Classification Clear as Air

Emmanuel Pedernal*

Masters of Science in Data Science, College of Computer Studies, De La Salle University, 2401 Taft Ave., Manila Philippines

Corresponding author: emmanuel_pedernal@dlsu.edu.ph

ABSTRACT

Global estimate suggests that nearly all of the population breathes polluted air that is beyond the recommended levels. Worryingly, outdoor pollution is responsible for annual millions of premature deaths, while pollution from household cooking with fuels like wood or coal attribute to increasing damages to respiratory system. This study aims to create a predictive model that could classify air quality levels through "Good", "Moderate", "Hazardous" based on climate variables; temperature(°C), humidity(%), environmental pollutants such as fine_matter(μg/m³) nitrogen_dioxide(ppb) sulfur_dioxide(ppb) carbon_monoxide(ppm), urban density from proximity_to_industrial_areas(km) and population_density(people/km²). The data obtained from 2022 World Health Organization on air quality dataset and World Bank data on population density consisting of 5,000 samples were fed in a Multinomial Logistic Regression (MLR). The results showed the model correctly classifies the features 97% of the time with F1-score of 0.96 on test set. Carbon monoxide was identified as the most influential feature across all classes, additionally the odds of predicting "Good" air quality increased by 1423% (OR= 15.23) for every kilometer of distance from industrial area, while an increase in Sulfur oxide (OR=0.1362) and Carbon monoxide (OR=0.0019) reduces odds by 86.38% and 98.1% sequentially. To classify as "Moderate" the features that increases the odds are carbon monoxide (OR = 3.2811), sulfur dioxide (OR = 1.5340), and temperature (OR = 1.3041) by 228%, 53% and 30% respectively. For "Hazardous" air quality, carbon monoxide (OR = 156.2634) showed 14,626% increase in odds, makes it the most influential feature. Nitrogen dioxide (OR = 5.7920) increases the odds by 479%, and temperature (OR = 5.4548) contributes to a 445% increase in hazardous air quality. The results reveal the role of pollutants and urban proximity in air quality classification. In future research, more flexible models, such as Generalized Additive Models (GAMs) or Tree-based methods, should be considered along with polynomial transformation, extensive hyperparameters tuning and an expanded dataset to better generalize and address assumption violations.

Keywords: *Multinomial Logistic Regression, Environmental Pollutants, Odds Ratio*

INTRODUCTION

Air pollution is a major global health and environmental issue, driven by a variety of factors ranging from industrial emissions to household practices, contributing to respiratory diseases, cardiovascular conditions, and even premature death.

The World Health Organization (WHO) estimates that nearly 99% of the population breathes air that exceeds their recommended pollution levels, with most of its effects are felt in Southeast Asia and Western Pacific (*World Health Organization, 2024*).

Air pollution defined by *(Almetwally et al. 2020)* as presence of natural or manmade substance that has high concentrations to cause damage to human health, vegetation, property or environment. From World Health Organization (1999), the authors pointed out that air pollution is the byproduct of human activities that releases pollutants in the air through gas such as nitrogen oxides, sulfur dioxide or as particulate matter PM2.5, and PM10.

Air pollution is not graded through presences of pollutants but how high their concentration and interaction with the environment.

In 2021, UN Human Rights Council declared that a pollution-free environment is a fundamental human right, and reduced environmental quality is considered a threat to public health. Air pollution is responsible for millions of preventable deaths annually and has far-reaching economic and health impacts, particularly in vulnerable regions where damages caused billions, with young and elderly age groups are most affected. *(Khyber Medical University Journal, 2021)*

This study aims to predict air quality using pollutant concentrations and urban density as predictors through multinomial logistic regression, incorporating hyperparameter tuning and bootstrapping. In addition, the researcher examines how each feature affects the classification of air quality levels by interpreting the corresponding odds ratios.

Data Collection

Data was scrape from World health organization *(World Health Organization. 2024)* and World Bank Data *(World Bank. 2024)* with a total count of 5000 samples from ambient air quality database comprise of annual mean concentrations of the following; nitrogen dioxide (NO_2) a gas produced by household, factories and vehicles that can irritate lungs and add on respiratory conditions that is linked with asthma. Carbon monoxide (CO) odorless gas by product of incomplete combustion with vehicles being major source, that could affect the oxygen transportation in the body causing dizziness and fatigue (Bleecker, M. L. 2015). Sulfur dioxide (SO_2) a colorless gas that is easily dissolve by water, by product of household heating, and factory power consumption attributed to asthma. Particulate matter of a diameter equal or smaller than 10 µm (PM_{10}) that is common produce from construction and vehicle emissions while

2.5 µm ($PM_{2.5}$) a type of particulate matter that can easily penetrate deep inside lungs and bloodstream due to size. The data aims at representing an average for the city or town as a whole, rather than for individual stations. Pollutants mainly came from human activities related to the burning of fossil fuels *(Mujtaba Mateen. 2024)*. The data also utilize urban measurements through urban, residential, commercial and industrial areas close to communities. The data then merged with the population density estimate to form the dataset where the data is already cleaned and ready for modelling, the class separation is based on WHO's guideline on the limit of intake; minimum for "Good", average of the gas intake as "Moderate", max for "Poor" and over the limit group "Hazardous" enumerated in Table 1.

Table. 1 Features and Description

Feature	**Type**	**Non-Null**	**Description**
temperature(°C)	float64	5000	Average outdoor air temperature at measurement time, influences dispersion and chemical reactions of pollutants.
humidity(%)	float64	5000	Moisture percentage in air; affects particle formation and pollutant behavior.
fine_matter(µg/m^3)	float64	5000	PM2.5 concentration; fine particles that penetrate deep into lungs, linked to respiratory health risks limited to .5 – 15 µg/m^3 per 24 hrs.
coarse_matter(µg/m^3)	float64	5000	PM10 concentration; larger particles from dust and construction sources limited to 15 -45 µg/m^3 per 24 hours
nitrogen_dioxide(ppb)	float64	5000	Emitted by vehicles and industry; precursor to ozone and fine particulates limited to 10 – 25 µg/m^3 per 24 hours.
sulfur_dioxide(ppb)	float64	5000	Emitted by burning fossil fuels; contributes to acid rain and respiratory issues limited to 40 µg/m^3 per 24 hours.

Feature	Type	Non-Null	Description
carbon_monoxide(ppm)	float64	5000	Toxic gas from incomplete combustion; harmful at high concentrations limited to 4 mg/m^3 per 24 hours.
proximity_to_industrial_areas(km)	float64	5000	Distance to nearest industrial area; proxy for localized emissions exposure.
population_density(people/km²)	int64	5000	Residents per km²; higher density may indicate greater urban emissions.
air_quality	object	5000	Categorical air quality label (Good, Moderate, Poor, and Hazardous).

METHODS

Data Transformation

Upon checking of class distribution, the researcher found class imbalance of Class "Good" 2000 samples to "Hazardous" of 500 samples. Since Class "Poor" and "Hazardous" are both on the labels for dangerous levels of gas present in the air pollution the researcher combined these classes in an attempt to level the class distribution enumerated in Appendix A.

Table 2. Classification and Corresponding Values

Classification	Value (n)
Good	2000
Moderate	1500
Hazardous	1500

Statistical Analysis

All analysis was conducted using JASP (0.19.3.0) and Python 3.12 with the following packages; for data analysis, mathematical computation and visualization, pandas (v2.2.3), numpy (v1.26.4), matplotlib (v3.9.2), seaborn (v0.13.2), and plotly (v6.0.0). Statistical modeling was conducted using statsmodels (v0.14.4), and machine learning workflows were implemented with scikit-learn (v1.5.1). The scipy package (v1.14.1) was used for additional statistical and numerical operation.

Generated descriptive statistics for the air quality data with Count, Mean, Standard deviation (std), min, 25%, 50%, 75% and Max Values.

Variable Relationships

As part of statistical analysis, both univariate and bivariate analysis were conducted.

Univariate Analysis

Univariate analysis examines each variable individually, with focus on the variables' descriptive, summarized and understanding of data distribution *(Cleff, T. 2025).* for this paper we'll focus on the frequency counts per feature that could provide the researcher of insights in general structure of the dataset.

Bivariate Analysis

Bivariate analysis focuses on the possible relationships between variables and or target class, whether they are associated, correlated or dependent on each other *(Köhler, T. 2023).* With numerical features, the researcher tested it using pairwise scatter plot with distinction per target class. Results are used to aid the researcher for feature analysis and bring about additional insights from the dataset.

Model

A multiple logistic regression model (MLR) was used to identify significant features that determine the air quality. The researcher fit the data with Elastic net regularization with GridSearch on 5-fold cross-validation. The model was evaluated using scikit-learn's comprehensive report where it focuses on the model's accuracy, precision, sensitivity, and f1 score.

Assumptions for Model Diagnostic

Prior fitting, the multinomial logistic regression key assumptions were assessed *(Hua, Y. et. al, 2025)* for multiple categorical outcomes, independence of observations, sufficient large sample size validity, absence of multicollinearity, linearity of logits, no extreme outliers and Independence of irrelevant alternatives (IIA). First, to test multiple outcomes, dependent variable must be categorial and classes are greater than 2 (> 2) that satisfies the multinomial response. The dataset should not have duplicate rows to verify that each observation is independent from each other. Multicollinearity among features was tested using Variance Inflation Factor (VIF). Features with VIF greater than 10 (>10) are flagged for dropping or feature transformation with another variable due to having high linear relationship will result to distorted coefficient estimation. Linearity of logits was tested by checking continuous features' logits for linear relationship through scatter plot. Lastly, Independence of Irrelevant Alternatives (IIA), assumes that the relative odds between any two outcome categories are unaffected by the presence or absence of other alternatives. The diagnostic tests mentioned are done before the model is fitted.

Preprocessing

Separated the dataset into feature matrix (X) and target (y). The target dataset is then transformed using LabelEncoder to change Classes {0:'Good',1:'Moderate',2:'Hazardous'}. This is done to make the labels usable with the model. Next, sklearn's train_test_split was used to create subsets for 70% of the data for training, 20% for validation and 10% for test set, the research also applied stratified sampling to ensure that all subsets retained the original class distribution to avoid sampling bias as shown in Appendix A. Finally, standardized the numerical features for them to have mean of 0 and standard deviation of 1, this is done to mitigates the impact of differing feature scales and improves the numerical stability of gradient-based algorithms.

Model Training

Regularization techniques are widely used to reduce overfitting, and improve generalizability, by shrinking coefficient estimates (Jecinta & Obi, 2023). The model utilizes elastic net regularization which combines both L1(lasso) and L2(ridge) penalties that lets the model do simultaneous feature selection and model generalization. The model also got the parameter class_weight='balanced' ensures that class(es) are not underrepresented during training. An optimization algorithm through 'saga' solver is applied, an extension of Stochastic average Gradient descent solver, that minimizes the full multinomial loss for multiclass problems *(Arefin & Asadujjaman, 2016)*.

To optimize the model's performance a grid search was done with the following hyper parameters:

C: inverse regularization strength that controls model complexity, smaller C value increases regularization, encouraging the model to find a simpler solution by penalizing large coefficients, while a larger C allows for a more complex model by reducing the regularization strength.

tol: tolerance for solver convergence, the model training stops when the improvement in the objective function between iterations is smaller than tolerance.

l1_ratio: ratio of regularization between L1 and L2.

max_iter: maximum iterations for solver convergence.

The grid search resulted in the following optimal parameters enumerated in Appendix B.

Evaluation

The model's performance during hyperparameter tuning was assessed using balance accuracy score, which considers the imbalanced in class distribution by getting the average recall across all classes. This is done to avoid preference over major class by the model.

GridSearch of 5-fold cross-validation partitions the training data into 5 stratified folds. For every iteration the model was trained for 4 folds and

validated on the remaining fold. The purpose of the search is to get the more generalized performance estimate and avoid the risk of overfit to any single fold. The model is then input in bootstrap for 1000 permutations to better estimate statistic interest (Standard deviation, z, p-value, Confidence interval).

In addition, we can extract the coefficient for each feature then exponentiate to get the odds ratio. Odds ratio are odds of an event happening in one group compared to odds of occurring in another group. An OR >1 indicates an increase in being classified into specific class, example an OR of 1.71 suggest that the odds increase by 71% for each one-unit increase in the covariate. On the other hand, an OR < 1 means reduction in being classified, example an OR of 0.58 means for each increase in the covariate reduces the odds to be classified by 42% *(Hancock, M., & Kent, P. 2016)*.

Variable analysis, assumption tests, model evaluations and statistic interest are gathered and results are interpreted.

RESULTS AND DISCUSSION

Table 3. Descriptive Statistics

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
temperature(°C)	5000	30.03	6.72	13.4	25.1	29	34	58.6
humidity (%)	5000	70.06	15.86	36	58.3	69.8	80.3	128.1
fine_matter (µg/m³)	5000	20.14	24.55	0	4.6	12	26.1	295
coarse matter (µg/m³)	5000.	30.22	27.35	-.20	12.3	21.7	38.	315.8
nitrogen_dioxide (ppb)	5000	26.41	8.9	7.40	20.1	25.3	31.9	64.9
sulfur_dioxide (ppb)	5000	10.01	6.75	-6.2	5.1	8	13.73	44.9
carbon_monoxide (ppm)	5000	1.5	0.55	0.65	1.03	1.41	1.84	3.72
proximity_to_industrial_areas	5000	8.43	3.61	2.5	5.4	7.9	11.1	25.8
population_density	5000	497	152.8	188	381	494	600	957

Descriptive Statistics

Descriptive statistics of the air quality dataset are shown in *Table 3*. There is an average of 30 °C with standard deviation of 6.72 with minimum value of 13.4°C with maximum value of 58.6°C, which coincides with our findings with univariate analysis that data was taken from a warm location, IQR spans from 25.1°C to 34°C which indicates that middle values fall within warm temperatures, with max temperature exceeding 50 °C that could result to heath related illnesses such as heatstroke or heat exhaustion on top of the risk in hazardous air quality.

Based from WHO guidelines features used are in excess of safety levels. *(World Health Organization, 2021)*. The mean humidity level is beyond the normal levels by (30-60%) this could mean that most of the datapoints are near an industrial zone or population per km² are overcrowded. Having max value of 128 could mean data entry errors or possible extreme outliers that should be cleaned.

Fine matter 75th percentile of 26.1 exceeds WHO guideline threshold and a maximum value of 295 µg/m³ point to possible extreme pollution or extreme outlier. Sulfur Dioxide having negative values are not plausible and requires pruning. Data distribution from 25th to 75th percentile indicates normal within the guidelines.

Proximity to industrial Areas (km) with an IQR of 5.4km to 11.1km suggests that most data points location are close to industrial zones.

Population density mean of 497 people/km² and standard deviation of 152.8, the regions from the dataset ranges from moderately populated to densely populated zones (up to 957 people/km²). The IQR (381 to 600) suggests that half of the data are clustered in urban to semi-urban population densities.

Univariate Results

The researcher found that the air quality data suggests that the location where the it was gathered came from warm regions as indicated by temperatures ranging from 25 to 30°C, while humidity levels are in the 70 to 80% implies a location where moistures is high that could bring respiratory stress or lung discomfort. Particulate matter is within the range of 0-12 µg/m³ which falls within the "Good" category from WHO guidelines that reflects the researcher's class distribution analysis that class has the greatest number of counts. Nitrogen dioxide concentrations, observed between 10 and 40 ppb, are typical of urban environments and may not be Hazardous but still risk for respiratory illnesses. Sulfur dioxide levels below 20 ppb under minimum guidelines that suggests may only impact individuals who are respiratory sensitive. Carbon monoxide concentrations appear to be within background levels commonly seen in non-industrial settings and are not expected to cause discomfort. Spatial Data suggest that most of the dataset are near industrial zones while the average population is ~500 people/km². (Appendix C)

Bivariate Results

An initial analysis of pair plots shows correlations among features shown in Appendices D and E, such as fine_matter and coarse_matter resulted with high positive correlation (r =0.973), the results coincides with *(Chen et al. 2018)* where fine and coarse matter $PM_{2.5}$ to PM_{10} respectively, was significantly associated with increases in cause-specific mortality. Proximity_to_industrial_areas showed strong negative correlation with carbon_monoxide (r= -0.7) that aligns with the expectation that the near an individual to industrial zone the more dangerous the air is. Population_density showed moderate correlation with carbon_monoxide (r=0.59) suggests that highly populated areas see an increased in CO levels possibly due to use of household heating, vehicles and human related combustion sources. Results are shown on Appendix

Assumption Results

Number of Classes

Checking for class distribution as shown in Table 2, the dataset has 3 (three) kinds of classes namely, “Good”, “Moderate”, and” Hazardous”.

Independence of Observations

The researcher found that all observations are independent, no rows are duplicated as shown in Appendix F.

Sufficient Sample Size

As shown in Appendix G, the dataset should have 10 to 20 observations per feature multiplied by number of classes (3). The number of observations should be at least 270 rows. The air quality dataset consist of 5000 observations satisfies this assumption.

Variance Inflation Factor

Upon evaluating multicollinearity using Variance Inflation Factor (VIF), resulted in (2) features above the threshold of >10 VIF, namely fine_matter and coarse_matter with 29 and 34 score respectively shown in Appendix H. The researcher decided to drop the coarse_matter feature due having high VIF means a strong linear relationship that can disrupt the coefficient estimation of the model.

Linearity of Logits

The plots presented in Appendix I, shows the features maintained linear relationship with the logits of the classes. Features such as temperature, carbon monoxide and proximity to industrial area, only fine_matter was not perfectly aligned with, which is not substantial to do feature transformation. Thus, the assumption of linearity was considerable met.

Outliers

As illustrated in Appendix J, outliers exist for all features but most pronounced with fine_matter, carbon_monoxide and sulfur_dioxdie. These outliers could disrupt the model estimates and violates the assumption for our MLR model. To mitigate, the researcher filtered the data with IQR that resulted in 577 observations removed. Post filter shown no extreme data points outside the whiskers of box plot.

Independence of Irrelevant Alternatives (IIA)

Conducted the Pseudo Hausman-McFadden test, which assess the relative probability of choosing one alternative over another is unaffected by the presence or absence of other irrelevant alternatives in the categories. This was done by comparing the model’s weight changed through dropping one of the classes in comparison with a complete model. The test shown in Appendix K resulted in p-value < 0.0001 implies that we reject the null hypothesis that the model is not affected by new alternatives.

The air quality dataset was tested with reference to *Hua, Y. et. al. 2025*, with added assumption check of Independence of Irrelevant Alternatives. Most assumption checks were satisfied with the exception of the (IIA) assumption.

Model Classification Report

Appendix L represents the multinomial logistic regression results based on model

accuracy, precision, recall and f1-score which are standard performance metrics for classification tasks *(Sokolova & Lapalme 2009)*.

The" Good" class achieved precision of 1, recall of 0.99 and F1-score of 1, this result suggests strong effectiveness in predicting instances of "Good" air quality. On the other hand, "Moderate" and "Hazardous" class showed slightly lower results with Moderate class's precision of 0.92, recall of 0.94 and F1-score of 0.93, while Hazardous class results show precision at 0.95, recall of 0.94 and F1-score of 0.95, suggests potential overlap with both classes.

The model has an overall accuracy score of 0.97, which demonstrates the predictive performance of the model and an outstanding outcome for precision and recall across all categories. These results affirm the model's strength in predicting air quality through environmental pollutants and urban density related features.

Confusion Matrix

The confusion matrix in Appendix M display's the model's ability to predict among the 3 classes of air quality, while showing detailed insights to performance of the model from true positive, false positive, true negative and false negatives that allows the researcher to see trade-offs on different types of classification errors, useful in dealing with class imbalance or multiple classes *(Sokolova & Lapalme 2009)*.

First, the "Good" class (0) shows remarkable accuracy of 197/198 or TPR (True positive rate) of 99.4%, correctly classifying the test set with an FN (False negative) misclassification of "Hazardous" air quality with 0 FP (False positive) from other classes and TN (True Negative) of 245. Having a recall of 0.99 means that the model is almost certain that features are from "Good" class are actually from that class.

Next is "Moderate" class (1) correctly predicted 97/103 or 94% TPR of the test set with few FN misclassifications of 6 "Hazardous" air quality with FP and TN of 8 and 332 respectively. Despite the great performance results having 0.92 precision and 0.94 recall suggests that overlapping of patterns between the classes.

Lastly, the "Hazardous" class (2) resulted in 134/142 or TPR 94% correct classification of observations where 8 FN from the test set are misclassified as "Moderate" air quality with FP and TN of 8 and 294 respectively.

High scores of macro average F1-score and model accuracy 0.96 and 97% indicates that the model can be used to predict air quality specially with "Good" class, the overlap between "Moderate" and "Hazardous" class suggests possible changes to data such as feature engineer to better capture relationship or change in model parameters.

ROC and AUC

Appendix N shows an outstanding result for classification across the 3 classes. The multiclass ROC curve computed using One-vs-Rest strategy demonstrates strong model performance; the "Good" Class has a value of AUC 1 that suggests that the model can perfectly classify good air

quality while a 0.99 score for "Moderate" and "Hazardous" mean classification are near-perfect.

The model is reliable in differentiating between types of air quality based on environmental factors. As for near-perfect scores Moderate and Hazardous class indicates that the data for both classes used have similar discriminative power.

Decision Boundaries

Utilizing PCA to properly visualize the decision boundaries shown in Appendix O the researcher reduced the original 9 features to 2 (two) principal components. The results show 3 (three) decision regions, with "Good" class having clear boundaries while few data points are mixed between "Moderate" and "Hazardous" classes.

Odds Ratio

Appendix Q highlights the features that are statistically significant for our model with p value > 0.05 while coefficients are displayed through Equations.

For the Good air quality class, an increase in temperature drastically reduces the odds of air quality to be classified as "Good" by 86.05%. Similarly, increases in humidity, fine particulate matter (PM2.5), nitrogen dioxide and sulfur dioxide all decreases the likelihood of "Good" air quality by 47.7%, 38.84%, 85.13% and 86.38% respectively. An increase in carbon monoxide reduces the odds by 98.1%. In contrast, being outside industrial zones has positive impact on "Good" air quality, while a km increase in distance from industrial areas boost the odds of having "Good" air quality by 1422.84%, highlighting the effects of industrial activity to air quality, the population density indicates negative impact of having higher density on air quality, reducing the odds of having "Good" air by 52.4% as stated in *(Kleinman, M. T. 2020)*,*(Ozgok-Kangal, K. et. al. 2016)* carbon monoxide is prone to areas where human imposed heating, vehicle emissions and industrial activity are rampant that may lead to abrupt respiratory illness.

$$Equation\ 1$$

$$\begin{aligned}\eta Good = &-1.9694 \cdot temperature - 0.6483 \\ &\cdot humidity - 0.4916 \\ &\cdot fine\_matter - 1.9056 \\ &\cdot nitrogen\_dioxide - 1.9939 \\ &\cdot sulfur\_dioxide - 6.2472 \\ &\cdot carbon\_monoxide + 2.7232 \\ &\cdot Prox\_Industrial\_area \\ &- 0.7424 \cdot Pop\_density\end{aligned}$$

For the Moderate air quality category, the odds ratio results are mixed effect of environmental variables. An increase in temperature leads to a 30.41% increase in the odds of having moderate air quality, meanwhile an additional unit to humidity decreases the odds by 22.08%. Similarly, an increase in sulfur dioxide increases the odds of moderate air quality by 53.40%, while proximity to industrial areas reduces the odds by 52.08%, results suggest that being in close proximity of industrial area and high humid levels lean on having heightened levels of pollutants in "Moderate" air quality. Moreover, with an increase with carbon monoxide levels increases the odds of having moderate air quality by 228.11% this means that carbon monoxide has a strong relation with "Moderate" level of

air quality, reported in *(Levy, R. J. 2015)*, *(Ozgok-Kangal, K. et. al. 2016)* where exposure to carbon monoxide even at small amounts could lead to symptomatic effect on general public and impair neurodevelopment, particularly in children, that causes disruption of critical development of brain and auditory systems.

Equation 2

$$\eta Moderate = 0.2655 \cdot temperature - 0.2494 \cdot humidity + 0.4279 \cdot fine\_matter + 0.1178 \cdot nitrogen\_dioxide + 0.1775 \cdot sulfur\_dioxide + 0.0466 \cdot carbon\_monoxide - 0.4884 \cdot Prox\_Industrial\_area - 0.0789 \cdot Pop\_density$$

For "Hazardous" air quality, the model results suggest a linear relationship, where higher level of pollutants linked to increase of likelihood of having hazardous air quality. An increase in temperature increases the odds of hazardous air by 445.48%, and an increase in humidity raises the odds by 147.37%, while an increase with fine matter particles results in 45.36% increase in the odds that the air quality is "Hazardous". Moreover, an increase in nitrogen dioxide and sulfur dioxide both increases the chances of hazardous air quality. In particular, carbon monoxide has strong impact with an increase in ppm levels showed an immense 14,626.34% increase in the odds of hazardous air quality *(World Health Organization, 2021)* reported that once entered in respiratory system, carbon monoxide spreads across lung tissues and bloodstream that leads to deprivation of oxygen with damages to respiratory system, results in symptoms such as trouble in breathing, exhaustion, dizziness, and other flu-like symptoms and in high-levels may cause death. In contrast, an km increase from industrial area yields, to an 86.18% decrease in the likelihood of having dangerous levels of air, the findings stay consistent with other classes that the more distance from industrial areas results in better air quality *(Ozgok-Kangal, K. et. al. 2016)*. In similar manner, the population density increases lead to 110.76% likelihood of having hazardous air quality. The result remains linear, the more crowded an area the greater the emission levels leading to degrade in air quality *(Levy, R. J. 2015)*.

Equation 3

$$\eta Hazardous = 1.6965 \cdot temperature + 0.9057 \cdot humidity + 0.3741 \cdot fine\_matter + 1.7565 \cdot nitrogen\_dioxide + 1.5580 \cdot sulfur\_dioxide + 5.0515 \cdot carbon\_monoxide - 1.9794 \cdot Prox\_Industrial\_area + 0.7456 \cdot Pop\_density$$

CONCLUSION

This study used a Multinomial Logistic Regression model to classify air quality levels into three categories Good, Moderate, and Hazardous, based on environmental pollutants and urban density variables.

The dataset of 5,000 observations was subjected to assumptions; most tests was met except for Independence of irrelevant alternative (IIA) that future research could work upon. Based on the univariate analysis, the data suggests that moderately populated, warm and humid area are likely to have safe air quality levels according to WHO guideline as opposed to living near industrial areas or highly dense neighborhood. While, feature relationships are shown in bivariate analysis with fine matter and coarse matter show strong correlation as confirmed by Variance inflation factor (VIF) which can be fix by feature transformation. Also, both analyses showed increase in population density and close proximity with industrial area are linked increase carbon monoxide levels that leads to decrease in air quality.

Despite this, the model managed to have an accuracy of 0.97 and F1-average of 0.96. The" Good" class has the highest accuracy while "Moderate" and "Hazardous" had some overlap but still reliably predicted as supported by ROC/AUC and decision boundaries.

Odds ratio revealed that greater distance from industrial zones, lower levels of carbon monoxide and sulfur dioxide were the strongest variables to determine as "Good" air quality. On the other hand, hazardous conditions were associated by increased levels of carbon monoxide, nitrogen dioxide, and temperature.

The results points to industrial emission and byproduct produced by manmade combustion as key features in air quality classification. While the model proved it can predict air quality the violation of IIA assumption showed that it needs a flexible model approach.

Future Work

Future research should aim to boost model robustness by exploring alternatives that may capture underlying patterns from nonlinear relationships. Tree-based classifiers, such as Decision Trees, Random Forest, or Gradient Boosting may better account for complex feature interactions than MLR model.

While the observed violation of Independence of irrelevant Alternatives (IIA) calls for flexible models that can be use instead, such as nested logits, mixed logits or Generalized Additive Models (GAMs) that are tailored for nonlinear relationship between features.

Moreover, the use of expanded dataset that includes more data and broader environmental conditions would aid the model to generalize better and reduce the risk of overfitting and/or class imbalance.

Finally, further improvements of current MLR model through polynomial transformation of features, fine tuning Elastic Net (L1/L2 ratio) or Bayesian approaches for better coefficient stability and interpretability with respect to data sparsity.

REFERENCES

Almetwally, A. A., Bin-Jumah, M., & Allam, A. (2020). Ambient air pollution and its influence on human health and welfare: An overview. *Environmental Science and Pollution Research, 27*(Suppl 1), 1–17. https://doi.org/10.1007/s11356-020-09042-2

Arefin, M. R., & Asadujjaman, M. (2016, July). Minimizing average of loss functions using gradient descent and stochastic gradient descent. *Dhaka University Journal of Science, 64*(2), 141–145. https://doi.org/10.3329/dujs.v64i2.54490

Bleecker, M. L. (2015). Carbon monoxide intoxication. In Occupational neurology (pp. 191–203). Elsevier. https://doi.org/10.1016/b978-0-444-62627-1.00024-x

Chen, R., Yin, P., Meng, X., Wang, L., Liu, C., Niu, Y., Liu, Y., Liu, J., Qi, J., You, J., Kan, H., & Zhou, M. (2019). Associations between coarse particulate matter air pollution and cause-specific mortality: A nationwide analysis in 272 Chinese cities. *Environmental Health Perspectives, 127*(1), 017008. https://doi.org/10.1289/EHP2711

Cleff, T. (2025). Univariate data analysis. In *Applied statistics and multivariate data analysis for business and economics* (pp. 33–78). Springer. https://doi.org/10.1007/978-3-031-78070-7_3

Hancock, M., & Kent, P. (2016). Interpretation of dichotomous outcomes: Risk, odds, risk ratios, odds ratios, and number needed to treat. *Journal of Physiotherapy, 62*(3), 172–174. https://doi.org/10.1016/j.jphys.2016.02.016

Hua, Y., Stead, T. S., George, A., & Ganti, L. (2025). Clinical risk prediction with logistic regression: Best practices, validation techniques, and applications in medical research. *Academic Medicine & Surgery*. https://doi.org/10.62186/001c.131964

Jecinta, I. C., & Obi, J. C. (2023, January). A review of techniques for regularization. *International Journal of Research in Engineering and Science (IJRES), 11*(1), 360–367. http://www.ijres.org/papers/Vol11-Issue1/111360367.pdf

Khyber Medical University Journal. (2021). Health impacts of air pollution and climate change. *KMUJ: Khyber Medical University Journal, 14*(4), 254–256. https://doi.org/10.35845/kmuj.2021.22287

Kleinman, M. T. (2020). Carbon monoxide. In Environmental toxicants (pp. 455–486). Wiley. https://doi.org/10.1002/9781119438922.ch12

Köhler, T. (2023). Quantitative data analysis: Descriptive, univariate, and bivariate statistics. In *Business research methods* (Chapter 26). Oxford University Press. https://doi.org/10.1093/hebz/9780198869443.003.0026

Levy, R. J. (2015). Carbon monoxide pollution and neurodevelopment: A public health concern. *Neurotoxicology and Teratology, 49*, 31–40. https://doi.org/10.1016/j.ntt.2015.03.001

Mateen, M. (2024). Air Quality and Pollution Assessment [Data set]. *Kaggle*. https://doi.org/10.34740/KAGGLE/DS/6197184

Ozgok-Kangal, K., Arziman, İ., Uzun, G., & Yildiz, Ş. (2016). Carbon monoxide poisoning in the workplace: A hidden danger. *Apollo Medicine, 13*(3), 196–197. https://doi.org/10.1016/j.apme.2016.02.002

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

World Bank. (2024). Population density (people per sq. km of land area). *The World Bank*. https://data.worldbank.org/indicator/EN.POP.DNST?end=2022&start=1961&view=chart

World Health Organization. (2021.). Types of pollutants. World Health Organization. Retrieved August 4, 2025, from https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/health-impacts/types-of-pollutants

World Health Organization. (2021). What are the WHO air quality guidelines? https://www.who.int/news-room/feature-stories/detail/what-are-the-who-air-quality-guidelines

World Health Organization. (2024, October 24). Ambient (outdoor) air quality and health. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

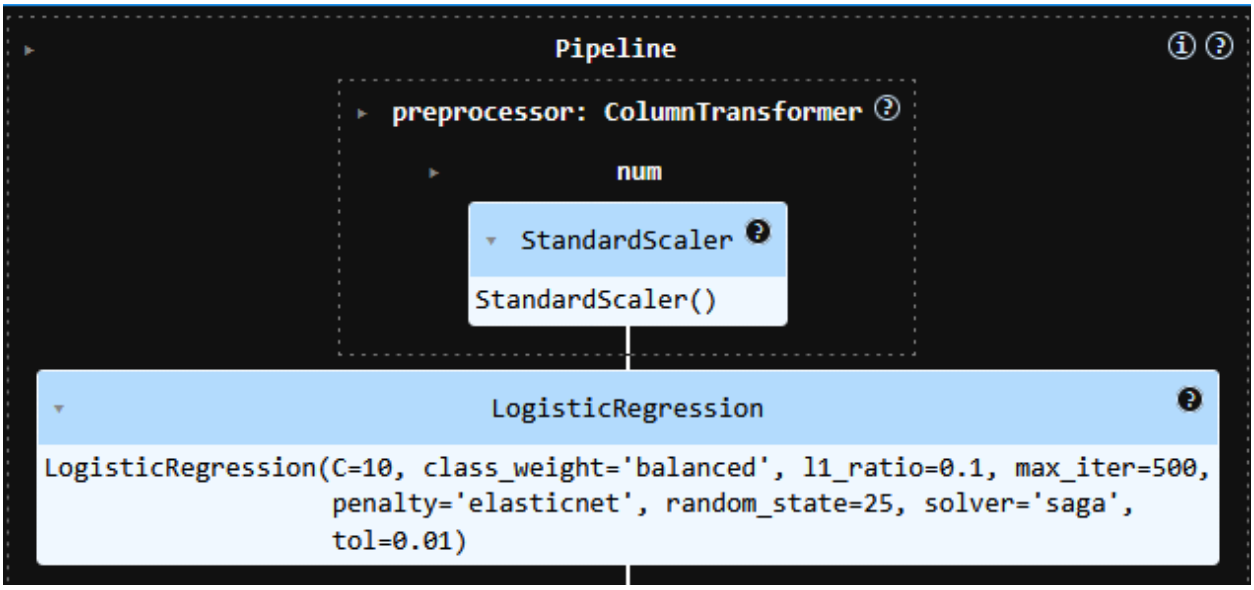
World Health Organization. (2024). WHO air quality database. *World Health Organization*. https://www.who.int/data/gho/data/themes/air-pollution/who-air-quality-database

APPENDIX

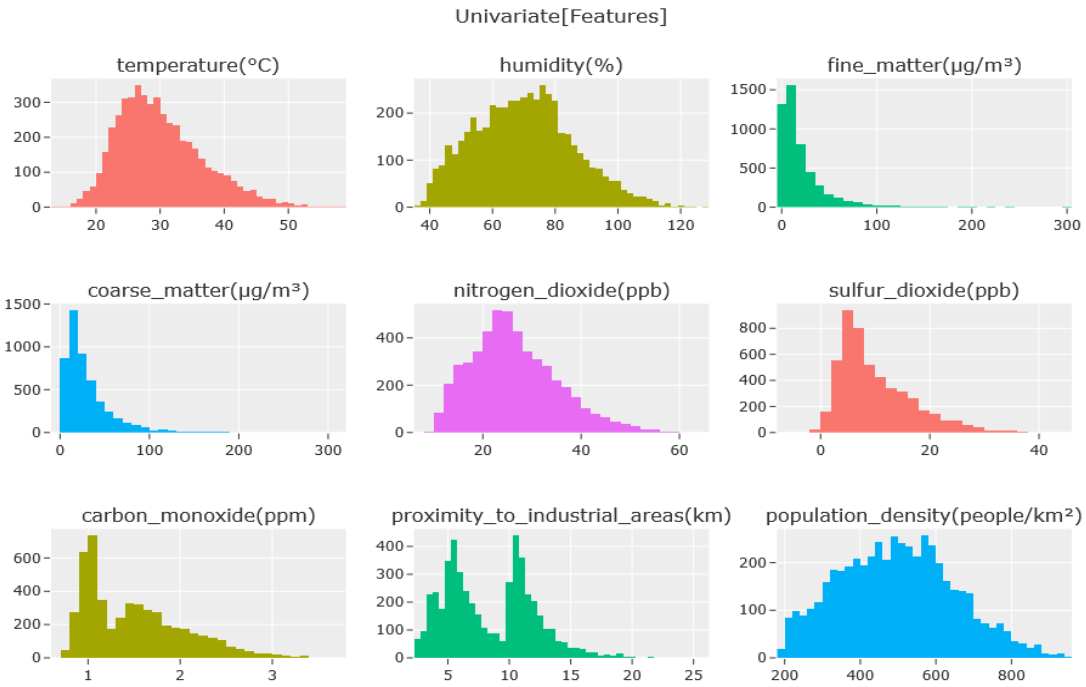
Appendix A. Class Distribution

	Class	Full_Count	Train_Count	Val_Count	Test_Count	Full_Prop	Train_Prop	Val_Prop	Test_Prop
0	Good	1981	1387	396	198	0.447886	0.447997	0.447964	0.446953
1	Hazardous	1024	717	204	103	0.231517	0.231589	0.230769	0.232506
2	Moderate	1418	992	284	142	0.320597	0.320413	0.321267	0.320542

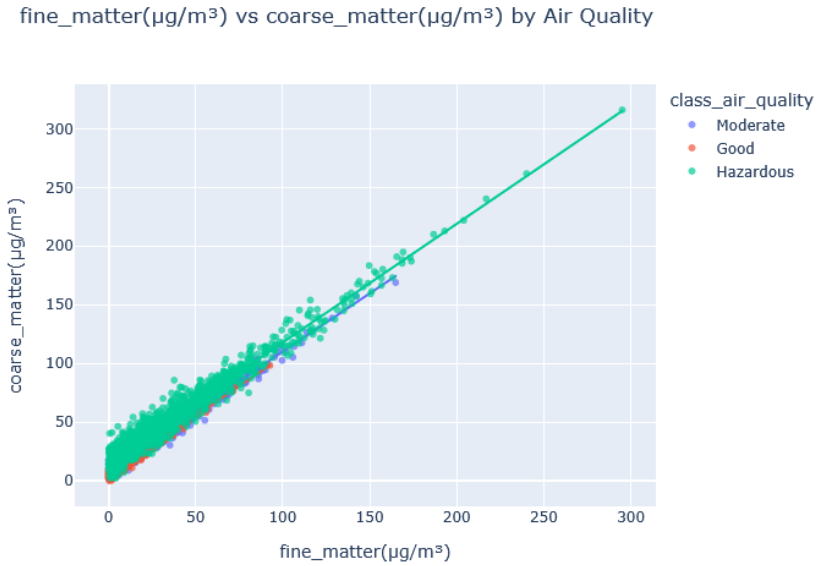
Appendix B. Model Pipeline

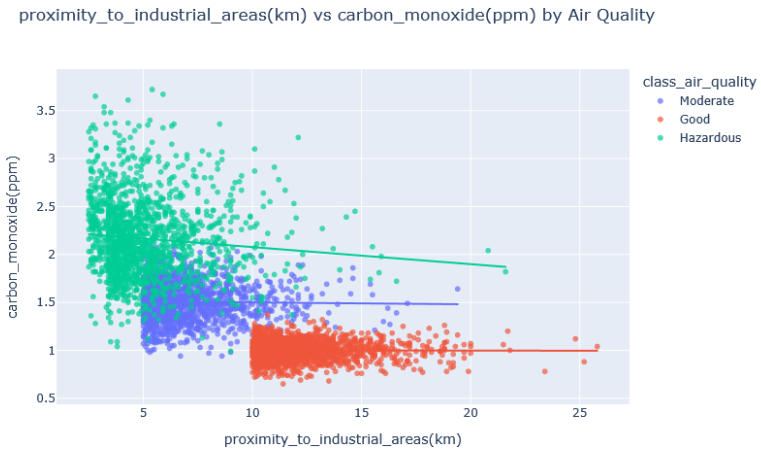


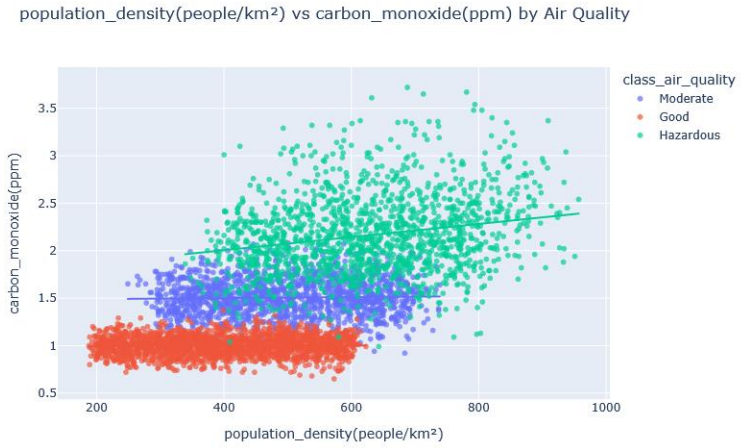
Appendix C. Univariate Analysis



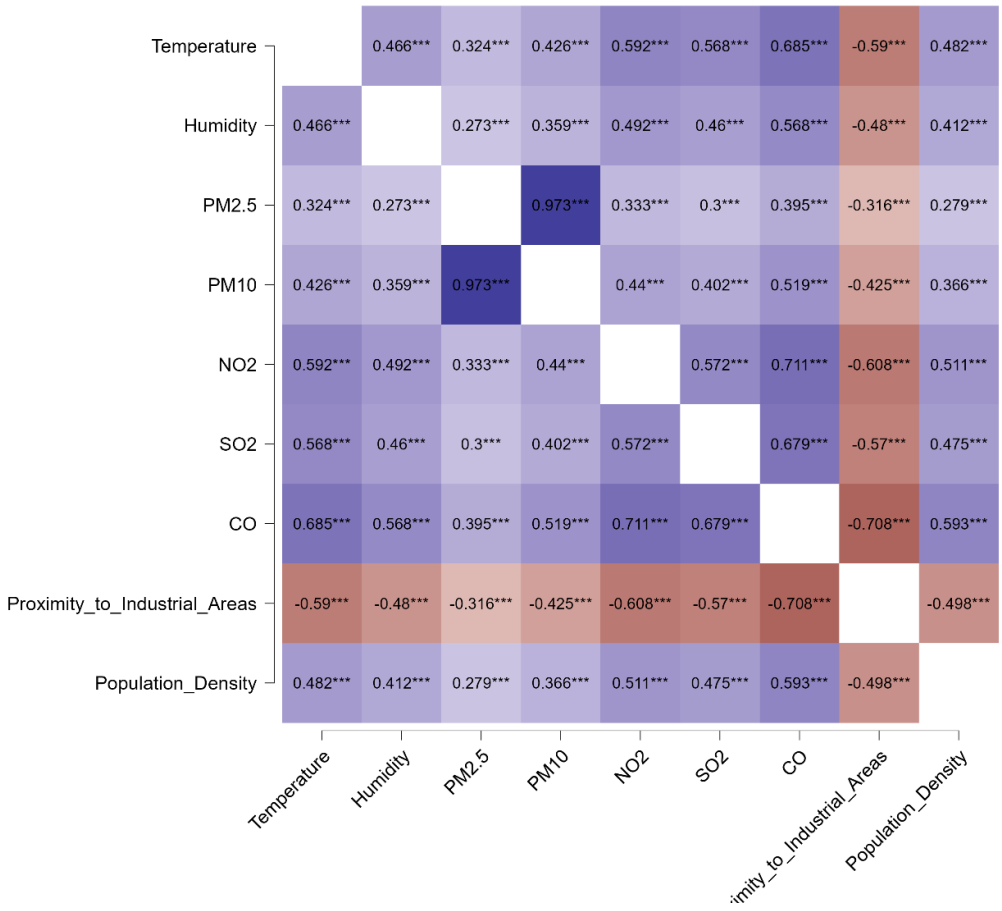
[Appendix D](). Bivariate Scatter Plots







Appendix E. Heat Map



Appendix F. Independent Observations

2. Independence of Observation [PASSED]

```
df.shape
```

```
(5000, 10)
```

```
df.duplicated().sum()
```

```
0
```

Appendix G. Sample Size

3. Sufficient Sample Size 10–20 observations per feature [PASSED]

```
feat_count = len(df.columns[:-1])
print(f'Number of features: {feat_count}')
print(f'Number of classes: {df.class_air_quality.nunique()}')
print('\n20 Observations needed per feature')

required_observation = (feat_count * df.class_air_quality.nunique() * 20)

print(f'\nNumber of Observations needed: {required_observation}')
print(f'\nTotal Obervations (Dataset*) {df.shape[0]}')

print(f'\n{'PASSED' if df.shape[0] > required_observation else 'FAILED'}')
```

```
Number of features: 9
Number of classes: 3

20 Observations needed per feature

Number of Observations needed: 540

Total Obervations (Dataset*) 5000

PASSED
```

Appendix H. Variance Inflation Factor (After and Before)

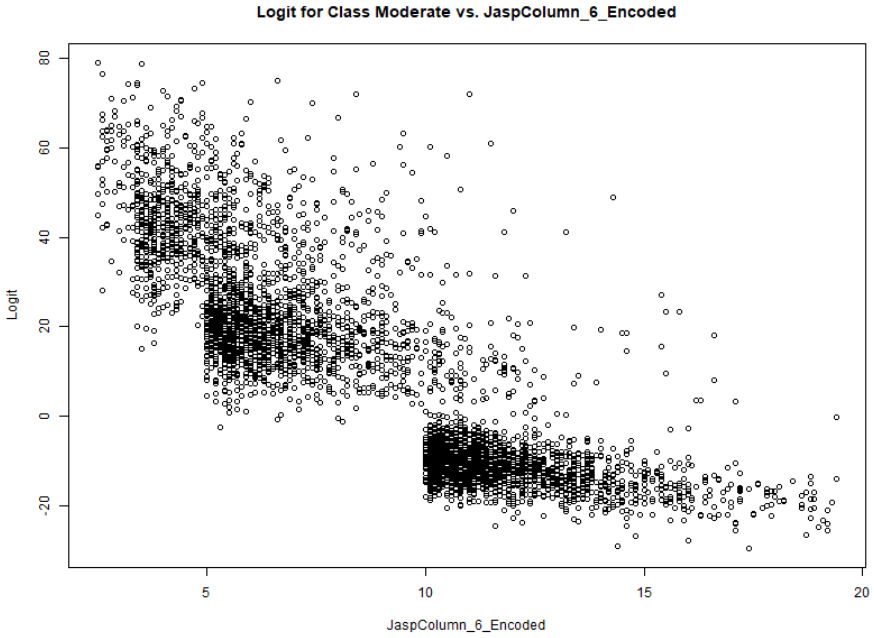
4. No Multicollinearity Among Predictors [PASSED] ¶

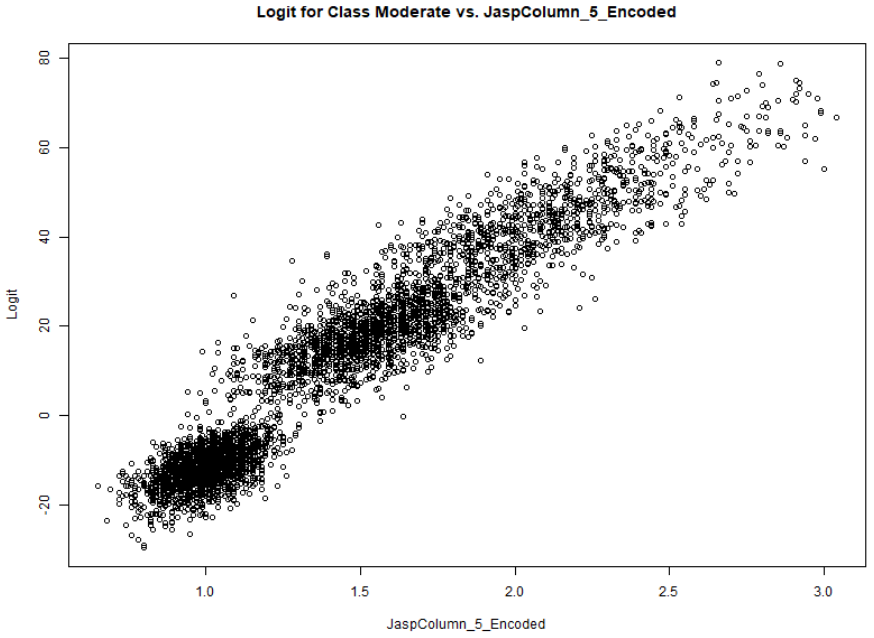
```
X_reduced = X_scaled.drop(columns=['coarse_matter(µg/m³)'])
vif_data = pd.DataFrame()
vif_data['Feature'] = X_reduced.columns
vif_data['VIF'] = [variance_inflation_factor(X_reduced.values, i) for i in range(X_reduced.shape[1])]
print(vif_data)
```

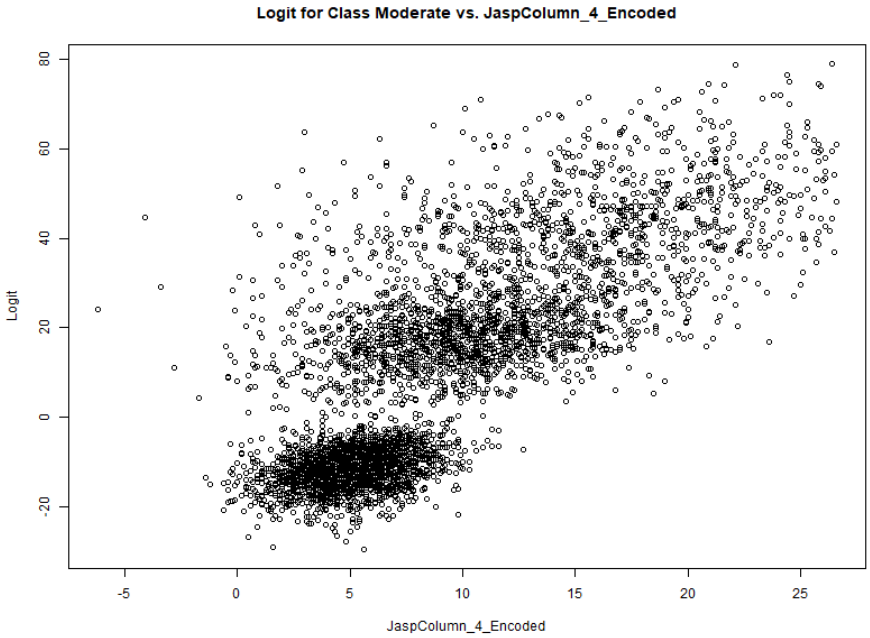
```
                          Feature       VIF
0                 temperature(°C)  2.092231
1                     humidity(%)  1.557295
2              fine_matter(µg/m³)  1.203047
3           nitrogen_dioxide(ppb)  2.260311
4             sulfur_dioxide(ppb)  2.013364
5            carbon_monoxide(ppm)  3.727626
6  proximity_to_industrial_areas(km)  2.215939
7    population_density(people/km²)  1.630867
```

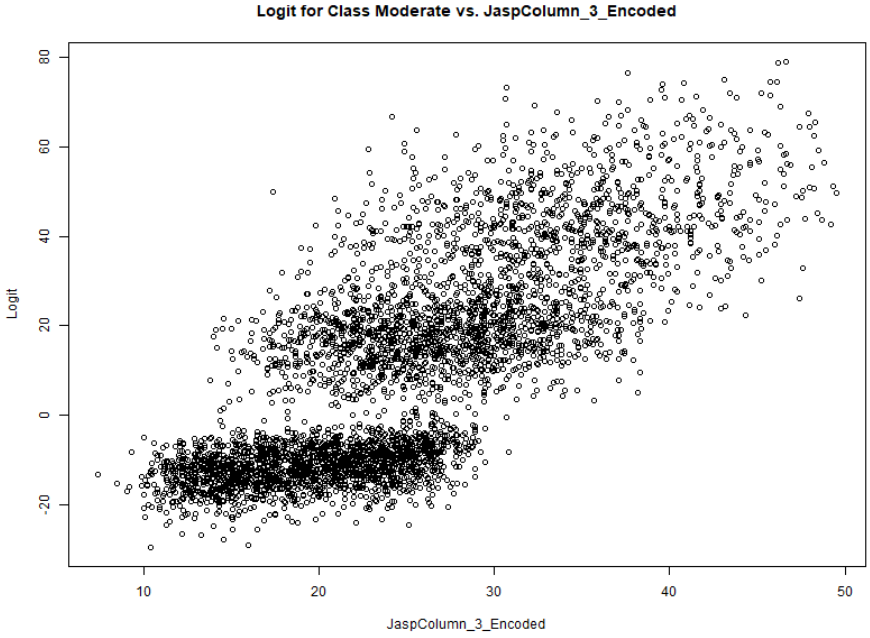
	Feature	VIF
0	temperature(°C)	2.108506
1	humidity(%)	1.566619
2	fine_matter(µg/m³)	29.401193
3	coarse_matter(µg/m³)	34.566851
4	nitrogen_dioxide(ppb)	2.287701
5	sulfur_dioxide(ppb)	2.029253
6	carbon_monoxide(ppm)	3.913178
7	proximity_to_industrial_areas(km)	2.250473
8	population_density(people/km²)	1.636513

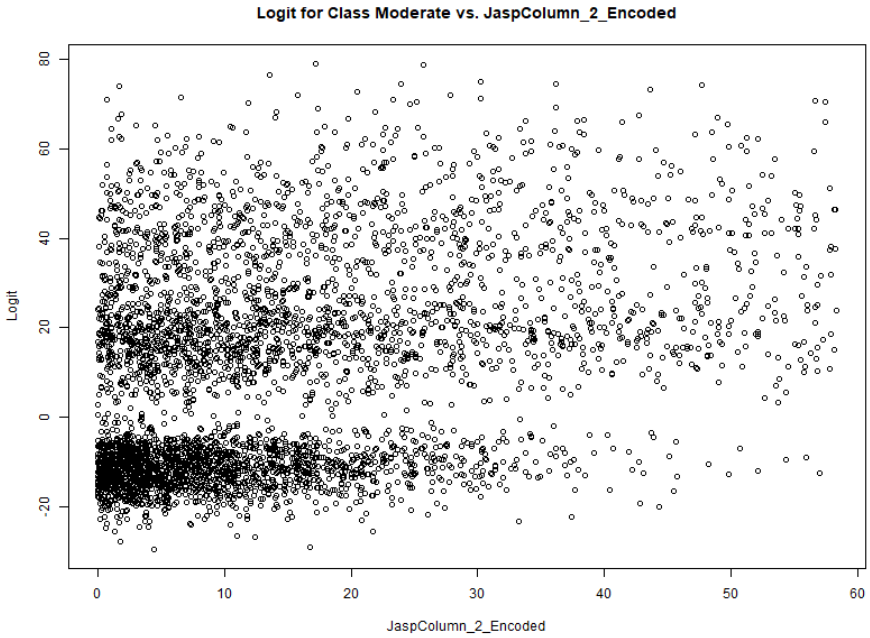
Appendix I. Linearity of Logits

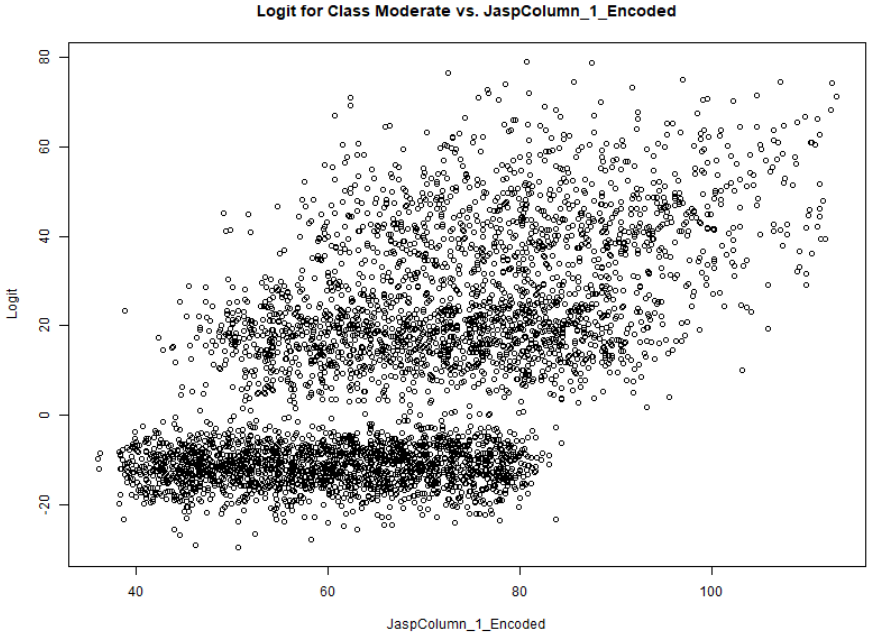


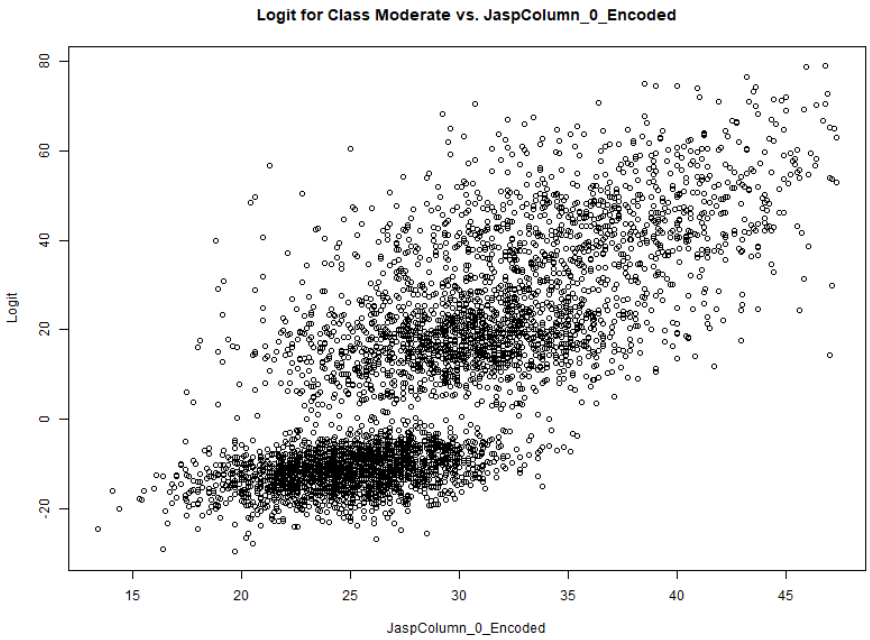


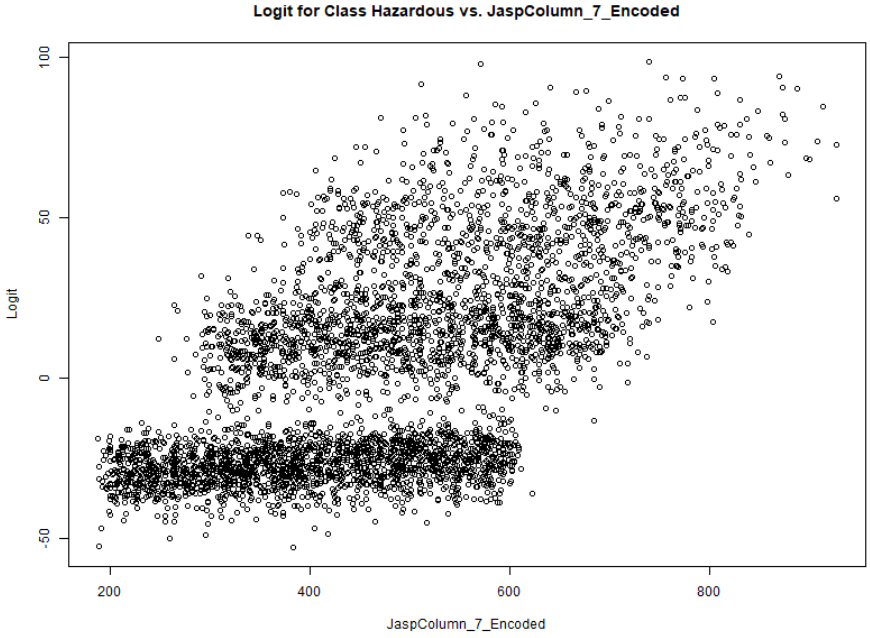

Logit for Class Moderate vs. JaspColumn_4_Encoded
Logit
JaspColumn_4_Encoded
-20
0
20
40
60
80
-5
0
5
10
15
20
25

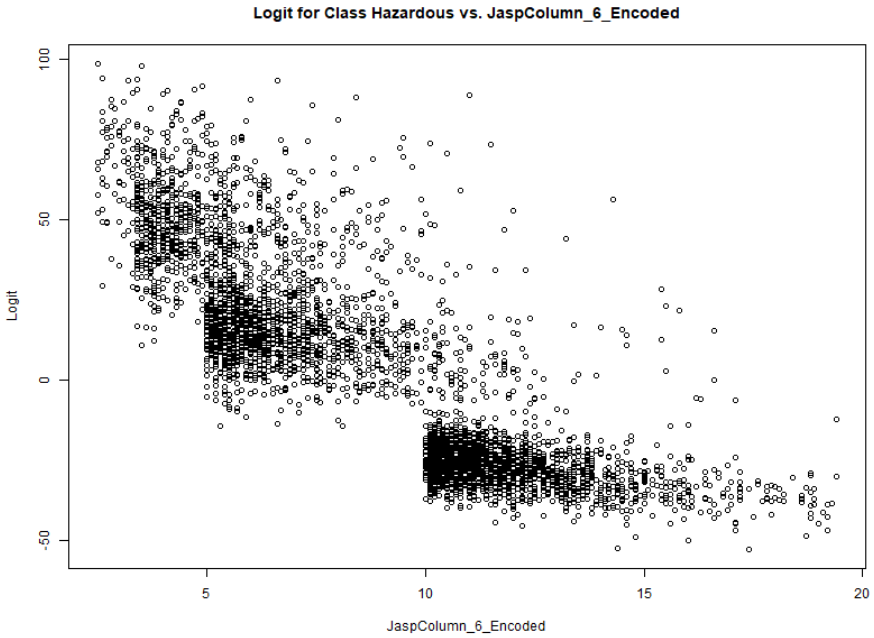

Logit for Class Moderate vs. JaspColumn_3_Encoded
Logit
JaspColumn_3_Encoded
-20
0
20
40
60
80
10
20
30
40
50

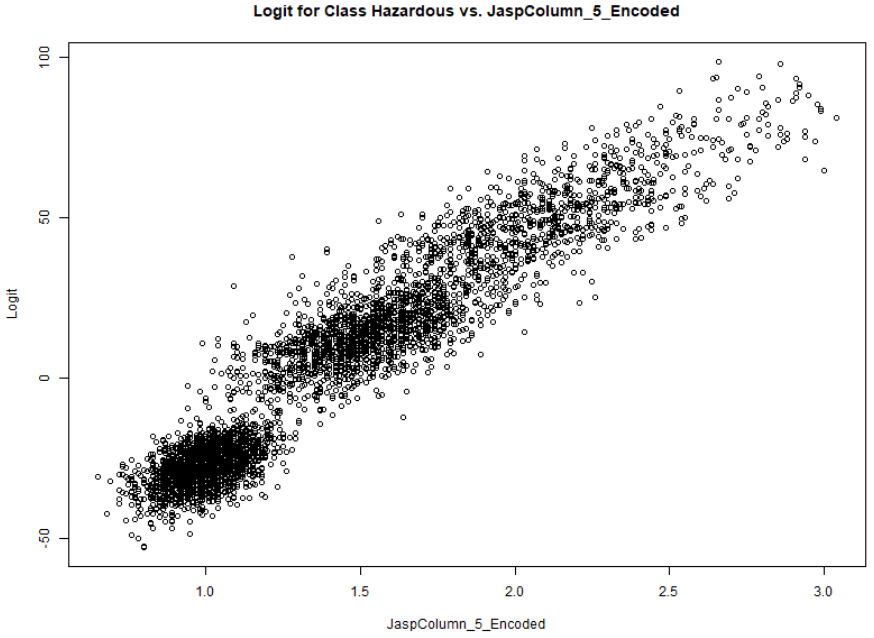

Logit for Class Moderate vs. JaspColumn_2_Encoded
Logit
JaspColumn_2_Encoded
80
60
40
20
0
-20
0
10
20
30
40
50
60

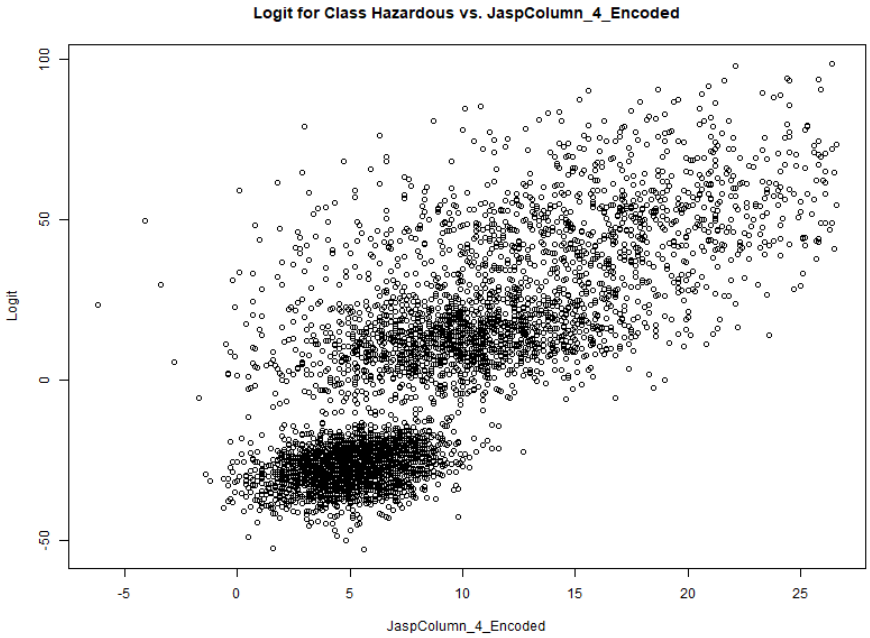

Logit for Class Moderate vs. JaspColumn_1_Encoded
Logit
JaspColumn_1_Encoded
80
60
40
20
0
-20
40
60
80
100

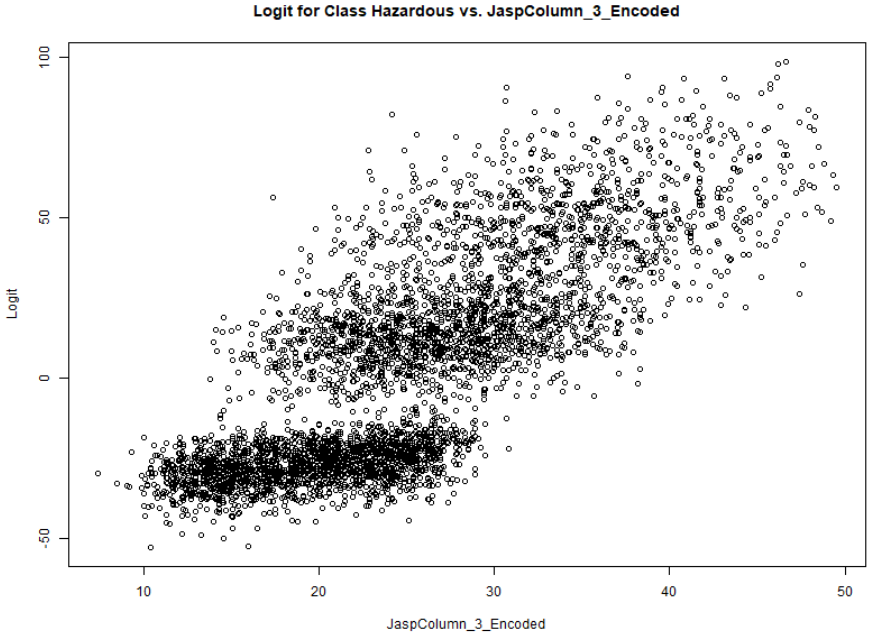

Logit for Class Moderate vs. JaspColumn_0_Encoded
Logit
JaspColumn_0_Encoded
-20
0
20
40
60
80
15
20
25
30
35
40
45

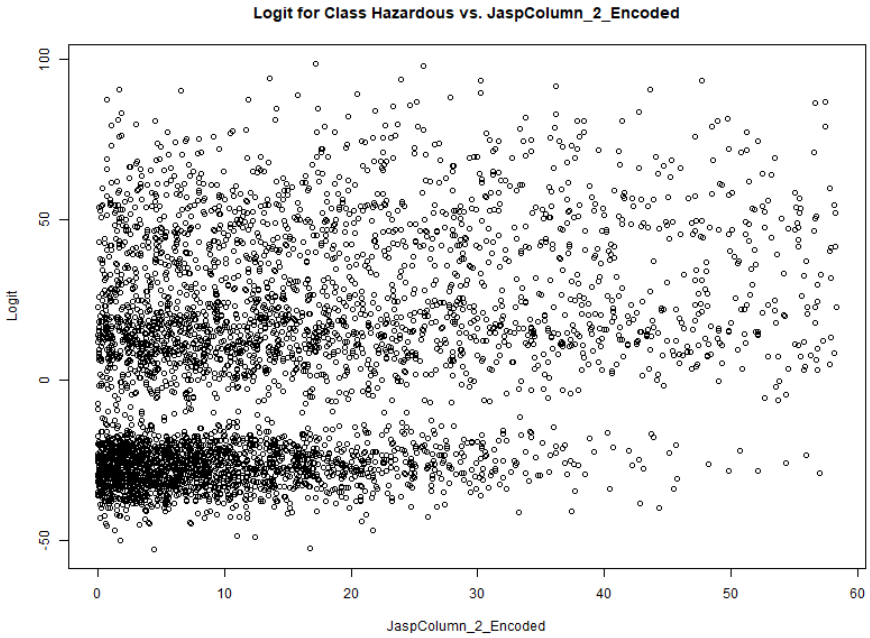

Logit for Class Hazardous vs. JaspColumn_7_Encoded
Logit
JaspColumn_7_Encoded
-50
0
50
100
200
400
600
800

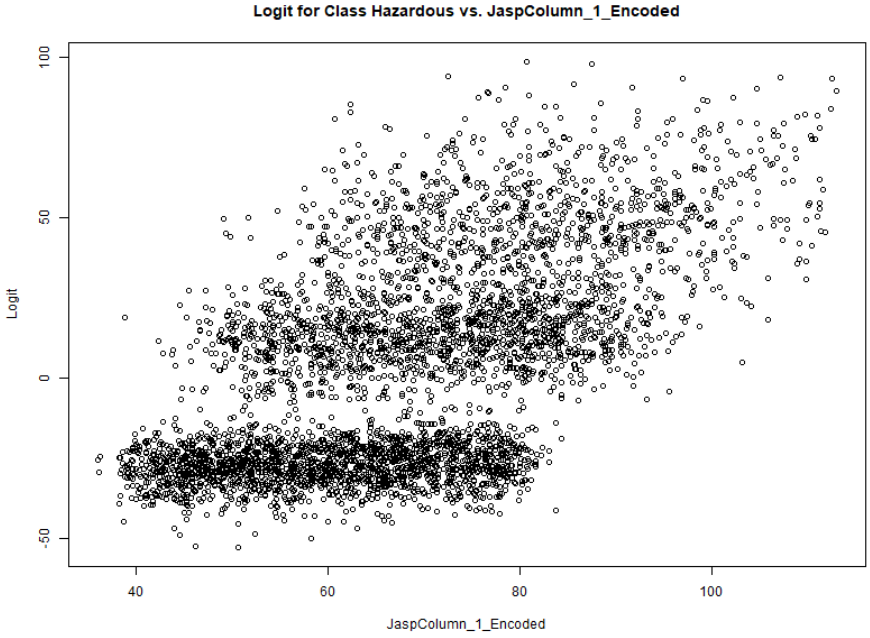

Logit for Class Hazardous vs. JaspColumn_6_Encoded
Logit
100
50
0
-50
5
10
15
20
JaspColumn_6_Encoded

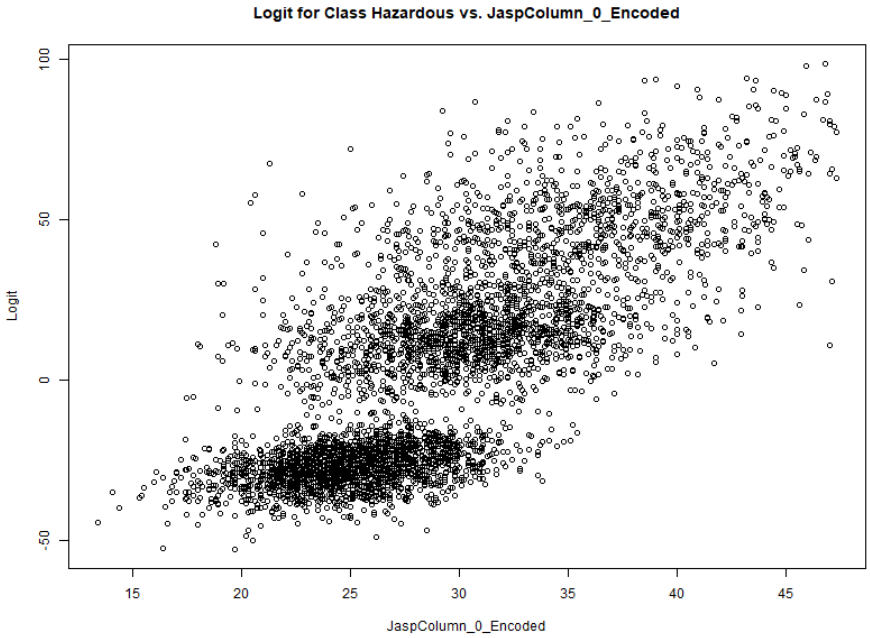

Logit for Class Hazardous vs. JaspColumn_5_Encoded
Logit
100
50
0
-50
1.0
1.5
2.0
2.5
3.0
JaspColumn_5_Encoded

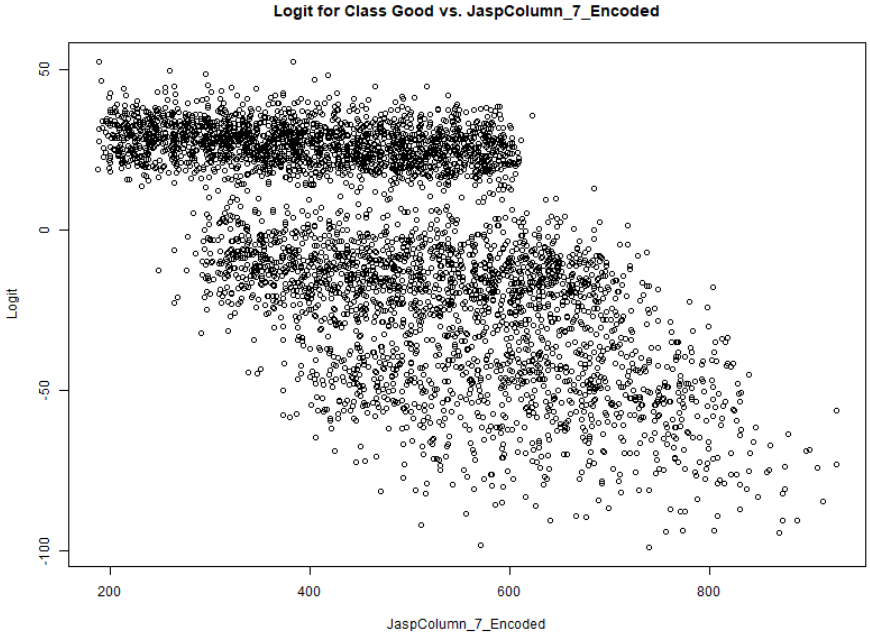

Logit for Class Hazardous vs. JaspColumn_4_Encoded
Logit
JaspColumn_4_Encoded
100
50
0
-50
-5
0
5
10
15
20
25

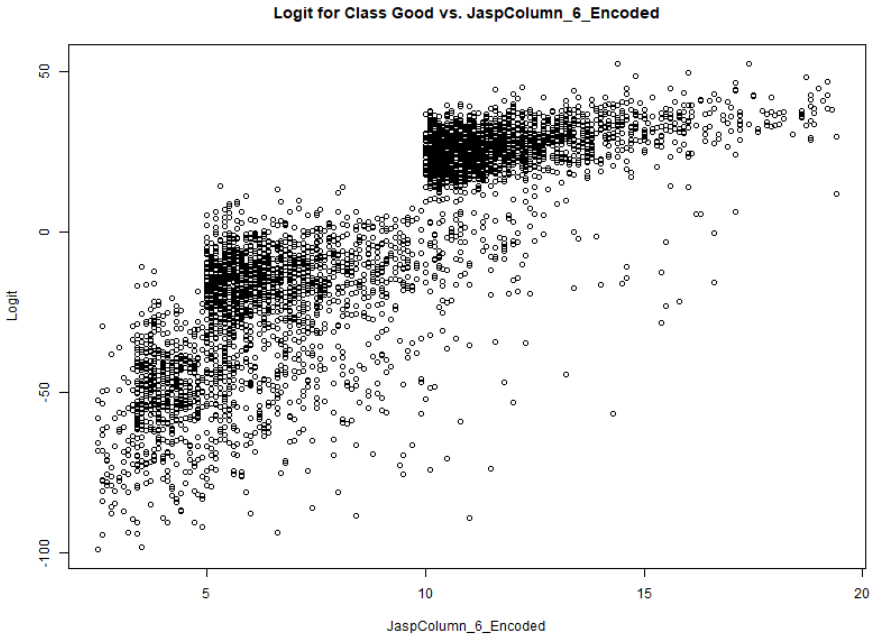

Logit for Class Hazardous vs. JaspColumn_3_Encoded
Logit
JaspColumn_3_Encoded
100
50
0
-50
10
20
30
40
50

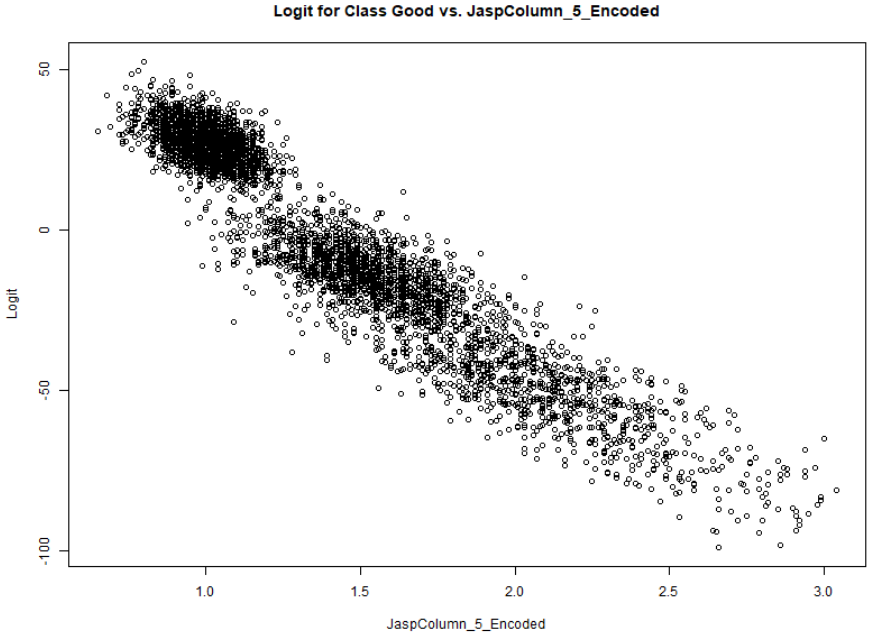

Logit for Class Hazardous vs. JaspColumn_2_Encoded
Logit
100
50
0
-50
0
10
20
30
40
50
60
JaspColumn_2_Encoded

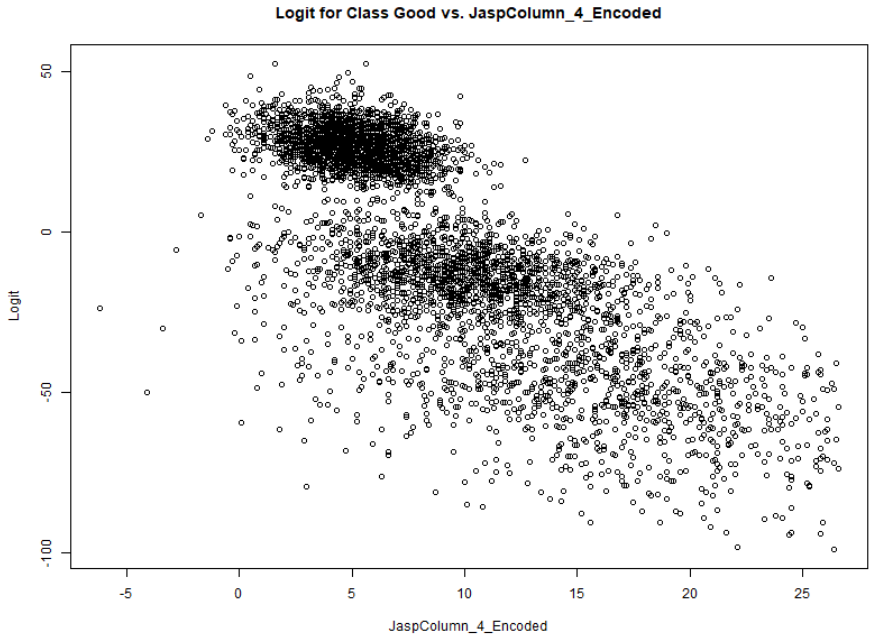

Logit for Class Hazardous vs. JaspColumn_1_Encoded
Logit
100
50
0
-50
40
60
80
100
JaspColumn_1_Encoded

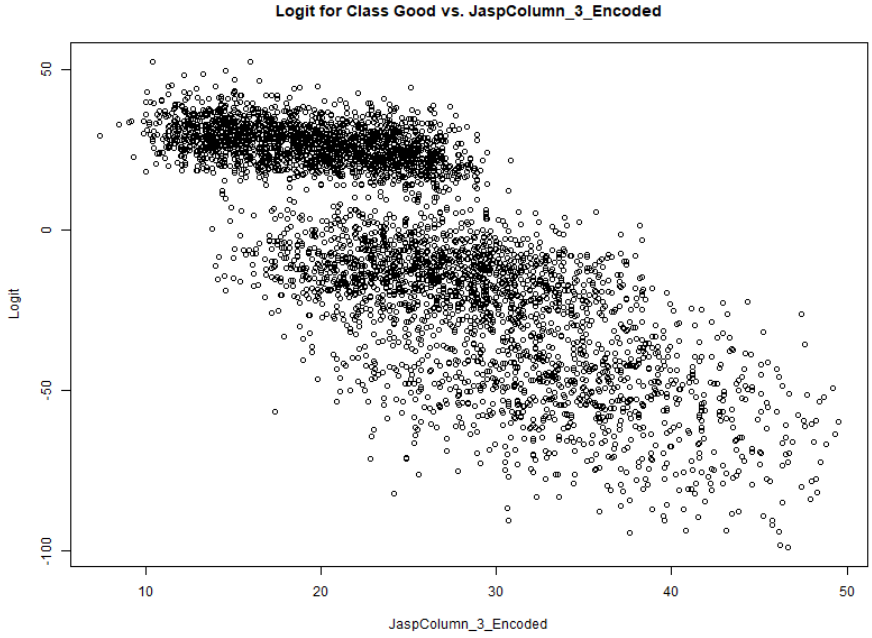

Logit for Class Hazardous vs. JaspColumn_0_Encoded
Logit
JaspColumn_0_Encoded
100
50
0
-50
15
20
25
30
35
40
45

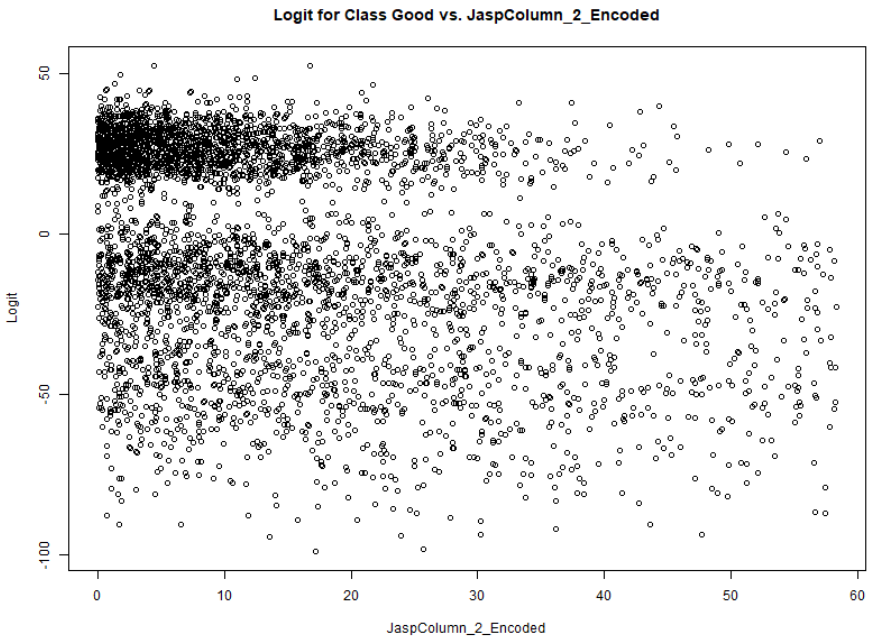

Logit for Class Good vs. JaspColumn_7_Encoded
Logit
JaspColumn_7_Encoded
50
0
-50
-100
200
400
600
800

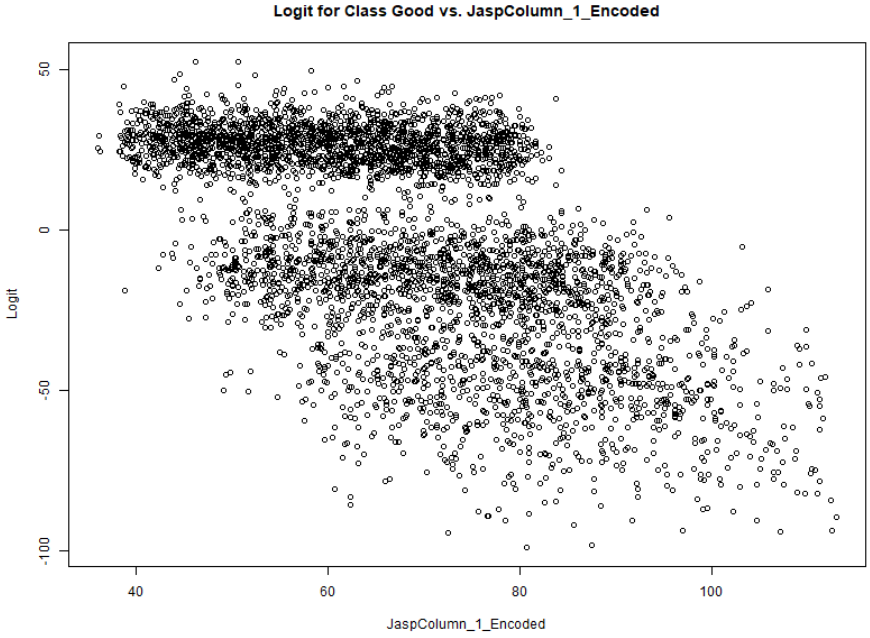

Logit for Class Good vs. JaspColumn_6_Encoded
Logit
50
0
-50
-100
5
10
15
20
JaspColumn_6_Encoded

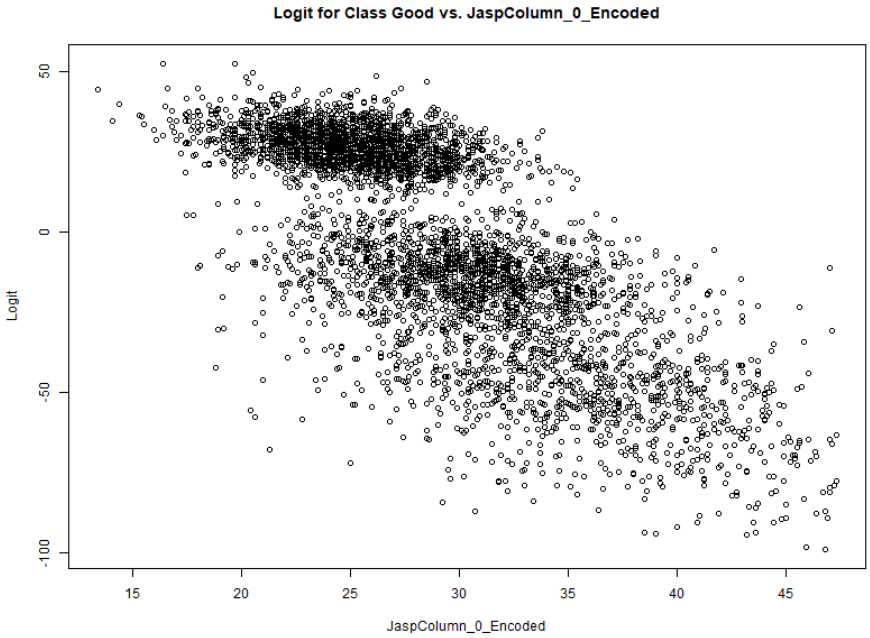

Logit for Class Good vs. JaspColumn_5_Encoded
Logit
50
0
-50
-100
1.0
1.5
2.0
2.5
3.0
JaspColumn_5_Encoded

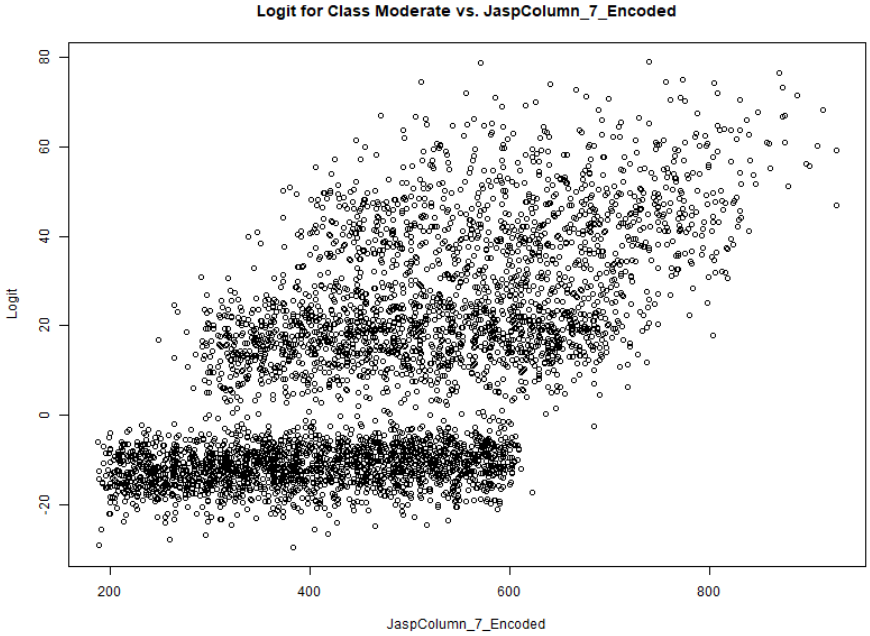

Logit for Class Good vs. JaspColumn_4_Encoded
Logit
50
0
-50
-100
-5
0
5
10
15
20
25
JaspColumn_4_Encoded


Logit for Class Good vs. JaspColumn_3_Encoded
Logit
50
0
-50
-100
10
20
30
40
50
JaspColumn_3_Encoded

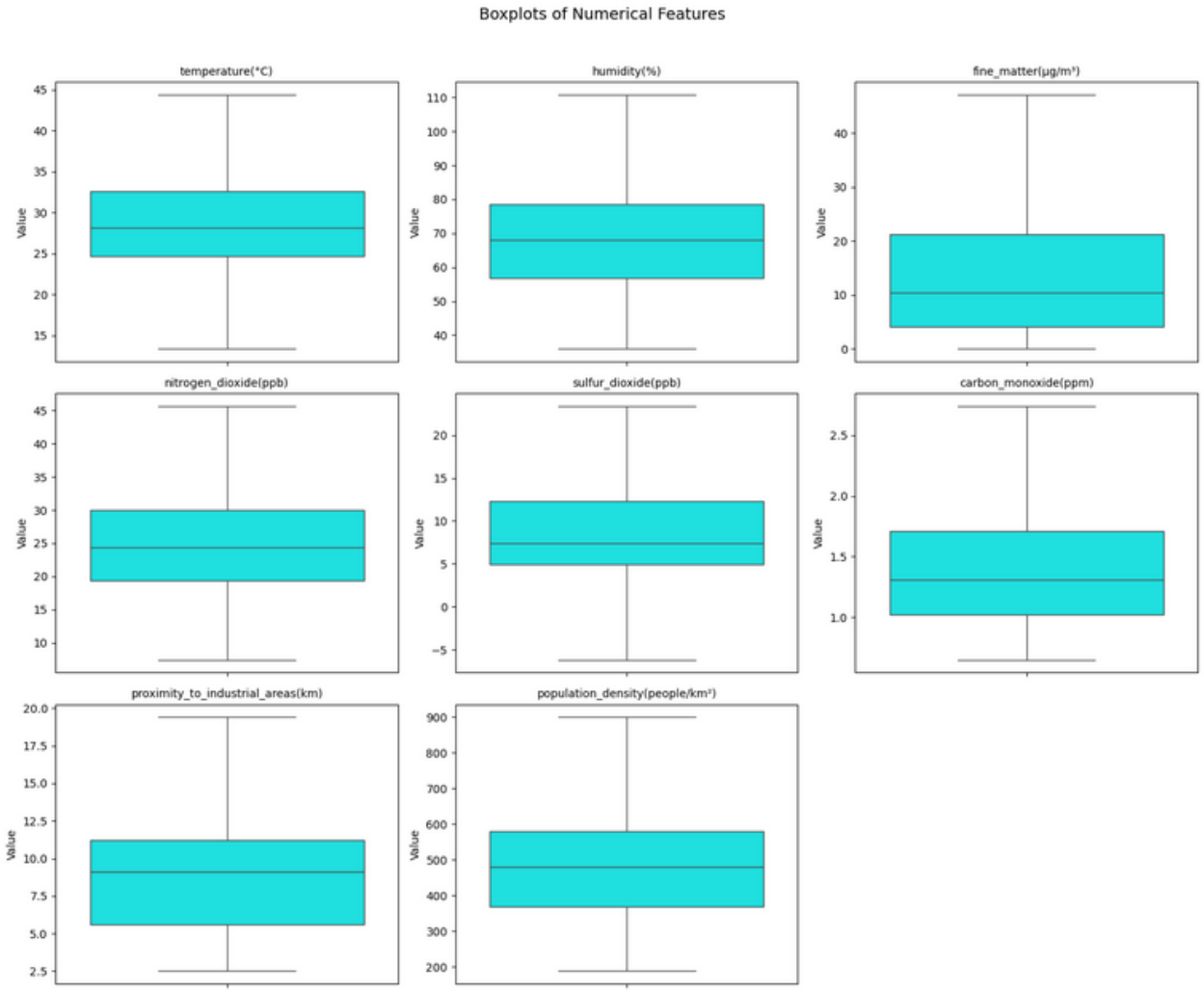

Logit for Class Good vs. JaspColumn_2_Encoded
Logit
50
0
-50
-100
0
10
20
30
40
50
60
JaspColumn_2_Encoded

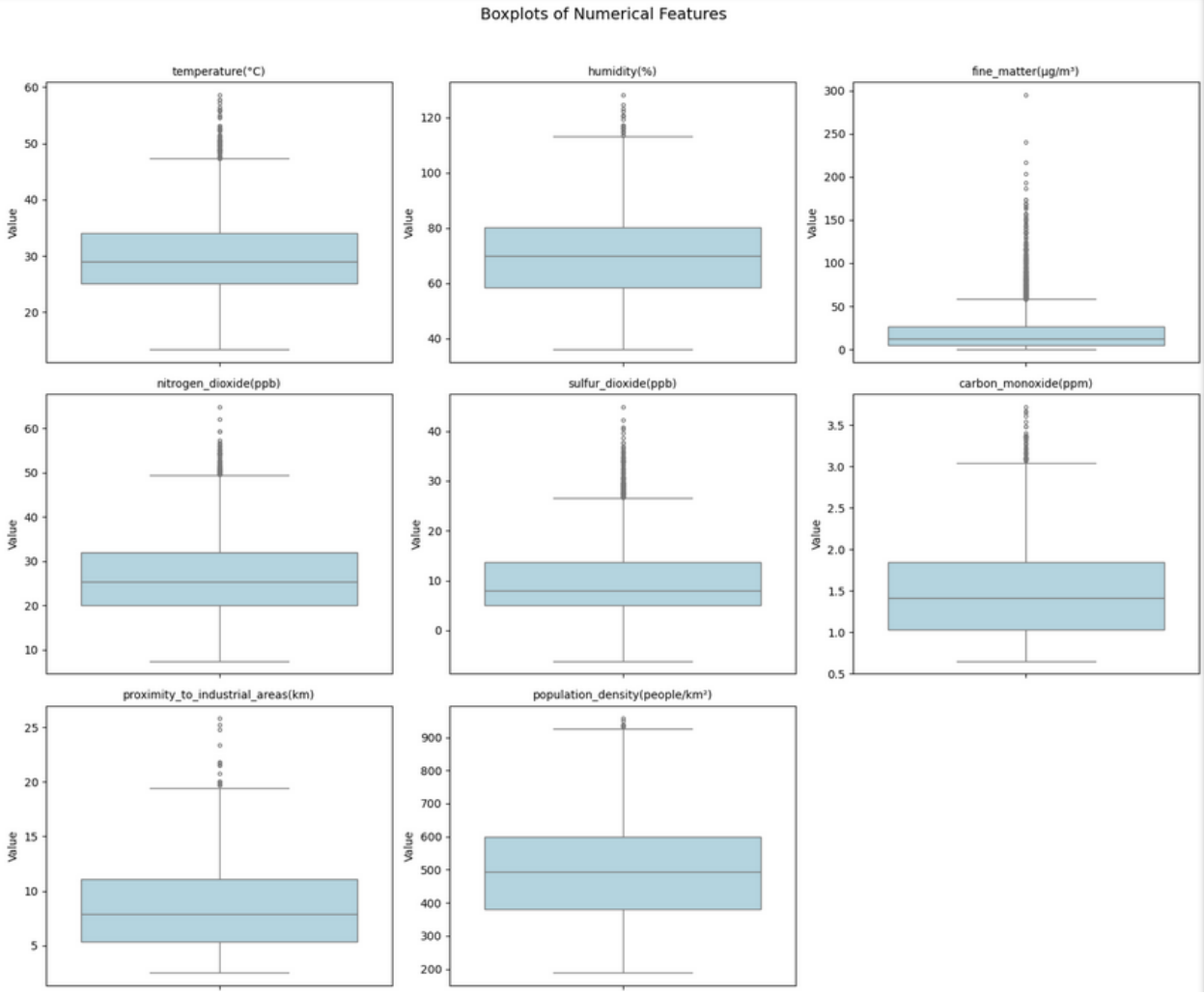

Logit for Class Good vs. JaspColumn_1_Encoded
Logit
50
0
-50
-100
40
60
80
100
JaspColumn_1_Encoded


Logit for Class Good vs. JaspColumn_0_Encoded
Logit
50
0
-50
-100
15
20
25
30
35
40
45
JaspColumn_0_Encoded


Logit for Class Moderate vs. JaspColumn_7_Encoded
Logit
80
60
40
20
0
-20
200
400
600
800
JaspColumn_7_Encoded

Appendix J. No Strong Outliers (After and Before)





Appendix K. IIA

7. Independence of Irrelevant Alternatives Using R through JASP [Failed]

IIA p val less than 0.05 reject the null hypothesis the alternatives are independent of irrelevant alternatives, presence of other alternatives does affect the odds between the original alternatives

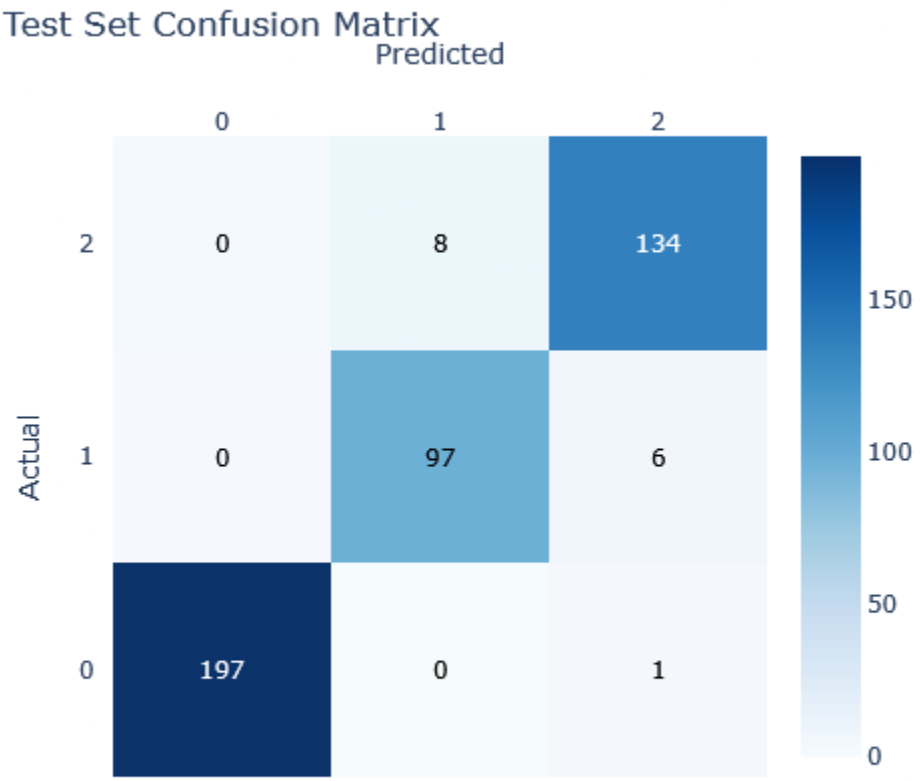
```
# weights:  30 (18 variable)
initial  value 4859.162153
iter  10 value 2766.733448
iter  20 value 1134.105295
iter  30 value 426.886728
iter  40 value 418.437222
iter  50 value 415.234925
iter  60 value 402.414118
final  value 401.993597
converged
Warning: group 'Good' is empty# weights:  10 (9 variable)
initial  value 1692.665415
iter  10 value 553.953624
iter  20 value 358.112407
iter  30 value 358.102500
final  value 358.102464
converged
Likelihood Ratio Test Statistic:  87.78227
Degrees of Freedom:  9
p-value:  4.551914e-15
```

Appendix L. Model Classification Report (Test set)

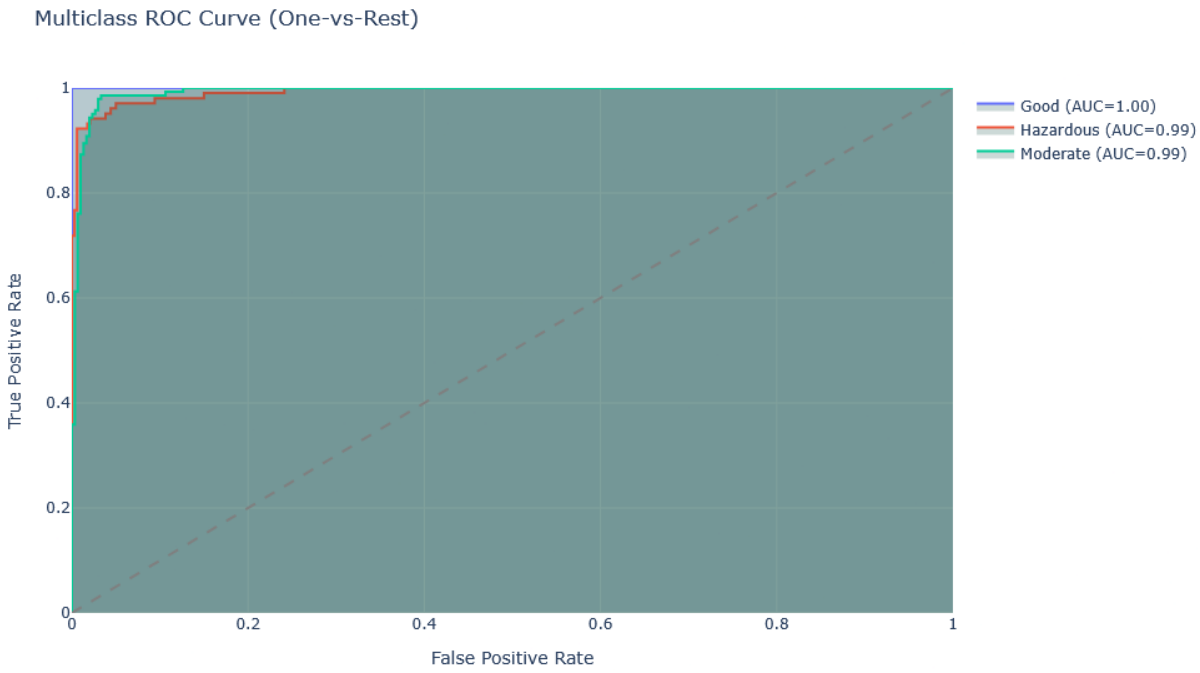
Test Classification Report:

	precision	recall	f1-score	support
0	1.00	0.99	1.00	198
1	0.92	0.94	0.93	103
2	0.95	0.94	0.95	142
accuracy			0.97	443
macro avg	0.96	0.96	0.96	443
weighted avg	0.97	0.97	0.97	443

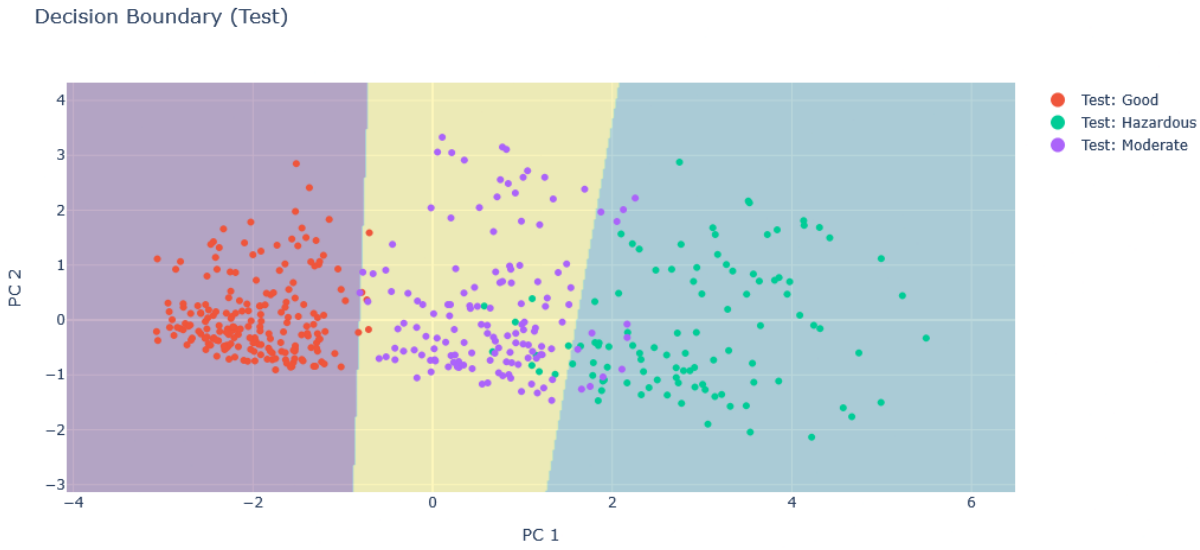
Appendix M. Confusion Matrix (Test set)



Appendix N. ROC and AUC



Appendix O. Decision Boundaries



Appendix P. Regression Coefficients

	Class	Feature	Coef	SE	z	p	CI_lower	CI_upper	OddsRatio
0	Good	temperature(°C)	-1.9694	0.2043	-9.6389	0.0000	-2.3699	-1.5689	0.1395
1	Good	humidity(%)	-0.6483	0.1701	-3.8104	0.0001	-0.9817	-0.3148	0.5230
2	Good	fine_matter(μg/m³)	-0.4916	0.1413	-3.4784	0.0005	-0.7686	-0.2146	0.6116
3	Good	nitrogen_dioxide(ppb)	-1.9056	0.2096	-9.0904	0.0000	-2.3165	-1.4947	0.1487
4	Good	sulfur_dioxide(ppb)	-1.9939	0.2247	-8.8727	0.0000	-2.4344	-1.5535	0.1362
5	Good	carbon_monoxide(ppm)	-6.2472	0.3089	-20.2240	0.0000	-6.8526	-5.6417	0.0019
6	Good	proximity_to_industrial_areas(km)	2.7232	0.2822	9.6493	0.0000	2.1700	3.2763	15.2284
7	Good	population_density(people/km²)	-0.7424	0.1633	-4.5460	0.0000	-1.0625	-0.4223	0.4760
8	Hazardous	temperature(°C)	1.6965	0.1548	10.9580	0.0000	1.3931	1.9999	5.4548
9	Hazardous	humidity(%)	0.9057	0.1438	6.2991	0.0000	0.6239	1.1875	2.4737
10	Hazardous	fine_matter(μg/m³)	0.3741	0.0982	3.8095	0.0001	0.1816	0.5665	1.4536
11	Hazardous	nitrogen_dioxide(ppb)	1.7565	0.1677	10.4732	0.0000	1.4278	2.0852	5.7920
12	Hazardous	sulfur_dioxide(ppb)	1.5580	0.1478	10.5439	0.0000	1.2683	1.8476	4.7491
13	Hazardous	carbon_monoxide(ppm)	5.0515	0.2415	20.9159	0.0000	4.5782	5.5249	156.2634
14	Hazardous	proximity_to_industrial_areas(km)	-1.9794	0.2329	-8.4984	0.0000	-2.4359	-1.5229	0.1382
15	Hazardous	population_density(people/km²)	0.7456	0.1250	5.9659	0.0000	0.5006	0.9905	2.1076
16	Moderate	temperature(°C)	0.2655	0.1151	2.3062	0.0211	0.0399	0.4912	1.3041
17	Moderate	humidity(%)	-0.2494	0.0976	-2.5569	0.0106	-0.4406	-0.0582	0.7792
18	Moderate	fine_matter(μg/m³)	0.1106	0.0760	1.4555	0.1455	-0.0383	0.2596	1.1170
19	Moderate	nitrogen_dioxide(ppb)	0.1436	0.1171	1.2266	0.2200	-0.0859	0.3732	1.1545
20	Moderate	sulfur_dioxide(ppb)	0.4279	0.1189	3.6002	0.0003	0.1949	0.6608	1.5340
21	Moderate	carbon_monoxide(ppm)	1.1882	0.1758	6.7580	0.0000	0.8436	1.5328	3.2811
22	Moderate	proximity_to_industrial_areas(km)	-0.7356	0.1615	-4.5540	0.0000	-1.0522	-0.4190	0.4792
23	Moderate	population_density(people/km²)	-0.0013	0.0920	-0.0138	0.9890	-0.1816	0.1790	0.9987

Appendix Q. Regression Coefficients with p-value < 0.05

	Class	Feature	Coef	SE	z	p	CI_lower	CI_upper	OddsRatio
0	Good	temperature(°C)	-1.9694	0.2043	-9.6389	0.0000	-2.3699	-1.5689	0.1395
1	Good	humidity(%)	-0.6483	0.1701	-3.8104	0.0001	-0.9817	-0.3148	0.5230
2	Good	fine_matter(μg/m³)	-0.4916	0.1413	-3.4784	0.0005	-0.7686	-0.2146	0.6116
3	Good	nitrogen_dioxide(ppb)	-1.9056	0.2096	-9.0904	0.0000	-2.3165	-1.4947	0.1487
4	Good	sulfur_dioxide(ppb)	-1.9939	0.2247	-8.8727	0.0000	-2.4344	-1.5535	0.1362
5	Good	carbon_monoxide(ppm)	-6.2472	0.3089	-20.2240	0.0000	-6.8526	-5.6417	0.0019
6	Good	proximity_to_industrial_areas(km)	2.7232	0.2822	9.6493	0.0000	2.1700	3.2763	15.2284
7	Good	population_density(people/km²)	-0.7424	0.1633	-4.5460	0.0000	-1.0625	-0.4223	0.4760
8	Hazardous	temperature(°C)	1.6965	0.1548	10.9580	0.0000	1.3931	1.9999	5.4548
9	Hazardous	humidity(%)	0.9057	0.1438	6.2991	0.0000	0.6239	1.1875	2.4737
10	Hazardous	fine_matter(μg/m³)	0.3741	0.0982	3.8095	0.0001	0.1816	0.5665	1.4536
11	Hazardous	nitrogen_dioxide(ppb)	1.7565	0.1677	10.4732	0.0000	1.4278	2.0852	5.7920
12	Hazardous	sulfur_dioxide(ppb)	1.5580	0.1478	10.5439	0.0000	1.2683	1.8476	4.7491
13	Hazardous	carbon_monoxide(ppm)	5.0515	0.2415	20.9159	0.0000	4.5782	5.5249	156.2634
14	Hazardous	proximity_to_industrial_areas(km)	-1.9794	0.2329	-8.4984	0.0000	-2.4359	-1.5229	0.1382
15	Hazardous	population_density(people/km²)	0.7456	0.1250	5.9659	0.0000	0.5006	0.9905	2.1076
16	Moderate	temperature(°C)	0.2655	0.1151	2.3062	0.0211	0.0399	0.4912	1.3041
17	Moderate	humidity(%)	-0.2494	0.0976	-2.5569	0.0106	-0.4406	-0.0582	0.7792
18	Moderate	sulfur_dioxide(ppb)	0.4279	0.1189	3.6002	0.0003	0.1949	0.6608	1.5340
19	Moderate	carbon_monoxide(ppm)	1.1882	0.1758	6.7580	0.0000	0.8436	1.5328	3.2811
20	Moderate	proximity_to_industrial_areas(km)	-0.7356	0.1615	-4.5540	0.0000	-1.0522	-0.4190	0.4792