Practical Technique in Data science
Coursework - Regression and classification

Vaamini Mathimaran

April 2025

1 Abstract

This report explores how regression and classification techniques can be applied to two real-world datasets. The first involves predicting pollutant levels in Delhi using linear regression. The second applies classification models to detect pulsar stars based on signal features. Results showed that PM10 is an effective predictor of PM2.5, and Logistic Regression performed best in identifying pulsars. These findings demonstrate the usefulness of data science in environmental monitoring and astrophysics.

2 Introduction

This report explores how regression and classification techniques can be applied to real-world datasets. Two datasets were used: one focused on air pollution in Delhi, and the other on pulsar star classification. For the Delhi dataset, simple and multiple linear regression were used to predict pollutant levels based on other pollutants. This is useful in real-world scenarios such as sensor redundancy systems, where missing sensor values can be estimated using available data.
For the pulsar dataset, three classification models — Naive Bayes, Linear Discriminant Analysis (LDA), and Logistic Regression — were used to classify neutron stars based on eight input features. This approach supports automated discovery in astronomy by helping identify real pulsars from large amounts of signal data.

3 Delhi Dataset

The first dataset is a daily air quality dataset from Delhi. It includes nine columns: one for the date and eight for different pollutants — PM2.5, PM10, CO, NO, NO2, NH3, SO2, and O3.
This data is important because air pollution is a serious health risk, especially in heavily polluted cities like Delhi. Monitoring pollutant levels supports public health responses, environmental planning, and sensor redundancy systems.
The goal for this dataset was to use regression models to predict the level of one pollutant using the others.
Two main approaches were used:

- Simple linear regression to test whether PM10 could be predicted using PM2.5
- Multiple linear regression to predict CO using all other pollutants

The method involved first running a correlation analysis to select useful predictors. The models were evaluated using R-squared and adjusted R-squared values, as well as visual inspection of predicted vs actual plots.

3.1 Dataset Description

The Delhi air quality dataset contains hourly data collected between 25/11/2020 and 24/01/2023. It includes a total of 18,776 observations with no missing values for any of the pollutants. The dataset has a shape of 18,776 rows $\times$ 9 columns, with one column for the date and eight columns representing pollutant levels: PM2.5, PM10, CO, NO, NO2, NH3, SO2, and O3. All pollutant concentrations are measured in $\mu g/m^3$. A summary of the dataset's statistical properties is shown in Table 1.

Pollutant	Mean	Std Dev	Min	Max
PM2.5	238.13	226.53	11.87	1708.09
PM10	300.09	247.16	15.66	1788.93
CO	2020.23	2854.52	200.35	21148.68
NO	33.66	62.13	0.0	500.68
NO_2	66.22	48.35	4.28	400.66
SO_2	6.60	49.44	5.25	579.83
NH_3	25.11	21.64	0.0	208.77
O_3	50.80	80.46	0.0	801.09

Table 1: Summary statistics for pollutants in the Delhi dataset

Table 1 shows clear differences in the distribution and variability of each pollutant.
CO has the highest mean and the largest standard deviation, which may reflect high traffic volumes or industrial emissions that fluctuate over time. Both PM2.5 and PM10 also have large maximum values and wide ranges, suggesting frequent spikes in fine particulate matter — a major health risk, particularly for individuals with respiratory conditions.
In contrast, gases such as NH3 (ammonia) and SO2 (sulfur dioxide) have lower mean values and narrower distributions. While their concentrations are lower, they remain important due to their roles in acid rain formation and agricultural impact.
NO and NO2 are nitrogen oxides commonly produced by combustion processes, especially from vehicle emissions, and are key indicators of urban air pollution.
Finally, O3 (ozone) levels vary significantly, often peaking due to photochemical reactions — chemical reactions triggered by sunlight. These peaks typically occur in the afternoon when sunlight is strongest.
Understanding the behaviour and spread of these pollutants is essential for developing accurate regression models and determining which variables are most suitable as predictors.

3.2 Correlation Analysis

From the dataset we can see that all 8 pollutant are somewhat correlated but to see how a correlation heatmap was produced to see the relationship between them.
The dataset shows that all eight pollutants are somewhat correlated, but to understand the strength and nature of these relationships more clearly, a correlation heatmap was produced.
The heatmap (Figure 1) visualizes the linear relationships between pollutants. A strong correlation is indicated by values close to 1 (positive) or –1 (negative), suggesting that one variable may be predicted using another. Figure 1 shows that CO is strongly correlated with PM2.5 (0.94), PM10 (0.95), and NO (0.91). This is expected, as these pollutants are typically co-emitted from sources such as traffic and industrial combustion. Their strong correlation makes them good candidates for predicting CO.
PM2.5 and PM10 also have a very high correlation (0.99). Since they are both forms of particulate matter, this is not surprising. However, because they are so similar, including both in a regression model may cause multicollinearity — a condition where the model struggles to distinguish the effect of each variable, which can lead to overfitting or unstable predictions.
In contrast, O3 (ozone) shows weak or even negative correlations with most other pollutants. This is also expected, as ozone is formed through photochemical reactions involving sunlight, which typically occur at times when other pollutants (like NO and CO) are lower, such as mid-afternoon rather than during traffic peaks.

3.3 Simple Linear Regression

A simple linear regression model was used to investigate whether PM10 can be used to predict PM2.5 concentrations. This is based on the fact that PM2.5 consists of finer particles, which are harder to detect directly, and are often estimated from PM10 sensors. Both pollutants are highly correlated, as they originate from similar sources and both represent forms of particulate matter. Therefore, PM10 was chosen as the predictor variable.
Table 2 presents the key regression metrics obtained from the model, including the coefficient, significance level (p-value), confidence interval, and overall model fit. The model shows a very strong linear relationship between PM10 and PM2.5, with an R-squared value of 0.9787. This means that 97.87% of the variation in PM2.5 can be explained by PM10. The coefficient for PM10 is 0.8388, showing that for

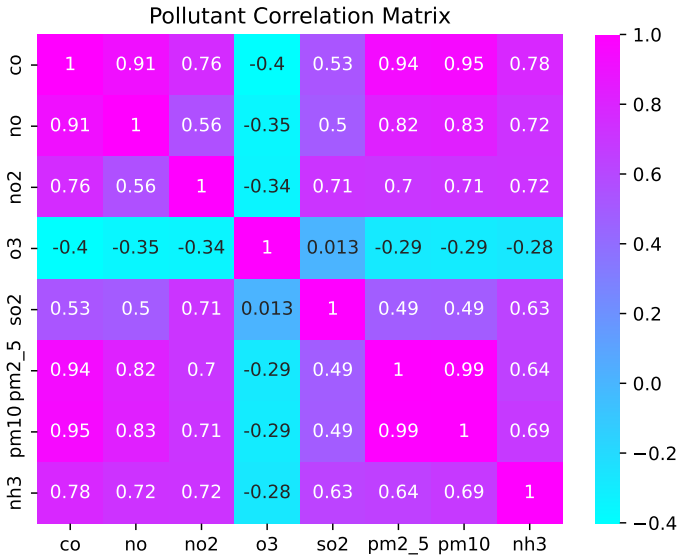


Figure 1: Pollutants Correlation Heatmap

Metric	Coefficient	p-value	95% CI Lower	95% CI Upper
PM10	0.8388	0.0000	0.8371	0.8406
R-squared	0.9787	–	–	–
Adjusted R-squared	0.9787	–	–	–

Table 2: Simple linear regression results predicting PM2.5 using PM10

every 1 unit increase in PM10, PM2.5 increases by approximately 0.84 units on average. The p-value is effectively zero, confirming that PM10 is a statistically significant predictor of PM2.5.
The adjusted R-squared is also 0.9787, indicating that the model maintains a strong fit without overfitting. The 95% confidence interval for the PM10 coefficient ranges from 0.8371 to 0.8406, showing a high level of precision in the estimate. These results support the assumption that the two pollutants are closely related in their behavior and that PM10 can be a reliable estimate for PM2.5 in cases where direct sensor data may be unavailable.
After fitting the model, predicted PM2.5 values were obtained using the regression equation. These predicted values were then compared with the actual PM2.5 values from the dataset.
Figure 2 shows a scatter plot comparing the predicted PM2.5 values from the regression model to the actual values from the dataset. Each point represents an hourly observation, and the diagonal line serves as a reference indicating perfect prediction — where predicted values equal actual ones.
The points are tightly clustered around the reference line, indicating a strong agreement between predicted and actual values. This visual alignment supports the model's high accuracy and corresponds with the previously reported R-squared value of 0.9787, confirming that the model captures nearly all variation in PM2.5 using only PM10 as a predictor.
There are no visible patterns, major deviations, or significant outliers, which further suggests that the model assumptions — such as linearity and homoscedasticity (when the variance of the residual is constant) — are met.
Overall, this graph reinforces the conclusion that PM10 is a reliable and effective predictor of PM2.5, making it particularly valuable in situations where PM2.5 sensor data is unavailable, incomplete, or too difficult to detect due to the smaller size of the particles compared to PM10.

3.4 Multiple Linear Regression

A multiple linear regression model was used to predict CO concentrations based on the other pollutants in the dataset. CO was selected as the dependent variable due to its strong correlation with several other pollutants, as shown in the heatmap, and because of its relevance in air quality monitoring. Being able

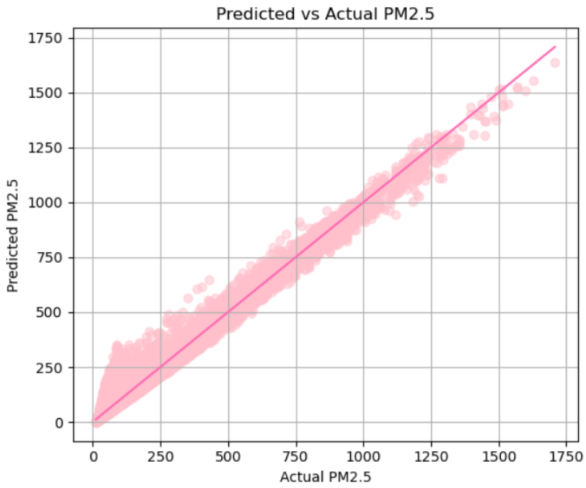


Figure 2: Predicted vs Actual Values of PM2.5

to estimate CO using other pollutants is valuable in cases where CO data is missing or the sensor fails. The initial model included all other pollutants: PM2.5, PM10, NO, NO2, NH3, SO2, and O3. Table 3 presents the regression metrics for this full model. All variables were statistically significant, with p-values below 0.05. The model achieved an R-squared value of 0.982, indicating that 98.2% of the variation in CO levels could be explained by the other variables. The adjusted R-squared was also 0.982, confirming a strong overall model fit.

Variable	**Coefficient**	**p-value**	**95% CI Lower**	**95% CI Upper**
NO	18.3717	0.0000	18.173	18.570
NO_2	13.1684	0.0000	12.918	13.419
O_3	-1.0327	0.0000	-1.118	-0.947
SO_2	-5.5642	0.0000	-5.756	-5.373
PM2.5	1.6229	0.0000	1.443	1.803
PM10	3.2663	0.0000	3.106	3.426
NH_3	10.1942	0.0000	9.797	10.591
R-squared	0.982	–	–	–
Adjusted R-squared	0.982	–	–	–

Table 3: Multiple linear regression results for predicting CO using all other pollutants

Figure 3 shows the predicted vs actual CO values. The points are closely clustered along the diagonal, suggesting that the model accurately predicts CO levels. The corresponding residual plot (Figure 4) shows no visible patterns, confirming that the assumptions of linearity is satisfied.

While the full model achieved a very high R-squared of 0.982, this may indicate overfitting, especially given the number of predictors used. Overfitting occurs when a model captures not only the underlying relationship but also small fluctuations or noise that do not generalise well to new data. In this case, although all variables were statistically significant, some like O3, SO2, and NH3 contributed minimally and had weaker physical relevance. In addition, the regression summary reported a high condition number (1.2e+03), suggesting the presence of multicollinearity, where highly correlated predictors can make the model unstable or harder to interpret. These issues made the full model more complex than necessary, without a substantial gain in accuracy.

To improve model simplicity and reduce multicollinearity, backward elimination was applied. Variables were removed based on a combination of weak correlation with CO, high standard error, and redundancy due to strong correlation with other predictors. O3 and SO2 were removed due to weak correlation with CO. PM10 was dropped to reduce multicollinearity with PM2.5, as both were highly correlated (0.99). NH3 was excluded due to having the highest standard error in the model.

The reduced model included only PM2.5, NO, and NO2 — pollutants that are strongly correlated with CO and are also physically linked to combustion and vehicle emissions. Despite using fewer variables, the reduced model achieved an R-squared of 0.970, with improved stability, reduced multicollinearity, and

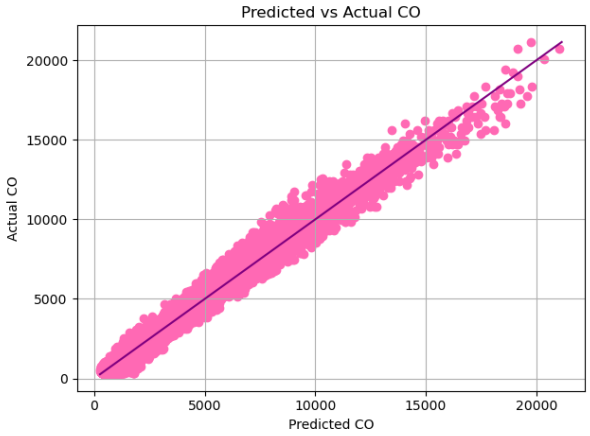


Figure 3: Predicted vs Actual CO values (Full Model)

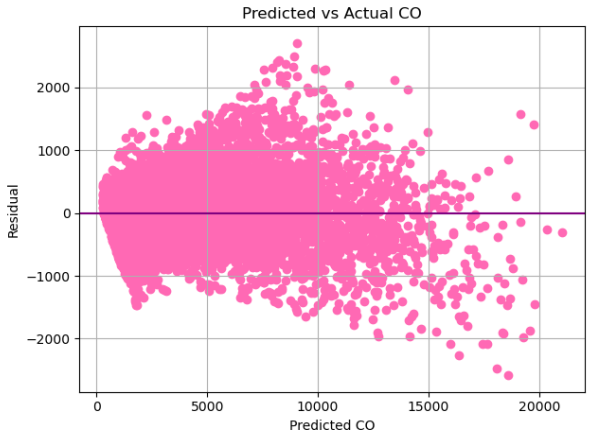


Figure 4: Residual Plot for CO (Full Model)

lower risk of overfitting.

Variable	**Coefficient**	**p-value**	**95% CI Lower**	**95% CI Upper**
NO	21.0411	0.0000	20.844	21.239
NO_2	13.1809	0.0000	12.976	13.386
PM2.5	5.1119	0.0000	5.049	5.175
R-squared	0.970	–	–	–
Adjusted R-squared	0.970	–	–	–

Table 4: Reduced multiple regression model after backward elimination

Table 4 shows the final regression metrics after backward elimination. All three remaining predictors—PM2.5, NO, and NO2 —had very small p-values, confirming their statistical significance in the model. The coefficients were also much higher compared to some of the previously removed variables, showing that these three variables had the strongest individual relationships with CO. The 95% confidence intervals were also narrow, suggesting precise estimates for each coefficient.

The R-squared and adjusted R-squared values were both 0.970, meaning the reduced model could still explain 97% of the variation in CO levels using just three variables. Compared to the full model (which had an R-squared of 0.982), this is a very small drop in performance, while gaining a major improvement in model simplicity, interpretability, and stability.

Figure 5 shows the predicted vs actual CO values for the reduced model. The points remain tightly clustered along the diagonal, indicating the model continues to make highly accurate predictions. The residual plot (Figure 6) shows no clear pattern or heteroscedasticity, and residuals are evenly spread around zero, supporting that the model still satisfies key linear regression assumptions.

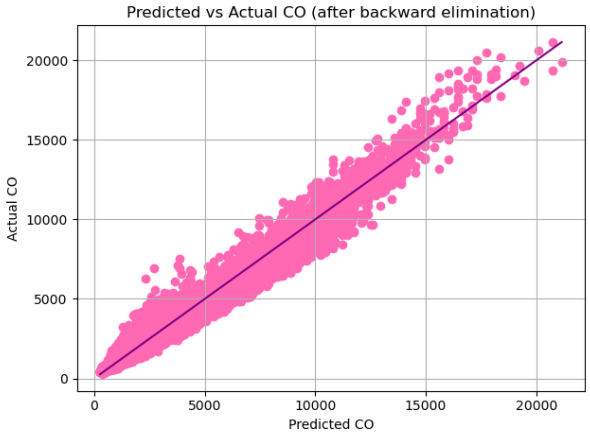


Figure 5: Predicted vs Actual CO values (Reduced Model)

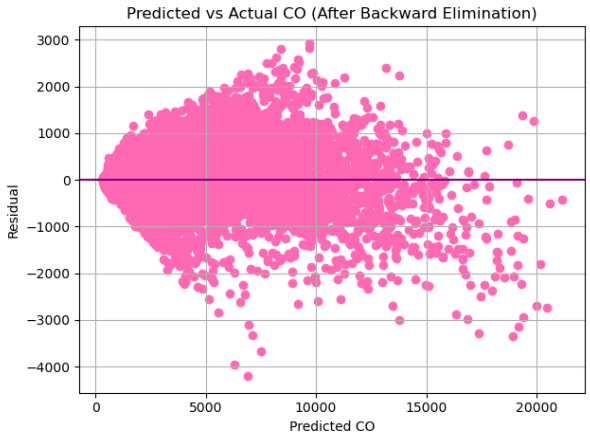


Figure 6: Residual Plot for CO (Reduced Model)

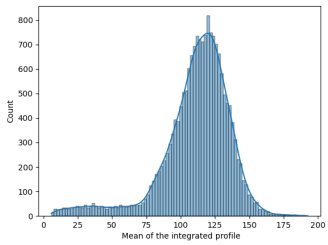
Overall, the reduced model is more efficient and just as reliable, avoiding the risks of overfitting and multicollinearity that were present in the full model. PM2.5, NO, and NO2 not only performed well statistically but also made sense in terms of physical sources, as they are all directly linked to traffic-related emissions. This makes the reduced model more robust and applicable to real-world prediction tasks where fewer, more reliable variables are preferred.This makes the reduced model not only more interpretable and stable, but also more practical for real-world applications where collecting data on every pollutant might not be feasible.

4 Pulsar Star Dataset

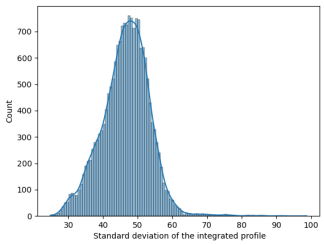
The second dataset focuses on the classification of pulsar stars. It contains 17,898 observations and 9 columns, with 8 continuous numerical features derived from integrated profile and DM-SNR curves, and one binary target column indicating whether the object is a pulsar star (1) or not (0). Pulsars are highly magnetised, rotating neutron stars that emit beams of electromagnetic radiation. Correctly identifying pulsars is important in astrophysics, as they are used in studying gravitational waves, testing theories of gravity, and exploring extreme states of matter.
The aim of this analysis was to classify whether an observation corresponds to a pulsar or a non-pulsar using the eight feature columns. Three supervised classification algorithms were used: Naive Bayes, Linear Discriminant Analysis (LDA), and Logistic Regression. Model performance was assessed using confusion matrices, ROC curves, and AUC values, and compared to determine which model gave the most accurate and reliable predictions.

4.1 Data Exploration

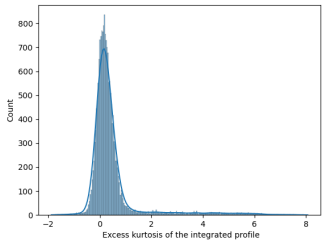
The pulsar dataset contains eight numerical features derived from two signal sources: the integrated pulse profile and the DM-SNR (Dispersion Measure–Signal-to-Noise Ratio) curve. Each of these is summarised by four statistical metrics: mean, standard deviation, skewness, and excess kurtosis. These help identify pulsars, which typically produce sharp, high-intensity bursts that create strong patterns in signal shape.



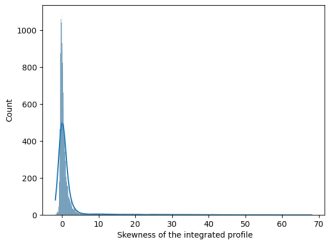
(a) Mean: Integrated Profile



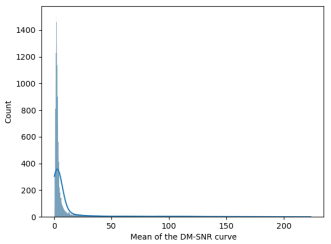
(b) Std Dev: Integrated Profile



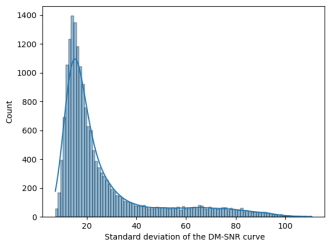
(c) Kurtosis: Integrated Profile



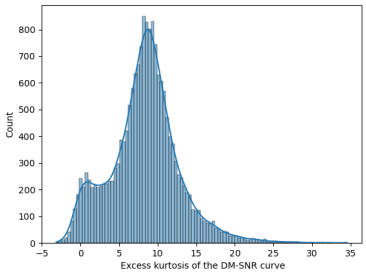
(d) Skewness: Integrated Profile



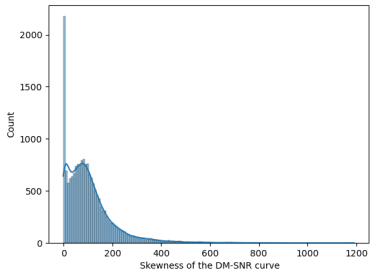
(e) Mean: DM-SNR Curve



(f) Std Dev: DM-SNR Curve



(g) Kurtosis: DM-SNR Curve



(h) Skewness: DM-SNR Curve

Figure 7: Distribution of all eight numerical features from the pulsar dataset

The histograms in Figure 7 reveal several important patterns across the features. The mean and standard deviation values for both the integrated profile and DM-SNR curve are fairly concentrated, with the integrated profile features showing roughly normal distributions and the DM-SNR curve being more right-skewed. This suggests most observations are non-pulsar signals with consistent, low-level structure. However, the shape-based features — skewness and kurtosis — show heavier tails and outliers, especially

in the DM-SNR curve. These long-tailed distributions indicate the presence of rare, sharp, or asymmetric signal patterns, which are consistent with the bursts expected from true pulsars. For example, the skewness of the DM-SNR curve reaches extreme values above 1000, highlighting just how different those signals are from the background noise.
This variation, particularly in the kurtosis and skewness metrics, is what makes these features valuable for classification. They help distinguish between normal noise and structured pulsar signals, especially when combined with statistical modelling techniques.

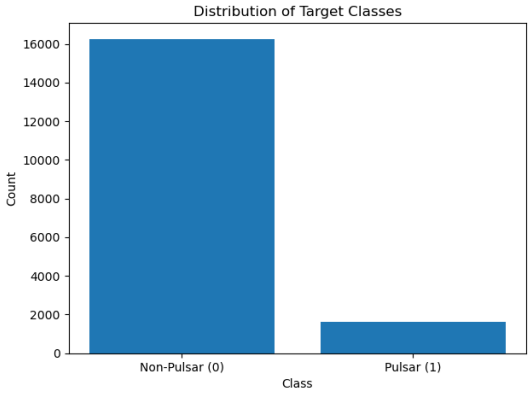


Figure 8: Class distribution: 0 = Non-Pulsar, 1 = Pulsar

Figure 8 shows a strong class imbalance in the dataset. The vast majority of observations are labelled as class 0 (non-pulsars), while only a small portion belong to class 1 (pulsars). This imbalance is expected in real-world astronomical data, where pulsars are rare compared to background signals or noise. As a result, standard accuracy is not a reliable metric when training classification models. Instead, performance will be evaluated using recall, precision, F1-score, and AUC, which are more informative when dealing with imbalanced datasets.

4.2 Classification of Pulsar Stars

To classify pulsars using the eight signal features, three models were used: Naive Bayes, Linear Discriminant Analysis (LDA), and Logistic Regression. An 80/20 train-test split was used, meaning 80% of the data was used to train the models and 20% to test them. These models were chosen because they are all commonly used for binary classification and are effective when the dataset has clear statistical patterns, like the ones found in pulsar signals.
Naive Bayes is a simple probabilistic model that works under the assumption that all features are independent. This assumption doesn't fully hold here, since the features are all calculated from the same two signals, but the model is fast and still performs reasonably well. LDA assumes that the features are normally distributed and that both classes share the same structure. This is mostly true for the dataset, and since many features follow normal-like distributions, LDA is a good fit. Logistic Regression doesn't assume normality or independence. It works by modelling the probability of each class and often performs well even when the dataset isn't perfect. It's also easier to interpret than other models.
This kind of classification is useful in physics because manually checking every signal for pulsar behaviour takes time. Automating the process helps speed up discovery and makes it easier to filter out irrelevant data. Also, since pulsars are rare, the models need to be especially good at catching as many true positives as possible without too many false alarms.

Table 5: Naive Bayes Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	3116	143
Actual 1	45	276

Naive Bayes identified 276 out of 321 pulsars, which gives a recall of 0.86. This means it successfully caught most of the pulsars. However, its precision was 0.66, which shows that a third of the predicted

pulsars were actually incorrect. This gives an F1-score of 0.75. The F1-score is calculated using the formula:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

This model has an AUC value of 0.9584, which is high, and shows that the model is generally good at separating pulsars from non-pulsars, even if it makes some mistakes.

Table 6: LDA Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	3243	16
Actual 1	76	245

LDA was more balanced. It correctly identified 245 pulsars with a recall of 0.76, and had a high precision of 0.94, which means most of the predicted pulsars were correct. The F1-score for class 1 was 0.84. It also had an AUC of 0.9735, which is slightly better than Naive Bayes. This shows that LDA can separate the classes well and makes fewer incorrect predictions, especially for non-pulsars.

Table 7: Logistic Regression Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	3241	18
Actual 1	57	264

Logistic Regression performed the best overall. It caught 264 pulsars with a recall of 0.82, and had a precision of 0.94. This gives it the highest F1-score of 0.88, meaning it was the most balanced model between catching pulsars and not overpredicting them. Its AUC was 0.9745, which was the highest among the three models.

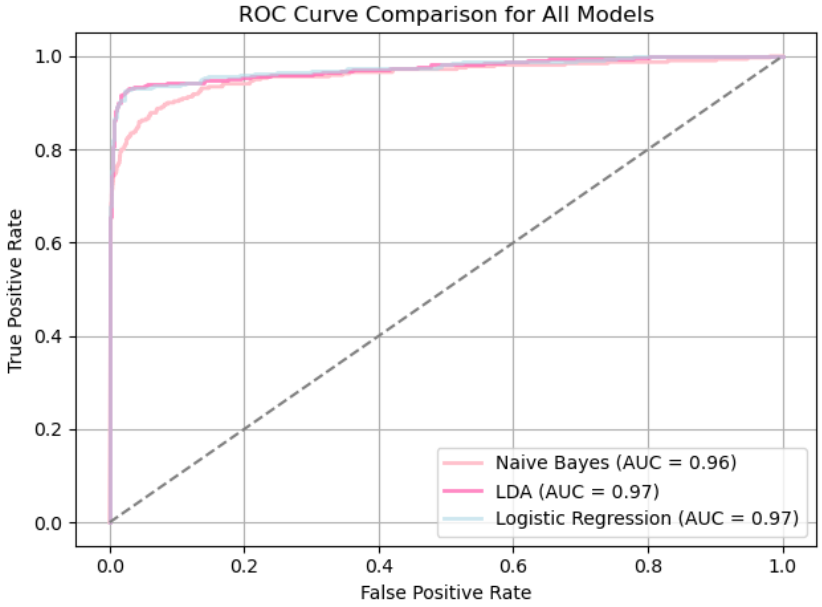


Figure 9: ROC Curve Comparison for All Models

Figure 9 shows the ROC curves for the three models. Logistic Regression and LDA had very similar curves, both close to the top-left, which shows high true positive rates with low false positives. Naive Bayes was slightly lower but still performed well. All three models had AUC values above 0.95, which confirms that the features in the dataset are useful for separating the classes.
Since pulsars are rare, the cost of missing one is high. This means recall is more important than overall

accuracy. However, precision also matters, because too many false positives would mean checking lots of signals manually. Logistic Regression gave the best overall results, with the highest F1-score and AUC, and a good balance between recall and precision. Based on this, it would be the best model to use for pulsar detection in this dataset.

5 Conclusion

This project applied regression and classification techniques to two real-world datasets, each with a different context and purpose. The first dataset focused on air pollution in Delhi and explored whether one pollutant could be used to predict another. A simple linear regression showed a strong relationship between PM10 and PM2.5, with an R-squared value of 0.9787. This suggests that PM10 can be used as a reliable estimate of PM2.5 when direct data is unavailable. A multiple linear regression was then used to predict CO using the other pollutants. The full model showed signs of multicollinearity and possible overfitting, so backward elimination was applied to reduce the number of predictors. The reduced model still achieved a high R-squared of 0.970 and was more stable and interpretable.
The second dataset involved detecting pulsar stars using signal-based features. Three models were tested: Naive Bayes, LDA, and Logistic Regression. All three achieved AUC scores above 0.95, showing that the features were useful for separating pulsars from non-pulsars. Naive Bayes had the highest recall but lower precision, while LDA was more balanced. Logistic Regression performed best overall, with the highest F1-score and AUC, making it the most reliable model for this task.
Overall, both parts of the project showed how statistical modelling can be used in physics and real-world applications. Regression was useful for estimating pollutants, which could support environmental monitoring and sensor redundancy. Classification helped automate pulsar detection, reducing the need for manual analysis. These methods show how data science can support scientific work by improving accuracy and efficiency.

6 Reference

- Kaggle. *Delhi Air Quality Dataset*. Available at: `https://www.kaggle.com/datasets/deepaksirohiwal/delhi-air-quality` [Accessed April 2025].

- Kaggle. *Pulsar Star Classification Dataset*. Available at: `https://www.kaggle.com/datasets/spacemod/pulsar-dataset` [Accessed April 2025].

- California Air Resources Board (CARB). *Inhalable Particulate Matter and Health*. Available at: `https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health` [Accessed April 2025].

- ScienceDirect. *Photochemical Smog*. Available at: `https://www.sciencedirect.com/topics/earth-and-planetary-sciences/photochemical-smog` [Accessed April 2025].

- V7 Labs. *F1 Score Guide – What It Is and How to Use It*. Available at: `https://www.v7labs.com/blog/f1-score-guide` [Accessed April 2025].

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, pp. 2825–2830. Available at: `https://scikit-learn.org/`

- Lorimer, D. R. (2008). Binary and Millisecond Pulsars. *Living Reviews in Relativity*, 11(1), 8. Springer. Available at: `https://link.springer.com/article/10.12942/lrr-2008-8`