IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

International Research Journal IJNRD Research Through Innovation

PREDICTION OF EMPLOYEE ATTRITION USING MACHINE LEARNING

Mrs. P. Swathi [1], A. Yasaswini [2], N. Lokesh Reddy [3], K. Tarun Sai [4], K.M.V.V. Prasad [5]

[1] Assistant Professor, Dept of Computer Science and Engineering, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India

[2,3,4,5] Students, Dept of Computer Science and Engineering, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India

ABSTRACT

Attrition is a term used to describe the process of reduction or decrease in the number of employees, customers, or participants over time. It can occur for a variety of reasons, including resignation, retirement, termination, or death. Employee attrition leads to a massive loss for the organization. This research study helps to predict employee attrition and helps HR Managers to understand the reason behind their employee's attrition using a machine learning model. This research study contains two datasets (IBM HR dataset and Healthcare dataset) to predict the attrition of employees working in two different organizations. Random Forest Classifier algorithm was used to predict employee attrition, this approach attained an accuracy score of 89% for the IBM HR dataset and 95% for the Healthcare dataset. To determine the factors that contributed to employee attrition, an Employee Exploratory Data Analysis (EEDA) was conducted. The key factors behind employee attrition, as per our study, were found to be Business Travel, Department, Education Field, Gender, Job Role, Marital status, and Over Time. To simplify the model complexity, we employed the data resampling technique called Synthetic Minority Oversampling Technique (SMOTE) on both datasets to balance them. The proposed approach aims to assist organizations in addressing employee attrition by identifying and enhancing the factors that contribute to it.

Keywords: Employee Attrition, Random Forest Classifier, Causes of Employee Attrition

1. INTRODUCTION

Employee attrition refers to the reduction of an organization's workforce resulting from various factors, including both voluntary and involuntary attrition. Employee attrition can cause significant losses for an organization. When an employee departs, the resulting vacancy may remain unfilled for an extended period, leading to delays in hiring a new person to fill the position. This can result in a loss of productivity and revenue for the organization. Additionally, the organization may incur significant costs in both time and money as newly hired employees typically require training for 5-6 months before they can fully contribute to the organization. Analyzing both the rate and causes of employee attrition can provide insight into an organization's level of progress, and can enable HR managers to take appropriate measures to address the issue before an employee departs. Identifying and managing the attrition rate and the factors that contribute to it can help to retain employees within an organization. The goal of using machine learning techniques is to achieve levels of accuracy that surpass those achievable by humans. The model was trained by using datasets (IBM HR and Healthcare) and the attrition of employees was predicted for new input data. To identify the factors responsible for employee attrition, the Employee Exploratory Data Analysis (EEDA) was employed to extract significant information from the dataset. To determine the attributes that significantly contribute to employee attrition, the feature engineering technique was employed, which involved examining feature correlation. The negatively correlated attributes are removed from the dataset through correlation analysis. The feature encoding was done to the refined dataset by using the one-hot encoding technique. After examining the dataset, it was discovered that it was imbalanced. To address this, the Synthetic Minority Oversampling Technique (SMOTE) data resampling technique was employed to balance the dataset. Following this, the dataset was split into a ratio of 85:15 for training and testing purposes. Balancing the datasets reduces the complexity of model prediction by ensuring an equal distribution of target variables within the dataset. The Random Forest Classifier model was employed to predict employee attrition.

2. LITERATURE SURVEY

A system was proposed to predict employee attrition and assist HR managers in decreasing the attrition rate of their organization. To train and test the model, the IBM HR employee dataset was employed. To reduce the dimensions of the dataset, a feature selection technique was employed. To predict employee attrition, the logistic regression technique was utilized, and the model attained an accuracy of 81%.

The IBM HR employee dataset was utilized to compare and apply different machine learning models for predicting employee attrition. The aim of this research study was to assist managers in identifying the challenges faced by their employees and updating their business strategies to reduce the organization's attrition rate. The study employed six machine learning models, and after comparing the results, the Random Forest algorithm was chosen as it provided the highest accuracy compared to the other algorithms. The model proposed in this study demonstrated an 85% accuracy in predicting employee attrition. The identified factors that lead to attrition can help HR managers to take proactive measures to reduce employee turnover in the organization.

3. DATASETS USED

Our research study involved the use of two distinct datasets to forecast employee attrition for two different organizations, one being a software organization and the other in the health sector.

IBM HR ANALYTICS EMPLOYEE ATTRITION & PERFORMANCE DATASET

The dataset used in our research study consists of 1470 rows and 35 columns. The variables include 'Age', 'Gender', 'Department', 'Distance from Home', and others. The 'Attrition' column contains two labels, 'Yes' or 'No', and the attrition rate of the organization is 16%.

EMPLOYEE ATTRITION FOR HEALTHCARE DATASET

The dataset used in this study contains 1676 rows and 35 columns. The columns include variables such as ‘Age’, ‘Business Travel’, ‘Daily rate’, ‘Department’, and more. The ‘Attrition’ column of the dataset consists of two class labels- ‘Yes’ or ‘No’. The attrition rate for this organization is 12%.

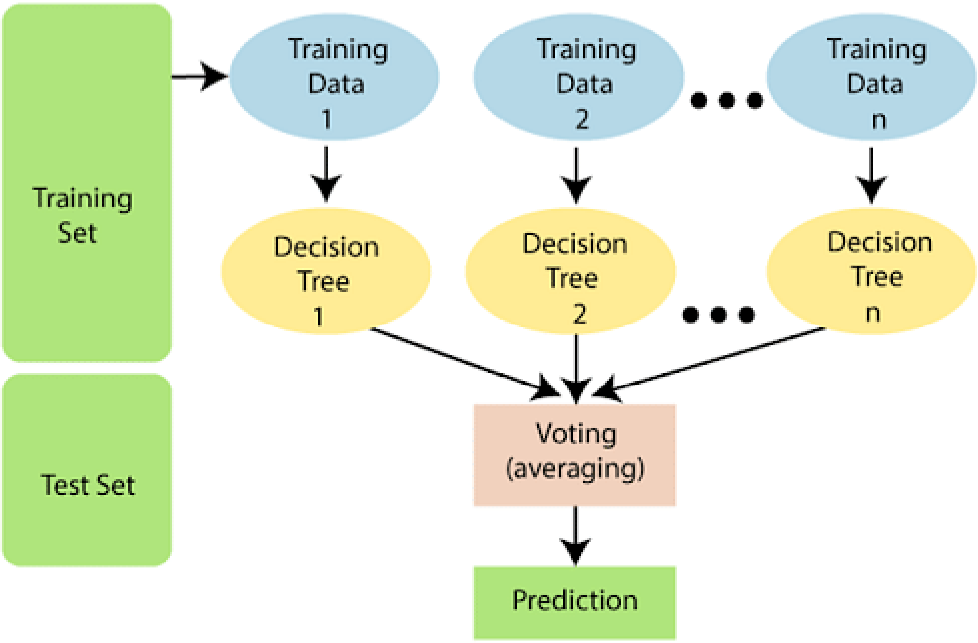
4. PROPOSED MODEL

Initially, two datasets are considered - the IBM HR dataset and the Healthcare dataset. The application of Employee Exploratory Data Analysis (EEDA) enabled the identification of significant attributes such as Business Travel, Department, Education Field, Gender, Job Role, Marital status, Over Time, etc., that contribute to employee attrition. The feature engineering technique was utilized to identify the attributes that have a high impact on employee attrition by analyzing their correlation with the target variable. Through correlation analysis, the dataset was analyzed to determine negatively correlated attributes, which were then removed from the dataset. Next, we balanced the dataset using SMOTE (Synthetic Minority Over Sampling Technique) to address the issue of class imbalance. To split the preprocessed dataset, an 85:15 ratio is utilized, where 85% of the data is allocated for training and 15% is allocated for testing. The Random Forest algorithm is utilized to predict employee attrition, enabling HR managers to concentrate on the factors that contribute to employee attrition.

4.1 RANDOM FOREST ALGORITHM

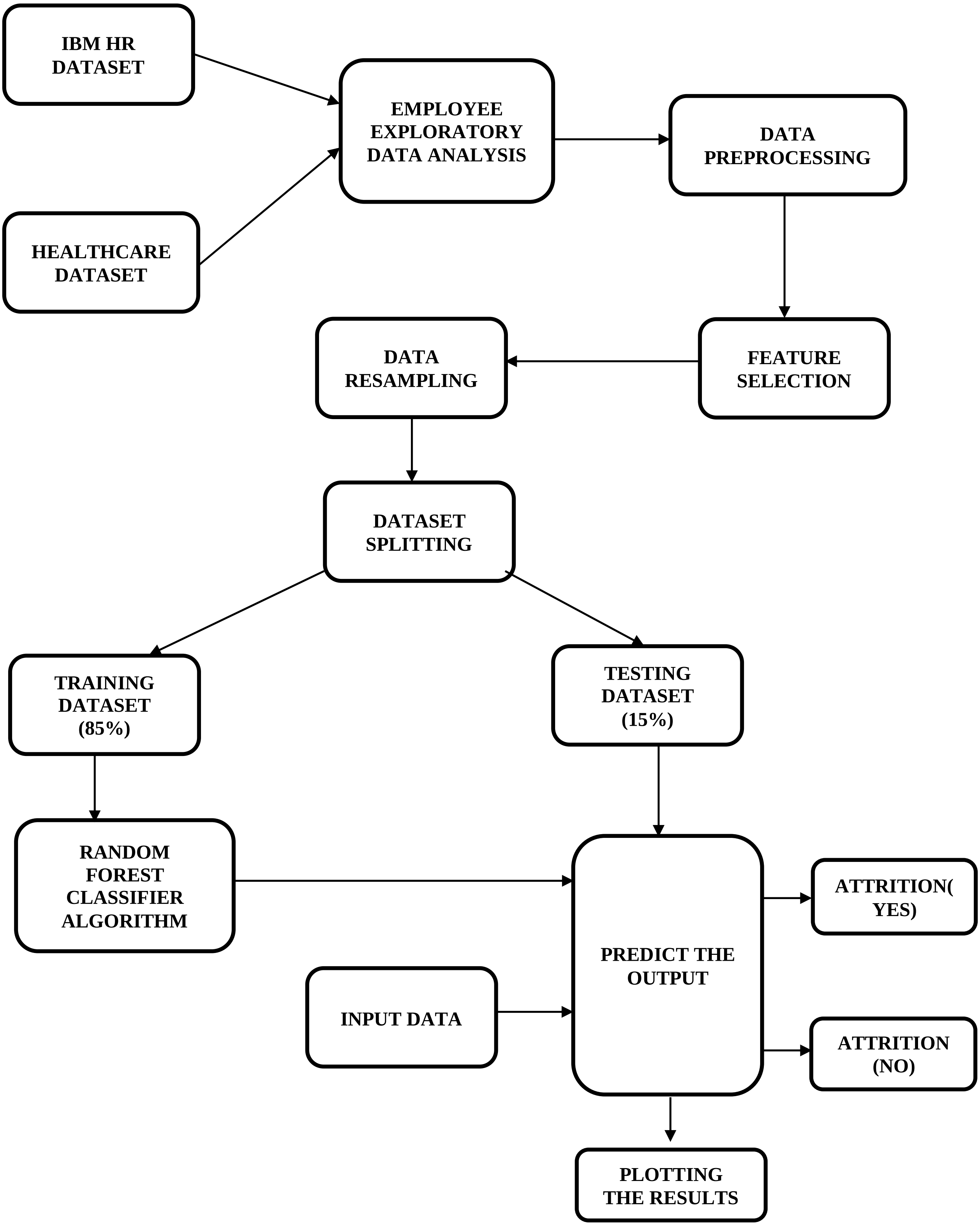
The Random Forest Algorithm is a commonly used supervised machine learning algorithm that is applied to solve problems related to classification and regression in the field of Machine Learning. Just like a forest consists of numerous trees and becomes more robust as the number of trees increases, the accuracy of a Random Forest Algorithm also increases with an increase in the number of trees. The Random Forest Classifier is comprised of multiple decision trees, each constructed based on different subsets of the given dataset, which results in increased predictive accuracy for a particular dataset. The Random Forest Classifier combines multiple decision trees and is known for its ability to improve the performance of the model and solve complex problems. Utilizing various subsets of the dataset, can increase the accuracy of the predictions and provide a more robust solution.

Random Forest Algorithm follows the steps given below:



Step 1: From the given training data, consider random instances of data.
Step 2: A decision tree is constructed for each of the training data.
Step 3: The decision trees' results will be aggregated through a voting process.
Step 4: The final prediction result is determined by considering the most voted prediction from the decision trees.

4.2 SYSTEM ARCHITECTURE



DATASETS:

- **IBM HR Analytics Employee Attrition & Performance Dataset**

This dataset comprises several attributes such as 'Work Life Balance', 'Gender', 'Job Role', 'Distance from Home', etc. The 'Attrition' column in the dataset has two labels, 'Yes' or 'No'. The organization's attrition rate is 16%.

- **Employee Attrition for Healthcare Dataset**

The dataset comprises attributes such as 'Years At Company', 'Gender', 'Department', 'Marital Status', etc. The 'Attrition' column contains two class labels, 'Yes' or 'No.' The organization's attrition rate is 12%.

EMPLOYEE EXPLORATORY DATA ANALYSIS (EEDA):

The Employee Exploratory Data Analysis (EEDA) is utilized to identify the key factors that contribute to employee attrition.

DATA PREPROCESSING:

Data preprocessing is a crucial step that needs to be performed before using the data for analysis. It involves converting a raw and unrefined dataset into a clean and structured dataset using various techniques. The information is analyzed in order to manage handle missing values, noisy data, and any other irregularities that may impede the efficacy of the algorithm. Preprocessing helps in improving the accuracy and efficiency of the model by making the data more suitable for analysis.

FEATURE SELECTION:

To improve the accuracy score, the feature selection technique is used to handle the features of the dataset and identify the significant attributes that lead to employee attrition.

DATA RESAMPLING:

Employing the method of resampling on the dataset is crucial to eliminate any partiality towards the majority category, thereby achieving a balance. This technique helps in reducing the model's complexity and improves the overall performance of the model.

DATASET SPLITTING:

To ensure that our model generalizes well, we use dataset splitting. The division of the dataset into two sections, specifically, a training set and a testing set, is implemented. The commonly used ratio for this splitting is 85:15, where 85% of the dataset is used for training the model, and 15% of the dataset is used for testing the model's performance. This approach helps in estimating the model's performance on new, unseen data.

RANDOM FOREST ALGORITHM:

Random Forest is a popular supervised learning algorithm used for both classification and regression tasks. Multiple decision trees are created by the algorithm, which are then merged to generate a prediction that is not only more precise but also reliable.

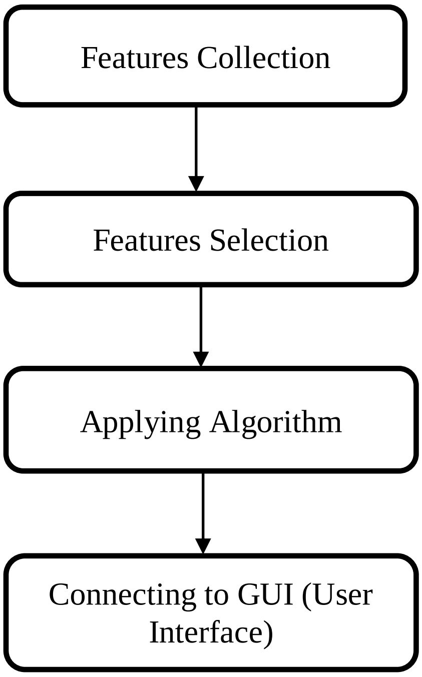
PREDICT THE OUTPUT:

The model generates an output in the form of a categorical value, i.e., either 'Yes' or 'No'.
Where,
Yes – Employee may leave the organization
No – Employee may not leave the organization

PLOTTING THE RESULTS:

Based on the selected attributes in the feature selection phase, the attrition value is predicted and a graph is plotted for the categorical values (Yes and No) for the employees in the organization. This helps the HRs of an organization to come up with a solution to reduce the attrition rate of their organization by understanding the reason behind a particular employee's attrition by observing the form filled by the employees.

4.3 MODULES DIVISION



FEATURES COLLECTION:

To gather and select relevant employee attributes that are significant for our analysis.

FEATURES SELECTION:

To develop an accurate employee attrition model, Feature Selection techniques are used to select the most relevant and impactful employee features. This process helps to filter out less important features and reduce the model's complexity.

APPLYING ALGORITHM:

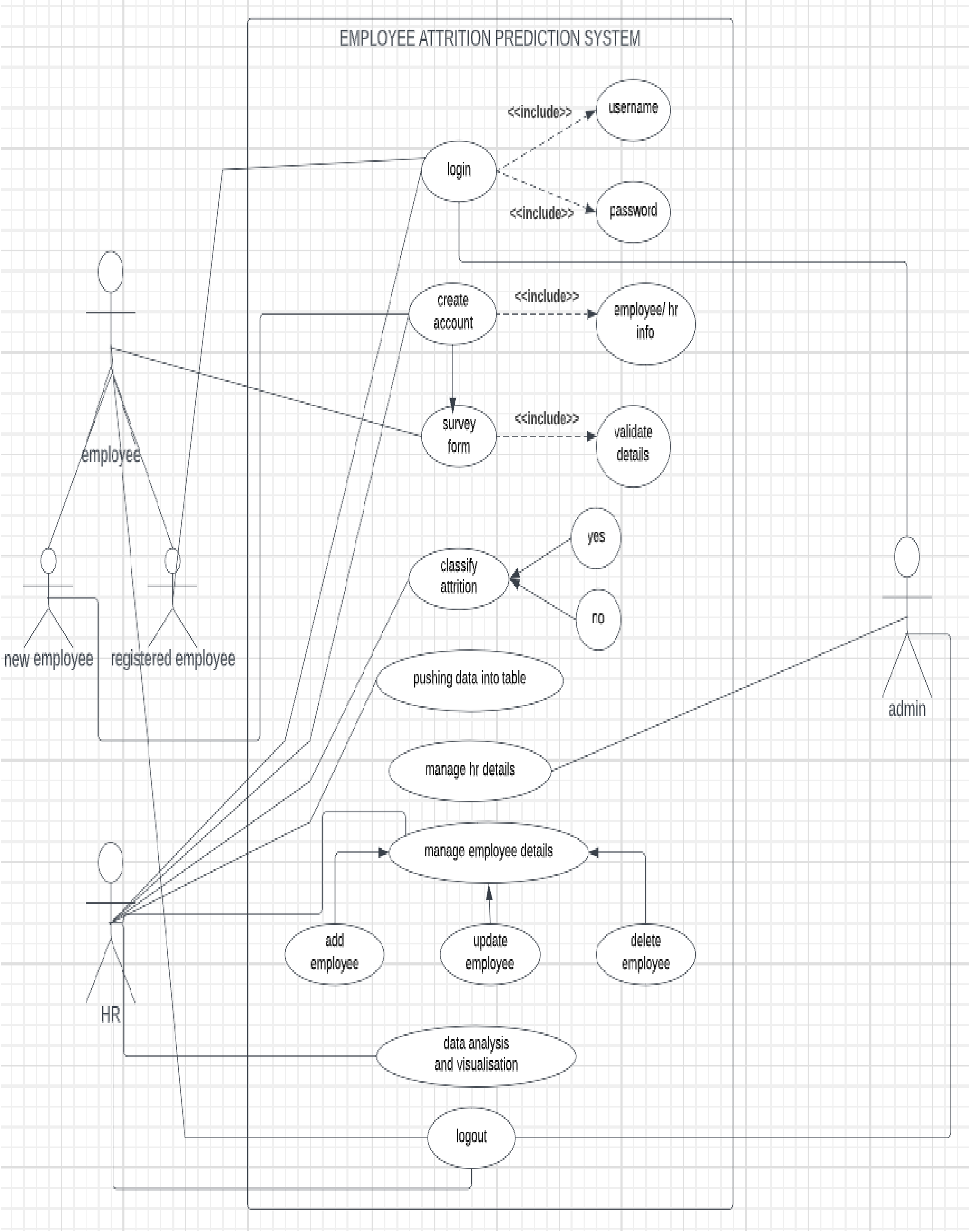
The next step involves training the model by utilizing the Random Forest Algorithm, which predicts the attrition value of each employee as either ‘Yes’ or ‘No’.

CONNECTING TO GUI (USER INTERFACE):

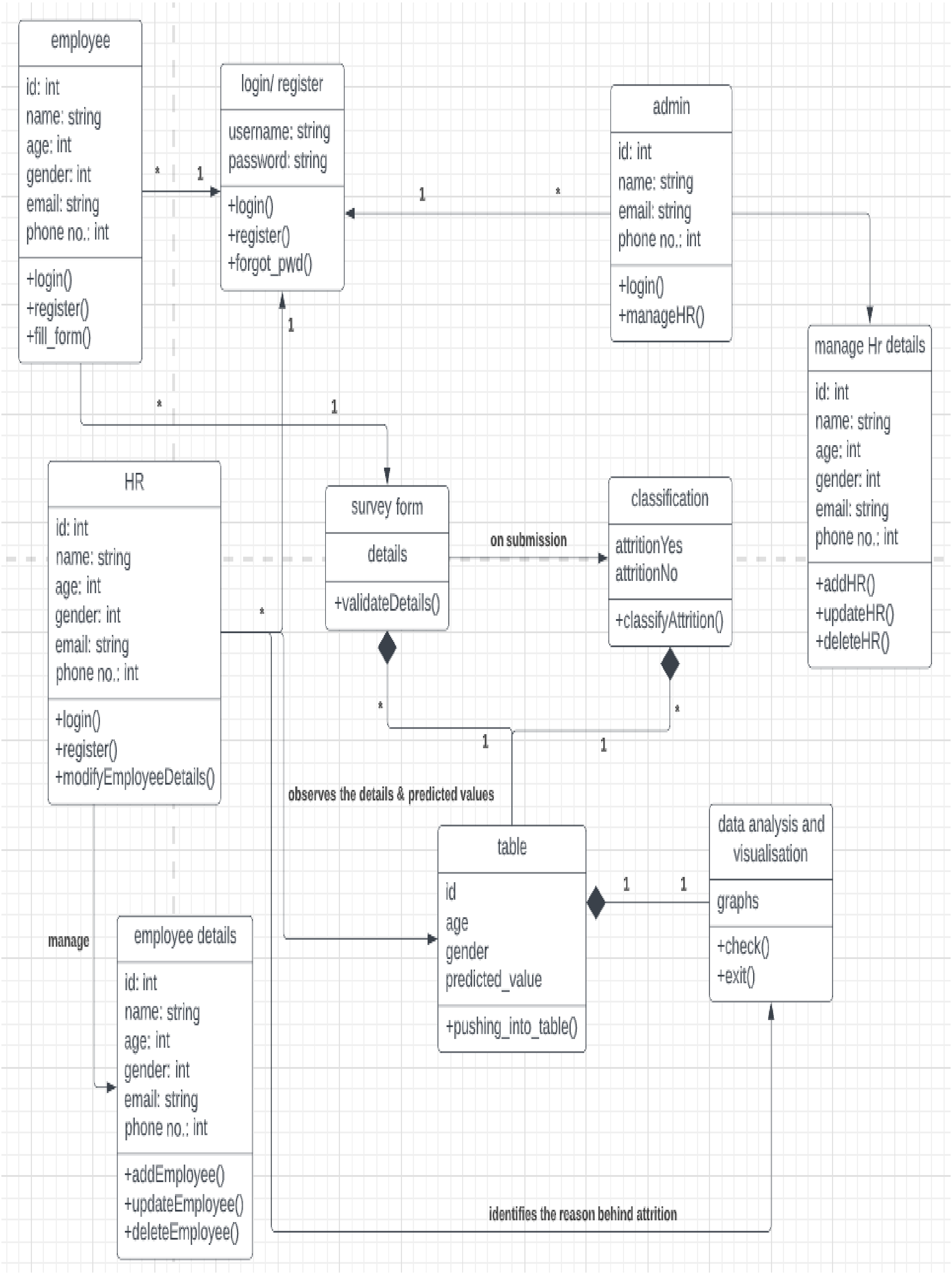
The employee will be allowed to submit a survey form that predicts the attrition value. So that HR can view the status of each employee and understand the reason behind their employee attrition by observing the plotted results.

4.4 UML DIAGRAMS

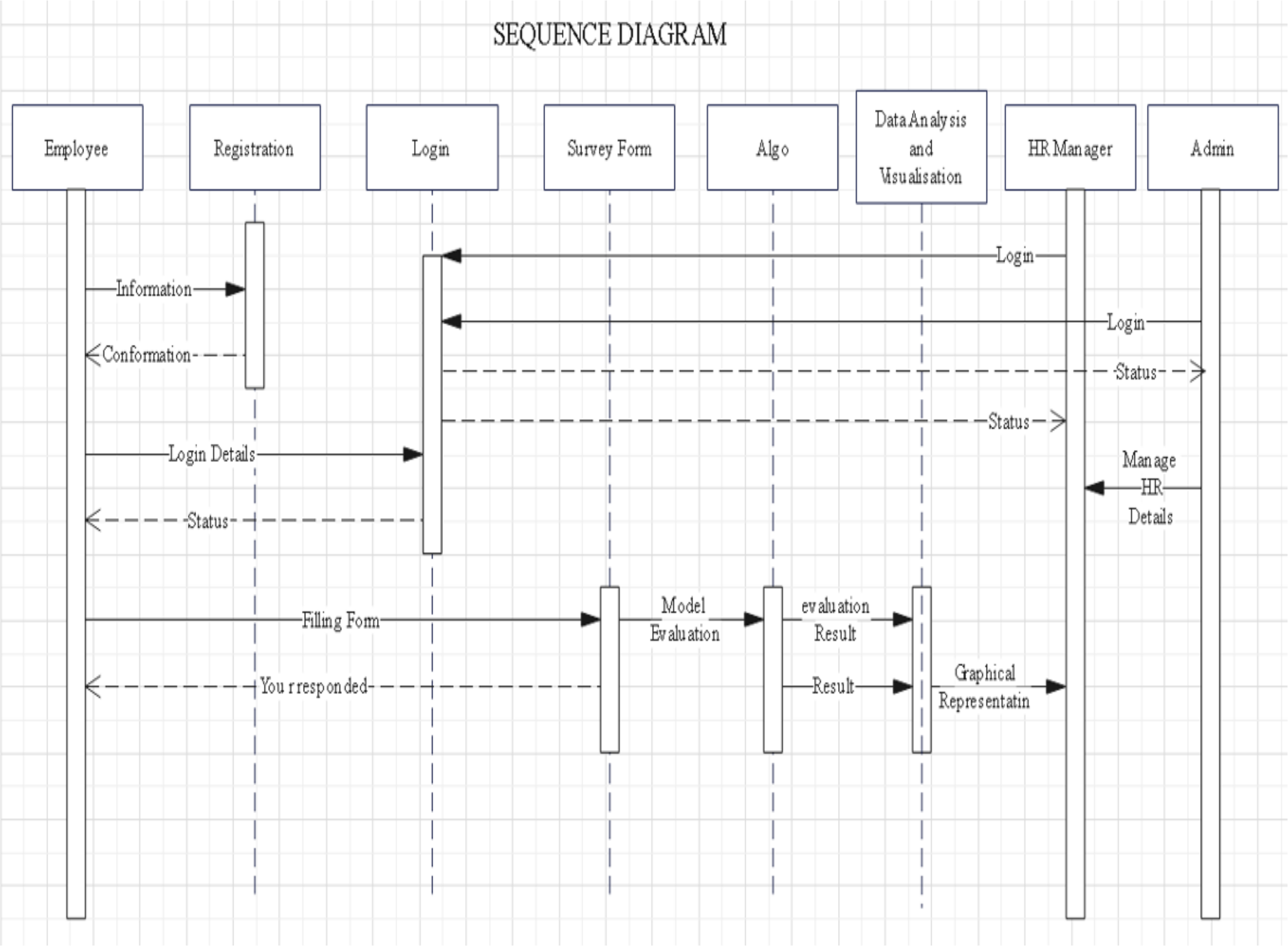
❖ USE CASE DIAGRAM:



❖ **CLASS DIAGRAM:**

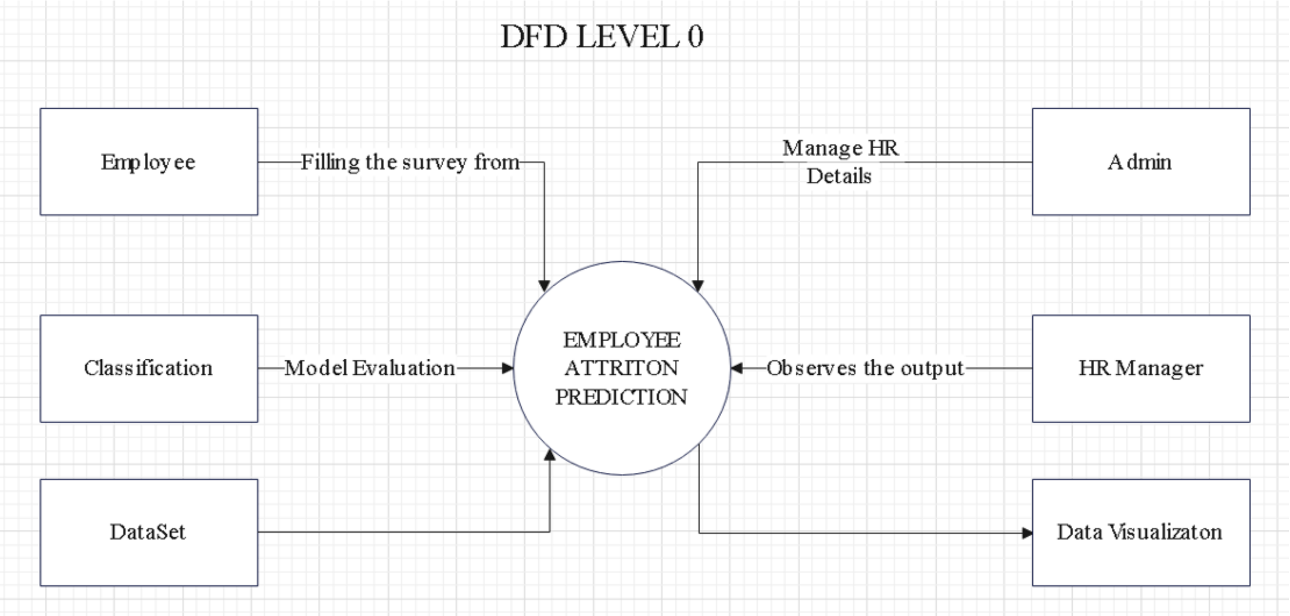


❖ SEQUENCE DIAGRAM:

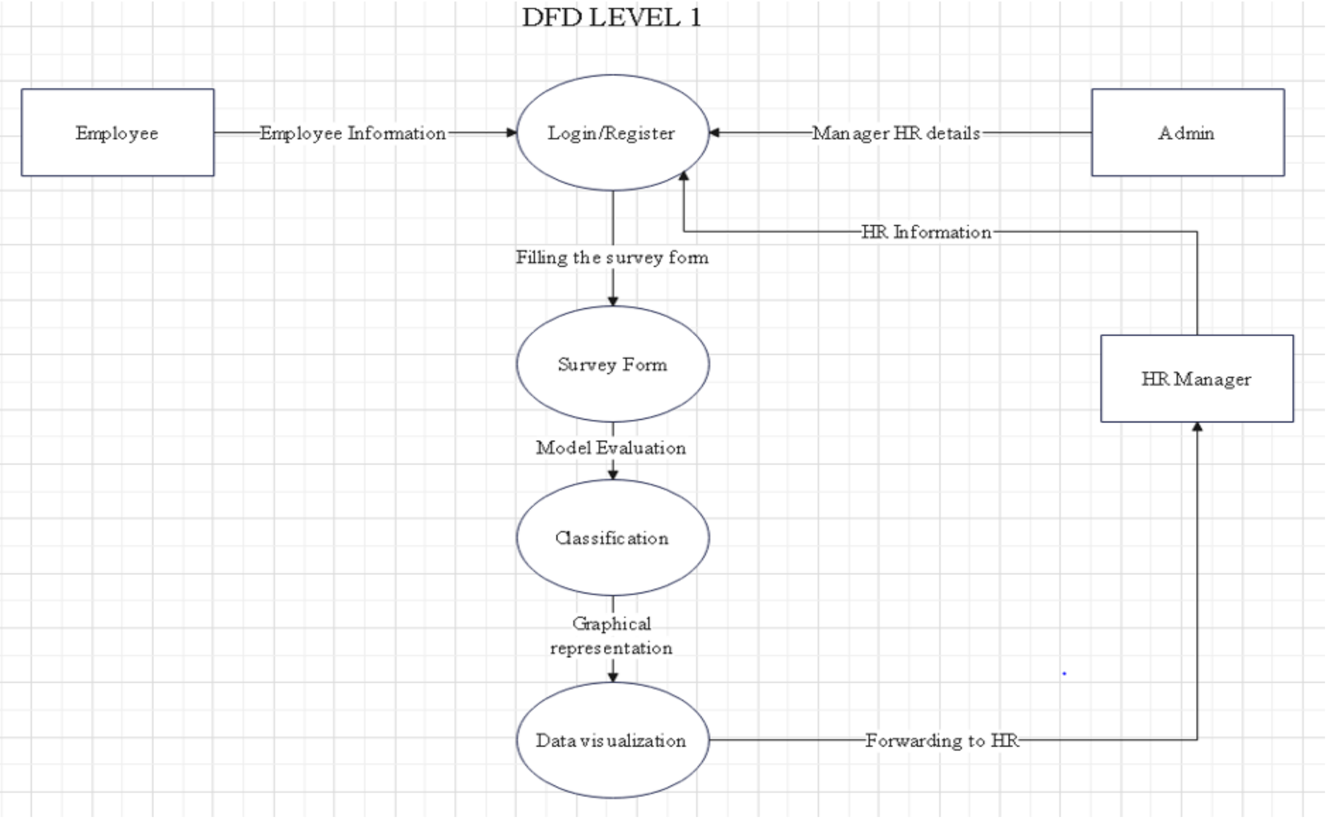


❖ DATA FLOW DIAGRAMS:

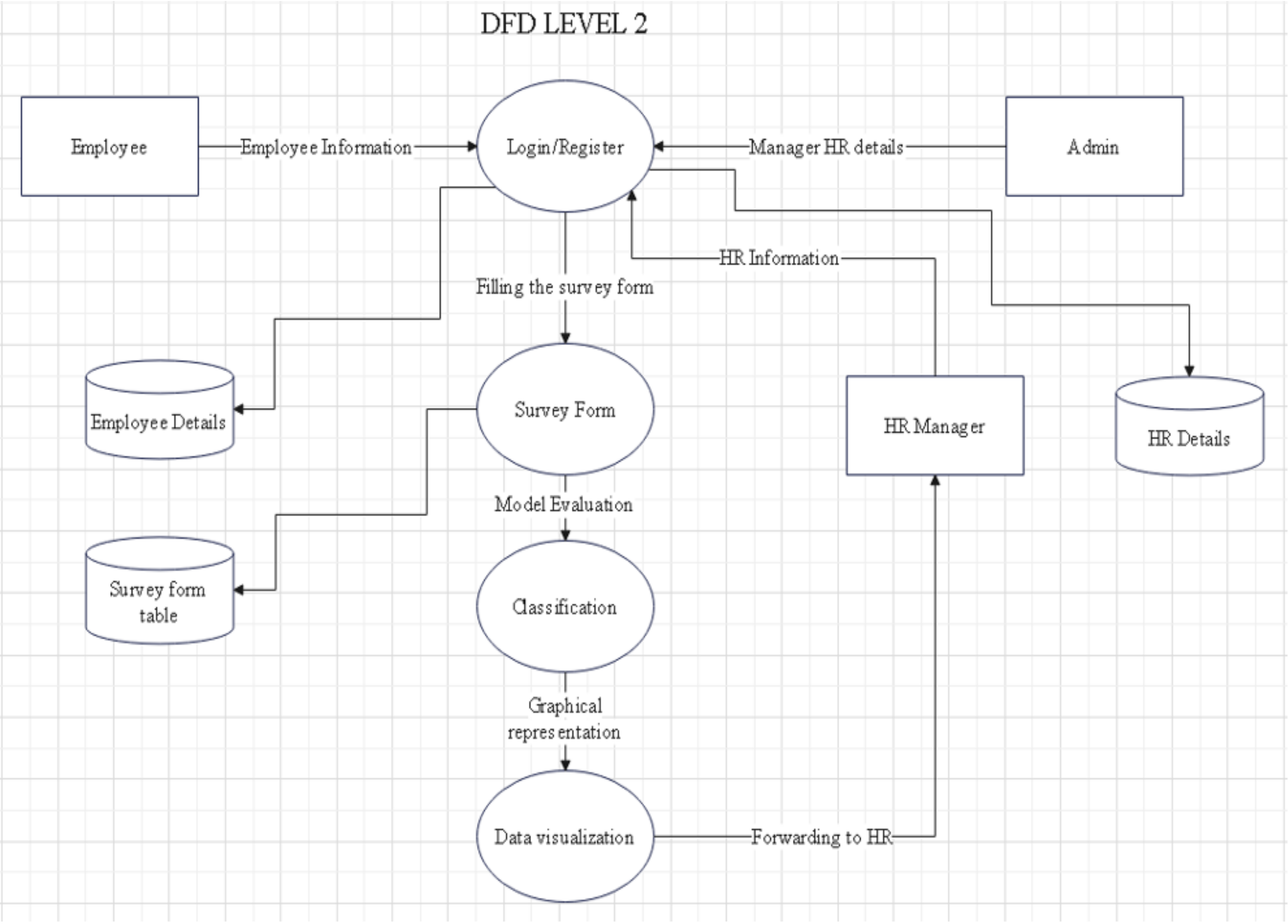
DFD Level-0:



DFD Level-1:



DFD Level-2:



4.5 INPUT

The employees will fill out a form that consists of details like Department, Education Field, Business Travel, Job Role, Gender, Marital Status, Years Since Last Promotion, etc.,

4.6 EXPECTED OUTPUT

a. The model will output a categorical value of either 'Yes' or 'No' based on the input details filled in by the employee where ‘Yes’ means the employee may leave the organization and ‘No’ means the employee may not leave the organization.

b. Our model helps HR to easily understand the reason behind their employee’s attrition through data visualization.

5. CONCLUSION

The use of the Random Forest Classifier model in predicting employee attrition has yielded favorable results, with an accuracy score of 89% and 95% for the IBM HR dataset and Healthcare dataset respectively. This approach is an effective way for organizations to identify and address the factors that contribute to employee attrition, ultimately helping to reduce it. Developed front-end web pages provide easy access to admin, HR managers, and employees and this helps the HR managers to understand the status of the organization. Our research findings help organizations (Software organizations and the Healthcare sector) overcome employee attrition.

6. FUTURE SCOPE

As the front-end pages are limited to only 2 organizations – IBM and Healthcare, we will enhance the project by including some other organizations like schools, colleges, etc., This helps the HR managers of organizations to overcome employee attrition and minimize the loss created by the employee attrition to the organization.

7. REFERENCES

1. The book titled "Data Science and Big Data Analytics - Discovering, Analyzing, Visualizing and Presenting Data" was authored by EMC Education Services and was published in July 2015.

2. In 2016, Pavan Subhash published a dataset on employee attrition and performance titled "IBM HR Analytics Employee Attrition & Performance" on the website Kaggle. The dataset can be accessed online at https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset.

3. In 2022, Shobhanam and Sumati published an article titled "HR Analytics: Employee Attrition Analysis using Random Forest" in the International Journal of Performance Engineering. The article can be found in Volume 18, starting from page 275.

4. The website apollotechnical.com published a list of 19 surprising employee retention statistics in 2022. This information can be accessed online at https://www.apollotechnical.com/employeeretention-statistics/ and was last retrieved on May 6th, 2022.