**1. Introduction**

Most research on pollutant distribution in the atmosphere use a variety of prediction methods, including gaussian models and numerical models. Emission inventories, meteorological data, and receptor positions are the primary inputs to analysis methods. The estimated ground-level concentration at specific receptor positions is the primary contribution from these models.

The models are mainly based on the mathematical formulation of the physics and chemistry of the atmosphere, which govern the dispersion of pollutants. This is to evaluate the level of 'Sulfur dioxide' in in Jeddah using machine learning algorithm with other parameters are inputs for air quality parameters.

**2. Literature review**

When the gaussian distribution does not adequately reflect the impact on horizontal dispersion caused by incremental shifts in wind direction over a one-hour span (**Collet and Oluyemi, 1997**). Even though dispersion models have some physical foundations, accurate knowledge about pollutant origins and other parameters is not widely available.

To address this constraint, mathematical models are also used to aid in pollutant concentration estimation (**Ziomass et al., 1995**; **Finzi and Tebaldi, 1982**). This models are based on the assumption that the interaction between variables is predictive in nature. However, such models necessitate knowledge of the data's distribution, which is not always accessible (**Comrie, 1997**). Not long after, Pollutant concentrations have also been predicted using neural network-based models.

Because of their computational efficiency and generalization ability, these models are a better alternative to statistical models. They will work on data that has a high dimensionality. SO2 concentrations at Sostanj were predicted using a neural network model (**Boznar et al., 1993)** Inputs provided historical data on measured SO2 concentrations as well as meteorological parameters including wind speed and direction.

A feedforward network was used, along with a backpropagation learning algorithm. The study's main finding is that neural network models can be used to forecast ground-level pollutant concentrations in diverse terrains. The findings, however, were not equivalent to those obtained using other modeling techniques. **Gardner and Dorling (1998**) provided an overview of the use of artificial neural networks in atmospheric sciences.

The implementations of neural network in atmospheric sciences is thoroughly explored. **Gardner and Dorling (1999)** used artificial neural networks to forecast NO2 and NOx concentrations using meteorological inputs that included and excluded the associated pollution component. In both of these experiments, there are two kinds of multilayer perceptrons: feedforward networks and feedback networks.

Feedback networks were used in both of the aforementioned experiments. For feature approximation and perceptrons may be used to simulate the device, feedforward networks are available. Pattern identification generally. Feedback networks can help in predicting time and spatial patterns for data series (**Beale 1997**) and they can enhance the precision of the feedback networks. In this research, the SO2 concentration in three weather monitoring sites in Delhi is forecast using meteorological data through a recurring network with feedback in the hidden layer.

Although the data was obtained from three locations, testing was carried out on only one network in place of three neuronal networks. The results are also compared using a regression model.

**3. Method and materials with all details**

In this work we will try to build an ANN model for our main variable Sulphur Dioxide (SO2 μg/m3) using machine learning algorithm called CART in Minitab software. Our response variable or target is a continuous variable. And the responses variables are also continues and a numeric values, so in Minitab we will choose CART regression algorithm. We choose the k-fold cross-validation as the appropriate method to test our model and also because we have a small dataset (≤5000 observation).

To use the CART algorithm in Minitab follow this steps:

1. Introduce the data: Air quality data.xlsx
2. Go to Stat> **Predictive Analytics** > **CART® Regression**.
3. In **Response**, choose the main variable Sulfur dioxide.
4. In  **Continuous predictors**, enter *the rest of the variables as Inputs*
5. In **Categorical predictors**, nothing because we don’t have categorical variable
6. Click **Validation**.
7. In **Validation method**, select **K-fold cross-validation**.
8. Click **OK** in each dialog box.

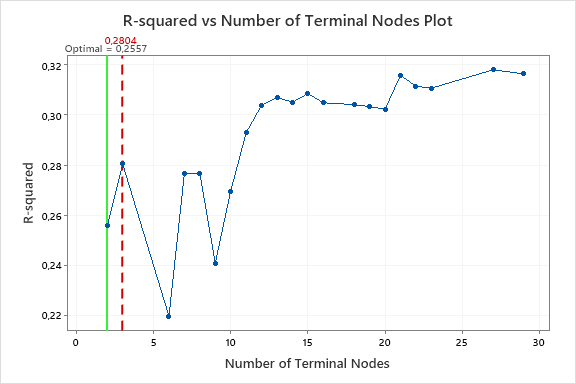
For more details see the Appendix

**4. Results and findings**

The CART algorithm shows the smallest tree that has an R-squared value within one std-error of the tree with the max r-squared. And in the settings we use the k-fold validation so the criterion is the maximum k-fold R-squared value. In the results we have 8 Nodes.

Before we interpret the tree, we should start first by examining the R-squared plot that shows its value from the cross-validation and also the number of nodes.

The Figure 1 shows the evolution of the R-squared statistic increase and stochastically evaluate for the first few nodes, after 11 nodes become to flatten. The figure also shows two possible number of nodes that we’ve tested in the green line we tested two nodes what give us an R-squared equal to 0.2557, and in the red line we had 3 nodes with a 0.2804 become more significant. We conclude that an increase in the number of nodes impact positively the value of the R-squared.

Figure 1: R-squared vs number of Terminal Nodes Plot

Source: Author’s calculation with CART algorithm in Minitab

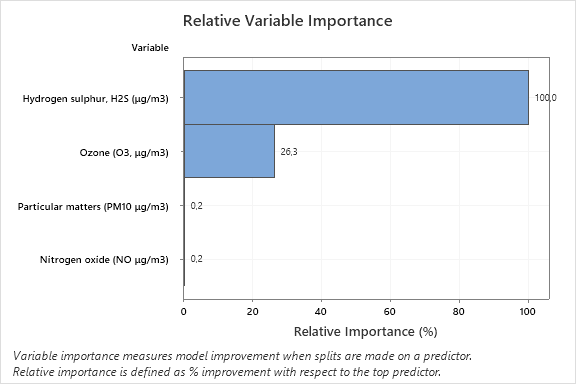
No we examine the model summary to evaluate the performance of the model. Table 1 shows that the value of the test and training statistics are close together, so the model does not show any overfit. And also the two results using 2 nodes and 3 nodes are very near to each other. So we will use the model with the 2 nodes to explore the relation between our main variable (Sulphur Dioxide, SO2 μg/m3) and the inputs variables for air quality measures (Carbon Monoxide, (CO, μg/m3), Hydrogen sulphur, H2S (μg/m3), Ozone (O3, μg/m3), Nitrogen oxide (NO μg/m3), Particular matters (PM10 μg/m3)).

Table 1: Model Summary

|  |  |
| --- | --- |
| Total predictors | 5 |
| Important predictors | 4 |
| Number of terminal nodes | 3 |
| Minimum terminal node size | 16 |
| **Statistics** | | | **Training** | **Test** |
| R-squared | | | 0,3616 | 0,2804 |
| Root mean squared error (RMSE) | | | 8,5776 | 9,1067 |
| Mean squared error (MSE) | | | 73,5747 | 82,9322 |
| Mean absolute deviation (MAD) | | | 4,8948 | 5,0807 |
| Mean absolute percent error (MAPE) | | | 0,8282 | 0,8383 |

Source: Author’s calculation with CART algorithm in Minitab

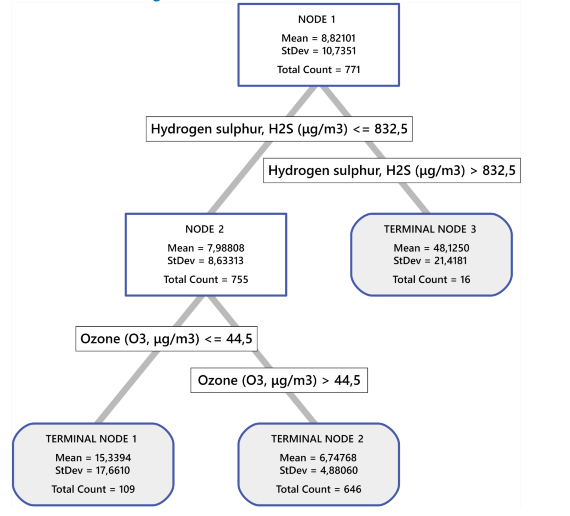
From the figure 2 we say that the Hydrogen Sulphur is the most important predictor of the relative importance variables. In figure we have the Hydrogen Sulphur as the 100% top predictor variable, as conclusion we can compare the other variable with the Hydrogen Sulphur to determine their impact. The figure also shows list describe the next most important variables in our model. The list determine that there 4 positive importance, the relative classification provide information about the variables that need to be controlled and monitored for the air quality in Jeddah.

****Figure 2: Relative Variable Importance.

Source: Author’s calculation with CART algorithm in Minitab

The figure 3 shows all the 3 possible combination from the all variables already mentioned. Including information about the fits and error statistics with terminal nodes info.

The first node is split using our most important variable in the data that records the level of Hydrogen Sulphur in the air was split to two interval (Hydrogen Sulphur<= and Hydrogen Sulphur=>) with 771 case count in total and mean and StDev equal respectively 8.82, 17.73. Node 2 include the level of the Ozone variable with 755 count case and 7.89 and 8.63 as mean and standard deviation respectively the mean and SeDev are less than the first overall mean and StDev of the node 1.

Figure 3: Tree Diagram

Source: Author’s calculation with CART algorithm in Minitab

**5. Conclusions:**

This work aim to study the level of Sulfur dioxide in in Jeddah using artificial intelligent methods. Such as the CART algorithm. As results we found that the level of our main variable is defined by the Hydrogen Sulphur as the most important contributing variable in our ANN model. This work was inspired by many other paper (see references) uses the same new techniques of machine learning. Also the findings shows that this new methods perform better for high dimensional data set and can give a perfect results with set of rules to be included in settings of the model specification.

**6. References:**

Boznar, M., Lesjak, M., Mlakar, P., 1993. A neural network-based method for short-term predictions of ambient SO2 concentrations in complex terrain. Atmospheric Environment 27B (2), 221–230.

Chelani, A. B., Rao, C. C., Phadke, K. M., & Hasan, M. Z. (2002). Prediction of sulphur dioxide concentration using artificial neural networks. Environmental Modelling & Software, 17(2), 159-166.

Collett, R. S., & Oduyemi, K. (1997). Air quality modelling: a technical review of mathematical approaches. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, *4*(3), 235-246.

Comrie, A. C. (1997). Comparing neural networks and regression models for ozone forecasting. *Journal of the Air & Waste Management Association*, *47*(6), 653-663.

Finzi, G., & Tebaldi, G. (1982). A mathematical model for air pollution forecast and alarm in an urban area. *Atmospheric Environment (1967)*, *16*(9), 2055-2059.

Gardner, M.W., Dorling, S.R., 1998. Artiﬁcial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric Environment 32 (14/15), 2627–2636

Gardner, M.W., Dorling, S.R., 1999. Neural network modeling and prediction of hourly NOx and NO2 concentrations in urban air in London. Atmospheric Environment 33, 709–719.

Taylan, O., Alkabaa, A., Alamoudi, M., Basahel, A., Balubaid, M., Andejany, M., & Alidrisi, H. (2021). Air Quality Modeling for Sustainable Clean Environment Using ANFIS and Machine Learning Approaches.

Ziomas, I. C., Melas, D., Zerefos, C. S., Bais, A. F., & Paliatsos, A. G. (1995). Forecasting peak pollutant levels from meteorological variables. *Atmospheric Environment*, *29*(24), 3703-3711.

**Appendix: Steps to use CART Algorithm**

