High Level Design(HLD)

Employee Attrition Prediction

Revision Number : 2.0

Last date of revision :12-06-2023

Praveen Kumar

Thokchom Mobisana

**Document Version Control**

|  |  |  |  |
| --- | --- | --- | --- |
| **Date Issued** | **Version** | **Description** | **Author** |
| 03-06-2023 | 1.0 | Added Introduction and General description | Thokchom Mobisana |
| 04-06-2023 | 1.1 | Added Design details and abstract | Thokchom Mobisana |
| 06-06-2023 | 1.2 | Created and organized whole document | Praveen Kumar |
| 07-06-2023 | 1.3 | Updated Design Details | Thokchom Mobisana |
| 09-06-2023 | 1.4 | Checked and added Performance | Praveen Kumar |
| 10-06-2023 | 1.5 | Added Conclusion and References | Praveen Kumar |
| 12-06-2023 | 2.0 | Updated diagrams in Design Details | Thokchom Mobisana |

**Table of Content**

Document Version Control

Abstract

Introduction

1. Why this High-Level Document?

2. Scope

3. Definition

General Description

1. Product Perspective

2. Problem Statement

3. Problem Solution

4. Proposed Methodology

5. Further Improvements

6. Data Required

7. Tools Used

8. Constraints

9. Assumptions

Design Details

1. Process Workflow

2. Error Handling

Performance

1. Reusability

2. Application compatibility

3. Resources Utilization

4. Deployment

Conclusion

Reference

**Abstract**

Employee attrition is a significant concern for organizations due to its associated costs and impacts on workforce stability. Predicting employee attrition using machine learning techniques can provide valuable insights for effective workforce management and retention strategies. This project focuses on using a Random Forest Classifier, an ensemble learning algorithm, to predict employee attrition. The approach involves collecting relevant employee data, preprocessing the data by handling missing values and encoding categorical variables, and splitting the data into training and testing sets. The Random Forest Classifier is trained on the training set, and hyperparameters are adjusted to optimize its performance. The model's performance is evaluated using appropriate evaluation metrics, and the feature importance provided by the model is analyzed to identify key factors influencing attrition. The trained model can then be deployed to make predictions on new, unseen data, aiding in workforce planning and implementing targeted retention strategies. The use of a Random Forest Classifier offers advantages such as handling categorical and numerical features, robustness to overfitting, and the ability to handle high-dimensional data. By following this approach, organizations can gain insights into attrition patterns, reduce costs associated with turnover, and create a positive work environment conducive to employee satisfaction and long-term success.

**Introduction**

**1. Why this High-Level Design Document?**

The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

* Present all the design aspects and define them in detail
* Describe the user interface being implemented
* Describe the hardware and software interfaces
* Describe the performance requirements
* Include design features and the architecture of the project

**2. Scope**

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly technical terms which should be understandable to the administrators of the system.

**General Description**

**1. Product Perspective**

The product prospect of the employee attrition prediction project using a Random Forest Classifier is to provide organizations with a data-driven tool that can assess the risk of employee attrition, identify key factors influencing attrition, and aid in the development of targeted retention strategies. This can help organizations reduce costs associated with turnover, improve workforce planning, and create a positive work environment conducive to employee satisfaction and long-term success.

**2. Problem Statement**

To create the machine learning based solution to predict employee attrition rate based on the parameters.

**3. Problem Solution**

Develop the web application to predict the air quality index and quality of air, which can help citizen to decide whether to go out or not and to alert the citizen in particular area if AQI is relatively high.

**4. Further Improvement**

Further improvements of the employee attrition prediction using a Random Forest Classifier can include enhancing feature engineering by incorporating additional relevant data, optimizing the model through advanced hyperparameter tuning techniques, and exploring ensemble methods to improve prediction accuracy and capture diverse aspects of attrition. These improvements aim to enhance the model's performance, robustness, and ability to provide actionable insights for effective workforce management and retention strategies.

**5. Data Required**

For training the model we need the Data is completely depending upon our problem statement. Here the given database(github) is used.

**6. Tools Used**

* Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn are used to build the whole model.
* PyCharm and Visual Studio Code is used as IDE.
* For visualization of the plots, Matplotlib and Seaborn are used.
* Heroku is used for deployment of the model.
* GitHub is used as version control system.

**7. Constraints**

Limited or incomplete data availability, poor data quality, and potential challenges in interpreting feature importance accurately.

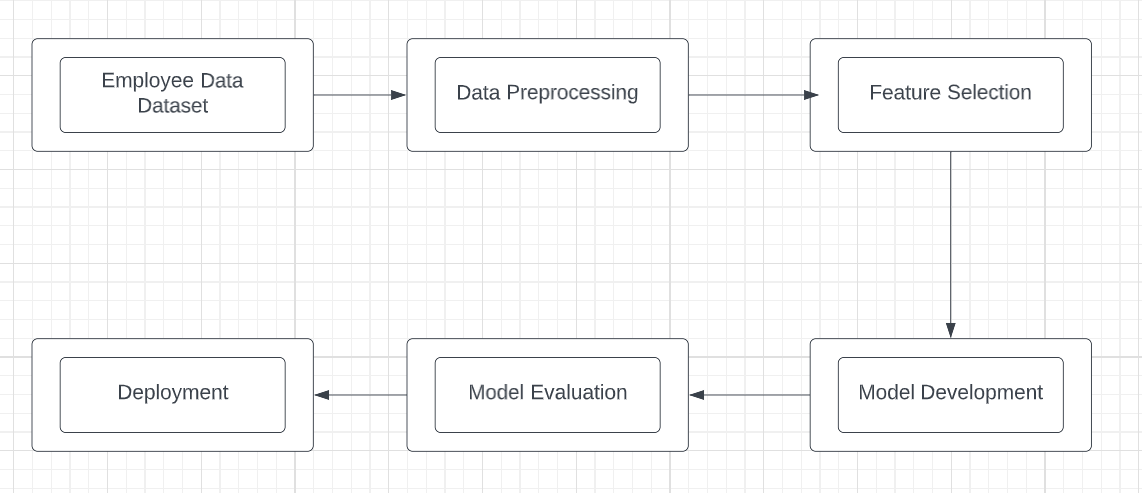
**8. Assumptions**

Assumptions made in employee attrition prediction using a Random Forest Classifier include independence of observations, stability of factors, feature relevance, and model generalization.

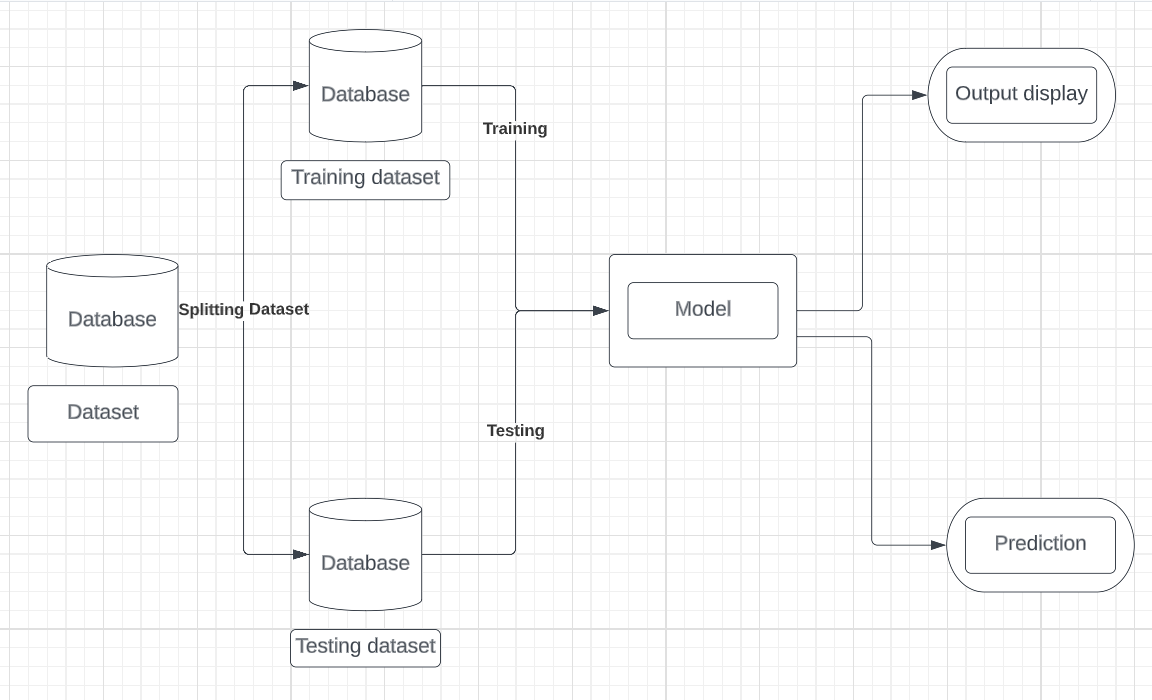
**Design Details**

**1. Process Workflow**

For identifying the different types of anomalies, we will use a machine learning model. Below is the process flow diagram.



Model Training and Evaluation



**2. Error Handling**

Error handling in employee attrition prediction using a Random Forest Classifier involves addressing potential issues that may arise during data preprocessing, model training, and prediction stages. One common issue is missing data, which can be handled by imputing values using techniques like mean imputation or multiple imputation. Outliers can be detected and handled by removing or transforming extreme values. In cases of imbalanced data, techniques like oversampling the minority class or undersampling the majority class can be employed. If there are errors in the data, data cleaning techniques such as outlier removal or error correction can be applied. Additionally, handling class imbalance can be addressed through techniques like stratified sampling or using weighted loss functions during model training. Regular model validation and evaluation should also be conducted to identify and address potential overfitting or underfitting issues.

**Performance**

**1. Reusability**

To enhance reusability, use modular code with well-defined functions and configurable parameters to enable easy integration into different projects or systems, allowing the model to be reused without extensive modifications.

**2. Application compatibility**

Ensuring application compatibility in the context of employee attrition prediction using a Random Forest Classifier involves making the model compatible with the technology stack, programming languages, and frameworks used in the target application. This includes ensuring that the model can be seamlessly integrated into the application's infrastructure, data pipeline, and user interface. Compatibility can be achieved through standardization of data formats, APIs, and adhering to industry standards. Additionally, providing clear documentation and support resources can facilitate the integration process for developers and users of the application.

**3. Resource utilization**

Resource utilization in employee attrition prediction involves efficiently utilizing computational resources such as CPU, memory, and storage during the data preprocessing, model training, and prediction stages. Optimizing resource utilization ensures efficient execution, minimizes costs, and allows for scalability and performance improvements in large-scale applications.

**4. Deployment**

The code is deployed in GitHub. The whole system is live and is hosted on Heroku.

**Conclusion**

Employee attrition prediction using a Random Forest Classifier is a valuable approach for organizations to proactively manage and address attrition challenges. By leveraging machine learning techniques, organizations can gain insights into the factors influencing attrition, develop targeted retention strategies, and optimize workforce planning. Despite potential constraints and assumptions, the project offers promising prospects for reducing costs, improving employee satisfaction, and creating a stable and productive work environment. Continued improvements in data quality, model optimization, and error handling can enhance the accuracy and effectiveness of the predictions, leading to better decision-making and long-term organizational success.

**References**

1. <https://docs.streamlit.io/en/stable/>
2. <https://scikit-learn.org/stable/user_guide.html>
3. <https://numpy.org/doc/>
4. <https://seaborn.pydata.org/examples/regression_marginals.html>
5. <https://seaborn.pydata.org/examples/scatterplot_matrix.html>
6. <https://matplotlib.org/>
7. <https://pandas.pydata.org/docs/>