UNIVERSITY OF MINES AND TECHNOLOGY (UMAT), TARKWA

SCHOOL OF POSTGRADUATE STUDIES
DEPARTMENT OF PETROLEUM AND NATURAL GAS ENGINEERING

A THESIS REPORT ENTITLED

PREDICTION OF HEATING VALUE OF NATURAL GAS FROM GHANA'S OIL FIELDS USING SUPERVISED MACHINE LEARNING TECHNIQUES

BY
SAMPSON OWARE

SUBMITTED IN FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE IN PETROLEUM ENGINEERING

THESIS SUPERVISOR

……

ASSOC PROF ERIC BRONI BEDIAKO

TARKWA, GHANA

NOVEMBER 2023

DECLARATION

I declare that this project work is my own work. It is being submitted for the degree of Master of Science in Petroleum Engineering in the University of Mines and Technology (UMaT), Tarkwa. It has not been submitted for any degree or examination in any other University.

…….

(Signature of candidate)

06th day of November 2023

ABSTRACT

The heating value of natural gas is used to determine the quality of the gas sample, hence accurate prediction of heating value helps in controlling the issue of under billing and overbilling between a gas aggregator and an off-taker. The current method of determining heating value with gas chromatograph comes with many limitations as there can be carrier gas leaks, calibration gas issue and many more which can affect the prediction accuracy. The research focused on predicting the high heating value of natural gas based on percentage gas compositions obtained from Ghana's offshore oil fields using different algorithms with the aid of Colab Notebook Software and selecting the best performing model from the algorithms used. Seven Algorithms namely Artificial Neural Networks (ANN), AdaBoost, XGBoost, Linear Regression, Random Forest, Bagging Regressor and Stacking Regressor (Hybrid model) were modelled to determine the best predictive model using 2021 sample data on Gas specifications obtained from Ghana's offshore field, of which 80% of the data was used for training and the remaining 20% was used for testing. The performance of each algorithm was evaluated using metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), R^2 and Adjusted R^2. Random Forest model performed better than all the other predictive models with an R^2 and adjusted R^2 of 91.66% and 91.43% respectively and RMSE, MAE and MAPE of 1.6821, 0.5517 and 0.57% respectively during the testing stage. The hybrid model and the XGBoost Model equally did very well during the testing and can be relied on for the prediction of heating values of natural gas. The incorporation of this method provides a diverse approach to the analysis of the pipeline dynamic results of the heating value of natural gas.

This thesis is dedicated to my late father, Mr Emmanuel Appiah

ACKNOWLEDGEMENTS

I thank the Almighty God, the true source of all wisdom and understanding for His Grace, and sustenance throughout this thesis work and my entire educational journey. I also express my profound gratitude to my parents and siblings for providing me with unfailing support and encouragements throughout my years of study and through the process of writing this thesis work. You are truly priceless.

I would also like to thank my supervisor, Assoc Prof. Eric Broni-Bediako of the Petroleum and Natural Gas Engineering Department for his patience and fatherly, supervision and guidance throughout this project not forgetting all the lecturers of the department especially Dr Eric Thompson Brantson, Assoc Prof. Richard Amorin and Dr Solomon Asante-Okyere for their relentless efforts and contributions to the success of this thesis work.

Finally, I am truly grateful to all my course mates for their contributions in the form of ideas, motivation and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

Quality	Units	Symbol
Temperature (t)	degree Celsius	°C
Density	kilogramme per cubic metre	kg/m^3
Mass	gramme (0.001 kg)	g
Volume	cubic metre	m^3
Pressure	pascal	Pa

CHAPTER 1

INTRODUCTION

1.1 Background of Study

Natural gas is a multi-component fossil fuel that is created underneath the earth's surface. (Holloway, 2001). Methane (CH4), the largest component of natural gas constitutes a carbon (C) atom and four hydrogen (H) molecules (Baker and Lokhandwala, 2008). Natural gas is also known to contain insignificant quantities of natural gas liquids (NGL). Gases made up of hydrocarbon liquids also form this NGL. "Father of Natural Gas" William Hart dug the first natural gas well in Fredonia, United States, in 1821 (Kidnay and Parrish, 2006; Tronci *et al*., 2020). Initially, natural gases were locally adopted as the energy source for light but was subsequently transported for utilisation widely as a result of current advancement in engineering after World War II that saw to safety and allowed for long-coverage pipelines transportation of gas (Faramawy, Zaki, and Sakr, 2016).

Natural gas is considered to be colourless, odourless and amorphous and gives off valuable amount of energy when it undergoes combustion (Economides, 2009; Faramawy *et al*., 2016). The combustion of fossil fuels like coal or oil emits large quantities of harmful compounds like nitrous oxide, carbon dioxide and sulphur oxide. Comparatively, during the combustion of natural gas the emission of sulphur oxide is negligible as well as lower emission of nitrous oxide and carbon dioxide which helps reduce the problem of acid rains, greenhouse effects (Faramawy *et al*., 2016).

The world shift in energy preference from fossil fuels to natural gas is a result that, natural gas serves as a cleaner source of energy (Perera, 2018).

With reference to the BP statistical review of World Energy 2022 edition, global natural gas demand grew 5.3% in 2021, recovering above pre-pandemic 2019 levels and crossing the 4 trillion cubic meter mark for the first time. Its share in primary energy in 2021 was unchanged from the previous year at 24% (Dale, 2022). Also in Ghana, the production of natural gas has increased significantly since 2014 when full commercial production of natural gas started. From a production volume of 2 trillion Btu in 2014 to 107.83 trillion Btu in 2021 at an annual average

growth of 76.3% (Oscar, 2022). This statistic shows the increasing demand for natural gas as a source of fuel over the years. Natural gas has evolved from its primarily used as local energy for heat and electricity to a more robust used in residential, industrial and commercial heating globally dominating in the world economic growth (Mokhatab, Poe and Speight, 2006).

Natural gas is utilised as a petrochemical industry fuel and feedstock for organic chemical industry operations in the manufacture of ethylene and propylene (Siirola, 2010). Natural gas is also used in fertiliser industry to produce ammonia. Natural gas may also be used to produce gases like carbon black, syngas, carbon sulfide, hydrogen, and sulphur (Faramawy *et al.*, 2016; Siirola, 2010). In Ghana, natural gas is obtained from the Jubilee, TEN and Sankofa Fields. Ghana's natural gas is predominantly used for domestic power supply for industries, transports and cooking. This has increased natural gas consumptions exponentially over the decades. Over two decades, Ghana's natural gas consumption increased by 52.6%. This is the result of increase in industrial and residential demands for natural gas as their source of energy. The country has ten thermal power plants out of which two run solely on natural gas and five uses gas/oil. In an effort to ensure cleaner and progressive supply of energy, the country anticipates a shift from more environmental unfriendly fuels to a relative cheaper and cleaner natural gas-based fuel for its energy supply (Ayaburi and Bazilian, 2020). It appears that most of the research carried out in Ghana are concerned with the use and safety of natural gas.

By typical laboratory measurement with a bomb calorimeter, the heating value of a natural gas is calculated based on the mass rather than the volume consumed. The quantity of heat released during the combustion of one volume of gas is referred to as the heating value and is measured in btu/scf. Total, gross, and net calorific values are used to illustrate how effective a natural gas is at heating a space based on the amount of water present or consumed. While the existence of vapour as the liquid is referred to as net calorific value, the heating value of a natural gas is referred to as having gross calorific value when there is water present. However, net calorific value is considered more efficient in energy calculation as it shows features of real operational situation (Armstrong, 1966; Lett and Ruppel, 2004; Ludtke, 1986).

The quality of a natural gas material may be determined by considering its composition as well

as its heating value. The same materials that make up the gas are used in analytical laboratory processes to determine both the heating value and composition of these components. (Ringen, Lanum and Miknis, 1979).

1.2 Problem Statement

It is important to consider a natural gas's heating value while assessing its quality. Tools such as the Gas Chromatography, Moisture analysers, Gravitometers, Hydrogen Sulphur Monitor are used but the most widely used instrument is the Gas Chromatography (GC). In the petrochemical industry, GC is used to determine the quality of the gas which is dependent on the selling price of the gas. Also, GC is used to predict the quality of a gas at any strategic position down a pipeline (Ewing, 2001). With reference to Ghana's gas industries, most terminal stations along the gas pipeline network where custody transfer takes place are equipped with an online GC at the end of the pipeline close to the customer or the off-taker. The GC is incorporated with a flow computer for the estimation of the heating value and energy of the natural gas; therefore, an accurate estimation of the heating value solely depends on the proper functioning of the GC.

In unusual situations, the GC develops fault due to corona (partial discharge), thermal heating and arching. This results in error GC reading consequently resulting in wrong diagnosis of gas quality (composition) and pricing. There are also situations where the auxiliaries, such as gas carrier leaks or calibration gas running out, lead to incorrect analyses of the gas composition and therefore influence the natural gas's heating properties.

The proper maintenance of the GC is essential as this ensures the availability and reliability of the equipment. However, in most cases, it takes a longer time to get the GC fixed and running when it breaks down due to constraints like; delayed procurement and the delivery of parts as well as the availability of skilled personnel to fix the issue. The prediction of the heating value of natural gas using Machine Learning models will not only be used when the GC is out of service but will also serve as a check for the accuracy of the heating values provided by the GC. The data required for the Machine Learning models can be either from historical data usually

over a period of years (for the trend/ pattern) or better still real time data of the gas composition obtained from upstream.

According to the ISO 6976:2016 standard, the conventional method of calculating the heating values of natural gas uses a correction of pressure and temperature at the reference point and addresses the assessment of the uncertainties related to the heating value. The estimation of the uncertainties associated with the heating values makes it very time consuming and laborious and even does not usually promise an accurate estimation as errors can occur, as such more time friendly and less tedious approach must be adopted. This therefore requires a better and more relying method of predicting the heating value instead of depending on the historical data (data picked from a particular day and time when the GC was working) for billing. This study seeks to propose an alternative approach to predict the heating values of natural gas from different oil/gas fields in Ghana using machine learning models.

1.3 Relevance of the Study

In relation to the issues associated with Gas Chromatograph, many gas companies select the heating value from a specific date and then a specific time when the GC worked well to serve as reference for billing when it comes to custody transfer. Contrary to this, the application of Machine Learning models rather use the trend/pattern of the heating value obtained from a selected period of years for the prediction of the heating value hence providing a much more accurate value when there is an issue with the GC or the auxiliaries to control under billing or overbilling between the aggregator and off taker, and determine the actual quality of the natural gas which is the basis of this study.

1.4 Research Objectives

The objectives of this research are to:

i. Determine the heating value of a commingled gas; and
ii. Prediction of heating value of natural gas using supervised machine learning techniques.

1.5 Methods Used

The research methods used include:

i. Review of relevant literature;
ii. Gathering of data at Ghana Gas Company at Atuabo;
iii. Predict the heating value of natural gas from Ghana's offshore oilfields.

1.6 Organisation of Thesis

This Thesis contain five chapters. Chapter one is an introductory chapter to the thesis work. It states the issue the work is aimed at addressing. Background of the Study, Problem Statement, Relevance of the Study, Objective of the research, Methods Used and Organisation of Thesis are the main contents of this chapter. Chapter two presents precise and thorough literature review on topics pertaining to this research. The materials and procedures used to construct the correlation for calculating the heating values of natural gas from various oil and gas fields are explained in the third chapter. The results and debates from the methodologies used are provided in the fourth chapter, and the conclusions and suggestions derived from the research effort are covered in chapter five.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

A naturally occurring gas that is high in carbon and hydrogen is known as natural gas. Most naturally occurring gas is found in coal beds, natural gas fields, and oil fields (Faramawy *et al*., 2016). It is formed deep underneath the earth's crust and is a rich fossil energy source. Natural gases are utilised to provide heat and power for businesses and industries. Natural gas liquids (NGLs) and gases composed of non-hydrocarbons such carbon dioxide ($C0_2$) and water vapor are present in trace amounts. It is basically used as fuel and in the production of chemicals and materials.

Both sweet and sour gases can be used to describe natural gas. This is determined by the quantity of sulfur present in the gas component. Sweet gas contains traces or smaller amounts of hydrogen sulfide whereas sour gas has large amounts of sulfur-containing hydrogen. Sweet natural gas is less acidic, non-corrosive, and less difficult to handle, unlike sour gas. Gas sweetening is a procedure that removes the acid component of sour gas to transform it into sweet natural gas (Faramawy *et al*., 2016).

Dry and moist gas are additional categories for natural gas. The wetness or dryness of natural gas is dependent on the concentration of methane compounds. Methane makes up at least 85% of dry natural gas. In the creation of liquid natural gas (LNG) and compressed natural gas (CNG), dry natural gas is crucial (Weaver and Miller, 2019). On the other hand, wet natural gas has a lower methane content (less than 85 %) than water vapor and natural gas components such as propane, butane, and pentane as part of its components.

Wet natural gas is processed in pipelines to remove water vapor and LNG to ensure that they are safe for home consumption and transportation. However, wet natural gas can be used to produce plastics and other products as well as for outdoor grills (Welker, 2015).

2.2 Properties of Natural Gas

Gas is a homogeneous fluid with low density and less viscous property and irregular volume but can easily expand completely to take the shape of the container it is occupying. Understanding the links between pressure, volume, and temperature as well as the gas' other chemical and physical characteristics is essential for understanding how natural gas behaves. Natural gases are also known to be colorless, odorless, and have a weight lighter than air (Pellegrini *et al.*, 2019). The characteristics of natural gas include gas-specific density, gas molecular weight, gas viscosity, gas compressibility, and gas pressure and temperature.

2.2.1 Specific Gravity/Density

The mass of a unit volume of a gas divided by the mass of a unit volume of dry air at absolute temperature (273.15 K) and pressure (60 °F) is known as the gas specific density. Gas specific density is also termed as Gas specific density (Dembicki, 2017). Gas density is the ratio of the weight of the gas molecules to the weight of the air molecules in relation to ideal gas behaviour. Finding the average molecular weight of the gas's constituent parts yields the molecular weight of the gas (Dembicki, 2016). Equation 2.1 is the formula for calculating natural gas's molecular weight.

$$Ma = XiMi \tag{2.1}$$

In the formula, Mi stands for the components' molecular weights, Xi is for the gas component's mole fraction, i for the gas component, and n for the total number of compounds in the gas component (Guo, 2011). Once the Molecular Weight (Ma) is determined, the gas specific gravity can be determined by using Equation 2.2.

$$\gamma = \frac{\mathrm{Ma}}{\mathrm{M}air} \tag{2.2}$$

Where M is the apparent molecular weight of the gas, γ is the gas-specific density, and M*air* is the molecular weight of air (Mokhatab *et al.*, 2018). A molecular weight of 28.96 is assigned

to air. The gas's chemical makeup may be used to compute the molecular weight. (Debye, 1947). This is usually determined in the laboratories and the results are reported in mole fractions of the gas component.

2.2.2 Molecular Weight

The mass of one mole of an element or compound is what is used to determine a gas's molecular weight. It is also termed molecular mass. The molecular weight of a substance is measured in grams per mole ($g\ mol^{-1}$). By measuring the gas particle's mass and dividing it by Avogadro's number (6.02×10^{23}), the molar mass of the gas particle is determined (Kolb, 1978). The molecular weight of a gas mixture may be calculated using the relative ratio and molar mass of each individual gas component. By combining the mole fraction of the gas mixture and multiplying it by the molar mass of each component, the average molecular weight of the gas mixture is determined. Equation 2.3 is used to determine the molecular weight of a gas

$$M = \sum XiMi \qquad (2.3)$$

Xi is the mole fraction, Mi is the molar mass, and M is the molecular weight of each gas component (Alamooti and Malekabadi, 2018).

The same molar volumes but different molar masses are seen for gases measured at the same temperature and pressure at STP. The molecular weight of an ideal gas may also be calculated using the ideal gas law (PV=nRT). According to Lautier and Garai (2007) and Lower (2011), the formula for calculating the number of moles is n=m/M, where n is the mole of the compound, m is the mass of the gas compound, and M is the molecular mass.

2.2.3 Wobbe Index

The Wobbe Index of a gas is calculated as the product of the gas's gross heating value and its specific gravity, squared. It provides a measurement of the heat contained in a material through a certain hole at a specified gas pressure. The Wobbe index serves as an important parameter during the exchange of gasses in an engine (Klimstra, 1986). It is also regarded as a measure of changes between gases when they are used as fuels. The Wobbe index gauges the power

production of various gases during combustions. Wobbe index is vital in examining the effect of changeover in fuel and it is required conditions that accompany devices that transport gases (Mokhatab *et al.*, 2018). A higher heating value (HHV) and the gas-specific gravity are needed to determine the Wobbe index of gas. The combustion properties of a gas can be determined by the Wobbe index (Zachariah-Wolff *et al.*, 2007). Equation 2.4 is used to determine the Wobbe Index of a gas.

$$\text{Wobbe index} = \frac{\text{Calorific value}}{\sqrt{\text{Relative density}}} \tag{2.4}$$

2.2.4 Gas Viscosity

Viscosity is defined as the frictional forces within gas because of cohesion in fluids that impedes the flow of the gas substance. The mechanism of viscosity is crucial in designing a pipeline for the transport of materials (Dembicki Jr, 2017). A measurement of the resistance to fluid flow is called viscosity. Fluids have the ability that enables them to resist the flow of objects that are immersed in them and exhibit self-resistance to the movement of layers with varying velocities within them. The scientific unit (SI) of viscosity is Pascal (Pa). In practical and scientific publications Pascal is rarely used. The more frequent unit used to represent viscosity is the Poise (P) which is named after Jean Louis Poiseille who was a French Physiologist and it is expressed as dyne second per square centimeter (dyne s/c). Viscosity comes into play as a result of the transition of molecules from one layer of gas to another. During the process, there is a transfer of energy of molecules from a faster surface to a slower surface and vice versa (Dimri *et al.*, 2012). The viscosity (ŋ) of a gas is determined by Equation 2.5.

$$\text{Ŋ} = \frac{Shear\ stress}{Velocity\ gradient} \tag{2.5}$$

According to Newton's equation of motion, the fluid's shear is inversely related to its viscosity and directly proportional to the force applied (Towell, 2020). The temperature, pressure, and gas composition all affect the viscosity of natural gas, which may also be determined by laboratory testing. However, empirical data that is already accessible can be utilised to estimate the viscosity of the gas in the absence of laboratory data. Low pressure causes the viscosity to

rise with temperature because the gas molecules are being stirred up. On the other hand, a gas's viscosity reduces with high pressure and falling temperatures. When the pressure is intermediate, the gas's internal movement increases more when the temperature rises than when the pressure is low (Hanafy *et al.*, 1997; Vazquez and Beggs, 1977).

2.2.5 Gas Compressibility

A compressibility factor is necessary for many petrochemical engineering calculations, including those involving a recently found formula that is being tested, a drop in gas flow pressure through a pipe, a pressure gradient in gas wells, gas processing, and compression. Additionally, the compressibility factor is used into formulae to calculate the initial gas material balance. The Z-factor, commonly known as the gas compressibility factor, is calculated experimentally as a percentage of any typical PVT report. Mostly (Shokir *et al.*, 2012), data on gas composition is used to determine the Z-factor in cases when the PVT report is missing. Compressibility is defined as the quantity of volume that decreases when placed under pressure.

Practically, gases are more compressible than liquid and solids due to the presence of larger spaces between the gas particles. It is estimated that at standard pressure and temperature, the spaces between gas molecules are ten times apart. This analogue is the reason why gas particles are easily compressed. When gas molecules are pressed against each other vertically, the pressure exerted by the gas mount on each other through linear and nonlinear mechanisms (Gabitto and Tsouris, 2010). Gas compressibility (Z-factor) is the quantity of deviation from the ideal gas behaviour. Equation 2.6 is used to estimate the compressibility of a real gas in terms of density, pressure, temperature and molecular weight of the gas.

$$Z = \frac{PM}{\rho RT} \tag{2.6}$$

The variable Z is the gas compressible factor, P represents the pressure, ρ is the density, R is the constant gas pressure, M is the molecular weight of gas and the T is an absolute temperature (Winter, 2014).

2.2.6 Relative Density

The density of a gas substance is divided by the density of a reference item expressed in the same unit to get its relative density. Standard pressure (101.325 KPa) and room temperature (20 °C) are often used to assess density. Relative density has no unit. Relative density is used to quantify the weight of a sample that is required for the preparation a solution with a specified concentration. Relative density can be defined in terms of both real and ideal situations. When the gas and the air particles are viewed as gaseous and follow the ideal gas law, the ideal relative density of a gas is measured. However, when they are seen as real fluids, they are known to be real relative density. If a relative density of a substance is less than 1 then the substance can easily flow on water and vice versa. A relative density of 1 means they are the same as water (Riazi, 2005; Webb, 2001).

Equation 2.7 is used to estimate the relative density (RD) of a substance.

$$RD = \frac{\rho substance}{\rho reference} \tag{2.7}$$

Where RD is the relative density of the substance, $\rho_{substance}$ represents density of the substance, and $\rho_{reference}$ is the compared density of the substance (Picard *et al.*, 2008; Skempton, 1986).

2.3 Behaviour of Natural Gas

Natural gas is used to describe a hydrocarbon combination that is often present under the earth's crust. Natural gas exhibits a variety of behaviours as a result of its composition and origins. Temperature and pressure have an impact on how natural gas behaves.

2.3.1 Ideal Gas Law

The connection between temperature, pressure, and gas volume is expressed by the ideal gas law. The laws of Boyle's Law, Charles' Law, and Gay-Lussac's Law are all combined in these interactions. These relationships were obtained from the laws of gases established by Gay-Lussac, Charles, and Boyle. Volume and temperature are directly proportional at constant pressure, as demonstrated by Charles' law. Boyle's Law asserts that pressure and volume are

inversely proportional at constant temperature, whereas the Gay-Lussac Law states that temperature and pressure are directly proportional at constant volume (Laugier and Garai, 2007; Woody, 2013). The Ideal Gas Law is made up of these rules in combination. In the equation PV=NRT, P stands for pressure, V for volume, N for the number of moles of gas, R for the universal gas constant, and T for absolute temperature (Laugier and Garai, 2007).

When the force between individual atoms or molecules is fully elastic and there is no attractive force between the molecules, gas is said to be ideal. In order for a gas to be considered "ideal", it must meet all four requirements which include, its volume must be negligible, the particles within the gas must be of equal size, and must have a non-existing force of attraction or repulsion between/within the molecules, the gas molecules must be fully elastic with no energy loss, and the gas particles must move in accordance with an existent Newton's Law of Motion. (Tenny and Cooper, 2017). The concepts of "ideal" gases are strictly theoretical and do not exist in reality. This is because all existing natural gases violate the governing assumptions of natural gas principles. This is true because there is an existing volume in any gas participles (Tenny and Cooper, 2017). Also, gas particles exist in different sizes and there is an intermolecular force of attraction or repulsion between neighbouring gas particles. It is clear gas molecules move randomly, there is periodic preservation of energy and forces within the gas system hence a non-existing perfect elastic collision (Levine, 1985). The relationship between pressure and volume of a gas at constant temperature was established by Robert Boyle in 1662. The formula for Boyle's law is presented in Equation 2.8 and Equation 2.9.

$$p \propto \frac{1}{V} \tag{2.8}$$

, where P stands for pressure and V stands for volume. Boyle's law may be applied to determine a gas's initial pressure and volume, which are known as;

$$P1V1 = P2V2 \tag{2.9}$$

Another French scientist, Joseph Louis Gay-Lussac, gave Jacques Charles credit for his unpublished work in 1970. The volume of a gas is directly proportional to its temperature while the pressure is constant, according to that statement. Equations 2.10 and 2.11 were derived from

Charles Law.

This is stated as:

$$V \propto T \tag{2.10}$$

Using Charles Law to calculate the volume and temperature of a known initial volume and temperature of a gas, it is presented as:

$$\frac{V1}{T1} = \frac{V2}{T2} \tag{2.11}$$

An extension of the Charles Law was done by Joseph Louis Gay-Lussac by relating temperature and pressure. Gay-Lussac also proved that a gas's pressure and temperature are directly related at constant volume. The pressure and temperature of a gas are computed because this is represented as PT. Equation 2.12 shows the formula proposed by Gay-Lussac

$$\frac{P1}{T1} = \frac{P2}{T2}. \tag{2.12}$$

Lastly, Amedeo Avogadro postulated in 1811 that a gas's volume is exactly proportional to the number of moles it contains. According to this law, gases with the same volume, temperature, and pressure have an equal number of molecules. The sum of the formulas mentioned above is the ideal gas law, which Emile Clapeyron first presented in 1834. The ideal gas law is written as PV = nRT, where n is the number of moles of gas, R is the universal gas constant, and T is the absolute temperature (Laugier and Garai, 2007; Levine, 1985; Tenny and Cooper, 2017).

2.3.2 Real Gas

Real gases deviate from the ideal gas behaviours. They can be defined as gases that assume the shape of their container and interacts with each other. A gas that under all standard pressure and temperature does not obey the real gas law is said to be a real gas. Real gases have mass, volume, and velocity. The extent of deviation of real gases from ideal gases is measured by the use of the compressibility factor.

Real gases at higher temperatures pull closer to each other in the container in which they are stored. This result in a higher intermolecular force of attraction between the gas molecules.

These intermolecular forces hold the gas molecules together reducing the force and the rigorist collision with the container walls. This presents a pressure value lowers than the ideal gas value. Additionally, at greater pressures, a bigger portion of the container's capacity is occupied by the volume of the gas. The pressure inside the container is too high due to the reduced volume present.

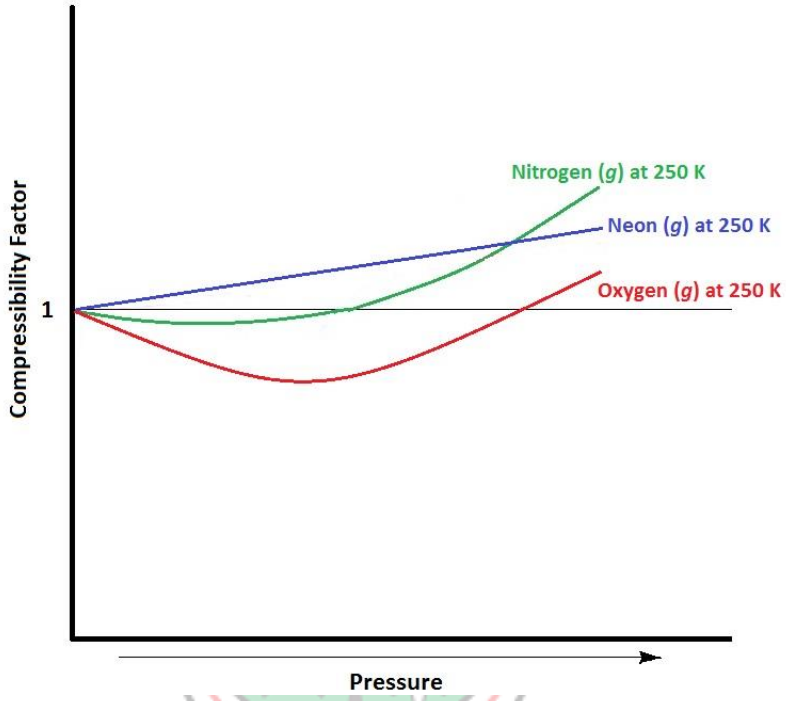


Figure 2.1 Compressibility factors of three different gasses at the same temperature of 250 K. (Key and Ball, 2014)

The temperature of gas also contributes to the deviation of ideal gas behaviour. The average kinetic energy of a gas decreases with a decreasing temperature. As a result, more gas molecules lack the kinetic energy necessary to repel the intermolecular interactions created by atoms in the same container. As a result of the gas molecules colliding with the container walls in this situation, the pressure of the gas decreases (Key and Ball, 2014).

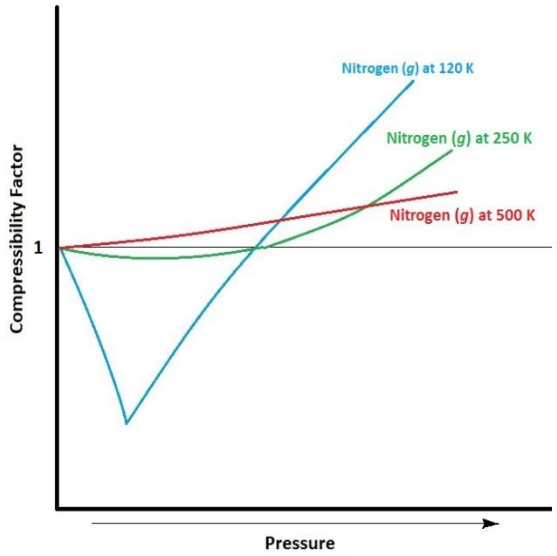


Figure 2.2 Compressibility Factor of nitrogen Gas with Different Temperatures. (Key and Ball, 2014)

Johannes Van der Waals a Dutch physicist 1973 proposed a formula that compensates for the deviations of the ideal gas behaviour by demonstrating the effects of the size of molecules and the force of attraction between the forces. The intermolecular forces and the volume of gas molecules are corrected using the Van der Waals equation, which combines two constants. The Van der Waals formula is shown in Equation 2.13.

$$\left(P + \frac{an^2}{V^2}\right)(V - nb) = n \tag{2.13}$$

Where P represent the gas pressure, a is the correction of the intermolecular forces between the gas molecules, V is the volume, n is the quantity of gas molecules, R is the constant of universal gas, and T is the temperature of the gas. The Van der Waals force has an ultimate impact on gas properties and results in the force of attraction between two or more objects that are separated by minute gaps. In establishing the colloid's stability and adhesion of a gas molecule, the concept of Van der Waals forces plays an important role.

2.4 Heating Value of Natural Gas

Natural gas's entire thermal energy output determines how effective it is as a heater. Caloric value is another name for the heating value. The energy generated during the combustion of one cubic meter of natural gas is referred to as the heating value and is expressed in mega joules

per cubic meter (MJ/m3). Gas chromatography is used to test natural gas's heating value. (Armstrong and Jobe, 1982). Heating values are directly dependent on the content of methane in the gas, that is the higher the methane content the lower the caloric value (Moharir *et al.*, 2019). Basically, two types of heating values of gas exist. These are gross heating values (higher heating values) and net heating values (lower heating values) (Gupta and Mondal, 2020).

2.4.1 Gross Heating Values

The amount of heat produced when a gas completely burns in the presence of oxygen under constant pressure is known as the gross heating value, and the entire combusted products are cooled at a specific temperature and the water present in the resultant gaseous products are condensed to a liquid state (Francis and Peters, 2013). The gross heating value is also termed as total heating value (Almarri *et al.*, 2013). The gross calorific value is presented in Equation 2.14.

$$\text{GHV of gas (MJ/m3)} = \frac{Weight\ of\ water\ (Kg)\ x\ temperature\ rise\ of\ water(°C) x\ 4.186}{Volume\ of\ gas\ (m3\ at\ STP)\ x\ 100} \quad (2.14)$$

The standard measurement of a gas's energy content is its gross or greater heating value. The gross heating value is crucial in determining the energy analysis of the system. Higher heating values are obtained using a bomb calorimeter by means of sophisticated and costly laboratory methods. However, more advanced and less costly methods such as ultimate and proximate analysis have been adopted for estimating higher heating values (Nhuchhen and Salam, 2012).

2.4.2 Net Heating Value

The inferior or lower calorific value is another name for the net heating value. The amount of heat emitted after the full combustion of a certain volume of gas under constant pressure is known as the net calorific or heating value and the all the burnt product are reversed to the same temperature as the reactants and assumes a gaseous state as the product is recovered. Net heating value is obtained under the same condition as gross heating value except that the quantity of heat that may be recovered from the water vapour released after the gas is burnt (Jenkins, 2020). The link between a fuel's weighted percentage (W) of hydrogen fuel (H) and

its weighted net heating value (LHV) and gross heating value (GHV) is established in Equation 2.15.

$$\mathrm{LHV} = \mathrm{GHV} - 10.55(\mathrm{W} + 9\mathrm{H}) \quad (2.15)$$

Where LHV is the Low Heating Value, GHV is the Gross Heating Value, W is the weighted percentage of fuel and H is the hydrogen fuel.

2.4.3 Inferior Heating Value

This is the amount of heat that would be generated following the full combustion of a certain amount of gas at constant pressure (p1), the result of combustion being reversed to a common specific temperature (t1), and the combustion producing a gaseous state as the end result.

2.4.4 Superior Heating Value

This is the amount of heat that would be generated following the complete combustion of a certain amount of gas under constant pressure (p1) and the outcome of combustion reversed to a common specific temperature (t1) except for water, which is formed and later condenses into a liquid state.

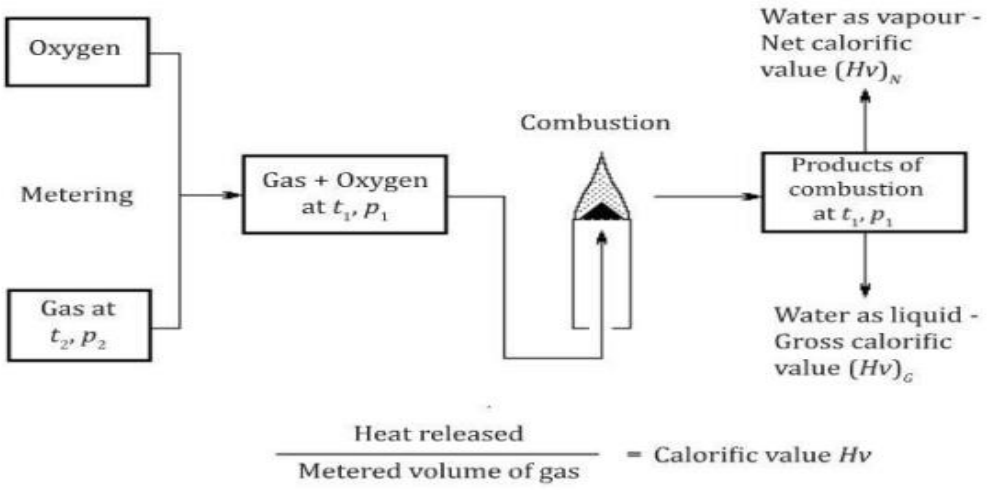


Figure 2.3 Gross Calorific Value and Net Calorific Value of Gas (ISO 6976:2016)

2.5 Standard for Estimating Heating Values of Natural Gas

The internationally recognised standard for estimating the heating values of natural gas is the International Organisation Standardisation (ISO). The technical committee for the ISO standard was created to carry out the standard and to represent the committee, government, and non-governmental organisations. To produce standards for electrotechnical, ISO collaborates closely with the International Electrotechnical Commission (IEC). Drafting of international standards is in accordance with regulations established by ISO and IEC. The drafted standards are circulated to the electorate by the technical committee who cast their vote and it requires approval from not less than 75 % of the electorate casting the vote (SS-1SO-6976, 1996).

Several versions of ISOs have existed since the 1980s which have served as an established standard for the manufacturing of gas. These standards also determine the prices of natural gas on a volumetric basis. The concept of using thermal energy as the basis of billing natural gas has become important due to the differences in the quality of gases produced in different parts of the world as well as the value of energy on the world market. Calorific values are calculated using a number of techniques for this purpose. Determining the energy of natural gas is always considered an important parameter during regulation, at the point of production, processing, and through to consumption. In determining the thermal energy of a gas, a measure of the product by volume or by mass or the calculation of its calorific value is required.

Six parts make up ISO-6974, according to the standardisation's text. It teaches how to use analysis to forecast how unpredictable natural gas is. This procedure is best to quantify heating values and other physical properties of the gas showing an uncertainty that is definable. Part two of the ISO-6947 outlines the measure of the system quality and qualitative approach to the handling of data and computation of errors. Gas chromatography is used to determine the level of uncertainty by using the 10 outlined procedures.

i. Defining the working range
ii. Defining the requirements of the analytical methods
iii. Selecting equipment and working conditions
iv. Performing type I (primary calibration) and type II (performance evaluation) analysis

v. Assigning relative response
vi. Applying quality assurance procedures
vii. Sample analysis
viii. Calculation of component mole fraction
ix. Calculating the uncertainty in the mole ratio
x. Calculating the expanded uncertainty in mole fraction.

The final part (part 3-6) postulate methodologies that target the analysis and can be applied in combination with section one and two of the ISO-6974.

There are several ISO standards that ensure quality system management requirements which sees to the design, production, processing, transportation, and services in the operation of the petroleum, petrochemicals, and natural gas industries. International standards give consumers or traders in gas commodities the confidence that the traded goods are reliable, good quality, and meet internationally accepted requirements. Natural gas products that certify ISO mean the product can operate beyond borders (Zawada, 2014).

2.6 Methods of Estimating Heating Values

In response to the global climate crisis, many countries worldwide have heavily promoted the use of alternate fuel sources than fossil fuels. This has become necessary due to the increased environmental effect and fossil fuel non-renewability (Han *et al.*, 2017). These are measures to mitigate global challenges on global warming and the depletion of fossil fuels (Yin, 2011).

The combustion of biomass such as natural gas modelling and operations heavily relies on several characteristics such as heating values, ash, moisture, and elementary compositions. The heating values of natural gas components are reported in either lower or higher heating values (Han *et al*., 2017). Through laboratory testing, the quantity of heat in gas components may be determined. A bomb calorimeter does this by calculating the difference in enthalpy between the reactants and the products (Xing *et al.*, 2019; Yin, 2011). Measuring heating values through experimentation using a bomb calorimeter is relatively easier and more accurate, however, due

to unavailability, alternative methods known as proximate or ultimate analysis are conducted. Results obtained from the proximate and ultimate analysis are subsequently used to determine heating values through empirical correlations (Sheng and Azevedo, 2005).

In contrast to ultimate analysis, which determines the weight percent of elements such as hydrogen, carbon, oxygen, nitrogen, and sulphur present in a natural gas product, proximate analysis determines the weight of moisture in percentage (wt%), fixed carbon (FC), volatile matter (VM), and ash content of the gas product (Yin, 2011). Both experimental and ultimate analysis determination of heating values require special instrumentation whiles the proximate analysis determination can be relatively easily obtained using simple equipment in the laboratories (Demirbaş and Demirbaş, 2004).

In most countries. Pricing for the quantity of gas is determined by calculating the thermal energy of the metric value product using a standard reference and a higher heating value depending on volume. Therefore, the heating value of the natural gas products is determined by various principles, volume base, molar base, and Mass base methods.

2.6.1 Molar Basis

In determining the heating value of natural gas using the molar base method, the individual molecular species present in the gas mixture are weighed in reference to its mole fraction, and all molecules in the compound are added to obtain the average mole fraction of the natural gas mixture on a molar basis. Subsequently, a conversion is done to change the average mole fraction of the gas mixture to its net caloric value. When determining the gross calorific value of a mixture with known composition at temperature t1, the ideal gas law is used to determine the mole of a gas ingredient in Equation 2.16.

$$n = \frac{PV}{RT} \tag{2.16}$$

2.6.2 Mass Basis

The mass of a gas also known as the molar weight is defined as the weight of one mole of a

sample. The mass of a substance is obtained by multiplying the mass of the equal volume of the substance by its relative density. During the calculating of density and relative density for various species of a gas mixture, the molar mass of the sample is weighed in line with its mole fraction, all the molecules present are summed up to obtain the average mole fraction of the gas quantity. The values are then converted into the relative density of the gas compound. Density and relative density values of the gas are then obtained by applying a volumetric correction factor or compression factor.

2.6.3 Volume Basis

The measure of the energy of natural gas is determined by its volume. The volume of a gas is measured in cubic meter or cubic feet. The volume or cubic foot of a gas is defined as the quantity of gas requires to fill one cubic volume under standard temperature and pressure. The volume of nature is dependent on the atmospheric temperature and pressure. People in gas industries have developed standards for measuring natural gases. Natural gas is traded by means of a cubic foot. Per the United States (U.S) standards the quantity of the occupies imaginary box one foot on either side at a temperature of 60° Fahrenheit and a pressure at sea level. Gas volume helps determine the amount of heat that can be generated from a gas. A cubic foot is equivalent to 1,020 British thermal units (BTU).

A BTU is referred to as the quantity of heat energy that is required to increase the temperature of one by 1 °Fahrenheit and pressure at sea level. Natural gas energy content varies in different locations. To ensure they are comparable, 1 cubic foot of natural gas is equal to 1,000 BTU of heat energy. This standard efficient trade between countries.

2.7 Mathematical Model for heating value of Natural Gas

Several formulae for determining the HHV or LHV of natural gas from the results produced during elementary analysis have been proposed. Dulong (1980) developed the first model for calculating the heating value of a coal sample. In Dulong's formula, the mass of ash-free dry weights of carbon, hydrogen, oxygen, and sulfur were combined linearly (Buckley, 1991; Vargas-Moreno *et al.*, 2012). Dulong's model has undergone several revisions by different

scientists and proposed variations have been derived to include new coefficients and new expressions. The original model proposed by Dulong was based on ultimate analysis. Dulong's formula for calculating the HHV of a coal was expressed as HHV = 8080C + 34,460H - 4,308O + 2250S. Where C, H, O, and S stand for carbon, hydrogen, oxygen, and sulfur, respectively, and HHV is the higher heating value. The unit of measure is expressed in kcal/kg (Kathiravale *et al.*, 2003).

The Institute for Gas Technology (IGT) and Lloyd's, respectively, presented models toward the end of the 1970s and the beginning of the 1980s that each contained a fraction of ash. All earlier models were conducted on coal samples (Channiwala and Parikh, 2002; Francis and Lloyd, 1983). Current models include a quadratic equation that included C, H, and N for biomass samples proposed by Friedl *et al.* in 2005. The equation is expressed as HHV= 0.00355[C] 2 − 0.232[C] − 2.230[H] + 0.0512[C·H] + 0.131[N] + 20.600 (Friedl *et al.*, 2005). In addition, two models were presented by Sheng and Azevedo in 2005 based on the findings of the elementary analysis. One of the equations is based on the concentration of carbon whiles the other into consideration of other variables such as carbon, hydrogen, and oxygen. The equations proposed by Sheng and Avezedo are;

Equation 1: HHV= 0.3259[C] + 3.4597 and
Equation 2: HHV= −1.3675 + 0.3137[C] + 0.7009[H] + 0.0318[O].
(Sheng and Azevedo, 2005)

In the same year, Thipkhunthod *et al.* (2005) reviewed earlier publications and proposed five models with novel and simplified coefficients. These experiments were done on sewage sludges. The five models are outlined as follows:
Equation 1: HHV= 0.4912[C] − 0.9119[H] + 0.1177[O]
Equation 2: HHV= 0.4925[C] − 0.9260[H] + 0.1176[O] + 0.0193[S]
Equation 3: HHV= 0.4148[C] − 0.1841[H] + 0.1789[O] − 2.1595
Equation 4: HHV= 0.4259[C] − 0.0698[H] + 0.1817[O] − 0.1805[N] − 2.2770
Equation 5: HHV= 0.4302[C] − 0.1867[H] − 0.1274[N] + 0.1786[S] + 0.1842[O] − 2.3799
(Thipkhunthod *et al.*, 2005).

Yin in 2011 recognised the advancement in the work of Sheng and Azevedo and proposed a new model through experimentation with several materials. Equation 2.17 was used by Yin to calculate heating value (Yin, 2011).

$$\mathrm{HHV} = 0.2949[\mathrm{C}] + 0.8250[\mathrm{H}] \qquad 2.17$$

Lastly, Callejón-Ferre *et al.* (2011) investigations examined the energy contained in intensive horticulture wastes that contributed to the production of greenhouse gases in specific regions of Spain. The results of the experiment produced six equations which were expressed are:

Equation 1: HHV= −3.147 + 0.468[C]

Equation 2: HHV= −2.907 + 0.491[C] + 0.261[H]

Equation 3: HHV= −3.393 + 0.404[C] − 0.341[H] + 0.067[N]

Equation 4: HHV= −3.440 + 0.517[C+N] − 0.433[H+N]

Equation 5: HHV= 5.736 + 0.006[C]

Equation 6: HHV= −5.290 + 0.493[C] + 5.052[H]

(Callejón-Ferre *et al.*, 2011; Vargas-Moreno *et al.*, 2012)

2.8 Natural Gas Analysis

Natural gas is analysed using gas chromatography introduced by Martin and James in 1952. Natural gas is examined for a variety of purposes, including as determining the quality of a gas and assessing its specificity, components, sources, and physical qualities (Bartle and Myers, 2002). Due to impurities of non-hydrocarbon components like carbon dioxide, hydrogen sulfide, and other components, as well as the existence of hydrocarbon components such gas condensate and natural gasoline, natural gas samples are also tested. In a similar vein, analytical techniques are used to determine the gas's heating value (McNair *et al.*, 2019; Snow and Slack, 2002).

2.8.1 Analytical Methods

In addition to understanding its physical and chemical characteristics, understanding the reactivity of pollutants in the gas component is crucial to ensuring that production and processing of natural gas meets the requirements for the sale of gas products. Natural gas differs in composition and properties because of derivation from different parts of the world. Due to this, the reactivity of the gas molecule can vary chemically and physically. The purity of a natural gas stream is dependent on its physical and chemical composition. In cases like condensate and gasoline which have a complex chemical mixture of hydrocarbon and non-hydrocarbon compounds present in the gas may not be seen in its composition. To ensure that the product meets standards, the gas goes through processes to determine its true boiling point, specific gravity, viscosity, density, water contents and sediments, and other tests (Shepherd, 1947; Wallis, 2013).

2.8.2 Gas Chromatograph Analysis

Gas chromatography is an analytical technique that is used to examine unstable substances when they are in the gaseous phase. It remains the primary technique to determine the distribution of carbon and hydrogen in a hydrocarbon liquid. Gas chromatography also serves the reason for determining the purity of the gas. Due to its inefficiency to determine absolute purity, distillation and solidification become the best method for determining absolute purity. Even though, gas chromatography lacks this quality it is the widely adopted technique that is used to determine and measure derivatives of hydrocarbon in crude oils products (Wallis, 2013).

Gas-liquid chromatography is mostly used to separate volatile components of organic compounds in solution. This method is in fact, the most adopted and efficient technique to separate organic compounds. It is mostly suitable for quantitative analysis of a compound with known components where each component is determined independently. Gas chromatography combined with mass spectrometry becomes an extremely useful tool to determine the compositions of an organic compound (Chemistry LibreTexts, 2020).

The operation of gas-liquid chromatography requires the injection of the compound of interest into the sample port where vaporisation takes place. An inert gas mostly helium or nitrogen

transport the injected vaporise sample where it travels through a gas chamber loaded with silica that is liquid coated. The liquid's solubility now controls how quickly results appear. This module's significance is in providing a clearer grasp of the measurement and separation technique and its use. Due to the gas's thermal expansion, the temperature of the gas chromatography oven affects how quickly the gas flows. The centre of the gas chromatography is the point where the separation of the gas molecules takes place. At this phase, the samples are separated into individual components and exit the column. The detector in the gas chromatographic results in an output signal. In a chromatogram, the signal results in the gas chromatography peak characteristics. The chromatogram's peaks show a proportionate depiction of the concentration of the target gas. Gas chromatography software and hardware are connected to the gas chromate graph to aid in diagnosing, reporting, and output of the products (Ying et *al.,* 2019).

2.9 Machine Learning (ML)

Machine learning has over the past two decades witnessed a dramatic progress in various facet of life. It is undoubtedly and undebatable one of currents rising subjects in the areas of technology and undeniably the most advanced areas in the study of computer and data sciences. The field of machine learning has advanced from laboratory novelty to a more robust technological practices for wider output and use commercially (Jordan and Mitchell, 2015). ML emerged as one of the important fields under artificial intelligence and has been useful in the arena of computer science disciplines and has played a crucial role in the development of practical software for controlling and visualisation of more advance technologies including robots, languages encryption and other beneficial application. There are two basic branches of machine learning which are the supervised and unsupervised machine learning (Zhou, 2021). The development of a mathematical model to evaluate the heating values of natural gas is the main goal of this thesis. The mathematical models that will enable making predictions for the heating value of natural and comingled gas will be developed as part of this thesis using a supervised machine learning method and inputs. These methods were selected because of its accuracy in predicting heating values. To buttress this claim, many studies have revealed the application of machine learning in predicting Higher Heating Values (HHV) of materials. Xing

et al., (2019) used ANN, Support Vector Machines (SVM), and Random Forest Regression (RFR) to predict the HHV of biomass based on their proximate and ultimate analyses. The authors used R^2 to compare the accuracy of the models and the RFR Algorithm performed better with $R^2 > 0.94$. Taki and Rohani (2022) used Radial Bias Function Artificial Neural Network (RBF-ANN), Multilayer Perceptron Artificial Neural Network (MLP-ANN), Support Vector Machine (SVM) and Adaptive Nero-Fuzzy Inference System (ANFIS) to predict the HHV of Municipal Waste (MW) for waste -to- energy evaluation. The authors used six different inputs which were carbon, water, hydrogen, oxygen, nitrogen, sulphur and ash. The results revealed that RBF-ANN can predict the HHV of MSW with higher accuracy than other models. Birgen *et al*. (2021), also used ML based modelling to predict the Lower Heating Value (LHV) of municipal waste. In their work, the Gaussian Processes Regression (GPR) was used.

2.9.1 Supervised Machine learning

Supervised or Inductive machine learning develops a target function that can be used to forecast the value of a particular values of class of interests (Muhammad and Yan, 2015). The main goal of supervised machine learning is to create a function that links an expected output to an input. Supervised machine learning is classified as the commonly used type of machine learning by classification because their aim is to enable the computer or technology in use to learn a pattern or classified system that has been generated. In data aggression using supervised machine learning, the first step include generating the datasets to be used for the programming to obtain the desired outputs. This involves an expert selection of appropriate input variables or measuring all variables available to obtained relevant data that will be of interest for the programming. The later process is mostly referred to as "brute-force" method. After the required data set is acquired, the data is then prepared and processed. This is an important level in supervised machine learning. There are numerous techniques proposed by a lot of researchers to help correct and deal with missing data during the data generation process. The next step is the selection of algorithm for obtaining the outputs (Muhammad and Yan, 2015; Osisanwo *et al*., 2017).

Generally, in order for a supervised learning works properly, the model must be trained to

produce reliable predictions. In order to avoid overfitting, it is usual practice to randomly partition the available data, using half for training and the remaining half for verification. Overfitting is the phenomena of fitting a model to the training data so closely that it does not function well in general (Saleh, 2022). Some of the machine learning techniques that will be considered in this paper includes AdaBoost, XG Boost, and Random Forest, Artificial Neural Network, Linear Regression and bagging regression.

2.9.2 Adaptive Boosting (AdaBoost)

This machine learning boosting model merged many weak and incorrect comparing rules to build a very accurate prediction rule. AdaBoost is the first boosting algorithm and widely used practical boosting model and it has serves numerous multifaceted purposes in different fields. AdaBoost algorithm was developed by Yoav Freund and Robert E Schapire (Freund, Schapire, and sciences, 1997; Schapire, 2013). The AdaBoost works on principle that the final outcome of interest is dependent on the repetitive measure of previous outcomes i.e the power of prediction is slightly increased based on output from previous test. AdaBoost's weak learners create a single-split decision tree called the decision stump from a single input attribute. Each observation receives the same amount of weight when creating the first decision stump. By adding weights to data and models, AdaBoost modifies the basic idea of boosting. AdaBoost uses the steps below during its operationalisation;

1. AdaBoost selects authentic learning data and weights each row by 1/n, where n represents the total number of data points;
2. It uses the aforementioned predictive model to make forecasts about the aforementioned subgroup;
3. It then uses the sample drawn at random from the initial data and builds the decision stump predictive model;
4. The modified weight of the original data is then increased by increasing the weight of incorrectly predicted row and decreasing the weight of successful prediction row;
5. It selects rows from the original datasets, giving those with more weight a higher priority. There rows reflect the errors in the prior forecasting model’
6. It then generates a new model using the previously chosen dataset; and
7. Then, steps (i) through (v) are repeated until a reliable accuracy is attained or the

maximum number of iterations is reached (Hashmi, 2019.

2.9.3 XG Boost

Extreme Gradient Boosting (XG Boost) is a large-scale machine learning system for tree boosting. The XG Boost comprises of an effective linear solver and a tree learning algorithms which enable several operations, such as ranking, classification and regression. The XG Boost package is designed to be extendable, allowing users to quickly specify their own goals. The scalable tree boosting machine learning technique is openly available and it is generally acknowledged in a variety of data mining and machine learning challenges (Chen and Guestrin, 2016; Chen *et al*., 2015). XG Boost has several features which includes;

1. Speed: With openmp, XG Boost can perform parallel computation instantly on Linus and Windows with its speed more than ten times quicker than gbm.
2. Inputs: It has the capacity to accommodate different types of input data which includes:
   a. Dense matrix
   b. Sparse matrix
   c. Data file
   d. xgb.Dmatrix
3. Sparsity: It can acknowledge sparse inputs from linear and tree booster with maximum optimisation for sparse inputs.
4. Customisation: It supports objective and evaluation functions of customisation.
5. Performance: When processed on different database, XG Boost produce better performance (Chen *et al*., 2015).

Generally, XG Boost is widely used in different discipline due to its ability to produce state-of-the-art results and provide solution to many problems. XGBoost works by enhancing computation and making boosting techniques faster. The following steps describe how XGBoost works:

i. Builds a prediction model X_1 using all of the Learning data and the target variable and then computes the distinction between the predicted and the original target variable as

G_1;

ii. Develops a new predictive model X_2 using G_1 as the goal variable, which represents the gradient of the errors made by model X_1. It then finds the distinction between the predicted and the original target variable as G_2;

iii. It then repeats these steps in (i) and ii until the error is zero or the highest number of iterations is achieved; and

The final prediction value will be the total of all the predictions made by the models X_1, X_2, … and X_n (Hashmi, 2019).

2.9.4 Random Forest

The Random Forest algorithm proposed by Breiman (2001) has been used widely for predictions in machine learning. This algorithm puts together different randomised decision trees and combining their prediction by finding averages of the combined predictions. Random decision forest has been identified as better prediction tool compared to linear regression because of its ability to adapt to nonlinear conformity and produces better results when dealing with medium to larger datasets. It utilises the aggregation of multiple machine learning algorithm by combining the unit prediction of the most frequent outcomes from a series of classified trees and combining the predictions to produce the final results. Results produce from random forest are accurate, shows tolerable outliers and sound well and do not produce overfitting (Biau and Scornet, 2016; Liu, Wang and Zhang, 2012; Schonlau and Zou, 2020) .. This supervised machine learning is a classifier composing of different tree-structured classifiers $\{h(x, \Theta_k), k=1\}$ where $\{\Theta_k\}$ are different and independent random vectors, x is the input cast by each decision tree and k is the position of each of the casted input on the decision tree (Wang and Zhang, 2012).

2.9.5 Artificial Neural Network

Artificial Neural Network (ANN) has over the years obtain some level of success in its application to solving problems related to engineering. ANN is an artificial intelligence tool which is designed to reduce the involvement of the human brain and nervous system from stress

and strain by solving complex problems where there is an unknown relationship between modeled variables. It is likened to the build-up of the human brain and the nervous system where networks of various connections called nodes are connected (weights) to each other to process information or inputs and produce desirable or meaningful output. In ANN, the connection of networks, the value of the weight, and other important functions determine the strength of the output produced. ANN is a complex and robust tool that is used in broad range of disciplines in solving computational, traditional and conventional mathematical problems (Abiodun *et al*., 2018; Shahin, Jaksa, and Maier, 2001; Wu and Feng, 2018).

2.9.6 Linear Regression and bagging analysis

Regression analysis is a statistical technique used to calculate potential connection variables of interest that demonstrate cause-and-effect relationships. Regression analysis can be univariate or multivariate depending on the number of independent variables being tested. A univariate regression has a dependent variable and one independent variable whereas a multivariate regression analysis considers an independent variable against more than one independent variables. To prove a linear relationship between the dependent and independent variables, univariate regression analysis is performed. Nevertheless, synchronic accounting for the variations between the dependent and independent variables is attempted in multivariate regression analysis (Maulud *et al.*, 2020; Uyanık *et al.*, 2013). A multivariate regression analysis is shown in Equation 2.18.

$$y = \beta_o + \beta_1 X_1 + \beta_\eta x_\eta + \mathcal{E} \qquad 2.18$$

where:

y is the dependent variable

$X_1 \ldots X_n$ are the independent variable

B_0, $B_1 \ldots B_n$ are the co-efficient variables

Ɛ is the error.

Linear regression analysis is very important that it allows the use of multiple independent variables (Uyanık *et al.*, 2013).

Bagging in machine learning is a method used to improve results during classification algorithm. Using random subsets of the original data set, a collective meta-estimator known as a bagging classifier updates basic classifiers one at a time, combining the individual predictions into a single prediction. To increase the system's overall performance, machine learning techniques like bagging typically combine the predictions of many models. This treats each model in a separate subset and then combines the prediction from each subset to gain the final prediction. The bagging model is important because by averaging the forecasts from different models, it can lower the variance of the model. By reducing the correlation between the models when the models are trained on different subsets of the data, overfitting may also be reduced (Dey, 2023; Machová *et al.*, 2006).

CHAPTER 3

MATERIALS AND METHODS USED

3.1 Data Acquisition

An unpublished secondary data on heating values and other related parameters by the Gas Chromatograph were obtained from Ghana's Offshore Oil Fields through Ghana National Gas Company for the prediction. Standard heating values at reference conditions of 20 °C temperature and 101.325 KPa were as well obtained. All data used were in their acceptable field units. Appendix A shows the first one hundred data points on average daily gas specification recorded by the Gas Chromatograph. All data used were in their standard unit.

3.2 Processing-Aided Tools

3.2.1 Microsoft Excel

Microsoft Excel Spreadsheet Software was used for all the calculations involved in this project. Microsoft produced the spreadsheet program Microsoft Excel for Windows, Android, and iOS. It includes computations, graphing tools, pivot tables, and Visual Basic for Application, a macro programming language. In many different sectors across the world, it has been used extensively. Together with Microsoft PowerPoint, Excel is a component of the Microsoft Office package. This software was selected because of its high accuracy, reliability, and user-friendliness. It also provides a wide range of computation and graphical applications. Figure 1 shows an excel interface.

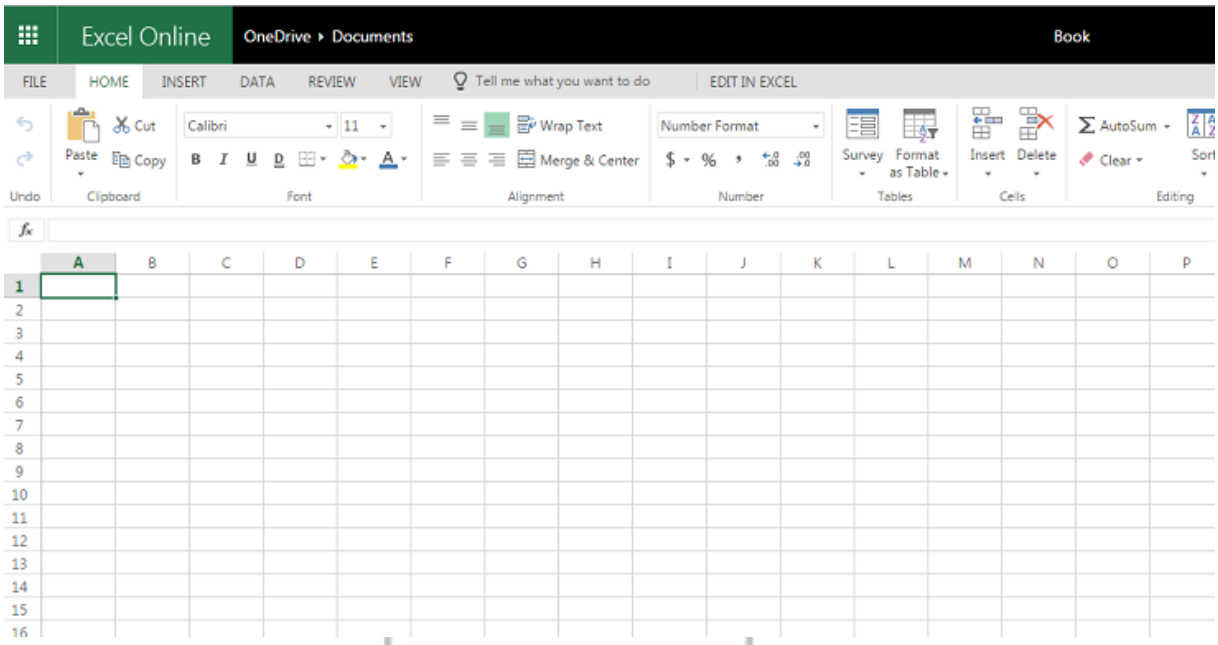


Figure. 3.1 Excel Spreadsheet Interface

3.2.2 Python Google Colab Notebook Tool

The Google colab notebook is a popular data science platform for analysing, processing, classifying, modelling, and visualising data. One of Google Research's products is called Colaboratory, or "Colab" for short. Colab is extremely useful for machine learning, data analysis, and teaching since it enables anybody to create and run arbitrary Python code through the browser. Technically speaking, Colab is a hosted Colab notebook service that requires no installation and offers free access to computational resources, including GPUs. Colab Notebook supports multiple programming languages like R, Julia, and Python. For this thesis, python language was employed. Figure 3.2 shows the interface of the Colab Notebook.

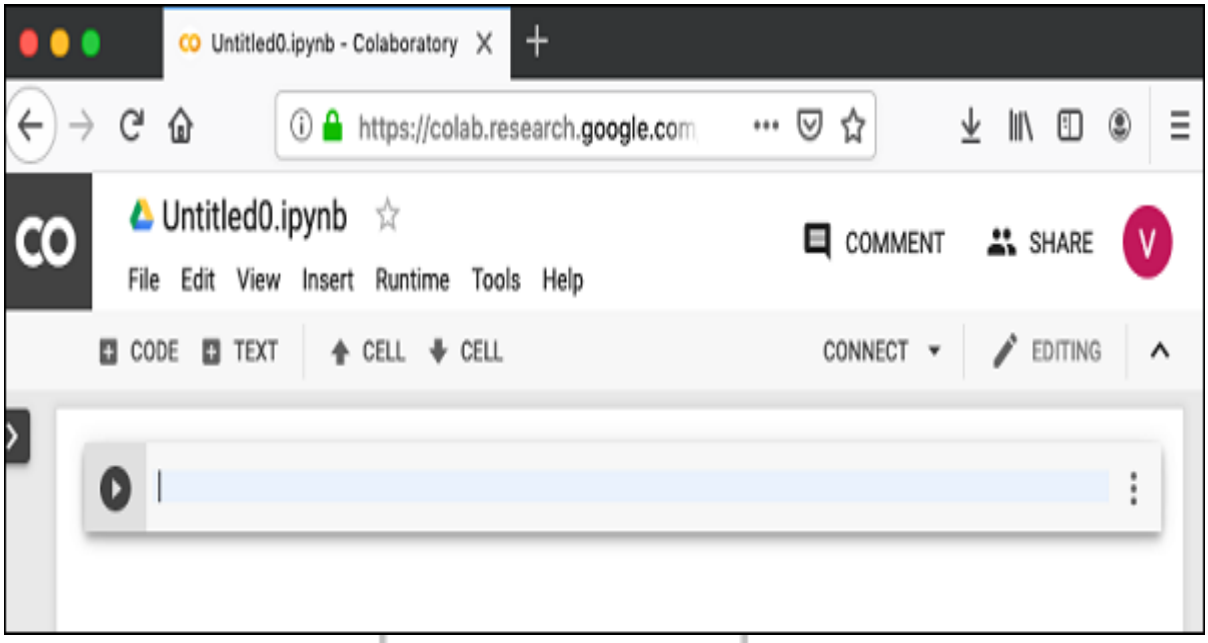


Figure. 3.2 Colab Notebook User Interface

3.3 Heating Value Estimation using ISO 6976:2016

The calculation of heating values was done at a reference condition of 20 °C temperature and 101.325 KPa which is specific to West Africa. The heating value was expressed in three forms namely molar basis, mass basis, and volume basis. This thesis focused more on the volume basis.

The execution of this was done in the following stages:

- ☐ Estimation of Commingled Gas Composition;
- ☐ Estimation of Heating Values (Molar, Mass, and Volume Basis);
- ☐ Calculation of Gas Compressibility Factor (Z);
- ☐ Estimation of Relative Gas Density;
- ☐ Estimation of Real Gas Density; and
- ☐ Estimation of Wobbe Index.

The chronological procedure presented above was done according to such stages, just like in mathematical calculation, the values obtained from the first is needed to calculate the next value. The commingled gas composition is needed to estimate the heating value as well as the

calculation the compressibility factor. This is the standard procedure for the calculation of the values presented above.

3.3.1 Estimation of Commingled Gas Composition

The constituents of a comingled gas sample was determined to ensure accuracy. The gas obtained is usually from two or three sources namely JUBILEE, TEN, and Sankofa Fields based on the location. The composition of the comingled gas was calculated as follows:

- The total volume of gas was multiplied by the individual composition. For example, Gas A with 85% methane and a total volume of 200 MMSCF will yield 170 MMSCF of Methane in the mixture (0.85x200MMSCF). This simply means out of 200 MMSCF of gas transported, 170 MMSCF is methane. Gas B with 50% methane and a total volume of 100 MMSCF will similarly have 50MMSCF methane in the mixture.
- Compositional volume of each gas stream is added and divided by the total g as in the two streams to determine the comingled composition. The comingled gas composition can be estimated using equation 3.1.

Mathematically:

$$Comingled\ Gas\ Composition = \frac{(Volume\ of\ Gas\ A * C\ ) + Volume\ of\ Gas\ B * D}{Total\ Volume\ of\ Gas\ in\ A\ and\ B} \quad (3.1)$$

Where;

C = Individual Composition of gas A (Methane, Ethane, Propane, *etc*.)

D = Individual Composition of gas B (Methane, Ethane, Propane, *etc*.)

3.3.2 Estimation of Heating Values

The heating value calculation was done simply by multiplying the individual gas compositions and their respective standard heating values using Equations 3.2, 3.3 and 3.4.

$$Volume\ Basis = \frac{Molar\ HV * P}{z * R * T} \quad (3.2)$$

Where;

P = Reference Pressure in KPa

R = Gas Constant (8.31451JMol^-1K^-1)

T = Metering Temperature in K

$$Molar\ Basis = \sum_{i=1}^{n} HVi * Xi \quad (3.3)$$

Where:

HVi = Component heating value in KJ/mol

Xi= Component Mole fraction

n= the number of constituents in gas sample

$$Mass\ Basis = \frac{HVmolar}{Molar\ Mass\ \ of\ Mixture} \quad (3.4)$$

$$Ma = \sum_{i=1}^{n} Mi * Xi \quad (3.5)$$

Where:

Ma = Molecular Mass of Mixture

Mi = Component Molecular mass

Xi = Component Mole Fraction

n= the number of constituents in gas sample

3.3.3 Estimation of Gas Compressibility Factor (Z)

The compressibility factor of the mixture was calculated using Equation 3.6.

$$Zmix = 1 - (\sum_{i=1}^{n} XiBi)\text{^}2 \quad (3.6)$$

Where:

Xi = Component Mole Fraction

Bi = Component Summation Factor

3.3.4 Estimation of Relative Gas Density

The relative gas density at reference conditions was calculated using Equation 3.7.

$$\rho g = \frac{MaZair}{MairZmix} \quad (3.7)$$

Where, M_a = Apparent Molecular weight of the gas mixture

Z_{air} = Compressibility of air

M_{air} = Molecular Weight of Air (28.9626 Kg/Kmol)

Z=Gas mixture Compressibility factor

ρg = Relative Gas Density

3.3.5 Estimation of Real Gas Density

The real gas density of the mixture was estimated at reference conditions of 101.325 KPa and a temperature of 20 °C using Equation 3.8.

$$\rho g = \frac{PMa}{ZRT} \quad (3.8)$$

Where:

P = Reference Pressure in KPa

M_a = Apparent Molecular weight of the gas mixture

Z=Gas mixture Compressibility factor

T = Reference temperature in K

R= Gas Constant

3.4 Pre-processing and Statistical Analysis of Data

Different algorithms were used to train the data set with the aid of the Colab notebook, and the best algorithm was selected based on the R-squared, Adjusted R-Squared value, Mean Absolute

Percentage Error (MAPE) value, Mean Absolute Error (MAE) value and Root Mean Square Error (RMSE). The methods used in preparing the data for the different algorithms are Exploratory Analysis of Data, Processing of Data, and finally Model Development and Evaluation.

3.4.1 Data Pre-processing and Approaches Used

The data set was explored to check if there are abnormalities within the data. The main objective of this analysis was to provide a statistical description of the data, determine outliers within the data set to check for missing values as well as to provide correlation analysis. For accuracy in results, this step was carried out before model development.

Statistical Description

For each column in the data set, the mean, standard deviation, minimum and maximum values, and percentiles were all calculated as part of the statistical description of the data. To achieve this, the "data. describe" coding function in python language was used.

Outliers Determination

Outlier determination is very crucial in data analysis. In this research, outliers were determined to know the values which are far from the normal sets of data. If a dataset has huge number of outliers, it affects the results negatively. In the determination of outliers, the box plot obtained was used in checking for the outliers. Further work on outliers was done during the Data Processing Stage.

Missing Values Determination

Missing values were determined to see if there were any of the data sets missing between the predictors and dependent variable which is the High Heating Value in this case. To verify this, the total count of values should be the same for all parameters. This is all part of the pre-

processing stage. To achieve this, “data.info” code function was entered.

Correlation Analysis

Another important thing that was checked before data processing was Correlation Analysis. This was done to check the issue of Multicollinearity, which is a strong correlation between predictors which affect the results when not resolved.

3.4.2 Statistical Analysis of Data and Approaches Used

Data processing is a very important part of the modelling. This is where the dataset used for the prediction of the heating value of natural gas is being processed to make important conclusions and findings. Data processing is made up of Outlier Determination and Fixing, Multicollinearity Test, Input and Output variable selection, Data Splitting, Data Normalisation, Model Development and Prediction, Evaluation on training and testing dataset, and Plotting of predicted heating value and Actual Heating Value.

Outlier Determination and Fixing

An outlier is a data point that substantially varies from other observations in statistics. An outlier might result from measurement variability, or it could be an indication of an experimental mistake; the latter is sometimes removed from the data set. An outlier in statistical analysis might result in significant issues. The dataset for this project was a real-time field of daily values recorded by a Gas Chromatograph, however, there were several outliers in the dataset which could have affected the prediction if not fixed. Outliers can skew the results of the model and lead us towards wrong interpretations, to identify the outliers in the dataset, visualisations with boxplots, a statistical approach using interquartile ranges, and imputation of the values of the outliers were used.

Due to the irregular distribution of the data from the Gas Chromatograph, outliers were calculated using the interquartile ranges rather than Z-scores. Equations 3.9, 3.10 and 3.11 were

used to estimate the outliers from the dataset using Inter-Quartile Range (IQR) Approach.

$$Q1 - 1.5 * IQR \quad (3.9)$$

$$Q3 + 1.5 * IQR \quad (3.10)$$

$$IQR = Q3 - Q1 \quad (3.11)$$

IQR is for interquartile range, while Q1 represents the lower quartile (25th percentile) and Q3 represents the upper quartile (75th percentile). Outliers are values that fall outside of the range of Equations 3.9 and 3.10. A straightforward Python function that accepts our column from the data frame and produces the outliers using the handy pandas can be built. In resolving the outliers in the dataset, an imputation approach was used whereby the mean value of each parameter was determined and used to replace the outlier.

Multicollinearity

When there is a significant connection between two or more independent variables in multiple linear regression, the dataset has multicollinearity. When a researcher or analyst tries to figure out which independent variable may be utilised to predict or comprehend the dependent variable in a statistical model, multicollinearity can result in skewed or misleading results.

A multiple regression model's collinearity may be identified and quantified using a statistical method known as the variance inflation factor (VIF). When the predictor variables are not linearly connected, the variance of the predicted regression coefficients is less inflated, as measured by the VIF. Variables that have a VIF of 1 are not correlated, whereas those that have a VIF between one and five are moderately linked, and those between five and ten are strongly correlated.

Input and Output Variable Selection

This is where the data is segregated into two. One part is the dependent variable or output

variable and the independent variable or predictors. In machine learning, there is always an input and output variable which is being studied and used for further predictions. For the sake of this thesis work, the predictors used were methane (C_1), Ethane (C_2), Propane (C_3), Isobutane (iC_4), Normal Butane (n-C_4), Isopentanes (iC_5), Normal Pentane (nC_5), Hexane (C_6+), Nitrogen (N_2) and Carbon dioxide (CO_2) and the independent or output variable was High Heating Value (HHV). In the coding, "Label" was used for the output variable which in our case is High Heating Value and "Features" was used for the predictors.

Data Splitting

At this stage, the data was randomly split into training and testing. The quality of the training data impacts directly the accuracy and reliability of the algorithms. . The dataset's training portion was used for 80% of the computations, and the testing portion for 20%. The total dataset used for the work was 2021, out of which 1617 representing 80% were used for training and 404 representing 20% were used for testing. The training dataset was used to train the various algorithms, and the testing dataset was used to evaluate their accuracy.

Data Normalisation

Data Normalisation was done to scale the dataset between 0 and 1 for easy prediction. This scaling was done so that there wouldn't be many outliers in the dataset. This was done to ensure fairness in the dataset for better prediction. In this project, The MinMax Scaler function in learns pre-processing library was used in normalizing the dataset in (0, 1) intervals. Equation 3.12 shows the formula for MinMax Scaler.

$$Y = \frac{Yactual - Ymin}{Ymax - Ymin} \qquad (3.12)$$

Where:

Y= Normalised value for the parameter (C_1, C_2, C_3…)

Y_{actual} = Value for individual parameter

Y_{min} = Minimum Value for the parameter

Y_{max} = Maximum Value for the parameter

3.4.3 Accuracy Measures

The model was then evaluated on the training and testing datasets using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R-Squared (Coefficient of determination) Adjusted R-Squared, and Mean Absolute Percentage Error.

After model development and evaluation, a crossplot of the estimated heating value and actual heating value was made and a line of best fit was drawn to determine the equation for prediction. The crossplot provides the visual representations of the relationship between the predicted and the actual heating value

Root Mean Square Error (RMSE)

The standard deviation of the errors is measured by the Root Mean Square Error (RSME). This measurement, which reveals how effectively a regression model can forecast an absolute value for a response variable, is crucial for prediction. The predicted heating values from each model were exported to excel to estimate the RMSE for both training and testing datasets using Equation 3.13.

$$RSME = \sqrt{\frac{1}{n}\sum_{1}^{n}\left(HHV_{act} - HHV_{pre}\right)^2} \quad (3.13)$$

Where RSME represent the root mean square error, n is the number of data samples, HHV_{pre} predicted heating value and HHV_{act} is the actual heating value.

Mean Absolute Error (MAE)

The total disparity between a dataset's actual and anticipated values is averaged out to get the Mean Absolute Error (MAE). The predicted heating values from each model were exported to

excel to estimate the Mean Absolute Error for both training and testing datasets using Equation 3.14.

$$MAE = \frac{1}{n}\sum_{1}^{n}\left|HHV_{act} - HHV_{pre}\right| \quad (3.14)$$

Where MAE is the mean absolute error, n is the number of data samples, HHV_{pre} predicted value and HHV_{act} is the actual value.

Coefficient of Determination (R^2)

The amount of the dependent variable's (Heating Value) variation that the linear regression model accounts for is represented by the coefficient of determination (R2). R^2 is always less than 1. The predicted heating values from each model were exported to excel to estimate the R^2 for both training and testing datasets using Equation 3.15.

$$R^2 = 1 - \frac{\sum_{1}^{n}(HHV_{act} - HHV_{pre})^2}{\sum_{1}^{n}(HHV_{act} - HHV_{avp})^2} \quad (3.15)$$

Where R^2 is the coefficient of determination, n is the number of data samples, HHV_{pre} predicted heating value, HHV_{act} is the actual value and HHV_{avp} is the average of the predicted heating value.

Mean Absolute Percentage Error (MAPE)

In regression analysis, the mean absolute percentage error is used to gauge how well a predicting approach will perform. MAPE is always in percentage (%). The predicted heating values from each model were exported to excel to estimate the MAPE for both training and testing datasets using Equation 3.16.

$$MAPE = \frac{\frac{100}{n}\sum_{1}^{n}\left|HHV_{act} - HHV_{pre}\right|}{HHV_{act}} \quad (3.16)$$

Adjusted R^2

A modified version of the coefficient of determination (R^2) known as "adjusted R^2" is primarily modified for the number of variables that are independent in the model. Mostly, Adjusted R^2 will be less than or equal to R^2. In this project, the predicted heating values from each model were exported to excel to estimate the Adjusted R^2 for both training and testing datasets using Equation 3.17.

$$Adj\ R^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1} \tag{3.17}$$

Where Adj R^2 is the Adjusted R^2, n is the number of data samples, k is the number of predictors, and R^2 is the sample R^2.

3.5 Model Development and Prediction of HHV

Using Colab Notebook and the Python programming language, several techniques were utilised to forecast the natural gas heating value. There were 2021 datasets utilised in all; 1617 of those were used to train the different models, and 404 were used to test and assess the model. Various algorithms, including Random Forest, Artificial Neural Networks (ANN), Multiple Linear Regression, Bagging Regressor, ADABOOST, and Extreme Gradient Boosting, were employed in this study.

3.5.1 Model Description

This portion of the research provides a brief explanation of the approaches taken by each model to forecast the heating value using the Colab Notebook. The models are Linear Regression, Bagging Regressor, Extreme Gradient Boosting, Adaboost, Artificial Neural Networks (ANN), Random Forest and Stacking Regressor for hybrid model.

3.5.2 Hyperparameters

A machine learning model's behaviour and performance are controlled by hyperparameters,

which are parameters that are established before the model is trained. Hyperparameters, in contrast to model parameters, are set by the practitioner and have to be carefully chosen to achieve the best outcomes. Model parameters are learnt from the data during training.
Some common hyperparameters in machine learning models include:

Learning rate: The step size of the optimisation technique utilised for updating the model's parameters during training is controlled by the learning rate. Faster convergence is produced by greater learning rates, whereas slower convergence but potentially superior solutions are produced by lower learning rates.

Regularisation: Regularisation involves modifying the loss function by including a penalty term to avoid overfitting. The L1 or L2 regularisation strength, for example, can be used as a hyperparameter to control the regularisation term, which regulates the model's complexity.

Number of hidden units: The number of hidden units in a neural network determines how complicated the model will be. A more complex model, which may capture more intricate interactions between the inputs and outputs, is produced by adding more hidden units, although this can also result in overfitting.

Number of trees: The number of trees determines the complexity of the model in tree-based algorithms like decision trees and random forests. A more complicated model is created by using more trees, but it also increases the risk of overfitting.

Depth of the tree: The depth of the tree influences the model's complexity in decision trees and random forests. A deeper tree produces a more complicated model, which can depict more intricate connections between inputs and outputs but may also result in overfitting.

3.5.3 Hyperparameter tuning.

It involves choosing the ideal hyperparameters for a particular machine learning issue. This is typically done through a combination of trial and error, guided by some performance metric such as accuracy or mean squared error. To adjust hyperparameters, methods including grid search, random search, and Bayesian optimisation are frequently employed. In summary,

hyperparameters are important components of machine learning models that control their behaviour and performance. Properly setting the hyperparameters is crucial for obtaining the best results for a given problem.

3.5.4 Hyperparameter Optimisation

Finding the optimal hyperparameters for a given machine learning algorithm that produces the best results when evaluated against a validation set is referred to as hyperparameter optimisation. The best hyperparameter combinations for training the model were discovered using the Bayesian optimisation approach. The Bayesian optimisation method entails creating a surrogate of the objective function, a probabilistic model, and using this model to iteratively choose the next point to assess based on an acquisition function that balances both exploration and extraction. Bayesian optimisation may quickly find the optimal function with just a few evaluations by iteratively adding fresh evaluations to the probabilistic model.

Random Forest

Random forest is an ensemble technique that utilises bagging, also known as Bootstrap Aggregation. Bagging involves randomly selecting subsets of data from the original dataset with replacement, a process known as bootstrap, to generate independent models. These models are trained separately to produce results, which are then combined using majority voting or averaging. This aggregation step generates the final output.

The steps involved in the random forest algorithm:

Data preparation: A training set and a validation set are created from the data. The validation set is used to assess the performance of the model after it has been built using the training set.

Tree building: A decision tree is constructed using a portion of the training data that is chosen at random. The best feature is used to split the data at each node of the tree, where a random subset of the characteristics is picked. Recursively repeating this method until the tree is fully formed.

Ensemble building: Various subsets of the training data and characteristics are used to construct a number of decision trees. Every tree is trained separately from the rest.

Prediction: To make a prediction, each decision tree in the forest is evaluated using the input data, and their predictions are combined to produce a final output. In regression problems, this can be the average of the tree predictions, while in classification problems, the majority vote of the tree predictions is used.

Performance evaluation: The validation set is used to assess the random forest model's performance. To evaluate the model's quality, measures like as accuracy, precision, recall, F1 score, and others are calculated.

The random forest model was imported from sklearn. ensemble package and was trained on the training set consisting of 10 features with 1617 records and a test set consisting of the HHV values also having 404 records. The data was normalised before training (models tend to perform badly when the data is in different ranges) using the MinMax Scaler, all of the data were balanced to be in the range of 0 and 1.

Bayesian Optimisation was used to obtain the most optimal hyperparameters, after several different ranges were explored, the most optimal combination was chosen to train the model. Table 3.1 shows the optimal hyperparameter for random forest model.

Table 3.1 Optimal hyperparameters for Random Forest Model

Number of Estimators	Random State	Max Depth
500	26	57

The training was completed in 1.69 seconds. Model performance was then evaluated for both training and testing using metrics like RMSE, MAE, MAPE, R Squared and Adjusted R Squared. Predictions made by the model on test data were further processed, tabulated, and visualised using scatter plots and line plots.

Multiple Linear Regression Model

The linear regression model was set up for training by importing the Linear Regression Model from sklearn. linear package. The training was done using the training set (consisting of 10 features with 1617 records representing 80% of the dataset) and training labels (HHV values). In the training process, the model coefficients were estimated from the data using the ordinary least squares method. Each feature in the dataset had a corresponding coefficient. The linear relationship in Equation 3.18 establishes the connection between the answer and predictions. To reduce the residual sum of squares between the observed targets in the dataset and the expected by linear approximation, the linear regression model fits a linear model with coefficients w = ($\beta_1 \ldots \beta_k$).

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta kXk + \varepsilon \quad (3.18)$$

Where:

The y-intercept (β_0), indicating what happens when all of the variables from x1 to xk are zero, determines the slope of y. The slope coefficient of all independent variables is 0; the regression coefficients β_1 and β_2 show the change in y due to one-unit changes in x1 and x2, respectively; and the term ε explains the random error (residual) in the model.

The training was completed in 58.7 milli seconds. Model performance was then evaluated for both training and testing using metrics like RMSE, MAE, MAPE, R Square and Adjusted R Square. Predictions made by the model on test data were further processed, tabulated, and visualised using scatter plots and line plots. Figure, 3.3 shows the linear regression model diagram for this work.

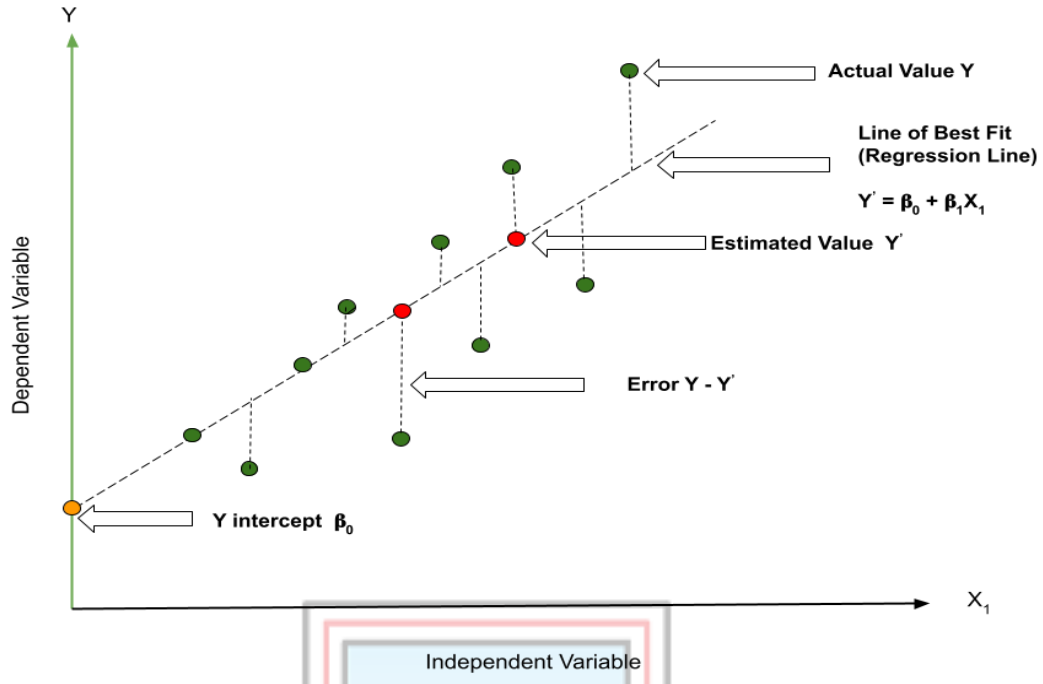


Figure 3.3 Diagram of a Linear Regression Model.

Artificial Neural Networks

A form of machine learning technique known as an Artificial Neural Network (ANN) is designed after the composition and operation of biological neurons in the human brain. Artificial neurons (ANNs) are networks of linked nodes that collaborate to process and analyse complicated data.

An ANN's structure is divided into layers, generally comprising an input layer, one or more hidden layers, and an output layer. Each neuron in the input layer represents a characteristic or characteristic of the data, and each neuron in the output layer represents a grouping or prediction based on the input data. Between the input and output layers are the hidden layers, where the neurons process the input data and extract useful features.

The neurons in an ANN are connected by weighted edges, which represent the strength of the relationship between the neurons. Each edge has a weight that determines how much influence the input from one neuron has on the output of another. The weights of the edges are changed

during training to improve the network's performance. The output of each neuron is determined by its activation function given the inputs it receives. A step function or another basic threshold function, like a sigmoid function or a rectified linear unit (ReLU) function, can serve as the activation function. The issue being addressed, and the properties of the data determine the activation function to be used.

The ANN model was constructed using a sequential model, with four layers of dense neurons stacked or connected. It comprises of an input, hidden, and output layer. The input layer included 60 neurons, a "relu" activation function, and the standard kernel_initialiser. The second hidden layer also contained 60 neurons, a "relu" activation function, and the standard kernel_initialiser. The output layer had one neuron for prediction and a linear activation function. Two dropout layers were included with a dropout of 0.2. Artificial Neural Network model followed a sequential model architecture as in figure 3.4, with an input, hidden and output layer.

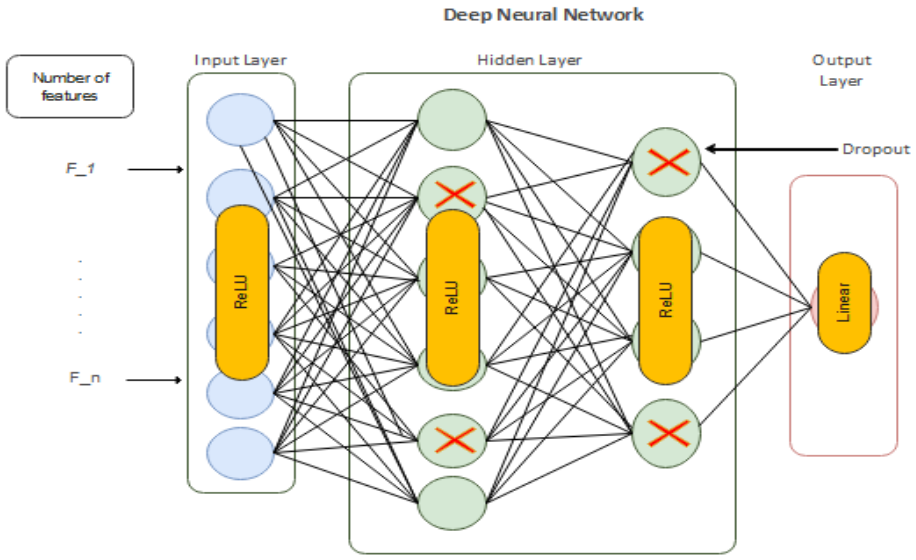


Figure 3.4 Diagram of a Deep Neural Network

The model was compiled using mean square error to measure the loss. With a learning rate of 0.01 and momentum of 0.9, Adam Optimiser was employed as an optimiser. These measures included MSE, MAE, and RMSE.

The model was trained on the training set (consisting of 10 features with 1617 records) and training labels (HHV values), for 150 epochs. The data were normalised using the StandardScaler before training. Validations were performed during training using the test_set and test_labels. Table 3.2 shows the ANN model compilation configuration.

Table 3.2 ANN Model Compilation Configuration

Optimizer	**Loss**	**Metrics**	**Epochs**
Adam	Mean Squared Error	MAE, Root Mean Squared Error	150

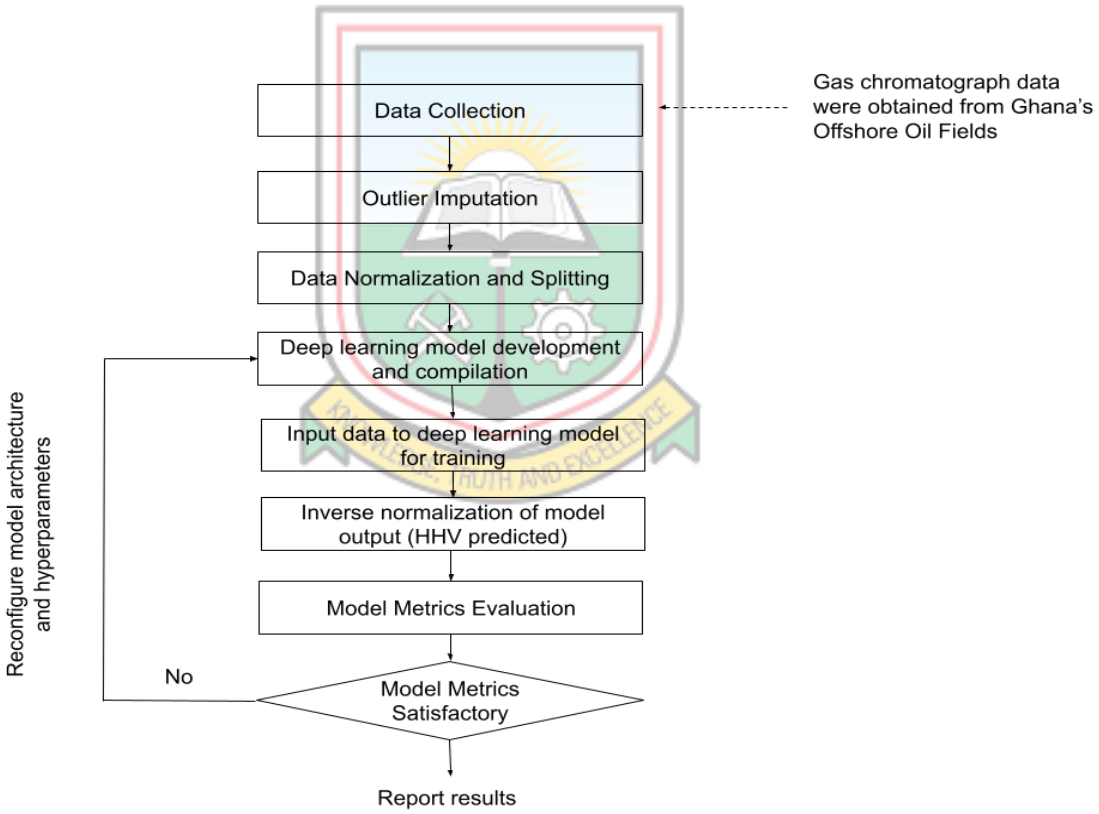


Figure 3.5 Flowchart of a Deep Neural Network Training Process.

The training was completed in 42.1 seconds. Both training and Validation loss computed during training were plotted. Loss is plotted concerning the number of epochs and shows how the

model loss changed while training the model. Model performance was evaluated for both training and testing using metrics like RMSE, MAE, MAPE, R Square, and Adjusted R Square. Predictions made by the model on test data were further processed, tabulated, and visualised using scatter plots and line plots. Figure 3.5 shows the flowchart for deep neural network training process.

AdaBoost

For classification and regression issues, a common machine learning technique called AdaBoost (Adaptive Boosting) is utilised. Several weak learners (base classifiers) are combined in this ensemble learning technique to create a strong classifier. AdaBoost is a boosting algorithm, which means it works by increasing the weight of samples that are misclassified by the previous base classifiers.

To be used for this project, the AdaBoost Regressor model was imported from sklearn.ensemble package and trained on the training dataset consisting of 10 features with 1617 records. The model was developed by fitting a regressor on the supplied dataset, followed by fitting further replicas of the regressor on the same dataset with the weights of the instances being changed in accordance with the error of the current prediction. To improve their predictions, succeeding regressors concentrated more on data points with incorrect predictions.

The hyperparameters used in training are the combination that provided the best model performance. The learning rate was very sensitive and affected the model performance to much extent, but the chosen values were found to give the best results.
Bayesian Optimisation was used to obtain the most optimal hyperparameters, after several different ranges were explored, the most optimal combination was chosen to train the model.

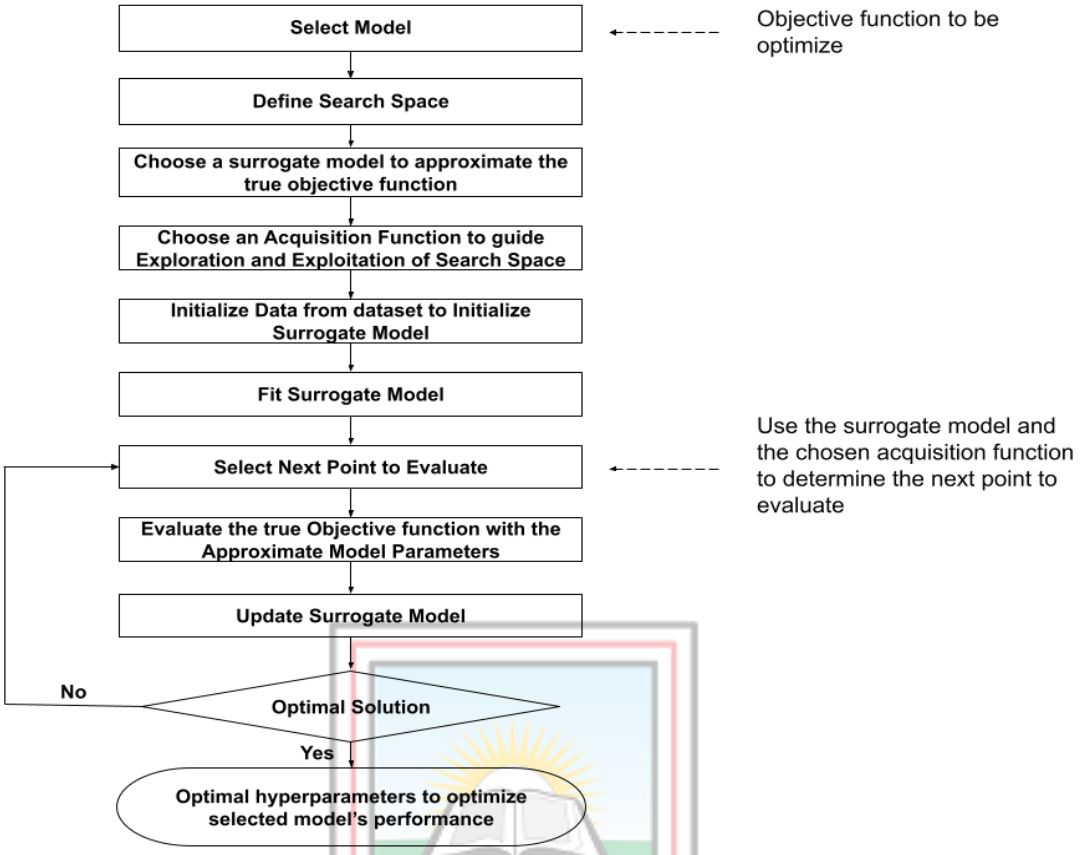


Figure 3.6 Flowchart of Bayesian Optimisation.

Table 3.3 Optimal hyperparameters for AdaBoost Model

Number of Estimators	Random State	Learning Rate
459	27	0.03

Adaboost algorithm's model training process is as described as follows:

Each feature in the training set is associated with a particular weight, which determines the extent to which that feature affects the HHV value to be predicted. At the beginning of the algorithm, the weights of the samples are initialised to be equal, which means that each sample has the same importance. The weights of the samples are updated at each iteration of the algorithm. A weak learner is a base classifier that is trained on the current sample weights. The goal is to find a weak learner that performs better than chance. The weak learner can be any simple classifier such as a decision tree or a linear regression model in this case. The sample weights are changed to reflect the performance of the weak learner after training. The samples

which the weak learner incorrectly predicts are given heavier weights than the ones that are properly predicted.The weight of the weak learner is calculated based on its accuracy. The weight is given by Equation 3.19.

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \tag{3.19}$$

Where: t is the iteration number and ϵ_t is the misclassification rate of the weak learner.The final classifier is updated by a combination of the weighted predictions of the weak learners. The prediction for a sample is given by Equation 3.20.

$$f(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t\, h_t(x)\right) \tag{3.20}$$

Where, T represent the number of weak learners, $h_t(x)$ is the prediction of the t^{th} weak learner, where "sign" refers to the sign function, which yields a value of 1 for values that are positive and a value of -1 for negative values.
The second and fifth phases of the method are repeated until a stopping requirement, such as a maximum number of iterations or a minimum accuracy level, is satisfied.

AdaBoost is a fast and effective algorithm that can handle imbalanced datasets and non-linearly separable classes. However, it is sensitive to outliers and can be prone to overfitting if the number of weak learners is too large.

The training was completed in 2.55 seconds. Model performance was then evaluated for both training and testing using metrics like RMSE, MAE, MAPE, R Squared and Adjusted R Squared. Predictions made by the model on test data were further processed, tabulated, and visualised using scatter plots and line plots. Figure 3.6 shows the Bayesian optimisation flowchart.

Extreme Gradient Boosting

The gradient boosting technique has been refined to produce Extreme Gradient Boosting

(XGBoost). For both regression and classification issues, it is a robust and scalable machine learning approach that is often utilised. XGBoost is a tree-based algorithm that works by constructing an ensemble of decision trees that are trained in a sequential manner to improve the predictive performance. Figure 3.7 shows the model structure for an XGBoost.

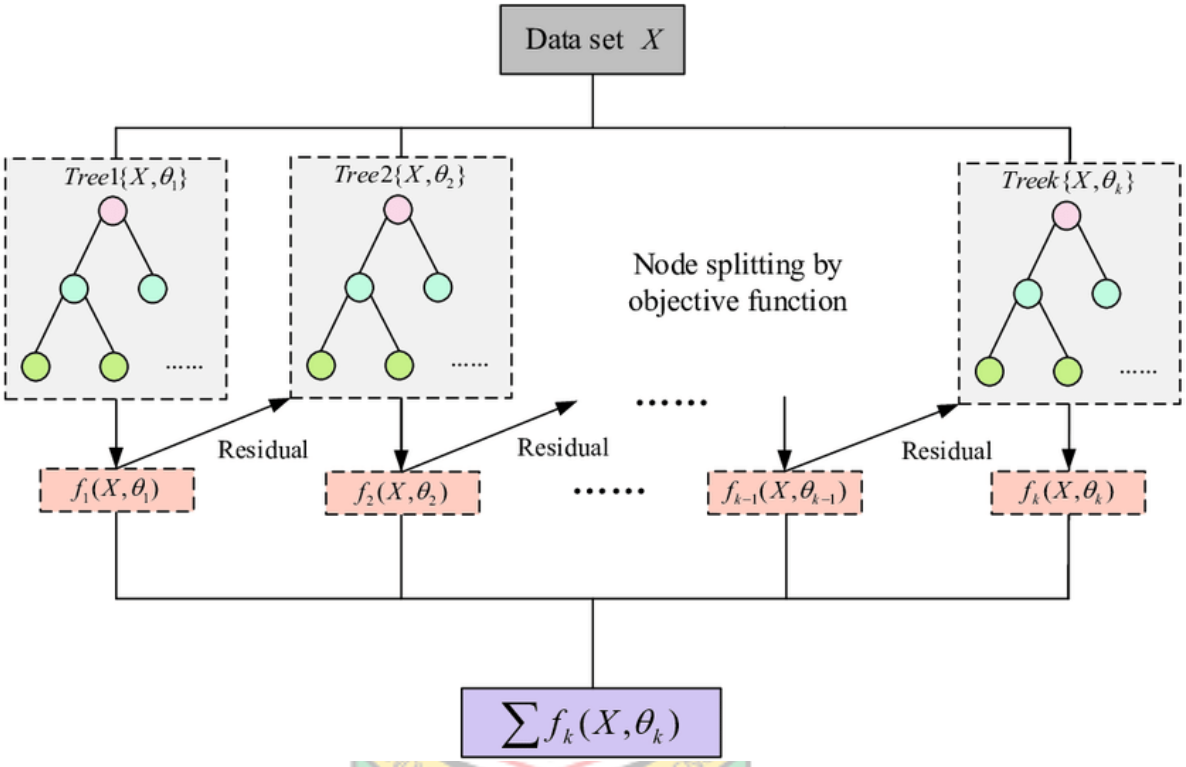


Figure 3.7 Diagram of the Model Structure of XGBoost (Guo *et al.*, 2020).

The hyperparameters used in training was the combination that provided the best model performance. The learning_rate, gamma, max_dept, min_child_weight was, max_delta_step very sensitive and affected the model performance to much extent, but the chosen combinations gave the best results. Table 3.4 shows the optimal hyperparameters for XGBoost Model.

Table 3.4 Optimal hyperparameters for XGBoost Model

Number of Estimators	Random State	Max Depth
339	388	6

Bayesian Optimisation was used to obtain the most optimal hyperparameters, after several different ranges were explored, the most optimal combination was chosen to train the model. The model training process of the Extreme Gradient Boosting algorithm is as described as follows:

The base learners are first set with a value that remains constant, such as the mean or median of the target variable before the start of the process. On the residuals (difference between predicted values and true values) of the prior base learners, a decision tree model is trained. The objective function, which is the sum of the squared residuals, is defined as the objective function to be minimised by the decision tree model. A new set of base learners is created by combining the forecasts of the decision tree model with the predictions of the prior base learners. For the next iteration, the new residuals are formed by the updated base learners.

Calculate the weight of the tree model: The weight of the tree model is calculated based on the reduction in the objective function after adding its predictions to the base learners. The weight of the tree model is given by Equation 3.21.

$$\eta = \text{learning rate}$$

$$\text{Objective Function} = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i + \eta f(x_i)\right) \tag{3.21}$$

Where: L is the loss function, y_i is the true value of the i^{th} sample, $\hat{y}_i$ is the prediction of the base learners for the i^{th} sample, and $f(x_i)$ is the prediction of the decision tree model for the i^{th} sample.

The second stages of the method are repeated until a stopping requirement, such as a maximum number of iterations or a minimal error threshold, is achieved. XGBoost is highly scalable and efficient, and it has several built-in regularisation methods to prevent overfitting, such as early stopping and pruning. XGBoost is also equipped with a parallel processing engine that enables it to handle large datasets efficiently. However, it can be sensitive to noisy data and may not perform well with highly correlated features. The Training completed in 1.07 seconds. Model performance was then evaluated for both training and testing using metrics like MSE, MAE, MAPE, R Square and Adjusted R Square. Predictions made by the model on test data were

further processed, tabulated, and visualised using scatter plots and line plots.

Bagging Regressor

An ensemble learning approach for regression issues is the bagging regressor. By merging different instances of a regression model, a technique known as bootstrap aggregating makes it easier and more efficient to analyse data. A regressor is trained on each of these samples via the Bagging Regressor using several bootstrapped samples of the training data. The average of the individual regressors' estimates makes up the final forecast.

These hyperparameters used in training as the combination that provided the best model performance. The n_estimators and random_state was more responsible for the variation in the model performance, but the chosen values were found to give the best results. Bayesian Optimisation was used to obtain the most optimal hyperparameters, after several different ranges were explored, the most optimal combination was chosen to train the model. Table 3.5 shows the optimal hyperparameters for Bagging regressor model

Table 3.5 Optimal hyperparameters for Bagging Regressor Model

Number of Estimators	Random State
13	16

The model training process of the Bagging Regressor algorithm is as described below:

A randomly selected portion of the training data which is obtained with replacement is referred to as a bootstrapped sample. Although duplicate samples may be present, the bootstrapped samples are the same size as the initial training data. Bagging Regressor creates multiple bootstrapped samples of the training data to train the individual regressors. A regressor was trained on each bootstrapped sample to predict the target variable. Any fundamental regression model, such as linear regression or decision trees, can be used as the regressor.

Calculate the final prediction: The final prediction for a sample is given by the average of the predictions of the individual regressors. The prediction for a sample is given by Equation 3.22.

$$f(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \quad (3.22)$$

Where B is the number of individual regressors, $f_b(x)$ is the prediction of the b^{th} regressor. Bagging Regressor is a simple and effective algorithm that is less prone to overfitting compared to other complex algorithms. It is also less sensitive to outliers and noisy data compared to individual regressors. Bagging Regressor is also highly scalable and efficient, making it suitable for large datasets. However, it may not perform well if the correlation between the features is high or if the data has a complex structure. The Training completed in 172 milli seconds. Model performance was then evaluated for both training and testing using metrics like MSE, MAE, MAPE, R Square and Adjusted R Square. Predictions made by the model on test data were further processed, tabulated, and visualised using scatter plots and line plot.

Hybrid Machine Learning

To benefit from their strengths and overcome their weaknesses, hybrid machine learning models combine two or more machine learning algorithms or methodologies.
Hybrid machine learning models are designed to combine the benefits of two or more machine learning algorithms, such as improved accuracy, better stability, and faster convergence. In general, hybrid models can be categorised into two types: ensemble models and integrated models.

Ensemble models combine multiple models to improve the prediction accuracy. The most popular ensemble model is the Random Forest, which combines multiple decision trees to provide better prediction results. Other ensemble models include boosting, bagging, and stacking.

Integrated models combine multiple techniques to address the limitations of individual algorithms. For example, hybrid models can combine rule-based systems and machine learning algorithms to take advantage of their complementary strengths. Another example is combining deep learning and reinforcement learning to create models that can learn from experience and

make complex decisions.

For this project a hybrid model was constructed using the mlxtend package.
The StackingRegressor module from mlxtend takes a list of regressors and a meta regressor to fit to the training set. Multiple Linear Regression, Adaboost Regressor, Bagging Regressor and RandomForest Regressor, were used as the regressors and the XGBoost Regressor was used as a meta regressor.

The training was completed in 9.17 seconds.

Named_regressoers
{'linearregression': LinearRegression(), 'adaboostregressor': AdaBoostRegressor(learning_rate=0.03, n_estimators=459, random_state=27), 'baggingregressor': BaggingRegressor(n_estimators=13, random_state=16), 'randomforestregressor': RandomForestRegressor(max_depth=57, n_estimators=500, random_state=26)}

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

The results obtained from this thesis are in two sections, the first section focuses on the estimation of heating value using ISO 6976: 2016 at reference conditions of 20 ^{o}C and 101.325 KPa. In this section, each composition of gas has its specific standard heating value at the stated reference condition, so the various percentage composition was key in determining the heating value of the gas mixture. The commingled gas composition from different sources in Ghana, namely TEN Field, Jubilee Field, and Sankofa Fields are estimated.

The second section focuses on the predictive models. This is where the results obtained from each model in the Colab notebook are presented and further discussions are made using the metrics (MAPE, RMSE, R^2, Adjusted R^2, MAE). The best method with the highest accuracy and low error margin was selected to predict the heating value of natural gas for Ghana's Offshore Fields.

4.1.1 Results from ISO 6976:2016

This section of the project provides the results obtained from the estimation of heating value at specified reference conditions used in Ghana. The reference condition used in the calculation of heating values is 101.325 KPa and 20^{o}C. This is the same approach used by the Gas Chromatograph in calculating the daily Heating Values in the Gas Industry. Table 4.1 shows the results from the Commingled Gas Composition Calculation.

Table 4.1 Results from Commingled Gas Composition Calculation

	ATUABO -LEAN			ENI (S)		COMMINGLED	
	VOLUMES						
	112.70 MMSCFD			147.83 MMSCFD		261 MMSCF	
	GAS COMPONENTS						
	FIELD	JUBILEE/ TEN		SANKOFA		AIS/S	AIS/S mix
	Component	A %	mmscfd	S %	mmscfd	mmscfd	xi
N2	Nitrogen	0.39	0.440	0.44	0.650452	1.090	0.00418
CO2	Carbon Dioxide	0.91	1.026	0.42	0.620886	1.646	0.0063
C1	Methane	87.42	98.522	91.41	135.1314	233.654	0.8968
C2	Ethane	5.84	6.582	4.21	6.223643	12.805	0.0492
C3	Propane	4.09	4.609	2.27	3.355741	7.965	0.0306
iC4	i-Butane	0.43	0.485	0.34	0.502622	0.987	0.0038
nC4	n-Butane	0.78	0.879	0.61	0.901763	1.781	0.0068
iC5	i-Pentane	0.08	0.090	0.14	0.206962	0.297	0.0011
nC5	n-Pentane	0.04	0.045	0.11	0.162613	0.208	0.0008
C6	Hexane	0.01	0.011	0.01	0.014783	0.026	0.0001
	TOTAL	100.0	112.7	100.0	147.8	260.5	1.0

In accordance with ISO 6976:2016, the results of the computation of the commingled gas composition were utilised as the final gas composition at the gas terminal stations to determine the heating value of natural gas under standard reference circumstances.

Table 4.2 shows the results obtained in calculating the heating value and other gas parameters. The heating values were calculated in three bases namely Mass basis, Molar Basis, and Volume basis; the focus was on the volume basis. A factor of 1.008 was used as compensation for the heating value (Volume basis). Equation 4.1 was used as a conversion factor in converting the heating value from MJ/m^3 to BTU/SCF.

$$1MJ/m3 = 26.83BTU/SCF \quad (4.1)$$

Table 4.2 Results for Heating Value Calculation using ISO 6976:2016

			Calculation of Calorific Value and other properties of Natural Gas at metering constant of 20°C										
	Enter Value Here (Mole%)	Component	Mole Fraction (X)	Summation Factor (B)	X*B	Gross Calorific Value MJ/m^3(H1)	X*H1	Gross Calorific Value KJ/mol(H2)	X*H2	Gross Calorific Value MJ/Kg(H3)	X*H3	Molecular Weight Kg/Kmol (M)	X*M
	0.41837	Nitrogen	0.00418	0.0173	0.00007238	**0.0000**	0	**0.0000**	0	**0.0000**	0	**28.0135**	0.1172
	0.63196	Carbon Dioxide	0.00632	0.0748	0.00047271	**0.0000**	0	**0.0000**	0	**0.0000**	0	**44.0098**	0.2781
	89.684	Methane	0.89684	0.0447	0.04008875	**37.7060**	33.81625	**891.5800**	799.604668	**55.5740**	43.517796	**16.0428**	14.3878
	4.9151	Ethane	0.04915	0.0922	0.00453173	**66.0660**	3.247213	**1562.1400**	76.7808209	**51.9500**	4.1787468	**30.0696**	1.4780
	3.0573	Propane	0.03057	0.1338	0.00409066	**93.9340**	2.87184	**2221.1000**	67.9055821	**50.3700**	3.6958847	**44.0970**	1.3482
	0.37893	i-Butane	0.00379	0.1789	0.00067791	**121.4000**	0.460024	**2870.5800**	10.8775513	**49.3900**	0.5920407	**58.1234**	0.2202
	0.68354	n-Butane	0.00684	0.1871	0.00127890	**121.7900**	0.832482	**2879.7600**	19.6842699	**49.5500**	1.0714151	**58.1234**	0.3973
	0.11405	i-Pentane	0.00114	0.2280	0.00026002	**149.3600**	0.170338	**3531.7000**	4.02773488	**48.6300**	0.2177799	**72.1500**	0.0823
	0.07972	n-Pentane	0.00080	0.2510	0.00020010	**149.6500**	0.1193	**3538.6000**	2.82095133	**49.0400**	0.1535151	**72.1500**	0.0575
	0.01	Hexane	0.00010	0.2510	0.00002510	**149.6500**	0.014965	**3538.6000**	0.35386	**48.7200**	0.0191313	**72.1500**	0.0072
Total	99.9730		0.9997		0.05169826		41.53241		982.055439		53.446309		18.3738

M_{air} = 28.9626 Kg/Kmol	
Z_{air} (293.15K, 101.325KPa	0.99963

Compressibility (Z_{mix}) =	**0.99732729**				**Relative Density of Real Gas**	**0.6358636**	
Superior Calorific Value on volumetric basis =	**41.5324**	**MJ/m^3(H) or**	**1114.696**	**Btu/ft^3**	**Density of Real Gas**	**0.7791545**	**Kg/m^3**
Superior Calorific Value on mass basis =	**53.4463**	**MJ/Kg**			**Wobbe Index of Real Gas**	**52.22368**	**MJ/m3**
Superior Calorific Value on molar basis =	**982.0554**	**KJ/mol or**	**931.4321**	**BTU/mol**			
					HHV with compensation	**1123.614**	**Btu/ft3**

4.2 Prediction Models

The results obtained from the models in the Google Colab Notebook are presented here for further discussion.

4.2.1 Pre-Processing Stage

The data for the prediction was first explored to check for outliers, statistical description of the data, missing values checks, correlation analysis, and multicollinearity were all done to ascertain the quality of the dataset for better prediction using Google Colab with python as the coding language.

Statistical Description

Table 4.3 shows the statistical description of the dataset used in the prediction of the heating value of natural gas. The statistical description indicates the lowest value, mean value, standard deviation, maximum value, 50th percentile or median, upper quartile (75th percentile), lower quartile (25th percentile), and the total number of the dataset, which was 2021. The maximum value in the dataset for heating value was 1143.27 and the minimum value was 1014.73, with a mean and standard deviation of 1122.318095 and 12.776741 respectively.

Table 4.3 Statistical Description of Dataset for Prediction

	C1	C2	C3	IC4	NC4	IC5	NC5	C6+	N2	CO2	HHV
count	2021.00	2021.00	2021.00	2021.00	2021.00	2021.00	2021.00	2021.00	2021.00	2021.00	2021.00
mean	88.46	5.93	3.12	0.34	0.62	0.12	0.10	0.07	0.42	0.81	1122.33
std	0.66	0.41	0.22	0.02	0.04	0.01	0.01	0.01	0.01	0.08	12.78
min	86.89	0.00	2.29	0.28	0.47	0.04	0.03	0.01	0.40	0.43	1014.73
25%	88.05	5.85	3.01	0.33	0.60	0.11	0.10	0.07	0.42	0.80	1119.90
50%	88.47	5.98	3.10	0.34	0.61	0.12	0.10	0.07	0.42	0.82	1122.74
75%	88.65	6.15	3.28	0.35	0.63	0.13	0.11	0.08	0.43	0.85	1127.54
max	91.97	6.75	4.96	0.44	0.78	0.16	0.13	0.11	0.47	1.57	1143.27

Outliers Determination in Dataset

Outliers have a major effect on prediction effectiveness. A histogram was developed to get hint of how the data is distributed. In the histogram, the data for each predictor and dependent variable (Heating Value) did not follow a normal distribution which showed either positive skewness or negative skewness. Figure 4.1 is the description of outliers in the dataset for the prediction of heating value. From Figure 4.1, HHV, CO_2, N_2, C_1, C_2, and C_3 has a large number of outliers in the dataset but IC_4, NC_4, IC_5, and NC_5 have small outliers in the data. Further processing was done to remove the outliers in the dataset for better prediction, the result will be presented in the Data Processing Section of this thesis.

Figure 4.1 Histogram for Dataset Explaining Outliers

Correlation Analysis

A correlation matrix was developed using Colab Notebook Software to determine the correlation that exists between the predictors and the dependent variable used in the prediction.

This is to show whether there's a positive or negative correlation between them. Figure 4.2 shows the correlation matrix obtained from the study. From the matrix, the correlation between CO_2, N_2, C_1, C_2, C3, iC_4, nC4, iC_5, nC_5, C_6+ and HHV are 0.28, -0.34, 0.18, 0.30, 0.29, 0.41, 0.38, 0.35, 0.32, -0.39 respectively. It can be inferred that CO_2, C_2, C_3, IC_4, C_6+, IC_5, nC_4 and nC_5 have a positive correlation with the heating value and C1, and N2 have a negative correlation with the heating value. The correlation values show that N_2 and C_1 have a weakly negative correlation with HHV.

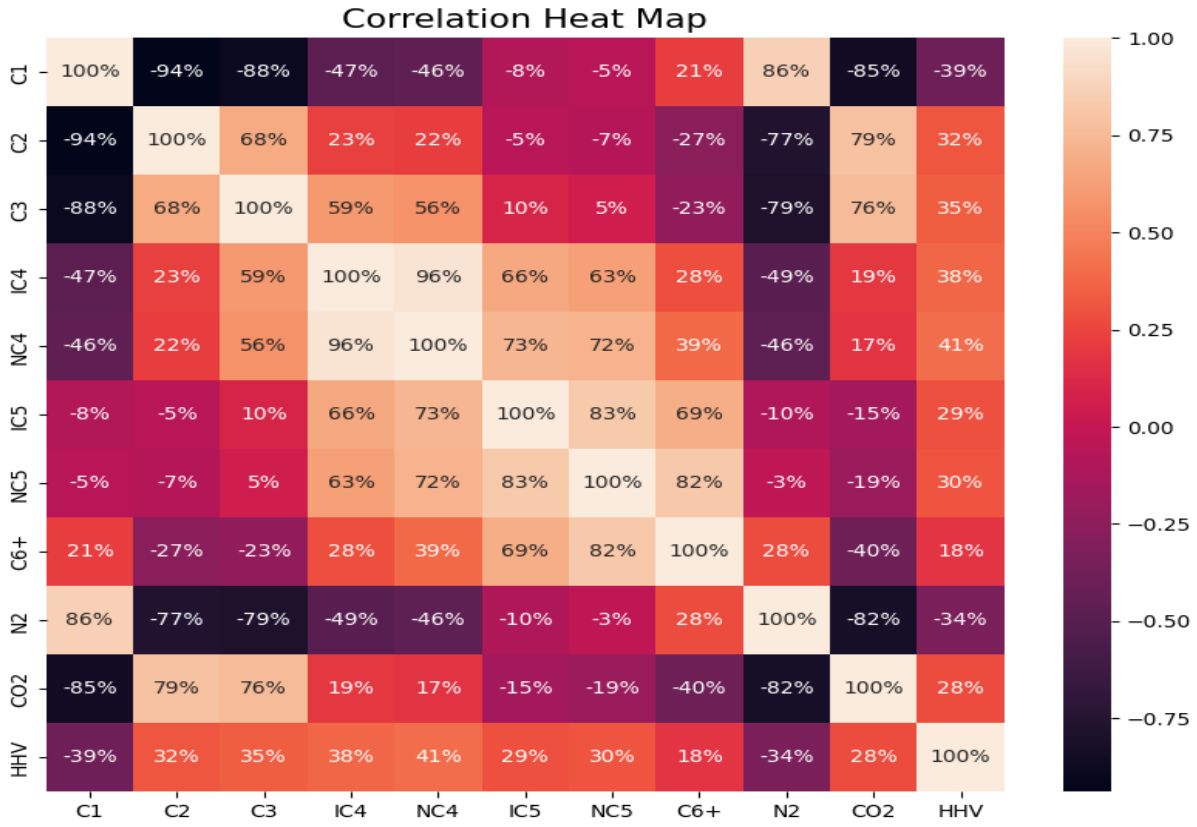


Figure 4.2 Correlation Matrix for Dataset

4.2.2 Data Processing

Data processing was done by selecting the input and output variable, checking for multicollinearity, detecting, and handling outliers for a better prediction using regression, data splitting into training and testing datasets, and normalisation using MinMax Scaler and Standard Scaler.

Input and Output Variable Selection

Table 4.4 shows a sample of the input or predictors used in the modelling. The output variable was heating value and was named "label" and the input variable in this project were CO_2, N_2, C_1, C_2, C_3, iC_4, nC_4, iC_5, nC_5, and C_6+ and were named "features" in the model.

Table 4.4 Input Variables for the Model

	C1	C2	C3	IC4	NC4	IC5	NC5	C6+	N2	CO2	HHV
0	88.80	5.74	3.00	0.33	0.60	0.12	0.10	0.08	0.43	0.80	119.80
1	88.88	5.75	2.92	0.33	0.59	0.12	0.10	0.08	0.43	0.80	1118.36
2	88.71	5.80	3.02	0.34	0.60	0.12	0.10	0.08	0.43	0.80	1120.71
3	88.68	5.80	3.04	0.34	0.61	0.12	0.10	0.08	0.43	0.80	1121.28
4	88.65	5.81	3.05	0.34	0.61	0.12	0.10	0.08	0.43	0.81	1121.59
…	…	…	…	…	…	…	…	…	…	…	…
2019	87.90	6.58	3.07	0.34	0.60	0.11	0.09	0.05	0.40	0.85	1125.71
2020	87.90	6.58	3.07	0.34	0.60	0.11	0.09	0.05	0.40	0.85	1125.71
2021	87.90	6.58	3.07	0.34	0.60	0.11	0.09	0.05	0.40	0.85	1125.71
2022	87.90	6.58	3.07	0.34	0.60	0.11	0.09	0.05	0.40	0.85	1125.71
2023	87.90	6.58	3.07	0.34	0.60	0.11	0.09	0.05	0.40	0.85	1125.71
	2024 rows X 11 columns										

Multicollinearity

Multicollinearity was checked among the predictors to achieve accurate prediction. This was done by using the Variance Inflation Factor (VIF). If the VIF is 1, the variables are not correlated; if it is between 1 and 5, the variables are moderately correlated; and if it is between 5 and 10, the variables are highly linked. Table 4.5 shows the VIF obtained for the predictors using Google Colab Software.

Table 4.5 VIF Scores of Predictors

Independent Features			VIF Scores
0		C1	13696.92
1		C2	593.04
2		C3	1310.56
3		IC4	4233.33
4		NC4	5653.67
5		IC5	449.24
6		NC5	875.56
7		C6+	229.38
8		N2	9327.98
9		CO2	617.54

From Table 4.5, it is observed that most of the variables have a high range of VIF scores which is above the normal range of VIF score for multicollinearity Although multicollinearity can skew the results of a model, it is also important to note that, its effect is not relevant if the model being developed is to be used solely for prediction purposes and the model parameter are not to be interpreted in a business context hence the impact would not be handled.

Detection and Handling of Outliers in Predictor Variables and Dependent Variables

Outliers have a major effect on prediction accuracy. For this study, outliers were determined within the predictors with the help of a box plot using Equations 3.9, 3.10, and 3.11. Figure 4.5 shows the box plot for the data with outliers. From Figure 4.3, points that were found outsides the box plot are termed outliers. Figure 4.4 shows the exact number of outliers in each predictor variable. The outliers in the predictor variables and dependent variables were removed using the method of imputation, this is where every outlier in each predictor variable and dependent variables were replaced by the mean value of that variable. Figure 4.5 shows the box plot obtained after handling the outliers. The data description for features before and after outliers was removed are also presented in Tables 4.6 and 4.7 respectively. After imputation, it was seen that the values for each predictor were scaled within a defined limit. For instance, before imputation, CO_2 values were between 0.43 and 1.57 but after imputation, the values were between 0.73 and 0.92. Table 4.8 also shows the data description for the dependent variable (HHV) after imputation.

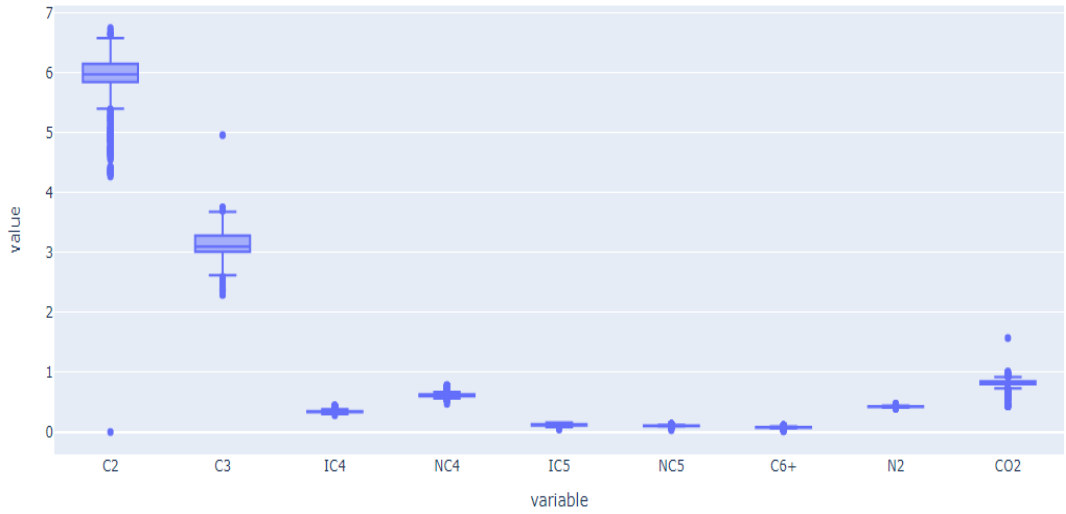


Figure 4.3 Detecting Outliers in the Feature (Predictor Variables)

```
C1
number of outliers:158
max outlier value:91.97
min outlier value:86.89
========================================
C2
number of outliers:162
max outlier value:6.75
min outlier value:0.0
========================================
========================================
C3
number of outliers:43
max outlier value:4.96
min outlier value:2.29
========================================
========================================
NC5
number of outliers:150
max outlier value:0.13
min outlier value:0.03
========================================
C6+
number of outliers:189
max outlier value:0.11
min outlier value:0.01
========================================
```

```
=======================================
IC4
number of outliers:214
max outlier value:0.44
min outlier value:0.28
=======================================
NC4
number of outliers:237
max outlier value:0.78
min outlier value:0.47
=======================================
IC5
number of outliers:2
max outlier value:0.07
min outlier value:0.04
=======================================
N2
number of outliers:155
max outlier value:0.47
min outlier value:0.4
=======================================
CO2
number of outliers:304
max outlier value:1.57
min outlier value:0.43
=======================================
HHV
number of outliers:74
max outlier value:1143.27
min outlier value:1014.73
=======================================
```

Figure 4.4 Number of Outliers in Each Predictor Variable

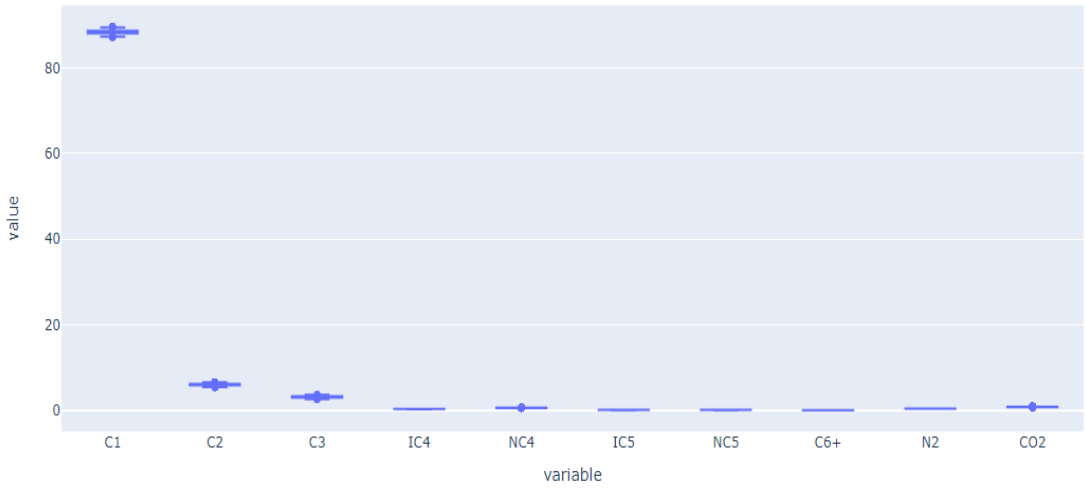


Figure 4.5 Distribution of Features without Outliers

Table 4.6 Data Description for Predictor Variables before Outliers were Removed

	C1	C2	C3	IC4	NC4	IC5	NC5	C6+	N2	CO2
count	2021.00	2021.00	2021.00	2021.00	2021.00	2021.00	2021.00	2021.00	2021.00	2021.00
mean	88.46	5.93	3.12	0.34	0.62	0.12	0.10	0.07	0.04	0.81
std	0.66	0.41	0.22	0.02	0.04	0.01	0.01	0.01	0.01	0.08
min	86.89	0.00	2.29	0.28	0.47	0.04	0.03	0.01	0.40	0.43
25%	88.05	5.85	3.01	0.33	0.60	0.11	0.10	0.07	0.42	0.80
50%	88.47	5.98	3.10	0.34	0.61	0.12	0.10	0.07	0.42	0.82
75%	88.65	6.15	3.28	0.35	0.63	13.00	0.11	0.08	0.43	0.85
max	91.97	6.75	4.96	0.44	0.78	0.16	0.13	0.11	0.47	1.57

Table 4.7 Data Description for Predictor Variables after Imputation

	C1	C2	C3	IC4	NC4	IC5	NC5	C6+	N2	CO2
count	2021.00	2021.00	2021.00	2021.00	2021.00	2021.00	2021.00	2021.00	2021.00	2021.00
mean	88.35	6.01	3.13	0.34	0.61	0.12	0.10	0.07	0.42	0.83
std	0.44	0.21	0.19	0.01	0.02	0.01	0.01	0.01	0.01	0.03
min	87.17	5.40	2.62	0.31	0.56	0.08	0.09	0.06	0.41	0.73
25%	88.06	5.89	3.02	0.33	0.60	0.11	0.10	0.07	0.42	0.81
50%	88.46	5.98	3.11	0.34	0.61	0.12	0.10	0.07	0.42	0.82
75%	88.61	6.15	3.28	0.35	0.62	0.13	0.11	0.08	0.43	0.84
max	89.55	6.58	3.68	0.37	0.67	0.16	0.12	0.09	0.44	0.92

Table 4.8 Data Description for dependent variable (HHV) after imputation

	HHV
count	2021.000000
mean	1123.625338
std	5.548656
min	1108.500000
25%	1120.300000
50%	1122.650000
75%	1127.400000
max	1138.620000

4.2.3 Model Development and Evaluation

The models were developed and evaluated using line plots, metrics, and scatter plots with a line of best fit, and feature importance in each model. The models used were the Linear Regression Model, Bagging Regressor Model, Artificial Neural Networks, Extreme Gradient Boosting Model, AdaBoost Model, Random Forest Model and Stacking Regressor.

Linear Regression

For the dataset, a linear regression model was created. The model's output is reported in this part as results. Figure 4.8 displays the line plot created using linear regression for the actual and anticipated heating levels.

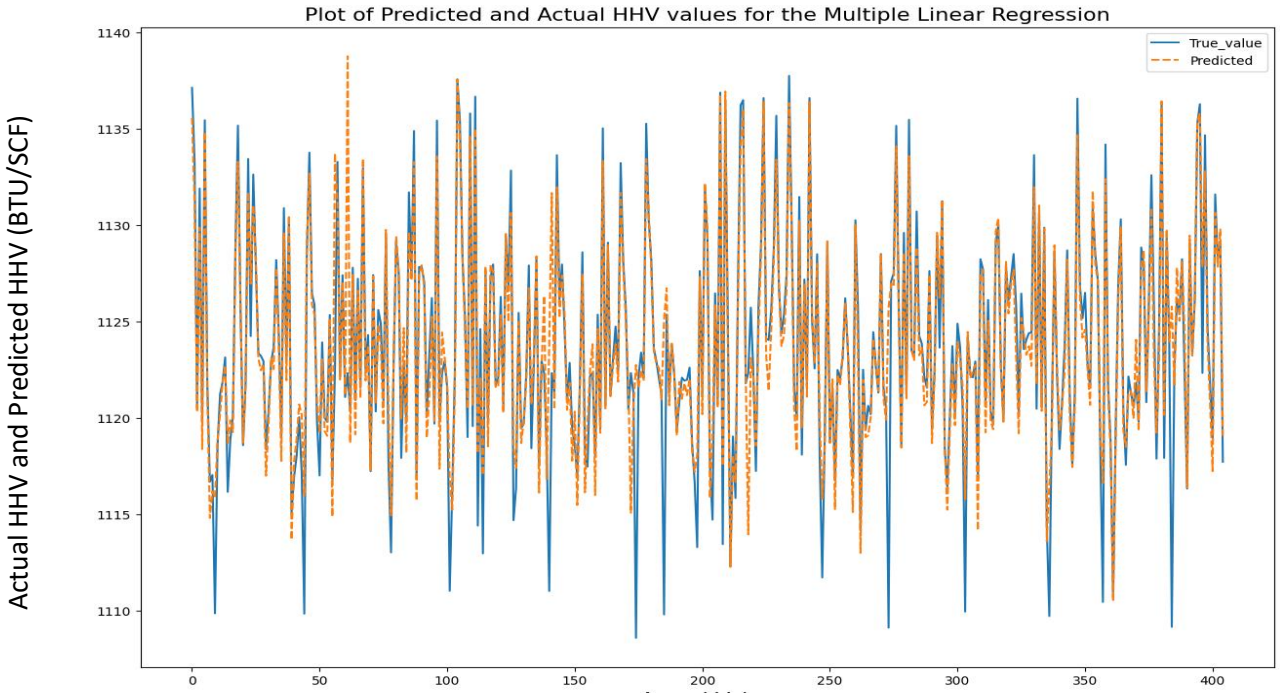


Figure 4.6 Line Plot for Actual and Predicted HHV in Multiple Linear Regression

From Figure 4.6, the model was not able to predict heating values lower than 1 110 BTU/SCF. As a result of this, the error margin for this model increased. Figure 4.7 shows the scatter plot for Actual HHV and Predicted HHV in the Linear Regression Model. A line of best fit was drawn with an R^2 value determined. An equation for prediction was also generated from the model. Equation 4.2 is the linear equation generated for the linear regression model.

$$HHV = 1116.7221 - 2.8079 \times C1 - 1.4472 \times C2 + 5.2744 \times C3 + 1.0681 \times IC4 - 1.1064 \times$$

$$NC4 + 4.2370 \times IC5 + 0.2595 \times NC5 \ + 2.8699 \times C6 + + 0.5632 \times N2 - 0.6045 \times C02 \quad (4.2)$$

Where (CO_2, N_2, C1…..C6) is the composition of the gas sample.

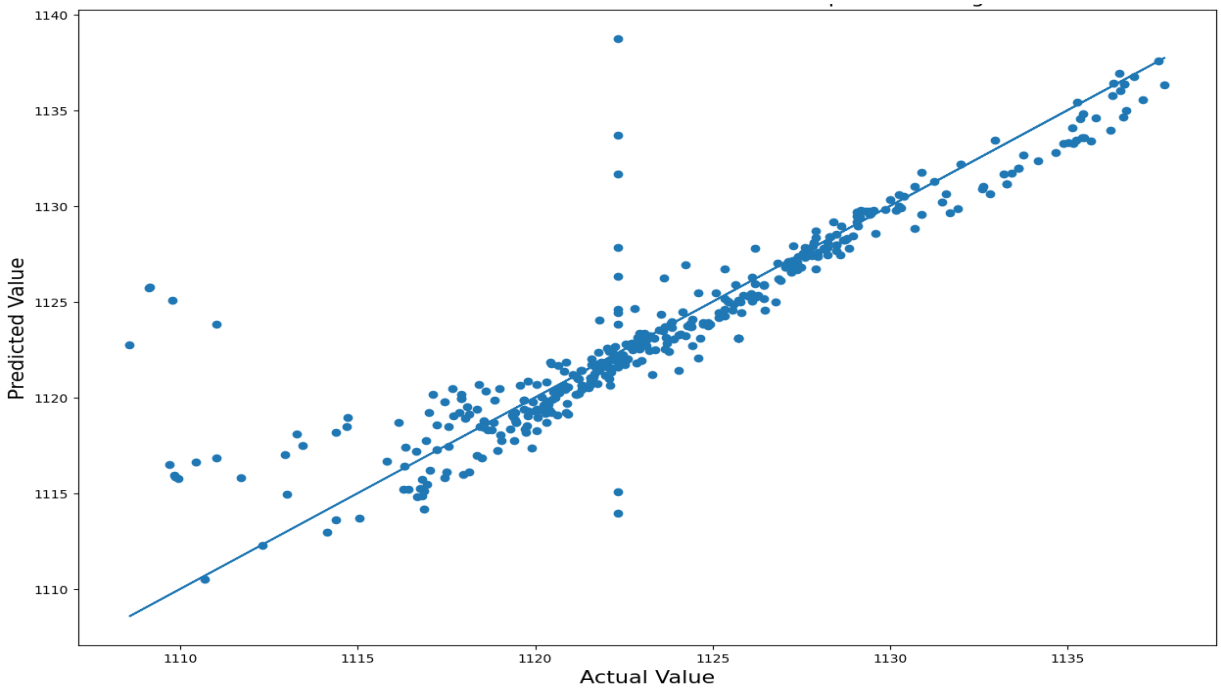


Figure 4.7 Scatter Plot for Actual HHV and Predicted HHV in Linear Regression

From Figure 4.7, it is seen that most of the values were scattered, and this led to a large difference between the predicted HHV and the actual HHV. Table 4.9 shows the metric values obtained for Linear Regression Model for both the training and testing dataset.

Table 4.9 Training and Testing Results for Linear Regression Model

Linear Regression	RMSE	MSE	MAE	Adjusted R^2	MAPE	R^2
Training	2.0116	4.0466	1.1343	0.8641	0.53%	0.8650
Testing	2.5343	6.4224	1.2971	0.8055	0.55%	0.8108

From Table 4.9, the errors for the training dataset were lower than that of the testing dataset. In

the training, An R^2 of 86.50% was recorded which shows that the predictor variables were able to explain 86.50% of the variations in the output variable (Heating Value), whereas in the testing the value of R^2 decreased to 81.08% which shows that the model developed can only explain about 81.08% of the output variable.

Random Forest Regression

A random Forest Regression model was developed for the dataset. The results obtained from the model are presented in this section. The line plot produced by the Random Forest Regression Model for the actual and anticipated heating levels is shown in Figure 4.10.

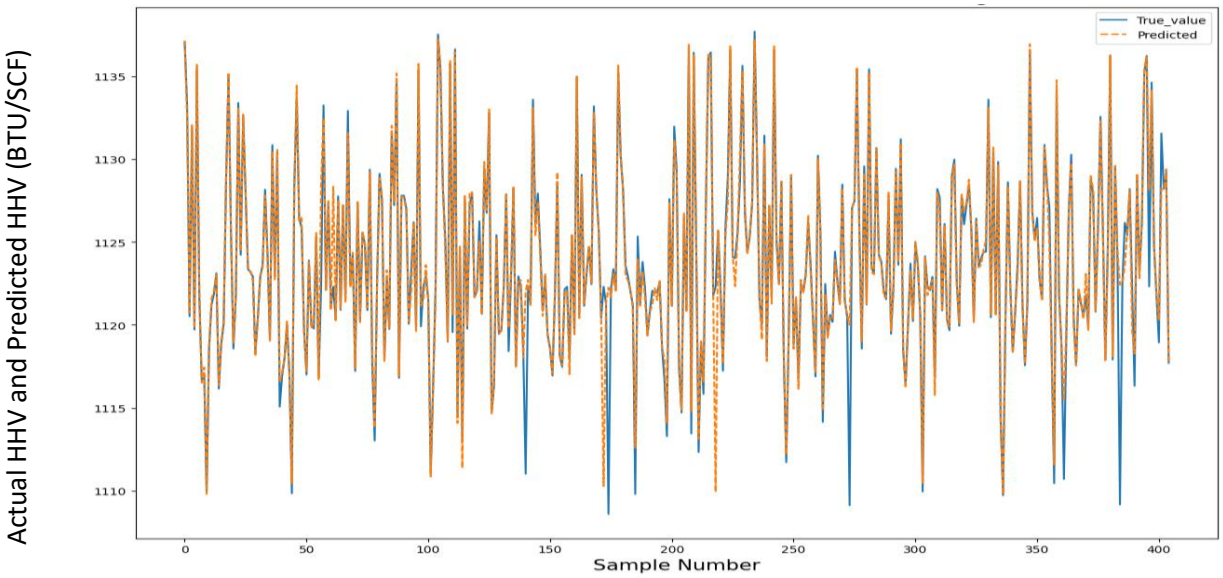


Figure 4.8 Line Plot for Actual and Predicted HHV in Random Forest Regression Model

From Figure 4.8, the predicted values and the actual values are very close together which shows a high connection between them, however, this model was not able to predict heating values less than 1 110BTU/SCF. Figure 4.9 shows the scatter plot for Actual HHV and Predicted HHV in the Random Forest Regression Model. A line of best fit was drawn with an R^2 value determined. From Figure 4.9, most of the points are on the line of best fit compared to that of Linear Regression and hence the error in Random Forest is less than that of Linear Regression.

Table 4.10 shows the metric values obtained for the Random Forest Regression Model for both the training and testing datasets.

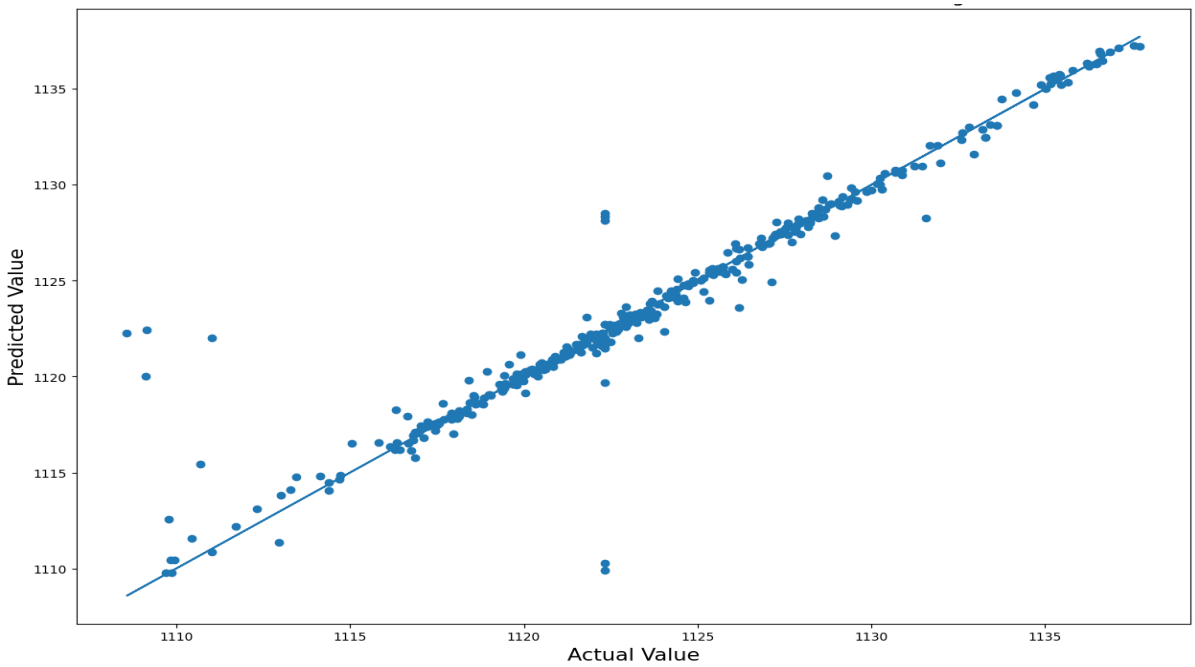


Figure 4.9 Scatter Plot for Actual HHV and Predicted HHV in Random Forest Regression

Table 4.10 Training and Testing Results for Random Forest Regression Model

Random Forest	RMSE	MSE	MAE	Adjusted R^2	MAPE	R^2
Training	0.5411	0.2928	0.1847	0.9902	0.54%	0.9902
Testing	1.6821	2.8295	0.5517	0.9143	0.57%	0.9166

According to Table 4.10, the training dataset's errors were fewer than those for the testing dataset. An R2 of 99.02% was obtained during training, indicating that the predictor variables could account for 99.02% of the fluctuations in the output variable (Heating Value). This indicates that the model did quite well during training, whereas in the testing the value of R^2 decreased to 91.66% which shows that the model developed can only explain about 91.66% of the output variable which is better than that of the Linear Regression.

Figure 4.10 shows the feature importance in the Random Forest Model. It can be seen from Figure 4.10 that, C3 had the highest importance in the prediction of heating value for the Random Forest Model whereas NC5 had the least importance.

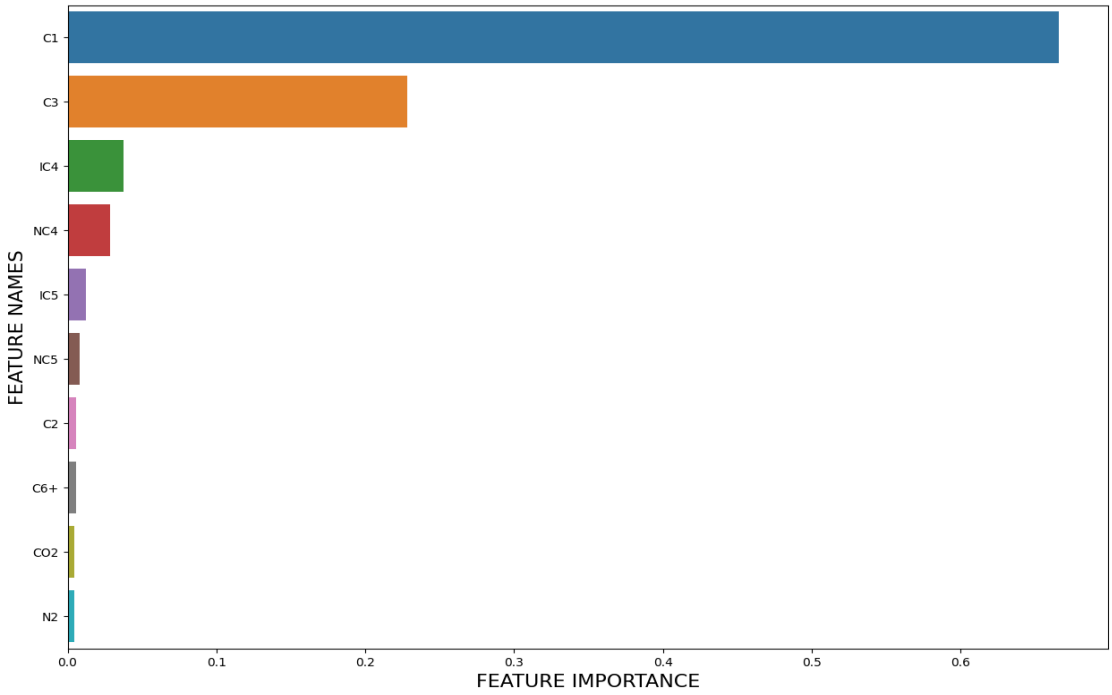


Figure 4.10 Feature Importance for the Random Forest Model

AdaBoost Regression

AdaBoost Regression model was developed for the dataset. The results obtained from the model are presented in this section. Figure 4.11 shows the line plot obtained for the actual heating value and predicted heating value using AdaBoost Regression Model

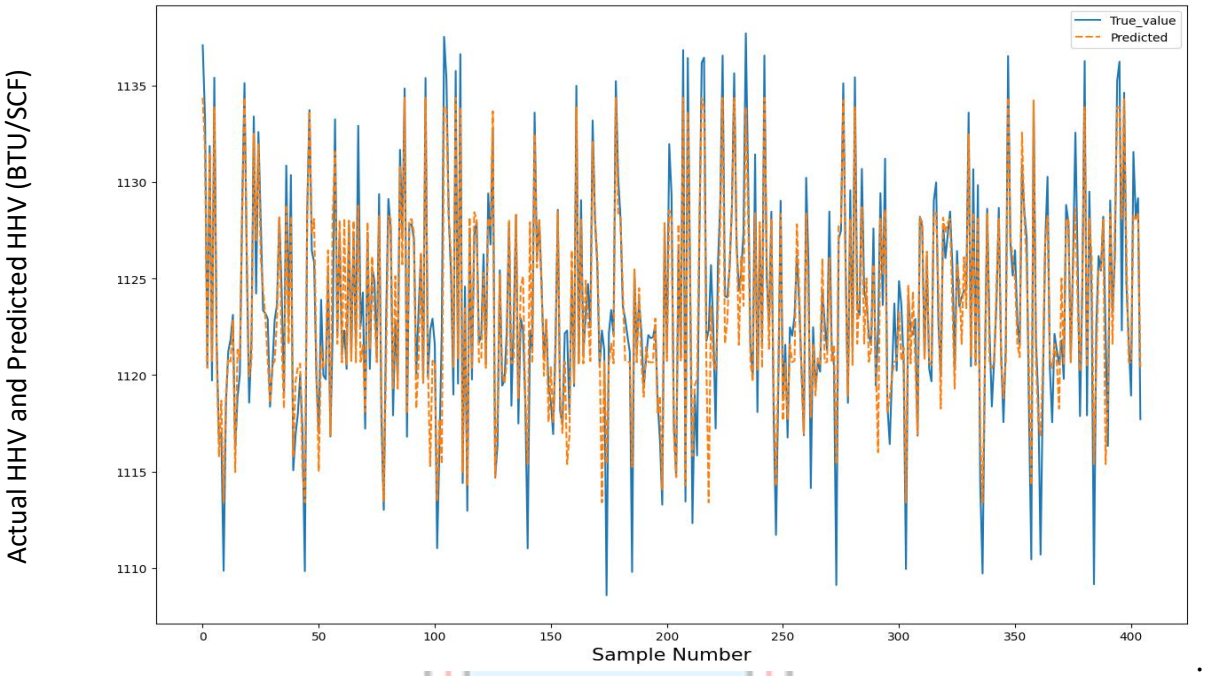


Figure 4.11 Line Plot for Actual and Predicted HHV in AdaBoost Regression Model

From Figure 4.11, the predicted values and the actual values were not very close together as compared to the Random Forest Model and this resulted in a high error margin between the predicted HHV and the actual HHV. Furthermore, the model was unable to forecast heating levels lower than 1110 BTU/SCF.

Figure 4.12 shows the scatter plot for Actual HHV and Predicted HHV in the AdaBoost Regression Model. A line of best fit was drawn with an R^2 value determined. From Figure 4.12, it is evident that most of the points are away from the line of best fit hence there is a high error since the predicted HHV and the actual HHV are not close to each other.

Table 4.11 shows the metric values obtained for the AdaBoost Regression Model for both the training and testing dataset.

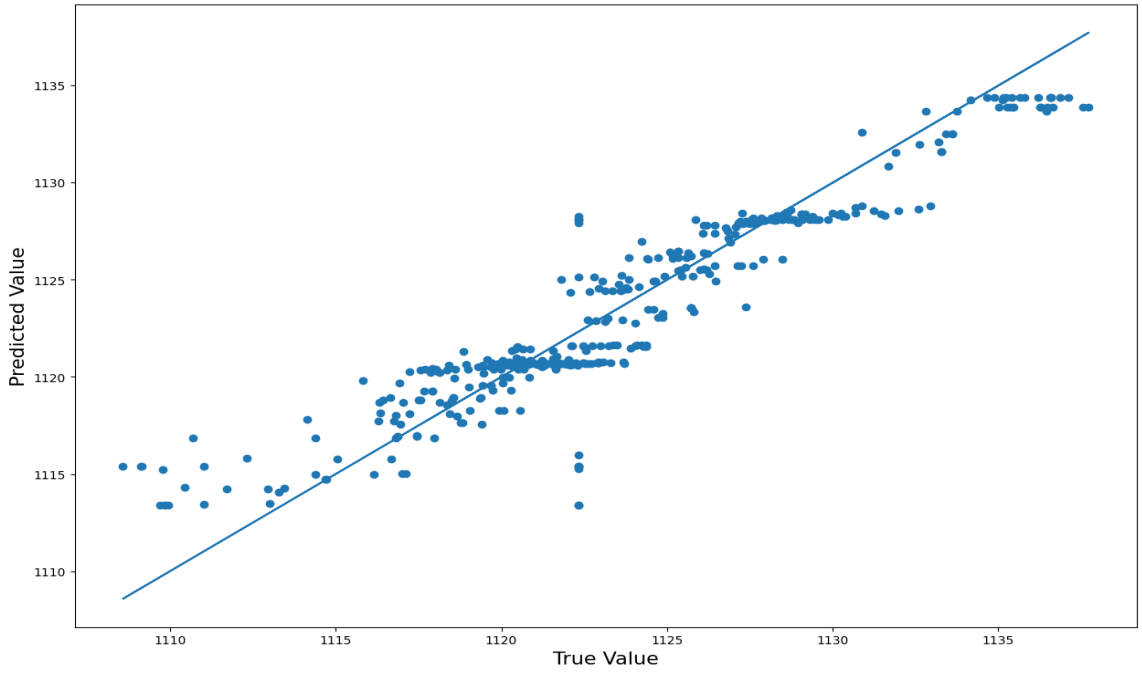


Figure 4.12 Scatter Plot for Actual HHV and Predicted HHV in AdaBoost Regression

Table 4.11 Training and Testing Results for AdaBoost Regression Model

AdaBoost Regression	RMSE	MSE	MAE	Adjusted R^2	MAPE	R^2
Training	1.9521	3.8105	1.4356	0.8720	0.52%	0.8729
Testing	2.0230	4.0926	1.4559	0.8761	0.55%	0.8794

From Table 4.11, the errors for the training dataset were lower than that of the testing dataset. In the training, an R^2 of 87.29% was recorded which shows that the predictor variables were able to explain 87.29% of the variations in the output variable (heating value) in the training of the model and this depicts that the model performed very well in the training, whereas in the testing the value of R^2 increased to 87.94% which shows that the model developed can only explain about 87.94% of the output variable which is better than that of the Linear Regression.

Figure 4.13 shows the feature importance of the AdaBoost Regression Model. It can be seen from Figure 4.13 that, C3 had the highest importance in the prediction of heating value for the AdaBoost Model whereas CO2 had the least importance.

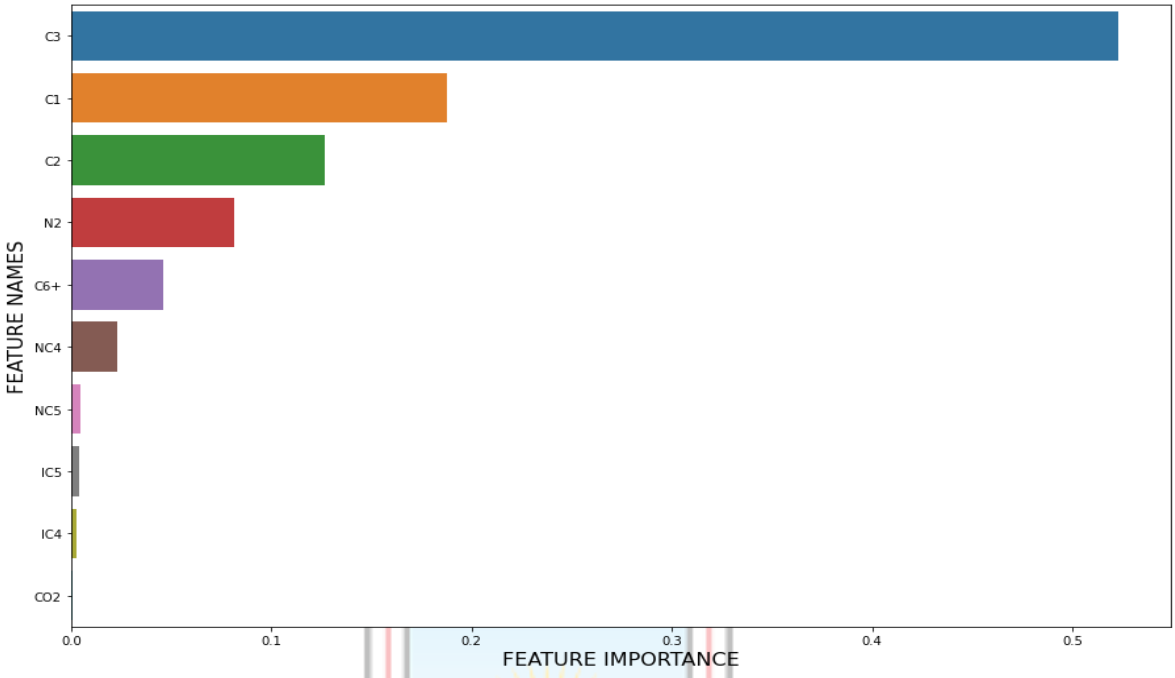
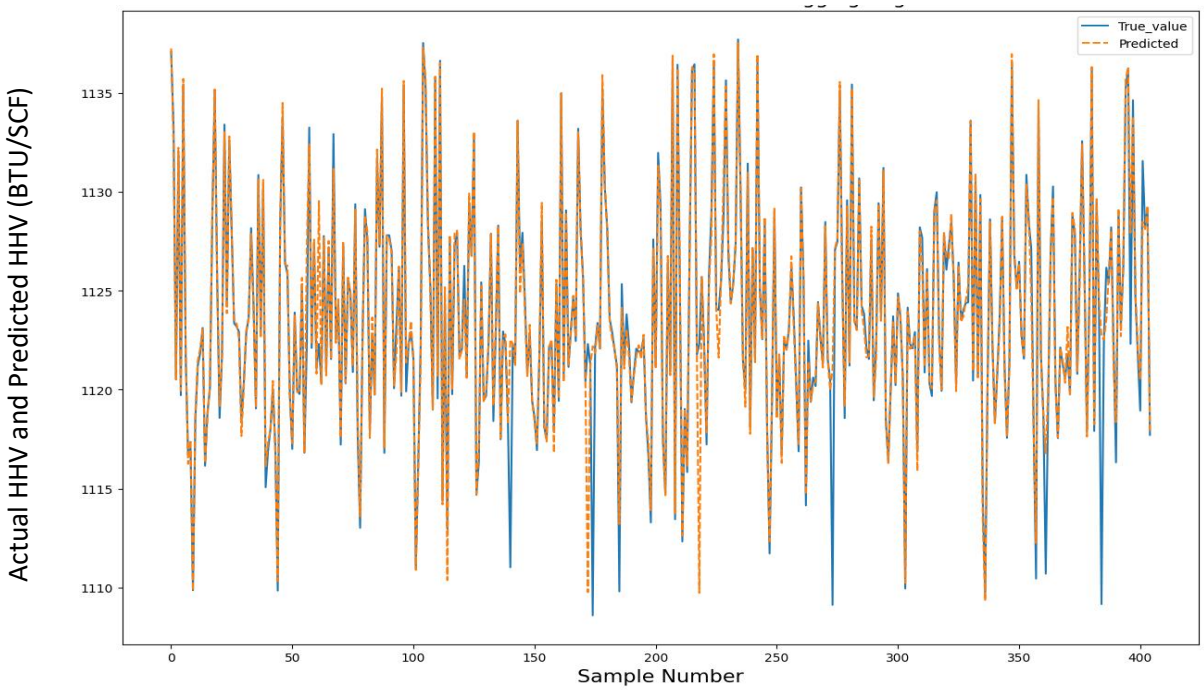


Figure 4.13 Feature Importance for the AdaBoost Regression Model

Bagging Regressor

A bagging Regressor model was developed for the dataset. The model's output is reported in this part as results. The line plot produced using the Bagging Regressor Model for the actual and anticipated heating levels is shown in Figure 4.14.



.

Figure 4.14 Line Plot for Actual and Predicted HHV in Bagging Regressor Model

From Figure 4.14, the predicted values and the actual values were a little close to each other as compared to the AdaBoost Regression Model and this resulted in a better model prediction than that of AdaBoost Regression. However, this model was not able to predict heating values less than 1 110 BTU/SCF.

Figure 4.15 shows the scatter plot for Actual HHV and Predicted HHV in the Bagging Regressor Model. A line of best fit was drawn with an R^2 value determined. From Figure 4.15, it is seen that most of the points lie on the line of best fit which in turn increased the R^2 for this model and hence made it a better predictor.

Table 4.12 shows the metric values obtained for the Bagging Regressor Model for both the training and testing dataset.

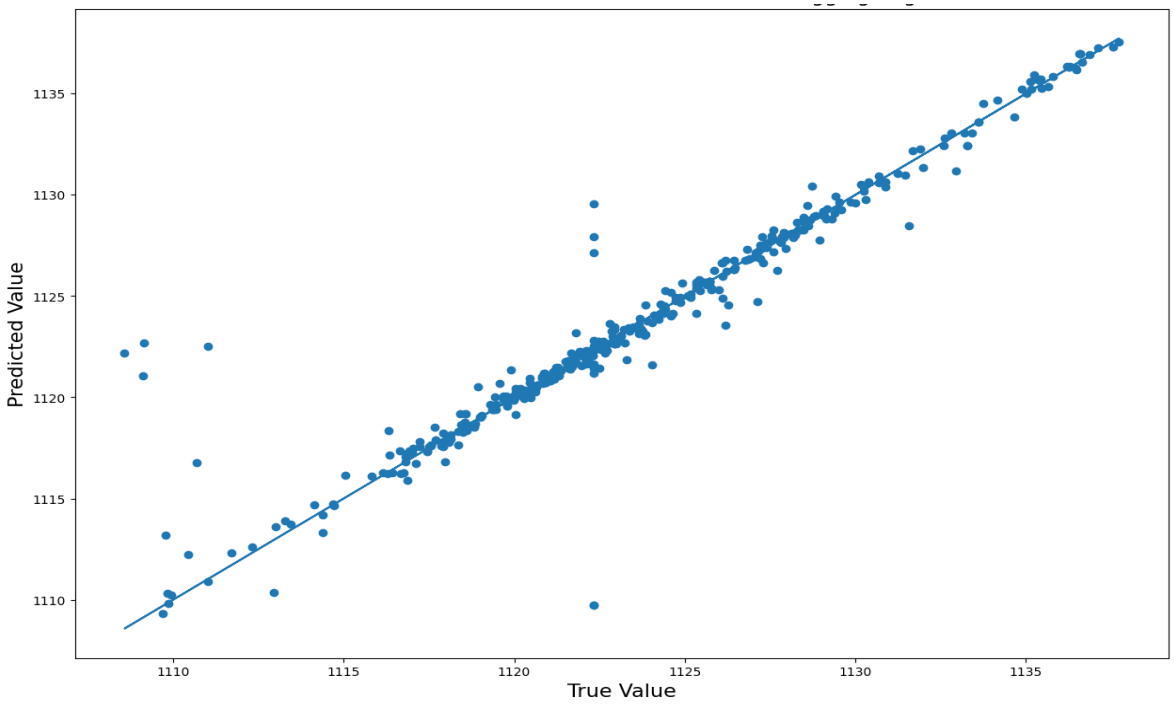


Figure 4.15 Scatter Plot for Actual HHV and Predicted HHV in Bagging Regressor Model

Table 4.12 Training and Testing Results for Bagging Regressor Model

Bagging Regressor	RMSE	MSE	MAE	Adjusted R^2	MAPE	R^2
Training	0.6176	0.3815	0.2062	0.9872	0.54%	0.9872
Testing	1.7434	3.0395	0.5837	0.9079	0.57%	0.9105

From Table 4.12, the errors for the training dataset were lower than that of the testing dataset. In the training, An R^2 of 98.72% was recorded which shows that the predictor variables were capable of explaining 98.72% of the variations in the output variable (heating value) during the training of the model and this concludes that the model performed very well in the training than in testing, whereas in the testing the value of R^2 decreased to 91.05% which shows that the model developed can only explain about 91.05% of the output variable which is better than the prediction model for AdaBoost Regression.

Extreme Gradient Boosting Regressor Model

XGBoost Regressor model was developed for the dataset. The results obtained from this model are presented in this section. Figure 4.16 shows the line plot obtained for the actual heating value and predicted heating value using XGBoost Regressor Model.

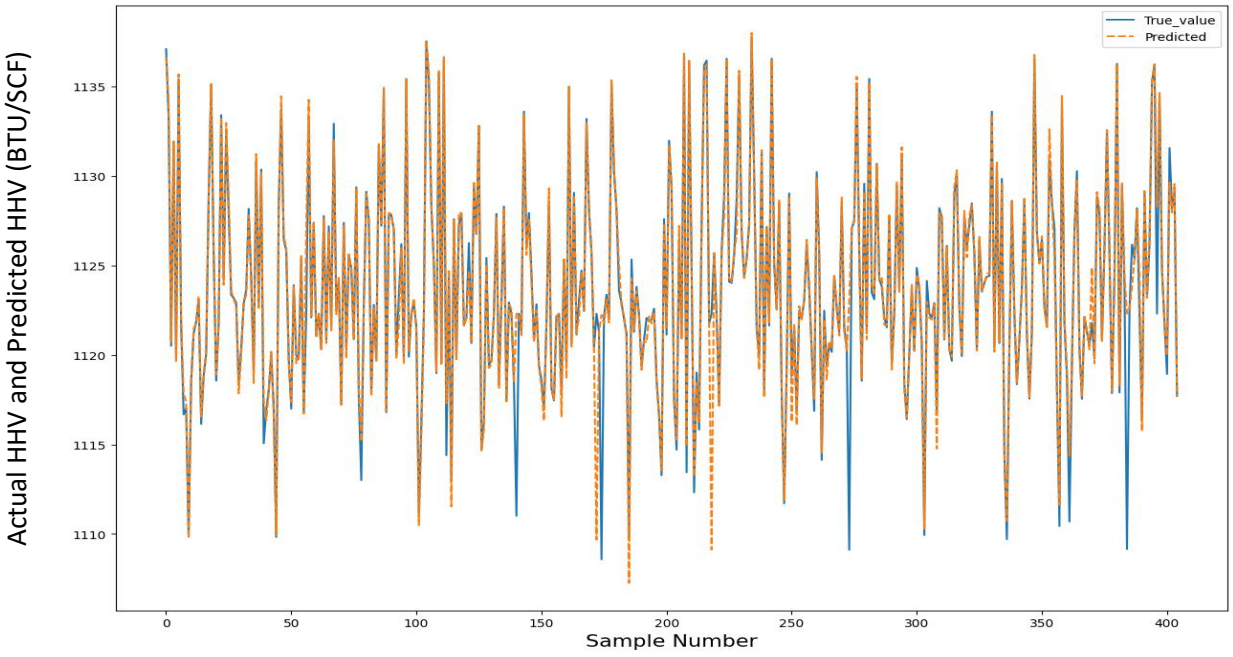


Figure 4.16 Line Plot for Actual and Predicted HHV in XGBoost Regressor Model

From Figure 4.16, the predicted values and the actual values were a little close together as compared to the AdaBoost Regression and Linear Regression and this resulted in a better prediction for this model than AdaBoost and Linear Regression models with comparatively small errors. However, this model was not able to predict heating values less than 1 110 BTU/SCF because the dataset contained values above that and practically on field, the average heating value recorded was above 1 110 BTU/SCF.

Figure 4.17 shows the scatter plot for Actual HHV and Predicted HHV in the AdaBoost Regression Model. A line of best fit was drawn with an R^2 value determined. From Figure 4.17, most points lay on the line of best fit which in turn increased the R^2 for this model and hence

made it a better predictor.

Table 4.13 shows the metric values obtained for the XGBoost Regressor Model for both the training and testing dataset.

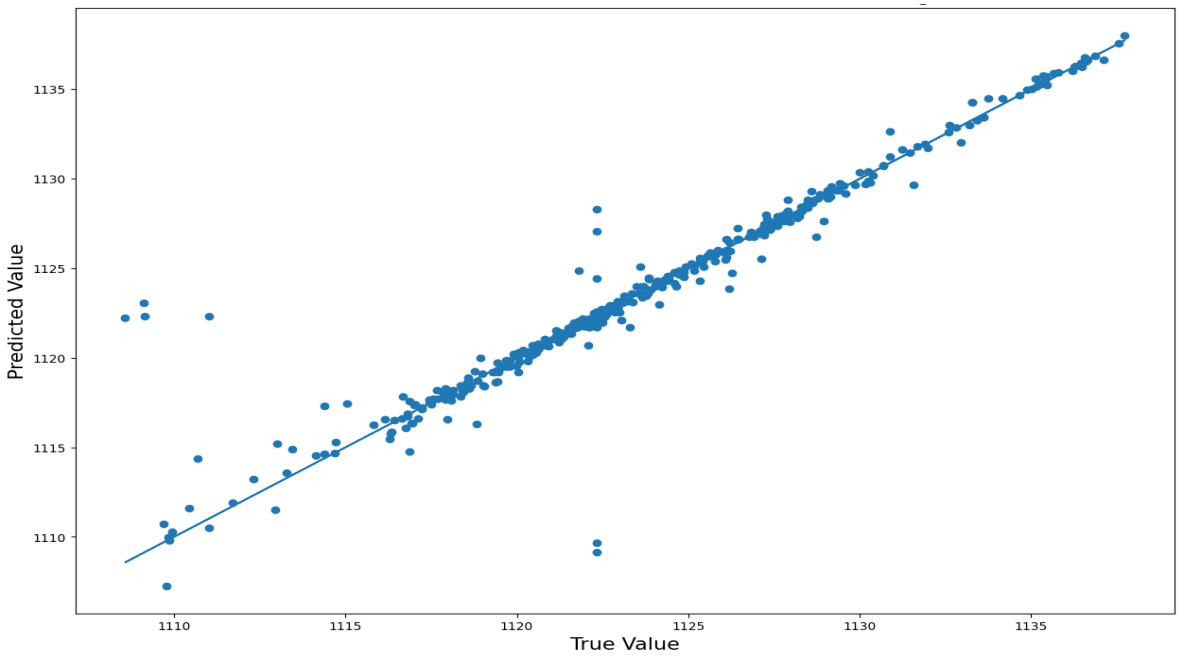


Figure 4.17 Scatter Plot for Actual HHV and Predicted HHV in XGBoost Regressor Model

Table 4.13 Training and Testing Results for XGBoost Regressor model

XGBoost Regressor	RMSE	MSE	MAE	Adjusted R^2	MAPE	R^2
Training	0.2761	0.0763	0.0234	0.9974	0.54%	0.9975
Testing	1.7302	2.9934	0.5393	0.9093	0.57%	0.9118

From Table 4.13, the errors for the training dataset were lower than that of the testing dataset. In the training, an R^2 of 99.75% was recorded which shows that the predictor variables were able to explain 99.75% of the variations in the output variable (heating value) in the training of the model and this means that the model performed very well in the training, whereas in the testing the value of R^2 decreased to 91.18% which shows that the model developed can only explain about 91.18% of the output variable which is better than that of the AdaBoost Regressor

Model.

Figure 4.18 shows the feature importance in the XGBoost Regressor Model. It can be seen from Figure 4.18 that, C3 had the highest importance in the prediction of heating value for the XGBoost Model whereas CO_2 had the least importance.

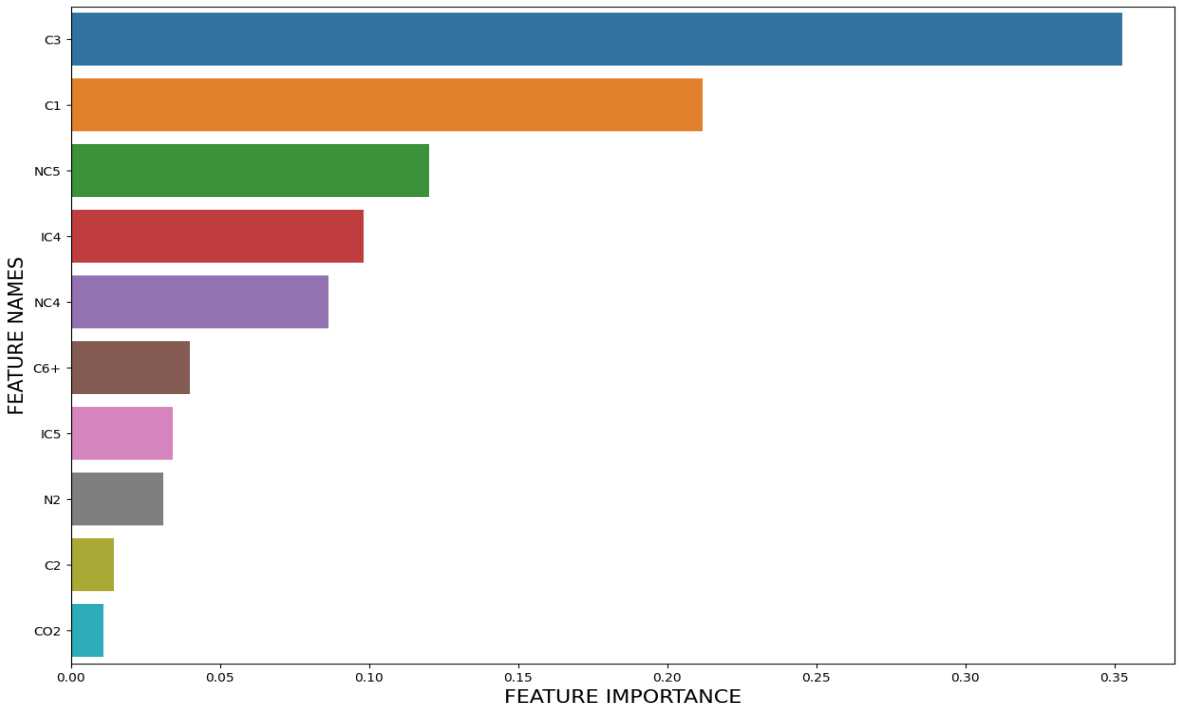


Figure 4.18 Feature Importance for the XGBoost Regressor Model

Hybrid Model (Stacking Regressor)

Hybrid model was developed for the dataset. The results obtained from this model are presented in this section. Figure 4.19 shows the line plot obtained for the actual heating value and predicted heating value using Hybrid Model achieved using the Stacking Regressor from the mlxtend.regressor package.

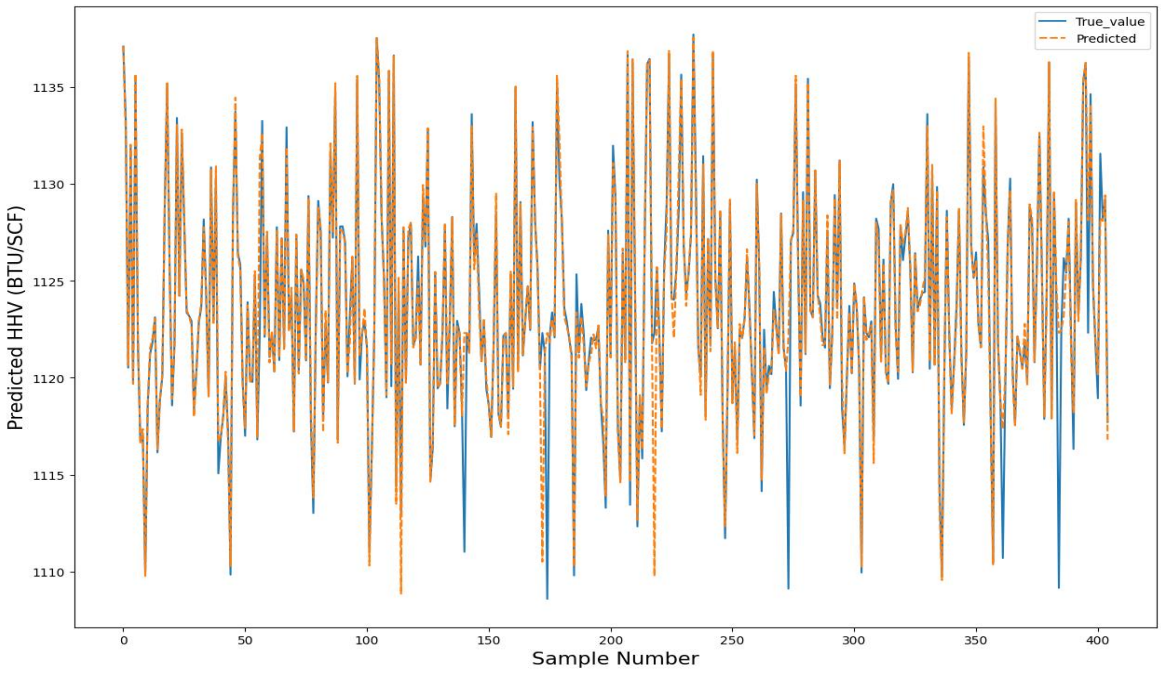


Figure 4.19 Line Plot for Actual and Predicted HHV in Stacking Regressor Model

From Figure 4.19, the predicted values and the actual values were a little close together as compared to the AdaBoost Regression and Linear Regression and this resulted in a better prediction for this model than AdaBoost and Linear Regression models with comparatively small errors.

Figure 4.20 shows the scatter plot for Actual HHV and Predicted HHV in the Stacking Regressor. A line of best fit was drawn with an R^2 value determined. From Figure 4.20, most of the points lie on the line of best fit which in turn increased the R^2 for this model and hence made it a better predictor.

Table 4.14 shows the metric values obtained for the Stacking Regressor for both the training and testing dataset.

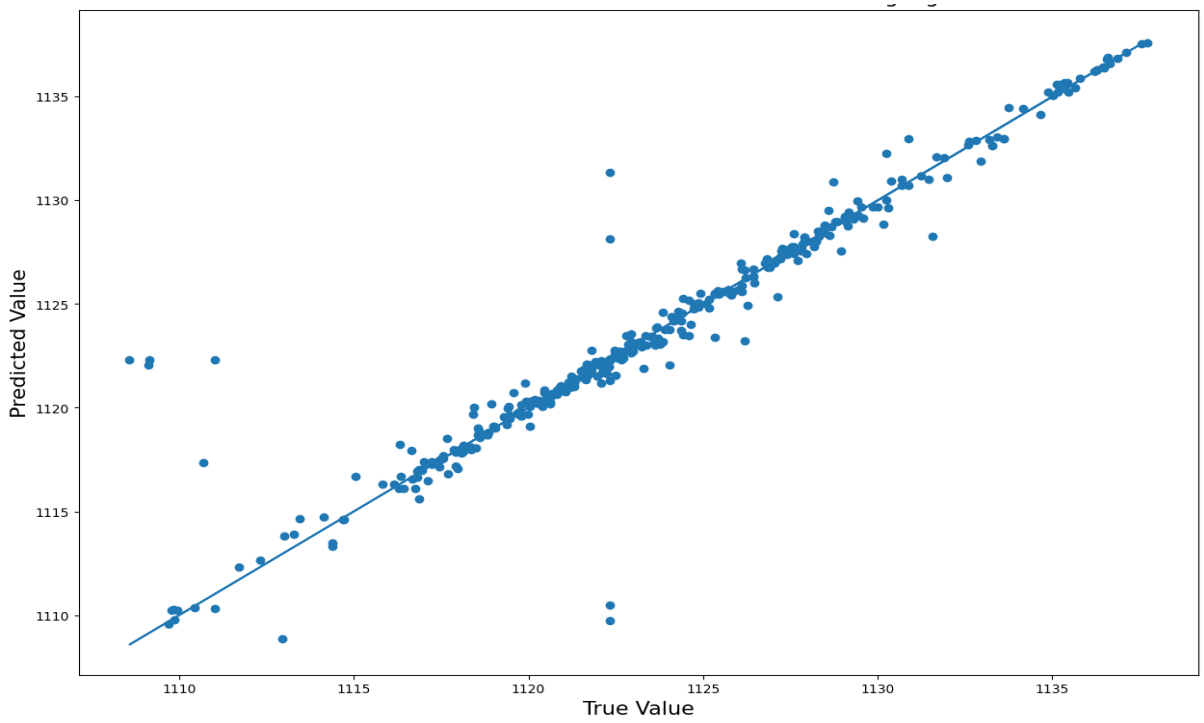


Figure 4.20 Scatter Plot for Actual HHV and Predicted HHV in Stacking Regressor Model

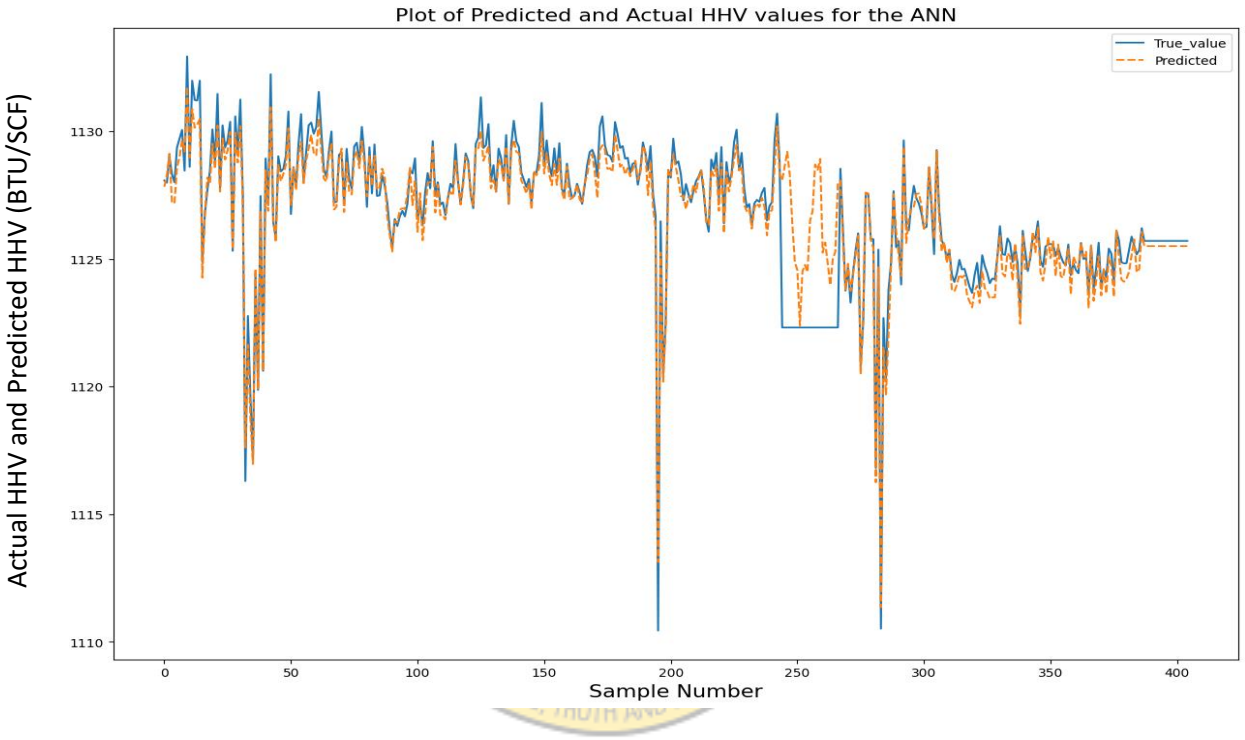
Table 4.14 Training and Testing Results for Stacking Regressor Model

Stacking Regressor	RMSE	MSE	MAE	Adjusted R^2	MAPE	R^2
Training	0.2825	0.0798	0.0569	0.9973	0.54%	0.9974
Testing	1.7543	3.0777	0.5684	0.9068	0.57%	0.9093

From Table 4.14, the errors for the training dataset were lower than that of the testing dataset. In the training, an R^2 of 99.74% was recorded which shows that the predictor variables were able to explain 99.74% of the variations in the output variable (heating value) in the training of the model and this means that the model performed very well in the training, whereas in the testing the value of R^2 decreased to 90.93% which shows that the model developed can only explain about 90.93% of the output variable which is better than that of the AdaBoost Regressor, ANN and Linear Regression.

Artificial Neural Networks (ANN)

Artificial Neural Networks were developed for the dataset which had many predictor variables and one output variable (heating value). The results obtained from this model are presented in this section. Figure 4.21 shows the line plot obtained for the actual and predicted heating values using Artificial Neural Networks.



.

Figure 4.21 Line Plot for Actual and Predicted HHV in ANN Model

From Figure 4.21, the `ANN model gave a moderate prediction for the heating values which were close to the actual heating values.

Figure 4.22 shows the scatter plot for Actual HHV and Predicted HHV in the Artificial Neural Networks Model. A line of best fit was drawn with an R^2 value determined. From Figure 4.22, it is seen that most of the points lie on the line of best fit but not as good as compared to other models.

Table 4.15 shows the metric values obtained for the ANN Model for both the training and testing dataset.

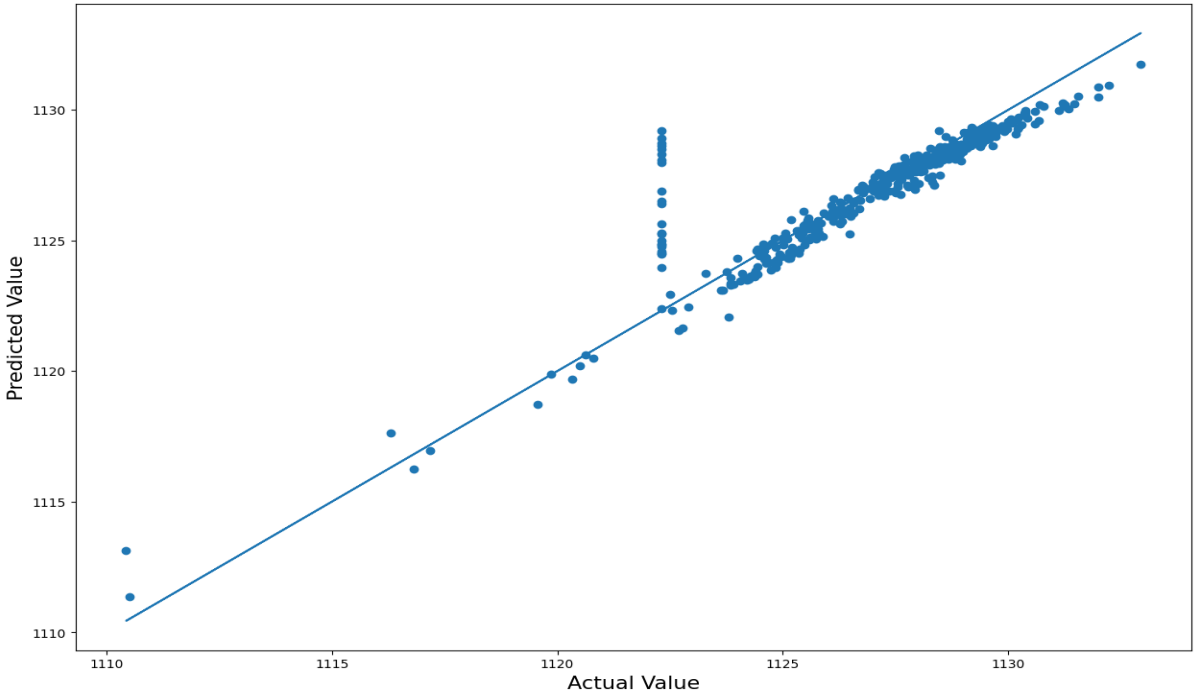


Figure 4.22 Scatter Plot for Actual HHV and Predicted HHV in ANN Model

Table 4.15 Training and Testing Results for ANN Model

ANN Model	RMSE	MSE	MAE	Adjusted R^2	MAPE	R^2
Training	0.8366	0.6999	0.3781	0.9789	0.03%	0.9790
Testing	1.1588	1.3425	0.6149	0.8229	0.05%	0.8273

From Table 4.15, the errors for the training dataset were lower than that of the testing dataset. In the training, an R^2 of 97.90% was recorded which shows that the predictor variables were able to explain 97.90% of the variations in the output variable (heating value) during the training of the model and this means that the model performed very well in the training than in testing, whereas in the testing the value of R^2 decreased to 82.73% which shows that the model developed can only explain about 82.73% of the output variable. Figure 4.23 shows the training and validation loss for ANN Model

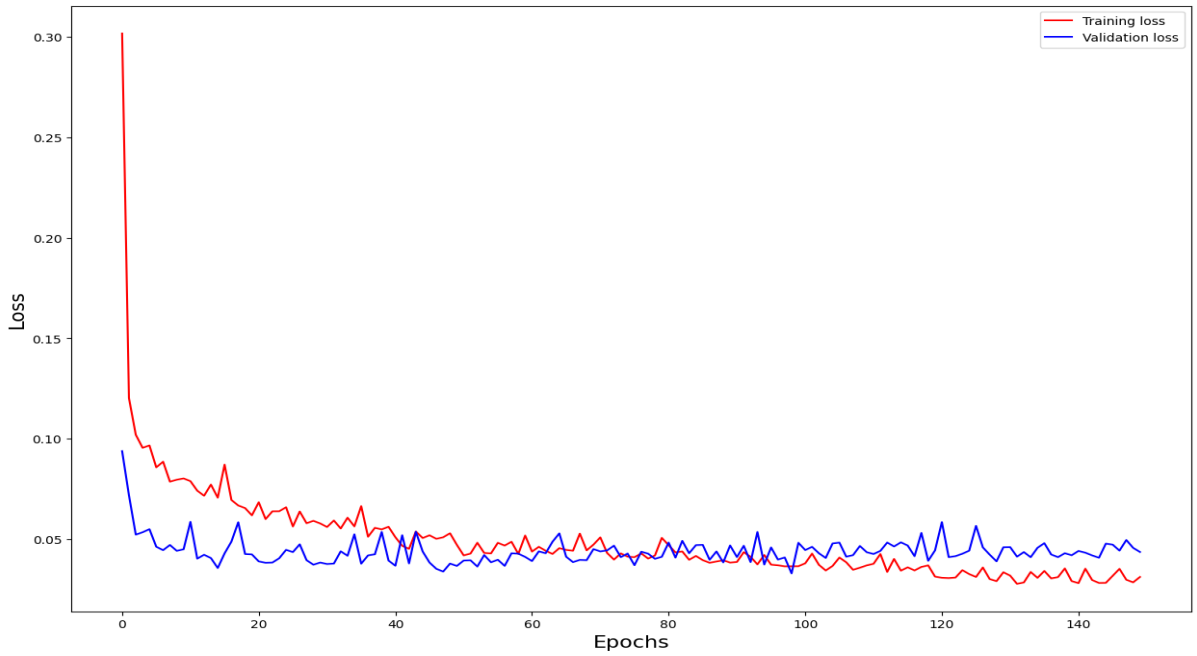


Figure 4.23 Training and Validation Loss in ANN Model

4.3 Comparison of Model Used

Because it is utilised in invoicing, determining the heating value of natural gas is crucial for the gas sector. Heating value determination is subject to the estimation of gas composition by the Gas Chromatograph. In the gas industry, ISO 6976:2016 and GPA 2172 are known for the estimation of heating value. There is instance whereby the Gas Chromatograph is faulty and can't be relied on in the determination of the gas composition. When this happens, many gas industries sometimes rely on previous data for billing which is not appropriate, as such there should be an alternative for the prediction of heating values when such issues come up. This project is geared towards resolving the issue at hand using supervised machine learning approach, of which seven different models were trained and tested on a secondary data obtained from a gas industry. The results obtained will not only serve as an alternative predictor for heating value but also as a verification or check even when the Gas Chromatograph is working perfectly.

Table 4.16 shows the metric values obtained for all seven models. Comparatively, all the models performed better during the training than the testing of the data. Since the industry's

economics heavily rely on the calculation of heating value, a model with the lowest possible error is preferred. For a model, the lower the RMSE, MAE, and MAPE, the higher the accuracy of the model. From Table 4.16, the models with the lowest Root Mean Square Error (RMSE) were XGBoost Regressor, Random Forest and Stacking Regressor and these gave very good R^2 values compared to the other model. A simple definition of the R^2 value is how well the independent variables can account for fluctuations in the dependent variable. Random Forest gave an R^2 of 99.02% and 91.66% respectively for both training and testing, this simply means that in training, Random Forest Model was able to explain almost all the variations in the heating value and for testing, 91.66% of variations were accounted for which is a good score. During the training, both Stacking Regressor and XGBoost Regressor performed better than Random Forest with higher R^2 values, however Random Forest was the best with highest R^2 Value in the Testing.

The Random Forest model will be chosen as the best model for this study since it recorded the least error and had the best coefficient of prediction of all the models when used to estimate our heating value. However, an equation that considers data mistakes and can be applied to prediction was created using the linear regression model.

Table 4.16 Metric Results for All Models

MODEL	Training						Testing					
	RMSE	MSE	MAE	Adjusted R2	MAPE	R^2	RMSE	MSE	MAE	Adjusted R^2	MAPE	R^2
Linear Regression	2.0116	4.0466	1.1343	0.8641	0.5252%	0.8650	2.5342	6.4224	1.2971	0.8055	0.5541%	0.8108
Random Forest Regression	0.5411	0.2928	0.1847	0.9902	0.5359 %	0.9902	1.6821	2.8295	0.5517	0.9143	0.5694%	0.9166
AdaBoost Regression	1.9521	3.8105	1.4356	0.8720	0.5193%	0.8729	2.0230	4.0926	1.4559	0.8761	0.5509%	0.8794
Bagging Regressor	0.6177	0.3815	0.2062	0.9872	0.5361%	0.9873	1.7434	3.0395	0.5837	0.9079	0.5702%	0.9105
XGBoost Regressor	0.2761	0.0763	0.0234	0.9974	0.5389%	0.9975	1.7302	2.9934	0.5393	0.9093	0.5740%	0.9118
Stacking Regressor	0.0798	0.2825	0.0569	0.9973	0.5388%	0.9973	1.7543	3.0777	0.5684	0.9068	0.5722%	0.9093
Artificial Neural Networks	0.8366	0.6999	0.3781	0.9789	0.0338%	0.9790	1.1587	1.3425	0.6149	0.8229	0.0546%	0.8273

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

This thesis offers an in-depth knowledge on the accurate prediction of heating value and how it plays a very important role in the gas industry as it helps in billing off-takers. As there cannot be overreliance in the GC due to a number of times it has failed and possibility of failing in the future, the focus of the thesis was to predict the heating value of comingled natural gas using the gas composition values and the ISO 6976:2016 approach and also supervised machine learning approach. In the event of an unavailable GC, gas industries sometimes select previous heating values from a specific date and specific time (historical) when the GC worked well to serve as reference for billing rather than real time values. The use of Machine Learning models will rather use the trend/pattern of the heating value obtained from a selected period of years for the prediction of the heating value hence providing a much more accurate value when there is an issue with the GC or the auxiliaries to control under billing or overbilling between the aggregator and off taker, and determine the actual quality of the natural gas which is the basis of this study. From the research, it can be concluded that:

i. Artificial Neural Network (ANN), AdaBoost Regressor, XGBoost Regressor, Linear Regression, Bagging Regressor, Random Forest and Stacking Regressor can be used to forecast the heating value with an accuracy of 82.73%, 87.94%, 91.18%, 81.08%, 91.05%, 91.66% and 90.93% respectively. Random Forest performed better with the highest accuracy while Linear Regression showed the least results.
ii. The mathematical formula obtained for linear regression can be used for predicting the heating value of natural gas by accounting for the error.
iii. Even though the ISO 6976:2016 approach of calculating heating value is laborious and time consuming, it can also be used to calculate the heating value of natural gas by accounting for uncertainties.

5.2 Recommendations

From the research work, it is recommended that:

i. Future studies can be done with other software such as R and MATLAB to predict the heating value of natural gas.
ii. More data set should be used in future studies to improve the accuracy of the prediction.
iii. Future works using linear regression should satisfy all assumption (Multicollinearity, Outliers, Homoscedasticity, and Normality etc.)

REFERENCES

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A. and Arshad, H. J. H. (2018), "State-of-the-Art in Artificial Neural Network Applications: A survey", *Heliyon Journal*, Vol. 4, No. 11, pp. 2-5.

Alamooti, A. M. and Malekabadi, F. K. (2018), "An Introduction to Enhanced Oil Recovery", *Fundamentals of Enhanced Oil and Gas Recovery from Conventional and Unconventional Reservoirs*, pp. 1-40.

Alamooti, A. M., and Malekabadi, F. K. (2018), "An Introduction to Enhanced Oil Recovery", *Fundamentals of Enhanced Oil and Gas Recovery from Conventional and Unconventional Reservoirs; Elsevier*, Amsterdam, the Netherlands, pp 1-40.

Almarri, S., Alrumah, M., and Bisanti, M. S. (2018), "Gross Heating Value Correlations for Kuwaiti Gases", *9th Mediterranean offshore conference*, Alexandria, Egypt, pp. 2-20.

Anon (2011), "Annual Energy Outlook 2011: With Projections to 2035", *U.S EIA Report*, United State, 88 pp.

Armstrong, G. T. (1966), "Calculation of the Heating Value of a Sample of High Purity Methane for use as a Reference Material ", *US Government Printing Office*, Vol. 299, pp. 1-4.

Armstrong, G. T. (1966), "Calculation of the Heating Value of a Sample of High Purity Methane for use as a Reference Material", *US Government Printing Office*, Vol. 299, pp. 1-4.

Armstrong, G.T. and Jobe Jr, T.L., (1982), "Heating values of natural gas and its components", *Technical report*, National Bureau of Standards, Washington DC (USA), 57 pp.

Ayaburi, J. and Bazilian, M. (2020), "Economic Benefits of Natural Gas Production: The Case of Ghana's Sankofa Gas Project", *Energy for Growth Hub*, pp. 1-2.

Baker, R. W. and Lokhandwala, K. (2008), "Natural Gas Processing with Membranes: An Overview", *Industrial and Engineering Chemistry Research*, Vol. 47, No. 7, pp. 2109-2121.

Bartle, K. D., and Myers, P. (2002), “History of gas chromatography”, *TrAC Trends in Analytical Chemistry*, Vol. 21, pp. 547-557.

Biau, G., and Scornet, E. (2016), "A random forest guided tour", *Journal of Spanish Society of Statistics and Operations Research*, Vol. 25, No. 1, pp. 197-227.

Birgen, C., Magnanelli, E., Carlsson, P., Skreiberg, O., Mosby, J. and Becidan, M.(2020),“*Machine learning based Modelling for Lower heating value prediction of municipal solid waste*”; Vol. 283, pp. 1-8.

Buckley, T. J. (1991), “Calculation of higher heating values of biomass materials and waste components from elemental analyses” *Resources, conservation and recycling*, Vol. 5, No. 4, pp 329-341.

Callejón-Ferre, A., Velázquez-Martí, B., López-Martínez, J., and Manzano-Agugliaro, F. (2011). “Greenhouse crop residues: Energy potential and models for the prediction of their higher heating value”. *Renewable and sustainable energy reviews*, Vol. 15, No. 2, pp 948-955.

Channiwala, S., and Parikh, P. (2002), “A unified correlation for estimating HHV of solid, liquid and gaseous fuels”, Fuel, Vol. 81, No. 1, pp. 1051-1063.

Chemistry LibreTexts (2020), “Gas Chromatography” *https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Supplemental_Modules(Analytical_Chemistry)/Instrumental_Analysis/Chromatography/Gas_Chromatography*. Accessed: March 23, 2022.

Chen, T., and Guestrin, C. (2016), "Xgboost: A scalable tree boosting system", *22nd acm sigkdd international conference on knowledge discovery and data mining*, San Francisco, USA, pp. 785-794.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., and Zhou, T. (2015), "Xgboost: extreme gradient boosting", *R Package Version*, Vol. 1, No. 4, pp. 1-4.

Dale, S. (2022), "BP Statistical Review of World Energy", *British Petroleum Company*, London, United Kingdom, pp. 1-60.

Debye, P. (1947), “Molecular-weight determination by light scattering”, *The Journal of Physical Chemistry*, Vol. 51, No. 1, pp. 18-32.

Dembicki Jr, H. (2017), “Interpreting crude oil and natural gas data”, Practical Petroleum Geochemistry for Exploration and Production; Elsevier, Amsterdam, pp. 135-188.

Dembicki, H. (2016), “Practical petroleum geochemistry for exploration and production”, *Elsevier Publishers*, Amsterdam, 342 pp.

Demirbaş, A., and Demirbaş, A. H. (2004), “Estimating the calorific values of lignocellulosic fuels”, Energy exploration and exploitation, *Sage Publications*, California, Vol. 22, No. 2, pp. 135-143.

Dey, D. (2023), "Bagging classifier", *https://www.geeksforgeeks.org/ml-bagging-classifier/* Accessed: Januaray 16, 2023.

Dimri, V., Srivastava, R., and Vedanti, N. (2012), “Reservoir geophysics: some basic concepts”, Handbook of Geophysical Exploration: Seismic Exploration, *Elsevier Publications*, Amsterdam, Vol. 41, pp. 89-118.

Dudley, B. (2015), "BP Statistical Review of World Energy 2016", *British Petroleum Company London UK*, pp. 10 -12.

Economides, M. (2009), "Advanced Natural Gas Engineering", *Gulf Publishing Company*, Houston, Texas, pp. 9-11.

Ewing, L. (2001), "Fundamentals of Gas Chromatography, Gas Quality and Troubleshooting", *Chandler Engineering Company LLC*, Indiana Avenue, Oklahoma, pp. 6-7.

Faramawy, S., Zaki, T. and Sakr, A. E. (2016), "Natural Gas Origin, Composition, and Processing: A Review", *Journal of Natural Gas Science and Engineering*, Vol. 34, pp. 34-54.

Francis, H. E., and Lloyd, W. G. (1983), “Predicting heating value from elemental composition”, *Analytica chimica acta*, Vol. 2, No. 2, pp. 192-198.

Francis, W., and Peters, M. C. (2013), “Fuels and Fuel Technology”, *Pergamon Publishers*, Oxford, 2nd Edition, 1069 pp.

Freund, Y., and Schapire, R. E. (1997), "A decision-theoretic generalisation of on-line learning and an application to boosting", *Journal of Computer and System Sciences*, Vol. 55, No. 1, pp. 119-139.

Friedl, A., Padouvas, E., Rotter, H., and Varmuza, K. (2005), "Prediction of heating values of biomass fuel from elemental composition", *Analytica chimica acta*, Vol. 544, No. 1-2. pp 191-198.

Gabitto, J. F., and Tsouris, C. (2010), "Physical properties of gas hydrates: A review", *Journal of Thermodynamics*, Vol. 2010, pp. 1-12.

Guo, B. (2011), "Petroleum production engineering, a computer-assisted approach", *Elsevier*, 288pp.

Guo, B., Liu, X. and Tan, X., (2017), "Petroleum production engineering", *Elsevier Publications*, Oxford, 780 pp.

Guo, R., Zhao, Z., Wang, T., Lu,i G., Zhao, J. and Gao, D., (2020), "Degradation state recognition of piston pump based on ICEEMDAN and XGBoost*"* Vol. 18 6593pp

Gupta, G. K., and Mondal, M. K. (2020), "Bioenergy generation from agricultural wastes and enrichment of end products*", refining biomass residues for sustainable energy and bioproducts*, Elsevier, pp. 337-356.

Han, J., Yao, X., Zhan, Y., Oh, S.-Y., Kim, L.-H., and Kim, H.-J. (2017), "A method for estimating higher heating value of biomass-plastic fuel", *Journal of the Energy Institute*, Vol. 990, No. 2, pp. 331-335.

Hanafy, H., Macary, S., ElNady, Y., Bayomi, A., and El Batanony, M. (1997), "Empirical PVT correlations applied to Egyptian crude oils exemplify significance of using regional correlation", *International Symposium on Oilfield Chemistry*, Houston, Texas, pp 122- 147.

Holloway, S. (2001), "Storage of Fossil Fuel-derived Carbon dioxide Beneath the Surface of the Earth", *Annual Review of Energy and the Environment*, Vol. 26, No. 1, pp.145-166.

ISO, B. (2005), "Natural Gas–Calculation of Calorific Values, Density, Relative Density and Wobbe Index from Composition", *International Standards Organisation*, 88pp.

ISO, E. (2016), "Natural gas-Calculation of calorific values, density, relative density and Wobbe indices from composition", *International Standards Organisation*, 56pp.

Jenkins, R. G. (2020), “Thermal gasification of biomass—a primer”, *Bioenergy*, pp. 293-324.

Jordan, M. I., and Mitchell, T. M. J. S. (2015), "Machine learning: Trends, perspectives, and prospects", *Journal of Science*, Vol. 349, No. 6245, pp. 255-260.

Kathiravale, S., Yunus, M. N. M., Sopian, K., Samsuddin, A. H., and Rahman, R. (2003), “Modeling the heating value of Municipal Solid Waste”, *Fuel Processing and Technology Journal*, Vol. 82, No. 9, pp 1119-1125.

Key, J. A., and Ball, D. W. (2014), “Introductory chemistry”, 1st-canadian edition, pp. 974-989.

Kidnay, A. J. and Parrish, W. R. (2006), “Fundamentals of Natural Gas Processing”, *CRC press*, Vol 3, pp. 10-36.

Klimstra, J. (1986), “Interchangeability of gaseous fuels-The importance of the Wobbe-index”, *SAE transactions*, pp. 962-972.

Kolb, D. (1978), “The Mole”, *Journal of Chemical Education*, Vol. 55, No, 11, 728 pp.

Laugier, A., and Garai, J. (2007), “Derivation of the ideal gas law”, *Journal of Chemical Education*, Vol. 88, No. 1, 1832 pp.

Lett, R. G., and Ruppel, T. C. (2004), Coal, *chemical and physical properties*. 1-3 pp.

Levine, S. (1985), “Derivation of the ideal gas law”, *Journal of Chemical Education*, Vol. 62, No. 5, 399 pp.

Li, B., Xue, J., Xia, K., Zhou, L., Qian, P., and Jiang, Y. J. (2021), "An Auto-Contouring Method for Kidney Using a Novel Semi-Supervised Leaning Extreme Learning Machine Method", *Journal of Medical Imaging and Health Informatics*, Vol. 11, No. 8, pp. 2267-2273.

Liu, Y., Wang, Y., and Zhang, J. (2012), "New machine learning algorithm: Random forest", *Information Computing and Applications third International Conference*, Chengde, China pp. 246-252.

Lower, S. (2011), “Moles, Mixtures, and Densities Properties of Gases”, *Chem1 virtual textbook*, Simon Fraser University, Burnaby, 50 pp.

Lower, S. (2011), “Moles, Mixtures, and Densities Properties of Gases”, *chem1 virtual textbook*, Simon Fraser University, Burnaby, 60 pp.

Ludtke, P. R. (1986). Natural gas handbook. Vol 2, pp 987 - 1023

Machová, K., Barcak, F., and Bednár, P. (2006), "A bagging method using decision trees in the role of base classifiers", *Acta Polytechnica Hungarica,*Vol. 3, No.2, pp. 121-132.

Maulud, D., and Abdulazeez, A. M. (2020), "A review on linear regression comprehensive in machine learning", *Journal of Applied Science and Technology,* Vol. 1, No. 4, pp. 140-147.

McNair, H. M., Miller, J. M., and Snow, N. H. (2019), Basic gas chromatography, John Wiley and Sons Publications, Hoboken, 288 pp.

Moharir, R. V., Rena, P. G., and Kumar, S. (2019), “Bio-drying of solid waste”, *Biological Processing of Solid Waste*, 129 pp.

Moharir, R. V., Rena, P. G., and Kumar, S. (2019), Biological Processing of Solid Waste, *CRC Press*, Florida, 1st Edition, 18 pp.

Mokhatab, S., Poe, W. A., and Speight, J. (2006), "Natural gas compression", *Handbook of Natural Gas Transmission and Processing; Gulf Professional Pub*.: Burlington, MA, USA, pp. 295-322.

Mokhatab, S., Poe, W. A., and Mak, J. Y. (2018), “Handbook of natural gas transmission and processing: principles and practices”, *Gulf Professional Publishing*, Cambridge, 862 pp.

Mokhatab, S., Poe, W. A., and Mak, J. Y. (2018), “Handbook of natural gas transmission and processing”, *Gulf professional publishing*, Houston, Texas, 4th Edition, 862 pp.

Muhammad, I., and Yan, Z. (2015), "SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY", *ICTACT Journal on Soft Computing*, Vol. 5, No. 3, pp. 946-952.

Nhuchhen, D. R., and Salam, P. A. (2012), “Estimation of higher heating value of biomass from proximate analysis: A new approach” *Fuel Processing Technology*, Vol. 99, pp 55-63.

Nhuchhen, D. R., and Salam, P. A. (2012), “Estimation of higher heating value of biomass from proximate analysis: A new approach”, *Fuel*, Vol. 99, pp. 55-63.

Oscar AN. National Energy Statistics. Ghana: Energy Commission; 2022.

Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O., and Akinjobi, J. (2017), "Supervised machine learning algorithms: classification and comparison", *International Journal of Computer Trends and Technology*, Vol. 48, No, 3, pp. 128-138.

Pellegrini, L. A., De Guido, G., and Lange, S. (2019), “Handbook of Natural Gas Transmission and Processing-Principles and Practices”, *Elsevier-Gulf Professional Publishing*, 739 pp.

Perera, F. (2018), "Pollution from Fossil-Fuel Combustion is the Leading Environmental Threat to global Pediatric Health and Equity: Solutions Exist", *International Journal of Environmental Research and Public Health*, Vol. 15, No. 1, 16pp.

Picard, A., Davis, R., Gläser, M., and Fujii, K. (2008), “Revised formula for the density of moist air (CIPM-2007)”, *Metrologia*, Vol. 45, No. 2, pp. 149-155.

Riazi, M. (2005), “Characterisation and properties of petroleum fractions”, ASTM *International Publishers*, USA, 401 pp.

Ringen, S., Lanum, J., and Miknis, F. P. (1979), "Calculating Heating Values from Elemental Compositions of Fossil Fuel", United Kingdom, Vol 1, No. 68pp.

SALEH, H. (2022), "Machine Learning–Regression", 4th year Seminar Thesis Report, Higher Institute for Applied Sciences and Technology, Damascus, Syria, 23 pp.

Schapire, R. E. (2013), "Explaining adaboost", *Journal of Machine Learning*, pp. 1-14.

Schonlau, M., and Zou, R. Y. (2020), "The random forest algorithm for statistical learning", *The Stata Journal*, Vol. 20, No. 1, pp. 3-29.

Shahin, M. A., Jaksa, M. B., and Maier, H. R. (2001), "Artificial neural network applications in geotechnical engineering", *Australian Geomechanics Journal*, Vol. 36, No. 1, pp. 49-62.

Sheng, C., and Azevedo, J. (2005), “Estimating the higher heating value of biomass fuels from basic analysis data”, *Biomass and bioenergy*, Vol. 28, No. 5, pp 499-404.

Shepherd, M. (1947), “Analysis of Natural Gas”, *Analytical Chemistry*, Vol. 19, No. 9, pp. 635-640.

Shokir, E. M. E.-M., El-Awad, M. N., Al-Quraishi, A. A., and Al-Mahdy, O. A. (2012), “Compressibility factor model of sweet, sour, and condensate gases using genetic programming”, *Chemical Engineering Research and Design*, Vol. 90, No. 6, pp. 785-792.

Siirola, J. J. (2010), "Natural Gas as a Chemical Industry Fuel and Feedstock: Past, Present, Future", *http://egon.cheme.cmu.edu/esi/docs/pdf/SiirolaNaturalGas.pdf* Accessed: June 20, 2022.

Skempton, A. (1986), “Standard penetration test procedures and the effects in sands of overburden pressure, relative density, particle size, ageing and over consolidation” *Geotechnique*, Vol. 36, No. 3, pp. 425-447.

Snow, N. H., and Slack, G. C. (2002), “Head-space analysis in modern gas chromatography” *TrAC Trends in Analytical Chemistry*, Vol. 21, No. 9-10, pp. 608-617.

Taki, M., and Rohani, A. (2022), “*Machine learning models for prediction of the Higher Heating Value (HHV) of Municipal Solid Waste (MSW) for waste-to-energy Evaluation, Case Stud. Therm. Eng.”*,31pp.

Tenny, K. M., and Cooper, J. S. (2017), “Ideal Gas Behaviour”, *https://europepmc.org/article/nbk/nbk441936*. Accessed: June 14, 2022.

Tenny, K. M., and Cooper, J. S. (2022), Ideal Gas Behaviour, *StatPearls Publishing*, Florida, 30 pp.

Thipkhunthod, P., Meeyoo, V., Rangsunvigit, P., Kitiyanan, B., Siemanond, K., and Rirksomboon, T. (2005), “Predicting the heating value of sewage sludges in Thailand from proximate and ultimate analyses”, *Fuel*, Vol. 84, No. 7-8, pp. 849-857.

Towell, G. (2020), “Viscosity: Definition, Unit and Formula”, *https://sciencing.com/convert-centistoke-centipoise-8279085.html*.Accessed: November 21, 2022.

Towell, G. (2020), “Viscosity: Definition, Unit and Formula”, *www. sciencing.com/convert-centistoke-centipoise-8279085.html.* Accessed: October 19, 2022.

Tronci, S., Chebeir, J. A., Mandis, M., Baratti, R. and Romagnoli, J. A. (2020), "Control Strategies for Natural Gas Liquids Recovery Plants", *Computer Aided Chemical Engineering*, Vol. 48, pp. 1291-1296.

Uyanık, G. K., and Güler, N. (2013), "A study on multiple linear regression analysis*", 4th international conference on new horizons in education*, Sakarya, Turkey, Vol. 106, pp. 234-240.

Vargas-Moreno, J., Callejón-Ferre, A., Pérez-Alonso, J., and Velázquez-Martí, B. (2012), "A review of the mathematical models for predicting the heating value of biomass materials" *Renewable and Sustainable Energy Reviews*, Vol. 16, No. 5, pp. 3065-3083.

Vazquez, M., and Beggs, H. D. (1977), "Correlations for fluid physical property prediction" *SPE Annual Fall Technical Conference and Exhibition*, Denver, Colorado, pp. 234-421.

Vazquez, M., and Beggs, H. D. (1980), "Correlations for fluid physical property prediction" *Journal of petroleum Technology*, Vol. 30, No. 6, pp. 968-970.

Wallis, P. (2013), "Natural gas analysis", Trends in Analytical Chemistry, Vol. 5, No. 3, 63 pp.

Weaver, C., and Miller, C. (2019), "A framework for climate change-related research to inform environmental protection", *Environmental management*, Vol. 64, No. 3, pp. 245-257.

Webb, P. A. (2001), "Volume and density determinations for particle technologists", *Micromeritics Instrument Corp*, Vol. 2, No. 16, pp. 1 – 16.

Welker, C. (2015), "Flipping pancakes: how gas inflows and mergers shape galaxies in their cosmic environment*", Doctoral Dissertation*, Université Pierre et Marie Curie-Paris VI, Paris, 53pp.

Welker, C. (2015), "Flipping pancakes: how gas inflows and mergers shape galaxies in their cosmic environment", *Published PhD thesis Report*, Université Pierre et Marie Curie-Paris VI, Paris, 198 pp.

Winter, M. (2014), "Benchmark and validation of Open Source CFD codes, with focus on compressible and rotating capabilities, for integration on the SimScale platform",

https://docplayer.net/54188410-Benchmark-and-validation-of-open-source-cfd-codes-with-focus-on-compressible-and-rotating-capabilities-for-integration-on-the-simscale-platform.html. Accessed: April 12, 2022.

Woody, A. I. (2013), "How is the ideal gas law explanatory?", *Science and Education*, Vol. 22, No. 7, pp. 1563-1580.

Xing, J., Luo, K., Wang, H., Gao, Z., and Fan, J. (2019), "A comprehensive study on estimating higher heating value of biomass from proximate and ultimate analysis with machine learning approaches", *Energy*, Vol. 188, 116077 pp.

Yin, C.-Y. (2011), "Prediction of higher heating values of biomass from proximate and ultimate analyses", *Fuel*, Vol. 90, No. 3, pp. 1128-1132.

Ying, G., Lian-Kui, D., Hua-Dong, Z., Yun-Liang, C., and Li, Z. (2019), "Quantitative analysis of main components of natural gas based on Raman spectroscopy", *Chinese Journal of Analytical Chemistry*, Vol. 47, No. 1, pp. 67-76.

Zachariah-Wolff, J. L., Egyedi, T. M., and Hemmes, K. (2007), "From natural gas to hydrogen via the Wobbe index: The role of standardised gateways in sustainable infrastructure transitions", *International journal of hydrogen energy*, Vol. 32, No. 9, pp. 1235-1245.

Zawada, B. (2014), "The practical application of ISO 22301", *Journal of business continuity and emergency planning*, Vol. 8, No. 1, pp. 83-90.

APPENDICES

APPENDIX A

FIRST ONE HUNDRED DATA POINTS USED IN THE PREDICTION

C1	C2	C3	I-C4	N-C4	I-C5	N-C5	C6+	N2	CO2	HHV
88.8	5.74	3	0.33	0.6	0.12	0.1	0.08	0.43	0.8	1119.8
88.88	5.75	2.92	0.33	0.59	0.12	0.1	0.08	0.43	0.8	1118.36
88.71	5.8	3.02	0.34	0.6	0.12	0.1	0.08	0.43	0.8	1120.71
88.68	5.8	3.04	0.34	0.61	0.12	0.1	0.08	0.43	0.8	1121.28
88.65	5.81	3.05	0.34	0.61	0.12	0.1	0.08	0.43	0.81	1121.59
88.88	5.66	2.99	0.34	0.61	0.12	0.11	0.08	0.43	0.77	1120.33
89.18	5.47	2.9	0.34	0.62	0.13	0.11	0.09	0.44	0.72	1118.71
89.31	5.39	2.86	0.34	0.61	0.13	0.11	0.08	0.44	0.72	1117.1
88.85	5.76	2.92	0.34	0.6	0.12	0.1	0.08	0.43	0.8	1118.95
88.64	5.83	3.04	0.33	0.61	0.13	0.1	0.08	0.43	0.81	1121.48
88.6	5.87	3.04	0.33	0.61	0.13	0.1	0.08	0.43	0.82	1121.58
88.53	5.87	3.11	0.33	0.61	0.13	0.1	0.08	0.43	0.82	1122.76
88.5	5.87	3.12	0.34	0.62	0.12	0.1	0.08	0.43	0.82	1123.36
88.61	5.84	3.05	0.34	0.61	0.13	0.1	0.08	0.43	0.82	1121.78
88.55	5.89	3.03	0.34	0.62	0.12	0.11	0.08	0.42	0.83	1122.22
88.56	5.87	3.03	0.35	0.62	0.13	0.11	0.08	0.42	0.83	1122.72
88.59	5.84	3.03	0.34	0.63	0.13	0.11	0.09	0.42	0.82	1122.73
88.74	5.77	2.96	0.34	0.62	0.13	0.11	0.09	0.42	0.81	1120.88
88.69	5.81	2.98	0.34	0.62	0.13	0.11	0.09	0.42	0.82	1121.36
88.63	5.86	2.99	0.33	0.62	0.13	0.11	0.09	0.42	0.82	1121.91
88.61	5.86	3.01	0.34	0.62	0.12	0.11	0.09	0.42	0.82	1122.08
88.62	5.83	3.02	0.34	0.62	0.12	0.11	0.09	0.42	0.82	1122.24
88.58	5.84	3.05	0.34	0.62	0.12	0.11	0.09	0.42	0.83	1122.5
88.51	5.87	3.08	0.34	0.63	0.12	0.11	0.09	0.42	0.84	1123.36
88.5	5.88	3.07	0.35	0.63	0.13	0.11	0.09	0.42	0.84	1123.55
88.55	5.85	3.06	0.34	0.63	0.13	0.11	0.09	0.42	0.84	1123
88.48	5.91	3.07	0.34	0.63	0.13	0.11	0.09	0.42	0.84	1123.53
88.46	5.94	3.07	0.34	0.62	0.12	0.11	0.09	0.42	0.84	1123.49
88.43	5.92	3.1	0.35	0.62	0.13	0.11	0.09	0.42	0.84	1124.26
88.82	5.68	2.99	0.34	0.63	0.13	0.11	0.09	0.42	0.78	1121.48
88.98	5.64	2.89	0.34	0.61	0.13	0.11	0.09	0.43	0.79	1118.91
88.77	5.89	2.89	0.33	0.58	0.12	0.1	0.08	0.43	0.82	1118.6
88.58	5.93	3.03	0.33	0.59	0.12	0.1	0.07	0.43	0.83	1120.93
88.66	5.79	3.08	0.33	0.6	0.12	0.1	0.07	0.43	0.82	1121.07
88.74	5.78	3.03	0.32	0.6	0.12	0.1	0.07	0.43	0.81	1119.9
88.89	5.73	2.94	0.32	0.59	0.11	0.1	0.07	0.43	0.81	1117.83
88.76	5.84	2.96	0.33	0.59	0.11	0.1	0.07	0.43	0.81	1119.03
88.65	5.85	3.06	0.33	0.59	0.11	0.1	0.07	0.43	0.82	1120.7
88.65	5.83	3.07	0.33	0.6	0.11	0.1	0.07	0.43	0.82	1120.71

88.59	5.8	3.14	0.34	0.6	0.11	0.1	0.07	0.43	0.82	1121.93
88.64	5.8	3.09	0.33	0.6	0.11	0.1	0.07	0.43	0.82	1121.15
88.62	5.84	3.07	0.34	0.6	0.11	0.1	0.07	0.43	0.82	1121.11
88.67	5.81	3.06	0.33	0.6	0.11	0.1	0.07	0.43	0.82	1120.58
88.62	5.83	3.09	0.33	0.6	0.11	0.1	0.07	0.43	0.82	1121.21
88.65	5.8	3.09	0.33	0.6	0.11	0.1	0.07	0.43	0.82	1121
88.57	5.84	3.12	0.33	0.6	0.11	0.1	0.07	0.43	0.82	1121.89
88.63	5.82	3.09	0.33	0.6	0.11	0.1	0.07	0.43	0.81	1121.41
88.64	5.85	3.06	0.33	0.6	0.11	0.1	0.07	0.43	0.81	1120.8
88.63	5.84	3.07	0.33	0.6	0.11	0.1	0.07	0.43	0.82	1121.09
88.61	5.86	3.07	0.33	0.6	0.11	0.1	0.07	0.43	0.82	1121.05
88.6	5.86	3.08	0.33	0.6	0.11	0.1	0.07	0.43	0.82	1121.24
88.65	5.83	3.07	0.33	0.6	0.11	0.1	0.07	0.43	0.82	1120.8
88.62	5.83	3.09	0.33	0.6	0.12	0.1	0.07	0.43	0.82	1121.29
88.63	5.83	3.07	0.33	0.6	0.12	0.1	0.07	0.43	0.82	1121.17
88.98	5.67	2.93	0.32	0.58	0.11	0.1	0.07	0.43	0.81	1116.74
88.92	5.71	2.94	0.32	0.58	0.12	0.1	0.07	0.43	0.81	1117.27
88.77	5.78	3.01	0.32	0.59	0.11	0.1	0.07	0.43	0.82	1119.13
88.69	5.81	3.05	0.33	0.59	0.12	0.1	0.07	0.43	0.82	1120.21
88.62	5.83	3.09	0.33	0.6	0.11	0.1	0.07	0.43	0.82	1120.96
88.63	5.85	3.07	0.33	0.59	0.11	0.1	0.07	0.43	0.82	1120.81
88.63	5.84	3.08	0.33	0.59	0.11	0.1	0.07	0.43	0.82	1120.85
88.65	5.82	3.07	0.32	0.6	0.11	0.1	0.07	0.43	0.82	1120.52
88.62	5.84	3.08	0.33	0.6	0.11	0.1	0.07	0.43	0.82	1120.79
88.59	5.87	3.08	0.33	0.6	0.11	0.1	0.07	0.43	0.82	1121.15
88.58	5.87	3.07	0.33	0.6	0.12	0.1	0.07	0.43	0.82	1121.44
88.61	5.85	3.07	0.33	0.6	0.11	0.1	0.07	0.43	0.82	1121.17
88.6	5.88	3.06	0.33	0.6	0.11	0.1	0.07	0.43	0.82	1121.3
88.62	5.83	3.07	0.34	0.6	0.12	0.1	0.08	0.43	0.82	1121.46
88.63	5.84	3.05	0.33	0.6	0.12	0.1	0.08	0.43	0.82	1121.21
88.63	5.84	3.06	0.34	0.6	0.11	0.1	0.08	0.43	0.81	1121.36
88.6	5.85	3.07	0.34	0.61	0.12	0.1	0.08	0.43	0.81	1121.96
88.67	5.8	3.06	0.34	0.61	0.12	0.1	0.08	0.43	0.81	1121.36
88.41	5.97	3.14	0.34	0.62	0.11	0.1	0.08	0.42	0.82	1123.88
88.19	6	3.28	0.36	0.63	0.11	0.1	0.08	0.42	0.82	1127.16
88.15	6.02	3.3	0.35	0.64	0.12	0.11	0.08	0.42	0.82	1127.81
88.15	6.03	3.29	0.35	0.64	0.12	0.11	0.08	0.42	0.82	1127.61
88.12	6.02	3.32	0.36	0.64	0.12	0.11	0.08	0.42	0.82	1128.16
88.11	6	3.34	0.36	0.64	0.12	0.11	0.08	0.42	0.82	1128.41
88.12	6.02	3.32	0.35	0.64	0.12	0.1	0.08	0.42	0.82	1128.17
88.19	6	3.28	0.35	0.63	0.12	0.1	0.08	0.42	0.82	1127.13
88.1	6.02	3.34	0.35	0.64	0.13	0.1	0.08	0.42	0.82	1128.48
88.12	6.03	3.32	0.36	0.64	0.12	0.1	0.08	0.42	0.82	1127.91
88.21	5.96	3.3	0.35	0.63	0.11	0.1	0.08	0.42	0.82	1127.01
88.46	5.86	3.18	0.34	0.62	0.12	0.1	0.08	0.43	0.81	1124.07
88.65	5.81	3.07	0.33	0.61	0.12	0.1	0.08	0.43	0.81	1121.47

88.72	5.8	3.02	0.33	0.6	0.12	0.1	0.08	0.43	0.8	1120.62
88.1	6.05	3.31	0.35	0.64	0.12	0.1	0.08	0.42	0.83	1127.96
88.25	5.98	3.25	0.35	0.64	0.12	0.11	0.08	0.42	0.81	1126.98
88.21	5.99	3.27	0.35	0.64	0.12	0.11	0.08	0.42	0.8	1127.58
88.32	5.94	3.22	0.35	0.63	0.12	0.11	0.08	0.42	0.8	1126.08
88.23	5.95	3.29	0.35	0.64	0.12	0.11	0.08	0.42	0.8	1127.58
88.41	5.88	3.2	0.35	0.63	0.12	0.11	0.08	0.42	0.8	1125.06

Index