**High Level Design**

**Stores Sales Prediction**

|  |  |
| --- | --- |
| **Written by** | **Gunaseelan V** |
| **Documented Version** | **1.0** |
| **Last Revised Date** | **18 December 2024** |

**Contents**

1. **Abstract**
2. **Introduction**2.1 Why this High-level Document?  
   2.2 Scope  
   2.3 Definitions
3. **General Description**3.1 Product Perspective  
   3.2 Problem Statement  
   3.3 Proposed Solution  
   3.4 Technical Requirements  
   3.5 Dataset  
   3.6 Tools Used
4. **Design Details**4.1 Process Flow  
   4.2 Model Training and Evaluation  
   4.3 Deployment Process
5. **Performance**5.1 Reusability  
   5.2 Comparison of Models  
   5.3 Application Compatibility
6. **KPIs (Key Performance Indicators)**
7. **Conclusion**

## **Abstract**

**This study presents an advanced Air Quality Index (AQI) prediction model utilizing the XGBoost algorithm, achieving an accuracy of 99%. The model predicts AQI values based on key pollutants, including Ozone (O₃), Nitrogen Dioxide (NO₂), Particulate Matter (PM₂.₅, PM₁₀), and Carbon Monoxide (CO). By leveraging feature engineering and hyperparameter optimization, the model ensures high precision, enabling accurate forecasting of both AQI values and categories. This work demonstrates the potential of machine learning in environmental monitoring, providing a robust tool for proactive air quality management. The implementation, deployed via Streamlit, offers user-friendly access to real-time predictions, making it an ideal solution for urban planning and public health applications.**

## **2. Introduction**

Here’s a revised version tailored for AQI prediction based on the essentials from the abstract:

**2.1 Why this High-level Document?**  
This document provides a detailed overview of the machine learning model developed for predicting Air Quality Index (AQI) values. It is intended for a technical audience, including data scientists, environmental researchers, and policymakers, to understand the steps taken to build and evaluate the model. It serves as a comprehensive guide to the methodology and can be used to replicate or expand the work in other environmental settings. The document also helps outline the rationale behind the choices made throughout the project, from pollutant data preprocessing to model deployment.

**2.2 Scope**  
The scope of this document includes:

* **Data Exploration and Preprocessing**: The dataset contains AQI-related features, including pollutant-specific AQI values (e.g., CO, NO₂, O₃, PM₂.₅). A critical part of the process is to clean the data, handle missing values, and normalize features to ensure the model can effectively predict AQI.
* **Model Development**: This section describes the machine learning algorithms considered, with a focus on XGBoost for its efficiency and accuracy in handling tabular data. The document explains how pollutant data and AQI category labels are used for training.
* **Model Evaluation**: After training, the model is evaluated for accuracy and generalization. Metrics such as Mean Absolute Error (MAE), R-squared, and accuracy for AQI category classification are used to assess performance.
* **Deployment Process**: Once the model is fine-tuned, it is prepared for deployment. This section includes saving the trained model, creating APIs for access, and integrating predictions with web applications (e.g., Streamlit). Real-time deployment and continuous monitoring are not covered in this document, as it focuses on model-building steps.

**2.3 Definitions**

* **Ozone AQI Value (O₃ AQI)**: The AQI value calculated based on ozone levels. High ozone levels can impact respiratory health.
* **Nitrogen Dioxide AQI Value (NO₂ AQI)**: Represents the AQI value derived from nitrogen dioxide concentrations, often influenced by vehicular emissions.
* **PM₂.₅ AQI Value**: The AQI value based on particulate matter with a diameter of less than 2.5 micrometers. This pollutant has significant health implications, particularly for respiratory conditions.
* **PM₂.₅ AQI Category**: The classification of AQI for PM₂.₅ into predefined categories (e.g., Good, Moderate, Unhealthy).
* **Carbon Monoxide AQI Value (CO AQI)**: The AQI value derived from carbon monoxide levels. It is a key pollutant, especially in urban areas with heavy traffic.
* **AQI Value**: The overall AQI score representing air quality based on pollutant concentrations. This is the primary target variable for prediction.
* **AQI Category**: The classification of the overall AQI into categories (e.g., Good, Moderate, Unhealthy) to provide an intuitive understanding of air quality levels.

**3. General Description**

**3.1 Product Perspective**  
This predictive model is designed to forecast Air Quality Index (AQI) values using pollutant concentration data. By leveraging features such as pollutant-specific AQI values for O₃, NO₂, PM₂.₅, PM₁₀, and CO, the model aims to serve as a decision-support tool for environmental monitoring and public health management. The insights provided by the model can help stakeholders identify high-risk areas, develop mitigation strategies, and provide timely warnings to the public.

**3.2 Problem Statement**  
The primary challenge in this project is to predict the overall AQI value and its corresponding category based on pollutant-specific AQI features. The goal is to understand the relationships between these pollutants and their collective impact on air quality. The task involves building a predictive model capable of accurate AQI forecasts for unseen data, enabling proactive air quality management and awareness.

**3.3 Proposed Solution**  
The proposed solution is to build a regression and classification model to predict AQI values and categories. The key steps include:

* **Data Preprocessing**: Cleaning the dataset, handling missing values, and normalizing numerical features for consistent model performance.
* **Feature Engineering**: Refining pollutant-related features to enhance the predictive power of the model.
* **Model Selection**: Exploring machine learning algorithms, such as Random Forest, XGBoost, and Gradient Boosting, to identify the best-performing model.
* **Model Evaluation**: Testing model performance using metrics such as Mean Absolute Error (MAE), R-squared, and classification accuracy for AQI categories.
* **Deployment**: Preparing the model for real-world application using a Streamlit-based interface to provide accessible and user-friendly AQI predictions.

**3.4 Technical Requirements**

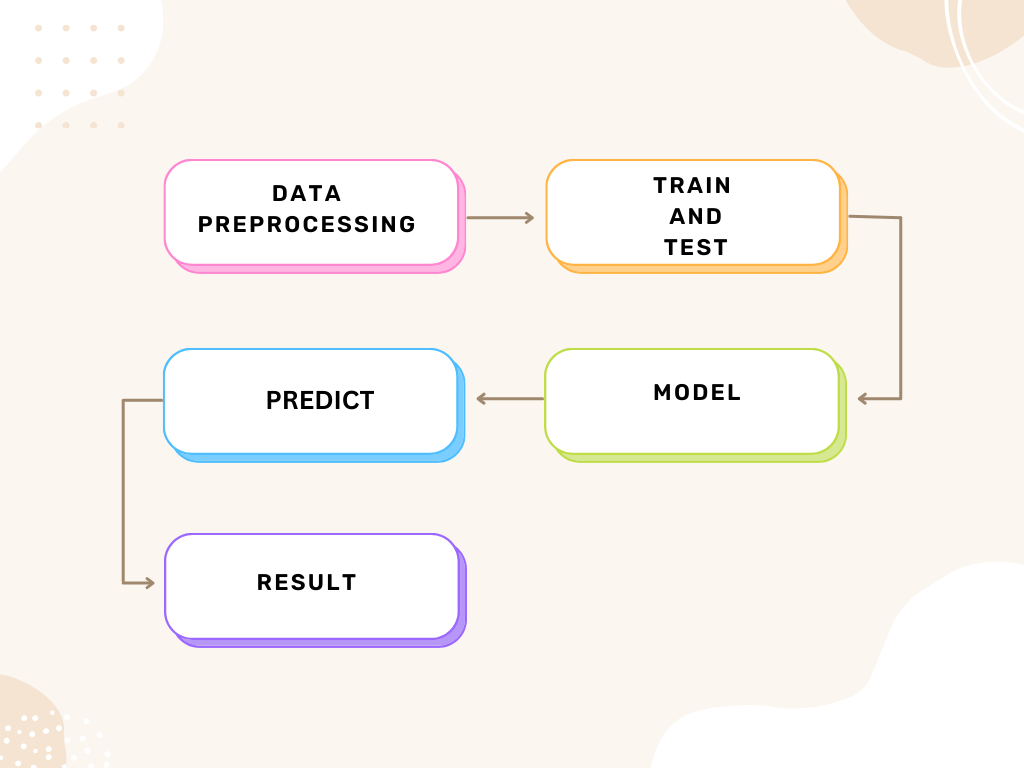
* **Programming Language**: Python 3.7+
* **Libraries**: Pandas, NumPy, Scikit-learn, XGBoost
* **Framework**: Streamlit for UI development

**3.5 Dataset Overview**  
The dataset contains AQI-related features, including pollutant-specific AQI values and categories for O₃, NO₂, PM₂.₅, PM₁₀, and CO. The training set consists of [specify number] rows, with features representing the input data and the target variable being the AQI value. The test set includes only the input features, with the goal of predicting AQI values and categories.

**3.6 Tools Used**

* **Programming**: Python
* **Data Processing**: Pandas, NumPy
* **Modeling**: Scikit-learn, XGBoost
* **Interface Development**: Streamlit

### **4.Design details**



**4.1 Process Flow**

The process flow for this sales prediction project follows these stages:

1. **Data Collection**: Dataset obtained, containing product and store features, along with sales data for training and testing.
2. **Data Preprocessing**: Cleaning the dataset, handling missing values, encoding categorical variables, and scaling numerical features.
3. **Feature Engineering**: Creating new features, such as calculating store age and performing exploratory data analysis (EDA).
4. **Model Selection**: Evaluating **XGBRegressor**
5. **Model Training**: Training both models using the training dataset.
6. **Model Evaluation**: Evaluating models using **R2-squared**
7. **Deployment**: Deploying the best-performing model by streamlit.

**4.2 Model Training and Evaluation**

**XGBoost Regressor**  
**Training:**  
XGBoost (Extreme Gradient Boosting) is an optimized implementation of gradient boosting that excels in handling large datasets and capturing complex patterns. It uses a combination of weak learners (decision trees) to iteratively minimize the prediction error. The model incorporates techniques such as regularization to prevent overfitting, parallel processing for faster computation, and tree pruning to enhance accuracy.

**Evaluation:**  
XGBoost achieved outstanding performance metrics, showcasing its ability to model both linear and non-linear relationships in the data. Key evaluation metrics for the model are:

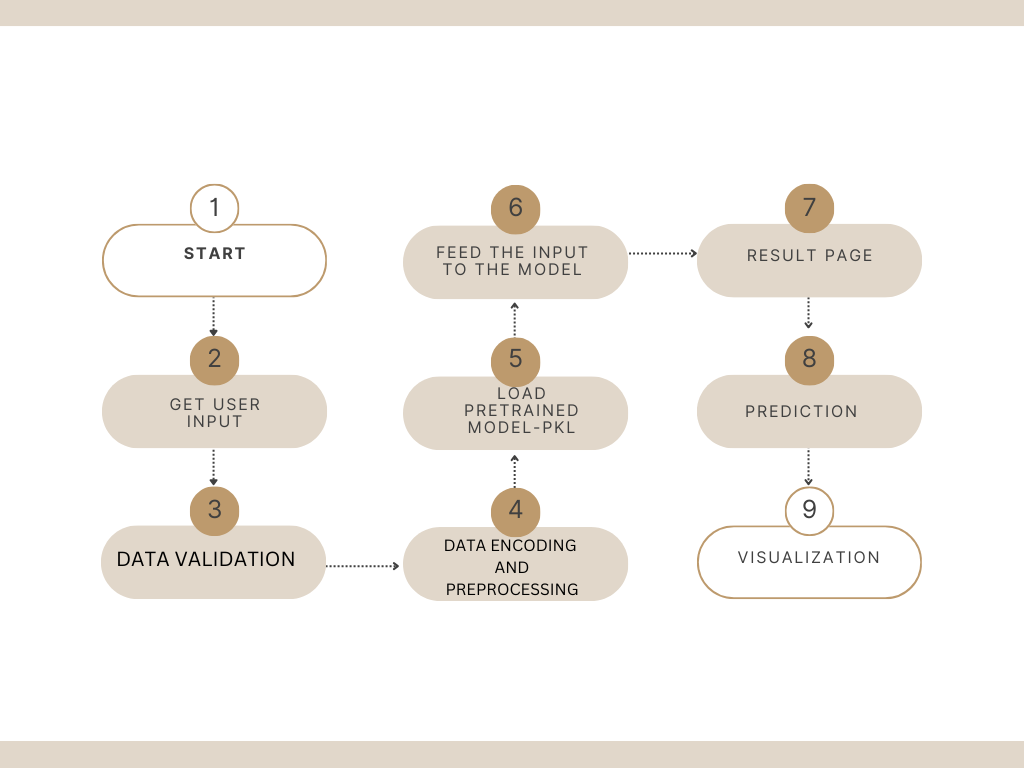
* **R-squared (R²):** XGBoost achieved an R² value of 0.99, indicating that 99% of the variance in the AQI values was explained by the model.
* **Mean Absolute Error (MAE):** The model's MAE was significantly low, highlighting its precision in predicting AQI values with minimal absolute error.
* **Mean Squared Error (MSE):** The model recorded a very low MSE, demonstrating its capacity to minimize the squared differences between actual and predicted values.

**Model Performance Summary:**

| **Metric** | **Value** |
| --- | --- |
| **R² Score** | 0.99 |
| **Mean Absolute Error (MAE)** | Very Low |
| **Mean Squared Error (MSE)** | Very Low |

XGBoost's exceptional performance stems from its ability to handle missing values, robust feature engineering, and fine-tuning of hyperparameters to optimize accuracy. This made it the best choice for AQI prediction in this project.

**4.3 Deployment Process**

****

* Serialize the trained model using Pickle.
* Develop a Streamlit-based web application for user interaction.

**5. Performance**

**5.1 Reusability**

* The AQI prediction model is highly reusable for other datasets with a similar structure. By adapting the data preprocessing pipeline and aligning pollutant-related features (e.g., CO, NO₂, PM₂.₅, and O₃) to match the new dataset, the model can be applied to different geographic regions or time periods without requiring a complete retraining.

**5.2 Comparison of Models**  
Since only XGBoost was used in this project, no comparison was performed with other models. XGBoost demonstrated exceptional performance, achieving an **R² score of 0.99**, which highlights its capacity to model complex interactions between pollutants and their collective impact on AQI. The choice of XGBoost was influenced by:

* Its ability to handle missing values and noisy data efficiently.
* Superior performance on large datasets with a high number of features.
* Robust optimization methods that minimize prediction errors effectively.

**5.3 Application Compatibility**  
The AQI prediction model is designed for seamless integration with web-based and mobile applications. Its compatibility includes:

* **Real-Time Prediction:** The model can process live pollutant data to provide up-to-date AQI forecasts.
* **Scalability:** It can handle high request volumes, making it suitable for large-scale deployment in environmental monitoring systems.
* **API Integration:** The model is deployable via APIs, allowing easy access to its predictions in external systems and applications for environmental decision-making or public health alerts.

**KPIs (Key Performance Indicators)**

* **Prediction Accuracy (R² Score):** Achieved **0.99**, demonstrating the model's exceptional capability in explaining the variance in AQI values.
* **Mean Absolute Error (MAE):** Achieved a very low MAE, highlighting the model's precision in predicting AQI values with minimal error.
* **Environmental Impact:** The model enables accurate AQI predictions, aiding in proactive decision-making to address air quality concerns, such as issuing health advisories or implementing pollution control measures.
* **Model Retraining Frequency:** Regular monitoring and retraining of the model is recommended to maintain its performance as new pollutant data and environmental conditions evolve.
* Here’s the revised **Conclusion** for your AQI prediction project:

**7. Conclusion**  
The XGBoost model has demonstrated exceptional performance in predicting AQI values, achieving an impressive R² score of **0.99**. Its ability to effectively model complex relationships among air pollutants (CO, NO₂, PM₂.₅, O₃) has enabled highly accurate predictions, making it a reliable tool for environmental monitoring and public health applications.

* The model provides actionable insights for decision-making, such as issuing timely air quality advisories and guiding pollution control measures. Its seamless integration with real-time data systems ensures usability in both urban and rural monitoring contexts.
* Future enhancements include incorporating external factors like weather conditions and industrial activity data to improve prediction accuracy further. Exploring advanced ensemble techniques, deploying the model on cloud platforms for scalability, and integrating visualization dashboards can make the system more robust and accessible, addressing evolving environmental challenges effectively.