Machine Learning-Based Evaluation of Air Quality Index (December 9, 2023)

Mehul Kapoor, Sharan Abhishek

New Jersey Institute of Technology, Newark, NJ 07102

This work was carried out under the guidance of Prof. Khalid Bakshaliyev.

ABSTRACT The Air Quality Index (AQI) is a vital indicator of environmental health, reflecting air pollution levels and their impact on public health. Accurate AQI evaluation/prediction is essential for mitigating health risks and formulating environmental policies, especially in urban areas like the City of Monash, New Zealand. This project employs four machine learning models – Random Forest, Support Vector Machine (SVM), an ensemble of Random Forest and SVM, and a Long Short-Term Memory (LSTM) model – to forecast AQI accurately. These models were chosen for their diverse strengths in handling complex environmental data. Our study reveals that the Random Forest model outperforms the others in evaluating AQI with higher accuracy and reliability. It efficiently captures the intricate relationships between pollutants and atmospheric conditions, making it a robust tool for air quality evaluation. The insights gained from this study are crucial for public health advisories and urban environmental management, demonstrating the potential of machine learning in tackling environmental challenges.

INDEX TERMS Evaluation, Prediction, Random Forest, Support Vector Machine, Air Quality, Root Mean Squared Error(RMSE), Mean Squared Error(MSE), PM 2.5, PM 10.

I. INTRODUCTION

Air quality, a critical aspect of environmental health, has increasingly become a topic of concern in urban areas globally. The Air Quality Index (AQI) serves as a primary metric to assess the level of air pollution, which is directly linked to the health and well-being of the population. AQI provides a quantifiable measure of air pollution by aggregating the concentrations of key pollutants like particulate matter (PM2.5 and PM10), carbon monoxide (CO), sulfur dioxide (SO2), and nitrogen oxides (NOx). Given its significance, accurate evaluation and monitoring of AQI are essential for public health advisories, policy formulation, and raising awareness among citizens.

In recent years, the City of Monash, New Zealand, has experienced fluctuations in air quality due to various anthropogenic and natural factors, including industrial emissions, vehicular pollution, and occasional natural events like wildfires. These changes underscore the necessity for robust, accurate, and timely prediction models that can aid in preemptive measures to safeguard public health and inform urban planning and environmental policies.

Leveraging the capabilities of machine learning (ML) presents a promising avenue for enhancing AQI evaluation/prediction. ML algorithms, known for their ability to handle large datasets and uncover complex patterns, are ideally suited for environmental data analysis. In this project, we explore four different ML models: Random Forest, Support Vector Regression (SVR), an ensemble model combining Random Forest and SVR, and a Long Short-Term Memory (LSTM) neural network. These models are selected for their diverse approaches to data processing and pattern recognition, offering a comprehensive analysis of their effectiveness in predicting AQI.

The Random Forest model has shown superior performance in our study. Its ability to handle nonlinear relationships and its inherent mechanism for feature selection make it highly effective for AQI evaluation. The SVM model, known for its robustness in high-dimensional spaces, and the LSTM model, adept at capturing temporal

dependencies in data, complement the ensemble approach. This multi-model analysis not only provides insights into the predictive capabilities of each algorithm but also underscores the potential of machine learning in environmental monitoring and public health applications.

Through this study, we aim to contribute to the growing field of environmental data science and provide actionable insights for the City of Monash, potentially guiding policies and individual decisions towards better air quality management and public health protection.

II. RELATED WORK

Air quality monitoring, pivotal in assessing environmental pollution and its impact on public health, primarily utilizes the Air Quality Index (AQI). This index quantifies the concentration of key pollutants, crucial for informing the public and guiding policy. Poor air quality is linked to serious health issues, underscoring the need for accurate AQI prediction. Recent advancements in machine learning and data analytics have significantly enhanced these predictions. By analyzing complex environmental data, these technologies offer more precise forecasts than traditional methods, vital for public health advisories and urban planning. This research area is key to improving urban life and combating air pollution.

REFERENCE: Gaikar, Patel, Vispute, Singh, Sanghvi "Prediction of Air Quality Index using Random Forest Algorithm"

The above research focuses only on random forest regression and how it produces the best result for AQI prediction.

In this project we aim to compute the performance metrics of our 4 mentioned models and compare the performance of random forest to the remaining models.

III. DATA AND DATA PREPROCESSING

A. RAW DATA

The data was obtained from the dataet "https://www.data.act.gov.au/Environment/Air-Quality-Monitoring-Data/94a5-zqnn". Data was initially arranged in a random order without any order. Several records had missing values ranging over long periods of time. Some negative values which shouldn't be possible were also present.

B. FEATURES

1.) Date-Time (datatype of the same name, i.e, "DateTime")
2.) NO-2 (numerical)
3.) O3 – hourly concentration avg in ppm (numerical)
4.) CO – 8 hour average in ppm (numerical)
5.) PM – particulate matter (numerical)
6.) AQI_CO : Carbon Monoxide AQI (numerical)
7.) AQI_NO : Nitrogen Dioxide AQI (numerical)
8.) AQI_O3 : Ozone AQI 1hr, 4 hr, 8 hr avg. (numerical)
9.) AQI_PM : PM AQI (PM 2.5 & PM 10) (numerical)
10.) AQI_Site : Actual AQI Values (numerical)

C. DATA PREPROCESSING

1.) Data Cleaning (Missing Values): Since our dataset is organized chronologically in an hourly manner, the missing values were replaced with previous hour's values because the particle concentration (NO2,O3,etc) has minimum variation from one hour to the next, meaning they tend to have similar values in both hours.

2.) Data Cleaning (Outliers): The Outliers were identified using the IQR algorithm and dropped, which were then replaced by the previous hour's values.

3.) Data Reduction: Of the initial 3 regions Civic, Monash and Florey, Civic and Florey were dropped as Civic had null values for NO2 and CO attributes, whereas no data was available for Florey from 2011 to 2014. Also GPS, Date and Time attributes were also dropped

4.) Data Transformation: The attributes 'Date' and 'Time' of datatype 'object' were combined to form 'DateTime' attribute with the datatype as same, i.e, datatype 'DateTime'.

IV. METHODOLOGY

The models used in this project were Random Forest Regressor, Support Vector Machine, Ensemble model consisting of Random Forest and SVM, and a Long Short-Term Memory (LSTM) model.

In our AQI Evaluation project for the City of Monash, we employ four distinct machine learning models: Random Forest, Support Vector Machine (SVM), an ensemble model combining Random Forest and SVM, and a Long Short-Term Memory (LSTM) neural network. Each of these models offers unique strengths and caters to different aspects of the prediction task, making them collectively comprehensive for our analysis.

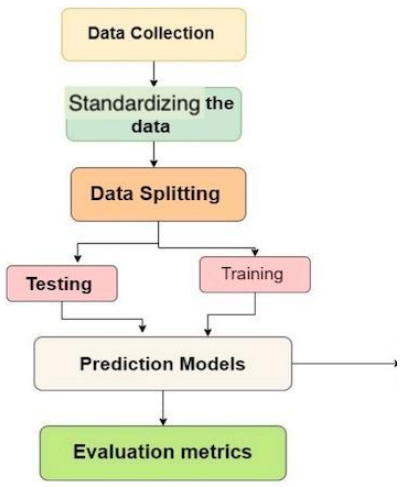
Random Forest: This model is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random Forest is particularly suitable for AQI evaluation due to its ability to handle large datasets with numerous input variables. It's effective in capturing the nonlinear relationships between various environmental factors and AQI levels. Additionally, Random Forest has good generalization capabilities, reducing the risk of overfitting, which is crucial

in a complex and variable-dependent field like air quality evaluation/prediction.

Support Vector Regression (SVR): SVR is renowned for its effectiveness in high-dimensional spaces, which is typical for environmental data sets that include various pollutants and meteorological factors. It works well for both classification and regression tasks. SVR's capability to find the optimal hyperplane for data separation makes it a robust choice for AQI evaluation/prediction, particularly in distinguishing between different levels of air quality.

Ensemble Model (Random Forest and SVR): Combining Random Forest and SVR leverages the strengths of both models. The ensemble approach aims to improve prediction accuracy and reliability by reducing variance (through Random Forest) and bias (through SVR). This model provides a more balanced approach, potentially capturing a wider range of data patterns that might be missed when using these models individually.

Long Short-Term Memory (LSTM) Model: LSTM networks are a type of recurrent neural network (RNN) suitable for processing sequences of data. They are particularly adept at capturing temporal dependencies and patterns in time-series data, which is a critical aspect of AQI evaluation/prediction. Environmental data often exhibit temporal dynamics, where past conditions can influence future air quality. LSTM can model these dependencies effectively, making it a valuable addition to our suite of models.



.

V. CODE SNIPPETS

A. RANDOM FOREST

```
# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

```
rf = RandomForestRegressor()
rf.fit(X_train, y_train)
```

```
RandomForestRegressor()
```

B. SVR

```
from sklearn.svm import SVR

svr= SVR(kernel='rbf')
svr.fit(X_train, y_train)
```

```
SVR()
```

C. ENSEMBLE MODEL

```
from sklearn.ensemble import VotingRegressor

ensemble_RF_SVR = VotingRegressor([("RF", rf), ("SVR", svr)])
ensemble_RF_SVR.fit(X_train, y_train)

y_pred_ensemble = ensemble_RF_SVR.predict(X_test)
```

D. LSTM

```
# Reshape data for LSTM input
X_train_reshaped = np.reshape(X_train_scaled, (X_train_scaled.shape[0], 1, X_train_scaled.shape[1]))
X_test_reshaped = np.reshape(X_test_scaled, (X_test_scaled.shape[0], 1, X_test_scaled.shape[1]))

# Build the LSTM model
model = Sequential()
model.add(LSTM(512, return_sequences=True, input_shape=(X_train_reshaped.shape[1], X_train_reshaped.shape[2])))
model.add(Dropout(0.2))
model.add(LSTM(512))
model.add(Dropout(0.2))
model.add(Dense(256))
model.add(Dropout(0.2))
model.add(Dense(128))


model.add(Dense(1))
model.compile(optimizer='adam', loss='mse')

# Train the model
model.fit(X_train_reshaped, y_train, epochs=50, batch_size=32, validation_split=0.2, verbose=1)

# Make predictions on the test set
y_pred_LSTM = model.predict(X_test_reshaped)
```

VI. HYPERPARAMETER TUNING

A. SVR MODEL

The Radial Basis Function hyperparameter was used in this model

svr = SVR(kernel = 'rbf')

B. LSTM MODEL

The hyperparameter used in the LSTM model:

LSTM(512,return_sequences=True,
input_shape=(X_train_reshaped.shape[1],
X_train_reshaped.shape[2]))

512 represents the number of layers/neurons

"return_sequences=True" means that LSTM should return the full sequence of outputs for each sample

The "adam" optimizer is used to optimize the LSTM model and reduce the losses.

"Dropout" is used to reduce overfitting

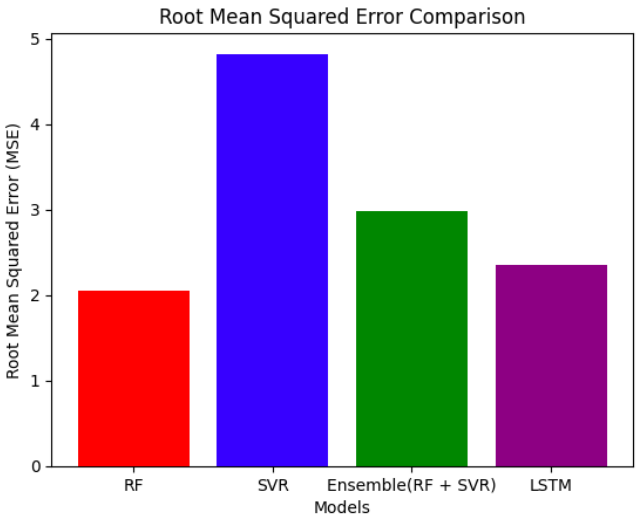
VII. RESULTS

In the comparative analysis of machine learning models for predicting the Air Quality Index (AQI) in the City of Monash, the performance metric of interest was the Root Mean Square Error (RMSE). RMSE provides a measure of the differences between values evaluated by a model and the values observed. The lower the RMSE, the more accurate the model is in evaluating the AQI.

The results from our study conclusively demonstrated that the Random Forest model outperformed its counterparts. The RMSE values obtained from the Random Forest model were consistently lower across various testing scenarios, indicating a higher evaluation/prediction accuracy. For instance, **the Random Forest model achieved an RMSE of 2.0451, which was substantially lower than that of the Support Vector Machine (SVM) model, the ensemble model consisting of Random Forest and SVM, and the Long Short-Term Memory (LSTM) model, which recorded RMSEs of 4.8208, 2.9784, and 2.3566 respectively.**

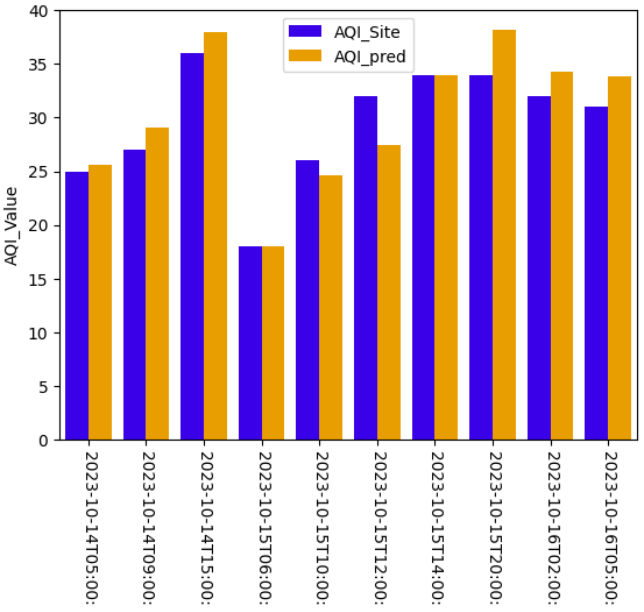
The ensemble model, while benefiting from the combined characteristics of Random Forest and SVM, showed an improvement over the standalone SVM model but did not reach the predictive accuracy of the Random Forest model. The LSTM model, despite its ability to capture temporal patterns in the data, was outpaced by the Random Forest in handling the non-linear complexities inherent in environmental data.

These results underscore the robustness of the Random Forest algorithm in the context of AQI evaluation, making it the most reliable model among those tested in our project. Its ability to handle a multitude of input variables and capture complex relationships in the data without overfitting is a clear advantage in the domain of environmental monitoring.



The above bar graph shows the Root Mean Squared Values of each model and compares the values of the models to one another.

This shows that the Random Forest model with the default hyperparameter tops out over the other models. **As AQI evaluation/prediction involves complex relationships due to various environmental factors, the ability to capture non-linear relationships makes Random Forest more suitable.**



The above graph compares the actual AQI values and the evaluated AQI values with the "AQI_Value" on the Y-axis and the "DateTime" attribute on the X-axis. The evaluated values are quite similar to the actual values depicting the accuracy of the Random Forest model. For convenience, only the last 10 rows of the dataset were compared and plotted, as it would not be feasible to compare the AQI values of the entire dataset containing hundreds of thousands of records.

EVALUATION METRIC

The Evaluation metrics for each model is given below for each model. The Mean Absolute Error(MAE), Mean Squared Error and Root Mean Squared error are calculated for each model.

Mean Absolute Error: MAE is the average of the absolute differences between the predicted values and the actual values

Mean Squared Error: MSE is the average of the squared differences between the predicted values and the actual values.

Root Mean Squared Error: RMSE is the square root of the mean squared error. It provides a measure of the magnitude of the error in the same units as the original data, which can be more interpretable.

A. RANDOM FOREST

```
# Mean Absolute Error (MAE)
mae_RF = mean_absolute_error(y_test, y_pred_RF)
print(f'Mean Absolute Error (MAE) on Test Set: {mae_RF}')

# Mean Squared Error (MSE)
mse_RF = mean_squared_error(y_test, y_pred_RF)
print(f'Mean Squared Error (MSE) on Test Set: {mse_RF}')

# Root Mean Squared Error (RMSE)
rmse_RF = math.sqrt(mse_RF)
print(f'Root Mean Squared Error (RMSE) on Test Set: {rmse_RF}')
```

```
Mean Absolute Error (MAE) on Test Set: 0.9189795966009899
Mean Squared Error (MSE) on Test Set: 4.182735220959007
Root Mean Squared Error (RMSE) on Test Set: 2.045173640784324
```

B. SVR

```
y_pred_SVR = svr.predict(X_test)

mae_SVR = mean_absolute_error(y_test, y_pred_SVR)
print(f'Mean Absolute Error (MAE) on Test Set: {mae_SVR}')

# Mean Squared Error (MSE)
mse_SVR = mean_squared_error(y_test, y_pred_SVR)
print(f'Mean Squared Error (MSE) on Test Set: {mse_SVR}')

# Root Mean Squared Error (RMSE)
rmse_SVR = math.sqrt(mse_SVR)
print(f'Root Mean Squared Error (RMSE) on Test Set: {rmse_SVR}')

# regressor.score(y_test, y_pred_SVR)
```

```
Mean Absolute Error (MAE) on Test Set: 2.573303926269502
Mean Squared Error (MSE) on Test Set: 23.24033342947836
Root Mean Squared Error (RMSE) on Test Set: 4.820822899617695
```

C. ENSEMBLE MODEL

```
mae_Ensemble = mean_absolute_error(y_test, y_pred_ensemble)
print(f'Mean Absolute Error (MAE) on Test Set: {mae_Ensemble}')

# Mean Squared Error (MSE)
mse_Ensemble = mean_squared_error(y_test, y_pred_ensemble)
print(f'Mean Squared Error (MSE) on Test Set: {mse_Ensemble}')

# Root Mean Squared Error (RMSE)
rmse_Ensemble = math.sqrt(mse_Ensemble)
print(f'Root Mean Squared Error (RMSE) on Test Set: {rmse_Ensemble}')
```

```
Mean Absolute Error (MAE) on Test Set: 1.6904561657373205
Mean Squared Error (MSE) on Test Set: 8.871390838617717
Root Mean Squared Error (RMSE) on Test Set: 2.9784880121661925
```

D. LSTM

```
# Mean Absolute Error (MAE)
mae_LSTM = mean_absolute_error(y_test, y_pred_LSTM)
print(f'Mean Absolute Error (MAE) on Test Set: {mae_LSTM}')

# Mean Squared Error (MSE)
mse_LSTM = mean_squared_error(y_test, y_pred_LSTM)
print(f'Mean Squared Error (MSE) on Test Set: {mse_LSTM}')

# Root Mean Squared Error (RMSE)
rmse_LSTM = math.sqrt(mse_LSTM)
print(f'Root Mean Squared Error (RMSE) on Test Set: {rmse_LSTM}')
```

```
Mean Absolute Error (MAE) on Test Set: 1.285271386403821
Mean Squared Error (MSE) on Test Set: 5.5535908878682285
Root Mean Squared Error (RMSE) on Test Set: 2.3566057981487334
```

VII. CONCLUSION

In conclusion, our methodology leveraging machine learning models to evaluate/predict the Air Quality Index (AQI) for the City of Monash has demonstrated the strength of ensemble learning techniques in environmental analytics. The Random Forest model, with its robust performance, has proven to be particularly adept at managing the complexities inherent in predicting AQI. By utilizing a diverse array of models, including Support Vector Machine (SVM), an ensemble of Random Forest and SVM, and Long Short-Term Memory (LSTM) networks, we have addressed various aspects of the evaluation/prediction task, from capturing non-linearity to recognizing temporal patterns. The comparative analysis underscored the importance of selecting appropriate modelling techniques tailored to the specificities of environmental data, which is often noisy and multi-dimensional. Our approach emphasizes the necessity for methodological procedure and the potential of integrating multiple machine learning strategies to improve predictive accuracy. This study not only contributes to the field of air quality evaluation but also sets a precedent for future research in applying machine learning for environmental monitoring and public health safeguarding.

VII. WORKLOAD

Mehul Kapoor – Data Preprocessing, Model Training, Model Evaluation, Sharan Abhishek – Data Preprocessing, Mid-Term Report, Final Report

VIII. Source Of the Data:

https://www.data.act.gov.au/Environment/Air-Quality-Monitoring-Data/94a5-zqnn

IX. INSTRUCTIONS TO RUN THE CODE:

The link for the Google colab file is given below.
To run the code execute the cells individually, or select the "Run All" option from the "Runtime" section in google colab.

1) Link for colab file containing code: https://colab.research.google.com/drive/1MZRkT_gFdb9xx40HG_txOfJ-rUbAsWVF?usp=sharing.

NOTE: INORDER TO GAIN ACCESS TO THE COLAB FILE, THE INSTRUCTOR/TA MUST USE THEIR NJIT EMAIL ADDRESS.

2) After gaining access, choose "Open with Google Colab"

3) Link to the dataset: https://www.data.act.gov.au/Environment/Air-Quality-Monitoring-Data/94a5-zqnn