Attrition & Salary Insights

ON

Submitted in partial fulfillment of the requirements of the degree of

Bachelor of Engineering

(Information Technology)

By

Aditya Kirtane (26)

Sairam Konar (27)

Bhushan Kor (28)

Under the guidance of

Dr. Ravita Mishra

Department of Information Technology

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY,

Chembur, Mumbai 400074

(An Autonomous Institute, Affiliated to University of Mumbai) April 2024

Abstract

This project, "Employee Attrition Prediction and Salary Anomaly Detection," develops a robust machine learning framework to enhance human resource management by predicting employee turnover and identifying salary irregularities. Utilizing a comprehensive dataset with features such as satisfaction levels, years of experience, and salary tiers, the system employs classification models like Logistic Regression, Decision Tree, Random Forest, KNN, and XGBoost, achieving accuracies up to 0.8334 with XGBoost. Additionally, the Isolation Forest algorithm detects salary outliers, promoting fairness in compensation. Exploratory data analysis reveals a 60% retention rate, with attrition concentrated among lower-paid employees with less than 20 years of experience. The solution integrates data visualization and performance metrics (precision, recall, F1-score) to provide actionable insights, enabling HR to proactively address retention challenges and ensure equitable pay practices. This data-driven approach bridges traditional HR analytics with predictive modeling, fostering a stable and satisfied workforce.

Keywords- **Employee Attrition, Salary Anomaly Detection, Machine Learning, HR Analytics, Predictive Modeling, Fair Compensation**

Table of Contents

CHAPTER: 1

INTRODUCTION

Chapter 1

Introduction

1.1. Introduction

Employee retention is crucial to organizational success, especially in today's competitive job market. Understanding why employees leave and detecting irregularities in salary distribution can help companies make informed decisions. This project leverages machine learning to predict employee attrition and identify salary anomalies using classification and outlier detection models. By analyzing key features such as salary, department, years of experience, and job satisfaction, the system provides insights into employee behavior. The project not only enhances HR analytics but also aids in proactive decision-making. The solution is data-driven, interpretable, and aimed at increasing organizational efficiency through predictive modeling.

1.2. Objectives

The project aims to enhance decision-making in Human Resource Management by applying machine learning techniques for employee attrition prediction and salary anomaly detection. It addresses challenges in retaining employees and ensuring fair compensation through data-driven analysis. By identifying key attrition factors and salary outliers, the project enables HR teams to take proactive and informed actions.

- **Perform EDA –** Explore and visualize data to uncover patterns, correlations, and trends.
- **Analyze employee data –** Understand employee demographics, job roles, and satisfaction metrics influencing attrition.
- **Build and compare ML models –** Implement various classification algorithms to predict attrition and evaluate their performance.
- **Detecting salary anomalies –** Use outlier detection techniques to identify irregularities in employee salaries.

1.3. Motivation

Employee turnover and salary mismanagement have long-term repercussions on a company's growth and culture. Traditional HR analytics are often reactive and lack the precision of predictive models. The motivation behind this project is to empower organizations with proactive tools that anticipate attrition and flag inconsistencies in salary data. By leveraging machine learning, companies can adopt a scientific approach to workforce management, improving satisfaction and transparency. The increasing availability of employee-related data provides a strong foundation to drive impactful insights, reduce churn, and foster a more stable work environment.

1.4. Scope of the Work

The scope of this project outlines the core components and deliverables required to develop a comprehensive, intelligent system for predicting employee attrition and detecting salary anomalies using machine learning techniques. The focus is on robust model development, insightful data visualization, and actionable HR analytics.

- **Data Collection and Preprocessing:** Gather employee data from the provided dataset and perform preprocessing steps such as handling missing values, encoding categorical variables, scaling features, and preparing the dataset for modeling.
- **Exploratory Data Analysis (EDA): Perform in-depth EDA to uncover trends and relationships within the data, such as salary distributions, department-wise attrition, and feature correlations. Visualizations will guide further modeling efforts.**
- **Attrition Prediction Modeling:** Implement multiple classification models including Logistic Regression, Decision Tree, Random Forest, SVM, KNN, and XGBoost. Each model will be trained, tested, and optimized using appropriate metrics.
- **Anomaly Detection for Salary Outliers:** Use the Isolation Forest algorithm to detect unusual salary patterns, aiding HR departments in identifying potential inconsistencies or outliers in compensation.
- **Performance Evaluation and Visualization:** Evaluate model performance using metrics such as accuracy, precision, recall, and F1-score. Visualize the confusion matrix for each model to compare effectiveness in classifying attrition.
- **Result Analysis and Interpretation:** Analyze outcomes of both classification and anomaly detection tasks. Provide detailed inferences supported by quantitative results and graphical insights.

CHAPTER: 2

LITERATURE SURVEY

Chapter 2

Literature Survey

2.1. Review of Literature Survey

1. **Francesca Fallucchi, Marco Coladangelo, Romeo Giuliano, Ernesto William De Luca (2020) – Predicting Employee Attrition Using Machine Learning Techniques**

- **Problem Statement:** The study aims to analyze objective factors influencing employee attrition by identifying the primary causes that lead employees to leave an organization. Using a real-world dataset from IBM HR Analytics containing approximately 1500 samples and 35 features, the research applies various machine learning models to predict whether a specific employee is likely to resign. By uncovering patterns related to factors such as salary, age, overtime, distance from home, and job satisfaction, the study enables the development of predictive tools that support HR departments in making informed, data-driven decisions. Ultimately, this approach helps organizations proactively address attrition risks and implement targeted retention strategies.
- **Models used and accuracies:**
    1. Gaussian Naive Bayes – 82.5%
    2. Bernoulli Naive Bayes – 84.5%
    3. Logistic Regression – 87.5%
    4. K-Nearest Neighbors (K-NN) – 85.2%
    5. Decision Tree – 82.3%
    6. Random Forest – 86.1%
    7. Support Vector Classifier (SVC) – 85.9%
- **Conclusion:**
  The study effectively demonstrates how machine learning techniques can be leveraged to predict employee attrition by analyzing key objective factors from HR data. By identifying the most influential variables—such as monthly income, age, overtime, and job involvement—the research provides valuable insights into the reasons behind employee turnover. The implementation of predictive models, particularly the Gaussian Naive Bayes classifier with the highest recall rate, equips HR departments with a powerful decision-support tool. This enables early intervention, allowing organizations to implement targeted retention strategies and ultimately reduce the costs and disruptions associated with high attrition rates.

2. Rukiye Kaya, Mehtap Saatçi, Mehmet Gökhan Bakal (2024) – "Improving Salary Offer Processes With Classification Based Machine Learning Models"

- **Problem Statement:** This study addresses the challenge of offering fair and data-driven salary estimates during the hiring process. By analyzing employee attributes—such as education, experience, and competencies—using classification-based machine learning models, the aim is to predict appropriate salary categories for job candidates. The goal is to support HR departments in improving the efficiency, accuracy, and fairness of salary offers while mitigating potential biases and wage inequalities.

- **Models used and Accuracies:**
    1. Decision Tree Classifier – 49.7%
    2. Random Forest Classifier – 56.3%
    3. Multinomial Logistic Regression – 51.9%
    4. K-Nearest Neighbors (KNN) – 56.8%
    5. Support Vector Machine (SVM) – 54.1%
    6. XGBoost Classifier – 54.6%
    7. Artificial Neural Network (ANN) – 58.2%
- **Conclusion:** The study demonstrates that machine learning models, particularly Artificial Neural Networks and ensemble techniques, can enhance salary prediction accuracy in HR decision-making. Despite moderate accuracy due to data imbalance, the system supports fairer and more consistent salary offers. Future improvements with larger datasets can further boost predictive performance and reliability.

CHAPTER: 3

DESIGN AND IMPLEMENTATION

Chapter 3

Design and Implementation

3.1. Introduction

This chapter presents the design and implementation details of the AI-powered chatbot system tailored for educational use. The system architecture integrates a Next.js-based frontend with a Flask backend, ensuring a responsive and interactive user interface. Handwritten and typed PDFs are processed using an OCR API, and relevant content is extracted and chunked using an efficient vector database. These chunks are then fed into a retrieval model to provide context-aware answers, with prompt-based support for summarization and expansion. The implementation also includes text-to-speech conversion and chat session sharing, focusing on usability, modularity, and real-time performance through external APIs.

3.2. Requirement Gathering

The initial phase of the project focused on gathering requirements by engaging with HR professionals to understand core challenges in employee management. Two primary issues emerged: rising attrition rates and employee dissatisfaction related to perceived salary discrepancies. These concerns shaped both functional and non-functional system needs. Functionally, the system must preprocess employee data, handle missing or categorical values, and classify employees based on their likelihood of leaving the organization. In addition, it should detect salary anomalies using outlier detection algorithms such as Isolation Forest. On the non-functional side, the system should be efficient, scalable, and interpretable, ensuring that outputs are accessible to non-technical HR stakeholders. Visualization features like correlation heatmaps and confusion matrices were considered essential for interpretability. Python was chosen for its robust data science ecosystem, with libraries like scikit-learn, pandas, and matplotlib identified for model training, data handling, and visualization, respectively. This foundation ensures a reliable and insightful analytical solution.

3.3. Proposed Design

The proposed design starts with collecting a structured employee dataset containing features such as salary, department, satisfaction level, years at the company, evaluation scores, and promotion history. This data provides a complete profile for analyzing attrition and salary-related anomalies. Stored in CSV format, it can be easily integrated into the machine learning pipeline.

Data preprocessing ensures data quality through handling missing values (via imputation or removal), encoding categorical variables using label or one-hot encoding, and scaling numerical features through normalization or standardization. These steps are crucial for consistent model performance.

Exploratory Data Analysis (EDA) is conducted to uncover trends and correlations. Salary distributions are analyzed across departments, while attrition is broken down by satisfaction, experience, and departmental groups. A correlation heatmap highlights feature relationships, guiding model selection and feature engineering.

Multiple classification models—Logistic Regression, Decision Tree, Random Forest, SVM, KNN, and XGBoost—are trained using an 80-20 train-test split. Hyperparameters are optimized, and their performances are compared to determine the best model for predicting employee attrition.

To complement this, Isolation Forest is employed as an anomaly detection method to flag salary outliers. Applied to numerical features like salary and experience, it identifies irregular patterns, helping HR detect potential issues in payroll fairness.

Evaluation metrics such as accuracy, precision, recall, and F1-score are used to assess model performance. Confusion matrices provide a visual breakdown of results. The best-performing model is selected based on these metrics, and a final report is prepared to summarize insights and guide improvements for future development or deployment.

3.4. Proposed Algorithm

- **Logistic Regression**

Logistic Regression is a statistical model used for binary classification problems like employee attrition. It calculates the probability of an employee leaving the organization based on input features by fitting data to a logistic curve. The model outputs values between 0 and 1, interpreted as the probability of attrition. It works well with linearly separable data and offers a baseline for comparison with more complex models. Due to its simplicity, speed, and interpretability, logistic regression serves as a foundational algorithm in this project's attrition prediction pipeline.

- **Decision Tree**

Decision Tree is a supervised learning algorithm that splits the data into subsets based on feature values, forming a tree-like structure. Each internal node represents a decision based on an attribute, and each leaf node corresponds to an outcome (e.g., "Leave" or "Stay"). It is easy to visualize and interpret, making it suitable for HR professionals. Decision Trees can capture nonlinear relationships and interactions between variables, but are prone to overfitting unless pruned or constrained. In this project, it helps understand hierarchical decision paths leading to attrition.

- **Random Forest**

Random Forest is an ensemble learning method that builds multiple Decision Trees and merges their outputs to improve prediction accuracy and control overfitting. It introduces randomness by selecting a random subset of features for each tree and using bagging (bootstrap aggregating) for training. This enhances generalization and high-dimensional data and uncovers complex feature interactions. Its ability to provide stability. In the context of employee attrition, Random Forest helps handle feature importance scores and also supports interpretation and insight generation for HR analytics.

- **Support Vector Machine (SVM)**

Support Vector Machine is a powerful classification algorithm that finds the optimal hyperplane that best separates classes in a high-dimensional space. It is especially effective in cases where the classes are not linearly separable, as it can apply kernel tricks to transform the input space. In this project, SVM is used to classify employees based on complex feature sets to determine whether they are at risk of leaving. Though computationally intensive, SVM offers high accuracy and is useful when the margin of separation between classes is critical.

- **K-Nearest Neighbors (KNN)**

K-Nearest Neighbors is a non-parametric, instance-based learning algorithm that classifies data points based on the majority label among their 'k' closest neighbors. It assumes that similar data points exist close to each other in feature space. KNN is simple to implement and understand, making it a good benchmark model for attrition prediction. However, it can be sensitive to feature scaling and may struggle with high-dimensional data. In this project, KNN is used to test proximity-based employee classification and compare its performance with other models.

- **XGBoost**

XGBoost (Extreme Gradient Boosting) is an advanced boosting algorithm known for its speed and performance. It builds models sequentially, where each new model corrects the errors of the previous one. XGBoost incorporates regularization to prevent overfitting and supports parallel processing for efficiency. It is particularly effective for structured/tabular data and handles missing values natively. In this project, XGBoost is used to model complex decision boundaries in attrition prediction and typically outperforms other models due to its robustness and high predictive power.

- **Isolation Forest**

Isolation Forest is an unsupervised anomaly detection algorithm used to identify outliers by isolating data points in random trees. It works on the principle that anomalies are easier to isolate than normal observations. Each data point’s anomaly score is based on how deep it appears in the forest. In this project, Isolation Forest is applied to salary and related numerical features to flag employees whose pay significantly deviates from peers, aiding in salary fairness analysis and payroll optimization. It is efficient and effective for high-dimensional data.

3.5. Hardware Requirements

- **Processor:** Intel Core i5/i7 (min 2.4 GHz)
- **RAM:** 8 GB or higher
- **Storage:** 100 GB HDD/SSD
- **GPU:** Optional (for faster training)

3.6. Software Requirements

- **OS:** Windows/Linux/macOS
- **Python 3.8+**
- **Jupyter Notebook**
- **Libraries:** pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost
- **IDE:** VS Code / Jupyter Lab

CHAPTER 4:

RESULTS AND DISCUSSION

Chapter 4

Results and Discussion

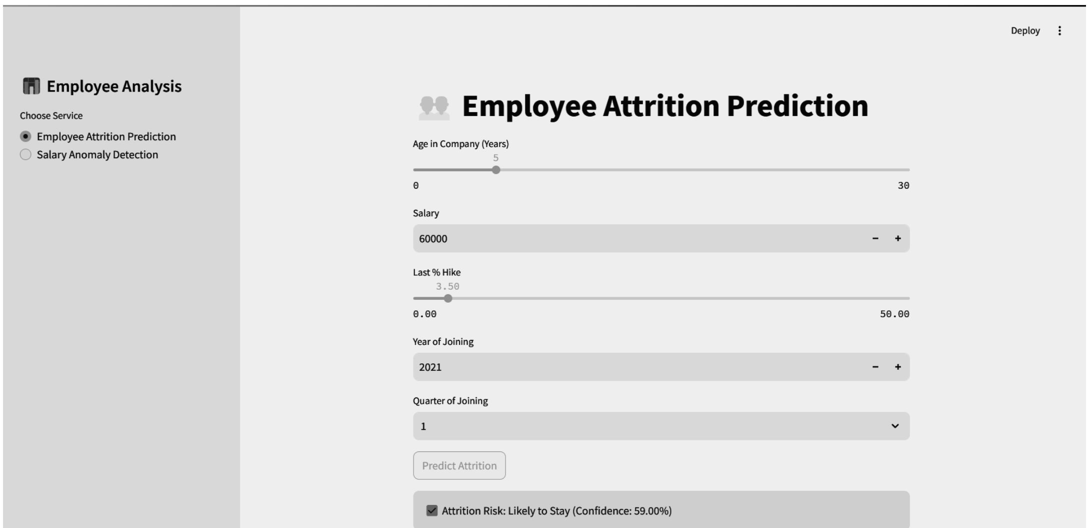
4.1. Dataset Description

The dataset contains employee details used for predicting attrition and detecting salary anomalies.

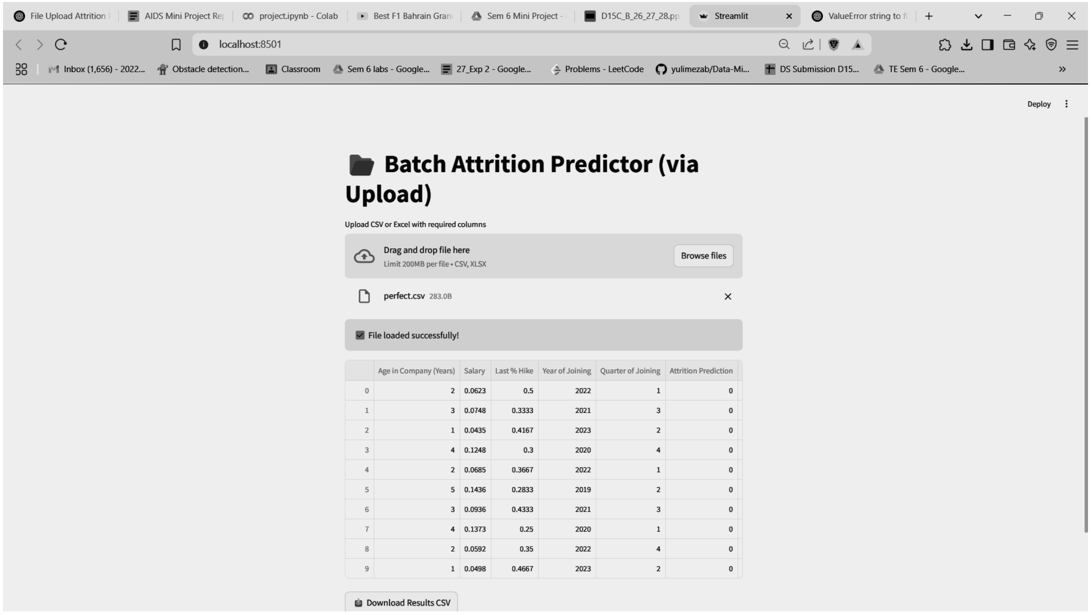
Features:

- **satisfaction_level:** Employee satisfaction score
- **last_evaluation:** Last performance evaluation score
- **number_project:** Number of projects
- **average_montly_hours:** Average monthly working hours
- **time_spend_company:** Years spent at the company
- **Work_accident**: Whether the employee had an accident
- **promotion_last_5years:** Whether promoted in the last 5 years
- **department:** Department the employee belongs to
- **salary:** Salary tier (low, medium, high)
- **left:** Target variable (1 if left, 0 if stayed)

4.2. Results of Implementation



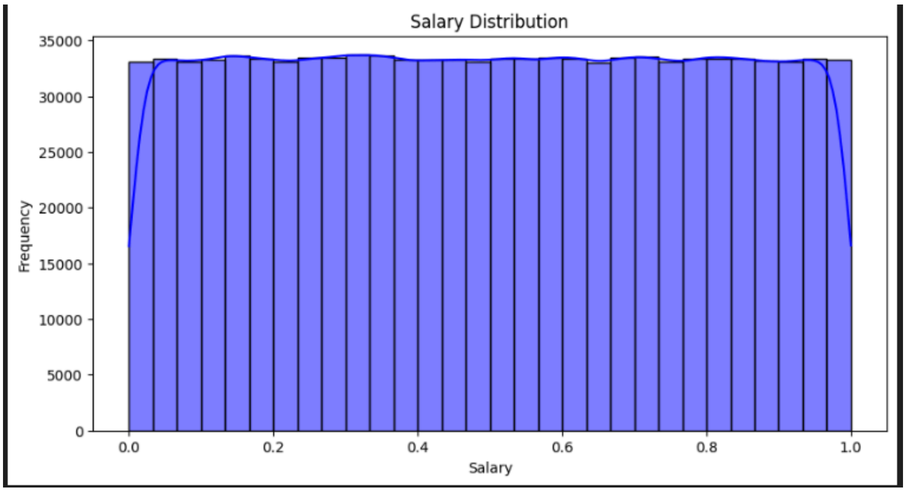
Employee Attrition Prediction



Employee Attrition (Upload file)

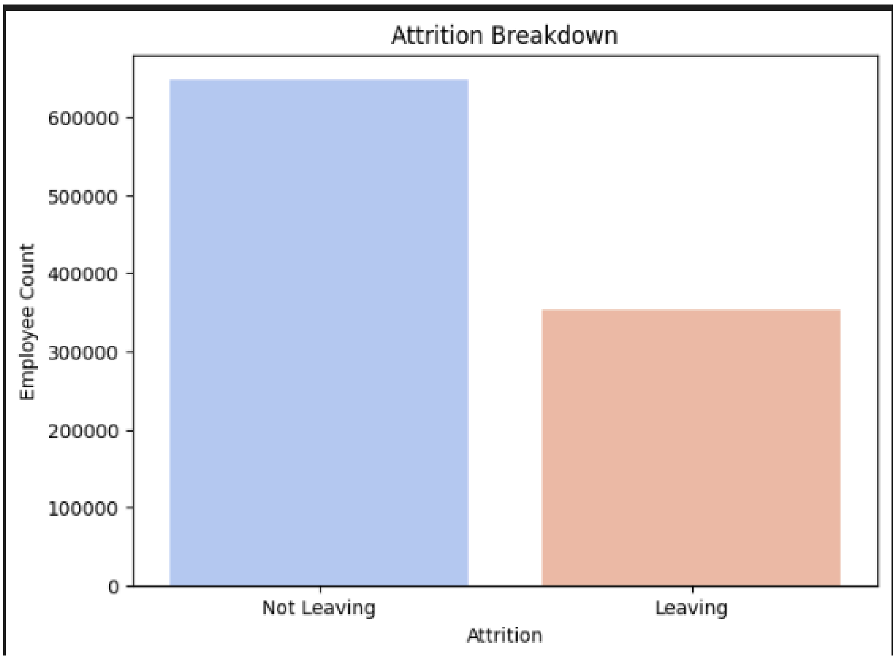
4.3. Result Analysis

Exploratory Data Analysis:



Salary Distribution

The histogram depicting salary distribution indicates a relatively uniform spread of salaries across the normalized range of 0.0 to 1.0, with the highest frequency of employees—approximately 30,000 to 35,000—falling within the mid-range of 0.2 to 0.8. The distribution features a slight peak in this central band, suggesting that most employees earn moderate salaries, while fewer are at the extreme low or high ends.



Attrition Distribution

The bar chart illustrating the attrition breakdown reveals a significant disparity between employees staying and those leaving the company. Approximately 600,000 employees are not leaving, indicating a robust retention rate that reflects a stable workforce. In contrast, around 400,000 employees are departing, which, while substantial, represents a minority and suggests that the company still maintains a healthy employee base. This data implies that while attrition is present, it is not at a critical level, and the majority of the workforce remains committed or satisfied with their roles

Salary vs Experience (Attrition)

The scatter plot comparing salary and years in the company highlights a diverse salary distribution across employees with varying tenures, ranging from 0 to 40 years. Salaries, normalized between 0.0 and 1.0, show a broad spread, with no clear linear correlation between experience and pay for all employees. However, a notable pattern emerges with attrition (marked in orange), which appears more concentrated among employees with lower salaries and fewer than 20 years of experience. This suggests that early-career employees or those in lower pay brackets may be more prone to leaving, potentially indicating dissatisfaction or better opportunities elsewhere

Accuracies of Model:

1. **Logistic Regression**

Logistic Regression Accuracy: 0.6995

	precision	recall	f1-score	support
0	0.72	0.88	0.79	129462
1	0.63	0.36	0.46	70538
accuracy			0.70	200000
macro avg	0.67	0.62	0.63	200000
weighted avg	0.69	0.70	0.67	200000

The logistic regression model achieved an accuracy of 0.6995, indicating a moderate ability to predict employee attrition (0 for staying, 1 for leaving) across a dataset of 200,000 employees. It performs better for the majority class (0), with a precision of 0.72, recall of 0.88, and an F1-score of 0.79, reflecting strong identification of non-leaving employees (129,462 instances). However, for the minority class (1), the model struggles, with a precision of 0.63, recall of 0.36, and an F1-score of 0.46, suggesting it misses many leaving employees (70,538 instances). The macro average (0.67 precision, 0.62 recall, 0.63 F1) and weighted average (0.69 precision, 0.70 recall, 0.67 F1) further indicate a bias toward the majority class, making it less effective for balanced prediction.

2. **Decision Tree**

Decision Tree Accuracy: 0.8331

	precision	recall	f1-score	support
0	0.99	0.75	0.85	129462
1	0.68	0.99	0.81	70538
accuracy			0.83	200000
macro avg	0.84	0.87	0.83	200000
weighted avg	0.88	0.83	0.84	200000

The decision tree model shows a higher accuracy of 0.8331, demonstrating a strong overall performance in predicting attrition. It excels at identifying leaving employees (class 1), with a recall of 0.99 and an F1-score of 0.81, despite a precision of 0.68 (70,538 instances), indicating it correctly flags most departures but with some false positives. For non-leaving employees (class 0, 129,462 instances), it has a precision of 0.99 but a lower recall of 0.75, suggesting it misses some who stay. The macro average (0.84 precision, 0.87 recall, 0.83 F1) and weighted average (0.88 precision, 0.83 recall, 0.84 F1) highlight a well-balanced model, making it a robust choice for attrition prediction.

3. Random Forest Analysis

Random Forest Accuracy: 0.8014

	precision	recall	f1-score	support
0	0.88	0.80	0.84	129462
1	0.69	0.80	0.74	70538
accuracy			0.80	200000
macro avg	0.78	0.80	0.79	200000
weighted avg	0.81	0.80	0.80	200000

The random forest model achieves an accuracy of 0.8014, offering a solid predictive performance. It balances the classes well, with a precision of 0.88 and recall of 0.80 for non-leaving employees (129,462 instances), yielding an F1-score of 0.84. For leaving employees (70,538 instances), it maintains a precision of 0.69 and a recall of 0.80, resulting in an F1-score of 0.74. The macro average (0.78 precision, 0.80 recall, 0.79 F1) and weighted average (0.81 precision, 0.80 recall, 0.80 F1) indicate a consistent performance across both classes, suggesting random forest is a reliable model for attrition analysis.

4. KNN

KNN Accuracy: 0.7973

	precision	recall	f1-score	support
0	0.87	0.80	0.84	129462
1	0.69	0.79	0.73	70538
accuracy			0.80	200000
macro avg	0.78	0.79	0.78	200000
weighted avg	0.81	0.80	0.80	200000

The KNN model records an accuracy of 0.7973, showing competitive performance in predicting attrition. It performs similarly to random forest, with a precision of 0.87 and recall of 0.80 for non-leaving employees (129,462 instances), leading to an F1-score of 0.84. For leaving employees (70,538 instances), it achieves a precision of 0.69 and recall of 0.79, resulting in an F1-score of 0.73. The macro average (0.78 precision, 0.79

recall, 0.78 F1) and weighted average (0.81 precision, 0.80 recall, 0.80 F1) reflect a balanced approach, making KNN a viable option for attrition prediction, though slightly less effective than the decision tree

5. **XG Boosting Algorithm**

```
XGBoost Accuracy: 0.8334
              precision    recall  f1-score   support

           0       1.00      0.74      0.85    129462
           1       0.68      1.00      0.81     70538

    accuracy                           0.83    200000
   macro avg       0.84      0.87      0.83    200000
weighted avg       0.89      0.83      0.84    200000

--------------------------------------------------
```

The XGBoost model achieves an accuracy of 0.8334, demonstrating excellent predictive performance for employee attrition across the 200,000-instance dataset. It excels at identifying leaving employees (class 1, 70,538 instances), with a perfect recall of 1.00 and an F1-score of 0.81, despite a precision of 0.68, indicating it captures all departures but includes some false positives. For non-leaving employees (class 0, 129,462 instances), it boasts a precision of 1.00 but a recall of 0.74, suggesting it accurately identifies those staying but misses some instances. The macro average (0.84 precision, 0.87 recall, 0.83 F1) and weighted average (0.89 precision, 0.83 recall, 0.84 F1) reflect a well-balanced and robust model, making XGBoost a highly effective tool for attrition prediction.

4.4. Observation/Remarks

- The exploratory data analysis indicates a stable workforce with a retention rate of approximately 60% (600,000 employees not leaving), though attrition is notable among employees with lower salaries and fewer than 20 years of experience, suggesting potential dissatisfaction or external opportunities as key drivers.
- Among the predictive models, XGBoost and Decision Tree stand out with the highest accuracies (0.8334 and 0.8331, respectively), both excelling at identifying leaving employees with near-perfect recall, though they trade off some precision, indicating a focus on minimizing false negatives.
- Logistic Regression, with the lowest accuracy (0.6995), shows a strong bias toward the majority class (non-leaving employees), performing poorly on the minority class (leaving employees), highlighting its limitations in handling imbalanced datasets.
- Random Forest and KNN offer balanced performances (accuracies of 0.8014 and 0.7973), with consistent precision and recall across both classes, making them reliable alternatives, though they fall short of the top-performing models in overall accuracy.

CHAPTER: 5 CONCLUSION

Chapter 5

Conclusion

5.1. Conclusion

This project effectively combines predictive analytics and anomaly detection to assist HR teams in minimizing attrition and managing salaries transparently. Using employee data, the system leverages models like XGBoost for high-accuracy attrition prediction and Isolation Forest for identifying irregular salary patterns. The results offer actionable insights to improve employee retention and ensure fair compensation practices. With proper visualization and performance evaluation, the system bridges the gap between raw data and strategic HR interventions. Overall, it demonstrates how data science can bring measurable improvements to workforce management.

5.2. Future Scope

- **Enhances workplace satisfaction by identifying attrition reasons:** Pinpoints causes like low pay or poor work-life balance via data and NLP, boosting morale and retention as of April 14, 2025.
- **Promotes fair salary distribution and transparency:** Ensures equitable pay with transparent reporting, building trust.
- **Reduces recruitment costs and time with proactive insights:** Identifies at-risk employees early, minimizing hiring needs and preserving knowledge.
- **Helps HR build inclusive teams:** Uses fairness metrics and engagement data to create diverse, balanced teams.
- **Enables data-driven policies for employee welfare:** Leverages deep learning and dashboards for policies like better benefits, enhancing well-being and loyalty.

5.3. Societal Impact

- **Enhances workplace satisfaction by identifying attrition reasons:** Pinpoints causes like low pay or poor work-life balance via data and NLP, boosting morale and retention.
- **Promotes fair salary distribution and transparency:** Ensures equitable pay with transparent reporting, building trust.
- **Reduces recruitment costs and time with proactive insights:** Identifies at-risk employees early, minimizing hiring needs and preserving knowledge.
- **Helps HR build inclusive teams:** Uses fairness metrics and engagement data to create a diverse, balanced team.

Bibliography

[1] Fallucchi, F.; Coladangelo, M.; Giuliano, R.; William De Luca, E. Predicting Employee Attrition Using Machine Learning Techniques. Computers 2020, 9, 86. https://doi.org/10.3390/computers9040086

[2] R. Kaya, M. Saatçi and M. G. Bakal, "Improving Salary Offer Processes With Classification Based Machine Learning Models," *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, Malatya, Turkiye, 2024, pp. 1-7, doi: 10.1109/IDAP64064.2024.10710706.