Employee Attrition Prediction using Machine Learning

Parv N Gandhi
Sinhgad Academy of Engineering
Savitribai Phule Pune University
Pune, India

Aishwarya Jangale
Sinhgad Academy of Engineering
Savitribai Phule Pune University
Pune, India

Ganesh S Mhaske
Sinhgad Academy of Engineering
Savitribai Phule Pune University
Pune, India

Abhilasha Kadlag
Sinhgad Academy of Engineering,
Savitribai Phule Pune University,
Pune, India.

***Abstract*— Every organization has its own productivity and strength which stands of the legs of the employees. Keeping regular employee is a great challenge for all organization in the competitive world. Employee Attrition is one of the biggest business problems in HR Analytics. Companies invest a lot in the training of the employees keeping in mind the returns they would provide to the company in the future. If an employee leaves the company, it is the loss of opportunity cost to the company. These study interpreters the employee's attrition rate through the related attributes like Job Role, overtime, job level affect the attrition largely. The paper contain the survey of various classification algorithms like logistic regression, LDA, SVM, KNN, Random Forests to predict the probability of attrition of any new employee.As a result, training balanced dataset with Random Forest achieved the second highest performance, with 0.269 F1-score but has achieved the highest accurancy with the algoritms we have used.**

Keywords*— Employee attrition; Support vector machine; random forest; K nearest neighbours; Feature selection;Attrition Rate; HR; Classifier; Preprocessing; Employment Features.

I. Introduction

The outcome of many research shows that the most valuable asset and important resource in organizations are their employees. Now a day due to increased competition and improved requirement in employees' proficiency determines the attrition rate. The employee attrition is considered to be a serious issue for organizations. The cost of searching and training employees is very high.

Organizations need to search, hire and train new employees. Loss of experienced workers especially high performers is difficult to manage and is negatively related to the success and performance of organizations. The study focuses on the variables that may lead to control the attrition rate of the employee.

The problem of employee turnover has turn to eminence in organizations because of its pessimistic impacts on issues on work place self-esteem and efficiency. The organizations deal with this problem is by predicting the risk of attrition of employees using machine learning techniques thus giving organizations to take proactive action for retention.

II. Litrature Review

In this paper, modified approaches using various data mining techniques are collected to analyze the employee attrition rate at various levels. The study related to data mining for extracting the employee's attrition rate used in various models and the comprehensive literature review of various researcher's works are stated below;

Qasem A, A.Radaideh and Eman A Nagi, has applied data mining techniques to build a classification model to predict the performance of employees [3]. They adopted CRISP-DM data mining methodology [4] in their work. The Decision tree was the main data mining tool used to build the classification model, where several classification rules were generated. They validated the generated model; several experiments were conducted using real data collected from several companies. The model is intended to be used for predicting new applicants' performance.

Amir Mohammad EsmaieeliSikaroudi, [5] RouzbehGhousi and Ali EsmaieeliSikaroudi et al, implemented knowledge discovery steps on real data of a manufacturing plant. They chew over many characteristics of employees such as age, technical skills and work experience. They used to find out importance of data features is measured by Pearson Chi-Square test.

John M. Kirimi and Christopher Moturi et al, [6] proposed a prediction model for employee performance forecasting that enables the human resource professionals to refocus on human capability criteria and thereby enhance the performance appraisal process of its human capital.

RohitPunnoose and PankajAjit et al, explored [7] the application of Extreme Gradient Boosting (XGBoost) technique which is more robust because of its regularization formulation. [8] Data from the HRIS of a global retailer is used to compare XGBoost against six historically used supervised classifiers and demonstrate its significantly higher accuracy for predicting employee turnover.

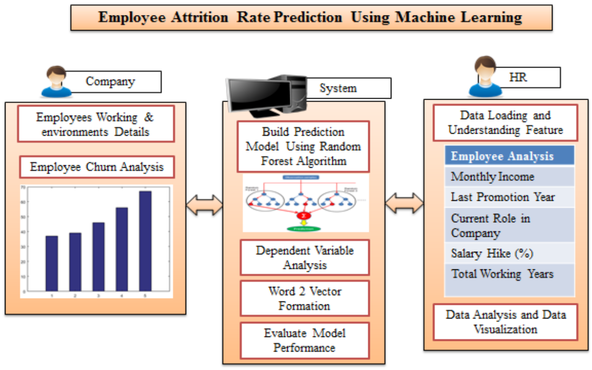
Research Authors	Problem studied	Techniques Studied	Recommend
Jantan, Hamdan and Othman[9]	Data Mining techniques for performance prediction of employees	C4.5 decision tree, R.andom Forest, Multilayer Perceptron(MLP) and Radial Basic Function Network	C4.5 decision tree
Nagadevara, Srinivasan and Valk[10]	Relationship of withdrawal behaviors like lateness and absenteeism, job content, tenure and demographics on employee turnover	Artificial neural networks, logistic regression, classification and regression trees (CART), classification trees (C5.0), and discriminant analysis)	Classification and regression trees (CART)
Hong, Wei and Chen[11]	Feasibility of applying the *Logit*and *Probit*models to employee voluntary Turnover predictions	Logistic regression model (logit), probability regression model (probit)	Logistic regression model (logic)
Marjorie Laura Kane Sellers[12]	To explore various personal, as well as work variables impacting employee voluntary turnover	Binomial logit regression	Binomial logit regression
Alao and Adeyemo[13]	Analyzing employee attrition using multiple decision tree algorithms	C4.5, C5, REPTree, CART	C5 decision tree
Saradhi and Palshikar[14]	To compare data mining techniques for predicting employee churn	Naïve Bayes, Support Vector Machines, Logistic Regression, Decision Trees and Random Forests	Support Vector Machines

III. Proposed System

Initially the data is downloaded from Kaggle is pre-processed first so that we can extract important features like Monthly Income, Last Promotion Year, Salary Hike and etc. that are quite natural for employee attrition. Dependent variables or Predicted variable are the one that helps to get the factors that mostly dependent on employee related variables. For example the employee ID or employee count has nothing to do with the attrition rate.

Exploratory Data Analysis is an initial process of analysis, in which you can 0summarize characteristics of data to can predict who, and when an employee will terminate the service. The system builds a prediction model by using random forest technique. It is one of the ensembles learning technique which consists of several decision trees rather than a single decision tree for classification.

The techniques perform dependent variable analysis and word formation vector to evaluate the employee churn. Hence, by improving employee assurance and providing a desirable working environment, we can certainly reduce this problem significantly.


Employee Attrition Rate Prediction Using Machine Learning
Company
Employees Working & environments Details
Employee Churn Analysis
70
60
50
40
30
20
10
0
1
2
3
4
5
System
Build Prediction Model Using Random Forest Algorithm
Dependent Variable Analysis
Word 2 Vector Formation
Evaluate Model Performance
HR
Data Loading and Understanding Feature
Employee Analysis
Monthly Income
Last Promotion Year
Current Role in Company
Salary Hike (%)
Total Working Years
Data Analysis and Data Visualization

System Architecture

IV. DATASET

We have taken this dataset from the IBM HR Analytics in order to help to find a solution to this problem. This dataset contain all in total 35 attributes out of **Attrition** is the dependent attribute. We have come to a conclusion that with the help of this dataset we might be able to find a solution this problem. These are the features that are present in our dataset.

- *Age- Numeric Discrete*
- *Attrition-Categorical*
- *Business Travel- Categorical*
- *Daily Rate- Numeric Discrete*
- *Department- Categorical*
- *DistanceFromHome- Numeric*
- *Education- Categorical*
- *Education Field- Categorical*
- *Employee Count- Numeric Discrete*
- *Employee Number- Numeric Discrete*
- *Environment Satisfaction- Categorical*
- *Gender- Categorical*
- *Hourly Rate- Numeric Discrete*
- *Job Involvement- Categorical*
- *Job Level- Categorical*
- *Job Role- Categorical*
- *Job Satisfaction- Categorical*
- *Marital Status- Categorical*
- *Monthly Income- Numeric Discrete*
- *Monthly Rate- Numeric Discrete*
- *NumCompaniesWorked- Numeric Discrete*
- *Over18- Categorical*
- *OverTime- Categorical*
- *PercentSalaryHike- Numeric Discrete*
- *PerformanceRating- Categorical*
- *RelationshipSatisfaction- Categorical*
- *StandardHours- Numeric Discrete*
- *StockOptionLevel- Categorical*
- *TotalWorkingYears- Numeric Discrete*
- *TrainingTimesLastYear- Numeric Discrete*
- *WorkLifeBalance- Categorica*
- *YearsAtCompany- Numeric Discrete*
- *YearsInCurrentRole- Numeric Discrete*
- *YearsSinceLastPromotion- Numeric Discrete*
- *YearsWithCurrManager- Numeric Discrete*

IV. Feature Selection

Feature Selection is considered as the most crucial theory in the fields of machine learning which has a significant amount of impact on the actual performance of the model your building. These features can be simply used to coach your model and have an enormous influence on the performance.

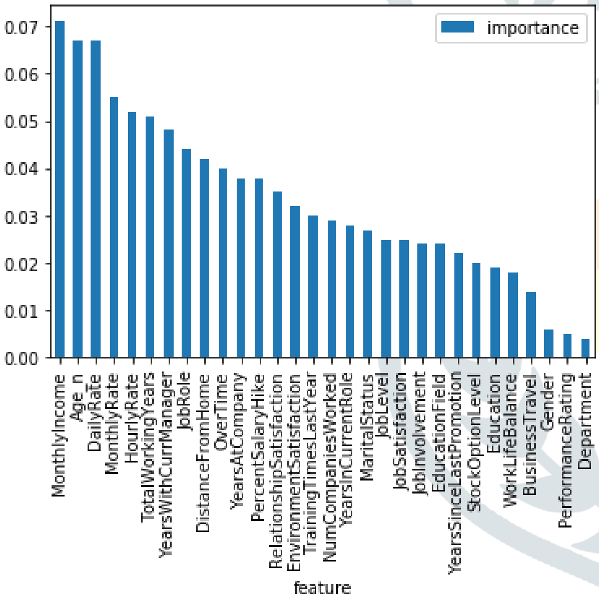
Trivial and unrelated features can have a negative impact on the performance of the model. Feature selection and Data cleaning should be the first and most significant step of your model designing.

Feature Selection is the process where you automatically or manually select those features on the basis of some various techniques like Univariate Selection Feature Importance Correlation Matrix which contribute most to your dependent variable or output variable in which you are interested in.

After analysing the dataset manually we came to a conclusion that these features ***Employee Count, Employee Number, Over18*** have no direct impact on our output variable ***Arttrition.*** Therefore, these features have been completely neglected before applying any feature selection methods**.**

Feature Importance

Feature importance gives you a score for each feature of your data, the higher the score more important or relevant is the feature towards your output variable.
Feature importance is an inbuilt class that comes with Tree Based Classifiers, we will be using Extra Tree Classifier for extracting the top features for the dataset.



Feature Importance

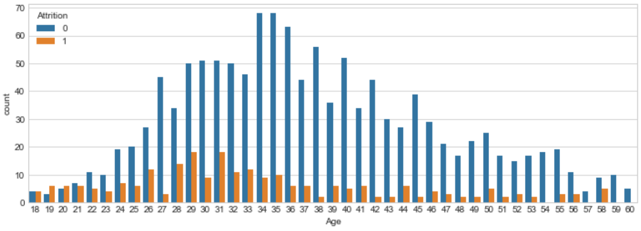
The diagram above represents the feature importance of each feature of our dataset with the help of this feature importance method we could analyse that the features like ***Monthly income, Age , Daily rate*** , Hourly rate etc are some of the significant attributes . Along with that we came to conclusion that the features ***like Business travel Gender, Department, Performance rating*** are having least impact on our output variable ***Atrrition.*** Therefore we can neglect these features beforehand.
After applying feature selection methods these following attributes are selected for model designing

- Monthlyncome
- Age
- Monthlyrate
- Hourlyrate
- Totalworkingyears
- Yearswithcurrmanager
- Jobrole
- Distancefromhome
- Overtime
- Yearsatcompany
- Percentsalaryhike

- Realationshipsatifaction
- Environmentsatisfaction
- Trainintimelastyear
- Nocompaniesworker
- Yearcurrentrole
- Yearsincurrentrole
- Martialstatus
- Joblevel
- Jobsatisfaction
- Jobinvolvement
- Educationfield
- Yearssincelastpromotion

V. Exploratory Data Analysis

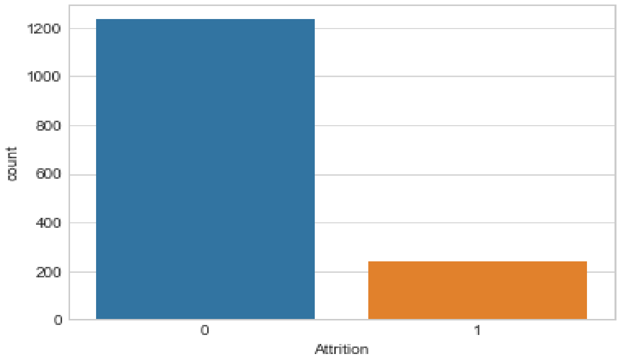
In this part we are going to analyse the relationship between the attributes and the output variable. As there are many attributes available to us we cannot show each and every attributes relation. So for simplicity we are going to use the Age attribute as an example and show the relation.



As shown in the figure above we can clearly see that at the age of 29-31, the rate of attrition is the highest among all ages. At the age of 50 and above the chances of attrition are at the lowest.

V. Imbalanced Dataset

In the dataset 90% records are labelled with class **YES** and remaining 10% records are labelled with class **NO**. This type of datasets are called as imbalanced datasets and can have adverse effect on the performance of the model it makes the model biased towards majority class of output variable. Therefore handling imbalanced dataset becomes a necessary task for this type of problem statement.

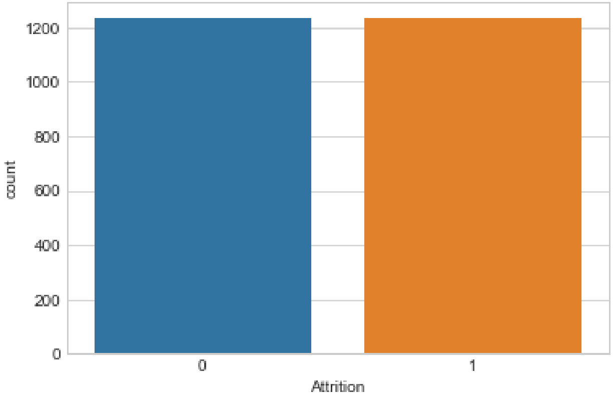


Imbalanced Data

Following are some of the methods to handle imbalanceness of dataset

- Random Under Sampling
- Random Over Sampling
- Custer Based Over Sampling

For our dataset we are using over sampling method to handel the imbalancness of the dataset. Before over ssmpling ***1233 records*** were labelled with class **NO** and ***only 237 records*** were labelled with class **YES.** After performing over sampling we similarzed the number of records of both classes to **1233 records** as shown in the diagram below.



Data Distribution after Over sampling

IV. Compartive Analysis

Employee attrition prediction problem, statement comes under the classification type of machine learning. To solve such a classification problem there are multiple choices available. Comparative analysis is a study of choosing the best algorithm for the problem statement in this section we illustrate the results of models based on their accuracy, precision, recall, and F1 score.

Here we are comparing between multiple classification algorithms like Random forest(RFA) , Decision tree ,K-Nearest neighbour(KNN) ,Logistic regression(LR), Stochastic Gradient Descent (SGD) over evaluation metrics like Accuracy, Precision, Recall, and F1 score.

A: Performance Evaluation

All trained models were evaluated by measuring their accuracy, precision, recall and F1 score which are described below :

- $\text{Accuracy} = \frac{\text{TP+TN}}{TP+TN+FP+FN}$
- $\text{Precision} = \frac{\text{TP}}{TP+FP}$
- $\text{Recall} = \frac{\text{TP}}{TP+FN}$
- $\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$

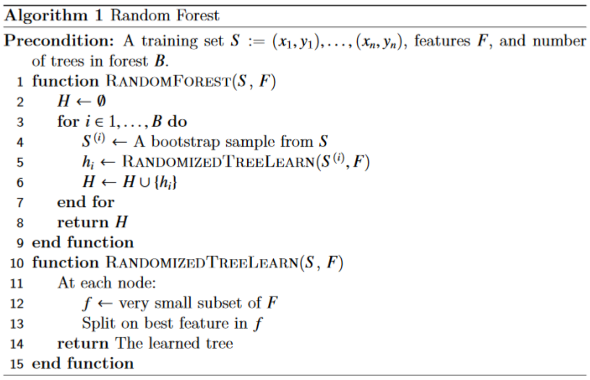
Random Forest (RFA)

Random Forest algorithm is a supervised classification algorithm. The decision tree is a decision support tool.

Random forest (RF) is one of the most powerful supervised machine learning algorithms for generating classifications and regressions. RF uses multiple decision trees to train data [9]. Each tree votes for a classification label for a certain dataset, then the RF model chooses which class had the most votes from the decision trees [10].

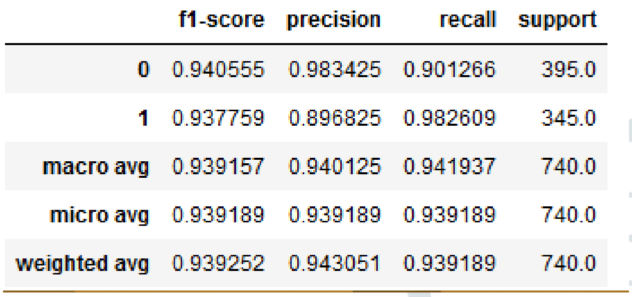
It uses a tree-like graph to show the possible consequences. If you input a training dataset with targets and features into the decision tree, it will formulate some set of rules. These
Rules can be used to perform predictions.

There are two stages in Random Forest algorithm, one is random forest creation, and the other is to make a prediction from the random forest classifier created in the first stage.



```
Algorithm 1 Random Forest
Precondition: A training set $S := (x_1, y_1), \dots, (x_n, y_n)$, features $F$, and number
    of trees in forest $B$.
function RANDOMFOREST($S$, $F$)
    $H \leftarrow \emptyset$
    for $i \in 1, \dots, B$ do
        $S^{(i)} \leftarrow$ A bootstrap sample from $S$
        $h_i \leftarrow$ RANDOMIZEDTREELEARN($S^{(i)}$, $F$)
        $H \leftarrow H \cup \{h_i\}$
    end for
    return $H$
end function
function RANDOMIZEDTREELEARN($S$, $F$)
    At each node:
        $f \leftarrow$ very small subset of $F$
        Split on best feature in $f$
    return The learned tree
end function
```

Performance of RFA Model

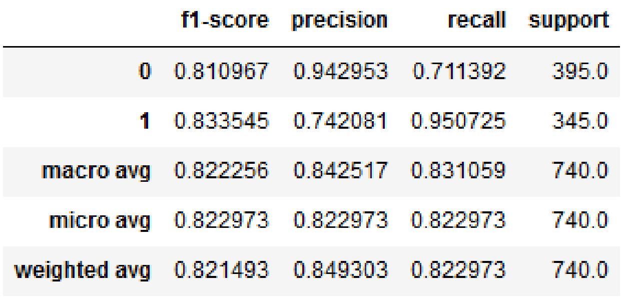


	f1-score	precision	recall	support
0	0.940555	0.983425	0.901266	395.0
1	0.937759	0.896825	0.982609	345.0
macro avg	0.939157	0.940125	0.941937	740.0
micro avg	0.939189	0.939189	0.939189	740.0
weighted avg	0.939252	0.943051	0.939189	740.0

K-Nearest Neighbour

K-nearest neighbours (KNN) is one of the simplest machine learning algorithms and is used for both classification and regression. KNN works by specifying the value of K, which indicates the number of closest training points for a single data point. Each new data point will be classified based on the majority of votes collected from its neighbours

Performance of KNN Model



	f1-score	precision	recall	support
0	0.810967	0.942953	0.711392	395.0
1	0.833545	0.742081	0.950725	345.0
macro avg	0.822256	0.842517	0.831059	740.0
micro avg	0.822973	0.822973	0.822973	740.0
weighted avg	0.821493	0.849303	0.822973	740.0

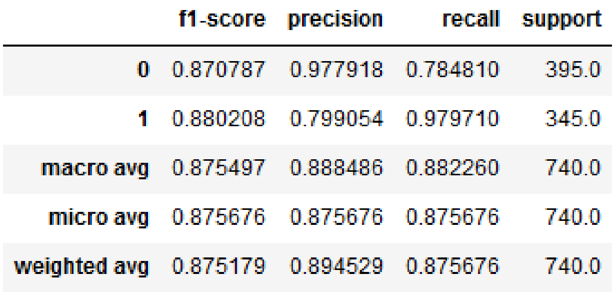
Decision Tree

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).

In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record’s attribute. On the ***basis of*** comparison, we follow the branch corresponding to that value and jump to the next node

Performance of Decision Tree Model

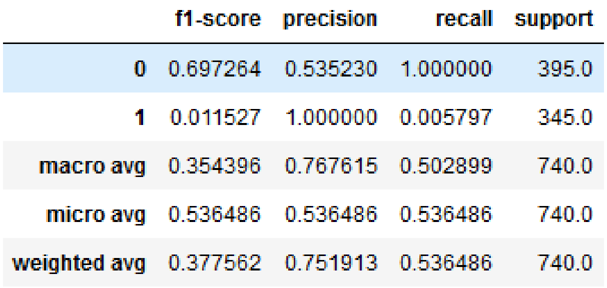


	f1-score	precision	recall	support
0	0.870787	0.977918	0.784810	395.0
1	0.880208	0.799054	0.979710	345.0
macro avg	0.875497	0.888486	0.882260	740.0
micro avg	0.875676	0.875676	0.875676	740.0
weighted avg	0.875179	0.894529	0.875676	740.0

Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent (SGD) is a simple yet efficient optimization algorithm used to find the values of parameters/coefficients of functions that minimize a cost function. In other words, it is used for discriminative learning of linear classifiers under convex loss functions such as SVM and Logistic regression. It has been successfully applied to large-scale datasets because the update to the coefficients is performed for each training instance, rather than at the end of instances.
Stochastic Gradient Descent (SGD) classifier basically implements a plain SGD learning routine supporting various loss functions and penalties for classification. Scikit-learn provides SGDClassifier module to implement SGD classification.

Performance of SGD Model



	f1-score	precision	recall	support
0	0.697264	0.535230	1.000000	395.0
1	0.011527	1.000000	0.005797	345.0
macro avg	0.354396	0.767615	0.502899	740.0
micro avg	0.536486	0.536486	0.536486	740.0
weighted avg	0.377562	0.751913	0.536486	740.0

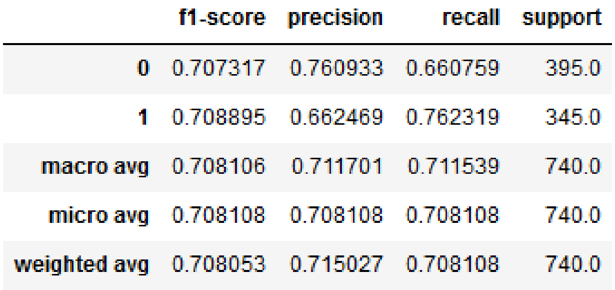
Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

Performance of LR Model



	f1-score	precision	recall	support
0	0.707317	0.760933	0.660759	395.0
1	0.708895	0.662469	0.762319	345.0
macro avg	0.708106	0.711701	0.711539	740.0
micro avg	0.708108	0.708108	0.708108	740.0
weighted avg	0.708053	0.715027	0.708108	740.0

By comparative analysis of above 5 algorithms we came to a firm conclusion that that Random Forest Algorithm has the best accuracy. Therefore RFA is used for model building for our system.

VI. CONCLUSION

Human Resource is the main pillar for any organization. The growth level as well as market penetration are duly depends on the strength of the employees. Now a day due to increased population and people with high competency makes great success for any firm. But the prime issues which are normally addressed in any organization are only the attrition. This is a great challenge as well as retention is also the prime task. In this paper we have studied about different techniques and methods used by the various researchers for employee prediction strategy

VI. REFERENCES

[1] S. Jahan, "Human Resources Information System (HRIS): A Theoretical Perspective", Journal of Human Resource and Sustainability Studies, Vol.2 No.2, Article ID:46129, 2014.

[2] C. Cortes and V. Vapnik, Support-vector networks. Machine learning, 20(3), 273-297, 1995

[3] RenukaAgrawal, Jyoti Singh, and Zadgoankar .S, "Formative Assessment For Performance Evaluation Of Faculty Using Data Mining", *International Journal Of Advances In Electronics And Computer Science*, ISSN: 2393-2835.

[4] HosseinAlizadeh, Buinzahra Branch and Islamic, 2016 ,"Introducing A Hybrid Data Mining Model ToEvaluate Customer Loyalty", *Engineering, Technology & Applied Science Research Volume. 6,No.6,1235-1240.*

[5] Amir Mohammad EsmaieeliSikaroudi ,Rouzbehghousi and Ali Esmaieelisikaroudi, 2015 "A Data Mining Approach To Employee Turnover Prediction" (Case Study: Arak Automotive Parts Manufacturing), *Journal Of Industrial And Systems Engineering* Volume. 8, No. 4.

[6] Anjali A. Dudhe and SachinSakhare .R, January 2018, "Teacher Ranking System To Rank Of Teacher As Per Specific Domain" ,*Journal On Soft Computing ICTACT,* Volume: 08, Issue: 02, Issn: 2229-6956.

[7] Rohit Punnoose and Pankaj Ajit, 2016 "Prediction Of Employee Turnover In Organizations Using Machine Learning Algorithms", *International Journal Of Advanced Research In Artificial Intelligence*(IJARAI) Volume. 5, No. 9.

[8] Dilip Singh Sisodia, SomduttaVishwakarma, AbinashPujahari" Evaluation of Machine Learning Models for Employee Churn Prediction", Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017) IEEE [13] Xplore Compliant - Part Number: CFP17L34-ART, ISBN: 978-1-5386-4031-9.

[9] H. Jantan, A. R. Hamdan, and Z. A. Othman, "Towards Applying Data Mining Techniques for Talent Managements", 2009 International Conference on Computer Engineering and Applications, IPCSIT vol.2, Singapore, IACSIT Press, 2011

[10] V. Nagadevara, V. Srinivasan, and R. Valk, "Establishing a link between employee turnover and withdrawal behaviors: Application of data mining techniques", Research and Practice in Human Resource Management, 16(2), 81-97, 2008.

[11] W. C. Hong, S. Y. Wei, and Y. F. Chen, "A comparative test of two employee turnover prediction models", International Journal of Management, 24(4), 808, 2007.

[12] L. K. Marjorie, "Predictive Models of Employee Voluntary Turnover in a North American Professional Sales Force using Data-Mining Analysis", Texas, A&M University College of Education, 2007.

[13] D. Alao and A. B. Adeyemo, "Analyzing employee attrition using decision tree algorithms", Computing, Information Systems, Development Informatics and Allied Research Journal, 4, 2013.

[14] V. V. Saradhi and G. K. Palshikar, "Employee churn prediction", Expert Systems with Applications, 38(3), 1999-2006, 2011.