2019 5th International Conference for Convergence in Technology (I2CT)
Pune, India. Mar 29-31, 2019

Employee Attrition Prediction Using Classification Models

Namrata Bhartiya
Mukesh Patel School of Technology Management and Engineering,
NMIMS University
bhartiya_namrata@yahoo.com

Sheetal Jannu
Mukesh Patel School of Technology Management and Engineering,
NMIMS University
sheetal.jannu123@gmail.com

Purvika Shukla
Mukesh Patel School of Technology Management and Engineering,
NMIMS University
pshukla97@gmail.com

Radhika Chapaneri
Mukesh Patel School of Technology Management and Engineering,
NMIMS University
radhika.chapaneri@nmims.edu

***Abstract*- The term Attrition refers to the voluntary or involuntary discontinuation of employees in an organization. This paper focuses on discussing a systematic flow for predicting Attrition using Data Analysis and Machine Learning techniques. The steps include Data Acquisition, Data Conditioning, Visualization, and Classification by applying the following Classification Models: Support Vector Machine, Decision Tree, K-Nearest Neighbor, Random Forest, and Naive Bayes algorithms in the Python environment. The resulting predictions of classification were evaluated using three performance metrics: Accuracy Score, Confusion Matrix and the ROC Curve. Based on the obtained results, we inferred that Random Forest classifier delivered the highest accuracy being 83.3% whereas Naive Bayes and Support Vector Machine was better in terms of classifying True Positives and indicated greater Area Under Curve values. This paper intends to be of great use to the Organizations aiming at detecting the key causes of Attrition and minimizing them using the power of data.**

Keywords- Attrition; Machine Learning; Supervised Learning; Data Analysis; Classification Models.

I. INTRODUCTION

In today's world data is being created at an ever-increasing rate. The analysis of this stored data has proved to be beneficial in gaining insights and creating general awareness about any business or organization. Data analysis is the process of collecting, inspecting, cleansing, transforming, and modeling raw data with the aim of deriving valuable insights and retrieving relevant information to reach a conclusion for good decision making. Machine Learning is the process of using algorithms to train a machine to predict accurately using the existing data.

Employees are a crucial resource for any organization, and hence withdrawal of productive employees might affect an organization with respect to various aspects. Some of the consequences of Employee Attrition are: Investing in staffing and training new employees, increased burden on existing employees and a decline in the performance of the organization. In this paper, we intend to classify employees with respect to previous 'Attrition' patterns and other relevant attributes. Section II of the paper provides a brief description of the dataset used in the process and Section III discusses the acquisition of the data and importing the required libraries. Some common problems encountered during the analysis such as poorly formatted data files, inconsistent and incomplete data, duplicate entries, different value representation and misclassified data can cause difficulties in predicting the desirable informative pattern in the data. In Section IV of the paper, we discuss Data Exploration steps which include exploring the data, preparing it for further utilization and Section V is a brief summary of visualizations to get an overview of the data set. The error-free data set from the exploration step is then used to anticipate valuable indications by training classification models using selected features of the data set. Section VI of the paper will focus on Feature Selection, Training, and Testing Classification Models.

II. DATASET INFORMATION

The dataset used for the process is the 'IBM HR Employee Analytics Attrition and Performance', which was obtained from Kaggle, an online source for datasets and a platform for data science-related competitions [13]. This dataset comprises 1470 records and 35 attributes. The data columns consist of independent variables such as 'Age', 'Gender', 'Department', 'Distance from Home', etc. and in this paper we consider 'Attrition' to be the dependent variable. 'Attrition' data column consists of two class labels, 'Yes' or 'No'. The 'Attrition' rate in the organization is 16%.

III. DATA AND TOOLS ACQUISITION

A. Import Libraries: For data analysis and machine learning in the Python environment, we consider the following powerful and effective libraries for the analysis and predicting Attrition:

1) *Numpy:* It is one of the most important python libraries for Mathematical and Scientific Computations.

978-1-5386-8075-9/19/$31.00 ©2019 IEEE

2) *Pandas:* A library designed for easy and agile manipulation of data frames.

3) *Matplotlib:* A python library designed for creating visualizations by the powerful generation of graphs and charts such as bar graphs, pie charts, scatter plot and more.

4) *Scikit-Learn:* A number of supervised and unsupervised machine learning algorithms are offered by SciKit-Learn library. Data Modelling is the primary focus of machine learning libraries.

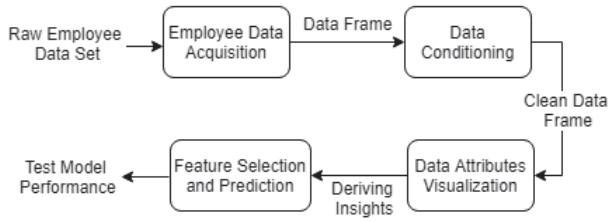

Fig.1: Block Diagram of the process.

B. Import Dataset: We carried out the analysis steps on the existing dataset of the .csv format by reading it using the Pandas library function read_csv().

C. Storing Dataset as Data Frame: The dataset read was stored in the form of a Data Frame for further analysis and predictions. The alterations made during the data conditioning step will affect the data frame, the original data remains unchanged.

IV. DATA CONDITIONING

Data Conditioning refers to the process of cleaning data, normalizing datasets and performing transformations on the data. These steps are performed to get the datasets into a state that enables analysis in further phases [1]. In Data Conditioning the following steps were performed:

A. Explore Dataset Properties:

The objective of data exploration was to understand the relationships among the variables and to analyze the problem domain [1]. This exploration stage is valuable for detecting frequent issues such as Null values, Outliers, Redundancies, etc. encountered in datasets.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
Age                       1470 non-null int64
Attrition                 1470 non-null object
BusinessTravel            1470 non-null object
DailyRate                 1470 non-null int64
Department                1470 non-null object
DistanceFromHome          1470 non-null int64
```

Fig.2: A part of output with the number of records and data type.

B. Data Preparation:

It comprises of the steps to explore, preprocess, and condition data prior to modeling and analysis. This process was carried out in order to learn facts about our dataset and become familiar with it. It involved transforming data into a format to facilitate subsequent analysis. This is typically the most labor-intensive and most iterative step in the analytics lifecycle [1]. The major steps followed for preparing data are:

1) Feature Reduction: This step was crucial in making decisions about which of the features in the dataset will be useful for analysis in the later stages - deciding on which features of the dataset are to be retained and which features to transform or discard. This was done based on the discretion of which feature is relevant and which is irrelevant for attrition prediction. Examples of some properties based on which features were discarded from the further analysis are:

a. *Multiple Employee Ratings:* Attributes such as Hourly Rating, Monthly Rating, and Daily Rating provide similar kind of ratings for an employee. Having multiple employee ratings not only increases the time and space overhead for analysis but also does not add any valuable insight into the prediction. Hence, we merged them into a single column attribute, Employee Rating which was replaced by the mean value of the Daily ratings. This results in a single measure for employee performance on a monthly basis.

b. *Attributes having Non-Unique values:* The following attributes consist of non-unique values: 'Employee count', this attribute gives the value of employee count which is '1' for every employee. 'Standard working hours', this attribute gives the value of the standard working hours of an employee which is '80' for every record. 'Over 18 yrs of age', this attribute gives the age criteria validation (to be over 18) of an employee which is 'Yes' for every record.

2) Data Cleaning: Removal of data anomalies, typically missing values, redundant data and for detection of outliers to ensure an improved data quality.

3) Categorical to Numerical: Since standard libraries do not take categorical variables as input, these values need to be converted to numeric form. This was done using the Label Encoder Method to transform non-numerical labels into numerical labels or nominal categorical variables as shown in Figure 3. Numerical labels are always between 0 and n_classes-1.

	Attrition
0	Yes
1	No
2	Yes

Attrition
1
0
1

Fig.3: (a)Attrition values before converting, (b)Attrition values after conversion.

4) Data Binning: In this step, we group a number of more or less continuous values into a smaller number of so-called "bins". These bins are assigned labels to refer to each particular range of values. It was performed to constrict the range of values so that the values are represented in a concise yet informative form. Since the range is constricted, the visualizations performed on those attributes are much clearer and interpretable. Figure 4 shows the resulting 'Age Category' after binning.

Age	AgeCategory
41	3
49	4
37	3
33	2
27	2

Fig.4: Resulting 'Age Category' attribute.

5) Balancing Dataset: In the Employee dataset, the number of records having 'Attrition' label as '0' is greater than the number of records having 'Attrition' label as '1', thus creating an imbalance in the dataset. Figure 5 is a bar graph which shows the count of each label in the dataset.

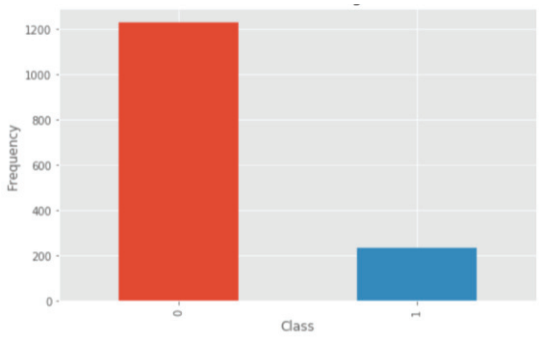

Fig.5: Bar chart for Class Distribution.

Synthetic Minority Oversampling Technique (SMOTE) was performed to synthetically generate records for the class with a lesser count. SMOTE is a technique for the oversampling of the minority class and was preferred over undersampling as it might result in the loss of crucial information[12].

V. VISUALIZATION

This step was performed to gain an overview of the data and examining the data quality. Visualizations help in understanding the data characteristics, including its trends - consistencies or inconsistencies, outliers, skewness and the relationships among data variables. It enables assessing the data granularity, the range of values and the level of aggregation of the data. It provides a high-level view of the data and a great deal of information about a given dataset in a relatively short period of time [1]. This step also plays a key role in the further filtering of the relevant features from the irrelevant ones. Some of the realizations that we encountered in the Visualization steps that aided in feature filtering process:

A. Attrition with respect to Education Field:

In Figure 6, it can be observed how different education fields influence the attrition rate differently. For instance, the attrition rate for the "Technical" field is 24.2% while it is only 13.6% for the "Medical" field.

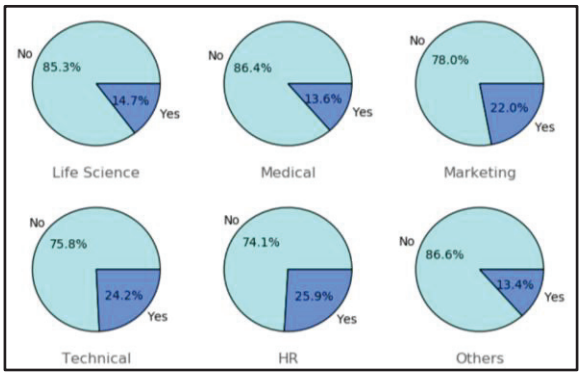

Fig.6: Pie-Chart Visualization for 'Education Field'.

From Fig. 6 representation, the following information can be inferred:

- Which professions are trending in the market?
- Which area requires a salary hike to reduce attrition?
- Which field needs to work on ways in order to retain skilled employees?

B. Attrition with respect to Gender:

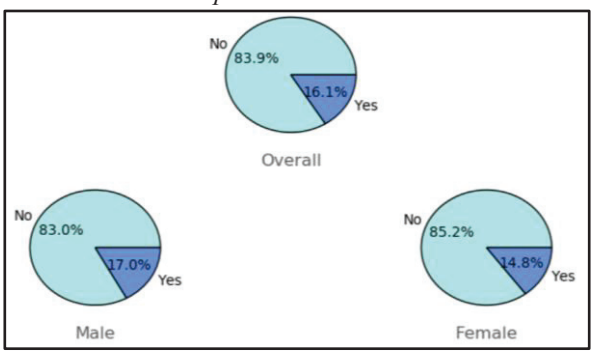

Fig.7: Pie-Chart Visualization for 'Gender'.

In Figure 7, we observe that the gender of the employee has no significant influence on the attrition rate. The attrition rate remains similar in each case. This shows that Gender is not a potential feature to be included in the further attrition prediction processes. Such visualizations make the feature selection and reduction much clearer and easier.

C. Attrition with respect to Performance Rate:

In Figure 8, we observe that higher the performance rating, higher is the attrition rate. For instance, for a one-star rating, the attrition rate is 13.7% while for a five-star rating it is 25.9%. Such cases are not really a benefit to the organization. This is so because even though ideally any organization would desire a 100% employee retention, but at the same time would not want to retain employees with poor performance.

One other element observed in the figure is a pie chart which shows 100% employee retention when the performance rating is zero stars. This is a specific case of an outlier.

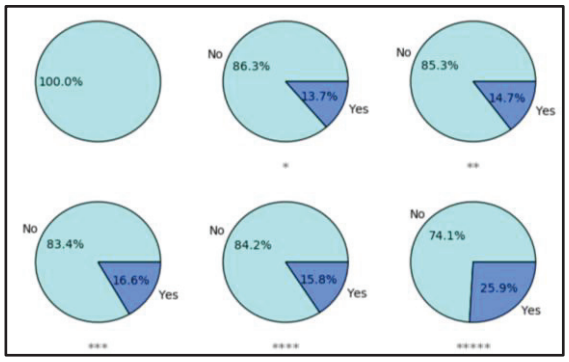

Fig. 8: Pie-Chart Visualization for 'Performance Rate'.

VI. FEATURE SELECTION AND PREDICTION

A. Splitting Data into Train and Test: The dataframe 'EDataFrame' is split into Train and Test data. The machine was trained by employing the Train set to a machine learning algorithm and based on the knowledge gained from the train set, the required attribute will be predicted for the Test set. The size of the Train set has to be greater than the Test set as this will ensure better learning of the data. Usually, the Train set is 75-80% of the entire dataset. In this case, the Train set is 75% of the entire data frame which is 1249 records, and the remaining 15% or 221 records is the Test set.

B. Feature Selection: Selection of relevant attributes and the reduction of redundant attributes is crucial for determining accurate results [7]. Some of the advantages of Feature selection are prediction performance improvement, reduces training utilization time, results in a better understanding of data concepts, and reduces the complexity of the process.

C. Using Classification Models for Prediction: An iterative process post data preparation which is run on models for repetitive improvements to ensure maximum accuracy.

1) Decision Tree: Decision Tree Classifier is one of the suitable methods for multistage decision making. In any multistage decision-making process, a complex decision is fragmented into several elementary decisions, this has proved to be an important feature of Decision Tree Classifiers. Fragmentation of decision-making process makes it easy for the user to interpret the solution [1][2].

2) Support Vector Machine: Developed in 1995 by Cortes and Vapnik, Support Vector Machines are used for both classification and regression. The SVM classifier creates a maximum-margin hyperplane that lies in a transformed input space and splits the example classes while maximizing the distance to the nearest cleanly split examples. Support vectors are the points lying on the boundaries and the center of the margin is the required hyperplane. The parameters of the solution hyperplane are derived from a quadratic programming optimization problem [3][4]

3) k-Nearest Neighbor: k-Nearest Neighbour (k-NN), uses an instance-based learning approach, in which it considers the minimum distance from the query instance to the training data set to fetch k-nearest neighbors. After fetching the k-nearest neighbors, the majority of these are taken into consideration to predict the output of the query instance[14]. The training records are vectors in a dimensional feature space, each with a class label [6][5].

4) Random Forest: Random Forest is a type of supervised learning algorithm, that is used for both classification and regression. The method followed under this algorithm is that it builds multiple decision trees and merges them together to get a more accurate and stable prediction. It is a popularly referred to as an ensemble learning algorithm, i.e., it takes a bunch of randomly implemented decision trees as input and generates one strong Python Machine Learning predictor, random forest [9].

5) Naive Bayes: Naive Bayes is a classification technique based on Bayes' theorem, which is based on the assumption that predictor variables will be independent of other variables. This method is intuitive in nature and makes use of probabilities of each attribute of each class for making predictions. The specialty of this technique is that the presence or absence of a certain feature in a class is not related to the presence or absence of any other feature in the same class. Generally, the input variables are categorical in nature. The Naive Bayes classification algorithm is best suited for handling large datasets as it is easy to implement and can effectively undergo execution without any prior knowledge of data required [1].

D. Performance Testing: In order to assess the correctness of the predicted values by the classification models to evaluate the performance of a classifier, we made use of 3 metrics, which are:

1) Accuracy: The 'accuracy_score' function compares the data set values of 'Attrition' with the predicted values, it calculates the number of exact matches in the labels predicted by the classification models. The accuracy attained using the Decision tree, SVM, k-NN, Random Forest and Naive Bayes models in predicting 'Attrition' are shown in Table 1.

TABLE I: ACCURACY RATE OBTAINED FOR DIFFERENT CLASSIFICATION MODELS.

Classification Model	Accuracy with Feature reduction	Accuracy without Feature reduction
Decision Tree	81.0%	69.7%
SVM	76.5%	38.9%
k-NN	76.8%	61.1%
Random Forest	83.3%	41.6%
Naive Bayes	71.0%	45.7%

2) Confusion Matrix: Confusion Matrix is a specific table layout that allows visualization of the performance of a classifier [1]. The diagonal elements indicate the number of records classified as its true label whereas the remaining

elements indicate the number of misclassified records [11]. In a Confusion Matrix M, for binary classification, $M_{0,0}$ represents the number of True negatives(TN), $M_{1,0}$ represents the count of False positives(FP), $M_{0,1}$ represents the number of False negatives(FN) and $M_{1,1}$ represents the number of True positives(TP).

A good classifier should have large TP and TN and small (ideally zero) numbers for FP and FN [1][10]. The confusion matrices generated as a result of the predictions obtained using each of the classification models are shown in Figure 9. From these results, we can infer that SVM and Naive Bayes are better in terms of True Positive Rate, which implies that they classified a major number of records with label '1' as '1'. Whereas Random Forest, k-NN, and Decision Tree show better results in classifying the True Negatives, which indicates they classified a decent number of records with label '0' as '0'. A certain cost is assigned to the misclassified records, it depends on the specifications of an organization, whether classifying a record with label '1' as '0' is acceptable or classifying a record with label '0' as '1'.

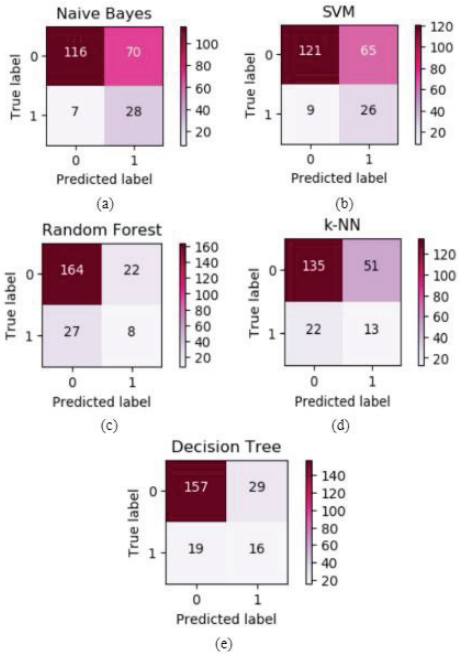

Fig. 9: Confusion Matrix for (a)Naive Bayes, (b)SVM, (c)Random Forest, (d)k-NN, (e)Decision Tree

3) ROC curve: Receiver Operating Characteristic (ROC) curve is another metric used to evaluate the performance of classification models. In a ROC curve, two operating characteristics, true positive rate(TPR) is plotted against false positive rate(FPR). An ideal classifier is expected to perfectly separate positives from negatives. Thus, this will help achieve the top-left corner(TPR=1, FPR=0). However, a better classifier is supposed to lie closer to the top left. The area under the curve (AUC) is the area which is related to the ROC curve. Therefore, the value of AUC is calculated by measuring the area under the ROC curve. Higher the AUC scores mean the classifier performs better. The AUC score ranges from 0.5 (TPR=FPR) to 1(ROC passing through the top-left corner)[1]. Figure 10 describes the resulting ROC for the 5 classification models used to classify 'Attrition' values. From these results, it is evident that the AUC values of SVM and Naive Bayes algorithm are much higher as opposed to the other classification models, they clearly separate the positives from the negatives, lying closer to the top left.

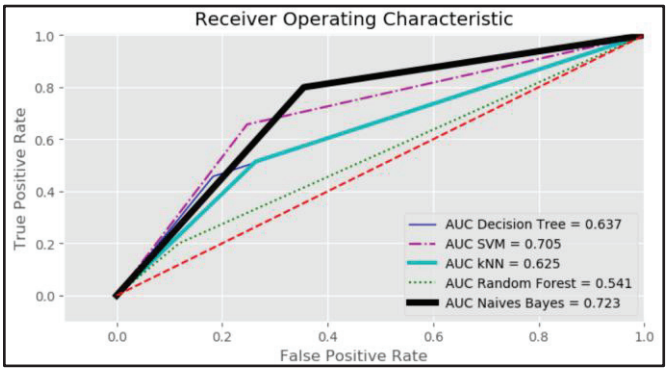

Fig. 10: ROC Curve Comparison for the Classification models used.

VII. CONCLUSION

On evaluating the performance of the five classification models, a major observation is that, if the feature selection for prediction is properly done then no matter which classification model is used, the accuracy rate will always be better in comparison to classification without feature reduction. The accuracy score obtained from Random Forest classifier with feature reduction was 83.3% and that of Decision tree classifier was 81% as shown in Table 1. These two classifiers have higher accuracy scores due to greater values for True Negatives as compared to others. From the Confusion Matrix results, it was observed that SVM and Naive Bayes algorithms provided best classification of True Positives. Naive Bayes, and SVM also delivered greater Area Under Curve(AUC) values as opposed to the other three models. The analysis and classification steps discussed in the paper can be a stepping stone towards advancements of a number of data-driven decision making, harnessing the power of data for deriving new insights in the corporate sector in order to develop and improve the work ethics and corporate standing of an organization.

REFERENCES

[1] EMC Education Services, "Data Science and Big Data Analytics - Discovering, Analyzing, Visualizing and Presenting Data", July 2015.

[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] David Meyer, "Support Vector Machines", R-News, Vol.1/3, 9.2001.

[6] Vincent Garcia and Eric Debreuve and Michel Barlaud. "Fast k Nearest Neighbor Search using GPU", arXiv:0804:1448Vi, April 2009.

[7] Jayalekshmi J, Tessy Mathew, "Facial Expression Recognition and Emotion Classification System for Sentiment Analysis", 2017

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/random-forests-explained.html
[10] http://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html
[11] http://scikit-learn.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
[14] Kardi Teknomo, "How K-Nearest Neighbor (KNN) Algorithm works?", people.revoledu.com,2017[Online].Available:https://people.revoledu.com/kardi/tutorial/KNN/HowTo_KNN.html