**PREDICTIVE MODELLING OF BIOOIL AND BIOCHAR PRODUCTION FROM MICROWAVE PYROLYSIS OF VARIOUS BIOMASS**

PROJECT

By

## **RUTVIK KUMAR**



**SCHOOL OF INFRASTRUCTURE**

**INDIAN INSTITUTE OF TECHNOLOGY BHUBANESWAR**

**BHUBANESWAR -751013, ODISHA**

**NOVEMBER 2023**

**PREDICTIVE MODELLING OF BIOOIL AND BIOCHAR PRODUCTION FROM MICROWAVE PYROLYSIS OF VARIOUS BIOMASS**

**A PROJECT**

Submitted in partial fulfillment of the

requirement for the award of the degree

of

MASTER OF TECHNOLOGY

in

ENVIRONMENTAL ENGINEERING

By

## **RUTVIK KUMAR (19CE02026)**

Under the supervision of

**Dr. REMYA NEELANCHERRY**



**SCHOOL OF INFRASTRUCTURE**

**INDIAN INSTITUTE OF TECHNOLOGY BHUBANESWAR**

**BHUBANESWAR -751013, ODISHA**

**NOVEMBER 2023**



**INDIAN INSITUTE OF TECHNOLOGY BHUBANESWAR**

# **CANDIDATE’S DECLARATION**

We hereby certify that the work which is being presented in the project entitled “**PREDICTIVE MODELLING OF BIO-OIL AND BIO-CHAR PRODUCTION FROM MICROWAVE PYROLYSIS OF VARIOUS BIOMASS**”, in the partial fulfillment of the requirement for the award of the Degree of Master of Technology in Enviornmental Engineering and submitted in the School of Infrastructure, Indian Institute of Technology Bhubaneswar, is an authentic record of my own work carried out during a period from July 2022 to May 2023 under the supervision of **Dr. REMYA NEELANCHERRY**, IIT Bhubaneswar. The matter presented in the project has not been submitted by me for the award of any degree of this or any other Institute/ University

**RUTVIK KUMAR (19CE02026)**

This is to certify that the above statement made by the candidates is correct to the best of our knowledge.

**Signature of Supervisor (S) Signature of H.O.S.**

# ABSTRACT

Accurately predicting the properties of bio-char and bio-oil holds paramount importance for their effective utilization. However, these predictions are significantly swayed by the diverse compositions of biomass and the specific conditions of pyrolysis. Biomass, encompassing crops, wood, agricultural residues, and forestry remnants, stands as a pivotal renewable resource for the generation of sustainable bioenergy and eco-friendly materials. Microwave heating presents itself as a superior alternative to conventional heating due to its myriad advantages. Moreover, the judicious inclusion of suitable catalysts exerts a profound impact on the product selectivity during microwave pyrolysis. This research endeavour offers a novel perspective on forecasting the attributes of bio-oil and bio-char derived from various biomass compositions subjected to distinct pyrolysis conditions. It employs a range of machine learning and deep learning methodologies to facilitate these predictions. By doing so, it not only enhances our understanding of bio-char and bio-oil behaviour but also contributes to the optimization of biomass-based energy production and environmentally conscious materials manufacturing.

Keywords: bio-char; machine learning; linear regression; random forest regression, support vector machine, adaptive boosting, gradient boosting

# ACKNOWLEDGEMENT

In successfully completing this project, many people have helped me. I would like to thank all those who are related to this project.

I express my sincere gratitude towards my guide **Dr. Remya Neelancherry** for guiding me through the project. Her consistent guidance helped me to understand the project work and execute it. I would like to acknowledge her support, timely advice, and suggestion over the project, which immensely helped me complete this project.

I want to thank my friends for their consistent help in understanding some terms and for motivating me throughout the project tenure.

### **CONTENTS**

Page No.

Candidate’s Declaration i

Abstract iii

Acknowledgement v

Content vii

List of Figures ix

List of Tables xi

Notations xiii

**CHAPTER**

1. **INTRODUCTION**

* 1. General 1
  2. Objective 2

2. **LITERATURE REVIEW** 3

* 1. General 3
  2. Microwave pyrolysis 3

2.2.1 Principle of microwave pyrolysis

2.2.2 Advantages of microwave pyrolysis 3

2.2.3 Process and Products 4

2.2.4 Applications 4

* 1. Biomass feedstock dataset 4
  2. Machine Learning Algorithm 7

2.4.1 Linear Regression 7

2.4.2 Random Forest Regression 8

2.4.3 Support Vector Regressor 8

2.4.4 Decision Tree 9

2.4.5 Adaptive Boosting Regressor 9

2.4.6 Gradient Boosting Regressor 10

2.3.5 Predictive Model Performance Evaluation Matri 11

1. **METHODOLOG** 13
   1. General 13
   2. Prediction of Biochar yield (%), SSA and HHV 14

4. **RESULTS AND DISCUSSION** 15

4.1 General 15

4.2 Effect of Process Parameters 15

4.3 Comparisons of various ML predictive Models 15

4.3.1 Prediction of Yield (%) 18

4.2.2 Prediction of SSA 22

4.2.3 Prediction of HHV 26

5. **CONCLUSION 31**

5.1 Conclusion 31

**REFERENCES** 33

# LIST OF FIGURES

Figure. No. Details of Figure Page No.

4.1 Scatter plot of various parameters and target variable 16

4.2 Correlation between process parameters and target variable 17

4.3 Regression plot for yield (%) using proximate and process parameters 19

4.4 Regression plot for yield (%) using ultimate and process parameters 20

4.5 Feature importance plot for yield (%) prediction 22

4.6 Regression plot for SSA using proximate and process parameters 23

4.7 Regression plot for SSA using ultimate and process parameters 24

4.8 Feature importance plot for SSA prediction 26

4.9 Regression plot for HHV using proximate and process parameters 27

4.10 Regression plot for HHV using ultimate and process parameters 29

4.11 Feature importance plot for HHV prediction 30

# LIST OF TABLES

Table No. Details of Table Page No.

2.1 sample dataset of process parameters and experimental yield % 5

4.1 Error value for yield (%) prediction using proximate and process parameters 20

4.2 Error value for yield (%) prediction using ultimate and process parameters 21

4.3 Error value for SSA prediction using proximate and process parameters 24

4.4 Error value for SSA prediction using ultimate and process parameters 25

4.5 Error value for HHV prediction using proximate and process parameters 28

4.6 Error value for HHV prediction using ultimate and process parameters 30

# NOTATIONS

|  |  |  |
| --- | --- | --- |
|  |  |  |
| FW |  | Food Waste |
| MAE |  | Mean Absolute Error |
| MSE |  | Mean Squared Error |
| RMSE |  | Root Mean Squared Error |
| R2 |  | R-squared |

## **Chapter 1**

## **INTRODUCTION**

**1.1 General**

The excessive depletion of non-renewable fossil fuels has led to profound concerns regarding energy security and environmental degradation. Consequently, the focus has shifted towards harnessing biomass energy, which is abundant and sustainable. Thermochemical conversion stands out as an efficient and cost-effective means of transforming biomass into biofuels, which can be further synthesized into desired chemicals or directly utilized. This conversion process encompasses combustion, gasification, and pyrolysis, with pyrolysis, a thermally induced decomposition method conducted in an oxygen-free environment, emerging as a promising technique due to its simplicity and rapidity of operation. Pyrolysis holds great potential for biomass conversion, offering a pathway to address energy and environmental challenges.

Assessing the suitability of various biomass feedstocks for energy conversion processes, including combustion, pyrolysis, gasification, liquefaction, and hydrothermal processing, necessitates comprehensive characterization and experimental investigations. Among the critical properties for these thermal conversions, the Higher Heating Value (HHV), also known as gross calorific value or gross energy, plays a pivotal role. HHV quantifies the heat released when a unit mass of fuel undergoes complete combustion, producing water in a condensed state.

Traditionally, researchers measure the HHV of biomass and other fuel sources using an adiabatic oxygen bomb calorimeter, a widely accepted, precise, but time-consuming and cost-intensive method. Unfortunately, this method isn't readily accessible to all researchers interested in HHV determination for biomass. Pyrolysis, a process involving thermal decomposition in the absence of oxygen, constitutes the initial step in generating primary products for combustion and gasification processes, followed by the total or partial oxidation of intermediate products. Biomass pyrolysis yields solid char, liquid bio-oil, and gaseous fuel, with their proportions influenced by operational variables. Lower process temperatures and extended vapor residence times favour char production, while higher temperatures and longer vapor residence times enhance gas production. Optimal conditions for liquid production involve moderate temperatures and shorter vapor residence times. Much research on pyrolysis employs conventional heating sources like electric and gas heaters. However, utilizing microwave irradiation for biomass pyrolysis shows promise due to several advantages over conventional methods.

Evaluating biomass suitability for energy conversion processes necessitates characterizing properties like HHV, while pyrolysis serves as a critical step in producing biofuels. Microwave-assisted pyrolysis presents a promising avenue for its efficiency and potential benefits compared to traditional methods. In this work, a comprehensive analysis of the bio-oil and bio-char yield, HHV and specific surface area prediction was presented. Using machine learning tools, the quality will be predicted according to biomass compositions and pyrolysis conditions, and the correlation analysis will be carried out. This study provides feasible thinking for the prediction of the characteristics of bio-oil and biochar obtained by biomass with different compositions under different pyrolysis conditions.

**1.2 Objective**

The primary goal is to examine the production of bio-oil and biodiesel through microwave pyrolysis of biomass. This involves scrutinizing proximate analysis parameters, ultimate analysis parameters, and various process factors like reaction temperature, reaction time, particle size, volatile matter, etc. Subsequently, machine learning models will be developed to forecast yield percentage, specific surface area, and high heating value. The aim is to identify optimal values for process parameters that result in the highest yield. Through a comprehensive analysis of these factors, the study seeks to enhance the efficiency of bio-oil and biodiesel production. By leveraging machine learning, the research aims to not only understand the intricate relationships within the pyrolysis process but also to predict and optimize crucial outcomes, paving the way for advancements in sustainable biofuel production.

## **Chapter 2**

## **LITERATURE REVIEW**

**2.1 General**

Predictive modelling of bio-oil and bio-char production will require the use of machine learning algorithms and the process involved for experimentation. In this chapter we have covered the method involved for production of biofuel. It has describe about the Microwave pyrolysis process, the different feedstocks used and what features are considered important to predict the yield, HHV and specific surface area. This chapter also provides the information about the basic machine learning algorithms like linear regression and some advanced algorithms like random forest, adaptive boosting regression and artificial neural network also provides a detail overview of how this algorithms work.

**2.2 Microwave Pyrolysis**

Microwave pyrolysis is an innovative and efficient technique employed in the production of bio-oil and bio-char from biomass. This method leverages microwave radiation to induce thermal decomposition of biomass in the absence of oxygen, leading to the creation of valuable products.

**2.2.1 Principle of Microwave Pyrolysis**

Microwave pyrolysis relies on the principle of selective heating. When microwave radiation is applied to biomass, water molecules within the material absorb the microwave energy and rapidly heat up, leading to the formation of steam. As the temperature increases, the biomass undergoes pyrolysis, breaking down into its constituent components: solid bio-char, liquid bio-oil, and gaseous products.

**2.2.2 Advantages of Microwave Pyrolysis**

1. Speed: Microwave pyrolysis is significantly faster compared to conventional heating methods. The selective heating of water molecules within the biomass allows for rapid and uniform heating throughout the material.
2. Energy Efficiency: It is more energy-efficient because it directly targets the water content within the biomass, reducing energy wastage.
3. Enhanced Product Quality: The controlled and uniform heating in microwave pyrolysis can lead to higher yields of bio-oil and bio-char with improved quality.
4. Reduced Environmental Impact: Microwave pyrolysis can be carried out in an oxygen-free environment, reducing the production of harmful by-products and minimizing emissions.

**2.2.3 Process and Products**

In a typical microwave pyrolysis process:

1. Biomass is loaded into a microwave-reactive vessel.
2. Microwaves are applied, causing rapid heating of the biomass.
3. As the temperature increases, the biomass undergoes pyrolysis, releasing volatile compounds that can be condensed into bio-oil.
4. The remaining solid residue is bio-char, which is rich in carbon and can be used as a valuable soil conditioner or for carbon sequestration.
5. Gaseous by-products can also be collected and used for energy generation or other applications.

**2.2.4 Applications**

* Bio-Oil: The liquid bio-oil obtained from microwave pyrolysis can be a valuable feedstock for the production of biofuels, chemicals, and other high-value products. Its composition can vary depending on the biomass source and pyrolysis conditions.
* Bio-Char: Bio-char, the solid residue, finds applications in agriculture as a soil amendment to improve soil fertility, water retention, and carbon sequestration. It can also be used as a precursor for activated carbon production.

Microwave pyrolysis is a promising method for the sustainable production of bio-oil and bio-char from biomass. Its advantages in terms of speed, energy efficiency, and product quality make it an attractive option for biomass conversion, contributing to the development of renewable energy and environmentally friendly materials.

**2.3 Biomass feedstocks dataset**

The experimental data of bio-oil and bio-char were collected from the literature. The dataset consists of a wide range of biomass feedstocks such as commercial fuels, industrial wastes, forest wastes (including branches and leaves), energy crops and cereals, which all can be used to build a model valid in a wide range of biomass feedstocks to predict the yield, specific surface area and HHV. Table 2.1 gives the sample dataset of process parameters from the original dataset which were used for prediction.

**Table 2.1. Sample Dataset of Process Parameters and target parameters**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **biomass type** | **C (%)** | **H (%)** | **N (%)** | **O (%)** | **VM (%)** | **Ash (%)** | **FC (%)** | **T (°C)** | **RT (min)** | **HR (°C/min)** | **N-char (%)** | **Yield-char (%)** | **SSA-char (m2/g)** | **HHV (MJ/KG)** |
| Durian Wood | 33.77 | 11.40 | 6.06 | 48.77 | 85.80 | 10.52 | 3.68 | 350 | 180 | 1.94 | 5.12 |  | 2.57 | 14.31 |
| Durian Wood | 33.77 | 11.40 | 6.06 | 48.77 | 85.80 | 10.52 | 3.68 | 450 | 180 | 2.5 | 2.01 |  | 45.68 | 14.31 |
| Durian Wood | 33.77 | 11.40 | 6.06 | 48.77 | 85.80 | 10.52 | 3.68 | 550 | 180 | 3.05 | 0.55 |  | 221.00 | 14.31 |
| marine macroalgae | 39.78 | 6.35 | 5.40 | 48.46 |  | 24.51 |  | 200 | 120 | 1.66 | 5.24 | 78.34 | 1.91 | 0.00 |
| marine macroalgae | 39.78 | 6.35 | 5.40 | 48.46 |  | 24.51 |  | 400 | 120 | 3.33 | 4.74 | 63.64 | 70.29 | 0.00 |
| marine macroalgae | 39.78 | 6.35 | 5.40 | 48.46 |  | 24.51 |  | 600 | 120 | 5 | 4.47 | 37.96 | 61.81 | 0.00 |
| empty fruit bunch (EFB) | 53.04 | 6.39 | 0.87 | 39.37 | 82.07 | 4.88 | 13.05 | 500 | 60 | 8.33 |  |  | 1.41 | 16.66 |
| empty fruit bunch (EFB) | 53.04 | 6.39 | 0.87 | 39.37 | 82.07 | 4.88 | 13.05 | 300 | 120 | 2.5 |  |  | 1.42 | 16.66 |
| empty fruit bunch (EFB) | 53.04 | 6.39 | 0.87 | 39.37 | 82.07 | 4.88 | 13.05 | 700 | 120 | 5.83 |  |  | 1.13 | 16.66 |
| rice husk (RH) | 52.85 | 6.19 | 0.44 | 40.22 | 70.78 | 12.00 | 17.22 | 500 | 60 | 8.33 |  |  | 7.84 | 16.23 |
| rice husk (RH) | 52.85 | 6.19 | 0.44 | 40.22 | 70.78 | 12.00 | 17.22 | 300 | 120 | 2.5 |  |  | 0.39 | 16.23 |
| corn straw pellets | 40.02 | 6.01 | 0.88 | 52.86 | 72.29 | 10.02 | 17.68 | 400 | 30 | 13.3 | 1.29 | 50.17 | 4.26 | 16.61 |
| corn straw pellets | 40.02 | 6.01 | 0.88 | 52.86 | 72.29 | 10.02 | 17.68 | 450 | 30 | 15 | 1.20 | 49.57 | 12.61 | 16.61 |
| Saccharina japonica algae | 40.18 | 5.55 | 2.00 | 51.39 |  |  |  | 300 | 60 | 5 |  |  | 1.36 | 0.00 |
| Saccharina japonica algae | 40.18 | 5.55 | 2.00 | 51.39 |  |  |  | 450 | 60 | 7.5 |  |  | 210.00 | 0.00 |
| Saccharina japonica algae | 40.18 | 5.55 | 2.00 | 51.39 |  |  |  | 600 | 60 | 10 |  |  | 266.00 | 0.00 |
| rice husk (RH) | 40.65 | 5.95 | 1.07 | 52.19 |  | 16.53 |  | 400 | 60 | 6.66 | 8.65 | 60.57 | 5.49 | 0.00 |
| rice husk (RH) | 40.65 | 5.95 | 1.07 | 52.19 |  | 16.53 |  | 700 | 60 | 11.6 | 6.59 | 52.58 | 7.12 | 0.00 |
| sewage sludge (SS) | 45.74 | 10.76 | 7.04 | 34.91 | 49.36 | 42.02 | 8.63 | 400 | 60 | 6.66 | 4.83 | 53.84 | 4.39 | 10.26 |
| sewage sludge (SS) | 45.74 | 10.76 | 7.04 | 34.91 | 49.36 | 42.02 | 8.63 | 700 | 60 | 11.6 | 3.35 | 40.35 | 10.71 | 10.26 |
| bamboo sawdust (BS) | 49.51 | 3.42 | 0.77 | 46.02 |  | 2.04 |  | 400 | 60 | 6.66 | 4.43 | 44.51 | 2.78 | 0.00 |
| Raw maize straw | 42.19 | 7.21 | 1.28 | 49.19 |  | 5.31 |  | 250 | 60 | 4.16 |  |  | 5.20 | 0.00 |
| Raw maize straw | 42.19 | 7.21 | 1.28 | 49.19 |  | 5.31 |  | 300 | 60 | 5 | 1.68 | 36.70 | 7.80 | 0.00 |
| Raw maize straw | 42.19 | 7.21 | 1.28 | 49.19 |  | 5.31 |  | 500 | 60 | 8.33 | 1.71 | 28.90 | 33.20 | 0.00 |
| Raw sugarcane bagasse | 48.88 | 7.11 | 1.02 | 42.88 |  | 4.98 |  | 500 | 60 | 8.33 | 0.68 | 23.70 | 97.80 | 0.00 |
| Raw sugarcane bagasse | 48.88 | 7.11 | 1.02 | 42.88 |  | 4.98 |  | 500 | 480 | 1.04 |  | 20.30 | 105.40 | 0.00 |
| Raw sugarcane bagasse | 48.88 | 7.11 | 1.02 | 42.88 |  | 4.98 |  | 600 | 480 | 1.25 |  | 18.80 | 108.70 | 0.00 |
| Raw pine sawdust | 50.61 | 6.18 | 0.05 | 43.10 |  | 3.76 |  | 500 | 60 | 8.33 | 0.15 | 25.80 | 68.40 | 0.00 |
| Raw pine sawdust | 50.61 | 6.18 | 0.05 | 43.10 |  | 3.76 |  | 500 | 480 | 1.04 |  | 24.30 | 79.30 | 0.00 |
| sewage sludge | 43.97 | 13.67 | 7.24 | 32.84 | 55.80 | 33.70 | 10.50 | 300 | 10 | 30 | 8.64 |  | 14.00 | 11.84 |
| sewage sludge | 43.97 | 13.67 | 7.24 | 32.84 | 55.80 | 33.70 | 10.50 | 500 | 10 | 50 | 6.76 |  | 55.72 | 11.84 |
| sewage sludge | 43.97 | 13.67 | 7.24 | 32.84 | 55.80 | 33.70 | 10.50 | 600 | 10 | 60 | 9.20 |  | 92.45 | 11.84 |
| soybean oil cake | 44.48 | 6.28 | 8.21 | 40.49 | 77.50 | 6.47 | 16.03 | 600 | 60 | 10 | 8.48 | 29.72 | 9.47 | 16.89 |
| soybean oil cake | 44.48 | 6.28 | 8.21 | 40.49 | 77.50 | 6.47 | 16.03 | 800 | 60 | 13.3 | 5.51 | 27.77 | 5.37 | 16.89 |
| Switchgrass | 44.70 | 7.24 | 1.19 | 46.87 | 94.16 | 2.47 | 3.37 | 600 | 60 | 10 | 0.60 | 23.30 | 47.50 | 15.49 |
| Switchgrass | 44.70 | 7.24 | 1.19 | 46.87 | 94.16 | 2.47 | 3.37 | 800 | 60 | 13.3 | 0.46 | 21.34 | 62.50 | 15.49 |
| Water Oak | 53.65 | 7.45 | 1.05 | 37.84 | 85.80 | 1.87 | 12.33 | 200 | 60 | 3.33 | 0.87 | 74.88 | 2.92 | 17.01 |
| Water Oak | 53.65 | 7.45 | 1.05 | 37.84 | 85.80 | 1.87 | 12.33 | 400 | 60 | 6.66 | 0.74 | 38.86 | 17.71 | 17.01 |
| Water Oak | 53.65 | 7.45 | 1.05 | 37.84 | 85.80 | 1.87 | 12.33 | 600 | 60 | 10 | 0.62 | 27.04 | 31.87 | 17.01 |
| Biosolids | 58.49 | 7.71 | 5.70 | 28.11 | 54.67 | 38.19 | 7.14 | 400 | 60 | 6.66 | 7.29 | 48.74 | 67.29 | 10.61 |
| Biosolids | 58.49 | 7.71 | 5.70 | 28.11 | 54.67 | 38.19 | 7.14 | 600 | 60 | 10 | 7.54 | 33.01 | 55.42 | 10.61 |
| Biosolids | 58.49 | 7.71 | 5.70 | 28.11 | 54.67 | 38.19 | 7.14 | 800 | 60 | 13.3 | 6.36 | 30.12 | 53.12 | 10.61 |
| goat manure | 45.02 | 6.57 | 2.19 | 46.22 | 76.12 | 18.95 | 4.93 | 400 | 30 | 13.3 | 2.74 | 44.50 | 3.27 | 13.22 |
| goat manure | 45.02 | 6.57 | 2.19 | 46.22 | 76.12 | 18.95 | 4.93 | 500 | 30 | 16.6 | 2.75 | 40.60 | 1.68 | 13.22 |
| goat manure | 45.02 | 6.57 | 2.19 | 46.22 | 76.12 | 18.95 | 4.93 | 600 | 30 | 20 | 1.78 | 37.90 | 13.92 | 13.22 |
| Municipal biosolids | 45.18 | 10.63 | 6.95 | 35.71 |  | 41.30 |  | 400 | 60 | 6.66 | 8.61 | 60.57 | 5.49 | 0.00 |
| Municipal biosolids | 45.18 | 10.63 | 6.95 | 35.71 |  | 41.30 |  | 450 | 60 | 7.5 | 9.13 | 59.55 | 7.21 | 0.00 |

**2.3 Machine Learning Algorithms**

**2.3.1 Linear Regression**

Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable also change accordingly (increase or decrease) [13].

Mathematically, the relationship can be represented per equation 2:

**𝑌 = 𝑚𝑋 + 𝑏** (2)

Where, 𝑌 is the dependent variable (to be predicted)

𝑋 is the dependent variable used to make predictions.

𝑚 is the slope of the regression line which represents the effect 𝑋 on 𝑌

𝑏 is a constant, known as the 𝑌-intercept. If 𝑋 = 0, 𝑌 would be equal to 𝑏.

In case the best fit line passes through origin, the equation becomes

**𝑌 = 𝑏𝑋**  (3)

Generalized linear regression model for any response (y) and its predictors (𝑥1, 𝑥2, ..., 𝑥n), i.e., biodiesel production process parameters in this case can be stated as:

**Y = 𝛽0 + 𝛽1𝑥1 + 𝛽2𝑥2 + 𝛽n𝑥n**  (4)

**2.3.2 Random Forest Regression**

The Random Forest is based on applying bagging to decision trees, with one important extension: in addition to sampling the records, the algorithm also samples the variables. To determine how to create a sub partition of a partition A, the algorithm chooses the variable and split point by minimising a criterion such as Gini impurity with random forests, the choice of variable is limited to a random subset of variables at each stage of the algorithm [14]. Compared to the basic tree algorithm, the random forest algorithm adds two more steps: the bagging and the bootstrap sampling of variables at each split:

1. Take a bootstrap (with replacement) subsample from the *records*.
2. For the first split, sample *p* < *P variables* at random without replacement.
3. For each of the sampled variables *Xj*1 , *Xj*2 , ..., *Xjp* , apply the splitting algorithm:
4. For each value *sjk* of *Xjk* :
5. Split the records in partition *A*, with *Xj*(*k*) < *sj*(*k*) as one partition and the remaining records where *Xjk* ≥ *sjk* as another partition.
6. Measure the homogeneity of classes within each subpartition of *A*.
7. Select the value of *sjk* that produces maximum within-partition homogeneity of class.
8. Select the variable *Xjk* and the split value *sjk* that produces maximum with in partition homogeneity of class.
9. Proceed to the next split and repeat the previous steps, step 2 onwards.
10. Continue with additional splits, following the same procedure until the tree is grown.
11. Go back to step 1, take another bootstrap subsample, and start the process over again.

**2.3.3 Support Vector Regressor**

A Support Vector Regressor (SVR) is a machine learning algorithm used for regression tasks. It is based on the same principles as the Support Vector Machine (SVM), which is primarily used for classification. SVR, however, is designed to predict continuous values, making it suitable for tasks such as stock price forecasting, house price prediction, and more. SVR works by finding a hyperplane that best fits the data, while also minimizing the margin of error. In SVR, this margin is called the "tube," and the goal is to fit as many data points as possible within this tube while limiting violations (data points that fall outside the tube). The data points that touch the edges of the tube are the support vectors, and they play a critical role in determining the hyperplane.

To achieve this, SVR uses a mathematical optimization technique that involves minimizing the L2 norm of the weight vector (hyperplane) while subject to the constraint that the error on each data point is within a specified epsilon (ε). The choice of kernel function, such as linear, polynomial, or radial basis function, allows SVR to handle nonlinear relationships between input features and the target variable. SVR is a powerful regression algorithm that uses support vectors to find the best-fitting hyperplane and can handle both linear and nonlinear relationships in the data, making it a valuable tool in predictive modelling.

**2.3.4 Decision Tree**

A Decision Tree Regressor is a machine learning algorithm used for regression tasks. It works by recursively partitioning the data into subsets based on the values of input features to predict continuous output values. Each split in the tree represents a decision or a rule, and the outcome is obtained by averaging the target values of data points within a leaf node. The tree-building process begins with the entire dataset and aims to find the feature and split point that best separates the data based on a chosen criterion, often mean squared error. This process continues recursively until a stopping criterion is met, such as a maximum depth of the tree or a minimum number of data points in a leaf node.

Decision trees are intuitive and easy to interpret, allowing users to understand the model's decision-making process. However, they can be prone to overfitting, meaning they capture noise in the training data. Techniques like pruning and setting appropriate hyperparameters help mitigate overfitting. Decision Tree Regressors are useful for modelling both linear and nonlinear relationships between input features and target variables. They are the building blocks for more complex ensemble methods like Random Forests and Gradient Boosting, which combine multiple decision trees to improve predictive accuracy. Decision trees are valuable tools in various fields, including finance, healthcare, and environmental modelling, where interpretable and accurate regression models are required.

**2.3.5 Adaptive Boosting Regressor**

Adaptive boosting, also known as AdaBoost, is a sequential ensemble technique that develops several weak learners using different training subsets drawn at random from the original training dataset [13-14]. Weights are assigned during each training and are used when learning each hypothesis. The weights are used to compute the hypothesis error on the dataset and act as an indicator of the relative importance of each instance. After each iteration, the weights are recalculated so that instances incorrectly classified by the previous hypothesis receive higher weights. This allows the algorithm to concentrate on more difficult-to-learn instances. The algorithm's most important task is to assign revised weights to incorrectly classified instances. In regression, unlike in classification, the instances are not correct or incorrect; rather, they constitute area-value error. The computed error can be labelled as an error or not error by comparing it to a predefined threshold prediction error, and thus the AdaBoost classifier can be used. Instances with larger errors on previous learners are more likely (i.e., have a higher probability) of being chosen to train the next base learner. Finally, the weighted average or median of the individual base learner predictions is used to provide an ensemble prediction.

**2.3.6 Gradient Boosting Regressor**

A Gradient Boosting Regressor is a powerful machine learning technique used for regression tasks. It is an ensemble learning method that combines the predictions of multiple weaker models, typically decision trees, to create a robust and accurate predictive model.

Here's how Gradient Boosting Regressor works:

1. Sequential Learning: Gradient boosting builds an ensemble of decision trees sequentially. Each tree corrects the errors made by the previous ones. The process starts with an initial, simple model, and subsequent trees are trained to minimize the residual errors of the ensemble.
2. Gradient Descent: The term "gradient" in gradient boosting refers to the use of gradient descent optimization to minimize the loss function. It adjusts the weights (or parameters) of the individual trees in each iteration to minimize the prediction errors.
3. Weak Learners: Gradient boosting typically employs shallow decision trees as weak learners. These trees are often referred to as "stumps" and are combined to create a strong predictive model. The depth of these trees is a hyperparameter that can be tuned.
4. Regularization: To prevent overfitting, gradient boosting incorporates regularization techniques, such as tree depth control, learning rate adjustment, and feature subsampling
5. Predictive Power: The final prediction is a weighted sum of predictions from all the weak learners, where the weights are determined during training. This ensemble model can capture both linear and nonlinear relationships in the data, resulting in a highly accurate regression model.

Gradient Boosting Regressor is widely used in various fields, including finance, healthcare, and natural language processing. It is known for its ability to handle complex datasets, robustness against outliers, and excellent predictive performance, making it a popular choice for regression tasks when high accuracy is required.

**2.3.7 Predictive Model Performance Evaluation Metrics**

The residue, 𝜀i between the ith original value, 𝑦i and predicted value, y0 is calculated as,

**𝜀i = 𝑦i – 𝑦0** (5)

The coefficient of determination, 𝑅2 is computed using the 𝑦i, 𝑦0 and the mean of the dataset, ym.

**𝑅2 = 1 − ∑ (𝑦i− 𝑦0i)2/∑ (𝑦i − 𝑦m)2**  (6)

The mean-absolute-error is computed based on the number of the samples, 𝑛 as

**MAE = ∑ |𝑦i − 𝑦0i|/𝑛** (7)

The mean-squared-error, MSE, is computed as

**MSE = ∑ (𝑦i − 𝑦0i)/𝑛** (8)

The root-mean-squared-error, RMSE, is computed as

**RMSE = (∑ (𝑦i – y0i)2/𝑛 )0.5** (9) The maximum error is computed as

**Max. Error (𝑦, 𝑦0) = max. (|𝑦i − 𝑦0i|)** (10)

The median error is computed as

**Med. Error (y, y0) = median (|yi – y01|, ... , |yn – y0n)** (11)

## **Chapter 3 METHODOLOGY**

**3.1 General**

In this research endeavour, the primary focus revolves around the meticulous analysis of experimental data with the aim of discerning underlying patterns and trends. The data in question holds the key to understanding and potentially predicting the bio-oil and bio-char production processes. To harness the full potential of this data, a crucial preliminary step involves data pre-treatment to render it suitable for utilization within predictive models. The foundation of this research rests on the utilization of cutting-edge Machine Learning algorithms, which serve as the cornerstone of our predictive endeavours. These algorithms are strategically applied to predict the outcomes of the bio-oil and bio-char production processes. The predictive models are designed to consider various process parameters, which play a pivotal role in determining the final outcomes. Among the ensemble of Machine Learning algorithms at our disposal, we will employ linear regression and random forest regression. These techniques have been chosen based on careful observation and analysis of the experimental data. Their application is rooted in the belief that they possess the capability to effectively capture and extrapolate the intricate relationships between the process parameters and the production responses. To gauge the effectiveness and accuracy of these predictive models, several error metrics are employed. These metrics serve as benchmarks against which the performance of the Machine Learning regression models is evaluated. By quantifying the disparities between predicted values and actual outcomes, these error metrics provide invaluable insights into the predictive power and reliability of the models. The significance of this research extends beyond its immediate objectives. It holds the potential to revolutionize the bio-oil and bio-char production processes. The ability to predict these processes with a high degree of accuracy can pave the way for enhanced efficiency, reduced costs, and environmentally friendly production methods. Moreover, the insights gleaned from this research can be extrapolated to other domains, demonstrating the broader applicability of Machine Learning in industrial and scientific endeavours. In conclusion, this research embarks on a journey into the realm of data-driven analysis and prediction in the context of bio-oil and biochar production processes. Ultimately, this research has the potential to reshape the landscape of biofuel production and contribute to a more sustainable and eco-conscious future.

**3.2 Prediction of Bio-char yield (%), SSA and HHV**

The objective is to forecast the yield (%), specific surface area (SSA), and higher heating value (HHV) of Biochar using a minimal set of parameters. This prediction task is divided into three segments, one for each target variable (yield, SSA, HHV). Furthermore, within each segment, there are two approaches: (a) using proximate analysis and process parameters and (b) utilizing ultimate analysis and process parameters. This results in a total of 12 distinct models, each tailored to predict one of the three properties, depending on the chosen parameter set.

**Chapter 4**

## **RESULTS AND DISCUSSION**

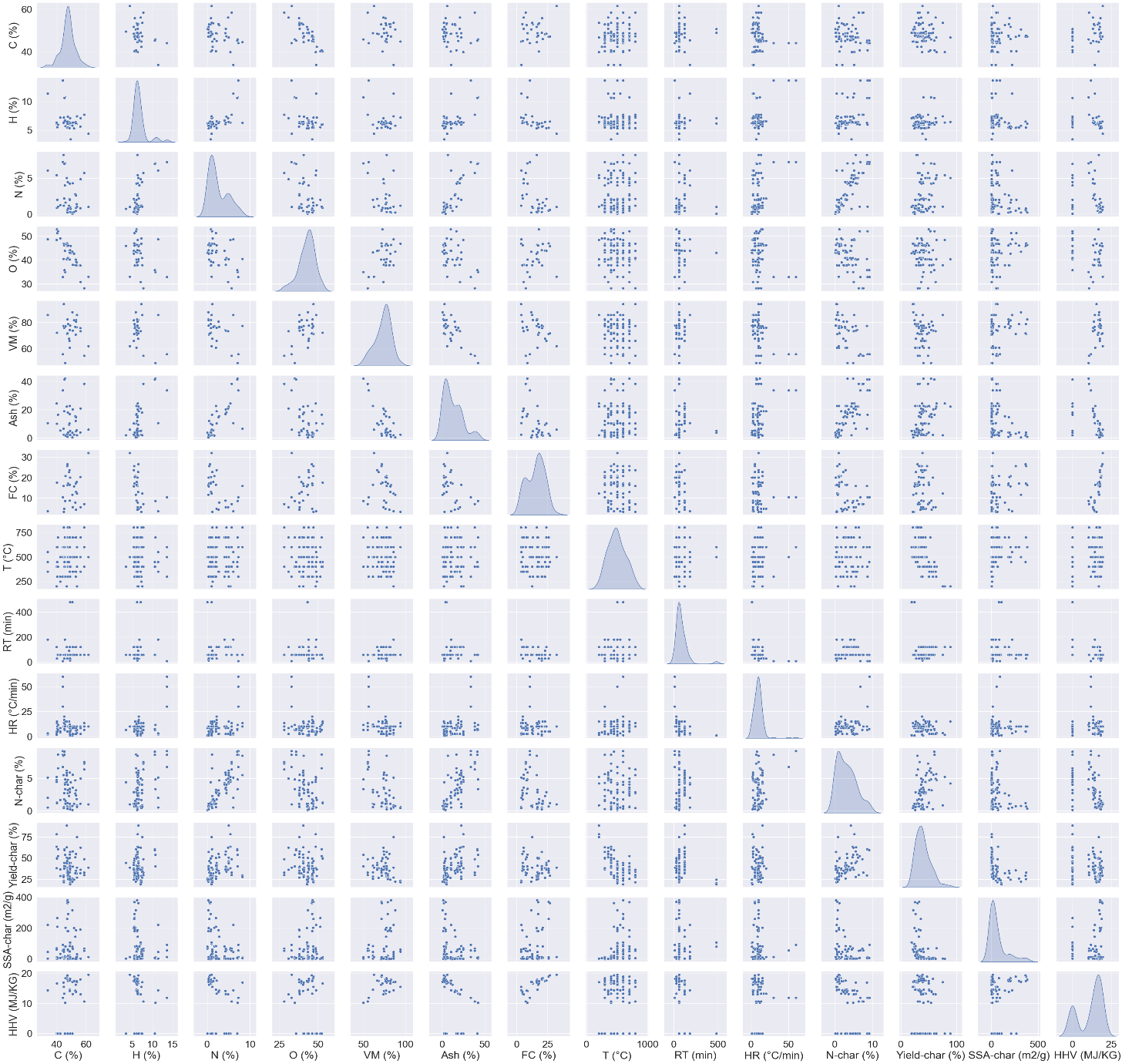
**4.1 General**

Evaluating the performance of different experimentation with model is must to get the best one for prediction. This chapter includes the information about how much these parameters are correlated with each other and the performance of different machine learning algorithms using this process parameters for predicting the bio-char yield, SSA and HHV. This chapter also includes the best process parameters to predict the accurate target variable.

**4.2 Effect of Process Parameters**

Figure 4.1 and Table 2.1 serve as visual representations of our training data, showcasing scatter plots that reveal the relationships between various features and their impact on the yield (%), specific surface area (SSA), and higher heating value (HHV). The scatter plots, thoughtfully organized, exhibit how each feature's variation corresponds to these critical variables. A significant observation can be made when examining the histograms displayed along the diagonal of the scatter plots. These histograms indicate that the distribution of our features closely resembles a normal distribution. This normality in feature distribution is a crucial assumption in many statistical analyses and machine learning algorithms, enhancing the reliability of our results.

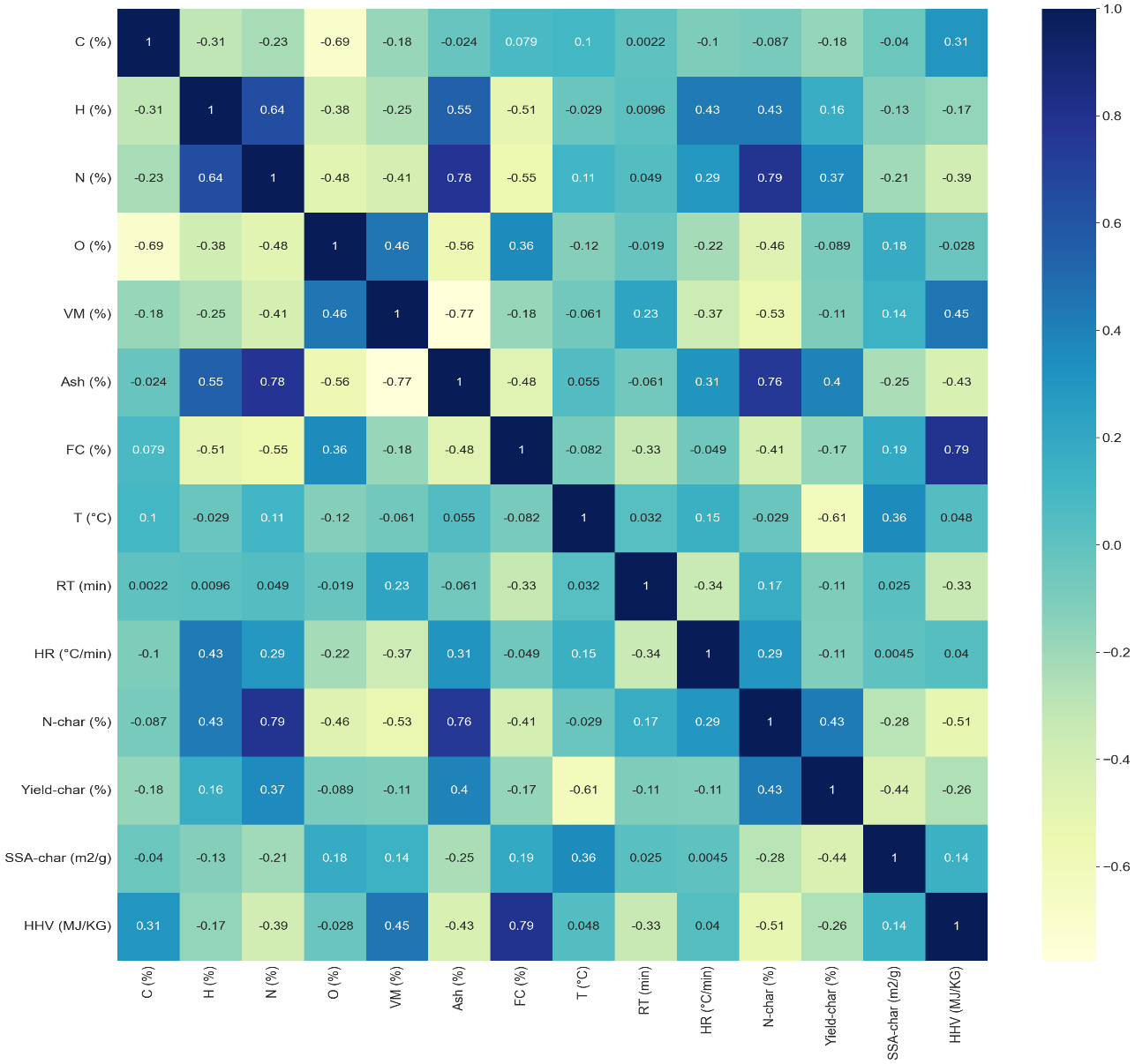
When focusing on the relationships between the yield (%) and our features, it becomes apparent that C (%), H (%), N (%), and O (%) demonstrate a positive correlation with yield (%). In contrast, SSA exhibits a positive association with Ash (%), FC (%), and T (°C). Moreover, HHV displays a positive relationship with RT (min) and HR (°C/min). The yield histogram emphasizes that the central tendency of all experimentally determined yields hovers around 40%. This statistic provides a pivotal insight into the overall performance and characteristics of our dataset, serving as a reference point for further analysis and modelling.



**Figure 4.1 Scatter plot of the various process parameters and Target Variable.**

Figure 4.2 delves into the intricate relationship between various process parameters and the target variable. A notable observation here is that the process parameters exhibit negligible correlation, confirming the absence of multicollinearity. Multicollinearity, if present, can undermine the statistical power of regression analyses, and its absence in this context is a favourable characteristic. Yield (%) exhibits strong positive correlations with H (%), N (%), and Ash (%) with correlation coefficients of 0.16, 0.37, and 0.4, respectively. Conversely, it demonstrates a substantial negative correlation with T (°C), registering a correlation coefficient of -0.61. SSA shows robust positive correlations with VM (%), FC (%), and T (°C), boasting correlation coefficients of 0.14, 0.19, and 0.36, respectively. Conversely, it exhibits noteworthy negative correlations with N (%) and Ash (%) with correlation coefficients of -0.21 and -0.25, respectively. HHV, in its relationship with the process parameters, displays significant positive correlations with C (%), VM (%), and FC (%), yielding correlation coefficients of 0.31, 0.45, and 0.79, respectively. Conversely, it showcases substantial negative correlations with N (%), Ash (%), and RT (min), featuring correlation coefficients of -0.39, -0.43, and -0.33, respectively.

Despite these correlations, the scatter plots in Figure 4.2 lack a discernible trend, suggesting that an exhaustive experimental analysis of various process parameter combinations may be necessary. This underscores the pivotal role of advanced machine learning algorithms in constructing predictive models capable of quantifying yield as a function of these multifaceted process parameters.



**Figure 4.2 Correlation between the various process parameters and Target Variable**

**4.2 Comparison of Various ML Predictive Models**

The evaluation and comparison of machine learning predictive models will be executed independently for each of the three distinct segments. This assessment will be conducted through the utilization of regression plots (regplots), graphically illustrating the relationships between actual values and predicted values. Moreover, a comprehensive model evaluation will be performed, incorporating essential metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R-squared, R2). The model selection process will be based on the minimization of prediction errors (MAE, MSE, and RMSE) and the maximization of the goodness-of-fit (R2). The chosen model, exhibiting the least error and the highest R2 value, will be subject to hyperparameter tuning to further enhance predictive accuracy. Hyperparameter optimization is a crucial step in refining the model's performance.

Subsequently, after achieving the final optimized model, an analysis of feature importance will be conducted. This analysis will be represented in the form of a bar graph, showcasing the relative significance of each feature in the prediction process. Identifying feature importance is instrumental in understanding the variables that have the most substantial influence on the predictive model's outcomes, providing valuable insights for model interpretability and feature selection.

**4.2.1 Prediction of Yield (%)**

Based on the training dataset, predictive models have been constructed to estimate the yield percentage (%), considering ultimate analysis, proximate analysis, and process parameters. In Figure 4.3, the performance of these predictive models is evaluated using a test dataset that relies on proximate analysis and process parameters. Remarkably, the Support Vector Regressor exhibited the least favourable predictive performance, as indicated by its predicted line. In contrast, the Gradient Boosting and Random Forest models emerged as the most effective models within this dataset evaluation. Their predictive lines closely resemble the ideal scenario where the predicted values match the actual values (Y=X), highlighting their superior performance and predictive accuracy.

This outcome underscores the significance of model selection, as the choice of an appropriate algorithm can significantly impact the quality of predictions. In this context, Gradient Boosting and Random Forest models have demonstrated their potential to provide precise and reliable yield predictions in relation to the specified parameters, offering promising avenues for further analysis and application.

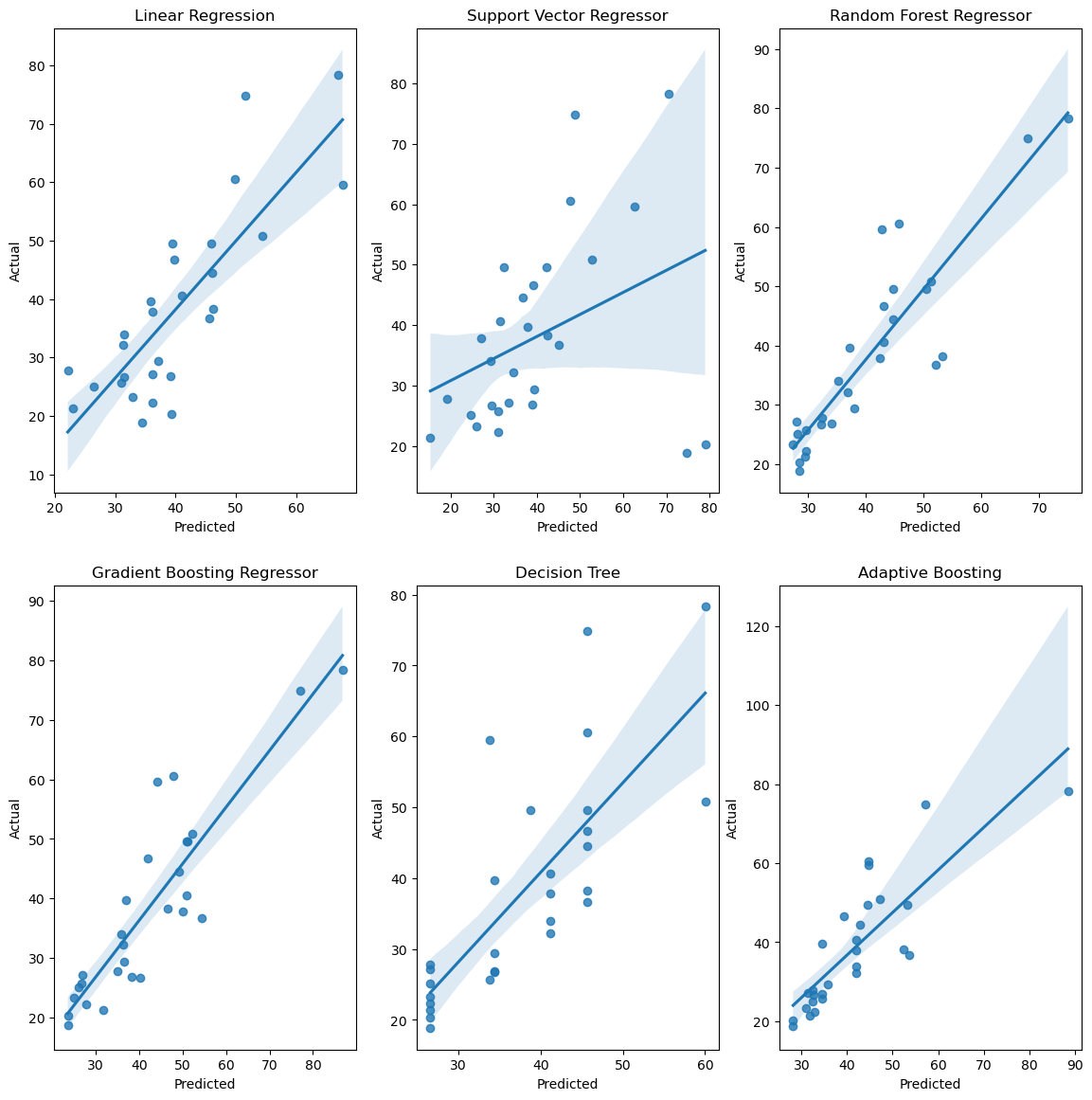
**Figure 4.3 Regression plot for Yield (%) using proximate and process parameters**

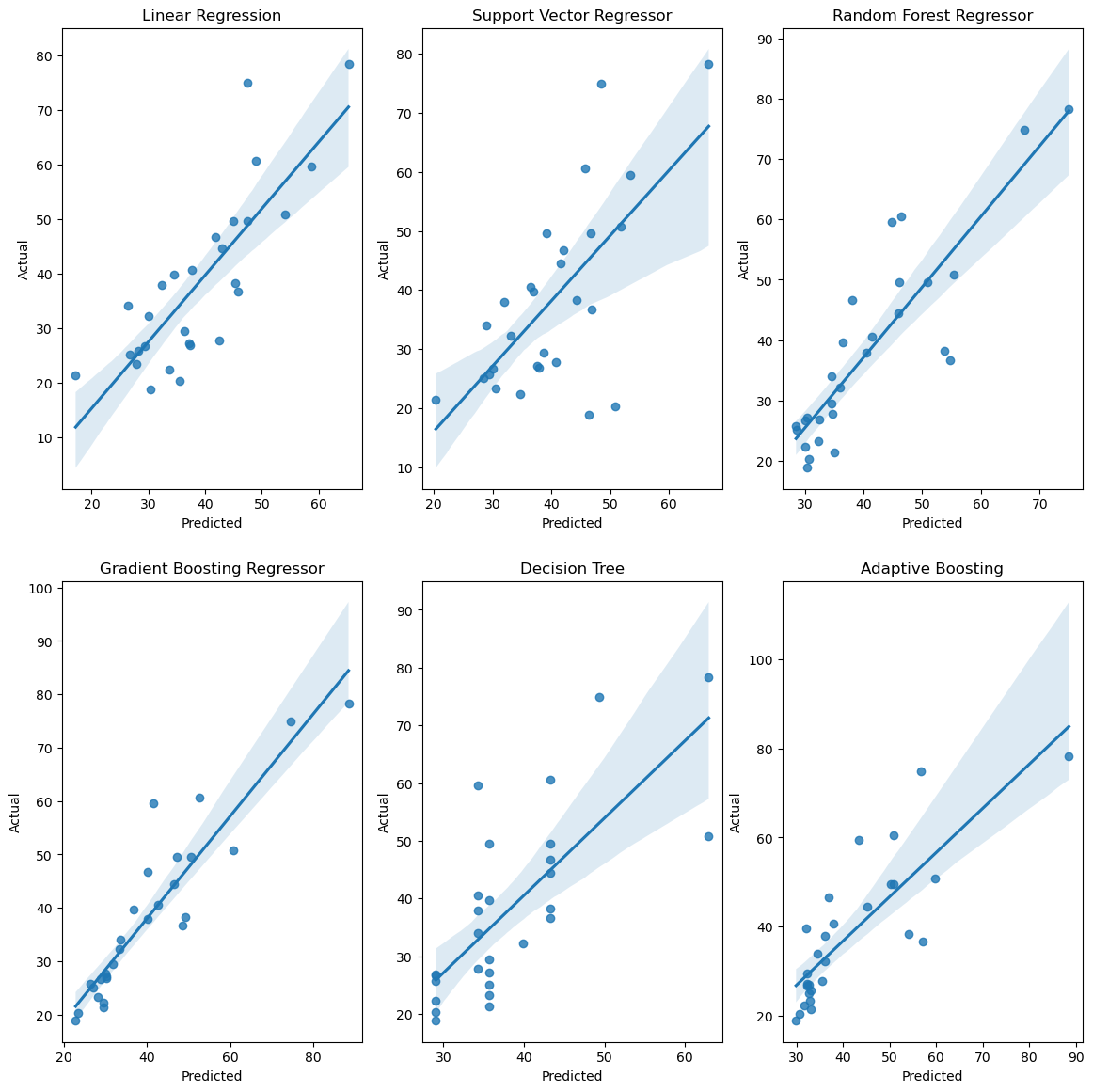
Table 4.1 provides a comprehensive overview of model performance as assessed through various evaluation metrics. Notably, the Support Vector Regressor stands out with the highest error metrics, boasting a Mean Absolute Error (MAE) of 11.09 and a notably negative R-squared value (R2) of -0.28 in comparison to all other models. Conversely, the Gradient Boosting and Random Forest models shine as the top-performing options within this dataset, characterized by their exceptional predictive accuracy. The MAE for Gradient Boosting is reported at 6.33, along with an R2 of 0.74, demonstrating its robust performance. The Random Forest model achieves an even lower MAE of 6.01 and an impressive R2 of 0.76.

This analysis conclusively establishes that when employing proximate analysis and process parameters, the model offering the least prediction error, with an MAE of 6.01 and the highest R2 value of 0.76, is the Random Forest model. These results underscore the efficacy of this model in providing accurate yield predictions based on the specified input parameters.

**Table 4.1. Error value for yield (%) prediction using proximate and process parameters**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Linear Regression | SVR | Random Forest | Gradient Boosting | Decision Tree | AdaBoost |
| MAE | 7.56 | 11.09 | 6.01 | 6.33 | 7.65 | 7.5 |
| MSE | 88.58 | 313.49 | 57.57 | 63.73 | 105.8 | 78.69 |
| RMSE | 9.41 | 17.71 | 7.59 | 7.98 | 10.29 | 8.87 |
| R2 | 0.64 | -0.28 | 0.76 | 0.74 | 0.57 | 0.68 |

In Figure 4.4, the performance of machine learning models on a test dataset, utilizing ultimate analysis and process parameters, is presented. Notably, all models displayed commendable performance on this dataset. However, it's worth noting that the Support Vector Regressor, while not performing as well as the other models, exhibited superior predictive capabilities compared to its performance with proximate and process parameters. The standout performers in this evaluation were the Gradient Boosting and Random Forest models. These models excelled, with a significant number of data points closely aligning with the ideal Y=X line, reinforcing their effectiveness and accuracy in predicting the target variable within this specific dataset.



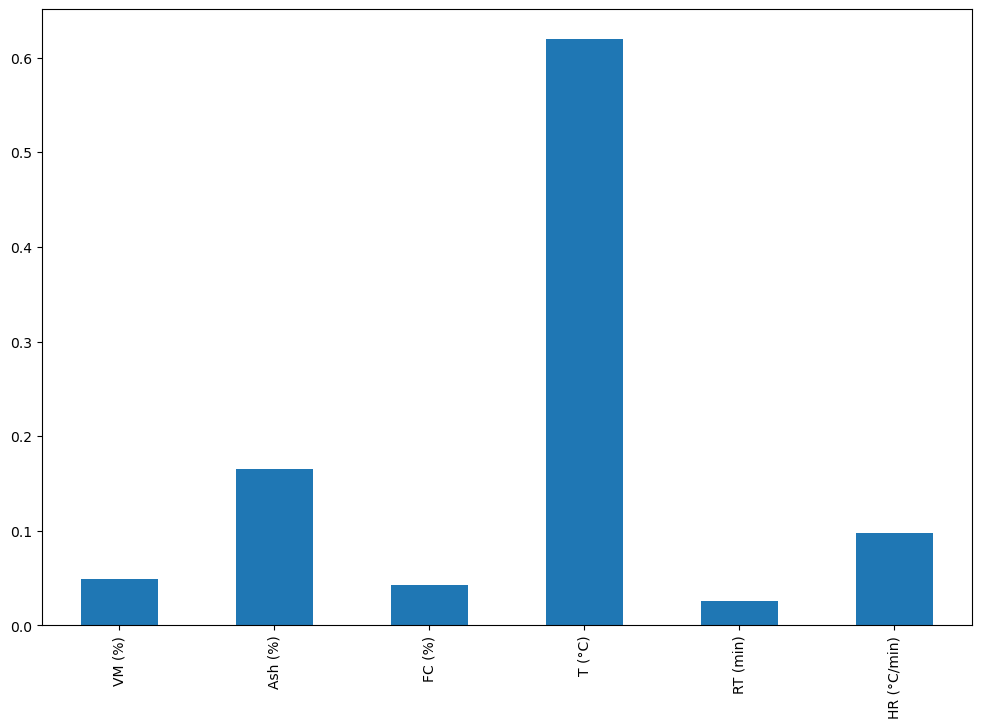
**Figure 4.4 Regression plot for Yield (%) using ultimate and process parameters**

Table 4.2 presents an extensive analysis of model performance utilizing evaluation metrics. The Support Vector Regressor stands out with the highest error metrics, featuring a Mean Absolute Error (MAE) of 8.89 and an R-squared (R2) value of 0.43 compared to all other models. However, it's noteworthy that its performance surpasses that observed with the proximate and process parameters. Conversely, the Gradient Boosting and Random Forest models demonstrated superior performance within this dataset. Specifically, the Gradient Boosting model achieved a remarkable MAE of 4.82 with an outstanding R2 of 0.83, indicating its high accuracy. The Random Forest model, while having a slightly higher MAE at 6.68, still displayed a commendable R2 of 0.72. These results highlight the exceptional predictive capabilities of the Gradient Boosting model, making it the preferred choice for this dataset when utilizing ultimate analysis and process parameters.

**Table 4.2. Error value for yield (%) prediction using ultimate and process parameters**

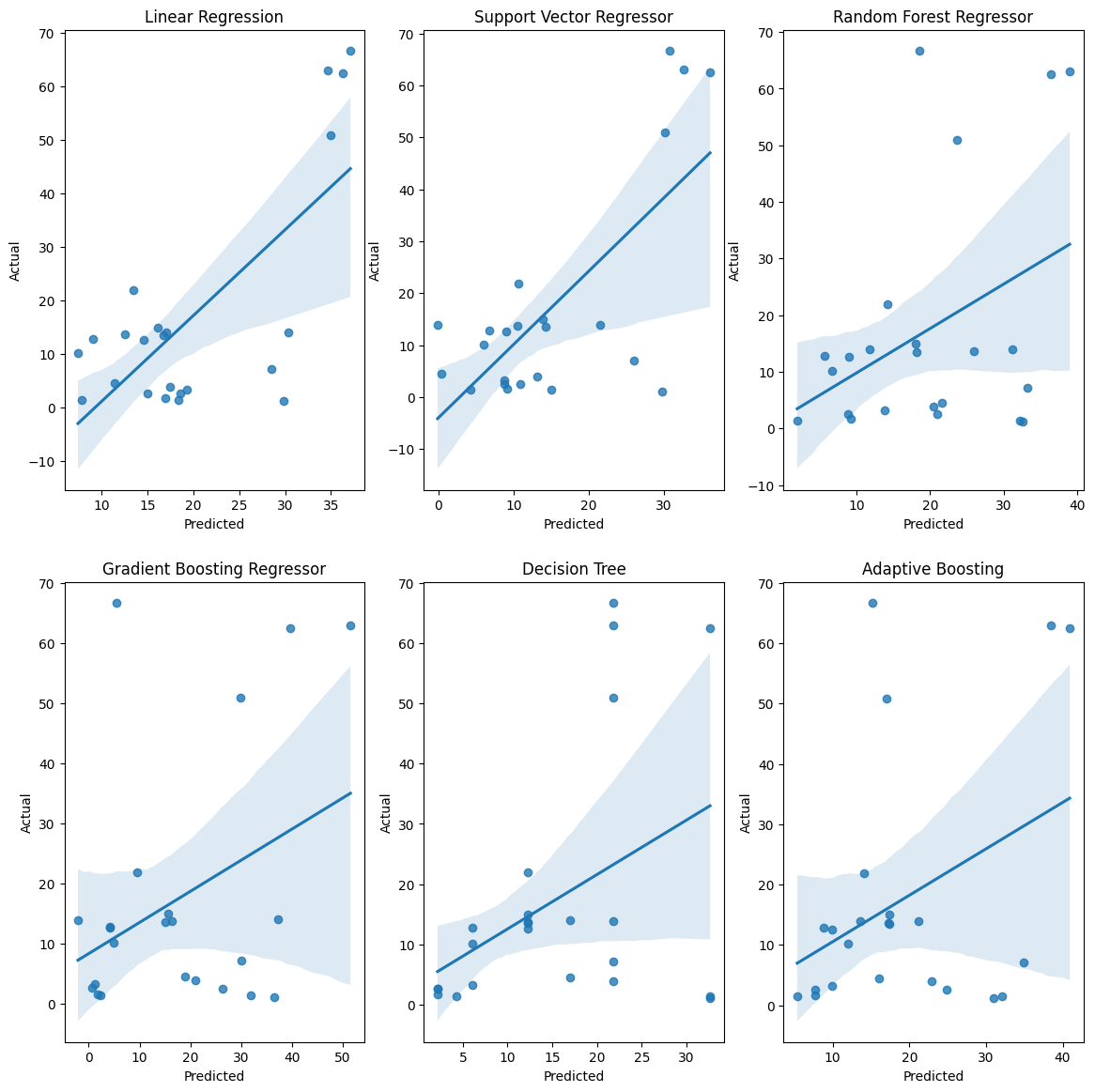
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Linear Regression | SVR | Random Forest | Gradient Boosting | Decision Tree | AdaBoost |
| MAE | 7.31 | 8.89 | 6.68 | 4.82 | 8.8 | 8.32 |
| MSE | 85.83 | 139.02 | 68.5 | 41.36 | 117.54 | 96.24 |
| RMSE | 9.26 | 11.79 | 8.28 | 6.43 | 10.84 | 9.81 |
| R2 | 0.65 | 0.43 | 0.72 | 0.83 | 0.52 | 0.61 |

Following a comprehensive model comparison for both segments, the Gradient Boosting model emerged as the superior choice when utilizing ultimate analysis and process parameters. Subsequent hyperparameter tuning further improved its performance, resulting in a reduced error, with a Mean Absolute Error (MAE) of 4.24 and an impressive R-squared (R2) value of 0.85. In Figure 4.5, the analysis of feature importance for yield (%) prediction is depicted. Notably, T (°C) was identified as the most critical feature for accurate predictions, underscoring its pivotal role in the predictive model's performance.

**Figure 4.5 Feature Importance plot for Yield (%) prediction**

**4.2.2 Prediction of SSA**

Utilizing the training dataset, predictive models were developed to estimate specific surface area (SSA), considering ultimate analysis, proximate analysis, and process parameters. In Figure 4.6, the performance of these models was assessed using a test dataset based on proximate analysis and process parameters. Notably, the Random Forest models and Adaptive Boosting models exhibited suboptimal predictive performance, as indicated by their respective predicted lines. Surprisingly, the Support Vector Regressor emerged as the most effective model during this dataset evaluation, with its predicted line closely resembling the ideal scenario where predicted values align with actual values (Y=X). This underscores the crucial importance of selecting the right algorithm, as it significantly influences prediction quality. In this context, the Support Vector Regressor models have demonstrated their potential to offer precise and reliable SSA predictions based on the specified parameters, indicating promising prospects for further analysis and real-world application.



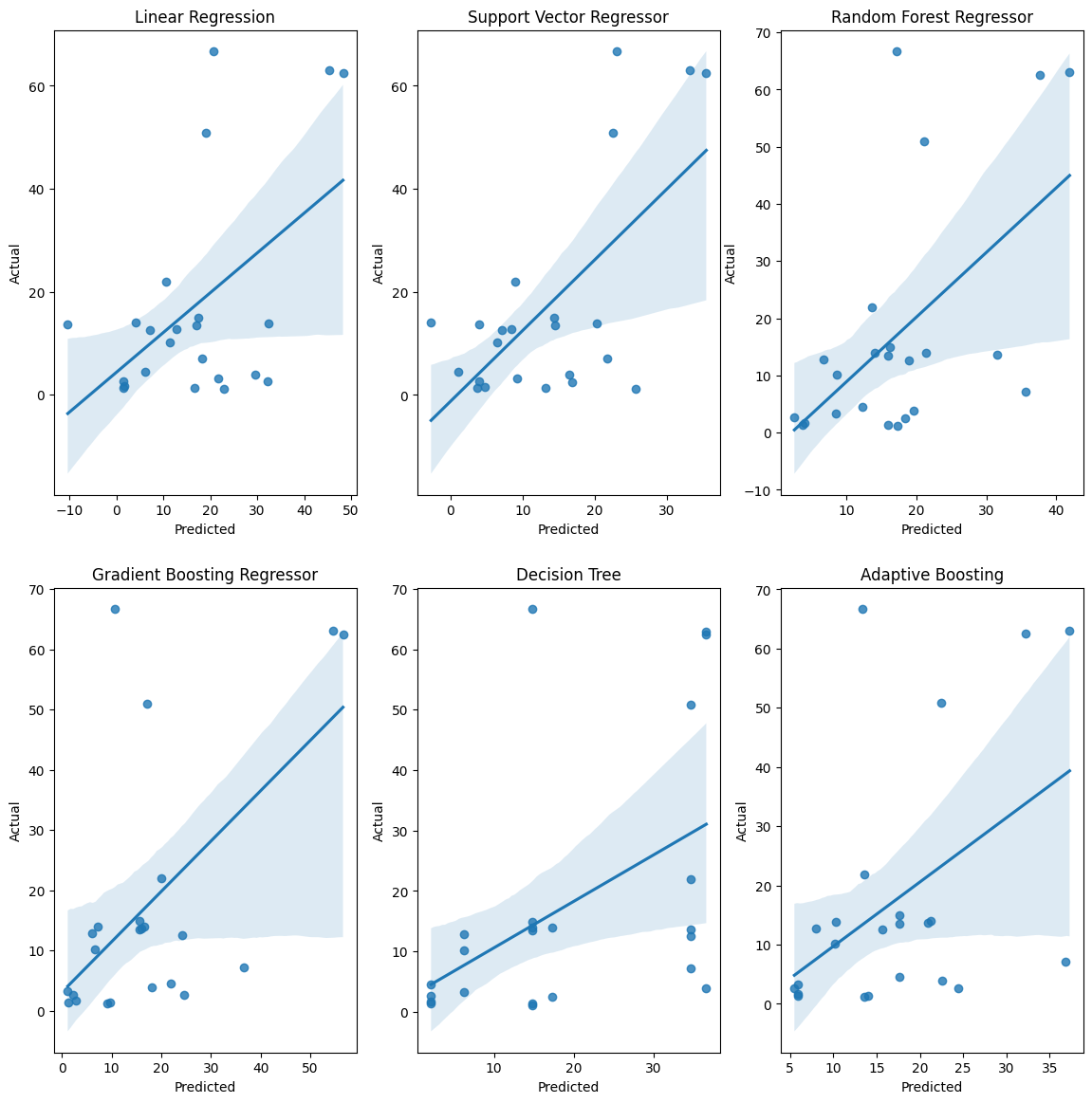
**Figure 4.6 Regression plot for SSA using proximate and process parameters**

Table 4.3 furnishes a comprehensive depiction of model performance, evaluated through a spectrum of pertinent metrics. Strikingly, the Random Forest models and Adaptive Boosting models surface with the most glaringly adverse error metrics. The Random Forest model exhibits a towering Mean Absolute Error (MAE) of 17.3 and a conspicuously negative R-squared (R2) value of 0.31, while the Adaptive Boosting model features a prominent MAE of 11.75 and a notably negative R2 of 0.54 in contrast to other models. Conversely, the Support Vector Regressor model emerges as the standout performer within this dataset, characterized by its exceptional predictive accuracy. Notably, it attains a significantly lower MAE of 11.74 and achieves a relatively positive R2 of 0.74, underscoring its robust performance. This thorough analysis unequivocally establishes that when employing proximate analysis and process parameters, the Support Vector Regressor model offers the least prediction error with an MAE of 11.74 and the highest R2 value of 0.74. These results accentuate the model's effectiveness in furnishing accurate SSA predictions grounded in the specified input parameters.

**Table 4.3. Error value for SSA prediction using proximate and process parameters**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Linear Regression | SVR | Random Forest | Gradient Boosting | Decision Tree | AdaBoost |
| MAE | 12.84 | 11.74 | 17.3 | 15.16 | 12.91 | 11.75 |
| MSE | 289.72 | 208.05 | 376.42 | 434.57 | 363.3 | 318.91 |
| RMSE | 17.02 | 14.43 | 19.4 | 20.85 | 19.06 | 17.85 |
| R2 | 0.42 | 0.74 | 0.31 | 0.40 | 0.46 | 0.54 |

In Figure 4.7, the performance of machine learning models on a test dataset, employing ultimate analysis and process parameters, is illustrated. Impressively, all models showcased commendable performance in this dataset. It is important to note that while the Decision Tree and Adaptive Boosting models did not perform as strongly as some other models, they exhibited improved predictive capabilities compared to their performance with proximate and process parameters. The standout performers in this evaluation were the Support Vector Regressor models. These models excelled, with a substantial number of data points closely aligning with the ideal line, underscoring their effectiveness and precision in predicting the target variable within this specific dataset.



**Figure 4.7 Regression plot for SSA using ultimate and process parameters**

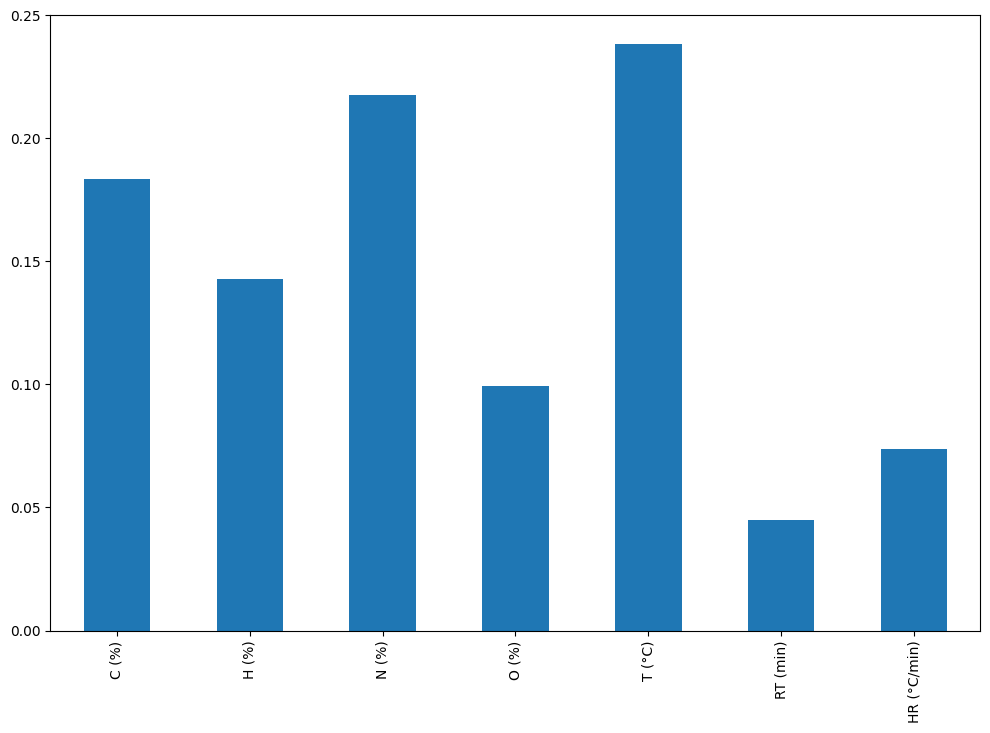
Table 4.4 offers a comprehensive analysis of model performance, assessed through various evaluation metrics. Notably, the Decision Tree and Adaptive Boosting models stand out with the highest error metrics, featuring a considerable Mean Absolute Error (MAE) of 52.74 and an R-squared (R2) value of –1.52 for the Decision Tree, and an MAE of 50.88 with an R2 of –1.0 for the Adaptive Boosting model in comparison to other models. However, it's essential to recognize that their performance outperforms what was observed with proximate and process parameters. Conversely, the Support Vector Regressor models exhibited superior performance within this dataset. Specifically, the model achieved an impressive MAE of 26.67 with a remarkable R2 of 0.13, underscoring its high accuracy and predictive capabilities. These outcomes underscore the exceptional performance of the Support Vector Regressor model, positioning it as the preferred choice for this dataset when utilizing ultimate analysis and process parameters.

**Table 4.4. Error value for SSA prediction using ultimate and process parameters**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Linear Regression | SVR | Random Forest | Gradient Boosting | Decision Tree | AdaBoost |
| MAE | 13.53 | 12.33 | 12.4 | 10.49 | 13.09 | 10.98 |
| MSE | 328.65 | 276.39 | 296.06 | 274.99 | 343.99 | 270.07 |
| RMSE | 18.13 | 16.63 | 17.21 | 16.58 | 18.55 | 16.43 |
| R2 | 0.40 | 0.68 | 0.32 | 0.36 | 0.21 | 0.38 |

After conducting a thorough model comparison for both data segments, the Support Vector Regressor model emerged as the optimal choice when employing proximate analysis and process parameters. Subsequent fine-tuning of its hyperparameters further enhanced its performance, resulting in a notable reduction in prediction errors. The Support Vector Regressor model achieved an impressive Mean Absolute Error (MAE) of 10.42, coupled with an outstanding R-squared (R2) value of 0.81, highlighting its superior accuracy in predicting the target variable.

In Figure 4.8, an analysis of feature importance for predicting yield (%) is presented. Notably, two features, T (°C) and N (%), were identified as the most critical contributors to accurate predictions. This underscores the pivotal role played by these features in driving the overall performance of the predictive model, emphasizing their significance in achieving precise yield predictions.

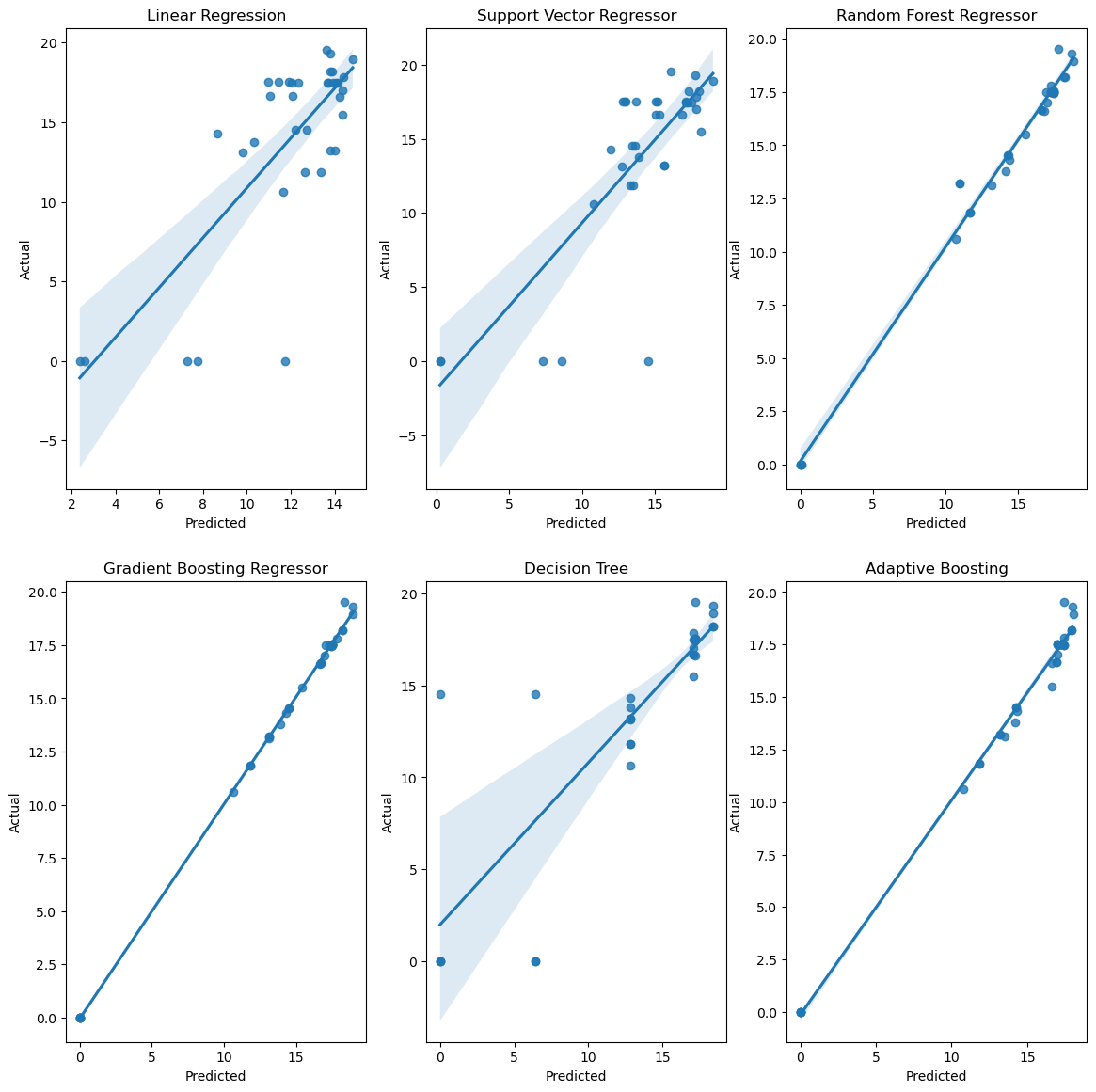


**Figure 4.8 Feature Importance plot for SSA prediction**

**4.2.3 Prediction of HHV**

Using the training dataset, predictive models were constructed to estimate High Heating Value (HHV), taking into account ultimate analysis, proximate analysis, and process parameters. In Figure 4.9, the performance of these models was evaluated using a test dataset that relied on proximate analysis and process parameters. Notably, the Linear Regression model displayed suboptimal predictive performance, as evidenced by its predicted lines.

Surprisingly, the Random Forest and Adaptive Boosting models emerged as the most effective options during this dataset evaluation, with their predicted lines closely resembling the ideal scenario where predicted values align perfectly with actual values (Y=X). This highlights the pivotal role of algorithm selection, which significantly impacts the quality of predictions. In this context, the Adaptive Boosting models have proven their potential to provide precise and reliable SSA predictions based on the specified parameters. This finding suggests promising avenues for further analysis and practical applications, underscoring the importance of choosing the right model for the task at hand.



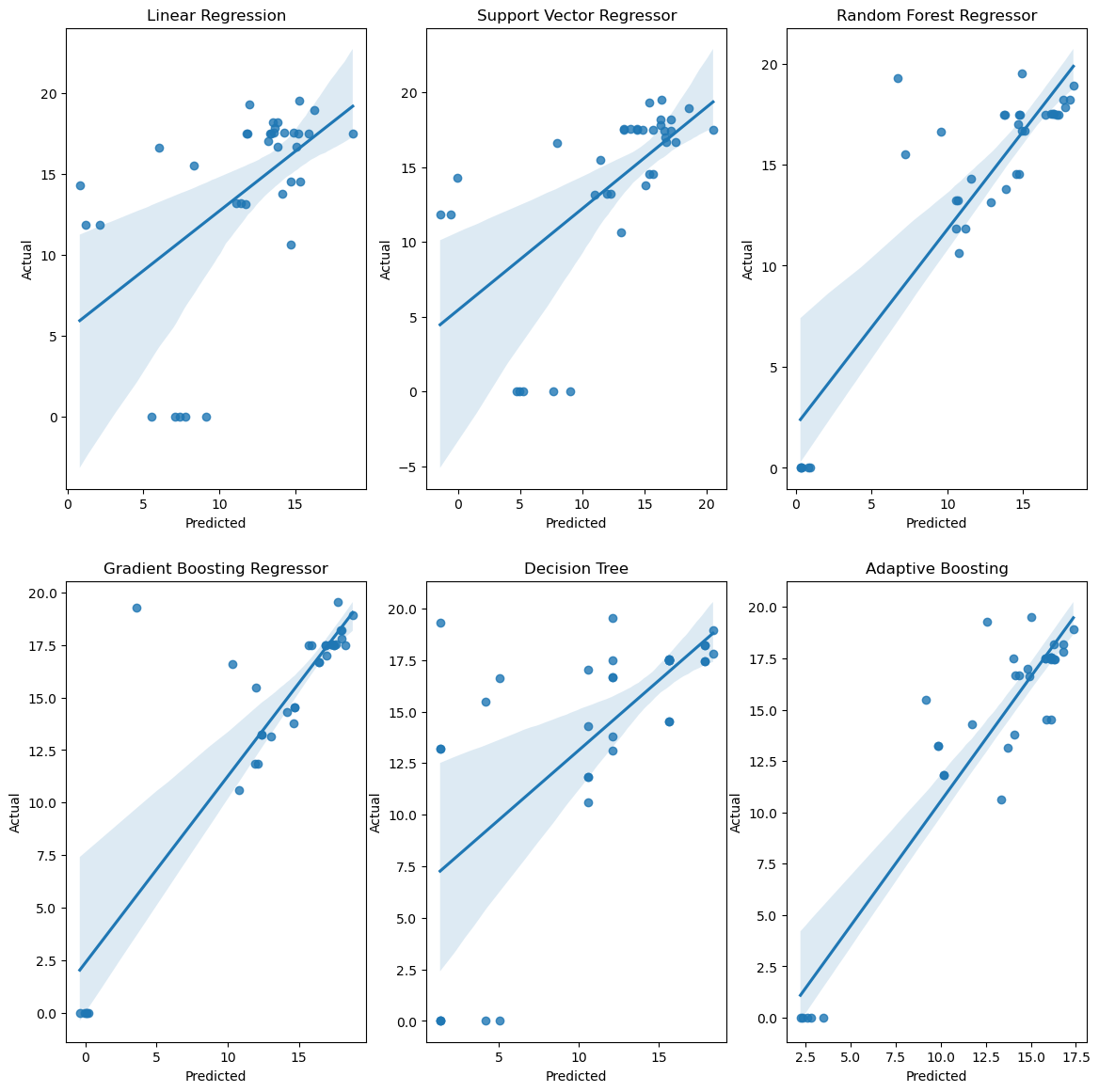
**Figure 4.9 Regression plot for HHV using Proximate and process parameters**

Table 4.5 provides a comprehensive overview of model performance, assessed through a range of key metrics. The most striking observation is that the Linear Regression model exhibits notably unfavourable error metrics. Specifically, it displays a substantial Mean Absolute Error (MAE) of 3.99 and a conspicuously negative R-squared (R2) value of 0.41 compared to the other models. Conversely, the Adaptive Boosting model emerges as the standout performer within this dataset, marked by its exceptional predictive accuracy. Impressively, it achieves a substantially lower MAE of 0.29 and a notably positive R2 of 0.92, underscoring its robust and highly accurate performance. This comprehensive analysis unequivocally establishes that when utilizing proximate analysis and process parameters, the Gradient Boosting model excels by offering the least prediction error. These findings emphasize the model's effectiveness in providing accurate predictions for HHV, based on the specified input parameters.

**Table 4.5. Error value for SSA prediction using proximate and process parameters**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Linear Regression | SVR | Random Forest | Gradient Boosting | Decision Tree | AdaBoost |
| MAE | 3.99 | 2.22 | 0.3 | 0.08 | 1.51 | 0.29 |
| MSE | 21.15 | 13.35 | 0.41 | 0.05 | 10.57 | 0.25 |
| RMSE | 4.6 | 3.65 | 0.64 | 0.22 | 3.25 | 0.5 |
| R2 | 0.41 | 0.63 | 0.91 | 0.95 | 0.7 | 0.92 |

Figure 4.10 presents the performance of machine learning models on a test dataset, leveraging ultimate analysis and process parameters. Impressively, all models demonstrated commendable performance in this dataset. Notably, while Linear Regression didn't perform as strongly as some other models, it doesn’t exhibit improved predictive capabilities compared to its performance with proximate and process parameters. The standout performers in this evaluation were the Random Forest and Gradient Boosting models. These models excelled, with a substantial number of data points closely aligning with the ideal line, underscoring their effectiveness and precision in predicting the target variable within this specific dataset. This emphasizes the critical role of model selection, as these models proved their worth by delivering highly accurate predictions. Ultimately, the dataset, incorporating ultimate analysis and process parameters, benefited significantly from the Random Forest and Gradient Boosting models, which offer promising avenues for advanced analysis and real-world applications.



**Figure 4.10 Regression plot for HHV using ultimate and process parameters**

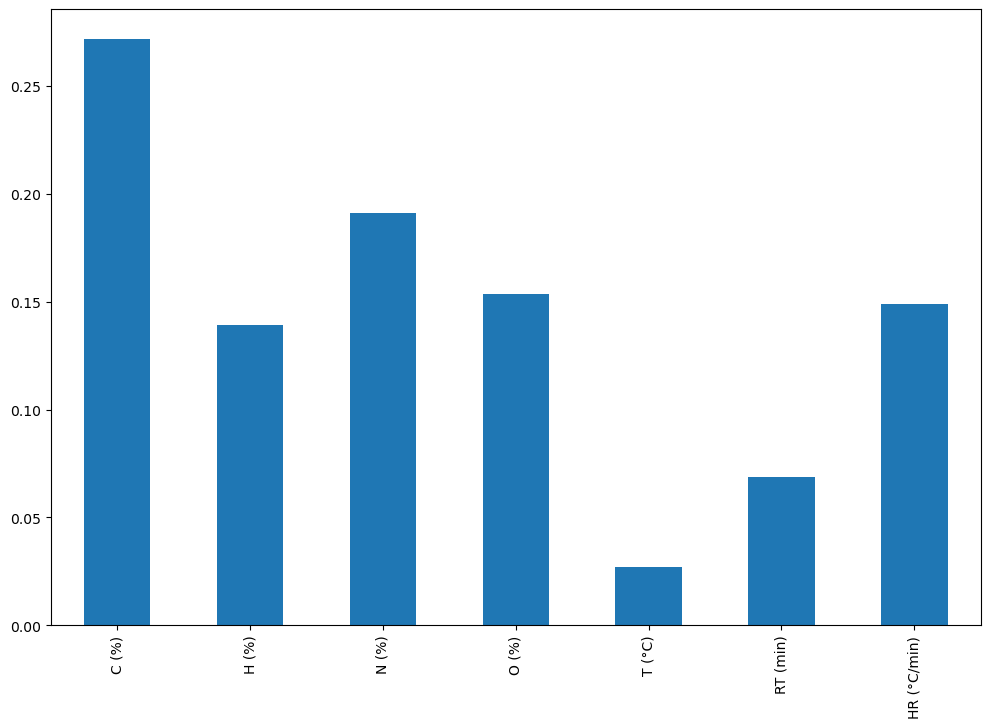
Table 4.6 provides a comprehensive analysis of model performance, assessed through various evaluation metrics. Notably, the Linear Regression model stands out with the highest error metrics, featuring a considerable Mean Absolute Error (MAE) of 4.71 and an R-squared (R2) value of 0.1, when compared to the other models. However, it's crucial to acknowledge that its performance surpasses what was observed with proximate and process parameters.

Conversely, the Gradient Boosting models demonstrated superior performance within this dataset. Specifically, the model achieved an impressive MAE of 1.06 with a remarkable R2 of 0.77, underscoring its high accuracy and predictive capabilities. These results underscore the exceptional performance of the Gradient Boosting model, positioning it as the preferred choice for this dataset when utilizing ultimate analysis and process parameters. This signifies the model's effectiveness in providing accurate predictions, highlighting its potential for advanced analysis and practical applications.

**Table 4.6. Error value for HHV prediction using ultimate and process parameters**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Linear Regression | SVR | Random Forest | Gradient Boosting | Decision Tree | AdaBoost |
| MAE | 4.71 | 3.74 | 1.92 | 1.06 | 3.7 | 1.98 |
| MSE | 32.26 | 27.21 | 10.56 | 8.38 | 31.5 | 9.0 |
| RMSE | 5.68 | 5.22 | 3.25 | 2.9 | 5.61 | 3.0 |
| R2 | 0.1 | 0.24 | 0.71 | 0.77 | 0.12 | 0.75 |

Following an extensive model comparison for both data segments, the Gradient Boosting model emerged as the optimal choice when utilizing proximate analysis and process parameters. The Gradient Boosting model achieved exceptional results, with a strikingly low Mean Absolute Error (MAE) of 0.26 and an impressive R-squared (R2) value of 0.96, underscoring its outstanding accuracy in predicting the target variable. Figure 4.11 delves into the analysis of feature importance for predicting HHV. Notably, two key features, C (%), N (%), and HR (°C/min), were identified as the most critical contributors to accurate predictions. This underscores the pivotal role played by these features in driving the overall performance of the predictive model, underscoring their significance in achieving highly precise yield predictions. These findings highlight the potential of the Gradient Boosting model and the pivotal role of these specific features in enhancing the accuracy and reliability of HHV predictions, offering promising prospects for further analysis and practical application.



**Figure 4.11 Feature Importance plot for HHV prediction**

**Chapter 5**

## **CONCLUSION**

**5.1 Conclusion**

Developing predictive models to quantify biochar production based on process parameters represents a practical and impactful pursuit in the realm of sustainable development. This research paper illustrates this endeavor by creating models for biochar yield, specific surface area (SSA), and higher heating value (HHV) as functions of proximate analysis parameters, ultimate analysis parameters, and process parameters. The study encompasses a comprehensive comparative analysis to evaluate the effectiveness of six distinct machine learning techniques: linear regression, Support Vector Regressor, Decision Tree, Random Forest Regression, AdaBoost Regression, and Gradient Boosting Regressor in the context of biochar production.

A wide array of accuracy and error metrics is employed to gauge the performance of these machine learning algorithms. The findings reveal important insights. When predicting yield (%), the Gradient Boosting Regression approach achieved only moderate accuracy when utilizing ultimate analysis and process parameters. For SSA prediction, the Support Vector Regressor, incorporating ultimate analysis and process parameters, demonstrated the highest accuracy. In the case of HHV, the Gradient Boosting model, combined with proximate parameters and process parameters, achieved the least prediction error.

These outcomes hold significant implications for biochar production. By implementing these predictive models, significant time and effort can be saved in the pursuit of identifying optimal process parameters to enhance yield. This approach aligns with the broader goals of sustainable development, contributing to more efficient and eco-friendly biochar production processes. The study underscores the value of advanced machine learning techniques in addressing crucial challenges in sustainable development and resource optimization.

## **REFERENCES**

**Journal Paper**

1. A.J. Ragauskas, C.K. Williams, B.H. Davison, G. Britovsek, J. Cairney, C.A. Eckert, et al.(2006), “The path forward for biofuels and biomaterials”
2. D. Tilman, R. Socolow, J.A. Foley, J. Hill, E. Larson, L. Lynd, et al.(2009) “Beneficial biofuels–the food, energy, and environment trilemma”
3. M.E. Himmel, S.Y. Ding, D.K. Johnson, W.S. Adney, M.R. Nimlos, J.W. Brady, et al.(2007) “Biomass recalcitrance: engineering plants and enzymes for biofuels production”
4. A.V. Bridgwater, G.V.C. Peacocke(2000) “Fast pyrolysis processes for biomass”
5. A.A. Peterson, F. Vogel, R.P. Lachance, M. Froling, M.J. Antal Jr., J.W. Tester(2008) “Thermochemical biofuel production in hydrothermal media: a review of sub- and supercritical water technologies”
6. M. Sekar, T. Mathimani, A. Alagumalai, N.T.L. Chi, P.A. Duc, S.K. Bhatia, et al.(2021)“A review on the pyrolysis of algal biomass for biochar and bio-oil - Bottlenecks and scope”
7. Hemal Chowdhury, Tamal Chowdhury, Pranta Barua(2021), “Biofuel production from food waste biomass and application of machine learning for process management”
8. J.C. Solarte-Toro, J.A. González-Aguirre, J.A. Poveda Giraldo, C.A. Cardona Alzate(2021) “Thermochemical processing of woody biomass: A review focused on energy-driven applications and catalytic upgrading”
9. M. Bartoli, L. Rosi, P. Frediani, M. Frediani(2020) “Bio-oils from microwave assisted pyrolysis of kraft lignin operating at reduced residual pressure”
10. Q. Tang, Y. Chen, H. Yang, M. Liu, H. Xiao, Z. Wu, et al. (2020)“Prediction of Bio-oil Yield and Hydrogen Contents Based on Machine Learning Method: Effect of Biomass Compositions and Pyrolysis Conditions”
11. Bisheswar Karmakar, Sumit H.Dhawane, Gopinath Halder (2018), “Optimization of biodiesel production from castor oil by Taguchi design” ScienceDirect S2213343718301957
12. Monzur Alam Imteaz, A.B.M. Sharif Hossain, Maryam Bayatvarkeshi (2022), “A mathematical modelling framework for quantifying production of biofuel from waste banana”
13. Min-Yuan Cheng, Doddy Prayogo(2012), “Optimizing Biodiesel Production from Rice Bran Using Artificial Intelligence Approaches”
14. Rouf Ahmad Dar, Mudasir Yaqoob, Manisha Parmar, Urmila Gupta Phutela(2019), “Biofuels from Food Processing Wastes”
15. Navdish Arora(2021), “Biofuels From Food Waste – A Sustainable Approach”
16. Kambiz Tahvildari, Yasaman Naghavi Anaraki, Reza Fazaeli, Sogol Mirpanji, Elham Delrish(2015), “The study of CaO and MgO heterogenic nano-catalyst coupling on transesterification reaction efficacy in the production of biodiesel from recycled cooking oil”
17. Leung, D, Guo, Y(2006), “Transesterification of neat and used frying oil: Optimization for biodiesel production”
18. Lertsathapornsuk, Pairintra, Aryusuk, Krisnangkura(2008), “Microwave assisted in continuous biodiesel production fromwaste frying palm oil and its performance in a 100 kW diesel generator”
19. Karmakar, B.; Dhawane, S.H.; Halder, G(2018), “Optimization of biodiesel production from castor oil”
20. Kumar, R.; Tiwari, P.; Garg, S.(2013), “Alkali transesterification of linseed oil for biodiesel production”
21. Ahmad, T.; Danish, M.; Kale, P.; Geremew, B.; Adeloju, S.B.; Nizami, M.; Ayoub, M. (2020), “Conversion of flaxseed oil into biodiesel using KOH catalyst: Optimization and characterization dataset”.
22. Hazra, A.; Gogtay, N(2016), “Correlation and linear regression”.
23. Seo, D.K.; Kim, Y.H.; Eo, Y.D.; Park, W.Y.; Park, H.C.(2017), “Generation of Radiometric, Phenological Normalized Image Based on Random Forest Regression for Change Detection”.

**Website References**

Jupyter Notebook https://github.com/rutvikkumar2108/MTP