HR Analytics for Employee Attrition using ML

Submitted in partial fulfillment of the requirements of the degree of

BACHELOR OF COMPUTER ENGINEERING

by

Virag Hote - 21102057

Deepak Battula - 21102006

Akshada Gupta - 21102190

Vinay Hirap - 21102100

Guide:

Prof. Shamika Mule

Department of Computer Engineering

A. P. SHAH INSTITUTE OF TECHNOLOGY, THANE

(2024-2025)

A. P. SHAH INSTITUTE OF TECHNOLOGY, THANE

CERTIFICATE

This is to certify that the Major Project entitled "**HR Analytics for Employee Attrition using ML**" is a bonafide work of **"Virag Hote (21102057), Deepak Battula (21102006), Akshada Gupta (21102190), Vinay Hirap (21102100)"** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **Bachelor of Engineering** in **Computer Engineering.**

Prof. Shamika Mule
Guide

Prof. Deepak Khachane
Project Coordinator

Dr. Sachin H.Malve
Head of Department

Dr. Uttam D. Kolekar
Principal

A. P. SHAH INSTITUTE OF TECHNOLOGY, THANE

Project Report Approval for Major Project

This project report entitled "**HR Analytics for Employee Attrition using ML*"*** by ***Virag Hote, Deepak Battula, Akshada Gupta, Vinay Hirap*** is approved for the partial fulfillment of the degree of ***Bachelor of Engineering*** in ***Computer Engineering***, ***2024-25***.

Examiner Name Signature

1. _____

2. _____

Date:

Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, We have adequately cited and referenced the sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Virag Hote- 21102057

Deepak Battula- 21102006

Akshada Gupta- 21102190

Vinay Hirap- 21102100

Date:

Abstract

Employee attrition is a critical issue for organizations as it disrupts business operations, increases hiring costs, and diminishes workforce morale. The ability to predict which employees are at risk of leaving is crucial for human resource managers to implement proactive retention strategies. Traditional methods of managing attrition often rely on surveys and manual assessments, which are time-consuming and lack accuracy. Thus, there is a growing need for automated systems that leverage data-driven approaches to forecast employee turnover with greater precision.

This project proposes an Employee Attrition Prediction System that employs the Random Forest machine learning algorithm, known for its robustness in handling complex datasets and achieving high accuracy. Using the IBM HR Analytics Employee Attrition dataset, which contains various employee attributes such as job role, monthly income, job satisfaction, and years of experience, the model is trained to identify patterns that indicate potential resignation. The dataset's features are analyzed to determine their correlation with attrition, ensuring that the system captures both job-related and personal factors affecting turnover.

The developed system provides HR managers with actionable insights into which employees are most at risk and why, allowing for timely intervention. By understanding critical factors such as low job satisfaction, lack of career advancement, and work-life imbalance, companies can tailor retention programs to individual needs, improving overall employee satisfaction. This predictive approach not only reduces turnover but also enhances organizational productivity, reduces recruitment costs, and supports long-term workforce planning.

Keywords:

Employee Attrition, Machine Learning, Random Forest, IBM Employee Dataset, Turnover Prediction, Employee Retention, HR Analytics, Workforce Optimization.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Employee attrition is a growing concern for many organizations. When talented employees leave, it not only disrupts operations but also incurs substantial costs associated with recruiting and training replacements. More importantly, high turnover rates can negatively impact team morale and productivity, making it essential for companies to identify and address the root causes early on. This is where data-driven solutions can play a transformative role in proactively managing employee retention.

Traditional methods of understanding employee turnover, such as surveys and exit interviews, often fall short as they are conducted post-departure and lack real-time insights. In contrast, leveraging machine learning models can help organizations predict and mitigate attrition risks before they become a serious issue. By analyzing patterns in employee data, machine learning can identify which factors—be it job role, compensation, or work environment—contribute the most to employees' decisions to leave.

This project focuses on building an Employee Attrition Prediction System using the Random Forest algorithm, which is particularly effective in handling complex data. By utilizing the IBM HR Analytics Employee Attrition dataset, the system aims to pinpoint the characteristics of employees at risk of leaving, providing HR teams with valuable insights. This approach empowers organizations to shift from reactive to proactive strategies, reducing attrition rates and promoting a healthier work environment. Ultimately, such predictive systems not only help retain top talent but also foster a more engaged and satisfied workforce.

Chapter 2

Literature Survey

The paper examines employee attrition prediction using HR analytics and machine learning techniques, comparing Decision Trees, Random Forests, and XGBoost. XGBoost is highlighted as the most effective model, significantly enhancing prediction accuracy for better retention strategies. [1]

The paper compares machine learning models for predicting employee attrition, evaluating algorithms like Decision Trees and Logistic Regression, highlighting the most effective model for workforce management based on key performance metrics. [2]

The paper "Employee Attrition Analysis Using XGBoost" (2024) by Isha Nitin Thapliyal and Sheetal Solanki applies the XGBoost algorithm to predict employee attrition, identifying key factors and improving retention strategies. [3]

The paper by Rahman, Islam, and Bala (2024) analyzes employee retention factors using machine learning models, identifying key predictors and patterns to improve retention strategies in organizations. [4]

The paper explores predicting employee attrition rates using various machine learning classifiers, comparing their accuracy to identify factors influencing attrition and optimize retention strategies for organizations. [5]

The paper explores employee attrition prediction using HR analytics and various machine learning models, including Decision Trees, Random Forests, CNNs, SVMs, and XGBoost, identifying XGBoost as highly effective for enhancing prediction accuracy and retention strategies. [6]

Research Paper	Analysis
Predicting Employee Attrition through HR Analytics: A Machine Learning Approach [1] *Dr. Pooja Nagpal, Dr. Avinash Pawar, Dr. Sanjay*	The paper examines employee attrition prediction using HR analytics and machine learning techniques, comparing Decision Trees, Random Forests, and XGBoost. XGBoost is highlighted as the most effective model, significantly enhancing prediction accuracy for better retention strategies.
Comparative Analysis of Machine Learning Models on Employees' Attrition Prediction [2] *Christianah O. Akinduyite, Abiodun Oguntimilehin, Bukola Badeji-Ajisafe, Stephen E. Obamiyi*	The paper compares machine learning models for predicting employee attrition, evaluating algorithms like Decision Trees and Logistic Regression, highlighting the most effective model for workforce management based on key performance metrics.
Employee Attrition Analysis Using XGBoost [3] *Isha Nitin Thapliyal, Sheetal Solanki*	The paper "Employee Attrition Analysis Using XGBoost" (2024) by Isha Nitin Thapliyal and Sheetal Solanki applies the XGBoost algorithm to predict employee attrition, identifying key factors and improving retention strategies.
Analyzing Employee Retention Factors using Machine Learning [4] *Moshiur Rahman, Md Rashedul Islam, Partho Bala*	The paper by Rahman, Islam, and Bala (2024) analyzes employee retention factors using machine learning models.

Classification and Prediction of Employee Attrition Rate using Machine Learning Classifiers [5] *Umang Garg, Neha Gupta, Mahesh Manchanda*	The paper explores predicting employee attrition rates using various machine learning classifiers, comparing their accuracy to identify factors influencing attrition and optimize retention strategies for organizations.
Prediction of Employee Attrition Using Stacked Ensemble Method [6] *Sanjay Gowdru, Suyash Kumar Dubli, Pooja Agarwal, Bhoomika*	The paper explores employee attrition prediction using HR analytics and various machine learning models, including Decision Trees, Random Forests, CNNs, SVMs, and XGBoost, identifying XGBoost as highly effective for enhancing prediction accuracy and retention strategies.

Chapter 3

Limitations of the Existing System

1. Limited Model Generalization:

Many existing employee attrition models lack the ability to generalize effectively across diverse industries or job roles, leading to reduced prediction accuracy in varying organizational contexts.

2. Inadequate Handling of Imbalanced Datasets:

Most models struggle to manage class imbalances, where the number of employees staying significantly outweighs those leaving, resulting in biased predictions and overlooking minority classes.

3. Focus on Static Factors Only:

Current systems often focus on static features like salary and job role, neglecting dynamic factors such as employee engagement or real-time performance metrics, limiting comprehensive analysis.

4. Absence of Contextual Analysis:

Existing systems fail to incorporate contextual analysis, such as personal reasons or external market trends, reducing the effectiveness of predictions and insights.

5. Lack of Proactive Retention Strategies:

Most models only predict attrition but do not provide recommendations or strategies for HR managers to proactively address the identified risk areas.

6. High Dependency on Predefined Datasets:

Models heavily rely on specific datasets like the IBM HR dataset, leading to limited applicability and reduced accuracy when applied to real-world employee data.

7. Limited Interpretability of Predictions:

The models' decisions are often complex and not easily interpretable by HR personnel, making it difficult to justify interventions based on predicted attrition risks.

8. Poor Integration with HR Management Systems:

There is a lack of seamless integration with existing HR tools, hindering real-time data updating and analysis, which reduces the utility of the prediction system.

9. Insufficient Legal and Ethical Safeguards:

Current models do not account for ethical considerations and data privacy regulations, posing compliance challenges when handling sensitive employee data.

Chapter 4

Problem Statement, Objective & Scope

Problem Statement: -

Employee attrition poses a significant challenge for organizations, impacting productivity, morale, and financial stability. The increasing turnover rates necessitate effective predictive measures to identify employees at risk of leaving. Current attrition prediction models often rely on traditional statistical methods and do not fully harness the power of advanced machine learning techniques. Furthermore, many existing systems struggle with class imbalance and do not effectively analyze dynamic factors influencing employee decisions to resign. This inadequacy hinders organizations from implementing proactive retention strategies. The IBM HR Analytics Employee Attrition dataset offers a rich source of employee-related information, but existing models lack the sophistication to analyze it thoroughly. This project aims to utilize the Random Forest algorithm to develop a robust employee attrition prediction model that can accurately identify key factors influencing turnover. By addressing these gaps, organizations can better understand their workforce dynamics and implement targeted interventions to enhance employee retention and engagement.

Objective: -

- Develop a predictive model for employee attrition using the Random Forest algorithm.
- Utilize the IBM Employee Dataset to analyze key employee attributes.
- Identify significant factors influencing employee turnover through advanced data analysis.
- Enhance prediction accuracy and reliability compared to existing methods.
- Provide actionable insights for HR managers to implement effective retention strategies.
- Enable organizations to tailor interventions for at-risk employees.

Scope: -

- **Real-time Attrition Prediction**: Provides immediate insights into employee attrition risks, facilitating timely interventions by HR managers.
- **Effective Feature Identification**: Utilizes the Random Forest model to automatically identify key factors influencing attrition, such as salary and job satisfaction.
- **Data Cleaning and Preprocessing**: Ensures the IBM Employee dataset is preprocessed for accurate analysis, enhancing the model's performance.
- **Dynamic Risk Assessment**: Adapts to real-time changes in employee behavior, enabling proactive management of potential turnover.
- **Comprehensive Employee Insights**: Combines quantitative data analysis with qualitative insights from employee feedback, offering a holistic view of attrition causes.
- **Ethical Considerations and Data Privacy**: Addresses legal and ethical safeguards in handling sensitive employee data, ensuring compliance with regulations.

Chapter 5

Proposed System Architecture

The proposed system architecture for the Employee Attrition Prediction System integrates the Random Forest machine learning model with the IBM HR Analytics Employee Attrition dataset to address the limitations of existing attrition prediction methods. This architecture aims to provide accurate predictions and actionable insights to HR managers to enhance employee retention strategies.

1. Data Collection and Preprocessing:

The system begins with the collection of employee data from various sources, including the IBM HR dataset. This data undergoes preprocessing, which includes cleaning, normalizing, and handling missing values. Features such as job role, monthly income, years at the company, and other relevant attributes are extracted to ensure the dataset is ready for analysis.

2. Feature Selection:

Feature selection techniques are applied to identify the most significant predictors of employee attrition. This process reduces dimensionality and focuses on the most impactful variables, such as job satisfaction, overtime, and performance metrics, ensuring that the model learns from the most relevant information.

3. Random Forest Model Implementation:

The Random Forest algorithm is implemented to create a robust predictive model. This ensemble learning method builds multiple decision trees based on random subsets of the data and features, aggregating their predictions to improve accuracy and reduce overfitting.

4. Model Training and Validation:

The model is trained using historical employee data, and its performance is validated using techniques such as cross-validation to ensure generalizability. Metrics such as accuracy, precision, and recall are used to evaluate the model’s effectiveness in predicting attrition.

5. Predictive Insights and Reporting:

Once trained, the system generates predictive insights, identifying employees at risk of leaving. These insights are presented in a user-friendly dashboard for HR managers, allowing for easy interpretation and informed decision-making.

6. Integration with HR Management Systems:

The prediction system is designed to integrate seamlessly with existing HR management tools, facilitating real-time data updates and ongoing analysis. This integration enables HR personnel to monitor employee trends and adjust retention strategies proactively.

7. Legal and Ethical Compliance:

The architecture ensures that all data handling and analysis comply with legal and ethical standards, protecting employee privacy and maintaining data integrity throughout the process.

8. User Accessibility and Support:

Finally, the system includes user support features, such as documentation and training resources, to ensure that HR managers can effectively utilize the predictive model and insights to enhance employee engagement and retention efforts.

- **Architecture / Block Diagram**

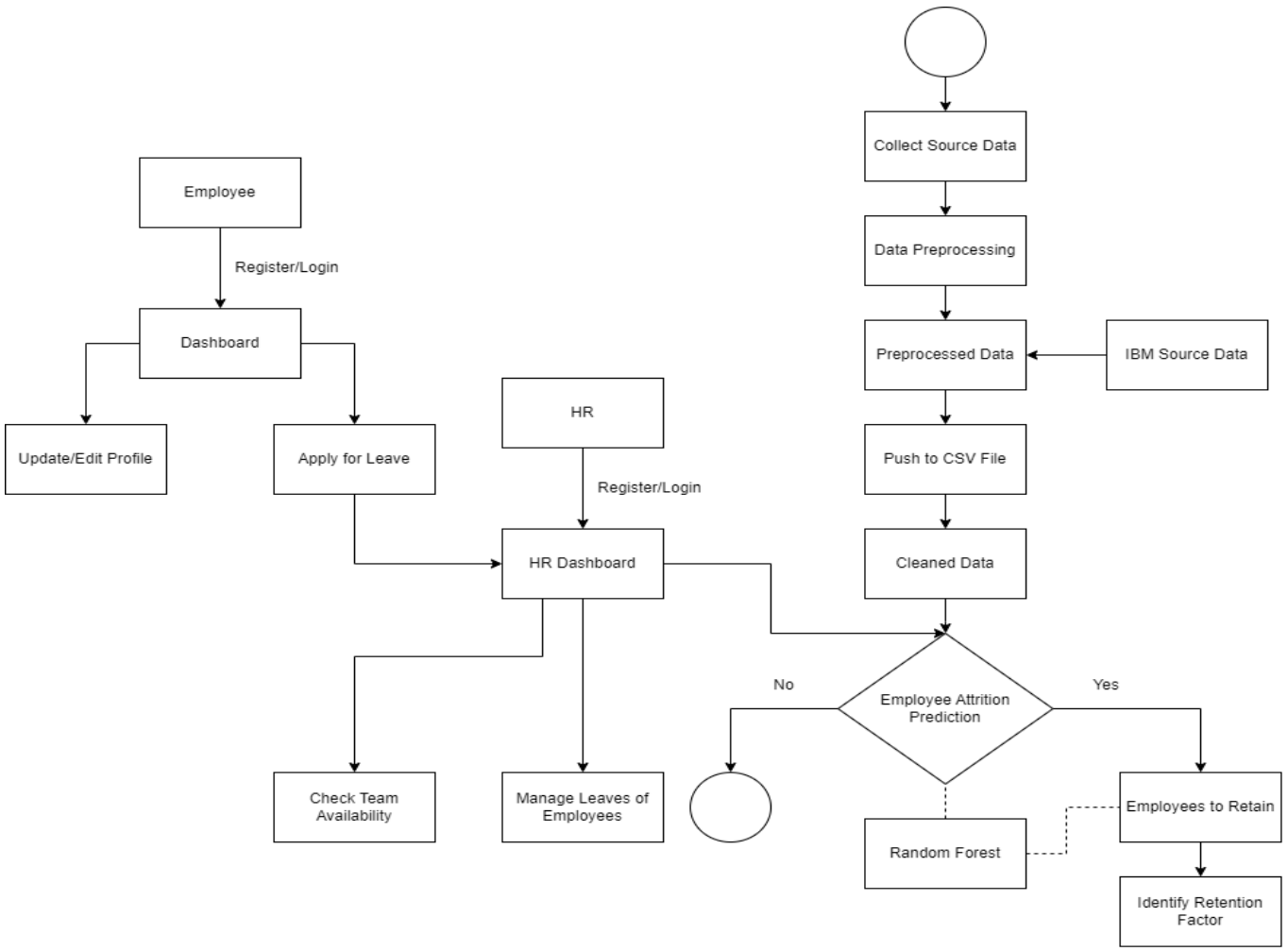

Fig 5.1: Architecture Diagram

The diagram shows how data is preprocessed and features are selected before being fed into the Random Forest model. It also highlights the integration of HR management systems and how predictive insights are generated and displayed to HR managers. This figure emphasizes the modular and scalable nature of the system, ensuring a seamless flow of information.

- **Data Flow Diagram (Level 0, Level 1 & Level 2)**

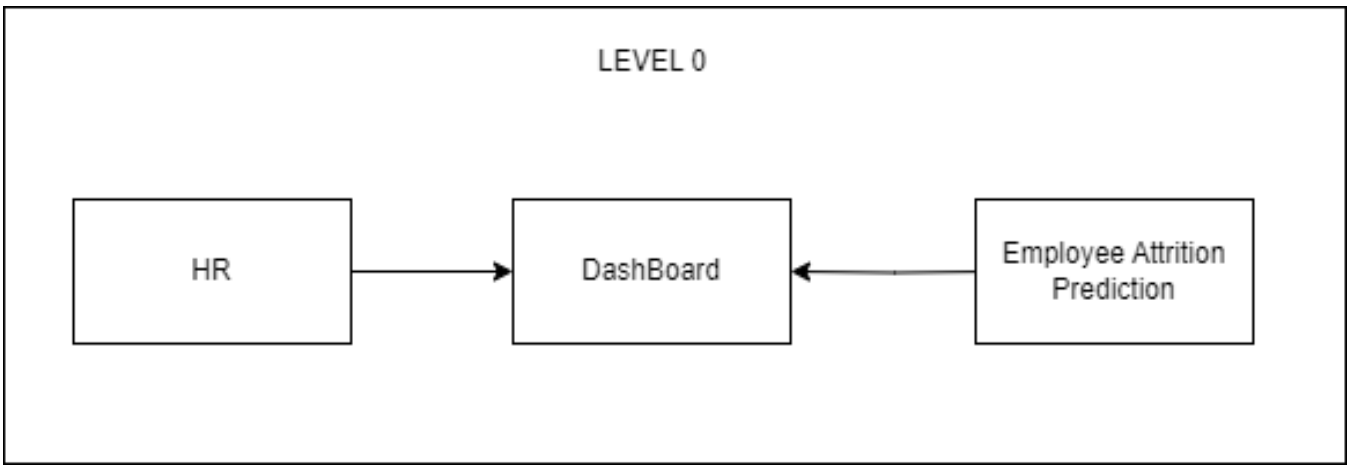

Fig 5.2: Level 0

The Level 0 Data Flow Diagram provides an overview of the entire system's functionality. It shows the major data inflows and outflows, including how employee data is collected, processed, and used to predict attrition. The diagram simplifies the process into its primary components—data input, processing, and output—offering a high-level view of how the system operates without getting into technical specifics.

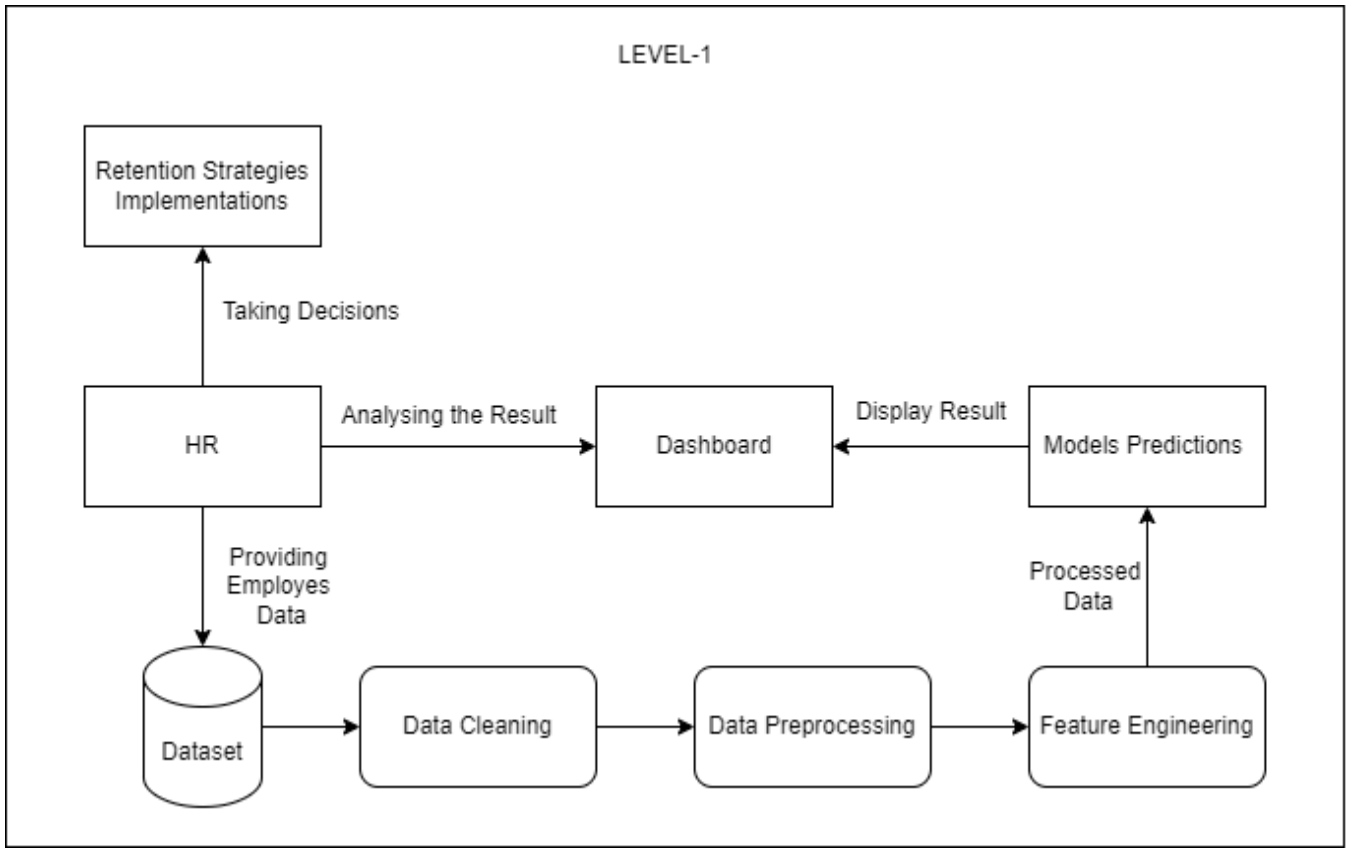

Fig 5.3: Level 1

This diagram takes a closer look at the data flow, offering more granularity compared to the Level 0 DFD. It details how different components within the system interact, such as how data is passed between modules like feature selection, model training, and prediction generation. By breaking down the main processes, it helps to understand how each part of the system contributes to the overall functionality.

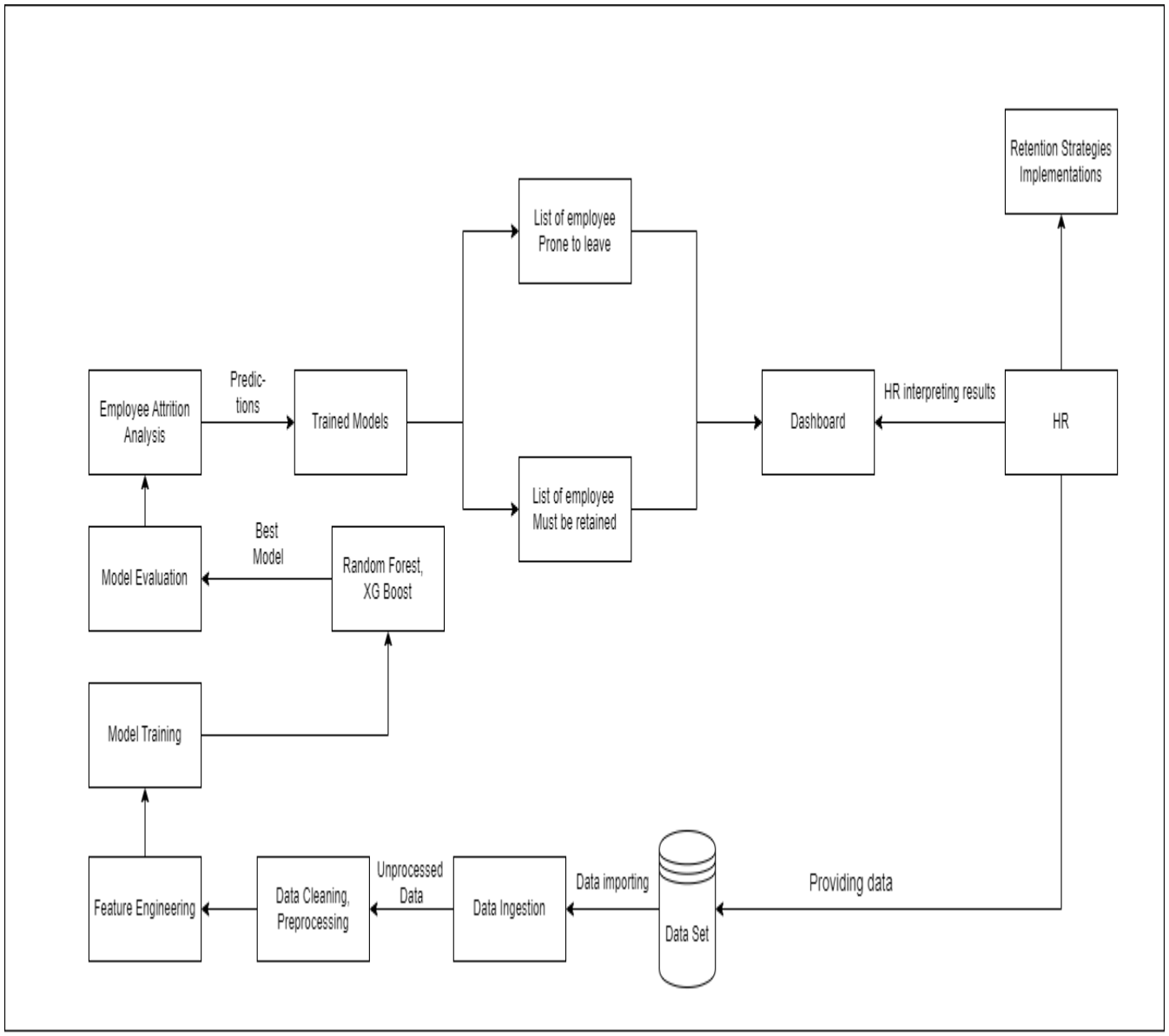

Fig 5.4: Level 2

The Level 2 DFD provides an in-depth look at the system's internal operations. It decomposes the activities into even smaller subprocesses, illustrating the fine details of how data is validated, cleaned, and analyzed before generating a prediction. This level of detail is useful for developers and system architects to understand the technical workflow and ensure every aspect of the system is covered efficiently.

- **Use Case Diagram**

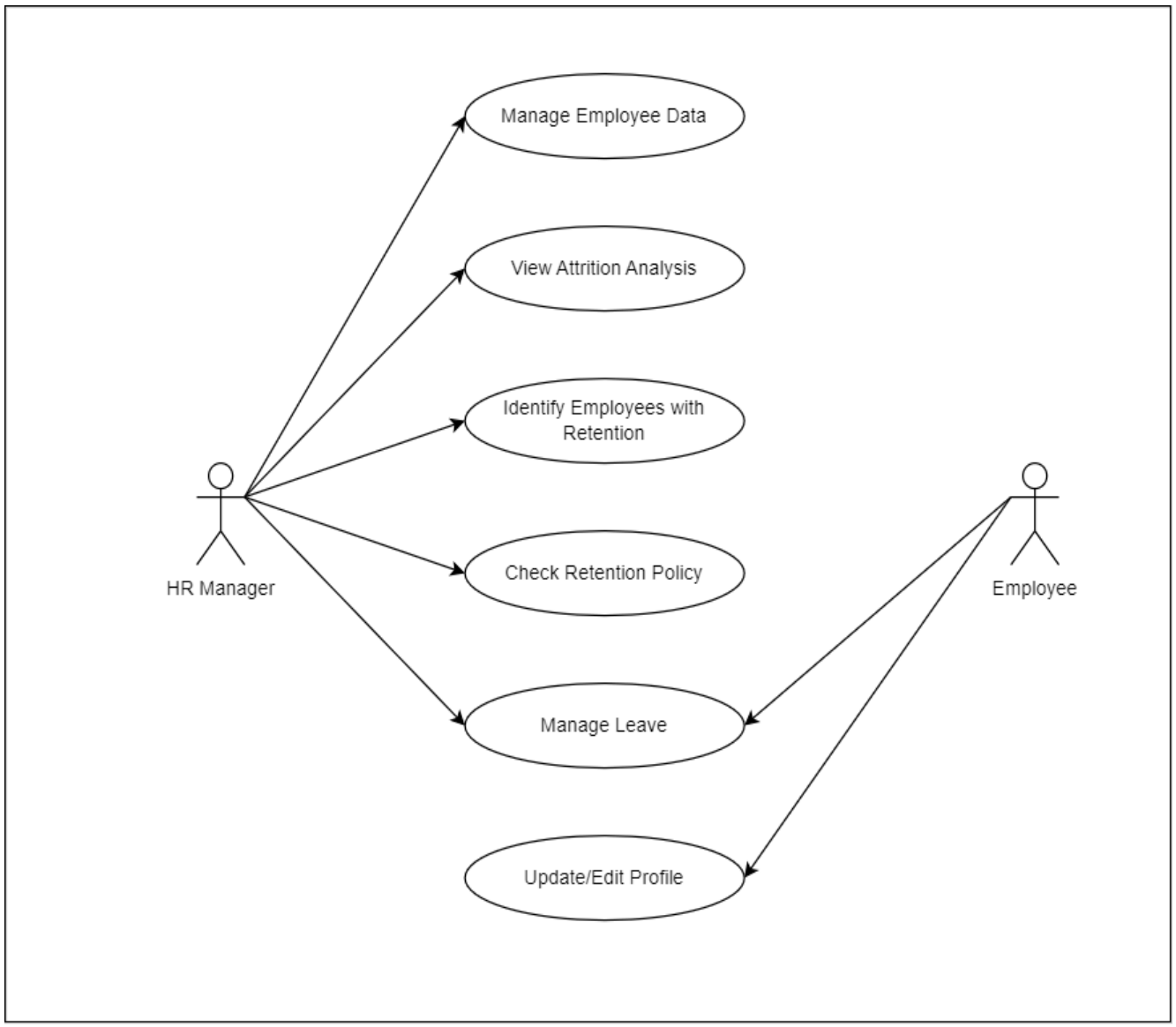

Fig 5.5: Use Case Diagram

The Use Case Diagram visualizes how the system interacts with different actors, specifically HR managers. It outlines the various actions that users can perform within the system, such as uploading employee data, generating attrition reports, and viewing predictive insights. This figure emphasizes the functionalities available to the user, helping to clarify the system's requirements from a user perspective.

- **Sequence Diagram**

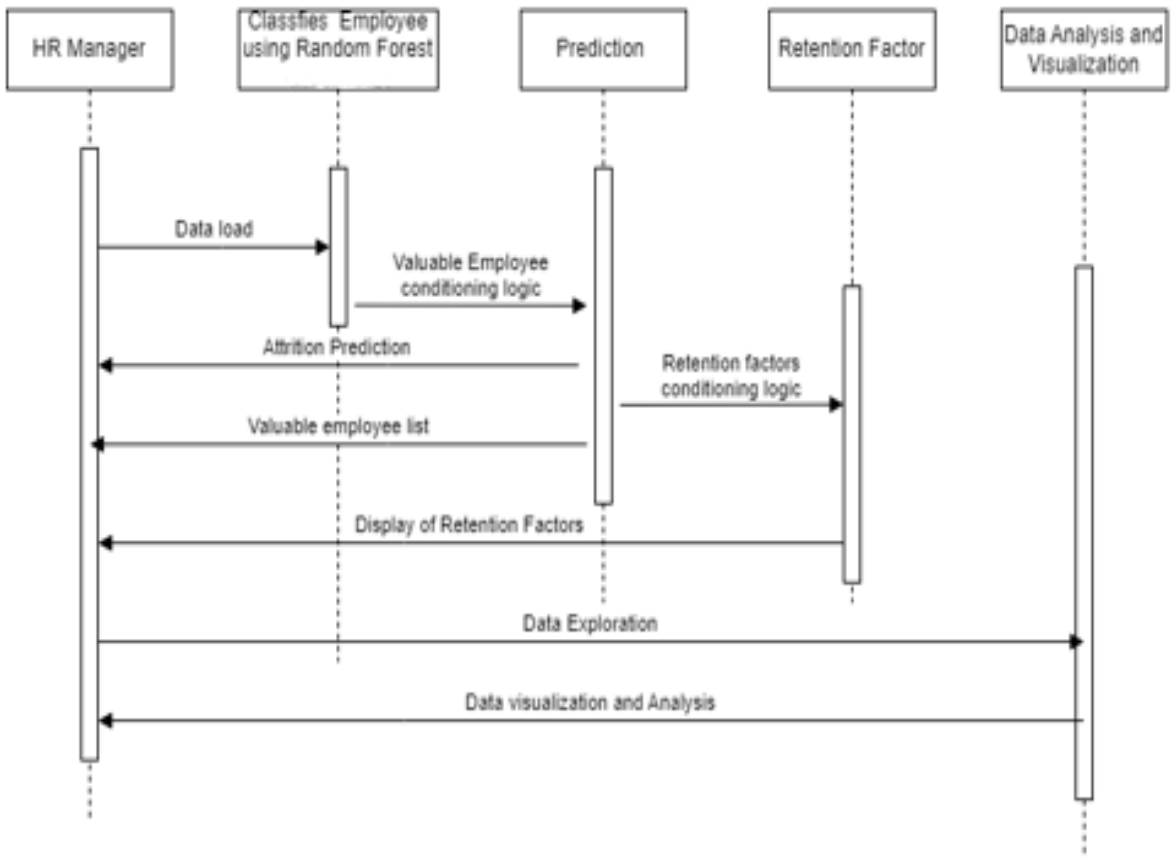

Fig 5.6: Sequence Diagram

The Sequence Diagram shows the order in which operations take place within the system when interacting with the user. It explains how data flows between components in a step-by-step manner, from user input to the system processing the request and generating a prediction. The diagram is useful for understanding how the system behaves dynamically, ensuring that the interaction between components happens in the correct sequence.

- **Activity Diagram**

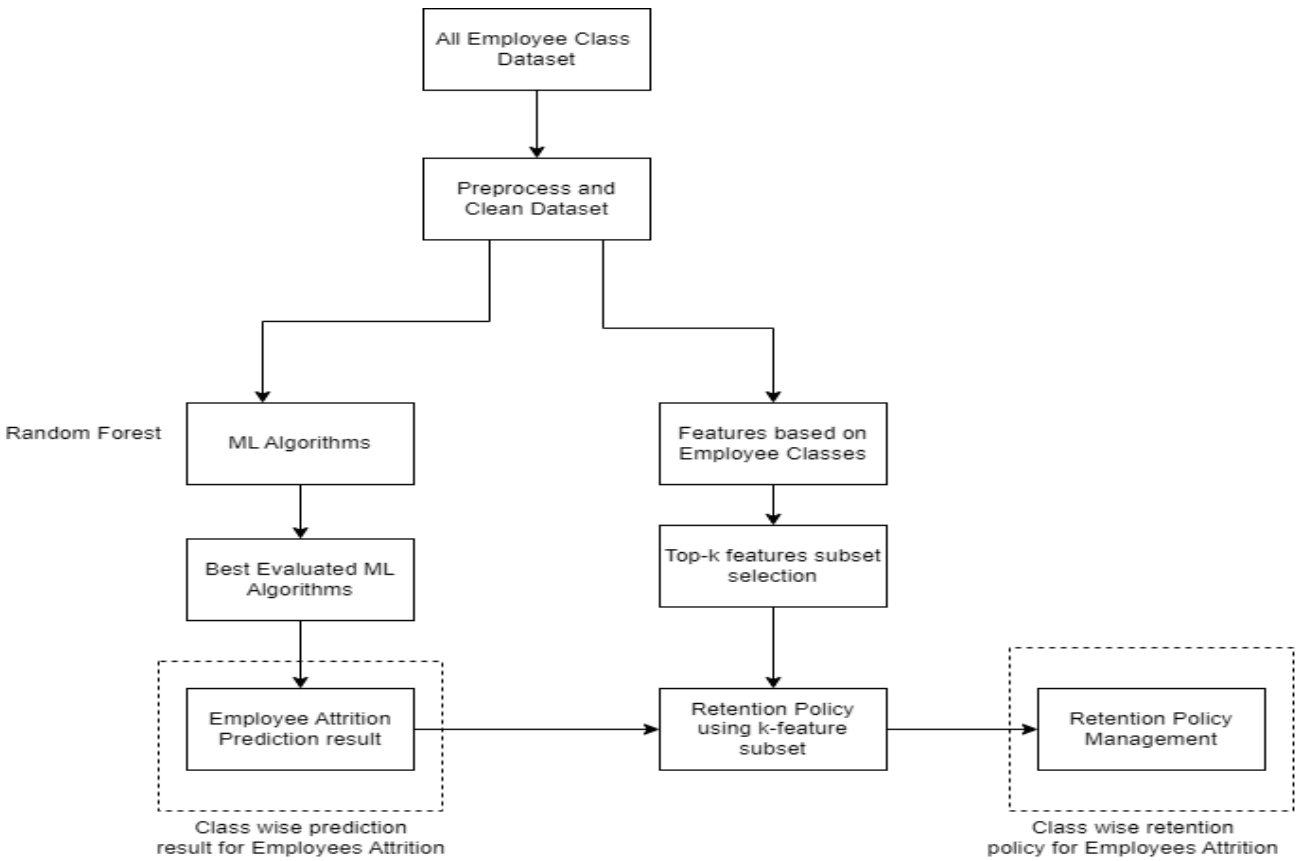

Fig 5.7: Activity Diagram

The Activity Diagram illustrates the various activities involved in the process of predicting employee attrition. It provides a flow of the tasks performed, such as data preprocessing, model execution, and result reporting. The diagram represents the flow of control from one activity to the next, offering a dynamic view of the system's operations. This helps ensure that all steps are connected logically and that the workflow is smooth.

Chapter 6

Experimental Setup

Software Requirements: -

The software requirements for HR Analytics for Employee Attrition using ML include the following:

1) Code Editor (VS Code)

2) Flask

3) React JS

4) Python

5) Javascript

Hardware Requirements: -

Processor - Dual Core Minimum

Ram - 2GB Minimum

Storage - 1GB

Chapter 7

Project Planning

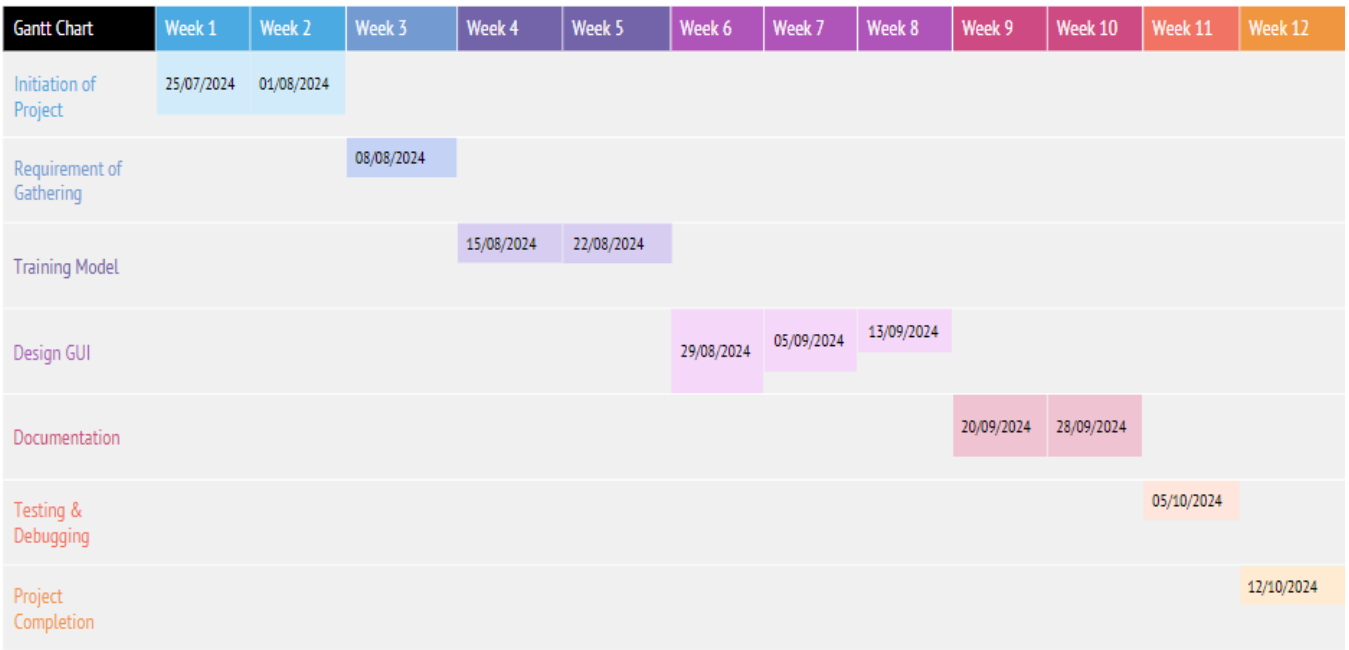

Fig 6.1: Gantt Chart

Chapter 8

Expected Outcomes

Accurate Attrition Prediction: The system will leverage the Random Forest algorithm to predict employee attrition with high accuracy, identifying individuals at risk of leaving based on historical data. This enables HR departments to focus retention efforts on the right employees.

In-depth Employee Insights: The system will analyze key employee attributes such as job satisfaction, monthly income, job role, and overtime hours to provide detailed insights into the factors driving attrition. HR managers can use these insights to understand which specific variables contribute most to turnover.

Proactive Retention Strategies: By identifying at-risk employees early, the system will allow organizations to take preemptive actions, such as offering career development opportunities, improving work-life balance, or adjusting compensation, to reduce turnover and enhance employee satisfaction.

Enhanced Decision-Making for HR: The predictive model will empower HR managers with data-driven decision-making capabilities. It will shift their approach from reactive measures, such as exit interviews, to proactive interventions, improving overall workforce stability and long-term retention.

Reduction in Recruitment Costs: Organizations will be able to lower their recruitment and training expenses by minimizing attrition rates. By retaining more employees, they will save on costs associated with hiring and onboarding new staff, leading to greater financial efficiency.

Real-Time Data Integration: The system will seamlessly integrate with existing HR management platforms, allowing for real-time updates and continuous monitoring of employee data. This ensures that retention strategies can be adjusted dynamically based on the latest workforce trends.

Scalable and Adaptive Model: The Random Forest-based model will be adaptable to various organizational contexts and employee datasets, making it scalable across different industries and job roles. This versatility ensures that the system remains relevant as employee data and attrition factors evolve over time.

Chapter 9

References

[1] Dr. Pooja Nagpa, Dr. Avinash Pawar, Dr. Sanjay, published by: IEEE Explore, "**Predicting Employee Attrition through HR Analytics: A Machine Learning Approach"** (2024).

[2] Christianah O. Akinduyite, Abiodun Oguntimilehin, Bukola Badeji-Ajisafe, Stephen E. Obamiyi, by IEEE Explore, "**Comparative Analysis of Machine Learning Models on Employees' Attrition Prediction"** (2024)

[3] Isha Nitin Thapliyal, Sheetal Solanki, published by: IEEE Explore, "**Employee Attrition Analysis Using XGBoost"** (2024)

[4] Moshiur Rahman, Md Rashedul Islam, Partho Bala, published by: IEEE Explore, "**Analyzing Employee Retention Factors using Machine Learning"** (2024)

[5] Umang Garg, Neha Gupta, Mahesh Manchanda, published by: Research gate, "**Classification and Prediction of Employee Attrition Rate using Machine Learning Classifiers"** (2024)

[6] Sanjay Gowdru, Suyash Kumar Dubli, Pooja Agarwal, Bhoomika, published by: Research gate, "**Prediction of Employee Attrition Using Stacked Ensemble Method"** (2023)