e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

A RANDOM FOREST CLASSIFICATION ON EMPLOYEE ATTRITION ANALYSIS AND HANDLING OF IMBALANCED DATA

Tandrangi Venkatadri Naidu*1, Dr. Vanitha Kakollu*2

*1PG Student, Department Of Computer Science, GITAM (Deemed To Be University), Visakhapatnam, Andhra Pradesh, India.

*2Assistnat Professor, Department Of Computer Science, GITAM (Deemed To Be University), Visakhapatnam, Andhra Pradesh, India.

ABSTRACT

Every organization's major valuable assets are employees. The recent information states that the present pandemic laid the foundation to the very high attrition rates in all the organizations particularly it showed a huge impact in the IT industry compared to the other industry. The persistence process of hiring the new employees and training and development of employees is cost-effective. The recruitment team of all the organizations did tons of research to analyse to understand the right reason behind the employees leaving the corporate or shifting to other organizations for better career growth. In this project here a sample dataset is maintained about the employee's data related to attrition that caused the employees to leave the organization. By extracting the required features from the dataset, the data pre-processing techniques are implemented to remove anomalies and the noise free data is considered for exploratory data analysis using different data visualization techniques. As the featured data is highly imbalanced the under-sampling technique is implemented to make it balanced to minority class labels. The Random Forest Classification model is used to predict the accuracy of the balanced dataset of the employee attrition.

Keywords: Employee Attrition, Machine Learning, Exploratory Data Analysis, Imbalanced Data, Random Forest Classification.

I. INTRODUCTION

In recent years it had a lot of the impact over the people and even the corporate world mostly on the IT industry. As per the reports which are released recently states that there is the highest attrition rate recorded since the previous collective of centuries. Most of the people know employees are the most valuable resource to any corporate sector, the process of hiring new employees, training and development is cost-effective. Every single individual has their own choice of reasons to leave the companies whether it might be professional growth or the impact of work-life balance or issues related to job satisfaction. The recruiting team in any company is unaware of the proper reason for the increase of the attrition rate as well as reasons behind leaving the organization and by knowing the majority of the reasons which are enlisted can help the recruiting team to improve it to lower the attrition rate. So to make it easy I have taken the sample data set which contains the reasons behind the employee attrition and use the random forest classification machine learning model which supports to detect the attrition rate precisely to make sure the organization implements new policies or an employee retention strategies to overwhelm the mentioned reasons and help the employee to stay loyal to the organization by satisfying their needs and requirements to lower down the cost of a new hiring and training as well as the attrition rate.

The most significant number of researchers was analysed the dataset by visualization graphs and illustrative curves, tables, and others. Especially in this concept, an ML model was trained, optimized and evaluated to predict whether a certain employee will leave the company or not and according to this predication the company will improve different retention strategies on targeted employees.

Employee attrition is when an employee leaves the corporate through any method, including voluntary resignations, layoffs, failure to return from a leave of absence, or maybe illness or death. Whenever anyone ceases working for the corporate for any reason and isn't replaced for an extended time if ever, that might be employee attrition. Here are two main sorts of employee attrition:

A. Voluntary Attrition

B. Involuntary Attrition

A. Voluntary Attrition is the most common type of attrition, where employees decide to simply quit their jobs. There are often many reasons for voluntary attrition (more thereon later) and most of them are in your control. You should proactively attempt to curb voluntary attrition among high-value talent, as this will bring down your productivity over time. For example, if a corporation sees its marketing experts moving out of various business units, it's a transparent cause for concern.

B. Involuntary attrition the corporate and not the worker that initiate the exit. For example, the worker may have shown instances of misconduct within the workplace – a standard reason for involuntary attrition. Structural reasons could also cause attrition. Mergers and acquisitions are frequently followed by a wave of involuntary attrition. Involuntary attrition through position elimination is that the commonest sort of attrition, because the company decides proactively to eliminate an edge. For other sorts of termination, the corporate usually decides after the termination to go away the work vacant.

In this paper mainly focused on Involuntary Attrition and the most common reasons for the employee attrition at any Organization are retirements, transfers, a resignation or a termination, deaths, layoffs, personal health issues etc.

The major causes for the employee attrition in any organization are being overworked or if work-life balance is lacking in the organization, maybe if an employee hates travelling frequently from one place to another place or if an employee hates his/her boss, or if an employee hates to work in the current department and if the distance from the working place to the home is more or if an employee salary is low and there is lack of recognition or a promotion in the organization, even the employee leaves the organization due to personals problems or even for the better career growth. The retention reasons which an employee considers staying in an organization if the employee is satisfied with the work and if the work-life balance is good in the organization, if the employee's salary is good enough according to his work and if the employee work is recognized by the organization.

II. METHODOLOGY

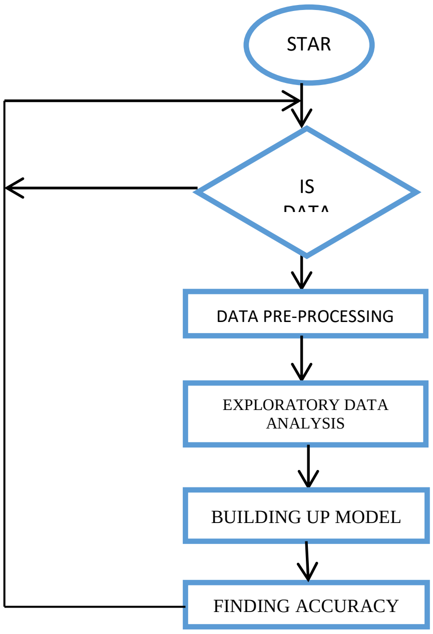


Figure 1: METHODOLOGY.

WORKING OF THE FLOW CHART:

- Import the required packages in the python platform.
- Import the dataset which we wanted to analyse and implement machine learning model.
- Data pre-processing is performed to reduce the missing values, noise etc.
- In exploratory data analysis we start comparing different reasons for leaving the organization.

- We use machine learning technique to balance the data and implement a model which predicts the accuracy of the data which is present in the dataset.

III. PROPOSED SYSTEM

As we know the data which companies maintain regarding the attrition reasons when the employees leave to contain a large amount of entries. To evaluate all the reasons for the data will be difficult work for the Human Resource team. So this paper helps the team to find the major reasons by analysing the complete data and also using machine learning classification models to find the accuracy. Here are the requirements to complete data analysis and implement a learning model: Windows/Linux/macOS any version, hence it can run on any platform. The latest version of the python to be installed in your system to run it successfully along with certain packages such as Pandas, Seaborn, Matplotlib, Ploty.express, sklearn. In terms of hardware requirements there is not much required at all but still certain requirements are a Laptop or the PC with a minimal configuration which can run the python efficiently.

IV. SYSTEM ARCHITECTURE

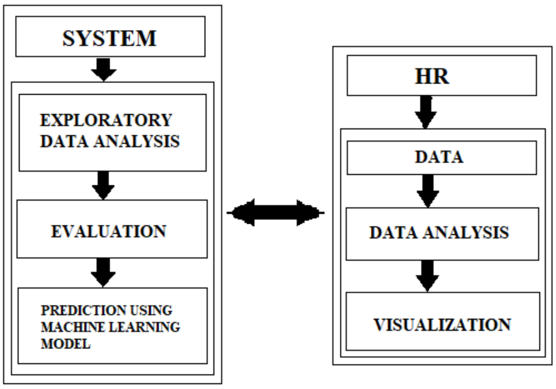

SYSTEM
EXPLORATORY DATA ANALYSIS
EVALUATION
PREDICTION USING MACHINE LEARNING MODEL
HR
DATA
DATA ANALYSIS
VISUALIZATION

Figure 2: SYSTEM ARCHITECTURE

From the above Figure2, the Human Resources Team of any organizations maintain the data regarding each and every employee who are working in the organization as well as who left the organization. This data is used for employee analysis like identifying total working hours, monthly income, last promotion year, salary hikes, reason leaving the organization etc. By using the data the data analysis is performed and the data visualization charts are prepared to analyse the data. The similar data can be used to import into the system where the Exploratory Data Analysis is performed for the reasons which an employee leave the organization and the implementation of the machine learning model used to predict the data in the dataset.

V. EVALUATION PROCESS

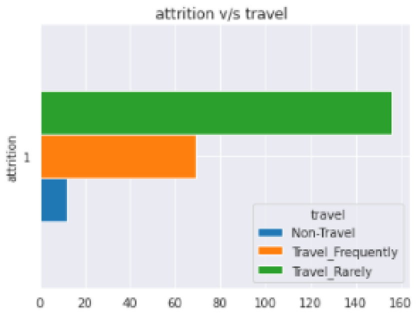
In this paper, we have taken the dataset related to employee attrition from the Kaggle. The dataset consists of 2397 observations with 35 attributes related to employee information. Among the all the attributes primarily we are focused on certain attributes such as business travel, distance from home, environment satisfaction, daily rate, percent salary hike, performance rating, relationship satisfaction in the organization, number of years taken for promotion or for getting the current role associated with the organization. All these attributes will help us to find out the major reason for the employee attrition in the organization as these might be major part of the attrition reason which is considered by employees while leaving the organization.

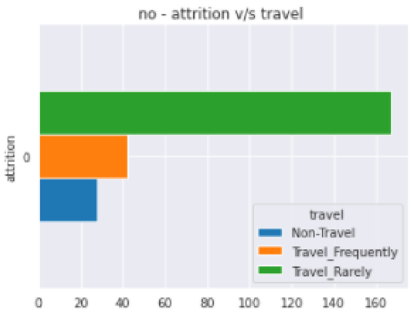
Once we import the dataset into the program, the data pre-processing is the initial step which helps us to modify the data which is suitable for building proper machine learning model related to the data. Before using data pre-processing techniques we change the attributes name to improve the readability while analyzing the data from the dataset. As part of data pre-processing we try to eliminate the missing values but in the dataset which we using doesn’t have the missing values and we check whether the data is balanced or not. When we check the current the dataset it clearly showed that the data is imbalanced. To make it balanced data we used under sampling machine learning technique to decrease the majority class label up to count of minority class label. In the Exploratory Data Analysis we try to perform evaluation on the data to discover patterns or spot anomalies or to check the assumptions with the help of summary and graphical representation of the data.

During this process we are primarily focused on what caused an employee to leave his or her organization. For this as we mentioned earlier we had taken few attributes as the reason and summarize the data in graphical form to find the result. As part of data analysis process the we tried to compare different type of attributes such as travel, Department, Distance from home, environmental satisfaction, job satisfaction, gender, job involvement, relationship satisfaction, work life satisfaction, hikes along with the attrition. Once the data analysis is completed we try to balance the data using under sampling technique and build the model with predict the accuracy of the reason which are taken in the data analysis process. In this paper we implemented the random forest classification model to predict the accuracy of the data by training the model.

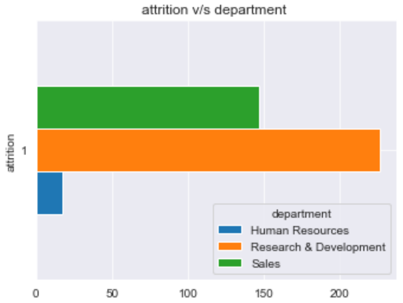
VI. RESULTS AND DISCUSSION

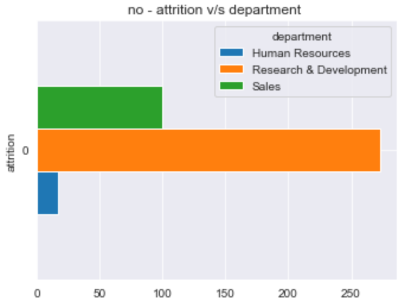
EXPLORATORY DATA ANALYSIS OUTCOMES:



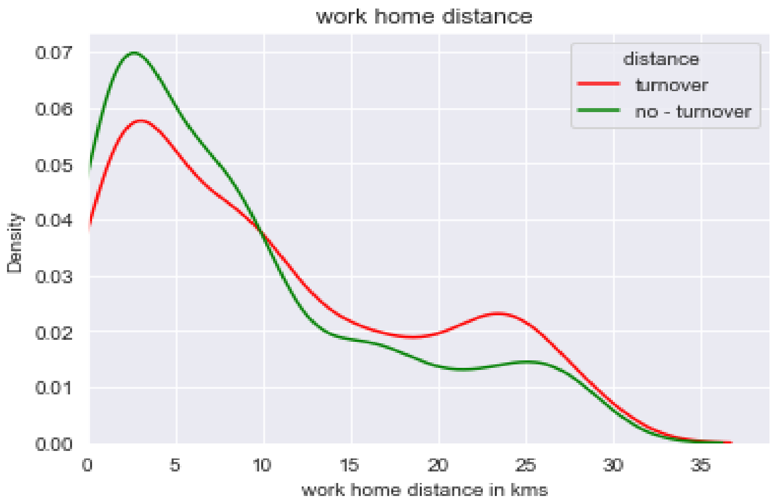


According to above chart, the resulted outcome is that most of the people who leaves the company are one of who travel frequently compared to people who don't travel.

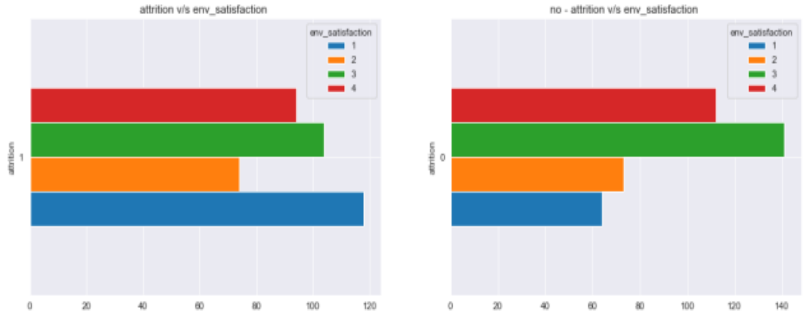




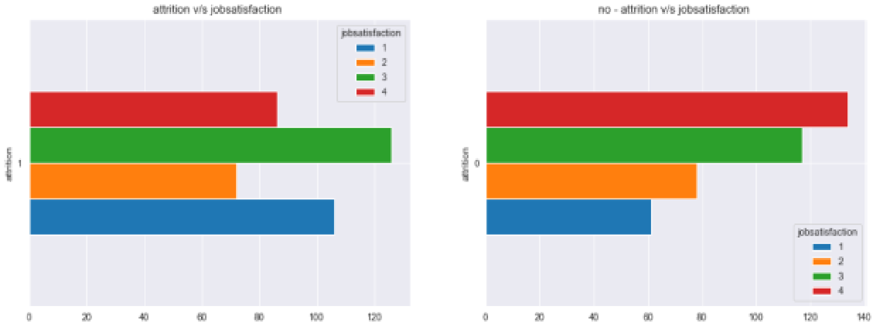
According to the above chart, the resulted outcome is most of the people who left the company are from the research &development department and most of the employees who are working in the organization are from the research & development department itself.



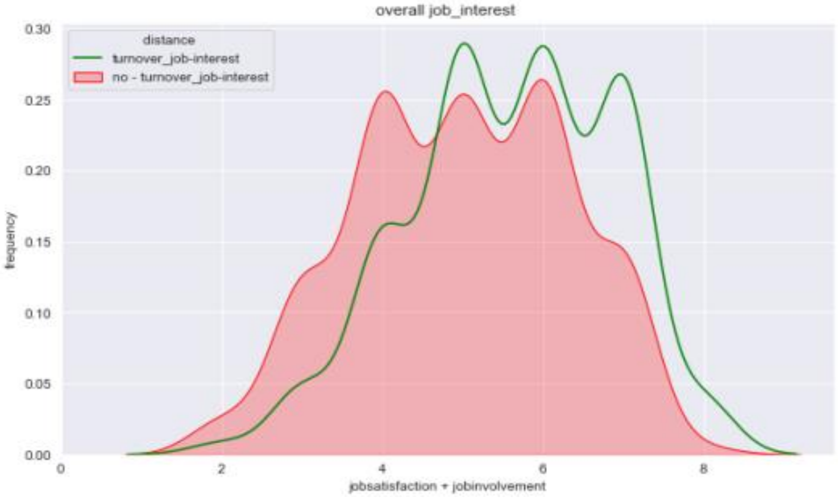
According to the above chart, the resulted outcome is that most of the employees distance between company and the home is less than 10kms.



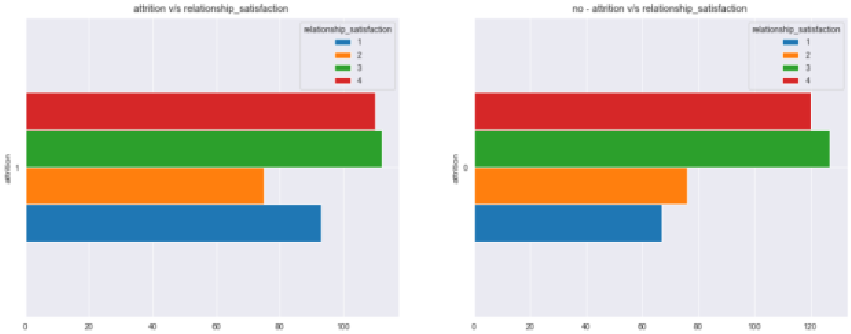
According to the above chart, the resulted outcome is that employees with environment satisfaction with 1 (being lowest) are the one who left the organization.



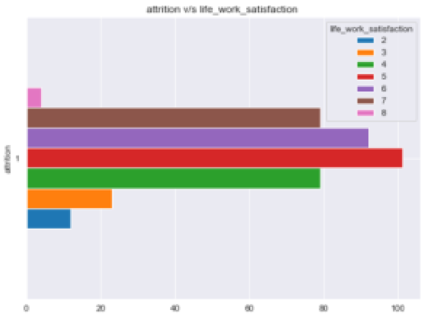
According to the above chart, the resulted outcome is that employees with environment satisfaction with 1 (being lowest) are the one who left the organization.

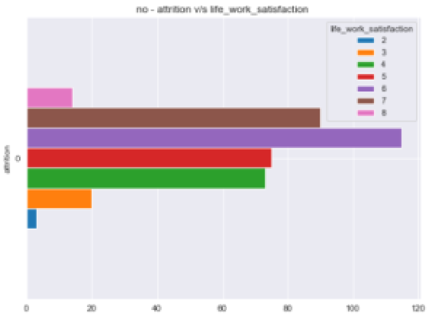


According to the above curves, the resulted outcome is that the employees with high job interest are less likely to leave the organization.

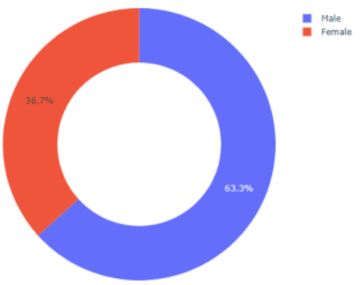


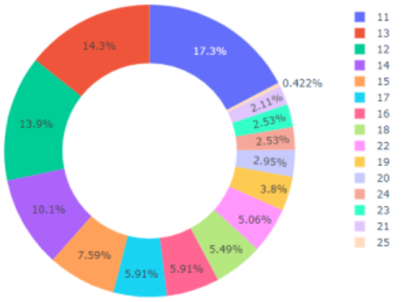
According to above chart, the resulted outcome is that employees with relationship satisfaction value are less likely to leave the organization.





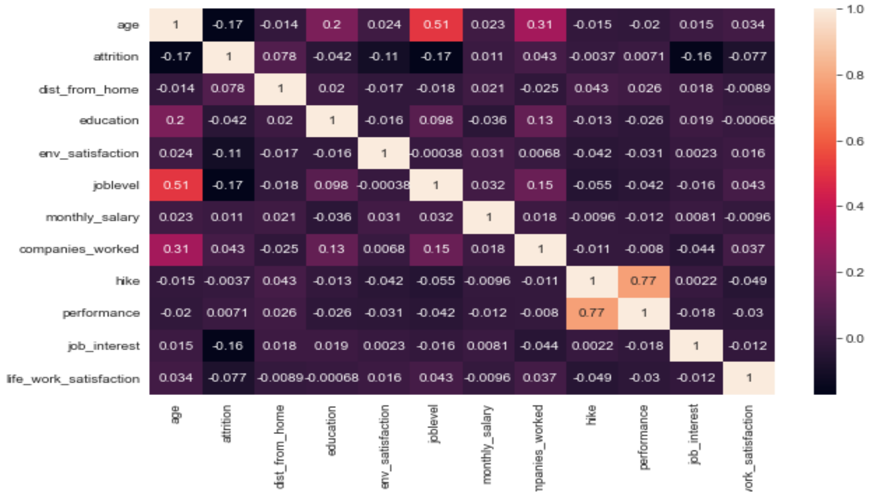
According to the above chart, the resulted outcome is that the employees with low work-life balance satisfaction are likely to leave the organization.



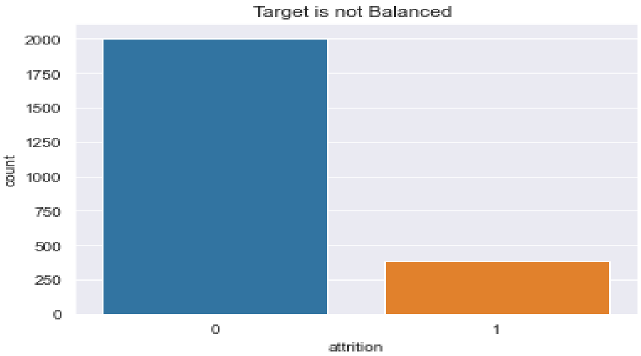


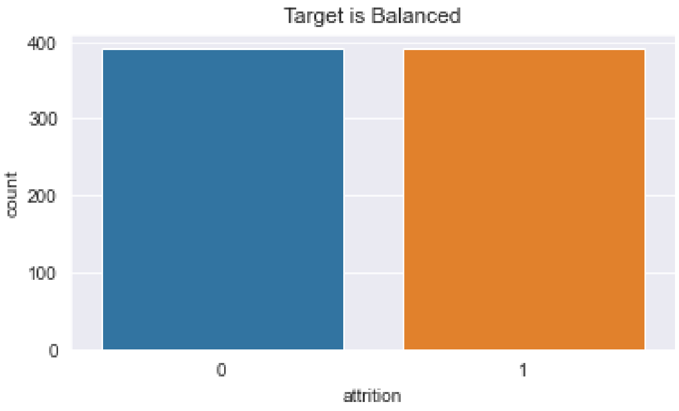
According to the above chart-1, the resulted outcome is that the male attrition level is higher the females who are working in the organization.

According to above chart-2, the resulted outcome is that most of the employees who left the organization are due to low salary as well as low hike provided during the service period.



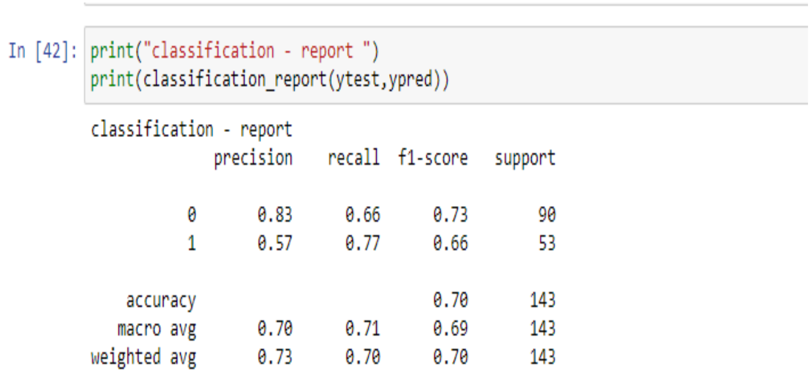
According to above chart, the resulted outcome is that the salaries hike respective to performance is high correlation and job level or position respective to age has moderate correlation.





As we know our data is imbalance, to make it balanced data before providing to our machine learning model we use under sampling technique which helps to decrease the majority class labels up to count of minority class label.

The exploratory data analysis helps us to find the major reasons and number of employees leaving the organization. The dataset which is used in this project is highly imbalanced data. To make it highly balanced data the under sampling technique is used where the data is balanced to the least target. After the data is balanced, the data is used for implementing random forest classification machine learning model which helps us to predict the accuracy of the data. Once the model building executed the predicted accuracy is more than 70% of the above data. Even the accuracy prediction output of the machine learning model is kept on changing for every execution of the code.



```
In [42]: print("classification - report ")
         print(classification_report(ytest,ypred))

classification - report
              precision    recall  f1-score   support

           0       0.83      0.66      0.73        90
           1       0.57      0.77      0.66        53

    accuracy                           0.70       143
   macro avg       0.70      0.71      0.69       143
weighted avg       0.73      0.70      0.70       143
```

VII. CONCLUSION

The employee attrition became huge problem in the current IT Sector and even other sectors also where the organizations are unable to find out the reasons behind the employees leaving the organizations. Through this project the resulted outcome is that the major reasons behind the employees quitting the jobs from the organization are mostly due to lack of recognition in the company and even lack of salary hike similar to the work and effort which they put for the growth of the organization where above factors affect the satisfaction levels of the employee while staying in the organization. Here in this project the data is analyzed by considering satisfaction related attributes and the outcome defines attrition level of the employees from the organization. The implementation of the Random Forest machine learning model helps us to predict the accuracy of the data in the dataset. We can consider this project as a reference for finding the attrition reasons and the organizations can bring up new retention policies and follow new retention strategies to retain the talent in the organization.

VIII. REFERENCES

[1] G. Louppe, “Understanding random forests from theory to practice”, PhD dissertation, University of Liege, 2014.

[2] J. L. E. E. Liu, “main causes of voluntary employee turnover: a study of factors and their relationship with expectations and pretences”, PhD thesis, Univ Chile, 2014.

[3] R. Y. Zou and M. Schonlau, The Random Forest Algorithm for Statistical Learning with Applications in Stata The Random Forest algorithm, pp. 1-20, 2016.

[4] Tanya Attri "Why an Employee Leaves: Predicting using Data Mining Techniques", 2017/2018.

[5] M. P. Debono, "Are organisations doing enough to retain their talent?" The Importance of Employee Retention, 2018.

[6] G. V. Sridhar, "Employee attrition and employee retention-challenges & suggestions employee attrition and employee retention-challenges & suggestions", Rajalakshmi Eng. Coll. Dept. Manag. Stud. January 2018.

[7] HananAlghamdi, "Prediction of Employee Attrition Using Machine Learning and Ensemble Methods " , International Journal of Machine Learning and Computing, Vol. 11, No. 2, March 2021.

[8] Francesco Falluchi, "Predicting Employee Attrition Using Machine Learning Techniques" Department of Innovation & Information Engineering, Guglielmo Marconi University, 00193 Roma, Italy, 3 November 2020.