Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.4, May : 2024

PREDICTION OF EMPLOYEE ATTRITION USING MACHINE LEARNING

Mr.K.V.Narasimha Reddy Assistant Professor Computer Science and Engineering Narasaraopeta Engineering College (Autonomous) Narasaraopet,Andhra Pradesh narasimhareddynec03@gmail.com
P. Bhavani Manjunadha Student Computer Science and Engineering Narasaraopeta Engineering College (Autonomous) Narasaraopet,Andhra Pradesh nadhamanju45@gmail.com
G.venkata Sai Student Computer Science and Engineering Narasaraopeta Engineering College (Autonomous) Narasaraopet,Andhra Pradesh saigadiparthi24@gmail.com
A.Narendra Reddy Student Computer Science and Engineering Narasaraopeta Engineering College (Autonomous) Narasaraopet,Andhra Pradesh narendrareddya77@gamail.com

Abstract-
Attrition encompasses employee departures from the company, including voluntary actions like retirement, resignation, as well as involuntary events such as termination or death. The attrition rate signifies the frequency at which employees depart from a company within a defined timeframe, usually expressed as a percentage. This rate is computed by dividing the number of employees who leave (whether voluntarily or involuntarily) by the average workforce size during that period, then multiplying by 100. Various factors, such as job dissatisfaction, unfavorable working conditions, or seeking better compensation, can contribute to attrition. These reports are typically reviewed by the human resources department and the respective department managers of the departed employees. Multiple Machine Learning Algorithms have been deployed to forecast the attrition rate within organizations. We have followed various processes such as Dataset Collection, Data Preprocessing, Data Visualization, apply SMOTE on the data and training the model by using the various ML algorithms like Extra Tree Classifier, Random forest algorithm and decision tree algorithm, Logistic Regression, K-Nearest Neighbors, Gradient Boost and XG Boost Classifier algorithms. The results are evaluated using accuracy score and confusion matrix. XG Boost classifier algorithm giving the best accuracy i.e., 91.16% compared to the other evaluating models. This work will help organizations to better understand the attrition causes.

Keywords- Attrition, SMOTE, random forest, SVM, decision tree, Extra Tree Classifier, XG Boost Classifier, dissatisfaction

I. INTRODUCTION

Data is being created in the modern world at a never-before-seen rate and is expanding quickly. This wealth of information is an invaluable tool for learning about and increasing awareness of different businesses or organizations. However, preprocessing the data is crucial before moving on to data modeling. In order to uncover useful patterns and retrieve pertinent results to support well-informed decision- making, the data must be arranged and refined. Preprocessing basically trains computers to make precise predictions based on historical data patterns, readying the data for analysis and prediction.

In today's interconnected world, employee attrition remains a significant and complex challenge within organizational cultures. Attrition, characterized by the gradual reduction in the workforce due to factors such as resignations, retirements, or unfortunate events, is an inevitable aspect of any business. Various forms of attrition exist, including demographic-specific trends, internal shifts, and voluntary departures. Extensive research has highlighted diverse factors contributing to employee attrition, ranging from financial concerns and limited career advancement opportunities to workplace dynamics and inadequate work-life balance. 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. Thispaper 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 unwantedattributes[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.

In this paper, research has been performed on the IBM HR analytics dataset consists of 35 attributes containing both categorical and numerical features and 1471 records are trained by machine learning classification models – Random Forest algorithm, Decision Tree, Extra Trees, Logistic Regression, K-Nearest Neighbors, Gradient Boost and XG Boost Classifier algorithms. – Trained to assess the correlation between the attributes of both retained and terminated employees, the models utilize a heatmap visualization. Subsequently, each model undergoes testing to determine its accuracy, aiding in the selection of the optimal model for predicting employee attrition. Performance metrics are employed to describe the results, revealing that the XG Boost Classifier machine learning approach achieved the highest accuracy of 91.16% for the dataset.

II. LITERATURE SURVEY

According to Barron's Business Dictionary,
attrition is "the normal as well as an uncontrollable decline of a workforce due to retirement, death, sickness, as well as relocation." It's a tactic to reduce labor size without explicitly involving upper management. Attrition-based decline has the drawback of frequently being unpredictable, which can cause gaps within an organization. The literature has numerous studies on attrition prediction analysis. Predicting staff attrition was the main area of concentration. To examine the factors influencing the attrition rate, researchers have used machine learning classification models such as SVM, random forest, logistic regression. precise projections derived from historical data trends.

Srivastava et al. introduced a framework that utilizes machine learning techniques to forecast employee turnover by examining employee behaviors and various attributes. Setiawan et al through their work found variables that have a major impact on employee attrition. They implemented the CRISP-DM 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.

Amir Mohammad Esmaieeli Sikaroudi,
Rouzbeh Ghousi, and among others, conducted knowledge discovery processes on real manufacturing plant data, examining employee attributes such as age, salary, and work experience. They utilized the Pearson Chi- Square test to ascertain the significance of data.

John M. Kirimi, Christopher Moturi, and colleagues proposed a predictive model for forecasting employee performance, enabling HR professionals to focus on human capability criteria and thereby improve the performance appraisal process.

Rohit Punnoose, Pankaj Ajit, and their team explored the application of Extreme Gradient Boosting (XG Boost) technique, renowned for its robustness due to its regularization formulation. They leveraged data from a global MNC’s HRIS to compare XG Boost against six traditional supervised algorithms, demonstrating its notably higher accuracy in predicting employee attrition rate.

III. METHODOLOGY

Machine learning techniques were employed to construct a predictive model capable of anticipating whether a worker would leave their job voluntarily or involuntarily. The model was built upon data mining methodologies for data collection, preprocessing, and refinement. Figure 1 illustrates the conceptual framework of the model.

- Dataset Collection

- Data Pre-Processing
- Data Visualization
- Model Development

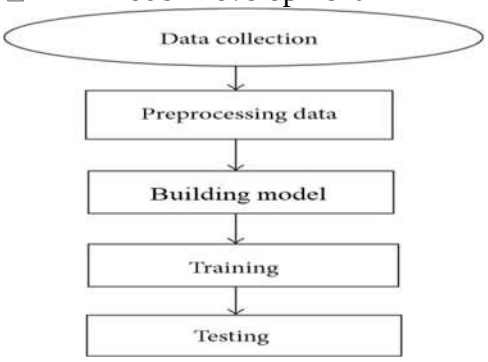


Figure.1 Methodology

3.1. DATASET COLLECTION

"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.

Dataset Features		Dataset Features	
Attributes	*Data-types*	*Attributes*	*Data-types*
Age	Numeric	MaritalStatus	String
Attrition	Boolean	MonthlyRate	Numeric
BusinessTravel	String	NumCompaniesWorked	Numeric
DailyRate	Numeric	Over18	Boolean
Department	String	OverTime	Boolean
DistanceFromHome	Numeric	PercentSalaryHike	Numeric
Education	Numeric	PerformanceRating	Numeric
EducationField	String	RelationshipSatisfaction	Numeric
EmployeeCount	Numeric	StandardHours	Numeric
EnployeeNumber	Numeric	StockOptionLevel	Numeric
EnvironmentSatisfac	Numeric	TotalWorkingHours	Numeric
tion	String	WorkLifeBalance	Numeric
Gender	Numeric	YearsAtCompany	Numeric
HourlyRate	Numeric	YearsInCurrentRole	Numeric
JobInvolvement	Numeric	YearsSinceLastPromotio	Numeric
JobLevel	String	nYearsWithCurrManager	Numeric
JobRole	Numeric		
JobSatisfaction			

Fig.2 Dataset Features

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 visualizations like bar plots, pie charts, and more.

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

2) Seaborn: Seaborn library helps you to visualize the data using pair plots that produce a matrix of relationships between each variable in the dataset.

3) Read Dataset: Read the dataset of .csv format using pandas function read_csv().

4) 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

Data preprocessing is a set of techniques and procedures applied to raw data before it is fed into a machine learning algorithm for analysis or modeling.

It involves cleaning, transforming, and organizing the data to make it suitable and efficient for the intended analytical process. Typical data preprocessing steps include data cleaning (handling missing values, outliers, etc.), data transformation (scaling, normalization, encoding categorical variables), and feature selection or extraction. The goal is to ensure that the data is in a consistent, reliable, and usable format, ultimately improving the performance and accuracy of machine learning models. In Data preprocessing the following steps were performed:

1) Investigate Dataset Properties: The goal of the research was to comprehend connections between factors and examine issue at hand. This research step is useful for spotting commondataset problems like Null values, Outliers, Redundancies, etc.

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 data will helpful for analysis. The decision as to which trait is important and which is not for attrition forecast wasmade. Following are some examples of characteristics based on which elements wereexcluded 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.

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.

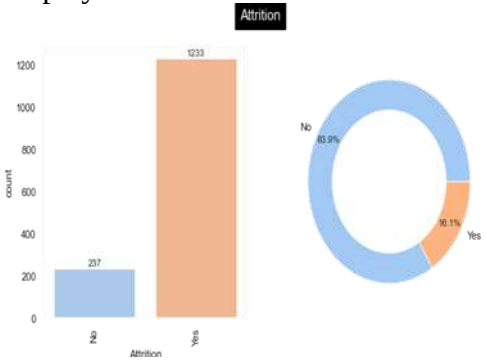


FIG.6 BAR PLOT AND PIE CHART FOR TARGET ATTRIBUTE DISTRIBUTION

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 used in data analysis that enables us to quicklygain a high- level understanding of the data and make effective decisions about features 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.

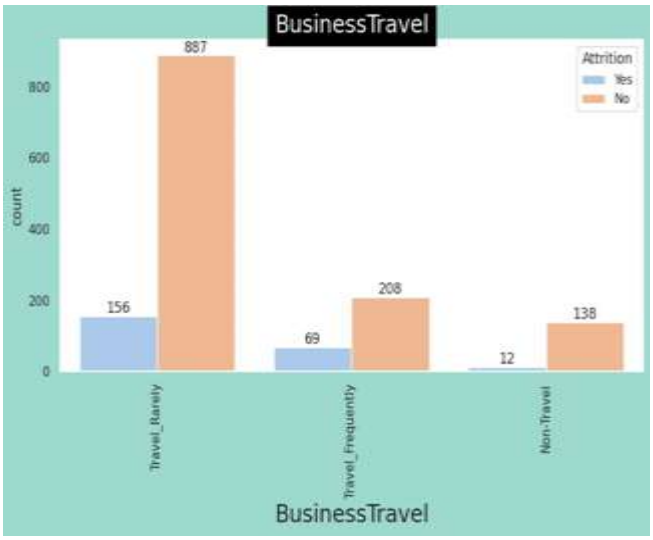


FIG.7 BAR PLOT REPRESENTATION FOR BUSINESS TRAVEL

2) Attrition vs Education Field

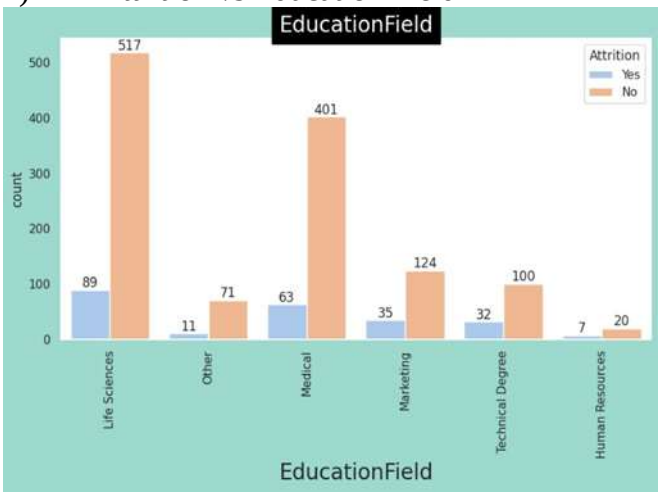


Fig.7 Attrition Vs Education Field

The greatest percentage of attrition, or 37.5% of all attrition, is attributed to employees with a life science degree level (89 out of 237). However, that only makes up 14.7% of attrition in the field of life sciences. The second-highest attrition rate, accounting for 13.57% of total attrition (63 out of 237), is seen at the medical education level. However, that only makes up 14.7% of attrition in the field of life sciences. In addition, the domains of marketing, technical degrees, and human resources are the most impacted by attrition, in that order. Approximately 22- 26% of their workforce departed the business.

3) Attrition vs Marital Status

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.4, May : 2024

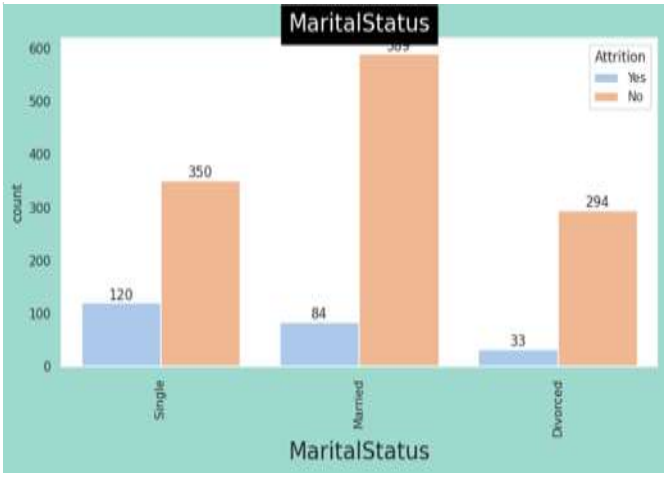


FIG.8 Attrition Vs Marital Status

Employees who are single exhibit the highest likelihood of departing from the company, comprising 50.6% of the total attrition. Following single employees, married and divorced individuals represent the subsequent largest groups in terms of attrition within the company.

4) Attrition on basis of gender:

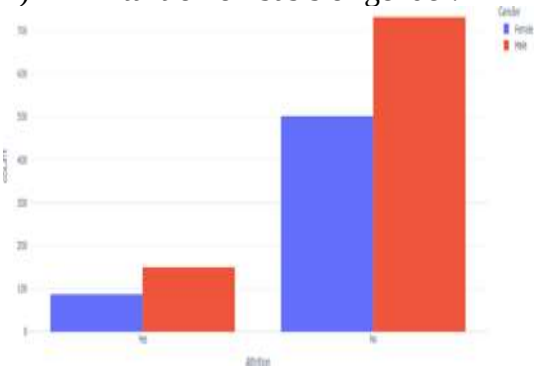


FIG: 9 BAR CHART REPRESENTATION FOR 'GENDER'

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 reductionmuch more understandable and simple.

5) MOST JOB ROLES FOR ATTRITIONS

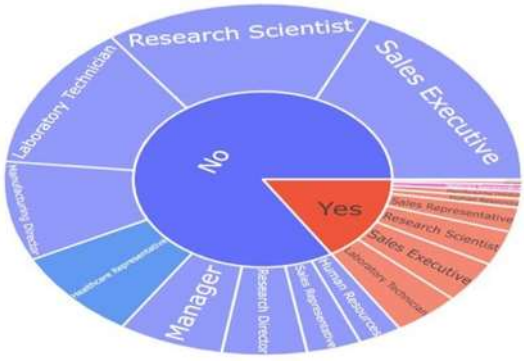


Fig.10 Pie chart of job role

The above figure represents the which job role is causing more attritions. We observed that the Laboratory technician is the attritions. Laboratory technician attrition may stem from factors like

inadequate compensation, limited career advancement opportunities, stressful work environments, and insufficient support for professional development. Addressing these issues through improved benefits, clearer career pathways, and better work-life balance initiatives can help organizations retain skilled technicians and mitigate turnover.

V. FEATURE IMPORTANCE AND TRAIN MODEL

A). Make a dataset into training and testing data: To prepare data for machine learning, 'Data Frame' was separated into two subsets:

Train and Test. The Train set was used to train the ML model, and 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, thetrain data should be around 70-85% of the dataset. In particular case, the train data consists of 75% of the 'Data Frame', i.e., 1249 rows, where other15% or 221 rows are from test data.

B) Feature Importance: In machine learning, feature importance is the process of determining the relative importance of different input variables, 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 is helpful tooptimize the performance of the algorithm by focusing on the most important features and reducing the dimensionality of the data.

C) SMOTE:

When one class predominates over another in a situation like predicting loan acceptance, SMOTE is an essential technique for resolving class imbalance in datasets. SMOTE helps balance the dataset, reducing bias towards the majority class and enhancing model performance by creating synthetic examples for the minority class. In our example, the class imbalance bias is successfully reduced by using SMOTE, resulting in a more representative and trustworthy dataset that we can use to train our machine learning models. By using this method, the model's capacity to generalize across both groups is improved, which eventually leads to forecasts of loan acceptance that are more fair and accurate.

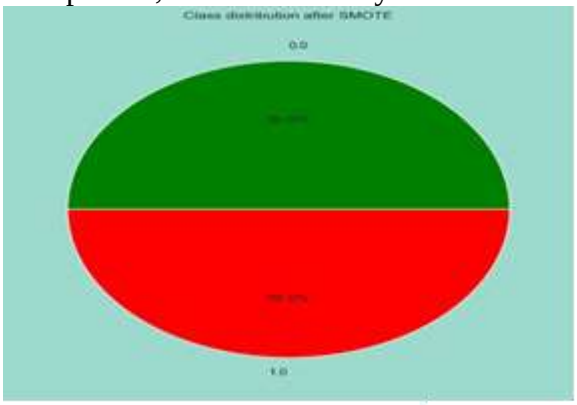


Fig.11 attrition status after smote

D) 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.

$IG(t) = 1 - \sum (i =1)^c p(i \mid t)^2$

2) Extra Tree Classifier: Instead of selecting the best split, Extra Trees randomly select splits at each node, leading to faster training and reduced variance. The key formula for prediction is the same as

Random Forest. 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.

3).Logistic Regression: This 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.

4).Random Forest: It is used for both classification and regression tasks. It's an ensemble learning method, meaning it builds multiple decision trees during training and merges them together to get a more accurate and stable prediction. The key formula involves aggregating predictions from individual trees:

$\hat{y}\_RF(x) = 1/N\_trees \sum (i=1)^{(N\_trees)} f\_i(x)$

5). Gradient Boost Classifier: Gradient Boosting is an ensemble learning technique that builds models sequentially, where each model corrects the errors of its predecessor. It minimizes a loss function using gradient descent. The key formula involves updating the model's predictions based on the gradient of the loss function:

$$L = \frac{1}{n}\sum_{i=0}^{n}\left(y_i - \gamma_i\right)^2$$

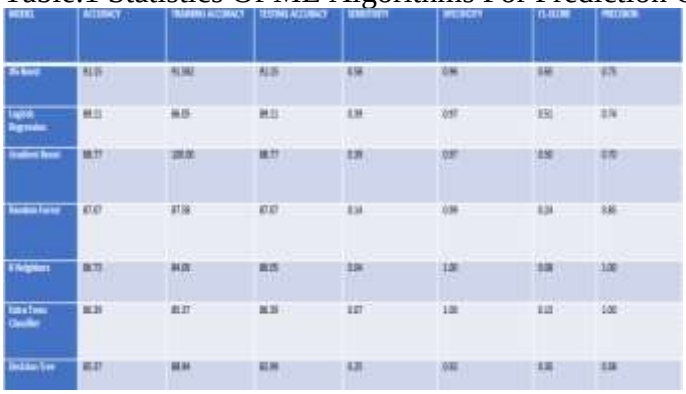
6). K-Nearest Neighbors (KNN): It is a non- parametric, lazy learning model that classifies data points based on the majority class of their k nearest neighbors. The key formula for prediction involves calculating the distance between data points and selecting the majority class among the k nearest neighbors.

7) XG Boost Classifier: XG Boost is a powerful gradient boosting classifier used for supervised learning tasks, particularly for classification and regression. It sequentially builds a series of decision trees, each attempting to correct the errors of the previous tree.

E) Result Analysis:

We found that the models that were used are Random Forest, Extra Trees, XG Boost Classifier Decision Trees, Gradient Boosting, Logistic Regression, and K-Nearest Neighbors performed differently. Every algorithm performed differently, as shown below:

Table.1 Statistics Of ML Algorithms For Prediction Of Employee Attrition



From table1, we depicts that It is clear from the given data that the XG Boost algorithm outperformed all other models in terms of accuracy, coming in at 91.15%. The discrepancy between the test and training accuracies, however, highlights the fact that it also shows the largest overfitting. Gradient Boost and Logistic Regression come a close second and third, with accuracies 88.77% and 89.11%, respectively. The ability of the Random Forest model to accurately categorize non-attrition instances is demonstrated by its high specificity of 0.99, while the best specificity among the top models is exhibited by Logistic Regression. In comparison to other models, K Neighbors and Extra Trees

Classifier perform relatively poorly, particularly when it comes to sensitivity and F1-score. Overall, XG Boost stands out as the top performer in the table, which offers a clear comparison of several machine learning algorithms in forecasting staff attrition.

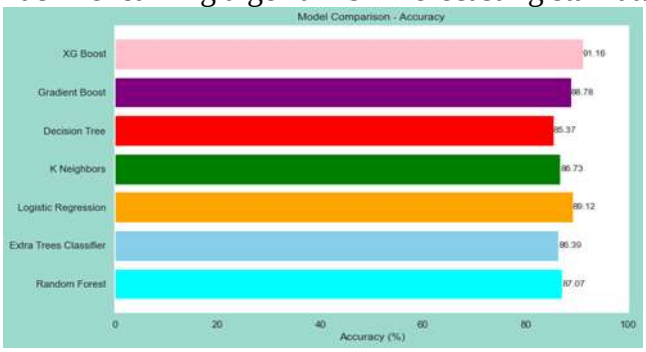


Fig.12 Performance Analysis of ML Algorithms

The graph plotted in Figure 12; using the same values as given in Table-1 illustrates the accuracy comparison of various machine learning models for predicting employee attrition. XG Boost stands in top with an accuracy 91.16%, followed by Gradient Boost at 88.77%. Decision Tree and K Neighbours demonstrate comparable accuracies around 85%, while Logistic Regression and Extra Trees Classifier exhibit accuracies exceeding 80%. Random Forest, although slightly lower in accuracy compared to the top models, still achieves a respectable accuracy of 87.07%. This visualization underscores the effectiveness of XG Boost and Gradient Boost in predicting employee attrition, offering valuable insights for workforce management strategies.

VI. CONCLUSION AND FUTURE SCOPE

The paper describes the impact of voluntary attrition on organizations and underscored the significance of accurately predicting. Additionally, it outlined several classification algorithms rooted in supervised learning as potential solutions to address the prediction challenge.

The findings of the conducted study revealed that several key factors contribute to employee attrition, including age, salary, total working years, tenure with current manager, previous employers, salary hike, Marital Status, Job Level and lack of Work-Life balance. The continuous turnover of employees within an organization can lead to increased hiring costs, reduced productivity, and the loss of valuable knowledge. The study emphasizes the necessity for organizations to focus on enhancing their Human Resources department to address the issue effectively. To achieve this, it is imperative to assess factors such as job satisfaction, workload, employee- manager interaction, and the working environment.

In Conclusion, it is evident that XG Boost emerges as the most accurate model for predicting attrition rate, achieving an impressive accuracy 91.16%. This suggests that organizations looking to optimize their attrition prediction systems should prioritize the implementation of XG Boost. However, it's worth noting that Gradient Boost also performs strongly with an accuracy of 88.77%, making it a viable alternative for companies seeking high accuracy in attrition prediction. Ultimately, leveraging advanced machine learning algorithms like XG Boost or Gradient Boost can significantly enhance workforce management strategies by providing more accurate insights into employee attrition patterns.

Further research could explore the integration of these models with additional data sources, such as employee sentiment analysis or external market trends, to improve predictive accuracy and identify early warning signs of attrition. Additionally, the development of interpretability techniques for these complex models could facilitate better understanding and actionable insights for organizational decision-makers. Overall, the future scope involves continuous innovation and refinement of predictive models to better support workforce management and employee retention efforts.

VII. REFERENCES

[1] Dataset for employee attrition is available at - https://www.kaggle.com/datasets/pavansubhasht/ibm- hr-analytics-attrition-dataset

[2] Ponnuru, S.; Merugumala, G.; Padigala, S.; Vanga, R.; Kantapalli, B.(2020) Employee Attrition Prediction using Logistic Regression. Int. J. Res. Appl. Sci. Eng. Technol. 8, 2871–2875.

[3] Srivastava, D. K., & Nair, P. (2017, March). Employee attrition analysis using predictive techniques. In International Conference on Information and Communication Technology for Intelligent Systems (pp. 293-300). Springer, Cham.

[4] Alao, D. A. B. A., and A. B. Adeyemo (2013). "Analyzing employee attrition using decision tree algorithms." Computing, Information Systems, Development Informatics and Allied Research Journal 4.1 :17- 28.

[5] Yadav, Sandeep, Aman Jain, and Deepti Singh. "Early Prediction of Employee Attrition using Data Mining Techniques." (2018) IEEE 8th International Advance Computing Conference (IACC). IEEE.

[6] Bhuva, K., & Srivastava, K. (2018). Comparative Study of the Machine Learning Techniques for Predicting the Employee Attrition. IJRAR- International Journal of Research and Analytical Reviews (IJRAR), 5(3), 568-577.

[7] Cotton, J.L. and Tuttle, J.M., 1986. “Employee turnover: A meta-analysis and review with implications for research” Academy of management review, pp.55-70.

[8] 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.

[9] Isabelle Guyon, Andre Elisseeff, “An Introduction to Variable and Feature Selection”, Journal of Machine Learning Research 3 (2003) 1157-1182.

[10] Ilan Reinstein, “Random Forest(r), Explained”, kdnuggets.com, October 2017[Online]. Available: https://www.kdnuggets.com/2017/10/randomforests- explained.html

[11] http://scikitlearn.org/stable/modules/generated/sklearn .metrics.confusion_matrix.html

[12] http://scikitlearn.org/stable/auto_examples/model_sele ction/ plot_confusion_matrix.html

[13] 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

[14] Pavan Subhash, “IBM HR Analytics Employee Attrition & Performance”, www.kaggle.com,2016[Online].Available:https://ww w.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

[15] 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