ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

COAL GAS CONTENT PREDICTION ON KINIK COALFIELD, SOMA BASIN WITH MACHİNE LEARNING METHODS

M.Sc. THESIS

Satuk Bugra AKDAS

Department of Mining Engineering

Mining Engineering Programme

JUNE 2024

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

COAL GAS CONTENT PREDICTION ON KINIK COALFIELD, SOMA BASIN WITH MACHINE LEARNING METHODS

M.Sc. THESIS

Satuk Bugra AKDAS
(505211010)

Department of Mining Engineering

Mining Engineering Programme

Thesis Advisor: Prof. Dr. Abdullah FİŞNE

JUNE 2024

ISTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

SOMA HAVZASI KINIK KÖMÜR YATAĞI'NDA MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE KÖMÜR GAZI İÇERİĞİ TAHMİNİ

YÜKSEK LİSANS TEZİ

Satuk Buğra AKDAŞ
(505211010)

Maden Mühendisliği Anabilim Dalı

Maden Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Abdullah FİŞNE

HAZİRAN 2024

Satuk Bugra AKDAS, a M.Sc. student of İTU Graduate School student ID 505211010 successfully defended the thesis/dissertation entitled "COAL GAS CONTENT PREDICTION ON KINIK COALFIELD, SOMA BASIN WITH MACHINE LEARNING METHODS", which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor : Prof. Dr. Abdullah FİŞNE .....
İstanbul Technical University

Jury Members : Prof. Dr. Türker Hüdaverdi .....
Istanbul Technical University

Prof. Dr. Ümit Özer .....
Istanbul University-Cerrahpasa

Date of Submission : 06 May 2024
Date of Defense : 12 June 2024

To my beloved mom,

FOREWORD

Firstly, I want to express my gratitude to my thesis advisor, Professor Dr. Abdullah FİŞNE, for his unwavering support throughout my years as a graduate student, project assistant, and engineer. I am truly thankful for his trust in me and his confidence in selecting the distinctive machine learning topic for my thesis.

The second significant appreciation goes to my parents, who nurtured and supported me throughout my education, preparing me for life's challenges since childhood. They are undeniably crucial contributors to this thesis, and I thank them wholeheartedly for everything.

Acknowledgment is also due to the university and the Department of Mining Engineering at Istanbul Technical University, which play a pivotal role in my current engineering endeavors. I extend my personal thanks to all the professors in the department.

I am particularly grateful to Professor Dr. Yasin ARSLANOĞLU for allowing me access to his research on "machine learning" and for assisting me in addressing my inquiries.

Special thanks go to Samet Can ÖZER, Olgun ESEN, and İrem Gizem MENEVŞE for their constant support and for always being there for me.

This research is dedicated to whom lost in Soma Accident in 2014, and all the mining workers.

June 2024

Satuk Bugra AKDAS
(Petroleum Engineer)

TABLE OF CONTENTS

Page

ABBREVIATIONS

USBM : United States Bureau of Mines
EIA : Energy Information Administration
UN : United Nations
U.S. : United States
CBM : Coalbed Methane
AI : Artificial Intelligence
ML : Machine Learning
SVM : Support Vector Machine
LR : Linear Regression
MLR : Multiple Linear Regression
RBF : Radial Basis Function
LS : Least Square
RBF : Radial Basis Function
lbfgs : Limited Memory Broyden-Fletcher -Goldfarb-Shanno Algorithm
MSE : Mean Squared Error
MAE : Mean Absolute Error
MAEP : Mean Absolute Error Percentage
TDG : Total Desorbed Gas

SYMBOLS

CO	**:** Carbon monoxide
CO_2	**:** Carbon dioxide
N_2	**:** Nitrogen
CH_4	**:** Methane
C_2H_6	**:** Ethane
Q1	**:** Lost Gas
Q2	**:** Measured Gas
Q3	**:** Residual Gas
m^3	**:** Cubic meter
t	**:** Ton
i, n	**:** Index Values
x	**:** Input Vector
y	**:** Output Vector
$\boldsymbol{w^T}$	**:** Scalar Product of Weight
b	**:** Bias
ϕ	: Kernel Function
r	: Radius of Kernel Function
u	: Mean of the Input Data
s	: Standard Deviation of the Input Data
z	: Standardized Data
C	: Regularization Parameter
ϵ_{abs}	**: Absolute error**
$\hat{y}$	: Predicted Desorbed Gas Values
y	: Original Desorbed Gas Values
%	: Percentage
R^2	: Regression Score
w_i	: Linear Regression Intercept
w_0	: Linear Regression Weight
a	: Learning Rate

LIST OF TABLES

Page

LIST OF FIGURES

Page

COAL GAS CONTENT PREDICTION ON KINIK COALFIELD, SOMA BASIN WITH MACHINE LEARNING METHODS

SUMMARY

Coal has been used by humanity since ancient times, becoming widespread with steam engines, and is now a complex energy source that is beginning to be replaced by alternatives. While coal is used as a direct source of energy, it also serves as a source rock that produces various fluids, primarily carbon-based gases. Its multiple functions and the presence of various quality and quantity reserves in many countries fundamentally extend the lifetime of coal. Moreover, the presence of methane and other natural gas components in coal makes it capable of contributing to natural gas reserves as an effective alternative when coal is phased out. This study introduces a novel data-driven methodology for interpreting the nonlinear challenge of analyzing the total desorbed gas content in coal seams. The investigation is centered on a low-rank coal reserve situated in the Kınık coalfield, where the United States Bureau of Mines (USBM) direct desorption method was employed to project the total desorbed gas content for underground mining operations. Utilizing core samples obtained during the reserve and gas content analysis, machine learning models were developed. These models were trained with coal properties data, including depth, moisture, ash, volatile matter, and calorific value, in correlation with the total desorbed gas content. Various machine learning algorithms, namely multiple linear regression, support vector machine, and artificial neural network, were utilized to predict the total desorbed gas content in the Kınık coalfield. Hyperparameter tuning was applied to optimize the machine learning models, and the most effective model was chosen based on its regression accuracy and computational efficiency. The raw data analysis, facilitated by pairplot, revealed associations between parameters and their direct influence on the total gas content in coal seams. Sensitivity analysis was performed to assess the impact of coal properties on total desorbed gas content. The selected model was then applied to predict the total desorbed gas content at a specific location in the coalfield. The study's outcomes offer valuable insights and recommendations for the analysis of unconventional reservoirs and the prediction of petrophysical systems using machine learning techniques. In essence, this research underscores the potential of machine learning in tackling nonlinear challenges within the geological domain and proposes a promising avenue for future investigations in this field.

SOMA HAVZASI KINIK KÖMÜR YATAĞI'NDA MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE KÖMÜR GAZI İÇERİĞİ TAHMİNİ

ÖZET

Kömür eski çağlardan bu yana insanlık tarafından kullanılmış, buharlı makineler ile yaygın hale gelmiş, günümüzde ise yerine alternatiflerine bırakmaya başlayan kompleks bir enerji kaynağıdır. Kömür doğrudan enerji sağlayan bir kaynak olarak kullanılırken, aynı zamanda bünyesinde bulundurduğu başta karbon kökenli gazlar olmak üzere, bir çok akışkanın üretildiği kaynak kayaç olma görevine sahiptir. Birden fazla fonksiyonu, temelde yaygın olarak bir çok ülkede çeşitli kalite ve miktarlardaki rezervleri olması kömürün son kullanım ömrünü sürekli ertelemektedir. Ayrıca içinde bulundurduğu başta metan ve diğer doğalgaz bileşeni gazların varlığı kömürün doğrudan kullanımdan çıkarken alternatifi olan doğalgaz rezervlerine katkı sağlama noktasında etkin rol alabilecek durumdadır. Bu çalışma, kömür damarlarından toplam desorbe edilmiş gaz içeriğini analiz etmenin doğrusal olmayan zorluğunu yorumlamak için yeni, veriye dayalı bir metodoloji sunmaktadır. Araştırma, Kınık kömür yatağında yer alan düşük dereceli bir kömür rezervine odaklanıyor; burada Amerika Birleşik Devletleri Maden Bürosu (USBM) doğrudan desorpsiyon yöntemi, yer altı madencilik operasyonları için toplam desorbe edilmiş gaz içeriğini tahmin etmek için kullanıldı. Rezerv ve gaz içeriği analizi sırasında elde edilen karot örneklerinden yararlanılarak makine öğrenimi modelleri geliştirildi. Bu modeller, toplam desorbe edilen gaz içeriğiyle bağlantılı olarak derinlik, nem, kül, uçucu madde ve kalorifik değer dahil olmak üzere kömür özellikleri verileriyle eğitildi. Kınık kömür sahasında toplam desorbe edilen gaz içeriğini tahmin etmek için çoklu doğrusal regresyon, destek vektör makinesi ve yapay sinir ağı gibi çeşitli makine öğrenme algoritmaları kullanılmıştır. Makine öğrenimi modellerini optimize etmek için hiperparametre ayarı uygulandı ve regresyon doğruluğu ve hesaplama verimliliği temel alınarak en etkili model seçildi. Çift grafiğiyle kolaylaştırılan ham veri analizi, parametreler arasındaki ilişkileri ve bunların kömür damarlarındaki toplam gaz içeriği üzerindeki doğrudan etkisini ortaya çıkardı. Kömür özelliklerinin toplam desorbe edilmiş gaz içeriği üzerindeki etkisini değerlendirmek için duyarlılık analizi yapıldı. Seçilen model daha sonra kömür sahasında belirli bir konumda toplam desorbe edilmiş gaz içeriğini tahmin etmek için uygulandı. Çalışmanın sonuçları, geleneksel olmayan rezervuarların analizi ve makine öğrenme tekniklerini kullanarak petrofiziksel sistemlerin tahmini için değerli bilgiler ve öneriler sunmaktadır. Temelde bu araştırma, makine öğreniminin jeolojik alandaki doğrusal olmayan zorlukların üstesinden gelme potansiyelinin altını çiziyor ve bu alanda gelecekteki araştırmalar için umut verici bir yol öneriyor.

1. INTRODUCTION

Coal stands as one of the earliest commercial energy sources, playing a pivotal role in the progression of human civilization for various purposes. Its full integration into the industrial age began in the 17th century when it became a vital component in energy generation through steam turbines. From that point onward, coal consumption has steadily increased, marking a notable upward trend that persisted until the last decade (EIA, Monthly Energy Review:Coal, 2021). Climate change, the policies of the world's largest economies, and international agreements have collectively influenced the energy market to prioritize lower carbon emissions, a commitment established in the Kyoto Protocol (UN, 2009). Following the United Nations Copenhagen meeting in 2009, many major countries transitioned from coal-based power plants to alternative sources as part of efforts to reduce greenhouse gas emissions (German Institute for Economic Research et al.,2019). However, the proven coal reserves worldwide are documented to be approximately 1.16 trillion short tons (EIA, Coal explained, 2021). Considering the U.S. coal production in 2020, which amounted to approximately 0.535 billion short tons, it is estimated that the recoverable coal reserves would last for about 470 years. However, if we specifically focus on the recoverable reserves at currently active mines, they would only last for around 25 years (EIA, Coal explained, 2021). Similar examples can be extended globally to major coal producers, including China (Zhou et al., 2016) and Australia (Tang et al., 2018). It is evident that the significance of remaining coal reserves should not be underestimated. However, coal is considered a prime candidate for reduction among available energy alternatives to mitigate greenhouse gas emissions. This dilemma creates a conflict—whether to utilize the coal reserves to meet growing energy demands, resulting in additional greenhouse gas emissions, or to leave them untapped, potentially leaving the market vulnerable to energy shortages as demands continue to rise. Consequently, an alternative approach is imperative for an effective energy transition.

The current strategy involves promoting investments in geothermal, wind, and solar power plants. Since 2009, there has been a notable increase in the number of such

plants. This global energy transition is gaining momentum, even in countries traditionally reliant on oil and gas exports (e.g., Sweihan Photovoltaic Independent Power Project, Abu Dhabi, 2017; Dumat Al Jandal Wind Farm, Al Jouf, Saudi Arabia, 2020 etc.).

In late 2019, the global COVID-19 pandemic spread worldwide, prompting heightened precautions and bringing everyday mobility to a standstill, except for essential transportation. This unprecedented situation led to a significant surge in electricity consumption on a global scale, fundamentally altering energy consumption patterns. Consequently, a substantial gap emerged between the vision of achieving zero carbon emissions and the heightened demand for energy.

This shift in dynamics, coupled with an escalation in energy commodity prices, resulted in challenges for countries reliant on energy imports, creating a struggle to balance economic needs with environmental goals.

Turkey, as a country heavily reliant on energy imports, is significantly influenced by global changes in the energy market. However, its unique geopolitical position, characterized by key energy transportation lines, maritime routes, and abundant coal resources, plays a crucial role in shaping its energy supply perspective. Despite being part of the Paris Agreement, with a commitment to reducing and eventually achieving net-zero carbon footprints in collaboration with European partners, Turkey still depends on carbon-based sources. The nation, however, aspires to increasingly incorporate renewable energy within its geographic boundaries.

Turkey possesses substantial coal reserves, offering a range of alternatives to align with global objectives. In-situ combustion and methane extraction from underground coal mines are highlighted as promising solutions for local energy supply. Among these alternatives, the extraction of methane from underground coal reserves stands out as a particularly viable option, providing a pathway for future energy policies and contributing to Turkey's energy transition objectives.

1.1 What is Coalbed Methane

The concept of Coalbed Methane (CBM) revolves around the entrapment of methane within either the cleats or pore system of coal seams. This phenomenon is categorized under unconventional resources, as discussed by Beaton et al. in 2006. While coal

primarily serves as an organic raw material for fuel, coal seams are also considered reservoir rocks with substantial gaseous contents.

In coal samples, various gas compounds are present, including CO, CO_2, N_2, CH_4, C_2H_6, and others. Notably, methane (CH_4) is found in significant amounts. Consequently, the concentration of methane can be regarded as a natural gas source within the coal formation. This entire concept is encapsulated in the Coalbed Methane system, as elucidated by Gao, Mastalerz, and Schimmelmann in 2020.

1.1.1 How the coalbed methane system works

Coalbed Methane (CBM) is a byproduct of the coal formation process, arising from the organic and physical transformation of plant-based fossils. During the coal formation stage, various types of gases are produced. Most of the gas content generated during coal formation migrates towards lower pressure zones or escapes. Since coal formation is a time-consuming process, the reservoir formation of gas follows the maturation of coal. This leads to continuous cycles of coal production and burial.

The gas content in coal seams becomes trapped within the coal, and eventually, underground coal formations are discovered and exploited. In the 1980s, coal mines faced significant challenges due to hazardous gas problems, including both poisonous and flammable gases. This situation gave rise to an alternative market aiming to eliminate these gas issues. Initially, efforts focused on gasifying coal seams, but over time, the emphasis shifted to the production of gas content through CBM operations.

Drilling operations from the Earth's surface were employed, investigating suitable coal beds with production wells. The extraction from initially drilled coal formations commenced in natural fractures, cleats, and connected pores, leveraging natural permeabilities. Furthermore, the Oil and Gas Industry introduced techniques such as hydraulic fracturing, in-situ gasification, and injections of CO_2 and steam into the CBM system to enhance production.

1.1.2 What is machine learning

In the 20th century, the global population experienced rapid growth, leading to subsequent changes driven by the emerging needs. The instinct for survival, advancements in supply chain management, and efforts to overcome geographical distances resulted in more complex challenges. As human needs surpassed individual

capacity, alternatives were sought, leading to the development of machines to address these intricate demands.

Among various machines, computers emerged as a transformative force in daily life, experiencing rapid development. The computational power of computers facilitated the application of classical mathematical approaches at unprecedented speeds. This increased efficiency in calculations played a crucial role in integrating the science of statistics into every facet of engineering.

Driven by evolving needs, statistical methods themselves underwent various transformations. The application of specific statistical methods paved the way for the growth of Artificial Intelligence (AI). Artificial Intelligence, in turn, embraced these statistical methods, giving rise to the field of Machine Learning. Machine Learning utilizes statistical techniques to construct diverse tools, contributing to the advancement of AI technology.

Machine Learning falls under the broader concept of Artificial Intelligence within the field of computer science. To grasp Machine Learning, it's essential to first understand Artificial Intelligence. Artificial Intelligence, in simple terms, is founded on human intelligence, encompassing rationality, the ability to make sound decisions, thought processes, reasoning, and behavior, among other facets. Intelligence can be expressed through one or a combination of these attributes. In the pursuit of larger processing capacities, quicker responses, and sustained performance, human intelligence has been replicated in computers or machines. This replication is known as Artificial Intelligence (AI), as described by Russell and Norvig, 2020.

Within the AI framework, when a computer engages in a learning process to observe data, constructs models based on that data, and then utilizes the designed models to generate meaningful outcomes or as part of software to solve problems, it is termed as Machine Learning (ML). Both AI and ML are dynamic concepts that evolve over time. Therefore, the boundaries and definitions must be updated to keep pace with advancements in the field (Russell and Norvig, 2020).

In this study, Machine Learning methods and related concepts are employed to enhance the efficiency of methane desorption on coal samples, serving as a valuable tool in this context.

1.2 Coalbed Methane Desorption Methods

Gas desorption involves extracting gaseous substances from solid or liquid domains. In the specific context of your study, methane (CH_4) desorption is being investigated in the Kınık field, Soma region. The primary purpose of methane desorption in this case is to remove explosive gas content from the exploitation section of underground mines, thereby minimizing health and safety risks and reducing the need for extensive ventilation efforts. If the goals of risk management and cost efficiency are achieved, a third objective is to capitalize on the methane gas removed from coal seams. Depending on the amount of extracted methane, what was once a problematic gas source focused on explosivity and ventilation can potentially transform into valuable natural gas reserves.

Various methods for measuring desorption exist in the literature, and in this study, both Direct and Indirect methods are considered, as field data have been collected using these approaches. However, the primary focus of this research is to assess the agreement of Machine Learning (ML) methods with experimental results, aiming to understand how well ML techniques align with the observed data in the context of methane desorption.

1.2.1 Direct method

In accordance with the U.S. Bureau of Mines standard, the direct desorption method involves sealing solid or crushed coal samples within a leakproof container made of steel or plastic. This sealed container is then filled with a fluid, typically water or inert gases. The amount of desorbed gas is determined by observing the fluid motion from the container to the measurement apparatus. This process, known as the direct desorption method, provides insights into the gas diffusion from the core sample (Diamond and Levine, 1981).

1.2.1.1 Core desorption measurement-U.S. Bureau of Mines

To measure the amount of gas desorbed from coal samples, the process typically involves drilling operations as the initial step. Subsequently, core samples are extracted from the formation, and measurements are initiated. As the core sample moves from the bottom to the atmosphere, gas content is released due to changes in pressure and temperature. The gas exposed at this stage is defined as lost gas (Q1).

This step provides valuable information about the desorbed gas content during the extraction and handling of core samples (Szlązak et al., 2021). Following the extraction of core samples, the core holder, now at atmospheric conditions, is promptly transferred into a sealed container. This container is then filled with inert fluids, sealed, and the time elapsed from the bottom to the closure of the leakproof container is calculated to determine the amount of lost gas. In the subsequent phase, the desorption process occurs, either through a slow or fast desorption method based on the specified preference or experimental design. This step allows for the controlled release and measurement of gases desorbed from the coal samples.

1.2.1.2 Core desorption measurements-other

After the core sample is brought to the surface and sealed, the researcher must decide on the measurement time. If the goal is to obtain rapid but less accurate results, the gas desorption from the canister can be measured within a few days. This method is referred to as fast desorption. On the other hand, if the experiment continues until the measured gas content diminishes to an untraceable level and stops diffusing, it is categorized as slow desorption. The gas obtained throughout this process is evaluated under measured gas content (Q2). The choice between fast and slow desorption depends on the specific objectives of the research and the desired level of precision in the results (Zhang, et al., 2022).

1.2.1.3 Drilling cuttings desorption measurements

Following the completion of the measured gas desorption stage, coal samples that were investigated from the field to the laboratory are removed from the leakproof canister. Subsequently, these coal samples, previously saturated with an inert fluid, are prepared to be set dry. In the final step, different sections of the coal samples are crushed. During the process of pulverization, the last gas molecules are released, and these become trapped in the secondary pore volumes of the coal samples. These residual gases, released during pulverization, are referred to as residual gas (Q3). The analysis of residual gas provides insights into the remaining gas content within the coal samples after desorption.

1.2.2 Indirect method

Determining the gas content of a coal reserve is a complex task, and several factors contribute to this complexity. The organic structure of coal, ongoing chemical reactions, errors occurring during core sampling, and variations in temperature and pressure conditions both in the atmosphere and in-situ are among the challenges. These conditions make it difficult to accurately detect the actual gas content. Additionally, a significant portion of the gas produced during coal formation tends to migrate to the atmosphere.

Given these challenges, a reverse engineering approach must be considered to assess the gas content capacity of coal samples. This involves evaluating the potential gas-holding capacity of the coal, essentially estimating how much gas it could contain. This reverse engineering perspective becomes crucial in overcoming the inherent difficulties in directly measuring and quantifying the gas content of coal reserves.

1.2.2.1 Sorption isotherm/pressure

In laboratory conditions, coal samples are prepared for methane gas injection. The experimental setup is adjusted to simulate reservoir conditions, and core samples are then filled with methane. Depending on the quality of coal, at specific pressure and temperature conditions, methane injection is carried out until a certain pressure is reached. The amount of methane absorbed by the coal is characterized by a sorption or adsorption isotherm. This isotherm represents the relationship between the quantity of gas adsorbed by the coal and the equilibrium conditions, such as pressure and temperature. The sorption isotherm provides valuable insights into the methane storage capacity of the coal under simulated reservoir conditions (Guan et al., 2018).

1.3 Types of Machine Learning Methods

Over centuries, an immense volume of data, measured in zettabytes, has been accumulated. Diverse types of data, including surface and subsurface datasets, have been scrutinized to address various problems. Each data point and dataset poses unique challenges inherent to its nature. Throughout history, methods and solutions have been developed to comprehend the biological, chemical, physical, and mathematical aspects underlying the collected data. However, some problems lack precise solutions or universally applicable methods across all datasets in a similar environment.

Subsurface data evaluation serves as a notable example of the challenges encountered in the same field with similar types of datasets. As the volume of available data for underground energy exploration and exploitation continues to grow, various tools have been employed for analysis. Statistical approximations, history matching, numerical solutions, and simulation software are well-known examples of these tools. The significant enhancement in computational power has opened up substantial opportunities for problem-solving in energy resource evaluation. State-of-the-art hardware and computational tools enable the exploration of relationships between different datasets and simultaneous evaluation.

Machine learning (ML) has emerged as a pivotal link between input data and predicted output (Sircar et al., 2021). ML algorithms can be categorized into supervised learning, unsupervised learning, and reinforcement learning, each with its subcategories. Figure 1.1 illustrates some of the ML algorithms used in this context.

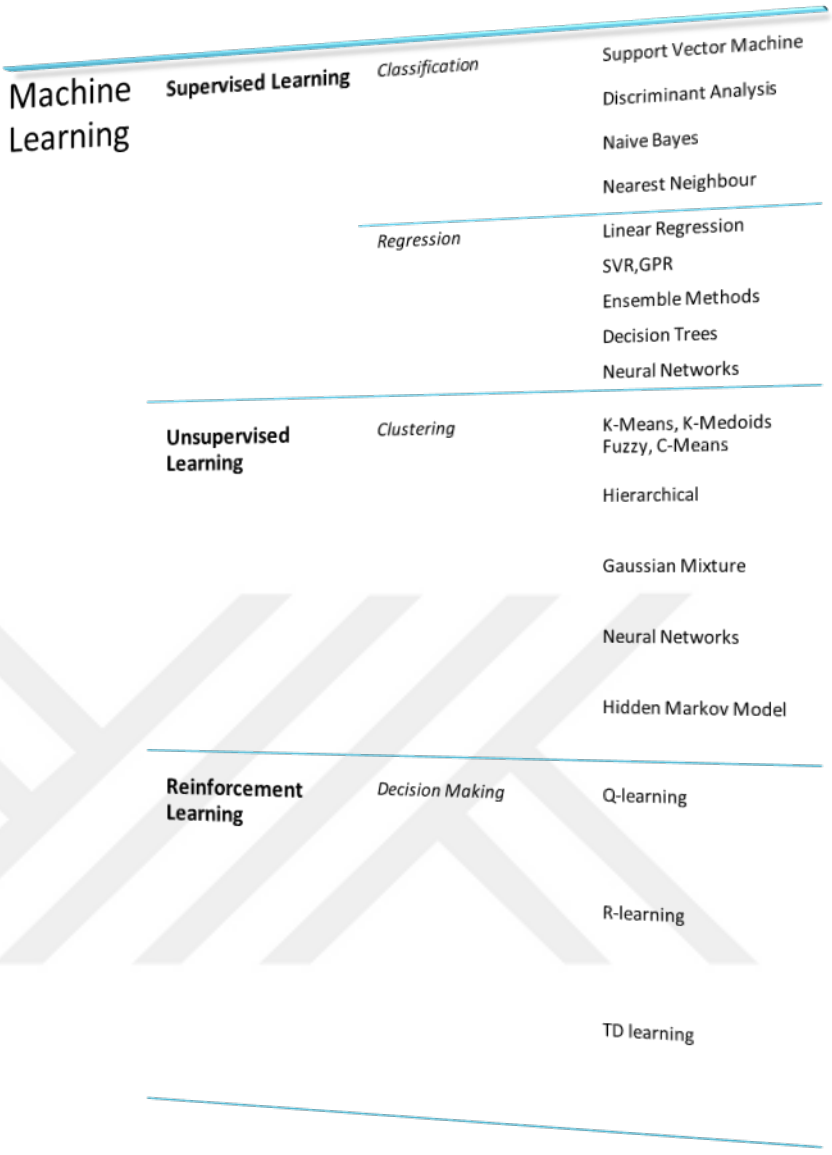

Figure 1.1 : Machine Learning Classification

1.3.1 Supervised learning

Machine learning methods are grounded in the logic of learning from experience, and one of the fundamental branches is supervised learning, which is a culmination of decades of experience. Supervised learning is categorized based on the nature of the labeled training data, and it is broadly divided into two main branches: categorical (classification) and continuous (regression) variables (Maleki et al., 2020). In categorical variables, the algorithm is trained to predict discrete classes or labels, while in continuous variables, the focus is on predicting values within a range. This fundamental division provides a framework for applying supervised learning

techniques to a wide array of problems, contributing to the versatility and applicability of machine learning in various domains.

1.3.1.1 Classification

A straightforward answer necessitates a clear question, and structuring the complexity around the variables involved is essential. In the domain of machine learning (ML), classification techniques are employed (Kononenko and Kukar, 2007). The fundamental distinction lies in the fact that classification starts with the labeling of input data. The title "supervised learning" originates from this process of labeling and supervising the investigated data from the beginning to the end.

Within the domain of supervised and labeled datasets, various methods have been developed to solve problems systematically. Support vector machines, discriminant analysis, naive Bayes, and nearest neighbors are examples of categorical methods falling under the classification umbrella. This study focuses on investigating support vector machines, specifically in terms of their representation within the classification methods, to provide a classification approach to regression problems.

1.3.1.2 Regression

The second primary grouping within the supervised learning domain is regression methods. To facilitate a more thorough analysis of the given data, inputs need to be grouped and graded. This process results in the creation of a pattern based on these groups and grades, establishing a gap. Regression, as an analogy, utilizes this difference to generate meaningful outputs (Flach, 2012). Within the regression category, several methods stand out, including linear regression, support vector regression, ensemble methods, decision trees, and neural networks. These methods are employed to model relationships between variables and make predictions based on the observed patterns in the data.

1.3.2 Unsupervised learning

Supervised learning involves a significant time and resource investment when it comes to labeling large datasets. Correct labeling requires expertise, making it a challenging task. In cases where dealing with complex and unlabeled datasets is necessary, supervised learning becomes impractical. In contrast to supervised learning, some situations do not permit establishing clear input-output relationships from the start.

This requires searching through massive datasets to derive meaningful outputs. If it's not possible to identify all the data through a specific methodology, unsupervised learning comes into play to identify similar data patterns and group them into categories. Unsupervised learning operates on a cluster-based logic, and depending on how the clustering is set up, it can associate different objects with a common output. Another advantage of unsupervised learning is its ability to reduce the dimensionality of dense data structures (Maleki et al., 2020).

1.3.2.1 Clustering

To extract meaningful patterns from given datasets, unsupervised learning employs a strategy of local grouping. It creates clusters that gather data points with a tendency to behave collectively as a single unit. These clusters serve as the foundation for establishing relationships in unsupervised learning. Among the most commonly used algorithms in unsupervised learning are K-Means, K-Medoids, Fuzz C-Means, Hierarchical, Gaussian Mixture, Hidden Markov Model, and Neural Networks. These algorithms help identify and organize data into clusters, facilitating the discovery of underlying structures and patterns within the data.

1.3.3 Reinforcement learning

The final machine learning method, known as reinforcement learning, is quite intricate. In situations where there is no pre-existing classification or neighbor datasets that can be interpreted, and where most everyday problems lack direct connections, a reward-based approach has been developed to address complex issues. This approach is particularly useful for dealing with uncertain outcomes and intricate environments, as it involves a sequence of decisions made in a specific order. These decisions are examined through trial and error, with each successful decision in the chain being rewarded to instruct the algorithm in achieving the desired logic.

An essential component of reinforcement learning is the creation of a simulated environment, followed by an initial starting point for trial and error. It then progresses with a series of feasible, practical, and ethical decisions. In general, this method has not been fully embraced, primarily due to safety concerns stemming from a lack of sufficient experience. Nevertheless, the application of reinforcement learning is gaining traction, particularly in autonomous driving, robot technologies, and complex

problems where supervised and unsupervised learning approaches have proven inadequate.

In this study, the investigation of underground coal reserves was conducted using supervised regressive machine learning methods. There are two main reasons for selecting this method. First, the prediction of coal gas content requires obtaining specific and continuous results, rather than a classification approach. It doesn't make sense to merely determine the presence or absence of gas when it's expected to have at least a trace of gas. Therefore, classification is not aligned with the objectives of this study. Second, since the datasets used in this work were labeled and limited, unsupervised methods would not be suitable.

However, it's worth noting that both supervised classification and unsupervised methods were considered in the selection of regression models used in this study, aiming to create a reliable foundation for further development of alternative machine learning methods for coal gas prediction. As a result, common machine learning techniques were employed to explore the fundamental relationships between methane content in the coal samples.

1.4 Statement of Problem

Over the past century, there has been a significant increase in energy consumption, which has led to concerns about carbon emissions and their impact on the environment. In response to this, the world has been compelled to address its carbon footprint. Under the Paris Agreement, the largest global economies have committed to a goal of achieving a zero-carbon emission policy. This has led to a shift in focus towards reducing coal consumption, while still meeting the growing demand for energy and striving to achieve zero emissions.

To strike a balance between energy demand and emission reduction objectives, a middle ground must be established. In support of the transition to lower carbon emissions, energy companies have been investing heavily in renewable energy over the past decade. Simultaneously, significant technological advancements in unconventional oil and gas reservoirs have enabled coal miners to extract methane gas from underground coal reserves. The unexpected COVID-19 pandemic disrupted the energy transition plans, and sudden shifts in the energy market led to a reevaluation of

coal reserves. Increased demand and limited supply of carbon-based energy commodities resulted in a surge in prices.

In this context, the importance of Coalbed Methane (CBM) resources and predictions has come to the forefront. Determining the presence and potential of Coalbed Methane is a process that involves drilling operations and laboratory experiments, incurring both time and financial costs. Therefore, it is proposed that Machine Learning techniques be used to detect methane potential quickly, reliably, and at a lower cost, reducing the overall expenses during the pre-development stage of CBM content analysis.

In summary, all of these designs and investigations are conducted with the aim of achieving a net-zero carbon emission transition, maximizing the efficiency of coal reserve development, and increasing profitability in research and development stages.

1.5 Scope of The Study

Since the dramatic increase in carbon emissions over the past century, reducing carbon consumption has become a primary objective. However, coal remains a common and unrivaled energy source for some countries, leading them to explore alternative methods to improve their coal reserves. Methane content in underground coal reserves is a significant concern when it comes to coal exploitation. Beyond addressing carbon emissions, this methane can be converted into alternative natural gas reserves. Therefore, it has become crucial to determine methane content for purposes such as de-gassing or methane production.

The determination of Coalbed Methane (CBM) content is an expensive process, mainly due to the high volume of drilling operations and the heterogeneous nature of coal formations. A single well experiment may not accurately reflect the entire coal system in the studied area when it comes to CBM content. To mitigate these issues, reduce research costs, and address the unpredictability of coal systems, the idea of applying Machine Learning (ML) methods has emerged. ML techniques can be valuable in accelerating the investigation of large underground mining leases and compensating for missing data, ultimately facilitating the efficient utilization of coal reserves.

To delineate the extent of this research, a diagram displayed below has been designed to illustrate the workflow used for generating data-driven solutions for gas content analysis.

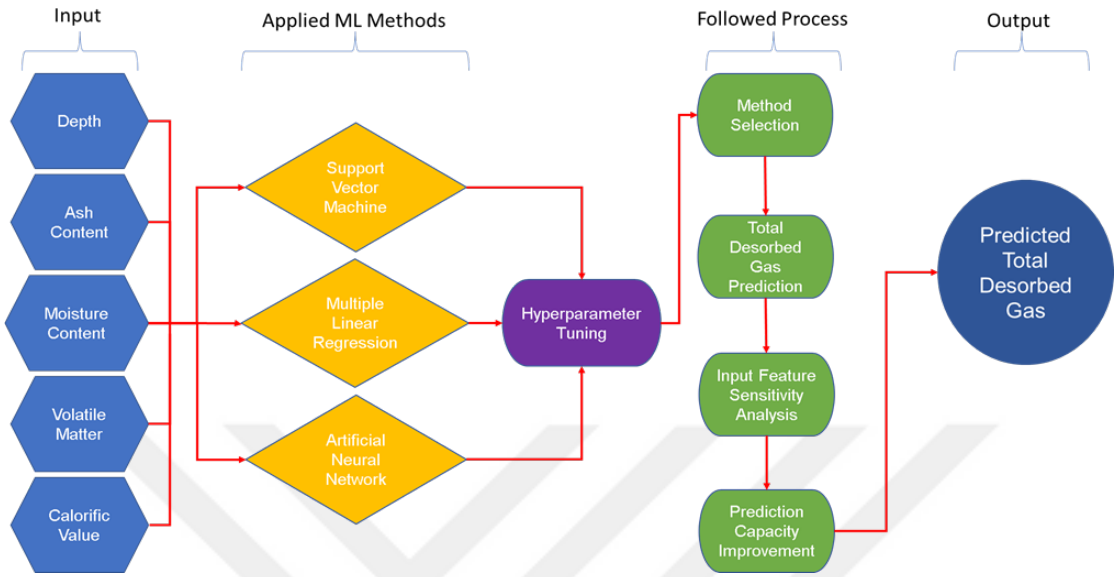

Figure 1.5 : Investigated Method diagram for prediction of coal gas content. (Akdaş and Fişne, 2023)

This research presents an innovative method that provides an alternative means to circumvent the lengthy experimental setup and its associated delays. By using fundamental parameters to assess the quality of coal samples, independently of gas content measurements, this approach broadens the scope of predictions and addresses issues arising from disparities and various factors influencing the uneven distribution of coal gas in underground systems. Furthermore, the resemblances between coal and shale reserves enable the application of the same approach in the exploitation of source rocks for unconventional oil and gas systems like shale oil and shale gas (Akdaş and Fişne, 2023).

2. GEOLOGICAL SYSTEM and ASSUMPTIONS

Turkey possesses a distinctive geological structure that brings both a wide range of geological riches and challenges. The region's geological history is closely tied to the Tethyan Sea (ancient Mediterranean), which played a crucial role in the formation of Anatolia. This process involved the submergence of the African continent beneath the Asian landmass. Located at the crossroads between Asia and Europe, Anatolia's geological system has become a complex and challenging puzzle for geologists and scientists to interpret due to its unique history and geological features (Şengör and Yılmaz, 1981).

Complex geological structures have led to the formation of diverse underground resources. Recent discoveries made by the General Directorate of Mineral Exploration and Research have significantly increased Turkey's coal reserves, bringing the total to 19.1 billion tonnes. These findings highlight the wealth of geological resources in the country and have important implications for energy and resource planning (Leleoglu, 2020). Lignite is the most prevalent type of coal found in Turkey. Within these lignite reserves, the calorific values vary, ranging from 1,000 kcal/kg to 4,200 kcal/kg. Notably, around 90% of Turkey's lignite reserves have a calorific value of less than 3,000 kcal/kg. This information underscores the predominance of lower-calorific-value lignite in Turkey's coal resources.

2.1 Kınık Coalfield, Soma Basin Physical System Description

2.1.1 History of the coalfield

The underground coal mine license being investigated is located on the border between the cities of Izmir and Manisa in the Soma Coal Basin. Coal production in this region has been ongoing since 1950, primarily through open-pit mining operations. However, with the discovery of new coal reserves extending from the north to the south, there has been a shift towards underground coal mining methods.

Recent discoveries of coal seams have prompted the transition from open-pit mining to underground mining. Records indicate that Turkey's deepest underground coal mine is planned to be established in the Kınık Coalfield. The expected coal reserve in this area is approximately 250 million tonnes. This development signifies a significant

expansion in underground coal mining in the region, reflecting changes in mining methods to access deeper coal seams (Esen, 2021) . Location of the studied area, coal mining boundaries respect to depth and investigated portion of the Soma Basin showed in Figure 2.1.

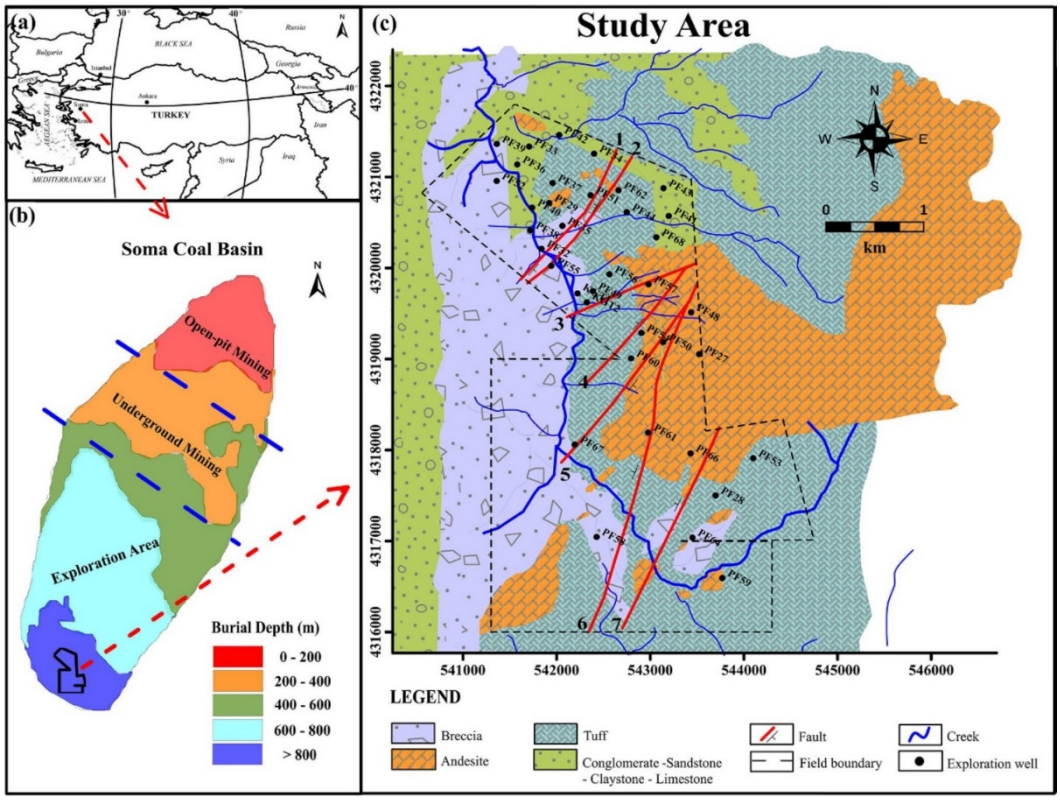

Figure 2.1 : (a) Soma Coalfield location in Turkey. (b) Location of Exploitation area in Soma Coalfield Basin. (c) Geological structures in studied area (Esen, 2021).

2.1.2 Geological aspect of the coalfield

The Kınık coalfield is characterized by a geological composition that includes Mesozoic carbonates and Miocene volcanic rocks. During the Early to Middle Miocene period, the Soma Formation in this region was filled with a combination of lacustrine-based rocks, including limestone-clay, sandstone-siltstone, and claystone. Within the Soma Formation, there are two notable coal seams referred to as kM2 and kM3.

As the Miocene period came to a close, the Deniş Formation was deposited in the area. This formation contains a relatively shallow coal seam known as kP1. These geological details provide insights into the composition and history of the Kınık coalfield and the coal seams found within it (Oskay et al., 2019). Investigated area and coal seam formations showed in Figure 2.2.

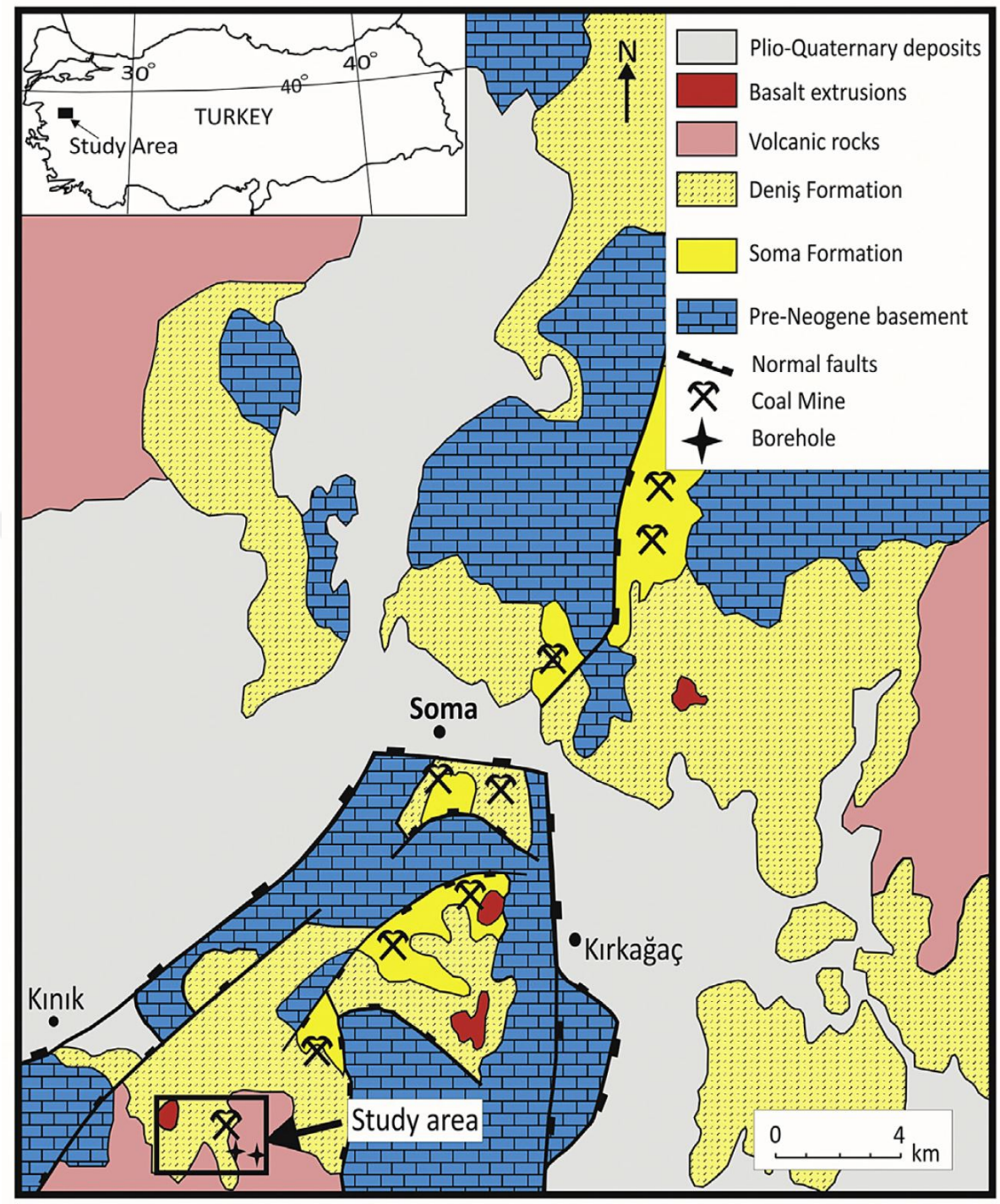

Figure 2.2 : Geological map of the Soma Basin (Oskay et al., 2019).

Expected most shallow coal seam is kP1 and relatively second thickest coal seam after kM2. kM2 also the deepest coal reserve according to previous studies. Kınık Coalfield stratigraphy showed in Figure 2.3.

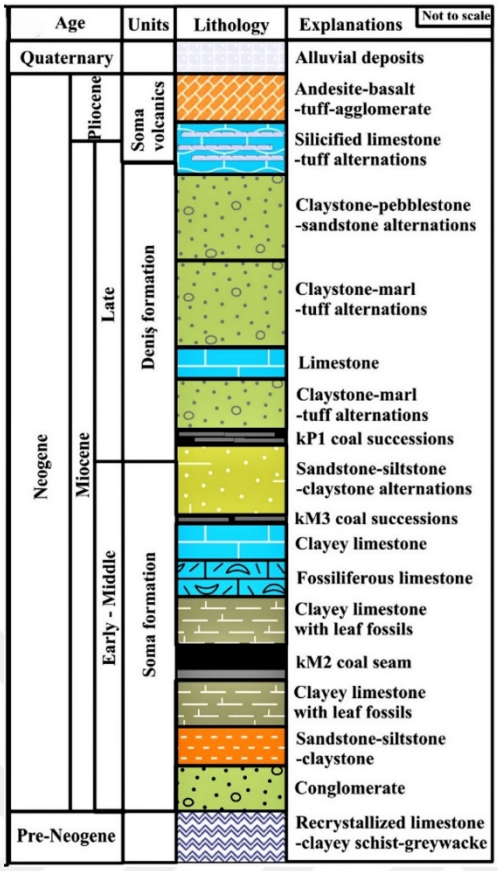

Figure 2.3 : Stratigraphic column of the Soma Basin (Oskay et al., 2019).

The Soma Formation is dated to the lower-middle Miocene period. This geological formation exhibits varying thickness, typically ranging from 50 to 100 meters. It covers a substantial area with a width of approximately 4 kilometers and a length of around 25 kilometers. The Soma Formation is characterized by its deposit orientation, which is predominantly from the northeast to the southwest.

Within this geological structure, there are five foundation layers, which are aligned from upper levels to lower levels. These layers play a crucial role in understanding the geological history and composition of the Soma Formation and the coal reserves it contains;

- Middle Coal Series (kM3)
- Limestone Series
- Marn Series
- Lower coal Series (kM2)
- Foundation formation

At the conclusion of the described formation, the foundation is comprised of a mixture of shale, graywacke series, and recrystallized limestone rocks. These geological components provide valuable insights into the composition of the lower layers of the formation.

Another significant underground rock structure in this area is the Deniş Formation. The Deniş Formation was formed on top of the Soma Formation and is characterized by clastic rock deposits. Units within the Deniş Formation are further detailed as follows, likely providing information about the geological layers and composition within the Deniş Formation;

- Andesite – Tuff – Agglomerate
- Conglomerate – Marn - Tuff
- Upper Lignite Level (kP1)
- Sandstone – Siltstone – Shale Level

The Deniş Formation plays a crucial role in guiding the exploration of the deeper portions of the coal distribution in the Kınık Coal field. Its significance is notably associated with the presence of the kP1 coal seam, which typically ranges in thickness from 0.5 to 4.5 meters. While kP1 may not hold significant economic importance for underground coal mining, it holds a special significance as it is the first coal layer to be encountered in the field with respect to depth. This discovery can provide valuable information and insights for further exploration and development in the area (Esen, 2021). To identify the Kınık coal field and assess its coalbed methane potential, this study has utilized data from the kP1 and kM2 coal series found within both the Deniş and Soma Formations.

2.2 Data-Driven Model Assumptions of Kınık Coalfield

Kınık Coalfield was investigated by Dr. Olgun Esen in his dissertation (Esen, 2021). This research has conducted investigations at 37 different locations, and a total of 84 core samples have been collected from these sites. Within these collected core samples, 24 of them pertain to the kP1 coal seam, while the remaining 60 core samples were extracted from the kM2 coal seam. Notably, the smaller kM3 coal seam was excluded from the study. This decision was made in light of the irregular distribution of coal within the locations and the fact that some drilled wells contained insufficiently sized

samples that could not support continued experimentation. These considerations helped streamline the study and focus on the more representative and feasible data.

2.2.1 Expected coalbed methane content

To assess the methane content in the core samples obtained from the Kınık Coalfield, a series of distinct stages were carried out. The initial step involved the successful coring of core samples from the two primary coal seams, followed by the measurement of the preliminary gas desorption under atmospheric pressure and at room temperature. The second phase of the process involved periodic desorption analysis conducted in a laboratory setting until all free gas content was fully extracted from the core samples. The final assessment was performed using dried coal samples to analyze the waste gas (Q3) in relation to crushing samples smaller than 212 μm in size.

The results of these analyses revealed that the total gas content varied, with values falling between 0.9 and 2.61 m^3/t for the kP1 coal seam and approximately ranging from 0.51 to 3.86 m^3/t for the kM2 coal seam. This data provides essential insights into the methane content within these specific coal seams in the Kınık Coalfield (Esen, 2021).

2.2.2 Initial and boundary conditions

It is a well-acknowledged fact that all underground structures exhibit variations in mineral deposits and heterogeneity in the distribution of solid and fluid components. Depending on factors like discontinuities and tectonic activity, the structure may display different alignments within the same region. Consequently, each coal reserve must be studied individually to account for these variations.

It is crucial to note that coal is a unique carbon resource that can be used directly as fuel or serve as a reservoir for gases, primarily methane. In essence, coal consists of both solid and fluid components, often in the form of gaseous compounds like methane. This duality creates multiple potential carbon sources.

If coal samples contain bacterial activity that produces additional gas, it adds an extra layer of complexity to the assessment of coal reserves in terms of methane and fuel resources. In addition to the factors mentioned above, several other criteria contribute to the intricacy of underground systems.

- A geological structure (faults, fissures, synclinal, anticlinal, etc.)
- Thickness of cap rock above the coal seam
- Depth of coal seam
- Thickness of coal seam

Indeed, beyond the aspects mentioned earlier, understanding the initial formation of coal and the processes that transformed it into its final state are critical. These questions are essential for both defining coal reserves with a more comprehensive approach and conducting more accurate analyses of gas potential within these reserves. Desorbed gas concentration of Kınık coal field was detected by Esen's dissertation (Esen, 2021). Moreover, this research has yielded essential findings, and detailed field explanations have been provided. The aim of this study is to streamline the processing of data obtained from field and laboratory conditions, seeking a more simplified approach to handling this valuable information. This simplification is crucial in making the data more accessible and practical for analysis and decision-making.

3. MACHINE LEARNING METHODS RESULTS for PRODUCED GAS DISTRIBUTION and ITS VERIFICATION

3.1 Machine Learning Algortihm Selection Problem

Machine learning algorithms underwent thorough examination. To gain a more comprehensive understanding, two primary approaches were explored. Firstly, in Introduction, a general machine learning (ML) algorithm was introduced, highlighting the necessity of presenting results that encompass three key categories to encompass fundamental machine learning techniques:

Supervised Learning Algorithms: This group included methods like Support Vector Machine (SVM) and Linear Regression (LR). These methods serve as distinct techniques, with SVM being used for classification and LR for multiple linear analysis (MLR).

Additionally, alongside supervised learning algorithms, unsupervised learning was explored using a Neural Network model to address the same problem.

This approach ensured a thorough exploration of various machine learning techniques, providing a more comprehensive perspective on the research problem.

In introduction, emphasized that reinforcement learning is a more complex algorithm that demands dynamic learning through multiple trial and error iterations for a single judgment. Consequently, it consumes more time and computational power to yield results. Given the goal of achieving accurate predictions for the desorbed total gas in the Kınık Coalfield, the decision was made to refrain from using the reinforcement learning method.

Apart from reinforcement learning algorithms, various branches and subcategories of machine learning techniques were selected after conducting preliminary analyses to enhance the accuracy of error assessments. In line with scientific curiosity, additional analysis will be detailed in future works.

The second fundamental idea is to develop practical approaches. The primary objective of applying machine learning techniques to coal desorption analysis is to reduce the need for unnecessary fieldwork and save time during the development stage. To gain

a deeper understanding of the distribution of total gas in the Kınık coal field, data from both the kP1 and kM2 coal seams were used in conjunction.

In this chapter, machine learning (ML) algorithms will be utilized using three specified methods, all within the context of supervised regression techniques. The objective is to provide a detailed extension of these methods to understand how the calculations are performed and how the results obtained are established based on specific approaches.

For the ML application, the "Python Spyder 3.9" integrated development environment (IDE) and the "sklearn" machine learning libraries were employed. These tools were chosen to facilitate the implementation of machine learning techniques and the analysis of data in a user-friendly and efficient manner.

To begin, we've opted for supervised learning using the Support Vector Machine (SVM) because it can effectively handle both classification and regression tasks. SVM is a versatile technique within supervised learning, accommodating various methods such as classification, regression, and outlier detection.

SVM functions by creating hyperplanes that reconfigure multidimensional data into distinct classes by establishing layers of hyperplanes to effectively separate the data. In the context of supervised classification, SVM utilizes this data separation to construct a classifier that is readily applicable to new data sets. In classification scenarios, SVM determines an optimal separating hyperplane to partition the data set into two categories. The selection of the optimal hyperplane is driven by the objective of minimizing losses in the analysis of the desired data set (Gunn, 1998).

To optimize prediction efficiency, the approach involves the construction of an optimal hyperplane that intersects with two closest data points, which are referred to as "support vectors." This optimal hyperplane is positioned in such a way that it maximizes the margin, which represents the minimum Euclidean distance between the hyperplane and the support vectors. This configuration ensures that the predictive model is fine-tuned to achieve the most accurate and efficient results by utilizing these key elements. Optimum hyperplane, support vectors, and margin are showed into Figure 3.1.

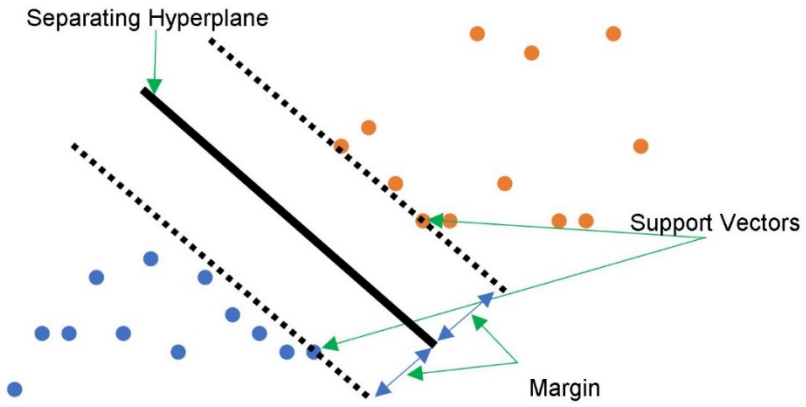

Figure 3.1 : Support Vector Classification Hyperplane.

In a classification problem, the process involves aligning labeled data with specific classifiers. To accomplish this, Support Vector Machine (SVM) employs a two-class classifier, which corresponds to one of the two labels representing the two classes. More precisely, the labeling is typically done with +1 for positive examples and -1 for negative examples. The labeled data is organized and controlled with index values ranging from "i" to "n."

The input vector, denoted as "x," consists of individual components labeled as "x_i." During the training phase, both the input data and the corresponding outputs are represented as (x_i, y_i). Initially, SVM assumes that the input data is in vector form and follows a linear pattern. However, to accommodate a broader range of data patterns, kernels are introduced to relax this initial assumption, allowing for the inclusion of data points in either linear or nonlinear fashion. This expansion in flexibility is a significant development in SVM's capabilities (Ben-Hur and Weston, 2010).

The linear classifier in SVM classification defined with scalar product of weight (w^T) and inputs (x_i) shown as,

$$w^T x = \sum_i w_i x_i , \tag{3.1}$$

and linear classifier can be written form of discriminant function as adapted Ben-Hur and Weston (2010),

$$f(x) = w^T x + b. \tag{3.2}$$

Apart from the weight and input variables, "b" represents the bias or intercept. In simpler terms, the bias can be described as the distance between the origin and the hyperplane. As previously mentioned, the hyperplane serves to divide the data set into two parts, defined by the discriminant function, which distinguishes between the negative region (below the hyperplane) and the positive region (above the hyperplane).

The equations presented in Eqs. 3.2 were initially constructed based on linear input parameters, guiding SVM to create a linear classification model. However, in this study, the input parameters exhibit nonlinear behavior. To accommodate this nonlinearity, nonlinear parameters were introduced to the SVM model. As a result, the Eqs. 3.2 were extended to form a nonlinear classifier, enabling SVM to effectively handle nonlinear patterns and relationships within the data, as shown,

$$f(x) = w^T \phi(x) + b. \tag{3.3}$$

Nonlinear input parameters come with several disadvantages, including increased computational cost, higher dimensionality, and the potential for errors. To mitigate these issues, kernel functions are employed. Kernel functions are a powerful tool in machine learning that enable the transformation of data into a higher-dimensional space, where it may be easier to work with and analyze.

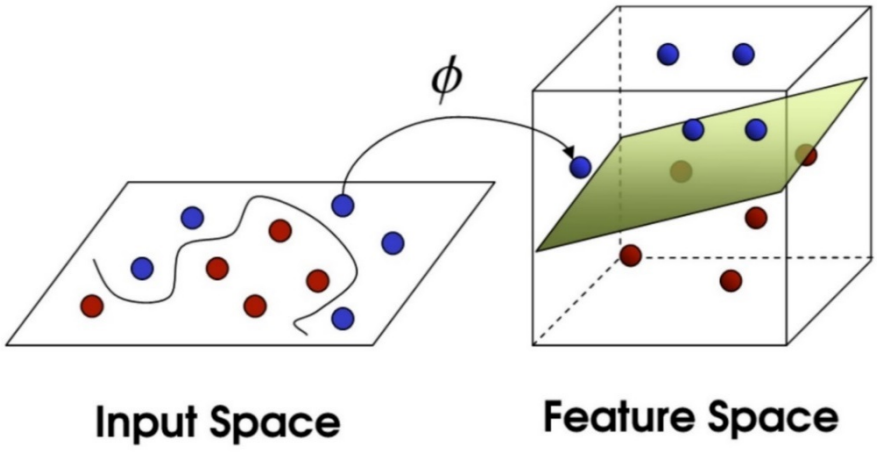

Figure 3.2 : Kernel Function use for Mapping adapted from (Wilimitis, 2022).

The kernel functions are implemented to transform input data into desired feature space by using linear or nonlinear mapping as seen Figure 3.2. Kernel functions summarized by Kor (2021) as shown,

$$K(x, x') = \langle \phi(x), \phi(x') \rangle, \tag{3.4}$$

and multiple variables extended as,

$$\phi(x) = \left(\phi_i(x), ..., \phi_n(x)\right)^T \quad (3.5)$$

In literature, different type of kernel functions is developed. In this study, radial basis function (RBF) used. Reason behind the RBF selection is having better data processing with controlled radius approach. RBF type of kernel function shown as,

$$K\left(x, x'\right) = x^T x' \quad (3.6)$$

and

$$K\left(x, x'\right) = \exp\left(-r \left\|x - x'\right\|^2\right). \quad (3.7)$$

In Eqs. 3.7, “r” parameter is defined as variable is controlling the radius of kernel function.

Determining the quantity of desorbed gas is a unique challenge. To address this, the Support Vector Machine (SVM) classifier was tasked with the identification of binary classifications and support vectors for prediction. Similar to the SVM classifier approach, there was a desire to implement the SVM Regression method to achieve more sensitive predictions. To construct a prediction model, the "sklearn" library in Python was utilized. Within this library, the "svm.SVR" method was employed.

The "svm.SVR" method essentially operates as a classical regression method. This is particularly important because the data from the Kınık coalfield is highly variable. Therefore, preprocessing was necessary to standardize the feature values in the Kınık coalfield dataset. To standardize this data, the following equation was utilized.

$$z = \frac{x-u}{s}. \quad (3.8)$$

In the equation, "x" represents the input data that needs to be standardized. "u" stands for the mean of the input data, while "s" represents the standard deviation of the input data. The result, denoted as "z," is the standardized data. This standardized data is employed within the "svm.SVR" method to facilitate more effective and efficient analysis. Standardization ensures that the data is transformed to a common scale, making it suitable for use in the SVM Regression method. Transformation of input data from “x” to “z” is showing in Figure 3.3,

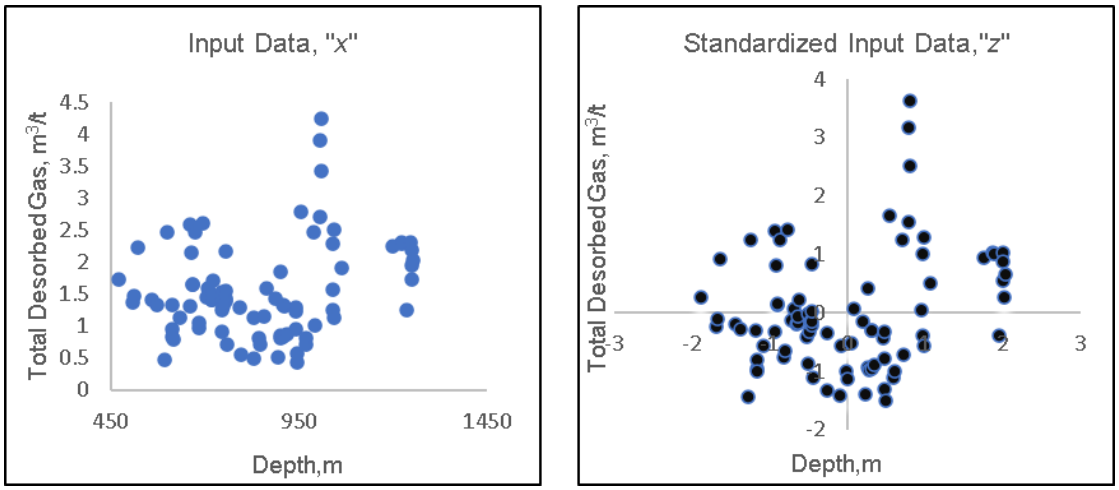

Figure 3.3 : Standardization of Kınık Coalfield Data set sample respect to Depth versus Total Desorbed Gas content.

Algorithm used the default parameters except, “kernel”, “tol” tolerance for the stopping criterion, “epsilon” distance which is determines the radius of no penalty associated in the training loss function with predicted points. And finally, regularization parameter “C” was used to normalize set of dispersed data. Optimized parameter and values showed in Table 3.1.

Table 3.1 : SVM Optimized variables and values.

Kernel Function	Tol (Tolerance)	Epsilon	C
RBF	0.1	0.1	325

Kınık coal field data which is consist of kP1 and kM2 info is used to process with Support Vector Machine method under the selected parameter in Table 3.1. Predicted and original results are shown in Figure 3.4.

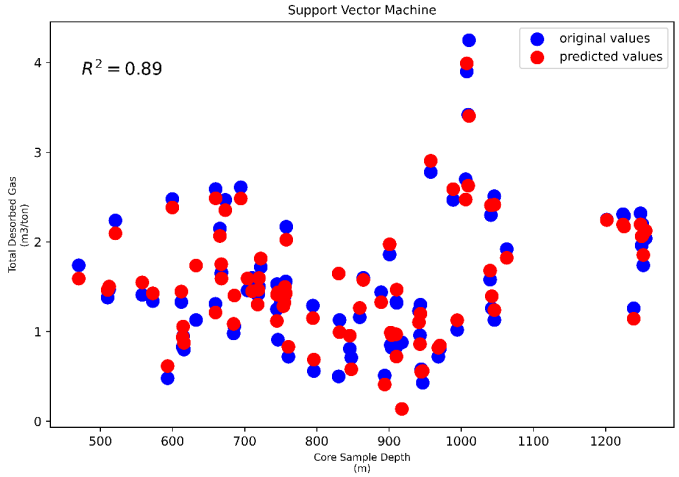

Figure 3.4 : The comparison of SVM prediction and original data obtained for kP1 and kM2 coal seams.

In order to see error distribution on predicted data with SVM Regression, it is desired to calculate absolute error with predicted values ($\hat{y}$) and original desorbed gas values (y) is shown as,

$$\epsilon_{abs} = \left| \frac{\hat{y} - y}{y} \right| \times 100. \tag{3.9}$$

Absolute error (0_{abs}) calculated from Eqs. 3.9 and unit is percentage (%). This point and forward, ML algorithm that have used in this study completed with error calculation and distribution with Eqs. 3.9.

The Support Vector Machine (SVM) achieved an average error of 14.43% when compared to the actual data collected from fieldwork. To provide a general overview, both the training data and the predicted data set were utilized without any outlier detection. This evaluation helps to assess the performance of the SVM model in predicting desorbed gas quantities in the Kınık Coalfield. Error distribution regarded to SVM Regression is showed in Figure 3.5,

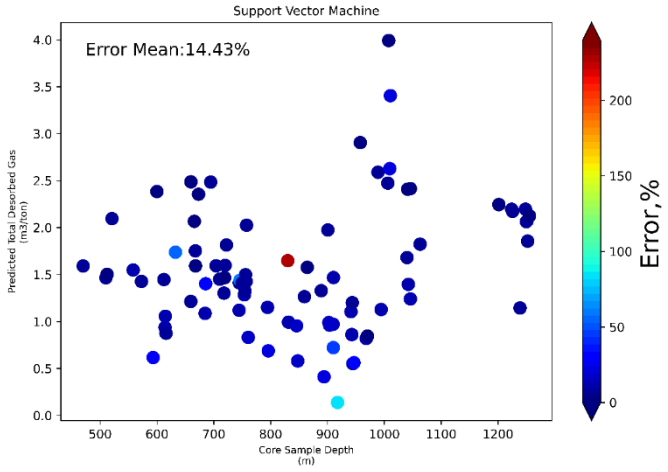

Figure 3.5 : Error distribution of SVM on Kınık coal field data prediction.

Multiple linear regression is a thoroughly studied method that relies on a geometrical approach. In this approach, the training data exists within a Cartesian coordinate system. A straightforward way to connect these data points is by drawing a linear line that aims to capture the majority of the data points as closely as possible. This linear line represents the regression model, which seeks to establish a relationship between the independent variables and the dependent variable in a way that minimizes the overall error.

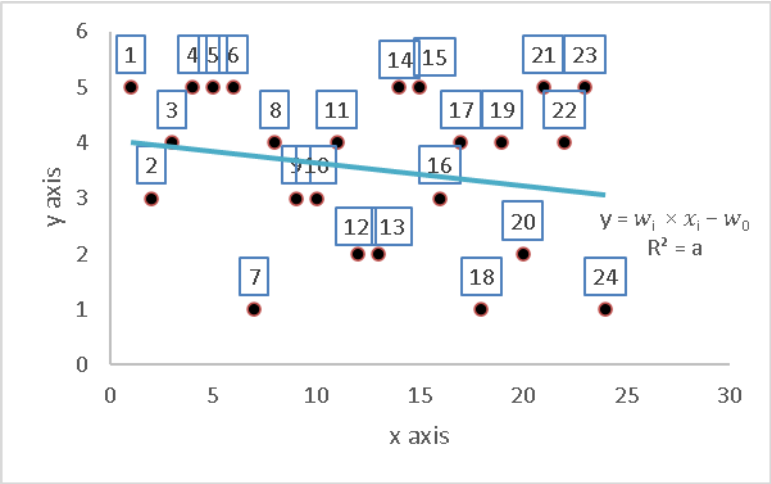

Figure 3.6 : Example of linear-regression training analysis.

The linear line or trendline (showed with blue color in is Figure 3.6) is the key factor that creates the equation to predict further input data results. In Figure 3.5, it is shown that prediction equation is $y = w_i \times x_i - w_0$ and score of the created analogy is shown with $R^2 = a$.

Linear models require numerical parameters that are learned from the training data, and these parameters are not predetermined. The nature of the input data shapes the linear approach, reducing the variation between the linear model and the input data, making it a simple and effective representation. Outliers, which are data points significantly different from the majority of the data, have a relatively small impact on the linear model when compared to the bulk of the data points in the inputs.

One of the key advantages of the linear model is its reliance on a limited number of numerical parameters to make predictions for future data points. However, the model may not perform well with points located at the outer boundary that lack input data or fall outside the scope of prediction based on the majority of the training data. This limitation can lead to underfitting, where the model doesn't capture the full complexity of the data (Flach, 2012).

In the extent of supervised learning algorithms, regression analysis was conducted using a linear regression model. This model is built by fitting a linear trendline to the data, which involves determining coefficients. These coefficients are calculated in such a way that they minimize the residual sum of squares between the training data and the predicted data sets when using a linear approximation. In essence, the linear regression model aims to find the line that best represents the relationship between the

input variables and the target variable, minimizing the errors or residuals between the observed data points and the values predicted by the model.

$$y = w_i \times x_i - w_0 . \tag{3.10}$$

Equation 3.10 illustrates a simple linear regression equation as an example. The goal here is to train the data represented by "x_i" and obtain the weight "w_i" and the intercept "w_0" to reach the target value "y." In the context of linear regression training, "w_0" and "w_i" are the model coefficients that are determined. Once the model coefficients have been established, they can be used to estimate new target values when provided with new input data. The key rule is that both the training data and the data under investigation must be labeled (James et al., 2014).

In this study, Multiple Linear Regression (MLR) was employed in addition to SVM. MLR involves the analysis of the relationship between multiple independent input parameters and an output dependent parameter, where the dependent parameter responds to various independent parameters. Unlike simple linear regression, MLR presents a unique challenge in that there can be relationships between multiple independent variables and the dependent variable. The MLR method can be expanded as follows,

$$y = w_0 + w_1 \times x_1 + w_2 \times x_2 + ... + w_n \times x_n . \tag{3.11}$$

In equation 3.11, “w” is representing the weight or regression coefficients which is changes in respond to independent variable (i.e., $w_1 \times x_1$), “x” is the independent variable, “y” is the dependent output from regression analysis. “w_0” is the model intercept. “i” in weight and independent variable which is started from 0 to n represent the number of variables.

Ordinary least square principle works as;

$$\min \left\| w_n \times x_n - y \right\|_2^2 . \tag{3.12}$$

In extension to multiple linear regression, Eqs. 3.13 can be display as;

$$\text{Least Square } (LS) = \sum_{i=1}^{n} \left(y - w_i \times x_i\right)' \times \left(y - w_i \times x_i\right), \tag{3.13}$$

and variables can be shown as (Montgomery and Runger, 2003);

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, x = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \dots & x_{nk} \end{bmatrix}, w = \begin{bmatrix} w_0 \\ \vdots \\ w_k \end{bmatrix}. \quad (3.14)$$

The principle of least square method is to minimize the “LS” parameter in Eqs. 3.13. In order to do that, least square estimator $\hat{w}$ is the solution for w in the equations.

$$\frac{\partial LS}{\partial w} = 0. \quad (3.15)$$

Result of derivation of Eqs. 3.15 shown as;

$$x' x \hat{w} = x' y. \quad (3.16)$$

Eqs. 3.16 seven is scalar form of matrix solution and both are identical. In order to solve Eqs. 3.16 both sides dived with same variable ($x' x$) than equation became,

$$\hat{w} = (x' x)^{-1} x'y \quad (3.17)$$

In final form, trained model with coefficient matrix “$\hat{w}$” is applied into new independent variables “x_{new}” to find objective function “$\hat{y}$” shown as,

$$\hat{y} = x_{new} \times \hat{w} \quad (3.18)$$

In this study, Linear Regression was implemented using the linear model from the "sklearn" library. To train the algorithm, the Kınık dataset was employed, and the linear regression model was developed based on the regression correlation between the input features and the target variable. Similar to the SVM analysis, Linear Regression was trained using the Kınık field data to enable predictions within the Kınık Coal system, specifically for estimating desorbed gas quantities. Then results showed in Figure 3.7,

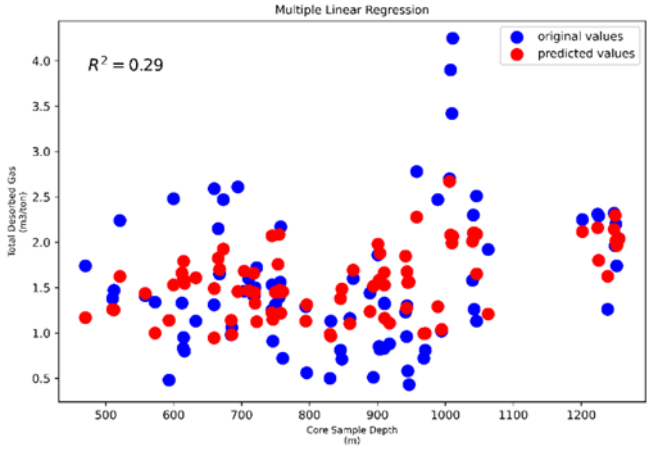

Figure 3.7 : The comparison of Linear Regression prediction and original data obtained for kP1 and kM2 coal seams.

The Linear Regression model achieved an average error of 39.99% when compared to the actual data collected from fieldwork. Similar to the SVM analysis, the purpose was to gain a general understanding of the algorithm's application in gas desorption. Consequently, both the training data and the predicted data set were utilized without any outlier detection. This approach helps provide an overall assessment of the performance of the Linear Regression model in predicting desorbed gas quantities in the Kınık Coalfield. Error distribution of linear regression method showed in Figure 3.8,

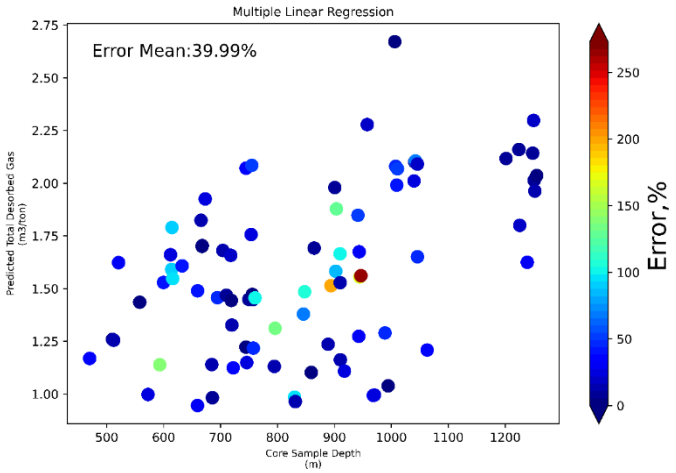

Figure 3.8 : Error distribution of Linear Regression on Kınık coal field data prediction.

In this section, another popular approach was employed because it is available in both supervised and unsupervised learning contexts. This approach is the Neural Network, which is a comprehensive method with the capability to operate in a supervised learning manner. The term "Neural Network" is derived from the structure and

function of neurons in the human body. Neuron cells have intricate networking capabilities, functioning through the transmission of signals back and forth. The complexity of neural networks, their ability to coordinate commands throughout the body, produce multiple responses, and process numerous inputs into a single action, has opened up extensive and versatile possibilities for solving problems that were once considered unsolvable. The intricate nature of human neural networks has inspired the development of the Artificial Neural Network (ANN) method in the field of artificial intelligence (AI) (Belyadi and Haghighat, 2021). The simple ANN architecture is the perceptron. Perceptron is using single input layer and output node. It is visualized in Figure 3.8.

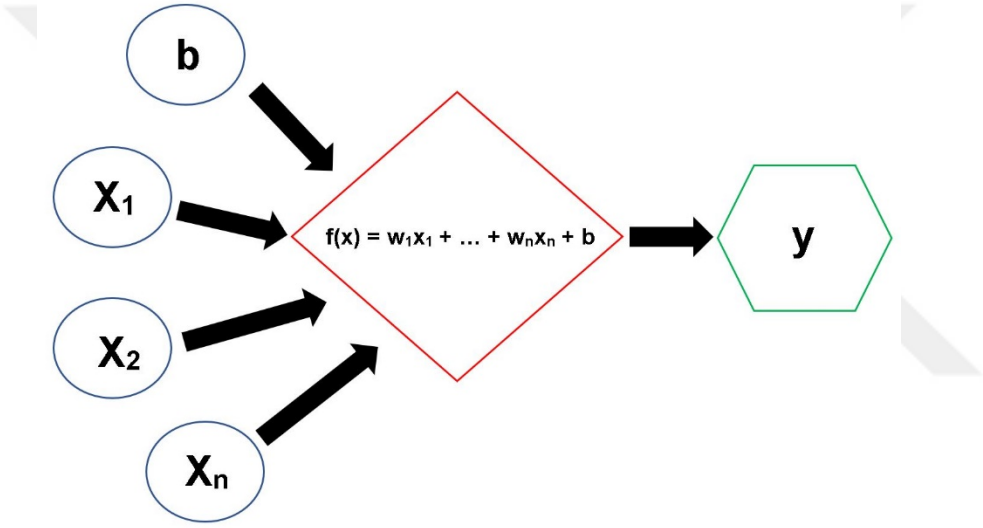

Figure 3.9 : Artificial Neural Network structure: The perceptron.

As it is shown in Figure 3.9 input data (x_n) introduced into system with external input (b). There is a reason behind the introduction of "b". If the given inputs will not enough to produce reasonable outcome in the evaluation stage between back and forth propagations, the external input or bias (b) is providing a result that can be used by another neurons in order to retain computation. Same as the supervised learning algorithm, ANN has weight variables (w) whereas difference with existence of hidden layer.

The Artificial Neural Network (ANN) is a suitable choice for addressing complex problems where traditional mathematical approaches prove impractical. Beyond its ability to handle complexity, ANN is well-suited for dynamic learning, continuously adapting and improving as it ingests new data. Furthermore, ANN establishes

connections between unlabeled data sets through hidden layers positioned between the input and output data layers.

In general, ANN identifies inputs through its algorithm, assigns weights to each data point, and subsequently cross-validates these data sets by correlating the assigned weights. If the weight adjustment calculations become overly complex within the computational sequence, multiple hidden layers can be introduced between the correlated weights and the final output. The number and density of these hidden layers are determined by the user with the goals of reducing computational costs, enhancing output accuracy, and creating a nuanced relationship between linear and nonlinear inputs during the processing stage. This flexibility allows ANN to effectively address a wide range of problems.

Artificial Neural Network structure showed in Figure 3.9 is extended with cross validation of input variables in hidden layer in Figure 3.10.

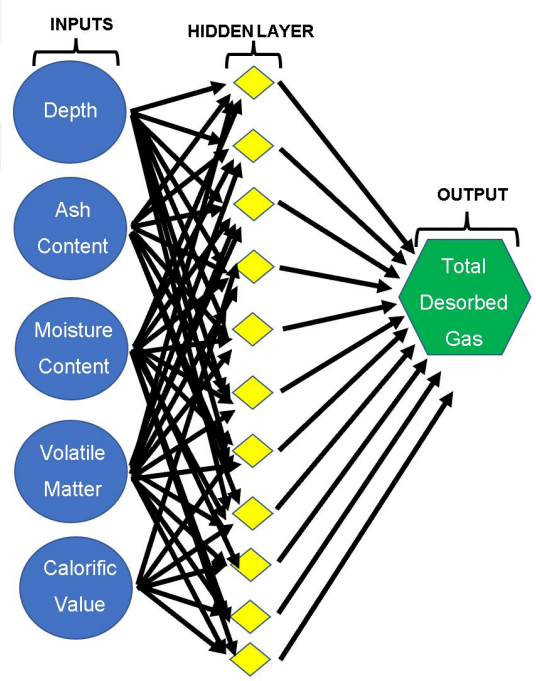

Figure 3.10 : Hidden layer demonstration of Artificial Neural Network.

ANN is using streamline that begins with input layer send a signal into the action potential of the node "j" in the hidden layer (h_j). Since the signal is integrated with bias (b) and selected activation function, second phase which is weight correlation with back and forth propagation starts then final results send into output layer. The calculation made in action potential in hidden layer shown as,

$$h_j = \sum_{i=1}^{n} w_{ij}\, x_i \; + \; b_j \,. \tag{3.19}$$

In Eqs. 3.19 number of inputs shown with “i” indices and number of nodes shown with “j” indices. The bias is an arbitrary term that makes output value 1 in case of input values are 0. In order to calculate output or output layer, activation functions are using. Activation function is determining the amplitude of the output signal within ranges of “0 and 1”, or “-1 to 1” (Belyadi and Haghighat, 2021). In general, there are four different type of activation functions.

In this study, tanh function was used by Neural Network algorithm. Thus, simplification made in activation function section to focus on general performance of the ANN. Tanh function is stated as,

$$F_{tanh}\left(h_j\right) = \frac{1\text{-}\exp(\text{-}2\, h_j)}{1+\exp(\text{-}2\, h_j)} \,. \tag{3.20}$$

In order to have output, tanh function can be applied on node calculation in hidden layer as,

$$y = F_{tanh}\left(h_j\right) \,. \tag{3.21}$$

ANN is using minimum error to correspond weight of the hidden layers. Thus, error calculated in predicted weight is minimize with least-square as,

$$\min L_w = \sum_{x,y} (y - \hat{y}\;)^2 \,. \tag{3.22}$$

In Eqs. 3.22 “Lw” is the loss function and “$\hat{y}$ ” is the target variable. The least square method is applicable on continuous target variables. Some instances are prevented to differentiate loss function. So that, smoothing applied on Eqs. 3.22, then equation became into,

$$\nabla LS_{smooth} = \sum_{x,y} (y - \hat{y}\;)x \,. \tag{3.23}$$

As it is pointed out in upper sections, best outputs can be achieved by finesse weight adjustment. In order to have better weights in hidden layer backpropagation made with,

$$w \leftarrow w + a\left(y - \hat{y}\;\right) x. \tag{3.24}$$

The term “a” in Eqs. 3.24 is representing the learning rate.

This section completed with “sklearn” library multi-layer perceptron regressor (MLPRegressor) method. Some of the parameter explained in ANN section above, rest of the tuned parameter are explained as; “solver” is the function that optimize weights in Eq. 3.24. “Alpha” is the normalization parameter. “Tol” is tolerance for the optimization. If the loss or score is not improved within given tolerance iteration is stops. “Random State” is the determines the random number for weights and bias initialization. When “Random State” set in specific number, it does maintain the inputs are used in ANN algorithm to train and test. Used parameter and values are shown in Table 3.2.

Table 3.2 : ANN optimized parameter used in MLPRegressor method.

Hidden Layer Size	Maximum Iteration	Activation Function	Tol	Random State	Solver	Alpha	Learning Rate
375	10000	Tanh	10^{-3}	7	lbfgs	10^{-4}	0.001

ANN trained with Kınık field data to make prediction on Kınık Coal system.

Out of the four activation functions available in the "Sklearn Library" for Artificial Neural Network (ANN) methods, the "tanh" function was specifically chosen for this study using the GridSearchCV optimization algorithm. This decision was made after a thorough examination of various activation functions, as it was observed that linear activation functions were not suitable for addressing the non-linear nature of coal gas content predictions. The preference for the "tanh" activation function stems from its continuity and differentiability within the range of -1 to 1, with a central value of 0. Compared to other activation functions, "tanh" exhibits a steeper and less restrictive behavior when transitioning between activated (1) and deactivated (-1) gradients (Sharma et. al., 2020). Additionally, the success of the "tanh" function can be attributed to the interdependence and mutual effects of the features within the neural network. These features often interact in complex ways while contributing to the connections between neurons. The distinct gradient characteristics of the "tanh" function enable a more comprehensive analysis of the information flow between neurons, resulting in more effective error handling and improved output accuracy (Akdaş and Fişne, 2023).

MLPRegressor method results with optimized parameter taken from Table 3.2 shown in Figure 3.11.

The used solver method in MLPRegressor is “Limited Memory Broyden-Fletcher - Goldfarb-Shanno Algorithm” (lbfgs). This algorithm is minimizing the cost function by using inverse Hessian matrix while the cost function is average of loss or error gathered by training dataset.

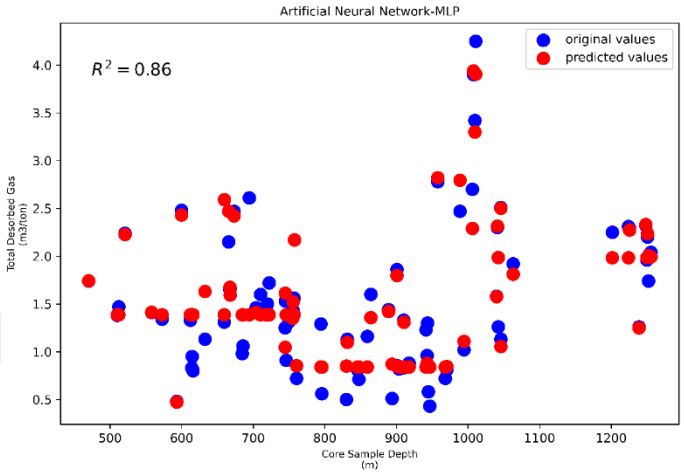

Figure 3.11 : The comparison of ANN prediction and original data obtained for kP1 and kM2 coal seams.

The Artificial Neural Network (ANN) achieved an average error of 15.67% when compared to the actual data collected from fieldwork. Similar to the supervised learning methods, the purpose was to gain a general understanding of the algorithm's applicability to gas desorption. As a result, no outlier detection was performed, and neither smoothing was applied to either the training or prediction data or the ANN algorithm. This approach was intended to provide an overall assessment of how well ANN can predict desorbed gas quantities in the Kınık Coalfield without any additional data processing. Error distribution of ANN algorithm is showed in Figure 3.12,

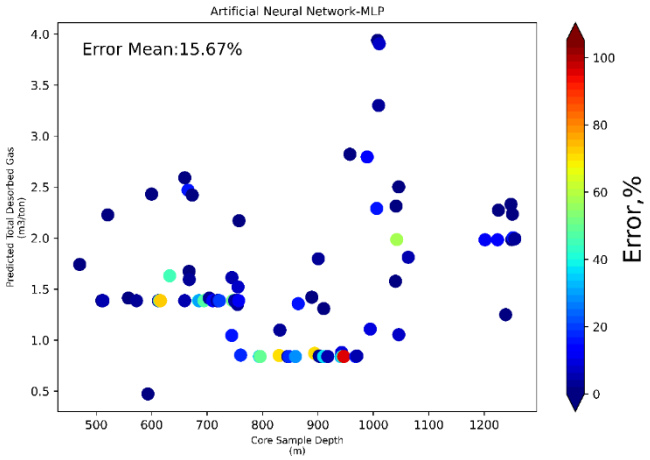

Figure 3.12 : Error distribution of ANN on Kınık coal field data prediction.

In this study, three different machine learning (ML) methods were considered, and for two of them, hyperparameter optimization was necessary. The optimization process involved the configuration of various parameters, including data point assessments, neighborhood functions, distance metrics between data points, tolerance values for weight optimization, activation functions for modeling relationships, and the number of iterations required to achieve the final model setup. To carry out this hyperparameter optimization, the GridSearchCV algorithm was employed. This algorithm systematically assesses the model's performance using a range of hyperparameter combinations. It calculates the model's score during the training phase and identifies the most suitable combination of hyperparameters for the chosen ML methods. The algorithm explores all provided hyperparameter combinations and selects the one that produces the best performance. This optimal combination is then used to build the actual model and make predictions on the target variables.

3.2 Verification of ML Resutls with Experimental Results

This research utilized the Kınık Coalfield dataset, comprising 84 data points related to five distinct features (Depth, Moisture, Ash, Volatile Matter, and Calorific Value), resulting in a total of 420 data entries. In section 3.1, machine learning algorithms were applied, with 10% of the data allocated for testing and the remaining 90% utilized for training (Akdaş and Fişne, 2023). It is equivalent to 9 different depth and 45 data points for testing out of 420 data. Used test and training data points showed in Figure 3.13.

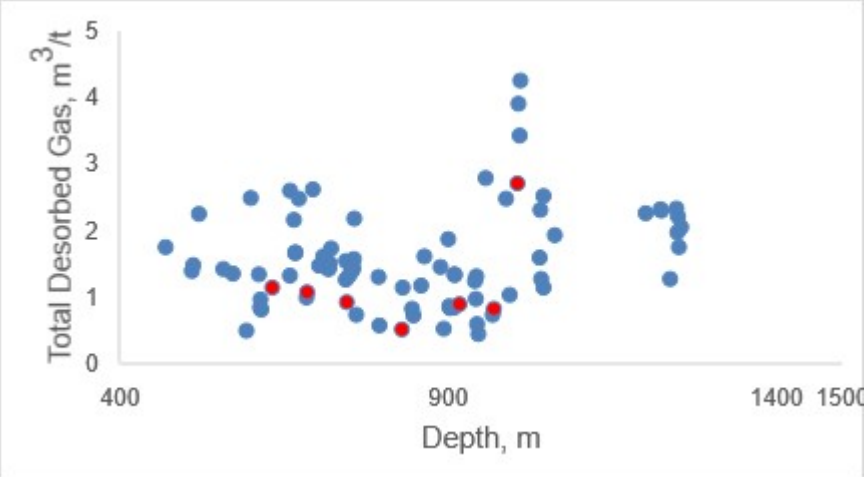

Figure 3.13 : Kınık Coalfield data point for Depth vs Total Desorbed gas.

In Figure 3.13, Red dots shows test data (9 different depth, in total 45 points for 5 different features) and blue dots shows the training data set. Test dataset is showing in Table 3.3 in below.

Table 3.3 : Kınık Coalfield test data.

Index	Depth, m	Moisture, %	Ash, %	Volatile Matter, %	Calorific Value, MJ/kg	Total Desorbed Gas, m^3/t
12	632.6	16.69	24.09	28.06	16.65	1.13
20	685.7	14	39.31	38.16	6.23	1.06
30	746	14.68	33.94	36.15	9.64	0.91
32	753.7	11.41	30.71	31.8	13.75	1.43
40	830.1	14.39	42.35	33.64	6.74	0.5
52	910.2	15.9	18.7	34.7	17.68	1.33
54	917.8	15.53	26.55	39.8	11.75	0.88
62	970.3	8.33	43.57	47.4	3.85	0.81
65	1006	9.65	10.1	25.72	15.22	2.7

Models created with Kınık field data and response withs predictions displayed in following figures.

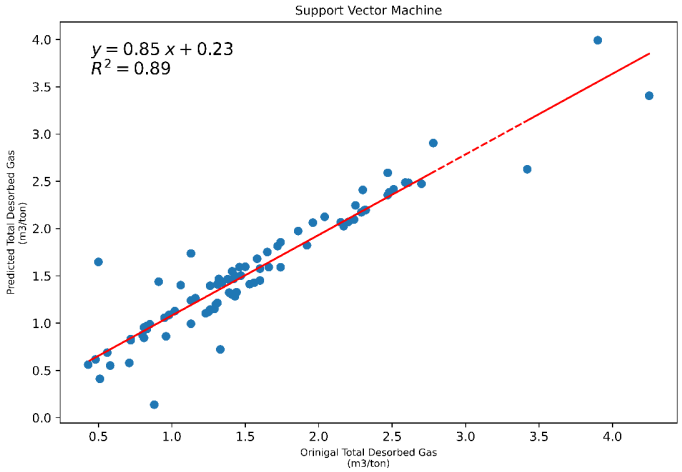

Figure 3.14 : Kınık field original and predicted Total Desorbed Gas fitting on SVM with R^2.

This chapter is dedicated to see which ML algorithm is genuinely gives better results in specifically Kınık coalfield gas content prediction. First ML model is Support Vector Machine. SVM working principles was briefly mentioned above. Since it is known that SVM is trying to find a correct path to train given system, it was worked with high percentage of training dataset smoothly handled by SVM. It is resulted to have best agreement by SVM amongst to other two method. The conspicuous detail in this method is having “RBF” kernel and high regularization parameter “C=325”. According to RBF in radial base of each training set, data points were carefully investigated by the algorithm. Then, model is strongly processed with high regularization parameter which has caused extremely small penalties while the model was less effected by outliers and parameters that has weak contribution to model training. Overall, Figure 3.14 showed that prediction capacity of SVM in response to original field data which are taken by USBM methodology, is satisfactory. Overall, Figure 3.14 illustrates that SVM's prediction capacity in response to the original field data, obtained through the USBM methodology, is satisfactory.

The second method employed is Multiple Linear Regression (MLR). It's the oldest and simplest approach for comparing two data sets with the same quantity. Figure 3.15 illustrates that MLR is strongly influenced by the uniqueness of the input parameter features. Unlike other algorithms used in this study, the key difference with MLR is that Linear Regression doesn't employ any normalization or penalties to prevent the training process from going off course. Consequently, MLR obtained the lowest " R^2" score among the methods evaluated.

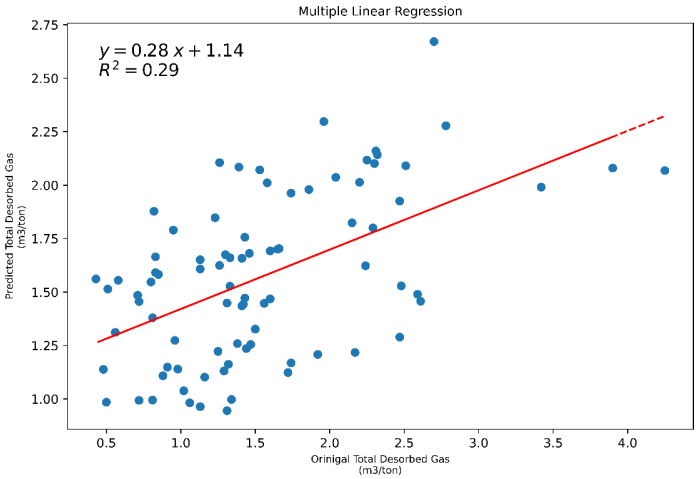

Figure 3.15 : Kınık field original and predicted Total Desorbed Gas fitting on LR with R^2.

The final method employed is Artificial Neural Network (ANN). In comparison to other methods, ANN is relatively like a closed box. The relationships between hidden layers, neurons, and nodes are well-controlled through the use of activation functions, solvers, regularization parameters, learning rates, and other techniques. However, one point of contention with ANN systems is the computational cost involved in their operation.

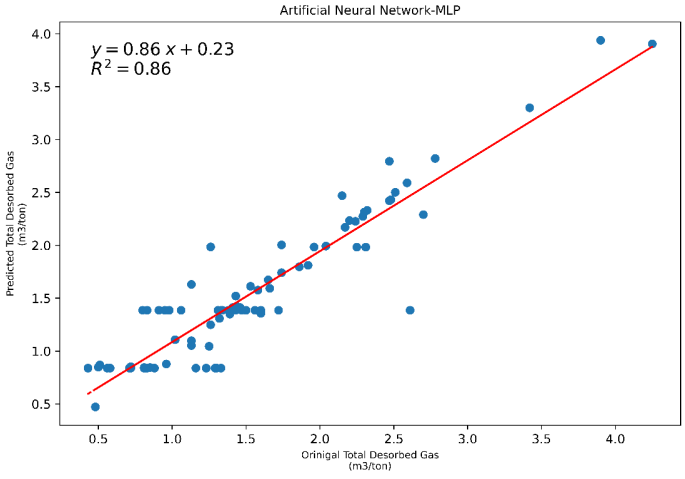

Figure 3.16 : Kınık field original and predicted Total Desorbed Gas fitting on ANN with R^2.

ANN with Multilayer Perceptron was used to predict Total desorbed gas in response to Kınık coalfield data taken from Esen (2021) PhD. dissertation. Figure 3.16 is showing to original and predicted data agreement.

The Neural Network algorithm yielded good agreement in its predictions. While Multiple Linear Regression (MLR) showed a weak connection between data points, and Support Vector Machine (SVM) provided results that were relatively close, the success of the Multilayer Perceptron method can be attributed to several key parameters. These parameters include a high iteration capacity (max_iter=10,000), a large number of hidden layer nodes (hidden_layer_sizes=375), a small regularization parameter (alpha=0.0001, which works inversely to the "C" parameter in SVM), the use of the tanh activation function, the lbfgs solver, and the consistent training and test categories (random_state=7).

After the model's performance showed in Figs. 3.14-15-16, it is desired to check the methods performance regarding to statistical perspective. To start with the discussion of the R^2 score, it's essential to understand the concept of variance. Variance is a measure of the distance between the observed values and the mean of the predicted values. Consequently, the R^2 score represents the magnitude of the variance of target values in relation to input values. It's used in regression models to assess the strength of the relationship between two sets of variables, and its values typically fall within the range of 0 to 1.

In this context, if the R^2 score reaches 1, it signifies an identical relationship between the two variables being analyzed. As mentioned earlier, SVM has been found to provide the strongest relationship between the field and predicted data sets, as reflected by its R^2 score. It is compared in Figure 3.17 in below.

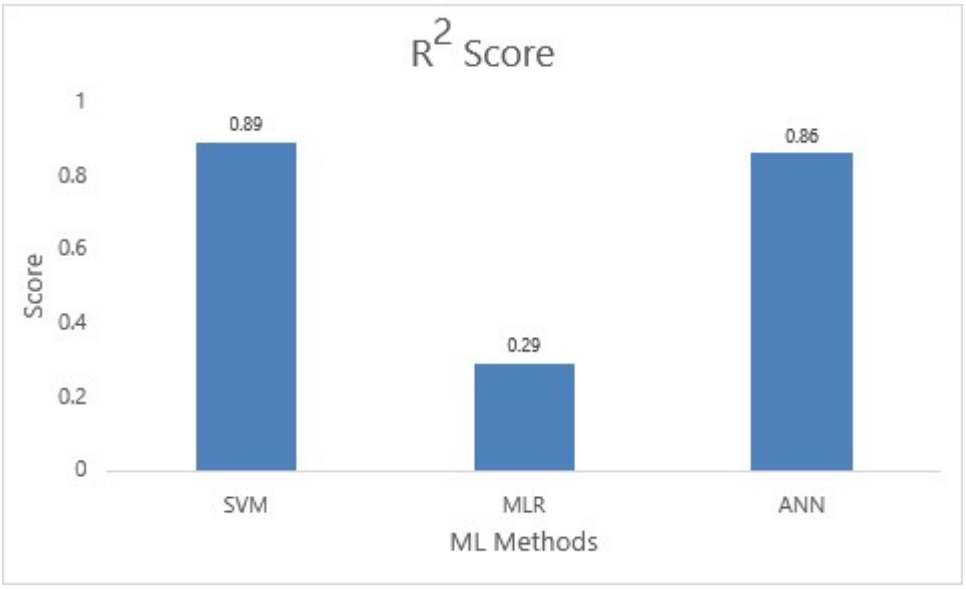

Figure 3.17 : Investigated ML methods prediction R^2 score regarding to field data.

In general, R^2 is not evaluated by itself. In support to two data set agreement, alternative relationships must be considered. Second criterion is Mean Squared Error (MSE). The Mean Squared Error (MSE) is calculated by taking the squared differences between observed and predicted values, summing these squared differences, and then dividing the sum by the number of observed and predicted value pairs. It is a metric used to assess the stability of a system, with the aim of achieving a small MSE, ideally close to 0, which represents a perfect match in terms of MSE comparison.

In Figure 3.18, three machine learning methods are compared in terms of MSE. Similar to the R^2 score, Support Vector Machine (SVM) yielded the lowest MSE value, indicating its strong performance in minimizing prediction errors. The Artificial Neural Network (ANN) model followed SVM with the second-lowest MSE value.

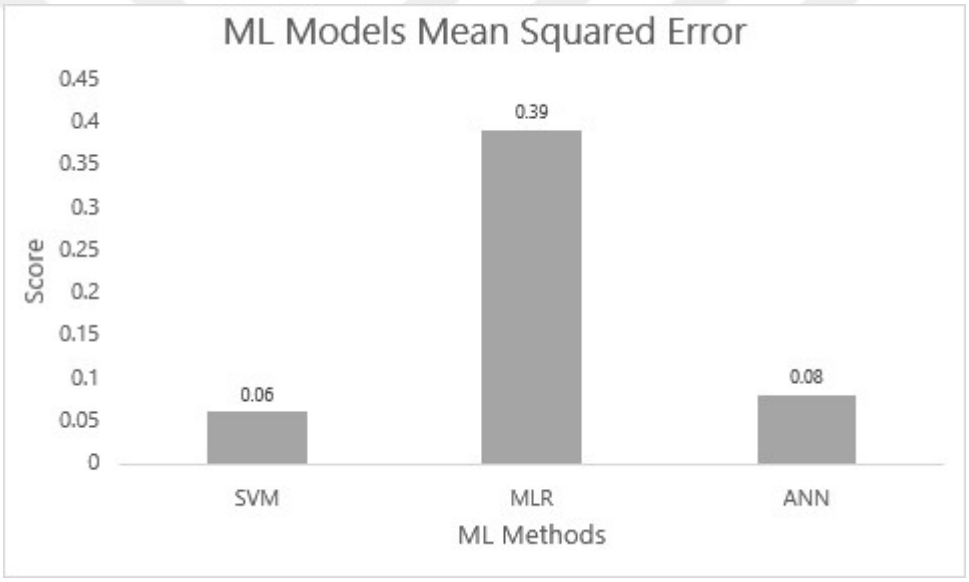

Figure 3.18 : Investigated ML methods mean squared errors based on field data.

In the Machine Learning ecosystem, the performance of models is typically evaluated based on their training quality. While the R^2 and MSE scores can indicate that a specific method provides satisfactory results in comparison to a particular target, it doesn't necessarily guarantee that the same approach will work equally well with a different set of test samples in a different section of the investigation. Therefore, a method that achieves the highest training score with statistical evidence of good agreement, such as high R^2 and low MSE, may be extended to create a system for further predictions in a designed system.

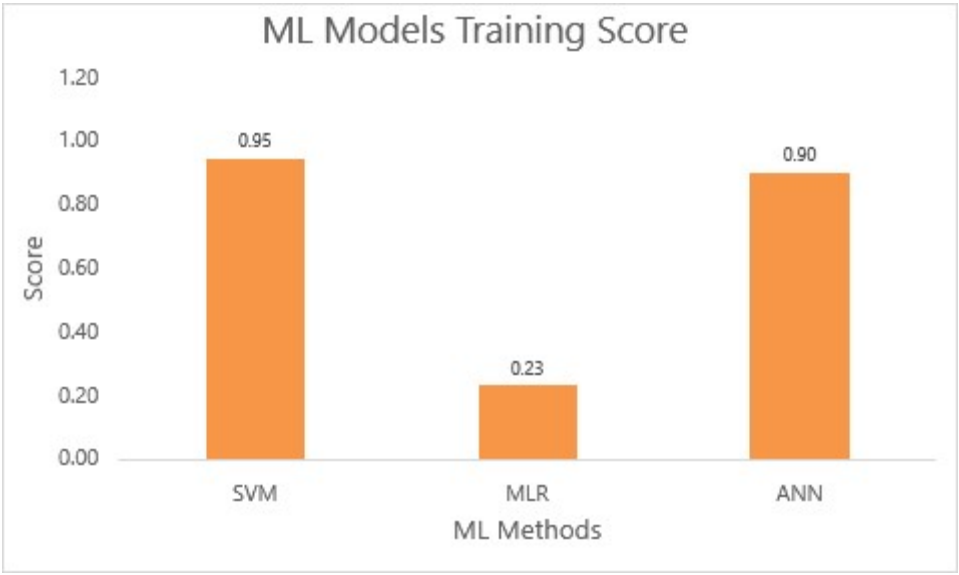

Figure 3.19 : Investigated ML methods training scores based on field data.

The training scores of the three Machine Learning methods were compared in Figure 3.19. It is evident that Support Vector Machine (SVM) offers the best training capability among the ML methods used. Furthermore, Artificial Neural Network (ANN) closely follows SVM's training score, consistent with the findings regarding R^2 and MSE scores. Each of the three comparisons, whether in terms of R^2, MSE, or training scores, reinforces the overall integrity of the system. Therefore, the reliability of SVM methods for predicting desorbed gas in the Kınık field has become evident.

4. GAS CONTENT PREDICTION PERFORMANCE of THE INPUT FEATURES

4.1 Effect of Input Features

Predicting coalbed gas content is a challenging problem due to the heterogeneity of coal seams and various factors discussed in geological system and assumptions. This study aims to develop an alternative approach to augment some of the required data collection for mining activities. Given the uniqueness of each coal sample and the impact of features associated with coal, the goal is to improve the accuracy of coal gas content prediction. To achieve this, the Kınık Coalfield dataset was examined, considering features such as Depth, Moisture, Ash, Volatile Matter, and Calorific Values.

This section is dedicated to identifying the features that have an impact on gas content prediction. In the literature, there are various findings regarding the specific features that influence the amount of produced gas, and it also depends on the combination of these features. In the Machine Learning (ML) algorithms used, the features are individually processed in relation to the "Total Desorbed Gas Content." Subsequently, the results for regression scores (R^2), model training scores, mean squared error (MSE), mean absolute error (MAE), and mean absolute error percentage will be presented in the following sections.

This section of the study was carried out using SVM methods due to their superior performance in the field investigation, as demonstrated in the previous chapter. This is clearly related to the best R^2, MSE, and training scores achieved with SVM. Additionally, SVM offers the advantage of lower computational cost when compared to Multilayer Perceptron. Therefore, the analysis of coal sample features for predicting desorbed gas in the Kınık field will be conducted using Support Vector Machine Regression.

First, the training scores will be evaluated. As Chapter 3 has made clear, having better statistical scores is not useful if there is no reliable training process. The training scores for the five features are shown in Figure 4.1.

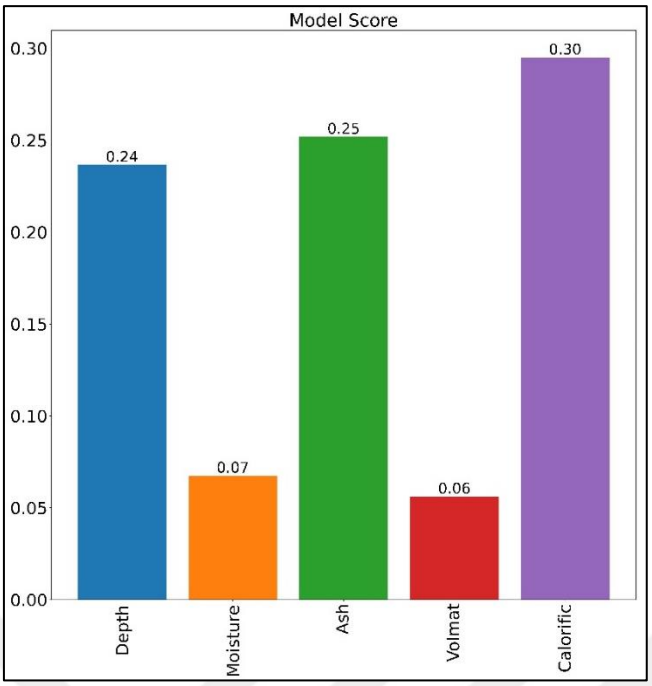

Figure 4.1 : Kınık Coalfield Training Scores of Features in response to Total Desorbed Gas prediction.

Figure 4.1 illustrates the training scores of the coal sample features. It shows that the most reliable parameter in terms of SVM training is Calorific value. This is because the calorific value determines the rank of the coal. High-ranking coals have a better affinity for retaining carbon material. Therefore, a high calorific value has a strong relationship with better coal and a higher carbonaceous gas content. On the other hand, the least reliable parameter is Moisture. Coal does not have a strong affinity for retaining hydrogen or oxygen except during the burning process. Thus, there is no strong relationship between the porosity of coal and moisture content.

The second unreliable parameter is Volatile Matter. It has been observed that volatile matter behaves similarly to methane during the coal formation process. Both molecules tend to rise to lower pressurized sections, initially the atmosphere. Permeability systems then allow low-viscosity fluids such as gaseous matter (volatile matter and methane) to change their path to less pressurized cap rock systems. Therefore, the distribution of light molecules changes twice compared to relatively low-mobility features like moisture and ash content. This difference in light molecule mobility in the geological system challenges the linear analogy that deeper reservoirs should have more gas content. The simple logic related to depth does not apply in this case. The traditional concept that deeper reservoirs are better in gas reservoir modeling is challenged by the mobility issue. The main difference between natural gas reservoirs

and coalfield gas content lies in the multiple features of coal. Coal does not behave like natural gas reservoirs.

The second analysis involved comparing the R^2 scores of the investigated features in relation to the prediction and experimental (USBM) data for the desorbed gas content of Kınık coalfield.

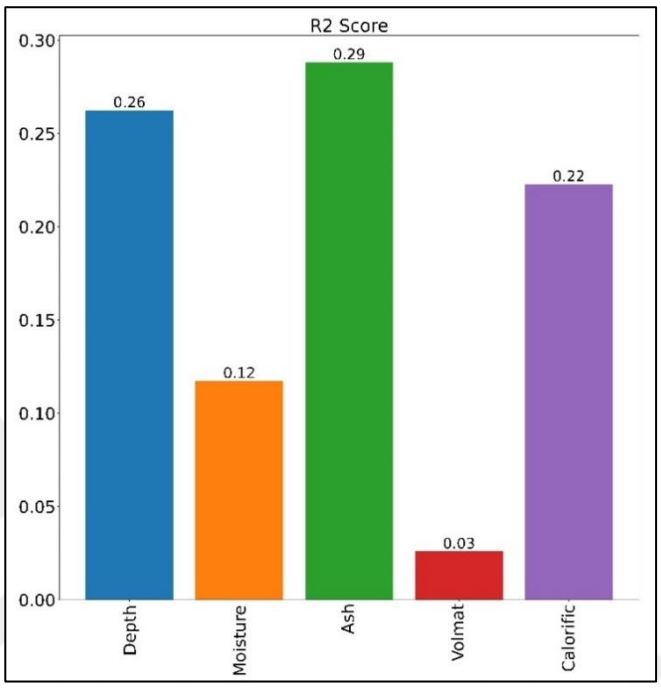

Figure 4.2 : Kınık Coalfield R^2 Scores of Features in response to Total Desorbed Gas prediction.

Comparison of the regression scores (R^2) for the features is shown in Figure 4.2. The best training score had created an expectation of better regression scores for Calorific value. However, it was found that Ash content dominates the prediction of total desorbed gas. This suggests that when ash content covers most of the porous media, the prediction is shifted towards Ash content, overshadowing the influence of Calorific value.

Moisture and volatile matter exhibit the expected prediction performance by the regression score themselves. Interestingly, Depth also has a significant influence on the total desorbed gas prediction, with the second-best regression score. The uniqueness of coal reservoir permeability and untouched coal reserves in the deepest part of the coal seams plays a role in the gas prediction analysis in terms of regression score.

In general, the deeper parts of coal seams contain more gas compared to the shallow sections of coal reserves. However, this does not hold true for every coal structure, especially those with double and continuous permeability. The Calorific value can be generalized across the reservoir, but depth cannot. The coal formation contributes to the overall reservoir with similar parameters, but the burial and accumulation of fresh coal samples change the entrapping of gas molecules. Therefore, depth plays a significant role in coalfield gas contents.

The error analyses for the effect of coal features on total desorbed gas content in the Kınık coalfield are presented in Figure 4.3. This analysis provides a different perspective on the error evaluation and helps understand the variations and accuracy of predictions for each feature.

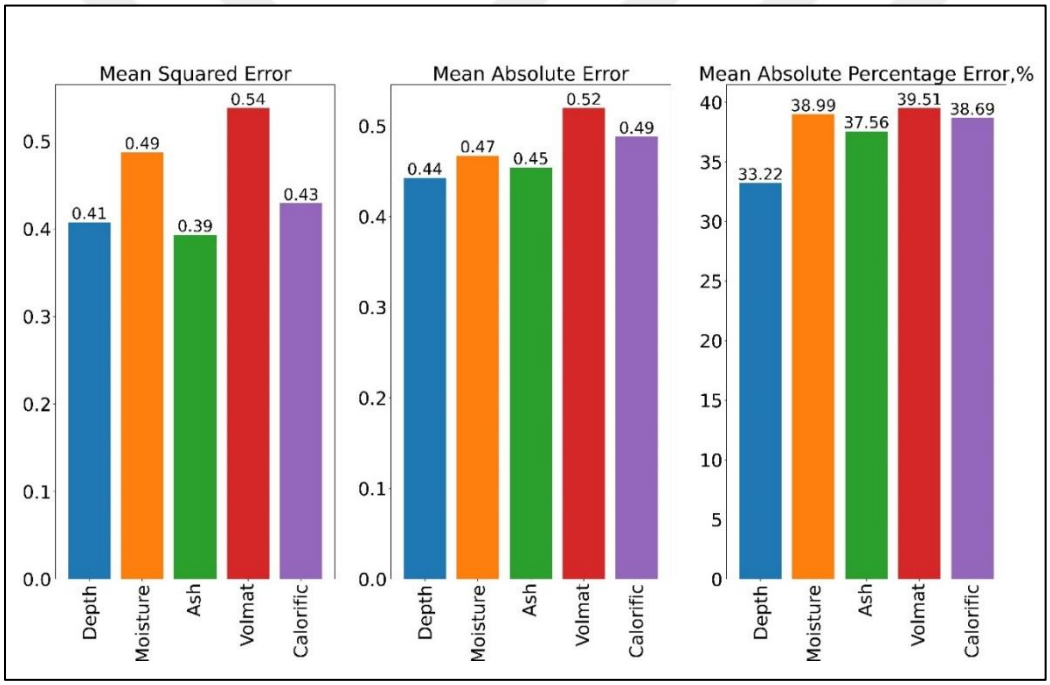

Figure 4.3 : Kınık Coalfield Error Scores of Features in response to Total Desorbed Gas prediction.

The mean squared error (MSE) was assessed, and the results were consistent with the R^2 score. A successful regression analysis was expected to have lower error values, which is reflected in Figure 4.3. The Ash content feature had the smallest error fraction, corresponding to its high position in the regression score. The Depth feature had the second lowest error, and its position was reversed compared to the regression chart.

After MSE, the study investigated absolute error in the second bar chart in Figure 4.3. Then, the mean absolute error (MAE) was considered to assess the amount of error

encountered in each feature analysis in relation to total desorbed gas content. The analysis revealed that Ash content had more error compared to Depth analysis.

In the last bar chart of Figure 4.3, Mean Absolute Error Percentage (MAEP) is examined to assess the relationship between the mean absolute error (MAE) and the data. Both analyses complemented each other. In conclusion, the regression analysis and its verification through MSE showed that the strongest relationship to the prediction of total desorbed gas (TDG) in Kınık coalfield is related to Ash content. Meanwhile, Depth was found to be the most accurate feature for TDG prediction at a specific point when comparing the absolute error analyses.

In this section, sensitivity analyses of the features were conducted to assess the contribution of each feature to the overall prediction, as described in Chapter 3. The comparison of regression scores and mean squared errors between the SVM analysis using all five features (R^2: 0.89, MSE: 0.06) and the best individual feature scores (e.g., Ash content with R^2: 0.29, MSE: 0.39) highlights the importance of considering all features together for a more accurate prediction.

4.2 Cost Analysis of ML Techniques

Machine learning has become a widely used and practical approach for analyzing data, making predictions, and establishing relationships between seemingly unrelated features. This is largely due to the tremendous increase in computational power over the years. Problems that were once challenging due to the large volumes of random data, numerous categories and features, and the need for continuous input have become more manageable by feeding algorithms with as much data as possible.

It is desired to show that, if ML methods run with the specified parameters for the computational cost evaluation, result showing in below,

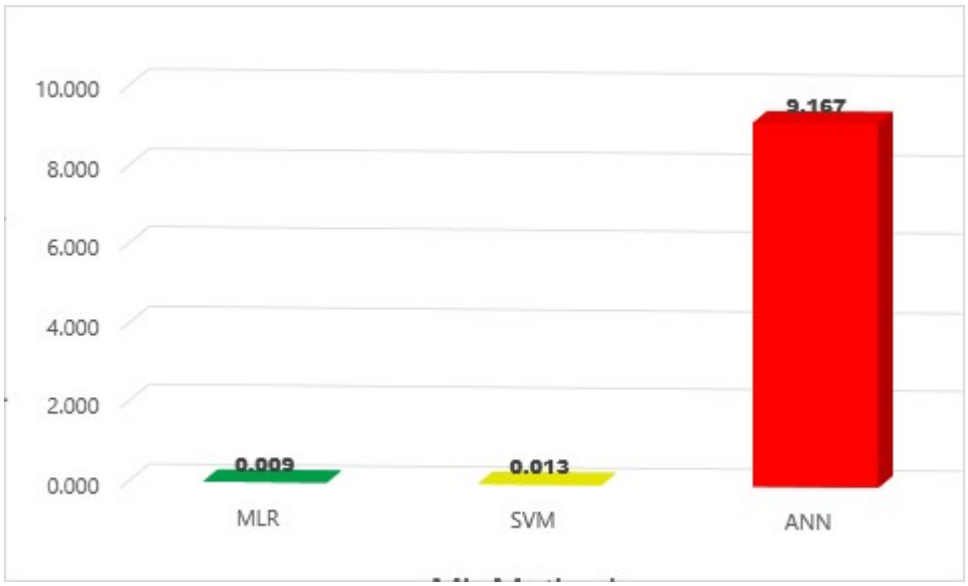

Figure 4.4 : Three ML method computational cost evaluation.

Figure 4.4 demonstrates that the Artificial Neural Network (ANN) method is highly computationally inefficient. This is because ANN employs backpropagation through hidden layers, nodes, and neurons to enhance the accuracy of results, resulting in significantly increased time and energy consumption to obtain valid outputs.

SVM has demonstrated superior performance in regression and training scores for predicting Kınık coalfield data. It also exhibited the second-best computational efficiency among the methods used. However, when compared to Linear Regression (LR), SVM is three times (3X) more computationally expensive. LR, while less computationally expensive, did not provide acceptable results for this specific field. Consequently, the comparison was between SVM and ANN. In this comparison, SVM is 705 times (705X) faster or less computationally expensive than ANN while still providing reasonable results for Kınık coalfield prediction.

5. TEST SAMPLE ANALYSIS

5.1 Prediction Capacity of Optimized ML for Test Samples

The study of Kınık coalfield involved the application of machine learning methods to assess various features and their predictive capabilities. These analyses were conducted within specific constraints, including limited field data and a limited number of features. Under these conditions, the best performance was achieved using Support Vector Machine (SVM). One of the key distinctions between SVM and other methods is its suitability for small data sets.

While SVM proved to be the best and faster method for interpreting the Kınık coalfield data set, it's generally not ideal to have almost the same data set for both training and testing. It's essential to emphasize that without a thorough analysis of the entire data set, the performance of individual predictions will be weak, similar to the situation where the entire data set is used for both training and testing. To address this concern, the "train_test_split" function from the "Sklearn" library was employed to split the 84 data samples taken from kP1 and kM2 coal seams. This division into training and testing data sets helps ensure robust model evaluation.

In the previous chapters, a division of 10% of the data set was used for testing, and 90% of the data were utilized for training. Figure 3.13 and Table 3.3 presented these data. All the machine learning training analyses were conducted without considering these test data. Subsequently, predictions were made using the entire data set.

The intention behind selecting the test data was to minimize the impact on the training system by excluding less predictable data points while retaining the most predictable ones. As a result, the test data set was chosen from the most densely populated sections in the depth and total desorbed gas distribution, as illustrated in Figure 3.13. These selected test data points were further processed for analysis.

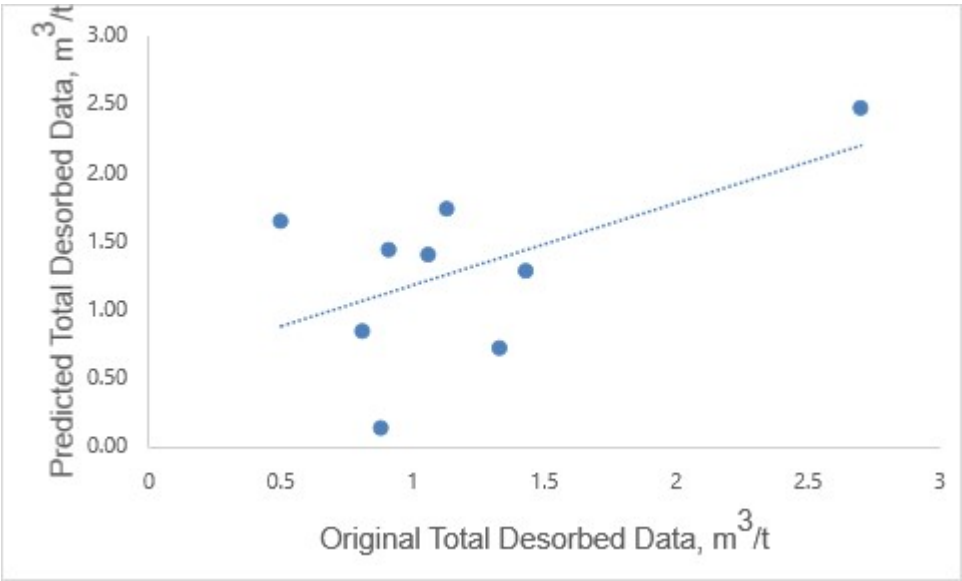

Figure 5.1 : Test Data Set prediction from 75 Training Data.

As shown in Figure 5.1, the prediction performance significantly deteriorated when using only 75 training data points from the entire data set. Furthermore, a single assessment of each desired point to obtain Total Desorbed Gas content generally resulted in high absolute errors. These comparisons are detailed in Table 5.1.

Table 5.1 : Test Data Prediction Results.

Index	Depth, m	Total Desorbed Gas, m^3/t	Predicted Total Desorbed Gas, m^3/t	Absolute Error, %
12	632.6	1.13	1.74	53.77
20	685.65	1.06	1.4	32.34
30	746.03	0.91	1.44	58.13
32	753.65	1.43	1.28	10.18
40	830.1	0.5	1.65	229.55
52	910.2	1.33	0.72	45.73
54	917.75	0.88	0.14	84.33
62	970.3	0.81	0.84	4.32
65	1006.1	2.7	2.47	8.38

At the outset of this study, it was emphasized that a reliable tool must be developed to achieve accurate predictions of gas content in coal seams. Therefore, the designed ML prediction algorithm was applied to predict individual data points in the same test data set. In this scenario, 83 training data points were used out of the 84 available to predict

a single data point. The results of single data point predictions are presented in Figure 5.2.

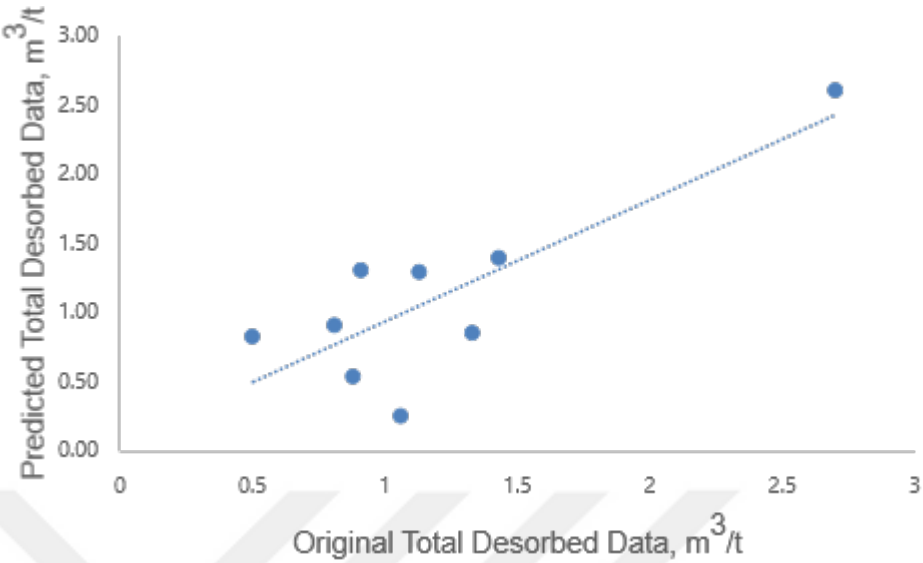

Figure 5.2 : Test Data Set prediction from 83 Training Data.

The analysis began with 84 data points, which were then divided into a 10% test data set and a 90% training data set. The SVM method was applied to define Kınık coalfield based on the 10% test data. Following the field definition, the test data was predicted at the same time, but the results were unsatisfactory. Subsequently, 9 data points in the test data set were individually removed and predicted based on the remaining data from the entire Kınık coalfield dataset. This process doubled the " R^2 score" by performing single-point predictions within the Kınık coalfield data set. Figure 5.2 was extended with the absolute error of each single data point prediction, as shown below,

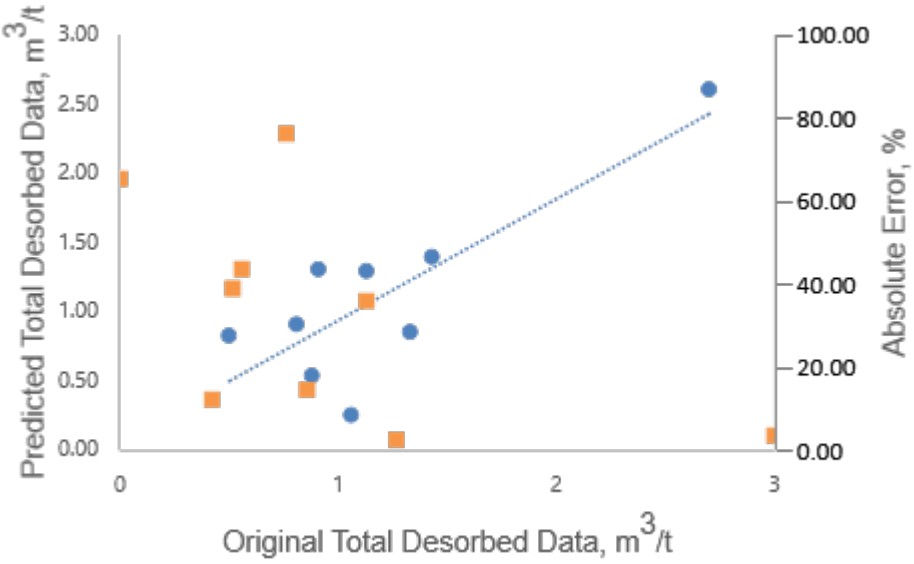

Figure 5.3 : Test Data Set prediction from 83 Training Data and absolute error analysis between original and individually predicted data points.

5.2 Additional Field Data Implementation into ML Environment

Machine learning is a versatile tool with a wide range of applications that can be extended to address various problems. In this study, the goal is to provide a local solution for predicting coal sample gas content in the Kınık coalfield. While the initial analysis was conducted with 84 data points from different locations and depths, it is now desired to expand the dataset with additional data points. The new data points will be used to further train the SVM model and enhance its ability to make predictions based on different coal sections and types.

The new dataset includes 10 different sections and various types of coal samples. However, for the new entries, the "Calorific Value" feature was excluded due to missing information about these data points. In total, there will be 368 data points and 4 features, resulting in 1472 data points to predict one at a time, similar to the approach in Chapter 5.1. The purpose of this section is to assess whether the inclusion of this new data improves the prediction accuracy of the test dataset as discussed in the previous subchapter.

The absence of one feature and the increase in the total number of data points from 420 to 1472 have led to the following outcomes,

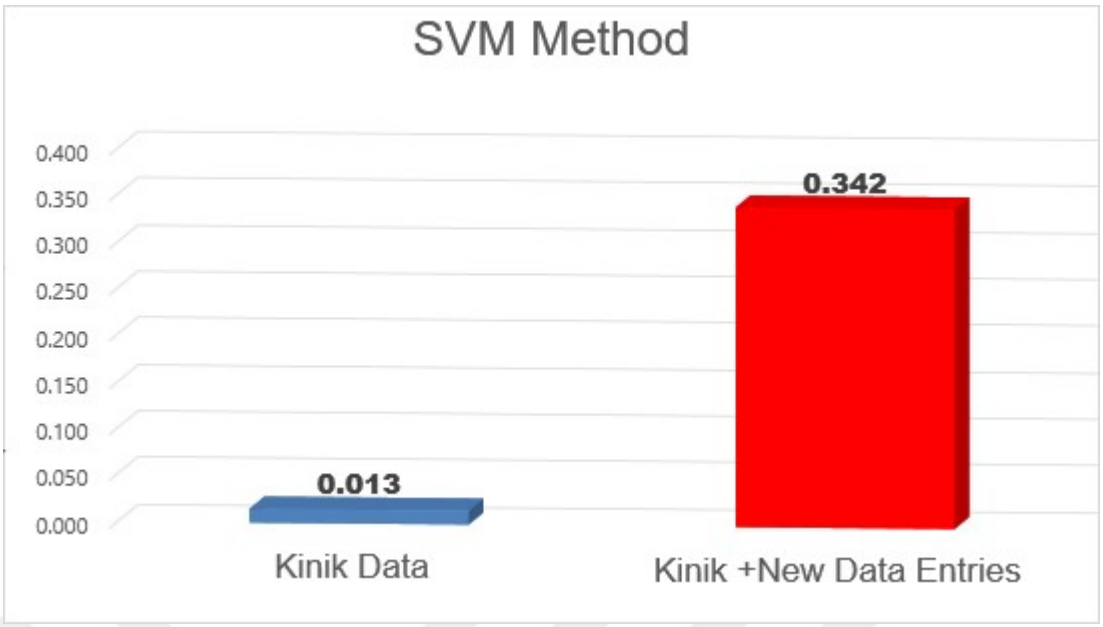

Figure 5.4 : Kınık coalfield dataset and additional data set computational cost comparison in respect to SVM analysis.

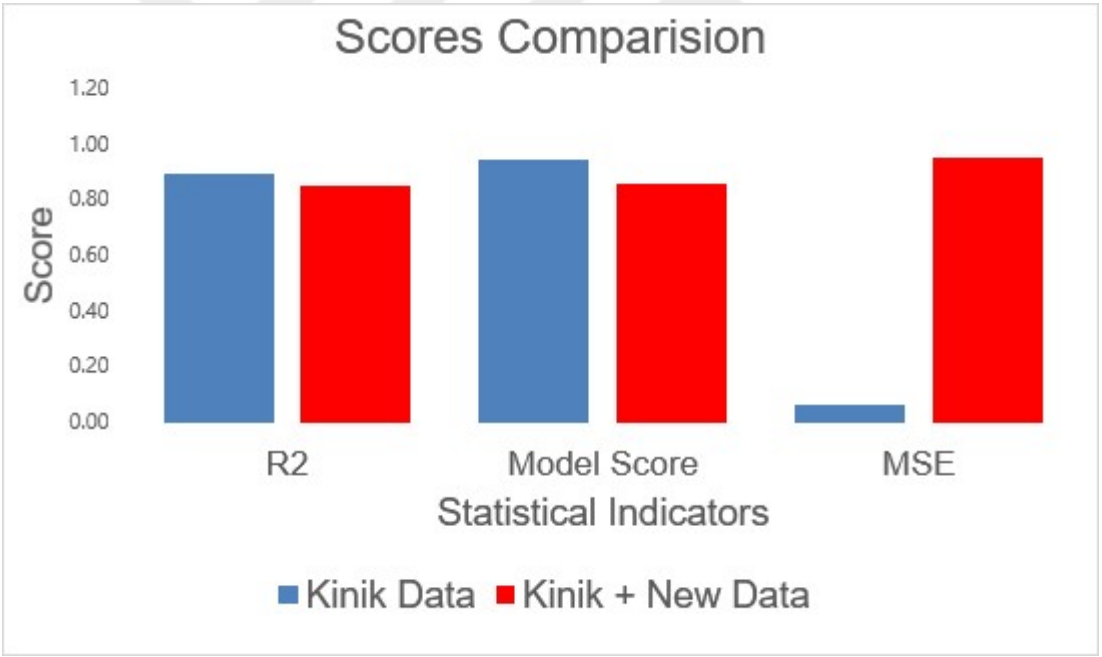

Figure 5.5 : Kınık Coalfield Data set and New Data set statistical scores comparison.

It is apparent that the computational cost has increased with the larger data set compared to the original Kınık coalfield data set, as illustrated in Figure 5.4. Furthermore, Figure 5.5 reveals that the statistical scores of the data set prediction and training performance have decreased, while the mean squared error has increased. The introduction of new data and coal samples from various fields and different coal ranks has led to less successful results in terms of analyzing the field data sets. It's important

to emphasize that the main purpose of this section was to examine the advantages and disadvantages of a larger data set.

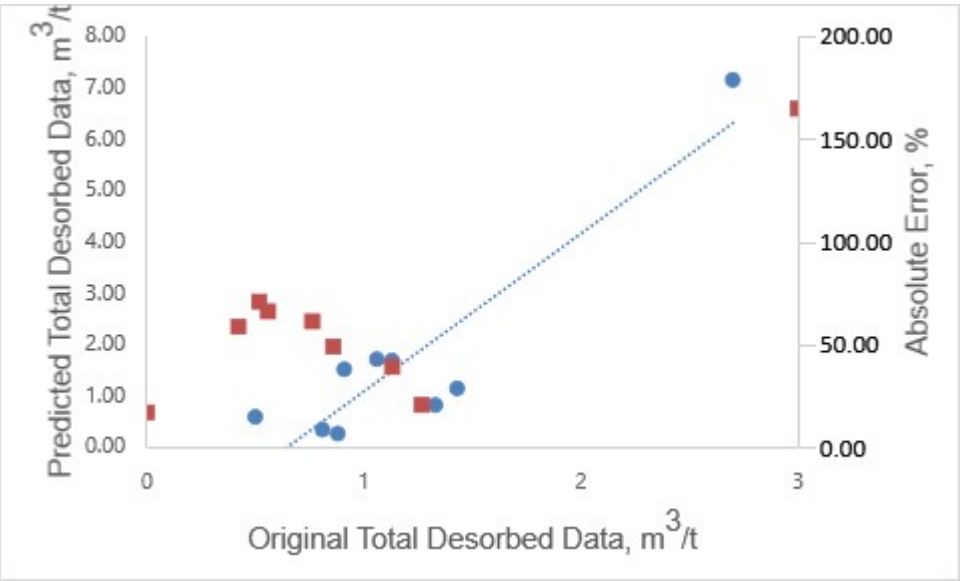

Figure 5.6 : New Data set prediction performance on Test Data set taken from Kınık coalfield.

In Figure 5.6, the test data taken from Kınık coalfield were individually processed, and the results indicated that the larger data set exhibited better agreement in terms of the "R^2 score". However, the mean absolute error of the test data increased. The main reasons for both field data analysis and error problems were primarily attributed to the absence of one feature, namely "calorific value". Additionally, various factors, including different locations, various coal types, and the geological characteristics of different coal sections, had an impact on the results. Nevertheless, the overarching goal was to process similar data and create a framework for coal gas content prediction. Despite the challenges posed by the addition of more data, even with weak relationships to the original field data, this integration was successfully implemented within the created ML environment.

6. SUMMARY and CONCLUSIONS

6.1 Summary and Conclusion

The desorbed gas originating from coal beds is a critical factor in both coal mining and subsequent operations in underground exploitation. Whether it's underground coal mining or the use of gas for production purposes, particularly methane and carbon dioxide storage, the gas content of the coal matrix plays a vital role. In essence, coal cannot be properly assessed without taking into account its gas content, and it is essential to determine this gas content as thoroughly as possible before proceeding with any further actions.

In this study, the focus was on investigating machine learning methods to find the most suitable approach for predicting coalbed gas content. As discussed in earlier chapters, gas content research is time-consuming and costly. To expedite the prediction process and reduce ineffective methods, computational power has become a viable solution over the past two decades. The primary goal of this study was to serve as a guide for classifying the type of prediction problem and selecting the most appropriate machine learning technique.

To achieve this goal, three popular machine learning methods were chosen based on conventional categories within machine learning, and they are among the most widely used methods. These methods include Support Vector Machine, Linear Regression, and Neural Network. The performance of these methods was verified using data from the desorption analysis of Kınık Coalfield. The verifications were conducted by comparing the machine learning predictions with field and laboratory results obtained using the USBM method.

Among these three methods, the comparison was based on the statistical scores generated for the Kınık Coalfield data. It was found that the Support Vector Machine (SVM) produced the best prediction and training scores. Furthermore, a cost-efficiency analysis was conducted, which confirmed that SVM also outperformed the other methods in terms of computational cost. This makes SVM the most suitable alternative method for predicting coalbed gas content.

Based on the findings of this thesis study, the following conclusions can be drawn:

- Coal gas, or coalbed desorbed gas content, is fundamental to the underground coal mining system, and its accurate prediction is essential.
- Coal seams exhibit linear distributions between formation layers, but the gas content within coal matrices does not have a uniform distribution.
- The nonlinear distribution of coal matrix properties makes the prediction of gas content exceptionally challenging, necessitating a significant number of core drilling operations and gas content analyses.
- Advanced computational power and powerful machine learning methods have proven to be valuable tools for analyzing complex problems like gas content prediction in coalfields.
- Gas content determination in underground coalfields is complex but relevant, and there are relationships between coal samples, depth, geological aspects, and gas content despite the intricate nature of the dataset.
- Conventional approaches are insufficient to address the complexity of the problem, especially considering the uniqueness of each coalfield.
- Numerical analyses offer promising results as they can adapt to different coalfields.
- Machine Learning is a dynamic and evolving field that continuously adapts to various challenges. Teaching machines to find solutions to complex problems simplifies the problem-solving process.
- Sensitivity analysis was performed to determine which features contributed the most to overall prediction accuracy.
- It was demonstrated that increasing the amount of input data leads to better outcomes with machine learning algorithms.

6.2 Recommendations for Future Works

It's important to outline potential directions for future research that can improve upon the predictive models and analyses presented in this study. Future studies in this field might explore the following suggestions:

• Machine Learning has the potential to be expanded to model the 3D distribution of gas within coal seams through improved prediction methods.

• Enhancing Neural Network Analysis can be achieved by optimizing the preprocessing of nonlinear data sets, making the analysis more efficient and faster.

• To broaden the scope of gas content prediction in various coalfields, it is essential to incorporate more data into the Machine Learning environment.

• The cost optimization process described in Chapter 4 was based on selecting the best parameters obtained from individual analyses. To simplify the parameter optimization of Machine Learning algorithms, the use of optimization tools like Genetic Algorithms can be considered. This approach enables the optimization of all parameters in a single run, leading to a more automated prediction system.

• This study has encompassed both supervised and unsupervised approaches. However, there is a more intricate and dynamic approach within the domain of Machine Learning known as reinforcement learning. In cases where gas content prediction lacks a clear intermolecular relationship with the features, aside from "Depth," reinforcement learning would have been the sole choice to implement in order to utilize Machine Learning effectively in this study. Nevertheless, there may be scenarios where feature data is limited, and thus, the prediction of underground systems will require a self-adaptive mechanism to tackle varying challenges within different geological layers. Therefore, new methods should be developed using reinforcement learning to enhance control over underground parameters.

REFERENCES

Akdaş, S. B., & Fişne, A. (2023). A data-driven approach for the prediction of coal seam gas content using machine learning techniques. *Applied Energy*, 347, 121499, https://doi.org/10.1016/j.apenergy.2023.121499.

Beaton, A., Langenberg, W., & Pană, C. (2006). Coalbed methane resources and reservoir characteristics from the Alberta Plains, Canada. *International Journal of Coal Geology, 65*(1-2), 93-113. doi:https://doi.org/10.1016/j.coal.2005.04.013.

Belyadi, H., & Haghighat, A. (2021). *Machine Learning Guide for Oil and Gas Using Python: A Step-by-Step Breakdown with Data, Algorithms, Codes, and Applications.* Gulf Professional Publishing, https://doi.org/10.1016/B978-0-12-821929-4.00001-9.

Ben-Hur, A., & Weston, J. (2010). A User's Guide to Support Vector Machines. *Clifton, N.J.: Methods in molecular biology*. Humana Press, 609,223-239,

Diamond, W., & Levine, J. (1981). *Direct Method Determination of the Gas Content of Coal.* Washington, D.C.: United States Department of the Interior, Bureau of Mines, 8515.

EIA. *Coal explained*. eia.gov: Retrieved, November 15, 2021, from https://www.eia.gov/energyexplained/coal/how-much-coal-is-left.php.

EIA. *Monthly Energy Review:Coal. 116*. Washington, DC, USA: U.S. Energy Information Administration. Retrieved, October 2021, from https://www.eia.gov/totalenergy/data/monthly/pdf/sec6.pdf.

Esen, O. (2021). *Soma Kömür Havzası Kömür Damarlarının Gaz İçeriği,Gaz Depolama Kapasitesi ve Gaz Akış Özelliklerinin Araştırılması*. (Doctoral dissertation) Retrieved from https://itu.edu.tr/.

Flach, P. (2012). *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge: Cambridge University Press.

Guan, C., Liu, S., Li, C., Wang, Y., & Zhao, Y. (2018). *The temperature effect on the methane and CO2 adsorption capacities of Illinois coal*. Fuel, 211, 241-250.

Gunn, S. (1998). *Support Vector Machines for Classification and Regression*. University of Southampton, ISIS technical report, 14(1), 5-16.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to Statistical Learning*. Springer, New York, Statistical Theory and Related Fields, 6(1), 87. https://doi.org/10.1080/24754269.2021.1980261.

Kononenko, I., & Kukar, M. (2007). *Machine Learning and Data Mining: Introduction to Principles and Algortihms*. Chichester,UK: Horwood Publishing.

Kor, K. (2021). *Penetration Rate Optimization in Heterogeneous Formations with Support Vector Machines Method*. (Doctoral dissertation) Retrieved from https://itu.edu.tr/.

Leleoglu, S. (2020). *Kömür ve Enerji Raporu*. Ankara: Tmmob Chamber of Mining Engineers.

Maleki, F., Ovens, K., Najfian, K., Forghani, B., & Reinhold, C. F. (2020). *Overview of Machine Learning Part 1 Fundamentals and Classic Approaches* Neuroimaging Clinics of North America, 30, e17-e32, https://doi.org/10.1016/j.nic.2020.08.007.

Oskay, R., Bechtel, A., & Karayiğit, A. (2019). *Mineralogy, petrography and organic geochemistry of Miocene coal seams in the Kınık coalfield (Soma Basin-Western Turkey): Insights into depositional environment and palaeovegetation*. International Journal of Coal Geology, 210, https://doi.org/10.1016/j.coal.2019.05.012.

German Institute for Economic Research. (2019). *Phasing out Coal in the German Energy Sector: Interdependencies, Challenges and Potential Solutions*, Wuppertal Institute, Ecologic Institute.

Russell, S. J., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach (4 b.)*, Boston: Pearson.

Şengör, A., & Yılmaz, Y. (1981). *Tethyan evolution of Turkey: a plate tectonic approach*. Tectonophysics, 75, 181-241.

Sharma, S., Sharma, S., & Athaiya, A. (2020). Activation Function in Neural Networks. *International Journal of Engineering Applied Sciences and Technology*, 4(12), 310-316, https://doi.org/10.1016/0040-1951(81)90275-4.

Sircar, A., Yadav, K., Rayavarapu, K., Bist, N., & Oza, H. (2021). *Application of machine learning and artificial intelligence in oil and gas industry*. Petroleum Research,6(4), 379-391, https://doi.org/10.1016/j.ptlrs.2021.05.009.

Szlązak, N., Obracaj, D., & Korzec, M. (2021). *Estimation of Gas Loss in Methodology for Determining Methane Content of Coal Seams*. Energies, 14(4), 982, https://doi.org/10.3390/en14040982.

Tang, Y., Gu, F., Wu, X., Ye, H., Yu, Y., & Zhong, M. (2018). *Coalbed methane accumulation conditions and enrichment models of Walloon Coal measure in the Surat Basin, Australia*. Natural Gas Industry B, 5(3), 235-244, https://doi.org/10.1016/j.ngib.2017.11.007.

UN. (2009). Copenhagen Accord. COP 15 (s. 6). *Copenhagen: United Nations. Wilimitis, D.* Retrieved, February 2022, from Towards Data Science. towardsdatascience.com: https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f

Wilimitis, D. *Towards Data Science. towardsdatascience.com*, Retrieved, February 2022, from https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f.

Zhang, B., Zhang, Y., Zhao, S., He, W., Tao, S., Pan, Z., & Cui, Y. (2022). *Coalbed methane desorption characteristics controlled by coalification and its implication on gas co-production from multiple coal seams.* Frontiers of Earth Science, 17, 121–134 (2023). https://doi.org/10.1007/s11707-022-0974-y.

Zhou, F., Xia, T., Wang, X., Zhang, Y., Sun, Y., & Liu, J. (2016). *Recent developments in coal mine methane extraction and utilization in China: A review*. Journal of Natural Gas Science and Engineering, 31, 437-458, https://doi.org /10.1016/j.jngse.2016.03.027.

CURRICULUM VITAE

Name Surname **:** Satuk Buğra Akdaş

EDUCATION **:**

- **B.Sc.** **:** 2016, Istanbul Technical University, Mining Faculty, Petroleum and Natural Gas Engineering Department
- **M.Sc.** **:** 2020, University of Tulsa, Graduate School, McDougall School of Petroleum Engineering Department

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2015-2018 Istanbul Technical University at the Ventilation Laboratory.
- 2018-2020 International Graduate Education Scholarship – Rebuplic of Türkiye Ministry of National Education
- 2020- Present Senior Petroleum Engineer at General Directorate of Mineral Research and Exploration

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Akdas S.B. and Fisne A.** 2017: Coalbed Methane Potential of Zonguldak Coal Basin in Turkey. nternational Energy Raw Materials and Energy Summit, September 27-30, 2017 Istanbul, Turkey

- **Akdas S.B. and Onur M.** 2022: Analytical solutions for predicting and optimizing geothermal energy extraction from an enhanced geothermal system with a multiple hydraulically fractured horizontal-well doublet. *Renewable Energy*, 181, 567-580

- **Akdas S.B. and Fisne A.** 2023: A data-driven approach for the prediction of coal seam gas content using machine learning techniques. *Applied Energy*, 347