Journal of Data Analysis and Information Processing, 2023, 11, 217-239
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Application of Regularized Logistic Regression and Artificial Neural Network Model for Ozone Classification across El Paso County, Texas, United States

Callistus Obunadike, Adekunle Adefabi, Somtobe Olisah, David Abimbola, Kunle Oloyede

Department of Computer Science, Austin Peay State University, Clarksville, USA
Email: callistusobunadike@gmail.com, aadefabi@my.apsu.edu, solisah@my.apsu.edu, dabimbola@my.apsu.edu, koloyede@my.apsu.edu

How to cite this paper: Obunadike, C., Adefabi, A., Olisah, S., Abimbola, D. and Oloyede, K. (2023) Application of Regularized Logistic Regression and Artificial Neural Network Model for Ozone Classification across El Paso County, Texas, United States. *Journal of Data Analysis and Information Processing*, **11**, 217-239.
https://doi.org/10.4236/jdaip.2023.113012

Received: April 25, 2023
Accepted: July 8, 2023
Published: July 11, 2023

Abstract

This paper focuses on ozone prediction in the atmosphere using a machine learning approach. We utilize air pollutant and meteorological variable datasets from the El Paso area to classify ozone levels as high or low. The LR and ANN algorithms are employed to train the datasets. The models demonstrate a remarkably high classification accuracy of 89.3% in predicting ozone levels on a given day. Evaluation metrics reveal that both the ANN and LR models exhibit accuracies of 89.3% and 88.4%, respectively. Additionally, the AUC values for both models are comparable, with the ANN achieving 95.4% and the LR obtaining 95.2%. The lower the cross-entropy loss (log loss), the higher the model's accuracy or performance. Our ANN model yields a log loss of 3.74, while the LR model shows a log loss of 6.03. The prediction time for the ANN model is approximately 0.00 seconds, whereas the LR model takes 0.02 seconds. Our odds ratio analysis indicates that features such as "Solar radiation", "Std. Dev. Wind Direction", "outdoor temperature", "dew point temperature", and "PM10" contribute to high ozone levels in El Paso, Texas. Based on metrics such as accuracy, error rate, log loss, and prediction time, the ANN model proves to be faster and more suitable for ozone classification in the El Paso, Texas area.

Keywords

Machine Learning, Ozone Prediction, Pollutants Forecasting, Atmospheric Monitoring, Air Quality, Logistic Regression, Artificial Neural Network

1. Introduction

Ozone is created in the atmosphere from gases that are released through smokestacks, tailpipes, and a variety of other sources. These gases react when exposed to sunlight, thereby creating ozone pollution. Ozone is a key component of the Earth's atmosphere; it plays a vital role in protecting life on our planet + by absorbing harmful ultraviolet radiation. However, excessive levels of ozone can have negative impacts on human health and the environment. Ozone prediction is an important task that helps us to better understand and manage the effects of ozone on our planet. Application of Machine learning serves as a powerful mechanism that helps to predict ozone levels in the atmosphere. This could be achieved by training a machine learning model on historical data, to make predictions about future ozone levels based on various factors such as temperature, wind speed, and emissions. These predictions can then be used to inform decision making and mitigate the negative effects of excessive ozone levels. Ozone starts off as an invisible pollution when not properly monitored it combines with other contaminants to cause lots of health challenges [1]. Ozone happens to be one of the most dangerous elements on earth. For the past several decades, researchers have been examining how ozone affects human health. In El Paso, Texas, United States, ozone level has been recorded as the highest affected city across the United States. Three oxygen atoms make up the gas molecule known as ozone (O_3). Ozone, also known as "smog", is dangerous to breathe. By chemically interacting with lung tissue, ozone actively damages it.

2. Literature Review

Three oxygen atoms make up the gas molecule known as O_3 (see **Figure 1**). Another name for ozone (O_3) is "smog", which is very dangerous when inhaled. Ozone becomes very harmful when it chemically interacts with the lung tissue, thus causing severe damage. **Figure 1** illustrates ozone molecules.

2.1. Formation of Ozone (O_3)

The same processes that produce ozone also produce other dangerous pollutants when O_3 is present. Although, we are protected from the majority of the sun's UV radiation by the ozone layer, which is located high in the stratosphere (*i.e.*, upper atmosphere). However, O_3 air pollution poses major health risks when it is present at ground level where we may breathe it (*i.e.*, within the troposphere). Nitrogen oxides (NOx) and volatile organic compounds (VOCs) are the two main raw materials that produce ozone. In addition, burning of fossil fuels like gasoline, oil, or coal or the evaporation of certain chemicals like solvents also contributes to ozone production. Power plants, automobiles, and other high-heat combustion sources all emit NOx whereas vehicles, chemical plants, refineries, factories, petrol stations, paint, and other sources all release VOCs [1]. **Figure 2** shows the reaction that leads to ozone formation pattern.

2.2. Risk of Ozone Exposure

Anyone who spends time outside in an area with high levels of ozone pollution

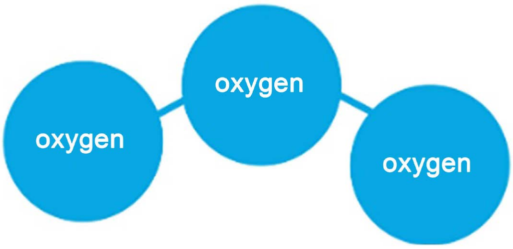


Figure 1. Showing ozone (O_3) molecule.

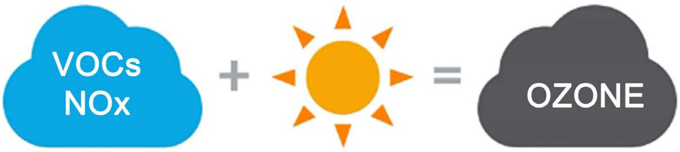


Figure 2. Showing ozone formation.

could be in danger. The effects of inhaling ozone are particularly harmful to four types of people:

- Children and teenagers [2].
- Everyone above the age of 65 [2].
- Those who already have lung conditions including asthma and chronic obstructive pulmonary disease (COPD), which also encompasses emphysema and chronic bronchitis [2].
- Those who work or exercise outside [2].
- People living with obesity [2].

2.3. Implications of Ozone Exposure

People with allergies may respond more strongly to allergens after inhaling ozone. Children were more likely to experience hay fever and respiratory allergies when ozone and PM2.5 levels were high, based on research study that was published in 2009 [3].

2.3.1. Premature Death

When exposed to the ozone layer, one's life may be shortened. From several research carried out in cities across the U.S., Europe, and Asia, it is obvious that ozone has a devastating effect on people's health and life span. Over time, researchers have discovered that exposure to increasing ozone levels raised the chance of premature death [4]. Even when other pollutants are also present, ozone raises the chance of premature mortality, according to more recent research [1].

2.3.2. Inhalation Problems

In major counties across the United States (like: El Paso, Texas), ozone level increases over the summer thus leading to increase in health challenges [5]. In addition to a higher risk of premature mortality, inhalation challenges like wheezing, coughing, and shortness of breath; asthma episodes; increases the need for hospitalization and medical care for persons with lung disorders including asthma or chronic obstructive pulmonary disease (COPD), as well as higher risk of respiratory infections, susceptibility to pulmonary inflammation, and risk of respiratory infections [2].

2.3.3. Risk from Long-Term Exposure

Recent research alerts us to the negative consequences of prolonged exposure to ozone. Scientists are discovering that prolonged exposure (*i.e.*, radiation exposure > 8 hours as well as days, months, or years) increases the chance of premature mortality. Researchers have discovered that high levels of ozone are linked to an increased risk of respiratory disease which leads to a high mortality rate [4]. Also, New York researchers examined hospital data for pediatric asthma patients and discovered that exposure to ozone over an extended period increased the probability of hospital admission for asthma patients. Recent studies show that kids from low-income households were more likely to be hospitalized due to high levels of ozone exposure as against kids from high-income households [6].

2.3.4. US Environmental Protection Agency (EPA) Findings

In February 2013, EPA published a comprehensive review of their most recent findings on ozone pollution [7]. EPA had asked the "*Clean Air Scientific Advisory Committee*", a group of distinguished scientists, to assist them in evaluating the evidence that was gathered by EPA; in particular, they looked at research published between 2006 and 2012. The EPA and the committee's experts concluded that ozone pollution posed numerous, substantial health risks. Based on that evaluation in 2015, the EPA firmly supports the "*National Ambient Air Quality Standard*" (*i.e.*, the official ozone acceptable limit). However, recent studies show that ozone can be dangerous even at much lower concentrations. In a scientific paper published in 2017, researchers presented additional proof that confirms that older adults face a higher risk of premature death even with low ozone levels beyond the national acceptable level [8].

2.4. Features or Variable Types

According to [9], the predictor variables could otherwise be known as "PIE (predictor, independent or explanatory) variables" while the response variables could otherwise be termed "DORT (dependent, observatory, response or target) variables". Features (variables) importance enables the ML algorithm to train faster as well as reduces cost and time required for training the dataset, therefore making it simpler to interpret. It also reduces the variance of the model and improves the accuracy, provided the right subset is chosen [9].

Odds Ratio

Generally, the intensity of the odds ratio is called the "***strength of the association***". The further away an odds ratio is from 1, the more likely it is that the relationship between the exposure and the disease is causal. For instance, an odds ratio of 1.25 is above 1, but is not a strong association while that of > 9.5 suggests a stronger association [9].

2.5. Selection of Logistic Regression and Artificial Neural Network Model

It's important to note that the choice between LR and ANN models depends on the ***specific problem***, ***dataset***, and ***desired outcome***. LR is suitable for simpler tasks and when interpretability is crucial, while ANN models excel in more complex problems where high accuracy is the priority.

2.5.1. Advantages of LR and ANN

The Logistic Regression model is straightforward and interpretable. It's easy to understand and implement, making it a good choice for simple classification problems [10]. Training an LR model is computationally efficient compared to complex ANN models. It can handle large datasets with relative ease [11]. LR provides meaningful insights into the impact of each feature on the predicted outcome. It assigns weights to features, indicating their importance in the decision-making process [12].

Artificial Neural Networks (ANNs) can model complex and nonlinear relationships between features and the target (DORT) variable. They can learn intricate patterns that may be difficult for LR models to capture. ANN models can automatically extract relevant features from raw data, reducing the need for manual feature engineering [13]. ANN models, especially deep learning models, have achieved state-of-the-art performance on various tasks, including image and speech recognition, natural language processing, and recommendation systems [14].

2.5.2. Disadvantages of LR and ANN

The Logistic Regression model assumes a linear relationship between features and the target variable. It may struggle to capture complex patterns and nonlinear relationships in the data [15]. Logistic Regression relies heavily on manual feature engineering. Thus, choosing relevant features and transforming them appropriately is crucial for its performance. LR performs well in certain scenarios, it may underperform when faced with highly complex datasets or problems that require high predictive accuracy [16].

Artificial Neural Network models, especially deep neural networks, require significant computational resources and can be time-consuming to train and they often require specialized hardware like GPUs [17]. ANN models can be challenging to interpret and understanding how the model arrives at its predictions can be difficult, making it less transparent compared to LR models [18]. In addition, ANN models are prone to overfitting, especially when working with limited training data [19]. Regularization techniques and careful hyperparameter tuning are necessary to mitigate this risk [5].

2.6. Factors that Influence Accuracy of LR and ANN

It's important to consider these factors and carefully optimize them to achieve the best classification accuracy for LR and ANN models.

- ***Dataset quality and size***: The quality and size of the dataset used for training and evaluation play a crucial role. A larger dataset with a diverse range of samples can help both LR and ANN models generalize better and achieve higher accuracy [18].
- ***Feature selection and engineering***: The choice and preparation of input features can significantly affect model performance, proper feature selection and engineering can improve the discriminative power of the features and lead to better accuracy for both LR and ANN models [20].
- ***Model complexity***: The complexity of the model can impact classification accuracy. LR assumes a linear relationship, while ANN models, especially deep neural networks, can capture complex nonlinear relationships [17]. In general, more complex models like ANNs have the potential to achieve higher accuracy, but they are also more prone to overfitting [19].
- ***Regularization techniques***: Regularization methods, such as L1 or L2 regularization, can help prevent overfitting in both LR and ANN models. Regularization adds a penalty term to the model's objective function, discouraging overly complex models and improving generalization [5].
- ***Hyperparameter tuning***: Both LR and ANN models have various hyperparameters that need to be tuned for optimal performance. Examples include ***learning rate***, ***regularization strength***, ***number of hidden layers***, and ***number of neurons***. Proper hyperparameter tuning can significantly affect classification accuracy [18].
- ***Training duration and convergence***: The duration and convergence of the training process can impact final accuracy. Training for too few iterations may result in ***underfitting***, while training for too many iterations may lead to ***overfitting*** [19]. Finding the right balance and ensuring convergence is essential for achieving high accuracy.
- ***Class imbalance***: Class imbalance occurs when one class has significantly more or fewer samples than others. This can affect the model's ability to accurately predict the minority class. Techniques like oversampling, under sampling, or class weighting can help address class imbalance and improve accuracy [21].
- ***Preprocessing and normalization***: Proper preprocessing steps, such as handling missing values, scaling features, and handling outliers, can impact the accuracy of both LR and ANN models. Different preprocessing techniques may be more suitable for different models, and their proper application can enhance accuracy [22].
- ***Model evaluation and validation***: The choice of appropriate evaluation metrics and validation techniques can affect the reported accuracy. Metrics such as ***accuracy***, ***precision***, ***recall***, and ***F*1 *score*** provide different perspectives on model performance, and using appropriate validation methods like ***cross-validation*** can give a more reliable estimate of the model's accuracy [23].

- ***Computational resources***: ANN models, especially deep learning models, can be computationally intensive and may require specialized hardware, such as GPUs, for efficient training. The availability of computational resources can impact the size and complexity of the ANN models used, which can, in turn, affect their accuracy [17].

3. Methodology

The aim and objective of this research is to examine the prediction ability of Logistic Regression and Artificial Neural Network Models in correctly classifying ozone levels into high and low categories, considering other predictor variables. The dataset consists of 973 rows and 14 variables (features). Among these variables, ozone was selected as the response (dependent) variable, with “low ozone” assigned as 0 and “high ozone” assigned as 1. Therefore, we are dealing with a binary classification problem. The dataset was analyzed using the R programming language. The first step in the analysis involved checking the variable types, identifying missing values, outliers, and potentially incorrect records, and conducting exploratory data analysis (EDA), including frequency distribution of the target variables and the association between them.

3.1. Descriptive Analysis of the Dataset

The dataset contains 14 variables (features), ‘ozone’ is assigned to be the target variable otherwise known as *DORT or Y* (dependent, observatory, response, or target variables) while the remaining 13 variables represents *PIE or X* (Predictor, Independent or explanatory variables) (see **Table 1**).

3.2. Data Pre-Processing

The following steps were adopted during the data prr-processing to ensure the accuracy of our dataset. We performed exploratory data analysis on the dataset by cleaning the data and checking for missing values (refer to section 3.2.1). Additionally, we applied a for-loop to iterate over the dataset and cross-check for other types of missing values, such as “na”, “NA”, or an empty string (refer to **Figure 3**). Based on the output or results, our dataset showed no missing values.

3.2.1. Checking for Missing Values

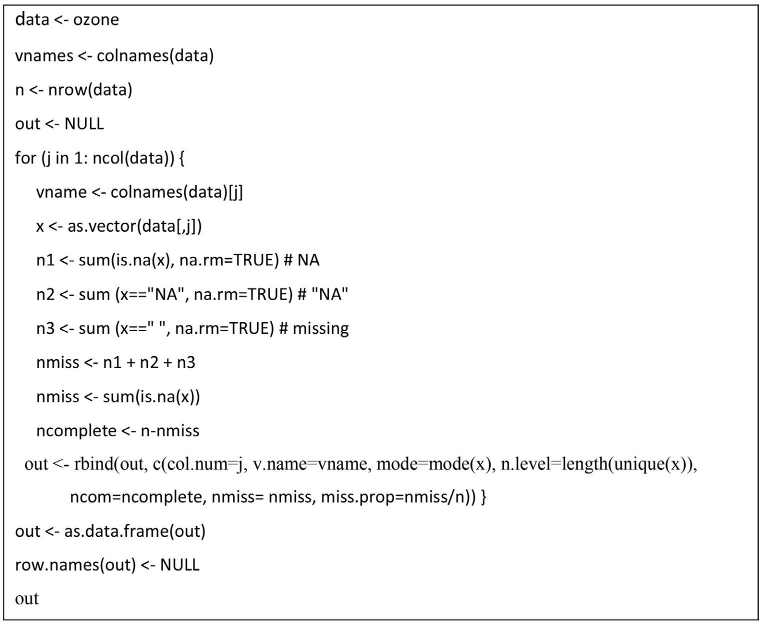
The anyNA () function was used to check for missing variables in our dataset. The outcome was “False”. Thus, it implies that we did not have any missing data. In addition, we went further to visualize if there was any sort of missing data using the naniar package. **Figure 4** shows a bar plot depicting there are no missing variables in our dataset.

3.2.2. Intensive Cross Checking of Other Missing Values

To ensure that our analysis and model would be free from errors. It is very important to thoroughly loop through the whole dataset to check for other missing values that may occur in other forms apart from “NA”.

Table 1. Description of the variable's data types.

S/No	Variables	Data Type
1	Nitric Oxide	Numeric
2	Nitrogen Dioxide	Numeric
3	Oxides of Nitrogen	Numeric
4	Wind Speed	Numeric
5	Resultant Wind Speed	Numeric
6	Resultant Wind Direction	Integer
7	Maximum Wind Gust	Numeric
8	Std. Dev. Wind Direction	Integer
9	Outdoor Temperature	Numeric
10	Dew Point Temperature	Numeric
11	Relative Humidity	Numeric
12	Solar Radiation	Numeric
13	PM10	Numeric
14	Ozone	Integer



```
data <- ozone
vnames <- colnames(data)
n <- nrow(data)
out <- NULL
for (j in 1: ncol(data)) {
    vname <- colnames(data)[j]
    x <- as.vector(data[,j])
    n1 <- sum(is.na(x), na.rm=TRUE) # NA
    n2 <- sum (x=="NA", na.rm=TRUE) # "NA"
    n3 <- sum (x==" ", na.rm=TRUE) # missing
    nmiss <- n1 + n2 + n3
    nmiss <- sum(is.na(x))
    ncomplete <- n-nmiss
  out <- rbind(out, c(col.num=j, v.name=vname, mode=mode(x), n.level=length(unique(x)),
          ncom=ncomplete, nmiss= nmiss, miss.prop=nmiss/n)) }
out <- as.data.frame(out)
row.names(out) <- NULL
out
```

Figure 3. Showing the (for-loop) code iteration for missing values.

Table 2 shows V.name or the variable names, Mode (data types), N. level (number of occurrences out of the total observations), Ncom (number of total observations), Nmiss (number of missing observations), and Miss. Prop (percentage of missing observations).

3.2.3. Frequency Distribution of Target Variable (Ozone)

Analyzing the ozone level, transformed the ozone level from binary-numerical to

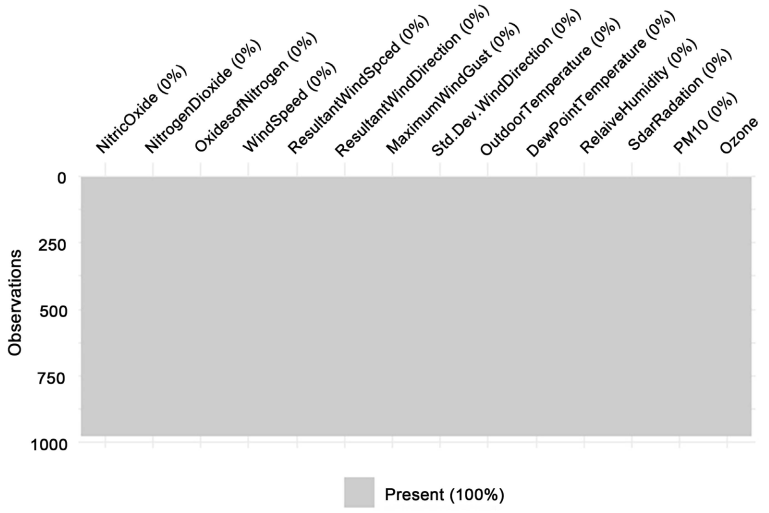


Figure 4. Showing barplot of the dataset using vis_mis () function in naniar package.

Table 2. Iteration through the dataset using for loop to check for other missing values.

Col.num	V. name	Mode	N.level	ncom	nmiss	Miss.prop
1	Nitric Oxide	numeric	82	973	0	0
2	Nitrogen Dioxide	numeric	228	973	0	0
3	Oxides of Nitrogen	numeric	238	973	0	0
4	Wind Speed	numeric	146	973	0	0
5	Resultant Wind Speed	numeric	153	973	0	0
6	Resultant Wind Direction	numeric	270	973	0	0
7	Maximum Wind Gust	numeric	242	973	0	0
8	Std. Dev. Wind Direction	numeric	68	973	0	0
9	Outdoor Temperature	numeric	367	973	0	0
10	Dew Point Temperature	numeric	442	973	0	0
11	Relative Humidity	numeric	459	973	0	0
12	Solar Radiation	numeric	549	973	0	0
13	PM10	numeric	451	973	0	0
14	Ozone	numeric	2	973	0	0

binary-categorical such that values of 1 are given "high level" while values of 0 are given "low level". From the frequency distribution plot below, the days with low ozone levels occur more frequently than those of high ozone level. Comparing the difference between both rates however, we can say the distribution is a bit balanced since the difference is not significantly large (see **Figure 5**).

4. Results

The first approach to building a model with high accuracy is to properly investigate

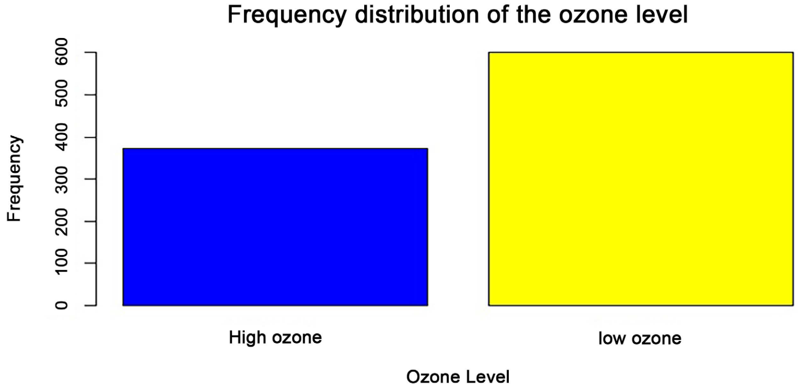


Figure 5. Frequency distribution of the target variable (ozone).

data quality, coherence, association, and correlations between the *DORT* and *PIE* variables. This will thus enable us to correctly predict areas with high ozone and low ozone effectively. Analyzing the output, we observe that all the variable types are continuous (quantitative) except for target (ozone) variable which is binary (categorical).

4.1. Data Exploration and Correlation

It is important to understand the degree of correlation and association between the predictor variables with the target variable (ozone). The correlations were computed with the data, and it shows different degrees of correlations ranging from strong negative to strong positive correlation (see Figure 3). Out of the 13 variables only "solar radiation", "outdoor temperature" and "std. dev. wind direction" show positive correlation with the target variable (ozone). Figure 6 illustrates predictor variables that are either positively or negatively correlated with the target variable (ozone).

4.2. Box Plots Predictor Variables

Figure 7 and Figure 8 show the boxplots of the predictor variables that are positively and negatively correlated with the target variable (ozone). In general, it could be seen that the negative correlated predictor variables have outliers.

4.2.1. Box Plots of Ozone and Selected Predictor Variables

The effect of selected predictor features (*i.e.*, "Solar Radiation", "Nitric Oxide", "Nitrogen Dioxide" and "PM10") differences used in determining the ozone level of a particular day (see Figures 9-12). The "high ozone" rate is nearly normal in most of the distributions as against the "low ozone level" with a negative skewness in distribution (see Figures 13-16). The histograms of "Solar Radiation", "Nitric Oxide", "Nitrogen Dioxide" and "PM10" show right-skewed distribution (see Figures 13-16). This already suggests that the distribution of the "Solar Radiation", "Nitric Oxide", "Nitrogen Dioxide" and "PM10" is not normal. Aside from Solar radiation, the other selected variables show presence of outliers.

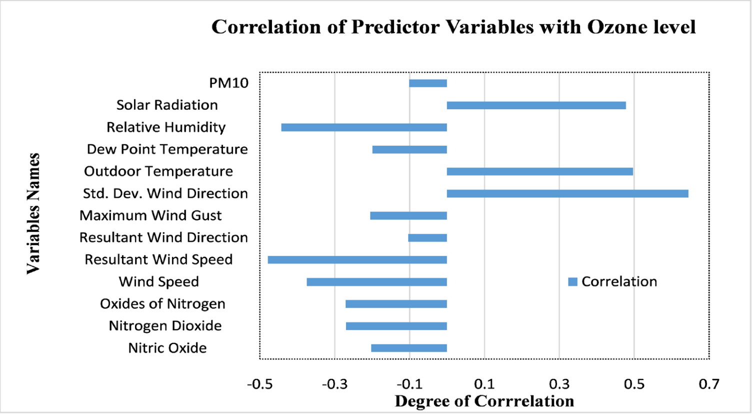


Figure 6. Correlation Coefficient between predictor variables and target (ozone) variable.

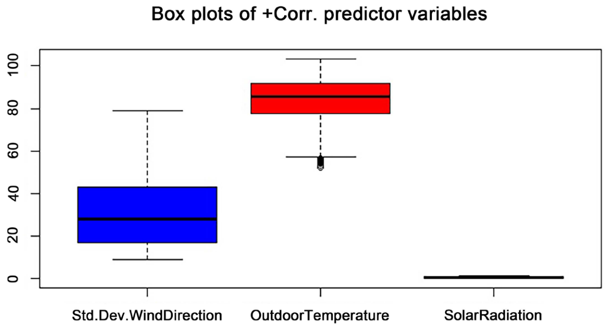


Figure 7. Boxplots of +Corr. predictor variables.

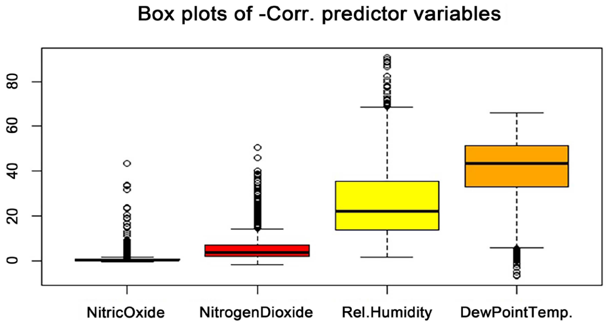


Figure 8. Boxplots of -Corr. predictor variables.

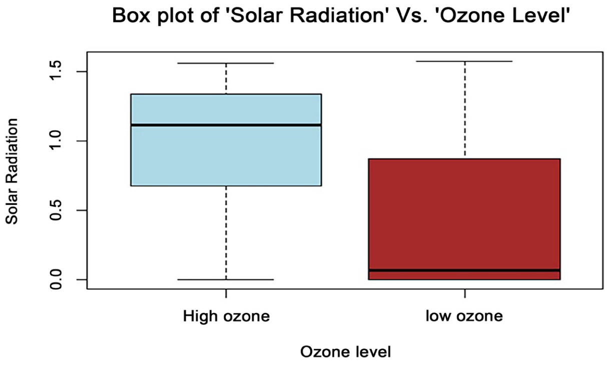


Figure 9. Boxplots of solar radiation vs ozone.

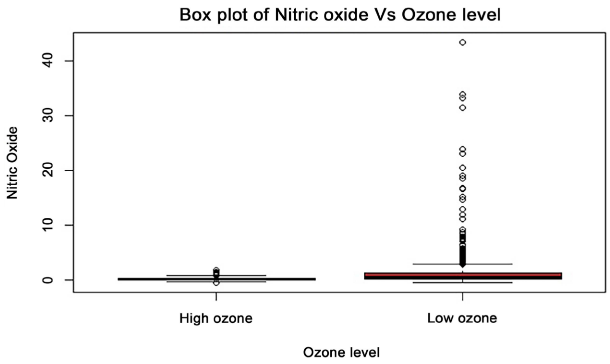


Figure 10. Boxplots of nitric oxide vs ozone.

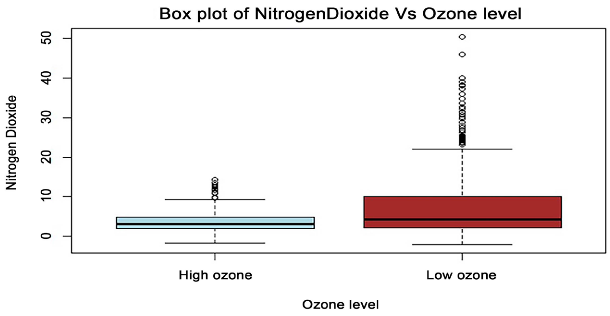


Figure 11. Boxplots of nitrogen dioxide vs ozone.

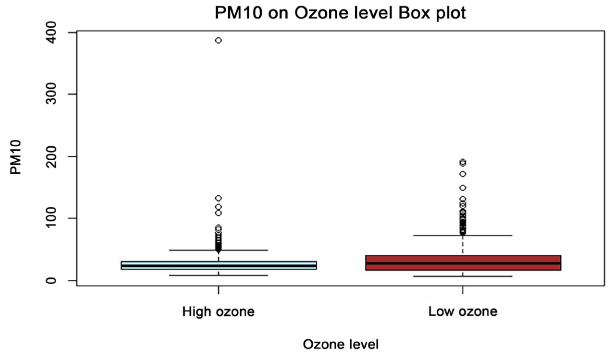


Figure 12. Boxplots of PM10 vs ozone level.

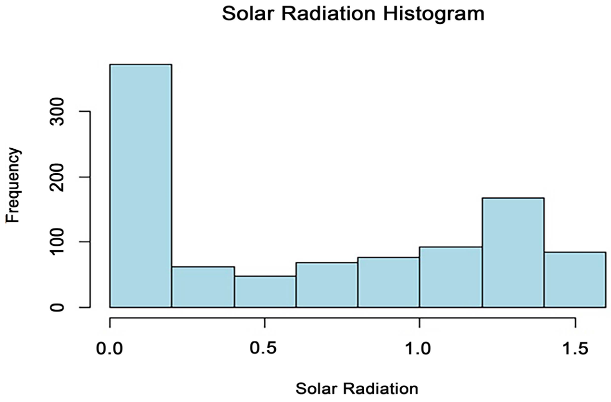


Figure 13. Solar radiation histogram.

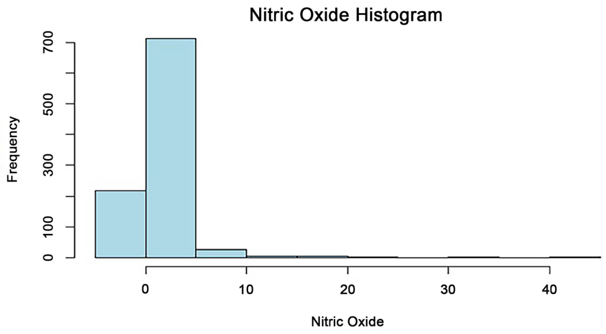


Figure 14. Nitric oxide histogram.

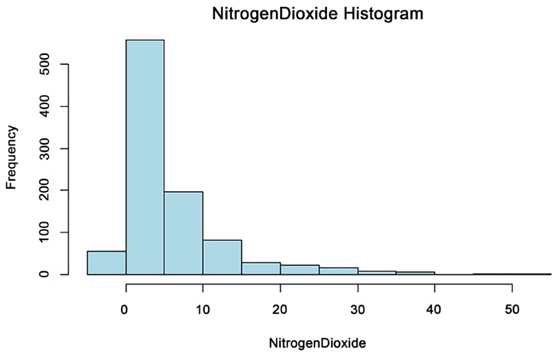


Figure 15. Nitrogen dioxide histogram.

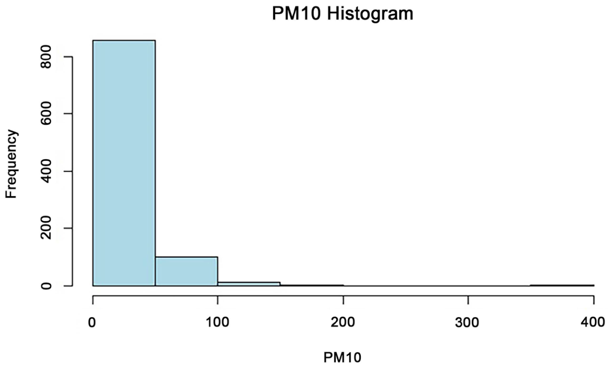


Figure 16. Particulate matter 10 histogram.

Since some of the box plots for the selected predictor variables have outliers and skewness, we would proceed to test for normality using the Anderson Darling and Shapiro-Wilk tests (see Table 3).

4.2.2. Normality Test and Wilcoxon Test

The normality test shows that the p-value of Anderson-Darling and Shapiro test are less than the significance level (0.05), which signifies that the distribution is not normal. Since the assumption of t-test is violated, we apply Wilcoxon rank sum test (non-parametric alternative test) to examine the association between Solar Radiation, Nitric Oxide, Nitrogen Dioxide and PM10 on ozone level (see Table 3).

Table 3. Association between target variable (ozone) and response variable (solar radiation).

	Solar Radiation Vs Ozone Level	Nitric Oxide Vs Ozone Level	Nitrogen Dioxide Vs Ozone Level	PM10 Vs Ozone Level
Anderson-Darling normality test	A = 54.466 p-value < 2.2e−16	A = 210.08 p-value < 2.2e−16	A = 76.434 p-value < 2.2e−16	A = 63.252 p-value < 2.2e−16
Shapiro-Wilk normality test	W = 0.84817 p-value < 2.2e−16	W = 0.31033 p-value < 2.2e−16	W = 0.7364 p-value < 2.2e−16	W = 0.64781 p-value < 2.2e−16

Wilcoxon rank test	W = 177795
	p-value < 2.2e−16
		W = 61823
		p-value < 2.2e−16
			W = 84314
			p-value = 9.622e−11
				W = 100056
				p-value = 0.005453

Comments	The p-value of the Wilcoxon rank sum test above is lower than the alpha value (0.05) indicating that there is significance relationship between ozone level and Solar Radiation/Nitric Oxide/Nitrogen Dioxide/PM10.			
	data: Alternative hypothesis (H_A): true location shift is not equal to 0			

4.3. Data Splitting

The dataset was partitioned into two parts with a ratio of 2:1, where the training data (D_1) has 67%, and the test data (D_2) takes 33%. Logistic regression technique was applied to the train data to build a predictive model. Firstly, we adopted the lasso regularization (L_1) with penalty to obtain the tuning parameter (λ) with cross validation. The logistic regression model was fitted with lasso regularization method using our trained data, D_1. The lasso method was applied because our aim is to build a parsimonious model which will properly explain our target (ozone) feature.

4.3.1. MSE and Tuning Parameter

The best lambda to regularize our model is evaluated using the MSE and miss-classification rate. The result of the first six rows of the Lambda's, miss classification rate and mean square error is shown in Table 4. Using MSE (Mean Square Error) as the evaluation metric. Our Best Lambda (tuning parameter) is 0.0026.

4.3.2. Mean Square Error vs Lambda

The plot below indicates that as the tuning parameter increases, the Mean Square Error increases as well. Therefore, it is important to keep Lambda very minimal to obtain low MSE. However, at 0.3 Lambda, the MSE becomes constant (see Figure 17).

4.4. Model Fitting and Odds Ratio of LR Model

Having gotten the best lambda. We fit the final Lasso Logistic regression model with the Training and Validation data pooled together. The important features can be seen from Table 5. After fitting the model with the best lambda, both "*Nitrogen Dioxide*" and "*Resultant Wind Speed*" happen to be the unimportant variables in our LR model (see Table 5).

Table 4. Lambda's, misclassification rate and MSE matrix.

	[1]	[2]	[3]
[1]	0.00010	0.355	0.2500
[2]	0.00261	0.126	0.0891
[3]	0.00512	0.135	0.0930
[4]	0.00764	0.138	0.0982
[5]	0.01015	0.145	0.1017
[6]	0.01266	0.145	0.1041

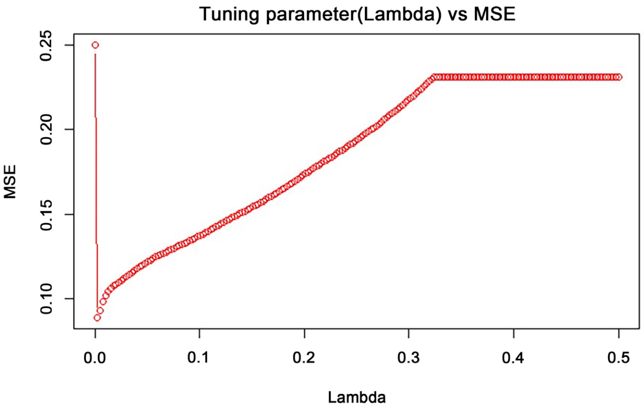


Figure 17. Showing plot of lambda vs MSE.

Table 5. Coefficients of important predictors using LGR (l_1) model.

Variables	Coefficients
Nitric Oxide	−1.98759
Nitrogen Dioxide	.
Oxides of Nitrogen	−0.00638
Wind Speed	−0.25833
Resultant Wind Speed	.
Resultant Wind Direction	−0.00391
Maximum Wind Gust	−0.01733
Std. Dev. Wind Direction	0.06886
Outdoor Temperature	−0.00652
Dew Point Temperature	0.04620
Relative Humidity	−0.13025
Solar Radiation	1.23665
PM10	0.01135

The negative coefficients of "Nitric Oxide" indicates that a slight increase in "Nitric Oxide" multiplies the odd ratio by a number < 1 which effectively increases the probability of the output being labeled as low ozone level (0). In addition, the positive coefficients of "Solar Radiation" suggests that a unit increase in the variable "Solar Radiation" multiplies the odd ratio by a number greater than one which effectively increases the probability of the output being labeled as high ozone level (1). We will then use the best-fitted model on our test data. Table 6 below presents the odds ratio of important predictor variables based on the best fit model.

4.4.1. Logistic Regression Model Evaluation

From Table 7, the AUC of 0.952 indicates that our fitted model has 95% ability to correctly classify ozone level into high or low. The confidence interval also indicates the true AUC falls within the interval (0.929, 0.975). Therefore, we are 95% confident that our AUC is accurate. From Table 7, we obtained an MSE value of 0.0833 which generally indicates a good performance for our model. We further computed the miss-classification rate since this is a logistic regression model and MSE is not the best evaluating method. The miss-classification rate value is 0.116 which means that our model correctly predicts the ozone levels into high and low ozone at a rate of 88.4% which suggests that our model performs well. The AUC of 0.952 implies that our best fitted model has 95.2% accuracy to predict if the ozone level is either high or low for a particular day. The C.I also indicates the true AUC falls within the interval (0.929, 0.975). Therefore, we are 95% confident that our AUC falls within this interval.

4.4.2. Receiver Operating Characteristic (ROC) Curve for LR Model

The ROC curve stands for Receiver Operating Characteristic curve. It is a graphical representation used in evaluating the performance of binary classification models. It illustrates the relationship between the hit rate (also known as sensitivity or true positive rate) and the false alarm rate (also known as the false positive rate). The hit rate refers to the proportion of correctly identified positive instances (true positives) out of all actual positive instances. It represents the model's ability to correctly classify positive cases. On the other hand, the false alarm rate represents the proportion of incorrectly identified negative instances (false positives) out of all actual negative instances.

The ROC curve plots the hit rate on the y-axis and the false alarm rate on the x-axis. It shows how the trade-off between these two rates changes as the classification threshold of the model varies. The threshold determines the point at which the model classifies instances as positive or negative based on the predicted probabilities or scores. Ideally, a good classification model would achieve a high hit rate and a low false alarm rate, resulting in a curve that hugs the upper-left corner of the ROC space. The closer the curve is to this corner, the better the model's performance. The diagonal line from (0, 0) to (1, 1) represents the performance of a random classifier.

In addition to the ROC curve itself, the area under the curve (AUC) is often calculated to provide a single metric summarizing the model's performance. The AUC represents the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. A higher

Table 6. Odds ratio of important predictor variables based on the best fit model.

	Odds Ratio	Implications	Target Variable (Ozone)	
			Low (0)	High (1)
Nitric Oxide	0.16	Nitric Oxide might not be a protective factor for high ozone level	✓	
Oxides of Nitrogen	0.981	Oxides of Nitrogen might lead to high ozone level subsequently	✓	
Wind Speed	0.733	Wind speed might lead to high ozone level subsequently	✓	
Resultant Wind Direction	0.997	Resultant wind direction might lead to high ozone level subsequently	✓	
Maximum Wind Gust	0.976	Maximum wind gust might lead to high ozone level subsequently	✓	
Std. Dev. Wind Direction	1.06	Std. Dev. wind direction is a risk factor for high ozone level		✓
Outdoor Temperature	1	Outdoor temperature is a risk factor for high ozone level		✓
Dew Point Temperature	1.04	Dew point temperature is a risk factor for high ozone level		✓
Relative Humidity	0.885	Relative humidity might lead to high ozone level subsequently	✓	
Solar Radiation	4.19	Solar radiation is certainly a major risk factor for high ozone level		✓
PM10	1.01	Particulate matter 10 is a risk factor for high ozone level		✓

Table 7. Results of the LR model evaluation.

Miss-classification rate	MSE	cvAUC	SE	CI	Confidence
0.116	0.0833	0.952	0.0115	0.929, 0.975	0.95

AUC value indicates better discrimination power of the model. Our model has a very high discriminatory power for correct prediction of high ozone and low ozone levels at any given day. Figure 18 shows the ROC curve (*i.e.*, trade-off between sensitivity (or TPR) and False Positive Rate [1 − Specificity]). It further indicates that the model performs better against the benchmark (50%) with total area of 0.952 (95.2%).

4.4.3. Performance of LR Model using Confusion Matrix

Metrics such as ***accuracy***, ***precision*** (***positive prediction value***), ***recall*** (***sensitivity***) ***and f1 score*** provide different perspectives on model performance. The confusion matrix also helps in the interpretation of model performance. The Sensitivity or Recall (TP rate) of 0.8761 (87.6%) indicates that the model has a higher % of detecting high ozone level of a particular day. The Specificity (TN rate) of 0.8878 (88.8%) which is relatively high indicates that the model has a higher % of detecting low ozone level of a particular day. Therefore, our fitted Model has an accuracy of 88.4% with respect to performance and a precision of 81.2% which implies that our Model has a low FP rate. Confusion matrix and other statistical prediction parameters for logistic regression model are shown in Table 8.

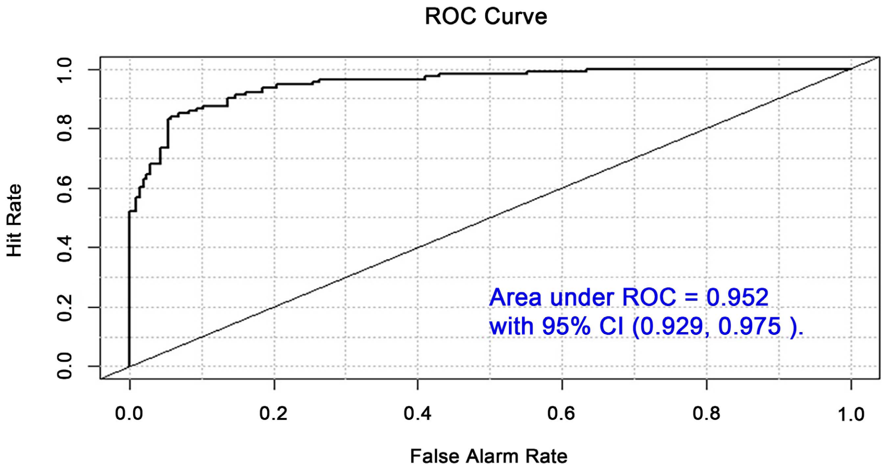


Figure 18. Showing the receiver operators curve of the hit rate vs false alarm.

Table 8. Confusion matrix and other statistical prediction parameters for LR.

Confusion Matrix and Statistics For LR	
Accuracy	0.884
95% CI	(0.843, 0.917)
Sensitivity/Recall	0.876
Specificity (True Negative Rate/TNR)	0.888
Pos Pred Value/Precision	0.811
F1 Score	0.843
Prediction Time	0.02 secs
Binary Cross Entropy	6.03

4.5. Artificial Neural Network (ANN) Model

To fit an Artificial Neural Network (ANN) model with our trained dataset D_1, to find the desired model. It is necessary to scale our training data, thereby creating a data frame with the target variable. After scaling, we then build our ANN structure which has 4 hidden layers containing 9, 7, 5, and 3 neurons respectively together with input and output layers.

4.5.1. Scaling and MSE of ANN Model

After scaling the test data D_2, we proceeded to predict the target "ozone" variable using our ANN model. Computing the Mean Squared Error (MSE) of our model, we obtained a value of 0.0833 which indicates that the model performed well. However, MSE alone is not an optimal evaluation technique for our model, hence we need to further calculate the misclassification rate and the confusion matrix.

4.5.2. Misclassification Rate of ANN Model

The miss-classification rate value is 0.107 which means that our model correctly predicts both high ozone and low ozone level at a rate of 89.3%. This suggests that our model performs well. Nevertheless, for better evaluation, we would further calculate the AUC and the confusion matrix of our ANN model (see Table 9).

4.5.3. ANN Model Evaluation

The AUC of 0.954 implies that our best fitted model has 95.4% accuracy to predict if the ozone level is either high or low for a particular day. The C.I also indicates the true AUC falls within the interval (0.929, 0.979). Therefore, we are 95% confident that our AUC falls within this interval.

4.5.4. ROC Curve

The AUC of 0.954 implies that our best fitted model has 95.4% accuracy to predict if the ozone level is either high or low for a particular day. The C.I also indicates the true AUC falls within the interval (0.929, 0.979). Therefore, we are 95% confident that our AUC falls within this interval (see Figure 19).

4.5.5. Performance of ANN Model using Confusion Matrix

The accuracy of our model is 0.893 (89.3%) which is relatively high indicating that our model performs well in predicting the ozone level for a day. The Sensitivity or Recall (TP rate) of 0.802 (80.2%) indicates that the model has a higher % of detecting high ozone level of a particular day. The Specificity (TN rate) of 0.957 (95.7%) which is relatively high indicates that the model has a higher % of detecting low ozone level of a particular day. With the high accuracy and a precision of 92.9%, these results imply that our Model has a low False Positive (FP) rate. Confusion matrix and statistical prediction parameters for artificial neural network model are shown in Table 10.

Table 9. Results of the ANN model.

Miss-classification rate	MSE	cvAUC	SE	CI	Confidence
0.107	0.0833	0.954	0.0126	0.929, 0.979	0.95

Table 10. Confusion matrix and other statistical prediction parameters for ANN.

Confusion Matrix and Statistics	
Accuracy	0.893
95%CI	(0.854, 0.925)
Sensitivity/Recall	0.802
Specificity (True Negative Rate/TNR)	0.957
Pos Pred Value/Precision	0.929
F1 Score	0.861
Prediction Time	0.00 secs
Binary Cross Entropy/Log Loss	3.74

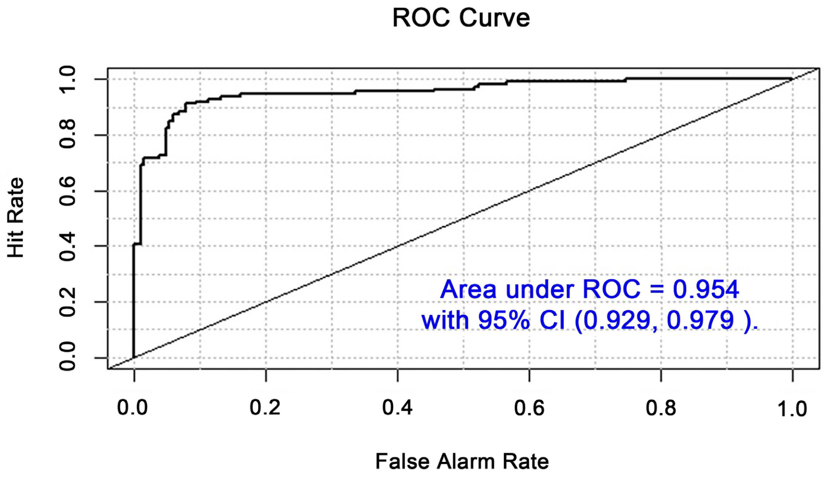


Figure 19. Showing the ROC of the hit rate vs false alarm.

4.6. *F*1 Score

The *F*1 score is a metric commonly used in classification tasks to evaluate the overall performance of a model. It combines both precision and recall into a single value, providing a balanced measure of the model's accuracy.

$$F1\text{ Score} = \frac{2 * (\text{precision} * \text{recall})}{(\text{precision} + \text{recall})}$$

*F*1 score considers both FP (false positive) and FN (false negatives), making it a ***useful metric*** when dealing with imbalanced datasets or when both precision and recall are equally important. When comparing different models or algorithms, a higher *F*1 score indicates better performance in terms of both precision and recall. Based on our results, ANN performs better than LR with *F*1 score of 0.861.

5. Conclusion and Recommendations

The accuracy of our model is 89.3% which is relatively high, thus it indicates that our model performs well in predicting the ozone level for a given day. Also, the Sensitivity or Recall (TP rate) of 80.2% indicates as well that our model has a higher chance of detecting the high ozone rate of a particular day. The Specificity (TN rate) of 95.7% indicates that the model has a higher chance of detecting the low ozone rate on a given day as well. With the high accuracy stated above and a precision of 92.9%, these results imply that our model has a low False Positive (FP) rate.

In addition, from our evaluation metrics for both models, Our ANN model performs slightly better than the LR model with the ANN model having higher accuracy 89.3% compared to LR's 88.4% and AUC 95.4% compared to LR's 95.2% while also having a lower miss-classification rate (10.7% compared to LR's 11.6%).

Furthermore, when we consider the precision and recall of our models' performance, both models perform very well with very high precision and high recall, meaning that our model has a high true positive (TP) rate and a low false positive (FP) rate. When the sensitivity is high, we also tend to have a lower false negative rate meaning that our model would most likely avoid a wrong prediction of a negative (low ozone level) outcome any day.

With regards to the prediction time, while both models show very small-time complexity for prediction execution, the ANN model has a lower prediction time. Also looking at the binary cross entropy, the ANN model has the lower binary cross entropy indicating that it performed better than the LR model in terms of classification.

We recommend that subsequent research should consider the following points:

Application of other types of supervised machine learning models, such as the Random Forest Model, Support Vector Machine, K-Nearest Neighbors, Decision Trees, and Naïve Bayes, for the classification of ozone.

Other researchers could try to expand the scope of the paper by using different datasets from regions affected by ozone pollution in various areas.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., Choirat, C., Dominici, F. and Schwartz, J.D. (2017) Air Pollution and Mortality in the Medicare Population. *The New England Journal of Medicine*, **376**, 2513-2522. https://doi.org/10.1056/NEJMoa1702747

[2] Lin, S., Liu, X., Le, L.H. and Hwang, S.-A. (2008) Chronic Exposure to Ambient Ozone and Asthma Hospital Admissions among Children. *Environmental Health Perspectives*, **116**, 1725-1730. https://doi.org/10.1289/ehp.11184

[3] Jerrett, M., Burnett, R.T., Pope, C.A., Ito, K., Thurston, G., Krewski, D., Shi, Y., Calle, E. and Thun, M. (2009) Long-Term Ozone Exposure and Mortality. *The New England Journal of Medicine*, **360**, 1085-1095. https://doi.org/10.1056/NEJMoa0803894

[4] Parker, J.D., Akinbami, L.J. and Woodruff, T.J. (2009) Air Pollution and Childhood Respiratory Allergies in the United States. *Environmental Health Perspectives*, **117**, 140-147. https://doi.org/10.1289/ehp.11497

[5] Bhuiyan, M.A.M., Sahi, R.K., Islam, M.R. and Mahmud, S. (2021) Machine Learning Techniques Applied to Predict Tropospheric Ozone in a Semi-Arid Climate Region. *Mathematics*, **9**, Article No. 2901. https://doi.org/10.3390/math9222901

[6] U.S. EPA. Nonattainment Areas for Criteria Pollutants (Green Book). https://www.epa.gov/green-book

[7] U.S. Environmental Protection Agency. Integrated Science Assessment (ISA) for Ozone and Related Photochemical Oxidants. https://www.epa.gov/isa/integrated-science-assessment-isa-ozone-and-related-photochemical-oxidants

[8] Medina-Ramón, M. and Schwartz, J. (2008) Who Is More Vulnerable to Die from

Ozone Air Pollution? *Epidemiology*, **19**, 672-679. https://doi.org/10.1097/EDE.0b013e3181773476

[9] Olufemi, I., Obunadike, C., Adefabi, A. and Abimbola, D. (2023) Application of Logistic Regression Model in Prediction of Early Diabetes across United States. *International Journal of Scientific and Management Research*, **6**, 34-48. https://doi.org/10.37502/IJSMR.2023.6502

[10] Tran, B., Sudusinghe, C., Nguyen, S. and Alahakoon, D. (2023) Building Interpretable Predictive Models with Context-Aware Evolutionary Learning. *Applied Soft Computing*, **132**, Article ID: 109854. https://doi.org/10.1016/j.asoc.2022.109854

[11] Issitt, R.W., Cortina-Borja, M., Bryant, W., Bowyer, S., Taylor, A.M. and Sebire, N. (2022) Classification Performance of Neural Networks versus Logistic Regression Models: Evidence from Healthcare Practice. *Cureus*, **14**, e22443. https://doi.org/10.7759/cureus.22443

[12] Valluri, C., Raju, S. and Patil, V.H. (2022) Customer Determinants of Used Auto Loan Churn: Comparing Predictive Performance Using Machine Learning Techniques. *Journal of Marketing Analytics*, **10**, 279-296. https://doi.org/10.1057/s41270-021-00135-6

[13] Xie, X., Wang, L. and Wang, A. (2010) Artificial Neural Network Modeling for Deciding If Extractions Are Necessary Prior to Orthodontic Treatment. *The Angle Orthodontist*, **80**, 262-266. https://doi.org/10.2319/111608-588.1

[14] Abiodun, O.I., Jantan, A., Omolara, A.E., Dada, K.V., Mohamed, N.A. and Arshad, H. (2018) State-of-the-Art in Artificial Neural Network Applications: A Survey. *Heliyon*, **4**, e00938. https://doi.org/10.1016/j.heliyon.2018.e00938

[15] Sarker, I.H. (2021) Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, **2**, Article No. 160. https://doi.org/10.1007/s42979-021-00592-x

[16] Couronné, R., Probst, P. and Boulesteix, A.-L. (2018) Random Forest versus Logistic Regression: A Large-Scale Benchmark Experiment. *BMC Bioinformatics*, **19**, Article No. 270. https://doi.org/10.1186/s12859-018-2264-5

[17] Sarker, I.H. (2021) Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, **2**, Article No. 420. https://doi.org/10.1007/s42979-021-00815-1

[18] Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M. and Farhan, L. (2021) Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *Journal of Big Data*, **8**, Article No. 53. https://doi.org/10.1186/s40537-021-00444-8

[19] Montesinos López, O.A., Montesinos López, A. and Crossa, J. (2022) Multivariate Statistical Machine Learning Methods for Genomic Prediction. Springer, Cham. https://doi.org/10.1007/978-3-030-89010-0

[20] Albaradei, S., Thafar, M., Alsaedi, A., Van Neste, C., Gojobori, T., Essack, M. and Gao, X. (2021) Machine Learning and Deep Learning Methods That Use Omics Data for Metastasis Prediction. *Computational and Structural Biotechnology Journal*, **19**, 5008-5018. https://doi.org/10.1016/j.csbj.2021.09.001

[21] Duan, F., Zhang, S., Yan, Y. and Cai, Z. (2022) An Oversampling Method of Unbalanced Data for Mechanical Fault Diagnosis Based on MeanRadius-SMOTE. *Sensors*, **22**, Article No. 5166. https://doi.org/10.3390/s22145166

[22] Karrar, A.E. (2022) The Effect of Using Data Pre-Processing by Imputations in

Handling Missing Values. *Indonesian Journal of Electrical Engineering and Informatics*, **10**, 375-384. https://doi.org/10.52549/ijeei.v10i2.3730

[23] Bin Rafiq, R., Modave, F., Guha, S. and Albert, M.V. (2020) Validation Methods to Promote Real-World Applicability of Machine Learning in Medicine. 2020 3*rd International Conference on Digital Medicine and Image Processing*, Kyoto, 6-9 November 2020, 13-19. https://doi.org/10.1145/3441369.3441372