Name: *Dario Cavada*
Course: *Ban668DE_FA23 Python Programming/Data*
Assignment: *Final Project*
Due Date: *Saturday, October 14th*

Table of Contents

Abstract

The research findings have broad practical applications in weather forecasting, renewable energy, construction, agriculture, environmental monitoring, disaster response, transportation, urban planning, and climate change research. These insights empower better decision-making and resource management across various sectors. The analysis revealed that the K-Nearest Neighbors (KNN) regression model outperformed other models in predicting wind speed. This model achieved a Mean Absolute Error (MAE) of approximately 1.63 units, signifying its superior accuracy in forecasting wind speed. For the owner of a wind farm, the superior accuracy of the KNN regression model in predicting wind speed directly translates into increased energy production, reduced operational costs, and a more reliable and efficient wind energy sector. This finding has a tangible contributes to the broader goals of sustainability and clean energy production (Example).

Data Origin and Overview

This research involved the analysis of a weather-related dataset comprising 14 variables and nearly 16,743 observations, sourced from the CORGIS Dataset Project. The summary statistics give an overview of the data's distribution and characteristics, including the count (number of data points), mean (average value), standard deviation (a measure of data spread), minimum, 25th percentile (Q1), median (50th percentile), 75th percentile (Q3), and maximum values for each column. The dataset contains weather-related measurements, such as precipitation, temperature, and wind data, likely collected over a span of time that includes the years 2016 and 2017.

2. Data Preprocessing

2.1 Data Cleaning

I didn’t need to clean too much the data since it was kind of clean. However, the names of the variables were confusing and wordy. I renamed the columns in my DataFrame to make them more concise and user-friendly. I started renaming each column one by one using this formula:

df = df.rename(columns={‘NAME OF THE OLD VARIABLE’: ‘NAME OF THE NEW VARIABLE’})

With these lines of code, I renamed each of the specified columns in my DataFrame to more concise and descriptive names, making the DataFrame easier to work with and understand. This was super helpful due I was dealing with a large dataset, and with this I could ensure clarity and consistency in my data analysis.

2.2 Missing Values

I tried to find any missing values but didn’t find any missing values in my dataset.

2.3 Extreme Outliers

I tried to find extreme outliers in the three different new variables I created through "Mutation". For two of them, 'AVG Temp' & 'Wind Direction' I didn't find any outliers. However, for 'Wind Speed' I found many, a total of 57 rows in my DataFrame that contain extreme outliers in the 'Wind Speed' column, according to the provided upper and lower limits. Then I decided to replace all those extreme values through the second method, using the median of 'Wind Speed'. I proved later that there were not outliers remaining.

2.4 Data Mutation

For data mutation I was looking for adding two new variables to my dataset that provide me with useful information for my analysis. First, I calculated the mean of 'Wind Speed' to set up the conditions accordingly.

I added meaningful categorizations to my dataset based on the 'Wind Speed' and 'AVG Temp' columns. For 'Wind Speed,' I established two conditions:

- The first condition identifies 'Dangerous' wind conditions when 'Wind Speed' exceeds 6.33.
- The second condition labels conditions as 'Safe' when 'Wind Speed' is at or below 6.33.

To achieve this, I created a new column, 'SafetyWind,' in my DataFrame using the np.select function. This column now contains values of 'Dangerous' or 'Safe' corresponding to the defined conditions.

Moving on to temperature categorization, I set up three conditions for 'AVG Temp':

- 'Extremely Cold' when 'AVG Temp' is less than -27.
- 'Good Temperature' for values between -27 and 80.
- 'Extremely Hot' for temperatures above 80.

Just like 'SafetyWind,' I used np.select to create a new column called 'MeasureTemp' in my DataFrame, which now holds values such as 'Extremely Cold,' 'Good Temperature,' or 'Extremely Hot' based on the temperature conditions I specified.

In summary, I enriched my DataFrame with two new columns: 'SafetyWind' and 'MeasureTemp,' making it easier to assess wind safety and temperature conditions with the respective categorizations 'Dangerous'/'Safe' and 'Extremely Cold'/'Good Temperature'/'Extremely Hot.'

3. Exploratory Data Analysis

3.1 Numerical Variables

There were a few numerical variables in the dataset, such as Precipitation, Month, Week, Ye ar, AVG Temp,

MAX Temp, MIN Temp, Wind Direction, Wind Speed. I calculated in first the average temper ature in between all cities and resulted 56.1. I decided to get in a scatterplot 'Wind Speed' a nd 'Wind Direction' and got meaningful results as find it in the "Appendix" in the last page. I found out that when the direction of the wind is between 23 and 29 approximately is when gets the highest speed.

3.2. Categorical Variables

In the dataset there were two categorical variables 'City' and 'State'. For 'City' I found out 307 different cities, being Norfolk, Jackson, Springfield, Charleston and Newark the most frequent ones with 106. For 'State', the ones more frequent were Alaska (1719), Texas (1272) & California (999). I plotted the new variable created, 'MeasureTemp', which has the three categories of 'Extremely cold', 'Good Temperature' and 'Extremely Hot'. The resulted showed that most observations were in 'Extremely Hot' and 'Good Temperature', 'Extremely Hot' getting close to 8000 and 'Good Temperature' around 7500. For 'Extremely Hot' resulted a total of 1500 observations approximately.

4. Comparison of Supervised Machine Learning Models

I conducted a regression analysis to forecast wind speed. The initial step involved transforming the dataset into a numerical format. Subsequently, I systematically categorized variables into numeric and categorical types, proceeding to encode the categorical variables, specifically 'City' and 'State,' to make them compatible for modeling.

Prior to model evaluation, I rigorously ensured that all variables were in numerical form. The dataset was then divided into distinct training and testing sets for subsequent analysis. Three regression models were subjected to evaluation:

- **Linear Regression:** This model yielded a Mean Absolute Error (MAE) of approximately 1.67 units.
- **Decision Tree Regression:** The decision tree regression model was found to have an MAE of roughly 1.87 units.
- **K-Nearest Neighbors (KNN) Regression:** Among the models considered, the KNN regression model demonstrated superior accuracy, with an MAE of approximately 1.63 units.

The exceptional performance of the KNN regression model suggests that, on average, its predictions deviate by approximately 1.63 units from the actual wind speed values. In essence, this signifies that the KNN regression model is the optimal choice for forecasting wind speed within the scope of this analysis.

5. Implications

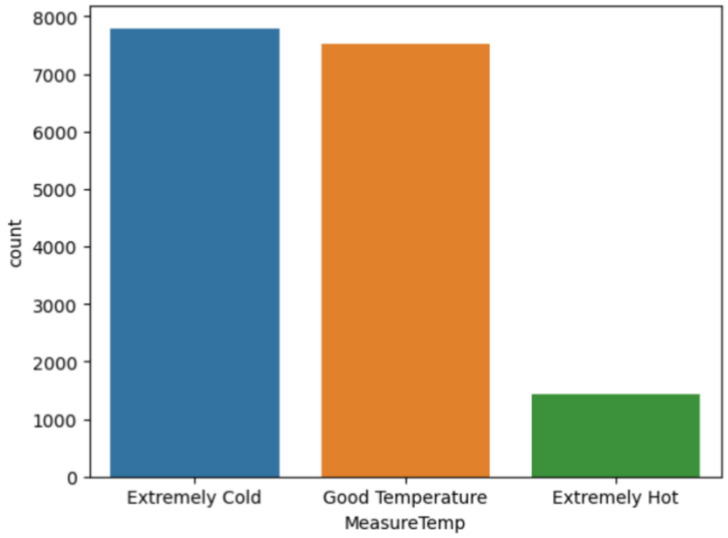
In this research, I conducted a comprehensive analysis of a weather-related dataset with 14 variables and nearly 16,743 observations. The dataset was obtained from the CORGIS Dataset Project, and it provided valuable insights into weather conditions during the years

2016 and 2017. The research findings, centered around the analysis of weather-related data and the precise prediction of wind speed, have far-reaching practical implications across multiple real-life sectors. Weather forecasting agencies can enhance the accuracy of their forecasts, ensuring better flight planning and safety in aviation. Wind power and renewable energy companies can optimize energy generation and reduce costs, while construction and engineering projects can rely on accurate wind speed data for safety and structural integrity. Agriculture benefits from informed decisions on planting and crop management, while environmental agencies can use this information for public health concerns. Emergency management, transportation, outdoor activities, urban planning, and climate change research all stand to gain valuable insights from this research. In sum, the findings offer a comprehensive resource that can drive improved decision-making, safety, and efficiency across a spectrum of industries and services, ultimately contributing to better resource management and outcomes.

6. Appendix

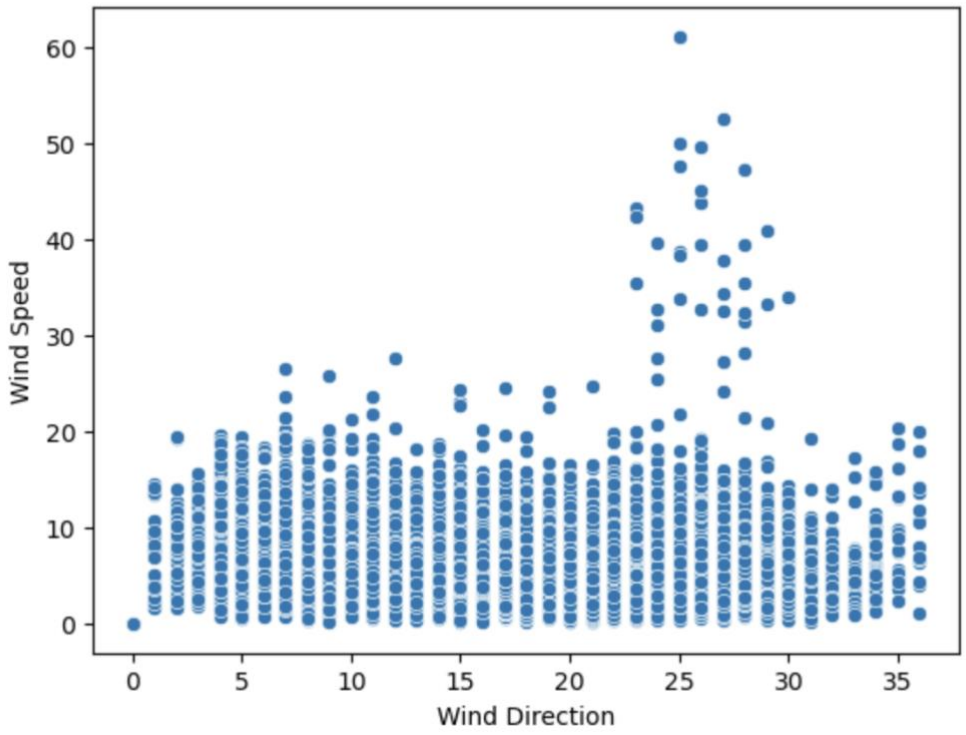
```
[60]: sns.countplot(data=df, x='MeasureTemp', order=['Extremely Cold', 'Good Temperature', 'Extremely Hot'])
```

```
[60]: <Axes: xlabel='MeasureTemp', ylabel='count'>
```



[138]:
```
sns.scatterplot(data=df, x='Wind Direction', y='Wind Speed')
```

[138]:
```
<Axes: xlabel='Wind Direction', ylabel='Wind Speed'>
```



[139]:
```
sns.histplot(data=df, x='AVG Temp')
```

[139]:
```
<Axes: xlabel='AVG Temp', ylabel='Count'>
```