

# PREDICTION OF EMPLOYEE ATTRITION USING MACHINE LEARNING

M.Adilakshmi <sup>1</sup>, N.Bhuvaneswari <sup>2</sup>, E.Madhavi <sup>3</sup> and Dr. M.Sireesha <sup>4</sup>

*1,2,3 student of Department of Computer Science and Engineering, Narasaraopet Engineering College*

*4 Faculty of Department of Computer Science and Engineering, Narasaraopet Engineering College, Narasaraopet*

*<sup>1</sup>marellaadilakshmi2001@gmail.com, <sup>2</sup>neelambhuvi@gmail.com, <sup>3</sup>madhaviereti@gmail.com, <sup>4</sup>sireeshamoturi@gmail.com*

**Abstract-** In today's IT world, the major concern is employee attrition rate. Attrition rate can be defined as the percentage of employees who left from the organization. The aim of this project is to analyse a particular employee will continue in the organization or not. The discontinuous of an employee can be done by either up to the individual or due to organization force. To predict attrition rate we have used different machine learning techniques. The steps are dataset collection, pre-processing the data, training model using machine learning classification models like Random Forest, decision tree classifier etc and result analysis . The results are evaluated using accuracy score and confusion matrix. Random forest algorithm giving the best accuracy i.e 85% compared to decision tree. This work will help organizations to better understand the attrition causes.

**Keywords-** Attrition, classification models, random forest, SVM, decision tree classifier

## I. INTRODUCTION

In these days, data produced at an exponential pace. This data has been useful in gaining knowledge and spreading awareness about any company or group. Before modelling data we have to pre-process the data with the goal of gaining insightful conclusions, recovering pertinent data to make wise decisions. It is a way of making a computer to make correct predictions using historical information.

Employees are playing major role for any company, so losing effective employees could have a negative impact on the business in a number of ways. Employee attrition has a number of negative effects, including increased costs for hiring and training new workers[1]. This will effect the well being of existing employees in the organization. This paper consists of 3 sections. Dataset collection is the first step and it is discussed in next step. section II discuss the data pre-processing steps. This step is crucial for any machine learning project before building model[9]. Dataset consists of inconsistent data , imbalanced class labels and unwanted attributes[8]. All these problems lead to poor model construct. We are supposed to find important attributes which impacts target attribute. For doing this step we do feature importance on all attributes.

Third section will discuss on model training here we pass more consistent data to different classification models.

## II. LITERATURE SURVEY

A lot of studies have been made on attrition prediction analysis in the literature. The major focus was on predicting employee attrition. Researchers have applied machine learning classification models like logistic regression, random forests, support vector machine, and others to analyze the attributes that impact the attrition rate. For instance, Srivastava[1] et al presented a framework that predicts employee churn by analyzing the behaviors of employees and attributes with the help of machine learning techniques. Setiawan[5] et al through their work found variables that have a major impact on employee attrition. Qasem A, A.Radaideh, and Eman A Nagi have utilized data mining techniques to construct a classification model that can anticipate employees' performance. They implemented the CRISP-DM data mining methodology in their research and employed the decision tree as the primary data mining tool to build the classification model. Multiple classification rules were created as a result of this. The generated model was validated through a series of experiments using actual data obtained from various businesses. The purpose of the model is to forecast the performance of new job applicants.

## III. DATASET COLLECTION

The "IBM HR Employee Analytics Attrition and Performance" dataset was acquired from Kaggle, a website that provides datasets and serves as a venue for data science-related contests [13]. There are 35 attributes and 1470 entries in this collection. The data categories include independent factors like "Age," "Daily Rate," "Education Field," "Number of companies worked," etc.; however, in this study, "Attrition" is regarded as the dependent variable. Two class names, "Yes" or "No," make up the "Attrition" data field.

#### IV. DATA PRE-PROCESSING

##### A).LIBRARIES USED:

1) *Libraries for Import:* We take into consideration the following potent and useful tools for the analysis and prediction of attrition rate. The libraries are:

a) *Numpy:* It ranks among the most significant Python tools for computational mathematics and science.

b) *Pandas:* A tool made for quick and simple data frame processing.

c) *Matplotlib:* A Python package that produces complex graphs and charts like bar charts, pie charts, and more.

d) *Scikit-Learn:* The SciKit-Learn package provides a variety of supervised and unsupervised machine learning methods. The main goal of machine learning tools is data modeling.

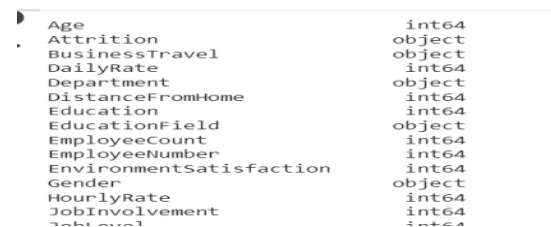
2) *Read Dataset:* Read the dataset of .csv format using pandas function read\_csv().

3) *Create dataset as Data Frame:* Now create data frame using read dataset object. This data frame will be used in further pre-processing steps.

##### B) DATA PREPROCESSING

Pre-processing means cleaning data, normalizing datasets and operate the changes in the data. These steps are performed to get the datasets into a state that enables analysis in further phases [1]. In Data preprocessing the following steps were performed:

1) *Investigate Dataset Properties:* The goal of data research was to comprehend the connections between the factors and to examine the issue at hand [1]. This research step is useful for spotting common dataset problems like Null values, Outliers, Redundancies, etc. Below figure 1 picture depicts the columns of dataset and its datatype.



Age	int64
Attrition	object
BusinessTravel	object
DailyRate	int64
Department	object
DistanceFromHome	int64
Education	int64
EducationField	object
EmployeeCount	int64
EmployeeNumber	int64
EnvironmentSatisfaction	int64
Gender	object
HourlyRate	int64
JobInvolvement	int64
JobLevel	int64
JobRole	object

FIG: 1 ATTRIBUTES AND DATA TYPES

2) *Data preparation:* This includes the procedures for exploring, pre-processing, and configuring data before data modeling. It was carried out in order to familiarize ourselves with our information and learn

more about it. It required converting the data into a structure that would make further research easier. This is usually the stage in the analytics lifecycle that requires the most effort and iterations [1]. The following are the main processes taken for data preparation:

a) *Feature Reduction:* This phase was important in deciding which features in the dataset should be kept and which features should be transformed or removed in order to make decisions about which attributes in the data will be helpful for analysis in the later stages. The decision as to which trait is important and which is not for attrition forecast was made. Following are some examples of characteristics based on which elements were excluded from further analysis:

i) *Attributes with numbers that are not unique:* There are non-unique numbers for the following attributes:

The number of the property "Employee count," which is "1" for each employee, is given by the employee count attribute.

"Standard working hours" is an attribute that provides the number of an employee's standard working hours, which is "80" for each entry. "Over 18 yrs of age" is an attribute that confirms whether an employee meets the age requirement (to be over 18), which is "Yes" for every entry.

All the above mentioned attributes having only one unique value .So, we are ignoring these attributes from dataset.

ii) *Data cleaning:* To guarantee better data quality, abnormalities, usually missing values, duplicate data, and outliers are removed. In our dataset there are no missing values and outliers.

iii) *Categorical to Numerical:* Since categorical variables are not accepted as input by normal libraries, these values must be transformed into numeric form. As seen in Figure 2, this was accomplished by using the Label Encoder Method to convert nominal category variables or categorical data into numerical labels. The range of a numerical identifier is always 0 to n\_classes-1.

Attrition	Attrition
Yes	1
No	0
Yes	1

FIG: 2 ATTRITION ATTRIBUTE LABEL ENCODING

iv) *Dataset balancing*: In given dataset, there are more records with the label "Attrition" set to "0" than there are records with the label "Attrition" set to "1," causing an unbalance. Figure 3 is a bar graph that displays the number of labels in the collection for each label.

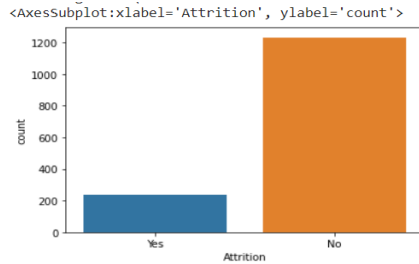


FIG: 3 BAR PLOT FOR TARGET ATTRIBUTE DISTRIBUTION

Using the Synthetic Minority Oversampling Technique (SMOTE), entries for the class with a lower total were artificially generated. SMOTE, a method for oversampling the minority class, was chosen over under sampling because the latter could lead to the removal of important data[12].

### C) VISUALIZATION

This process provides valuable insights into the dataset and helps to distinguish important features from irrelevant ones. Overall, visualization is a crucial step in data analysis that enables us to quickly gain a high-level understanding of the data and make informed decisions about feature selection.

#### 1) Attrition vs Business Travel:

From Figure 4, we can clearly knowing that Non-Travel employees having low attrition rate. In other way, employees who travel from one place to other place on business purpose, having high attrition rate.

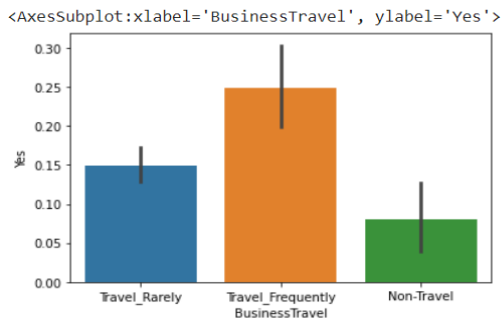


FIG: 4 BAR PLOT REPRESENTATION FOR BUSINESS TRAVEL

#### 2) Attrition on basis of gender:

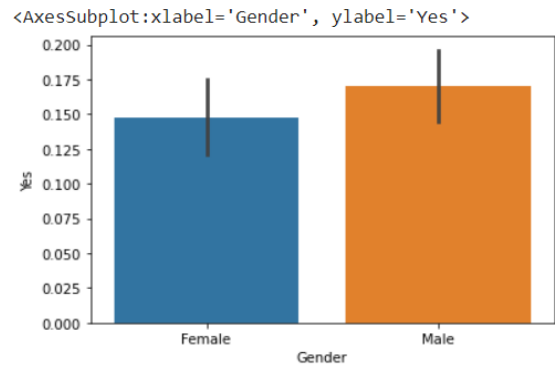


FIG: 5 BAR CHART REPRESENTATION FOR 'GENDER'

Figure 5 shows that the turnover rate is not significantly influenced by the employee's gender. In each instance, the turnover rate stays about the same. This demonstrates that Gender is not a characteristic that should be considered for inclusion in future attrition forecast methods. These graphic representations make feature selection and reduction much more understandable and simple.

## V. FEATURE IMPORTANCE AND TRAIN MODEL

A) *Divide dataset into Train and Test*: To prepare the data for machine learning, the 'DataFrame' was divided into two subsets: Train and Test. The Train set was used to train the machine learning algorithm, and the knowledge gained was used to predict the required attribute for the Test set. It is important to have a larger Train set than Test set as this helps the machine learn better from the dataset. Typically, the train data should be around 70-85% of the entire dataset. In particular case, the train data consists of 75% of the 'DataFrame', i.e 1249 rows, where other 15% or 221 rows are from test data.

B) *Feature Importance*: In machine learning, feature importance refers to the process of determining the relative importance of different input variables[7], or features, in predicting the output of a model.

Feature importance is useful because it helps to identify which features are most relevant to the problem being solved, and which features can be ignored or removed to simplify the model without sacrificing accuracy. This information can be used to optimize the performance of the model by focusing on the most important features and reducing the dimensionality of the data.

### C) Machine Learning Models for prediction:

After preparing the data, the next step in using machine learning models for prediction involves an loop process that aims to improve the accuracy of the models. There are several classification models that can be used for this purpose:

1).*Decision Tree Classifier*: This method is suitable for multistage decision-making and breaks down complex decisions into elementary ones for easy interpretation[2][3].

2).*Support Vector Machine (SVM)*: This approach can be utilized for both classification and regression tasks, and it involves constructing a hyperplane with maximum margin in a transformed input space to separate different classes of examples. The goal is to ensure that the hyperplane is as far as possible from the nearest correctly classified examples, which results in a well-separated and accurately classified dataset[4].

3).*Logistic Regression*: It is one of the simplest supervised machine learning algorithms. The logistic regression technique employs a linear model to convert the predictor variables into a probability value between 0 and 1. The logistic function parameters are estimated by the model using a technique known as maximum likelihood estimation, that involves determining the parameter values that affect the probability of observing the data. [14][6]

4).*Random Forest*: This method is an ensemble learning algorithm that generates multiple sub decision trees and merges them to generate better accurate and stable prediction[9].

In this project we train model using Random Forest algorithm, logistic regression[6], SVM and Decision tree. In these algorithms random forest gives best accuracy when compared to other classification models.

### D) Result Analysis:

When evaluating a machine learning model, it is important to use appropriate metrics to measure its performance. Three commonly used metrics in machine learning are:

1).*Accuracy*: This metric used to measure the proportion of correct predictions made by the model over the total number of predictions. It is calculated by dividing the number of correct predictions by the total number of predictions made.

2) *Confusion Matrix*: It is a matrix representation of TP, TN, FP and FN values. Using this matrix we can

also find out the accuracy score by  $(TP+TN)/(TN+TP+FN+FP)$ [10][11]

Among all classification algorithms, we observe that Random Forest algorithm giving best accuracy score and also predicting accurately on unknown data.

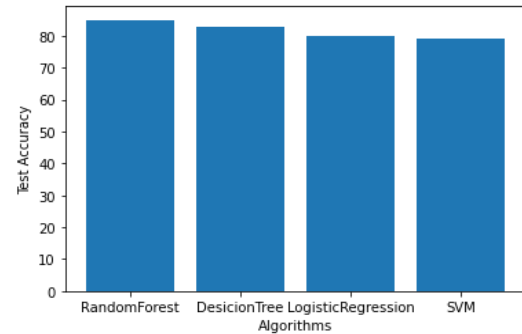


FIG:6 BAR GRAPH FOR TEST ACCURACIES

Figure 6 represents test accuracies of algorithms. From above figure we clearly observe that random forest algorithm gives best accuracy compared to other.

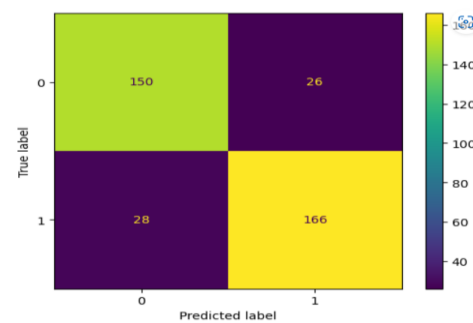


FIG:7 CONFUSION MATRIX FOR RANDOM FOREST

Figure 7, shows how correctly the random forest algorithm predict the target class.

## VI. CONCLUSION

After assessing the execution of four classification models, a significant finding was that if feature reduction for prediction is appropriately conducted, the accuracy rate of the classification models always be better compared to classification with feature selection. In particular, the Random Forest classifier with feature reduction achieved an accuracy score of 85.3%, while the Decision tree classifier achieved 83%. Random Forest model giving best classification for True positives and True negatives data. The methods described in the paper for analyzing and categorizing data can form a basis for improving data-driven decision-making processes. These techniques can unlock new insights from data and

help organizations improve their operations. Implementation of these methods can also contribute to a positive work culture and improve an organization's reputation in their respective industry.

## REFERENCES

- [1] Srivastava, Devesh Kumar, and Priyanka Nair. "Employee attrition analysis using predictive techniques." *International Conference on Information and Communication Technology for Intelligent Systems*. Springer, Cham, 2017.
- [2] S. S. Gavankar and S. D. Sawarkar, "Eager decision tree," 2017 2nd International Conference for Convergence in Technology (I2CT), Mumbai, 2017, pp. 837-840.
- [3] Safavian, S.R. Landgrebe. D, "A survey of decision tree classifier methodology", *IEEE Transactions on Systems, Man, And Cybernetics*, Vol. 21, No. 3, May-June 1991.
- [4] Shmilovici A. (2009) Support Vector Machines. In: Maimon O., Rokach L. (eds) *Data Mining and Knowledge Discovery Handbook*. Springer, Boston.
- [5] Setiawan, I., et al. "HR analytics: Employee attrition analysis using logistic regression." *IOP Conference Series: Materials Science and Engineering*. Vol. 830. No. 3. IOP Publishing, 2020
- [6] Schober, Patrick MD, PhD, MMedStat\*; Vetter, Thomas R. MD, MPH†. *Logistic Regression in Medical Research. Anesthesia & Analgesia* 132(2):p 365-366, February 2021. | DOI: 10.1213/ANE.0000000000005247
- [7] Jayalekshmi J, Tessy Mathew, "Facial Expression Recognition and Emotion Classification System for Sentiment Analysis", 2017 5 Authorized licensed use limited to: University College London. Downloaded on May 23,2020 at 00:07:22 UTC from IEEE Xplore. Restrictions apply. *International Conference on Networks & Advances in Computational Technologies (NetACT) |20-22 July 2017| Trivandrum*.
- [8] Isabelle Guyon, Andre Elisseeff, "An Introduction to Variable and Feature Selection", *Journal of Machine Learning Research* 3 (2003) 1157-1182.
- [9] Ilan Reinstein, "Random Forest(r), Explained", *kdnuggets.com*, October 2017[Online]. Available: <https://www.kdnuggets.com/2017/10/randomforests-explained.html>
- [10] [http://scikitlearn.org/stable/modules/generated/sklearn.metrics.confusion\\_matrix.html](http://scikitlearn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html)
- [11] [http://scikitlearn.org/stable/auto\\_examples/model\\_selection/plot\\_confusion\\_matrix.html](http://scikitlearn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html)
- [12] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", *Journal of Artificial Intelligence Research* 16 (2002), 321 – 357
- [13] Pavan Subhash, "IBM HR Analytics Employee Attrition & Performance", [www.kaggle.com,2016\[Online\]. Available:https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset](https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset)
- [14] Sperandei S. Understanding logistic regression analysis. *Biochem Med (Zagreb)*. 2014 Feb 15;24(1):12-8. doi: 10.11613/BM.2014.003. PMID: 24627710; PMCID: PMC3936971.