Classification Task Report

Student Id	: 2408985
Student Name	: Salon Prajapati
Group	: L5CG20
Module Leader	: Mr. Siman Giri
Tutor	: Mr. Bibek Khana

Contents

1 Introduction

1.1 Problem Statement

This project tries to predict the target variable HealthImpactClass using air quality dataset. The dataset used for analyzing has independent variable like AQI, PM10, PM2.5, NO2, and other factors. This task is done to categorize health risks into classes on the basis of pollution levels in the range from 0 to 1, aiding public health decision making.

1.2 Dataset

The dataset named Aqi.csv has 5811 records and 15 features. These features include specialized data concerning air quality measures such as PM2.5 and O3, along with meteorological indicators like temperature and humidity, and health outcomes such as respiratory and cardiovascular cases. The important dependent variable in this study is HealthImpactClass, which categorizes health impacts into five classes. This analysis contributes to the goals and targets defined by the United Nations Sustainable Development Goal (SDG) 3: Good Health and Well-being by providing insights into feasible approaches to reducing health risks associated with pollution.

1.3 Objective

Model to determine how HealthImpactClass might predict healthcare disparity and the percentage contributory major air quality attributes of the health-of the values of HealthImpactClass.

---

2 Methodology

2.1 Data Preprocessing

To handle the missing values, numerical values for PM10, PM2.5, and other variables were completed by replacing them with the mean of their respective columns. Duplicate entries in the dataset were carefully identified and removed to ensure the accuracy and consistency of the data. Additionally, for better readability and consistency, all column names were standardized by converting them into lowercase letters and replacing any spaces with underscores.

To handle the missing values, numerical values for PM10, PM2.5, and other variables were completed by replacing them with the mean of their respective columns. Some of the entries that appear multiple times are removed from the dataset. Some of the column names are converted into lowercase and underscore are added for space for having better reading in code.

2.2 Exploratory Data Analysis (EDA)

The AQI values available within the dataset has a minimum of 0 and a maximum of 499.86, and they are right-skewed which imply frequent moderate-to-high levels of pollution. Among health impact classes, Class 0 most prevalent, accounting for 71.9% of samples and indicating the lowest health risk. Classes corresponding to higher health risk (1-4) were rather sparsely represented, thus indicating a very noticeable class imbalance. Also, there were moderate correlations between the AQI and PM10 and PM2.5 levels and with incidences in hospital admissions and cardiovascular cases as well.

2.3 Model Building

Two models were employed for the purpose of analysis. The first of these is a logistic regression-trained model, acting in the capacity of a baseline probabilistic classifier. The second model, a decision tree classifier, was employed because it efficiently-captures complex feature interactions in a non-linear way. The data division was 80-20 for training and testing, respectively.

2.4 Model Evaluation

The performance of the models, however, was measured in terms of quite a few important crucial indicators: Accuracy, denoting the ratio of right predictions as compared to total predictions; Precision, mostly concerned with false positive occurrences in the health risk categories; Recall, aimed at finding out as much as possible true positives; and, lastly, the F1-Score, which ties together both precision and recall.

GridSearchCV was used for hyperparameter tuning. The best hyperparameters for the logistic regression were C=0.1 and 'l2' penalty. In a similar way, the decision tree classifier would use maximum depth of 5 and minimum 10 samples per node split for the model to avoid rising overfitting.

2.6 Feature Selection

Recursive Feature Elimination (RFE) identified critical features:

1. **AQI** (most significant)
2. **PM2.5**
3. **NO2**
4. **O3**
5. **CardiovascularCases**

---

3 Conclusion

3.1 Key Findings

According to standard performance metrics, logistic regression outperform its counterpart, the decision tree model, with an accuracy level of 88% against 82%. Given the bulk of negative examples, both models found it hard for the precision and recall scores of the minority classes (3 and 4). On the other hand, AQI and PM2.5 turned out to be significant predictors of health impact in both models.

3.2 Final Model

The Logistic Regression model achieved the highest performance:

- Accuracy: 88%
- F1-Score: 0.85

3.3 Challenges

The analytical process faced certain issues. One had to do with class imbalance, wherein the minority classes (3 and 4) had few observations with respect to the other classes. This resulted in poor sensitivity of the models in most cases to predict these classes. The second issue was major multicollinearity between features PM10 and PM2.5, which needed to be handled carefully to avoid problems concerning the stability of the model and the interpretation of the results.

3.4 Future Work

Various strategies may be employed to enhance the models' potential and toughness. One method is by responding to dissimilarities in classes through SMOTE (Synthetic Minority Over-sampling Technique), or employing a weighted loss function during training which would prioritize lesser represented classes. Another option would involve improving various advanced algorithms such as Random Forests or Gradient Boosting that utilize these approaches' potentials for detecting numerous non-linear patterns and associations among data. Expand the dataset's universe in order to include time and place-related reasons, which can lead to obtaining more intriguing insights

4 Discussion

4.1 Model Performance

The Logistic Regression model performed well overall, though precision dropped considerably for minority classes. For example, Class 4 had lower precision, only 62%, leaving scope for improving detection of severe health impacts.

4.2 Impact of Hyperparameter Tuning

GridSearchCV improved Decision Tree accuracy by **9%** by limiting depth to 5, reducing overfitting.

4.3 Interpretation of Results

Higher AQI and PM2.5 levels directly correlated with severe health impacts (Classes 3–4), aligning with known environmental health risks.

4.4 Limitations

As in the case with the dataset regarding the limitations of the dataset, the possible steps for future research on this dataset may be as follows. Firstly, it has temporal granularity, weekly data was available, but almost no hourly or daily data, thus more precise insight about pollution levels over time and effects on health would arise from that. Secondly, self-reported health outcomes include possible biases and affect the validity of the results. In general, introducing more detailed temporal measurements and utilizing objective health outcome measures would increase the dependability and validity of the analysis.

4.5 Future Research

To improve prediction by the models it is proposed that real time sensor data be integrated in order to make space for dynamic and more responsive predictions. Under this approach, immediate insights and modifications are also allowed based on current pollution levels to be included in the prediction. Neural

networks, along with real-time sensors, will enable researchers to capture complex relations between different pollutants and their cumulative exposure effects on health outcomes, thereby harnessing deep learning technology to unveil the very noise patterns that other models failed to capture.