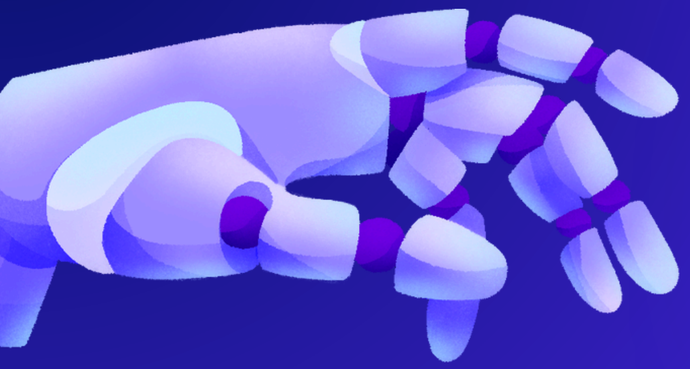
# DATA MINING PROJECT

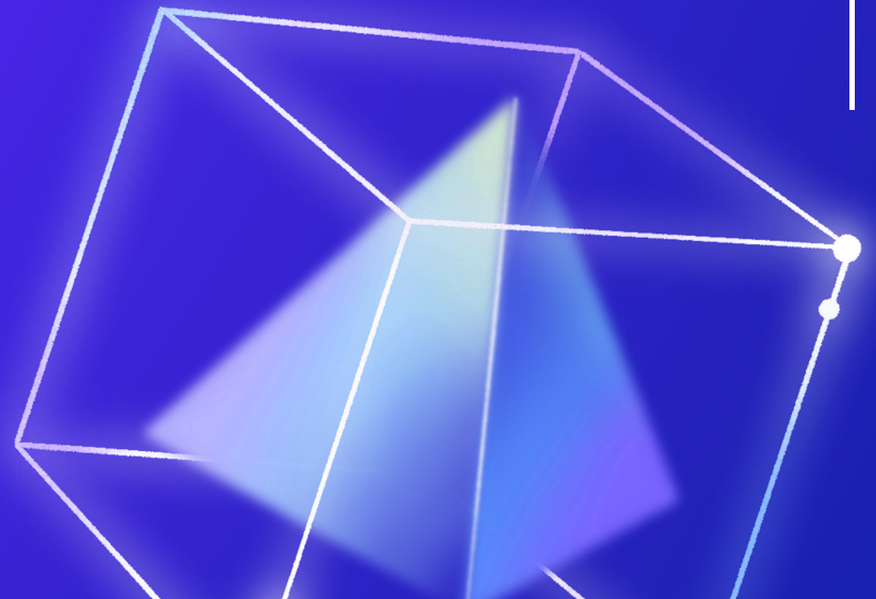
Prof. Dr. ZAFER ASLAN


Yusuf Çetin - B2005.010056
Eren Karagöz - B2005.010046
Yusuf Karakaş - B2005.010034
Ahmet Bedirhan Arvas - B2005.010018
Suzan Simay Topalyan - B2005.010068



# TABLE OF CONTENTS


- Introduction 01
- Literature Review 02
- Methodology 03
- Conclusion 04

# INTRODUCTION

01

- Definition of The Topic: This project analyzes wind speed and direction data from Beşkonak and Manavgat, focusing on temporal variations. Linear interpolation addresses missing data, while min-max normalization prepares data for machine learning. Key patterns are identified using statistics and visualization, supporting applications in weather forecasting, renewable energy, and disaster management.

02

- Definition of The Problem: Wind speed and direction data are crucial for applications like renewable energy and disaster management but often suffer from missing values due to sensor issues, disrupting analysis. Effective preprocessing, including handling missing data and normalization, is essential for accurate forecasting and machine learning applications.

# LITERATURE REVIEW

Wind speed and direction analysis is vital in meteorology, renewable energy, and planning, emphasizing the need for complete and accurate data. Linear interpolation effectively addresses missing data, while min-max scaling ensures uniformity for machine learning models. Descriptive statistics, visualization, and advanced algorithms like time series forecasting enhance accuracy, highlighting the importance of AI and preprocessing in wind data analysis.

# METHODOLOGY

01 Data Preprocesses

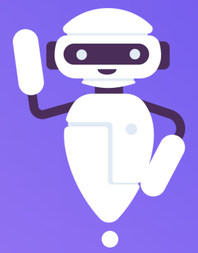
02 Descriptive Statistics

03 Time Series Analyses

04 Clustering and Classification

05 Future Simulations

# DATA PREPROCESSES



## HISTOGRAM

A histogram shows the distribution of numerical data in machine learning.

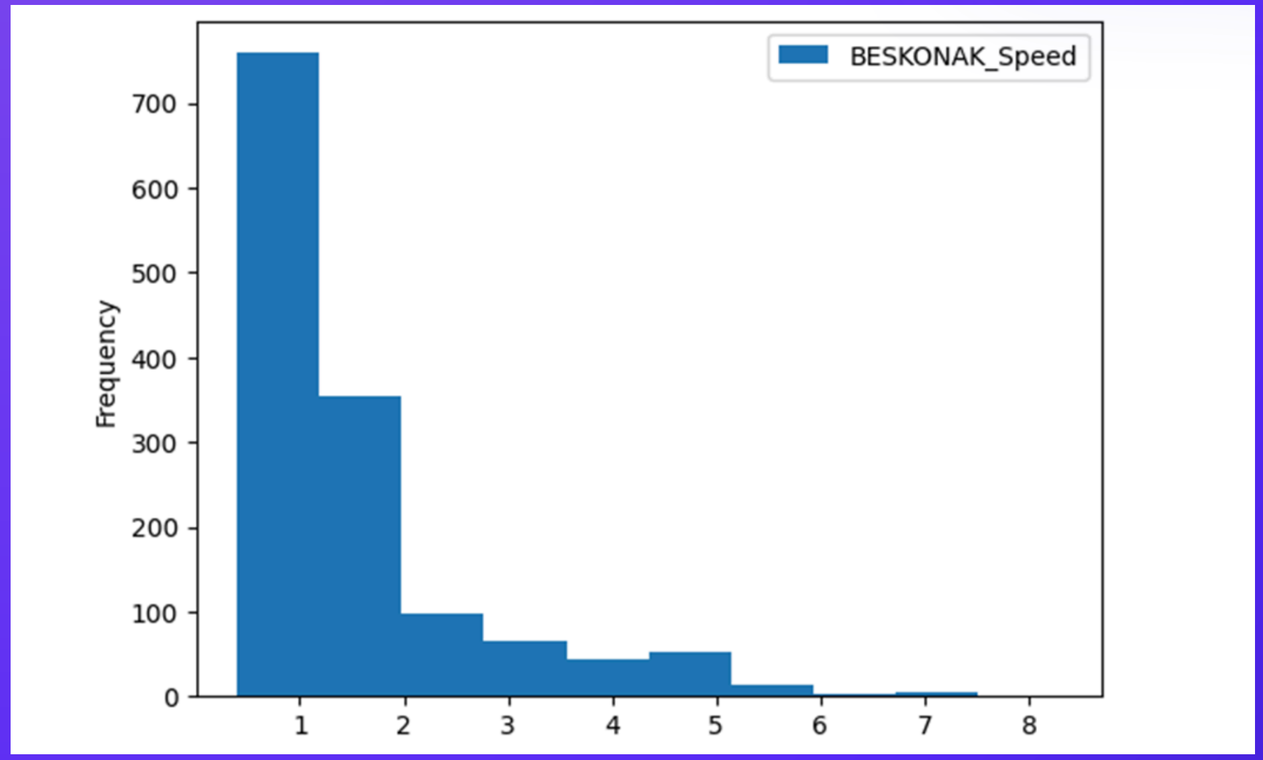

## QQ-PLOT

A QQ-Plot compares data distribution to a theoretical distribution.

## SHAPIRO-WILK TEST

The Shapiro-Wilk Test checks data for normality.

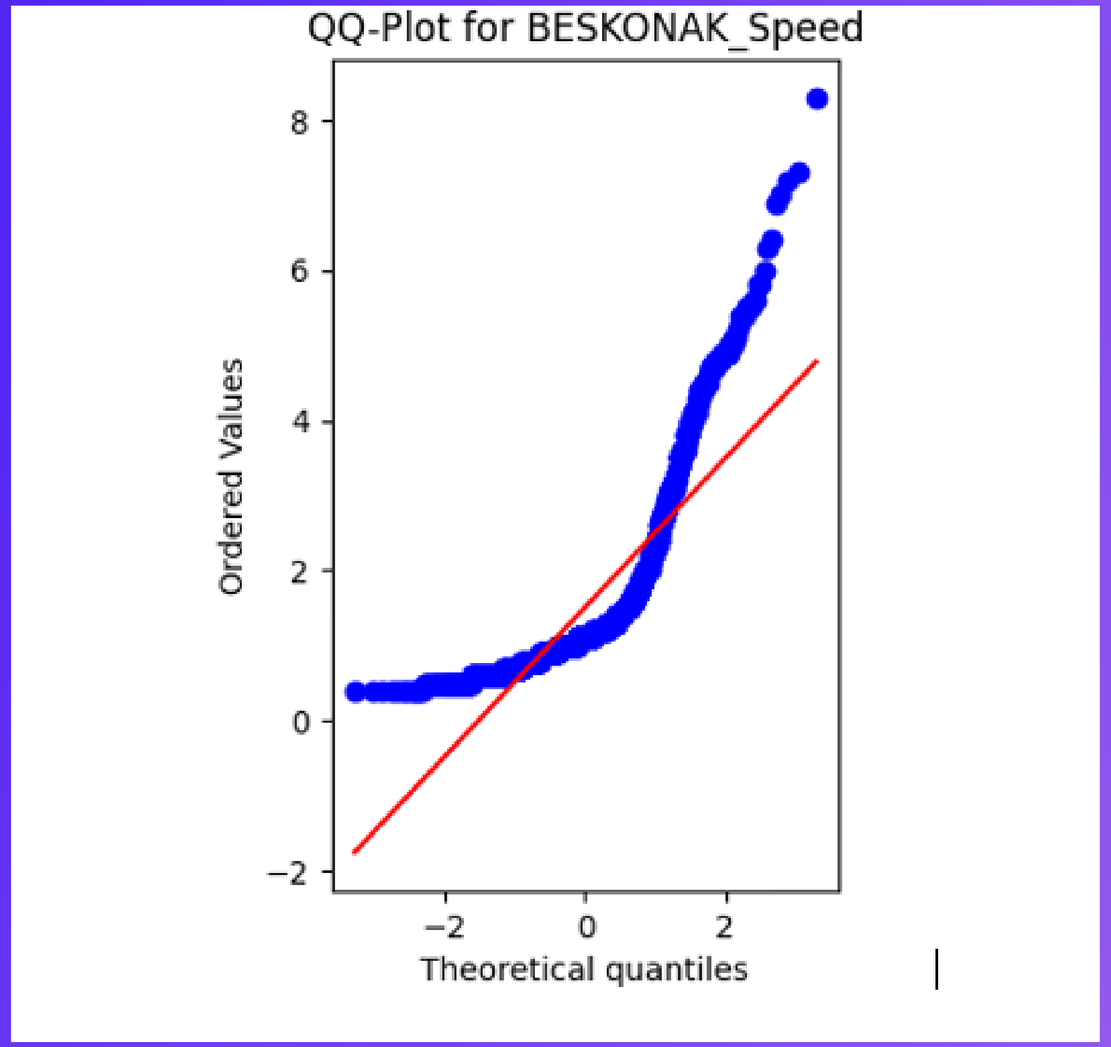
# HISTOGRAM




This histogram shows the frequency distribution of the BESKONAK_Speed data. The majority of the data is concentrated at lower speed values, and the frequency decreases as the speed increases. This indicates that the data exhibits a positively skewed distribution.

# QQ-PLOT

This QQ plot compares the observed values of the BESKONAK_Speed data with the theoretical quantiles of a normal distribution. The significant deviation of the points, especially at the upper tails, from the red reference line indicates that the data does not follow a normal distribution and exhibits positive skewness. The clustering of points below the reference line in the lower quantiles also supports the deviation from normality. This clearly demonstrates that the data does not conform to a normal distribution.



# SHAPIRO-WILK TEST

```
beskonak_speed = df['BESKONAK_Speed']
manavgat_dir = df['MANAVGAT_Dir']

shapiro_beskonak = stats.shapiro(beskonak_speed)
shapiro_manavgat = stats.shapiro(manavgat_dir)

print("Shapiro-Wilk Test Results:")
print(f"BESKONAK_Speed: Statistic={shapiro_beskonak.statistic}, p-value={shapiro_beskonak.pval
print(f"MANAVGAT_Dir: Statistic={shapiro_manavgat.statistic}, p-value={shapiro_manavgat.pvalue
```

```
Shapiro-Wilk Test Results:
BESKONAK_Speed: Statistic=0.7343169450759888, p-value=1.816082809764963e-42
MANAVGAT_Dir: Statistic=0.8875799179077148, p-value=1.1891153210827797e-30
```

The normality of the BESKONAK_Speed and MANAVGAT_Dir variables has been analyzed. According to the test results, the p-values for both variables ($p < 0.05$) indicate that the normality assumption is not met. In other words, the data does not follow a normal distribution.

# DESCRIPTIVE STATISTICS

## MEAN MODE

The mean is a measure of central tendency obtained by dividing the sum of all values in a dataset by the number of elements in that dataset.

The mode is a measure of central tendency that represents the most frequently occurring value in a dataset.


## MIN - MAX QUARTILES

Min and Max represent the smallest (min) and largest (max) values in a dataset, which are fundamental indicators for understanding the distribution of the data.

Quartiles are values that divide a dataset into four equal parts; they include the 1st quartile (Q1), the median (Q2), and the 3rd quartile (Q3).

## STANDARD DEVIATION - VARIANCE

Standard deviation is a statistical measure that quantifies the amount of variation or dispersion of values in a dataset from the mean.

Variance is the average of the squared differences from the mean of a dataset and is equal to the square of the standard deviation.

Descriptive statistics summarize the main features of the dataset. Metrics like mean, median, mode, standard deviation, and range are used to understand wind speed and direction's central tendency and variability. Visualizations like histograms and boxplots help examine data distribution and identify outliers.

[228]:
```
df[["BESKONAK_Speed","MANAVGAT_Dir"]].describe().T
```

[228]:

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **BESKONAK_Speed** | 1392.0 | 1.511494 | 1.156781 | 0.4 | 0.8 | 1.1 | 1.6 | 8.3 |
| **MANAVGAT_Dir** | 1392.0 | 200.508980 | 122.433845 | 1.0 | 91.0 | 206.0 | 324.0 | 360.0 |

[230]:
```
mode_value = df["BESKONAK_Speed"].mode()[0]
mode_count = df["BESKONAK_Speed"].value_counts()[mode_value]
print("Mode value: ", mode_value, "Count of the Mode value: ", mode_count)
```

```
Mode value:  1.0 Count of the Mode value:  146
```

[232]:
```
mode_value = df["MANAVGAT_Dir"].mode()[0]
mode_count = df["MANAVGAT_Dir"].value_counts()[mode_value]
print("Mode value: ", mode_value, "Count of the Mode value: ", mode_count)
```

```
Mode value:  336.0 Count of the Mode value:  18
```

[236]:
```
var_bes =  df["BESKONAK_Speed"].var()
print("Variance of Beskonak_Speed: ", var_bes)
```

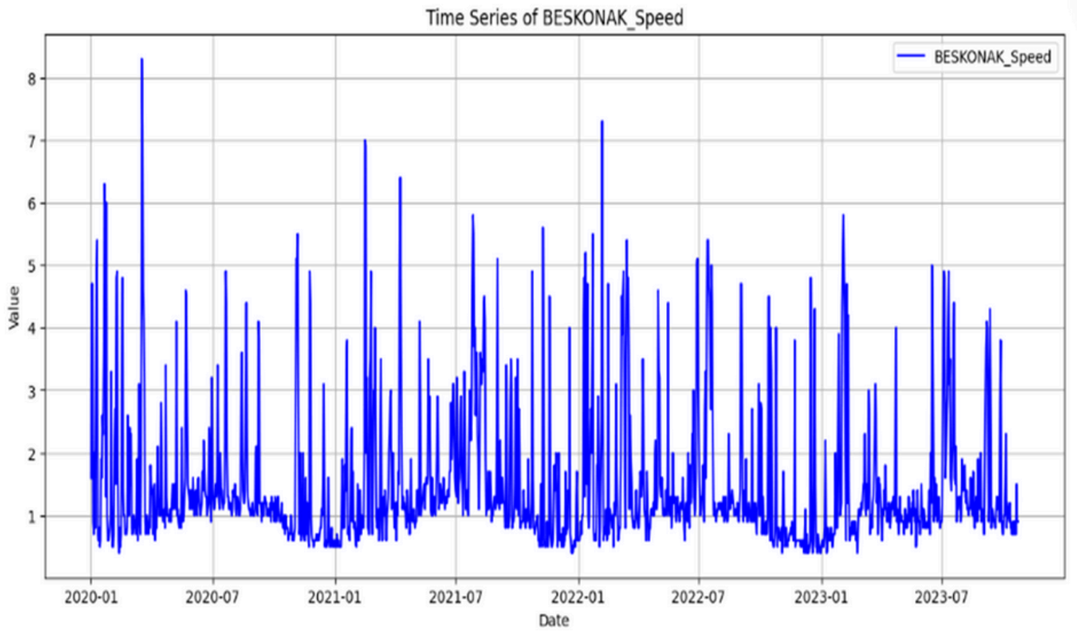
```
Variance of Beskonak_Speed:  1.3381424097440966
```

[238]:
```
var_man = df["MANAVGAT_Dir"].var()
print("Variance of Manavgat_Dir", var_man)
```

```
Variance of Manavgat_Dir 14990.046308791674
```

# TIME SERIES ANALYSES

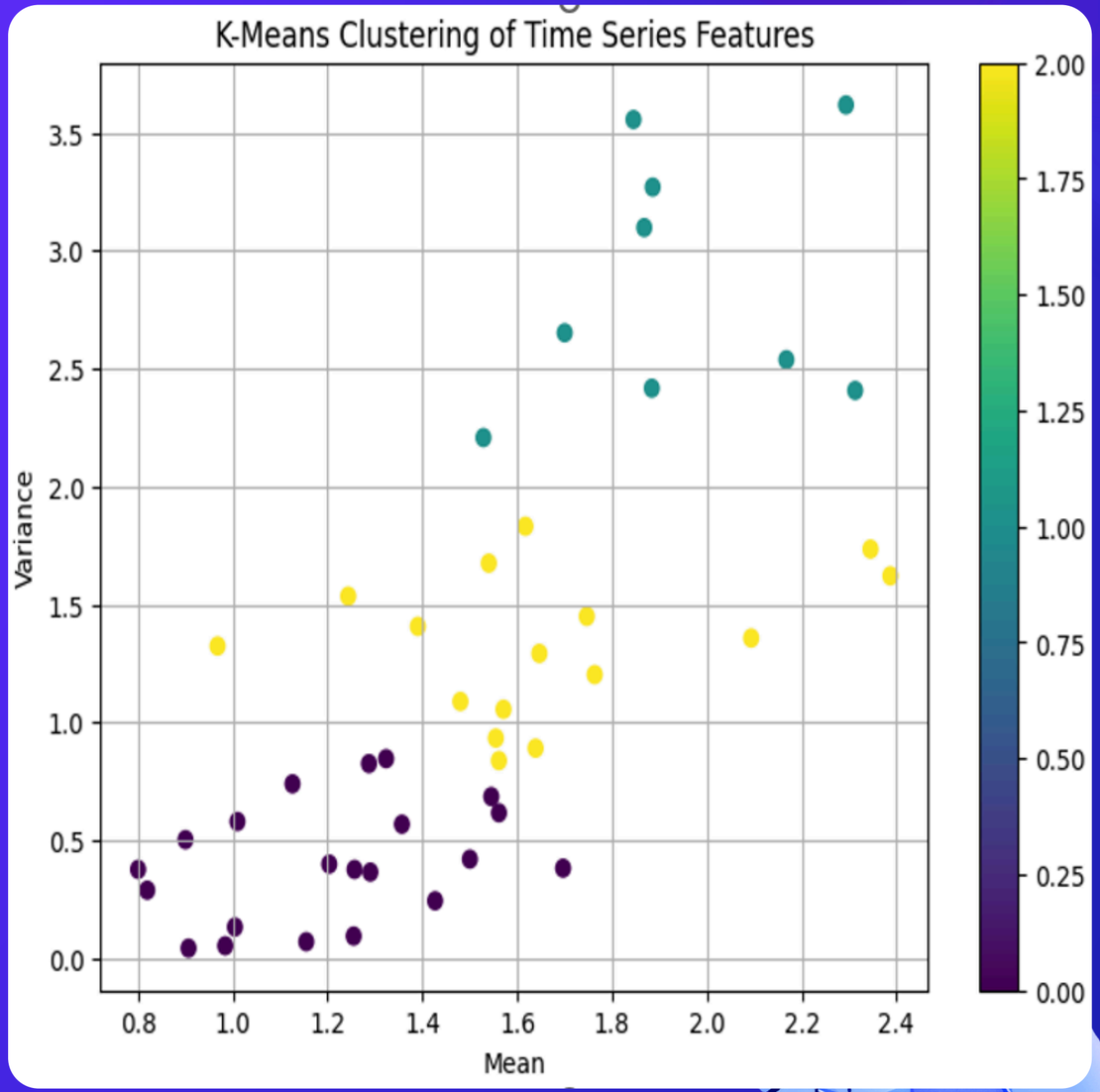
```
df['Date'] = pd.to_datetime(data[['Year', 'Month', 'Day']])

time_series_data = data[['Date', 'BESKONAK_Speed']].dropna()

plt.figure(figsize=(14, 6))
plt.plot(time_series_data['Date'], time_series_data['BESKONAK_Speed'], label='BESKONAK_Speed', color='blue')
plt.title('Time Series of BESKONAK_Speed')
plt.xlabel('Date')
plt.ylabel('Value')
plt.legend()
plt.grid(True)
plt.show()
```



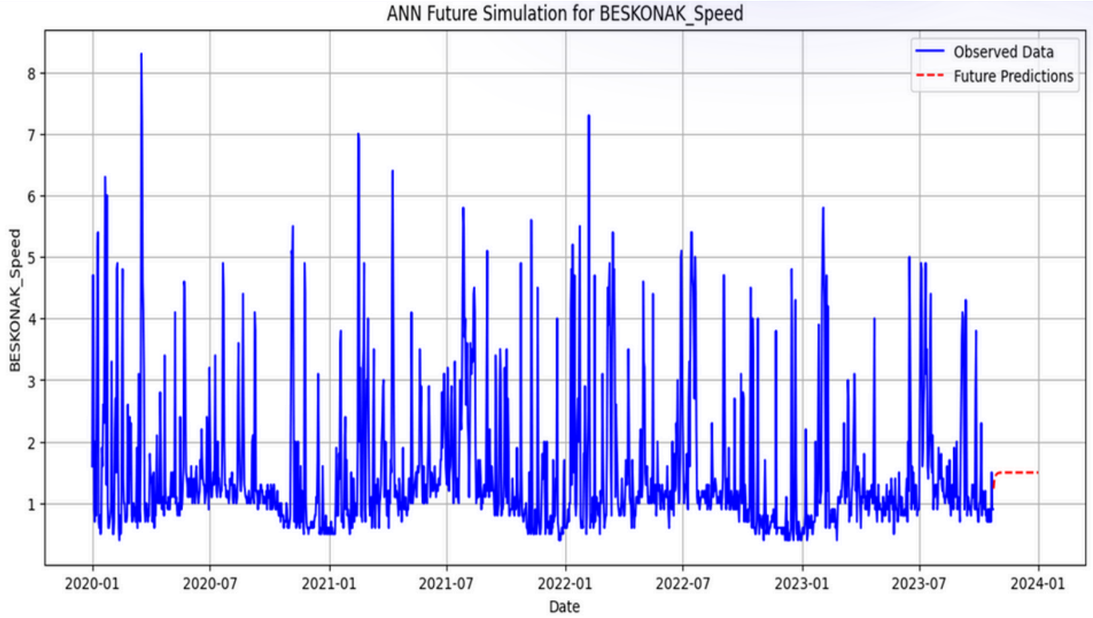

Time series analysis focuses on uncovering patterns and trends in the wind speed data over time. Techniques such as autocorrelation and moving averages are utilized to explore temporal dependencies. The aim is to detect seasonal patterns, anomalies, and potential long-term trends that could inform future forecasting models.

# CLUSTERING AND CLASSIFICATION

In this visualization, the mean and variance values of time series features are classified using the K-Means clustering method. Data points are divided into three distinct clusters, represented by different colors, and the color scale on the right indicates which cluster each data point belongs to. Each cluster reflects the grouping of data points based on their similarities.



# FUTURE SIMULATIONS




Model trained and predictions saved to 'future_predictions.xlsx'

[178]:
```
dff = pd.read_excel("future_predictions.xlsx")
dff.head(10)
```

[178]:

| | Date | Predicted_BESKONAK_Speed |
|---|---|---|
| 0 | 2023-10-24 | 1.235929 |
| 1 | 2023-10-25 | 1.354182 |
| 2 | 2023-10-26 | 1.402091 |
| 3 | 2023-10-27 | 1.443250 |
| 4 | 2023-10-28 | 1.462885 |
| 5 | 2023-10-29 | 1.475087 |
| 6 | 2023-10-30 | 1.484342 |
| 7 | 2023-10-31 | 1.489643 |
| 8 | 2023-11-01 | 1.492789 |
| 9 | 2023-11-02 | 1.494813 |

Machine learning models, such as regression and neural networks, are used to simulate and predict future wind speed and direction. The models utilize preprocessed and normalized data to generate forecasts, offering valuable insights for applications like energy planning and disaster preparedness. Simulations also assess the accuracy and reliability of predictive models in real-world scenarios.

# CONCLUSION

01
- Wind Behavior

02
- Data Analysis

03
- Preprocessing

04
- Machine Learning

05
- Prediction

This study analyzed wind behavior by addressing missing data with linear interpolation and preparing it for machine learning using min-max normalization.

The findings highlight the importance of preprocessing for accurate forecasting, with applications in renewable energy, weather prediction, and disaster management.

# THANK YOU!