# Classification Clear as Air

Emmanuel Pedernal

MSDS

# ABSTRACT

- Household cooking using wood or coal and Industrial area byproduct causes air pollution
- MLR Model to Predict Air quality with 97% accuracy
- Carbon monoxide (CO) is the strongest feature
- Every km away from industrial areas increases the odds of “Good” air quality by 1423%
- Consider using flexible models like Generalized Additive Models (GAMs) or Tree-based methods.
- Apply polynomial transformations, hyperparameter tuning, and a larger dataset for better generalization.

# Air pollution?

- Air pollution is defined not by the presence of pollutants, but by their concentration and interaction with the environment.

- Nearly **99%** of the population breathes air exceeding recommended pollution levels (*WHO, 1999)*

- Pollution-free environment is a fundamental human right (*UNHR, 2021*)

# Data set composition

- Scraped from WHO and World Bank, includes 5,000 samples
- Particulate matter (PM), nitrogen dioxide (NO2), sulfur dioxide (SO2), and carbon monoxide (CO)
- Temperature, Humidity, Urban density and Industrial Area proximity
- Classified as “Good”, “Moderate”, “Hazardous”

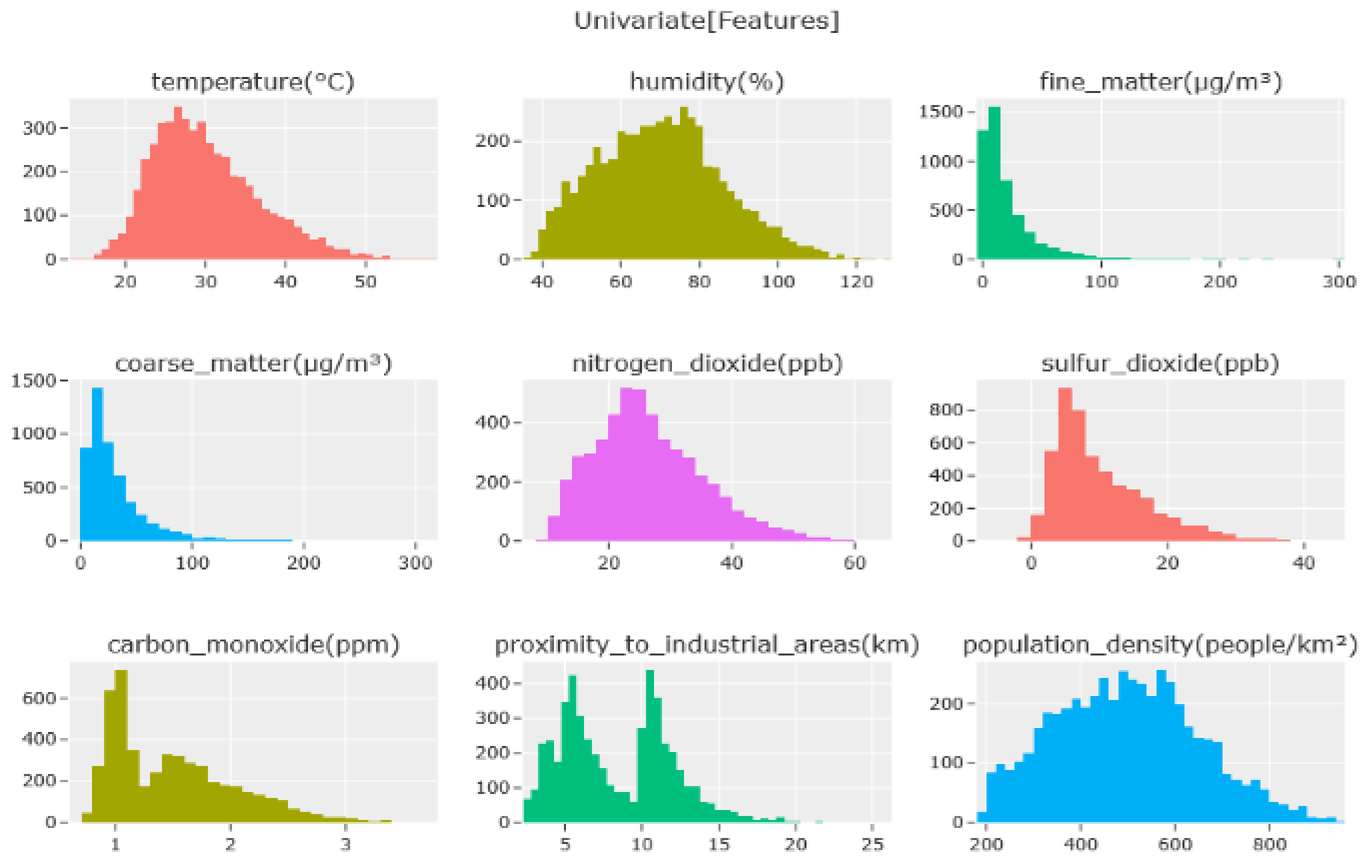
# METHODS

- Combined “Poor” and “Hazardous” classification as “Hazardous”
- JASP (0.19.3.0) and Python 3.12 for Data Analytics and Modeling
- Univariate and Bivariate Analysis
- Checked for Multinomial Logistic Regression (MLR) assumptions
- MLR with Elastic net regularization and 5-fold cross validation

# Results

## Univariate

- Data are gathered from warm region
- Mean Humidity level are HIGH
- Most PM levels are in “Good” Standing
- Nitrogen Dioxide are Moderate risk
- Carbon Monoxide are normal levels
- Near Industrial Zone and average of 500 people per km^2

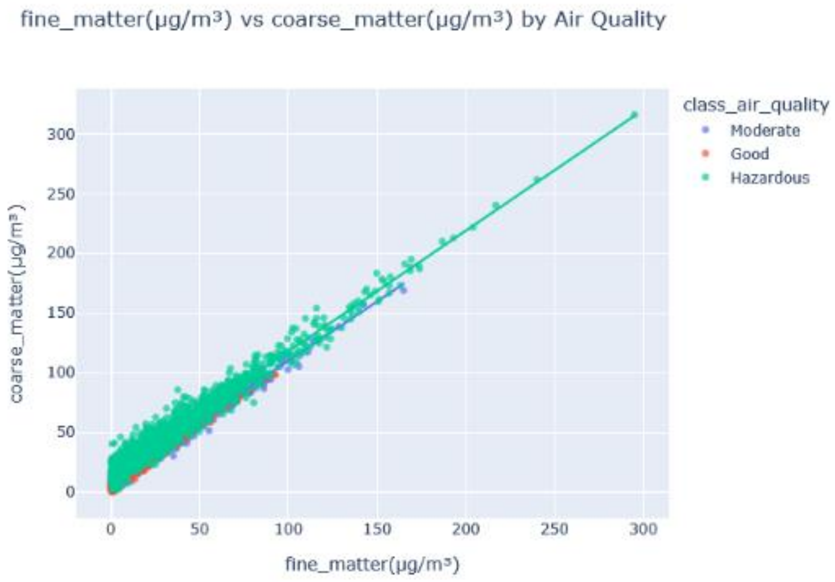


# Bivariate

- Fine_matter and coarse_matter have a high positive correlation (r = 0.973), consistent with findings from (*Chen et al. 2018)*, linking $PM_{2.5}$ and $PM_{10}$ to increased mortality.

- Proximity_to_industrial_areas shows a strong negative correlation with carbon_monoxide (r = -0.7), supporting the expectation that closer proximity to industrial zones results in higher air pollution risks.

- Population_density has a moderate correlation with carbon_monoxide (r = 0.59), indicating that more densely populated areas may experience higher CO levels due to household heating, vehicles, and combustion sources.



# Assumptions Results

- The Air quality Dataset Passed the assumption test
- Having Class > 2

**Table 2. Classification and Corresponding Values**

| Classification | Value (n) |
|---|---|
| Good | 2000 |
| Moderate | 1500 |
| Hazardous | 1500 |

# Independence of Observations

- All observations are independent, no rows are duplicated

## 2. Independence of Observation [PASSED]

```python
df.shape
```

```
(5000, 10)
```

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

```
0
```

# Sufficient Sample Size

- The Dataset needs at least 540 observations from (features) 9 x (classes) 3 * 10 – 20
- While the Air Quality Dataset supports 5000 samples

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

```python
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
```

# Variance Inflation Factor

## VIF (Before)

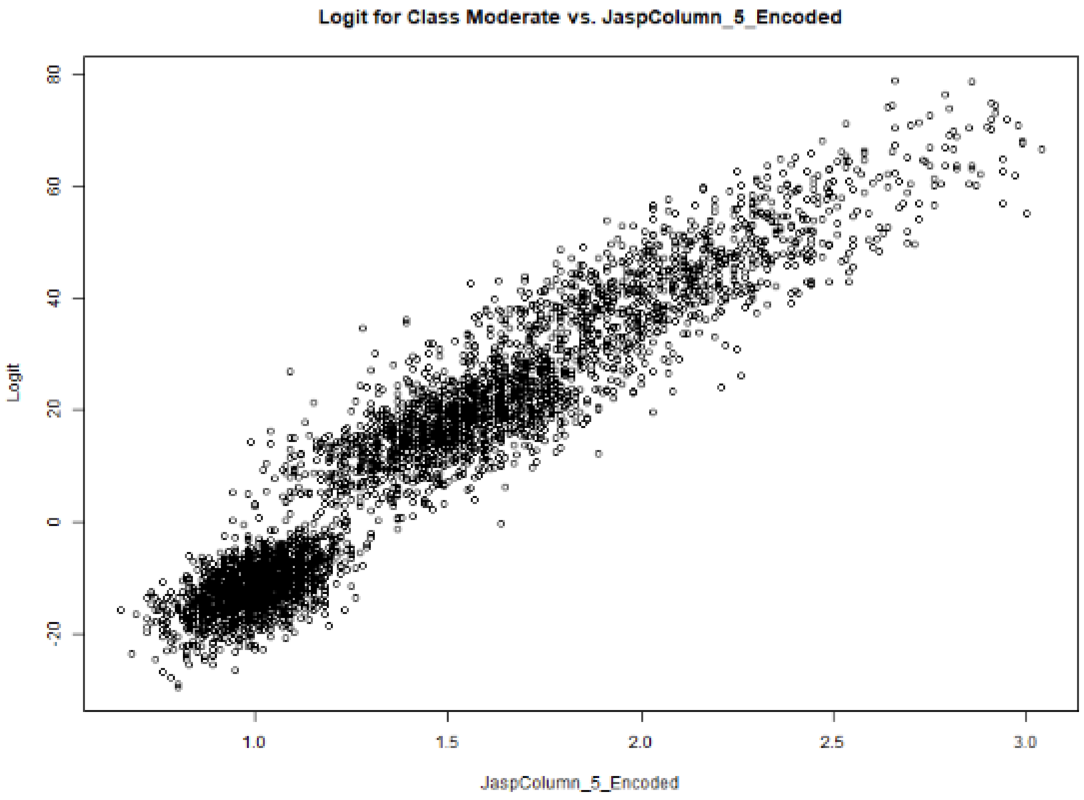
| | Feature | VIF |
|---|---|---|
| 0 | temperature(°C) | 2.108506 |
| 1 | humidity(%) | 1.566619 |
| 2 | fine_matter(µg/m³) | 29.401193 |
| 3 | coarse_matter(µg/m³) | 34.566851 |
| 4 | nitrogen_dioxide(ppb) | 2.287701 |
| 5 | sulfur_dioxide(ppb) | 2.029253 |
| 6 | carbon_monoxide(ppm) | 3.913178 |
|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

## VIF (After)

| | 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 |

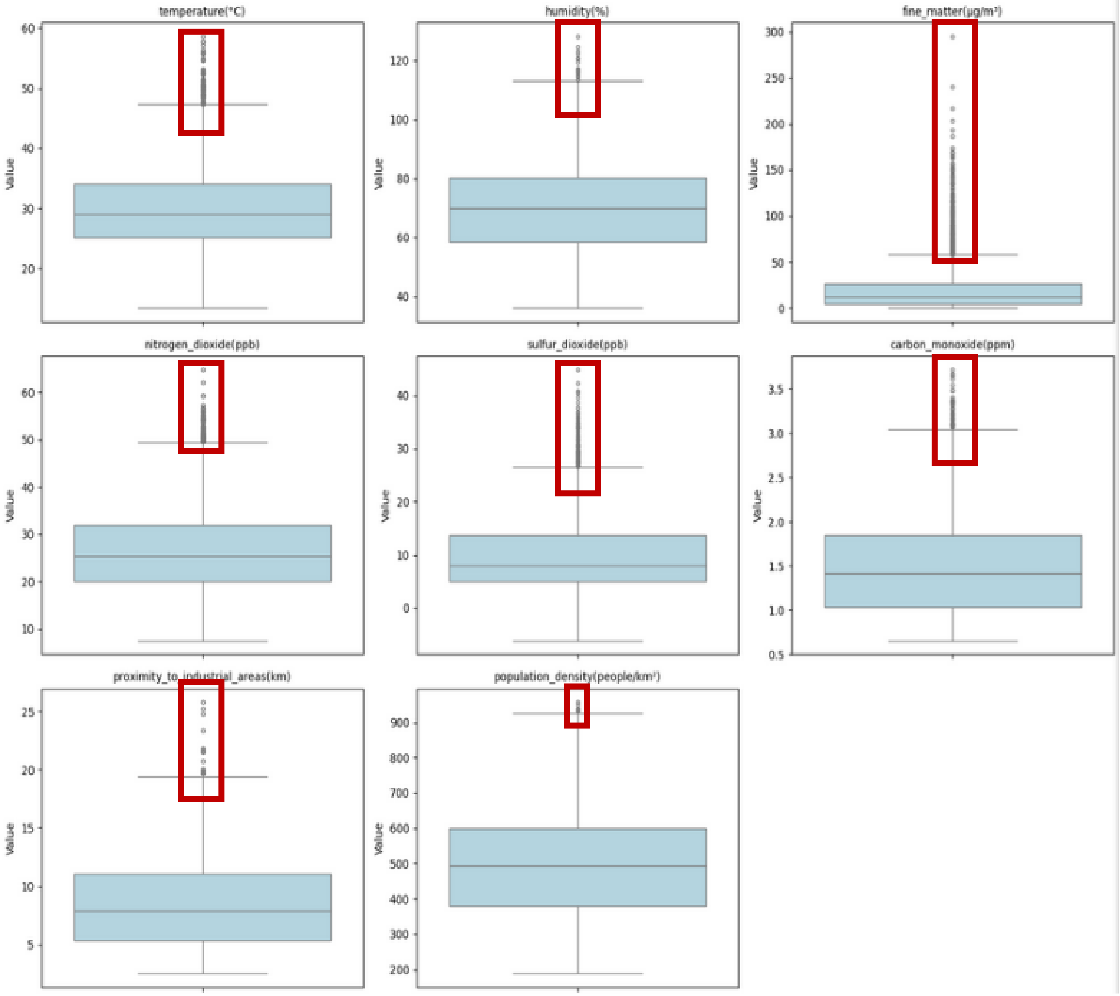
# Linearity of Logits

- All features shows linearity while only fine_matter was not perfectly aligned with, which is not substantial to do feature transformation

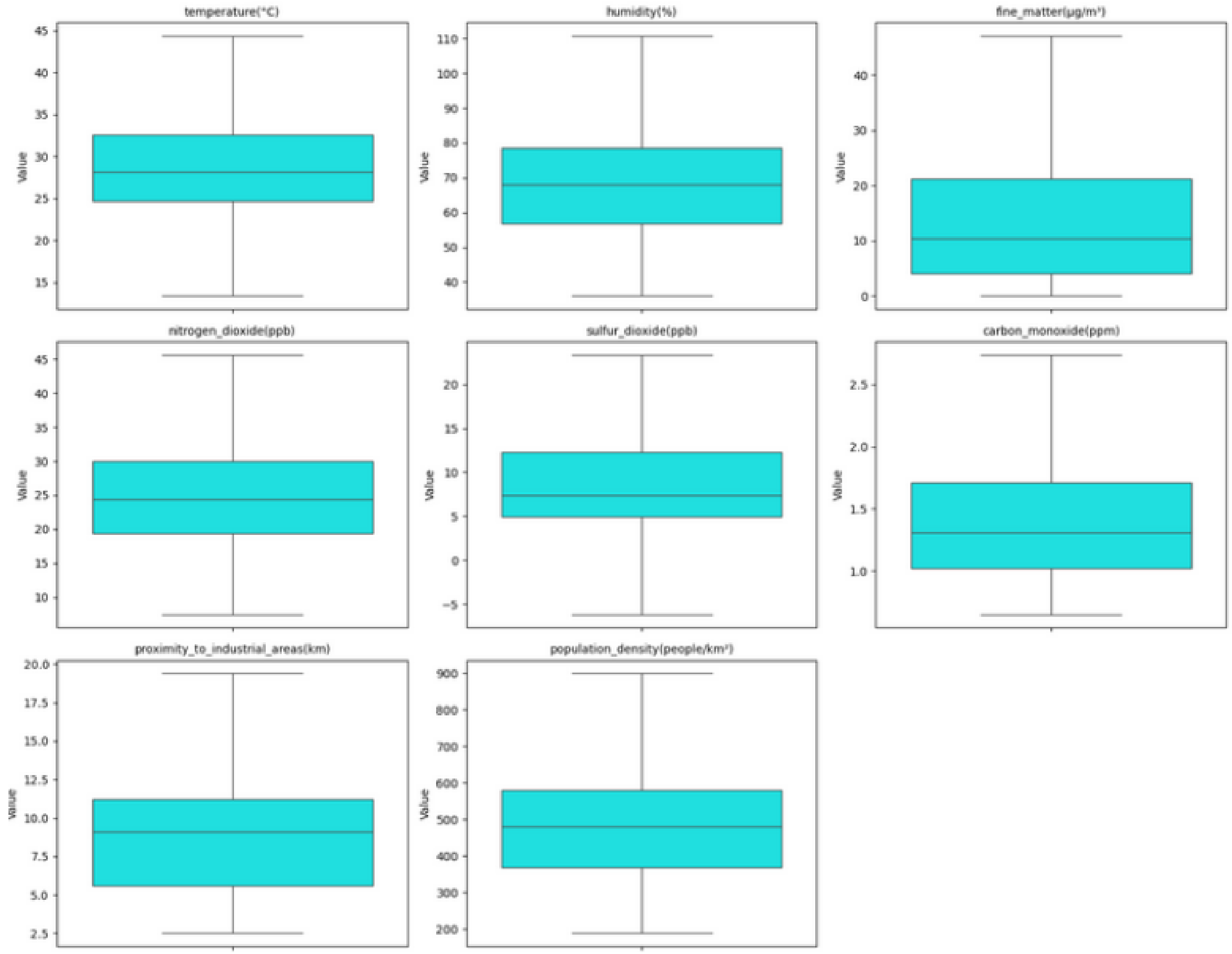


# Outliers

## Before




## After



# Independence of Irrelevant Alternatives (IIA)

- Only Test that failed
- Checked for weight changes if one class is dropped

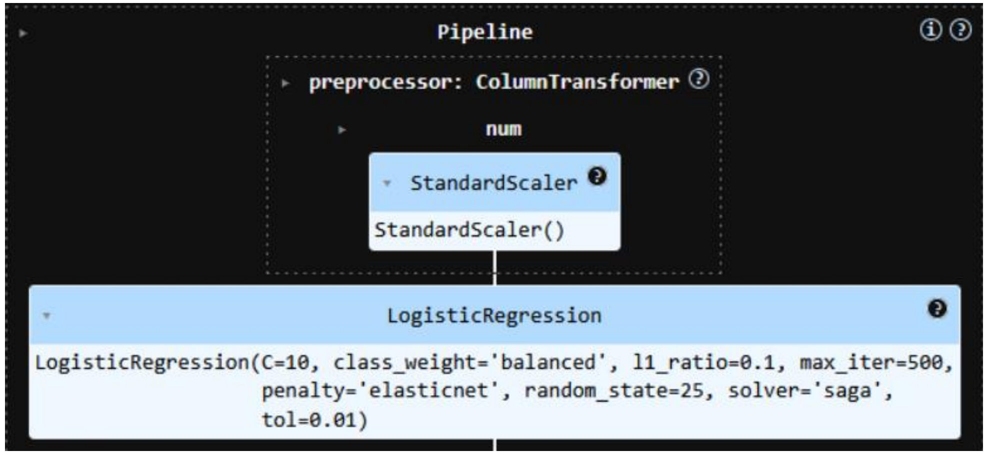
### 7. Independence of Irrelevant Alternatives Using R through JASP [Failed]

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
```

# Model

- LabelEncoder() on Classes
- StandardScaler()
  - Excluding Class
- 70/20/10 Train, Validation, Test split
  - Stratified sampling
- Grid Search Results
  - C =10
  - Tol = 0.01
  - L1_ratio = 0.1
  - Penalty = elasticnet
  - Solver='saga'

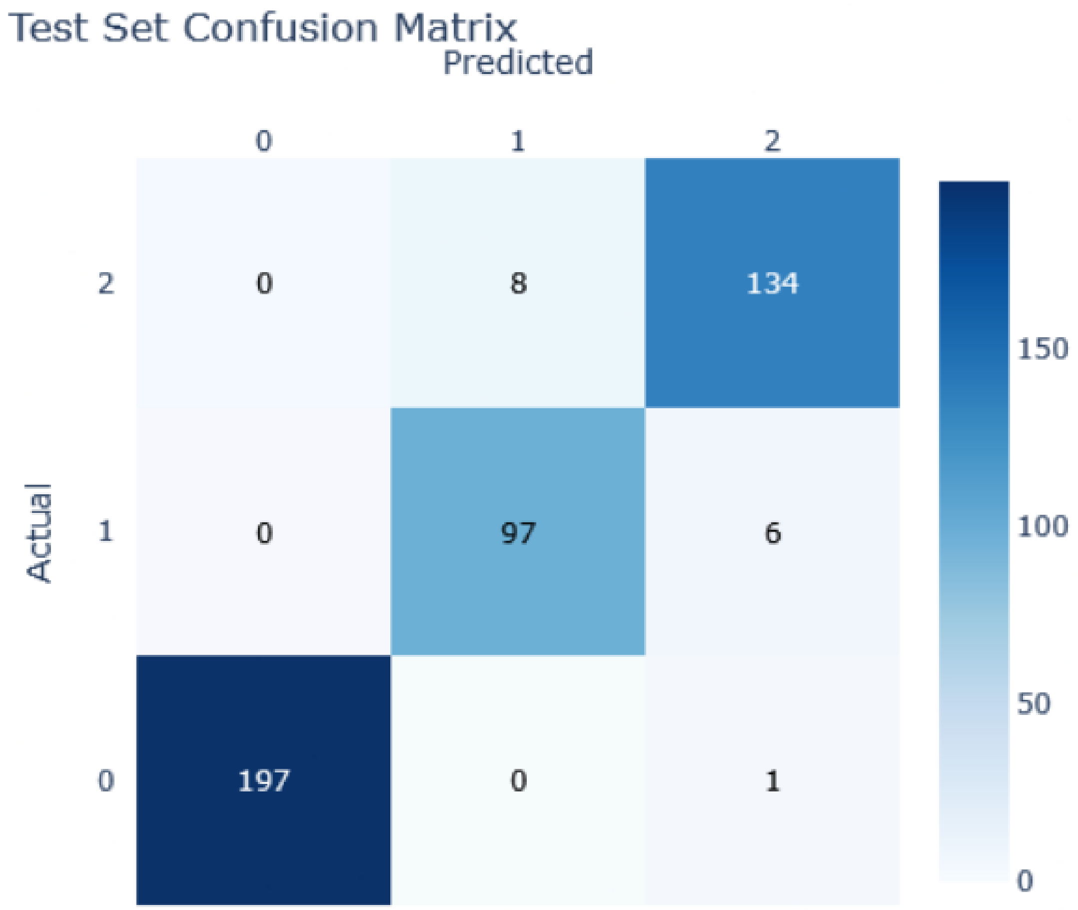


# Model Classification Report

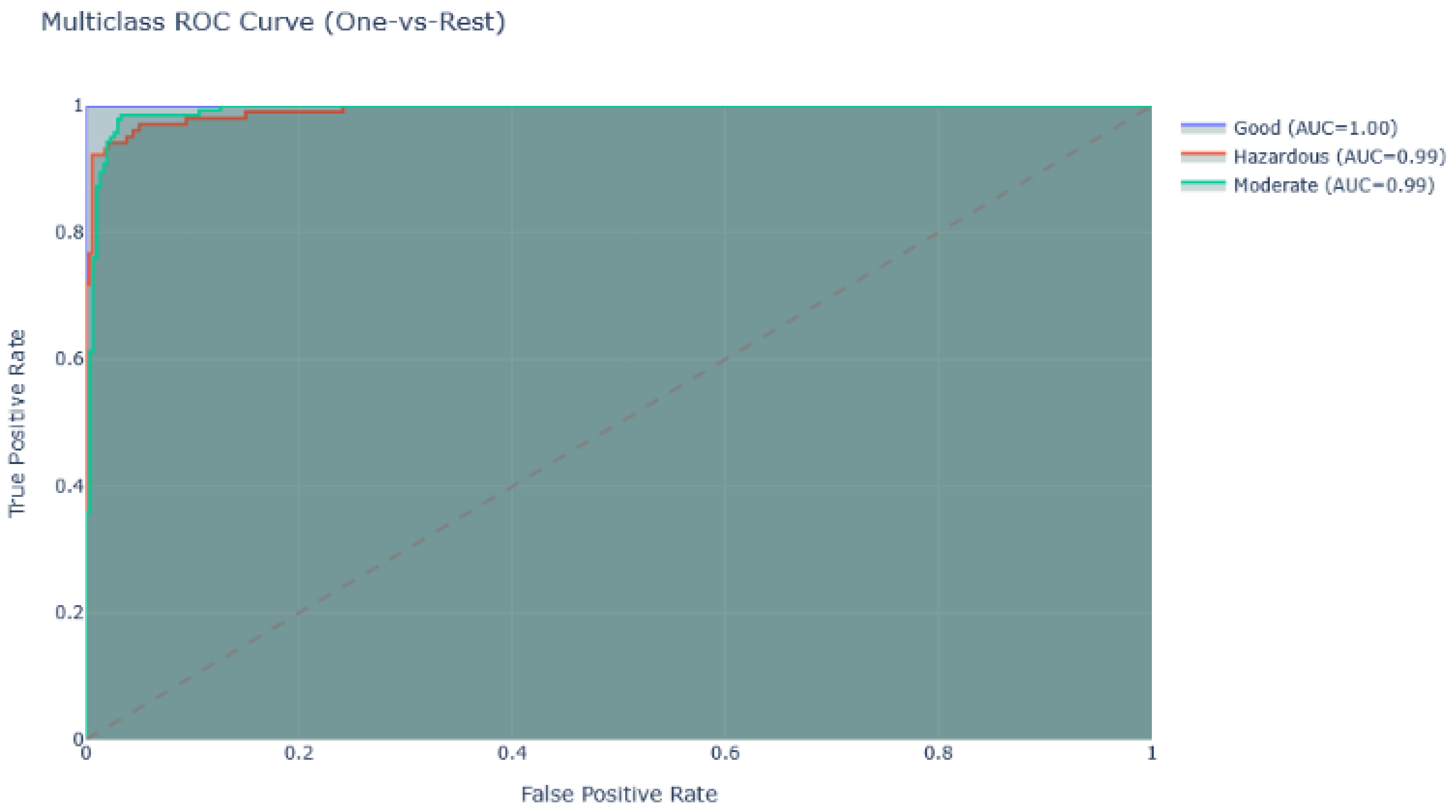
Test Classification Report:

| | 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 |

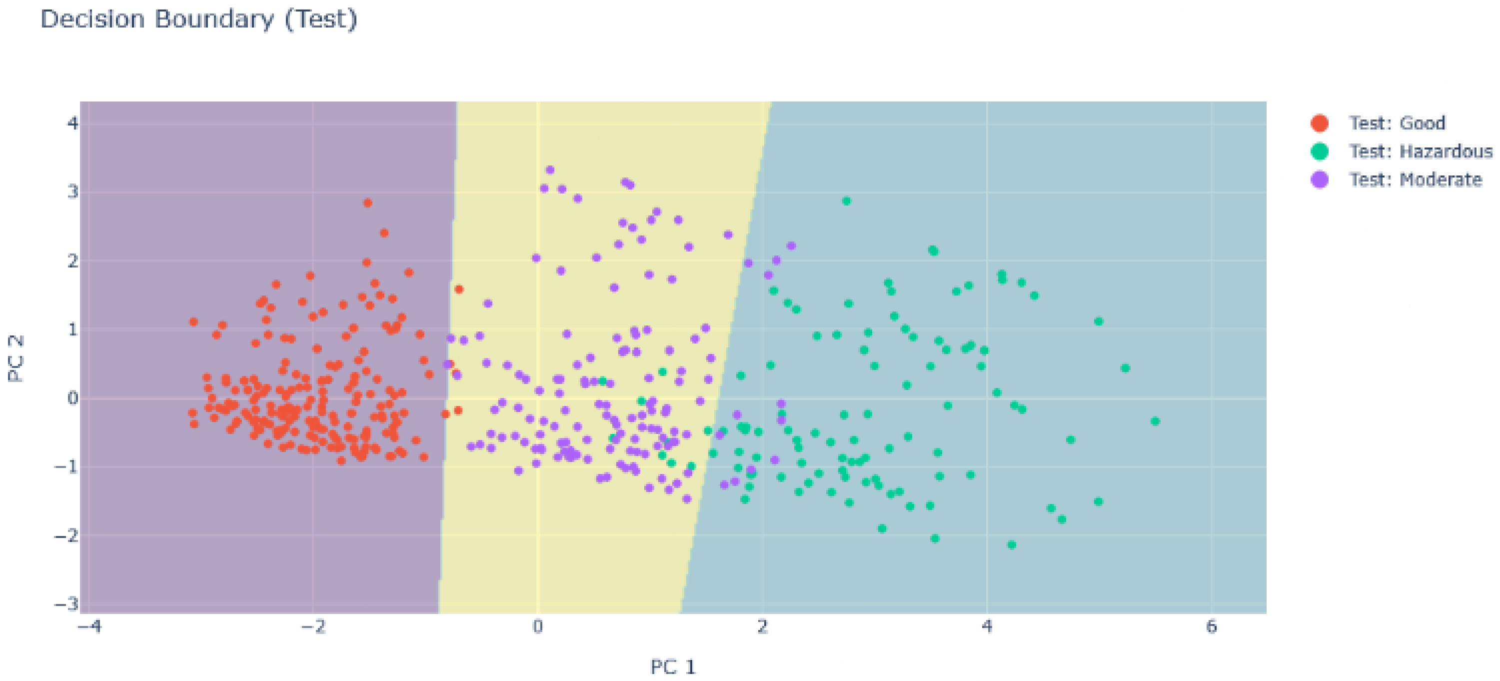
# Confusion Matrix



# ROC & AUC


Multiclass ROC Curve (One-vs-Rest)
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1
Good (AUC=1.00)
Hazardous (AUC=0.99)
Moderate (AUC=0.99)

# Decision Boundaries



# Odds Ratio (Good)

- Temperature increase reduces the likelihood of “Good” air quality by 86.05%.
- Increases in humidity, fine particulate matter (PM2.5), nitrogen dioxide, and sulfur dioxide decrease the odds of “Good” air quality by 47.7%, 38.84%, 85.13%, and 86.38%, respectively.
- Carbon monoxide reduces the odds by 98.1% increases the odds of “Good” air quality by **1422.84%**.
- Population density negatively impacts air quality, reducing the odds of “Good” air by 52.4%.

| | 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 |

# Odds Ratio (Moderate)

- Temperature increase raises the odds of “Moderate” air quality by 30.41%, while humidity increase decreases the odds by 22.08%.

- Sulfur dioxide increase boosts the odds by 53.40%, and proximity to industrial areas reduces the odds by 52.08%

- Carbon monoxide has a strong association with “Moderate” air quality, increasing the odds by **228.11%**

| | Class | Feature | Coef | SE | z | p | CI_lower | CI_upper | OddsRatio |
|---|---|---|---|---|---|---|---|---|---|
| 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 | $\boxed{3.2811}$ |
| 20 | Moderate | proximity_to_industrial_areas(km) | -0.7356 | 0.1615 | -4.5540 | 0.0000 | -1.0522 | -0.4190 | 0.4792 |

# Odds Ratio (Hazardous)

- Temperature increase raises the odds of "Hazardous" air quality by 445.48%, while humidity increase raises the odds by 147.37%, and fine particulate matter increases the odds by 45.36%.

- Carbon monoxide has a massive impact, with an increase in ppm levels leading to a **14,626.34%**

- Proximity to industrial areas decreases the odds of hazardous air quality by 86.18% with increased distance, while higher population density raises the odds by 110.76%.

| | Class | Feature | Coef | SE | z | p | CI_lower | CI_upper | OddsRatio |
|---|---|---|---|---|---|---|---|---|---|
| 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 |

# Sample

| Feature | Value | Scaled Value |
|---|---|---|
| temperature(°C) | 41.7 | 0.230276 |
| humidity(%) | 82.5 | -0.330412 |
| fine_matter(μg/m³) | 1.7 | 0.768669 |
| nitrogen_dioxide(ppb) | 31.1 | 0.804637 |
| sulfur_dioxide(ppb) | 12.7 | -0.018407 |
| carbon_monoxide(ppm) | 1.8 | 0.970496 |
| proximity_to_industrial_areas(km) | 4.6 | -0.978790 |
| population_density(people/km²) | 735 | 0.360127 |

# Logits (Good)

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

$$\eta Good$$

=(−1.9694)(0.230276)=−0.45339
+(−0.6483)(−0.330412)=0.21420
+(−0.4916)(0.768669)=−0.37792
+(−1.9056)(0.804637)=−1.53347
+(−1.9939)(−0.018407)=0.03672
+(−6.2472)(0.970496)=−6.06185
+(2.7232)(−0.978790)=−2.66561
+(−0.7424)(0.360127)=−0.26734
=**−11.1087**

# Logits (Moderate)

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

ηModerate

=(0.2655)(0.230276)=0.06115

+(−0.2494)(−0.330412)=0.08239

+(0.4279)(0.768669)=0.32884

+(0.1178)(0.804637)=0.09479

+(0.1775)(−0.018407)=−0.00327

+(0.0466)(0.970496)=0.04524

+(−0.4884)(−0.978790)=0.47777

+(−0.0789)(0.360127)=−0.02841=**1.0587**

# Logits(Hazardous)

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

ηHazardous

=(1.6965)(0.230276)=0.39069

+(0.9057)(−0.330412)=−0.29920

+(0.3741)(0.768669)=0.28755

+(1.7565)(0.804637)=1.41314

+(1.5580)(−0.018407)=−0.02868

+(5.0515)(0.970496)=4.90505

+(−1.9794)(−0.978790)=1.93627

+(0.7456)(0.360127)=+0.26847

**=8.8732**

# Exponentiate

$$odds_{Good} = e^{-19.9819} \approx 2.09\times10^{-9}$$

$$odds_{Moderate} = e^{-7.8145} \approx 0.000402$$

$$odds_{Hazardous} = e^{0} = 1$$

$$\mathrm{Z} = 2.09\times10^{-9} + 0.000402 + 1 \approx \mathbf{1.000402}$$

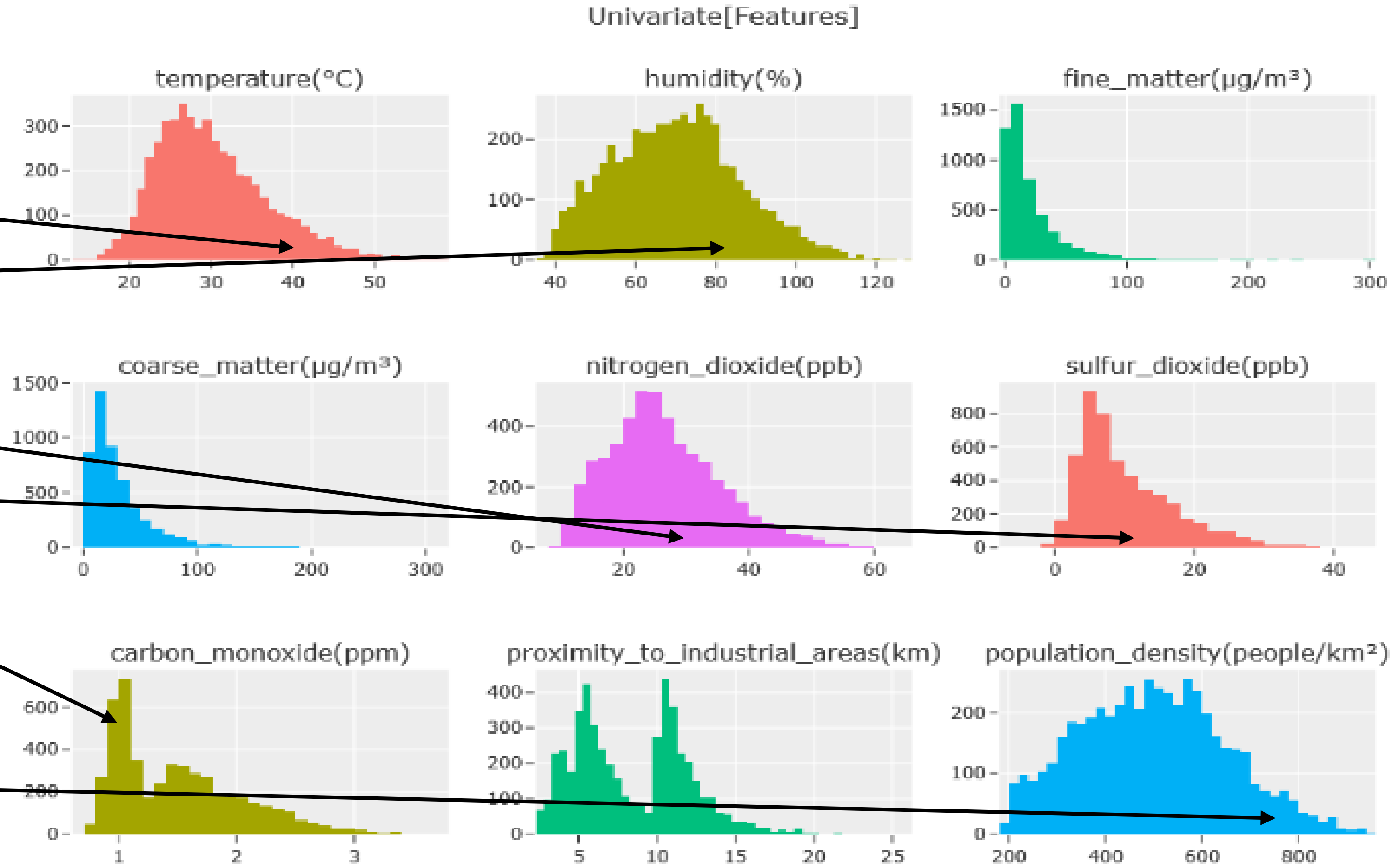
# SoftMax

$$P(Good) = \frac{2.09 \;*\; 10^{-9}}{1.000402} \approx 0.00000000209$$

$$P(Moderate) = \frac{0.000402}{1.000402} \approx 0.0004019$$

$$P(Hazardous) = \frac{1}{1.000402} \approx \mathbf{0.999598}$$

| Feature | Value |
|---|---|
| temperature(°C) | 41.7 |
| humidity(%) | 82.5 |
| fine_matter(μg/m³) | 1.7 |
| nitrogen_dioxide(ppb) | 31.1 |
| sulfur_dioxide(ppb) | 12.7 |
| carbon_monoxide | 1.8 |
| proximity_to_industrial_areas(km) | 4.6 |
| population_density(people/km²) | 735 |



# Conclusion

- **Model:** Multinomial Logistic Regression to classify air quality into Good, Moderate, and Hazardous.
- **Data:** 5,000 observations, assumptions mostly met except for Independence of Irrelevant Alternatives (IIA).
- **Findings:** Moderate populations, warm and humid areas likely have good air quality, while proximity to industrial areas and high population density worsen air quality.
- **Performance:** Model accuracy of 0.97 and F1-score of 0.96, with "Good" class having the highest accuracy.
- **Key Factors:** Distance from industrial zones, lower carbon monoxide and sulfur dioxide levels linked to "Good" air quality; higher carbon monoxide, nitrogen dioxide, and temperature linked to "Hazardous" air quality.
- Model classifies well, but with IIA assumption violation suggests a need for more flexible model approaches.

# Future Work

- Explore tree-based classifiers (Decision Trees, Random Forest, Gradient Boosting) to capture nonlinear relationships better than MLR.
- Address IIA violation with flexible models like nested logits, mixed logits, or Generalized Additive Models (GAMs) for nonlinear feature relationships.
- Use a larger, more diverse dataset to reduce overfitting and class imbalance and enhance the MLR model with feature transformations, Elastic Net fine-tuning, or Bayesian approaches for better stability and interpretability.

Full breakdown with code, reference and paper kindly visit

(Classification Clear as Air)

github.com/PedGit025