IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

EMPLOYEE ATTRITION CLASSIFICATION AND ANALYSIS USING MACHINE LEARNING APPROACH

[1]Prof. Rupali M. Bora, [2]Nikita Ughade, [3]Janhavy Bhalerao, [4]Ashish Deshpande and

[5]Aqusa Tabassum Syed Sadique Ali

[1]Assistant Professor, Department of Information Technolgy

[2,3,4,5] Students, Department of Information Technolgy

K. K. Wagh Institute of Engineering Education and Research, Nashik, India Affiliated under Savitribai Phule Pune University

Abstract: Employee attrition poses a significant challenge for organizations, impacting productivity, morale, and overall success. Predicting and mitigating attrition can be achieved through the application of machine learning techniques, leveraging the power of data-driven insights. This research paper explores the development of an attrition prediction model using machine learning algorithms, aiming to identify influential factors and provide actionable recommendations for talent management. A comprehensive dataset of employee information, encompassing demographics, job-related factors, performance metrics, and attrition status, was collected. Various machine learning algorithms, including random forests, XGBoost, Naive Bayes and Decision tree, were employed to develop attrition prediction models. The performance of these models was evaluated using accuracy, precision, recall, and F1 score. The results reveal significant predictors of attrition and provide valuable insights for organizations to proactively manage employee retention. This study contributes to the understanding of talent management by harnessing the potential of machine learning, enabling organizations to anticipate and address attrition risks effectively. The findings serve as a foundation for evidence-based decision-making and the development of tailored retention strategies. The outcomes of this research have implications for enhancing workforce stability and optimizing organizational performance. Future research directions include the exploration of advanced machine learning techniques, incorporation of additional data sources, and validation of the developed models across diverse industries and organizational contexts.

Index Terms **- Employee Attrition, Machine Learning, Gaussian Naive Bayes Classifier, Decision tree, XGBoost, Random Forest (RF), accuracy, precision, recall, and F1 score.**

I. INTRODUCTION

Employee attrition prediction using machine learning refers to the process of using data and machine learning algorithms to predict which employees are likely to leave an organization in the near future. This process involves analyzing various factors such as job satisfaction, work life balance, Environment satisfaction and Years since last promotion to identify patterns and trends that are associated with employee turnover.

By predicting which employees are likely to leave, organizations can take proactive measures to retain these employees, such as offering incentives, improving job satisfaction, and providing opportunities for career growth. This can help reduce the negative impacts of employee attrition on productivity, morale, and the bottom line. To predict employee attrition using machine learning, organizations can collect and analyze data from various sources such as employee surveys, performance evaluations, and HR records. Machine learning

algorithms can then be trained on this data to identify patterns and trends associated with employee attrition. Common machine learning algorithms used for employee attrition prediction include

1. Gaussian Nave Bayes Classifier
2. Decision Tree
3. XGBoost
4. Random Forrest (RF)

These algorithms can be used to create predictive models that assign a probability of employee attrition to each employee based on their individual characteristics and factors associated with employee turnover. Overall, employee attrition prediction using machine learning is a powerful tool that can help organizations take proactive measures to retain top talent and reduce the negative impacts of employee turnover. By leveraging the power of data and machine learning algorithms, organizations can gain insights into their workforce that were previously unavailable, leading to more effective workforce management and improved business outcomes.

II. LITERATURE REVIEW

There are several related works that have been done for employee attrition prediction. Some of the notable ones are:

1. *" A Hybrid Decision Tree-Based Ensemble Model for Predicting Employee Attrition"* by Ehsan Zarei and Mohd Fairuz Shiratuddin (2019)
2. " Predicting Employee Attrition using Decision Tree Algorithms" by Lijin John andArun John (2018)
3. *" An Empirical Study of Employee Turnover Predictive Models in IT Industry"* by Anusha M.K. and Shashidhar Kini (2017)
4. *" Predicting Employee Turnover using Machine Learning Techniques"* by Rhea Malviya and S. S. Bhattacharyya (2016)
5. *" Employee Attrition Prediction using Machine Learning Algorithms"* by Hemanth Kumar G. and Suresh Kumar K. (2015)

These works have used various machine learning algorithms such as decision trees, logistic regression, random forests, and support vector machines to predict employee attrition. They have also used different sets of features such as demographic data, job-related data, and performance data to build their models. The accuracy of the models reported in these works range from 75-95 percent. These related works provide valuable insights into the application of machine learning for employee attrition prediction and can be used as a reference for future research in this area.

III. STEPS FOR IMPLEMENTATION

Step 1: Data collection

Gathering of relevant employee data, here we have used the IBM HR analytics dataset [1] which includes features such as Age, Gender, Distance from home, monthly income etc.

Step 2: Data pre-processing

Cleaning the collected data, handle missing values, transform and normalize features to ensure consistency and prepare the data for analysis. Encode categorical variables into numerical representations.

Step 3: Feature engineering and selection

Deriving new features or select relevant features that are likely to be significant predictors of employee attrition. It is done by using domain knowledge and data exploration techniques to identify meaningful patterns and create informative features. Here, we have univariate feature selection using SelectKBest [2] and Recursive feature Selection techniques [3].

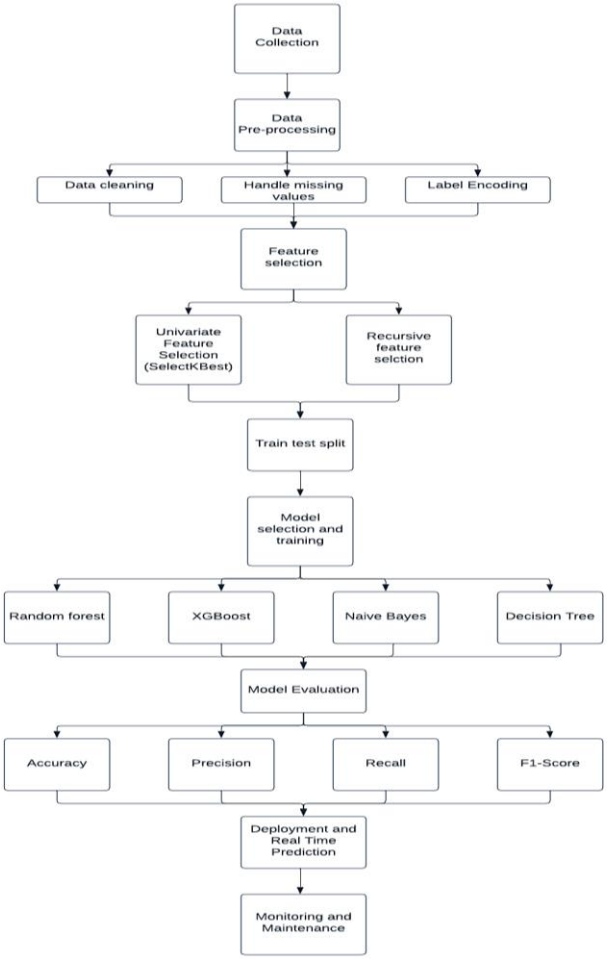

Fig.1. System Architecture

Step 4: Training and test data split

Splitting of the pre-processed dataset into training and test sets. The training set is used to train the machine learning models, while the test set is used to evaluate their performance.

Step 5: Model selection and training

Choosing of an appropriate machine learning algorithm for the attrition prediction task, such as random forests, XGBoost, Naïve Bayes and Decision Tree. Followed by training of chosen models using the training data and tune their hyperparameters to optimize performance.

Step 6: Model evaluation

1. Evaluation of the trained models using appropriate evaluation metrics, such as accuracy, precision, recall and F1 score.
2. Assessment of the models' performance on the test set to gauge their ability to generalize to unseen data.

Step 7: Deployment and real-time prediction

1. Deployment of the trained model into a production environment where it can make real-time attrition predictions.
2. Integration the model with the front end, to incorporate new employee data and generate predictions.

Step 8: Monitoring and maintenance

1. Continuous monitoring of the deployed model's performance, track its predictions, and assess its accuracy and reliability over time.
2. Regularly retrain the model using updated data to adapt to changing patterns and maintain its effectiveness.
3. Conducting periodic model audits and updates to ensure the model remains relevant and aligned with the organization's goals and evolving data.

IV. DATA ANALYSIS AND VISUALIZATION

1. Attrition pie chart

Fig.2. Attrition pie chart

Fig.2. is a pie chart of employee attrition. **"plt.pie()"** is a matplotlib library in python which is used to create this chart. The number of employees who left the company (Attrition = Yes) are represented by the blue color and the number of employees who stayed (Attrition = No) are represented by orange color in the above pie chart. **"explode=(0.2,0)"** parameter how far each slice of the pie chart is pulled away from the center. In this case, the first slice (No) is pulled away from the center by 0.2, while the second slice (Yes) is not pulled away from the center at all.

2.Gender wise attrition

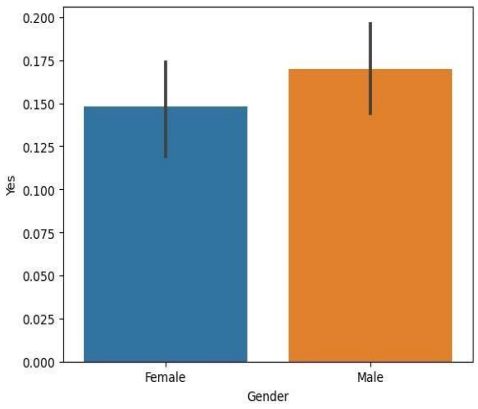

Fig.3. Gender wise attrition bar plot

Fig.3. graph is created using "**sns.barplot()"** which is a function provided by the seaborn library that creates a bar plot. **x='Gender'** specifies the variable to be plotted on the x-axis. In this case, it is the Gender variable, which is a categorical variable with two levels (Male and Female). **y='Yes'** specifies the variable to be plotted on the y-axis. In this case, it is the number of employees who left the company (Attrition = Yes). This bar plot compares the number of male and female employees who left the company. It can be a useful way to visualize the gender imbalance in attrition rates and identify any patterns or trends in the data.

3. Age wise attrition bar plot

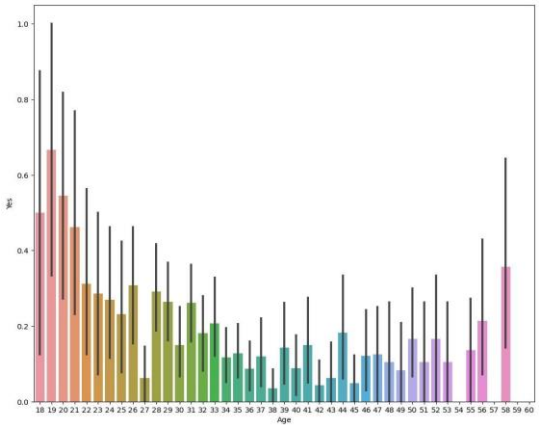

Fig.4. Age wise attrition

Fig.4. bar plot is created using seaborn's barplot function to visualize the relationship between age and a binary variable 'Yes' (which means the employees which were attrited) in the dataset 'df'. Here 'Age' is a continuous variable and 'Yes' is a binary variable. From the above visualization we can see that the employee below the age 35 have more attrition rate.

V. DATA ANALYSIS AND VISUALIZATION

1. Correlation matrix

	Age	Attrition	BusinessTravel	DailyRate
Age	1.000000	-0.159205	0.024751	0.010661
Attrition	-0.159205	1.000000	0.000074	-0.056652
BusinessTravel	0.024751	0.000074	1.000000	-0.004086
DailyRate	0.010661	-0.056652	-0.004086	1.000000
Department	-0.031882	0.063991	-0.009044	0.007109
DistanceFromHome	-0.001686	0.077924	-0.024469	-0.004985
Education	0.208034	-0.031373	0.000757	-0.016806
EducationField	-0.040873	0.026846	0.023724	0.037709
EmployeeNumber	-0.010145	-0.010577	-0.015578	-0.050990
EnvironmentSatisfaction	0.010146	-0.103369	0.004174	0.018355
Gender	-0.036311	0.029453	-0.032981	-0.011716

Fig.5. Correlation matrix

In Fig.5. a correlation matrix is used for the analysis of employee attrition to identify the relationships between various factors and the likelihood of employees leaving the organization. Some of the factors that can be included in the analysis are:

1. Demographic factors such as age, gender and education
2. Work-related factors such as job satisfaction, job performance, Overtime, and job involvement
3. Compensation and benefits factors such as salary, bonuses, and promotions
4. Career development and growth opportunities

By analyzing the correlation matrix, it is possible to identify the factors that are most strongly correlated with employee attrition. For example, if the correlation matrix shows a strong negative correlation between job satisfaction and attrition, it suggests that employees who are dissatisfied with their jobs are more likely to leave the organization. By identifying these factors, organizations can take steps to address them and reduce the rate of employee attrition.

2. Heatmap

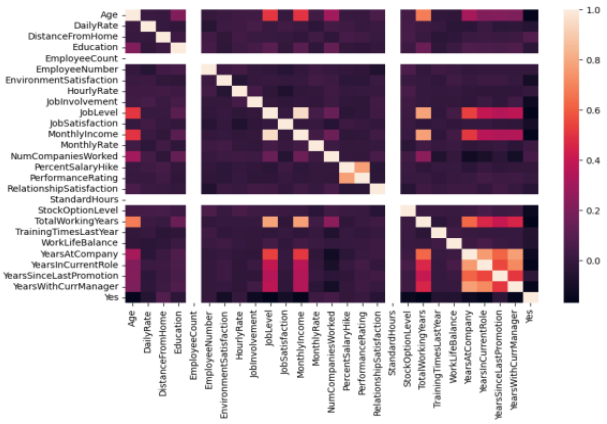

Fig.6. Heatmap

Fig.6. is a heatmap created using seaborn's "**heatmap**()" function to visualize the correlation matrix of the DataFrame 'df'. "**plt.figure()"** function is used to adjust the size of the figure to a width of 10 inches and a height of 6 inches. The **"heatmap()"** function takes the correlation matrix of the DataFrame as input and plots it as a heatmap, where each cell represents the correlation between two variables. The correlation coefficient ranges from -1 to 1, with -1 indicating a perfect negative correlation, 0 indicating no correlation, and 1 indicating a perfect positive correlation. By default, the **"heatmap()" function** uses a color map to visualize the correlation values, where darker colours indicate higher correlation values and lighter colours indicate lower correlation values. Heatmap is a useful way to quickly visualize the correlation matrix of a DataFrame and identify relationships between variables.

VI. MODEL BUILDING

1. Feature selection

Once the data pre-processing is done, a parameter set by using feature selection techniques is created. These parameters are clubbed under a set and displayed to the user. The features are selected using the **Univariate Selection** and **Recursive Feature Selection (RFE)** techniques respectively.

2. Model selection

A suitable machine learning model for employee attrition prediction is chosen. Some popular models for binary classification tasks like this include Random Forest (RF), XGBoost, Gaussian Naive Bayes and Decision Tree are used.

3. Model training and evaluation

Train the model on the training data and evaluate its performance on the test data. Use metrics such as *accuracy, precision, recall, F1-Score* to evaluate the model's performance and to get a more robust estimation of the model's performance. Here is the tabular representation of the performance metrics for Univariate using SelectKBest and Recursive feature selection.

4. Comparative analysis of models based on performance metrics

Comparative analysis of models based on performance metrics involves evaluating and comparing different models to determine their effectiveness for a specific task. It includes defining relevant performance metrics, preparing the data, selecting and training the models, evaluating their performance on test data, conducting statistical analysis, and making informed model selections based on the results. The analysis helps identify the best-performing models and guides further iterations or improvements as needed.

4.1 Univariate feature selection using SelectKBest

Table.1 Performance metrics for Univariate feature selection using SelectKBest

Sr No	Model	Accuracy	Precision	Recall	F1 Score
1.	Random Forest	**85.03%**	**82.35%**	25.45%	38.88%
2.	XG Boost	84.35%	65.51%	34.54%	45.23%
3.	Naïve Bayes	83.67%	57.14%	**50.90%**	**53.84%**
4.	Decision Tree	75.17%	37.17%	47.27%	41.60%

Table.1 summarizes the performance metrics for Univariate feature selection using SelectKBest. Here, is a brief description of it:

1. **Accuracy Comparison:** The Random Forest model achieves the highest accuracy (85.03%), followed closely by XG Boost (84.35%) and Naïve Bayes (83.67%). The Decision Tree model has the lowest accuracy at 75.17%.
2. **Precision Comparison:** The Random Forest model has the highest precision (82.35%), indicating a high proportion of correct positive predictions. XG Boost (65.51%) and Naïve Bayes (57.14%) also show reasonably good precision, while the Decision Tree model has the lowest precision (37.14%).
3. **Recall Comparison:** Naïve Bayes has the highest recall (50.90%), indicating its ability to identify a significant proportion of actual positive instances. The Decision Tree model (47.27%) and XG Boost (34.54%) also show moderate recall rates. The Random Forest model has the lowest recall (25.45%).
4. **F1 Score Comparison:** Naïve e Bayes achieves the highest F1 score (53.84%), which balances precision and recall. XG Boost (45.23%) and the Decision Tree model (41.60%) have moderately lower F1 scores, while the Random Forest model has the lowest F1 score (38.88%).

In summary, the Random Forest model performs well in terms of accuracy and precision but shows lower recall compared to other models. Naïve Bayes achieves a good balance between precision, recall, and F1 score. The Decision Tree model has relatively lower performance in most metrics. Considering the specific requirements and trade-offs Random Forest is the most suitable model for the employee attrition prediction system

4.2 Recursive Feature Selection

Table.2 Performance metrics for Recursive Feature Selection (RFE)

Sr No	Model	Accuracy	Precision	Recall	F1 Score
1.	Random Forest	**87.41%**	**60%**	15.38%	24.48%
2.	XG Boost	87.07%	52.17%	30.76%	38.70%
3.	Naïve Bayes	86.05%	46.42%	**33.33%**	**38.80%**
4.	Decision Tree	80.61%	29.54%	**33.33%**	31.32%

Table.2 summarizes the performance metrics for Recursive Feature Selection. Here, is a brief description of it:

1. **Accuracy Comparison:** The Random Forest model achieves the highest accuracy of 87.41%, followed closely by XG Boost at 87.07% and Naïve Bayes at 86.05%. The Decision Tree model has the lowest accuracy at 80.61%.
2. **Precision Comparison:** Random Forest has the highest precision at 60%, indicating a relatively high proportion of correct positive predictions. XG Boost follows with a precision of 52.17%, while Naïve Bayes has a precision of 46.42%. The Decision Tree model has the lowest precision at 29.54%.
3. **Recall Comparison:** Naïve Bayes has the highest recall at 33.33%, indicating its ability to identify a portion of the actual positive instances. XG Boost follows closely with a recall of 30.76%. The Decision Tree model and Random Forest both have a recall of 33.33%.
4. **F1 Score Comparison:** Naïve Bayes achieves the highest F1 score of 38.80%, which balances precision and recall. XG Boost follows with an F1 score of 38.70%. The Decision Tree model has the lowest F1 score at 31.32%, while Random Forest achieves an F1 score of 24.48%.

In summary, the Random Forest model performs well in terms of accuracy, but it has lower precision, recall, and F1 score compared to other models. Naïve Bayes and XG Boost show moderate performance across the metrics, with Naïve Bayes achieving the highest F1 score. The Decision Tree model has relatively lower performance in most metrics. Considering the specific requirements and trade-offs Random Forest is the most suitable model for the employee attrition prediction system.

Therefore, in accordance with both the feature selection techniques the following is the comparison in terms of accuracy:

Radom Forest > XG Boost > Naïve Bayes > Decision Tree

VII. CONCLUSION

After conducting a comprehensive analysis and developing a predictive model, the employee attrition aimed to understand and anticipate the factors contributing to employee turnover. Through data analysis and feature selection, influential variables were identified. Feature selection techniques used to identify the most influential features for employee attrition were univariate feature selection using SelectKBest and recursive feature selection (RFE). Followed by feature selection step, various machine learning models such as Random Forest, XGBoost, Naïve Bayes, and Decision Tree were trained and evaluated. To evaluate the performance of the developed model's, performance metrics such as the accuracy, precision, recall and F1-Score were used. Through the comparative analysis of the performance metrics of the models developed for both feature selection techniques it was observed that the Random Forest model achieved high prediction accuracy, outperforming baseline models i.e., 85.03% for univariate feature selection and 87.41% for recursive feature selection (RFE). Also, the Random Forest model has the highest precision i.e., 82.35% for univariate feature selection and 60% for recursive feature selection (RFE), indicating a high proportion of correct positive predictions. Hence the most suitable model for the employee attrition prediction system is Random Forest. In this paper key predictors of attrition were identified, providing actionable insights for HR departments. The findings enable organizations to implement proactive measures to reduce attrition rates and create a supportive work environment. We have successfully developed a robust predictive model for accurate attrition forecasting.

REFERENCES

[1] https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

[2] Univariate Feature Selection Techniques for Classification of Epileptic EEG Signals

[3] https://scikitlearn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

[4] A. Navada, A. N. Ansari, S. Patil and B. A. Sonkamble, "Overview of use of decision tree algorithms in machine learning," 2011 IEEE Control and System Graduate Research Colloquium, Shah Alam, Malaysia, 2011, pp. 37-42, doi: 10.1109/ICSGRC.2011.5991826.

[5] C. Jose and G. Gopakumar, "An Improved Random Forest Algorithm for classification in an imbalanced dataset," 2019 URSI Asia-Pacific Radio Science Conference (AP-RASC), New Delhi, India, 2019, pp. 1-4, doi: 10.23919/URSIAP-RASC.2019.8738232..

[6] L. Sun, "Application and Improvement of Xgboost Algorithm Based on Multiple Parameter Optimization Strategy," 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Harbin, China, 2020, pp. 1822-1825, doi: 10.1109/ICMCCE51767.2020.00400

[7] J. Ren, S. D. Lee, X. Chen, B. Kao, R. Cheng and D. Cheung, "Naive Bayes Classification of Uncertain Data," 2009 Ninth IEEE International Conference on Data Mining, Miami Beach, FL, USA, 2009, pp. 944-949, doi: 10.1109/ICDM.2009.90.