D682.1
Air Quality and Health Risk Prediction

Given Problem

The U.S. Environmental Protection Agency (EPA) have provided source data to educate a machine learning model in hopes of generating reports about air quality and possible associated health risks thereof.

Possible Solutions:

With the given data, approach methods for solving the given problem include *Linear Regression*, *K-Nearest Neighbors (KNN)*, and *Naive Bayes*. Out of these possible directions, the most suitable solution for the given problem is ***Linear Regression***. This is because it can provide a comprehensive approach to trend analysis while highlighting the relationship of continuous data to a key variable (in this case the 'Health Risk Score' index). Also, it is an adaptable model that can manageably be configured to analyze variable data.

Linear Regression as Eminence:

What makes a *Linear Regression* approach cater greatly to the given problem is, as aforementioned, its strengths of being adaptive and comprehensive in its analysis. The key architecture behind this machine learning model is the use of regression lines to compare trends in data. These lines, when constructed properly, can show the correlation between a specific category of data, or the whole data and the key variable that is trying to be approximated after the learning phase of the program.

The overarching design of a *Linear Regression* structure model enables connections to be found across different assignable categories of data, and synthesize outputs therefrom: thus showing its adaptive nature.

Similarly, the malleability and scaling that can be implemented with this algorithmic approach is a testament to the comprehensive nature of such a model. With the use of different segments capturing different concepts, there comes an ability to there synthesize and approximate a final solution with input of every relevant measure of data. Therefore, this allows a greater panoptic evaluation of data than other algorithms often provide.

More specifically, the *Naive Bayes* model attempts to make predictions based on naive analysis that doesn't incorporate calculations for all possibly relevant data. Similarly, *KNN* associates with similar data to congeal similarities into more meaningful outputs, but it also does not keep a greater holistic approach, as it focuses on the nearest correlations instead of the broader environment that may yet have different effects on certain sets of data.

Thus, a ***Linear Regression*** model is the optimal choice in this scenario.

Algorithm Limitations:

With the chosen machine learning model, there are some limitations that can arise based on the inherent architecture of such systems. Firstly, regression models analyze largely based on linear approximations: not all data trends can undergo accurate linear conversion. Data nuances can be unrealized because of this. Also with such models, outliers can be left unassessed and cause undue skewing to regression lines.

AI Evaluation:

For understanding the analysis done by the AI solution model for this given problem, different evaluation metrics are deployed to window the estimation errors in the program. The first of these to be implemented is *Average Error*, which tells the average deviation from the key variable (in this problem's context, the Health Score Risk index). This is helpful as it can inform the user of the standard error that the estimated output may have.

Secondly, the use of *Mean Absolute Percentage Error* (*MAPE*) provides a better vantage on what the overall difference ratio is between the estimated and actual value. This makes gadging the error relative to the key value simpler, as it is expressed in a direct percentage of difference rather than a constant value that needs greater context to be understood fully.

Model Test Results:

After training and testing an implementation of a Linear Regression Model, the following outputs were approximated:

- *Avg Air Quality:* **1.3799**
- *Average Error:* **6.7068**
- *Mean Absolute Percentage Error:* **68.2 %**

The *Average Air Quality* metric is synthesized from the atmospheric data that is provided in the reading report provided for this project. It quantifies air quality as a measurement that closely parallels the Health Risk Score metric while leaving out date/time data that is not directly correlated to atmospheric readings.

This *Average Error* reading indicates that estimated values for the Health Risk Score index are typically 6.7068 units off. This helps to show that while the approximation is close, it still has a notable error margin. The *MAPE* score furthers this assessment as it provides a more interpretable understanding of this model implementation’s error. By showing the error as a percentage of the predicted and actual value, the difference is able to be expressed not just as a constant value (as above), but proportionally.

(While these 2 metrics may at glance seem almost identical, they are not: the MAPE measures a similarity of estimates to the key variable, and Average Error shows the mean error of all variables: which outputs a similar number only because the Health Score Risk index is ~10, causing a base-10 aspect in the later metric.)

Avenues for Improvement:

To better this regression model system, the current methods of estimate calculation should be made better calibrated. This can be done by assessing the error values provided by the solution's current test results to implement a better regression line being established, tested, and utilized in the program. Namely, a cycle of gadging error, refining regression line estimations, and retesting the solution should be implemented to find a more precise prediction model.

This process can be further supported by the inclusion of additional evaluation metrics, such as *Root Mean Square Deviation* (*RMSE*) and *Coefficient of Determination* (R^2), which can help provide more perspectives for a more holistic system accuracy.