Data Analytics Project

Air Quality Analysis

# **Introduction**

Air pollution is a critical issue that has been a growing concern for public health and environmental sustainability. The impact of air pollution on human health and the environment has been well-documented, making it essential to understand the factors contributing to air quality and develop strategies to reduce air pollution. In this report, we aim to analyze an air quality dataset using various statistical and machine learning algorithms to compare and identify the most effective algorithm and improve air quality prediction. Air quality index (AQI) which is an important indicator of air quality as stated by Nigam et al [1] which is our focus in this study.

The air quality dataset provides information on several air quality parameters such as nitrogen oxides, carbon monoxide, sulfur dioxide, and particulate matter. Our analysis involves various steps such as data preprocessing, exploratory data analysis, feature selection, class resampling, and modeling using the R programming language. Previously techniques like kNN were implemented by Dragomir, E. G. [2] for predicting AQI values. Additional techniques of CatBoost Regression, Support Vector Regression, and Random Forest Regression were also implemented by Shivangi Nigam and Navneet Kumar, E. G. [3] for predicting AQI values.

We will explore the performance of various statistical and machine learning algorithms such as Support Vector Machine, Logistic Regression, Random Forest, and Ridge Regression in predicting air quality and identifying the key factors affecting it.

In this project our objective is to provide valuable insights into air quality analysis that can be leveraged to develop effective strategies to reduce air pollution and promote sustainable development. The findings of this study can be of great interest to policymakers, researchers, and stakeholders involved in public health and environmental sustainability. Overall, this report serves as a useful resource for understanding the impact of air pollution on public health and the environment and developing effective strategies to mitigate it.

# **Problem Description**

The specific problem addressed in this analysis is to identify the key factors affecting air quality and to develop a model that can accurately predict air quality levels based on these factors. The dataset provides information on various air quality parameters. By analyzing this data, we aim to identify the factors that have the greatest impact on air quality and develop a model that can accurately predict air quality levels based on these factors. This analysis can provide valuable insights into air quality management and help in developing effective strategies to reduce air pollution and promote a healthy environment.

# **Description of the Source Data**

The dataset utilized during this project came from a singular source collected from kaggle.com. It contains data sourced directly from the Central Pollution Control Board of India. The data contains information about a variety of particulate concentrations within the air of a given City, it also contains the date of collection.

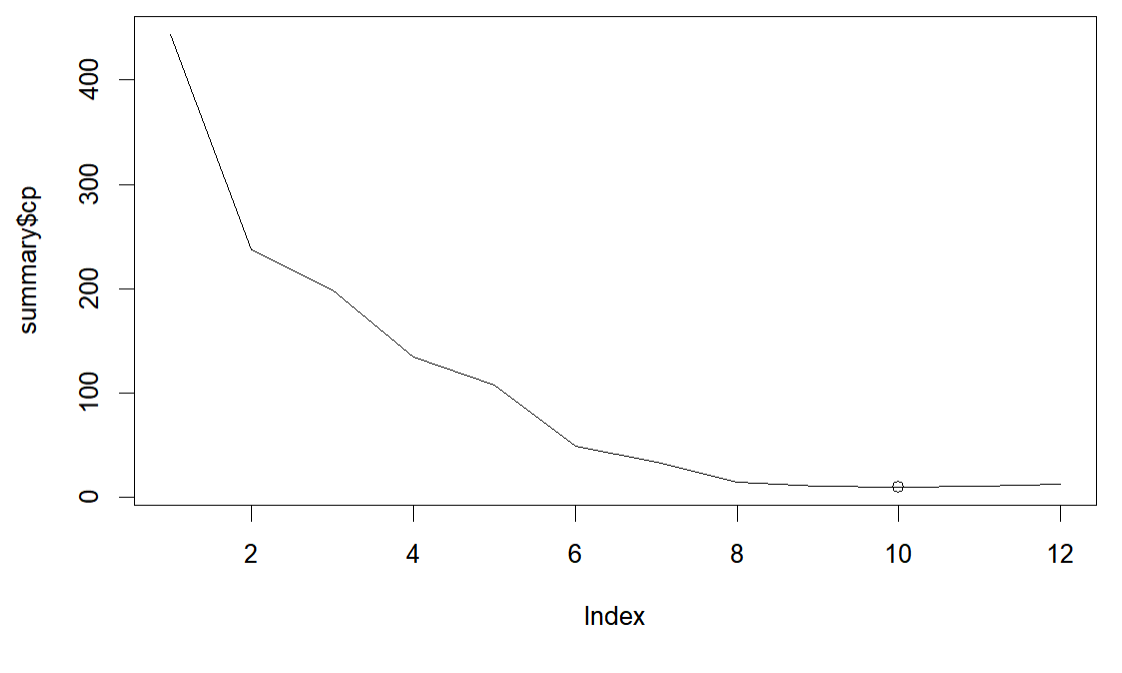
# **Data Preprocessing**

Before performing any feature selection or model creation. Data preprocessing had to be completed on the dataset. There were two main reasons for this: the first was that most of the records were unclean and had NULLs in some predictor fields. To resolve this, any given NULL value was replaced with the median value of the field. It was decided to take this approach due to the data being skewed, thus, replacing NULLs with a mean would make our data more susceptible to outliers. Also simply removing any NULLs with the na.omit function was not preferable due to over 75% being removed through this method.

A new field was also constructed named “AQI\_Bucket.” The field was built to convert the continuous prediction variable AQI into a categorical variable. Which allowed the use of classification models, in place of regression models. The reasoning for this is that a simple classification of the air quality for any given day, such as Very Poor, is much more informative to the average citizen than a more nuanced calculation of an Air Quality Index of 400.

# **Feature Selection**

Before testing any models, feature selection techniques were performed to find the most statistically significant variables when predicting the Bucketed AQI field. To do this the regsubsets function was used. This is a step wise function that recursively fits a linear regression model on all subsets of the predictor variables. Once the recursive function was completed, the model that minimized the CP value was selected. The CP value is a common value to determine the goodness-of-fit for any given model.



The outputted model when minimizing the CP value, kicked out 2 of the 12 variables. The two removed were “Benezene” and “PM2.5.” However, when plotting the graph of the CP-value to amount of included variables, we can see that the model that included all variables and the selected model are close in CP-value. Meaning that despite the selected model minimizing the CP-value, a model including all variables would still be a good fit.

Thus, it was decided to run two separate models for each method and compare the results. It is expected that the model that has includes all variables will be more accurate, since the models will have additional information while maintaining a good general fit for the data.

It was also tested to see if any Bucketed AQI type had a change in statistically important variables when considering the model. The reasoning for this is that there may have been certain cutoffs for particulate concentration that would then affect the AQI calculation more harshly. However, after subsetting the data and rerunning the regsubsets function on each, it was found that each model had the same statistically significant variables as when the function was run on the entire dataset.

# **Methods**

Logistic Regression Approach

The first method implemented to classify the AQI\_Bucket field was a Multinomial Logistic Regression approach. Multinomial logistic regression is a statistical technique used for classification problems with more than two categories. It models the probability of each outcome based on independent variables. It's useful for predicting outcomes and understanding the importance of independent variables. It's a widely used and powerful tool. The multinomial logistic regression model can be used to estimate the probabilities of each possible outcome of the dependent variable for a given set of values of the independent variables. These probabilities can then be used to make predictions about the most likely outcome.

The data preprocessing is done, and the multinomial logistic regression has been tuned to know the best decay value. The best tune decay is identified as 0.0, which indicates the amount of training data is very large relative to the number of predictor variables, and the model is already well-regularized due to the sheer amount of data, then adding additional regularization may not be necessary.

Experiment-1:

In experiment 1, the most relevant predictors are used to predict the AQI\_Bucket class. The overall accuracy and class-wise accuracy scores for this experiment are shown in the diagram below.

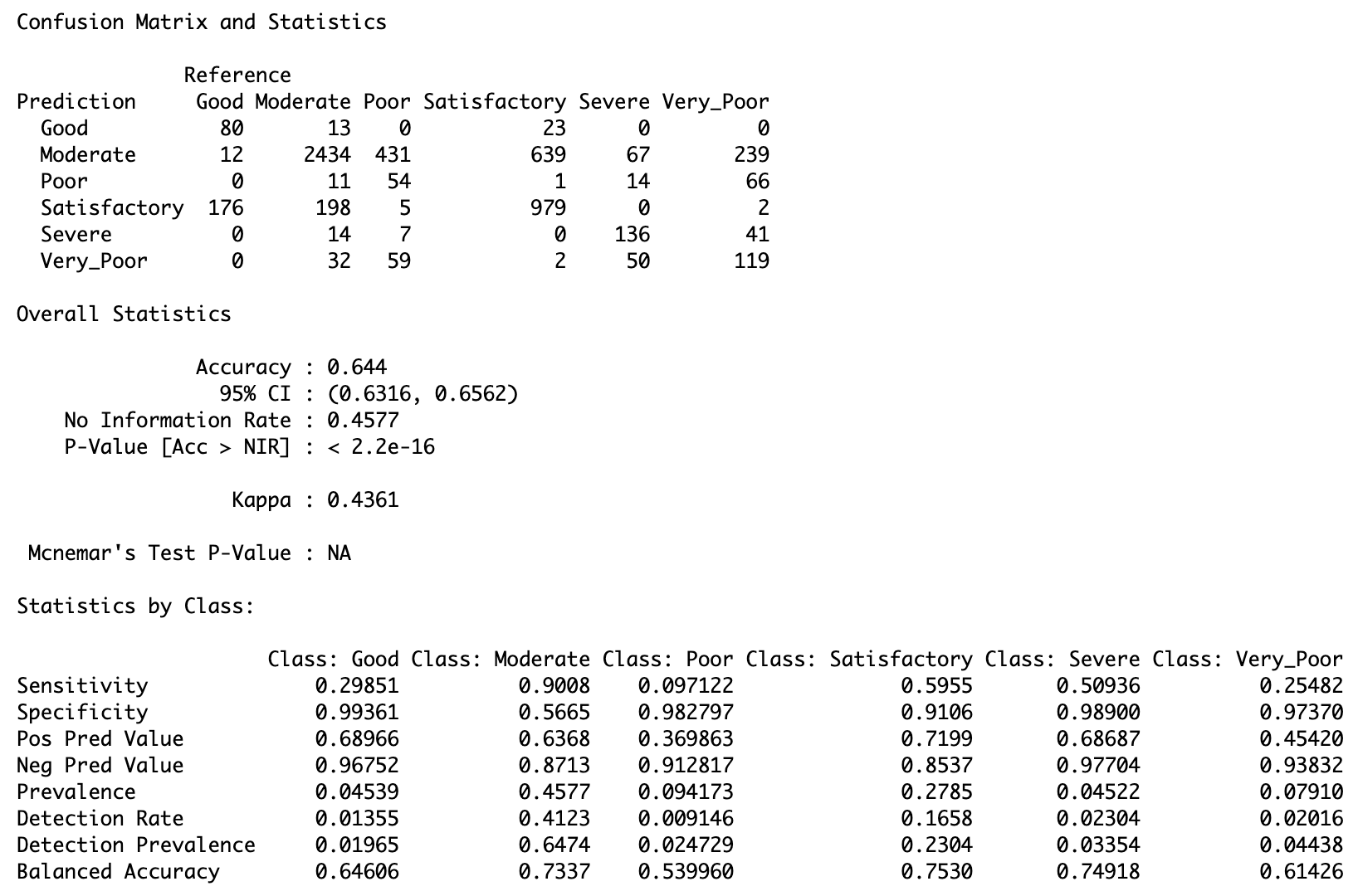


Fig: Multinomial Logistic Regression – Relevant Predictor Variables

Experiment-2:

In experiment-2, all predictors are used to predict the AQI\_Bucket class. The overall accuracy and class wise accuracy scores for this experiment are shown in the diagram below.

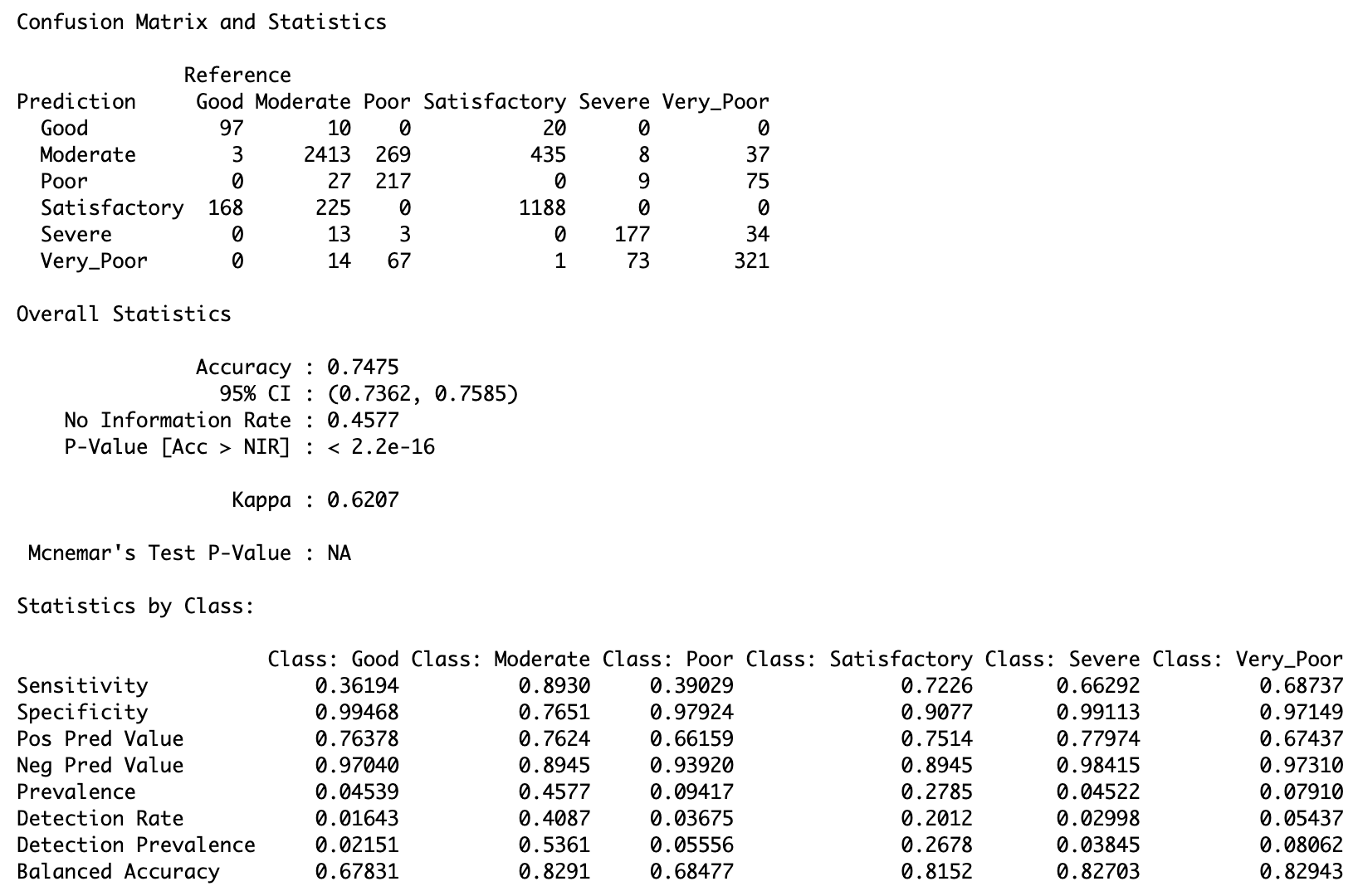


Fig: Multinomial Logistic Regression – All Predictor Variables

If we compare the results of both experiments the model with all predictor variables performed better than the model with only relevant predictor variables. This might be because the model is able to get some useful information from the other two variables which in turn helps the model to predict the AQI\_Bucket class accurately.

Ridge Regression

The second method implemented to predict the AQI\_Bucket field was a Ridge Regression approach. Ridge regression is used when the model has multiple predictors that are correlated with each other. It helps to prevent overfitting by penalizing the co-efficients of the variables. Ridge regression solves this problem by adding an extra parameter, alpha, which penalizes the coefficients for being too large. This makes it easier to find the best linear model and helps in predicting accurate results from the data.

The ridge regression has a tuning hyperparameter that can be used to make better predictions. In the below figures we can see that the model with all predictors has higher accuracy score than the model with the relevant predictors. In the below model, the model never predicts “Good” class as the predictor coefficients always recorded “Satisfactory” category.

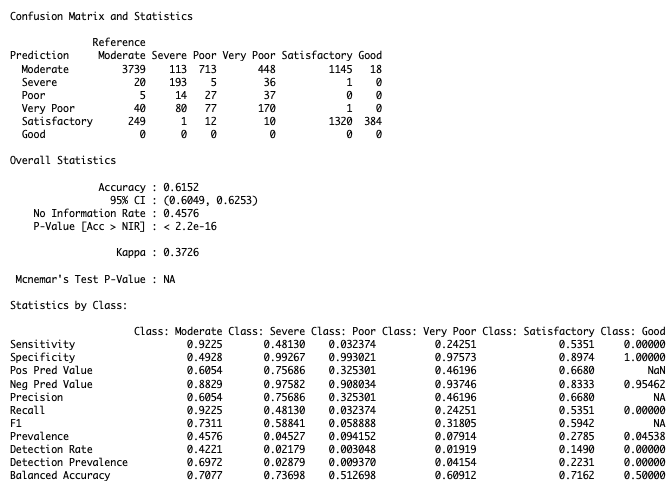


Fig: Ridge Regression Model- Relevant Predictors

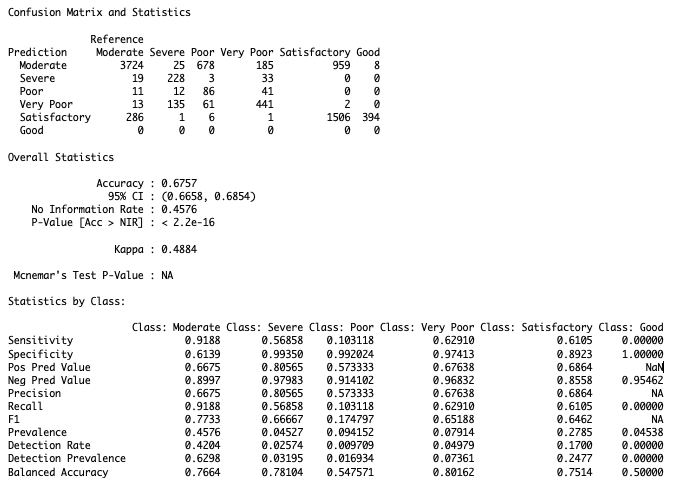


Fig: Ridge Regression Model- All Predictors

Random Forest Approach

The third method implemented to predict the AQI\_Bucket Field was a random forest approach. In a random forest, multiple decision trees are created from random subsets with replacement from the training data. Then based off majority vote from all created decision trees a prediction is made. This allows for a more accurate model than a singular simple decision tree.

The random forest algorithm also has many hyperparameters that can be tuned to allow for more accurate predictions. For this project two main hyperparameters were tuned. The first being the optimal ntree value. The ntree value represents the number of unique decision trees that are created. These trees then make a majority vote for the prediction. However, since all decision trees are based off training data, simply continuously increasing the number of trees will not always correlate to an increase in accuracy due to overfitting. Thus, the model was recursively run with a range of ntree values from 50 – 500 with 50 step intervals. Through this a best ntree value of 150 was determined. The same recursive method was also applied to the mtry hyperparameter with a range of 2 – 10 with 2 step intervals, which resulted in a value of 6. The mtry hyperparameter is the number of variables supplied to consider at each split in the created decision trees.

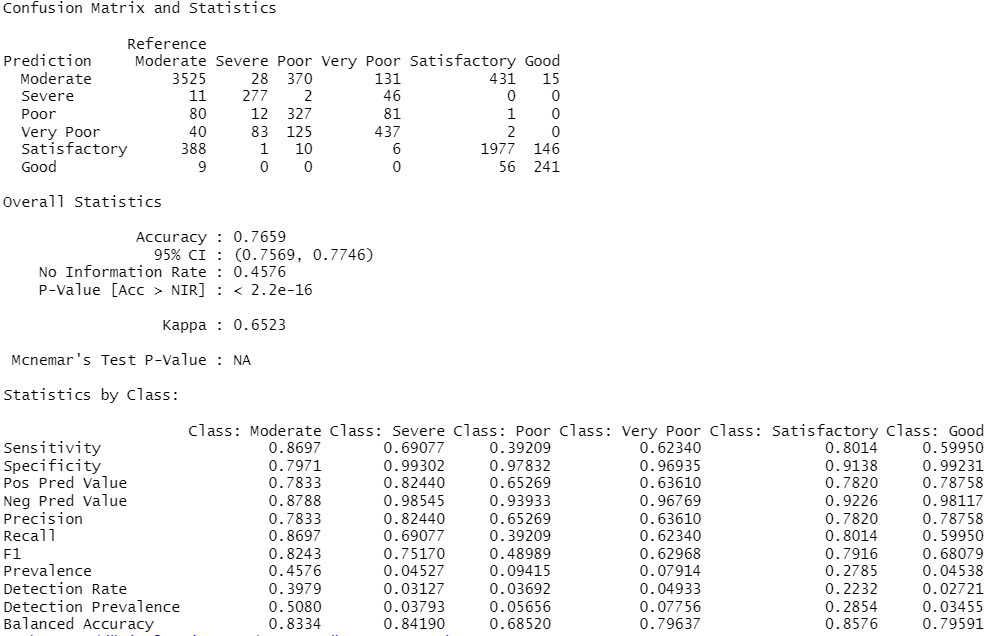


Fig: Random Forest Model- All predictor variables

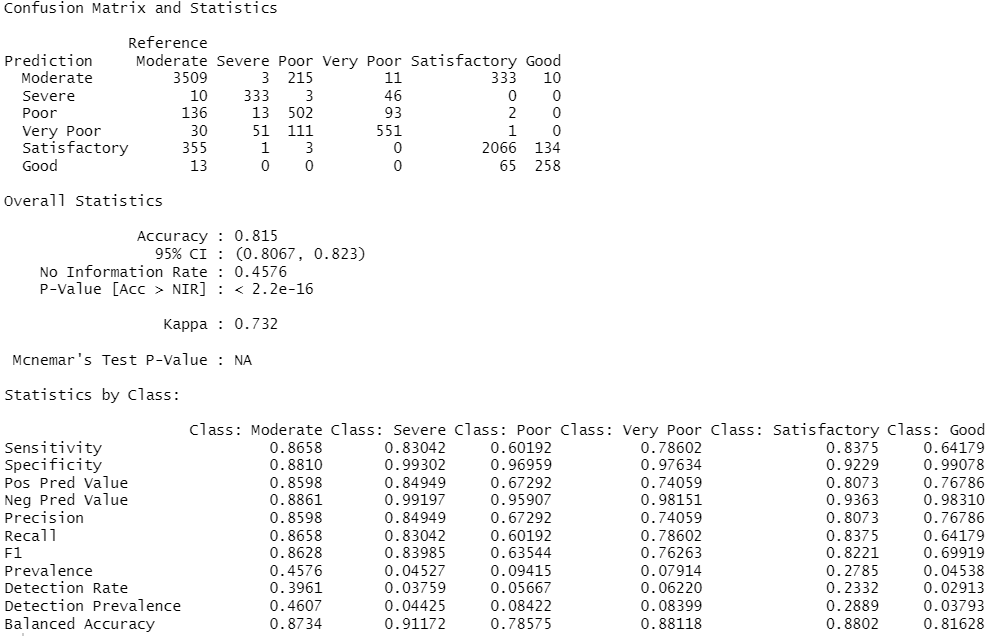


Fig: Random Forest Model - All predictor variables

Support Vector Machines Approach

The last method implemented to predict the AQI\_Bucket Field was a Support Vector Machine (SVM) approach. In this approach, the algorithm creates a hyperplane that separates the data into classes. The SVM approach is based on maximizing the margin between the hyperplane and the closest data points. This approach has been widely used for classification problems and can also be applied to regression problems like predicting AQI\_Bucket values.

The SVM algorithm has several hyperparameters that can be tuned to improve the prediction accuracy. For this project, the best combination of hyperparameters was determined by performing a grid search with a range of possible values for each hyperparameter. The two main hyperparameters that were tuned were the cost (C) parameter and the gamma parameter. The cost parameter controls the trade-off between the accuracy of the model and the complexity of the decision boundary, while the gamma parameter controls the smoothness of the decision boundary.

The grid search was performed using the caret package in R, and the best combination of hyperparameters was determined to be a cost value of 0.1 and a gamma value of 1. These values were used to train the SVM model on the training data, and the model was then used to make predictions on the test data. The performance of the SVM model was evaluated using various metrics, such as accuracy, precision, recall, and F1 score, which were calculated using the confusion matrix generated by comparing the predicted values to the actual values of the test data.

Experiment-1:

"e1071" and "caret" packages to implement the SVM algorithm and tune its hyperparameters. To tune the SVM hyperparameters, two functions named svm\_accuracy\_ntree() and svm\_accuracy\_mtry() were created. The svm\_accuracy\_ntree() function takes the gamma value and trains an SVM model using the radial basis function (RBF) kernel. It returns the accuracy of the prediction based on the confusion matrix. Similarly, the svm\_accuracy\_mtry() function takes the degree value, and trains the SVM model using the best gamma value and the degree value. It also returns the accuracy of the prediction based on the confusion matrix.

We ran the svm\_accuracy\_ntree() function on a set of gamma values ranging from 1 to 0.001 with 0.1 step increments, and recorded the accuracy of each prediction. Similarly, we ran the svm\_accuracy\_mtry() function on a set of degree values ranging from 0.1 to 100 with 1 step increments, and recorded the accuracy of each prediction.

We used the which.max() function to determine the gamma value and degree value that yielded the highest accuracy score. The best gamma value was found to be 1, and the best degree value was found to be 10.

Finally, we trained the SVM model using the best gamma and degree values, and predicted the AQI bucket for the test data. We used the confusionMatrix() function to create a confusion matrix that summarizes the model's performance. The confusion matrix showed that the SVM model predicted the AQI bucket with an overall accuracy of 78.36%. The model performed well for the Moderate and Satisfactory AQI buckets, but its performance for the Good, Poor, Very Poor, and Severe AQI buckets was poor.

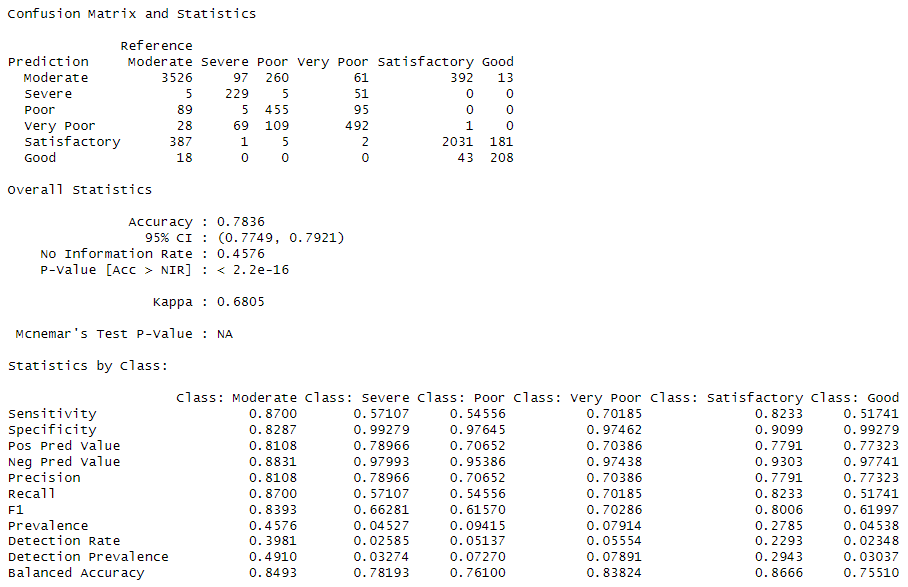


Fig: SVM Model- All predictor variables

Experiment-2:

In this particular analysis, the same dataset as in the previous experiment was utilized. However, this time, only the relevant predictor variables identified by feature selection techniques were considered, and 'PM2.5' and 'Benzen' variables were excluded from the model. The resulting model exhibited an accuracy of 72.38%, which was lower compared to experiment-1.

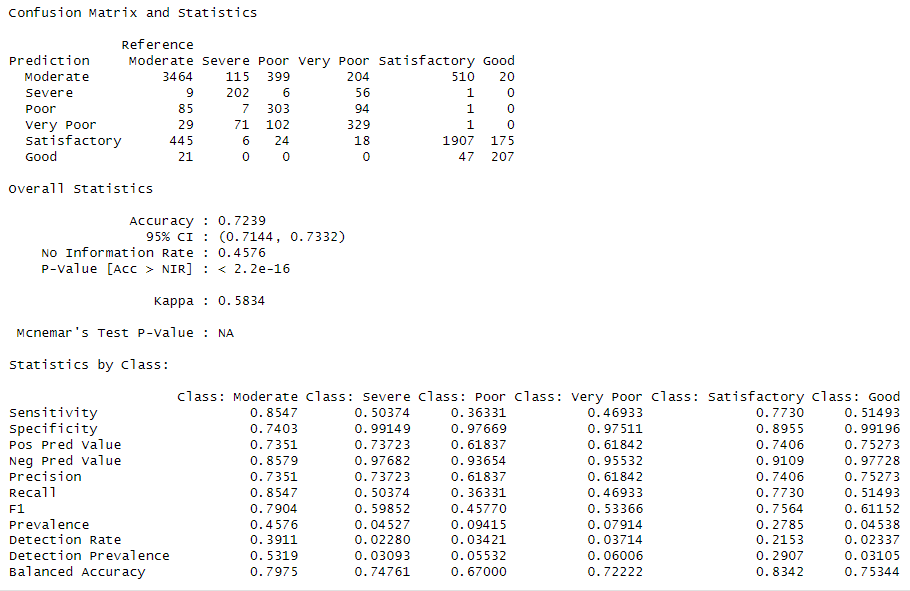


Fig: SVM Model- Relevant predictor variables

Experiment-3:

In this experiment, we performed k-fold cross-validation to train and evaluate the support vector machine (SVM) model on the same dataset. We used the caret package in R to implement k-fold cross-validation with a 5-fold split. The training data was used to train the SVM model with a radial kernel, and the testing data was used to evaluate the model's accuracy. The SVM model was trained with the following predictor variables: PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Toluene, and Xylene. We also used the best gamma and degree values that were found in a previous tuning step. The SVM model with a radial kernel achieved a mean accuracy of 72.05% in predicting AQI categories based on pollutant measurements. The k-fold cross-validation technique helped to evaluate the model's performance on multiple subsets of the data, and the results showed that the model was relatively stable.

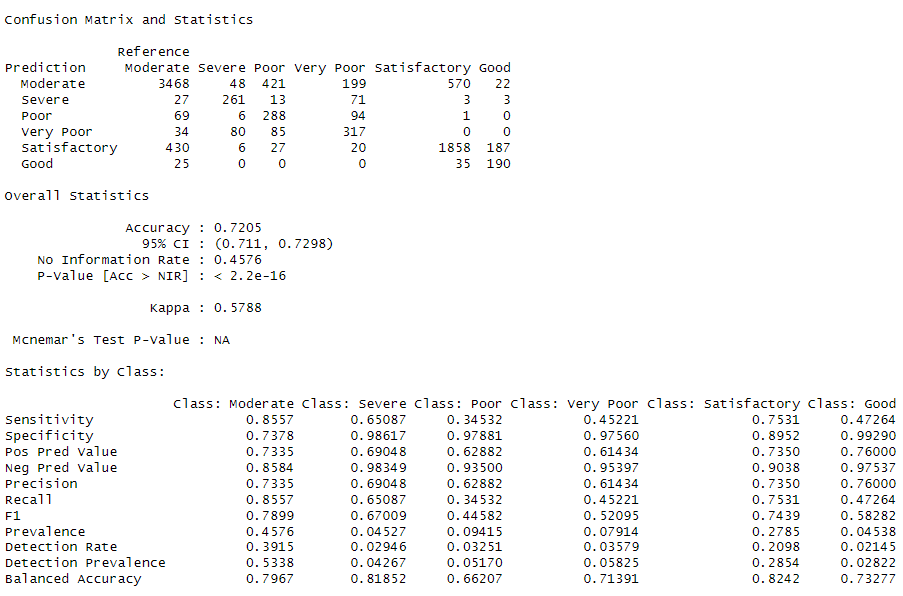


Fig: Kfold SVM Model- Relevant predictor variables

# **Results**

Models Most Statistically Relevant Predictor Variables

|  |  |  |
| --- | --- | --- |
| **Model Type** | **F1 Score [Good, Satisfactory, Moderate, Poor, Very Poor, Severe]** | **Accuracy** |
| Multinomial Logistic Regression | [0.64606, 0.7530, 0.7337, 0.53996, 0.61426, 0.74918] | 0.644 |
| Ridge Regression | [NA, 0.5942, 0.7311,0.0588,0.31805,0.58841] | 0.6152 |
| Random Forest | [0.75170, 0.68079, 0.7916, 0.8243, 0.48989, 0.62968] | 0.7659 |
| SVM | [0.7564, 0.7904, 0.45770, 0.53366, 0.59852, 0.61152] | 0.7239 |

Models With All Predictor Variables

|  |  |  |
| --- | --- | --- |
| **Model Type** | **F1 Score [Good, Satisfactory, Moderate, Poor, Very Poor, Severe]** | **Accuracy** |
| Multinomial Logistic Regression | [0.67831, 0.8152, 0.8291, 0.68477, 0.82943, 0.82703] | 0.7475 |
| Ridge Regression | [NA, 0.6462, 0.7733, 0.1747, 0.65188, 0.66667] | 0.6757 |
| Random Forest | [0.69919, 0.8628, 0.8221, 0.63544, 0.76263, 0.83985] | 0.815 |
| SVM | [0.61997, 0.8006, 0.8393, 0.61570, 0.70286, 0.66281] | 0.7836 |

When viewing the results, every model had worse accuracy when using only the statistically relevant predictors gathered from the regsubsets feature selection. This was to be expected, as mentioned previously, both the model that included all factors and the model that kicked out two predictors had similar low CP values. Meaning they both would be a good fit for the data; thus, the more complex model should result in higher accuracy. And, as prediction is the main goal and inference is not a priority, the complexity of the models is not of concern.  
  
The SVM and Random Forest were our best performing models for the accuracy metric. This is due to the data not resembling a traditional linear function that the Ridge Regression or Logistic Regression assumes. Thus, allowing a more accurate fit to the true nature of the dataset. However, it should be mentioned that despite the Multinomial Logistic Regression Model having a significantly smaller accuracy value, it does boast high accuracy for worse air quality conditions. Outperforming both the Random Forest and SVM when predicting Poor and Very Poor, while also having the second-best accuracy for Severe. Considering this it might be the model that best suits the purpose of the project, as warning citizens of potentially dangerous air quality conditions is of greater benefit than reporting satisfactory days.

# **Discussion**

The model performance achieved in this analysis was modest, as there were no models that would be accurate enough to allow for trustworthy reporting of air quality to citizens. However, it was demonstrated that an approach to training prediction models is feasible and could be improved to a state where accuracy would be high enough to confidently report on.

Additional future work could be done to improve the accuracy of the models further. One simple approach would be collecting additional data, especially for days when air quality was worse. As the source data was uneven, with most of the records being Satisfactory or Moderate. Class resampling could also be applied to handle this imbalance as well. There were also outliers within the dataset for days that had extreme amounts of particulates. Removing these outliers, thus reducing the skew of the model’s results, could also improve the model's accuracy as well.

Additionally, the approach to handle null values could be revised. In this project the Naïve surrogate's approach was implemented. This approach limited the model’s ability to capture complex relationships and artificially increased the correlation between variables. A more nuanced approach could be implemented, such as the Bayesian Method. The Bayesian method predicts the missing values by observing the distribution of the currently observable data and drawing conclusions from it. This would allow the replacement data to be much closer to what it was originally.

# References

1. Nigam, S., & Kumar, N. (2016, May). *Air Quality index – a comparative study for* a*ssessing the status of air quality* Researchgate. Retrieved May 6, 2023, from <https://www.researchgate.net/profile/Navneet-Kumar-18/publication/292643438_Air_Quality_Index_-_A_Comparative_Study_for_Assessing_the_Status_of_Air_Quality/links/56b092c708ae9c1968b774c8/Air-Quality-Index-A-Comparative-Study-for-Assessing-the-Status-of-Air-Quality.pdf>
2. Dragomir, E. G. (n.d.). *Air Quality index prediction using K-nearest neighbor technique - unde.ro*. Retrieved May 6, 2023, from <http://bmif.unde.ro/docs/20101/pdf_final_12%20EDragomir.pdf>
3. Gupta, N. S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Arulkumaran, G. (2023, January 30). *Prediction of air quality index using Machine Learning Techniques: A Comparative Analysis*. Journal of Environmental and Public Health. Retrieved May 6, 2023, from <https://doi.org/10.1155/2023/4916267>
4. Tripathi, T. (2022, May 25). Final notebook data analytics internship. Retrieved May 6, 2023, from <https://www.kaggle.com/code/tishyatripathi/final-notebook-data-analytics-internship/input>
5. Holllla. (2023, March 14). *Air Quality index prediction using MLR*. Kaggle. Retrieved May 6, 2023, from <https://www.kaggle.com/code/holllla/air-quality-index-prediction-using-mlr/notebook>
6. Jean-Yves Sgro, P. D., & Kristen Malecki, P. D. (2022, July 7). *Chapter 5 working with tabular data in R: Tabular Data Analysis with R and Tidyverse: Environmental Health*. Retrieved May 6, 2023, from <https://static-bcrf.biochem.wisc.edu/courses/Tabular-data-analysis-with-R-and-Tidyverse/book/5-workingwithtabulardata.html>
7. Yeon, J., & Wilber, J. (n.d.). Random Forest. Retrieved May 6, 2023, from <https://mlu-explain.github.io/random-forest/>

# Appendices

**Appendix A: Data Features and Descriptions**

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Range** | **Description** |
| City |  | City |
| Date | 12/31/2014 - 06/30/2020 | Date |
| PM2.5 | 0 - 950 | Particulate Matter 2.5-micrometer in ug / m3 |
| PM10 | 0 - 100 | Particulate Matter 10-micrometer in ug / m3 |
| NO | 0 - 391 | Nitric Oxide in ug / m3 |
| NO2 | 0 - 362 | Nitric Dioxide in ug / m3 |
| NOx | 0 - 468 | Any Nitric x-oxide in ppb |
| NH3 | 0 - 100 | Ammonia in ug / m3 |
| CO | 0 - 176 | Carbon Monoxide in mg / m3 |
| SO2 | 0 - 194 | Sulphur Dioxide in ug / m3 |
| O3 | 0 - 258 | Ozone in ug / m3 |
| Benzene | 0 - 455 | Benzene in ug / m3 |
| Toluene | 0 - 455 | Toluene in ug / m3 |
| Xylene | 0 - 170 | Xylene in ug / m3 |
| AQI | 13 - 2.05k | Air Quality Index |