T.C.

ISTANBUL AYDIN UNIVERSITY
SOFTWARE ENGINEERING DEPARTMENT

DATA MINING
PROJECT/REPORT

2024 / 2025

ACADEMICIAN

Prof. Dr. Zafer ASLAN

Prepared by.
Cihan ÖZÜN – B2005.090074 – Software Engineering
Elif Tuğçe KESLER - B2005.090061 – Software Engineering
Eylül DUMAN – B2005.090075 – Software Engineering
Göksu Sude GÜNDOĞDU – B2005.090059 – Software Engineering
Hacire HARMAN - B2005. 090039 – Software Engineering
Melike Büşra ŞENER - B2005.090098 – Software Engineering
Merve YILMAZ – B2005.090069 – Software Engineering

Contents

1. REPORT DESCRIPTION

Wind energy has the potential to reduce the world's energy problems. The following report, based on data mining, examines wind energy as one such renewable source of energy. This work makes use of data comprising measurement recordings of wind speed and direction between 2020 and 2023 for further analysis by data mining techniques that help in identifying and understanding energy production capacity.

The aims of the report can be listed as:

- Identify trends and patterns in wind behavior.
- Group similar wind conditions to better inform energy production strategies.
- Predict future wind speeds for operational planning.

2. LITERATURE REVIEW

Due to growing energy needs and the negative effects of fossil fuel combustion, renewable energy sources are taking an even greater importance. Regarding the cost of installation and the cost of generation, wind energy seems to be favorable. Nevertheless, the stochastic nature of wind power introduces complications and more accurate predictions are essential for control of turbines, load planning and energy markets.

Data mining, the process of finding patterns in large datasets, has been a big help in improving wind power forecasting. Here’s how it’s used in wind energy:

Key Applications

1. Data Preprocessing

There are often gaps, noise, or anomalies in the raw data from weather sensors and wind turbines. Preprocessing procedures are necessary to clean and get data ready for analysis. These steps basically include addressing missing values, lowering noise, and normalizing the data.

Here, gaps are filled by methods like interpolation and data dependability is preserved by outlier identification techniques.

2. **Descriptive Statistics**

Descriptive use measures like mean wind speed, std dev and frequency distributions to describe the wind data. Historical data analysis shows patterns in wind behaviour, turbine performance and seasonal energy variations.

3. **Time Series Analysis**

In order to carry out the forecasting it is important to first develop a concept of how wind varies with time. Wind speed and power prediction is done by employing various models and techniques such as ARIMA, LSTMs and Fourier transforms. The predictions are crucial for operational planning, grid integration and meeting the energy demands.

4. **Clustering and Classification**

In clustering, turbines or regions are grouped according to performance metrics such as capacity factors or maintenance requirements with the help of techniques like k-means and hierarchical clustering. Some of the classification models that are used include decision trees and SVMs to identify faults and classify the operational status of the turbines in order to enhance the efficiency and reliability of the turbines..

5. **Future Simulations**

Predictive simulations are used to determine energy output in view of different possible conditions that may exist such as variations in turbine design, weather conditions or even policy changes. Thus, simulations that incorporate data from diverse sources can aid in decision making as well as system optimization.

Tools and Techniques:

- *Supervised Learning*: Regression and classification models for energy forecasting and fault detection.
- *Unsupervised Learning*: Clustering for turbine performance and anomaly detection.
- *Time-Series Analysis*: ARIMA and machine learning models for wind speed and energy output.
- *Visualization Tools*: Tableau, Matplotlib, Seaborn to visualize and communicate results.

By using these data mining techniques the wind energy sector can improve performance, increase forecasting accuracy and tackle the challenges of integrating renewables into the grid.

3. METHODOLOGY

Importing necessary libraries

In [2]:
```
import pandas as pd
import numpy as np
```

Loading the dataset

In [3]:
```
file_path = 'wind_data.xlsx'
data = pd.read_excel(file_path)
```

Renaming columns

In [4]:
```
data.columns = ['Year', 'Month', 'Day', 'Direction', 'Speed']
```

Checking for missing values

In [5]:
```
missing_values = data.isnull().sum()
print("Missing values per column:", missing_values)
```

```
Missing values per column: Year         0
Month        0
Day          0
Direction    0
Speed        0
dtype: int64
```

Converting data types

In [6]:
```
data['Year'] = data['Year'].astype(int)
data['Month'] = data['Month'].astype(int)
data['Day'] = data['Day'].astype(int)
data['Direction'] = data['Direction'].astype(int)
data['Speed'] = pd.to_numeric(data['Speed'], errors='coerce')
```

Descriptive Statistics

Calculating basic statistics

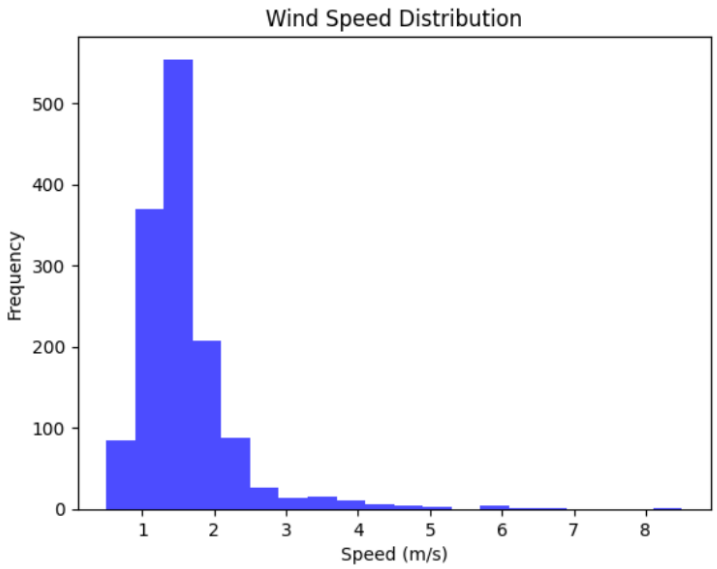
In [7]:
```
descriptive_stats = data.describe()
print(descriptive_stats)
```

```
              Year        Month          Day    Direction        Speed
count  1392.000000  1392.000000  1392.000000  1392.000000  1392.000000
mean   2021.424569     6.284483    15.660920   158.014368     1.569181
std       1.093764     3.356308     8.780982    60.259108     0.713850
min    2020.000000     1.000000     1.000000     1.000000     0.500000
25%    2020.000000     3.000000     8.000000   136.750000     1.200000
50%    2021.000000     6.000000    16.000000   160.000000     1.500000
75%    2022.000000     9.000000    23.000000   181.000000     1.800000
max    2023.000000    12.000000    31.000000   360.000000     8.500000
```

Visualizing data distribution

In [8]:
```
import matplotlib.pyplot as plt

plt.hist(data['Speed'], bins=20, color='blue', alpha=0.7)
plt.title('Wind Speed Distribution')
plt.xlabel('Speed (m/s)')
plt.ylabel('Frequency')
plt.show()
```

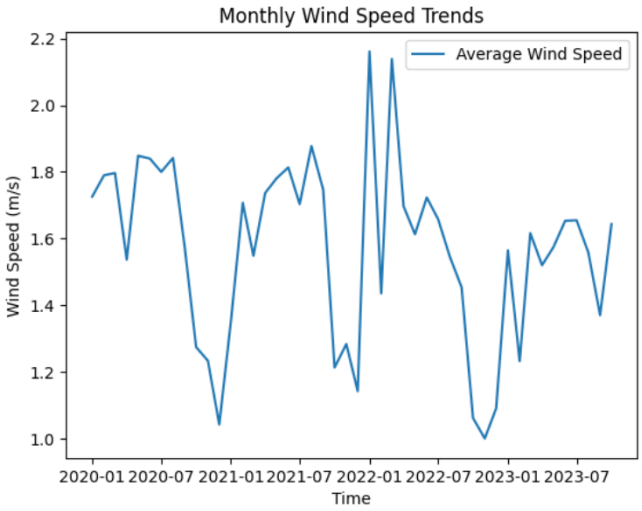


TIME SERIES ANALYSIS

Aggregating data for monthly analysis

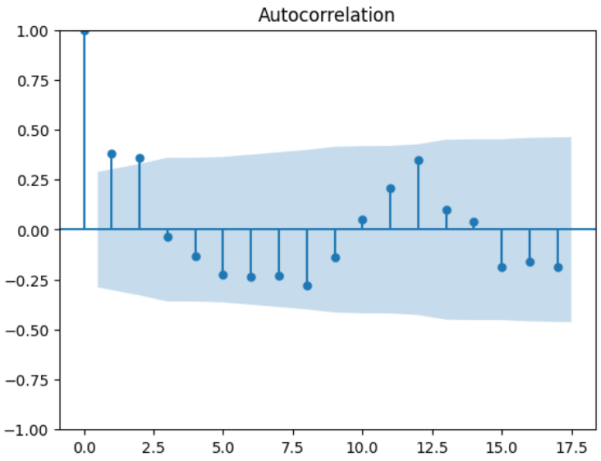
In [9]:
```
data['Date'] = pd.to_datetime(data[['Year', 'Month', 'Day']])
monthly_data = data.groupby(data['Date'].dt.to_period('M')).mean()

# Plotting wind speed trends
plt.plot(monthly_data.index.to_timestamp(), monthly_data['Speed'], label='Average Wind Speed')
plt.title('Monthly Wind Speed Trends')
plt.xlabel('Time')
plt.ylabel('Wind Speed (m/s)')
plt.legend()
plt.show()
```



Autocorrelation analysis

In [10]:
```
from statsmodels.graphics.tsaplots import plot_acf
plot_acf(monthly_data['Speed'])
plt.show()
```



CLUSTERING AND CLASSIFICATION

Clustering using K-means

In [11]:
```
from sklearn.cluster import KMeans
```

Selecting features

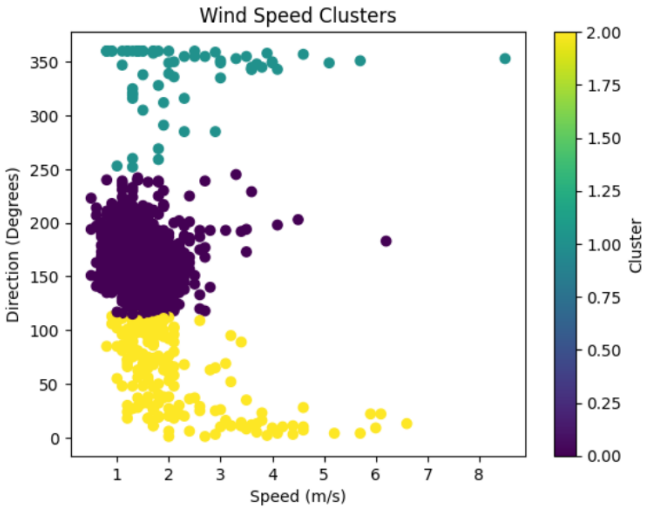
In [12]:
```
features = data[['Speed', 'Direction']]
```

Applying K-means clustering

In [13]:
```
kmeans = KMeans(n_clusters=3, random_state=42)
data['Cluster'] = kmeans.fit_predict(features)
```

Visualizing clusters

In [14]:
```
plt.scatter(data['Speed'], data['Direction'], c=data['Cluster'], cmap='viridis')
plt.title('Wind Speed Clusters')
plt.xlabel('Speed (m/s)')
plt.ylabel('Direction (Degrees)')
plt.colorbar(label='Cluster')
plt.show()
```



Decision tree classification

In [15]:
```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
```

Preparing data

In [16]:
```
X = data[['Speed', 'Direction']]
y = data['Cluster']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Training the model

In [17]:
```
dt = DecisionTreeClassifier()
dt.fit(X_train, y_train)
```

Out[17]:
```
▾ DecisionTreeClassifier

DecisionTreeClassifier()
```

Evaluating the model

In [18]:
```
y_pred = dt.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f"Decision Tree Accuracy: {accuracy * 100:.2f}%")
```

```
Decision Tree Accuracy: 100.00%
```

FUTURE SIMULATIONS

Linear regression model for prediction

In [19]:
```
from sklearn.linear_model import LinearRegression
```

Preparing data

```
In [20]: X = np.array(data.index).reshape(-1, 1)  # Using index as a feature
         y = data['Speed']
```

Splitting data

```
In [21]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Training the model

```
In [22]: model = LinearRegression()
         model.fit(X_train, y_train)
```

```
Out[22]: ▾ LinearRegression
         LinearRegression()
```

Making predictions

```
In [23]: y_pred = model.predict(X_test)
```

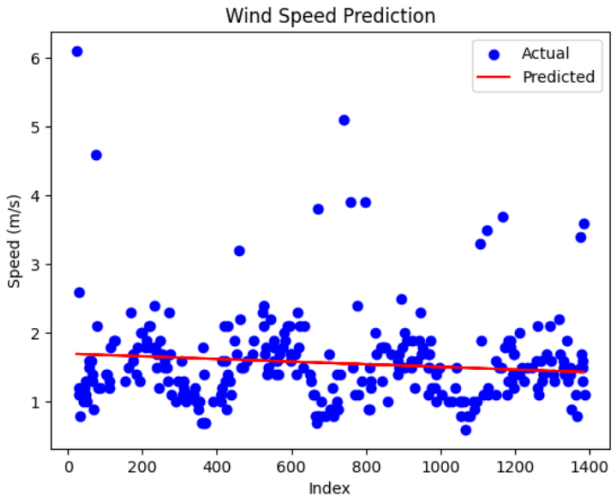
Evaluating performance

```
In [24]: from sklearn.metrics import mean_squared_error
         mse = mean_squared_error(y_test, y_pred)
         print(f"Mean Squared Error: {mse}")
```

```
Mean Squared Error: 0.45906594858285
```

Plotting predictions

```
In [25]: plt.scatter(X_test, y_test, color='blue', label='Actual')
         plt.plot(X_test, y_pred, color='red', label='Predicted')
         plt.title('Wind Speed Prediction')
         plt.xlabel('Index')
         plt.ylabel('Speed (m/s)')
         plt.legend()
         plt.show()
```



4. CONCLUSION

The wind speed data analysis reveals many important findings regarding the temporal evolution of the data, statistical characteristics, and the difficulties in making wind speed predictions, which has implications in a wide range of applications.

A comparison between the actual and the predicted wind speeds show that the current model does not capture the variability and the extreme values thereby causing under fitting of the data. This may indicate that the model does not incorporate all the dynamics of the change in wind speed. In order to enhance the performance, it is advisable to include more meteorological

variables including temperature, pressure or humidity besides using sophisticated forecasting models such as the ensemble model or LSTM.

Clustering analysis presents the distinct patterns of wind, which present potential for model focused on specific conditions. The most common circumstances include slow winds (1-2 m/s) with fixed directions and intermediate winds (2-5 m/s) with more random directions while high winds (>= 5m/s) are infrequent but important in describing catastrophic events. Thus, it will be possible to enhance the accuracy of predictions for the crucial applications, such as the extreme weather forecasting, by creating specific models for each of the identified clusters.

From the above analysis, it can be seen that there is the requirement for more advanced models and tailored methods for the effective utilization of wind speed data for different purposes. Thus, it is important to recognize the variations in the temporal as well as spatial wind dynamics in order to enhance the effective utilization of renewable energy sources, accurate weather forecasting, and solution to other wind related issues.

5. REFERENCES

[1] Colak, I., Sagiroglu, S., & Yesilbudak, M. (2012). Data mining and wind power prediction: A literature review. Renewable energy, 46, 241-247.

[2] Ding, Y. (2019). Data science for wind energy. Chapman and Hall/CRC.

[3] Jaber, S. (2013). Environmental impacts of wind energy. Journal of Clean Energy Technologies, 1(3), 251-254.

[4] Saidur, R., Rahim, N. A., Islam, M. R., & Solangi, K. H. (2011). Environmental impact of wind energy. Renewable and sustainable energy reviews, 15(5), 2423-2430.

[5] Wang, Z., Wang, C., & Wu, J. (2016). Wind energy potential assessment and forecasting research based on the data pre-processing technique and swarm intelligent optimization algorithms. Sustainability, 8(11), 1191.

[6] Liu, H., & Chen, C. (2019). Data processing strategies in wind energy forecasting models and applications: A comprehensive review. Applied Energy, 249, 392-408.