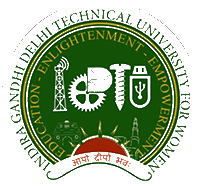
**INDIRA GANDHI DELHI TECHNICAL UNIVERSITYFOR WOMEN**

**(Established by Govt. of Delhi vide Act 09 of 2012)**

**KASHMERE GATE, DELHI-110006**

**ACADEMIC YEAR- 2024-26**

****

**PROJECT REPORT**

**IT WORKSHOP - II**

**MCA-110**

**IT DEPARTMENT**

**Submitted By:                                               Submitted To:**

Anam(07304092024) Ms. Manu Shree

Rashi Gupta(05304092024)

Semester - 2

REPORT

**INDEX**

|  |  |  |
| --- | --- | --- |
| **S no** | **Topics** | **Page No** |
|  | Title of the project | 3 |
|  | Problem Definition | 4 |
|  | Introduction | 5 |
|  | Aim & Objective | 6 |
|  | Tools & platform | 8-10 |
|  | System analysis | 11-13 |
|  | Software Model | 13-14 |
|  | Scheduling | 14-15 |
|  | DFD’S | 15 |
|  | DFD’S Design | 16-18 |
|  | ERD’S | 19 |
|  | ERD’S Design | 20 |
|  | Database Design | 20-24 |
|  | SQL Query |  |
|  | Input to the System | 24-25 |
|  | Output From The system | 25-26 |
|  | Module Description | 27 |
|  | User Interface |  |
|  | Coding |  |
|  | Testing/Security | 28 |
| 19. | Future Scope | 28-29 |
| 20. | Bibliography | 29 |

### **Title of the Project**

**Employee Attrition Prediction**

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### **Problem Definition**

**Problem Statement:** Employee attrition is a major concern for organizations as it impacts productivity, increases hiring costs, and affects team stability. Identifying the key factors contributing to employee attrition can help organizations take proactive measures to retain valuable employees.

This project aims to predict employee attrition using machine learning techniques by analyzing various factors such as salary, job satisfaction, overtime, and work-life balance. By leveraging predictive models, companies can develop better retention strategies and minimize turnover rates.

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### **Introduction**

Employee attrition, or workforce turnover, is a critical issue faced by organizations across various industries. Losing skilled employees not only disrupts workflow but also results in financial losses due to recruitment, training, and lost productivity. Understanding the patterns and factors that contribute to employee attrition is essential for businesses to develop effective retention strategies.

In recent years, **machine learning and data analytics** have played a crucial role in predicting employee attrition. By analyzing various employee-related factors such as **job satisfaction, salary, overtime, performance ratings, and years at the company**, organizations can identify employees who are likely to leave. This enables HR departments to take proactive steps, such as offering incentives or career development opportunities, to improve employee retention.

This project focuses on **predicting employee attrition using machine learning techniques**. By utilizing the IBM HR Analytics Employee Attrition dataset, we will explore the effectiveness of different models, including **K-Means Clustering, Naïve Bayes Classification, Decision Tree and Logistic Regression**, to determine the best approach for predicting attrition.

Through this study, organizations can **make data-driven decisions** to minimize attrition rates, enhance employee satisfaction, and optimize workforce management, ultimately leading to a more stable and productive work environment.

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### **Aim & Objective**

#### 

#### **Aim:**

The aim of this project is to develop a machine learning model that can accurately predict employee attrition, helping organizations identify employees at risk of leaving and take proactive retention measures.

#### **Objectives:**

* To analyze employee data and identify key factors influencing attrition, such as salary, job satisfaction, work-life balance, and career growth.
* To preprocess and clean the dataset to ensure high-quality input for machine learning models.
* To implement and compare different machine learning algorithms (**K-Means Clustering, Naïve Bayes Classification, Decision Tree and Logistic Regression**) for predicting attrition.
* To evaluate the performance of these models using accuracy, precision, recall, and F1-score.
* To provide actionable insights that HR professionals can use to reduce turnover and improve employee retention.

### 

### 

### **Tools & Platforms**

The implementation of this project requires various tools, programming languages, and software platforms to perform **data preprocessing, machine learning model training, and visualization of results**. The selection of these tools ensures efficient execution of predictive analytics and data science techniques.

### **Programming Language:**

* **Python** Python is widely used in the field of machine learning and data analytics due to its **simplicity, efficiency, and rich set of libraries** for data manipulation, visualization, and model training. It provides various built-in functions that make it easier to handle datasets and implement machine learning models effectively.

### **Development Environment:**

To develop and execute the machine learning models, the following development platforms are used:

* **Jupyter Notebook**
  + An open-source web-based application that allows interactive coding, visualization, and analysis.
  + It provides an efficient interface for executing Python code, debugging errors, and presenting findings in an organized manner.
* **Google Colab**
  + A cloud-based alternative to Jupyter Notebook, allowing execution of Python code without local installation.
  + Provides access to **GPUs and TPUs** for faster machine learning model training.

These development environments are essential for performing **data exploration, visualization, and model evaluation** efficiently.

**Libraries & Frameworks Used:**

To ensure smooth execution of data analytics and machine learning tasks, the following Python libraries are utilized:

* **NumPy** – Provides support for numerical computations and handling multi-dimensional arrays.
* **Pandas** – Used for data manipulation, handling missing values, and performing exploratory data analysis (EDA).
* **Matplotlib & Seaborn** – Visualization libraries for generating graphs, charts, and correlation heatmaps.
* **Scikit-learn (sklearn)** – Provides various machine learning algorithms for classification, clustering, and model evaluation.
* **HvPlot** – Used for interactive and high-quality visual analytics.

These libraries help in data preprocessing, model implementation, performance evaluation, and visualization of results.

### **Machine Learning Models Used:**

This project implements and compares multiple machine learning models to predict employee attrition. The following models are used:

* **K-Means Clustering (Unsupervised Learning):**
  + Groups employees into clusters based on similarities in attributes.
  + Helps in identifying patterns and trends in employee behavior.
* **Naïve Bayes Classification (Supervised Learning):**
  + A probabilistic classification technique based on Bayes’ theorem.
  + Efficient in handling categorical data and large datasets.
* **Decision Tree (Supervised Learning):**
  + A rule-based predictive model that determines attrition likelihood based on employee attributes.
  + Provides clear insights into which factors contribute most to employee turnover.
* **Logistic Regression (Supervised Learning):**

○ A statistical model that predicts the probability of employee attrition based on input features.  
 ○ Effective for binary classification tasks and interpretable in terms of feature importance.

These models help in identifying key patterns and predicting attrition rates effectively.

### **Dataset Used:**

The project is based on **IBM HR Analytics Employee Attrition Dataset**, which is obtained from **Kaggle**.

* **Source:** [IBM HR Analytics Attrition Dataset](https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/data)
* **Description:** The dataset consists of multiple employee attributes, including:
  + **Demographic Features** – Age, Gender, Marital Status.
  + **Job-Related Features** – Job Role, Department, Education, Monthly Income.
  + **Performance & Engagement** – Work-Life Balance, Overtime, Job Satisfaction.
  + **Historical Data** – Years at Company, Years Since Last Promotion, Number of Companies Worked.

This dataset helps in training machine learning models to identify employees likely to leave the company.

### **Platform Requirements:**

#### **Software Requirements:**

* **Operating System:** Windows 10 or later / Linux / macOS
* **Python Version:** 3.7 or later
* **Jupyter Notebook or Google Colab**

#### **Hardware Requirements:**

* **Processor:** Intel Core i5/i7 or equivalent AMD processor
* **RAM:** Minimum 8GB (Recommended: 16GB for better performance)
* **Storage:** At least 10GB free space for dataset storage and analysis
* **GPU (Optional):** Required for faster model training in deep learning applications

### 

### 

### 

### **System Analysis**

#### **Analysis of the Existing System**

Employee attrition has long been a challenge for organizations, with HR professionals relying on **traditional methods** such as employee surveys, exit interviews, and historical reports to understand why employees leave. While these methods provide useful insights, they suffer from several limitations:

* **Reactive Approach:** Most HR teams take action only after employees have resigned rather than predicting and preventing attrition.
* **Lack of Data-Driven Decision Making:** Traditional HR strategies rely heavily on intuition rather than objective, data-backed insights.
* **Inability to Handle Large Data:** With increasing workforce sizes, manually analyzing attrition trends is inefficient.
* **Imbalanced Decision Making:** Companies often fail to focus on critical factors like **overtime, salary hikes, or work-life balance**, leading to ineffective retention strategies.

### **Existing Employee Attrition Prediction Models**

Several machine learning models have been applied in HR analytics to predict attrition:

1. **Logistic Regression:** A probability-based model that predicts attrition but struggles with complex patterns.
2. **Random Forest:** An ensemble learning model that improves prediction accuracy but is computationally expensive.
3. **Support Vector Machine (SVM):** Effective for classification but slow on large datasets.
4. **Artificial Neural Networks (ANN):** Captures deep patterns but requires extensive training data.
5. **K-Means Clustering:** Segments employees into risk groups but lacks direct prediction capabilities.
6. **Naïve Bayes Classification:** Works well with categorical data but assumes feature independence.
7. **Decision Tree:** Provides interpretable insights but may overfit without proper tuning.

### **Limitations of Existing Models**

Despite the advancements in machine learning, most models have **narrow applications** in HR analytics. They either:

* Focus on **prediction accuracy** but do not provide insights into **why** employees leave.
* Work well with **balanced datasets** but fail in real-world HR data, where attrition cases are rare.
* Do not consider **feature selection techniques** to identify the most critical attrition factors.

### **Proposed System:**

To address these challenges, this project presents a **comprehensive and unique approach** to predicting employee attrition. The system is designed to **not only predict attrition but also identify key reasons behind employee turnover, enabling HR teams to take proactive actions.**

#### **Key Features of the Proposed System**

✔ **Comparative Machine Learning Models:**

* Unlike traditional research that focuses on a single model, this project **compares three models**—**K-Means Clustering, Naïve Bayes, and Decision Tree**—to determine the most effective approach for HR analytics.

✔ **Feature Importance Analysis for HR Decision-Making:**

* Instead of just predicting attrition, the system highlights **which factors contribute most to employee turnover** (e.g., **low salary, high overtime, poor work-life balance**).
* This allows HR managers to implement **data-driven retention strategies**.

✔ **Exploratory Data Analysis (EDA) with Visual Insights:**

* Uses **heatmaps, correlation matrices, and distribution plots** to explore key trends in attrition.
* HR professionals can visually understand which employee attributes contribute most to attrition.

✔ **Scalable and Practical Implementation:**

* The system can process **large employee datasets** and integrate into **HR management platforms** for real-time monitoring.

### **System Workflow**

1. **Data Collection:**
   * The project uses the **IBM HR Analytics Employee Attrition dataset** sourced from **Kaggle**.
2. **Preprocessing & Feature Engineering:**
   * Handles missing values, encodes categorical variables, and applies feature scaling.
3. **Exploratory Data Analysis (EDA):**
   * Identifies key attrition factors through visual analytics.
4. **Model Training & Evaluation:**
   * Applies and compares **K-Means Clustering, Naïve Bayes, and Decision Tree** models.
5. **Performance Metrics:**
   * Uses **accuracy, precision, recall, and F1-score** to evaluate model effectiveness.
6. **Predictions & HR Decision-Making:**
   * Generates actionable insights to help organizations retain employees.

### **SRS & Software Modules**

A **Software Requirements Specification (SRS)** document provides a complete description of the behavior and functionality of a system to be developed. It defines the scope, functional and non-functional requirements, constraints, and expected outcomes of the system. The SRS also includes **use cases** describing how users will interact with the software.

In addition to defining the system’s capabilities, the SRS outlines **non-functional requirements**, such as performance, quality standards, and design constraints, ensuring the software meets all necessary development criteria.

To develop an effective **Employee Attrition Prediction System**, it is essential to have a thorough understanding of the product requirements. This document is created based on detailed discussions with the development team and stakeholders. The SRS may be used as a **contract deliverable** or an **internal project guideline** depending on organizational needs.

The SRS should include:

**1. Functional Requirements**

* **Employee Data Management:**  
  The system should allow uploading, storing, and managing employee data including demographics, job roles, performance, and satisfaction levels.
* **Attrition Prediction using Machine Learning:**  
  The system should integrate multiple machine learning models (Logistic Regression, Decision Tree, Naïve Bayes, K-Means) to predict the likelihood of employee attrition.
* **Real-Time Prediction and Result Display:**  
  The system should provide real-time predictions and display results immediately after input submission.
* **Model Comparison and Analysis:**  
  Users should be able to view and compare performance metrics of the different ML models used.

**2. Non-Functional Requirements**

**System Features:**

* **Accurate Prediction and Model Integration:**  
  The system must ensure accurate predictions using well-trained ML models with proper evaluation metrics.
* **Real-Time Data Processing and Output:**  
  Inputs should be processed instantly, with predictions and insights returned without delay.
* **User-Friendly Interface:**  
  A clean and intuitive UI (HTML/CSS or optionally Jetpack Compose for Android version) should enable users to interact with the system effortlessly.
* **Secure Data Handling:**  
  Employee data should be securely handled with proper privacy measures and no unauthorized access.

**3. Use Cases**

* **HR Manager:** Upload employee data, view attrition predictions, analyze factors contributing to high risk.
* **Data Analyst:** Compare model performance, extract insights, and fine-tune models.
* **Management:** View dashboards for organizational attrition trends and make data-driven decisions.

### **Software Approach**

The **Spiral Model** is used for the development of this system, as it combines **iterative development** with structured **risk analysis** at each phase.

The **Spiral Model** is a combination of the **iterative development model** and the **sequential Waterfall model**, with a strong emphasis on **risk management and continuous improvement**. This approach allows for incremental refinements at each stage of the project, ensuring a well-structured and adaptable system.

Key characteristics of the **Spiral Model**:

* It follows a **phase-wise linear approach** while incorporating feedback loops.
* After each iteration, the system is evaluated, and improvements are made based on feedback.
* This process continues throughout the development lifecycle, ensuring **continuous enhancements** in accuracy and performance.

Each iteration of the **Spiral Model** includes:

1. **Planning:** Defining objectives, requirements, and constraints.
2. **Risk Analysis:** Identifying and addressing key challenges such as **data imbalance, feature selection, and model performance**.
3. **Development & Testing:** Implementing machine learning models (**K-Means, Naïve Bayes, Decision Tree, Logistic Regression**) and evaluating their effectiveness.
4. **Customer Evaluation & Feedback:** Analyzing results and refining the system based on performance metrics.

### **Scheduling**

Scheduling is an essential aspect of project development, ensuring that tasks are completed within the allocated time frame while maintaining efficiency and accuracy. The development of the **Employee Attrition Prediction System** follows a structured schedule, dividing the project into distinct phases.

The project timeline is planned using the **PERT (Program Evaluation and Review Technique) Chart**, which helps in tracking the progress of different activities.

### **PERT Chart Representation**

A **PERT chart** is used to visualize the project timeline and dependencies between tasks. It ensures that each phase is completed in a **logical sequence**, preventing delays and improving efficiency,

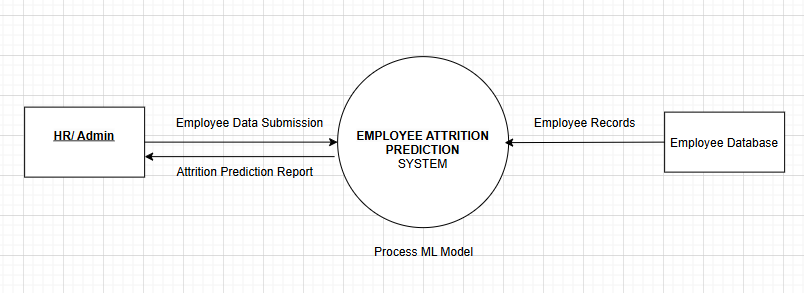
### 

### **Data Flow Diagrams (DFD)**

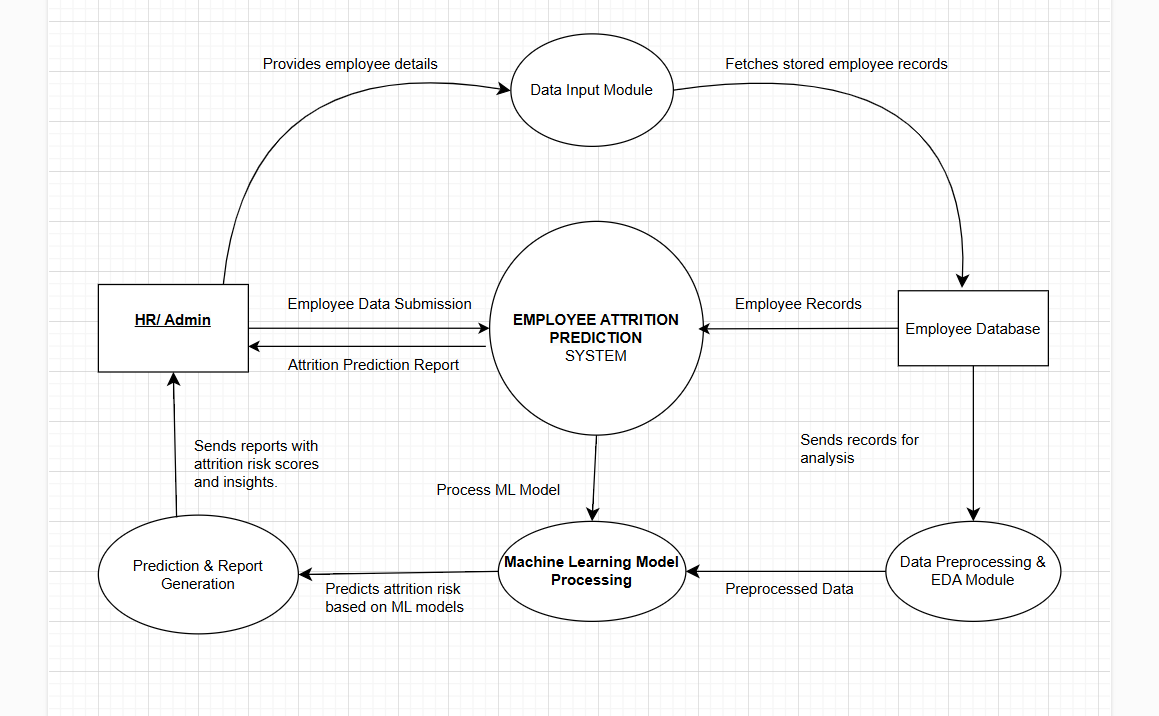
Data Flow Diagrams (DFDs) provide a **graphical representation** of how data moves through the **Employee Attrition Prediction System**. It helps in understanding the system’s components, processes, data storage, and interactions between users and the system. DFDs are designed at **different levels** to break down the system into simpler components.

### **DFD Designs**

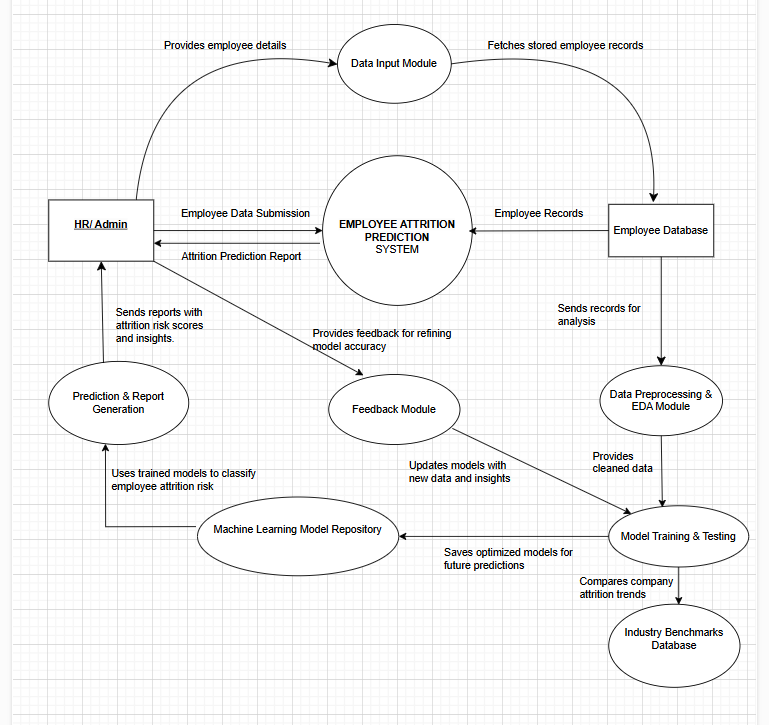
### **Level 0 DFD :**



### **Level 1 DFD :**



### **Level - 2 DFD :**



**Entity-Relationship Diagram (ERD)**

An **Entity-Relationship Diagram (ERD)** is a **data modeling technique** that provides a **graphical representation** of the relationships between different entities in a system. It helps in designing the database structure efficiently, ensuring data integrity and ease of access.

The ERD for the **Employee Attrition Prediction System** consists of multiple entities, their attributes, and the relationships among them. It visually represents how employee-related data is structured and how it interacts with other components of the system.

### **Main Components of ERD :**

* **Entities:** Represent objects or concepts in the system that store data.
* **Attributes:** Define the properties of each entity.
* **Relationships:** Show the associations between different entities.
* **Primary & Foreign Keys:** Ensure data uniqueness and maintain referential integrity.

### 

### 

### 

### 

### 

### 

### 

### **ERD Designs**

### 

### **Database Design**

Database design is a crucial step in structuring the **Employee Attrition Prediction System**, ensuring efficient data storage, retrieval, and management. It involves defining tables, attributes, primary keys, and relationships to support the machine learning model and HR decision-making.

### **Tables & Their Structure**

### **Employee Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Data Type** | **Constraints** | **Description** |
| EmployeeNumber | INT | PRIMARY KEY | Unique employee ID |
| Age | INT | NOT NULL | Employee's age |
| Gender | VARCHAR(10) | NOT NULL | Gender (Male/Female) |
| MaritalStatus | VARCHAR(20) | NOT NULL | Marital status (Single/Married) |
| Over18 | CHAR(3) | NOT NULL | Whether the employee is over 18 |
| Department | VARCHAR(50) | NOT NULL | Department name |
| JobRole | VARCHAR(50) | NOT NULL | Employee job role |
| JobLevel | INT | CHECK(JobLevel BETWEEN 1 AND 5) | Job seniority level |
| Education | INT | NOT NULL | Education level (1-5) |
| EducationField | VARCHAR(50) | NOT NULL | Employee's field of study |
| BusinessTravel | VARCHAR(30) | NOT NULL | Frequency of travel (Rarely/Frequently) |
| DistanceFromHome | INT | NOT NULL | Distance from home to office |
| StandardHours | INT | DEFAULT 40 | Standard working hours |
| OverTime | VARCHAR(5) | NOT NULL | Whether employee works overtime |
| NumCompaniesWorked | INT | NOT NULL | Number of companies worked for |

### 

### **Salary Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Data Type** | **Constraints** | **Description** |
| EmployeeNumber | INT | PRIMARY KEY, FOREIGN KEY REFERENCES Employee(EmployeeNumber) | Employee reference |
| MonthlyIncome | INT | NOT NULL | Monthly salary |
| MonthlyRate | INT | NOT NULL | Salary rate per month |
| HourlyRate | INT | NOT NULL | Hourly salary rate |
| DailyRate | INT | NOT NULL | Daily salary rate |
| PercentSalaryHike | INT | NOT NULL | Percentage salary hike |
| StockOptionLevel | INT | NOT NULL | Stock options granted |

### **Performance & Job Satisfaction Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Data Type** | **Constraints** | **Description** |
| EmployeeNumber | INT | PRIMARY KEY, FOREIGN KEY REFERENCES Employee(EmployeeNumber) | Employee reference |
| PerformanceRating | INT | CHECK(PerformanceRating BETWEEN 1 AND 5) | Performance rating (1-5) |
| JobSatisfaction | INT | CHECK(JobSatisfaction BETWEEN 1 AND 5) | Job satisfaction level |
| RelationshipSatisfaction | INT | CHECK(RelationshipSatisfaction BETWEEN 1 AND 5) | Relationship satisfaction |
| WorkLifeBalance | INT | CHECK(WorkLifeBalance BETWEEN 1 AND 5) | Work-life balance rating |

### **Work Experience Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Data Type** | **Constraints** | **Description** |
| EmployeeNumber | INT | PRIMARY KEY, FOREIGN KEY REFERENCES Employee(EmployeeNumber) | Employee reference |
| YearsInCurrentRole | INT | NOT NULL | Years in the current role |
| YearsAtCompany | INT | NOT NULL | Number of years at the company |
| YearsSinceLastPromotion | INT | NOT NULL | Years since last promotion |
| YearsWithCurrManager | INT | NOT NULL | Years with the current manager |
| TotalWorkingYears | INT | NOT NULL | Total years of work experience |
| TrainingTimesLastYear | INT | NOT NULL | Training sessions attended last year |

### **Attrition Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Data Type** | **Constraints** | **Description** |
| EmployeeNumber | INT | PRIMARY KEY, FOREIGN KEY REFERENCES Employee(EmployeeNumber) | Employee reference |
| Attrition | VARCHAR(10) | NOT NULL | Whether the employee left the company (Yes/No) |

### **SQL Query**

//DATABASE CODE FOR DATASET

-- Step 1: Create Database

CREATE DATABASE EmployeeDB;

USE EmployeeDB;

-- Step 2: Create Tables

-- Employee Table

CREATE TABLE Employee (

EmployeeNumber INT PRIMARY KEY,

Age INT NOT NULL,

Gender VARCHAR(10) NOT NULL,

MaritalStatus VARCHAR(20) NOT NULL,

Over18 CHAR(3) NOT NULL,

Department VARCHAR(50) NOT NULL,

JobRole VARCHAR(50) NOT NULL,

JobLevel INT CHECK(JobLevel BETWEEN 1 AND 5),

Education INT NOT NULL,

EducationField VARCHAR(50) NOT NULL,

BusinessTravel VARCHAR(30) NOT NULL,

DistanceFromHome INT NOT NULL,

StandardHours INT DEFAULT 40,

OverTime VARCHAR(5) NOT NULL,

NumCompaniesWorked INT NOT NULL

);

-- Salary Table

CREATE TABLE Salary (

EmployeeNumber INT PRIMARY KEY,

MonthlyIncome INT NOT NULL,

MonthlyRate INT NOT NULL,

HourlyRate INT NOT NULL,

DailyRate INT NOT NULL,

PercentSalaryHike INT NOT NULL,

StockOptionLevel INT NOT NULL,

FOREIGN KEY (EmployeeNumber) REFERENCES Employee(EmployeeNumber)

);

-- Performance & Job Satisfaction Table

CREATE TABLE PerformanceSatisfaction (

EmployeeNumber INT PRIMARY KEY,

PerformanceRating INT CHECK(PerformanceRating BETWEEN 1 AND 5),

JobSatisfaction INT CHECK(JobSatisfaction BETWEEN 1 AND 5),

RelationshipSatisfaction INT CHECK(RelationshipSatisfaction BETWEEN 1 AND 5),

WorkLifeBalance INT CHECK(WorkLifeBalance BETWEEN 1 AND 5),

FOREIGN KEY (EmployeeNumber) REFERENCES Employee(EmployeeNumber)

);

-- Work Experience Table

CREATE TABLE WorkExperience (

EmployeeNumber INT PRIMARY KEY,

YearsAtCompany INT NOT NULL,

YearsInCurrentRole INT NOT NULL,

YearsSinceLastPromotion INT NOT NULL,

YearsWithCurrManager INT NOT NULL,

TotalWorkingYears INT NOT NULL,

TrainingTimesLastYear INT NOT NULL,

FOREIGN KEY (EmployeeNumber) REFERENCES Employee(EmployeeNumber)

);

-- Attrition Table

CREATE TABLE Attrition (

EmployeeNumber INT PRIMARY KEY,

Attrition VARCHAR(10) NOT NULL,

FOREIGN KEY (EmployeeNumber) REFERENCES Employee(EmployeeNumber)

);

-- Step 3: Index Optimization

CREATE INDEX idx\_employee\_department ON Employee(Department);

CREATE INDEX idx\_salary\_employee ON Salary(EmployeeNumber);

CREATE INDEX idx\_performance\_employee ON PerformanceSatisfaction(EmployeeNumber);

CREATE INDEX idx\_work\_experience\_employee ON WorkExperience(EmployeeNumber);

-- Step 4: Backup & Recovery

BACKUP DATABASE EmployeeDB TO DISK = 'C:/Backup/EmployeeDB\_Backup.bak';

### **Input to the System**

The **Employee Attrition Prediction System** requires multiple inputs to process and analyze data effectively. These inputs come from **HR professionals, historical employee data, and machine learning models**, ensuring accurate attrition predictions.

### **Main Inputs to the System**

#### **1. Employee Data (HR Input)**

**Employee ID** – Unique identifier for each employee.  
**Name** – Employee’s full name.  
**Age** – Age of the employee.  
**Gender** – Male/Female/Other.  
**Department** – The department the employee works in.  
**Job Role** – Specific position held by the employee.  
**Monthly Income** – Salary details.  
**Overtime Status** – Whether the employee works overtime (Yes/No).  
**Years at Company** – Total number of years the employee has worked.  
**Job Satisfaction** – Employee’s job satisfaction level (scale of 1-5).  
**Attrition Status** – Whether the employee has left the company (Yes/No).

**Marital Status** – Single/Married/Divorced.

**Education Level** – Education level (1 = Below College, 2 = College, 3 = Bachelor, etc.).

**Business Travel** – Frequency of business travel (Rarely, Frequently, Non-Travel).

**Job Level** – The level of the job within the hierarchy.

**Total Working Years** – Total years of experience the employee has.

**Years in Current Role** – Time spent in the current position.

**Years Since Last Promotion** – Time since the employee was last promoted.

**Years with Current Manager** – Time spent with the current manager.

**Environment Satisfaction** – Satisfaction with work environment (scale of 1-4).

**Work-Life Balance** – Work-life balance score (scale of 1-4).

**Training Times Last Year** – Number of training sessions attended last year.

**Performance Rating** – Annual performance evaluation (scale of 1-4).

**Distance From Home** – Distance between employee’s home and office.

**Num Companies Worked** – Number of companies the employee has worked at before.

**Stock Option Level** – Whether stock options are provided (scale of 0-3).

**Hourly Rate / Daily Rate / Monthly Rate** – Payment structures (depending on dataset).

#### **2. Machine Learning Model Inputs**

✔ **Preprocessed Employee Data** – Cleaned and structured employee data.  
 ✔ **Encoded Categorical Variables** – Converting text-based attributes (e.g., job role, department) into numerical form for ML models.

#### **3. HR Queries & System Commands**

✔ **HR User Login** – Admin must enter credentials to access the system.  
 ✔ **Data Upload** – HR can upload new employee datasets for analysis.  
 ✔ **Attrition Prediction Request** – HR can request predictions for individual employees or bulk employee records.  
 ✔ **Report Generation Command** – The system generates a report based on employee attrition predictions.

### 

### **Output from the System**

The **Employee Attrition Prediction System** generates various outputs to help HR professionals analyze and manage employee turnover effectively. The outputs include **attrition predictions, key influencing factors, and detailed reports**, aiding in proactive decision-making.

### **1. Attrition Prediction Results**

✔ **Prediction Label:** Classifies employees as **"Likely to Stay"** or **"Likely to Leave"**.  
 ✔ **Probability Score:** Displays the likelihood of attrition.

### **2. Key Factors Influencing Attrition**

✔ Identifies **top reasons** for employee turnover, such as:

* Low salary
* High overtime
* Low job satisfaction

### **3. HR Report Generation**

✔ **Summary Report:** Shows attrition trends and risk levels.  
 ✔ **Department-Wise Analysis:** Highlights high-turnover areas.  
 ✔ **Charts & Graphs:** Visualize salary vs. attrition risk, job satisfaction impact, etc.

### **4. HR Decision-Making Support**

✔ **Retention Strategies:** Salary adjustments, work-life balance improvements, career growth plans.  
 ✔ **Downloadable Reports:** Export insights as **PDF/CSV files**.

### **Module Description**

The **Employee Attrition Prediction System** is divided into multiple modules, each handling a specific functionality to ensure efficient processing, prediction, and analysis of employee data. These modules work together to provide **accurate attrition predictions and HR insights**.

### **1. Data Preprocessing Module**

✔ Cleans raw employee data by handling **missing values, duplicates, and inconsistencies**.  
 ✔ Encodes categorical variables (e.g., Job Role, Department, Overtime) into numerical form.  
 ✔ Scales numerical features (Salary, Years at Company) for better model performance.

### **2. Exploratory Data Analysis (EDA) Module**

✔ Generates **heatmaps, correlation graphs, and distribution plots** to identify patterns in attrition.  
 ✔ Identifies key attrition factors such as **salary, job satisfaction, and overtime**.  
 ✔ Helps HR professionals understand **why employees leave** through visual analytics.

### **3. Machine Learning Module**

✔ Implements **four predictive models**:

* **K-Means Clustering** – Groups employees based on attrition risk.
* **Naïve Bayes Classification** – Uses probability-based classification for predictions.
* **Decision Tree** – Provides **rule-based attrition insights** and feature importance analysis.
* **Logistic Regression** – A statistical model used for binary classification that estimates the probability of attrition based on employee features.  
   ✔ Splits the dataset into **training and testing sets** to evaluate model accuracy.  
   ✔ Compares model performance using **precision, recall, F1-score, and accuracy**.

### **4. Prediction & Report Generation Module**

✔ **Predicts attrition risk** for each employee based on historical data.  
 ✔ Classifies employees as **“Likely to Stay”** or **“Likely to Leave”**.  
 ✔ Highlights the **most influential factors** contributing to attrition (e.g., low salary, excessive overtime).  
 ✔ Generates **HR-friendly reports** summarizing attrition risks and recommendations.

### **5. HR User Interface Module**

✔ Allows **HR users to upload employee data** for analysis.  
 ✔ Provides a **dashboard** displaying prediction results, attrition trends, and key insights.  
 ✔ Enables **report downloads** for decision-making and employee retention planning.

### **System Workflow**

1. **HR uploads employee data** → Data is preprocessed and analyzed.  
2. **Machine learning models predict attrition** → Results are generated.  
3. **HR views prediction reports** → Insights help in making retention decisions.

### **User Interface**



### **Coding**

**Machine Learning:**

**FrontEnd:**

**Index.html---------------🡪**

<!DOCTYPE html>

<html>

<head>

  <title>Employee Attrition Prediction</title>

  <link href="https://cdn.jsdelivr.net/npm/bootstrap@5.3.0/dist/css/bootstrap.min.css" rel="stylesheet">

  <link rel="stylesheet" href="style.css">

</head>

<body class="p-4">

  <div class="container">

    <h2>Predict Employee Attrition</h2>

    <form id="predictionForm" class="row g-3">

      <div class="col-md-4"><label>Age</label><input name="Age" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>DailyRate</label><input name="DailyRate" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>DistanceFromHome</label><input name="DistanceFromHome" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>Education</label><input name="Education" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>EnvironmentSatisfaction</label><input name="EnvironmentSatisfaction" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>HourlyRate</label><input name="HourlyRate" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>JobInvolvement</label><input name="JobInvolvement" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>JobLevel</label><input name="JobLevel" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>JobSatisfaction</label><input name="JobSatisfaction" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>MonthlyIncome</label><input name="MonthlyIncome" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>MonthlyRate</label><input name="MonthlyRate" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>NumCompaniesWorked</label><input name="NumCompaniesWorked" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>PercentSalaryHike</label><input name="PercentSalaryHike" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>PerformanceRating</label><input name="PerformanceRating" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>RelationshipSatisfaction</label><input name="RelationshipSatisfaction" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>StockOptionLevel</label><input name="StockOptionLevel" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>TotalWorkingYears</label><input name="TotalWorkingYears" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>TrainingTimesLastYear</label><input name="TrainingTimesLastYear" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>WorkLifeBalance</label><input name="WorkLifeBalance" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>YearsAtCompany</label><input name="YearsAtCompany" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>YearsInCurrentRole</label><input name="YearsInCurrentRole" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>YearsSinceLastPromotion</label><input name="YearsSinceLastPromotion" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>YearsWithCurrManager</label><input name="YearsWithCurrManager" type="number" class="form-control" required></div>

      <div class="col-md-4"><label>BusinessTravel</label>

        <select name="BusinessTravel" class="form-control">

          <option value="Non-Travel">Non-Travel</option>

          <option value="Travel\_Rarely">Travel\_Rarely</option>

          <option value="Travel\_Frequently">Travel\_Frequently</option>

        </select>

      </div>

      <div class="col-md-4"><label>Department</label>

        <select name="Department" class="form-control">

          <option value="Research & Development">Research & Development</option>

          <option value="Sales">Sales</option>

          <option value="Human Resources">Human Resources</option>

        </select>

      </div>

      <div class="col-md-4"><label>EducationField</label>

        <select name="EducationField" class="form-control">

          <option value="Life Sciences">Life Sciences</option>

          <option value="Medical">Medical</option>

          <option value="Marketing">Marketing</option>

          <option value="Technical Degree">Technical Degree</option>

          <option value="Human Resources">Human Resources</option>

          <option value="Other">Other</option>

        </select>

      </div>

      <div class="col-md-4"><label>Gender</label>

        <select name="Gender" class="form-control">

          <option value="Male">Male</option>

          <option value="Female">Female</option>

        </select>

      </div>

      <div class="col-md-4"><label>JobRole</label>

        <select name="JobRole" class="form-control">

          <option value="Sales Executive">Sales Executive</option>

          <option value="Research Scientist">Research Scientist</option>

          <option value="Laboratory Technician">Laboratory Technician</option>

          <option value="Manager">Manager</option>

          <option value="Human Resources">Human Resources</option>

        </select>

      </div>

      <div class="col-md-4"><label>MaritalStatus</label>

        <select name="MaritalStatus" class="form-control">

          <option value="Single">Single</option>

          <option value="Married">Married</option>

          <option value="Divorced">Divorced</option>

        </select>

      </div>

      <div class="col-md-4"><label>OverTime</label>

        <select name="OverTime" class="form-control">

          <option value="Yes">Yes</option>

          <option value="No">No</option>

        </select>

      </div>

      <div class="col-12 mt-4">

        <button type="submit" class="btn btn-primary">Predict</button>

      </div>

    </form>

    <div id="result" class="mt-4"></div>

  </div>

  <script>

    document.addEventListener("DOMContentLoaded", function () {

      document.getElementById("predictionForm").addEventListener("submit", async function(e) {

        e.preventDefault(); // prevent page reload

        const formData = new FormData(e.target);

        const jsonData = {};

        formData.forEach((val, key) => jsonData[key] = val);

        console.log("Sending data:", jsonData);

        try {

          const res = await fetch("http://127.0.0.1:5000/predict", {

            method: "POST",

            headers: { "Content-Type": "application/json" },

            body: JSON.stringify(jsonData)

          });

          const data = await res.json();

          console.log("Response from server:", data);

          document.getElementById("result").innerHTML = `

            <div class="alert alert-info">

              <strong>Prediction:</strong> ${data.prediction}<br>

              <strong>Probability:</strong> ${data.probability \* 100}%

            </div>

          `;

        } catch (err) {

          console.error("Error:", err);

          document.getElementById("result").innerHTML = `

            <div class="alert alert-danger">Something went wrong. Check console for details.</div>

          `;

        }

      });

    });

  </script>

</body>

</html>

**Style.css----------------🡪**

body {

    background: linear-gradient(to right, #f5f7fa, #c3cfe2);

    font-family: 'Segoe UI', Tahoma, Geneva, Verdana, sans-serif;

  }

  .container {

    background-color: white;

    padding: 30px;

    border-radius: 15px;

    box-shadow: 0 8px 16px rgba(0,0,0,0.15);

    margin-top: 40px;

  }

  h2 {

    color: #2c3e50;

    font-weight: 600;

    margin-bottom: 30px;

  }

  label {

    font-weight: 500;

    color: #34495e;

  }

  input, select {

    border-radius: 8px !important;

    box-shadow: none !important;

    border: 1px solid #ccc !important;

    padding: 8px;

  }

  button.btn-primary {

    background-color: #3498db;

    border: none;

    font-weight: bold;

    padding: 10px 20px;

    border-radius: 10px;

    transition: background-color 0.3s ease;

  }

  button.btn-primary:hover {

    background-color: #2980b9;

  }

  #result {

    font-size: 18px;

    margin-top: 20px;

  }

  .alert-info {

    background-color: #ecf6fc;

    border-left: 5px solid #3498db;

    padding: 15px;

    border-radius: 8px;

    color: #2c3e50;

  }

**Backend:**

**ML\_Server.py--------------🡪**

from flask import Flask, request, jsonify

from flask\_cors import CORS

import pickle

import numpy as np

app = Flask(\_\_name\_\_)

CORS(app)

# Load the trained model

with open('model.pkl', 'rb') as f:

    model = pickle.load(f)

# Label encoding maps (MUST match training)

label\_maps = {

    'BusinessTravel': {'Non-Travel': 0, 'Travel\_Frequently': 1, 'Travel\_Rarely': 2},

    'Department': {'Research & Development': 0, 'Sales': 1, 'Human Resources': 2},

    'EducationField': {'Life Sciences': 0, 'Medical': 1, 'Marketing': 2, 'Technical Degree': 3, 'Human Resources': 4, 'Other': 5},

    'Gender': {'Male': 1, 'Female': 0},

    'JobRole': {'Sales Executive': 7, 'Research Scientist': 6, 'Laboratory Technician': 4, 'Manager': 3, 'Human Resources': 2},

    'MaritalStatus': {'Single': 2, 'Married': 1, 'Divorced': 0},

    'OverTime': {'Yes': 1, 'No': 0}

}

@app.route('/predict', methods=['POST'])

def predict():

    data = request.get\_json()

    feature\_order = [

        'Age', 'DailyRate', 'DistanceFromHome', 'Education', 'EnvironmentSatisfaction',

        'HourlyRate', 'JobInvolvement', 'JobLevel', 'JobSatisfaction', 'MonthlyIncome',

        'MonthlyRate', 'NumCompaniesWorked', 'PercentSalaryHike', 'PerformanceRating',

        'RelationshipSatisfaction', 'StockOptionLevel', 'TotalWorkingYears',

        'TrainingTimesLastYear', 'WorkLifeBalance', 'YearsAtCompany',

        'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager',

        'BusinessTravel', 'Department', 'EducationField', 'Gender',

        'JobRole', 'MaritalStatus', 'OverTime'

    ]

    input\_data = []

    for col in feature\_order:

        if col in label\_maps:

            input\_data.append(label\_maps[col][data[col]])

        else:

            input\_data.append(int(data[col]))

    X = np.array([input\_data])

    pred = model.predict(X)[0]

    prob = model.predict\_proba(X)[0][1]

    return jsonify({

        'prediction': 'Yes' if pred == 1 else 'No',

        'probability': round(float(prob), 4)

    })

if \_\_name\_\_ == '\_\_main\_\_':

    app.run(debug=True)

### 

### **Testing & Security**

The **Employee Attrition Prediction System** undergoes rigorous testing to ensure **accuracy, reliability, and security** while maintaining data integrity.

### **Testing Approach :**

1. **Unit Testing** – Verifies individual modules like data preprocessing, model training, and predictions.  
 2. **Integration Testing** – Ensures smooth interaction between data processing, ML models, and report generation.  
 3. **System Testing** – Evaluates overall system performance and checks for inconsistencies.  
 4. **Performance Testing** – Measures speed, response time, and scalability for large datasets.  
 5. **User Acceptance Testing (UAT)** – Ensures usability and accuracy based on HR feedback.

### **Security Measures**

✔ **Data Encryption** – Protects sensitive employee data using secure encryption methods.  
 ✔ **Access Control** – Restricts system access to authorized HR personnel.  
 ✔ **Secure Data Storage** – Ensures safe handling of employee records and prediction results.  
 ✔ **Regular Audits** – Monitors system logs to detect unauthorized access or anomalies.

### 

### **Future Scope**

The **Employee Attrition Prediction System** has the potential for further enhancements to improve accuracy, usability, and integration with HR management systems. Future developments may include:

✔ **Integration with HRM Systems** – Connecting the model with HR software for real-time monitoring and decision-making.  
 ✔ **Improved Machine Learning Models** – Incorporating advanced techniques like deep learning for better prediction accuracy.  
 ✔ **Automated Retention Strategies** – Suggesting personalized employee retention plans based on attrition risk.  
 ✔ **Real-Time Data Processing** – Enabling continuous updates and instant attrition predictions.  
 ✔ **Interactive Dashboards** – Enhancing visual analytics with dynamic reports and predictive insights.

### 

### **Bibliography**

The following sources were referenced during the development of the **Employee Attrition Prediction System**:

1. <https://www.researchgate.net/publication/326029536_Employee_Attrition_Prediction>
2. <https://www.analyticsvidhya.com/blog/2021/11/employee-attrition-prediction-a-comprehensive-guide/>
3. <https://www.geeksforgeeks.org/machine-learning-models/>
4. <https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset>
5. Raza, A., Munir, K., Almutairi, M., Younas, F., & Fareed, M. M. S. (2022). Predicting Employee Attrition Using Machine Learning Approaches. *Applied Sciences*, *12*(13), 6424. <https://doi.org/10.3390/app12136424>
6. Minwir Al-Shammari, Yahya A. Ghanem, "A Systematic Literature Review of Quantitative Models for Predicting Employee Attrition", *2024 International Conference on Decision Aid Sciences and Applications (DASA)*, pp.1-7, 2024.
7. Vijay Kumar Kambhampati, K. B. V. Brahma Rao, "Advancing Employee Attrition Models: A Systematic Review of Machine Learning Techniques and Emerging Research Opportunities", *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp.1100-1107, 2024.