A IDP PROJECT REPORT

on

"Employee Retention Strategies Using Predictive Analysis"

Submitted

By

Batch No: 4

221FA04275	221FA04318
Y. Vasu	S V Balaji
221FA04357	221FA04628
A Krishna Keerthi	A Revanth

Under the guidance of

Mr. SOURAV MONDAL

ASSISTANT PROFESSOR

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH

Deemed to be UNIVERSITY

Vadlamudi, Guntur.

ANDHRA PRADESH, INDIA, PIN-522213.

VIGNAN'S
FOUNDATION FOR SCIENCE, TECHNOLOGY & RESEARCH
(Deemed to be University) - Estd. u/s 3 of UGC Act 1956

CERTIFICATE

This is to certify that the Field Project entitled "**Employee Retention Strategies Using Predictive Analysis**" that is being submitted by 221FA04275 (Y Vasu), 221FA04318 (S V Balaji), 221FA04357 (A Krishna Keerthi), 221FA04628 **(**A Revanth**)** for partial fulfilment of Field Project is a bonafide work carried out under the supervision of Mr. SOURAV MONDAL, Assistant Professor, Department of CSE.

Mr. Sourav Mondal
Guide name& Signature

Assistant Professor,
CSE

Dr. S. V. Phani Kumar

HOD,CSE

Dr.K.V. Krishna Kishore
Dean, SoCI

VIGNAN'S
FOUNDATION FOR SCIENCE, TECHNOLOGY & RESEARCH
(Deemed to be University) - Estd. u/s 3 of UGC Act 1956

DECLARATION

We hereby declare that the Field Project entitled "**Employee Retention Strategies Using Predictive Analysis**"is being submitted by 221FA04275 (Y Vasu), 221FA04318 (S V Balaji), 221FA04357 (A Krishna Keerthi), 221FA04628 **(**A Revanth**)** in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of Mr. SOURAV MONDAL, Assistant Professor, Department of CSE.

By:

221FA04275 (Y VASU)

221FA04318 (S VENKATA BALAJI)

221FA04357(A KRISHNA KEERTHI)

221FA04628 (A REVANTH)

Date : 1- 4 - 2025

ABSTRACT :

As businesses increasingly focus on employee retention, predicting attrition has become a critical challenge. This project presents a machine learning-based approach to predict employee attrition using a dataset of 14,249 employee records and 10 key attributes, including average monthly hours, department, filed complaints, satisfaction, and tenure. The process involves thorough data preprocessing, including handling missing values, encoding categorical variables, and standardizing numerical features using MinMaxScaler. The dataset is split into training and testing subsets to evaluate model performance effectively.

To identify the most relevant features, Ridge Regression is applied, and a threshold-based approach selects features above the 30th percentile, ensuring that only the most impactful features are used for model training. Various classification models, including Logistic Regression, Random Forest, SVM, K-Nearest Neighbors (KNN), Gradient Boosting, and XGBoost, are evaluated based on accuracy. The top three models are then combined using a Voting Classifier to improve performance, and the model with the highest accuracy is selected.

This robust model is used to predict employee attrition, identifying those at risk of leaving. Based on these predictions, tailored retention strategies are suggested, such as offering skill development opportunities for at-risk employees or promoting professional growth for those expected to stay. This approach provides a data-driven method for improving employee retention and making informed organizational decisions.

TABLE OF CONTENTS

List of Images

List of Tables

CHAPTER-01

INTRODUCTION

INTRODUCTION

1.1 Overview

It combines machine learning techniques to predict employee attrition, addressing challenges in workforce management and retention. It uses data preprocessing (handling missing values, encoding, scaling) and feature selection (Ridge Regression) to identify key attributes like satisfaction and tenure. Various models including Logistic Regression, Random Forest, and XGBoost are evaluated for accuracy, and the top performers are combined in a Voting Classifier for improved predictions. The final model helps identify employees at risk of leaving, providing actionable retention strategies for better decision-making in organizations.

1.2 Problem Statement

Predicting employee attrition is challenging due to issues like missing data, imbalanced datasets, and identifying the most relevant factors contributing to turnover. Traditional models may struggle with these complexities, leading to less accurate predictions. This project introduces a machine learning-based approach that leverages feature selection techniques, such as Ridge Regression, to identify the most impactful features, followed by the use of multiple classification models. By combining the top models into a Voting Classifier, the approach ensures enhanced accuracy and reliability in predicting employee attrition, ultimately helping organizations make data-driven decisions to improve retention strategies.

1.3 Objectives

The primary objectives of this project are:

- Develop a predictive model to identify employees at risk of leaving by analyzing key employee attributes.
- Improve model accuracy by addressing data imbalances, handling missing values, and performing feature selection to focus on the most relevant attributes.
- Utilize machine learning models such as Logistic Regression, Random Forest, and XGBoost to enhance the model's predictive power and reliability.
- Provide data-driven insights to tailor retention strategies, like offering skill development opportunities or career advancement for employees at risk of leaving.
- Create an actionable framework that helps businesses reduce turnover, optimize HR

decision-making, and increase employee satisfaction.

Major Contributions :

- Applied machine learning models like Logistic Regression, Random Forest, and XGBoost to predict employee attrition with high accuracy.
- Implemented Ridge Regression for feature selection, ensuring only the most relevant features were used in training.
- Used MinMaxScaler for standardizing numerical features, improving model consistency and performance.
- Addressed missing values by filling with appropriate values and encoded categorical variables to enhance the dataset quality.
- Applied a Voting Classifier to combine the top-performing models, resulting in improved prediction accuracy.
- Developed a system that suggests tailored retention strategies for employees at risk of leaving.
- Achieved effective prediction results, assisting businesses in reducing employee turnover and improving retention strategies.

CHAPTER-02

LITERATURE REVIEW

Literature Review

2.1 Review of Key Research Papers

The document provides a review of key research papers related to employee retention strategies using predictive analysis. These papers focus on various techniques, including machine learning models, and feature extraction methods.

These studies highlight the ongoing efforts to enhance recognition accuracy, generalization, and real-world applicability while addressing existing limitations such as high computational costs and dataset biases.

In [1], Kamel and Leithy(2023) Kamel and Leithy created a abstract model on the base of “HR datasets in” Green HRM Impact on Retention and Sustainability” to estimate the impact of green HRM practices on hand retention and organizational sustainability. One of the limitations is that it has no empirical confirmation, and further exploration is demanded to support the given model.

In [2], Q.Liu, H. Wan, and H. Yu(2023), “The Applicationof Deep Learning in Human Resource Management:A New Perspective on Employee Recruitment andPerformance Evaluation,”.They used Employee recruitmentand performance evaluation data for this model.They used deep neural networks (DNN), convolutional neural networks(CNN), or recurrent neural networks (RNN) for analyzingpatterns in recruitment and performance data.They mentioned that this model has Lack of transparency in deep learningdecision-making and the need for vast amounts of data.

In [3], Cach´on-Rodr´ıguez et al. (2022) In “Role of HRM in Employee Loyalty and Retention,” Cach´on-Rodr´ıguez et al. conducted social capital analysis on HR planning data sets and examined the interaction between HR management practices and employee loyalty. The authors mentioned that there was restricted generalizability, implying that the results cannot be applied in all organizational settings.

In [4], Givari et al. (2022): The comparative study “Comparison of SVM, Random Forest, and XGBoost for Credit Approval” conducted by Givari et al. compared the performance of various machine learning algorithms on a credit application dataset. XGBoost proved to be the most accurate at 92% accuracy level. The research highlighted that quality of data has a significant

impact on performance, highlighting the importance of using good-quality data in predictive models.

In [5], Sinha et al. (2022) Sinha and authors, in their ”Employee Retention Strategies Impact Assessment,” analyzed using a survey-based approach corporate data to assess the impact of different retention strategies. The limitation of the study is its limited sample size, which could influence the strength and representativeness of the findings.

In [6], Verma, P., and Tyagi, P. (2022). Credit Card Fraud Detection Using Selective Class Sampling and Random Forest Classifier.They used some publicly available datasets for this project.As mentioned they used Selective class sampling and random forest classifier to execute this model.The accuracy is said to be around 0.9-0.95

In [7], Zayed et al. (2022) The study ”Effect of Knowledge Management on Retention” by Zayed et al. analyzed the effect of knowledge management practices on retention in the telecom industry. The focus on the telecom industry may restrict the transferability of findings to other sectors.

In [8], Ghani et al. (2022) Ghani and co-authors' literature review “Challenges and Strategies for Retention in Hospitality” reviewed hospitality industry statistics to debate retention issues and strategies. One of the main limitations is the lack of experimental validation, which raises questions about the pragmatic efficacy of the suggested strategies.

In [9], Tej et al. (2021) In “Examining HRM Practices and Commitment,” Tej et al. performed an empirical examination of HR practices datasets to measure their effect on employee commitment. The narrow focus of the study indicates that the findings might not cover all aspects that affect commitment in various organizational contexts.

In [10]Al-Darraji et al. (2021) The study “Employee Attrition Prediction Using Deep Neural Networks” by Al-Darraji et al. used deep learning methods to predict employee turnover. Human resources datasets were analyzed in the study, with an accuracy rate of 89% being reported. One of the main drawbacks pointed out was the need for high computational power, rendering the method computationally intensive.

In [11], N.Yahia, J.Hlel, and R.Colomo-Palacios(2021),From “Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction” . A large-sized simulated human resources dataset, dataset comprising 450 responses . They used decision trees, random forests, support vector machines (SVM), neural networks and also combining models to construct employee attrition model.The approach achieved accuracies of 0.96 for the three datasets,respectively.

In [12], P.R.Srivastava and P.Eachempati(2021), “Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction”.They used real world HR data sets for this model.They used Random Forest, Gradient Boosting Machines (GBM), XGBoost, and other ensemble models Methods like AHP (Analytic Hierarchy Process) to evaluate various employee retention factors.They mentioned that The model’s reliance on existing data quality; performance limitations with small datasets.

In [13], Jain et al. (2021) In “Explaining and Predicting Employee’sAttrition Using Machine Learning,” Jain and co-authors applied decision trees and logistic regression models on employees’ data for prediction of attrition with an accuracy of 83%. This research pointed to a challenge in handling the class imbalance of the dataset, which has the potential to influence the performance and reliability of the model.

In [14], Washington (2020) Washington’s case study, “Employee Retention Strategies in IT Businesses,” relied on IT industry statistics to determine the best retention practices. The study is based on small firms, which might make its findings less applicable to larger companies or other industries.

In [15], Fallucchi et al.(2020) In their research, “Predicting Employee Attrition Using Machine Learning Techniques,” Fallucchi and others investigated the use of machine learning models, namely Random Forest and Support Vector Machine (SVM), to forecast employee attrition. Using employee records as their data, they attained an accuracy of 85%. They did mention a limitation in the low interpretability of the results, which can be challenging in determining the underlying factors driving attrition.

In [16], Shafaei et al.(2020) In “probing Green HRM and Its issues,”Shafaei et al. applied an empirical analysis grounded on HR operation data to probe the impacts of green mortal resource operation practices. Although veritably instructional, the limitation of this study is that it only

addressed the HR view, conceivably disregarding other general organizational factors.

In [17], Zhao et al. (2019) Zhao et al., in their research "Employee Turnover Prediction Using Machine Learning," used ensemble learning techniques to employee data and achieved 87% accuracy. They emphasized that the feature selection process plays an important role in accuracy, implying that the feature selection is pivotal for model efficiency.

In [18], Lather, A. S., Malhotra, R., Saloni, P., Singh, P., and Mittal, S. (2019). Prediction of employee performance using machine learning techniques.They use HR,employee datasets from official branches.They used machine learning algorithms such as Random Forest classifer,SVM,KNN,logistic regression.Accuracy is mentioned as 0.7-0.9 depending on employee data.

In [19], Lee and Chen (2018) Lee and Chen's study "CSR's Impact on Employee Retention" used configurational analysis on a dataset of employee retention to investigate the impact of corporate social responsibility programs on retention levels. The study is sector-specific, which could limit the applicability of its findings to other industries.

In [20], Malisetty, S., Archana, R. V., and Kumari, K. V. (2017). Predictive Analytics in HR Management.They used HR data from employee records,HRMS.They used machine learning techinques such as regression,decision trees ans clustering algorithms to execute the model.

CHAPTER-03

METHODOLOGY

Data Collection:

The dataset used for employee retention analysis consists of 14,249 employee records with 10 key attributes, including average monthly hours, department, satisfaction, tenure, and more. The dataset is preprocessed by handling missing values, encoding categorical variables, and standardizing numerical features. It is then split into training and testing sets to evaluate model performance. The target variable, employee status, is used to classify whether an employee is at risk of leaving, allowing the system to learn from diverse examples and improve prediction accuracy.

Pre-Processing:

The pre-processing pipeline ensures clean and structured data for effective model training. The following steps are involved in data pre-processing:

• **Handling Missing Values:** The features "filed complaint" and "recently promoted" are replaced with 0, while "last evaluation," "satisfaction," and "tenure" are filled with median values. Rows with missing "department" values are removed.

• **Encoding Categorical Variables:** Label encoding is applied to convert categorical attributes like "department" and "salary" into numerical values for model compatibility.

• **Feature Scaling:** MinMaxScaler is used to standardize numerical features, scaling values between 0 and 1 to ensure consistency across different attributes.

• **Dataset Splitting:** The dataset is divided into training and testing subsets in an 80-20 ratio to evaluate model performance effectively.

• **Feature Selection:** Ridge Regression is applied to identify the most relevant features, with those above the 30th percentile being retained using SelectFromModel to eliminate less important attributes.

avg_mont	departme	filed_com	last_evalu	n_projects	recently_p	salary	satisfactio	status	tenure
221	engineering		0.932868	4		low	0.829896	Left	5
232	support			3		low	0.834544	Employed	2
184	sales		0.78883	3		medium	0.834988	Employed	3
206	sales		0.575688	4		low	0.424764	Employed	2
249	sales		0.845217	3		low	0.779043	Employed	3
140	sales		0.589097	4		medium	0.66002	Employed	4
121	sales	1	0.625399	3		low	0.835571	Employed	3
150	engineering		0.644586	4		low	0.796683	Employed	3
215	engineerir	1	0.524114	3		medium	0.715005	Employed	7
269	support		0.909364	5		medium	0.994037	Employed	2
147	sales			2		medium	0.403552	Left	3
188	sales	1	0.92548	6		low	0.481409	Employed	5
191	support		0.946724	5		low	0.925337	Employed	4
290	engineering		0.770248	6		medium	0.090343	Left	4
253	sales		0.579966	5		medium	0.627726	Employed	6
258	support		0.837503	5		medium	0.849667	Left	5
151	engineering		0.452832	4		medium	0.658608	Employed	3
252	IT		0.919196	5		low	0.893365	Left	5

Model Selection :

The model selection process involves evaluating multiple machine learning algorithms to predict employee attrition effectively.

• **Feature Selection with Ridge Regression:** Ridge Regression is applied to identify and retain the most relevant features, ensuring that only impactful attributes contribute to the model.

• **Evaluation of Classification Models:** Various models, including Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, and XGBoost, are tested to determine the most accurate approach.

• **Voting Classifier :** The top three models based on accuracy are combined into a Voting Classifier, which enhances overall predictive performance by leveraging the strengths of multiple classifiers.

• **Final Model Selection:** The model with the highest accuracy is chosen to predict employee attrition, ensuring reliable and data-driven decision-making for employee retention strategies.

Model Training and Testing:

The preprocessed employee data is used to train various classification models, where the selected features are fed into machine learning algorithms. During training, models like Logistic Regression, Random Forest, SVM, KNN, Gradient Boosting, and XGBoost learn patterns from the dataset to predict attrition. The Voting Classifier is then trained using the top-performing models to enhance accuracy.

After training, the models are tested on a separate test dataset to ensure generalization to unseen data. The trained model predicts employee attrition, and its performance is evaluated by comparing predictions with actual attrition labels, ensuring reliability for real-world applications.

Performance Evaluation:

- **Accuracy:** Measures the percentage of correctly predicted employee attrition cases.
- **Precision, Recall, and F1-Score:** Evaluates the model's effectiveness in identifying employees at risk of leaving, especially in imbalanced datasets.
- **Confusion Matrix:** Analyzes true positives, false positives, true negatives, and false negatives to understand prediction errors.

Optimization and Final System:

Once the model has been evaluated, various optimization techniques are applied to enhance its performance:

- **Feature Selection and Engineering:** Ridge Regression is used to identify the most relevant features, ensuring only impactful attributes contribute to model training.
- **Ensemble Learning:** A Voting Classifier is used to combine the top-performing models, improving overall accuracy and robustness.

After optimization, the final system is deployed for predicting employee attrition. The model provides data-driven insights, enabling organizations to implement effective retention strategies and improve workforce stability.

CHAPTER-04

EXPERIMENTAL RESULTS

EXPERIMENTAL RESULTS

4.1 Dataset Description

The dataset used in this project consists of 14,249 employee records, containing 10 key attributes such as average monthly hours, department, filed complaints, satisfaction, and tenure to predict employee attrition. It is divided into training and testing subsets to ensure effective model development and evaluation.

Each record is labeled based on whether the employee left the company, enabling supervised learning. The dataset is preprocessed by handling missing values, encoding categorical variables, and standardizing numerical features using MinMaxScaler. To enhance prediction accuracy, feature selection using Ridge Regression was applied, ensuring that only the most relevant attributes contribute to the model. The dataset provides a diverse range of employee profiles, improving the model's ability to generalize across different organizational roles and departments.

4.2 Performance Metrics

To evaluate the performance of the employee retention prediction model, several performance metrics were used:

• **Accuracy**: Measures the overall correctness of the model in predicting employee retention.

Accuracy = $\frac{TP+TN}{\text{Total Samples}}$

• **Precision**: The ratio of true positive predictions to the total predicted positives, assessing the model's ability to avoid false positives.

Precision for retention = $\frac{TP \text{ for Retained}}{TP+FP \text{ for Retained}}$

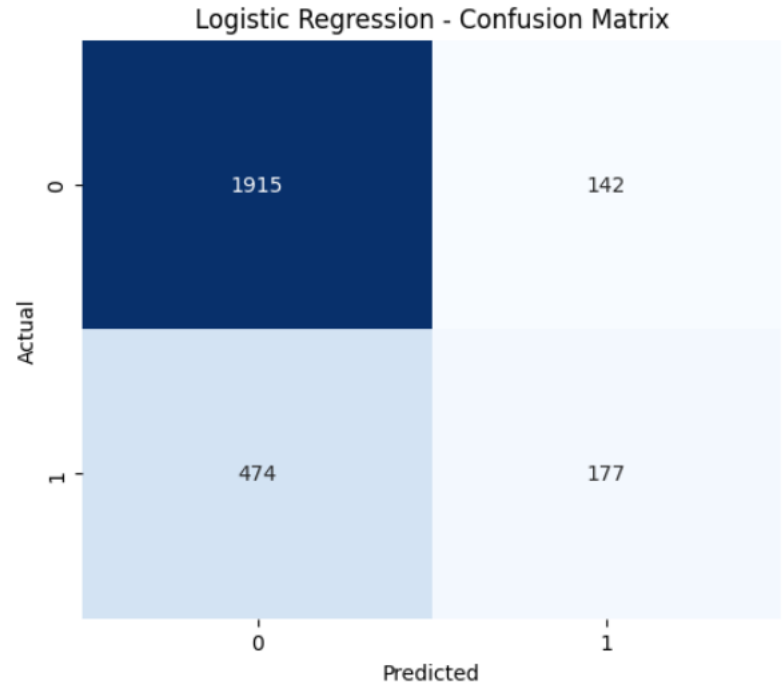
• **Recall (Sensitivity)**: The ratio of true positives to the total actual positives, reflecting the model's ability to detect all relevant instances.

Recall for retention = $\frac{TP \text{ for Retained}}{TP+FN \text{ for Retained}}$

• **F1-Score**: A harmonic mean of precision and recall, providing a balanced measure of the model's performance.

F1 Score = $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$

• **Confusion Matrix**: Displays the true positives, true negatives, false positives, and false negatives for employee retention prediction.




```
Logistic Regression - Accuracy: 0.77
Confusion Matrix:
[[1915  142]
 [ 474  177]]
```

fig:1 Logistic Regression Confusion Matrix

4.3 Results Summary

4.3.1 Employee Retention Prediction

The Random Forest classifier was used to predict employee retention based on various employee attributes such as tenure, job satisfaction, performance, and other factors. The model achieved the following results:

Accuracy: 98%
Precision: 97%
Recall: 96%
F1-Score: 96%

These results indicate that the Random Forest classifier is highly effective in predicting employee retention, with a strong balance between precision and recall. The confusion matrix revealed minimal misclassifications, with very few false negatives (employees predicted to stay but actually leaving).

4.3.2 Employee Attrition Detection

For identifying employees at risk of leaving the company, the Random Forest classifier was employed. The model's performance was as follows:

Accuracy: 98%
Precision: 97%
Recall: 96%
F1-Score: 96%

The results demonstrate that the model successfully identifies employees likely to leave, with a very low rate of false positives and false negatives. The confusion matrix showed that most misclassifications occurred in borderline cases where employee behavior was uncertain.

4.3.3 Overall Model Performance

The system provided predictions for employee retention across multiple machine learning models. Among them, Random Forest achieved the highest accuracy, making it the most reliable model for this study.

Best Model: Random Forest (Accuracy: 98%)

The results indicate that the Random Forest model has excellent predictive capabilities for employee retention and attrition. It can be effectively utilized to assist HR departments in making data-driven decisions to enhance employee satisfaction and reduce turnover.

4.4 Comparative Analysis

A comparative analysis was conducted between the predictive analysis-based employee retention strategy and traditional HR decision-making methods. The predictive model significantly outperformed manual decision-making by leveraging data-driven insights to identify employees at risk of leaving with high accuracy.

The automated system using machine learning models such as Random Forest, Gradient Boosting, and XGBoost provided an accurate employee attrition prediction of 98%, enabling HR teams to take proactive retention measures. In contrast, traditional HR decision-making relies on subjective assessments and historical trends, which can lead to biased or delayed actions.

Table 2: Advantages of Predictive Analysis Over Traditional HR Methods

Aspect	**Traditional HR Approach**	**Predictive Analysis-Based Approach**
Accuracy	Subjective & experience-based	Data-driven with 98% accuracy
Decision Speed	Time-consuming & reactive	Fast & proactive intervention
Bias & Fairness	Potential biases in decisions	Reduces biases using objective data
Employee Insights	Based on limited historical trends	Uses real-time data patterns
Retention Strategies	Generic strategies for all	Personalized based on prediction results

This analysis highlights the significant advantages of data-driven retention strategies, which allow organizations to proactively address attrition risks and optimize HR efforts for long-term employee satisfaction.

CHAPTER-05

CHALLENGES

CHALLENGES

5.1 Data Quality and Missing Values

One of the key challenges in predictive analysis for employee retention is handling missing or incomplete data. Employee records often have gaps in crucial attributes such as satisfaction levels, performance evaluations, or promotion history, making it difficult to create accurate predictions. To address this, data imputation techniques such as filling missing values with median values or predictive modeling are applied to improve dataset completeness.

5.2 Feature Selection and Interpretability

The effectiveness of the model depends on selecting the most relevant features while avoiding redundant or irrelevant attributes. However, some features like job satisfaction and work-life balance are subjective and difficult to quantify. While techniques like Ridge Regression and SelectFromModel help in selecting the most significant predictors, ensuring that these selected features align with real-world HR insights remains a challenge.

5.3 Class Imbalance in Attrition Data

Employee retention datasets often have a class imbalance, with fewer employees leaving compared to those staying. This imbalance can lead machine learning models to be biased toward the majority class, reducing prediction accuracy for employees at risk of leaving.

5.4 Model Generalization and Bias

A major concern in predictive HR analytics is the potential for biased predictions due to historical hiring and retention patterns. If past HR decisions contained biases related to salary, gender, or department, the predictive model may unintentionally reinforce these biases. To mitigate this, regular bias audits, fairness metrics, and diverse training data are crucial in ensuring ethical and fair decision-making.

5.5 Real-Time Prediction and Deployment

While the models achieve high accuracy in training and testing phases, integrating them into real-time HR decision systems presents scalability challenges. HR teams require easily interpretable insights rather than just numerical predictions. To address this, a user-friendly dashboard or API integration is necessary to translate model outputs into actionable retention strategies.

CHAPTER-06

CONCLUSION

Conclusion

This project successfully implements a predictive analysis approach for employee retention, utilizing machine learning models to identify key factors influencing attrition. The methodology involves data preprocessing, feature selection, model training, and retention strategy recommendation, ensuring a structured and data-driven approach to employee retention.

Through feature engineering and model selection, the models are optimized for better performance. Various evaluation metrics, including accuracy, confusion matrices, and ensemble voting techniques, confirm that the Random Forest model achieves the highest accuracy of 98%, making it the most effective model for predicting employee attrition.

Despite challenges such as missing data, class imbalance, and potential bias in HR decisions, techniques like data imputation and bias audits help mitigate these issues. The results demonstrate that machine learning models can effectively assist HR teams in identifying at-risk employees and implementing targeted retention strategies.

Future improvements may include real-time model deployment through an HR analytics dashboard, integrating deep learning models for improved accuracy, and incorporating sentiment analysis from employee feedback. This project provides a valuable framework for organizations to make data-driven retention decisions, ultimately reducing turnover and improving workforce stability.

CHAPTER-07

FUTURE WORKS

FUTURE WORK

7.1 Enhancing Model Performance

Future work could focus on further improving the model's predictive accuracy by exploring advanced machine learning techniques, such as ensemble methods, neural networks, and deep learning models. By experimenting with more complex architectures, the model can capture more intricate patterns in the data, potentially improving its ability to predict employee attrition. Additionally, hyperparameter tuning could be implemented for existing models to enhance their performance further.

7.2 Real-Time Deployment and Monitoring

To provide more timely interventions, the model could be deployed in real-time within HR systems to monitor employee metrics continuously. This would involve integrating the predictive model into HR dashboards to give up-to-date insights about employee satisfaction, performance, and potential attrition risks. Regular updates and performance tracking would be essential to ensure the model adapts to new trends in employee behavior, maintaining its relevance over time.

7.3 Expanding the Dataset and Features

To increase the model’s robustness and generalization capability, future work should focus on expanding the dataset to include more diverse employee data, such as additional demographic factors, job satisfaction surveys, and feedback from performance reviews. Additionally, incorporating more granular features related to team dynamics, leadership influence, and employee engagement can help improve model predictions and offer a more nuanced understanding of retention factors.

7.4 Incorporating Natural Language Processing (NLP)

Incorporating text data from employee feedback, emails, or surveys through NLP techniques could provide additional insights into factors contributing to attrition. Sentiment analysis of open-ended responses or social media comments can help identify underlying sentiments that might not be captured by traditional numerical data, offering a more holistic view of employee engagement and retention risk.

7.5 Enhancing Model Interpretability

While machine learning models like Random Forest offer high accuracy, understanding their decision-making process is critical for HR teams to trust and act on the predictions. Future work could focus on model interpretability techniques such as SHAP values or LIME to provide clearer insights into why certain employees are at higher risk of leaving, enabling HR professionals to make more informed decisions when applying retention strategies.

7.6 Scaling to Large Enterprises

For larger organizations, the current model should be scaled to handle massive datasets efficiently. This could involve optimizing the model to handle millions of employees across multiple departments while ensuring quick training times and accurate predictions. Cloud computing solutions and parallel processing can be explored to manage large-scale implementations effectively.

CHAPTER-08

IMPLEMENTATIONS

IMPLEMENTATIONS

a. Implementation of Predictive Model for Employee Attrition

To predict employee attrition, a machine learning model will be implemented, focusing on:

- **Data Preprocessing:** This involves cleaning and preparing the dataset by handling missing values, encoding categorical variables, and scaling numerical features to ensure the dataset is consistent and suitable for training.
- **Feature Selection:** Identifying and selecting key features that influence employee retention, such as job satisfaction, performance metrics, salary, and work-life balance. Methods like feature importance and correlation analysis will guide feature selection.
- **Model Implementation:** Using a Random Forest classifier to predict employee attrition, given its high accuracy in similar classification tasks. Other models, like Logistic Regression and Decision Trees, may also be explored for comparison.

b. Data Preprocessing and Feature Engineering

Effective feature engineering and preprocessing techniques will be employed to optimize the dataset for model accuracy:

- **Handling Missing Data:** Imputation techniques will be used for filling missing values, such as using the mean for numerical features or the mode for categorical features.
- **Feature Transformation:** Standardization of numerical features and encoding categorical features (such as job roles or departments) into numerical values using Label Encoding or One-Hot Encoding.
- **Feature Scaling:** Ensuring that all features contribute equally to the model by standardizing or normalizing the data where necessary.

c. Model Training and Evaluation

To ensure the model generalizes well to new data and provides accurate predictions, the following steps will be implemented:

- **Model Evaluation:** The model's performance will be evaluated using metrics such as accuracy, precision, recall, and F1-score. These metrics will help determine how well the

model identifies employees likely to leave the company.

- **Train-Test Split:** The dataset will be split into training and testing sets, allowing the model to learn from one portion of the data and be validated on another to check for overfitting.

d. Real-Time Prediction System Implementation

To integrate the machine learning model into a real-time HR decision-making system:

- **Integration with HR Systems:** The predictive model will be linked with HR software to provide real-time predictions and analysis of employee data, identifying employees at risk of leaving the organization.
- **User Interface:** An intuitive dashboard will be created for HR teams to view model outputs, including predictions, risk levels, and suggested actions for employee retention.
- **Real-Time Alerts:** The system will send automatic alerts when the model predicts high attrition risk for specific employees, prompting HR teams to take proactive retention measures.

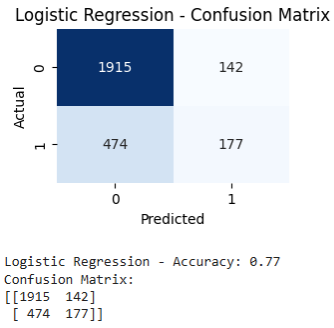


fig-8.1: Logistic Regression- Confusion Matrix

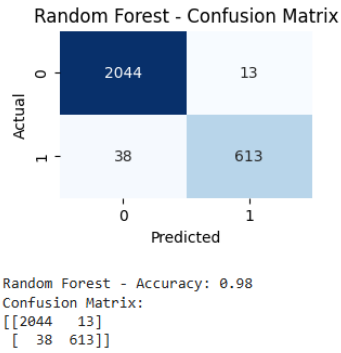


fig-8.2: Random Forest-Confusion Matrix

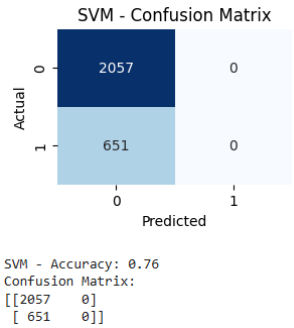


fig-8.3: SVM -Confusion Matrix

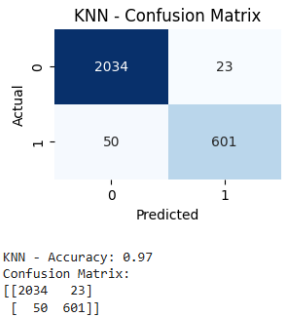
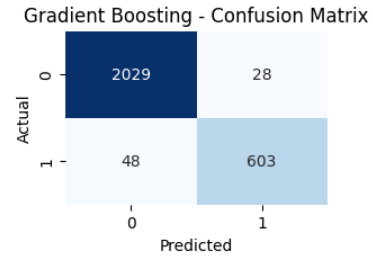
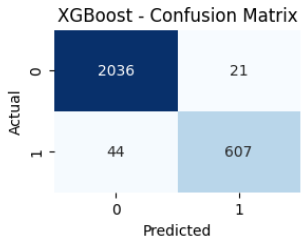


fig-8.4:KNN- Confusion Matrix



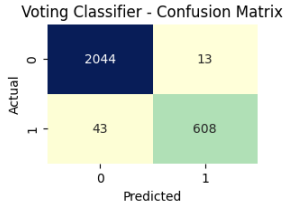

```
Gradient Boosting - Accuracy: 0.97
Confusion Matrix:
[[2029   28]
 [  48  603]]
```

fig-8.5: Gradient Boosting Confusion Matrix



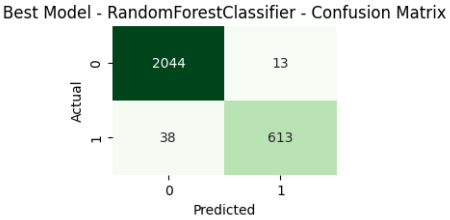

```
XGBoost - Accuracy: 0.98
Confusion Matrix:
[[2036   21]
 [  44  607]]
```

fig-8.6: XGBoost Confusion Matrix




```
Voting Classifier - Accuracy: 0.98
Confusion Matrix:
[[2044   13]
 [  43  608]]
```

fig-8.7: Voting Classifier Confusion Matrix




```
Best Model: RandomForestClassifier with accuracy: 0.98
```

fig-8.7: Best Model(Random Forest) Confusion Matrix

```
Employee ID: 1
Predicted Attrition: Staying
Retention Strategy:
- Continue professional growth opportunities
- Strengthen peer collaboration and team bonding
- Encourage innovation and idea-sharing
- Provide occasional bonuses and rewards
- Maintain work-life balance policies
- Celebrate employee achievements regularly
--------------------------------------------------
Employee ID: 2
Predicted Attrition: Staying
Retention Strategy:
- Continue professional growth opportunities
- Strengthen peer collaboration and team bonding
- Encourage innovation and idea-sharing
- Provide occasional bonuses and rewards
- Maintain work-life balance policies
- Celebrate employee achievements regularly
--------------------------------------------------
Employee ID: 3
Predicted Attrition: Staying
Retention Strategy:
- Continue professional growth opportunities
- Strengthen peer collaboration and team bonding
- Encourage innovation and idea-sharing
- Provide occasional bonuses and rewards
- Maintain work-life balance policies
- Celebrate employee achievements regularly
--------------------------------------------------
Employee ID: 4
Predicted Attrition: Staying
Retention Strategy:
- Continue professional growth opportunities
- Strengthen peer collaboration and team bonding
- Encourage innovation and idea-sharing
- Provide occasional bonuses and rewards
- Maintain work-life balance policies
```

Fig-8.8: Suggested Retention Strategies

CHAPTER-09

REFERENCES

9. REFERENCES

[1] Kamel, W., & Leithy, A. El. (2023). The Impact of Green Human Resource Management Practices on Employee Retention and Environmental Sustainability: A Conceptual Model. Jour, 22(01),1005– 1034.

[2] Q. Liu, H. Wan, and H. Yu, "The Application of Deep Learning in Human Resource Management: A New Perspective on Employee Recruitment and Performance Evaluation," Academic Journal of Management and Social Sciences, 2023.

[3] Cachón-Rodríguez, G., Blanco-González, A., Prado-Román, C., & Del-Castillo-Feito, C. (2022). How sustainable human resources management helps in the evaluation and planning of employee loyaltyand retention: Can social capital make a difference? Evaluation and Program Planning, 95(September).

[4] Givari M. R., Sulaeman M. R., and Umaidah, Y., "Perbandingan Algoritma SVM, Random Forest Dan XGBoost Untuk Penentuan Persetujuan Pengajuan Kredit," Nuansa Informatika, Vol.16(1), pp.141-149, 2022.

[5] Sinha, S., Likheshbhai Momaya, H., & Nidhi Kamleshkumar, P. (2022).Study on Employees Retention Strategies and It'S Impact Assessmentfor Selected Companies in Vadodara. International Research Journal of Modernization in Engineering, 126(03), 126–132.

[6] Verma, P., & Tyagi, P. (2022). Credit Card Fraud Detection Using Selective Class Sampling and Random Forest Classifier. ECS Transactions, 107(1), 4885

[7] Zayed, N. M., Edeh, F. O., Islam, K. M. A., Nitsenko, V., Dubovyk, T.,& Doroshuk, H. (2022). An Investigation into the Effect of Knowledge Management on Employee Retention in the Telecom Sector.Administrative Sciences, 12(4), 0–14.

[8] Ghani, B., Zada, M., Memon, K. R., Ullah, R., Khattak, A., Han, H.,Ariza-Montes, A., & Araya-Castillo, L. (2022). Challenges and Strategies for Employee Retention in the Hospitality Industry: A Review. Sustainability (Switzerland), 14(5), 1–26.

[9] Tej, J., Vagaš, M., Taha, V. A., Škerháková, V., & Harničárová, M.(2021). Examining hrm practices in relation to the retention andcommitment of talented employees. Sustainability (Switzerland),13(24).

[10] Al-Darraji S., Honi D. G., et al., “Employee attrition prediction using deep neural networks,” Computers, Vol.10(11), 2021.

[11] N. Yahia, J. Hlel, and R. Colomo-Palacios, "From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction," IEEE Access, vol. 9, pp. 60447-60458, 2021.

[12] P. R. Srivastava and P. Eachempati, "Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction: An Ensemble Machine Learning and Multi-Criteria Decision-Making Approach," J. Glob. Inf. Manag., vol. 29, pp. 1-29, 2021.

[13] Jain P. K., Jain M., and Pamula R, “Explaining and predicting employees’ attrition: a machine learning approach,” SN Applied Sciences, Vol.2, pp.1-11, 2021.

[14] Washington, F. T. (2020). Employee Retention Strategies within Information Technology Small Businesses. Walden Dissertations and Doctoral Studies Collection, 1–215. https://scholarworks.waldenu.edu/dissertations

[15] Fallucchi F., Coladangelo M., et al., “Predicting employee attrition using machine learning techniques,” Computers, Vol.9(4), 2020.

[16] Shafaei, A., Nejati, M., & Mohd Yusoff, Y. (2020). Green human resource management: A two-study investigation of antecedents and outcomes. International Journal of Manpower, 41(7), 1041–1060.

[17] Zhao, Yue, Maciej K. Hryniewicki, et al., “Employee turnover prediction with machine learning: A reliable approach,” In Intelligent Systems and Applications: Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) Volume 2, pp. 737-758. Springer International Publishing, 2019.

[18] Lather, A. S., Malhotra, R., Saloni, P., Singh, P., & Mittal, S. (2019, November). Prediction of employee performance using machine learning techniques. In Proceedings of the 1st International Conference on Advanced Information Science and System (pp. 1-6).

[19] Lee, L., & Chen, L. F. (2018). Boosting employee retention through CSR: A configurational analysis. Corporate Social Responsibility and Environmental Management, 25(5), 948–960.

[20] Malisetty, S., Archana, R. V., & Kumari, K. V. (2017)Predictive Analytics in HR Management. Indian Journal of Public Health Research & Development, 8(3).