Onur Kula

Final Report:

Wind Turbune Active Power Prediction

Problem Statement

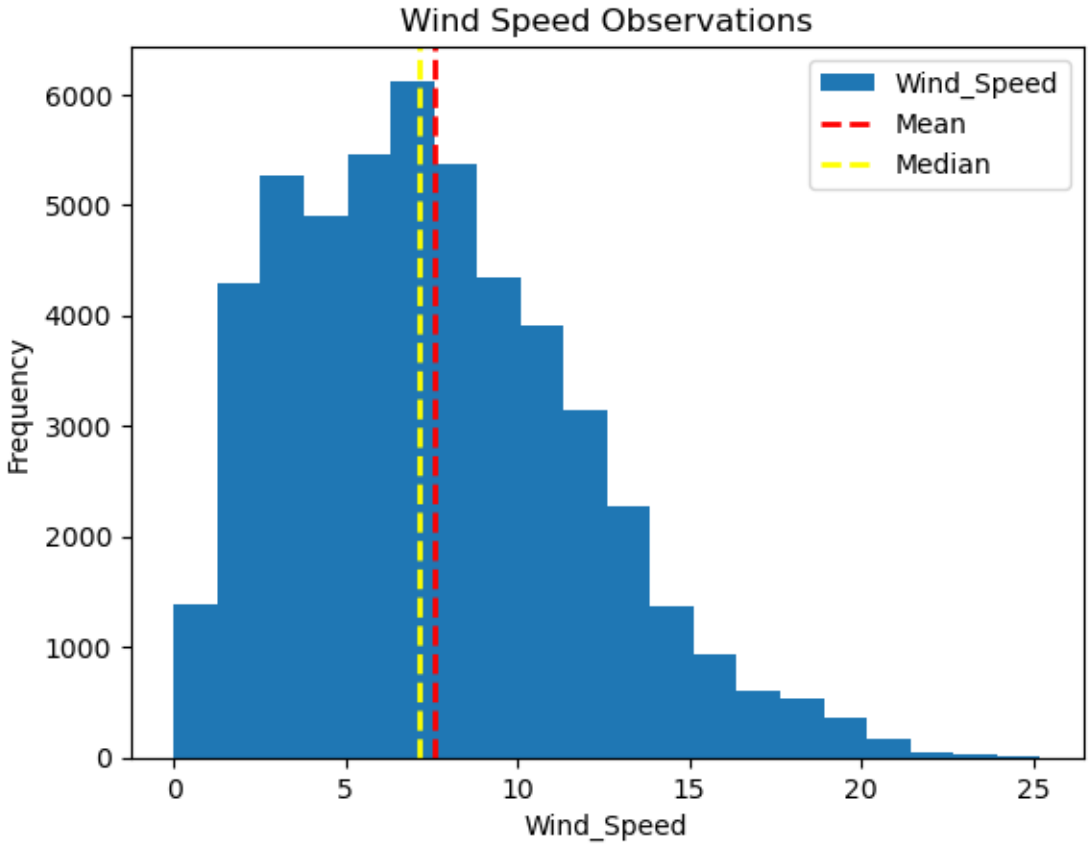
The primary objective of this project is to develop an effective predictive model for wind turbine energy generation within a 2-week timeline. Utilizing data sourced from an operational wind turbine's SCADA system in Turkey, recorded at 10-minute intervals throughout the year 2018, the project aims to create a robust predictive model. This model will be deemed successful if the predictions closely align with the real data, achieving an accuracy rate of 90% or more. Key variables, including Date/Time, LV ActivePower (kW), Wind Speed (m/s), Theoretical_Power_Curve (KWh), and Wind Direction (°), will be crucial in achieving this level of precision. The success criteria emphasize the importance of providing accurate and reliable predictions to contribute to a comprehensive understanding of factors influencing wind turbine performance and enhancing overall energy production efficiency.

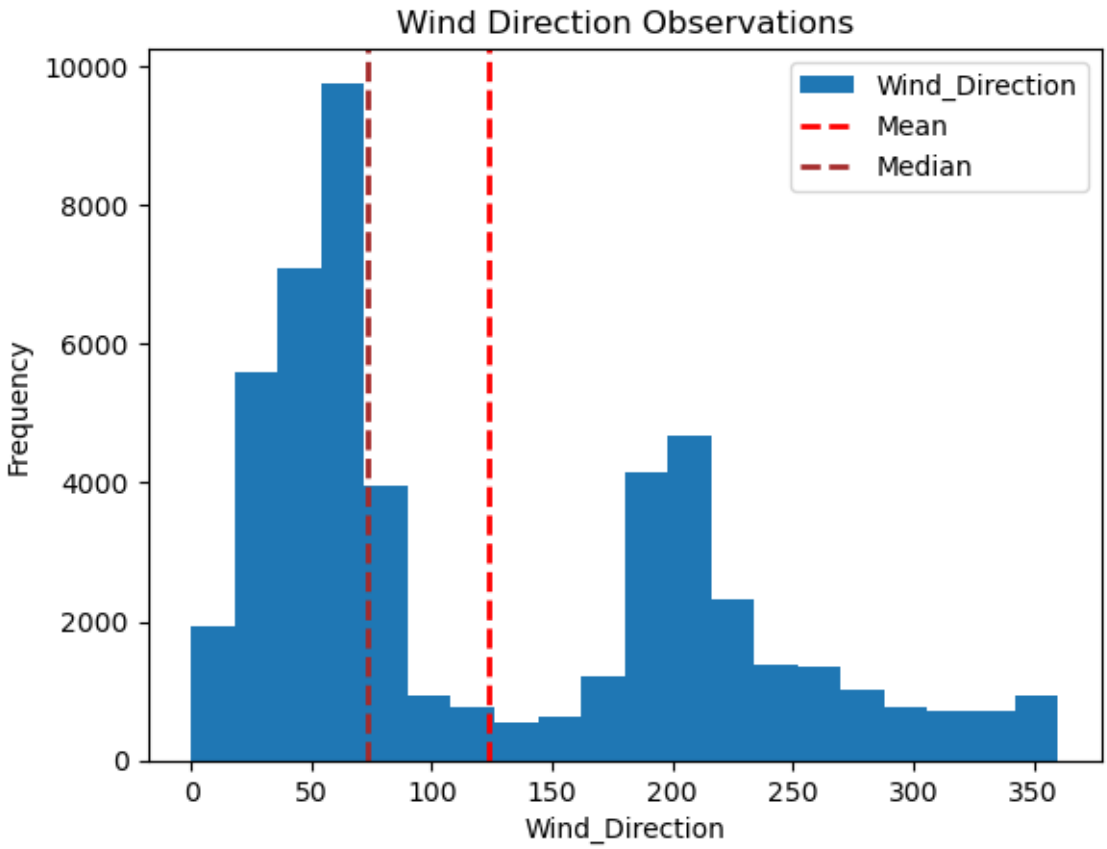
Data Wrangling

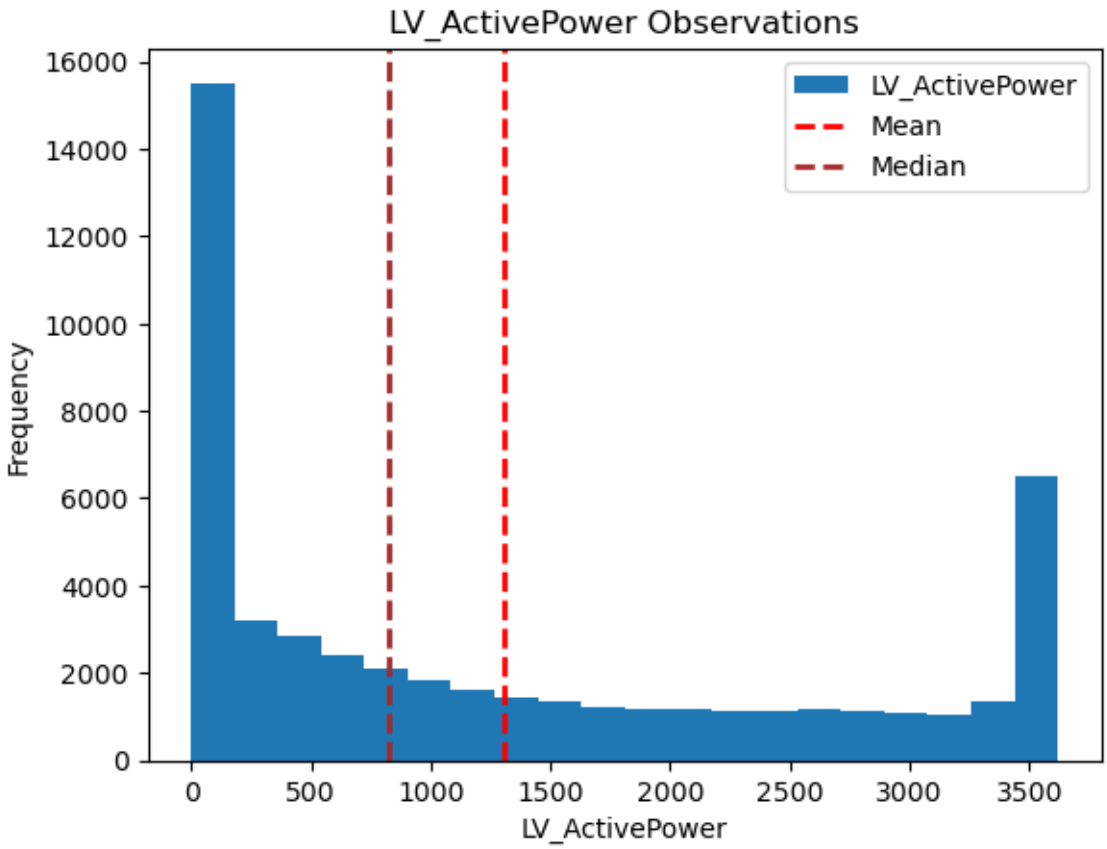
The wind turbine dataset comprises various measurements taken at 10-minute intervals throughout the year 2018. With 5030 rows and 5 columns, it includes data on wind speed, LV active power, theoretical power, and wind direction. Technically, values for wind speed, LV active power, theoretical power, and wind direction cannot be less than 0. Therefore, we've removed such entries from our dataset. After performing the necessary data wrangling steps, our refined dataset now consists of 50473 rows

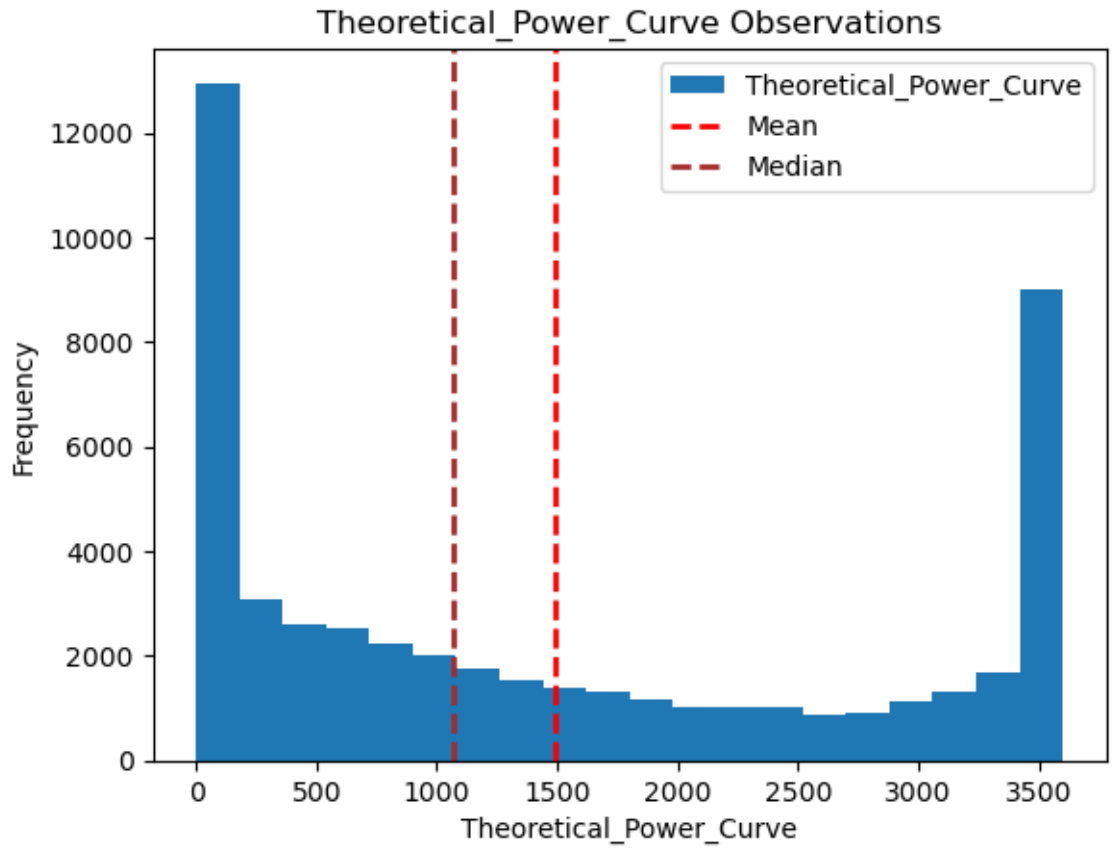
Exploratory Data Analysis

In this section, we visualized the distributions of the columns in our dataset through graphs, allowing us to observe their respective distributions and the average values they possess.




Wind Direction Observations
Wind_Direction
Mean
Median
Frequency
10000
8000
6000
4000
2000
0
0
50
100
150
200
250
300
350
Wind_Direction


LV_ActivePower Observations
LV_ActivePower
Mean
Median
Frequency
16000
14000
12000
10000
8000
6000
4000
2000
0
0
500
1000
1500
2000
2500
3000
3500
LV_ActivePower



In figures 1 and 2, we explored the relationships and correlation between these variables. We visualized how they interact with each other and displayed their correlation relationships.

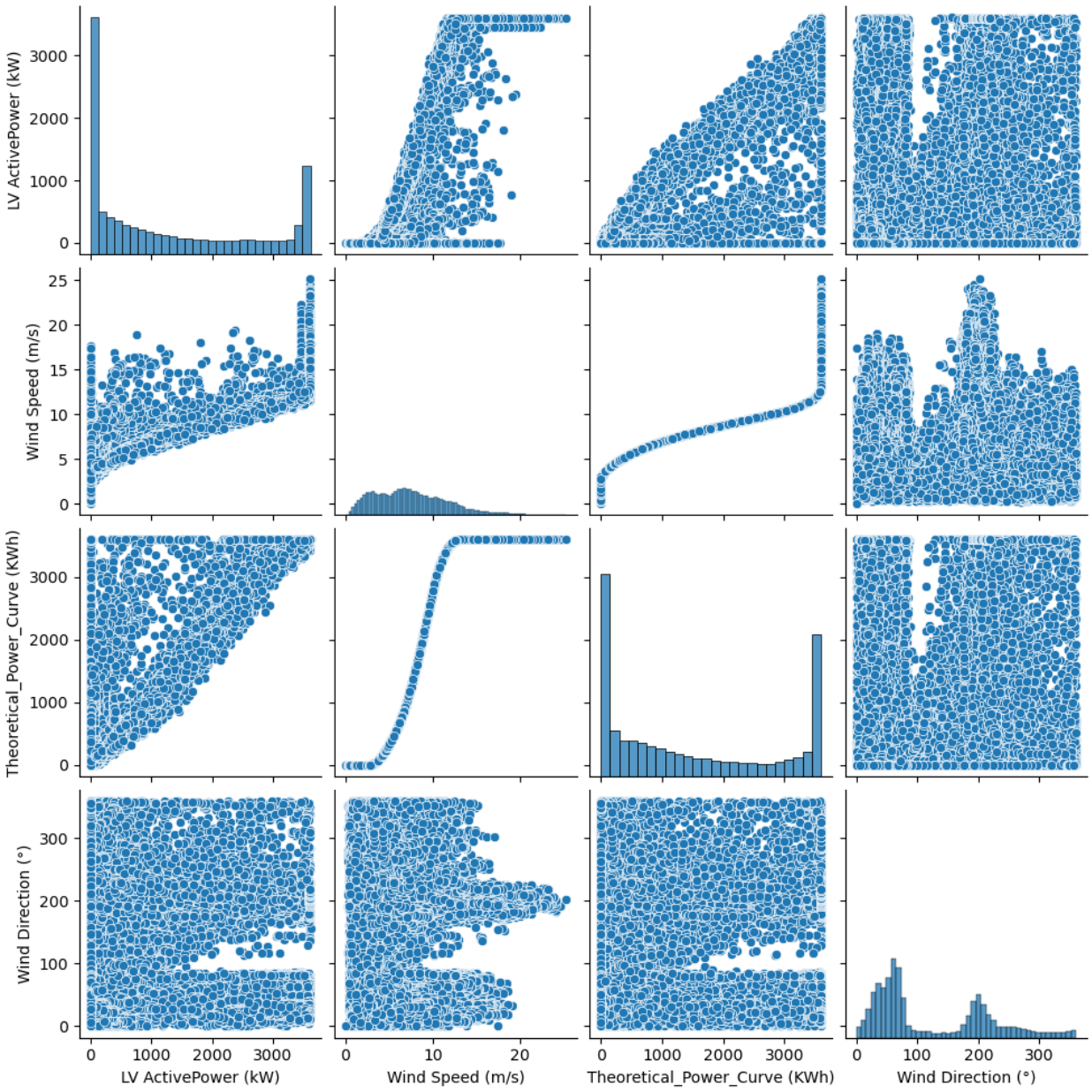

LV ActivePower (kW)
Wind Speed (m/s)
Theoretical_Power_Curve (KWh)
Wind Direction (°)
0
1000
2000
3000
0
5
10
15
20
25
0
100
200
300

Figure 1

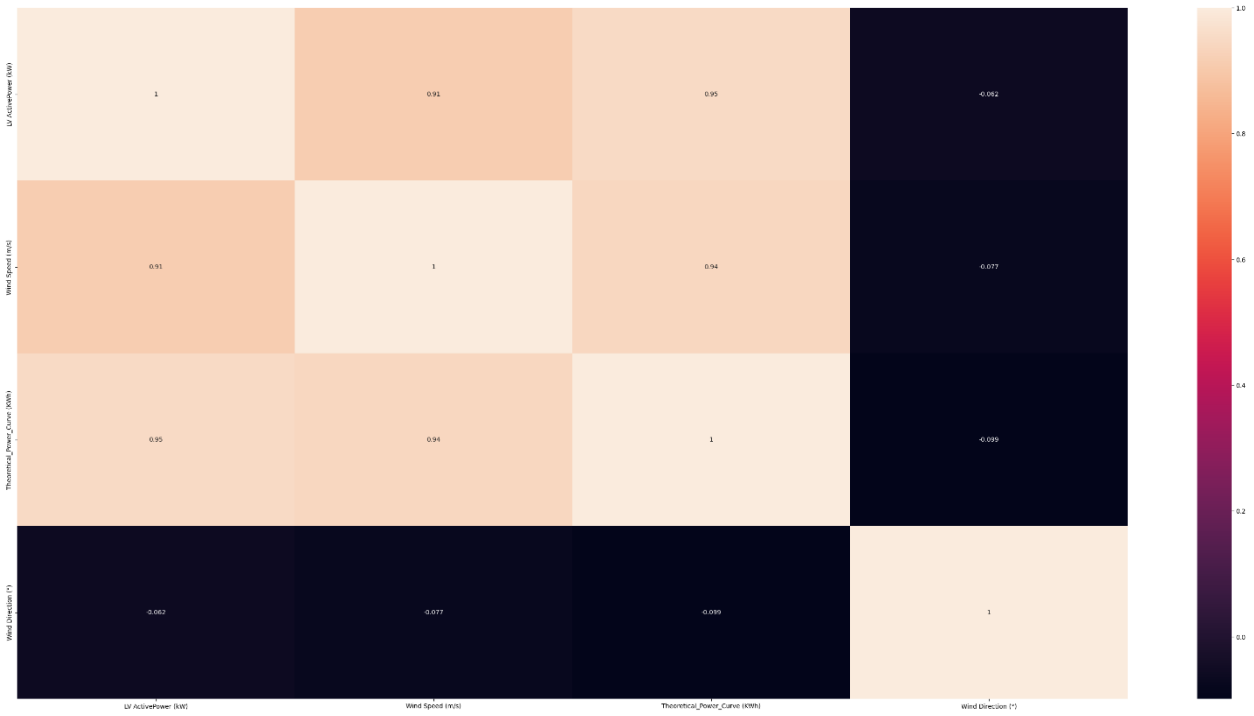


Figure 2

There is a relatively strong correlation between the Wind Speed (m/s) and LV ActivePower (kW) (Figure 3). Variables respectively. The values is approximately 0.91. if this value is close to 1, it means it is very strong correlation.

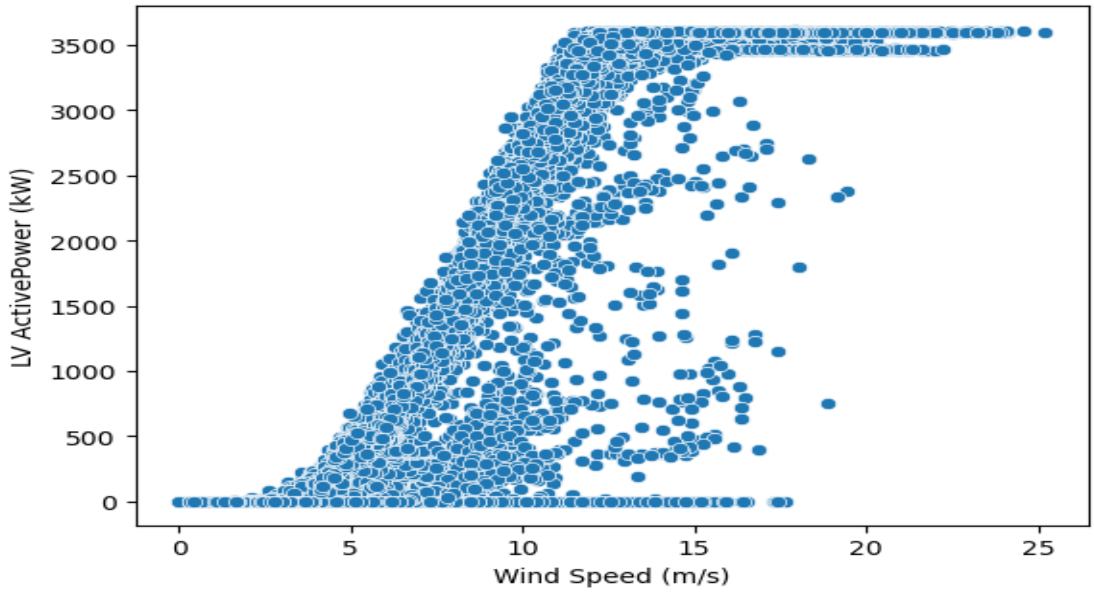


Figure 3

Pre-processing and Training Data Development

In this phase, derived features such as month, day of the week, and year were extracted from the dataset, as the date/time column will not be utilized for analysis. During this phase, we converted the date/time column into a more suitable format for analysis, as the original format is typically not readily usable. This transformation was achieved using the "get_dummies" function from the Pandas library, enabling us to utilize the data effectively.

The distributions and relationships between variables were visualized by creating graphs. Furthermore, correlation between variables was explored using scatter plots and correlation matrices.

To ensure uniformity across features, the StandardScaler was employed to standardize the dataset.

Following standardization, the data were split into training and testing subsets, with the target variable designated as "LV ActivePower (kW)". This partitioning scheme was chosen to facilitate the prediction of "LV ActivePower (kW)" values.

Model Selection

In this report, we present the process of selecting and evaluating regression models for predicting "LV ActivePower (kW)" based on the provided dataset. We explore various regression techniques including Linear Regression, Random Forests, Support Vector Regression, Ridge Regression, and Lasso Regression. The goal is to identify the most suitable model for predicting the target variable.

Linear Regression:

The Linear Regression model achieved an accuracy score of approximately 0.909 on the test data.

Cross-validation showed consistent performance with a mean score of 0.909.

Ordinary Least Squares (OLS):

OLS regression resulted in an R-squared value of 0.907 on the training data, indicating a good fit.

However, it's noteworthy that the design matrix might have multicollinearity issues or be singular.

Random Forests:

Random Forest Regression achieved the highest accuracy score of approximately 0.957 on the test data.

Cross-validation yielded a mean score of 0.943, indicating robust performance.

Support Vector Regression (SVR):

SVR achieved an accuracy score of approximately 0.934 on the test data.

Cross-validation resulted in a mean score of 0.926.

Ridge Regression:

Ridge Regression produced accuracy scores ranging from 0.908 to 0.909 for different alpha values.

The best score was 0.909 achieved with an alpha value of 0.0001.

Lasso Regression:

Lasso Regression performed comparatively lower with a maximum accuracy score of approximately 0.895.

Model Comparison:

Random Forests outperformed other models with the highest accuracy score of 0.957, followed by SVR, Linear Regression, Ridge Regression, and Lasso Regression.

The model accuracy scores were visualized, indicating the superiority of Random Forests.(Figure 4)

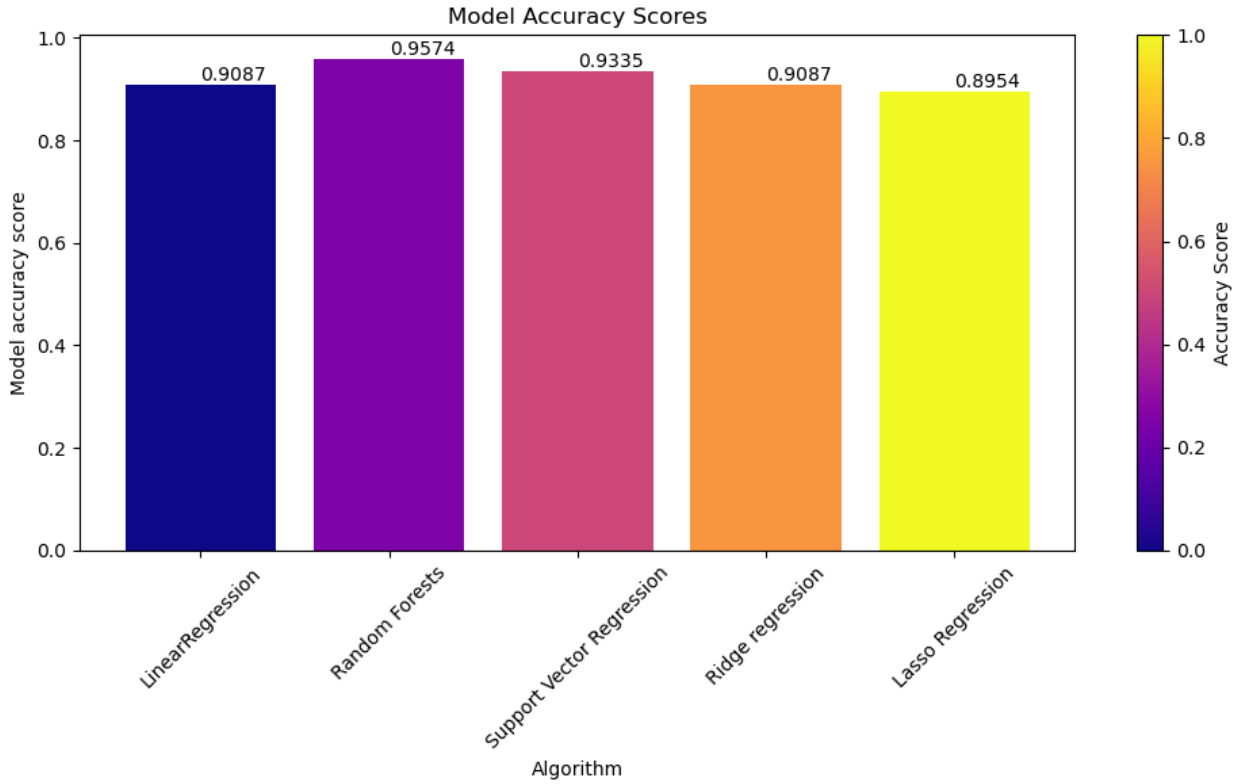


Figure 4

Hyperparameter Tuning:

GridSearchCV and RandomizedSearchCV were employed for hyperparameter tuning of Ridge Regression.

The optimal alpha value obtained was 0.0001 with both approaches, resulting in a test score of approximately 0.909.

Conclusion:

Based on the evaluation results, Random Forests proved to be the best model for predicting "LV ActivePower (kW)".

The performance of the models was consistent with the expectations, with Random Forests demonstrating superior accuracy.

Future Directions:

Further improvement can be achieved by incorporating data from additional years and exploring advanced feature engineering techniques.

Continuous monitoring and fine-tuning of the models can enhance their predictive capabilities over time.

Incorporate external data sources such as weather forecasts, geographical data, or historical energy production data to enrich the feature set and improve model performance.

Recommendations:

Use the predictive models, particularly the Random Forests model, to forecast future energy production with high accuracy.

Utilize the insights gained from the models to optimize operational strategies, such as scheduling maintenance activities during periods of predicted low energy production or adjusting turbine settings based on anticipated wind conditions.

Use the model results to inform resource allocation decisions, such as determining optimal locations for new wind farms based on historical wind patterns and expected energy generation potential.