IBM Innovation Center for Education

PHEMESOFT Software That Matters
TRUSTED DELIVERY PARTNER for IBM ICE

Sri Sri University

Project - Low Level Design

on

Employee Attrition Prediction

Team members:

Ayush Patil
Shrestha Agarwal
Samriddhi Jain
Akash Samantray

Guided by: Mr. Sudhir Mohapatra

Industry Mentor:

Table of Contents

Scope of Document

The document's scope encompasses a detailed analysis of employee attrition prediction through exploratory data analysis (EDA) and model building. It aims to provide insights into factors influencing attrition rates within an organization by leveraging data-driven techniques. The document covers importing essential libraries, data cleaning, visualization methods, statistical analysis, and model training using algorithms like Random Forest Regressor and SVM. It focuses on understanding the dataset structure, exploring correlations between variables, and building predictive models to forecast employee attrition accurately.

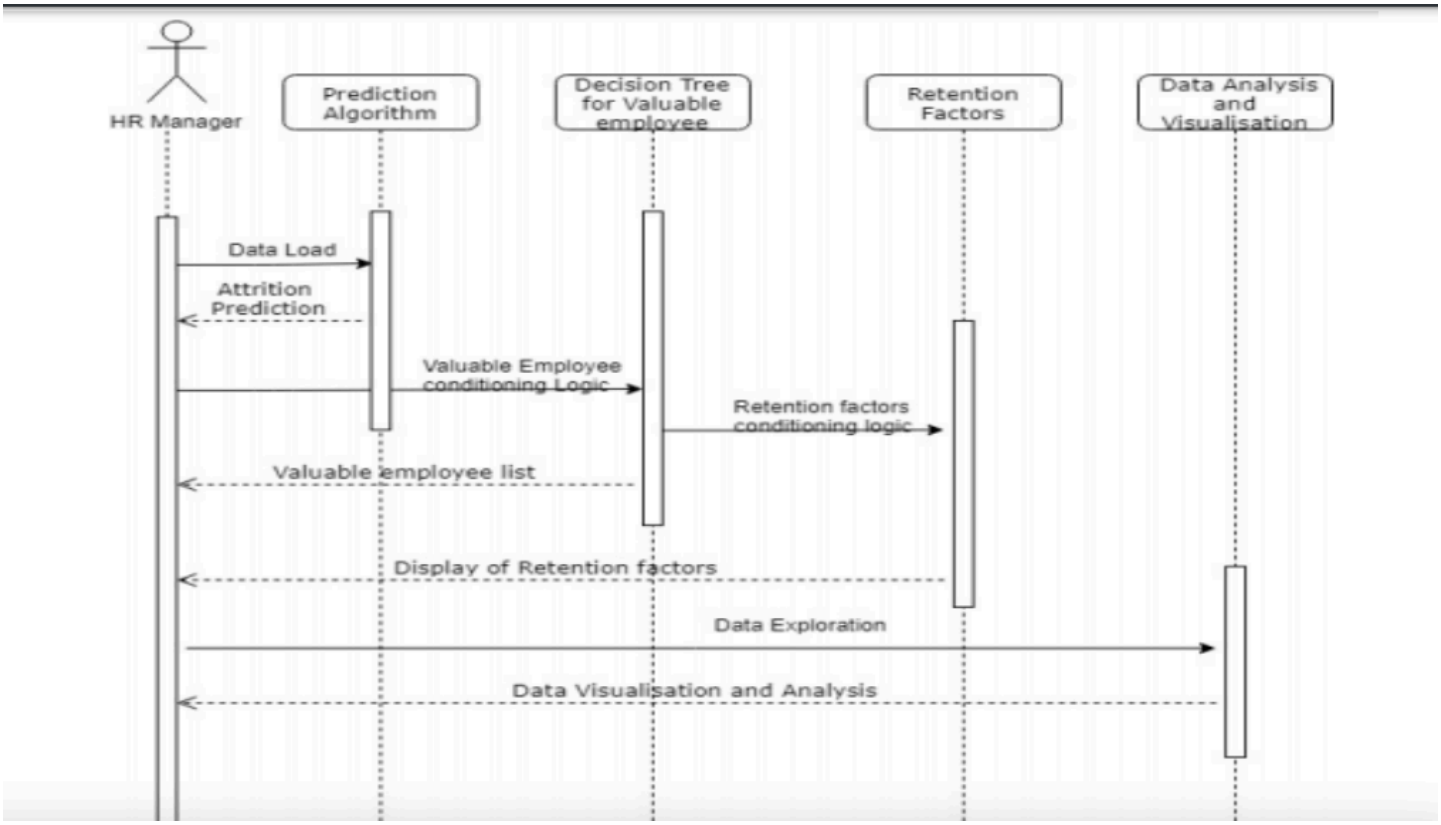
Intended Audience

The intended audience for this document includes data analysts, HR professionals, and individuals involved in human resource management and workforce planning. Data scientists seeking to understand employee attrition patterns and develop predictive models will find the document valuable. HR managers looking to enhance retention strategies based on data insights will benefit from the analysis presented. Additionally, stakeholders interested in leveraging machine learning for HR analytics and decision-making processes can gain valuable information from the document.

System Overview

The system overview provides a comprehensive insight into the process of importing libraries for data analysis, conducting exploratory data analysis (EDA) to gain a deeper understanding of the dataset, and building predictive models for employee attrition prediction. It highlights the importance of data cleaning, visualization techniques like heatmaps and pie charts, and statistical analysis to extract meaningful insights from the dataset. The overview emphasizes the significance of model building using algorithms tailored to predict employee attrition accurately based on the provided dataset structure.

Sequence Diagram



1. Data Import:

The commencement of the employee attrition prediction process involves the crucial step of data import. This initiates by incorporating fundamental Python libraries, such as pandas, matplotlib.pyplot, seaborn, and numpy, into the working environment. These libraries play a pivotal role in facilitating subsequent data manipulation, visualization, and numerical operations. Following the library import, the focus shifts to loading the employee dataset, leveraging the capabilities of pandas for efficient handling and analysis. This phase ensures that the necessary tools are in place for the comprehensive analysis that follows.

2. Data Cleaning:

Once the dataset is imported, the next pivotal step in the process is data cleaning. This intricate stage aims to enhance data quality by addressing issues such as missing values, irrelevant columns, and anomalies. Through systematic approaches like imputation, dropping unnecessary features, and outlier handling, the dataset is meticulously prepared for downstream analysis. The integrity of the dataset is

paramount in ensuring that the subsequent models are trained on reliable and representative information, thereby improving the accuracy of predictions.

3. Exploratory Data Analysis (EDA):

Following data cleaning, the process enters the exploratory data analysis (EDA) phase. This involves a comprehensive exploration of the dataset to glean insights into its structure and characteristics. Using matplotlib and seaborn, visualizations are generated to illustrate summary statistics, distributions, and correlations between different variables. EDA serves as a crucial step in understanding patterns and trends within the employee data, providing valuable context for subsequent modeling decisions.

4. Model Training:

With a refined and comprehensible dataset, the process advances to model training. This phase involves the division of the dataset into training and testing sets, enabling the application of machine learning algorithms. Algorithms like Random Forest Regressor and Support Vector Machines (SVM) are employed to build predictive models. These models utilize historical employee data to discern patterns and relationships, ultimately enabling accurate predictions of attrition rates.

5 .Model Evaluation:

Post-model training, the emphasis shifts to model evaluation, a critical step in assessing the performance and generalization capabilities of the trained models. Cross-validation techniques are employed to rigorously evaluate the model's accuracy, precision, recall, and other pertinent metrics. This ensures that the model is robust and can effectively make predictions on unseen data, reinforcing its reliability for practical applications.

6. Model Deployment:

The final stage of the sequence diagram involves the deployment of the trained model for practical use. This entails integrating the model into a production environment where it can receive new data inputs and provide real-time or batch predictions. The deployment phase ensures that the predictive capabilities of the model are harnessed for informed decision-making in workforce management, completing the end-to-end process of employee attrition prediction.

Components Design Implementation

1. **Feature Engineering Component:**
   Responsibility: Extract meaningful features from the raw data.
   Components:
   Feature Extractor: Identifies and extracts relevant features (e.g., performance metrics, satisfaction scores).
   Feature Transformer: Converts raw data into a format suitable for model input.
   Feature Selector: Chooses the most relevant features for prediction.

2. **Modeling Component:**

   Responsibility: Build and train machine learning models for attrition prediction.
   Components:
   Model Trainer: Trains the prediction model using historical data.
   Model Validator: Validates the model's performance using cross-validation techniques.
   Model Tuner: Fine-tunes hyperparameters to optimize model performance.

3. **Prediction Component:**

   Responsibility: Generate attrition predictions based on input data.
   Components:
   Prediction Engine: Applies the trained model to new data for attrition predictions.
   Confidence Estimator: Calculates confidence levels for each prediction.

References

1. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
2. https://scikit-learn.org/stable/modules/svm.html
3. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html
4. https://xgboost.readthedocs.io/