceFromHome


Attrition
0
1
count
Female
Male
Gender


Attrition
0
1
count
YearsWithCurrManager


Attrition
0
1
count
YearsSinceLastPromotion


Attrition
0
1
count
HourlyRate

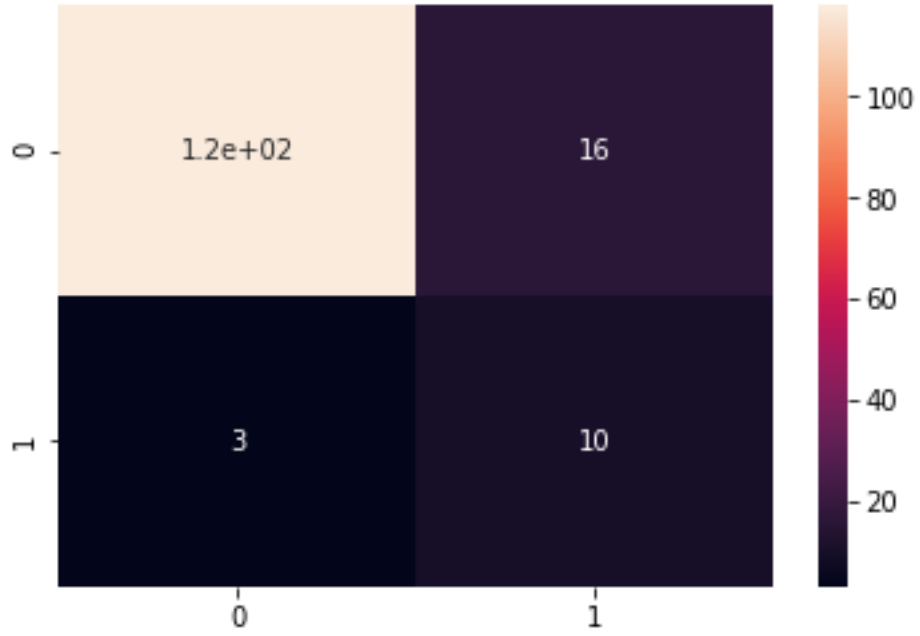
# Model Section

### Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model.

**Solver Used** liblinear

### Confusion matrix




### Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.98 | 0.88 | 0.93 | 134 |
| 1 | 0.38 | 0.77 | 0.51 | 13 |
| micro avg | 0.87 | 0.87 | 0.87 | 147 |
| macro avg | 0.68 | 0.82 | 0.72 | 147 |
| weighted avg | 0.92 | 0.87 | 0.89 | 147 |

### Accuracy Score

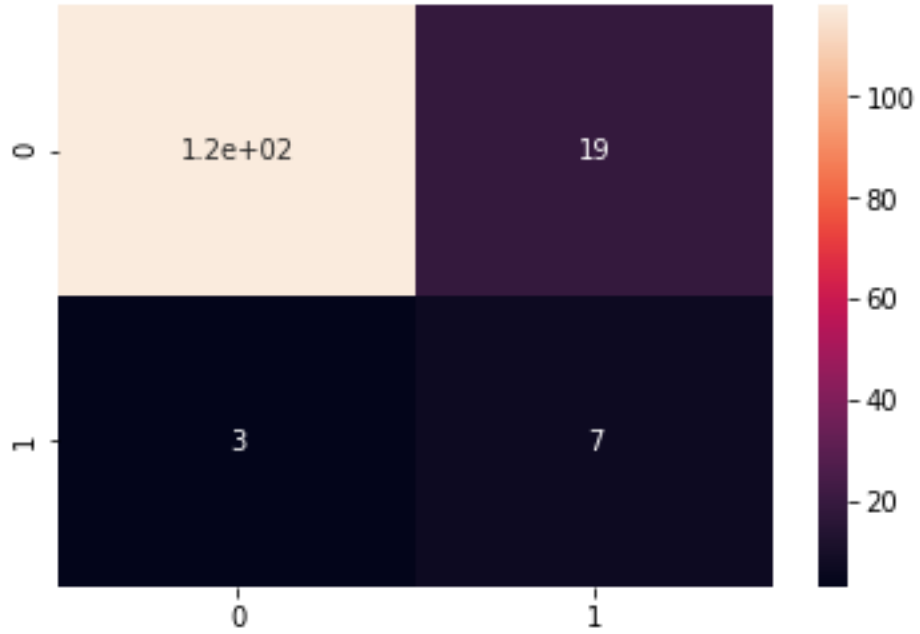
Accuracy Score Using Logistic Regression = `88.43537414965986%`

### Random Forest Classifier

Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

### Estimators Used 10

### Confusion matrix




### Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.98 | 0.86 | 0.91 | 137 |
| 1 | 0.27 | 0.70 | 0.39 | 10 |
| micro avg | 0.85 | 0.85 | 0.85 | 147 |
| macro avg | 0.62 | 0.78 | 0.65 | 147 |
| weighted avg | 0.93 | 0.85 | 0.88 | 147 |

### Accuracy Score

Accuracy Score Using Random Forest Classifier = 85.03401360544217%

# Conclusion Section

In this project I have separated the data into categorical, numerical and converted the categorical values to the numerical values using one hot encoding and also used min max scaling to normalize the data within a range of 0 and 1.

Taking Attritions from data set as y(target variable), I have trained the remaining data with the model selection such as Logistic Regression and Random Forest Classifier and got an accuracy score of 88 and 85 percentages respectively.

This project has shown me a practical application to resolve a real situation that has impacting personal and financial impact using Data Science tools.

BY,

BANDI KARTHIK,

20/07/2020.