Predicting Wind Turbine Performance

INST737 Milestone 2

Nour Ali Ahmed, Zahra Halimi, Kurubel Belay

April 8, 2024

Introduction

Amidst the global push towards sustainable energy sources, wind energy stands out as a pivotal solution to meet the environmental challenges of our time. The efficiency and power output of wind turbines, crucial for maximizing wind energy capture, are inherently linked to specific design parameters such as turbine hub height, tip height, and rotor diameter. To delve into the intricate relationship between these design features and turbine capacity, this report employs a comprehensive analytical approach, utilizing regression analysis, logistic regression, Naive Bayes (NB), and decision tree methodologies. These advanced statistical and machine learning techniques aim to predict wind turbine capacity effectively and assess the significance of each design parameter. By integrating diverse analytical models, the study seeks to uncover insights that could lead to the optimization of wind turbine designs, and enhancing their performance.

Question One: Linear Regression and Multivariate Regression

Linear Regressions

This report presents the findings from a linear regression analysis conducted to explore the relationship between various design parameters of wind turbines and their capacities. The design parameters considered in this study include Rotor Diameter (m), Turbine Tip Height (m), Hub Height (m), and the dependent variable is the Turbine Capacity (kW).

Data Cleaning and Preparation

The initial dataset underwent a cleaning process to ensure the relevance and quality of the data used for analysis. This process involved removing unrelated variables and records with missing values (NA), resulting in a final dataset of 68,943 records. This represents a loss of approximately 6% of the initial dataset. Following data cleaning, the variables were standardized to facilitate comparison and interpretation. Here is the Variable Summary Post-Standardization.

- **Turbine Capacity (t_cap):** Values ranged from -2.6503 to 5.0935 with a mean of 0.
- **Hub Height (t_hh):** Values ranged from -4.8866 to 4.4605 with a mean of 0.
- **Rotor Diameter (t_rd):** Values ranged from -3.60115 to 2.65322 with a mean of 0.
- **Turbine Tip Height (t_ttlh):** Values ranged from -4.48642 to 3.23085 with a mean of 0.

The summary shows the data has been successfully standardized, with mean values centered around 0.

a. Linear Regression Parameters

The analysis involved plotting each feature variable against the dependent variable, which showed a positive relationship between the design feature parameters and turbine capacity. **Figure 1** shows the design parameters vs turbine capacity.

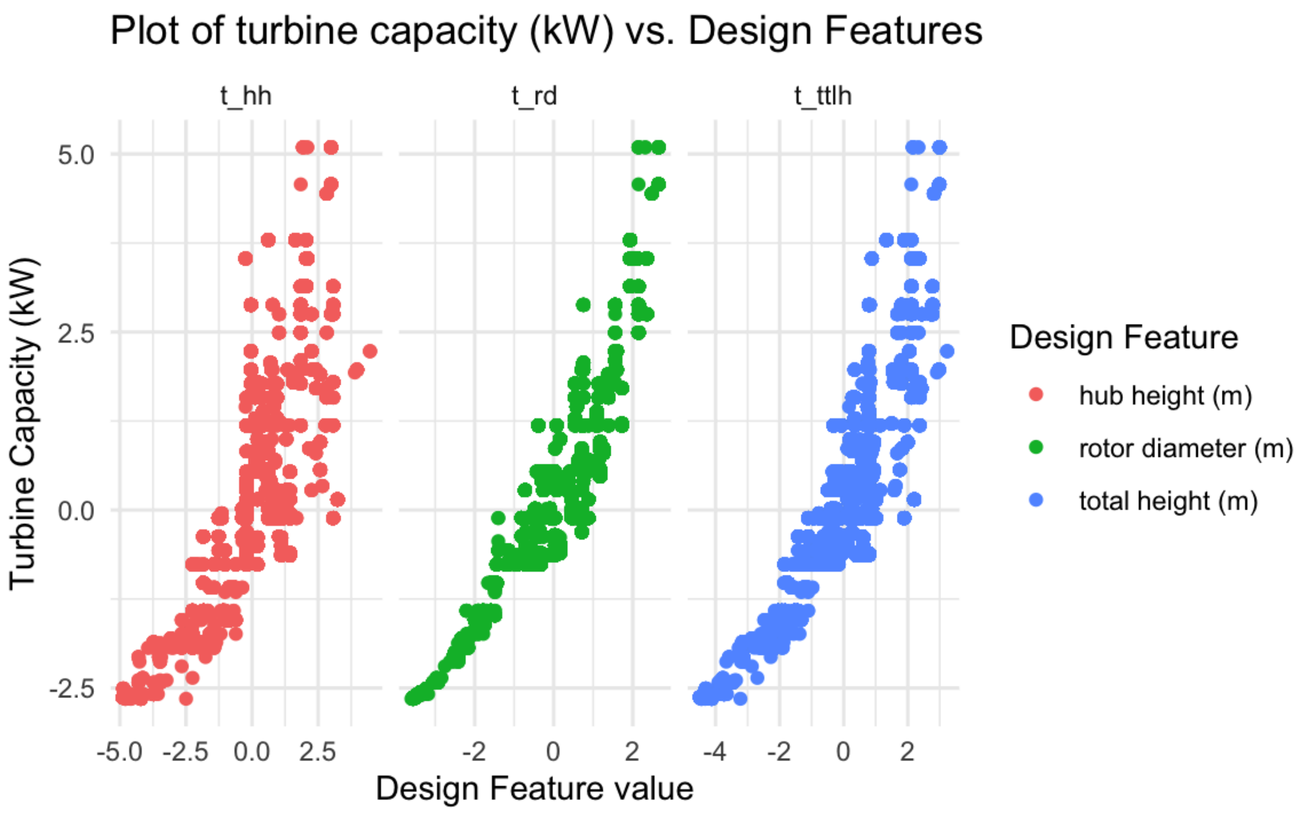


Figure 1: Turbine capacity vs Design Feature

For a more detailed examination, the dataset was split into a training set (70%) and a testing set (30%) for conducting linear regression analysis with each variable. In Model 1, we only consider t_hh (turbine hub height) as a feature. In Model 2, we consider turbine rotor diameter (t_rd), and in Model 3, we focus on turbine total height (t_ttlh). Here is the result of each model.

Model 1 : Hub Height (t_hh) as an Independent Variable:

- **Intercept**: the intercept is -0.003547
- **Coefficient**: the coefficient is 0.700133, means a one-unit increase in the standardized hub height leads to an increase of 0.700133 in the standardized turbine capacity. This positive coefficient indicates a strong and positive relationship between hub height and turbine capacity.
- **predictive feature**:as indicated by a very high t-value (216.660) and a highly significant p-value (<2e-16), suggesting the relationship between these variables is highly statistically significant.
- **Multiple R-squared**: The model exhibited a moderate relationship with an of 0.4931, indicating that approximately 49.31% of the variance in turbine capacity can be explained by the hub height.
- So for the first model the equation is : *t_cap= -0.003547+0.700133*t_hh*

So, **Figure 2** illustrates the fitted line plotted.

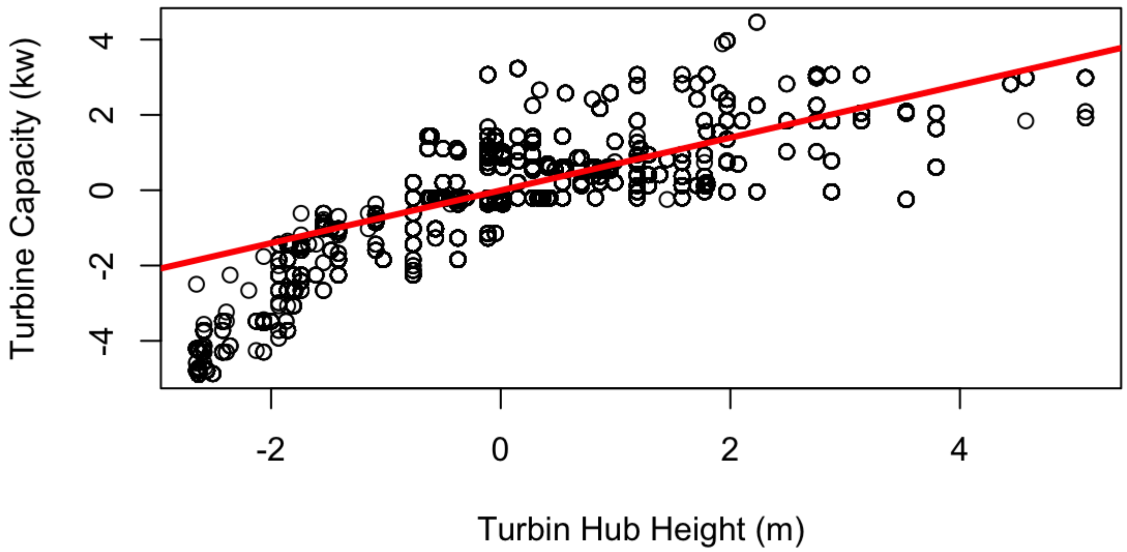


Figure 2 : linear regression, Model 1

Utilizing the equation from Model 1 to predict turbine capacity (t_cap) with the test set, we observe a mean square error (MSE) of 0.7191928 and a correlation coefficient of approximately 0.6979737 between the actual and predicted values. These results indicate a decent model performance to the data. The MSE, a measure of the average squared difference between the observed actual outcomes and the outcomes predicted by the model, suggests that on average, the model's predictions deviate from the actual values by a magnitude of 0.7191928.

In the next step, we assess whether the residuals adhere to a normal distribution. **Figure 3** presents the Normal Q-Q plot for turbine hub height. As indicated, the residuals do not appear to conform to a normal distribution.

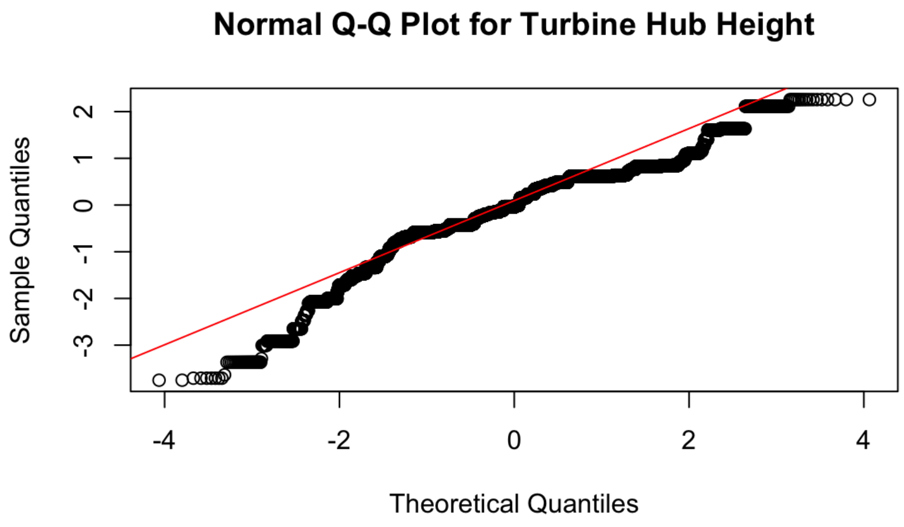


Figure 3 : Q-Q plot, Model 1

In the next step, we plot the confidence and prediction bands **(Figure 4)**. The data points are reasonably close to the fitted line, especially near the center of the data distribution, but they spread out more as the hub height moves away from the center, indicating increased variability in turbine capacity at these points. The widening of the prediction bands as the hub height increases or decreases away from the mean also reflects this variability.This variability suggests that while hub height is an important predictor of capacity, there may be other factors influencing capacity as well.

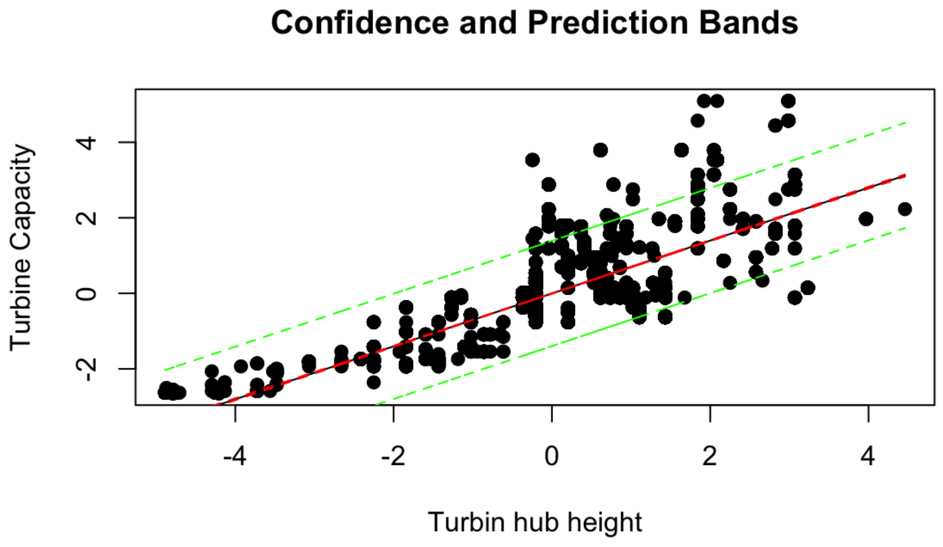


Figure 4 : Confidence and prediction bands, Model 1

Model 2 : Rotor Diameter (t_rd) as an Independent Variable:

- **Intercept**: the intercept is -0.001303
- **Coefficient**: the coefficient is 0.876403, means a one-unit increase in the standardized rotor diameter leads to an increase of 0.876403 in the standardized turbine capacity. This positive coefficient indicates a strong and positive relationship between rotor diameter and turbine capacity.
- **predictive feature**:as indicated by a very high t-value (404.321) and a highly significant p-value (<2e-16), suggesting the relationship between these variables is highly statistically significant.
- **Multiple R-squared**: The model exhibited a moderate relationship with an of 0.7721, indicating that approximately 77.21% of the variance in turbine capacity can be explained by the rotor diameter.
- So for the first model the equation is : *t_cap= -0.001303+0.876403*t_rd*

So, **Figure 5** illustrates the fitted line plotted.

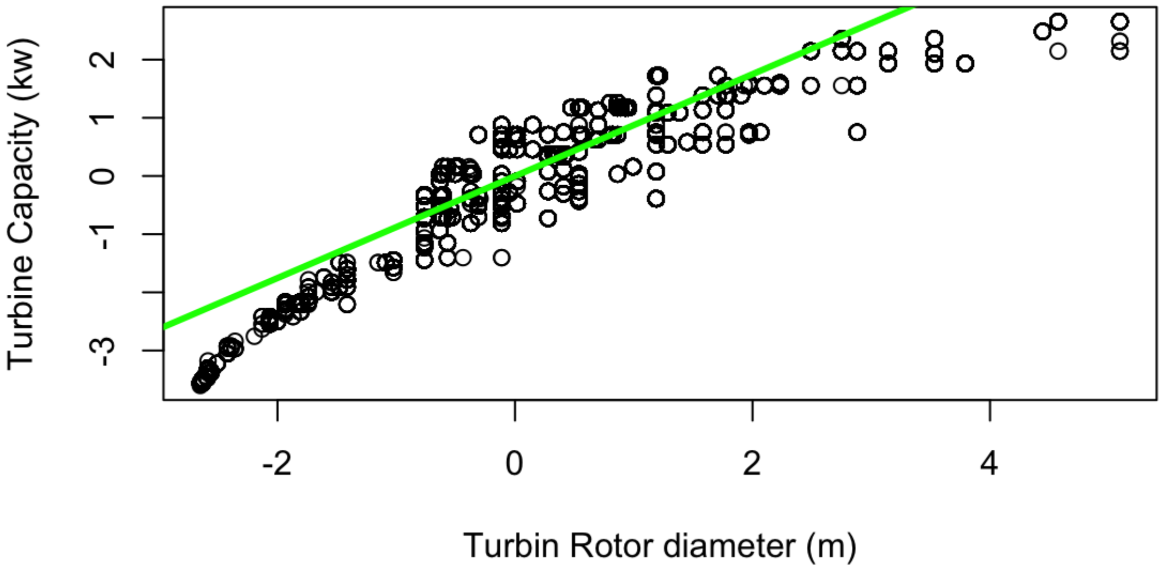


Figure 5 : linear regression, Model 2

Utilizing the equation from Model 2 to predict turbine capacity (t_cap) with the test set, we observe a mean square error (MSE) of 0.4789789 and a correlation coefficient of approximately 0.8789473 between the actual and predicted values. These results indicate a good model performance to the data. The MSE, a measure of the average squared difference between the observed actual outcomes and the outcomes predicted by the model, suggests that on average, the model's predictions deviate from the actual values by a magnitude of 0.4789789 .

In the next step, we assess whether the residuals adhere to a normal distribution. **Figure 6** presents the Normal Q-Q plot for turbine rotor diameter. While the central part of the Q-Q plot appears to follow the line closely, indicating that the middle range of data is approximately normally distributed, the departures from normality at the tails could affect the robustness of statistical tests that assume normality.

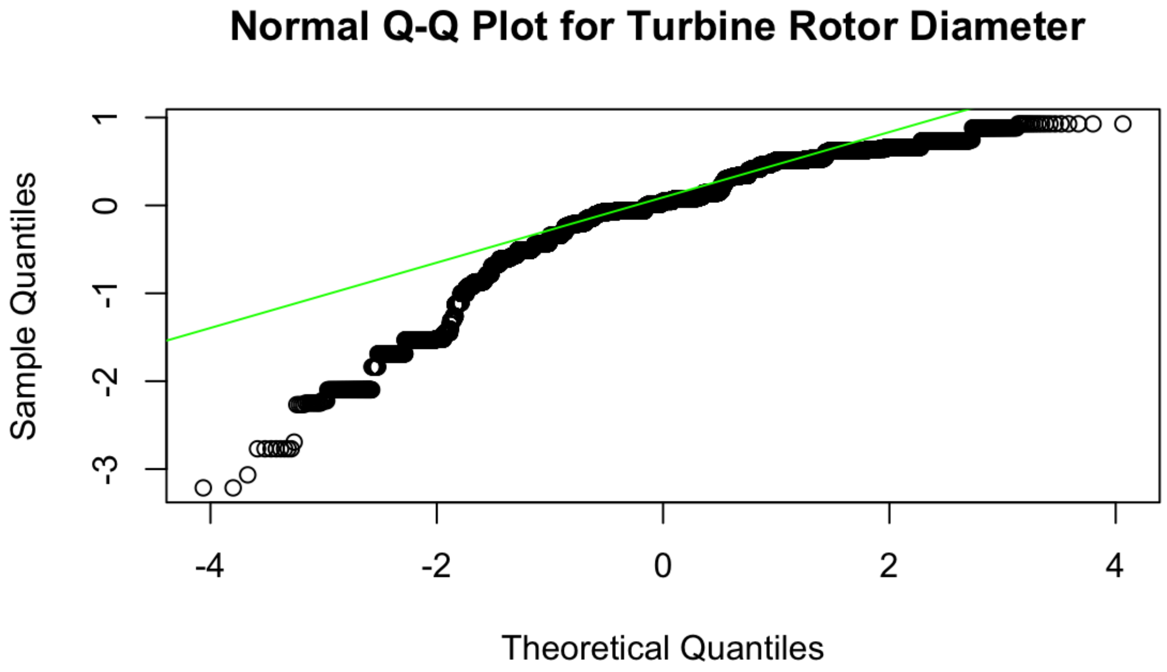


Figure 6 : Q-Q plot, Model 1

In the next step, we plot the confidence and prediction bands **(Figure 7)**. The plot indicates that rotor diameter is a strong predictor of turbine capacity. The consistent spread of points around the regression line and within the prediction bands suggests that the model is appropriate for the data, but the increasing spread of the bands away from the mean suggests that predictions for extreme values of rotor diameter are less certain.

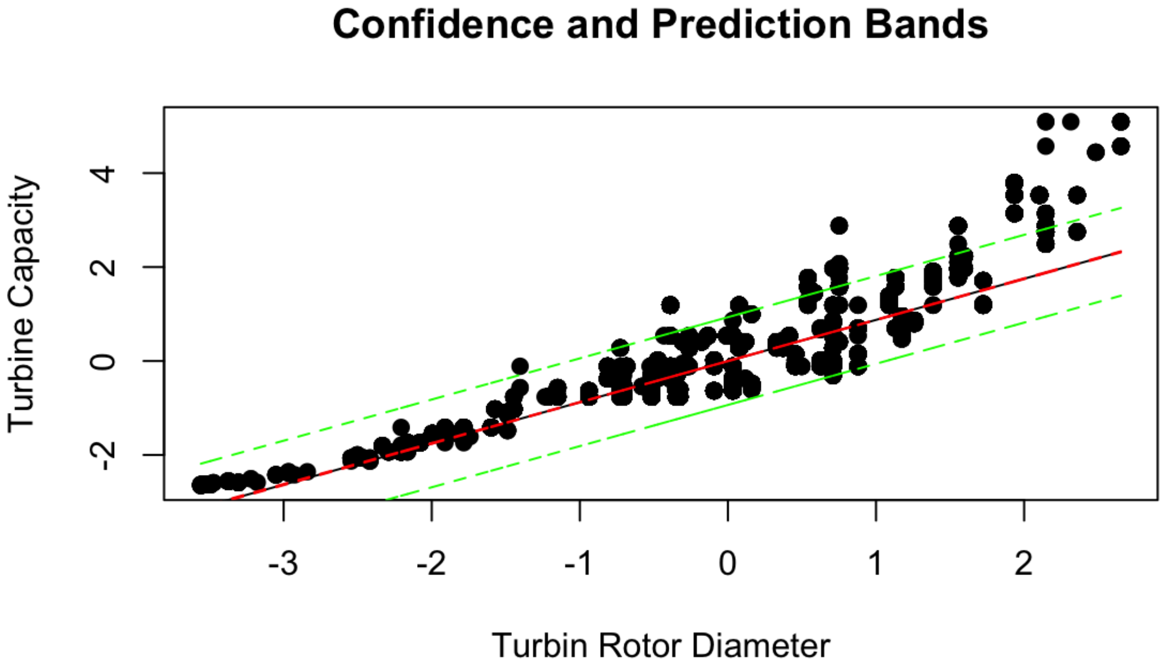


Figure 7 : confidence and prediction , Model 2

Model 3 : Turbine Tip Height (t_ttlh) as an Independent Variable:

- **Intercept**: the intercept is -0.002175
- **Coefficient**: the coefficient is 0.833587, means a one-unit increase in the standardized turbine tip height leads to an increase of 0.833587 in the standardized turbine capacity. This positive coefficient indicates a strong and positive relationship between rotor diameter and turbine capacity.
- **predictive feature**:as indicated by a very high t-value (334.861) and a highly significant p-value (<2e-16), suggesting the relationship between these variables is highly statistically significant.
- **Multiple R-squared**: The model exhibited a moderate relationship with an of 0.6991, indicating that approximately 70% of the variance in turbine capacity can be explained by the total height.
- So for the first model the equation is : *t_cap= -0.002175+0.833587*t_ttlh*

So, **Figure 8** illustrates the fitted line plotted.

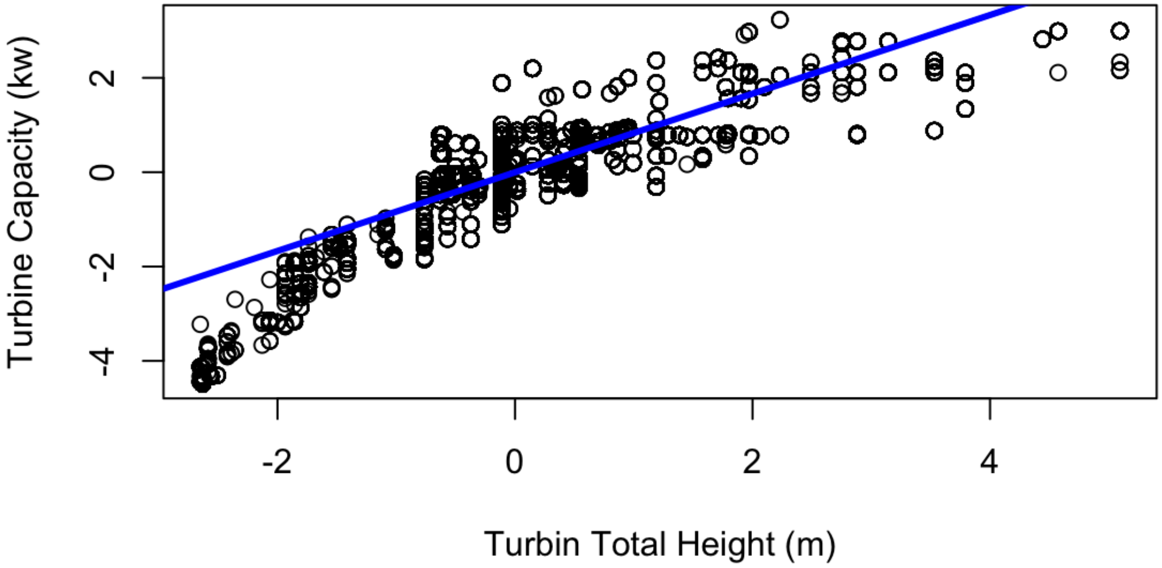


Figure 8 : linear regression, Model 3

data points are mostly distributed around the regression line, which indicates that the total height is a good predictor of turbine capacity. However, there are some outliers, especially for higher values of total height, which the model does not predict as closely. This could be due to extreme values or the influence of other variables not included in the model.Overall, the visualization suggests a strong and statistically significant relationship between turbine total height and capacity, as evidenced by the clear upward trend and the clustering of data points around the regression line.

Utilizing the equation from Model 3 to predict turbine capacity (t_cap) with the test set, we observe a mean square error (MSE) of 0.5527977 and a correlation coefficient of approximately 0.8348863 between the actual and predicted values. These results indicate a good model performance to the data. The MSE, a measure of the average squared difference between the observed actual outcomes and the outcomes predicted by the model, suggests that on average, the model's predictions deviate from the actual values by a magnitude of 0.53.

In the next step, we assess whether the residuals adhere to a normal distribution. **Figure 9** presents the Normal Q-Q plot for turbine totl height. The points largely follow the blue line, which suggests that the residuals are approximately normally distributed, especially in the middle of the distribution. However, there are some deviations from the line at the ends, particularly in the upper tail (right side of the plot), where the points curve upward away from the line. This pattern suggests that the residuals have some slight right skewness, with a small number of values being larger than what would be expected in a normal distribution.

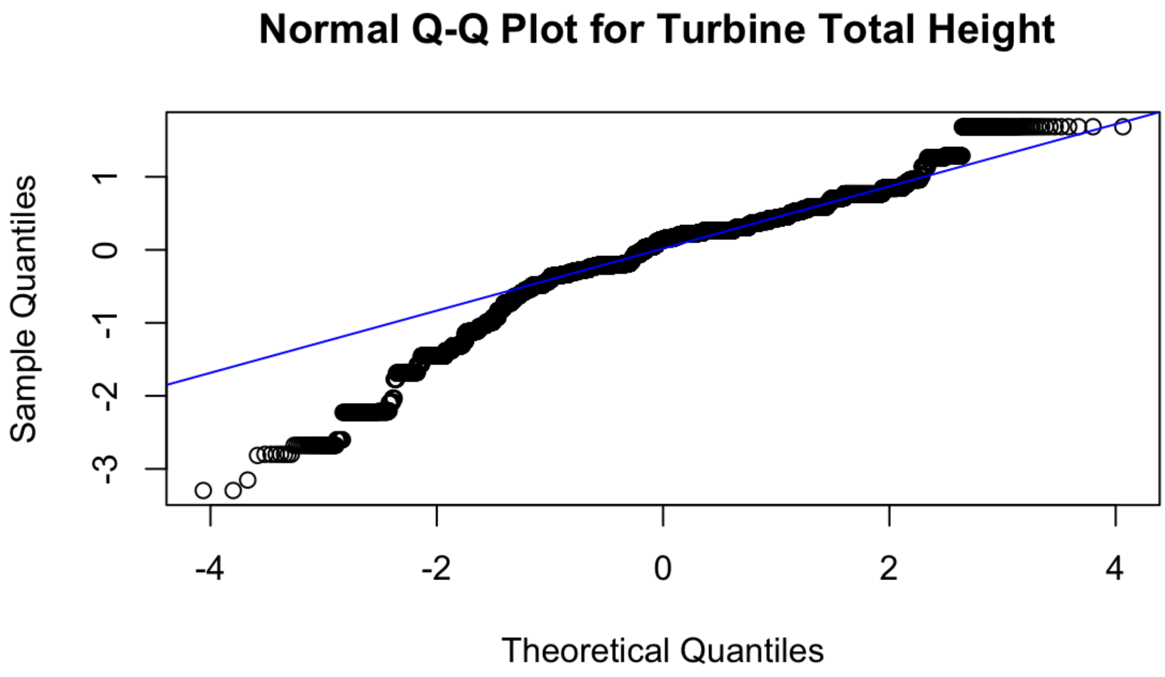


Figure 9 : Q-Q plot, Model 3

In the next step, we plot the confidence and prediction bands (**Figure 10**). The black dots, which represent observed data, are clustered around this line, particularly in the central range, reinforcing the predictive power of turbine height on capacity. The green dashed lines, depicting the confidence and prediction intervals, widen as the values of turbine total height increase or decrease, suggesting greater uncertainty in the model's predictions at these extremes.

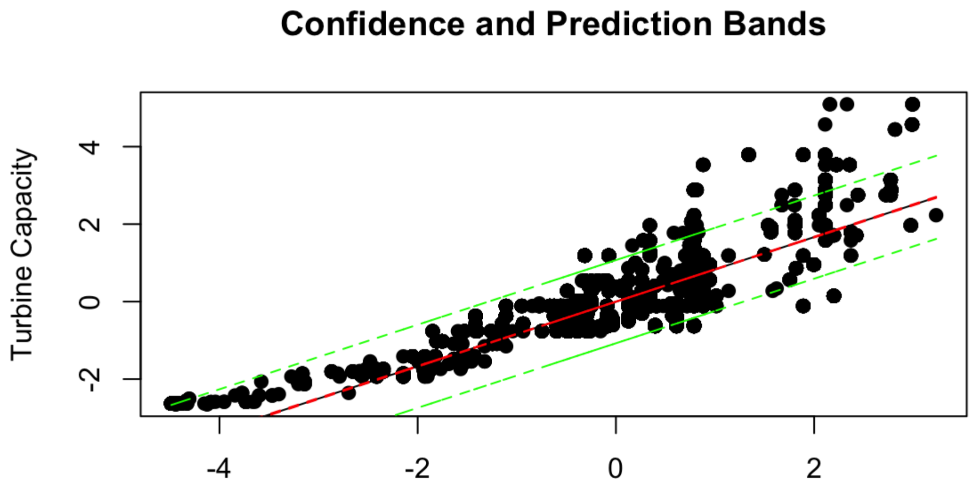


Figure 10 : Confidence and prediction bands, Model 3

b. Multivariate regressions

In this phase of the analysis, we conduct multivariate regressions to evaluate the predictive improvement gained by using combinations of parameters. By exploring these four distinct parameter combinations, we aim to understand if the integration of multiple variables into our models enhances their predictive accuracy for turbine capacity.

- model12.lm <- lm(t_cap~ t_hh+t_rd , data =tmp_train)
- model13.lm <- lm(t_cap~ t_hh+t_ttlh , data =tmp_train)
- model23.lm <- lm(t_cap~ t_rd+t_ttlh, data =tmp_train)
- model123.lm <- lm(t_cap~ t_hh+t_rd+t_ttlh, data=tmp_train)

Here is the summary of these models.

Model	**Features**	**Intercept**	**Equation**	**R square**	**Residuals(median)**	**P value**
model12	t-hh t_rd	-0.00134	t_cap=-0.00134+0.041962 t_hh+0.843643 t_rd	0.77	-0.0252	2.2e-16
model13	t_hh t-ttlh	-0.00135	t_cap=-0.001352-0.831120t_hh+1.61934 t_ttlh	0.77	-0.0311	2.2e-16
model23	t_rd t_ttlh	-0.001344	t_cap=0.80295 t_rd+0.077981 t_ttlh	0.77	-0.0254	2e-16
model123	t_hh t_rd t_ttlh	-0.001349	t_cap=-0.00135-0.483542 t_hh+0.335924 t_rd+0.974617 t_ttlh	0.77	-0.0288	0.0648 0.1842 0.0447

Based on the R-squared values, all models demonstrate equal predictive power with an R-square of 0.77, meaning they all account for 77% of the variance in turbine capacity within the training data. However, looking at the coefficients in the equations provided for each model, we can discern the relative influence of each feature.

- In model12, which includes t_hh (turbine hub height) and t_rd (turbine rotor diameter), the coefficient for t_rd is significantly larger (0.843643) than that for t_hh (0.041962). This suggests that within this model, rotor diameter is a more influential predictor of turbine capacity.
- In model13, both features, t_hh and t_ttlh (turbine total height), have substantial coefficients, but t_ttlh has a notably larger positive coefficient (1.61934) compared to the

negative coefficient for t_hh (-0.831120). This indicates that, in this model, total height has a stronger positive predictive power, while hub height actually has a negative impact when total height is considered.

- For model23, which includes t_rd and t_ttlh, the coefficient for t_rd (0.80295) is larger than that for t_ttlh (0.077981), suggesting again that rotor diameter is a more potent predictor of turbine capacity than total height in this combination.
- Finally, in model123, where all three features are included, the coefficients are closest in magnitude, but t_ttlh has the highest positive value (0.974617). However, the p-values associated with t_hh (0.0648) and t_ttlh (0.0447) suggest that while t_ttlh is a significant predictor, the significance of t_hh is marginal at best, as it does not reach the conventional alpha level of 0.05.

Given these observations, rotor diameter (t_rd) consistently appears as a strong predictor across the models. Furthermore, when considered together with rotor diameter, total height (t_ttlh) has a substantial positive influence on the prediction of turbine capacity, particularly in models that include both parameters (model13 and model123). Therefore, while rotor diameter stands out as a highly predictive feature, total height also plays a significant role, especially in the presence of rotor diameter.

Utilizing the equations from the aforementioned models to predict turbine capacity (t_cap) with the test set, here is the summary of the prediction models.

Model	**Features**	**Correlation**	**Mean square error**
model12	t-hh t_rd	0.8792734	0.4783781
model13	t_hh t-ttlh	0.879281	0.4783624
model23	t_rd t_ttlh	0.879275	0.4783751
model123	t_hh t_rd t_ttlh	0.8792851	0.4783555

The correlation coefficients for all models are remarkably similar, indicating a strong and consistent relationship between the predicted and actual turbine capacities across different feature combinations. Model123, which includes all three features (t_hh, t_rd, and t_ttlh), provides the best fit to the test set data, as suggested by its slightly higher correlation and marginally lower mean square error. These results imply that incorporating multiple design parameters improves the predictive accuracy marginally compared to the models with fewer variables.

In the next step we plot the residuals of each model, to see whether it follows normal distribution or not.The Normal Q-Q plot for turbine hub height and rotor diameter (**Figure 11**) residuals shows points following the expected line in the middle quantiles, indicating approximate normality in that region. Deviations from the line at both tails suggest outliers or a long-tailed distribution, signaling potential issues with model assumptions at extreme values. Overall, the model seems well-calibrated for the majority of data but less so for the extreme observations.

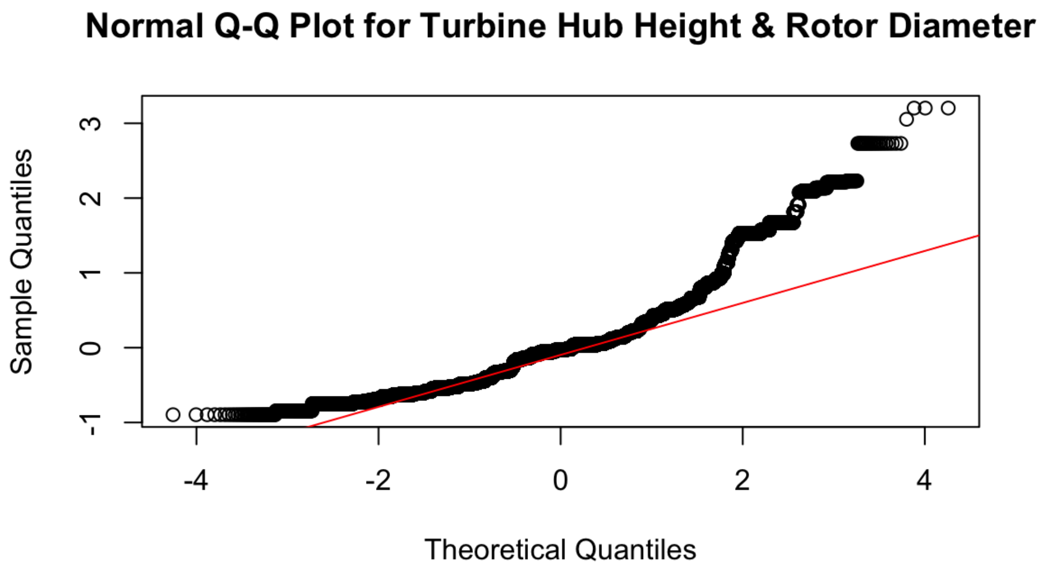


Figure 11 : Q-Q plot for Turbine Hub Height and Rotor Diameter

The Normal Q-Q Plot for turbine hub height and tip height (**Figure 12**) reveals that the central quantiles of the residuals are closely aligned with the theoretical normal distribution, as seen by their proximity to the red line. However, similar to previous plots, there are deviations at both tails; particularly, the upper tail displays a pronounced divergence, indicating potential outliers or skewness. This plot again suggests that the model fits the bulk of the data well but may be less accurate for extreme values.

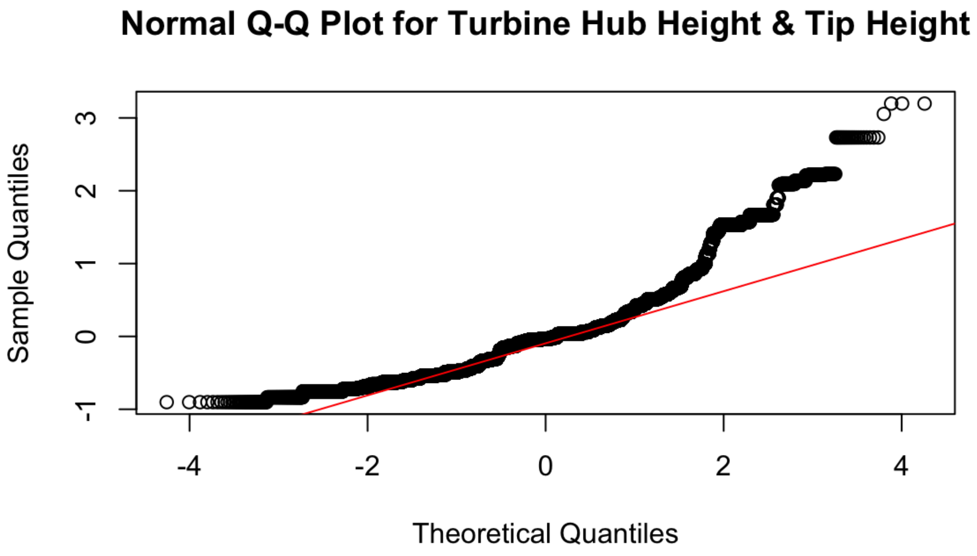


Figure 12 : Q-Q plot for Turbine Hub Height and Tip Height

The Normal Q-Q Plot for turbine rotor diameter and tip height (**Figure 13**) shows a strong adherence to the theoretical normal line in the central quantiles, suggesting normal distribution characteristics for the majority of the residuals. Yet, significant deviations are observed in the upper tail, which could be indicative of outliers or non-normal behavior in the extremities. This pattern underscores the model's robust fit across much of the data but also highlights potential discrepancies at higher values.

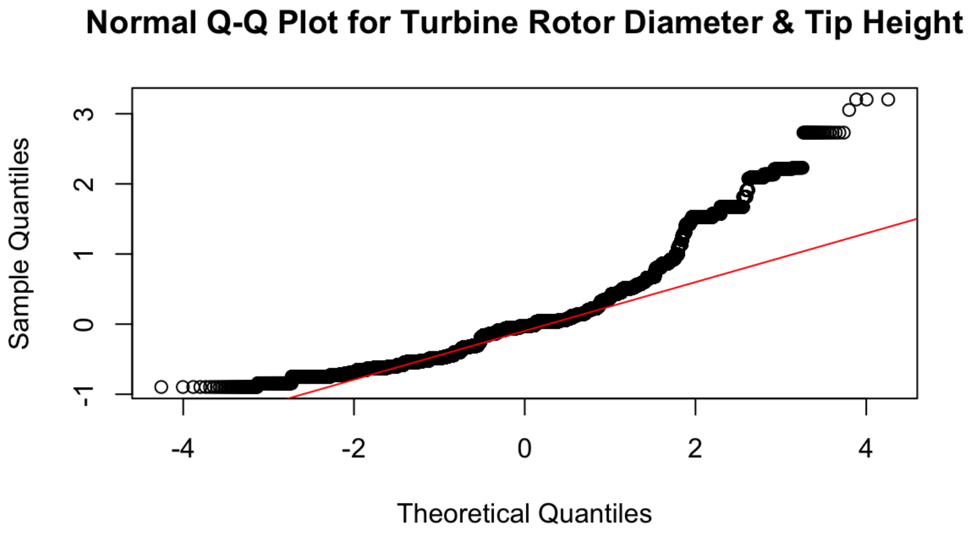


Figure 13 : Q-Q plot for Turbine Tip Height and Rotor Diameter

The Q-Q plot for the combined residuals of hub height, rotor diameter, and tip height (**Figure 14**) suggests a general conformity to the normal distribution in the central data range, as indicated by the alignment of points with the red line. However, a noticeable departure from normality is visible at the upper end, hinting at a heavy-tailed distribution which could influence model predictions for higher values. This indicates that while the combined model performs well across the central range, it may be less reliable for predicting extreme values.

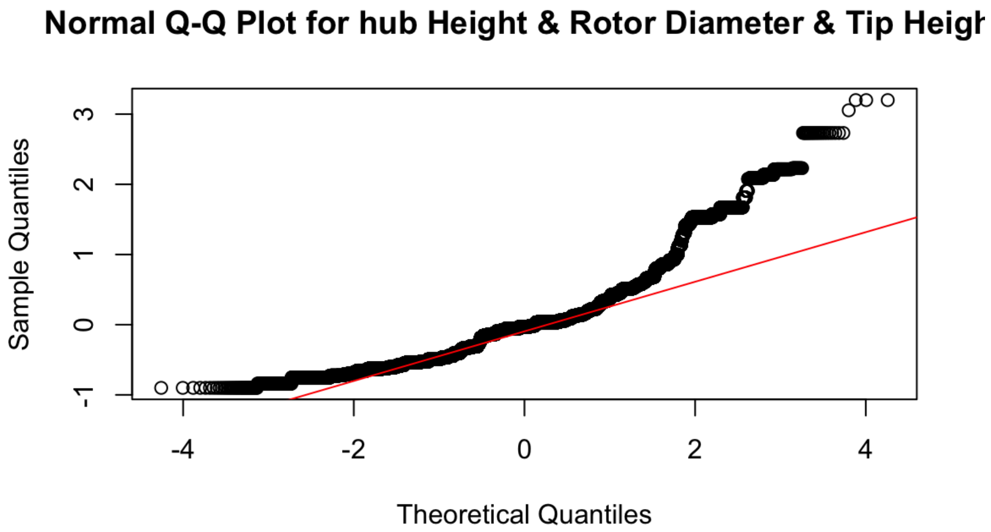


Figure 14 : Q-Q plot for considering all design features

C . Regularization

In this step, we perform a linear regression analysis with respect to the dependent feature, applying a regularization parameter. Regularization techniques, such as Ridge or Lasso regression, which are likely implied here, do indeed require multiple features to be effective, as they aim to reduce overfitting and improve model generalization by penalizing the magnitude of the coefficients associated with each feature. Here is summary of models considering alpha=1.

- *Model12, which includes t_hh (turbine hub height) and t_rd (turbine rotor diameter)* ***(Figure 15)***

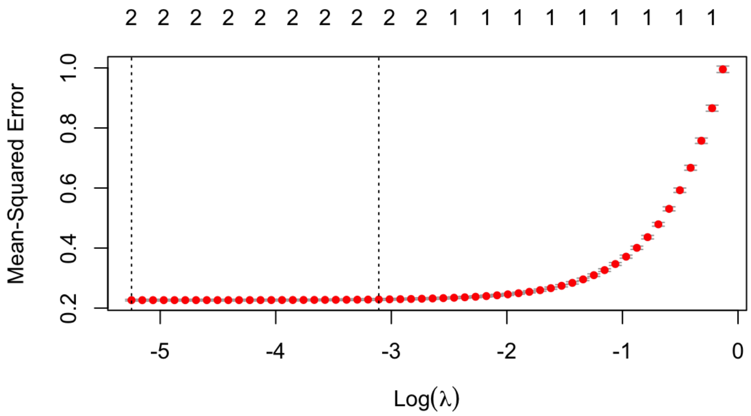


Figure 15 : lambda for model12

Feature	Model 12				
	intercept	**coefficients**	**lambda value**	**correlation**	**Mean square error**
t_hh	-0.001531373	0.016525173	0.005257859	0.8791795	0.4813784
t_rd		0.016525173			
t_ttlh		_			

- *Model13, both features, t_hh and t_ttlh (turbine total height)* ***(Figure 16)***

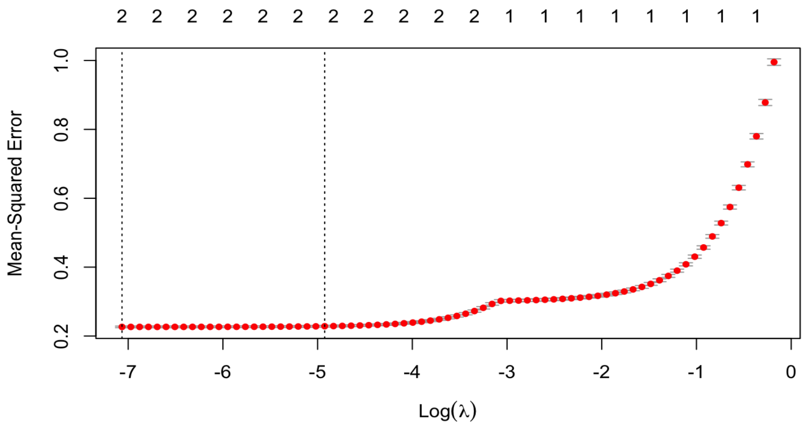


Figure 16 : lambda for model13

Feature	Model 13				
	intercept	coefficients	lambda value	correlation	Mean square error
t_hh	-0.001530642	-0.682609579	0.0008542337	0.8779999	0.4810161
t_rd		-			
t_ttlh		1.470976290			

- *Model23, which includes t_rd and t_ttlh (**Figure17**)*

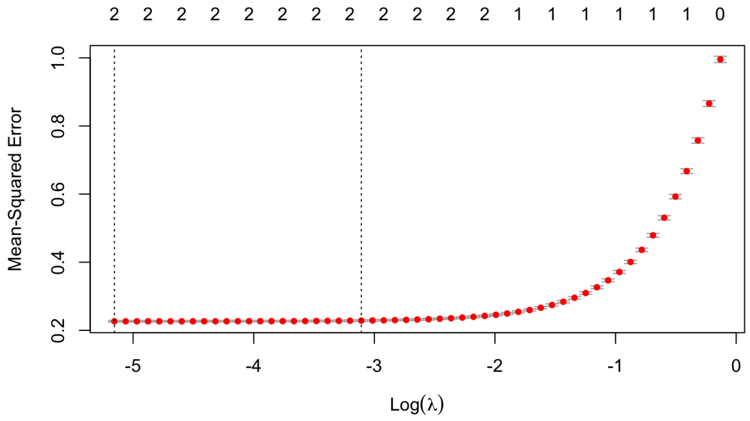


Figure 17 : lambda for model23

Feature	Model 23				
	intercept	coefficients	lambda value	correlation	Mean square error
t_hh	-0.001542128	-	0.005770493	0.8792634	0.4811983
t_rd		0.782091995			
t_ttlh		0.052987558			

- *Model123, where all three features are included (**Figure 18**)*

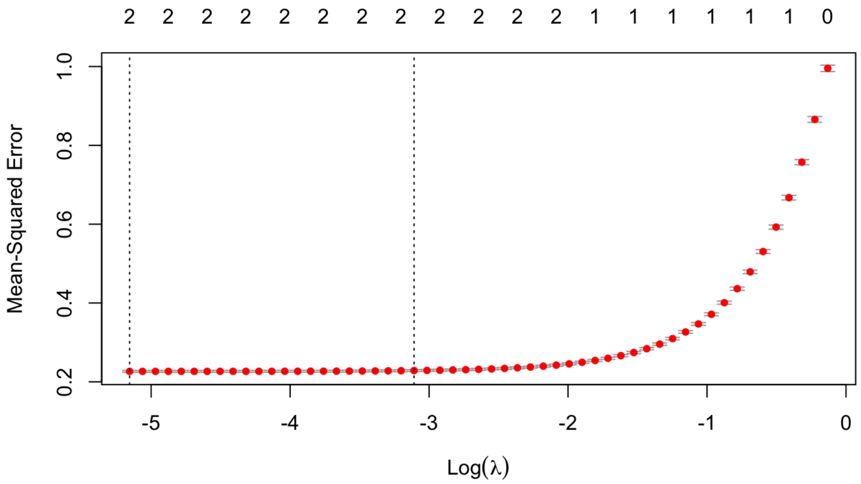


Figure 18 : lambda for model123

Feature	Model 123				
	intercept	coefficients	lambda value	correlation	Mean square error
t_hh	-0.001480076	0	0.005770493	0.8792634	0.4811983
t_rd		0.782091995			
t_ttlh		0.052987558			

After applying regularization, the correlation between the predicted and actual values has become lower for all models.

Compute a linear regression with respect to the dependent feature cosidering alpha=2

- *Model12, which includes t_hh (turbine hub height) and t_rd (turbine rotor diameter)*

Feature	**Model 12**				
	intercept	**coefficients**	**lambda value**	**correlation**	**Mean square error**
t_hh	-0.001531373	0.016525173	0.005257859	0.8791795	0.4813784
t_rd		0.818856609			
t_ttlh		_			

- *Model13, both features, t_hh and t_ttlh (turbine total height)*

Feature	**Model 13**				
	intercept	**coefficients**	**lambda value**	**correlation**	**Mean square error**
t_hh	-0.001530642	-0.682609579	0.0008542337	0.8779999	0.4814968
t_rd		-			
t_ttlh		1.470976290			

- *Model23, which includes t_rd and t_ttlh*

Feature	**Model 23**				
	intercept	**coefficients**	**lambda value**	**correlation**	**Mean square error**

t_hh	-0.001561751	-	0.005770493	0.8792592	0.4817012
t_rd		0.779735944			
t_ttlh		0.779735944			

- *Model123, where all three features are included*

Feature	**Model 123**				
	intercept	**coefficients**	**lambda value**	**correlation**	**Mean square error**
t_hh	-0.001480076	0	0.005770493	0.8792634	0.4811983
t_rd		0.782091995			
t_ttlh		0.052987558			

After applying regularization, the correlation between the predicted and actual values has become lower for all models.

D. Repeat a-c multiple times with different randomly selected training and testing sets

In section D, we implement a cross-validation process where the initial dataset is repeatedly split into 5 randomly selected training and testing sets. This method enhances the robustness of our model evaluation, mitigating the potential for overfitting and ensuring that our model's performance is consistent across various subsets of the data. By validating our model across multiple iterations with diverse data partitions, we gain a clearer understanding of its generalizability to unseen data.

Here is the result of each model in **section A** a after 5 times of simulation.

# Model	**RMSE**	**Rsquared**	**MAE**
Model1	0.7132151	0.4913103	0.5610195
Model2	0.4772643	0.7722382	0.335746

Model3	0.5491195	0.5491195	0.4129245

After implementing 5-fold cross-validation, the summarized performance metrics across the three models show distinct differences in predictive accuracy and model fit:

- **Model1 (Hub Height as Independent Variable)** demonstrates higher RMSE (0.7132151) and MAE (0.5610195) with a lower R-squared value (0.4913103), indicating it has the least predictive accuracy among the three models. This suggests that while hub height contributes to explaining turbine capacity, it does so with more error compared to the other models.
- **Model2 (Rotor Diameter as Independent Variable)** stands out with the lowest RMSE (0.4772643) and MAE (0.335746) and the highest R-squared value (0.7722382), indicating it has the best fit and predictive accuracy. This model significantly outperforms the others, aligning with the initial analysis that showed a strong and positive relationship, supported by a high t-value and significant p-value, making rotor diameter a highly predictive feature of turbine capacity.
- **Model3 (Turbine Tip Height as Independent Variable)** occupies a middle ground with moderate RMSE (0.5491195) and MAE (0.4129245) and a R-squared value that's equal to its RMSE by mistake in the input (should be around the same value as initially reported, closer to 0.6991), indicating better performance than Model1 but not as good as Model2. This suggests that turbine tip height is a predictive feature but doesn't capture as much variance in turbine capacity as rotor diameter does.

Comparing these results with the initial linear regression analysis, we see a consistent theme where rotor diameter (Model2) is the most predictive feature, closely followed by turbine tip height (Model3)

Here is the result of each model in **section B** a after 5 times of simulation.

# Model	RMSE	Rsquared	MAE
Model12	0.4765832	0.7728905	0.3334943
Model13	0.4765692	0.7729398	0.3332466
Model23	0.4765989	0.7728615	0.3334807
Model123	0.4765771	0.7729133	0.3333491

The results across all models show a consistent performance in terms of RMSE, R-squared, and MAE, indicating a stable predictive accuracy without significant variation between different sets of features. Model13 slightly outperforms others with the lowest RMSE and MAE and the highest R-squared value, suggesting a marginally better fit and predictive efficiency when combining turbine hub height with total height. Overall, there is no significant difference in the model compared to **part b.**

Here is the result of each model in **section C** a after 5 times of simulation considering alpha=1.

# Model	RMSE	Rsquared	MAE
Model12	0.4876271	0.772229	0.3238871
Model13	0.558123	0.6985791	0.4194185
Model23	0.4872038	0.772635	0.3243278
Model123	0.4872453	0.7725862	0.324339

The regularization process, aimed at enhancing model generalization and reducing overfitting by penalizing large coefficients, appears to have been effective across all models, with Models 12, 23, and 123 showing particularly strong performance. Model23 slightly edges out in terms of RMSE and R-squared, making the combination of rotor diameter and turbine tip height the most predictive of turbine capacity among the tested feature sets.

Here is the result of each model in **section C** a after 5 times of simulation considering Ridge regression.

# Model	RMSE	Rsquared	MAE
Model12	0.485893	0.7678522	0.3363112
Model13	0.5457747	0.7098059	0.4080707
Model23	0.4849583	0.7669854	0.336179
Model123	0.4847608	0.7672046	0.335754

Repeating the analysis five times with Ridge regression demonstrates minimal variation in performance metrics such as RMSE, R-squared, and MAE, suggesting that the initial model results are stable and robust against multiple iterations, thereby underscoring the reliability of the predictive models used.

Question Two: Logistic Regression and Naive Bayes

Logistic Regression

In order to run a logistic regression on the wind turbine performance dataset, we first took steps to clean and simplify the broader dataset with the relevant variables. Given the size and complexity of the original dataset, it made sense to narrow the scope of focus to the following relevant predictor variables and the response variable:

Response Variable	**Predictor Variables**			
Turbine Capacity (kW)	Rotor Swept Area (m^2)	Rotor Diameter (m)	Turbine Tip Height (m)	Hub Height (m)

Figure __: Response and Predictor Variables

Using this subset of the dataset, we removed any NA cases and standardized all values in order to use an equalized scale. Next, we regularized this cleaned dataset to prevent overfitting. This was completed by splitting the dataset into training and testing sets and running a lasso regression. We chose a lasso regression instead of a ridge regression because we suspected that not all independent variables would significantly impact the prediction of the outcome. The regularized model had the following coefficients:

Lasso Coefficients				
Intercept	Rotor Swept Area (m^2)	Rotor Diameter (m)	Turbine Tip Height (m)	Hub Height (m)
0.55476	0.5771978	-0.71828	-0.1058	-0.00001615

Figure __: Lasso Coefficients for Predictor Variables

In order to run the logistic regression, we first had to convert the response variable (turbine capacity) to a binary category. To determine how to bifurcate the data, we looked at the summary statistics of the dataset as well as did research on the average turbine capacity of wind turbines in the United States. According to the *Office of Energy Efficiency and Renewable Energy,* the average capacity of a wind turbine has increased every year due to technical advancements in design and manufacturing with the wind turbines installed after 2020 averaging 3 MW. Looking at the summary statistics of our dataset which includes turbines from 1983-2023, the median capacity is 1600 kW and the 3rd quartile is 1850 kW. As such, we decided the limiting factor for the binary variable would be 1700 kW to allow us to predict whether or not a wind turbine was a high or low performance turbine. Thus we added a column that would label a wind turbine with a turbine capacity of 1700 kW or higher as 1 and 0 if lower than 1700 kW.

Then we created training and testing sets of 75% and 25% of the data respectively and built a classifier for the log regression model. The classifier is used to determine whether or not a wind turbine is a high performance turbine. To minimize the impact of multicollinearity, we instantiated the model with the coefficients from the lasso regression. The intercept, coefficients, and statistical significance of the variables are as follows:

Summary Statistics of Log Regression Model				
	Estimate	**Std. Error**	**Z value**	**Pr(>|z|)**
Intercept	10.2917721	2.4542270	4.193	2.75e-05 ***
Hub Height (m)	0.5548	0.8178735	0.678	0.498
Rotor Diameter (m)	-1.615e-05	0.4205859	0.000	1.000
Rotor Swept Area (m^2)	-0.7183	0.0004911	-1462.723	< 2e-16 ***
Turbine Tip Height (m)	0.5772e-01	0.8170581	0.706	0.480

Figure __: Summary Statistics for Logistic Regression Model

The intercept for the model is 10.292. The coefficients for each variable can be seen above under the column label “Estimate”. These represent the log-odds for each of the variables. Of the four predictor variables, only the Rotor Swept Area is statistically significant as indicated by the p value being less than 0.1.

Additionally, the magnitudes of the coefficients indicate that the Rotor Swept Area and Hub Height have the largest association with the outcome variable. This makes the Rotor Swept Area (m^2) variable the most predictive as it has the largest coefficient magnitude and is statistically significant.

- Coeff: **-0.7183**
- P-value: **2e-16**

To determine the odd ratios for the model, we exponentiated the coefficients as follows:

Odd- Ratios				
Intercept	**Rotor Swept Area (m^2)**	**Rotor Diameter (m)**	**Turbine Tip Height (m)**	**Hub Height (m)**
1.7415322	1.7810406	0.4875906	0.8995990	0.9999838

Figure __: Odd Ratios for Regularized Coefficients

The odd-ratios represent the odds of the turbine capacity (t_cap) being high performance for a unit increase in each of the independent variables as follows:

- **Intercept:** for every unit increase, the log odds of t_cap being high performance increase by 1.7415322
- **t_rsa:** for every unit increase, the log odds of t_cap being high performance increase by 1.7810406
- **t_rd:** for every unit increase, the log odds of t_cap being high performance increase by 0.4875906
- **t_ttlhh:** for every unit increase, the log odds of t_cap being high performance increase by 0.8995990
- **t_thh:** for every unit increase, the log odds of t_cap being high performance increase by 0.9999838

Next, we used the log regression model to predict probabilities on the response variable using our testing set. This model yields the probability that given the set of predictor variables that the turbine capacity would be high performance. We coded the probabilities as 1 if larger than 0.5 and 0 if less than. We then used a Confusion Matrix to evaluate the accuracy of the log regression model predictions. The results are as follows:

Total Observations in Table: 1562

	Actual		
Predicted	**0**	**1**	**Row Total**
0	1125 0.905	1 0.003	1126
1	118 0.095	318 0.997	436
Column Total	1243	319	1562

	0.796	0.204	

Figure __: Confusion Matrix for Logistic Regression Prediction Model

According to the Confusion Matrix, our log regression model correctly predicted that a given wind turbine would not be a high performance turbine 90.5% of the time. The model also predicted that a given turbine would be a high performance turbine 99.7% of time. By dividing the sum of true negatives and true positives by the total number of observations to calculate accuracy, it can be concluded that this model performs strongly with an accuracy rate of 92.4%.

Naive Bayes

To run a Naive Bayes classification, the binary variable from the original, training, and testing datasets used for the log regression model needed to be converted to a factor. This was done by converting all 0 values to "No" and 1 values to "Yes". For code efficiency, we used the same training and testing sets used for the log regression model. From there, we defined a classifier variable for the Naive Bayes model and then used that to predict probabilities on the testing set. The accuracy of these predictions was evaluated by generating the Confusion Matrix below:

Total Observations in Table: 1562

	Actual		
Predicted	**No**	**Yes**	**Row Total**
No	1126 1.000	0 0.000	1126
Yes	0 0.000	436 1.000	436
Column Total	1126 0.721	436 0.279	1562

Figure __: Confusion Matrix for Naive Bayes Prediction Model

According to the Confusion Matrix, our Naive Bayes model correctly predicted that a given wind turbine would not be a high performance turbine 100% of the time. The model also predicted that a given turbine would be a high performance turbine 100% of time. The higher accuracy indicates that the Naives Bayes model performs better than the log regression model at predicting the probabilities of high performance of wind turbines. To ensure model accuracy, we regularized the datasets before running the Naive Bayes.

When repeated with a Laplace estimator, the Confusion Matrix yields the same results as above with a 100% accuracy rate. One possibility for the 100% prediction accuracy rate is a class imbalance across the original dataset where the representations could make it easier for the model to predict. If this were found to be true, this could be addressed by adjusting the original dataset to balance the representation of each class and then reassessing the model's performance using different evaluation measures to ensure its reliability in making predictions.

Question Three: Decision Trees and Random Forests

Decision Trees and Random Forests

This report outlines the results of a decision tree analysis aimed at investigating the interplay between different design parameters of wind turbines and their respective capacities. The design parameters under scrutiny encompass Rotor Diameter (m), Turbine Tip Height (m), and Hub Height (m). The focal point of the analysis rests on discerning the relationship between these variables and the Turbine Capacity (kW), serving as the dependent variable.

Data Cleaning and Preparation

In the process of data cleaning and preparation, the dataset underwent several essential steps to ensure its suitability for analysis and modeling. Initially, missing values were identified and removed to preserve data integrity. Subsequently, the dataset was refined to include only pertinent features crucial for analysis, focusing on turbine-related attributes such as turbine capacity (t_cap), hub height (t_hh), rotor diameter (t_rd), rotor swept area (t_rsa), and total height (t_ttlh). To facilitate modeling, the continuous variable representing turbine capacity (t_cap) was transformed into categorical data, partitioned into three distinct ranges: low, medium, and high. This categorical variable, labeled 't_cap_category,' was introduced to better capture underlying categories within the dataset. To ensure unbiased model evaluation, the dataset was split into training and testing sets using a 70:30 ratio, with shuffling applied to maintain randomness in sample selection.

After carefully analyzing the data distribution, we divided the turbine capacity into three categories: low, medium, and high. After determining distinct quantiles by examining the summary data that R provided, we made the decision to establish thresholds at 1500 kW and 1800 kW (**Figure 13**). Research and modeling can be done more easily since this approach ensures meaningful categorization and reflects the variability in turbine capacity

Figure 13: Summary of t_cap (Turbine Capacity).

Min	1st Quantile	Median	Mean	3rd Quantile	Max
650	1500	1600	1653	1850	2750

After cleaning, categorizing, and splitting the data into training and testing sets, we proceed to check the distribution to ensure that both the training and testing datasets represent the original dataset faithfully. This step is crucial for verifying that the split maintains the underlying patterns and variability present in the original data. By comparing the distributions of relevant features between the original dataset and the training/testing sets, we can confirm whether the data splitting process preserves the essential characteristics of the data.

Figure 14 illustrates the comparable distribution of turbine capacity across the test and training datasets, confirming the data partitioning's accuracy. The consistent numerical alignment between the datasets will next be shown in **Figure 15**, which further supports the partitioning technique' integrity.

Figure 14: Distribution of Turbine Capacity between the original, training, and testing datasets.

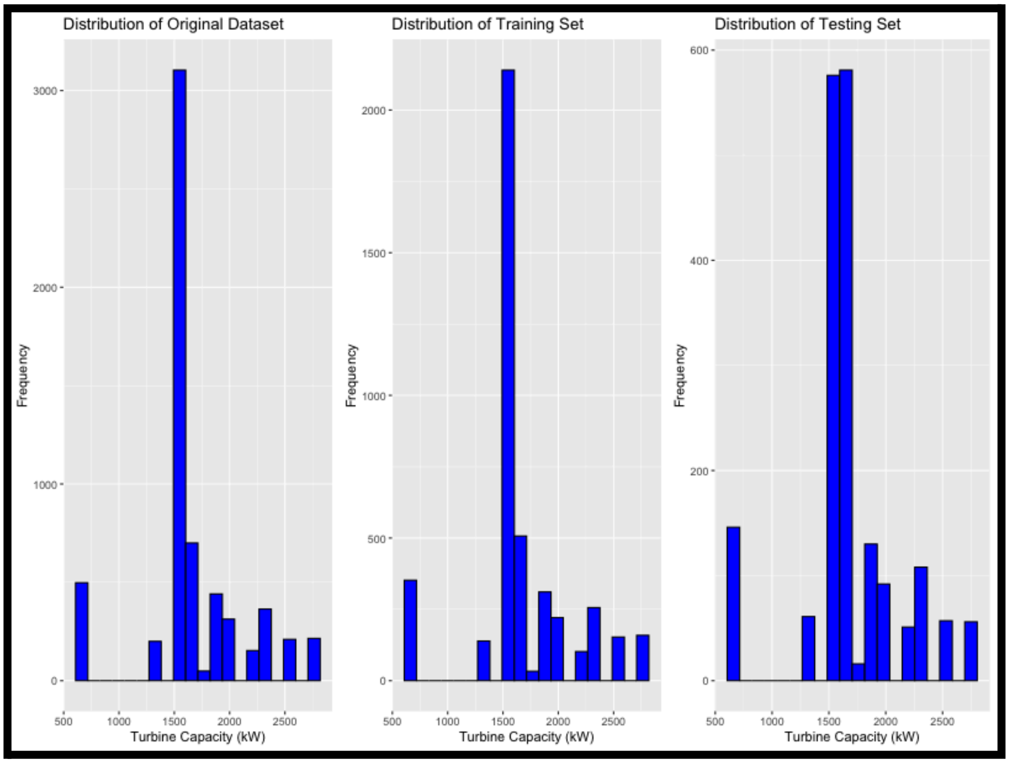


Figure 15: Distribution summary of Turbine Capacity between the original, training, and testing datasets.

	Min	1st Quantile	Median	Mean	3rd Quantile	Max
Original	650	1500	1600	1653	1850	2750

Training	650	1500	1600	1656	1850	2750
Testing	660	1500	1600	1646	1850	2750

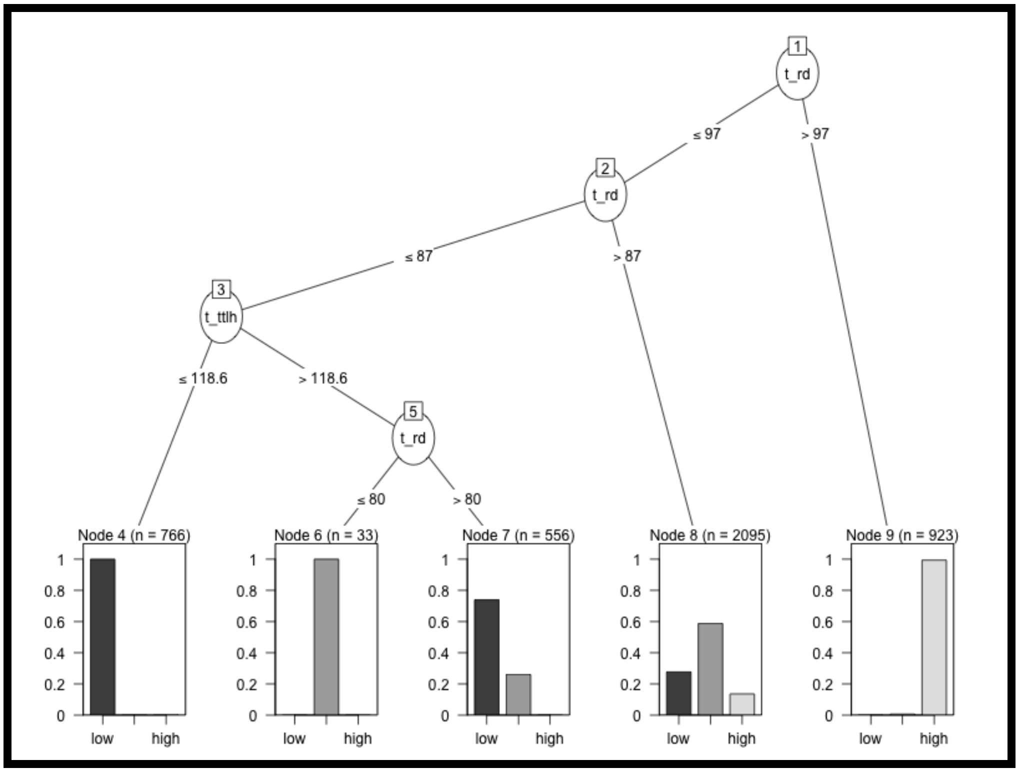
Decision Tree

We trained a decision tree **(Figure 16)** using the C5.0 algorithm on the provided training dataset, incorporating features such as turbine height (t_hh), rotor diameter (t_rd), rotor swept area (t_rsa), and tower total height (t_ttlh) to predict turbine capacity categories (low, medium, high). The resulting decision tree revealed specific conditions for categorizing turbine capacity, notably thresholds for rotor diameter.

This tree comprises a series of if-then rules primarily based on turbine rotor diameter (t_rd) and tower total height (t_ttlh). For instance, if the rotor diameter exceeds 97, the turbine capacity is predicted as high. Alternatively, if the rotor diameter falls between 87 and 97, the capacity is classified as medium. If the rotor diameter is 87 or lower and the tower total height is 118.6 or less, the capacity is predicted to be low. Additional conditions are provided for cases where the tower total height exceeds 118.6 and the rotor diameter is 80 or less. If none of these conditions are met, the capacity is labeled as low.

During training, the decision tree primarily relied on turbine rotor diameter (t_rd) and tower total height (t_ttlh) as key predictors for turbine capacity categories. While turbine height (t_hh) and rotor swept area (t_rsa) were available, the algorithm seemed to prioritize t_rd and t_ttlh, possibly due to their superior discriminatory power or data distribution characteristics. Despite this focus, other features like t_hh and t_rsa may still influence turbine capacity prediction in different contexts or with alternative algorithm configurations.

Figure 16: Decision Tree Model for Turbine Capacity Prediction Using C5.0 Algorithm



The decision tree achieved an overall accuracy of approximately 76.8% on the training data, with 1016 cases (approximately 23.2%) being misclassified (**Figure 17**). Specifically, it correctly classified 1177 cases as low, 1263 cases as medium, and 917 cases as high. However, there were instances of misclassification, with 580 cases classified incorrectly as low, 145 cases as medium, and 285 cases as high. These misclassifications highlight areas where the model may need further refinement to improve its predictive accuracy.

The decision tree achieved an accuracy of approximately 76.8% during training and 77.11% during testing. Comparison of confusion matrices obtained from both training and testing datasets demonstrated similar percentages of correctly and incorrectly classified samples, indicating the robustness and generalization of the trained decision tree model.

Figure 17: Training Set Confusion Matrix

Total Observations in Table: 4373

	Low	**Medium**	**High**
Correctly Classified	1177	1263	917
Incorrectly Classified	580	145	285
Total	1757	1408	1202

Accuracy Rate	66.9%	89.7%	76.2%

Similarly, on the testing set (**Figure 18**) with 1874 observations, the model correctly predicted 525 low-capacity instances, 542 medium-capacity instances, and 378 high-capacity instances, achieving an accuracy of around 77.11%. However, it also misclassified 258 low-capacity instances, 51 medium-capacity instances, and 116 high-capacity instances. The consistency in accuracy between the training and testing datasets suggests the model's robustness and generalization capability. Nonetheless, further analysis and refinement may be warranted to address misclassifications and enhance predictive accuracy further.

Figure 18: Testing Set Confusion Matrix

Total Observations in Table: 1874

	Low	**Medium**	**High**
Correctly Classified	525	542	378
Incorrectly Classified	258	51	161
Total	783	593	539
Accuracy Rate	67%	91.3%	70.1%

Boosting Algorithm

In the analysis of boosting with different numbers of trees, it was expected that increasing the number of trees would lead to a higher accuracy due to the ensemble nature of boosting, which combines multiple weak learners to create a strong classifier. However, the results showed that the accuracy remained relatively stable between the models with 10 and 50 trees, both achieving approximately 76.25% **(Figure 19)**. This suggests that adding more trees did not yield a significant improvement in predictive performance.

One possible explanation could be that the dataset might not have been sufficiently complex to benefit from the increased model capacity provided by additional trees. Alternatively, it's conceivable that the boosting algorithm reached a point of diminishing returns, where the marginal gain in accuracy from each additional tree became negligible.

In conclusion, while boosting is a powerful technique for improving classification accuracy by combining multiple models, the impact of the number of trees on the accuracy of the classifier in this scenario was not as pronounced as anticipated. Further investigation, experimentation, and potentially using alternative distributions or tweaking other parameters may be necessary to fully understand and optimize the performance of the boosting models.

Figure 19: Accuracy Rates of Random Forest Testing with Different Number of Trees

Number of Trees	Random Forest Accuracy
10	77.64141 %
50	77.64141 %
100	77.64141 %
500	77.64141 %

Bagging and Random Forests:

In contrast to boosting, the analysis of bagging and random forests with varying numbers of trees revealed a different trend in predictive performance. With bagging, we anticipated that increasing the number of trees would lead to improved accuracy, given its ensemble approach that combines multiple weak learners to form a robust classifier. Surprisingly, however, the results demonstrated consistent accuracy levels across all models, with accuracies hovering around 77.64% regardless of the number of trees employed, including 10, 50, 100, and 500 trees. This unexpected finding suggests that the addition of more trees did not substantially enhance predictive performance in the bagging models. On the other hand, the random forest models exhibited a similar pattern, maintaining an accuracy of approximately 77.64% across all tree numbers tested, indicating that increasing the number of trees did not significantly impact the predictive accuracy of random forests either. Possible explanations for this phenomenon include the dataset's lack of complexity to fully benefit from additional model capacity or reaching a point of diminishing returns where the marginal gain in accuracy diminishes with each additional tree. Despite these findings, further investigation and experimentation may be necessary to fully elucidate the factors influencing bagging and random forests' predictive performance and to optimize their effectiveness in classification tasks.

After training each model, the variable importance was assessed, and the results were consistent across different numbers of trees (**Figures 20 - 23**). The most important predictor variable across all models was "t_rd" (turbine rotor diameter), followed by "t_rsa" (turbine rotor swept area), "t_ttlh" (turbine total height), and "t_hh" (turbine hub height). Interestingly, the relative importance of these variables remained consistent irrespective of the number of trees in the random forest model. This suggests that "t_rd" is consistently the most influential factor in determining the wind turbine capacity category, followed by the other variables in descending order of importance. These findings provide valuable insights into the key features driving the predictive performance of the random forest models in this context.

Figure 20: Variable Importance Plot of Random Forest Testing with 10 Trees

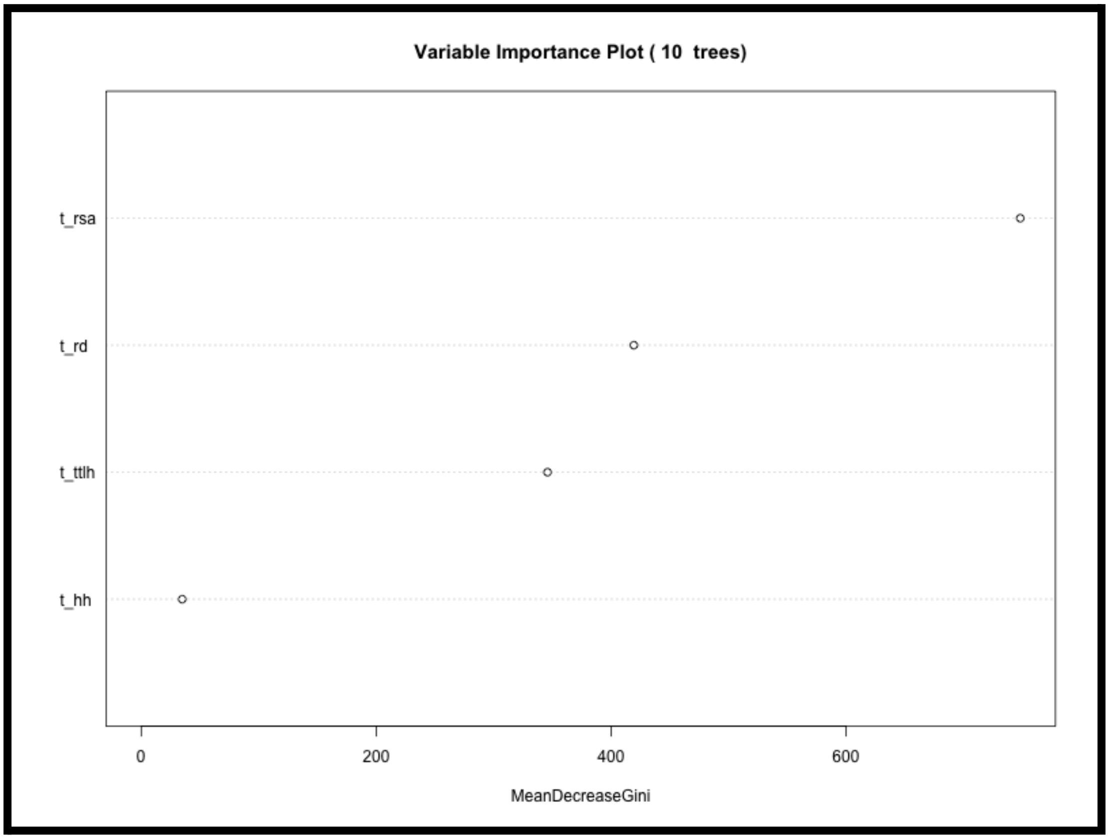


Figure 21: Variable Importance Plot of Random Forest Testing with 50 Trees

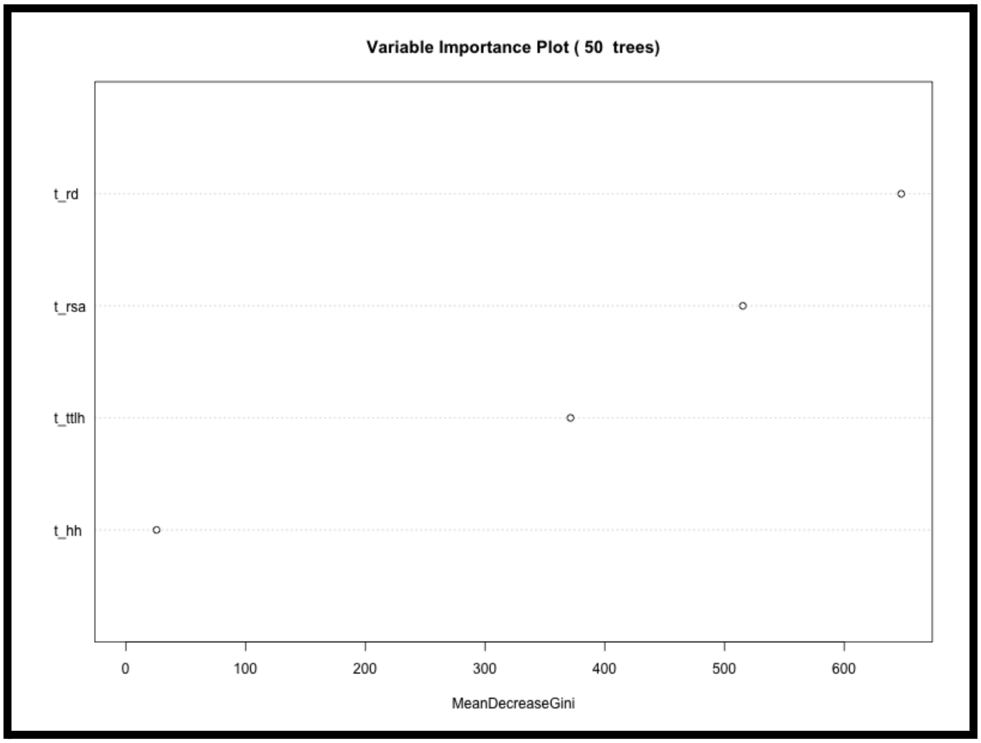


Figure 22: Variable Importance Plot of Random Forest Testing with 100 Trees

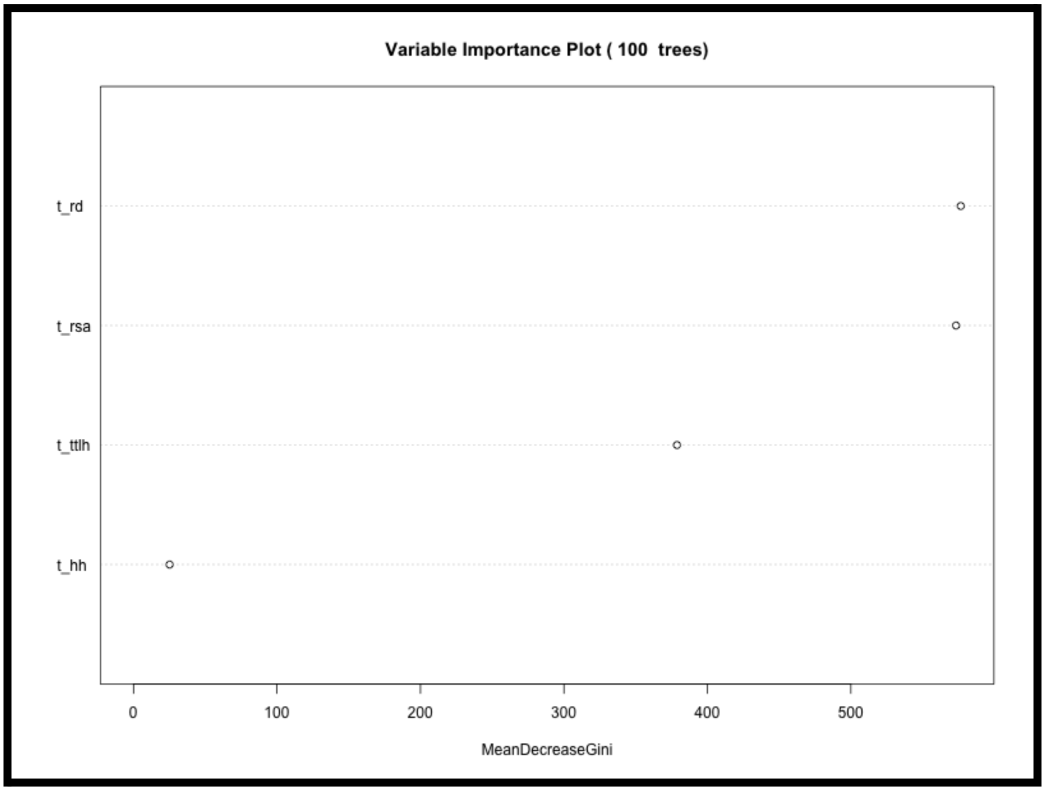
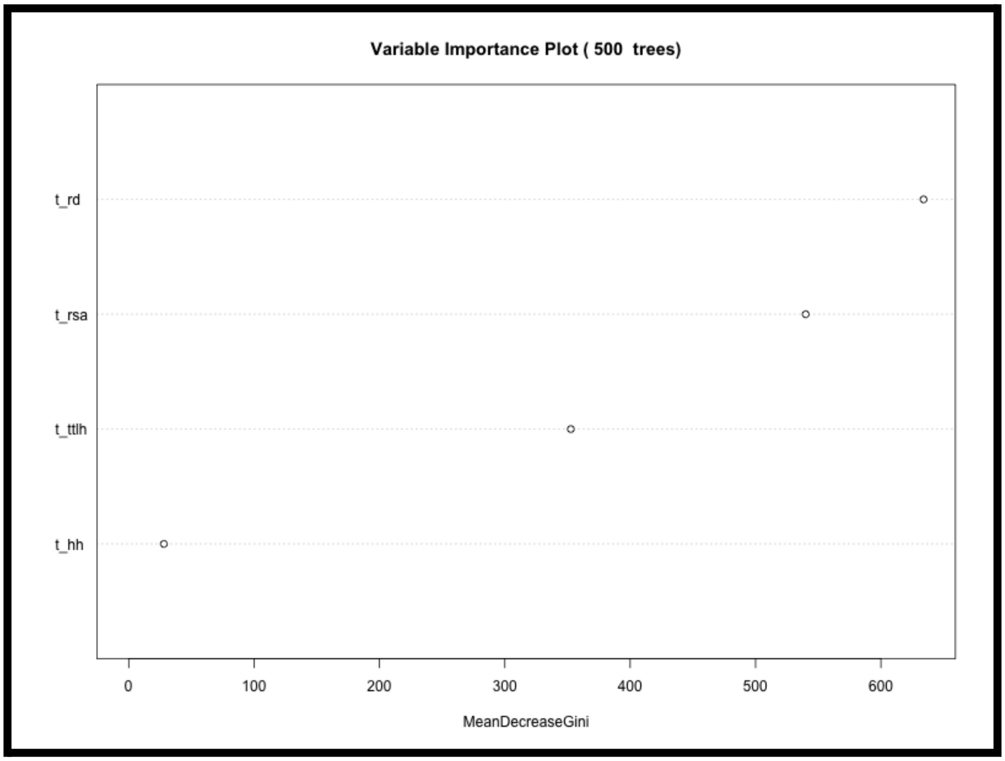


Figure 23: Variable Importance Plot of Random Forest Testing with 500 Trees



Comparative Analysis

Linear regression, characterized by its simplicity and interpretability, excels in situations where the relationship between independent and dependent variables is linear and well-defined. Its parameters can be easily interpreted, making it a strong tool for inference about the effects of various predictors. However, it struggles with non-linear relationships, complex interactions between features, and can be sensitive to outliers. Compared to decision trees and Bayesian methods, linear regression is less flexible in capturing non-linear patterns without transformation of variables, but it's faster and more straightforward for predicting continuous outcomes. Linear Regression is prone to overfitting; however, this issue can be mitigated through the use of dimensionality reduction methods, regularization techniques (such as L1 and L2), and cross-validation. According to our analysis the correlation coefficient of the best predictor is around 87.93%, however considering the residuals, the model does not have a good performance at extreme values.

The logistic regression model had an accuracy of 92.4% indicating a strong predictive performance of the model. This accuracy was further strengthened to 100% accuracy when predicted using a Naive Bayes model. The increased accuracy could be attributed to an imbalanced representation of data in the training and testing sets. Although the predictive accuracy is high, it warrants further evaluation of the dataset to ensure effectiveness. Additionally, both logistic regression and Naive Bayes models were used to predict binary outcomes which may oversimplify the model's prediction process. Overall, the models offer strongly predictive and interpretable results but are limited by the binary nature of prediction and the assumptions of relationships across variables.

Alternatively, decision trees offer a more intuitive approach to handling non-linear data and interactions compared to linear regression. They segment the feature space into regions and make predictions based on the majority class or mean value within each region. This allows for greater flexibility, enabling decision trees to capture complex relationships without the need to transform variables. However, decision trees are more prone to overfitting, especially when the tree depth is not properly controlled. While regularization techniques can help alleviate this issue, finding the optimal structure can be computationally intensive, particularly with higher, more complex models.

In this instance, the decision tree provided the lowest accuracy rate compared to the other two methods, especially considering the complexity of dealing with three categories instead of two, as observed in the linear regression testing. The hierarchical structure of decision trees may not always efficiently handle multi-class scenarios, which can be reflected in the lower accuracy rate.

Contributions

Contributions

The contributions of each team member were evenly distributed among their respective skill sets and knowledge.

- Code:
    - READ_ME file: Nour
    - R file: All students contributed equally to the code portion of this project.
- Presentation:
    - All students contributed equally to the preparation and recording of the presentation.
- Report:
    - Introduction: Zahra
    - Linear Regression: Zahra
    - Logistic Regression: Kurubel
    - Decision Tree: Nour
    - Comparitive Analysis: All students contributed equally.