Analysis of air pollution level in Delhi using Linear Regression and comparison of 3 pulsar classifier models.

Abstract

This report aims to create linear regression model based on the New Delhi pollution data and evaluate the performance of 3 classification models with the Pulsar dataset using Python programming language. First part of the report focuses on predicting the quantity of nitrogen dioxide (NO2) with linear regression, based on 8 different pollution sources in New Delhi dataset. In second part of report, Naïve Bayes, Linear Discriminant Analysis and Logistic Regression models were trained to classify whether the neutron star is a pulsar and then compared between each other. The analysis process for both parts included general inspection of the dataset, correlation analysis, creating and training models, and general evaluation of the model based on key metrics like R-squared and MRSE for linear regression and confusion matrices and AUC scores for classifier models.

Linear Regression using Delhi pollution data.

Dataset

The Delhi dataset consists of 18,776 records on air pollution levels, including particulate matter (PM2.5 and PM10) levels, nitric oxide (NO), nitrogen dioxide (NO2), sulphur dioxide (SO2), carbon monoxide (CO), ozone (O3), and Ammonia (NH3). The data was collected hourly from monitoring stations located in various areas of Delhi between January 1, 2020, and June 30, 2023. The data is stored in a CSV format, featuring a timestamp for each record. Occasionally the dataset has 0 values for some of the elements at random intervals, sometimes spanning for a week or more, which will later be described in correlation analysis.

Analysis

The analysis started with the general inspection of the dataset. The dataset consists of 18776 rows x 9 columns, which indicate date stamps and the values for pollutant at these date stamps. The dataset was checked for missing values, which weren’t present. The key metrics such as mean, standard deviation, and quartile values were calculated for each pollutant, which gives the idea on how pollutants in Delhi dataset are distributed.

	co	no	no2	o3	so2	pm2_5	pm10	nh3
count	18776	18776	18776	18776	18776	18776	18776	18776
mean	2929.23	33.66	66.22	60.35	66.69	238.13	300.09	25.11
std	2854.52	62.12	48.53	80.46	49.43	226.53	267.17	26.40
min	260.35	0.00	4.28	0.00	5.25	11.83	15.07	0.00
25%	1068.12	0.68	33.93	0.34	34.81	84.44	118.79	9.63
50%	1842.50	5.25	54.15	27.18	52.93	157.45	209.71	17.48
75%	3685.00	35.76	83.63	92.9	82.02	313.00	387.97	30.40

Table 1: Pollutant distribution statistics in Delhi dataset.

The correlation heatmap was then created to analyse the relationship between the variables in the dataset.

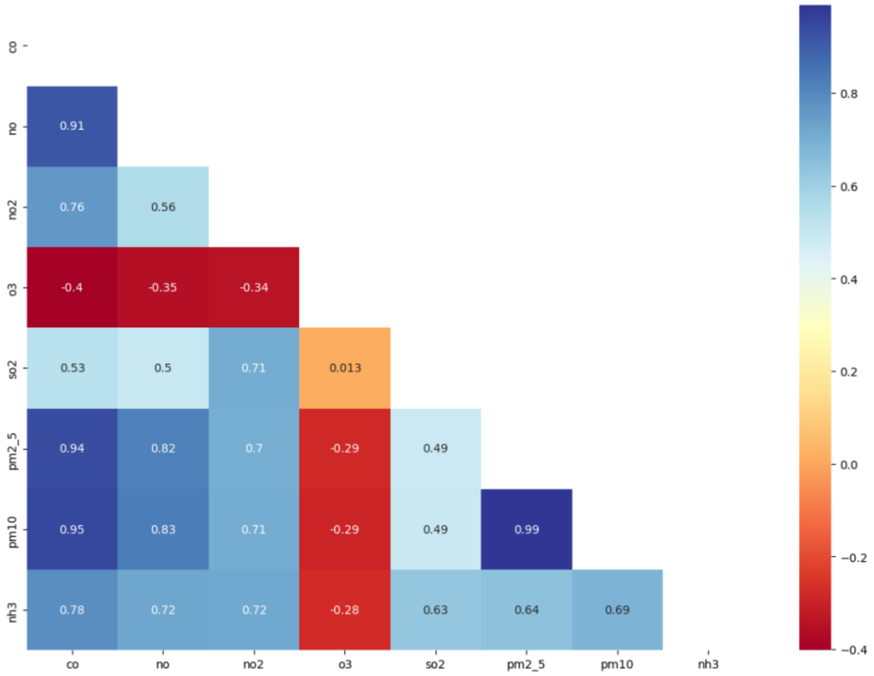

Figure 1: The correlation heatmap of the pollutants in Delhi dataset.

Most of the pollutants have positive correlation, with majority of them being above 0.5. This fact stems from several reasons, described below.

• **Common sources:** Pollutants originating from similar sources are likely to exhibit positive correlations. For example, pollutants emitted from vehicular exhaust, such as nitrogen oxides (NOx) and particulate matter (PM), tend to be positively correlated due to their common source.

• **Meteorological Conditions:** Certain weather conditions, such as temperature inversions or stagnant air masses, can trap pollutants near the ground. This can result in positive correlations among pollutants during periods of poor air quality.

• **Chemical Reactions:** Some pollutants can undergo chemical reactions in the atmosphere, leading to the formation of secondary pollutants.

• **Seasonal Variations:** Seasonal factors, such as changes in temperature, humidity, and atmospheric stability, can influence pollutant concentrations.

However, the only pollutant that has negative correlation value with all other pollutants except SO2, is ozone (O3). This happens due to the chemical reactions that ozone undergoes with other main pollutants in the atmosphere. Ozone formation primarily relies on photochemical reactions involving NOx and VOCs in sunlight; however, ozone is consumed in reactions with NOx, leading to its depletion. [1]

Ozone concentrations can also be influenced by volatile organic compounds VOCs, which contribute to ozone formation but subsequently react with ozone, depleting its levels. Additionally, CO indirectly affects ozone concentrations by altering NOx and VOC levels and directly reacting with ozone, reducing its quantity in the atmosphere.

The linear regression model with one variable was then created to predict the quantity of nitrogen dioxide (NO2), based on pollution data of particulate matter with 10 micrometres in diameter (PM10). The PM10 was chosen as a predictor variable based on the correlation between these two variables, which was 0.71. This positive correlation means that variables tend to move in the same direction most of the time, and same factors are likely to influence changes in this variable. The correlation coefficient wasn't chosen to be the highest from the heatmap, as the model needs to be able to learn patterns from the behaviour of the 2 correlated variables. If the variables are fully correlated, there would be nothing for the model to learn from. On the contrary, if the correlation between two variables is very low or almost zero, the behaviour of two variables is likely to be due chance and there is also little information for the model to learn.

After the model was trained, the main performance metrics were calculated, based on which the model's performance was assessed.

R-squared	RMSE	F-statistic	Log-Likelihood	AIC	P-value (NO2)
0.493	34.57	1.822e+04	-1.221e+05	2.442e+05	0.00

Table 2: Key metrics for evaluating linear regression model with PM10 as input and NO2 as output variables.

The R-squared value provides the information about the goodness of the fit of the model. The values of the R-squared range between 0 and 1, where zero represents lack of model's ability to predict data, and value of one means that the model can predict the relationship perfectly [2]. In this case, R-squared value of 0.493 means that the model was able to estimate the relationship for the 49.3% of the variation in independent values accounted for by dependant values.

One of the most important metrics is the Root Mean Squared Error (RMSE). It represents the average difference between values predicted by a model and the actual values. The RMSE is calculated by summing all of the model's errors squared, dividing by the number of points and taking a square root out of the sum. The predictive accuracy of the model is based on this metric, which shows the ability of the model to predict target values. The lower RMSE value is, the better the model. The RMSE value should be interpreted based on value range of the predicted variable. To evaluate it, we need to look at how the NO2 pollution data is distributed, as shown on table 1. 50% of the values lie between the first and 3rd quartiles, which represents the range between 33.93 and 83.63. The RMSE of 34.57 can be considered quite large in this case, as the it will account for more than 50% of this range.

The F-statistic metric shows whether linear regression provides a better fit to data than the model with no independent variables. It is used to understand how useful the selected variables in predicting the model's outcome. It is calculated by dividing the between-group variance and within-group variance. The between-group variance shows the dispersion of group means relative to the overall mean of the data. The within-group variance is a value which shows dispersion of the data within the group relative to the group mean. The higher the F-statistic, the better model's outputs estimate the data [3].

Log-Likelihood is another measure of how good the model fits the data. It can span from negative infinity to postive infinity. The log-likelihood is the product of the probability density function (PDF) or probability mass function (PMF) evaluated at each data point. The higher the value, the better model fits the data. It is only useful when two or more models are compared. Model that has the highest log-likelihood fits data the best [4].

Akaike Information Criterion (AIC) is a statistical measure used to compare models by assessing their performance on goodness of the fit and penalizing for model's complexity. It evaluates how well a model fits the data relative to the amount of information used in training, helping to identify the most suitable model for the given dataset. It uses the log-likelihood to calculate the goodness of the model's fit and adds a penalty term for models with higher parameter complexity. The lower the AIC value the better, which means the model with highest log-likelihood values and least parameters will have lowest AIC value [5].

P-value is a metric which shows how likely that the relationship between 2 variables could have occurred under the null hypothesis. Null hypothesis can be described as a statement that no relationship exists between two variables that are being analysed. This metric is used to identify if there is a relationship between two analysed variables. Usually, the p-value of less than 0.05 is considered significant, meaning that there is some kind of relationship present between two analysed variables.

After creating the model of PM10 as a predictor for NO2, all other pollutants were substituted one by one as a predictor variable, to check which single pollutant predicts a quantity of NO2 the best. After assesing all of the models on metrics described above, the model trained on carbon monoxide (CO) was found to have the highest R-squared and lowest RMSE values, with other key metrics displayed in the table below.

R-squared	RMSE	F-statistic	Log-Likelihood	AIC	P-value (NO2)
0.584	31.30	2.636e+04	-1.678e+0.5	3.356e+05	0.000

Table 3: Key metrics for evaluating linear regression model with CO as input and NO2 as output variables.

This model has better predictive ability in comparison to the last model. The R-squared value is bigger, meaning this model accounts for bigger part of the variation in the independent by the dependent values. It’s RMSE value is lower and at the same time the CO range is larger. It has bigger F-statistic, with larger values representing that the result occured not due chance. However, this model has a lower value of log-likelihood, because it is scale-dependent. Because carbon monoxide has bigger value range, the log-likelihood is lower, but it doesn’t mean that the model has a worse fit. The AIC value is also higher due to the spanning range of carbon monoxide, even the model was trained on the same data set but different variable.

The multivariate linear regression model was then created to predict NO2, using all other pollutants as predictors. From now on the RMSE statistic will not be calculated, as the weight for each of the predictors should be determined before calculating it, which lies outside of the scope of this report.

R-squared	F-statistic	Log-Likelihood	AIC	P-value (const)
0.870	1.792e+04	-80389	1.608e+05	0.005

Table 4: Key metrics for evaluating linear regression model with CO, NO, O3, SO2, PM2.5, PM10 and NH3 as input variables and NO2 as output variable.

This model has an R-squared value of 0.870, which is much higher than the values for the model with only CO as a predictor. An R-squared value above 0.75 is considered substantial, according to Henseler (2009) [6]. It has a high F-statistic, which means that variables in the model contribute to goodness of the fit to an NO2 predictor variable. Low log-likelihood with low AIC in comparison with previous models means that the model fits the data better and has a good fit/information used ratio. The model was then modified by deleting variables one by one, based on their p-values, which are shown on the table below, for each variable in the model.

	P>|t|
const	0.005
co	0.000
no	0.000
o3	0.000
so2	0.000
pm2_5	0.097
pm10	0.000
nh3	0.000

Table 5: The table showing p-values for each pollutant in linear regression model.

The variables with the p-values over 0.05 were deleted, in this case only PM2.5, which is considered non-significant and brings noise to the model. Metrics from the table 4 were analysed to understand which variables could be deducted from the model. After manually checking the statistics for various combinations of pollutants, the model with carbon monoxide (CO), nitrogen oxide (NO), ozone (O3), sulphur dioxide (SO2), ammonia (NH3) and particulate matter with 10 micrometres in diameter (PM10) was chosen to be the optimal, with the evaluation metrics presented in the table below.

R-squared	F-statistic	Log-Likelihood	AIC	P-value (const)
0.870	2.090e+04	-80390	1.608e+05	0.001

Table 6: Key metrics for evaluating linear regression model with CO, NO, O3, SO2, PM2.5, PM10 and NH3 as input variables and NO2 as output variable.

The k-fold cross-validation test was performed on the optimal model. K-fold cross-validation test splits the data into k equal parts, and then evaluates the performance of the model one these parts, one by one. This procedure makes evaluation of the model more reliable, because it evaluates its performance on 10 different parts of the data set, so there will be less chance that the model evaluation will be biased because of single evaluation sample [7]. The k-fold cross validation test was evaluated on a Root Mean Squared Error (MRSE) and Mean Absolute Error (MAE) values of 11.53 and 17.55 accordingly**.** The Mean Absolute Error represents the average of the absolute difference between actual and predicted values in the dataset. The scores for both metrics are relatively low compared to the value range of the nitrogen dioxide (NO2), and RMSE is lower by almost 2 times compared to the model with only carbon monoxide (CO) as a predictor, meaning this model fits data much better compared to the model with a single predictor variable.

Pulsar classifier models

Dataset

The Pulsar data set consists of 17,898 records for each of the 9 neutron star characteristics. Characteristics include mean, standard deviation, excess kurtosis, skewness of the integrated profile and DM-SNR curve, and a column which identifies whether neutron star is a pulsar or not. The data is stored in a CSV format and has no missing values.

Analysis

The analysis started with the general inspection of the dataset. The dataset was checked for missing values, which weren't present. The key metrics such as mean, standard deviation, and quartile values were calculated for each pollutant, which gave the idea on how variables are distributed in the dataset.

	Mean of int profile	Std of int profile	Excess kurtosis of int profile	Skewness of int profile	Mean of DM-SNR curve	std of DM-SNR curve	Excess kurtosis of DM-SNR curve	Skewness of the DM-SNR curve	target_class
count	17898	17898	17898	17898	17898	17898	17898	17898	17898
mean	111.08	46.55	0.48	1.77	12.61	26.33	8.30	104.86	0.09
std	25.65	6.84	1.06	6.17	29.47	19.47	4.51	106.51	0.29
min	5.81	24.77	-1.88	-1.79	0.21	7.37	-3.14	-1.98	0.00
25%	100.93	42.38	0.03	-0.19	1.92	14.44	5.78	34.96	0.00
50%	115.08	46.95	0.22	0.20	2.80	18.46	8.44	83.06	0.00
75%	127.09	51.02	0.47	0.93	5.46	28.43	10.70	139.31	0.00
max	192.62	98.78	8.07	68.10	223.39	110.64	34.54	1191.00	1.00

Table 7: Variable distribution in Pulsar dataset.

The number of pulsars were then determined from the dataset. The ratio of pulsars to the general quantity of neutron stars were calculated and visualized. The total number of neutron stars in this dataset is 17898, out of which 1639 are pulsars. It corresponds to a ratio of 9.16%.

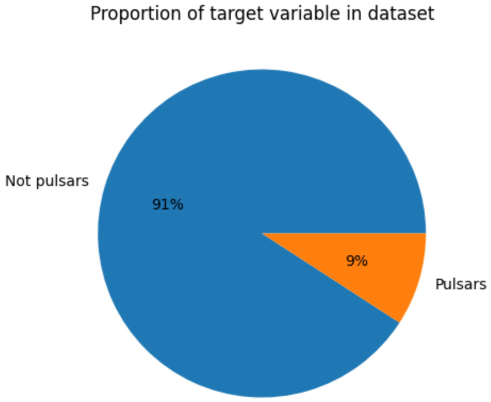

Figure 2: Ratio of pulsars to neutron stars.

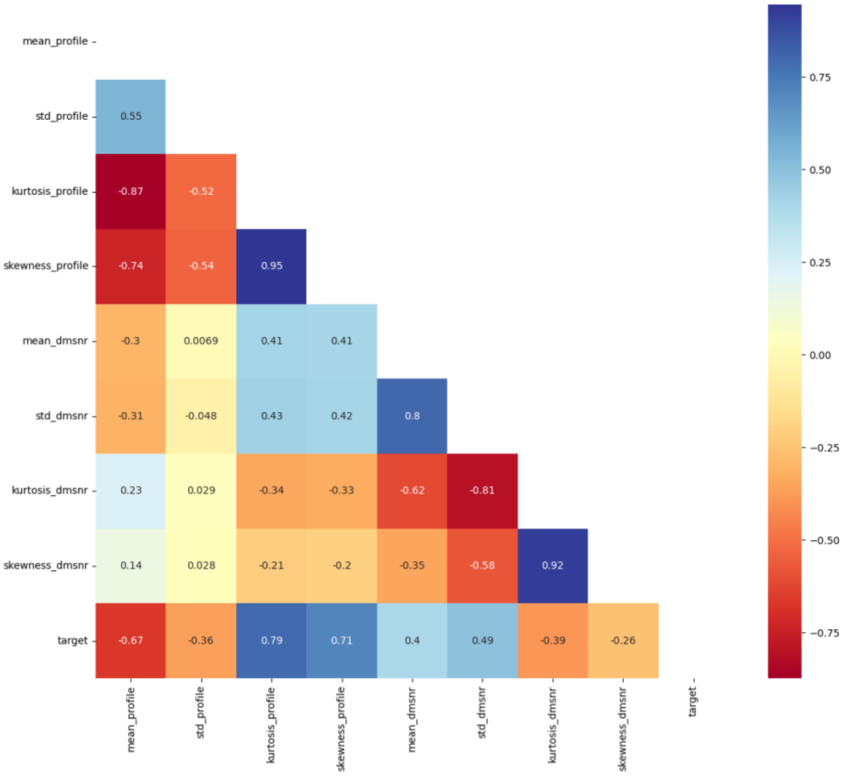

Figure 3: Correlation heatmap of variables in the Pulsar dataset.

The standard deviation of the integrated profile has almost zero correlation with majority of variables in the dataset, which means that changes in these two variables are not related to each other. Both DM-SNR and profile skewness and kurtosis have high positive correlations, as the kurtosis represents the tailedness of distribution and skewness is the measure of distribution’s asymmetry. Values which have highest correlation with a target class are kurtosis and skewness profile.

Naïve Bayes

Naïve Bayes classifier is a probabilistic machine learning algorithm which is used for classification tasks. It is based on the Bayes’ theorem for calculating probabilities and conditional probabilities. It is mainly used in text classification with high dimensional data, spam filtering and recommendation algorithms. Naïve Bayes is a simple and fast algorithm which works well with data with high dimensions [8]. However, it assumes independence between features which may not hold true in all situations, resulting in worse classifying ability in dataset with highly correlated features.

The distribution graph of each variable, along with corresponding QQ-plots were plotted for each predictor variable, to check if features are normally distributed. Some of the features like mean of integrated profile (mean_profile) and standard deviation of integrated profile (std_profile) have distributions close to normal. However, most of the features in the dataset have skewed distribution with long tails.

Since variables are continuous and have resemblance to normal distribution, Gaussian Naïve Bayes was chosen as a best guess for building a model.

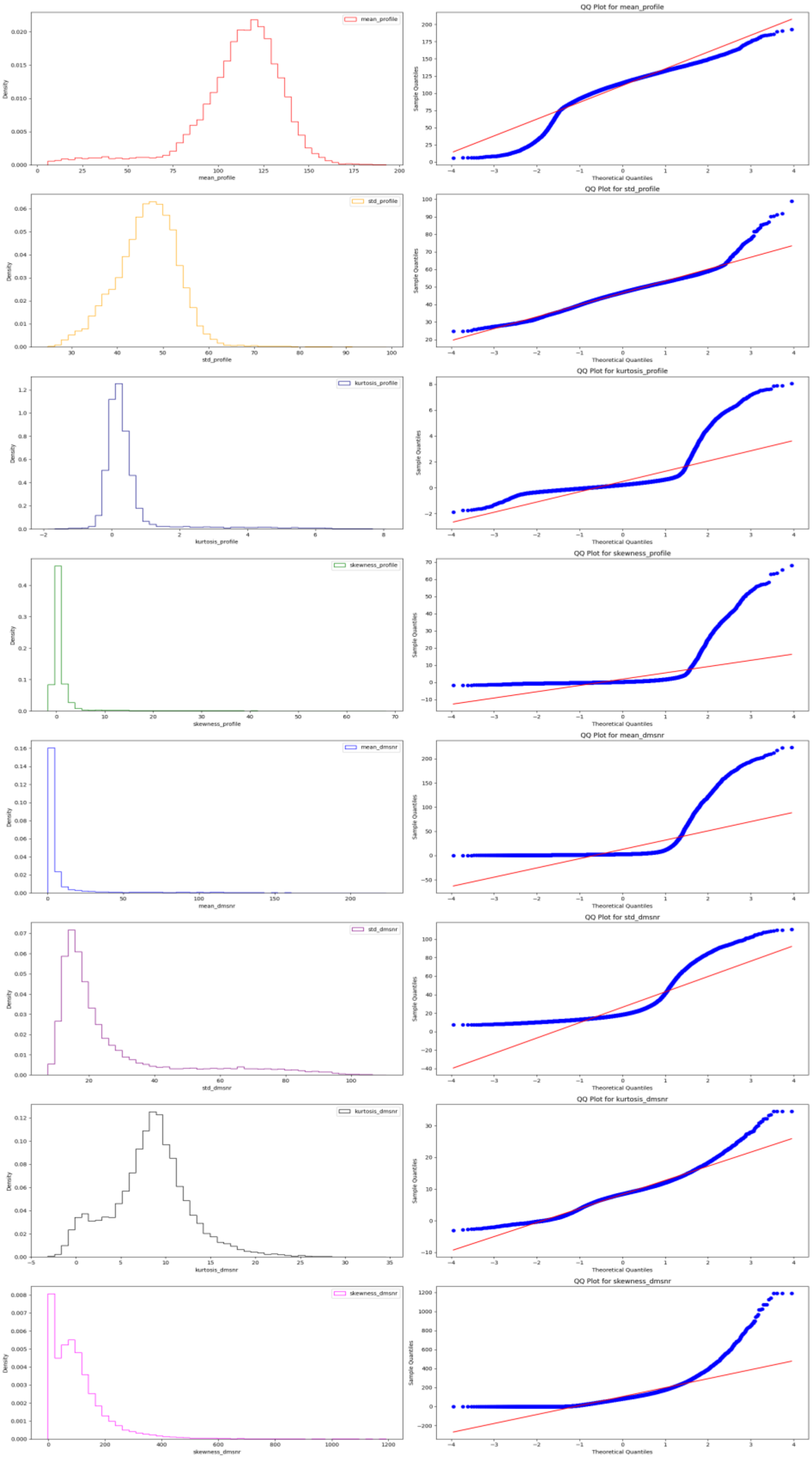

Figure 4: Histogram of every variable with corresponding QQ-plots to check whether the distribution is normal.

The data was then split into training and testing batches. The model will be trained on 85% and its performance tested on remaining 15% of the dataset, which accounts for 15213 and 2685 elements respectively. After the model was trained, the number of mislabelled points was calculated to be 116 out of 2685 of the training batch, accounting for 4.32% of mislabelled points. Confusion matrix was then created for further analysis of the model.

		Predicted	
		Not Pulsar	Pulsar
Actual	Not Pulsar	13154 (TP)	639 (FP)
	Pulsar	218 (FN)	1211(TN)

Table 8: Confusion matrix for Gaussian Naïve Bayes model.

A confusion matrix is a table that summarizes the performance of a classification algorithm. It shows the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions made by the model.

- True positive (TP): The number of instances correctly classified as positive by the model.

- True negative (TN): The number of instances correctly classified as negative by the model.

- False positive (FP): The number of instances incorrectly classified as positive by the model.

- False negative (FN): The number of instances incorrectly classified as negative by the model.

One of the main metrics upon which the performance of the model can be evaluated is accuracy. It is calculated by dividing the number of correct predictions by the overall number of measurements in confusion matrix.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Based on the values in confusion matrix, the Receiver Operating Curve (ROC) is plotted. ROC is a graph which shows of a classification model at classification thresholds. To build the ROC, 2 metrics should be calculated first, using values from confusion matrix. One of them is True Positive Rate (TPR), also known as recall, which is defined as follows:

$$Recall = \frac{TP}{TP + FN}$$

The second metric which is used to plot the ROC curve is the False Positive Rate (FPR), also known as specificity, which is defined as follows:

$$Specificity = \frac{FP}{FP + TN}$$

The False Positive Rate represents the x axis and 1 - True Positive Rate represents y axis on the ROC plot. The graph shows TPR and FPR rates at different classification thresholds. When lowering the classification threshold, more items are classified as positive, increasing False Positives and True Positives as a result [9].

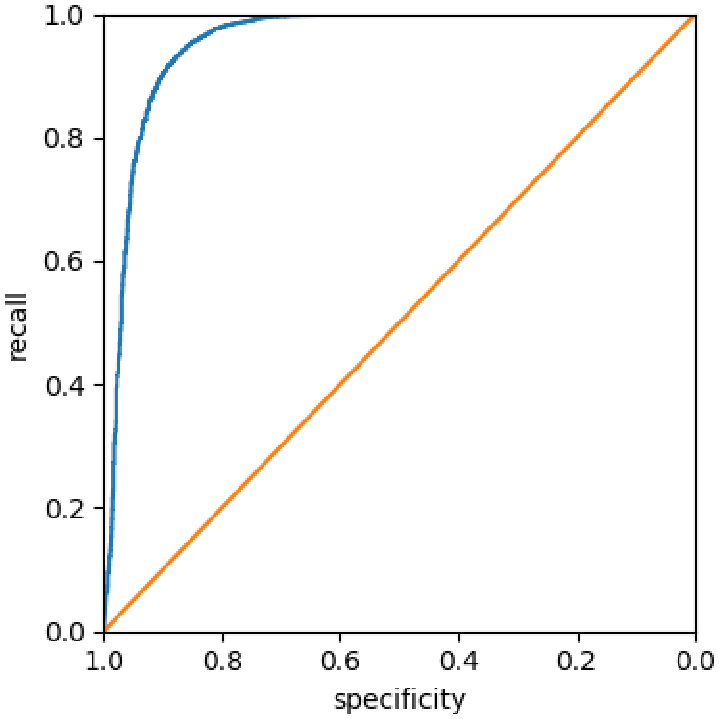

Figure 5: Receiver Operating Curve for Gaussian Naïve Bayes model.

The ROC curve visually shows how well a model discriminates between positive and negative classes across all possible threshold values. The closer the ROC curve is to the upper-left corner of the graph, the better the model at predicting positive over negative values.

The Area Under the Curve (AUC) is another metric used to evaluate the performance of a classification model. It quantifies the overall performance of the model across all possible threshold settings. The AUC represents the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

A higher AUC value indicates better model performance. A model with an AUC of 1.0 indicates perfect discrimination, while a model with an AUC of 0.5 suggests random guessing (no discrimination), which is represented by an orange diagonal line on the graph. Therefore, the AUC curve provides a single scalar value summarizing the model's performance across all possible classification thresholds.

In this report, models will be evaluated on the AUC and accuracy scores. The AUC value was calculated to be 0.956, meaning that model ranks positive examples more highly in 95.6% of cases, which is a good result. The accuracy for this model was calculated to be 0.944, meaning that model made correct prediction 94.4% of the times.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a statistical method used primarily for classification tasks, which can also be employed for dimensionality reduction. Its fundamental objective is to find a linear combination of features that best separates between classes in the dataset. Firstly, the original features are transformed into a new set of features, aiming to maximize the separation between classes while minimizing the variance within each class. This transformation seeks to locate a projection of the data where the means of different classes are as far apart as possible relative to the spread or variance within each class, ensuring well-separated classes in the transformed feature space.

LDA often leads to dimensionality reduction by projecting the data onto a lower-dimensional subspace while preserving as much discriminatory information as possible, which is particularly beneficial for high-dimensional datasets [10]. In the classification phase, LDA uses the discriminant functions learned during training to assign new data points to the class with the highest probability.

LDA assumes the data follows a multivariate normal distribution within each class and that the classes have identical covariance matrices. LDA is also sensitive to outliers, which are present for some variables in this dataset. Unlike logistic regression, LDA doesn't explicitly model the class probabilities; instead, it focuses on finding the optimal linear discriminant functions. It is widely used in pattern recognition and data mining.

The Linear Discriminant Analysis model was then created and trained with the same train/test split of 85% and 15% respectively. The scaling factors were then computed using linear discriminant factors for each variable that maximize class separation while minimizing within-class variance. The sign of scaling and magnitude of scaling factors indicate the direction of feature values relative to the overall dataset.

	Scaling factor
mean_profile	0.023
Std_profile	-0.012
Kurtosis_profile	3.121
Skewness_profile	-0.217
Mean_dmsnr	-0.007
Std_dmsnr	0.023
Kurtosis_dmsnr	-0.064

Table 9: Scaling factors for Linear Discriminant Analysis.

		Predicted	
		Not Pulsar	Pulsar
Actual	Not Pulsar	13769 (TP)	56 (FP)
	Pulsar	313 (FN)	1075 (TN)

Table 10: Confusion matrix for Linear Discriminant Analysis model.

This model has higher True Positive rate and lower True Negative rates than the Gaussian Naïve Bayes model. The biggest difference between two matrices lies in False Positive rate, which means this model wrongly classifies pulsars much less frequently. It has slightly bigger False Negative rate which implies that the model wrongly classifies neutron stars as not pulsars slightly more frequently. This model has smaller amount of falsely classified pulsars, represented by previous two metrics. The ROC was then built for the model.

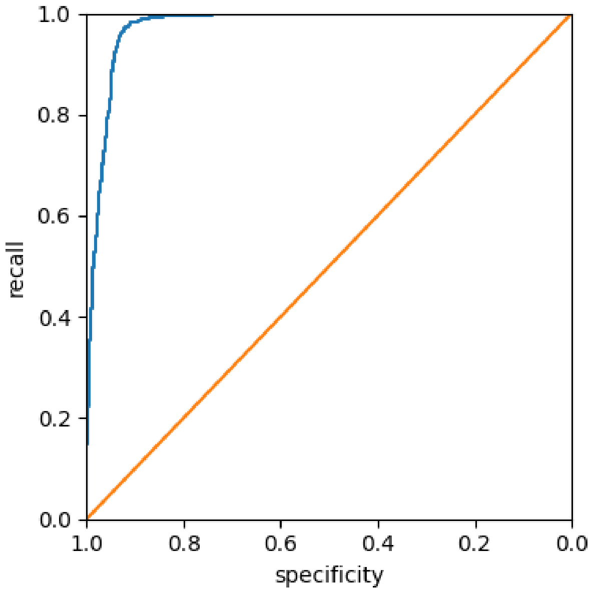

Figure 6: Receiver Operating Curve for Linear Discriminant Analysis model.

The ROC for this curve is slightly steeper, meaning that LDA achieves better true positive rates (recall) for a given false positive rate (1-specificity) compared to GNB. The accuracy was calculated to be 0.976, meaning it classifies neutron stars in 97.6% cases. The AUC value is also slightly higher, 0.975 as compared to 0.956. In practical terms, this implies that LDA has a more effective ability to correctly classify positive instances while minimizing false positives, leading to improved performance in binary classification tasks. The steeper ROC curve for LDA may result from its ability to capture more complex relationships and dependencies between features compared to Gaussian Naïve Bayes model.

Logistic Regression

Logistic Regression is a statistical method utilized for binary classification tasks, aiming to predict the probability that an observation belongs to one of two classes. Despite its name, logistic regression functions as a classification algorithm rather than a regression one. It models the relationship between one or more independent variables (features) and a binary outcome variable (target) using a logistic function. This function, also known as the sigmoid function, maps any real-valued number to a value between 0 and 1, representing the probability of the positive class [11].

During model training, logistic regression estimates the coefficients that minimize the difference between the predicted probabilities and the actual class labels in the training data, typically employing optimization algorithms such as gradient descent. Once trained, the model can predict the probability that a new observation belongs to the positive class, after which a common threshold (e.g., 0.5) is applied to convert these probabilities into class labels (0 or 1). Logistic regression assumes a linear relationship between the independent variables and the log-odds of the positive class and is considered a parametric model, making specific assumptions about the underlying distribution of the data [12].

		Predicted	
		Not Pulsar	Pulsar
Actual	Not Pulsar	13756 (TP)	69 (FP)
	Pulsar	233 (FN)	1155 (TN)

Table 10: Confusion matrix for Logistic Regression model.

By comparing the confusion matrices with previous models, confusion matrix for Logistic Regression has the best confusion matrix out of all three models. It has the lowest cumulative number of wrongly predicted values, False Negatives and False Positives, meaning it classifies neutron stars with higher accuracy compared to previous models.

ROC curves for Linear Discriminant Analysis and Logistic Regression look almost the same. In fact, they have the same AUC value of 0.976, meaning they have the exact same ability to correctly classify positive instances while minimizing false positives. However, this model has a better accuracy score of 0.980 compared to 0.976 for LDA, and overall can be considered the best model out of three classifier models.

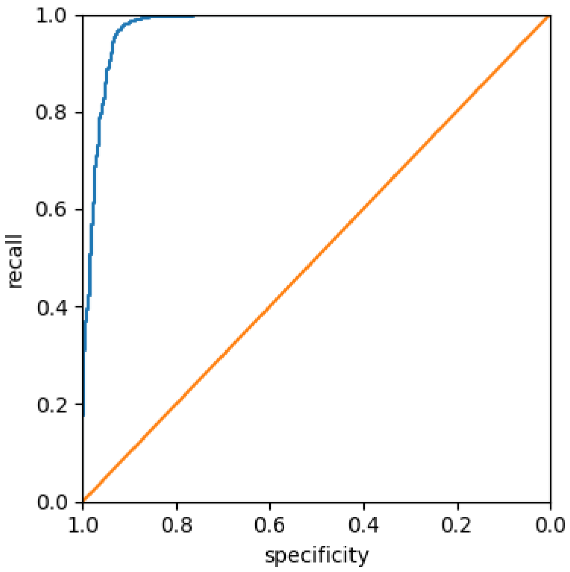

Figure 6: Receiver Operating Curve for Logistic Regression model.

Conclusion

This report focused on analysing two datasets. The first part of the report focused on analysing the air pollution levels in the city of Delhi using the linear regression model to predict the amine group (NH2) quantity based on pollution data from 8 different pollutants, which include carbon monoxide (CO), nitric oxide (NO), nitrogen dioxide (NO2), ozone (O3), sulphur dioxide (SO2) and particulate matter (PM2.5 and PM10) levels. It was found that carbon monoxide (CO) was the sole best predictor for the ammonia quantity, with the R-squared value of 0.584 and low RMSE value of 31.30, compared to value span of carbon monoxide.

Multivariate linear regression models were created and tested. They were compared against each other based on key metrics, including R-squared, F-statistic, Log-Likelihood, AIC and p-value criterions. The multivariate linear regression model with carbon monoxide (CO), nitric oxide (NO), ozone (O3), sulphur dioxide (SO2), particulate matter (PM10) level and ammonia (NH3) as predictors was found to be optimal, with R-squared value of 0.870 and F-statistic of 2.090e+04. The k-fold cross validation test was then employed to check how well the model fits data in different test batches. The MAE and RMSE values for k-fold test were calculated to be 11.53 and 17.55 respectively.

The second part of the report focused comparison of three classifier models trained on Pulsar dataset. The ratio of pulsar to neutron stars was calculated and the correlation analysis was performed for variables in the dataset. Three classification models, namely Naïve Bayes, Linear Discriminant Analysis (LDA), and Logistic Regression, were trained and evaluated for their performance. The models were created using the train/test split with 85% for training and 15% for testing data. Confusion matrices and ROC were plotted for every model and compared against each other. Models were then evaluated by accuracy and AUC scores. All models had accuracy values over 0.94 and AUC scores over 0.95, meaning they all have good predictive abilities. The best model out of three turned out to be a Logistic Regression model, with accuracy and AUC scores of 0.980 and 0.976 respectively.

Further refinement could involve investigating potential correlations between the neutron star characteristics to enhance model discrimination. By exploring more advanced feature engineering techniques and considering feature interactions, future iterations of the models could potentially yield even higher predictive accuracy. Additionally, conducting further studies to assess the impact of outliers and imbalanced data on model performance could contribute to improving the robustness and reliability of the classification models for pulsar detection.

Datasets

Pollution data in Delhi: https://www.kaggle.com/datasets/deepaksirohiwal/delhi-air-quality

Pulsar dataset: https://www.kaggle.com/datasets/spacemod/pulsar-dataset

References

[1] NASA Earth Observatory. (n.d.). *Chemistry in the sunlight*. https://earthobservatory.nasa.gov/features/ChemistrySunlight/chemistry_sunlight3.php

[2] Fernando, J. (2023, December 13). *R-Squared: Definition, Calculation Formula, uses, and Limitations*. Investopedia. https://www.investopedia.com/terms/r/r-squared.asp

[3] Frost, J. (2023, October 26). *How F-tests work in Analysis of Variance (ANOVA)*. Statistics by Jim. https://statisticsbyjim.com/anova/f-tests-anova/

[4] *Log-likelihood*. (n.d.). https://www.statlect.com/glossary/log-likelihood

[5] Kisslinger, C. (1996b). Aftershocks and Fault-Zone properties. In *Advances in Geophysics* (pp. 1–36). https://doi.org/10.1016/s0065-2687(08)60019-9

[6] Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. In *Advances in international marketing* (pp. 277–319). https://doi.org/10.1108/s1474-7979(2009)0000020014

[7] Brownlee, J. (2023, October 3). *A Gentle Introduction to k-fold Cross-Validation*. MachineLearningMastery.com. https://machinelearningmastery.com/k-fold-cross-validation/

[8] *What are naïve Bayes classifiers? | IBM*. (n.d.). https://www.ibm.com/topics/naive-bayes

[9] *Classification: ROC Curve and AUC*. (n.d.). Google for Developers. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

[10] Brownlee, J. (2020, August 14). *Linear Discriminant Analysis for Machine learning*. MachineLearningMastery.com. https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/

[11] *What is logistic regression? | IBM*. (n.d.). https://www.ibm.com/topics/logistic-regression

[12] Statistics Solutions. (2024, March 21). *Assumptions of Logistic Regression – Statistics Solutions*. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/