Data Science Project

Comparison between 4 different Machine Learning models for a Forecast regression based on CO2 sensors data

Leonardo Brighenti

September 17, 2024

Contents

1 Introduction

The objective of this project is to conduct a comparative analysis of four distinct forecast regression models employed in supervised machine learning. The evaluation aims to identify the most effective model based on the Mean Absolute Error (MAE) metric. This study applies these models to a dataset of CO2 concentration measurements collected by a sensor installed in my bedroom. The models under consideration include K-Nearest Neighbors (KNN), Least-squares linear regression, Neural Network and Random Forest.

2 Dataset

The initial dataset comprises data collected from multiple sensors installed in my bedroom, capturing various parameters such as PM1, PM2.5, PM10, CO2 sensor temperature, CO2 concentration, humidity, and ambient temperature. The data collection spans from January 2022 to July 2024, with measurements taken at four-minute intervals. In total, due some turn-off period, the dataset consists of 226135 rows (samples) and 14 columns (features).

Because of sensor faults and data inconsistencies, the analysis for forecasting was concentrated on CO2 concentration data, which demonstrated more reliability and data continuity.

CO2 concentration is primarily affected by two factors: the presence of person in the room, leading to an increase in concentration proportional to the number of people, and the opening of windows, which causes a decrease in concentration up to atmospheric levels. These factors show a recurring daily routine, such as night time increasing and seasonal patterns influenced by common behaviors, such as keeping windows open during the summer months.

2.1 Data Preprocessing

To ensure the effective utilisation of the available information, the raw data were subjected to a preprocessessing phase in order to handle anomaly, zero and NaN (Not a Number) values. The inconsistencies identified included occasional measurements below 400 parts per million (ppm), which represents the lower limit of the atmospheric concentration [1]. As a result of this cleaning process, approximately 1500 samples were removed. Figure 1 illustrates the distribution of the pre-processed data and Figure 2 plot 2 days of data.

2.2 Time Series Segmentation

The goal is to utilize a one hour window sample to forecast 20 minutes ahead while maintaining the data's granularity. Following the data cleaning phase, the next step involved slicing the data to prepare it for input into the forecasting model. Initially, an 80-minute time slicer was created to gather data arrays within that time frame. This approach was necessary due to sensor faults or inactivity that could ruin the continuity within that span. Upon examining the majority length of these arrays, which were 19 and 20, the decision was made to adopt 20 as the standard. Arrays with 19 samples

[1] Resource: https://climate.nasa.gov/vital-signs/carbon-dioxide/?intent=121

were interpolated to 20, resulting in a total of 10,938 arrays. This approach ensures that by taking the first 15 data points as the window and the last point as target, the data integrity is maintained.

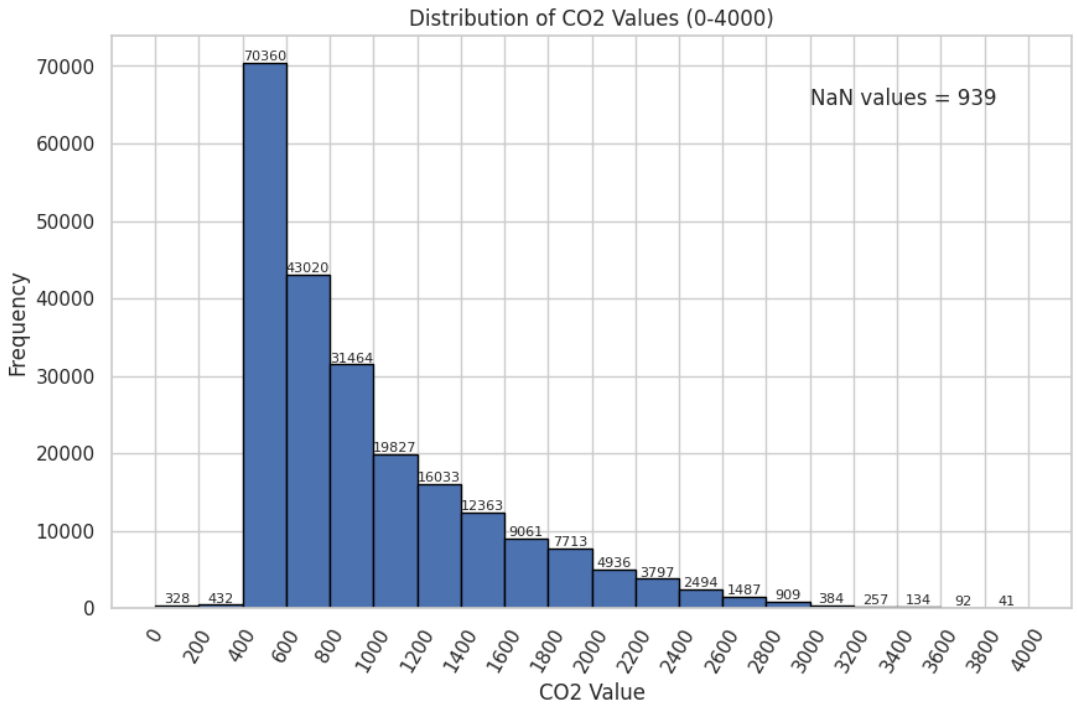

Figure 1: Data distribution of CO2 concentration, before preprocessing.

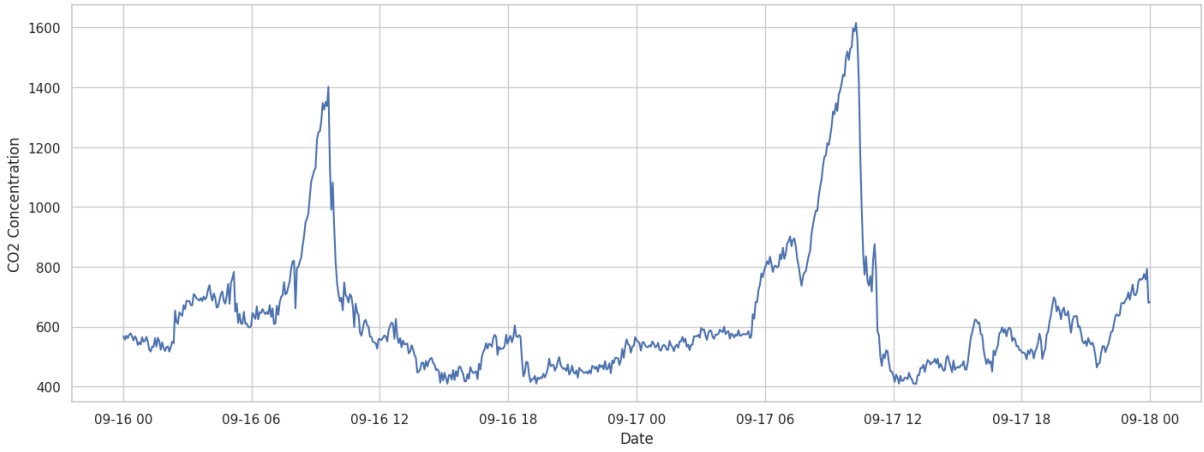

Figure 2: 2 days plot od CO2 concentration.

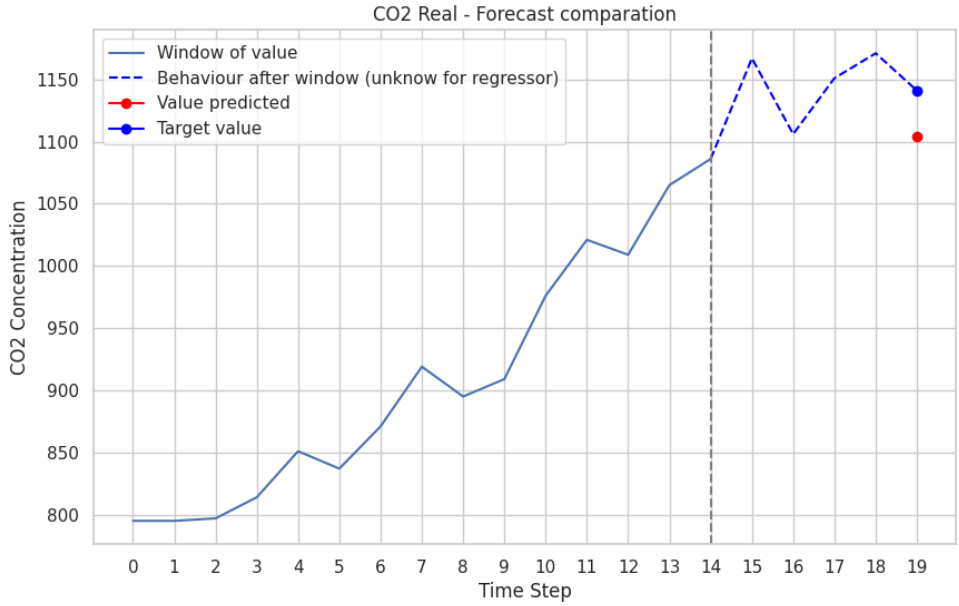

Figure 3: Graph of real behaviour of data and prediction, with Linear Regressor.

3 Model comparison

The data consists of a window segment (60 minutes) paired with the corresponding target (the data point at 80 minutes from the start of the window). In Figure 3 is show a graphical representation.

For the training and testing phases, the dataset was split:

- KNN, Least-squares linear regression, and Random Forest: 80% training and 20% testing sets.
- Neural Network: 70% training, 15% testing, and 15% validation (for the selection of the best model).

The models were trained and subsequently evaluated using their respective test sets, with Mean Absolute Error (MAE) as performance metric. The metrics are detailed in Table 1.

For KNN, the optimal result was chosen by varying the hyperparameter K from 0 to 30. A evolution of MAE over K is depicted in Figure 4

Model	**MSE Score**
Neural Network	89.11
Random Forest	90.94
Linear Regression	91.48
KNN Regression (N=27)	96.03

Table 1: Comparison of forecast models

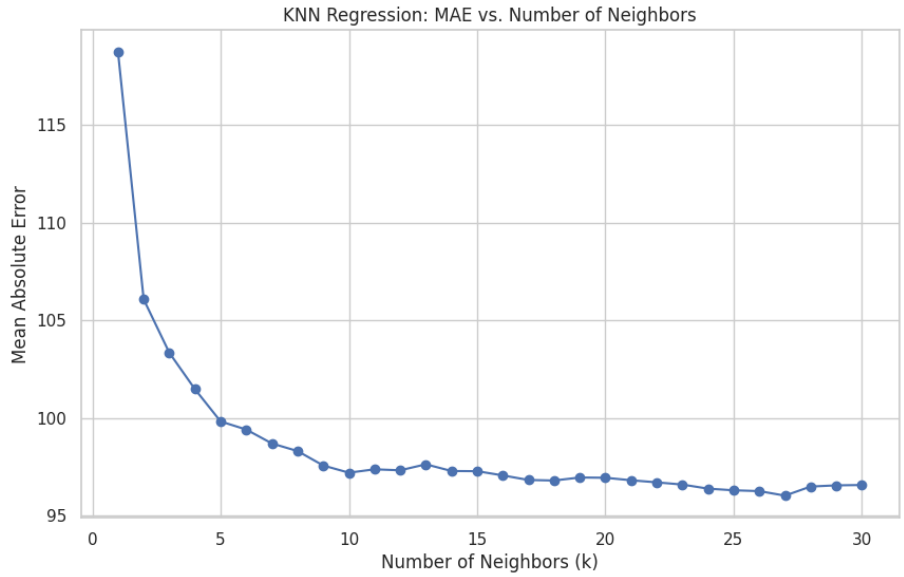

Figure 4: Variation of the Mean Absolute Error across different values of K in the K-Nearest Neighbors model.

4 Conclusion

The Neural Network model stands out as the best performer, achieving the lowest prediction error among the models. The Random Forest model is also highly effective, displaying closely competitive results with the Neural Network. While Linear Regression offers moderate accuracy, its performance is somewhat limited by its linear nature, although it highlights the linear behavior on these window data. The KNN model is less suitable for this dataset, as indicated by its higher error rate.

A final note regarding the preprocessing that was applied to all sensor data to address inconsistencies and missing values. Additionally, a correlation matrix was computed to identify relationships between features, which may potentially enhance forecast accuracy in future analyses.