Studying supervised machine learning techniques

Dafne Petrelli

Department of Physics, Queen Mary, University of London, Mile End Road, London, E1 4NS, UK

Abstract

This report focuses on the study of applying supervised machine learning techniques to two distinct data sets. The first data set involved examining the air pollution in the city Delhi, India by predicting the outcome of various pollutants depending on their correlation with the others, using linear least squares regression and stepwise regression techniques. The second data set involved the classification of neutron stars, to distinguish between pulsar and non-pulsars. Classification models such as Naive Bayes, Linear Discriminant Analysis and Logistic Regression were applied to the data set and the study of confusion matrices and the receiving characteristics curve (ROC) was also implemented.

1 Delhi data set

1.1 Data set exploration

Pollutant	Count	Mean	Std
CO	18 776.00	2929.23	2854.52
NO	18 776.00	33.66	62.13
NO_2	18 776.00	66.22	48.53
O_3	18 776.00	60.35	80.46
SO_2	18 776.00	66.69	49.44
NH_3	18 776.00	25.11	26.40
$PM_{2.5}$	18 776.00	238.13	226.53
PM_{10}	18 776.00	300.09	267.17

Table 1: Summary statistics of Delhi data set.

The first file handled throughout this study was about air quality in Delhi, India, and was labeled ”delhi-aqi.csv”. This file describes the concentration of the following pollutants: CO, NO, NO_2, O_3, SO_2, NH_3 , $PM_{2.5}$ and PM_{10}. The recordings took place from the 25th of November 2020 to the 24th of January 2023 at every hour.

”delhi-aqi.csv” was handled in Python language. To comprehend the data set, a table including the main statistics of it has been computed as in **Table 1**. The data set didn't contain any missing values, so what proceeded was its data analysis.

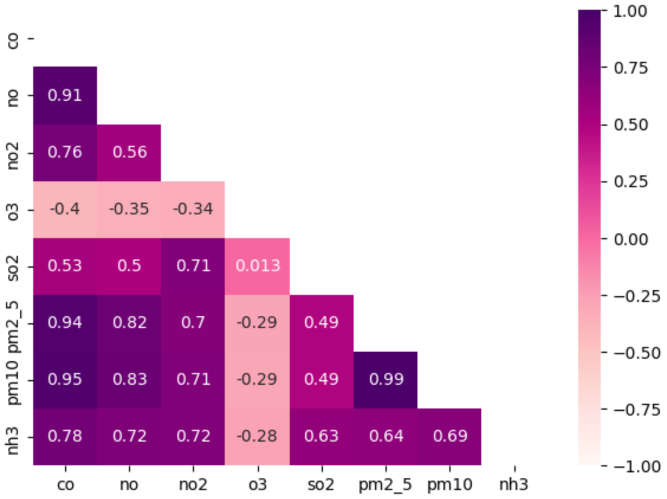

Figure 1: Heat map relating the pollutants.

For the Delhi dataset (**Figure 1**), a heat map was computed to visualize the correlations between different air pollutants, with darker colors indicating stronger correlations. A prominent observation is the strong positive correlation between CO and all other pollutants. This could be attributed to the predominantly urban setting in Delhi, particularly areas such as highways and industrial zones, where multiple pollutants often share common emission sources. The strongest correlation in the dataset is between PM_{10} and $PM_{2.5}$ (r = 0.99). This high correlation is likely due to the fact that $PM_{2.5}$ is a subset of PM_{10}, and both types of particulate matter often originate from the vehicular exhausts, particularly diesel vehicles, road dust, and industrial activities such as combustion processes [1]. In contrast, O_3 exhibits the weakest correlation with other pollutants. As a secondary pollutant, O_3 is not directly emitted into the atmosphere but rather forms through complex chemical reactions, which could explain its distinct correlation pattern compared to primary pollutants like CO and particulate matter.

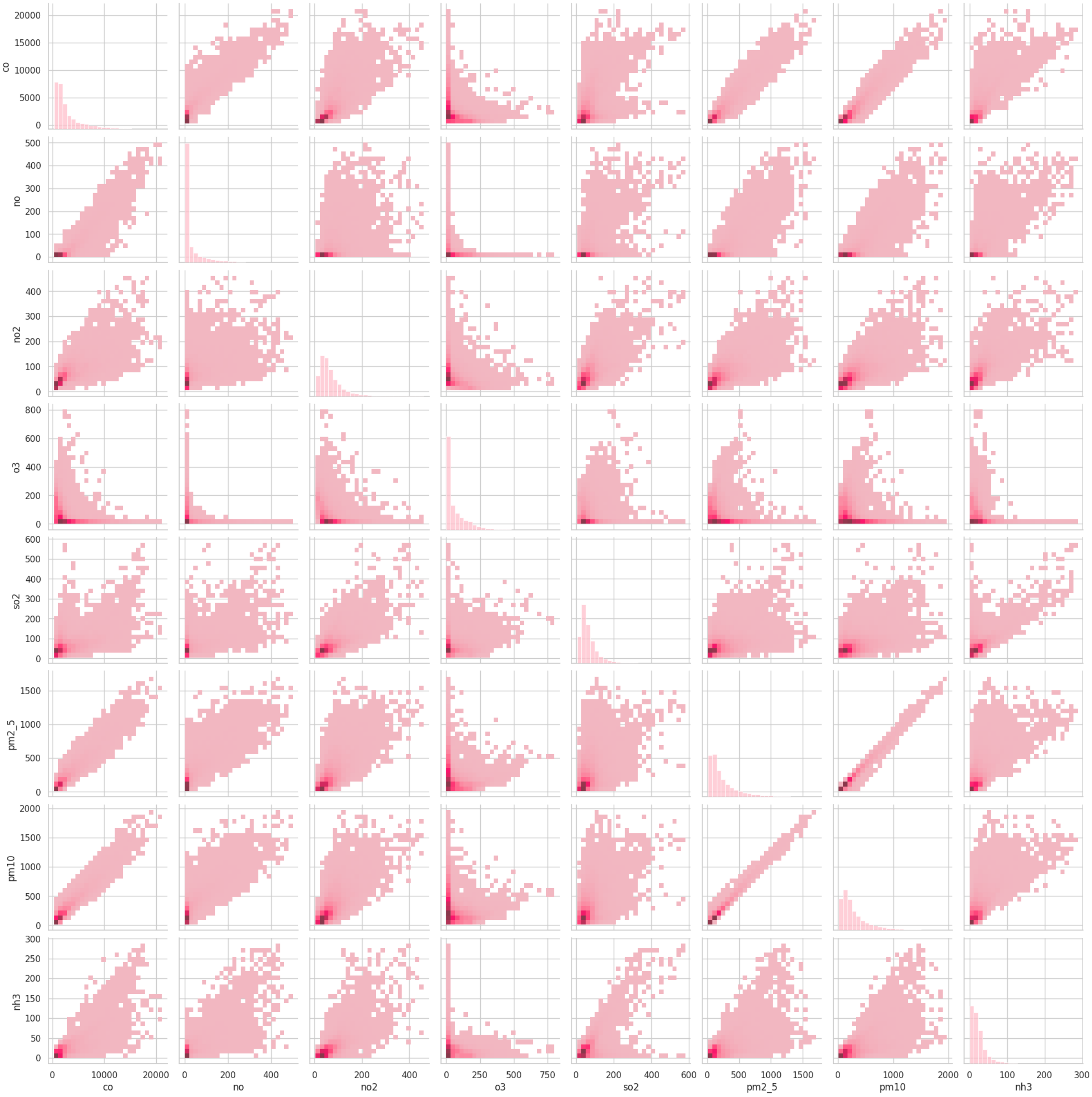

Figure 2: Pair plots relating the pollutants.

Figure 2 presents the pairwise scatter plots between various air pollutants, allowing for a more detailed visual assessment of their relationships. As previously noted, $PM_{2.5}$ and PM_{10} exhibit a strong positive correlation, clearly demonstrated by the tight upward-sloping distribution of their data points. This reinforces the understanding that both pollutants likely originate from common sources, such as vehicular emissions and construction activities. Carbon monoxide (CO) also shows a general positive correlation with several pollutants. This most evident in NO_2, $PM_{2.5}$, and PM_{10}, suggesting shared emission sources like combustion engines [1]. In contrast, its relationship with other pollutants appears weaker, as evidenced by the broader and more dispersed scatter patterns, indicating less consistent associations. Notably, the pair plots reveal weak or negligible correlations involving SO_2 and NH_3, which were less apparent in the heat map analysis. The data points in these plots appear widely scattered in a cloud-like formation, signaling minimal linear association with other pollutants. Furthermore, O_3 displays no discernible linear relationship with the remaining pollutants. The absence of any clear slope in its pairwise plots supports its classification as a secondary pollutant, with formation processes that differ significantly from those of the primary pollutants.

1.2 Least squares linear regression

1.2.1 Theory

To quantify the relationship nature between one of the pollutants to another, linear regression was performed, specifically ordinary linear least squares regression (OLS). This model predicts how much ”Y” will change when ”X” changes by a certain amount and is outlined by **Eq.1**.

$$Y = b_0 + b_1 X \tag{1}$$

where X is the predictor, Y is the outcome, b_0 is the intercept and b_1 is the slope or ”regression coefficient”.

The predicted values are then denoted by **Eq.2** as

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{2}$$

where X is the predictor, Y_i corresponds to the predicted value, β_0 and $\beta_1 X_i$ represent the estimated coefficients.

The regression line is the estimate that that minimizes the sum of the squared residual values and it takes the name of the residual sum of squares (RSS) and it showcased in **Eq.3** as

$$\text{RSS} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right)^2 \tag{3}$$

where β_0 and $\beta_1 X_i$ correspond to the values that minimize RSS and it is performed through linear least squares regression [2].

1.2.2 Results and Analysis

This methodology was applied to the two pollutants that were showcased to share the highest correlation throughout the data exploration, which corresponded to $PM_{2.5}$ and PM_{10}. By applying a Linear Regression model on Python, the following metrics were obtained

Metric	**Value**
R-squared	0.98
RMSE	38.99
Model Intercept	22.26
Model coefficient	1.17

Table 2: Least squares regression.

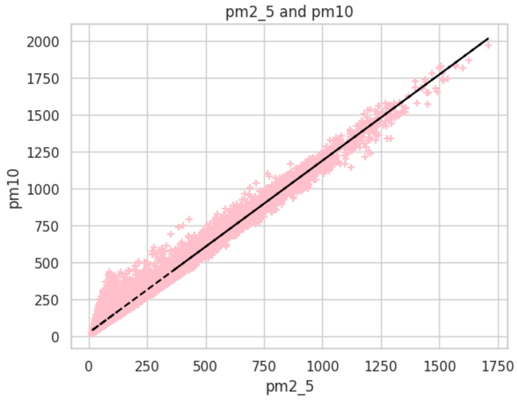

Figure 3: Scatter plot for linear least squares regression performed between $PM_{2.5}$ and PM_{10}, following **Eq. 2**.

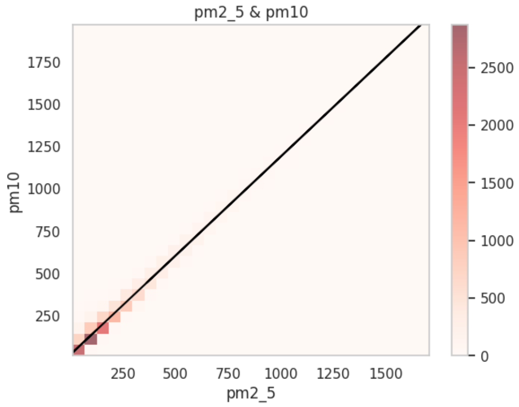

Figure 4: 2D histogram plot for linear least squares regression performed between $PM_{2.5}$ and PM_{10}.

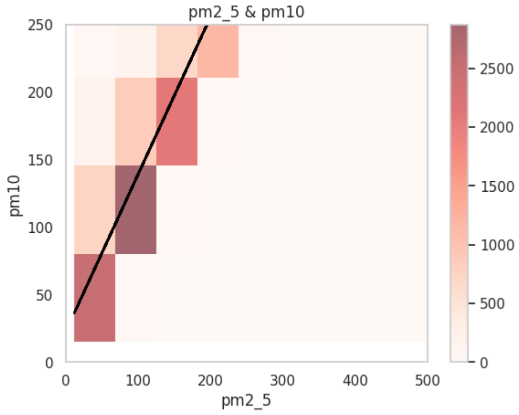

Figure 5: **Fig. 4.** zoomed in.

In this regression model the predictor and the outcome were selected to be $PM_{2.5}$ and PM_{10} respectively. By the metrics obtained in *Table 2* a correlation can be drawn between the pollutants. Firstly, the R-squared value, which is measured from 0 to 1 and it indicates the proportion of variation in the data that is accounted in the regression model. In this case how PM_{10} (the dependent variable) varies. A value for r-squared was obtained of 0.98, meaning that approximately 98% of the variance in PM_{10} is explained by model computed. It is a value very close 1, therefore resulting in the model being statistically significant and its performance is successful and it accounts for nearly all of the variability in the outcome variable. Secondly, the square root of the average squared error (RMSE), which measures the average magnitude of how much the the outcome data deviates from the original data and indicates how statistically significant the error is. The value of 38.99 compared to the pollutant's count being 18 776.00 it is very small, which implies that the model suits very well the data since on average the predicted data deviates from the original by approximately 39 units. Moreover, the model intercept (b_0) showcases that when the predictor $PM_{2.5}$ is equal to 0, then the predicted value for PM_{10} will correspond to 22.26. Finally the model's coefficient (b_1) being 1.17 for the predictor $PM_{2.5}$ implies that for every approximately 1 unit increase, the outcome will linearly increase as well. To visualize these previous results, a scatter graph was plotted (**Fig. 3**), which displays the fitted regression line based on the both the intercept and the coefficient and it quantifies the relation between the two pollutants. The regression model successfully re-affirmed that $PM_{2.5}$ and PM_{10} share a strong positive correlation. This is highlighted in the graph by the points tightly surrounding the regression line, showcasing that as $PM_{2.5}$ increases, PM_{10} increases linearly. In **Fig. 4**, a 2D plot was computed based on the previous results. **Fig. 5** provides a zoomed-in version of the 2D histogram, which gives a better visualization of the imperfections linked with using this regression model. The darker "rectangles" below the regression line showcase the original data. Therefore it can be stated that the predicted data from the regression model used tends to be higher than the original data [2].

$$Y_i < \hat{Y}_i \tag{4}$$

There is a slight overestimation of the values for PM_{10} for the values of $PM_{2.5}$. Nonetheless, this could be also due to the presence of outliers, which the model cannot account for. This showcases that there are still sources of errors affecting this model slightly. The error which was computed for intercept and for PM_{10} is showcased below

	std error
Intercept	0.363
pm10 coefficient	0.001

Table 3: Std error.

This showcases that the model has resulted overall to be quite successful since the PM_{10}'s coefficient standard error of 0.001 is much smaller than the intercept's standard error, indicating high precision when in the prediction that characterized this model.

1.3 Multiple linear regression

As previously identified, CO exhibited the strongest correlation with the other pollutants. To further investigate this relationship, a multiple linear regression model was implemented with CO as the dependent variable. The objective was to examine the full model and iteratively remove predictors to determine which variables significantly contribute to the prediction of CO concentrations.

1.3.1 Theory

The regression technique applied in this analysis follows the stepwise regression or backward selection approach. This approach are susceptible to overfitting, especially when the model begins to fit noise within the dataset rather than underlying patterns.

1.3.2 Results and Analysis

Pollutant	**coef**	**std error**	**t**	**P>t**
nh3	10.194	0.202	50.351	0.000
no	18.372	0.101	181.424	0.000
no2	13.168	0.128	103.158	0.000
so2	−5.564	0.098	−56.931	0.000
pm2_5	1.623	0.092	17.685	0.000
pm10	3.266	0.082	40.051	0.000
o3	−1.033	0.044	−23.595	0.000

Table 4: Regression coefficients with standard errors, t-statistics, p-values, and confidence intervals.

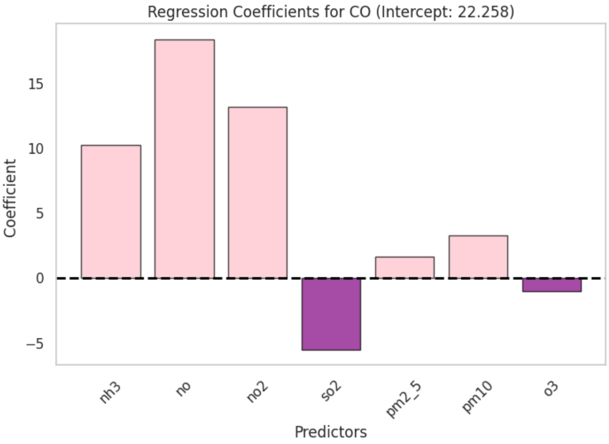

Figure 6: Histogram showcasing the regression coefficients for CO.

To determine which pollutants to exclude, a summary statistics table was generated (**Table 4**), and a histogram illustrating the various regression coefficients was created (**Fig. 6**). Based on these two analyses, the pollutants so_2 and O_3 were identified for removal. Both pollutants showed negative regression coefficients, indicating an inverse relationship with CO, and their small values suggested a weak correlation. The key statistics were then recalculated and are presented below.

Pollutant	coef	std error	t	P>t
nh3	7.075	0.231	30.624	0.000
no	17.926	0.108	166.448	0.000
no2	10.224	0.114	89.515	0.000
pm2_5	0.744	0.106	7.031	0.000
pm10	4.267	0.093	45.888	0.000

Table 5: Regression coefficients with standard errors, t-statistics, p-values, and confidence intervals.

After removing the previous pollutants, multicollinearity was no longer flagged in the code output. Previously, it had been detected due to a large condition number of 1.2e+03, which suggested the presence of strong multicollinearity or other numerical issues. The next step was based on the analysis of the t-statistics. The higher the t-statistic, the stronger the correlation between the predictor and the outcome. Consequently, the pollutants NH_3 and $PM_{2.5}$ were removed from the model, as they had the lowest t-values, 30.624 and 7.031, respectively. Finally, the standard errors for all the computed coefficients were below 0.1, indicating high precision in the model's predictions. This suggests that the model estimates the coefficients with a high degree of confidence. Given that these low standard errors typically correspond to statistical significance, we can infer that the relationships between the predictors and the outcome are meaningful and not due to random chance, as shown in the results below.

Pollutant	std error
no	0.097
no2	0.098
pm10	0.027

Table 6: std error.

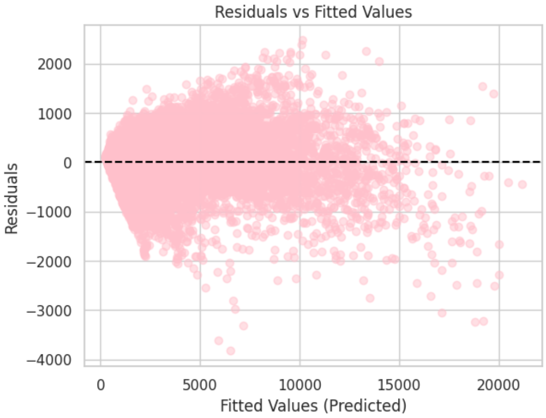

Figure 7: Scatter plot of residual values against fitted values for CO.

In **Fig. 7** the scatter plot showcases a "cone-like" dispersion pattern, indicating that the residuals are not uniformly distributed around the zero line. This spread suggests the presence of outliers where the model performs poorly, particularly at higher predicted values, which results as an unequal scatter. The phenomenon observed is referred to as heteroskedasticity and it is defined as "a lack of constant residual variance across the range of the predicted values"[2]. This is evident in the regression model that was performed (OLS), since it assumes that all residuals are drawn from a population which has a constant variance.

2 Neutron stars stars data set

2.1 Pulsars

Pulsars are highly fast neutron stars, which are extremely dense. Neutron stars are formed during supernovas when the core of a star collapses inwards. They are accelerated speeds such as the speed of light, which leads the decay particles to emit electromagnetic radiation as intense beams from the pulsar's magnetic poles. Pulsars eject periodic signals which are often very weak to detect .

2.2 Data set exploration

The second file handled throughout this study was about neutron stars and was labeled ”pulsar-stars.csv”. This file describes the various features measured to distinguish between a pulsar and a non pulsar as showcased in **Table 7**. Integrated pulsar signals, when aligned with respect to their rotational period, result in enhanced signal quality. This process yields what is known as the integrated pulse profile [3].

Feature	Count	Non-Null	Dtype	Mean	Std
Mean of the integrated profile	17 898.00	non-null	float64	111.08	25.65
Standard deviation of the integrated profile	17 898.00	non-null	float64	46.55	6.84
Excess kurtosis of the integrated profile	17 898.00	non-null	float64	0.48	1.06
Skewness of the integrated profile	17 898.00	non-null	float64	1.77	6.17
Mean of the DM-SNR curve	17 898.00	non-null	float64	12.61	29.47
Standard deviation of the DM-SNR curve	17 898.00	non-null	float64	26.33	19.47
Excess kurtosis of the DM-SNR curve	17 898.00	non-null	float64	8.30	4.51
Skewness of the DM-SNR curve	17 898.00	non-null	float64	104.86	106.51
Target class	17 898.00	non-null	int64	0.09	0.29

Table 7: Main statistics of ”pulsar-stars.csv”.

”pulsar-stars.csv” was handled in Python language. The data set didn’t contain any missing values, so what proceeded was its data analysis.

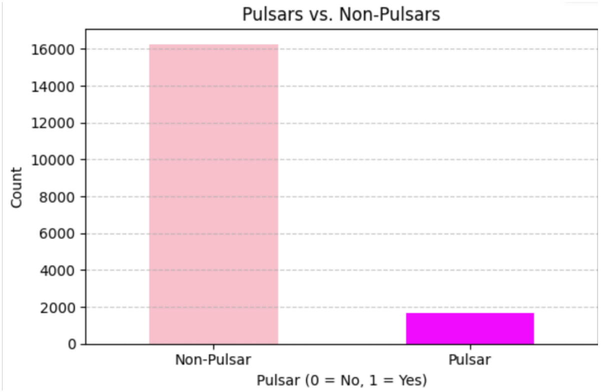

Figure 8: Class distribution categorizing pulsars and non pulsars.

Fig. 8 presents the class distribution of the dataset, providing insight into the nature of the data. As discussed previously, pulsars are difficult to detect due to their weak and subtle signals. Consequently, the majority of observations are more likely to represent irrelevant noise or non-pulsar sources. The class distribution, represented by the target class in the data set, clearly illustrates this imbalance. Specifically, 90.84% of the data corresponds to non-pulsars, while only 9.16% represents actual pulsars. This significant disparity highlights the challenges in accurately identifying pulsar signals. This classification for pulsars and non-pulsars was not taken account when plotting each different feature as a histogram in **Fig. 9**. To understand the different features of the neutron stars, the distribution of the histograms was taken into consideration.

Pulsars are often identified by the distinctive shapes of their integrated pulse profiles, which serve as unique identifiers. These profiles demonstrate remarkable consistency across multiple observations and are generally stable over time, reinforcing their reliability as a distinguishing feature.

Both the mean and the standard deviation of the integrated profile showcase a normal distribution with a bell-shaped curve. There is slight positive skewness in the plot for the mean, this could be due to the presence of extra noise. To further characterize the data distribution, excess kurtosis and skewness were utilized. Excess kurtosis quantifies the tailedness of a distribution relative to a normal distribution, indicating the presence of outliers or heavy tails. On the other hand, skewness measures the asymmetry of the distribution, highlighting any deviation from a symmetrical (Gaussian) shape. For the excess kurtosis and the skewness there is no defined distribution, but instead it is shifted towards the right, indicating the presence of outliers, specifically near 0. Another critical observational feature is the Dispersion Measure (DM). This arises due to the interaction of pulsar signals with free electrons in the interstellar medium. Observationally, this dispersion leads to a broadening of otherwise sharp pulses when data is collected over a finite bandwidth, thereby impacting signal clarity. The mean, the standard deviation and the skewness showcase strong right skewness. For the mean and standard deviation this could be due to the low presence of pulsars which if they would be present they would produce noticeable peaks through the distribution. Finally he excess kurtosis showcases a normal distribution which indicates a high presence of outliers [4].

These statistical features provide valuable insights into the structure and variability of pulsar signals.

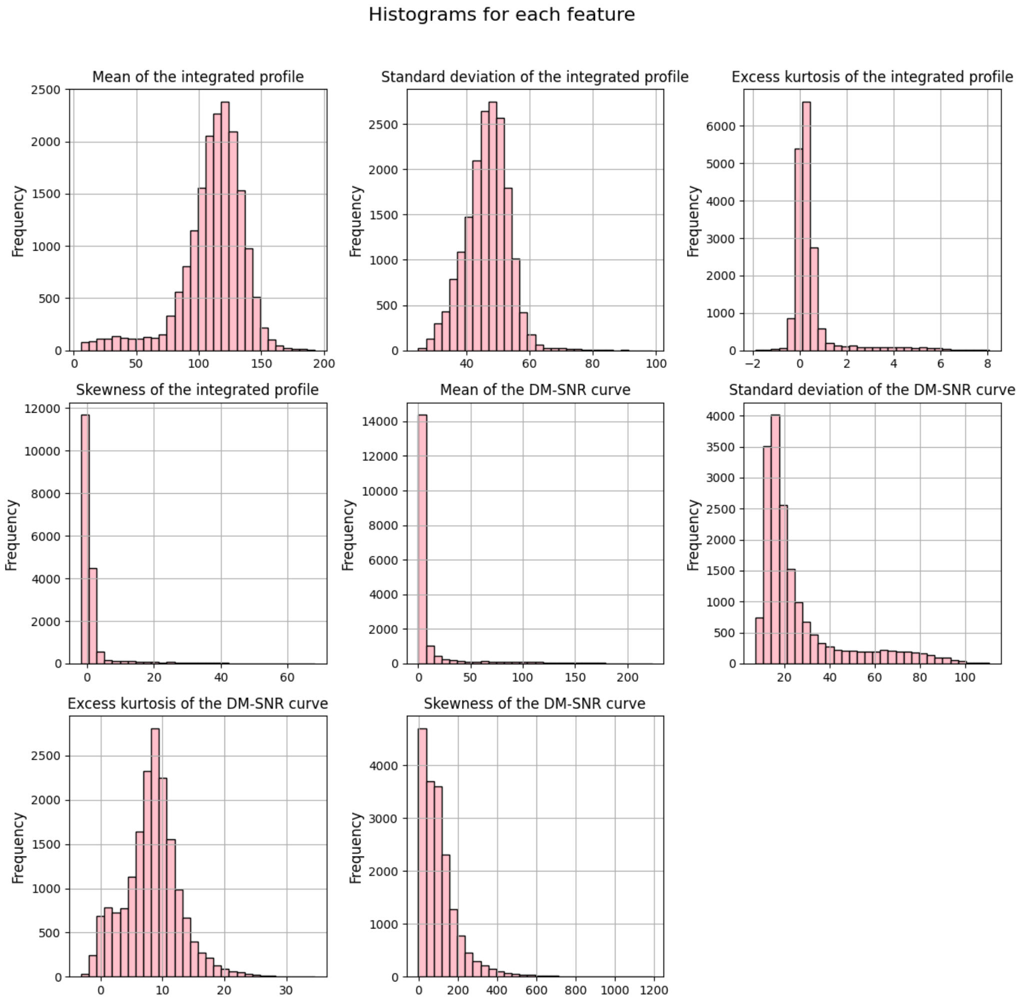

Figure 9: Histogram plots for each feature of the neutron stars data set.

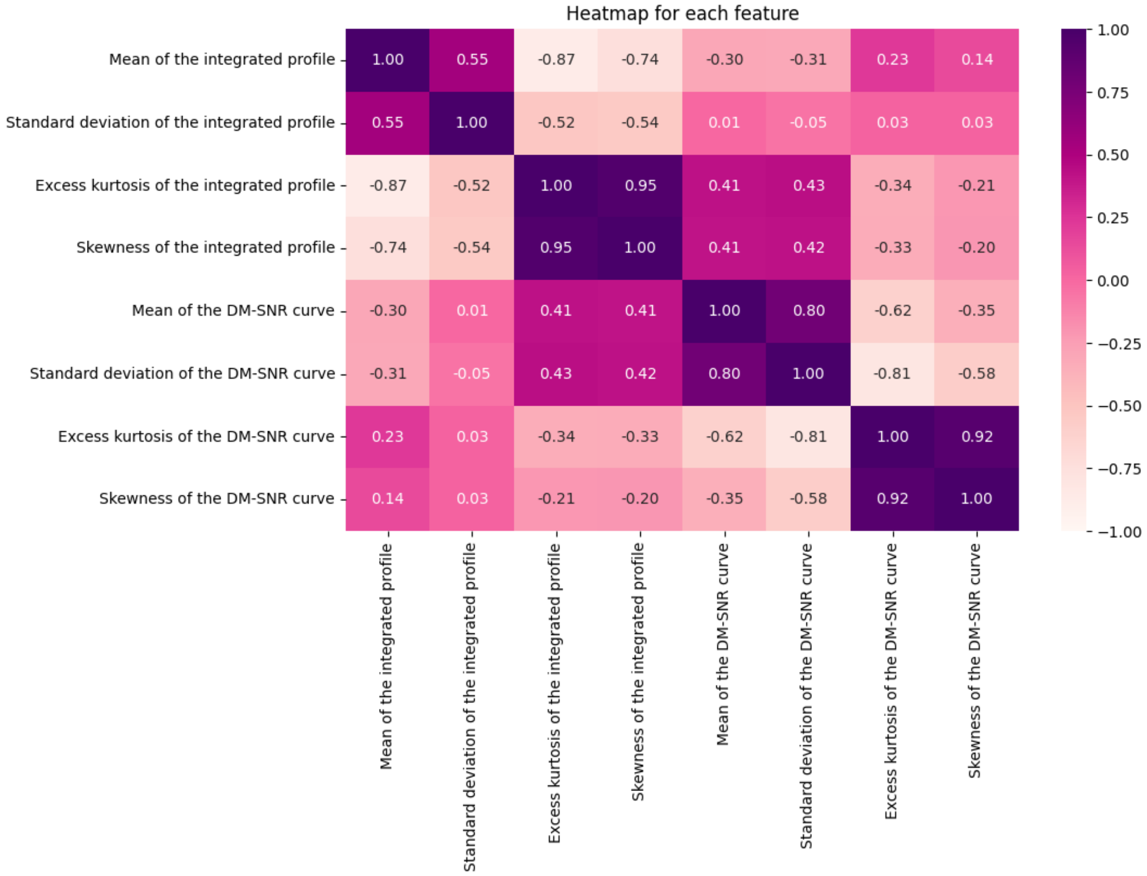

Figure 10: Heat map for the neutron stars features.

For the Neutron stars dataset (**Figure 10**), a heat map was computed to visualize the correlations between different features, with darker colors indicating stronger correlations. The most prominent positive correlation was observed between the skewness and excess kurtosis of the integrated profile, exhibiting a correlation coefficient of 0.95. This strong relationship may be attributed to their co-linearity in the context of pulsar detection. In many cases, a distribution that exhibits high skewness (asymmetry) also displays high kurtosis (peaked-ness), particularly in the presence of true pulsar signals. Both features are sensitive to outliers caused by sudden, which are more common in pulsar observations.

On the other hand, a strong negative correlation was observed between the mean and excess kurtosis of the integrated profile, with a coefficient of -0.87. This inverse relationship suggests that as the mean signal intensity increases, which indicates the presence of a strong, structured pulsar signal, the excess kurtosis decreases. This may be due to a reduction in noise and outlier influence as the signal stabilizes, leading to a flatter, more uniform distribution.

2.3 Classification models

2.3.1 Theory

Classification is a form of prediction, in this project the aim was to predict if the neutron star would correspond to a 0 (non-pulsar) or to a 1 (pulsar). In this report three types of classification models will be studied, Naive Bayes, Linear discriminant analysis and logistic regression. Confusion matrices were plotted for each model as they yield significant statistical importance as a metric. It is a table showing the number of correct and incorrect predictions categorized by the type of feature. In a confusion matrix, the rows represent the actual (true) class labels, while the columns represent the predicted class labels. The diagonal elements indicate the number of correctly classified instances, whereas the off-diagonal elements reflect the number of misclassifications.

Other fundamental metrics that will be looked at include the precision, the recall, the F1 score and the accuracy. The precision measures the accuracy of a predicted positive outcome as

$$precision = \frac{\Sigma TP}{\Sigma TP + \Sigma FP} \tag{5}$$

where TP refers to True positive predictions and FP refers to True negative predictions.

The recall measures the strength of the model to predict a positive outcome as

$$recall = \frac{\Sigma TP}{\Sigma TP + \Sigma FN} \tag{6}$$

where FN refers to False negative predictions. The F1 score measures the harmonic mean between the precision and the recall [5] as

$$F1 - score = 2\frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

The accuracy measures the total correct assignments as

$$accuracy = \frac{\Sigma TP + \Sigma TN}{Sample_t} \tag{8}$$

where TN refers to True negative predictions and Sample$_\mathrm{t}$ refers to the total sample size [4].

2.3.2 Naive Bayes

The Naive Bayes (NB) model provides to observe the probability of the outcome by using a set of predictor values. The number of possible combinations increases and often the mismatching of many records will increase, which reduces the probability of an exact matching taking place.This is because the entire data set is being used. The probability of observing the outcome $Y{=}i$, given by a set of predictors follows **Eq. 8** [4] as

$$P(Y = i|X_1, X_2, ..., X_p) \tag{9}$$

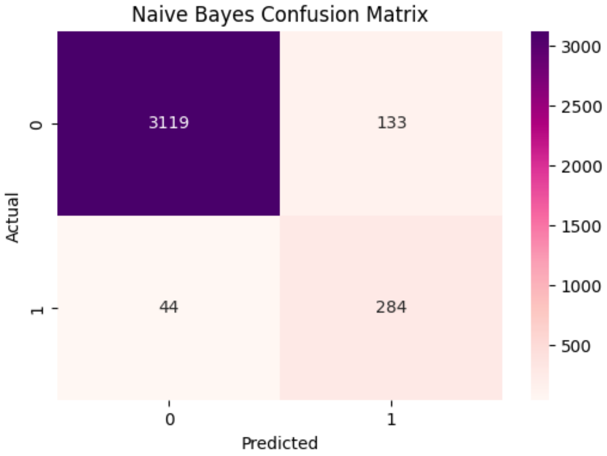

Figure 11: Confusion Matrix for Naive Bayes.

NB performed well when predicting True negative (TN=3119) and True Positive (TP=44). However, it fails when predicting false positives (FT=133), since it misclassified many class 0 predictions as class 1.

Class	Precision	Recall	F1- score	Accuracy
0	0.99	0.96	0.97	0.95
1	0.68	0.87	0.76	0.95

Table 8: Classification report for Naive Bayes.

NB showcases a very high precision (=0.99), showcasing that it is very good a predicting class 0 and recall (=0.96), meaning most TP were ”caught”. However its precision for predicting class 1 is quite low (=0.68), therefore many FP exhibited in the prediction. The F1 score (=0.76) for class 1, is a moderate result but still not enough when classifying pulsars and non-pulsars.

2.3.3 Linear Discriminant Analysis

The linear discriminant analysis (LDA) model assumes that the predictor variables are normal distributed continuous variables. It focuses on the ”between” sum of the squares, which measures the variation between the two groups and the "within" sum of the squares, which measures the variation within the group.

$$\frac{SS_b}{SS_w} \tag{10}$$

where $\mathrm{SS_b}$ is the sum in of the squares in "between" and $\mathrm{SS_w}$ is the sum of the squares "within".

This method therefore yields the greatest separation between the two groups [4].

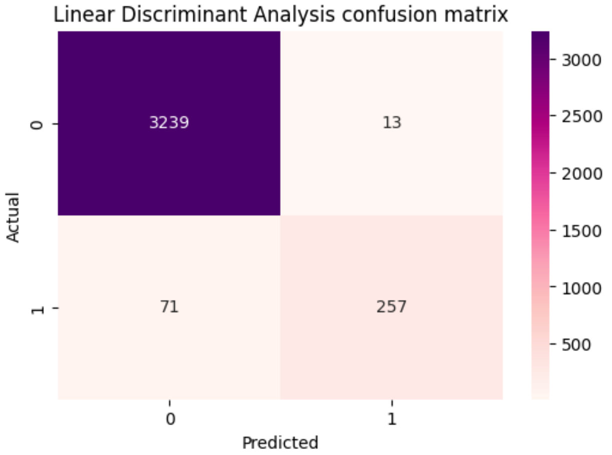

Figure 12: Confusion Matrix for Linear Discriminant Analysis.

LDA performed exceptionally well when avoiding false positives (FP=13), which suggests it is preventing the misclassifications of class 1.

Class	Precision	Recall	F1- score	Accuracy
0	0.99	1.00	0.99	0.98
1	0.95	0.78	0.86	0.98

Table 9: Classification report for Linear Discriminant Analysis.

The metrics are all balanced expect for the Recall for class 1, which means not as many true positives or false positives were predicted.

2.3.4 Logistic Regression

The logistic regression (LR) model uses the logistic response function, along with the logit transformation. This enables the mapping of probabilities, which are naturally constrained to a [0, 1] range, onto an unbounded continuous scale. This is showcased by **Eq. 11** as,

$$p = \frac{1}{1 + e^{-}(\beta_0 + \beta_1 x_2 + \beta_2 x_2 + \ldots + \beta_q x_q})\tag{11}$$

where p is modeled and this transformation allows it to stay between the values of 0 and 1 [4].

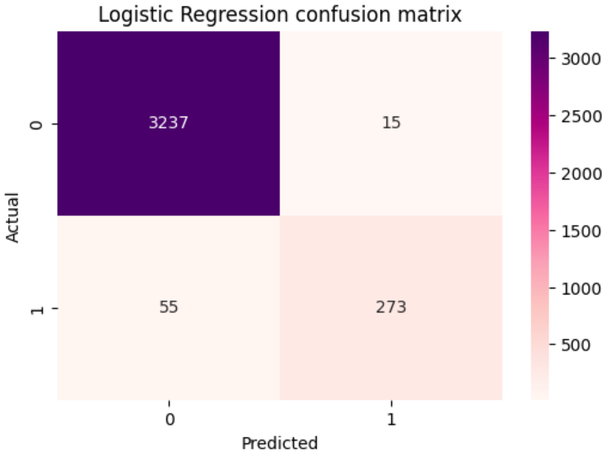

Figure 13: Confusion Matrix for Logistic Regression.

LR performed exceptionally well as but differently from LDA. Most true negatives (TN=3237) and true positives (TP=273) are predicted precisely. A key observation which the other models don't exhibit as much, is that LR reduces the number of false negatives which makes it more reliable when predicting class 1.

Class	Precision	Recall	F1- score	Accuracy
0	0.98	1.00	0.99	0.98
1	0.95	0.83	0.89	0.98

Table 10: Classification report for Logistic Regression.

The metrics showcase a great balance, with very high precision for both class 0 (=0.98) and class 1 (=0.95). Perfect recall for class 0 (=1.00) and well balanced F1 score between class 0 (=0.99) and class 1 (=0.89).

2.4 Discussion and Analysis

Classification model	AUC score
Naive Bayes (NB)	0.965
Linear Discriminant Analysis (LDR)	0.981
Logistic Regression (LR)	0.983

Table 11: AUC score for each classification model.

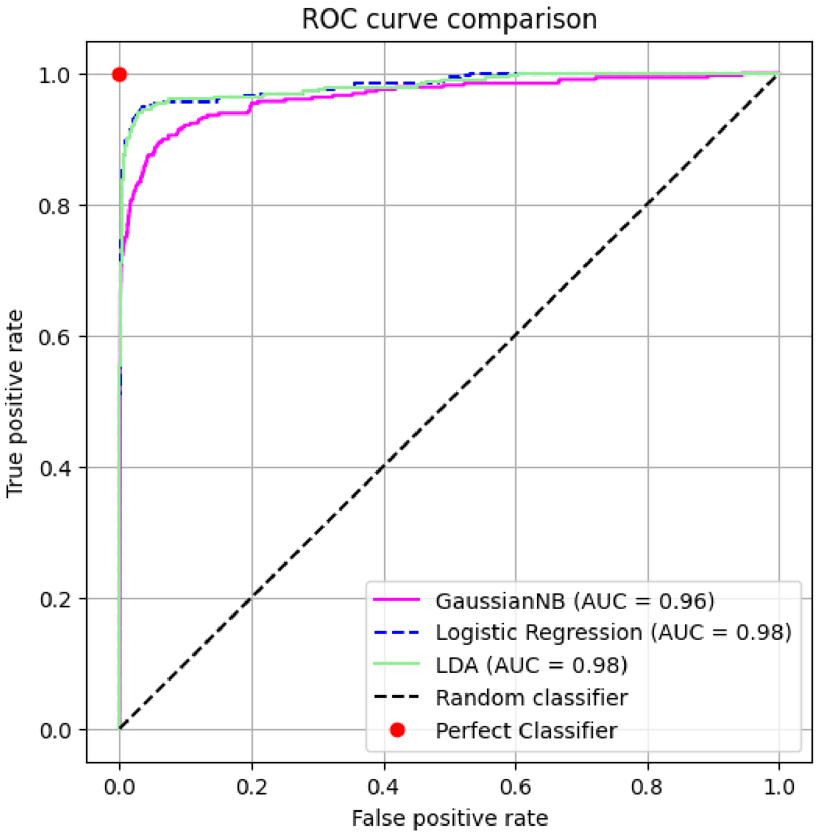

Figure 14: ROC curve comparison for all three classification models (Naive Bayes, Linear Discriminant Analysis, Logistic Regression).

In **Fig. 14** a receiving characteristics curve (ROC) curve comparison plot was performed to visually compare the three classification models. All three models exhibit to statistically significant with an Area under the curve (AUC) score of < 0.95 (**Table 11**). The closer the AUC score is to 1, which indicates the perfect classifier, the more accurate the classification model is. The model with the highest AUC score performance turned out to be LR (AUC=0.983) very close to LDR. It was decided to compute the AUC values to 3 significant figures to improve accuracy in the findings. This result is also exhibited by the ROC plot, which displays both the LDR and the LR curve to be the most far away from the random classifier line and NB being the closest, not by a lot but out of all the three models. This re-affirms NB's performance being the weakest out of the three models.

3 Conclusion

In this report, two data sets, one regarding the Air pollution in the city of Delhi, India and the other regarding Neutron stars and their different features. Two different supervised machine learning techniques were applied to both. Regression modeling techniques were applied to the Delhi data set, on the other hand classification techniques were applied to the Neutron stars data set. In the first data set, the exploration of the data which was performed by the use of heatmaps and pair plots, revealed which pollutants shared the highest correlation. $PM_{2.5}$ and PM_{10} shared the highest correlation coefficient (=0.99), therefore linear least squares regression was applied, with PM_{10} being the outcome and $PM_{2.5}$ being the predictor. This was done because $PM_{2.5}$ is a subset of PM_{10}, so this was validated to confirm this relation. Next, stepwise regression was performed on CO as the outcome, since it revealed itself during the data exploration to be the pollutant which shared on average the highest correlation with the others, so by looking at metrics such as regression coefficients, t-statistics and std error, step-by-step a predictor was dropped to finally form a model which upholds the most statistical significance. The predictors at the end of the model corresponded to NO, NO_2 and PM_{10} and they all had a std error of < 0.01 which made them all statistically significant.

In the second data set, the exploration of the data which was performed by the use of heatmaps and histograms (plotted for each feature), revealed the disparity in class between non-pulsars (=0) and pulsars (=1). To avoid class imbalance, the data was stratified in order to maintain the same test size for both predictors and outcomes throughout the classification modeling. Three classification models were performed including Naive Bayes (NB), Linear Discriminant Analysis (LDA) and Logistic Regression (LR). Confusion matrices were performed with use of each model and they all performed well, but the model which performed the best was LR, since it showed the best balance between recall and precision. NB had the highest recall but the lowest precision and finally LDA performed similarly to LR but not as well. A ROC curve comparison was performed and once again LR performed the best when predicting the AUC score(=0.983), with LDA being the second best (=0.981) and NB being the weakest (=0.965). This lead the best modeling performance to be LR when classifying neutron stars as pulsars and as non-pulsars. In conclusion, this report provides valuable insights into two distinct areas: air pollution in Delhi and the classification of neutron stars. The regression modeling applied to the air pollution dataset highlighted key correlations between pollutants, which could be used to inform future environmental monitoring. On the other hand, the classification of neutron stars using classification models demonstrated the effectiveness of machine learning techniques when solving complex astronomical classification tasks. Future research could build on these findings by exploring additional machine learning techniques or incorporating more data sources to further refine predictions and improve model accuracy. While linear models like LR proved effective, there is potential for even better performance with more advanced methods such as ensemble learning or deep learning. Furthermore, examining the impact of additional external factors on both air pollution and neutron star classification could provide a deeper understanding of these complex phenomena.

References

[1] ResearchGate. (n.d.). *AN ANALYSIS OF AIR POLLUTION AND ITS IMPACT ON HUMAN POPULATION IN DELHI.* [online] Available at: `https://researchgate.net/publication/328353976_AN_ANALYSIS_OF_AIR_POLLUTION_AND_ITS_IMPACT_ON_HUMAN_P` [Accessed 9 Apr.

2025]. [2] Qmul.ac.uk. (2024). *SPA5131 - 2024/25 — MyQMUL*. [online] Available at: https://qmplus.qmul.ac.uk/pluginfile.php/4295955/mod_resource/content/22/notes05.pdf [Accessed 9 Apr. 2025]. [3] Encyclopædia Britannica. (2019). *Pulsar — cosmic object*. [online] Available at: https://www.britannica.com/science/pulsar [Accessed 9 Apr. 2025]. [4] (Qmul.ac.uk. (2024). *SPA5131 - 2024/25 — MyQMUL*. [online] Available at: https://qmplus.qmul.ac.uk/pluginfile.php/4295955/mod_resource/content/22/notes05.pdf [Accessed 9 Apr. 2025]. [5]GeeksforGeeks. (2023). *F1 Score in Machine Learning*. [online] Available at: https://www.geeksforgeeks.org/f1-score-in-machine-learning/ [Accessed 9 Apr. 2025].