Employee Attrition Prediction using Various Machine Learning Algorithms

[1]Darshna Dalvadi
[1]Research Scholar
[1]Computer Science,
[1]C. U. Shah University , Wadhawan, India

Abstract: Employee attrition is the percentage of employees who leave a company and are replaced by new employees. A high rate of attrition in an organization leads to increased recruitment, hiring and training costs.
It is extremely demanding and trending research area in today's working culture. Most of the employees leaves the job due to dissatisfaction or low income or company political issue or reason can be enormous. This paper not only predicts the stay of an employee in a company but it also provides the key criteria which lead the employee to leave the job. Predicting Attrition has become essential need of Human Resources (HR) in many companies. Machine learning (ML) developments have made it possible to obtain both improved forecasting performance and clearer explanations of what essential features are associated to employee attrition. In this project, various machine learning approach for employee attrition prediction will be implemented to find out the best solution among them. This paper initially provides comparative analysis of various ML approaches for employee attrition prediction and then gives the best solution for employee attrition prediction and also provides critical features linked to employee attrition.

***IndexTerms* –Attrition, Prediction, Random forest, Logistic regression, ML**

I. INTRODUCTION

1.1. Employee Attrition

Employee is the key asset for the growth of any organization. If any employee leaves the company suddenly then it costs lot to the company. This loss are tangible and intangible both. Organization have to spend money on new recruitment, training and explanation of working culture. If company has lost their most efficient employee then it costs even in profit of organization. So if we can predict employee attrition in advance or in early stage then it can be blessings for a company and Human resource department. There are several benefits of employee attrition prediction which are listed below [1]:

- Assessing employee needs, as well as their strengths and weaknesses
- To reduce the cost of acquiring new talented employees based on employee profiling and organizational requirements.
- It can analyze and measure the loss of expertise and skill sets in terms of employees.
- Calculation of financial and productivity losses as a result of attrition
- The company will be able to plan ahead of time and minimize the damage.
- Has a thorough understanding of labour supply and demand.
- Capable of developing probable strategies based on the prediction model's insight and foresight

Those listed advantages have become the great motivation to develop such project. The main thing about project is that it not only predicts the employee attrition but also do the analysis of key features associated with attrition.

1.2 Machine Learning Algorithms

As per Wikipedia, "Machine learning is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence". In this paper various supervised learning algorithms are used for employee attrition prediction. Following algorithms were applied to predict employee attrition.

A. Logistic Regression

Logistic regression continues to be one of the most widely used methods in data mining in general and binary data classification in particular [2]. It is useful when you wish to perform binary classification. It performs even better if you remove the irrelevant columns from the dataset. Removing irrelevant data from the dataset gives the higher accuracy in the case of logistic regression [4].

B. Random Forest

Random Forest comes in the list of most efficient and powerful algorithm of machine learning. It is a similar to ML algorithm called Bootstrap Aggregation or bagging. Random forest gives more accurate and stable prediction because it builds multiple decision trees and then merges them together.

C. KNN (K Nearest Neighbor)

The KNN algorithm is very simple and very effective [4]. KNN means it finds the similar objects or say instances from the given dataset. K is the number of instances.

D. Naïve Bayes

Naive Bayes is a simple but astonishingly powerful algorithm for predictive modelling. It is very efficient on complex prediction problems. There are several methods of Naïve Bayes like:

- Gaussian Naive Bayes

- Multinomial Naive Bayes
- Complement Naive Bayes
- Bernoulli Naive Bayes
- Categorical Naive Bayes
- Out-of-core naive Bayes

E. Support Vector Machines (SVM)

SVM is the most popular name in the world of Machine learning algorithms. Here hyperplane is a line that distributes the input variable. In SVM, a hyperplane divides the input into mainly two categories here in our project it will be person leaving job and person not leaving job. These two classes are separated using hyperplane.

F. Decision Tree

Decision Trees are an important type of algorithm for predictive modeling machine learning. It is similar to the concept of data structure. Input variable is root and leaf node contains Y or N in case of binary classification.

II Literature Review

Many publications [1, 2] have demonstrated that human resource management (HRM) plays a significant role in current scenarios, personnel selection and training, and determining technological output. Indeed, the findings show that HRM's impact on productivity has a positive impact on a company's capital growth and intensity [3]. Most studies focus on evaluating customers and their activities, but they fail to address a company's most valuable asset, which is reflected in its people [4, 5]. Many studies have been conducted to investigate employee attrition. Existing research [6] discovered that employee demographics and job-related characteristics such as income and length of employment have the greatest influence on employee attrition. Another study [7] investigated the effects of demographics and employee absence on attrition. [8]'s writers focused primarily on job-specific variables. The authors of [9] used a Nave Bayes classifier and the decision tree method J48 to compare the likelihood of an employee quitting the organization. Two strategies were tested for each algorithm: tenfold cross-validation and percentage split 70. J48 results using tenfold cross-validation found an accuracy of 82.4 percent and an incorrect classification of 17.6 percent, whereas percentage split 70 revealed an accuracy of 82.7 percent and an incorrect classification of 17.3 percent. The Nave Bayes classifier produced an accuracy of 78.8 percent and an incorrect classification of 21.2 percent using tenfold cross-validation, whereas percentage split 70 achieved an accuracy of 81 percent and an incorrect classification of 19 percent [11]. Studied the use of Logistic Regression in forecasting employee turnover and discovered that it had an accuracy of 85% and a false negative rate of 14% [10].

III Implementation

3.1 Dataset Description

For implementation of various machine learning algorithms, a dataset was needed. Here in this paper we have used dataset from URL: https://www.kaggle.com/patelprashant/employee-attrition. The name of dataset is WA_Fn-UseC_-HR-Employee-Attrition.csv. This dataset consist of 1470 rows and 35 columns.

No	Attribute	Dtypes	No	Attribute	Dtypes	No	Attribute	Dtypes
1	Age	int64	13	HourlyRate	int64	25	PerformanceRating	int64
2	Attrition	object	14	JobInvolvement	int64	26	RelationshipSatisfaction	int64
3	BusinessTravel	object	15	JobLevel	int64	27	StandardHours	int64
4	DailyRate	int64	16	JobRole	object	28	StockOptionLevel	int64
5	Department	object	17	JobSatisfaction	int64	29	TotalWorkingYears	int64
6	DistanceFromHome	int64	18	MaritalStatus	object	30	TrainingTimesLastYear	int64
7	Education	int64	19	MonthlyIncome	int64	31	WorkLifeBalance	int64
8	EducationField	object	20	MonthlyRate	int64	32	YearsAtCompany	int64
9	EmployeeCount	int64	21	NumCompaniesWorked	int64	33	YearsInCurrentRole	int64
10	EmployeeNumber	int64	22	Over18	object	34	YearsSinceLastPromotion	int64
11	EnvironmentSatisfaction	int64	23	OverTime	object	35	YearsWithCurrManager	int64
12	Gender	object	24	PercentSalaryHike	int64			

[Fig1. Dataset columns and its data type]

3.2 Introduction to Google Colab

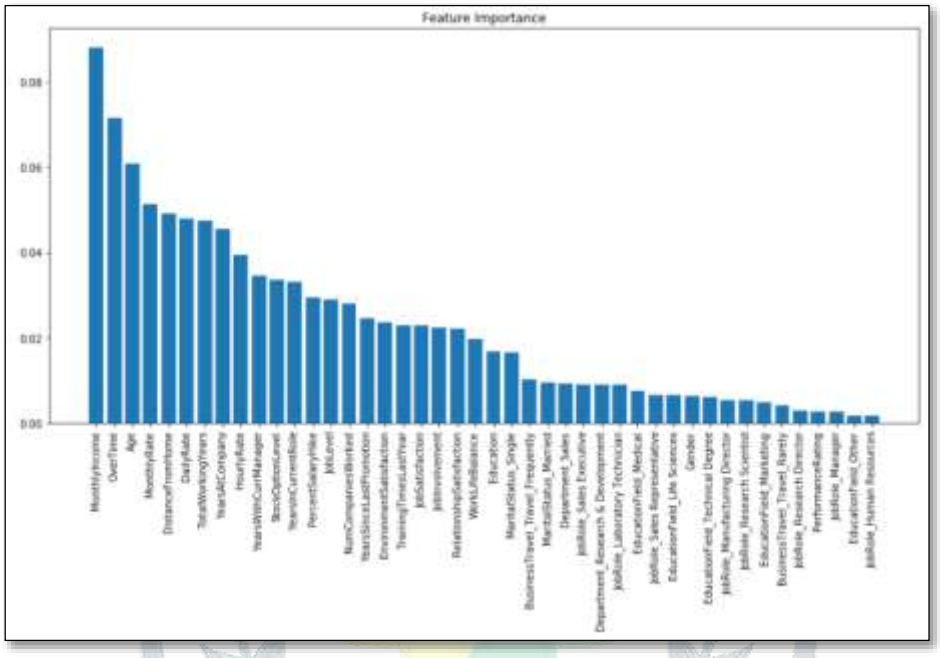
Google colab was utilized to develop multiple machine learning algorithms on the HR dataset. Google is adamant about AI research. Google spent many years developing TensorFlow, an AI framework, and Colaboratory, a development platform. TensorFlow is now open-source, and Google has made Colaboratory available to the public for free since 2017. Google Colab, or just Colab, has replaced Colaboratory. Another appealing feature that Google provides to developers is the utilization of GPU. Colab supports GPU and is completely free. One of the motivations for making it freely available to the public could be to make its software a standard in academics for teaching machine learning and data science. It may also have the long-term goal of establishing a customer base for Google Cloud APIs, which are sold on a per-use basis. Regardless of the causes, the introduction of Colab has made machine learning application learning and development easier [9].

3.3. Implementation Steps

Implementation steps are explained using following diagram. Diagram also explains the process which is going to occur in each step. The Fig.2 shows both: process name and process details.

3.3.1 Identifying feature importance for employee attrition

Feature of any project simply means the attribute. This section deals with the data that actually which features are more correlated for employee attrition which is our target attribute. Target attribute means the attribute which we wish to predict with our code. Here our target is to predict that which employee may leave the organization in near future which we also know as attrition. So to predict the attrition first of all it's important to check with every attribute that how much it is related to predict the target attribute.



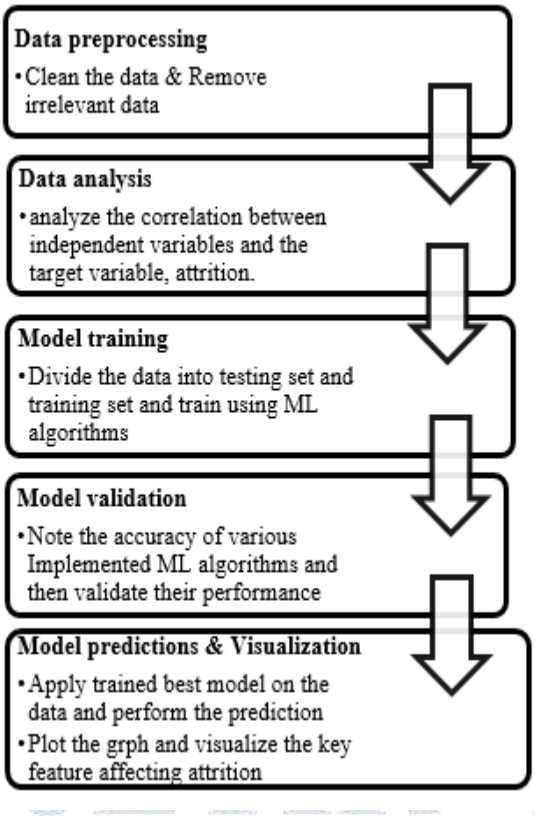
[Fig.3 Various features histogram for employee attrition prediction]

3.3.2 Implementing ML algorithm

In this paper, logistic regression, Decision tree classifier, Support Vector Machine (SVM), K-Nearest neighbor (KNN) and Gaussian NB algorithms were applied to comapare their results and finding out the best suitable algorithm for employee attrition prediction. With the help of sklearn library, various ML algorithms.

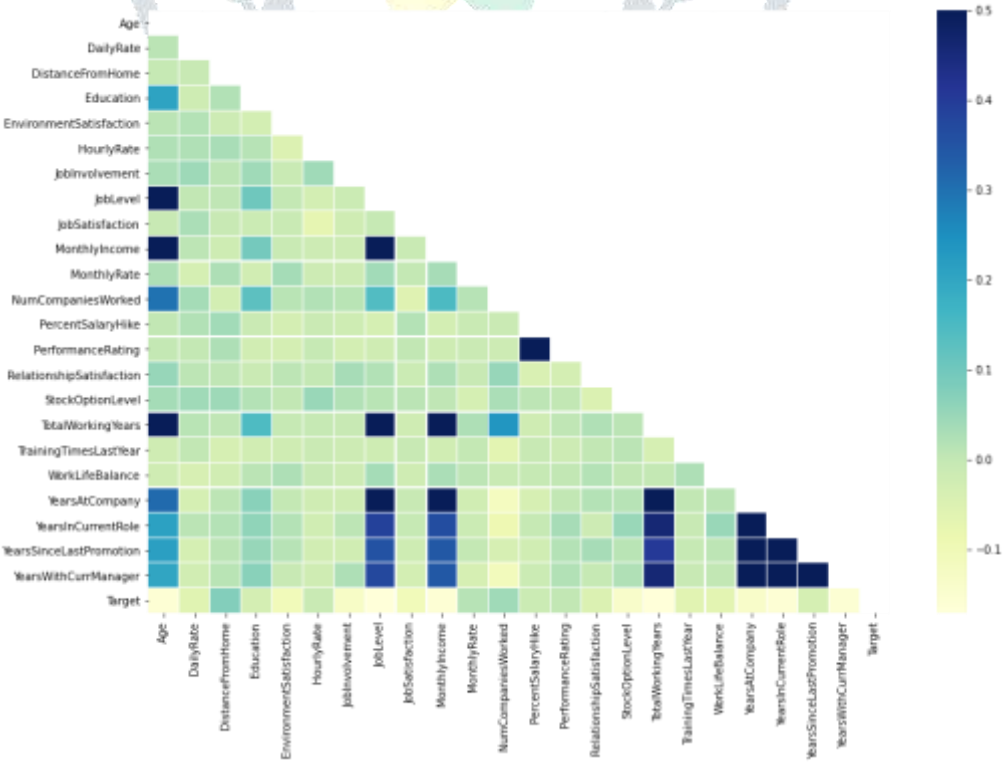
```
# selection of algorithms to consider and set performance measure
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB
models = []
models.append(('Logistic Regression', LogisticRegression(solver='liblinear', random_state=7,
                                                         class_weight='balanced')))
models.append(('Random Forest', RandomForestClassifier(
    n_estimators=100, random_state=7)))
models.append(('SVM', SVC(gamma='auto', random_state=7)))
models.append(('KNN', KNeighborsClassifier()))
models.append(('Decision Tree Classifier',
               DecisionTreeClassifier(random_state=7)))
models.append(('Gaussian NB', GaussianNB()))
```

[Fig 5 implementing various Machine learning algorithms]



[Fig.2 Implementation Steps]

IV Result Analysis



[Fig.4 Correlation of various features with attrition heatmap]

4.1 Feature correlation Analysis

Following a thorough examination of each column's link with the target attribute "Attrition," the following conclusion can be reached.

1. There are no missing or incorrect data values in the dataset, and all characteristics are of the correct data type.
2. The following factors have the most positive relationships with the goal features: Performance Rating, Monthly Rate, Number of Companies Worked, and Distance From Home.
3. The following factors had the largest negative associations with the objective parameters: total working years, job level, years in current role, and monthly income.
4. The dataset is skewed, with the vast majority of observations describing Currently Active Employees.
5. Several features (columns) are redundant for our study, including EmployeeCount, EmployeeNumber, StandardHours, and Over18.

Other observations include:

- When compared to married and divorced employees, single employees have the highest proportion of departures.
- When employees achieve their two-year anniversary with the company, around 10% of them leave.
- When compared to their contemporaries, loyal employees with greater incomes and more responsibilities have a lower rate of leavers.
- When compared to their counterparts, people who reside further away from their workplace have a higher rate of departures.
- Those that travel frequently have a higher proportion of departures than their counterparts.
- People who are required to work overtime have a larger rate of leavers than their counterparts.
- In the submitted dataset, employees who work as Sales Representatives have a substantial number of Leavers.
- Employees who have previously worked at multiple organizations (have "bounced" between workplaces) have a larger rate of leavers than their counterparts.

4.2 Best algorithm Analysis

	Algorithm	ROC AUC Mean	ROC AUC STD	Accuracy Mean	Accuracy STD
0	Logistic Regression	80.05	10.23	75.03	5.12
2	SVM	79.35	8.25	85.11	3.9
1	Random Forest	78.9	7.25	85.03	3.87
5	Gaussian NB	77.51	6.93	83.02	4.8
3	KNN	67.87	7.46	84.48	4.4
4	Decision Tree Classifier	62.44	6.09	78.86	5.25

[Table 1: Various Algorithm analysis for Employee attrition prediction]

V Conclusion

It can be concluded that there is huge requirement of such type of employee attrition software in companies. As it costs too much to hire new employees and to train them and when any company trains the employee and that employee again leaves the company so it becomes very costly for any company to hire and train again and again. So this project predicts the employee attrition based on training data. It is able to predict almost 80 percentage of accuracy. Thus, with this project it can be predicted that which employee is likely to leave the job, so accordingly when assigning manpower to the very crucial project this project helps to make a good team where there will be very less chances of employees to leave in between.

References

1. Marchington, M.; Wilkinson, A.; Donnelly, R.; Kynighou, A. Human Resource Management at Work; Kogan Page Publishers: London, UK, 2016.
2. Van Reenen, J. Human resource management and productivity. In Handbook of Labor Economics; Elsevier: Amsterdam, The Netherlands, 2011.
3. Deepak, K.D.; Guthrie, J.; Wright, P. Human Resource Management and Labor Productivity: Does Industry Matter? Acad. Manag. J. 2005, 48, 135–145.
4. Gordini, N.; Veglio, V. Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. Ind. Mark. Manag. 2016, 62, 100–107. [CrossRef]

5. Keramati, A.; Jafari-Marandi, R.; Aliannejadi, M.; Ahmadian, I.; Mozaffari, M.; Abbasi, U. Improved churn prediction in telecommunication industry using data mining techniques. Appl. Soft Comput. 2014, 24, 994–1012. [CrossRef]
6. Alao, D.; Adeyemo, A. Analyzing employee attrition using decision tree algorithms. Comput. Inf. Syst. Dev. Inf. Allied Res. J. 2013, 4, 17–28.
7. Nagadevara, V. Early Prediction of Employee Attrition in Software Companies-Application of Data Mining Techniques. Res. Pract. Hum. Resour. Manag. 2008, 16, 2020–2032.
8. Rombaut, E.; Guerry, M.A. Predicting voluntary turnover through Human Resources database analysis. Manag. Res. Rev. 2018, 41, 96–112. [CrossRef]
9. https://www.tutorialspoint.com/google_colab/index.htm
10. Dulari Bosamiya Mr. Akash K. Mehta, A Survey of the Farm Surveillance System for Animal Detection in Image Processing, International Journal Of Engineering Development And Research, 2014
11. Patel Milan, Dulari Bosamiya , Review Of Different Techniques For Ripe Fruit Detection, International Journal Of Engineering Development And Research, 2016
12. RACHANA PATEL, REEVA SONI, DULARI BHATT, Tumor Detection using Normalized Cross Co-Relation, IJRMEET, 2013
13. Bhatt, Dulari, Chirag Patel, and Priyanka Sharma. "Intelligent Farm Surveillance System for Bird Detection." (2011) in GRDET.
14. J. D. Fuletra and D. Bosamiya, “A survey on drivers drowsiness detection techniques,” International Journal on Recent and Innovation Trends in Computing and Communication, vol. 1, no. 11, pp. 816–819, 2013