KADIR HAS UNIVERSITY
SCHOOL OF GRADUATE STUDIES
DEPARTMENT OF ENGINEERING AND NATURAL SCIENCES

ESTIMATION OF ENERGY PRODUCTION IN BIOGAS PLANTS

ŞEVVAL AYŞE YURTEKİN

MASTER OF SCIENCE THESIS

ISTANBUL, JANUARY, 2023

ESTIMATION OF ENERGY PRODUCTION IN BIOGAS PLANTS

ŞEVVAL AYŞE YURTEKİN
ADVISOR: Assoc. Prof. Dr. ATİLLA ÖZMEN
CO-ADVISOR: Dr. BARAN TANDER

A thesis submitted to
the School of Graduate Studies of Kadir Has University
in partial fulfilment of the requirements for the degree of
Master of Science in Electronics Engineering

Istanbul, January, 2023

APPROVAL

This thesis titled ESTIMATION OF ENERGY PRODUCTION IN BIOGAS PLANTS submitted by ŞEVVAL AYŞE YURTEKİN, in partial fulfillment of the requirements for the degree of Master of Science in Electronics Engineering is approved by

Assoc. Prof. Dr. Atilla Özmen (Advisor) ……..
Kadir Has University

Dr. Baran Tander (Co-Advisor) ……..
Istanbul Aydın University

Prof. Dr. Metin Şengül ……..
Kadir Has University

Asst. Prof. Dr. Mesut Çevik ……..
Altınbaş University

I confirm that the signatures above belong to the aforementioned faculty members.

_____

Prof. Dr. Mehmet Timur Aydemir
Director of the School of Graduate Studies
Date of Approval: 05.01.2023

DECLARATION ON RESEARCH ETHICS AND PUBLISHING METHODS

I, ŞEVVAL AYŞE YURTEKİN; hereby declare

- that this Master of Science Thesis that I have submitted is entirely my own work and I have cited and referenced all material and results that are not my own in accordance with the rules;
- that this Master of Science Thesis does not contain any material from any research submitted or accepted to obtain a degree or diploma at another educational institution;
- and that I commit and undertake to follow the "Kadir Has University Academic Codes of Conduct" prepared in accordance with the "Higher Education Council Codes of Conduct".

In addition, I acknowledge that any claim of irregularity that may arise in relation to this work will result in a disciplinary action in accordance with the university legislation.

Şevval Ayşe Yurtekin

_____

Date (05/01/2023)

To My Dearest Family...

ACKNOWLEDGEMENTS

My thesis professors who were with me throughout the study, Assoc. Prof. Dr. Atilla ÖZMEN and Dr. Baran TANDER. I would like to thank all my professors who contributed to me. I would like to thank Berkin İMER, owner of Pales biogas plant, who provided us with data during this process. I would like to express my sincere thanks to my family, who never left me alone throughout the process, for all their help and motivation.

ABSTRACT

The importance of renewable energy sources is getting more and more significant day by day. Renewable energies are required since the world's energy consumption is rising in tandem with the human population. The gas produced from wastes such as biogas, agricultural waste, and animal dung is an example of biomass energy, a renewable energy source. Machine learning is a branch of computer science that tries to improve its performance with the data it accumulates over time by simulating a human learning mechanism. This research begins with a discussion of biogas generation and investigations. The discussion then shifts to artificial intelligence, machine learning, and neural networks. In the application portion, an application of feature selection utilizing data, wastes, and biogas production from the Pales biogas plant is developed, as well as an application that forecasts biogas output using regression and an artificial neural network model. In the Python-based model, machine learning and deep learning libraries are utilized, and the data is preprocessed to make it compatible with the model. In this 22-featured model, the elements that contribute to the model, specifically biogas generation, were chosen using feature selection algorithms. The regression model and neural network model were created with the selected features as Dairy cow manure, Wheat Juice, Potato peel, Potato whole, Mixed vegie, Weak vinasse and Poultry manure. There are three distinct digesters included in the data. Three separate analyses were performed on three distinct digesters, and the findings were compared to the cumulative data. In the study, 20% of the data were set aside as test data and 80% were used for training. In terms of model performance, the data feature selection approach was effective for regression models, but negatively effect of variable reduction in neural networks. The R^2 score was 52% in the neural network model trained without variable selection, and the mean of regression models trained with feature selection was 49%.

Keywords: biogas, energy production, feature selection, machine learning, neural networks

BİYOGAZ TESİSLERİNDE ENERJİ ÜRETİMİNİN TAHMİNİ

ÖZET

Yenilenebilir enerji kaynaklarının önemi her geçen gün daha da artmaktadır. Dünyadaki enerji tüketimi insan nüfusu ile paralel olarak arttığı için yenilenebilir enerjilere ihtiyaç duyulmaktadır. Yenilenebilir bir enerji kaynağı olan biyokütle enerjisine örnek olan biyogaz, tarımsal atıklar ve hayvan gübresi gibi atıklardan üretilen bir gazdır. Makine öğrenimi, bir insan öğrenme mekanizmasını simüle ederek zaman içinde topladığı veriler ile performansını artırmaya çalışan bir bilgisayar bilimi dalıdır. Bu çalışma, biyogaz üretimi ve biyogaz ile ilgili çalışmaların, araştırmalarının incelenmesi ile başlamaktadır. Daha sonra yapay zeka, makine öğrenimi ve sinir ağları kavramları tanımlanmıştır. Uygulama bölümünde, Pales biyogaz tesisinden elde edilen 650 günlük atık ve biyogaz üretimleri içeren veriler kullanılarak, öznitelik seçimi, regresyon ve yapay sinir ağı modeli kullanarak biyogaz çıkışını tahmin eden uygulamalar geliştirilmiştir. Python açık kaynak tabanlı bu modelde, makine öğrenimi ve derin öğrenme kütüphaneleri kullanılmış ve veriler modele uyumlu hale getirilmek için ön işlemden geçirilmişlerdir. 22 özniteliğe sahip olan bu modelde, öznitelik seçim algoritması kullanılarak, modele yani biyogaz üretimine katkı sağlayan öznitelikler seçilmiştir. İnek Gübresi, Buğday Suyu, Patates kabuğu, Patates, Karışık sebze, Şilempe ve Tavuk gübresi algoritma ile seçilmiş olup bu seçilen öznitelikler ile regresyon modeli ve sinir ağı modeli oluşturulmuştur. Veri de 3 farklı sindirici bulunmaktadır. 3 farklı sindirici için 3 farklı analiz yapılmış olup sonuçlar kümül veri ile karşılaştırılmıştır. Veri çalışmada %20'si test %80'i eğitim verisi olarak ayrılmıştır. Model performansı açısından, veri öznitelik seçimi yaklaşımı regresyon modelleri için pozitif etkili, ancak sinir ağlarında değişken azaltmanın olumsuz etkisi gözlemlenmiştir. Değişken seçimi olmaksızın eğitilen nöral ağ modelinde R^2 skor %52'i iken, değişken seçimi ile eğitilen regresyon modelleri ortalaması %49'dur.

Anahtar Sözcükler: biyogaz , enerji üretimi, öznitelik seçimi, makine öğrenimi, nöral ağlar

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS

y_i Actual Value

β Coefficient of Regression

ʎ Correction Parameter

$\hat{y}_i$ Expected Value

w Weight

LIST OF ACRONMYMS AND ABBREVIATIONS

AI	Artificial intelligence
ANOVA	Variance Analyses
ANFI	Adaptive Network-Based Fuzzy Inference System
ANN	Artificial Neural Networks
CO_2	Carbon Dioxide
C/N	Carbon Nitrogen Ratio
CD	Co-Digestion
CH_4	Methane
DA	Dual-Stage-Attention
FI	Feature Importance
FW	Food Waste
FVW	Fruit And Vegetable Waste
HRT	Hydraulic Retention Time
k-NN	k-Nearest Neighbors
LDA	Linear Discriminant Analysis
LSSVM	Least square support vector machine
LSTM	Long Short-Term Memory
L1	Regularization technique
L2	Regularization technique
MAD	Mean Absolute Difference
MAE	Mean Absolute Error
MSE	Mean Square Error
PCA	Principal Component Analysis
R^2	Coefficient Of Determination

ReLU	Rectified Linear Unit
RMSE	Root Mean Square Error
RSM	Response Surface Methodology Analysis
RMSProp	Root Mean Square Propagation
SVM	Support Vector Machine
VFA	Volatile Fatty Acids
VSNs	Variable Selection Networks
VS	Volatile Solid

1. INTRODUCTION

The demand for energy resources directly correlates with the growth in global population. Humans need energy for their life and have used fossil fuels such as natural gas and oil throughout their lives. However, following the oil crises of the 1970s, people are now aware that fossil resources will become scarce. Also, when fuels such as oil, natural gas and coal are burned dangerous gases like carbon dioxide, sulfur and nitrogen are discharged into the environment [1]. This threatens the functioning of the whole world, from the ecosystem to the structure of the atmosphere. Moreover, we may be faced with a big problem such as global warming and climate change that will reach global dimensions. Energy that is continuously or regularly accessed from the natural environment is referred to as renewable energy. Renewable energy is a very critical need on our planet. With renewable energy resources foreign dependency in energy production is reduced and it will help to increase energy efficiency, reduce air pollution. Also, it reduces carbon emission and provides an environmentally friendly energy consumption. Renewable energy sources are,

- Solar Energy,
- Wind Energy,
- Biomass Energy,
- Geothermal Energy,
- Hydraulic Energy,
- Hydrogen Energy,
- Wave Energy [2].

Biogas, which is an example for the biomass energy is a gas mixture composed primarily of methane, carbon dioxide and hydrogen sulfide, produced in anaerobic environment from raw materials such as agricultural waste, manure, municipal waste, plant materials, sewage, green waste and food waste [3]. Anaerobic environments mean

the production of energy without oxygen. Microorganisms disrupt biodegradable material through a procedure called anaerobic digestion without the presence of oxygen [4]. As a result of anaerobic digestion a precious two substances which are biogas and digestate are produced. Biogas contains an average of 55-80% methane (CH_4), 20-40% carbon dioxide (CO_2) and trace gases, including toxic hydrogen sulphide and nitrous oxide [5]. The materials used for biogas production can be substantially vegetable wastes (organic wastes), animal manure and industrial wastes. Following the separation and pressurization of the carbon dioxide, biogas can be utilized as a fuel for vehicles or in natural gas systems, as well as for the direct production of heat and electricity.

First and foremost, biogas production and biogas plants are discussed in this thesis. The literature is subsequently reviewed and contributions of the study is presented. Afterwards, the formation of biogas plants is introduced. The concepts of Feature selection, which is a machine learning algorithm in artificial intelligence, are given. Various techniques employed are discussed, as well as the findings are consulted.

1.1 Literature Review

According to statistics of renewable energy in Turkey since 2008, the capacity of bioenergy is increased more than tenfold, and as of 2021, it is exceeding 1.6 thousand megawatts [6]. Since the biogas has a great importance in energy production, many studies are carried out on this subject.

Anaerobic digestion is the system of microorganisms that split organic materials (which are made up of plants or animals). Bio-electrochemical anaerobic digestion consists of electrolysis of organic matters. The study in [7] shows the enhancement of methane production by using some machine learning models like tree based, regression and etc. for the bio-electrochemical anaerobic digestion (BEAD) systems.

[8] is about the prediction of biogas production from food waste with machine learning algorithms for adaptive network-based fuzzy inference system (ANFIS) and least square support vector machine (LSSVM). After training the model, outputs are compared with actual result by statistical analyses.

In [9], authors explore the different conditions that effect the biogas production by using the numerical models. Specifically, an artificial neural networks (ANN) is designed, where food waste (FW), fruit and vegetable waste (FVW), or blends of both in co-digestion (CD), reactor/feed type, volatile solid (VS), pH, OLR, hydraulic retention time, temperature, and reactor volume are input variables, and the cumulative biogas production is the output. A database is also built employing the values presented in the literature.

[10] is a study that shows the improvement of the biogas production with smart technologies. Anaerobic digestion parameters such as boot tank level, feed pump, temperature, pH, OLR, VFA, methane and carbon dioxide related sensors and actuators connected via IoT platform devices (Raspberry Pi) are designed, data are collected and stored. The regression analysis is performed and as a result, necessary actions are taken. Data storage needs to be upgraded to perform a comprehensive data analysis.

In [11], authors proposed a machine learning model from the data collected within 8 years for a system which is too complex and nonlinear to be modeled mechanistically. The goal of the model is to show how the different waste inputs and operating conditions affect the biogas yield. By the combination of feature significance and PD analysis a distinction is made between waste streams' impact due to larger incoming volumes and waste streams' impact to per unit waste taken into the digester. This experiments with various time delays also revealed information on how various wastes decompose after their load to the digester.

[12] is a study which offers the design of a hybrid deep learning model with anaerobic co-digestion (AcoD) in order to increase the biogas generation. The proposed model contains dual-stage-attention (DA), long short-term memory (LSTM) combined with variable selection networks (VSNs). In this model, 2-year data are collected from a municipal wastewater treatment plant. To increase the performance of the model, a hyper parameter optimization is used. Moreover, feature importance (FI) is utilized to predict input variables that affect the biogas output. The accuracy of the recent models

is compared with the hybrid model. As a result, the hybrid model is found to be considerably better.

In [13], a predictive model for biogas production at a Chinese biogas plant is presented. In order to distinguish the most significant inputs influencing the biogas production, a methodology employing prediction algorithms to daily production is involved. Machine learning models such as logistic regression, support vector machine, random forest, extreme gradient boosting and k-nearest neighbors regression are applied for two different plants. The best model for Hainan data is k-NN whereas, XGBoost in Shenzhen.

STATISTICA 10 software with artificial intelligence applications is used for the cost-effective optimization of biotechnological processes and the generation of biomethane in [14]. The established methods for converting biomass to biogas are combined with optimization techniques. Furthermore, it has been found that, by using the artificial intelligence models, more optimal outcomes can be obtained for more complicated inputs.

In [15], the authors investigates the biogas facilities in Poland that have the ability to work with agriculture. Analysis of biogas production and biomethane from animal manure is presented and the most significant wastes are identified.

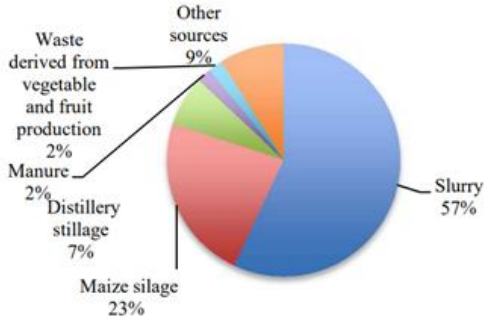


Figure 1.1 Most significant waste [15]

[16] focuses on selecting the most suitable resource for energy production among four different organic materials namely, cow dung, food waste, flower waste and fruit waste. Methane gas is released upon each degradation of trash. At this research, it is determined which waste can yield the largest amount of methane. The analyses conducted led to the conclusion that collecting methane from food and cow manure is faster and more economical, despite the greater methane output from fruit and flower fertilizers are being more expensive.

In [17], in order to increase the biomass capacity of Turkey to overcome the energy demand by the renewable sources , various analyses are performed. The impact of utilizing few trace elements on the production of biogas from leftover corn silage waste is studied. Consequently, although estimating the level of trace compounds is challenging, their presence increased the energy efficiency.

In [18], usage of trace elements from the iron and steel industries is found to boost the yield of biogas production. The study is divided into two sections. Different amounts of substances from each section are subjected to the response surface methodology analysis (RSM) and variance analyses (ANOVA).

1.2 Aim And Significance

Electrical energy is produced by the following 22 different wastes are applied to the biogas plant,

1. Dairy cow manure,
2. Weak vinasse,
3. Beef cattle manure,
4. Poultry manure,
5. Potato peel,
6. Maize silage,
7. Dairy cow manure seperated,
8. Mixed vegie,
9. Peaches,

10. DAF sludge,
11. Potato whole,
12. Fish waste,
13. Wheat juice,
14. Corn pomace,
15. Rumen,
16. Broiler chicken,
17. Chicken fat,
18. Wine pulp,
19. Digester-1,
20. Digester-2,
21. Digester-3,
22. Digestate dry.

The effect of these inputs on electricity production is examined with feature selections and extractions methods with machine learning algorithms. The feature importance technique is employed to check the important levels of the variables after feature selection. The data is scaled and made suitable for the model for the free source Python language and its machine learning packages. The model that provides the best results after being tested and compared with previous structures.

The followings are the contributions to the literature:

- Analyzing the waste in biogas production and identifying the most important wastes with machine learning models, therefore reducing the waste cost.
- Carrying out evaluations for real world, renewable energy sources.

2. BIOGAS PLANT AND MACHINE LEARNING

2.1 Process of Biogas Production

Biogas is a term used to describe gas production from organic materials that have undergone biological processing. It is naturally created as a result of anaerobic digestion, which is the biological conversion of organic carbon into carbon dioxide and methane. At this point, anaerobic microbes biochemically breakdown organic substances such as glucose into carbon dioxide and methane. It is represented in the following formula [4].

$$C6H12O6 \rightarrow 3CO2 + 3CH4 \quad (1.1)$$

2.2 Anaerobic Digestion

Four separate steps of anaerobic digestion are combined to produce methane: hydrolysis, acidogenesis, acetogenesis, and methanogenesis [4]. Hydrolysis of the input materials by bacteria is the initial stage in the digestion process. The decomposition of insoluble organic polymers, such as carbohydrates, produces soluble derivatives that are subsequently accessible to other bacteria [4]. The sugars and amino acids are subsequently converted by acidogenic bacteria into carbon dioxide, hydrogen, ammonia, and organic acids. These organic acids are then converted into acetic acid by bacteria during acetogenesis, along with extra ammonia, hydrogen, and other substances such as carbon dioxide. The compounds are certainly transformed into methane and carbon dioxide by methanogens. The colonies of methanogenic archaea are essential for treating anaerobic wastewater [4]. Physical confinement prevents gaseous oxygen from entering the reaction. Non-oxygen-based electron acceptors are utilized by anaerobes. These acceptors may consist of the organic substance itself or inorganic oxides from the input material. When the oxygen supply in an anaerobic system is sourced from the

organic material itself, the 'intermediate' end products are predominantly alcohols, aldehydes, and organic acids, in addition to carbon dioxide. In the presence of specialized methanogens, intermediates are transformed into the "final" end products of methane, carbon dioxide, and trace amounts of hydrogen sulfide. In an anaerobic environment, methanogenic archaea release the bulk of the chemical energy contained in the starting material as methane [4].

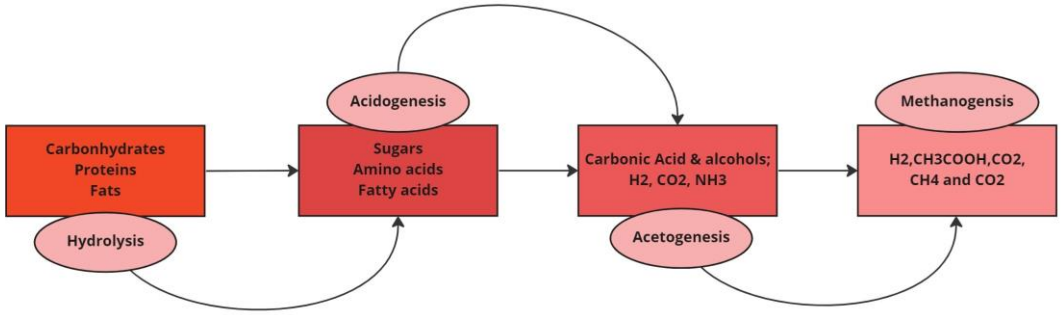


Figure 2.1 Steps of Anaeronic Digestion

2.3 Steps of Anaerobic Digestion

2.3.1 Hydrolysis

In the first stage, the microorganism-secreted cellular enzymes transform into soluble compounds in the sludge. They transform long-chain complex carbohydrates, proteins, lipids, and fats into short-chain compounds. The first process, hydrolysis, is concluded as a result of this conversion to simple organics [4].

2.3.2 Acidogenesis

Soluble organic compounds are transformed into chemicals with tiny structures, such as acetic acid, volatile fatty acids, hydrogen, and carbon dioxide. This process is conducted with anaerobic microorganisms. These bacteria facilitate the growth of methane-producing bacteria [4].

2.3.3 Acetogenesis

Acetogenesis is the third step of anaerobic digestion. In this phase, acetogens degrade simple molecules produced in the acidogenesis phase to yield acetic acid, carbon dioxide, and hydrogen [4].

2.3.4 Methanogenesis

It is the process of bacteria turning acetic acid or hydrogen and carbon dioxide into biogas. The creation of methane is a slower process when compared to other steps. Effective methane-forming bacteria are extremely sensitive to environmental circumstances [4].

2.4 Aspects Affecting Biogas Production

Every aspect that influences the microbiological microorganisms that contribute to biogas creation also influences the amount of biogas production. Biogas generation is influenced by raw material biogas potential, vaccines, substrate type, pH, temperature, loading rate, hydraulic retention time (HRT), carbon nitrogen ratio (C/N) ratio, volatile fatty acids (VFA), and inhibitory compounds, among others [19].

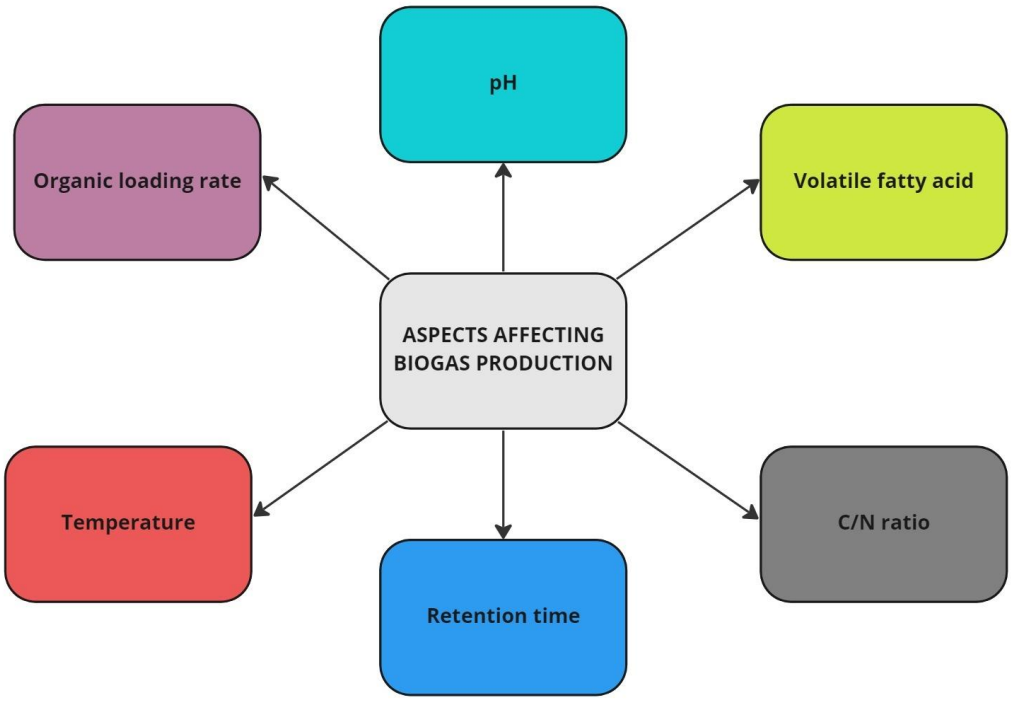


Figure 2.2 Aspects Affecting Biogas Production

Temperature: Methanogenic bacteria are inactive at both extremely high and extremely low temperatures. Therefore, the temperature of the reactor at which biogas will be

produced has a direct effect on the amount or rate of biogas produced. These bacteria are also extremely sensitive to variations in temperature. The dwell duration and the volume of the reactor are also determined by the temperature within the reactor [20].

pH: The pH value is one of the most important parameters in biogas generation. Neutral or slightly alkaline pH levels are a must optimal for methane-producing microorganisms. In the process of anaerobic digestion, the optimal pH range for maximal biogas generation is 6.8-7.2 [18]. When the pH value falls to 6.7, it becomes hazardous to the bacterium. There may be an increase in acid-forming bacteria, resulting in a decrease in pH and the cessation of methane synthesis. In such instances, organic matter is not added to the reactor therefore the acid ratio is decreased. The pH can also be stabilized with chemicals. One of these substances is calcium hydroxide, sometimes called slaked lime [21].

Carbon Nitrogen Ratio (C/N): C/N expresses the relative proportions of carbon and nitrogen in an organic compound. Carbon is used by anaerobic bacteria to produce energy, and nitrogen is required for bacterial growth and reproduction. If the C/N is too high, nitrogen is rapidly absorbed by methanogens to satisfy their protein requirements and lost in organic matter. Due to a lack of nitrogen, the use of carbon sources is inadequate, and biogas might inhibit its development [19].

Organic loading rate: The organic loading rate is the daily quantity of organic material delivered to bioreactors. It is an operational characteristic that influences methane production. The loading rate must be maintained at a somewhat optimal level, or the pH might be lowered to prevent gas production [19].

Parameters	Hydrolysis/ Acidogenesis	Methanogensis
Temperature(°C)	25-35	M : 32-42 T : 45-70
pH	5.2-6.2	6.7-7.5
C/N ratio	10-45	20-30
Dry matter (%)	<40	<30

Figure 2.3 Average Values of Parameters

2.5 Data Analysis and Machine Learning

2.5.1 Dimension reduction

Transferring data from a high-dimensional space to a low-dimensional space is dimension reduction. Thus, the low-dimensional representation maintains a number of essential characteristics of the original data [22]. Consider, for instance, a dataset with 100 features. The dimension reduction approach generates a new feature that encompasses all 100 features by taking into consideration the differences between the features. This allows us to describe the data set with 100 characteristics using five variables. Despite their similarities, dimension reduction and feature selection methods are distinct from one another. Size reduction techniques let the projection of the original features into a place with less dimensions, so creating a whole new set of features. Some size reduction algorithms are given below:

- Principal Component Analysis (PCA)
- Kernel PCA
- Linear discriminant analysis (LDA)

2.5.2 Feature selection

Feature selection is one of the most essential ideas in machine learning, since it has a significant influence on the model's performance. Occasionally, the data set may have 100 or 200 even more features. It can be observed that although some of these features contribute to the model, others negatively impact the model. The purpose of feature selection is to keep and reject a subset of the initial features. Feature selection approaches are utilized for a variety of purposes as listed below:

- Simplification of models to make them easier for researchers/users to comprehend,
- Faster training periods,
- Avoiding the curve of dimensionality,
- Enhancing data’s compatibility with a learning model class,
- Encoding natural symmetries existing in the input space.
- The possibility of overfitting is diminished. Too many variables can lead to an increase in model complexity and overfitting.

- Some variables in a data set may be highly correlated with one another (Multicollinearity). In this instance, we are aware that both variables will have the same explanatory power on the model's output variable (target). Consequently, eliminating one of them eliminates an unneeded variable while improving the model's performance [23].

There are essentially three distinct feature selection methods which are listed below:

- Wrapper Methods
- Filter Methods
- Embedded Methods

Wrapper Methods: Wrapper techniques are a way for preparing, evaluating, and comparing the performance of a variety of feature selection alternatives. During the modeling phase, this approach is utilized, and the model may be recreated based on variable selection. Its objective is to produce a model containing the factors that yield the greatest outcomes. Although this method produces accurate outcomes, it is more costly and time-consuming than the filter method [23]. Some techniques used in the wrapper method are as follows.

- Forward Feature Selection
- Backward Feature Elimination
- Exhaustive Feature Selection
- Recursive Feature Elimination

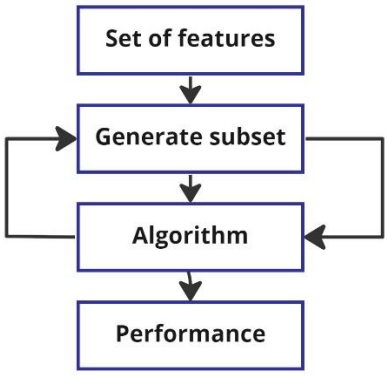


Figure 2.4 Wrapper Methods Algorithm

Filter Methods: Filter-type methods choose variables independently of the model. Only the dependent variable, or the target variable, is correlated. Filtering techniques are a strategy for eliminating ineffective variables and components. A classification or regression model that is used to classify or predict data will incorporate additional variables. These approaches are computationally efficient and overfitting-resistant [24]. Some statistical tests used in the filtering method are as follows.

- Information Gain: It is defined as the amount of relationship between an attribute and its target value and measures the decrease in entropy value. The information gain between each feature and the target variable is calculated by considering the target values for feature selection.
- Chi-square Test: This is a sort of non-parametric testing that employs hypothesis testing and the p-value of feature selection. Typically, the chi-square test is employed to investigate the association between categorical variables. It compares the observed values of a dataset's different attributes to the predicted value.
- Fisher’s Score: Fisher's Score chooses each feature individually according to Fisher criterion scores. This leads to inadequate feature content. The selected characteristic is superior as Fisher's score increases.
- Correlation Coefficient: This strategy suggests that only characteristics with a substantial correlation with the dependent variable should be included in the model. It shows that variables having a weak association are omitted from the model, or disregarded. Spearman's and Pearson’s methods are examples of correlation.
- Variance Threshold: It is a strategy that eliminates these traits if the variance falls below a specified level. This method advises excluding from the model any characteristics having zero variance. This technique implies that variables with high variance, i.e. characteristics, are more likely to contain greater amounts of information.
- Mean Absolute Difference (MAD): This approach computes the average absolute deviation from the mean value.

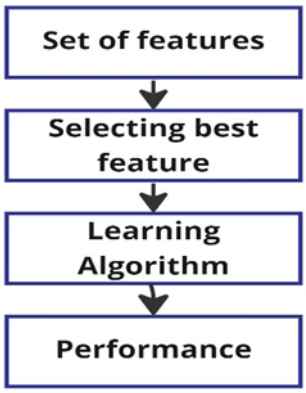


Figure 2.5 Filter Methods Algorithm

Embedded Methods: Embedded techniques show which features contribute most to the accuracy of the model when creating the model. These approaches include algorithms based on Lasso Regression or Decision Trees. The most common embedded techniques are as follows.

- Regularization: This technique penalizes many machine learning model parameters in order to prevent overfitting. Variable selection is given by the basic input variable penalization scheme. This approach uses Lasso (L1 regularization) and Elastic meshes to choose features (L1 and L2 regularization). By penalizing some of the coefficients, they are lowered to zero. Delete features with zero coefficients from the data set.
- Tree-based methods: These decision tree methods, such as Random Forest and Gradient Boosting, pick features based on their characteristics. The connection between a feature and the target variable reveals which traits are more crucial. By selecting the significance threshold, variables that fall below it are omitted from the model. Thus, the model's accuracy rate is enhanced [25].

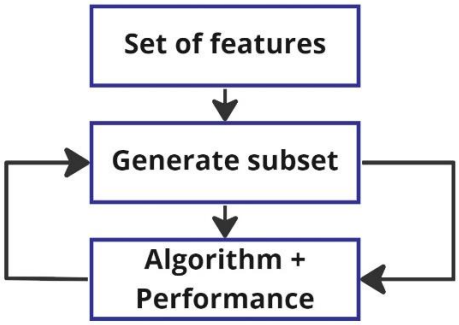


Figure 2.6 Embedded Methods Algorithm

2.6 Machine Learning

Artificial intelligence (AI) is the intelligence exhibited by machines, such as sensing, synthesizing, and inferring information, as opposed to the intelligence exhibited by animals and humans. According to the Oxford English Dictionary, artificial intelligence is the theory and development of computer systems capable of doing activities that ordinarily require human intellect, including visual perception, voice recognition, decision-making, and language translation [26] .

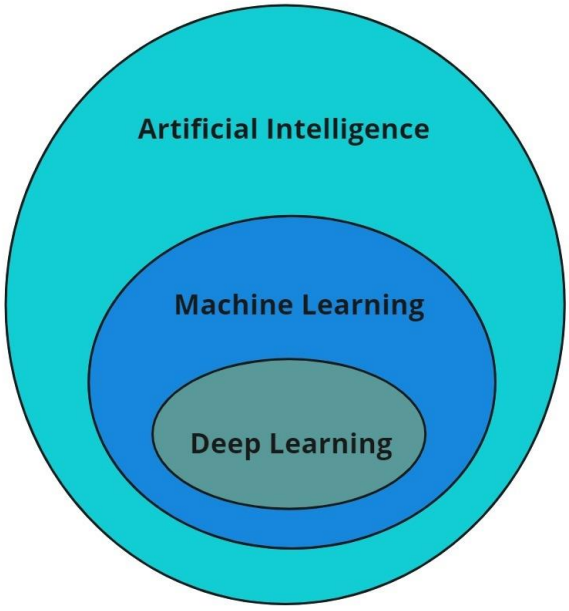


Figure 2.7 Artificial Intelligence

Machine learning is a subfield of artificial intelligence, computer science and research that has developed by simulating how humans learn and focused on increasing its accuracy through data and algorithms with mathematical and statistical methods. Essentially, machine learning utilizes historical data to develop a model and predict the future. There are essentially three distinct machine learning approaches which are supervised learning, unsupervised learning and reinforcement learning [27].

2.6.1 Supervised learning

The most prevalent machine learning methods are supervised machine learning algorithms. Supervised learning is a system for learning in which the target (in the literature target variable can be called dependent, output, response) variable is used to inform data-driven predictions. It is utilized in several sectors and academic disciplines, for example, disease diagnosis, fraud detection, etc. Classification and regression techniques are part of supervised learning. Some popular classification algorithms are:

- Logistic Regression,
- Naïve Bayes,
- K-Nearest Neighbors (KNN),
- Decision Tree,
- Support Vector Machine (SVM), etc.

Some popular regression algorithms are:

- Linear Regression,
- Ridge Regression,
- Lasso Regression,
- Neural Network Regression,
- Random Forest,
- Decision Tree, etc.

Classification models are not given because regression models were employed in this investigation.

2.6.2 Unsupervised learning

Unsupervised learning is a sort of learning in which the dependent variable is absent. In other words, it is a strategy that employs a function to anticipate an unknown data

structure. Unsupervised learning is more complex than supervised learning since there is litter to no data-related information. It is mostly utilized segmentation problems. The forms of unsupervised learning studied include clustering and dimension reduction. Most common clustering algorithms are as follows.

- K- Means Clustering,
- Dimensionally Reduction.

2.6.3 Reinforcement learning

Reinforcement learning is structured differently from supervised and unsupervised learning. It is a training strategy for machine learning based on rewarding and/or penalizing desired behavior. The objective of the machine is to identify the optimal path to the intended action; it draws conclusions from the errors it made along the route and operates on a reward-punishment system. The system then attempts to discover the correct action with the least amount of mistake based on its deductions. Autonomous vehicles and learning to play against a human opponent both employ reinforcement learning algorithms [28].

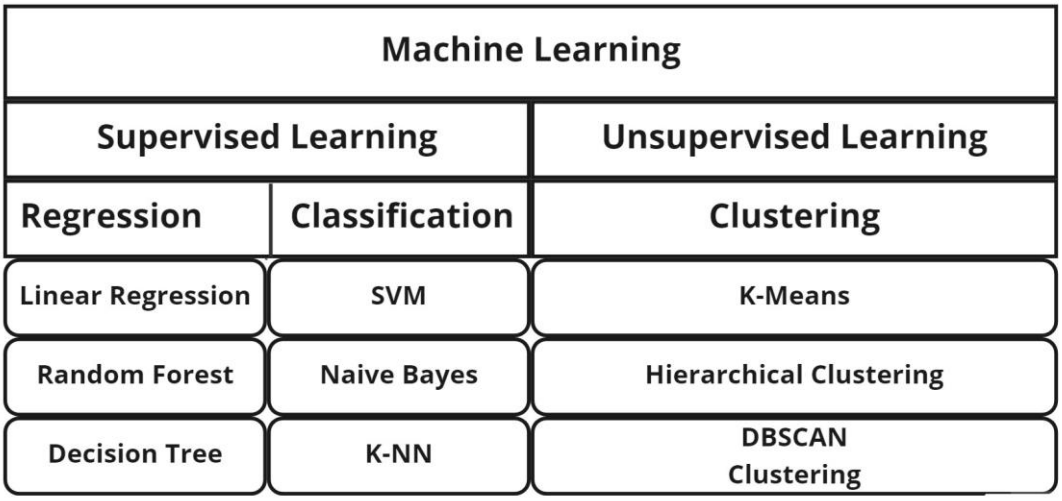


Figure 2.8 Types of Machine Learning Algorithms

2.7 Regression Models

Regression methods are utilized when there is a correlation between the input variable and the output variable. In general, regression analysis is utilized in the study of continuous variables such as weather forecasts, automobile price predictions, etc.

2.7.1 Linear regression

Linear regression is a method for modeling the connection between dependent (features) and independent (target) variables using a linear equation [29]. In the expression X are the features and Y represent the target.

$$Y = a + bX \tag{2.1}$$

Bias: Bias is the discrepancy between the predicted parameter's expected value and its actual value. High bias is referred to as underfitting. It is the circumstance of the model not learning its data set at all.

Variance: Variance is a statistical measure of the extent to which data points within a sample or dataset are spread. A significant variation may lead the model to learn and replicate the data exactly [30].

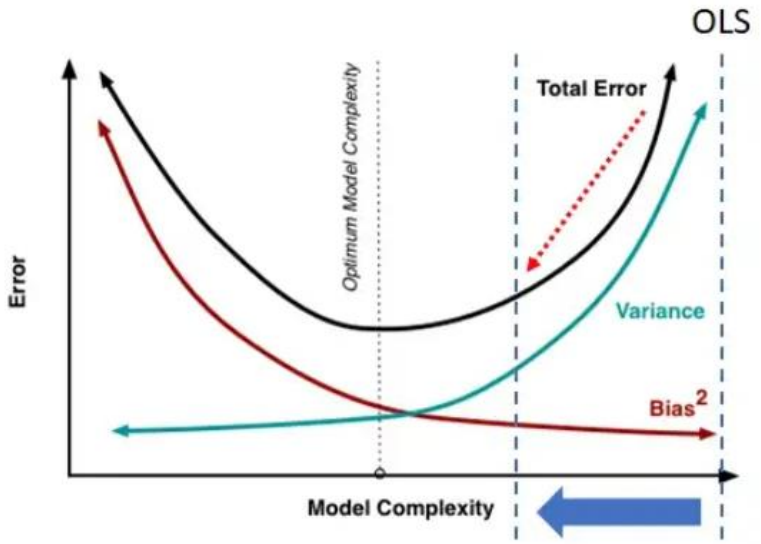


Figure 2.9 Model Complexity [31]

Adding more parameters to a model enhances the model's complexity. Figure 2.9 show that as model complexity rises, high variance has low bias. This paradigm is simultaneously complex and less complex. The right model, however, is one with minimal bias and variance. The capability to portray data is potent. Ridge and Lasso are regression techniques that can circumvent the complexity of the model.

2.7.2 Ridge regression

When the features, that is, the independent variables, are highly correlated, ridge regression is utilized to estimate the coefficients of multiple regression models. L2 regularization is a common form of regression model which means that called the penalty term, adds the "squared magnitude" of the coefficient to the loss function [32]. Estimating the coefficients now begins with the fundamental formula for the research sum of squares (RSS) and the penalty component is added. To determine the penalty time, λ (≥ 0) is defined as the correction parameter (λ) multiplied by the sum of the squares of the coefficients, as shown in the following equation [33].

$$\sum_{i=1}^{n}(y_i - \beta_1 * x_i + \beta_2)^2 + \lambda * {\beta_1}^2 = RSS + \lambda * {\beta_1}^2 \text{ [33]} \quad (2.2)$$

2.7.3 Lasso regression

Ridge regression and Lasso regression are quite similar. While L2 is edited using Ridge regression, L1 is regularization using Lasso regression. As a penalty term, L1 adds the "absolute magnitude value" of the coefficient to the loss function. During the L1 regularization, certain model attributes might be ignored. In other words, Lasso regression is crucial not just for decreasing over-learning, but also for feature selection.

2.7.4 Elastic net regression

The elastic net technique is a regularized regression approach that combines the L1 and L2 penalties of the lasso and ridge methods linearly [34].

2.7.5 Random forest regression

Random forests consist of numerous decision trees, this approach is termed ensemble technique. To quickly discuss decision trees, we may utilize a series of questions to arrive at an estimate and collect data to generate a question-and-answer flowchart in order to determine the most trustworthy outcome. It is utilized in both classification and regression situations. It is a bagging strategy. Returns the mean or average estimate of individual trees. Using a modified tree learning method, a random subset of characteristics is picked at each candidate split in the learning phase [35] .

2.8 Neural Networks

The interpretation and processing of information by the human brain is modeled by neural networks. Each process has several layers. A neural network typically consists of three layers: the input layer for communicating input variables, the hidden layer for processing hidden variables, and the output layer for receiving output variables. Data spreads from neuron to neuron in each layer. The network analyzes the data and develops a model, a prediction, for each data. Weights are utilized in the functions utilized during creation. When it makes an incorrect estimate, the weights are modified and it gains knowledge. After several rounds, the network's predictions improve [36].

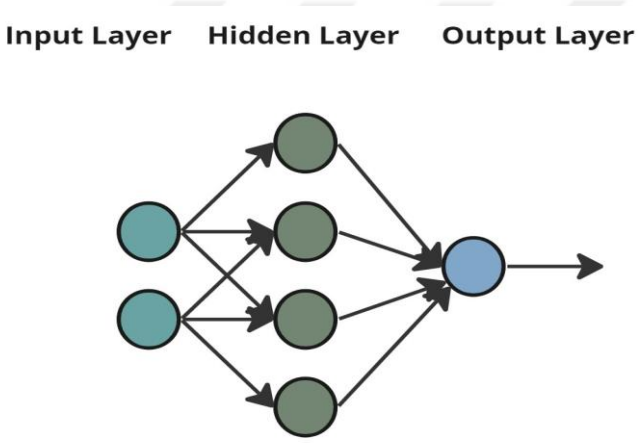


Figure 2.10 A Simple Neural Network

Figure 2.10 depicts two green perceptrons that make two distinct judgments based on the input. These judgments are sent to the subsequent four perceptrons. Complex procedures make advantage of additional hidden layer.

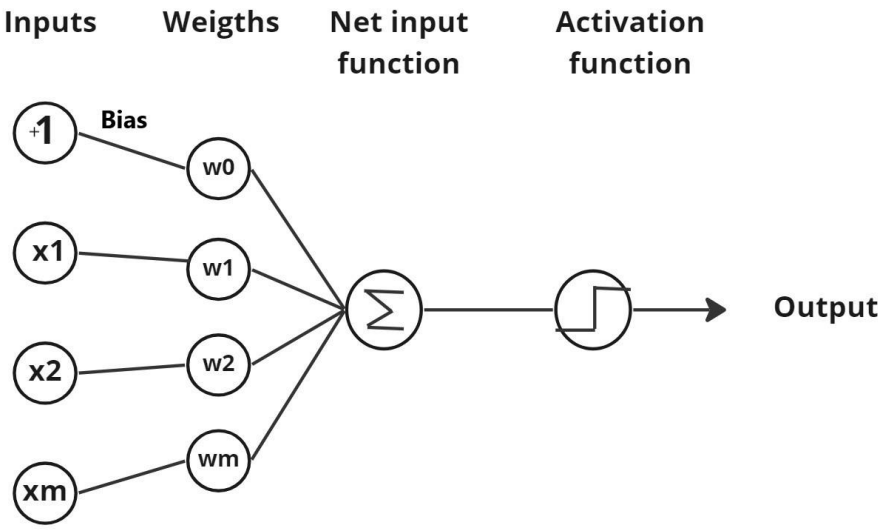


Figure 2.11 Structure of a Single Neuron

$$Output = f(\sum_{k=1}^{m} x_k * w_k + w_0) \quad (2.3)$$

Figure 2.11 shows that the nodes are the structures that compose layers. [37] According to 2.3 formula f represent an activation function. The figure's nodes determine the relevance and weight of the receiving inputs. Then, normalizes the output with activation function. If its output reaches a certain threshold, it "fires" (or activates) the node, which then forwards the data to the next network layer. The output of one node becomes the input of the subsequent node. This neural network is classified as a feedforward network since data is sent from one layer to the next. [38]

2.8.1 Parameters of neural networks

Creating neural networks requires the determination of some parameters. By modifying these parameters according to the model, the model's performance may be enhanced.
Number of neurons in the hidden layer: There are no hidden neurons in the input or output layer. Enhances processing power and system adaptability. Too many hidden neurons enhance the model's complexity, while too few might prevent the model from generating a fitting.

- Learning Rate: During modeling training, it is the parameter that controls deviations and changes in learning.

- Epoch: Number of iterations.

Activation Function: The selection of the activation function in the hidden layer serves to evaluate how successfully the network model acquires knowledge from the training set. The activation function changes based on the output layer prediction model. Without the activation function, the network is only a linear regression model, as the weight and bias would only have a linear equation, which is a polynomial of the first order. It is straightforward to solve, but insufficient for solving complicated and tough problems. Two types of Activation Functions can be distinguished, these are Linear Activation function and Non-linear activation functions.

- Linear Activation Functions: The linear activation function, often known as a straight line, is exactly proportional to the input weights and neuronal inputs.

$$f(x) = ax + b \tag{2.4}$$

  The inability to define the linear enabling function inside a certain range is a drawback. If the linear enable function is applied to each layer, the model will behave similarly to linear regression. The last layer of the neural network will act similarly to a linear function of the previous layer.

- Non-Linear Activation Functions: Nonlinear activation functions are the most commonly utilized in neural networks. It facilitates a neural network model's ability to differentiate between outcomes and adapt to a batch of data.

Sigmoid Activation Functions: The sigmoid function reduces any entered real number to a value within the specified interval (0,1).

Tanh Activation Functions: The nonlinear tanh function is the function between neural network layers. Unlike the sigmoid function, tanh evaluates values ranging from -1 to 1.

ReLU Activation Functions: Rectified Linear Units, also known as ReLU, is a nonlinear activation function. Similar to Sigmoid, but far more effective. Its formula is between Maxima (0,z).

Maxout: The ReLu function is generalized by this function. It is a piecewise linear function that returns the input with the greatest value. Because it is a specific instance of ReLu, it enjoys its benefits. As a result, more overall parameters must be taught, as the total number of parameters for each neuron is doubled.

ELU: The Exponential Linear Unit or ELU activation function is a function that tends to generate more accurate results and converges faster than other functions. ELU has a larger number of alpha constants than other functions and is comparable to ReLU except for its treatment of negative inputs. In the case of nonnegative inputs, both are identity functions. The ELU smoothes out gradually until its output equals -α, whereas the ReLU smoothes out abruptly.

Softmax Activation Functions: The Softmax activation function is used to compute the probability distribution of an event among 'n' distinct events. This function determines the probability of each target class over all possible target classes. The target class for the specified inputs is then determined using the computed probability.

3. APPLICATION OF FEATURE SELECTION FOR DATA ANALYSIS

As previously explained, the objective of this study was to develop a system that makes predictions using deep learning architecture and machine learning architectures, which are subfields of artificial intelligence and machine learning. The data set was obtained from the Pales biogas facility whose production of electricity is of global significance. The data set comprises variables such as the date, the amount of biogas generated, the digester, etc. This data set, which may have outliers because it represents actual production data, contains numerical variables. The objective of the study was to improve the model and reduce waste costs by identifying the most influential variable impacting the model, namely biogas generation, relative to earlier studies.

It was identified that the outputs and biogas production were mainly correlated to a subset of the facility's wastes, and the model was developed with the assumption that the predictive model would feed on this information and accurately predict the biogas production for the following days.

The established model was examined by trying different methods. The application was written in Python program, Numpy, Pandas and Sklearn feature selection for data editing, Seaborn, Matplotlib and Plotly for data visualization, Regression, Tensorflow and Keras libraries were used for machine learning model.

3.1 Dataset Analysis

In this study, machine learning and deep learning applications will be described for obtaining biogas, one of the most important renewable energy sources in the world, and electrical energy from biogas in the most efficient manner, with the goal of reducing waste cost by identifying the waste that contributes the most to the model with the developed algorithm.

A model was developed for the application by analyzing 18 variables of 654 days of biogas production data from the Pales biogas production plant. Also, the effect of waste from other three digesters was evaluated. For that as well, the Python programming language was utilized, and an artificial intelligence model was developed to discover the most significant waste contributing to the model in the creation of electrical energy.

Methane generation, carbon dioxide production, pH, etc. are included in the dataset. These variables were eliminated from the dataset because they were judged superfluous for the model. On some days in the dataset, certain wastes were not used; hence, the dataset contains null values. On the premise that the number was not excessive, certain remaining blank values were filled in using coding based on the literature. The data are displayed in the table.

Features	Explanation	Type
Date	Production day	Datetime
Dairy cow manure	Waste	Integer
Weak vinasse	Waste	Integer
Beef cattle manure	Waste	Integer
Poultry manure	Waste	Integer
Potatoe peel	Waste	Integer
Maize silage	Waste	Integer
Dair cow manure seperated	Waste	Integer
Mixed vegie	Waste	Integer
Peaches	Waste	Integer
DAF sludge	Waste	Integer
Potatoe whole	Waste	Integer
Fish waste	Waste	Integer
Wheat juice	Waste	Integer
Corn Pomace	Waste	Integer
Rumen	Waste	Integer
Broiler Chicken	Waste	Integer
Chicken Fat	Waste	Integer
Wine Pulp	Waste	Integer
biogas	Energy produced	Integer

Figure 3.1 Data review

The data contains 3 different digesters which are Digester 1, Digester 2 and Digester 3. In the analysis, each of these digesters was examined separately, assessed, and then investigated collectively.

After evaluating the data, the correlation between the features and the target variable was investigated, and the Pearson and Spearman feature selection methods were employed. The Pearson and Spearman correlation techniques were utilized since both the data input variable and the output variable are numerical data. It was discovered that the majority of the data set contained null values, which affects the selection of features. When examining all three digesters, it was discovered that the relationship between the “Dairy cow manure”, "Poultry manure" features and the target variable was particularly strong in the first digester, followed by the relationship between the “Dairy cow manure”, “Digester 1” , and “Digester 2” features and the target variable in the second digester, and finally the relationship between the “Dairy cow manure”, ”Poultry manure”, ”Potatoes’ whole”, ”Digester 2”, “Digestate dry” and ”Wheat Juice” features and the target variable in the third digester. The diagrams below show the association between each digester and the target variable. The branches in the figures 3.2 3.3, 3.4 and 3.5 show the length of the branches, referred to as dendrograms in the scientific community, and the distance between variables or sets of variables estimated using bivariate Pearson correlations.

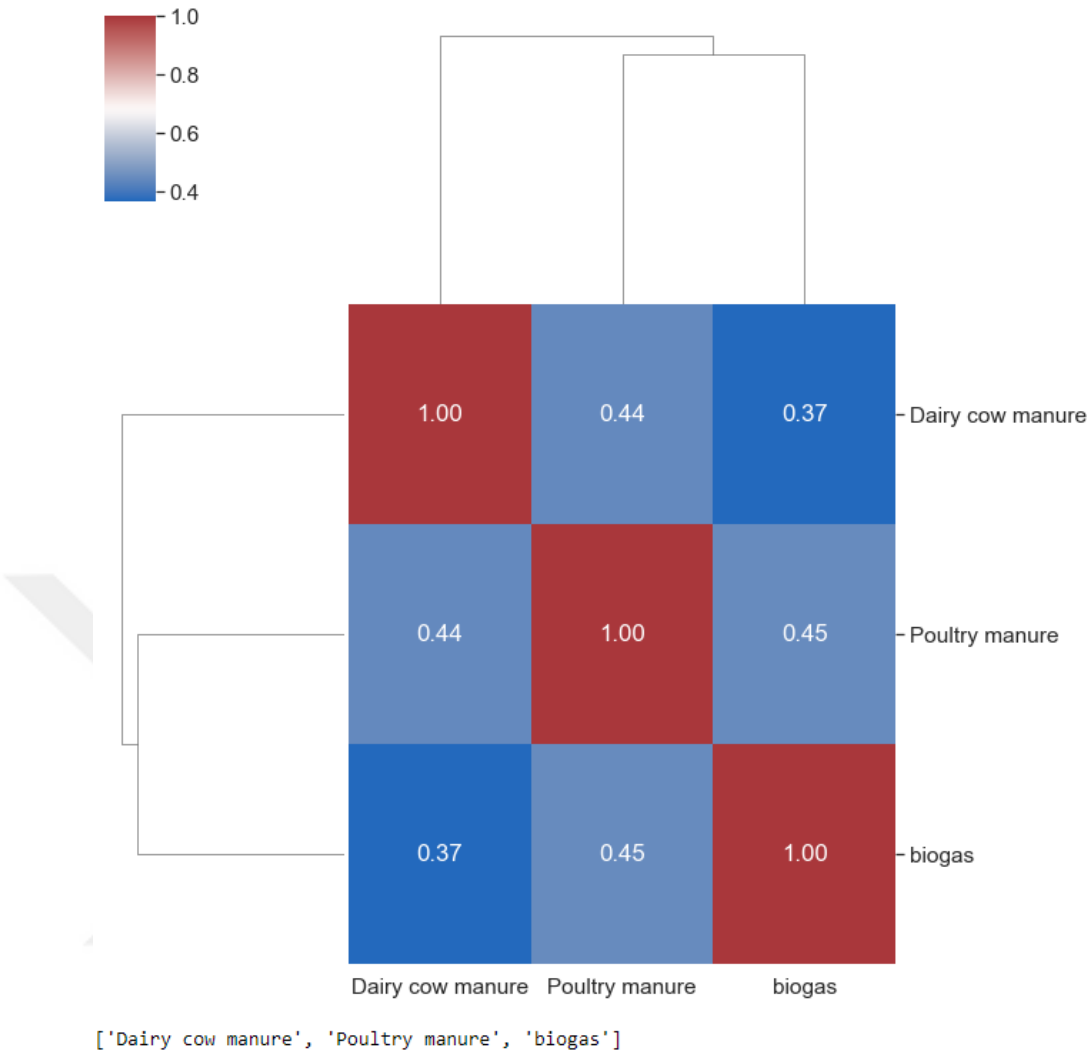


Figure 3.2 Digester 1 - Correlation matrix with target variable

Figure 3.2 represents the correlation between the features in the data and the dependent variable. While defining the correlation matrix, it was stated that the variables above a certain threshold (>= 0.30) should be found in the figure. The line on the figure, namely the dendograms, expresses the relationship between the features. In the data, variables having a strong correlation indicate each other. However, there are no significantly associated characteristics in this data.

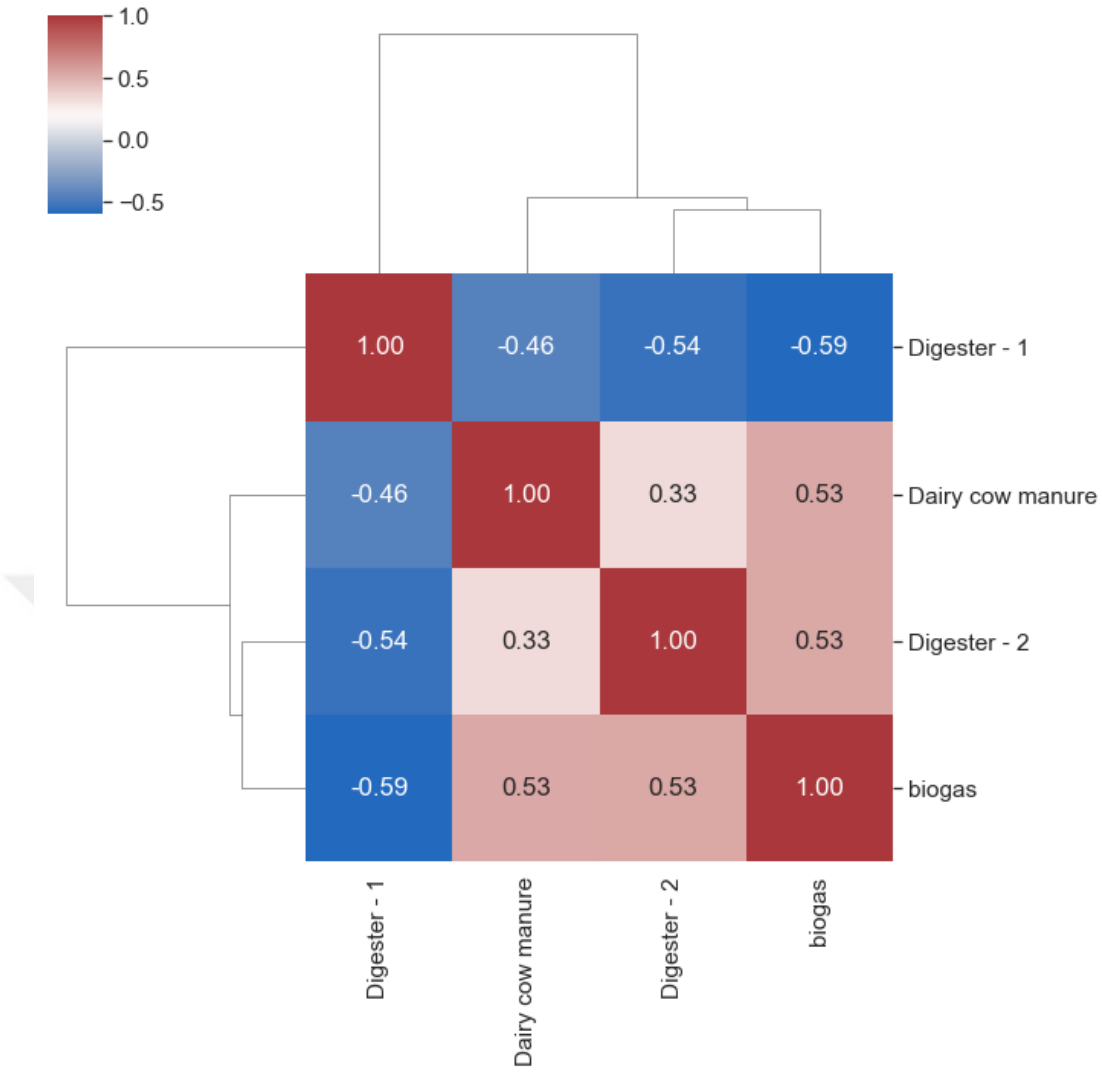
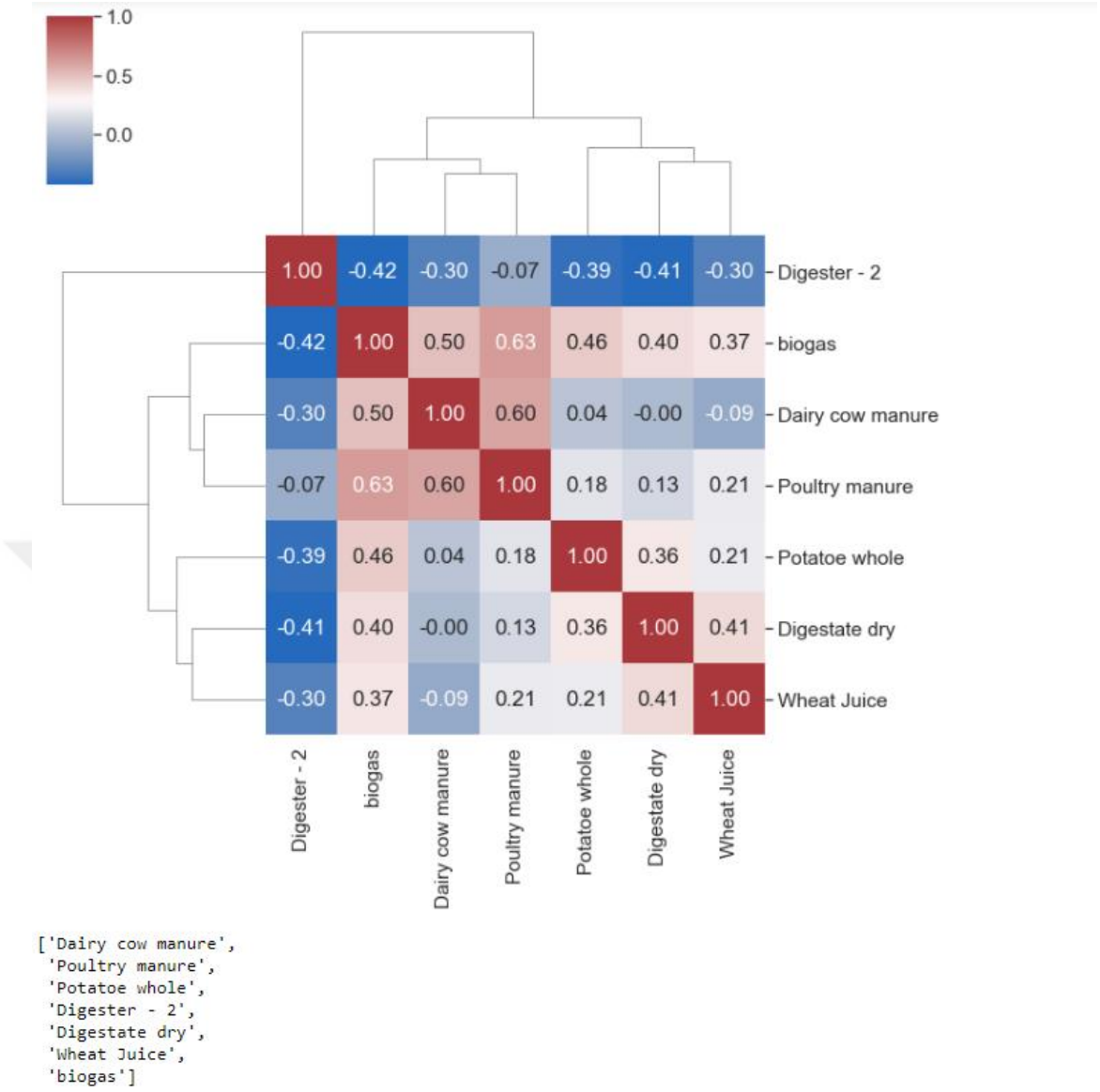


Figure 3.3 Digester 2 - Correlation matrix with target variable

As seen in Digester 2, the relationship between the variables is expressed. Some features have positive correlations, whereas others have negative correlations. In this instance, it impacts the performance of the model.




```
['Dairy cow manure',
 'Poultry manure',
 'Potatoe whole',
 'Digester - 2',
 'Digestate dry',
 'Wheat Juice',
 'biogas']
```

Figure 3.4 Digester 3 - Correlation matrix with target variable

As seen in Digester 3 in Figure 3.4, the relationship between the feature and the target variable is expressed. While the specified features positively impact biogas output, the Digester-2's waste negatively impacts biogas production.

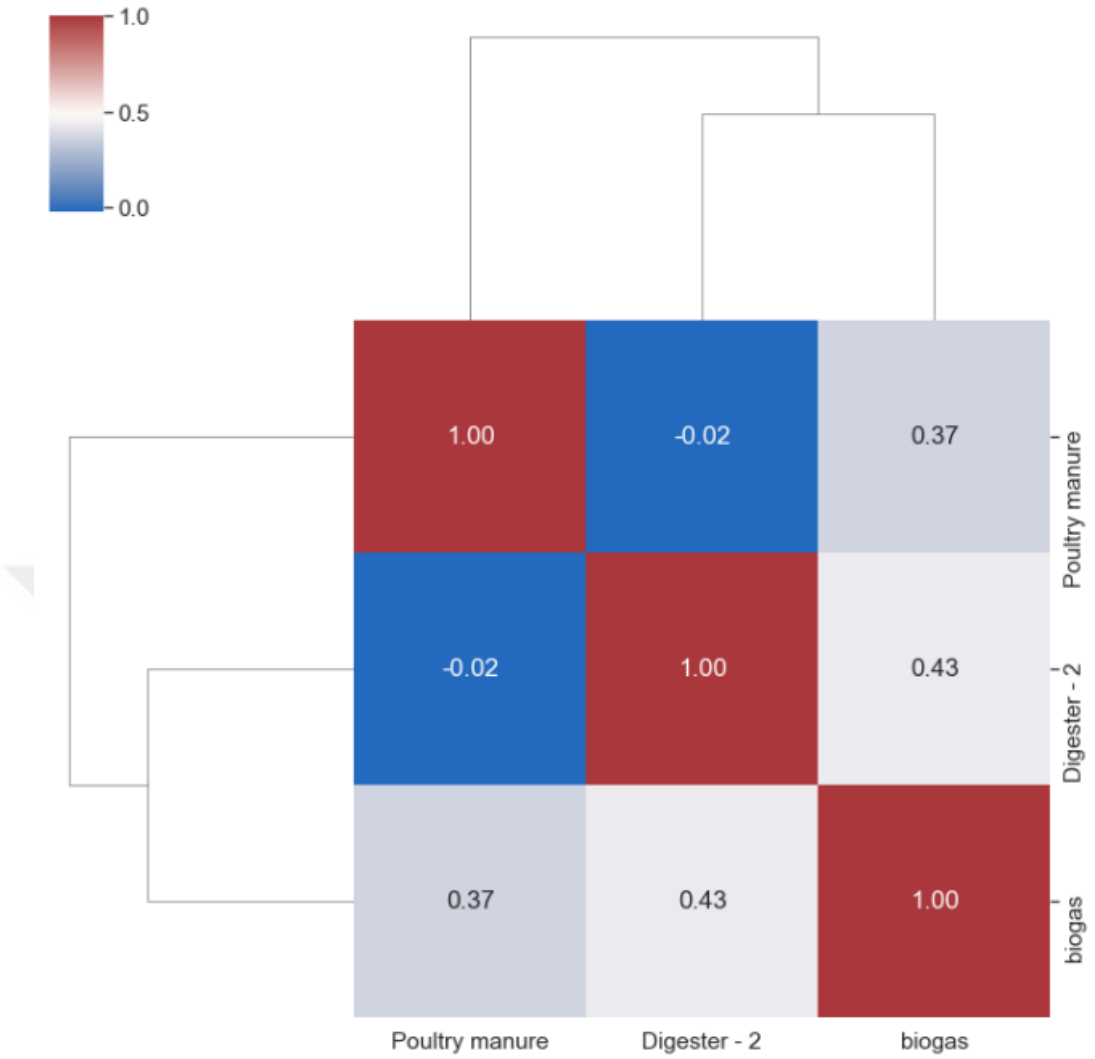


Figure 3.5 All data - Correlation matrix with target variable

After analyzing all digester data, it was revealed that 'Poultry manure' and 'Digester 2' were the most substantially linked characteristics with the dependent (target) variable.

3.2 Proposed Model

General terms, machine learning models in Python consist of the following stages:

- Explore the dataset
  - After reading the data, statistical summaries and visualization studies are performed.
- Preprocessing

- This stage finds and preprocesses data types, missing data, and outliers. Normalization and standardization processes are performed and also, feature selection methods. Moreover, feature engineering is used to generate additional variables that will contribute to the model.

- Split the data train and test
  - The train test split process is used to estimate the performance of machine learning algorithms and to train the model to make predictions on data not utilized in the model.
- Training the model
  - Training a model involves identifying (learning) appropriate values for all weights and bias from labeled examples.
- Model evaluation
  - Various metrics (depends on regression or classification model) are used to assess the performance of the model.

3.3 Model Selection

The performance measures for the regression analysis performed in the study include mean squared error, root square mean square error, absolute errors, and R^2.

- Mean Square Error (MSE)
  - It allows for the determination of the mean error in the prediction model. In the formula below, y_i represent the actual value and $\hat{y}_i$ represent the estimated value.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3.1)$$

- Root Mean Square Error (RMSE)
  - The root-mean-square error (RMSE) is a common measure of the variations between the values predicted by a model or estimator and the actual values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3.2)$$

- Mean Absolute Error (MAE)
  - It is calculated by adding the absolute differences of the observations and dividing by the total number of observations.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (3.3)$$

- R^2
  - Metric of the goodness of fit of a model. The R^2 score is a statistical metric that indicates, on a scale from 0 to 1, the accuracy of our model's predictions.

3.4 Structure of the Model

The data set was split into 20% test data and 80% training data, and a model was then developed. Null attributes are not included in the model, these null variables are Dair cow manure separated, Fish waste, Corn Pomace, Rumen, Broiler Chicken, Chicken Fat and Wine Pulp. In the analysis, the model was trained first using the remaining variables and then with the features identified during feature selection.

This study began with the development of regression models, the application of several regression algorithms which are Linear, Ridge, Lasso, Elastic Net, Random Forest, regression, the evaluation of the outcomes, and the selection of the optimal method. In the regression models used, the data were divided into training and testing. After the model was trained, prediction models were created with the training data. The mean squares error was found by comparing the training set with the predicted values. Then the model was estimated with the test data. The predicted data were compared with the test data.

Following the creation of the regression model, the artificial neural network model was developed. The model consists of 4 layers. In the model, the sequential model was utilized. There is one input layer, there are two intermediate layers and one output layer in the model. By raising the epoch number, the influence of the epoch on the model and the model's evolution were evaluated. This model also favors the 'ReLU' and 'tanh' activation functions, which provide the maximum accuracy by comparing the accuracy values of various activation functions. Since this is a regression model, the output layer utilizes a linear function.

4. RESULTS AND DISCUSSION

4.1 Visual Analysis

In the study, visual analyses were used to inspect and evaluate the data distribution.

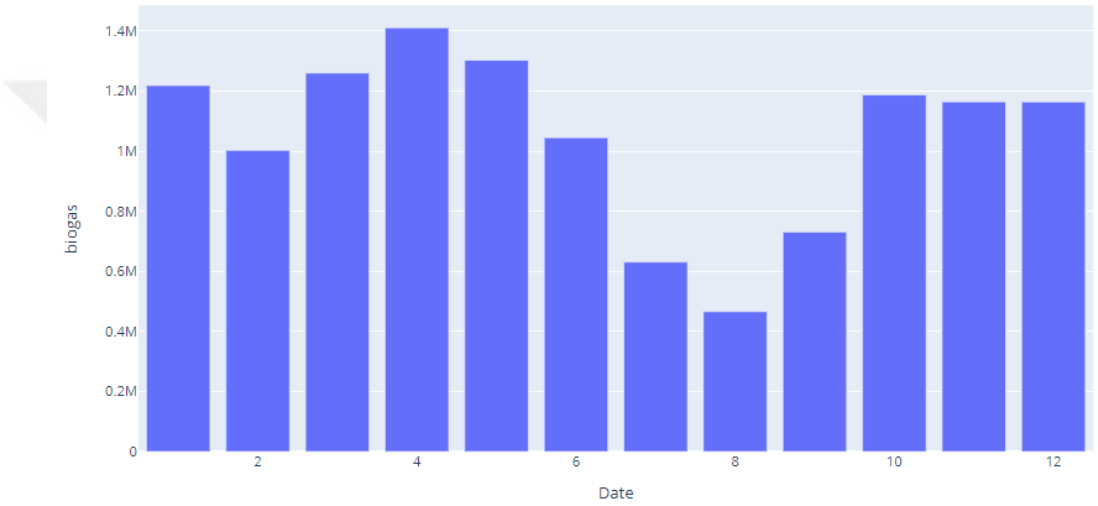


Figure 4.1 Monthly Biogas Production (m^3)

The volume of biogas generation each month is shown in cubic meters in the graph above. The production from September 2019 to June 2021 is depicted in the graph.

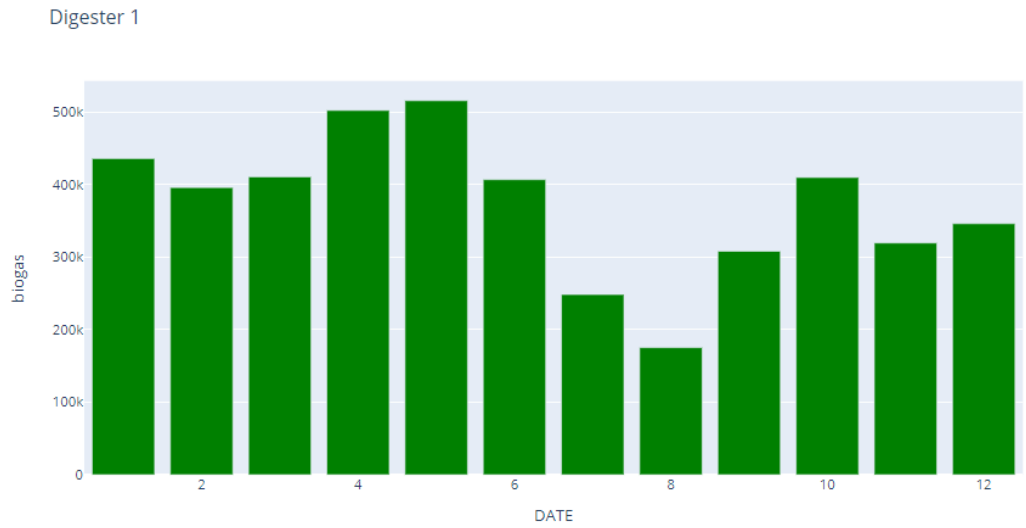


Figure 4.2 Digestion 1 Monthly Biogas Production (m^3)

The green graphic depicts the quantity of biogas produced by the first digester. April and May are shown to be the months with the maximum yield.

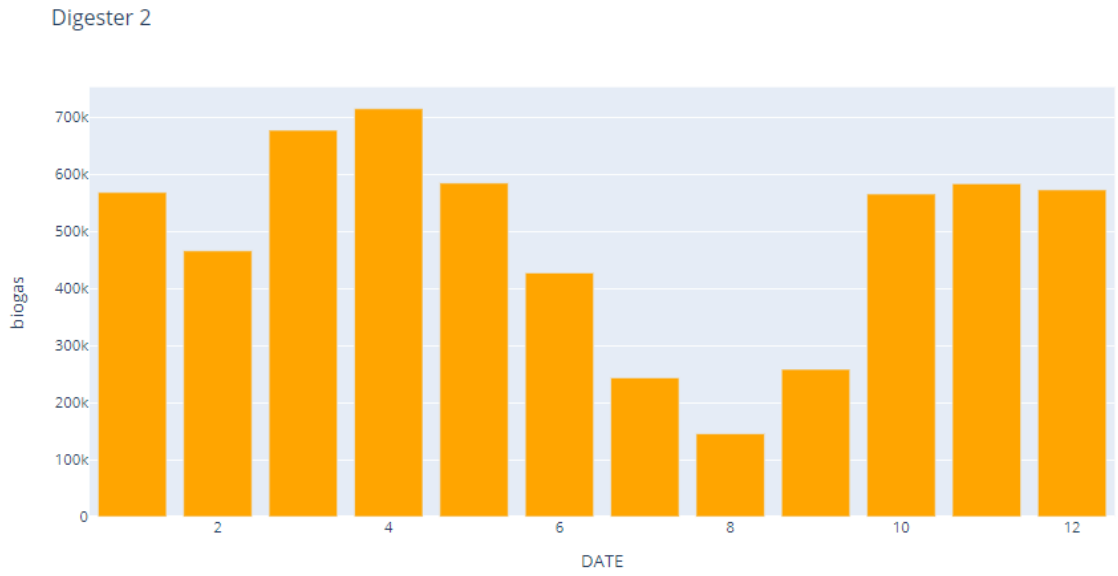


Figure 4.3 Digestion 2 Monthly Biogas Production (m^3)

The quantity of biogas generated by the second digester is seen in the orange bar graph located above this page. Here, the most productive months are March and April.

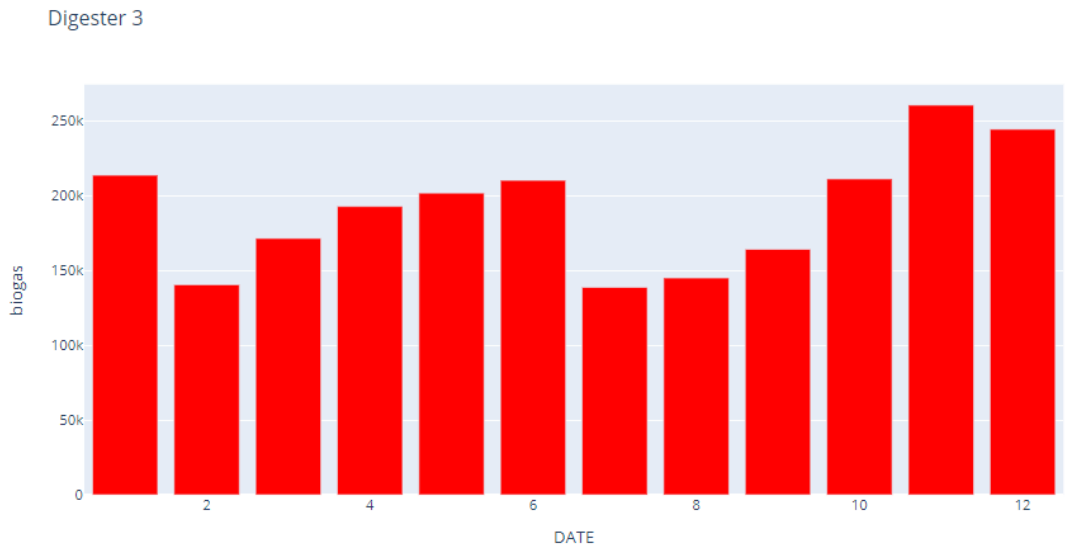


Figure 4.4 Digestion 3 Monthly Biogas Production (m^3)

In the third digester, biogas output increases in the first few months of the year, decreases in the middle of the year, and then increases in November and December as the year comes to a close.

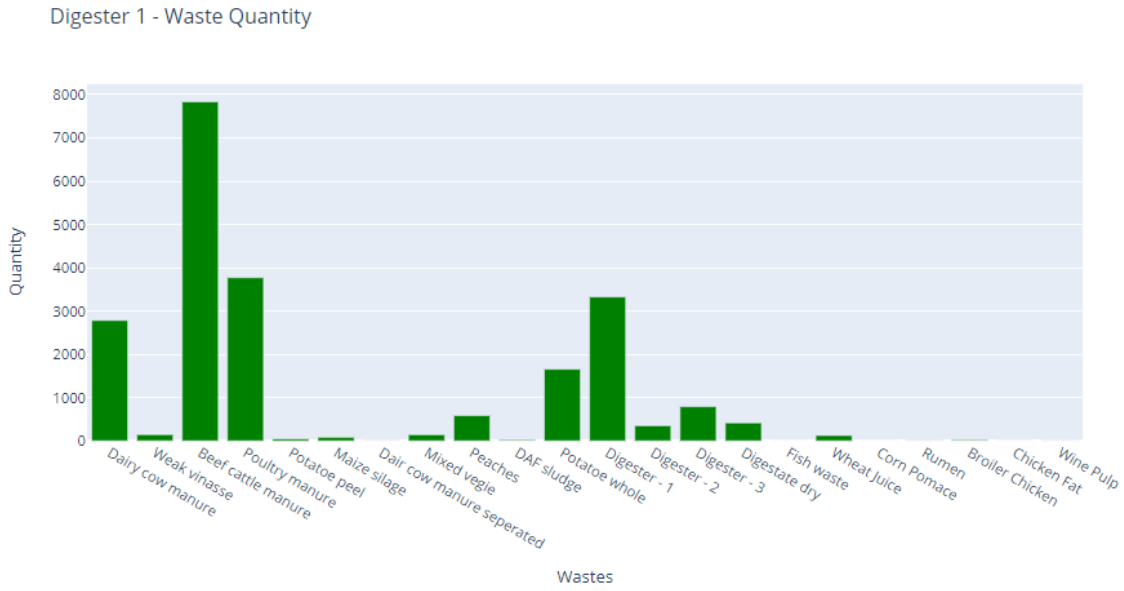


Figure 4.5 Digester 1 Waste amount data (tons)

The amount of waste in the first digester is shown in tons in the figure 4.5. The variables with the highest concentrations in the first digester include Beef cattle manure, Poultry manure, Dairy cow manure, and the first digester itself.

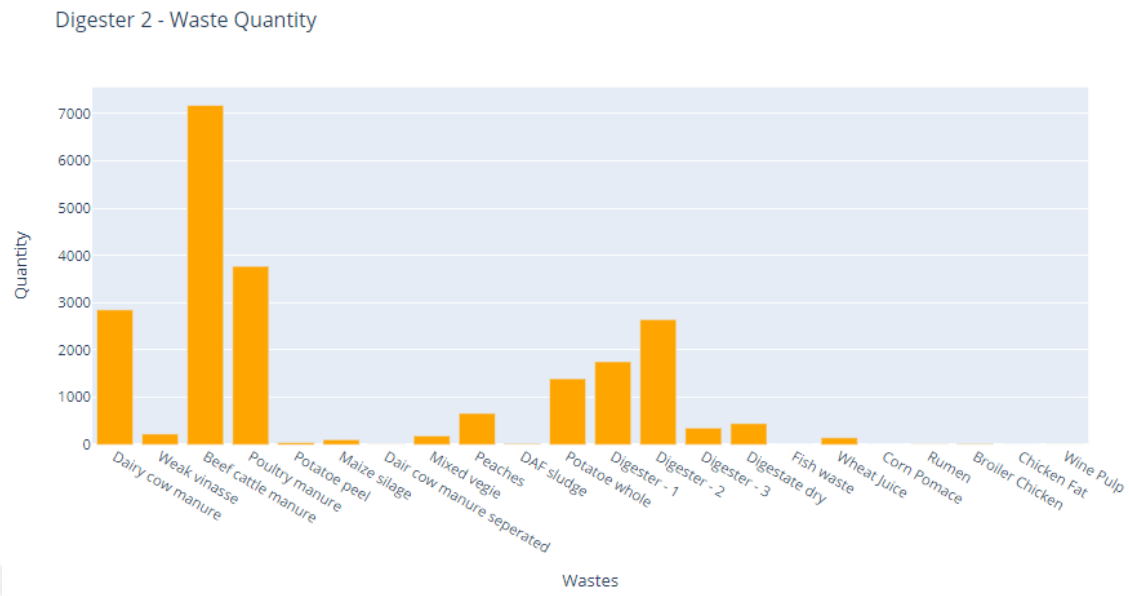


Figure 4.6 Digester 2 Waste amount data (tons)

The amount of waste in the second digester is shown in tons in the figure 4.6. Similar to the first digester, the second digester mostly contains Beef cattle manure, Dairy cow manure, and Digester-1. Additionally, the effect of Digester-2 is also observed.

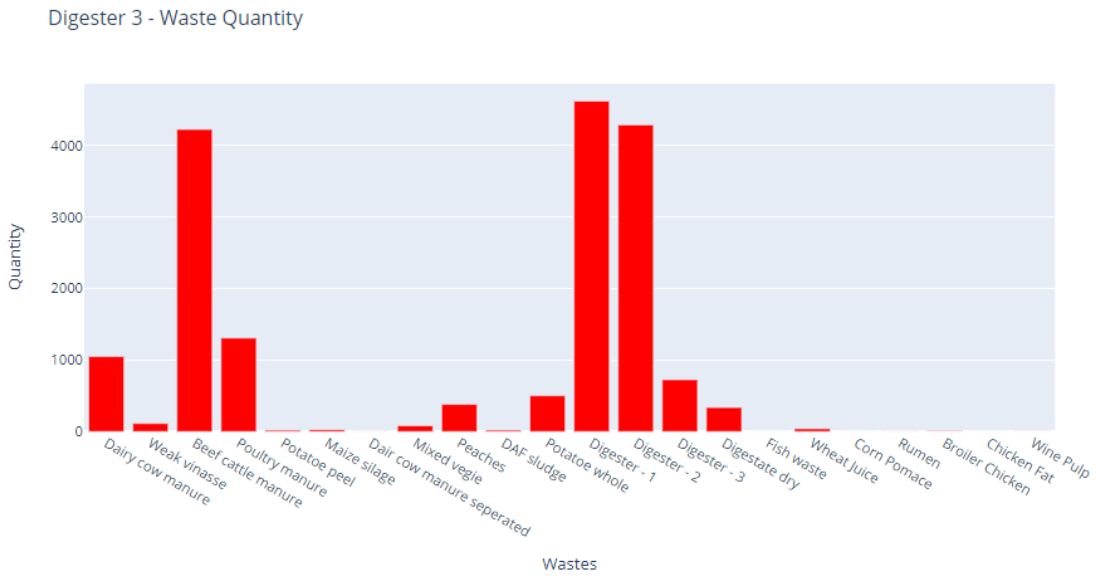


Figure 4.7 Digester 3 Waste amount data (tons)

The amount of waste in the third digester is shown in tons in the figure 4.7. This digester contains significantly less Beef cattle manure than the previous two digesters. Digester-1 and Digester-2 are largely influenced by Digester 3. Examining additional

common features reveals that they explain for nearly half of the waste from Digester-1 and Digester-2.

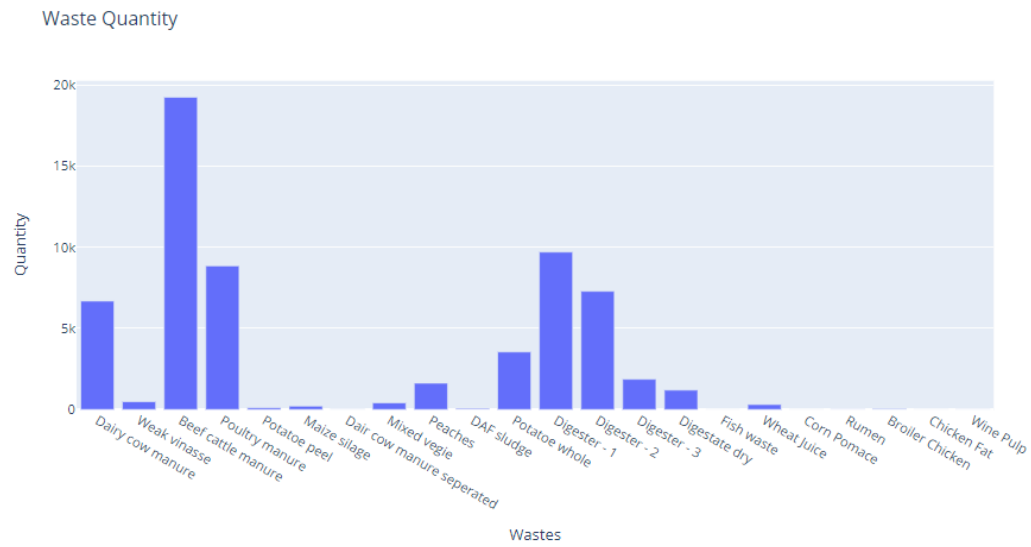


Figure 4.8 Waste amount all data (tons)

Waste amount data in the whole digester is displayed in figure 4.8. Examining the total quantity of waste in the data reveals that Beef cattle manure waste represents the majority of waste in each digester.

4.2 Proposed Models

The model outputs were examined after data analysis and correlation analyses were validated by visual analysis.

```
Feature: 0, Score: 239.41213
Feature: 1, Score: 493.31500
Feature: 2, Score: -1.02146
Feature: 3, Score: 128.15694
Feature: 4, Score: 227.03826
Feature: 5, Score: -1745.71899
Feature: 6, Score: -0.00000
Feature: 7, Score: 43.16844
Feature: 8, Score: 32.25010
Feature: 9, Score: -148.40539
Feature: 10, Score: 102.43222
Feature: 11, Score: -16.92903
Feature: 12, Score: 160.14923
Feature: 13, Score: -135.57326
Feature: 14, Score: -44.65830
Feature: 15, Score: -0.00000
Feature: 16, Score: 1798.33642
Feature: 17, Score: 0.00000
Feature: 18, Score: 504.41234
Feature: 19, Score: -275.42750
Feature: 20, Score: 345.69105
Feature: 21, Score: -283.97763
```

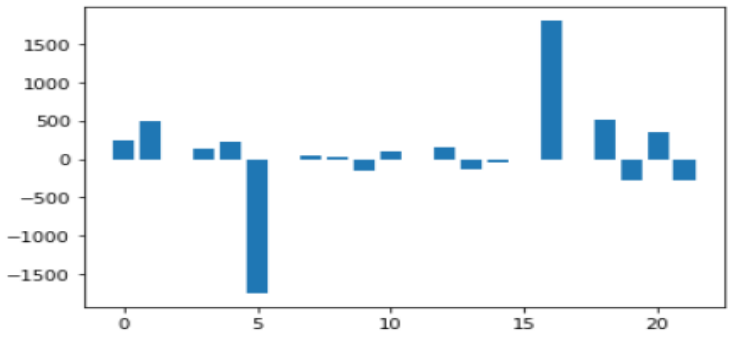


Figure 4.9 Feature Importance Graph

The graphic depicts the order of attribute significance produced for all data following regression model training. In the table below, a result was obtained by using the eli5 library, supporting the graph above. The variables that positively affect the model output are Wheat Juice, Rumen, Potatoes peel, Potatoes whole, Mixed vegie, Weak vinasse and Poultry manure while negatively affecting features are Maize silage, Digester 1 and Digestate dry. The variable y top features in figure 4.10 also represents one more positive is Beef cattle manure.

y top features

Weight[?]	Feature
+12830.700	<BIAS>
+1798.336	Wheat Juice
+504.412	Rumen
+493.315	Weak vinasse
+345.691	Chicken Fat
+239.412	Dairy cow manure
+227.038	Potatoe peel
+160.149	Digester - 2
+128.157	Poultry manure
+102.432	Potatoe whole
+43.168	Mixed vegie
+32.250	Peaches
+0.000	Corn Pomace
	… 1 more negative …
-0.000	Dair cow manure seperated
-1.021	Beef cattle manure
-16.929	Digester - 1
-44.658	Digestate dry
-135.573	Digester - 3
-148.405	DAF sludge
-275.428	Broiler Chicken
-283.978	Wine Pulp
-1745.719	Maize silage

Figure 4.10 Feature Importance Table

Contribution[?]	Feature	Value
+12830.700	<BIAS>	1.000
+4402.058	Digester - 2	27.487
+2612.222	Weak vinasse	5.295
+928.062	Dairy cow manure	3.876
-14.686	Beef cattle manure	14.377
-282.139	Digester - 3	2.081
-426.891	Digester - 1	25.216

Figure 4.11 Contribution of Features for Model

The variables that contribute to the learning of the test data as a result of linear regression are shown in the figure above. The conclusion gained from this is that the output of the model is unaffected by whether the quantity of waste is excessive or not. Although Beef cattle manure has a greater ratio than Digester-2, it has a negative effect on the model.

The factors that contribute to the model while applying the Lasso regression model to the data set are mentioned below.

y top features

Weight[?]	Feature
+12917.081	<BIAS>
+1609.841	Wheat Juice
+467.761	Weak vinasse
+232.791	Dairy cow manure
+192.929	Rumen
+192.737	Potatoe peel
+159.497	Digester - 2
+127.808	Poultry manure
+97.639	Potatoe whole
+35.434	Mixed vegie
+34.664	Peaches
-1.031	Beef cattle manure
-15.787	Digester - 1
-29.980	Digestate dry
-51.229	DAF sludge
-128.037	Digester - 3
-282.500	Broiler Chicken
-1546.628	Maize silage

Figure 4.12 Feature Importance Table with Lasso Regression

4.2.1 Regression models

Regression models included feature selection. Data was trained using 22 variables, and subsequently factors having negative effects were excluded from the model. Comparison was made between the prediction score produced with all variables and the prediction score after variable selection. Since the prediction score achieved after variable selection is higher, variable selection has been implemented in the subsequent models.

Several regression algorithms were trained in the study by separating the data into training and test sets. 650 days of biogas generation and waste entering the system are reflected in the statistics. This data was separated into training data at 80% and test data at 20%. The 520-day data therefore stayed in the training set and were trained. The remaining 130-day dataset was used to estimate it.

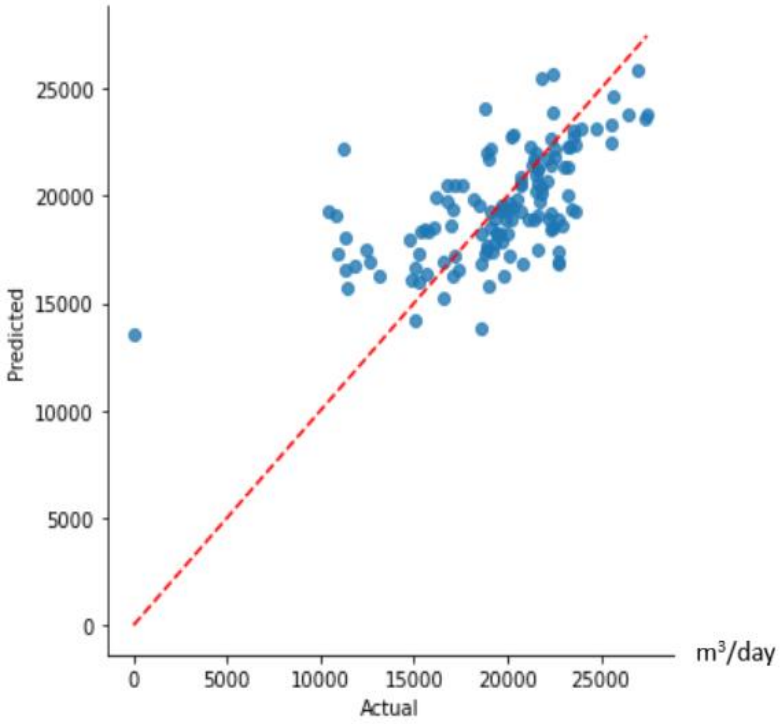


Figure 4.13 Linear Regression Plot

A graph of the estimated and actual values of the linear regression model is shown in Figure 4.13. The model proved incapable of predicting a value of 0 since there was hardly no production below 10000 in the actual data. Using a R^2 value of 48%, this model data set trained with variables was able to describe.

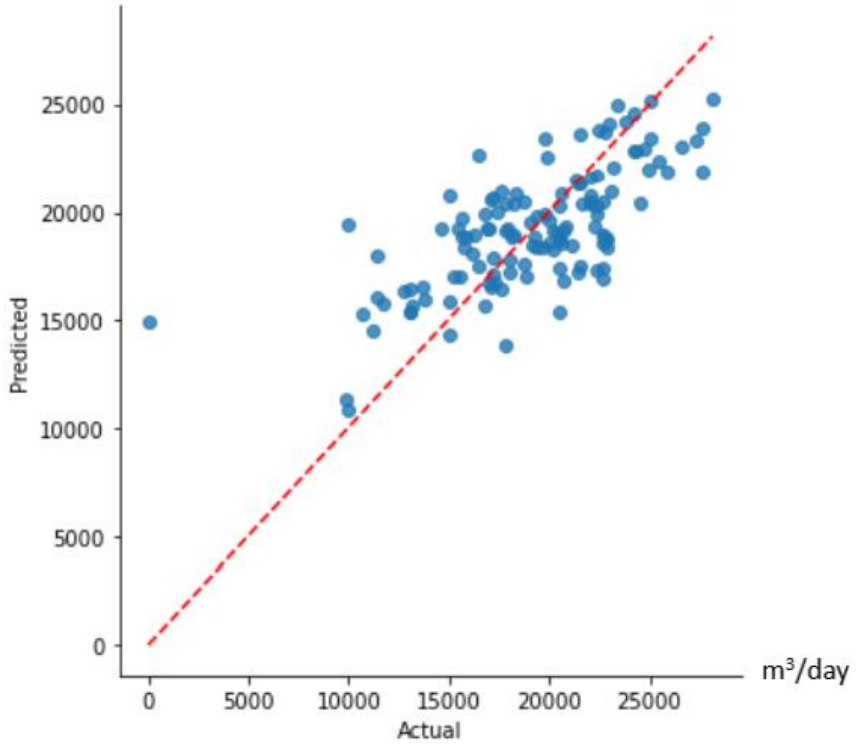


Figure 4.14 Ridge Regression Plot

Figure 4.14 represents the success of the ridge regression model graphically. In contrast to linear regression, it includes an alpha parameter. The alpha parameter is utilized for model regulation. The research suggests that ridge regression should perform better than linear regression, despite the fact that the model has the same R^2 score.

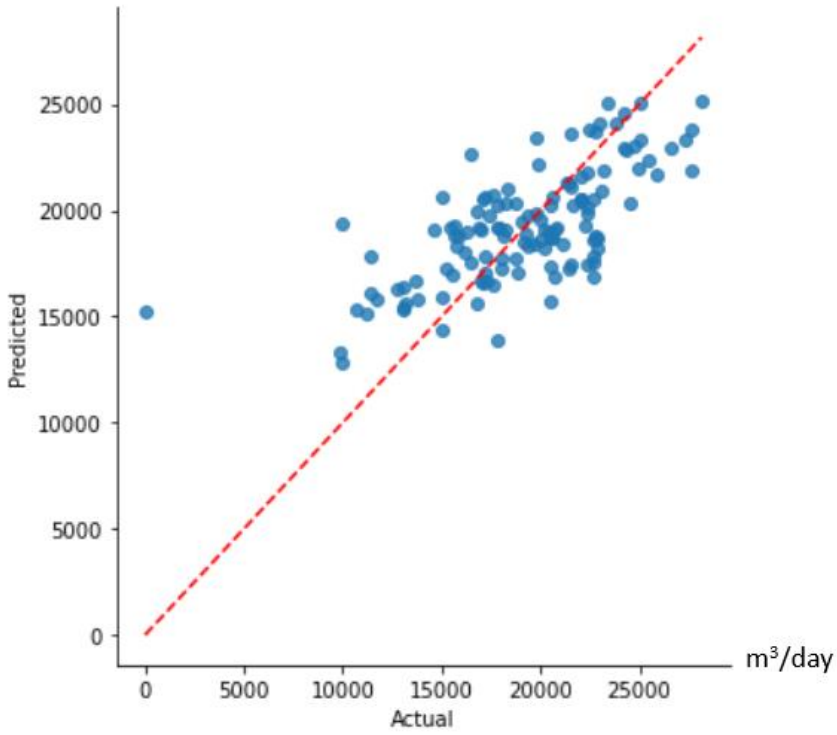


Figure 4.15 Lasso Regression Plot

Figure 4.15 depicts the Lasso Regression model's success graph. In contrast to Ridge regression, L1 regulation is implemented. Despite the fact that the model's graph resembles that of the ridge regression, a comparison of the actual and prediction values in the ridge regression graph reveals that it yields more accurate results.

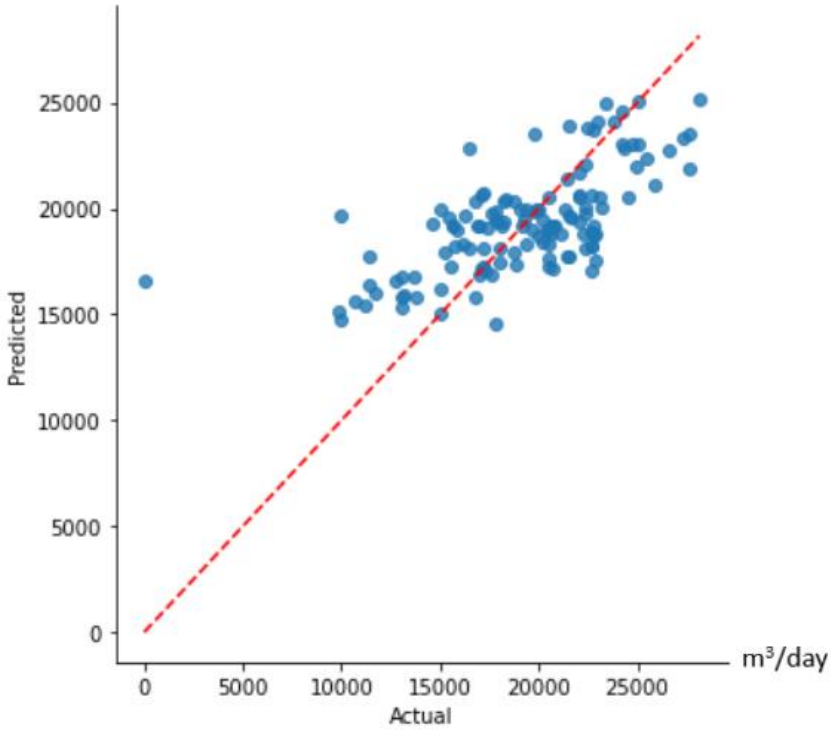


Figure 4.16 Elastic Net Regression Plot

The actual and predicted values of the Elastic Net Regression model are presented in Figure 4.16. L1 and L2 editing, unlike Ridge and Lasso regression, are applied concurrently. Parameters alpha and l1 ratio are present. Examining the regression line in the model's graph reveals points on the regression line. This indicates that the model accurately predicted those values. The R^2 score of the model is 44%.

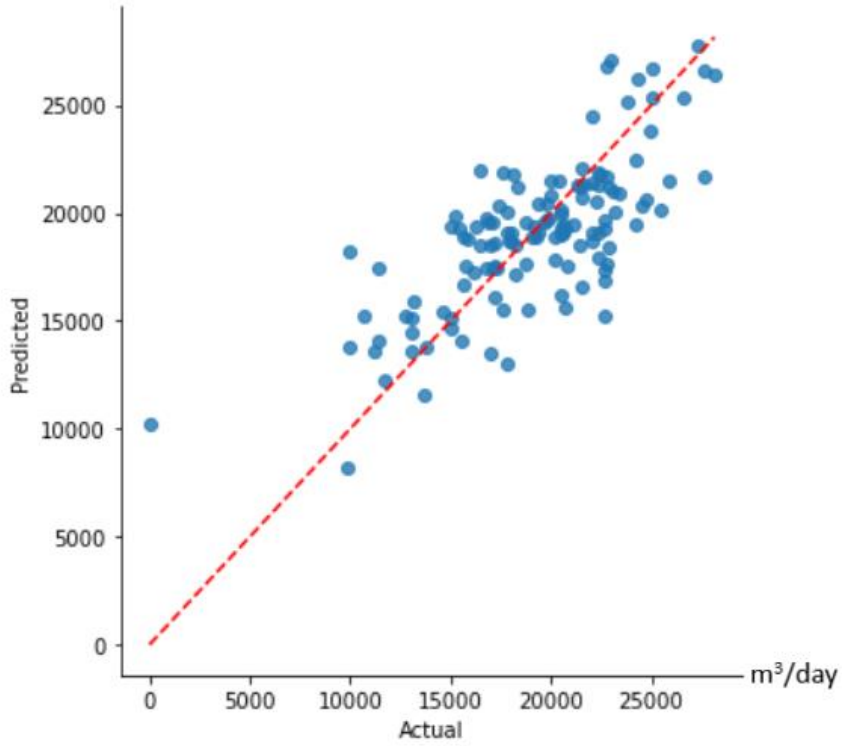


Figure 4.17 Random Forest Regression Plot

Actual and expected values of the random forest regression model are displayed in Figure 4.17. This model is more sophisticated than other models. As seen in the graph, the correlation between predicted and actual values is superior than those of competing models. 55% is the model's R2 score.

4.2.2 Neural network model

The sequential model was utilized in this study's neural network model. There is one input layer, two hidden layers, and one output layer in the model. Activation functions were evaluated and the optimal function of ReLU was determined. The output activation function is linear as a result of linear modeling. The Adam and RMSProp functions as optimization functions are evaluated. RMSProp estimated the optimal outcome. It was determined, based on the number of iteration trials, that the model will complete 500 iterations. The neural network model is initially trained with the whole dataset of 22 features. In the 22-variable model, the R^2 score is 52%. The figure below depicts the number of neurons utilized by the model in each layer.

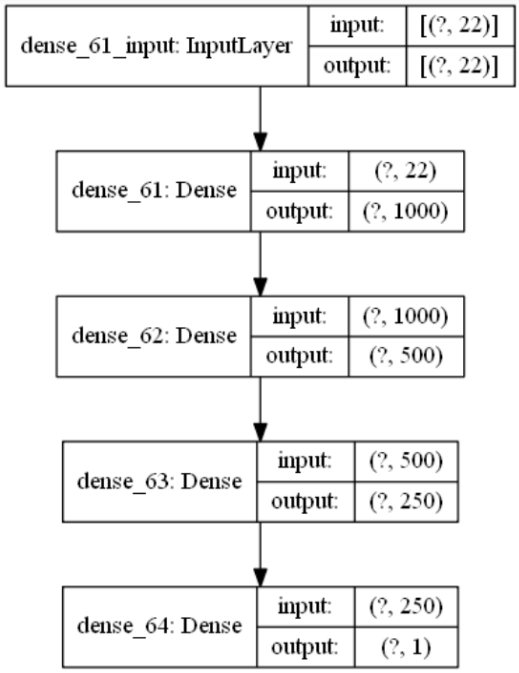


Figure 4.18 Neural Network model with 22 features

Delete the variables that have a negative impact on the model and retrain it using the same procedures as described previously. When we excluded features with negative effects from the model, R^2 decreased to 42%.

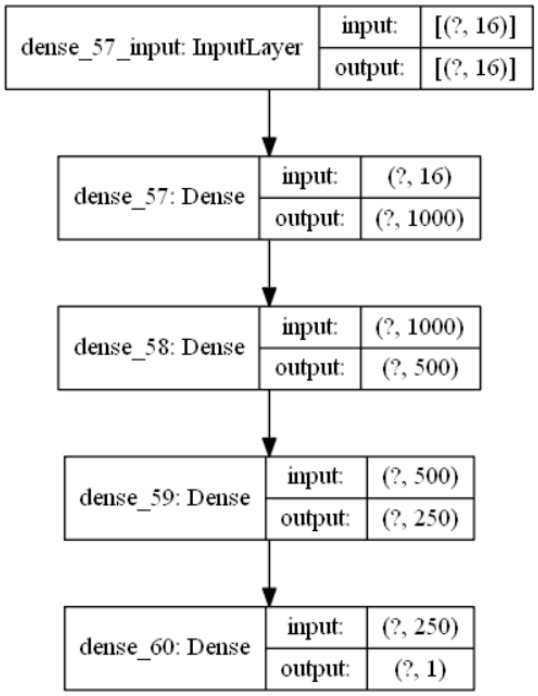


Figure 4.19 Neural Network model

The training and validation graphics of our artificial neural network model are as follows.

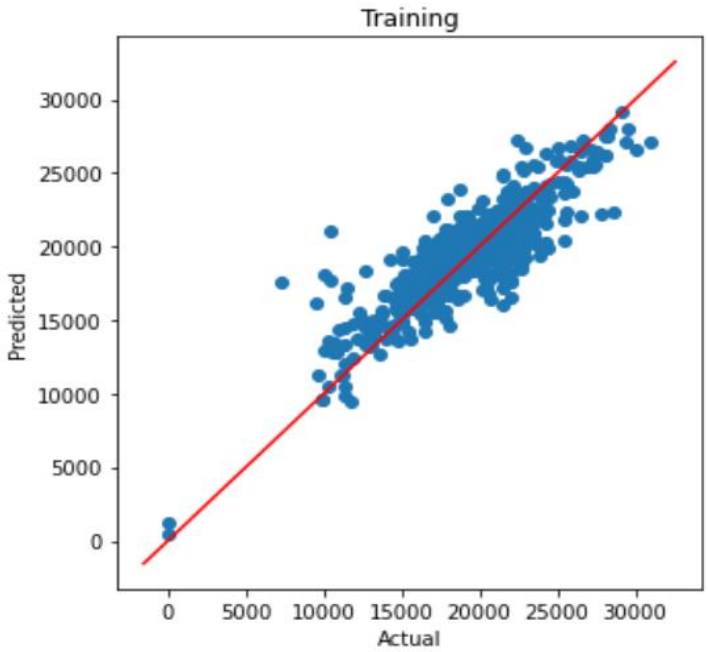


Figure 4.20 Graph of Neural Network Model Training Predicted-Actual Value

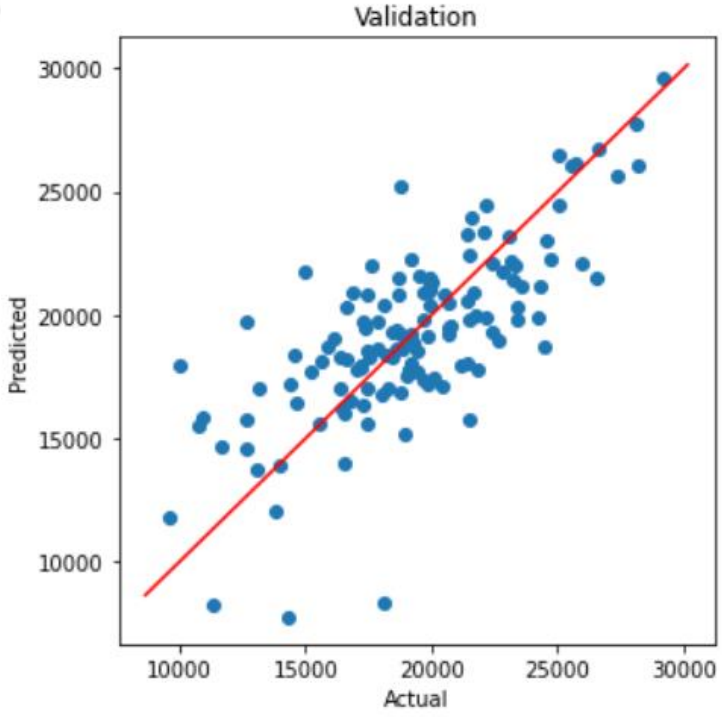


Figure 4.21 Neural Network Validation Curve

The actual and expected values of the neural network model are displayed in figures 4.20 and 4.21. In 4.20, the R^2 score for the training set is 73%. This indicates that the model accurately predicted the training data in the dataset, i.e. the parameters it has previously seen. Figure 4.21 depicts the validation data for the R^2 score, which reveals a 52% success rate. This demonstrates that the model cannot predict the test set, which it has not seen, as well as it can the training set. Examining both results demonstrates that variable reduction in neural network architecture has a negative impact on the model, but feature selection is essential in regression models.

5. CONCLUSION

As a necessary for human life, energy is a critically precious item in the universe. Although biogas systems are a viable choice for energy generation, they are complex and expensive to establish, hence there are very few production plants worldwide.
Several research have been conducted on biogas plants and energy production from biogas. In this study, machine learning and neural networks were used to construct a production forecasting model.

The primary objective of the study was to investigate the effect of system waste in biogas generation. With 650 days of waste and biogas production data from the Pales biogas facility, a model was developed. Each of the three digesters at the biogas plant was subjected to a unique set of tests. In the digester, the best variables contributing to the model were chosen. Following three digests, all data were evaluated. Initially, among 22 wastes, features that would contribute to the model were identified using feature selection techniques. After feature selection, the model was trained with the selected features. In the study, various regression models and a neural network model were evaluated.

As a result of the model, it was determined that the average success rate for regression models was 49%. The neural network model attained an average success rate of 48 percent. Observations revealed that while the significance of selecting features in regression models increased, variable reduction in neural networks had a negative effect on the model.

The data set can be expanded to boost the model's precision. The more historical data the model analyzes, the greater its forecasting ability. The existence of too many missing values in the data set from the biogas plant severely impacts the performance of the model. In the study, it is found that different regression models yield varying

degrees of precision when different parameters are employed. By optimizing the hyper parameters, research may be conducted to produce the most accurate model output using various hyper parameters. In neural networks, on the other hand, the optimal outcome may be sought by experimenting with different layers and varying the number of epochs.

The processor utilized in the study is Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz 2.11 GHz. When we expand the data set, the training time will increase proportional to the performance of the computer. In this procedure, it is projected that with a more powerful hardware, the model will exhibit stronger performance and produce predictions at faster rates.

BIBLIOGRAPHY

[1] G. W. Norton, “Economic and environmental impacts of IPM,” vol. 8, no. 3, pp. 271–277, 1994.

[2] “Türkiye’nin Geleceği İçin Yenilenebilir Enerji Kaynaklarının Önemi | Ekolojist.net.” https://ekolojist.net/turkiyenin-gelecegi-icin-yenilenebilir-enerji-kaynaklarinin-onemi/ (accessed Nov. 27, 2022).

[3] “Biogas - Wikipedia.” https://en.wikipedia.org/wiki/Biogas (accessed Nov. 27, 2022).

[4] “Anaerobic digestion - Wikipedia.” https://en.wikipedia.org/wiki/Anaerobic_digestion (accessed Nov. 27, 2022).

[5] I. Dekhtiar, T. Dyvak, and Y. Martsenyuk, “Features of biogas production process and methods of its modeling,” *2013 12th Int. Conf. Exp. Des. Appl. CAD Syst. Microelectron. CADSM 2013*, pp. 66–68, 2013.

[6] “Turkey: bioenergy capacity 2021 | Statista.” https://www.statista.com/statistics/878824/total-bioenergy-capacity-in-turkey/ (accessed Nov. 27, 2022).

[7] A. Cheon, J. Sung, H. Jun, H. Jang, M. Kim, and J. Park, “Application of Various Machine Learning Models for Process Stability of Bio-Electrochemical Anaerobic Digestion,” *Processes*, vol. 10, no. 1, pp. 1–14, 2022, doi: 10.3390/pr10010158.

[8] Y. Yang, S. Zheng, Z. Ai, and M. M. M. Jafari, “On the Prediction of Biogas Production from Vegetables, Fruits, and Food Wastes by ANFIS- And LSSVM-Based Models,” *Biomed Res. Int.*, vol. 2021, 2021, doi: 10.1155/2021/9202127.

[9] J. Gonçalves Neto, L. Vidal Ozorio, T. C. Campos de Abreu, B. Ferreira dos Santos, and F. Pradelle, “Modeling of biogas production from food, fruits and vegetables wastes using artificial neural network (ANN),” *Fuel*, vol. 285, no. August 2020, p. 119081, 2021, doi: 10.1016/j.fuel.2020.119081.

[10] R. Ravindra Pansari, S. Patil, and M. Khan, “Development of IoT based Monitoring of Biogas Plant PG Student [Embedded System], 2 HOD, E and TC Engg [Guide], 3 Professor [Co-Guide],” *Int. J. Sci. Eng. Res.*, vol. 11, no. 8, pp. 1326–1333, 2020, [Online]. Available: http://www.ijser.org.

[11] Y. Wang, T. Huntington, and C. D. Scown, “Tree-Based Automated Machine

Learning to Predict Biogas Production for Anaerobic Co-digestion of Organic Waste," *ACS Sustain. Chem. Eng.*, vol. 9, no. 38, pp. 12990–13000, 2021, doi: 10.1021/acssuschemeng.1c04612.

[12] K. Jeong, A. Abbas, J. Shin, M. Son, Y. M. Kim, and K. H. Cho, "Prediction of biogas production in anaerobic co-digestion of organic wastes using deep learning models," *Water Res.*, vol. 205, no. June, p. 117697, 2021, doi: 10.1016/j.watres.2021.117697.

[13] D. De Clercq, D. Jalota, R. Shang, K. Ni, and Z. Zhang, "US," 2019.

[14] C. Mateescu, E. Tudor, A. D. Dima, I. Chirita, V. Tanasiev, and T. Prisecaru, "Artificial Intelligence Approach In Predicting Biomass-to-Biofuels Conversion Performances," *2022 23rd Int. Carpathian Control Conf. ICCC 2022*, pp. 370–375, 2022, doi: 10.1109/ICCC54292.2022.9805871.

[15] M. Pawlita-Posmyk, M. Wzorek, and R. Gono, "Biogas and Biomethane from Animal Waste for Electricity Production," *Proc. 2022 22nd Int. Sci. Conf. Electr. Power Eng. EPE 2022*, 2022, doi: 10.1109/EPE54603.2022.9814162.

[16] S. A. Supti, N. Zaman, S. Islam, A. K. Podder, and A. S. M. Ibrahim, "Selection of Convenient Organic Matters for Bioenergy Generation: A Comparative Approach," in *2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, Feb. 2022, pp. 1–6, doi: 10.1109/ICAEEE54957.2022.9836387.

[17] B. E. Kas, E. Engineering, E. Sciences, and E. T. Bo, "EFFECT OF TRACE ELEMENTS ON BIOGAS PRODUCTION : EFFECT OF TRACE ELEMENTS ON BIOGAS PRODUCTION :," 2019.

[18] "ETKİSİNİN İNCELENMESİ Ahmet CANAN ENERJİ SİSTEMLERİ MÜHENDİSLİĞİ Tez Danışmanı Prof . Dr . Mehmet ÖZKAYMAK," 2021.

[19] S. S. Bhajani, "Review: Factors Affecting Biogas Production," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 2, pp. 79–88, 2022, doi: 10.22214/ijraset.2022.40192.

[20] R. Sindhu, P. Binod, A. Pandey, S. Ankaram, Y. Duan, and M. K. Awasthi, *Biofuel production from biomass: Toward sustainable development*. Elsevier B.V., 2019.

[21] A. Gashaw, "Anaerobic Co-Digestion of Biodegradable Municipal Solid Waste with Human Excreta for Biogas Production: A Review," *Am. J. Appl. Chem.*, vol. 2, no. 4, p. 55, 2014, doi: 10.11648/j.ajac.20140204.12.

[22] "Dimensionality reduction - Wikipedia." https://en.wikipedia.org/wiki/Dimensionality_reduction (accessed Nov. 27, 2022).

[23] "Feature Selection Methods and How to Choose Them - neptune.ai."

https://neptune.ai/blog/feature-selection-methods (accessed Nov. 27, 2022).

[24] A. Kewat, P. N. Srivastava, and D. Kumhar, “Performance Evaluation of Wrapper-Based Feature Selection Techniques for Medical Datasets,” no. March, pp. 619–633, 2020, doi: 10.1007/978-981-15-0222-4_60.

[25] “Simplilearn | Online Courses - Bootcamp & Certification Platform.” https://www.simplilearn.com/ (accessed Nov. 27, 2022).

[26] “Artificial intelligence - Wikipedia.” https://en.wikipedia.org/wiki/Artificial_intelligence (accessed Nov. 27, 2022).

[27] “Machine learning - Wikipedia.” https://en.wikipedia.org/wiki/Machine_learning (accessed Nov. 27, 2022).

[28] D. J. Joshi, I. Kale, S. Gandewar, O. Korate, D. Patwari, and S. Patil, “Reinforcement Learning: A Survey,” *Adv. Intell. Syst. Comput.*, vol. 1311 AISC, pp. 297–308, 2021, doi: 10.1007/978-981-33-4859-2_29.

[29] “Doğrusal regresyon - Vikipedi.” https://en.wikipedia.org/wiki/Linear_regression (accessed Dec. 06, 2022).

[30] “Bias–variance tradeoff - Wikipedia.” https://en.wikipedia.org/wiki/Bias–variance_tradeoff (accessed Dec. 06, 2022).

[31] “Ridge Regression for Better Usage | by Qshick | Towards Data Science.” https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db (accessed Dec. 08, 2022).

[32] “Ridge regression - Wikipedia.” https://en.wikipedia.org/wiki/Ridge_regression (accessed Dec. 06, 2022).

[33] “Building a ridge regressor | Python Machine Learning Cookbook - Second Edition.” https://subscription.packtpub.com/book/data/9781789808452/1/ch01lvl1sec14/building-a-ridge-regressor (accessed Dec. 06, 2022).

[34] “Elastic net regularization - Wikipedia.” https://en.wikipedia.org/wiki/Elastic_net_regularization (accessed Dec. 06, 2022).

[35] “Random forest - Wikipedia.” https://en.wikipedia.org/wiki/Random_forest (accessed Dec. 08, 2022).

[36] “The Neural Networks Model - IBM Documentation.” https://www.ibm.com/docs/en/spss-modeler/18.0.0?topic=networks-neural-model (accessed Nov. 30, 2022).

[37] “A Beginner’s Guide to Neural Networks and Deep Learning | Pathmind.” https://wiki.pathmind.com/neural-network (accessed Dec. 03, 2022).

[38] “What are Neural Networks? | IBM.” https://www.ibm.com/cloud/learn/neural-networks (accessed Dec. 03, 2022).

CURRICULUM VITAE

Personal Information

Name Surname : Şevval Ayşe YURTEKİN

Education

Undergraduate Education : Bachelor in Electrical and Electronics Engineering, Kadir Has University, Istanbul, Turkey, 2019

Graduate Education : Master of Electronics Engineering, Kadir Has University, Turkey

Foreign Language Skiils : Fluent in English

Work Experience

Name of Employer and Dates of Employment:
2019- Electronics Engineer at IoT Vision, Turkey
2021- Data Analytics and Machine Learning Asistant Specialist at Türkiye Sigorta, Turkey