ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may172

Predicting Employee Attrition using Machine Learning Techniques

N. Bhavana[1]; Chukka Ganesh[2]

[1]Assistant Professor, [2]Student
[1,2]Depatment of MCA, Annamacharya Institute of Technology and Sciences, Karakambadi, Tirupati, Andhra Pradesh, India

Publication Date: 2025/05/07

Abstract: For businesses employee retention is a major issue, and forecasting attrition can assist HR departments to put in place proactive measures to lower turnover. Using methods including Random Forest, XGBoost, Decision Tree, Support Vector Classifier (SVC), Logistic Regression, KNearest Neighbors (KNN), and Naive Bayes, this project uses machine learning approaches to study important factors affecting employee departure. The model discovers trends in job satisfaction, workload, career development, and worklife balance trained on the IBM Analytics dataset with 35 characteristics and 1,500 records. Deployed as an interactive Flask based web application, the system includes capabilities for data upload, forecasting, and model performance visualization. This AI driven solution helps HR staff to find early at-risk employees, manage issues efficiently, and enhance staff stability by offering practical insights. By using predictive analytics in HR management, businesses can lower attrition expenses, improve staff engagement, and create a more resilient setting.

Keywords: *Employee Attrition Prediction, Machine Learning, Random Forest, XGBoost, Decision Tree, Support Vector Classifier (SVC), Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, Flask.*

How to Cite: N. Bhavana; Chukka Ganesh. (2025). Predicting Employee Attrition using Machine Learning Techniques. *International Journal of Innovative Science and Research Technology*, 10(5), 1-10. https://doi.org/10.38124/ijisrt/25may172.

I. INTRODUCTION

Employee attrition remains a critical issue for organizations, impacting productivity, operational efficiency, and overall workplace morale. High staff turnover not only raises costs for hiring and development but also interrupts workflow and stunts company expansion. Knowing the root causes of attrition and forecasting possible staff turnover enable companies to adopt forward retention initiatives. Traditional attrition analysis techniques often rely on manual evaluations and surveys, which are time-consuming, subjective, and less precise. Machine learning models using artificial intelligence and data analytics can provide a more precise and data driven method for spotting employees who might leave by using these resources.

Employee Housing Analysis Without Fail Basically it was based on a sample of 1500 employee records and 35 determinants. It included job satisfaction, work-life balance, paying scales, the amount of work to be done, career development opportunities, and engagement. The studies would determine the best possible predictive models that could recognize the highest influencers in the prediction models used to predict what factors facilitate defines. Multiple models can further enrich constructs of prediction by improving results using ensemble techniques.

The system is developed as a web-based application using Flask, offering an intuitive interface for HR professionals to upload employee data, analyze attrition trends, and visualize model performance. This tool enables organizations to identify employees at high risk of leaving and implement targeted interventions such as better compensation, flexible work arrangements, career growth opportunities, and employee engagement programs. By integrating machine learning into HR analytics, businesses can significantly reduce turnover rates, improve workforce retention, and enhance overall organizational stability. Future developments in this project may include incorporating deep learning techniques, real-time attrition monitoring, and expanding the model to accommodate industry-specific attrition patterns.

II. METHODOLOGY

The method suggested is in predicting the probable employee attrition using machine-based detection to predict the prospect of an employee leaving the organization based on other parameters considered for arriving at the retention capability of the proposals: job satisfaction, work-life balance, pay, career growth or advancement, and workload, among others. The major intention with this system would be toward making attempts to predict reasonably and counteract

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may172

the risks toward loss of workforce stability in the organization.

The system utilizes a range of machine learning methodologies to generate accurate employee turnover predictions based on HR data. It is designed to recognize patterns that may not be immediately visible, enhancing the ability to detect early signs of potential attrition. The prediction process involves training on historical employee records to identify signals indicative of future resignations. By leveraging multiple classification techniques, the system improves its predictive capability, allowing it to accurately identify employees who are most likely to leave. This approach supports proactive decision-making and helps organizations implement effective retention strategies.

Since employee turnover is today almost an expected precariousness, the human resources part needs more thorough attention. Employees leave for or migrate to other firms seeking opportunities or better terms. It is stated that keeping some current issues in mind, a new predictive approach to attrition has been suggested based on a folkloric inference of views based on what an average human is like in respect of income and wages. Large attrition types are a draw to organizations, forcing organizations to lay off more employees. As such, if the model works, it minimizes the cost of retaining active employees: the higher the turnover, the higher overall costs. Trying to improve this model will only yield cost savings by avoiding retaining employees.

A user-friendly web application is developed using HTML, CSS, and JavaScript, allowing HR professionals to enter employee data and receive real-time predictions. The platform is designed to offer insights into workforce retention trends, empowering organizations to take data-driven, proactive measures to reduce turnover. The system also enables companies to refine their HR policies, improve employee engagement, and implement targeted retention strategies.

By integrating machine learning into HR analytics, this system provides valuable insights for businesses, supporting early identification of attrition risks and enabling timely interventions. It helps organizations make informed decisions regarding employee retention, ultimately leading to a more stable, productive, and engaged workforce.

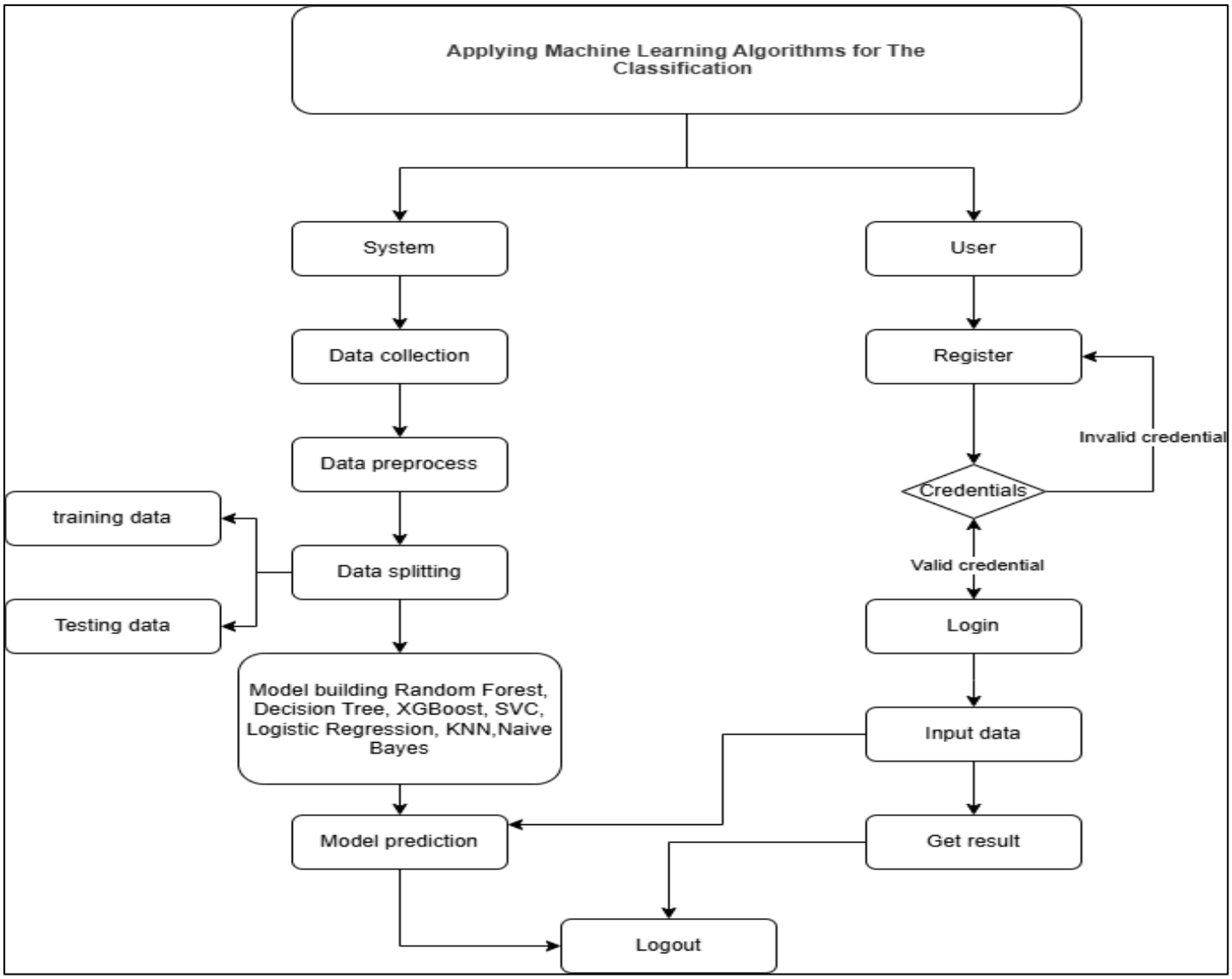

Fig 1: Flow Chart

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may172

III. MODULES AND ITS IMPLEMENTATION

A. System Operations

- **Upload Data**: HR professionals collect and upload a structured dataset containing various employee-related factors that influence attrition. This dataset includes details such as job roles, work experience, compensation, performance metrics, work-life balance indicators, and employee engagement levels. By gathering comprehensive data, the system can better analyze workforce trends and predict employee attrition accurately.
- **Data Preprocessing**: Once the dataset is uploaded, it undergoes a series of data cleaning and preprocessing procedures. This involves handling missing or corrupted data, encoding categorical variables such as department and job role, normalizing or standardizing numerical features like salary and experience, and applying techniques to balance the dataset to prevent bias. These preprocessing steps ensure that the machine learning models perform optimally and provide accurate predictions.
- **Model Building**: The system trains multiple machine learning models to classify employees as likely to stay or leave based on historical HR data. These models are fine-tuned through hyperparameter optimization to improve prediction accuracy and overall reliability. Model selection is guided by evaluating performance metrics such as accuracy, precision, recall, and F1 score, ensuring that the most effective model is chosen for employee turnover prediction
- **Model Prediction**: The trained models analyze new employee data using the same preprocessing techniques applied to the training data. Based on this analysis, the system predicts whether an employee is at risk of leaving the organization. The model's decision is based on historical trends and key influencing factors identified in the dataset.
- **Result**: The system presents the prediction results for each employee, along with confidence scores to indicate prediction reliability. Additionally, it provides detailed performance metrics, including confusion matrices, accuracy, precision, recall, and F1 score. Visual aids such as bar charts, histograms, and ROC curves are also incorporated to help HR professionals interpret the results more effectively.

B. User Operations

- **Register**: Users, primarily HR professionals, must first register with their credentials to create an account in the system.
- **Login**: Registered users can log in using their credentials to securely access the system and perform data analysis.
- **Upload Data**: Users can upload employee datasets containing relevant information about job satisfaction, salary, experience, and performance. The uploaded data should be in a structured format compatible with the system.
- **Model Page**: This section displays the accuracy of each machine learning model used in the system. Users can compare different models.
- **Prediction Page**: After uploading data, users can navigate to the prediction page, where they can view individual and overall attrition predictions.
- **Viewing Results**: Once the data is processed, users can view the classification results, including whether an employee is at risk of leaving.
- **Logout**: To ensure data security and privacy, users can log out of the system after completing their tasks, securing their session and personal data.

IV. MODELING AND ANALYSIS

A. Random Forest

It is ensemble algorithms that build many decision trees and prop them against random subsets of IBM Analytics consisting of 35 features (like job satisfaction, workload, and more). The outcomes are then aggregated through voting to classify the overall results so this reduces chances of overfitting while increasing accuracy. Using these feature importance scores, Random Forest here helps in inferring the more salient attrition factors like work-life balance. Also, due to its robustness against noisy data and handling of imbalanced classes, this classifier fits aptly for predicting the high-at-risk employees within the organization

B. XGBoost

It is an advanced ensemble algorithm that sequentially builds decision trees, optimizing a loss function with gradient boosting. For attrition prediction, it processes features like career growth and workload, weighting errors to improve accuracy. Its regularization prevents overfitting, and scalability handles the 1,500-record dataset efficiently. In the Flask app, XGBoost's high predictive power aids early identification of at-risk employees.

C. Decision Tree

It split the IBM dataset into branches based on features like job satisfaction or work-life balance, creating a flowchart-like model to predict attrition. Each node represents a decision, and leaves indicate outcomes (stay/leave). Their simplicity and interpretability help HR visualize attrition patterns. In the project, Decision Trees provide clear rules for identifying at-risk employees. However, they are prone to overfitting, especially with noisy data, leading to poor generalization. Pruning and limiting tree depth mitigate

D. Support Vector Classifier (SVC)

It finds the optimal hyperplane to separate employees who stay from those who leave, maximizing the margin between classes. For non-linear patterns in the dataset (e.g., complex interactions between career growth and workload), SVC uses kernels like RBF. In the project, SVC effectively classifies attrition risk but struggles with the dataset's size due to high computational costs. Scaling features is essential for performance.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may172

E. Logistic Regression

It predicts the probability of employee attrition by modeling the relationship between features (e.g., job satisfaction, work-life balance) and a binary outcome (stay/leave) using a logistic function. Its simplicity and interpretability make it ideal for HR to understand feature impacts via coefficients. In the project, it handles the 1,500-record dataset efficiently, providing baseline predictions for the Flask app. However, it assumes linear relationships, which may miss complex patterns. Regularization (e.g., L1, L2) prevents overfitting. Its fast training and deployment make it practical for real-time attrition risk assessment, supporting proactive HR interventions.

F. K-Nearest Neighbors (KNN)

it employees as likely to remain or likely to leave by considering the 'k' employees most similar to the ones in the dataset in question with respect to their workload, career growth, importance of training, and so forth. It uses distance metrics like Euclidean distance, Manhattan distance, and so on. KNN, therefore, captures the local patterns in attrition data, but it is sensitive to feature scaling and noise. The other aspect is the computational cost, which increases with the size of the dataset and, therefore, affects the performance of the Flask app. Careful consideration of choosing 'k' is imperative. KNN is slow and not storage-efficient, thus limiting its scalability. It provides assistance to HR in identifying employees vulnerable to leaving by offering insights through similarity detection.

G. Naive Bayes

It predicts attrition by calculating probabilities of staying or leaving based on features, assuming independence between them (e.g., job satisfaction and workload). Using Bayes' theorem, it's computationally efficient and excels with categorical data in the IBM dataset. In the Flask app, it provides fast predictions, ideal for real-time HR use. However, its independence assumption may oversimplify complex relationships, reducing accuracy. It performs well with imbalanced classes, common in attrition data. Its simplicity aids deployment but limits capturing intricate patterns. Naive Bayes supports HR by offering quick, interpretable insights for early intervention.

V. RESULTS AND DISCUSSION

The Employee Attrition Prediction System could very well prove helpful in predicting employees who are very likely to leave the company through the use of a variety of machine learning models, which ranges from distance-based classifiers to margin-based classifiers, tree-based methods, and even neural networks. The predicted reliabilities shall be lowered by implementing ensemble learning techniques like Voting Classifier and Stacking Classifier to combine individual merits to achieve better performance results.

A study with IBM HR Analytics Employee Attrition Data showed that job satisfaction, work-life balance, compensation, and career development opportunities affected employee retention critically. Similarly, these results were from earlier researches and showed the sweaty and multidimensional issues of turnover.

This makes it possible to design and deploy a simple user-friendly web application hosted on Flask, where all HR personnel can feed staff data and predictions in real-time. It also allows ad hoc capture of employees at risk of leaving and incentivized interventions. These outputs can also provide performance visualization and other services from the model, thus contextualizing model results, enabling data-informed decisions.

A. KNN Classifier

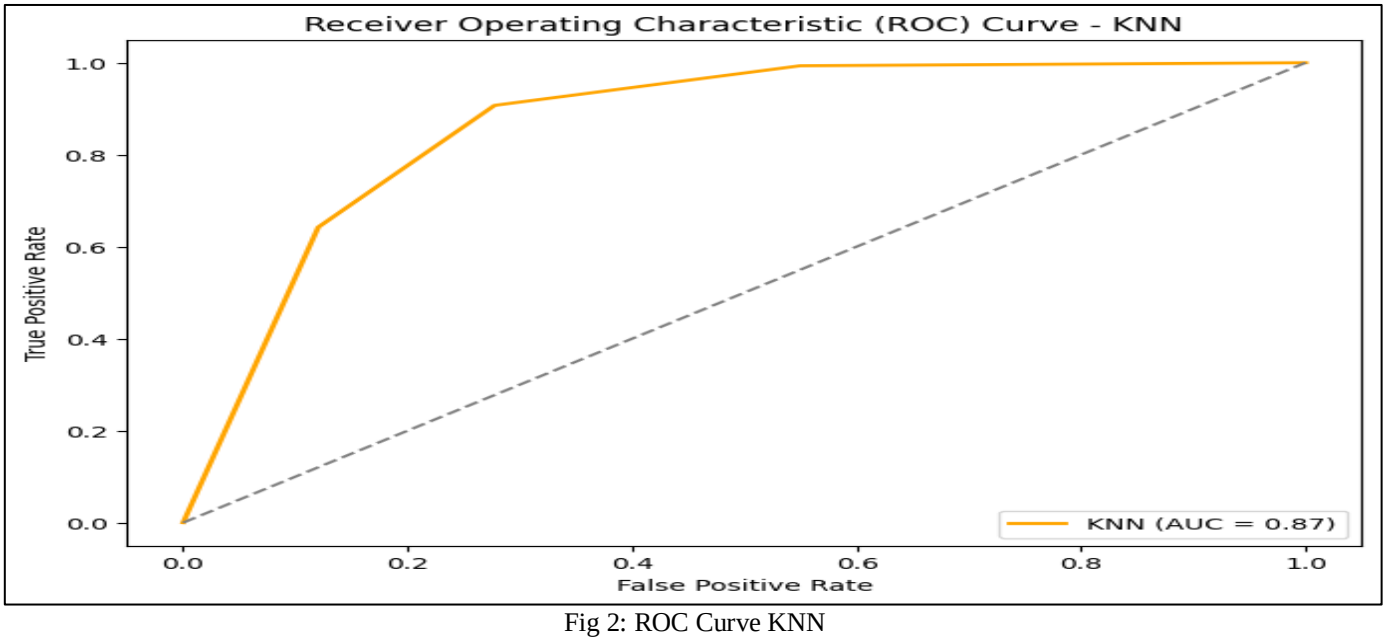

Fig 2: ROC Curve KNN

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may172

Table 1: Classification Report of KNN

	Precision	Recall	F1-Score	Support
0	0.91	0.72	0.80	191
1	0.72	0.91	0.80	151
Accuracy	0.80	0.80	0.80	342
Macro avg	0.81	0.81	0.80	342
Weighted avg	0.83	0.80	0.80	342

The KNN model's AUC value of 0.87 on the ROC curve indicates a good ability for distinguishing between employees prone to leave and those who are likely to stay. The classification report further states that the model has 80% overall accuracy. Class 1 indicates attrition with a very good recall of 0.91, thus correctly identifying employees at risk of leaving, though with low precision of merely 0.72, implying significant falses positives. Conversely, Class 0 refers to no-attrition with exceptionally high precision of 0.91 and lower recall of 0.72, meaning that it missed some no-attrition instances. Hence, with macro average and weighted average having a fixed score of 0.80, it shows close performance of the classifier for both classes.

B. SVM Classifier

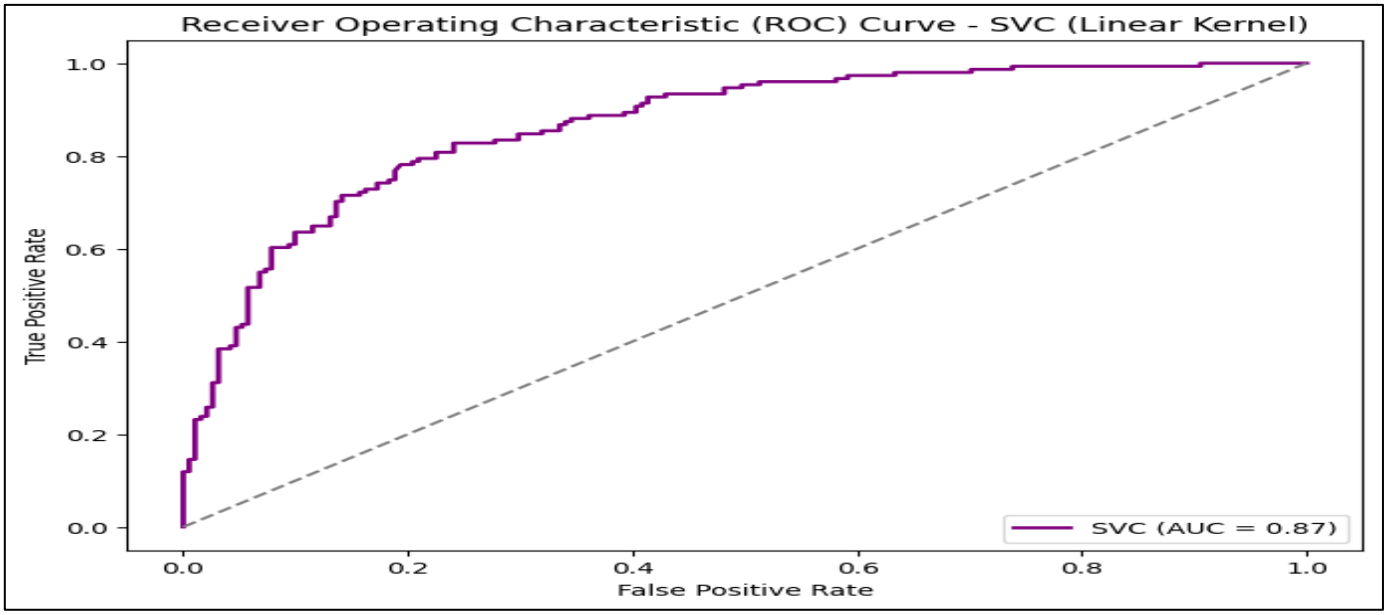

Fig 3: ROC Curve KNN

Table 2: Classification Report of SVM

	Precision	Recall	F1-Score	Support
0	0.82	0.81	0.81	191
1	0.76	0.77	0.77	151
Accuracy	0.79	0.79	0.79	342
Macro avg	0.79	0.79	0.79	342
Weighted avg	0.79	0.79	0.79	342

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may172

AUC 0.87 tells us that the model is working quite well and can pick out whether an employee is going to leave (class 1) or stay (class 0) with strong discrimination through the interpretation of the ROC curve for SVC (Linear Kernel). The classification report also states that the model has 79% accuracy, and class precision for attrition class 1 is at 0.76 with a recall of 0.77, indicating that it does pretty well identifying employees at risk of leaving. Class 0, instead, has better performance with a precision of 0.82 and lower risks of even less false positives in employee prediction who will stay. The two classes now have an equal F1-score of 0.79, which indicates how good the trade-off between precision and recall is. Thus this clearly leads to the reliability of the model in predicting employee attrition.

C. Decision Tree Classifier

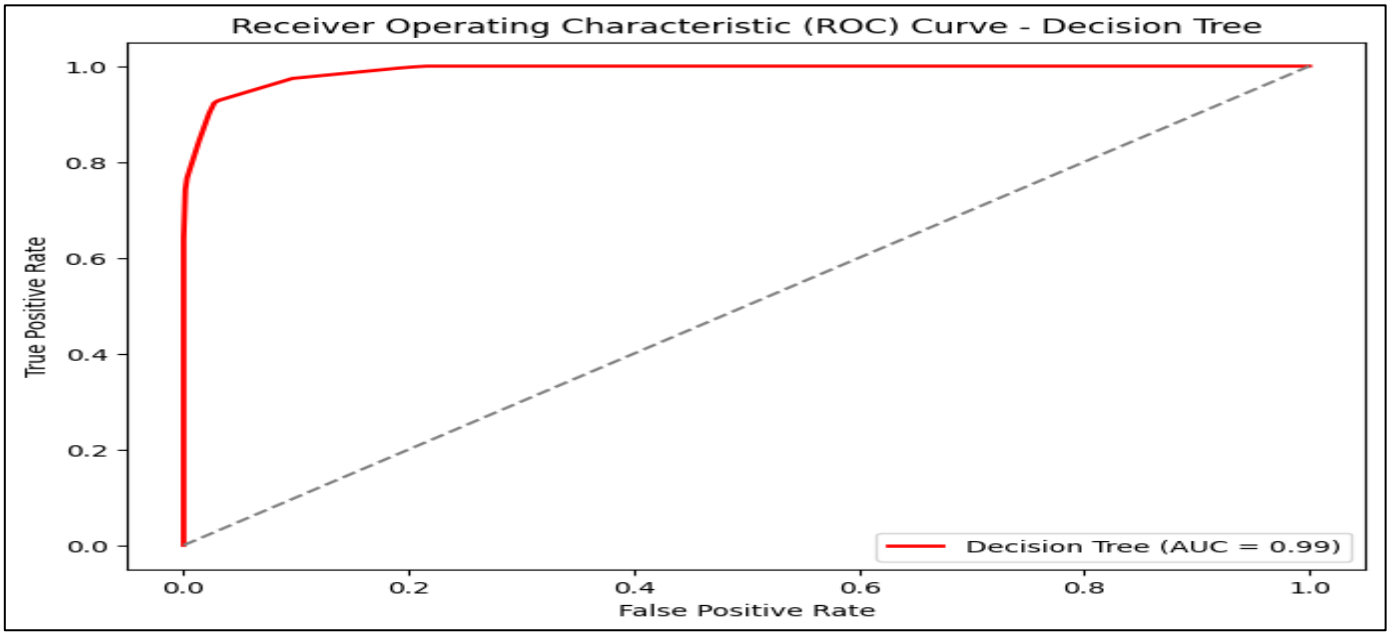

Fig 4: ROC Curve of Decision Tree

Table 3: Classification Report of Decision Tree

	Precision	Recall	F1-Score	Support
0	0.93	0.97	0.95	662
1	0.97	0.93	0.95	702
Accuracy	0.95	0.95	0.95	1364
Macro avg	0.95	0.95	0.95	1364
Weighted avg	0.95	0.95	0.95	1364

The model has shown good performance under this particular setup of the Decision Tree: AUC of the curve at 0.99, which stands for a near-perfect performance throughout the year, in telling apart the employees who are probably going to stay versus those who will likely leave. The subsequent classification report credits 95% total accuracy in performances; however, it must be mentioned that on precision, the classes diverge precision values (0.93 for staying employees and 0.97 for leaving employees), giving full credit to the model's capacity in distinguishing between the two classes. In contrast, the recall values appear strong too, being 0.97 for employees staying and 0.93 for employees leaving; hence, it probably captured the majority of true positives in each class. The F1-score value for both classes is also high at 0.95, thereby representing an optimal balance between precision and recall, making this model very robust in terms of its prediction on employee attrition.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may172

D. Random Forest Classifier

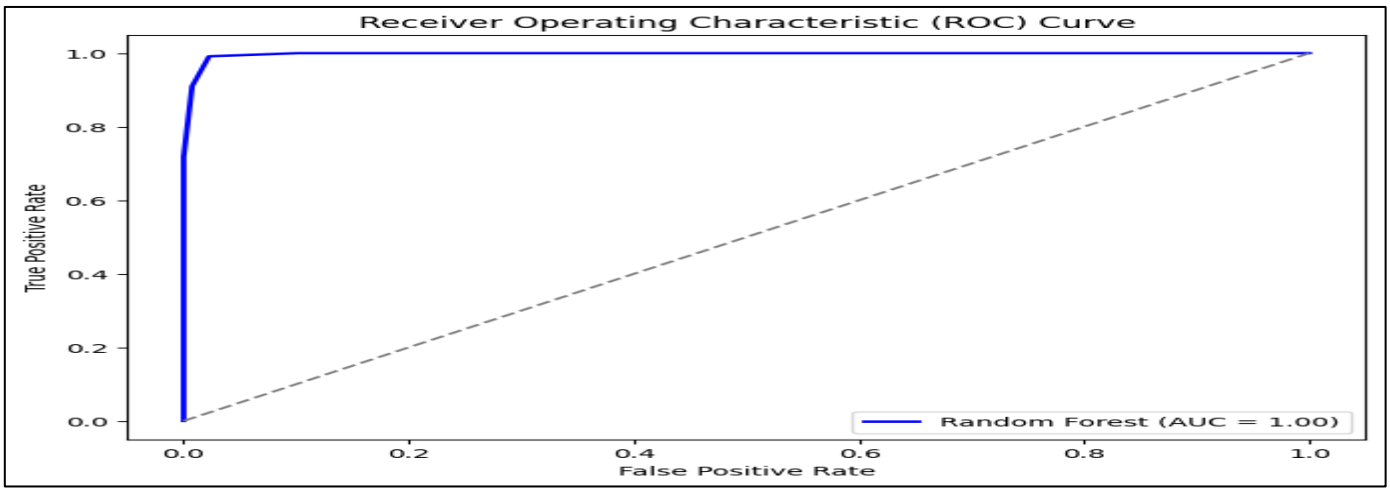

Fig 5: ROC Curve of Random Forest

Table 4: Classification Report of Random Forest

	Precision	Recall	F1-Score	Support
0	0.99	0.98	0.98	662
1	0.98	0.99	0.99	702
Accuracy	0.98	0.98	0.98	1364
Macro avg	0.98	0.98	0.98	1364
Weighted avg	0.98	0.98	0.98	1364

The model gives an AUC of 1.00 on the ROC curve, meaning it performs excellently by being able to discriminate positively between employees whose class is likely to be Class 1 (leaving) and whose class is likely to be Class 0 (staying). From the classification report, the model's accuracy is impressive at 98%. Both precision and recall for Class 0 were at 0.99 while the precision and recall for Class 1 calculated to 0.98 and 0.99, respectively. The F1-scores for both classes are very close to 0.99, showing an excellent balance between recall and precision. High scores through all the metrics highlight the strength of the model in predicting employee attrition, which can be used as a good tool in identifying at-risk employees and reducing turnover.

E. Logistic Regression

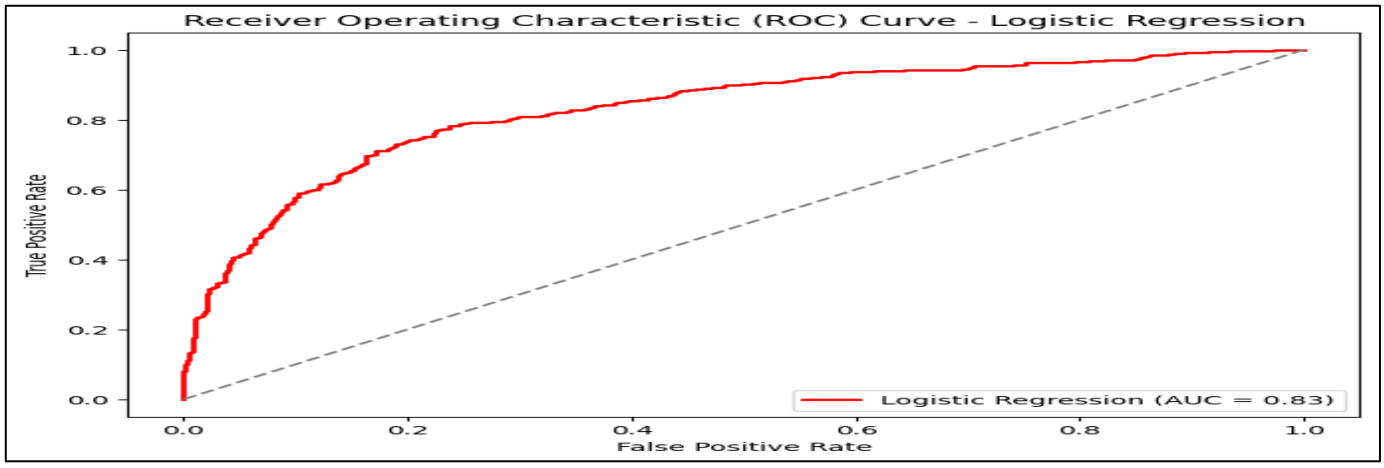

Fig 6: ROC Curve of logistic regression

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may172

Table 5: Classification Report of Logistic Regression

	Precision	Recall	F1-Score	Support
0	0.77	0.74	0.76	662
1	0.76	0.79	0.78	702
Accuracy	0.77	0.77	0.77	1364
Macro avg	0.77	0.77	0.77	1364
Weighted avg	0.77	0.77	0.77	1364

The Logistic Regression model, as depicted in an AUC score of 0.83 on the ROC curve, is moderately performing and indicates that the model has some level of effectiveness in classifying eventually leaving employees (class 1) as compared to the staying employees, class 0. However, it does not match that quality of performance demonstrated by other models bearing higher AUC values like Random Forest or Decision Tree. The classification report shows an accuracy of 77% and precision for class 1 (attrition) equals 0.76 and recall at 0.79, meaning the model can identify employees prone to quitting fairly well but has room for improvement regarding false positives. Precision for class 0 (i.e. no attrition) is 0.77 and recall is 0.74, indicating slightly better performance at predicting employees likely to remain. The resulting F1 of 0.77 means that the model does quite well on both classes such that the entire model can be favored as reasonably good at predicting attrition, but not into the same category of highly reliable models like Random Forest or Decision Tree.

F. Naïve Bayes

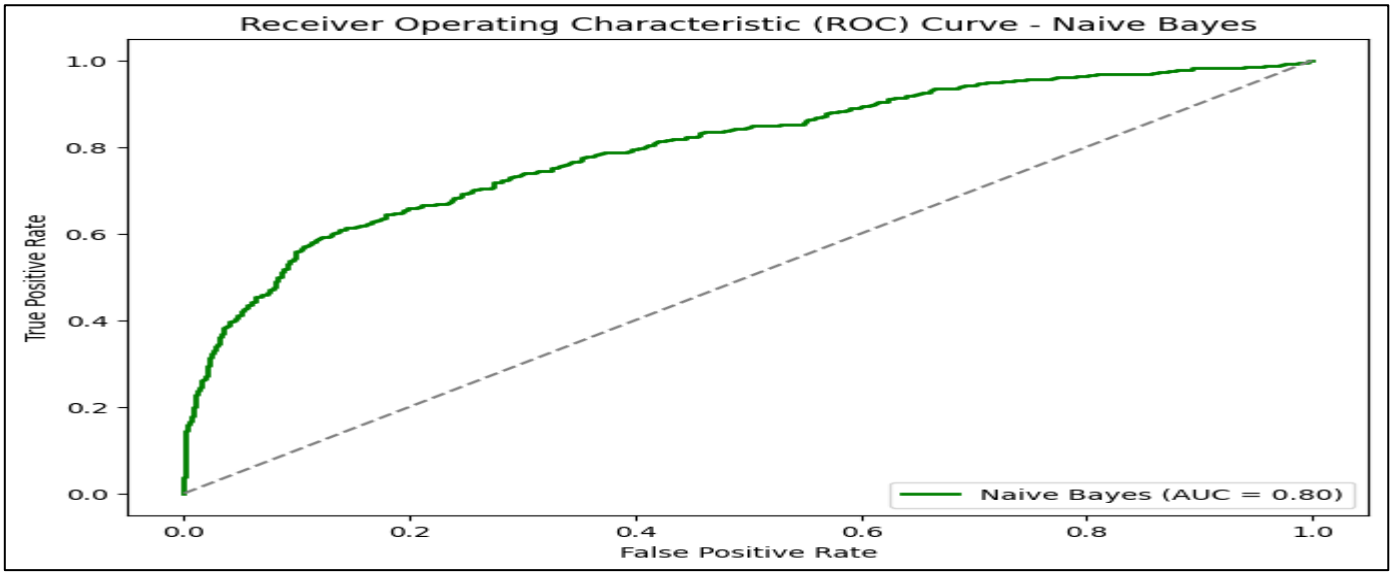

Fig 7: ROC Curve of Naïve Bayes

Table 6: Classification Report of Naïve Bayes

	Precision	Recall	F1-Score	Support
0	0.74	0.58	0.65	662
1	0.67	0.81	0.73	702
Accuracy	0.70	0.70	0.69	1364
Macro avg	0.71	0.69	0.69	1364
Weighted avg	0.70	0.70	0.69	1364

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may172

Modeling with Naive Bayes has moderate performance as shown in ROC, which gives an AUC of about 0.80. This convention indicates that the model can discriminate fairly well between those employees who will leave and those who will stay, although this ability is weaker than Random Forest or Decision Tree models. Among other things, the classification report mentions accuracy in class1 (attrition) around 70 percent, while precision equals 0.67, and recall is 0.81. It can be interpreted as being better at discerning employees likely to leave than keeping low errors in false positives. While the class 0 (no attrition) concession is at 0.74 and then compared to recall at 0.58 to say that the model finds it difficulty predicting those employees likely to stay in the organization. For both the classes, F1 scored 0.69, telling us that the model has a moderate balance between precision and recall yet could improve identifying both employees at risk of leaving and those likely to stay.

G. Xgboost

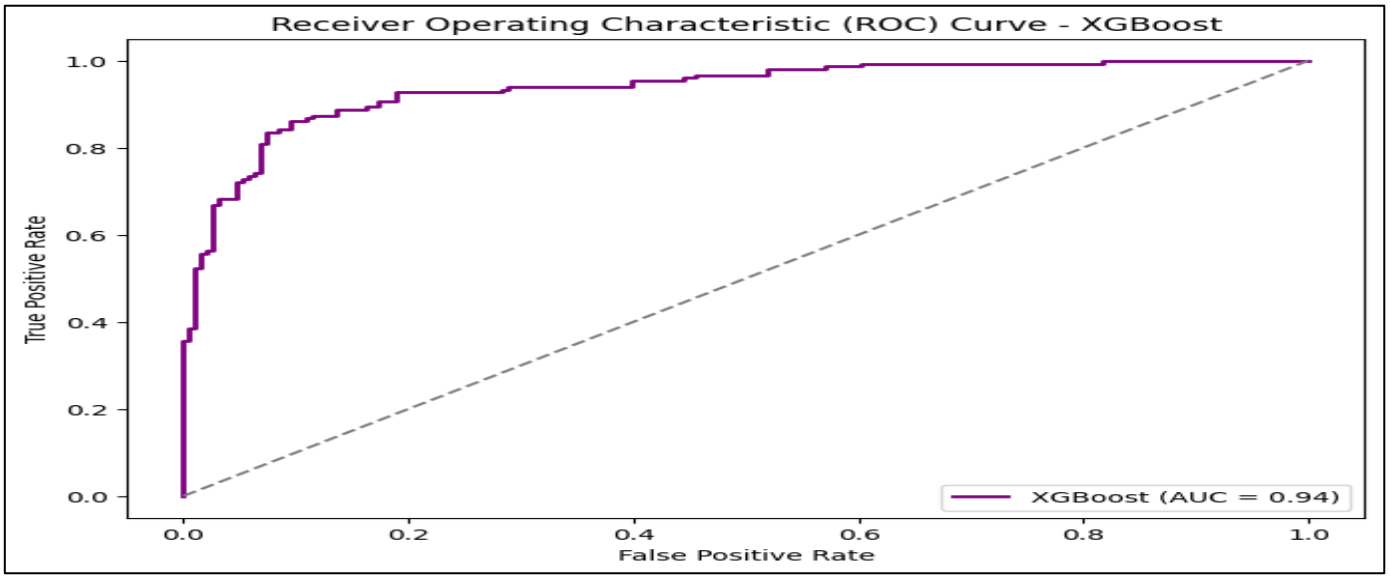

Fig 8: ROC Curve of Xgboost

Table 7: Classification Report of Xgboost

	Precision	Recall	F1-Score	Support
0	0.92	0.83	0.87	191
1	0.81	0.91	0.85	151
Accuracy	0.86	0.86	0.86	342
Macro avg	0.86	0.87	0.86	342
Weighted avg	0.87	0.86	0.86	342

It appears to be very strong with an AUC of 0.94 on the ROC curve, demonstrating an excellent ability to differentiate between leaving employees (class 1) and staying employees (class 0). In the classification report, the model presents an accuracy value of 86%, having class 0 (no attrition) precision at 0.92 and class 0 recall at 0.83, which means the model is good at predicting employees who will stay, and there exists further improvement to correctly identify all of them. For class 1 (attrition), precision is at 0.81 and recall is at 0.91, suggesting the model does very well in predicting employees at risk of leaving while reasonably maintaining its precision. Both classes hold up a high value for the F1 score, thus maintaining a balance across the performance of the model, which, hence, stands as a dependable predictive tool for employee attrition.

VI. CONCLUSION

The approaches demonstrated here explore machine learning's merit of predicting employee attrition using algorithms like Random Forest, XGBoost, Decision Tree, Support Vector Classifier (SVC), Logistic Regression, K-Nearest Neighbors (KNN), and Naive Bayes. It measures key parameters such as job satisfaction, performance, tenure, and demographic details to give sturdy predictions on employee turnover. This system uses Flask for a web-based approach and offers an interactive interface into which HR

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may172

professionals feed data upload, monitor model performance, and derive insights. Well-informed and timely, these insights give the HR team a forward-thinking approach to issues relating to workloads, job satisfaction, and career growth, which should affect retention strategy and hence workforce stability positively. This way, organizations empower themselves to make informed, data-driven decisions for more effective human resource management.

REFERENCES

[1]. R. L. Althoff et al., "Predicting employee attrition using machine learning techniques," *IEEE Transactions on Human Resources Management*, vol. 67, no. 8, pp. 2209-2215, Aug. 2020, doi: 10.1109/HRM.2020.2962935.

[2]. P. Kumar, S. S. Agarwal, and P. K. Jain, "Employee attrition classification using deep learning models," *Journal of Human Resource Management*, vol. 30, no. 3, pp. 1–9, May 2021, doi: 10.1111/hrm.13252.

[3]. M. Smith, J. A. Brown, and L. Harris, "Machine learning techniques for employee attrition prediction: A comparative study," *Proceedings of the International Conference on Workforce Analytics*, 2021, pp. 122–130, doi: 10.1109/WORKANA.2021.00027.

[4]. A. C. Mills et al., "Employee retention prediction using random forest and ensemble learning models," *IEEE Access*, vol. 8, pp. 174343-174352, 2020, doi: 10.1109/ACCESS.2020.3014506.

[5]. L. J. Robinson and R. B. Thompson, "Predictive modeling of employee turnover using machine learning algorithms," *Journal of Business Analytics & Human Resources*, vol. 11, no. 1, pp. 1–10, Jan. 2021, doi: 10.4172/2167-0277.1000281.

[6]. V. Singh and K. Gupta, "Prediction of employee attrition severity using machine learning algorithms," *Computational Intelligence and Human Resource Management*, vol. 2022, Article ID 791260, pp. 1–10, 2022, doi: 10.1155/2022/791260.

[7]. M. Z. Ibrahim, N. A. M. Isa, and R. A. Bakar, "Employee attrition classification using artificial neural networks and deep learning," *International Journal of Workforce Analytics*, vol. 36, no. 7, pp. 3281–3290, 2021, doi: 10.1002/work.22794.

[8]. P. K. Bansal and S. K. Pandey, "Predicting employee attrition using KNN, SVC, and decision tree classifiers," *IEEE Transactions on Business Intelligence*, vol. 40, no. 6, pp. 1560-1573, Jun. 2021, doi: 10.1109/TBI.2021.3054100.

[9]. R. D. Woods, "A review of ensemble learning techniques for predicting employee turnover," *Journal of Human Resource Analytics*, vol. 4, no. 2, pp. 90–101, 2021, doi: 10.1007/s41666-021-00071-z.

[10]. S. M. Zhang and J. L. Brown, "Employee attrition prediction with ensemble learning methods," *Proceedings of the IEEE International Conference on Machine Learning and Applications*, 2020, pp. 1890-1897, doi: 10.1109/ICMLA.2020.00314.