IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

OZONE LEVEL PREDICTION USING DATAMINING ALGORITHM

Dr.K.Merriliance[1] and S.Shunmuga priya[2]

Department of Computer Applications, Sarah Tucker College, Tirunelveli-7.

ABSTRACT

Air pollution is one of the most serious hazards to humans' health nowadays, it is an invisible killer that takes many human lives every year. There are many pollutants existing in the atmosphere today, ozone being one of the most threatening pollutants. It can cause serious health damage such as wheezing, asthma, inflammation, and early mortality rates. Because of the increasing attention on environmental issues, especially air pollution, predicting whether a day is polluted or not is necessary to people's health. In order to solve this problem, this work is classifying ground ozone level based on big data and machine learning models, where polluted ozone day has class 1 and non-ozone day has class 0. The dataset used in this work is derived from the UCI Website, containing various environmental factors that could possibly affect the occurrence of ozone pollution. This dataset is first filled up for further process, next standardized to ensure every feature has the same weight, and then split into training set and testing set. After this, two different machine learning models are used in the prediction of ground ozone level and their final accuracy scores are compared. In conclusion, among Linear Discriminant Analysis(LDA) and Linear Regression(LR). This work utilizes relatively simple methods of forecasting and calculates the first accuracy scores in predicting ground ozone level; it can thus be a reference for environmentalists. Moreover, the direct comparison among two different models provides machine learning field an insight to determine the most accurate model.

1. INTRODUCTION

Ground ozone pollution has been a serious air quality problem over the years and can be extremely harmful to people's health if no advanced forecasts are provided. However, the occurrence of an ozone polluted day depends on a lot of sophisticated chemical, physical, and geological factors, so it is too complicated and indirect to use simple math formula to calculate the ozone level. Fortunately, a group of scientists measured several environmental factors in Houston, Galveston and Brazoria area that might influence ground ozone pollution, which is the dataset used in this research. This group also utilized the same dataset to build models that can successfully predict the ground ozone level. Their paper, Forecasting Skewed Biased Stochastic Ozone Days : analysis , solutions and beyond, published in April 2007, researched about data mining techniques based on the collected dataset and had made huge improvements on the Huston ground ozone pollution forecasting system. On the other hand, in this research, five different machine learning models are trained to make binary predictions of ground ozone level (ozone day: 1, non-ozone day: 0). Besides providing precise forecasting system for the citizens, this research also contributes to the field of machine learning. With the employment of five different models, the accuracy scores can be compared and the conclusion that which method is the most accurate can be derived relatively simple and clear. During the construction of each method, variables are changed five times to maximize each model's test scores. These scores are eventually used to compare the accuracy of each method and select the most precise one as the best prediction method.

2. LITERATURE REVIEW

Air pollution is one of the major hazards to human health and the ecosystem nowadays. With the growing development of the economy, the rising population, the increase in industries and the growing need for transportation, this all leads to increased environmental pollution which includes air pollution. Air pollution is mainly caused by the emissions of factories, electrical facilities, vehicles that burn fuel, meteorological factors, etc. [1] Ground-level ozone is a major pollutant that is hazardous to humans′ health, unlike the stratospheric ozone layer that protects the earth. It is formed as a reaction between pollutants resulting from industrial emissions, vehicles, and electrical facilities. Health problems associated with ozone exposure include wheezing, coughing, asthma, chest pain, decreased capacity for exercise, inflammation, increased mortality rate and more [2]. It can have a severe impact not only on humans but also on vegetation and crops. It caused a €6.7 billion crop loss in the EU in 2007 [3]. Ozone is considered one of the green house gases that causes a reduction of carbon intake by plants which contributes to increased global warming [3]. Due to the dangers of air pollution, multiple air quality indices exist in different countries and they are used to determine if the pollutants′ concentrations are within the healthy range [4]. With the decreased cost of pollutants′ monitoring sensors, many projects are being carried out in different countries to monitor pollutants, e.g., stations measuring pollutants which are connected to the internet of things. Some of these projects have been collecting and storing data for several years which can lead to more knowledge about the problem of air pollution [5, 6]. Since the issue of air quality is of high significance, there have been multiple attempts to forecast air quality in different methods. Air quality forecasting systems are tools that can help describe the air quality problem and understand the relationship between pollutants, meteorological factors, emissions, and other atmospheric variables. They can help make future forecasts about air quality [1]. Types of forecasting systems include deterministic models that use mathematical equations to describe the atmospheric processes causing pollution. They are based on the physical and chemical nature of pollutants. The problem with these chemical and physical models is that they do not capture the behavior of pollutants very well and they tend to linearize a non-linear relationship between data in the natural world as well as having difficulties in pr cessing large amounts of data [7]. Due to the problems of the afore-mentioned models and the increase in available data, new methods were introduced to discover patterns in data and make better predictions for air pollutants. Machine learning, which is a subfield of artificial intelligence could be used in this case due to its ability to discover complex relationships between data and to analyze large datasets [5]. One of the most well-known machine learning algorithms is artificial neural networks (ANN) which is widely used by many authors due to its ability to discover non-linear relationships between variables [8]. Another algorithm that is used in this topic is the support vector machine (SVM) which is also called support vector regression (SVR) when used for regression purposes [9]. This algorithm is a good generalization algorithm that generalizes well to new data [10]. The third algorithm that we used is the decision tree (DT) which is a well-known machine learning algorithm with a graphical upside-down tree structure. When a decision tree is used for regression, it is called a decision tree regression (DTR) or regression tree [11]. Finally, extreme gradient boosting (XGBoost) which is a boosted tree with the gradient boosting method is being used for predictionpurposes due to its promising results and speed [12]. Despite the fact that many studies were conducted in various parts of the world to forecast ozone concentration using machine learning, the problem of air pollution prediction is not given significant importance in Jordan and thus no papers were done on this topic. Another point is papers in the context of air quality forecasting rarely focused on time-series smoothing filters and the time required for prediction. In this paper, we tackled the problem of ozone prediction in Jordan using machine learning techniques. We focused on three main topics in this paper: Firstly, we conducted a comparison between four machine learning algorithms which are multi-layer perceptron (MLP), SVR, DTR, and XGBoost to find the algorithm with the highest performance. Secondly, we explored the importance of adding a smoothing filter to the noisy time-series dataset. Previous research in [13] explained how using a denoising filter improved the results in the field of air pollution prediction. In this paper, we performed a comparison between three smoothing filters and compared their results with the original unfiltered data. Thirdly, we tried to decrease the time required for prediction by finding the most important features for predicting ozone concentration, since many variables in the dataset may not be relevant to the prediction process. Although ozone is highly affected by many complex atmospheric and meteorological variables, conducting feature selection and narrowing down the number of features proved efficient and greatly reduced the time and improved the results. This paper is structured as follows. The related work section contains a brief explanation of ANN, SVR, DTR, and XGBoost alongside previous research done in the field of air quality forecasting using these algorithms. The materials and methods section displays information about the dataset as well as describing some concepts about the smoothing filters and the performance evaluation metrics used in this research. The experimental results and discussion section illustrates each step of the experiments in detail such as the parameters configurations of the used algorithms, the results obtained, and a discussion of the results. Finally, the conclusion and related work section summarizes this research and gives ideas for future research work in this field.

3. METHODOLOGY

In this work, many applied techniques were tested for the subsequent stages of processing and analysis of the breast cancer dataset.

Stage 1: Pre-processing

As a part of this work, processing was performed on the raw breast cancer data to scale the features using the Standard Scaler module. Standardization of datasets is a common requirement for many machine learning estimators. It transforms the attributes to a standard Gaussian distributions based on (xi–mean(x))/ stdev (x) where stdev is the standard deviation. The Robust Scaler depends on the inter quartile range to transform the features using (xi–$Q1$(x))/($Q3$(x)–$Q1$(x)), where $Q1$, $Q2$, and $Q3$ represent quartiles.

Stage 2: Classification

Usually, ensemble machine learning algorithms allow better predictive performance compared with a single model. This can be considered machine learning competition, where the winning solution was used as a model for breast cancer diagnosis. In this work, the following heterogeneous ensembles machine learning algorithms were used to classify the given data set:

- Linear regression (LR) and
- Linear Discriminant Analysis (LDA).

Linear Regression

Linear regression assumes a linear or straight line relationship between the input variables (X) and the single output variable (y).

More specifically, that output (y) can be calculated from a linear combination of the input variables (X). When there is a single input variable, the method is referred to as a simple linear regression.

In simple linear regression we can use statistics on the training data to estimate the coefficients required by the model to make predictions on new data.

The line for a simple linear regression model can be written as:

1 y = b0 + b1 * x

where b0 and b1 are the coefficients we must estimate from the training data.

Once the coefficients are known, we can use this equation to estimate output values for y given new input examples of x.

It requires that you calculate statistical properties from the data such as mean, variance and covariance.

All the algebra has been taken care of and we are left with some arithmetic to implement to estimate the simple linear regression coefficients.

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is most commonly used as dimensionality reduction technique in the pre-processing step for pattern-classification and machine learning applications. The goal is to project a dataset onto a lower-dimensional space with good class-separability in order avoid overfitting ("curse of dimensionality") and also reduce computational costs.

LDA approach in 5 steps

Listed below are the 5 general steps for performing a linear discriminant analysis; we will explore them in more detail in the following sections.

1. Compute the dd-dimensional mean vectors for the different classes from the dataset.
2. Compute the scatter matrices (in-between-class and within-class scatter matrix).
3. Compute the eigenvectors (ee1,ee2,...,eedee1,ee2,...,eed) and corresponding eigen values (λλ1,λλ2,...,λλdλλ1,λλ2,...,λλd) for the scatter matrices.
4. Sort the eigenvectors by decreasing eigen values and choose kk eigenvectors with the largest eigen values to form a d×kd×k dimensional matrix WWWW (where every column represents an eigenvector).
5. Use this d×kd×k eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the matrix multiplication: YY=XX× WWYY = XX × WW (where XXXX is a n×dn×d-dimensional matrix representing the nn samples, and yyyy are the transformed n×kn×k-dimensional samples in the new subspace).

It works on a simple step-by-step basis. Here is an example. These are the three key steps.

(i) Calculate the separability between different classes. This is also known as between-class variance and is defined as the distance between the mean of different classes.

$$S_b = \sum_{i=1}^{g} N_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T$$

Between Class Variance

(ii) Calculate the within-class variance. This is the distance between the mean and the sample of every class.

$$S_w = \sum_{i=1}^{g}(N_i - 1)S_i = \sum_{i=1}^{g}\sum_{j=1}^{N_i}(x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^T$$

Within-Class Variance

(iii) Construct the lower-dimensional space that maximizes Step1 (between-class variance) and minimizes Step 2(within-class variance). In the equation below P is the lower-dimensional space projection. This is also known as Fisher's criterion.

$$P_{lda} = \arg\max_{P} \frac{\left|P^T S_b P\right|}{\left|P^T S_w P\right|}$$

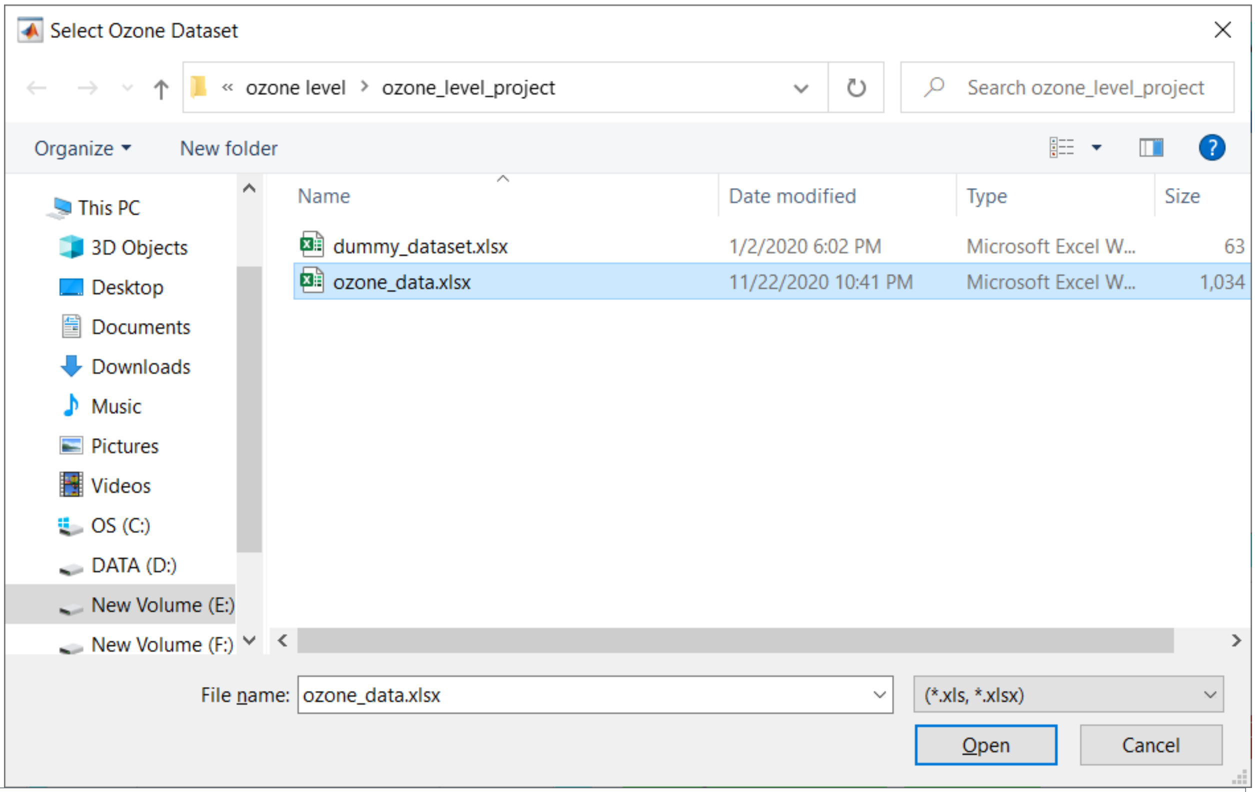
The representation of Linear Discriminant models consists of the statistical properties of the dataset. These are calculated separately for each class. For instance, for a single input variable, it is the mean and variance of the variable for every class.

If there are multiple variables, the same statistical properties are calculated over the multivariate Gaussian. This includes the means and the covariance matrix. All these properties are directly estimated from the data. They directly go into the Linear Discriminant Analysis equation.

The statistical properties are estimated on the basis of certain assumptions. These assumptions help simplify the process of estimation. One such assumption is that each data point has the same variance.

Another assumption is that the data is Gaussian. This means that each variable, when plotted, is shaped like a bell curve. Using these assumptions, the mean and variance of each variable are estimated.

4. RESULTS



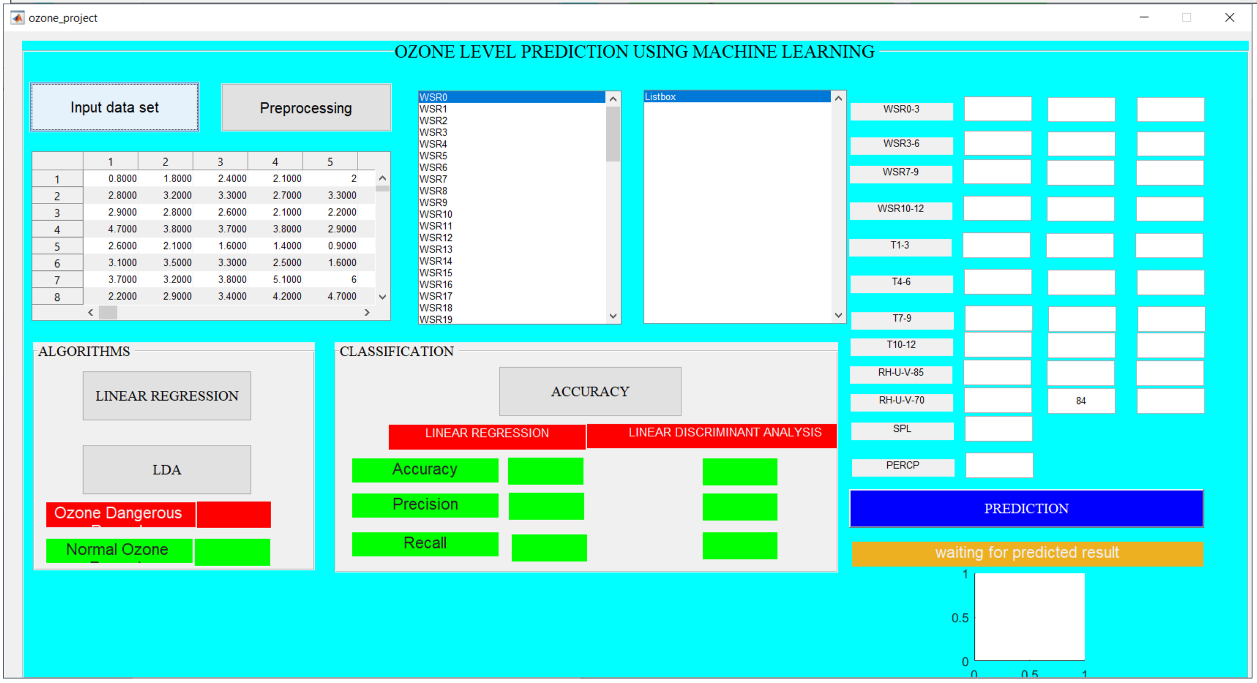
Select Ozone Dataset

« ozone level › ozone_level_project

Name	Date modified	Type	Size
dummy_dataset.xlsx	1/2/2020 6:02 PM	Microsoft Excel W...	63
ozone_data.xlsx	11/22/2020 10:41 PM	Microsoft Excel W...	1,034

File name: ozone_data.xlsx

(*.xls, *.xlsx)



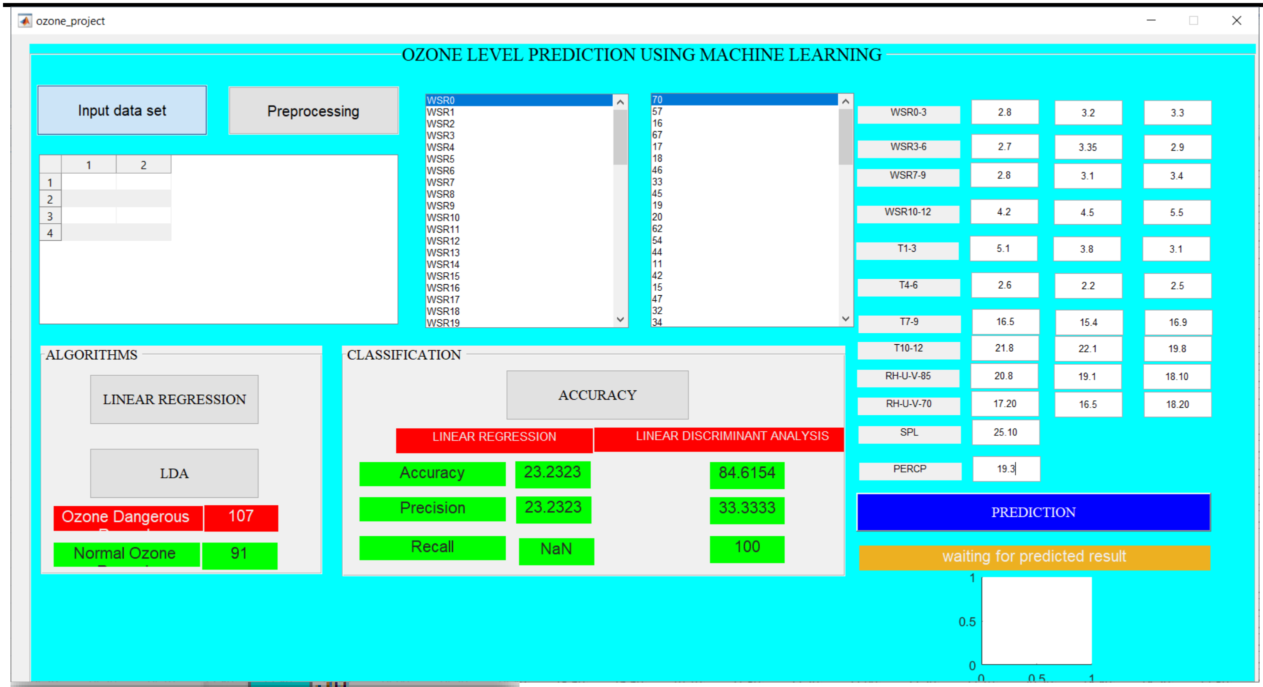
ozone_project

OZONE LEVEL PREDICTION USING MACHINE LEARNING

	1	2	3	4	5
1	0.8000	1.8000	2.4000	2.1000	2
2	2.8000	3.2000	3.3000	2.7000	3.3000
3	2.9000	2.8000	2.6000	2.1000	2.2000
4	4.7000	3.8000	3.7000	3.8000	2.9000
5	2.6000	2.1000	1.6000	1.4000	0.9000
6	3.1000	3.5000	3.3000	2.5000	1.6000
7	3.7000	3.2000	3.8000	5.1000	6
8	2.2000	2.9000	3.4000	4.2000	4.7000

RH-U-V-70: 84

waiting for predicted result

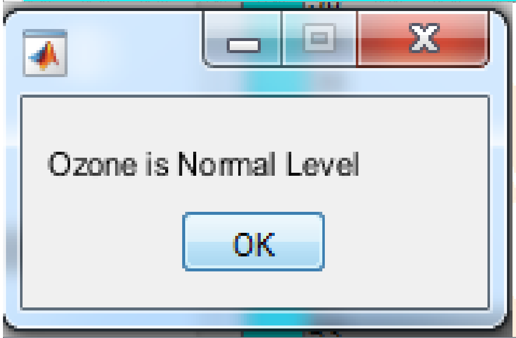


Ozone Dangerous	107	
Normal Ozone	91	

	LINEAR REGRESSION	LINEAR DISCRIMINANT ANALYSIS
Accuracy	23.2323	84.6154
Precision	23.2323	33.3333
Recall	NaN	100

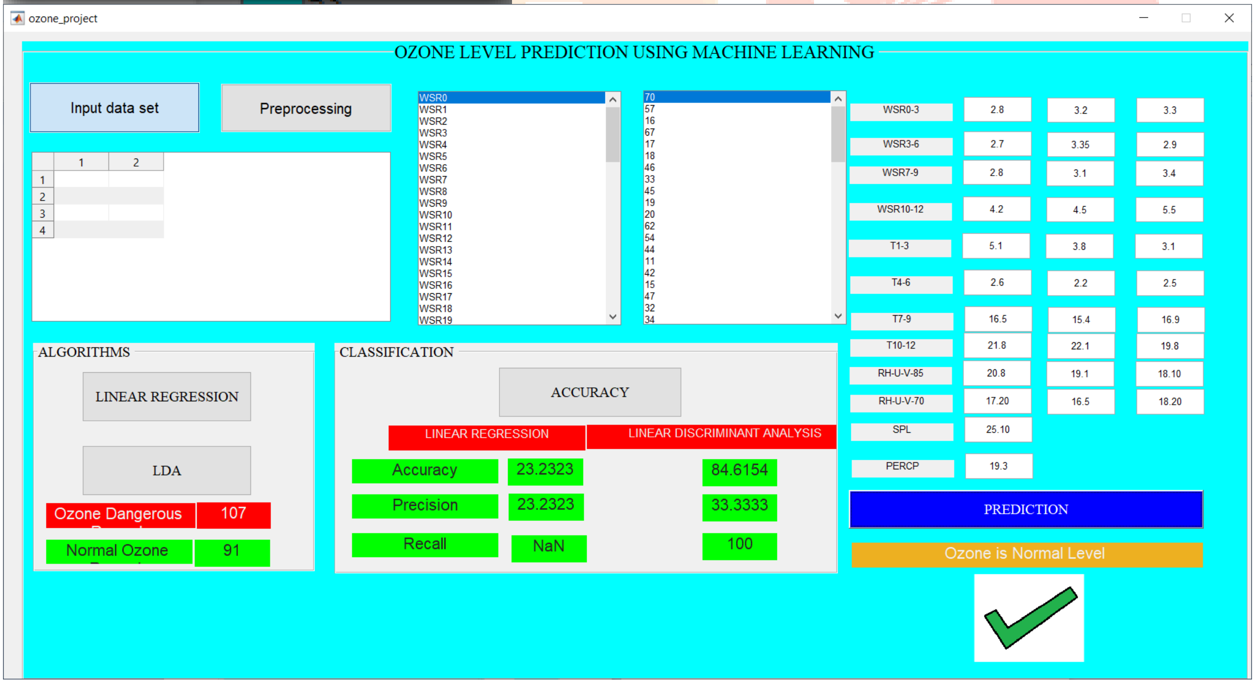
WSR0-3	2.8	3.2	3.3
WSR3-6	2.7	3.35	2.9
WSR7-9	2.8	3.1	3.4
WSR10-12	4.2	4.5	5.5
T1-3	5.1	3.8	3.1
T4-6	2.6	2.2	2.5
T7-9	16.5	15.4	16.9
T10-12	21.8	22.1	19.8
RH-U-V-85	20.8	19.1	18.10
RH-U-V-70	17.20	16.5	18.20
SPL	25.10		
PERCP	19.3		

waiting for predicted result



Ozone is Normal Level

OK



Ozone Dangerous	107	
Normal Ozone	91	

	LINEAR REGRESSION	LINEAR DISCRIMINANT ANALYSIS
Accuracy	23.2323	84.6154
Precision	23.2323	33.3333
Recall	NaN	100

WSR0-3	2.8	3.2	3.3
WSR3-6	2.7	3.35	2.9
WSR7-9	2.8	3.1	3.4
WSR10-12	4.2	4.5	5.5
T1-3	5.1	3.8	3.1
T4-6	2.6	2.2	2.5
T7-9	16.5	15.4	16.9
T10-12	21.8	22.1	19.8
RH-U-V-85	20.8	19.1	18.10
RH-U-V-70	17.20	16.5	18.20
SPL	25.10		
PERCP	19.3		

Ozone is Normal Level

5. CONCLUSION

In this research, we assessed the ability of machine learning techniques to predict ozone levels for the next day in Amman, Jordan, specifically, in King Hussein Public Parks and the surrounding area. We compared MLP, SVR, DTR, and XGboost and found that MLP outperformed the other algorithms. We also compared various smoothing filters for the time-series data and discovered that the Savitsky-Golay filter enhanced the results by 50% for R2 and 80% for both RMSE and MAE. The final contribution of this research is performing an intensive feature selection to reduce the number of features and thus decrease the time it takes to make the prediction since time is an important factor in the case of large datasets. Using the forward wrapper, we found that the previous day values of ozone, temperature, and humidity are the most influential features in our dataset for forecasting ozone concentration of the next day. The time is improved from before and after using the feature selection by about 91%. The final developed model scored R2 of 98.653%, RMSE of 1.016 ppb and MAE of 0.800 ppb which is a very promising result.

6. REFERENCE

1. H. P. Peng. Air Quality Prediction by Machine Learning Methods, Master dissertation, The University of British Columbia, Canada, 2015.
2. United States Environmental Protection Agency. Environments and contaminants: Criteria air pollutants. America’s Children and the Environment, 3rd ed., United States Environmental Protection Agency, Ed., Washington DC, USA: United States Environmental Protection Agency, 2015.
3. DEFRA. Air Pollution: Action in a Changing Climate, London, UK: Department for Environment, Food and Rural Affairs, 2010.
4. A. Plaia, M. Ruggieri. Air quality indices: A review. Re-views in Environmental Science and Bio/Technology, vol. 10, no. 2, pp. 165–179, 2011. DOI: 10.1007/s11157-010- 9227-2.
5. C. Bellinger, M. Shazan, M. Jabbar, O. Zaïane, A. Osornio-Vargas. A systematic review of data mining and ma- chine learning for air pollution epidemiology. BMC Public Health, vol. 17, no. 1, Article number 907, 2017. DOI: 10.1186/s12889-017-4914-3.
6. T. M. Chiwewe, J. Ditsela. Machine learning based estimation of Ozone using spatio-temporal data from air quality monitoring stations. In Proceedings of the 14th IEEE International Conference on Industrial Informatics, IEEE, Poitiers, France, pp. 58-63, 2016. DOI: 10.1109/INDIN.2016.7819134.
7. S. A. Abdul-Wahab, S. M. Al-Alawi. Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. Environmental Modelling & Software, vol. 17, no. 3, pp. 219–228, 2002. DOI: 10.1016/S1364-8152(01)00077-9.
8. W. Z. Lu, D. Wang. Learning machines: Rationale and application in ground-level ozone prediction. Applied Soft Computing, vol. 24, pp. 135–141, 2014. DOI: 10.1016/j.asoc.2014.07.008.
9. A. S. Sánchez, P. J. G. Nieto, P. R. Fernández, J. J. del Coz Díaz, F. J. Iglesias-Rodríguez. Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). Mathematical and Computer Modelling, vol. 54, no. 5-6, pp. 1453–1466, 2011. DOI: 10.1016/j.mcm.2011.04.017.
10. A. J. Smola, B. Schölkopf. A tutorial on support vector regression. Statistics and Computing, vol. 14, no. 3, pp. 199–222, 2004. DOI: 10.1023/B:STCO.0000035301. 49549.88.
11. G. G. Moisen. Classification and regression trees. Encyclopedia of Ecology, S. E. Jørgensen, B. D. Fath, Eds., Oxford, UK: Elsevier, 2008.
12. B. X. Zhai, J. G. Chen. Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China. Science of the Total Environment, vol. 635, pp. 644 658, 2018. DOI: 10.1016/j.scitotenv.2018.04.040.
13. M. R. Delavar, A. Gholami, G. R. Shiran, Y. Rashidi, G.R. Nakhaeizadeh, K. Fedra, S. Hatefi Afshar. A novel method for improving air pollution prediction based on machine learning approaches: A case study applied to the capital city of tehran. ISPRS International Journal of Geo Information, vol. 8, no. 2, Article number 99, 2019. DOI: 10.3390/ijgi8020099.