SHAPING THE FUTURE OF ENERGY:
A MACHINE LEARNING–BASED ANALYSIS OF TÜRKİYE'S REGIONAL RENEWABLE ENERGY POTENTIAL

Dr. Selen AVCI AZKESKİN
Prof. Dr. Zerrin ALADAĞ

ISBN: 978-625-5753-25-0
Ankara -2025

SHAPING THE FUTURE OF ENERGY: A MACHINE LEARNING–BASED ANALYSIS OF TÜRKİYE'S REGIONAL RENEWABLE ENERGY POTENTIAL

AUTHORS

Dr. Selen AVCI AZKESKİN[1]

Prof. Dr. Zerrin ALADAĞ[2]

[1] Kocaeli University, Faculty of Engineering, Kocaeli, Türkiye.
selen.avci@kocaeli.edu.tr
ORCID ID: 0000-0001-7433-5696

[2] İstanbul Nişantaşı University, Faculty of Engineering and Architecture, İstanbul, Türkiye.
zerrin.aladag@nisantasi.edu.tr
ORCID ID: 0000-0002-5986-7210

DOI: https://doi.org/10.5281/zenodo.17789639

E mail: ubakyayinevi@gmail.com
www.ubakyayinevi.org
It is responsibility of the author to abide by the publishing ethics rules.
UBAK Publishing House – 2025©

ISBN: 978-625-5753-25-0

December / 2025
Ankara / Turkey

PREFACE

Energy has always been a decisive factor in the economic and social development of humanity. Today, the globally increasing population, accelerated industrialization, and rapid urbanization have heightened the demand for energy more than ever before, compelling countries to pursue sustainable, reliable, and environmentally sound energy sources. In an era when energy supply security is directly linked to economic independence, the effective utilization of domestic and renewable energy resources has become a strategic imperative.

Türkiye, with its geopolitical position, climatic diversity, and rich natural resource potential, offers significant opportunities in the field of renewable energy. Nevertheless, the country's energy demand varies considerably across regions; geographical, socioeconomic, and industrial dynamics play a decisive role in shaping regional energy needs. For this reason, adopting a holistic and data-driven approach in Türkiye's sustainable energy planning is essential not only for enhancing energy efficiency but also for optimizing the strategic use of available resources.

This book aims to analyze Türkiye's regional energy potential using the powerful tools of modern data science. Through machine learning–based clustering approaches, the provinces of Türkiye are grouped according to their similar characteristics, and for each cluster, the most suitable renewable energy alternatives are identified. This enables not only a comprehensive understanding of the current energy

landscape but also an interpretation of regional needs and potentials from an integrated perspective.

The book first provides a conceptual framework for energy resources and an overview of Türkiye's current energy profile, followed by a detailed presentation of the findings obtained through machine learning methods. This approach seeks to offer policymakers and practitioners a scientific guide for determining regional priorities, designing investment strategies, and developing sustainable energy policies.

It is our hope that this study contributes to Türkiye's renewable energy transition and strengthens strategic perspectives regarding the nation's energy future.

With a firm belief in the importance of progressing toward a sustainable energy future under the guidance of science…

02.12.2025

Dr. Selen AVCI AZKESKİN

Prof. Dr. Zerrin ALADAĞ

TABLE OF CONTENTS

SHAPING THE FUTURE OF ENERGY: A MACHINE LEARNING–BASED ANALYSIS OF TÜRKİYE'S REGIONAL RENEWABLE ENERGY POTENTIAL

Dr. Selen AVCI AZKESKİN
Prof. Dr. Zerrin ALADAĞ

INTRODUCTION

Energy is regarded as one of the fundamental drivers of the economic, technological, and social development of modern societies, as it is extensively used not only for meeting daily needs such as heating, lighting, and transportation but also for supporting a wide range of industrial and agricultural activities. The rapidly growing global population, coupled with accelerated industrialization and urbanization, continues to increase the demand for energy; how this demand is met directly influences the economic independence of nations as well as their sustainable development goals. Consequently, countries are increasingly seeking new, reliable, and sustainable energy sources to address their rising energy needs.

Energy resources are generally classified into two main categories: "non-renewable" and "renewable." Non-renewable resources are those that cannot be replenished within a short period or require long geological processes for their regeneration. Crude oil, coal, and natural gas are considered "primary energy resources," as they are extracted directly from nature. The derivatives produced through the

transformation of these primary resources—such as electricity and refined fuels—are referred to as “secondary energy resources.” In contrast, renewable energy resources (RES), including solar, wind, hydroelectric, geothermal, wave, and biomass energy, are naturally replenished on a continuous basis. Although increasing the use of RES is essential for reducing environmental impacts and strengthening energy supply security, the share of renewables in total energy consumption remains below desired levels.

Türkiye is a country that meets a significant portion of its energy demand through imports and maintains an energy portfolio dominated by fossil fuels. However, Türkiye possesses substantial opportunities for expanding the use of renewable energy due to its high solar irradiance, strong wind potential, diverse geothermal fields, and considerable biomass capacity. The country also exhibits pronounced regional diversity in terms of geological features, climate zones, vegetation, and socioeconomic characteristics. While Türkiye is administratively divided into seven geographical regions, significant differences in energy demand and renewable potential exist even within the same region. Furthermore, densely populated provinces such as Istanbul, Ankara, and İzmir, as well as industrially strategic provinces like Kocaeli, possess energy profiles that must be evaluated independently from their broader regional characteristics. Accordingly, both energy demand and the most suitable renewable energy alternative can vary considerably across different parts of the country.

The main purpose of this book is to cluster Türkiye's provinces based on their geographical characteristics, renewable energy potential, and socioeconomic structure using machine learning–based methods, and to interpret the most suitable renewable energy source for each cluster.

In the first chapter, a general framework on energy resources is presented, followed by an examination of Türkiye's current energy profile. The second chapter provides a detailed explanation of the clustering and classification methods used in the methodological framework. In the third chapter, the methodology is introduced and the empirical findings are presented. The final chapter discusses the contributions that the proposed approach may offer to policymakers and practitioners, along with recommendations for future research. This book was developed by expanding a section of Selen AVCI AZKESKİN's doctoral dissertation, numbered 970827.

1. ENERGY RESOURCES AND AN OVERVIEW OF TÜRKİYE'S ENERGY LANDSCAPE

Energy sources used in power generation are classified into two main groups based on their availability in nature and their renewability characteristics: non-renewable energy sources and renewable energy sources (RES). Today, the increasing global energy demand and rising concerns regarding energy supply security make the diversification of energy resources a necessity. In this context, reducing dependence on non-renewable energy sources is critically important for preventing economic external dependence and minimizing environmental impacts.

1.1. Non-Renewable Energy Resources

Non-renewable energy resources are formed through geological processes that take millions of years and cannot be replenished at a rate comparable to their consumption. Since their depletion rate exceeds their natural regeneration rate, these resources are finite in nature. The major non-renewable energy sources are briefly described below.

Coal: Coal is one of the oldest and most widely used energy resources worldwide. Although it has a high energy density, its combustion releases significant amounts of carbon dioxide, contributing to air pollution and global warming. Due to its relatively low cost, coal continues to be extensively used, particularly in developing countries. In recent years, efforts have been directed toward reducing its environmental impacts through clean coal technologies and carbon capture systems.

Petroleum: Petroleum is a major fossil fuel with high economic value, widely used in the transportation and industrial sectors. However, its vulnerability to price fluctuations and the potential threat to supply security arising from geopolitical tensions increase its strategic risks. Additionally, the combustion of petroleum releases greenhouse gases, accelerating global climate change.

Natural Gas: Natural gas is transported in compressed or liquefied form and has a wide range of applications across various sectors. Owing to its relatively lower carbon emissions, it is considered a

cleaner fossil fuel. Nevertheless, because it contains methane, its leakage into the atmosphere produces a potent greenhouse gas effect, thereby contributing to global warming.

Nuclear Energy: Nuclear energy is produced through the fission of radioactive elements such as uranium and thorium. It offers advantages such as the absence of carbon emissions during electricity generation and high energy efficiency. However, issues related to radioactive waste management and the potential risks of nuclear accidents keep nuclear energy at the center of ongoing global debates.

1.2. Sustainable Energy and Renewable Energy Resources

Sustainable energy refers to energy production and consumption systems that meet the needs of the present generation without compromising the ability of future generations to meet their own needs. This approach emphasizes minimizing environmental impacts in energy production, ensuring economic feasibility, and adopting socially acceptable solutions. Therefore, sustainable energy is a multidimensional concept encompassing not only clean energy generation but also energy efficiency, energy conservation, the integration of technological innovations, and equitable access to energy. Within this framework, RES constitute one of the fundamental components of sustainable energy systems. These resources, which are naturally replenished through ecological cycles and do not carry the risk of depletion, include solar, wind, hydroelectric, geothermal, and biomass energy. Compared with fossil fuels, RES generate significantly lower greenhouse gas emissions, offering a major

advantage in terms of environmental sustainability. Additionally, since most renewable resources are domestically available, they enhance energy supply security and contribute to economic sustainability by reducing external dependence. The increasing environmental awareness in society and the growing social acceptance of renewable technologies further strengthen the position of RES within sustainable energy strategies. The main renewable energy resources are briefly described below:

Solar energy: Solar energy is obtained by converting sunlight into electrical or thermal energy through photovoltaic panels or solar collectors. The advantages and disadvantages of solar energy can be summarized as follows:

- It does not require complex technology.
- Operation and maintenance costs are low.
- It can be used in areas without electricity transmission lines.
- Since it does not require grid connection, it does not pose transmission-related constraints.
- As a weather-dependent resource, energy production significantly decreases during winter months and at night.
- Energy storage is difficult and overall efficiency is relatively low.

Wind energy: Wind energy is produced by converting the kinetic energy of moving air into mechanical energy through rotor blades

mounted on a shaft. Wind power plants (WPPs) typically operate efficiently for about 20 years, with a total system lifespan of approximately 30 years. WPPs begin generating electricity when wind speed reaches 3 m/s and continue operating until wind speeds reach approximately 25 m/s. Thanks to technological advancements and accurate feasibility studies, the cost of energy derived from wind has steadily decreased. The advantages and disadvantages of wind energy include:

- Investment costs have decreased due to technological improvements.
- Wind turbines are relatively easy to install, transport, and assemble; the risk of accidents during construction is minimal, and maintenance is straightforward.
- Large turbines require extensive land areas.
- Turbine height may pose risks to birds.
- Overall efficiency is generally lower compared to some other energy sources.
- There is a risk of turbine collapse or fire.
- Noise generated by turbines may disturb nearby residents (Erdoğan, 2020; Selçuklu et al., 2022; Movlyanov and Selçuklu, 2025).

Hydroelectric energy: Hydroelectric energy is produced by converting the potential energy difference between two points in a water source

into kinetic energy through a hydroelectric power plant. Hydroelectric power plants (HPPs) must be located at or near the source of water, meaning they cannot always be installed where energy demand exists. The advantages and disadvantages of hydroelectric energy are as follows:

- HPPs do not require fuel and experience minimal energy losses.
- Their efficiency is continuous, and the unit cost of energy is low.
- Maintenance costs are relatively low.
- HPP structures are simple and durable.
- Energy storage and transmission are relatively easy.
- HPPs can quickly respond to high energy demand when needed and can also be rapidly shut down in dangerous situations.
- Construction periods are long and initial investment costs are high.
- Dam construction may lead to submergence of land, displacement of local populations, and challenges during natural disasters.
- Energy production is dependent on precipitation levels.

- Water retention behind dams may cause reductions in agricultural productivity and result in microclimatic changes in the surrounding region.
- Water intake structures may disrupt river ecology, affect aquatic species' migration routes, and harm river ecosystems.

Geothermal energy: Geothermal energy is obtained from underground hot water sources with temperatures consistently above 20°C and with higher mineral and salt content than surrounding groundwater. The advantages and disadvantages of geothermal energy include:

- Geothermal power plants have shorter commissioning periods compared with other types of power plants.
- Continuous energy production is possible.
- The cost of electricity produced in geothermal plants is competitive with coal and natural gas power plants.
- Geothermal energy is not affected by climatic variations.
- Preparation and drilling costs are high.
- Energy transmission from geothermal sites is relatively inefficient.
- Some geothermal reservoirs contain potentially harmful chemical compounds, requiring reinjection techniques.
- The regeneration period of geothermal reservoirs is long once the resource is depleted.

Biomass energy: Biomass energy is produced by converting organic waste into energy through biochemical or thermochemical processes. Agricultural and forestry residues are among the key resources used in biomass energy production. The advantages and disadvantages of biomass energy are as follows (Erdoğan, 2020):

- Biomass crops can be grown in many different regions.
- Production and conversion technologies are well established.
- Low levels of sunlight are sufficient for biomass cultivation.
- Biomass is easy to store.
- Suitable temperatures for biomass production range between 5–35°C.
- Efficiency levels are generally lower compared with other energy sources.
- Biomass production may compete with agricultural land use.
- Significant water resources are required.

Marine current and ocean energy: Marine and ocean energies include wave energy, tidal energy, current energy, and ocean thermal energy conversion (OTEC). Wave energy is generated from the oscillatory motion of ocean waves and the pressure they create. Tidal energy is produced by converting the kinetic energy resulting from the movement of water masses caused by tides into electricity through turbines. For this purpose, water inlets suitable for tidal activity are blocked by constructing a barrage, and electricity is generated using

the height difference that occurs as water flows in and out. Current energy captures the kinetic energy of continuous water movement in seas and oceans using turbines installed on the seabed. Ocean thermal energy conversion utilizes the temperature difference between warm surface waters and cold deep ocean waters in tropical regions to generate electricity through a thermodynamic cycle. For the system to be effective, the temperature difference between the ocean surface and its depths must be at least 20°C (Soylu, 2019).

1.3. An Overview of Türkiye's Energy Profile

Türkiye's energy supply is predominantly based on fossil fuels; however, the transition toward RES has accelerated significantly in recent years. As of 2022, fossil fuels continue to dominate primary energy consumption, with natural gas and coal holding the largest shares in total demand. This dependence increases energy imports, thereby constituting one of the main factors deepening Türkiye's foreign trade deficit.

According to Figure 1, Türkiye's total primary energy supply reached 157.7 million tons of oil equivalent (Mtoe). As shown in Figure 1, petroleum accounted for the largest share of energy sources with 45.11 Mtoe, representing 28.6% of total supply. Petroleum was followed by natural gas with 43.54 Mtoe (27.6%) and coal with 42.02 Mtoe (26.8%). Within the RES category, geothermal energy provided the highest contribution at 11.51 Mtoe (7.3%). Hydropower accounted for 5.75 Mtoe (3.6%), biofuels for 4.51 Mtoe (2.9%), wind energy for

3.01 Mtoe (1.9%), and solar energy for 2.32 Mtoe (1.5%) (Republic of Türkiye Ministry of Energy and Natural Resources, 2023).

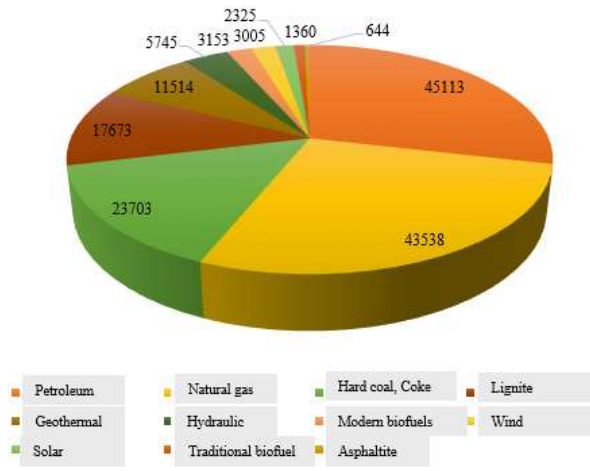

Figure 1: Quantities of Primary Energy Supply in Türkiye (Million TOE, 2022)

Reference: Republic of Türkiye Ministry of Energy and Natural Resources, 2023

According to Table 1, Türkiye's total installed electricity generation capacity reached 107,693 megawatts (MW) by the end of 2023. Hydropower ranked first with a 29.7% share in the installed capacity distribution. Natural gas power plants followed with 23.6%, while coal-based power plants held the third position with 20.3%. Notably, wind power capacity increased to 11%, surpassing that of lignite-fired power plants. Similarly, solar power plants reached an 11.5% share of total installed capacity, exceeding lignite capacity as well.

By the end of 2023, fossil fuel–based power plants had a combined installed capacity of 47,475.2 MW, accounting for 44.1% of total installed power. In contrast, RES-based power plants reached a total capacity of 60,217.6 MW, corresponding to 55.9% of total installed capacity and exceeding fossil fuel capacity (TMMOB Chamber of Mechanical Engineers, 2024).

Table 1: Installed Power Capacity by Energy Source (2023)

Primary Source	**Installed Capacity**		
	MW	**Share (%)**	**Cumulative Share (%)**
Imported Coal	10,373.80	9.63	
Hard Coal	840.80	0.78	
Asphaltite	405.00	0.38	
Lignite	10,194.00	9.47	44.08
Liquid Fuel	260.60	0.24	
Natural Gas	25,401.00	23.59	
Fossil Fuels Total	*47,475.20*	*44.08*	
Biomass + Waste	2404.00	2.23	
Wind	11,803.80	10.96	
Solar	12,354.30	11.47	
Hydropower	31,964.20	29.68	55.92
Geothermal	1691.30	1.57	
Renewables Total	*60,217.60*	*55.92*	
TOTAL	**107,692.80**	**100.00**	**100.00**

Reference: TMMOB Chamber of Mechanical Engineers, 2024

1.4. Renewable Energy in Türkiye

Türkiye's geographical location provides it with a remarkably high solar energy potential. According to the Solar Energy Potential Atlas (GEPA), Türkiye's annual total sunshine duration is 2,737 hours (7.5 hours/day), while the annual total solar irradiation reaches 1,527 kWh/m² (4.2 kWh/m² per day). Figure 2 presents the map illustrating Türkiye's total solar radiation. As shown in the map, solar potential decreases gradually from the southern regions toward the north. Owing to its geographical characteristics and high number of rainy days, the Black Sea Region receives the lowest level of solar irradiation. The Marmara and Aegean Regions receive moderate levels of irradiation, whereas Central Anatolia, Eastern Anatolia, the Mediterranean, and Southeastern Anatolia are the regions with high solar radiation values (Özgür, 2020).

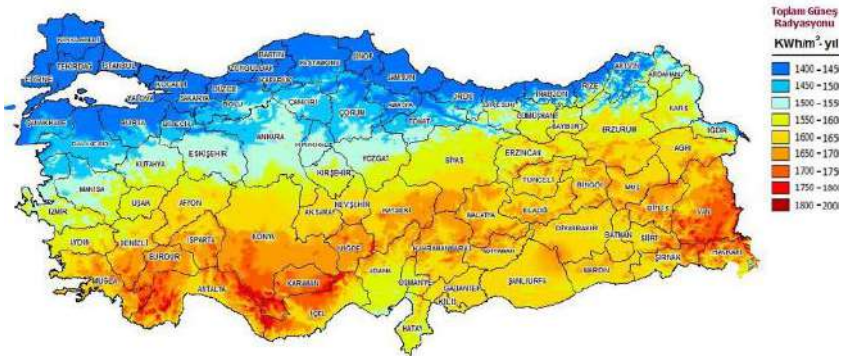

Figure 2: Türkiye Solar Energy Potential Atlas (GEPA)
Reference: Ministry of Energy and Natural Resources, 2024a

Türkiye possesses significant wind energy potential due to being surrounded by seas on three sides, its widespread mountainous terrain, and the presence of diverse climatic conditions. Considering annual average wind speeds, the Aegean and Marmara coastlines stand out as the most suitable regions for wind energy generation. The Türkiye Wind Energy Potential Atlas (REPA), presented in Figure 3, analyzes the country's wind characteristics and distribution, contributing to the identification of the most favorable areas for electricity production. According to calculations, when areas with wind speeds above 7 m/s at a height of 50 meters are considered, Türkiye's onshore wind energy potential is estimated to be approximately 48,000 MW (Kaya & Kaya, 2024).

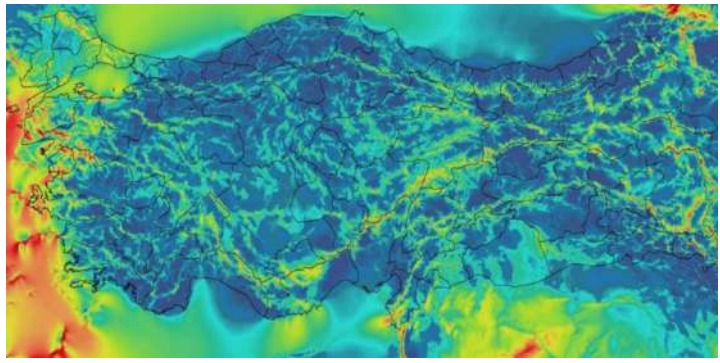

Figure 3: Türkiye Wind Energy Potential Atlas (REPA)
Reference: Ministry of Energy and Natural Resources, 2024b

Türkiye's annual average precipitation is approximately 574 mm, corresponding to an average of 450 billion m^3 of water per year. The country's gross surface water potential has been identified as 185

billion m³, while the groundwater potential is estimated at 18 billion m³. Figure 4 presents Türkiye's 25 drainage basins (Serdar, 2020).

Figure 4: Map of Türkiye's 25 Drainage Basins

Reference: Serdar, 2020

Türkiye, receiving substantial sunlight and possessing extensive agricultural land, abundant water resources, and diverse climatic conditions, offers considerable potential for biomass energy production. Figure 5 illustrates the distribution of biomass potential across Türkiye's provinces (İllez, 2020).

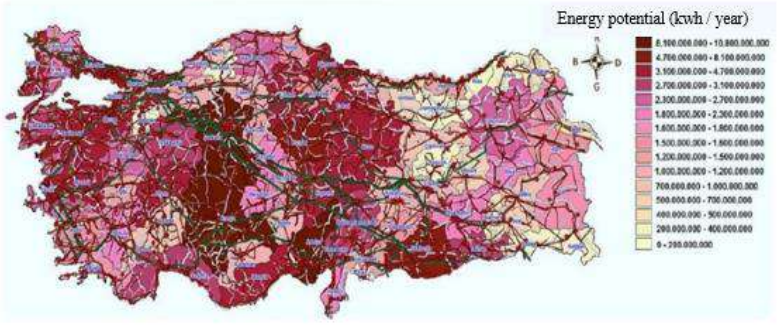

Figure 5: Distribution of Biomass Potential by Province in Türkiye

Reference: İllez, 2020

A significant portion of Türkiye's geothermal resources is concentrated in the Western Anatolia region. Approximately 78% of areas with geothermal potential are in this region, followed by Central Anatolia with 9% and the Marmara Region with 7%. The geothermal potential in Eastern Anatolia is around 5%, while other regions account for about 1%. Since nearly 90% of existing geothermal resources have low to medium temperature levels, they are predominantly used in heating systems, thermal tourism, and certain industrial applications. The remaining 10% is utilized in indirect energy applications such as electricity generation.

According to data from the International Energy Agency, Türkiye's installed geothermal electricity capacity increased from 94 MW in 2010 to 1,283 MW by 2018, while electricity generation rose from 668 GWh to 4,819 GWh over the same period. Most of this production capacity is concentrated in the Aegean Region. Provinces such as Aydın, Denizli, Manisa, and Çanakkale are among the leading locations with significant geothermal potential (Gürcün & Petek, 2021).

1.5. Türkiye's Renewable Energy Policies and Future Targets

Türkiye's long-term energy strategies aim to reduce carbon emissions and fulfill its commitments under the Paris Agreement. Within this framework, the Twelfth Development Plan (2024–2028) prioritizes the expansion of renewable energy generation capacity, the enhancement of energy efficiency, and the promotion of investments in this field. To

achieve these goals, it is planned to strengthen credit and incentive mechanisms for energy projects.

The Renewable Energy Resource Areas (YEKA) model is implemented to promote investments in renewable energy. YEKA projects offer specific incentives—particularly for large-scale wind and solar power plants—encouraging the private sector to invest in renewable energy. Another key mechanism supporting renewable energy production in Türkiye is the Renewable Energy Support Scheme (YEKDEM). YEKDEM provides financial assurance to the sector by guaranteeing the purchase of electricity generated from renewable sources at incentivized tariffs for a specified duration. In addition, regulatory reforms introduced by the Energy Market Regulatory Authority (EPDK) have facilitated market liberalization and supported private sector investments in the electricity market.

Through these initiatives, Türkiye aims to increase its renewable energy capacity by 2035 and raise the share of renewables to 65% of total electricity generation. Particular emphasis is placed on expanding wind and solar energy investments, with plans for these sources to reach a combined capacity of 120,000 MW. Furthermore, the development of energy storage systems and the modernization of grid infrastructure are considered critical. Investments in electricity transmission infrastructure are planned to strengthen the integration of renewable power plants into the national grid. Increasing green hydrogen production and promoting its use in industrial processes as a

substitute for fossil fuels also stand among Türkiye's strategic objectives.

However, the renewable energy sector in Türkiye faces several significant challenges. The lack of sufficient financing remains one of the main obstacles to implementing large-scale renewable energy projects. Sustainable financial models are required to support long-term investments. Additionally, Türkiye continues to depend on foreign technology for renewable energy systems. A large portion of critical equipment—such as wind turbines, solar panels, and energy storage systems—is imported, making it essential to strengthen domestic manufacturing capacity.

For these strategic goals to be effectively implemented, it is necessary not only to formulate energy policies at the national level but also to conduct a detailed analysis of energy potential at the regional level. Such an approach ensures the efficient use of public resources and enables the design of targeted incentive mechanisms tailored to the specific needs of each region. Based on this necessity, this book clusters Türkiye's provinces according to their sustainable energy potential using a range of variables and provides an in-depth analysis of the resulting clusters (Avcı Azkeskin & Aladağ, 2025).

2. MACHINE LEARNING METHODS

In this section, machine learning–based clustering and classification approaches are examined within a theoretical framework.

Machine learning is a branch of artificial intelligence that enables computer systems to learn autonomously by analyzing patterns within data, rather than relying on predefined rules. Machine learning is commonly categorized into three main types based on the learning paradigm: supervised learning, unsupervised learning, and semi-supervised learning. Each learning type differs depending on the structure of the data used and the nature of the model's learning process.

2.1. Supervised Learning

Supervised learning is a machine learning approach used when each input instance in the dataset is accompanied by a corresponding correct output (label). In this method, the model learns to make accurate predictions on new, unseen data by analyzing patterns within historical labeled data. The training process involves feeding the model with a labeled dataset, and the model parameters are optimized by minimizing the difference (error rate) between the model's predictions and the true labels.

Within supervised learning, the dataset is typically divided into training and test subsets. The training data are used during the learning phase, allowing the model to discover the relationships between inputs and outputs. The test data, which the model has not encountered previously, are used to evaluate the model's performance and assess its generalization ability. Supervised learning is broadly categorized into two main tasks: classification and regression. Classification involves assigning data points to predefined categories, whereas

regression focuses on predicting continuous numerical values (Kotsiantis, 2007).

The classification methods addressed within the scope of this book are explained in the following subsections.

2.1.1. Multinomial Logistic Regression (MLR)

Multinomial Logistic Regression (MLR) is a generalized version of logistic regression designed to analyze dependent variables that contain more than two categorical outcomes. This approach models the likelihood of each possible category of the response variable by relating it to a set of predictor variables. In MLR, these category probabilities are computed through a linear combination of the explanatory variables. The mathematical form of the model is given in Equation (1). In this formulation, $P(Y = k \mid X)$ denotes the probability that the response falls into class k; X represents the vector of predictors; K indicates the total number of outcome categories; and β_k is the parameter vector associated with class k (Coughenour et al., 2015).

$$P(Y = k \mid X) = \frac{e^{X^T \beta_k}}{1 + \sum_{j=1}^{K-1} e^{X^T \beta_j}} \quad (1)$$

2.1.2. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a straightforward yet powerful classification approach that predicts the class of an observation by examining the labels of the closest samples in the training dataset

(Tahtalı, 2020). Essentially, the algorithm assigns a class to a new data point by considering either the majority category among its k nearest neighbors in the feature space or a weighted voting process based on their distances. The choice of the parameter k plays a key role, as it directly influences both the model's predictive accuracy and overall performance.

2.1.3. Support Vector Machines (SVM)

Support Vector Machines (SVM) operate on the principle of identifying a hyperplane that can optimally separate two classes (Sasidharan, 2021). Initially designed for the classification of linearly separable binary problems, SVMs were later generalized to handle multiclass and non-linear datasets (Kaba & Kalkan, 2022).

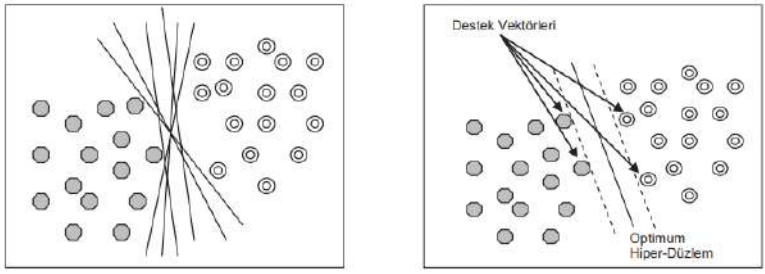

Figure 6: Support Vector Machines (SVM)

Reference: Kaba & Kalkan, 2022

As illustrated in Figure 6, multiple hyperplanes may be drawn to separate two classes. However, the primary objective of SVM is to identify the hyperplane that maximizes the margin—the distance between the hyperplane and the nearest data points from each class. To determine the optimal hyperplane, two parallel hyperplanes forming

the boundaries of the margin must be defined. The data points lying on these boundary hyperplanes are referred to as support vectors.

SVM determines the optimal separating hyperplane by solving the optimization problem shown in Equation (2) (Sasidharan, 2021). In this formulation, w represents the normal vector of the hyperplane; b denotes the bias term; y_i indicates the class label; and x_i represents the training data points.

$$\min_{\mathbf{w},\mathrm{b}} \frac{1}{2}|\mathrm{w}|^2 \tag{2}$$

Subject to:

$$\mathrm{y_i}(\mathrm{w^T x_i} + \mathrm{b}) \geq 1, \forall \mathrm{i}$$

The hinge loss function, used as the cost function of SVM, is defined in Equation (3). Here, y_i denotes the true class label of the i^{th} data instance, and $f(x_i) = w^T x_i + b$ represents the predicted value.

$$\mathrm{L}\big(\mathrm{y_i}, \mathrm{f(x_i)}\big) = \max\big(0, 1 - \mathrm{y_i\ f(x_i)}\big) \tag{3}$$

This function penalizes classification errors by assigning a loss to instances that lie within the margin or are misclassified. If a data point is correctly classified and lies outside the margin, the loss becomes zero. However, if a data point is misclassified or correctly classified but located within the margin, the function produces a positive penalty. Thus, SVM simultaneously aims to maximize the margin while minimizing the risk of misclassification and margin violations.

The SVM optimization problem can be transformed into its dual form using Lagrange multipliers. In Equation (4), α_i represents the Lagrange multipliers; C is the regularization (penalty) parameter; $K(x_i \cdot x_j)$ denotes the kernel function; and y_i and y_j indicate the class labels.

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (4)$$

Subject to:

$$0 \leq \alpha_i \leq C, \sum_{i=1}^{N} \alpha_i y_i = 0$$

This transformation facilitates the solution of the optimization problem and enables classification, particularly for datasets that are not linearly separable, using kernel functions. Kernel functions map data points into a higher-dimensional feature space, thereby increasing the likelihood of linear separability. The most used kernel functions are the Linear, Polynomial, Radial Basis Function (RBF), and Sigmoid kernels, which are given in Equations (5), (6), (7), and (8), respectively. In the polynomial kernel, *c* is a constant and *d* denotes the degree of the polynomial. In the RBF kernel, σ represents the kernel parameter. In the sigmoid kernel, α and *c* are constant parameters.

$$(x_i, x_j) = x_i^T \cdot x_j \quad (5)$$

$$K(x_i, x_j) = (x_i^T \cdot x_j + c)^d \quad (6)$$

$$K(x_i, x_j) = \exp(-\sigma|x_i - x_j|^2) \quad (7)$$

$$K(x_i, x_j) = \tanh(\alpha x_i^T . x_j + c) \quad (8)$$

The linear kernel is preferred when the data are linearly separable, offering a simpler and more interpretable model from a computational perspective. The polynomial kernel provides greater flexibility for modeling non-linear decision boundaries, with its behavior determined by the degree of the polynomial. The sigmoid kernel resembles the activation functions used in artificial neural networks and can be effective for certain data distributions.

SVM models using the RBF kernel are typically characterized by two main hyperparameters: C (the regularization parameter) and σ (the kernel width parameter). Compared to other non-linear kernel functions, the RBF kernel involves fewer hyperparameters and generally yields higher classification performance (Schölkopf & Smola, 2002). For these advantages, it was selected for use in this study.

2.1.4. Random Forest (RF)

Decision trees are among the most used classification algorithms due to their strong learning capabilities. However, this powerful learning ability often leads to the disadvantage of overfitting. The Random Forest (RF) method is a classification approach based on decision trees. Introduced to the literature by Breiman (2001), the algorithm randomly selects samples through resampling, constructs multiple

decision trees based on these samples, aggregates their predictions, and produces the final output through voting with high accuracy. Since each tree grows freely without pruning, the method avoids the overfitting problem (Güleryüz, 2022).

To construct an RF model, two main parameters are required: the number of trees (*ntree*) and the number of features considered at each split (*mtry*) (Andrade et al., 2020).

Decision trees typically split the dataset based on entropy as shown in Equation (9) or information gain as shown in Equation (10):

$$\mathrm{H(S)} = -\sum_{\mathrm{i=1}}^{\mathrm{c}} \mathrm{p_i} \log_2 \mathrm{p_i} \tag{9}$$

$$\mathrm{IG(T,a)} = \mathrm{H(T)} - \sum_{\mathrm{v \in a}} \frac{|\mathrm{T_v}|}{|\mathrm{T}|} \mathrm{H(T_v)} \tag{10}$$

In Equation (9), S represents the dataset, c denotes the number of classes, and p_i indicates the probability of the i^{th} class. In Equation (10), T is the root node, a is the feature on which the split is performed, $IG(T, a)$ denotes the information gain obtained by splitting on feature a, $H(T)$ is the entropy of the entire set, and T_v represents the subset of data for which feature a takes the value v. Thus, the formula calculates the reduction in entropy —i.e., the gain— when a feature is used for splitting.

2.1.5. Extreme Gradient Boosting (XGBoost)

In recent years, boosting-based methods have gained substantial popularity in the field of data science. These algorithms rely on the

sequential combination of multiple weak classifiers to construct highly accurate (strong) predictive models. The fundamental objective of the boosting approach is to enhance model performance by focusing particularly on observations that contribute most to prediction errors. The process begins with building a single weak learner; subsequently, each new model is constructed sequentially to minimize the errors made by the previous model. The final model is produced by weighting the weak learners based on their performance, giving greater influence on better-performing models. As a result, a highly generalizable and powerful ensemble model emerges.

XGBoost is a machine learning method based on Gradient Boosting Machines (GBM) and decision trees. The GBM algorithm was first introduced by Friedman (2002), and the XGBoost version, presented by Chen and Guestrin (2016) in a conference, quickly gained widespread adoption and became highly popular in the field of machine learning. XGBoost represents an optimized version of GBM, enhanced with various regularization techniques. In addition to its strong predictive power, XGBoost is superior to many traditional methods due to its ability to prevent overfitting, handle missing data effectively, and offer high computational efficiency.

The main objective of XGBoost is to minimize a loss function together with a regularization term, as expressed in Equation (11) (Chen & Guestrin, 2016):

$$L(\phi) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (11)$$

Here, l denotes the loss function (e.g., mean squared error – MSE); $\widehat{y_i}$ is the predicted value; and $\Omega(f_k)$ represents the regularization term, that is, the complexity penalty of the k^{th} weak learner (decision tree). The regularization term helps prevent overfitting by controlling the complexity of the model. It is typically defined as in Equation (12):

$$\Omega(\mathrm{f}) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{\mathrm{j}=1}^{\mathrm{T}} \mathrm{w}_{\mathrm{j}}^2 \tag{12}$$

Here, γ represents the fixed penalty coefficient for each leaf, T denotes the number of leaves, λ is the regularization parameter, and w_j indicates the weight of the j^{th} leaf. XGBoost improves the model by adding a new tree at each iteration based on the current predictions. This process is expressed in Equation (13):

$$\widehat{\mathrm{y}_{\mathrm{i}}^{(\mathrm{t})}} = \widehat{\mathrm{y}_{\mathrm{i}}^{(\mathrm{t}-1)}} + \mathrm{f}_{\mathrm{t}}(\mathrm{x}_{\mathrm{i}}) \tag{13}$$

Here, $\widehat{y_i^{(t)}}$ denotes the prediction for observation i at iteration t; $f_t(x_i)$, represents the new decision tree trained at iteration t; and $\widehat{y_i^{(t-1)}}$ refers to the prediction from the previous iteration.

XGBoost optimizes the loss function through a second-order Taylor expansion, as shown in Equation (14):

$$\mathrm{L}^{(\mathrm{t})} \approx \sum_{\mathrm{i}=1}^{\mathrm{n}} \left[\mathrm{l}\left(\mathrm{y}_{\mathrm{i}}, \widehat{\mathrm{y}_{\mathrm{i}}^{(\mathrm{t}-1)}}\right) + \mathrm{g}_{\mathrm{i}}\mathrm{f}_{\mathrm{t}}(\mathrm{x}_{\mathrm{i}}) + \frac{1}{2}\mathrm{h}_{\mathrm{i}}\mathrm{f}_{\mathrm{t}}^2(\mathrm{x}_{\mathrm{i}}) \right] + \Omega(\mathrm{f}_{\mathrm{t}}) \tag{14}$$

Here, $g_i = \frac{\partial L\left(y_i, \widehat{y_i^{(t-1)}}\right)}{\partial \widehat{y_i^{(t-1)}}}$ is the first derivative, and $h_i = \frac{\partial^2 L\left(y_i, \widehat{y_i^{(t-1)}}\right)}{\partial \left(\widehat{y_i^{(t-1)}}\right)^2}$ is the second derivative. Each iteration is formulated as in Equation

(15):

$$L^{(t)} = \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T \quad (15)$$

where $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$. The optimal weight of each leaf is computed using Equation (16):

$$w_j = -\frac{G_j}{H_j + \lambda} \quad (16)$$

2.1.6. Stacked Ensemble Learning Technique

Stacking is an ensemble strategy that integrates the outputs of several machine learning algorithms (base learners) by employing an additional predictive model known as a meta-learner. Unlike boosting, which sequentially improves a single model, stacking trains multiple diverse models in parallel and then uses a meta-model to determine the most effective way to combine their predictions, thereby improving overall predictive capability. In this framework, the meta-model receives the base models' predictions as input features and learns how to merge them to produce the final prediction.

Stacked ensemble methods are generally implemented in two phases:

Base Models: In the first phase, multiple base learners are fitted using the same training data. Each algorithm attempts to estimate the target variable based on its own modeling principles. These models produce predicted values for each observation, forming the initial layer of outputs.

Meta-Model: During the second phase, the outputs generated by the base models are used as new input variables for the meta-learner. This model identifies the most effective way to combine these predictions and generates the final outcome. Meta-learners with lower complexity—such as linear regression or logistic regression—are frequently preferred due to their stability and interpretability (Solomon et al., 2023; Shih et al., 2024, Avcı Azkeskin & Aladağ, 2025).

In this book, two different stacked ensemble models were utilized:

Ensemble Model 1: Predictions obtained from the RF and XGBoost algorithms were combined to create a stacked dataset. On this dataset, MLR, SVM, and KNN were trained as meta-models. The meta-models were trained using the training portion of the stacked dataset, and their accuracy values were evaluated on the test set.

Ensemble Model 2 (Majority Voting Model): The predictions from the RF and XGBoost models were combined using the majority voting approach.

2.1.7. Evaluation of Classification Performance

Classification performance can be assessed using several evaluation metrics, such as accuracy, Matthews Correlation Coefficient (MCC), and Cohen's Kappa.

Accuracy is one of the fundamental metrics used to evaluate the performance of classification models. It is calculated as the ratio of

correctly predicted instances to the total number of instances. However, accuracy may be misleading when the dataset contains class imbalance. For example, a model may achieve high accuracy simply by predicting the majority class consistently, even though it fails to distinguish between classes effectively. In multiclass models, accuracy is computed using Equation (17). Here, TP represents true positives, TN true negatives, FP false positives, and FN false negatives.

$$\text{Accuracy} = \frac{\sum_{i=1}^{K} TP_i}{\sum_{i=1}^{K} (TP_i + TN_i + FP_i + FN_i)} \tag{17}$$

MCC is a statistical metric used to evaluate the performance of classification models and is particularly useful for datasets with class imbalance, often providing a more reliable assessment than metrics such as accuracy. In multiclass problems, MCC is calculated using Equation (18). Here, C_{ii} represents the diagonal elements of the confusion matrix, while C_{ij} and C_{ji} denote the off-diagonal elements that represent misclassifications between classes.

$$\text{MCC} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{K} (C_{ii} \cdot C_{jj} - C_{ij} \cdot C_{ji})}{\sqrt{\left(\sum_{i=1}^{K} \left(\sum_{j=1}^{K} C_{ij}\right) \left(\sum_{j=1}^{K} C_{ji}\right)\right) \left(\sum_{i=1}^{K} \left(\sum_{j=1}^{K} C_{ji}\right) \left(\sum_{j=1}^{K} C_{ij}\right)\right)}} \tag{18}$$

Kappa measures how much the observed classification performance exceeds the performance expected by random chance. For multiclass classification, the Kappa statistic is calculated using Equation (19), where P_o is the observed agreement (the proportion of correctly classified instances), and P_e is the expected agreement under random classification.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (19)$$

2.2. Unsupervised Learning

Unsupervised learning refers to a category of machine learning techniques designed to reveal underlying structures, patterns, and relationships within data without relying on labeled examples. In this framework, the model examines only the input variables to detect similarities or distinctions among observations, as no predefined class information is supplied. In essence, the algorithm autonomously identifies which data points resemble one another and groups them according to the natural organization of the dataset.

Unsupervised learning approaches are primarily used for two tasks: “clustering” and “dimensionality reduction”. Dimensionality reduction methods aim to project high-dimensional data into a more compact form while maintaining its essential characteristics, allowing for more efficient and interpretable analyses. Clustering methods, on the other hand, focus on forming meaningful subgroups by gathering observations that share similar attributes.

Cluster analysis, a widely applied multivariate statistical technique, partitions datasets into groups based on specified similarity or distance metrics. This process helps reveal hidden structures, supports data organization, and contributes to generating more interpretable insights. The most commonly applied clustering algorithms in the literature are discussed in the subsequent sections (Zorlutuna & Erilli, 2018).

2.2.1. K-Means Clustering

The K-Means method is a popular clustering algorithm that aims to partition a dataset into k predefined, distinct clusters. The algorithm assigns each data point to the nearest cluster centroid, after which each centroid is updated based on the data points assigned to that cluster. This process continues until the cluster assignments stabilize or a predefined stopping criterion is met. Determining an appropriate value for k is critical for obtaining successful clustering results (Wu et al., 2021). The algorithm proceeds through the following steps (Jain, 2010):

1. Initialization: The number of clusters k is selected, and k initial cluster centers $\mu_1, \mu_{2,\dots}\mu_k$ are chosen randomly.

2. Assignment step: Each data point x_i is assigned to the nearest cluster center μ_j, as shown in Equation (20), where $|x_i - \mu_j|$ represents the distance between x_i and μ_j.

$$c_i = \arg\min_j |x_i - \mu_j| \tag{20}$$

3. Update step: Each cluster center is updated to the mean of the data points assigned to it, as formulated in Equation (21). Here, μ_j denotes the centroid of the j^{th} cluster, C_j is the set of points assigned to cluster j, and $|C_j|$ is the number of points in the cluster.

$$\mu_j = \frac{1}{|C_j|}\sum_{x_i \in C_j} x_i \tag{21}$$

4. Iteration: Steps 2 and 3 are repeated until the cluster centers change minimally or the stopping criterion is satisfied. The stopping criterion is typically defined using a threshold based on the degree of change in centroid locations or data point assignments.

In this study, the distance metrics used for clustering were Euclidean, Manhattan, and Minkowski distances, shown respectively in Equations (22), (23), and (24):

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (22)$$

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i| \quad (23)$$

$$d(x,y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{\frac{1}{p}} \quad (24)$$

As shown in Equation (24), the Minkowski distance can represent different distance metrics depending on the parameter p. For example, when p=1, Minkowski distance is equivalent to Manhattan distance; when p=2, it becomes Euclidean distance.

2.2.2. Hierarchical Clustering

Another widely used clustering method is hierarchical clustering. Hierarchical clustering algorithms construct clusters by successively merging or splitting the dataset into nested structures. This approach is based on two main strategies: agglomerative (bottom-up) and divisive (top-down). In the agglomerative approach, each data point initially starts as its own single cluster. At each step, the algorithm merges the two clusters that are closest to each other. This process continues until all data points are combined into a single cluster. Conversely, the

divisive approach begins with all data points grouped into one large cluster. This cluster is then recursively split into smaller subclusters at each step, and the process continues until each data point forms its own individual cluster. Both approaches produce a tree-like visualization known as a dendrogram. The dendrogram enables the examination of similarity levels between clusters and allows tracking of the clustering process visually. Because the dendrogram can be cut at any desired level to obtain different numbers of clusters, hierarchical clustering offers a flexible structure. Among hierarchical clustering techniques, one of the most preferred methods is Ward's method. Ward's method determines cluster merging decisions by minimizing within-cluster variance, which generally leads to the formation of balanced and homogeneous clusters. Unlike classical clustering methods such as K-Means and hierarchical clustering, which assign each data point to a single cluster absolutely, fuzzy clustering methods acknowledge that in some cases it may not be possible to assign observations to a single cluster definitively. Especially in datasets with ambiguous boundaries between clusters, fuzzy clustering techniques have been developed. In fuzzy clustering, an observation can belong to multiple clusters with different membership degrees ranging between 0 and 1. As in traditional clustering, distances are computed using distance metrics in fuzzy clustering as well. The Fuzzy C-Means (FCM) method, which was used in this study, is explained in the following subsection.

2.2.3. Fuzzy C-Means (FCM)

The Fuzzy C-Means (FCM) algorithm is one of the most widely used fuzzy clustering methods. Similar to many other fuzzy clustering algorithms, it is based on minimizing a specific objective function, and the algorithm terminates when the improvement in this function falls below a predetermined threshold (Güleryüz, 2022). Initially, the membership matrix $U^{(0)} = [u_{ij}]$ with dimensions $n \times c$ is randomly initialized. Here, u_{ij} represents the degree of membership of the i^{th} data point in the j^{th} cluster, subject to the conditions $0 \leq u_{ij} \leq 1$ and $\sum_{j=1}^{c} u_{ij} = 1$. Cluster centers are computed using Equation (25) (Bezdek, 1981):

$$v_j = \frac{\sum_{i=1}^{n} u_{ij}^{m} x_i}{\sum_{i=1}^{n} u_{ij}^{m}} \tag{25}$$

where v_j is the center of the j^{th} cluster, m is the fuzzifier parameter, and x_i is the i^{th} data point. Membership degrees are updated using Equation (26):

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d(x_i, v_j)}{d(x_i, v_k)} \right)^{\frac{2}{(m-1)}}} \tag{26}$$

where $d(x_i, v_j)$ denotes the distance between the i^{th} data point and the j^{th} cluster center. In this study, Euclidean, Manhattan, and Minkowski distance metrics were used to measure this distance, and the effects of these metrics on the clustering outcomes were analyzed. Finally, the stopping criterion shown in Equation (27) checks whether the change in the membership matrix U falls below a prescribed threshold. If the

change is smaller than this value, the algorithm terminates; otherwise, it returns to Equation (25). Here, ε is a small positive constant representing the stopping threshold.

$$|U^{(k+1)} - U^{(k)}| < \varepsilon \quad (27)$$

2.2.4. Evaluation of Clustering Performance

The Silhouette score is one of the most commonly applied measures for judging the quality of clustering results. It evaluates how well a data point fits into its assigned cluster while also considering how different it is from other clusters. The score varies between –1 and +1. Values approaching +1 indicate that the observation is well matched to its cluster, whereas scores near 0 suggest that the point is positioned close to a cluster boundary. The Silhouette value for each observation is computed using Equation (28), where $a(i)$ represents the average distance from the i^{th} point to all other points within the same cluster, and $b(i)$ is the average distance from the i^{th} point to the closest other cluster. The overall Silhouette score S is calculated by averaging the individual scores of all observations, as shown in Equation (29) (Rousseeuw, 1987):

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (28)$$

$$S = \frac{1}{N}\sum_{i=1}^{N} s(i) \quad (29)$$

The Calinski–Harabasz Index (CHI) is another well-established metric used for assessing clustering effectiveness. This index evaluates the

clustering structure by jointly considering within-cluster compactness and between-cluster separation. CHI is defined in Equation (30), where SSB denotes the between-cluster sum of squares, SSW is the within-cluster sum of squares, *K* is the number of clusters, and *N* is the total number of observations. The quantities SSB and SSW are derived from Equations (31) and (32), respectively (Caliński & Harabasz, 1974):

$$\text{CHI} = \frac{\text{SSB}/(\text{K}-1)}{\text{SSW}/(\text{N}-\text{K})} \tag{30}$$

$$\text{SSB} = \sum_{k=1}^{K} n_k \left(\mu_k - \mu\right)^2 \tag{31}$$

$$\text{SSW} = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{32}$$

In these expressions, n_k refers to the number of data points in cluster, k^{th} is the centroid of cluster, μ is the global mean, x_i represents the data point, and C_k denotes the set of observations that belong to the k^{th} cluster. In conclusion, the Silhouette score focuses on how clearly separated clusters are from one another, whereas the CHI index provides an evaluation based on both cluster compactness and inter-cluster distinctiveness.

2.3. Semi-Supervised Learning

Semi-supervised learning is a machine learning paradigm that incorporates both labeled and unlabeled data during model training. Because generating labeled datasets can be labor-intensive and expensive, this approach seeks to improve learning performance by

extending the information derived from a limited number of labeled samples to a substantially larger set of unlabeled observations. The model leverages the structural patterns recognized from the labeled portion of the data to make informed predictions about the unlabeled instances. As a result, semi-supervised techniques are capable of producing effective models while requiring far fewer labeled examples than traditional fully supervised methods.

This approach is frequently applied in domains such as healthcare, text analytics, and image processing—areas characterized by abundant data availability but limited feasibility for comprehensive labeling.

3. EVALUATION OF TÜRKİYE'S SUSTAINABLE ENERGY POTENTIAL USING MACHINE LEARNING METHODS

In this section, the provinces of Türkiye are grouped according to their sustainable energy potentials based on a set of selected indicators. To this end, a general assessment of the application areas of machine learning methods will first be presented, supported by examples from the literature. Subsequently, the analyses conducted within the proposed methodology will be examined, and the findings will be presented.

3.1. Related Work

Various studies in the literature have analyzed countries or regions in terms of their Sustainable Energy Potential (SEP). In some of these studies, countries or regions were grouped using clustering methods according to their SEP levels, and different analyses were conducted

based on the resulting groups. In others, SEP was treated as a decision-making problem, and Multi-Criteria Decision-Making (MCDM) methods were employed to rank alternatives and identify the most suitable option. Most of these studies either focus on the selection of a renewable energy source for a specific region (Saraswat & Digalwar, 2021; Şahin, 2021; Afsordegan et al., 2016; Abdullah & Najib, 2016; Seddiki & Bennadji, 2019) or aim to contribute to the development of policies to enhance SEP (Marinakis et al., 2017; Solangi et al., 2019; Dall'O' et al., 2013).

A summary of clustering-based studies is presented in Table 2. The table includes information about the clustering method used in each study, the application area, the variables considered, and how the number of clusters was determined or how clustering performance was evaluated.

Table 2: Summary of the Literature on SEP Assessment

Author(s)	Method	Application Area	Variables	Cluster Number Determination
Trappey et al. (2014)	Self-Organizing Map (SOM) and AHP	Assessment of renewable energy policies in Taiwan	Economic indicators: facility/installation costs, incentive and tax policies, energy supply; Environmental indicators: CO_2 emissions, fossil fuel usage, greenhouse gas impacts	Root Mean Square Error (RMSE)
Grigoras & Scarlatache (2015)	K-Means	Analysis of renewable energy potential in Romania	Installed capacity, voltage level, renewable technology type, geographic location	Silhouette index

Author(s)	Method	Application Area	Variables	Cluster Number Determination
Pelau & Chinie (2018)	Ward	Comparison of innovation and sustainability levels across European countries	Number of PhD graduates, scientific publications, R&D expenditures, patents, product/service exports, electricity use, waste generation, air pollution, GHG emissions, recycling rates	Dendrogram analysis, Elbow method
Tutak et al. (2020)	TOPSIS and Kohonen neural network	Sustainable energy development in EU countries	Total primary energy supply, final energy consumption, installed electricity capacity, energy efficiency, energy taxation, electricity prices, R&D spending, GHG emissions, air pollution, poverty rate	P_i values obtained via TOPSIS
Liu et al. (2020)	Multidimensional goal-oriented clustering, Gaussian Mixture Model (GMM)	Classification of energy investment options in Brisbane, Australia	Daily energy consumption, daily solar output, maximum temperature, energy tariff, demand tariff	Calinski–Harabasz Index (CHI)
Wang & Yang (2020)	Projection pursuit fuzzy clustering and accelerated genetic algorithm	Evaluation of renewable energy sustainability in 27 EU countries	Economic development, environmental pressure, energy conditions, social progress, governance and policy dimensions	XB (Xie–Beni), PE (Partition Entropy), PC (Partition Coefficient)
Kacperska et al. (2021)	Ward	Renewable energy usage patterns in EU and Visegrad Group countries	Share of renewables in transport, electricity generation, and heating/cooling	Dendrogram analysis
Gostkowski et al. (2021)	K-Means, DIANA (Hierarchical Clustering)	Energy consumption structure in Visegrad Group countries	Total primary energy supply (TPES), energy efficiency (energy/GDP), energy intensity (TPES/GDP), sectoral energy consumption	Silhouette index, Rand–Jaccard index
Matenga (2022)	K-Means	SDG7 performance comparison in developed (USA, China) and developing regions (Sub-Saharan Africa, South Asia)	Access to energy, energy production sources, energy consumption, energy losses, short-term debt, per-capita income	Elbow method

Author(s)	Method	Application Area	Variables	Cluster Number Determination
Quatro¬si (2022)	K-Medoids	Environmental performance and clean energy modelling in EU countries	Environmental performance scores, energy consumption, GDP per capita, industrial added value, population density, urbanization rate	Elbow method
Li et al. (2022)	Bootstrap DEA and K-Medoids	Regional sustainable development assessment in China	Energy use, labor, capital, R&D stock, SO_2 emissions, GDP, patent counts	Interpretation of bi-cluster graphs
Kosowski et al. (2023)	K-Means	Structure of primary energy consumption in Europe	Solid fossil fuels, crude oil, natural gas, nuclear energy, renewables, and other energy sources	SSW and Silhouette index

Table 2 shows that environmental and macroeconomic variables are frequently used in clustering studies. Unlike many others, Grigoras and Scarlatache (2015) also included "location" as an analysis criterion. Additionally, some studies incorporated social indicators such as the number of PhD graduates, scientific publications, R&D expenditures, and patent applications.

3.2. Methodology

3.2.1. Dataset

In this study, the provinces of Türkiye were clustered based on their SEP. The criteria affecting SEP were examined under three main categories: socioeconomic structure, geographical characteristics, and renewable energy potential (REP).

Within the socioeconomic category, the following variables were used: land area, population, annual population growth rate (‰), GDP per capita, total exports (thousand USD), total imports (thousand

USD), industrial volume (thousand TRY), and invoiced electricity consumption (MWh). Geographical position was incorporated into the model using latitude, longitude, northeastern boundary latitude, northeastern boundary longitude, southwestern boundary latitude, and southwestern boundary longitude. The REP category—directly related to sustainable energy planning—was defined more comprehensively. Numerous criteria associated with solar, wind, biomass, geothermal, and hydropower potential were included. These criteria are presented in Table 3.

Table 3: REP Criteria

No	REP Criterion	No	REP Criterion
REP1	Radiation value (kWh/m²-year)	REP14	Agricultural land (ha) – 2020
REP2	Average temperature (°C)	REP15	Total cultivated agricultural land (ha) – 2021
REP3	Average maximum temperature (°C)	REP16	Artificial areas (ha) – 2020
REP4	Average minimum temperature (°C)	REP17	Forests and semi-natural areas (ha) – 2020
REP5	Average sunshine duration (hours)	REP18	Wetlands (ha) – 2020
REP6	Average wind speed (m/s)	REP19	Water surfaces (ha) – 2010–2020
REP7	Average wind power density (W/m²)	REP20	Irrigated land (ha) – 2020
REP8	Average wind capacity factor (%)	REP21	Forest biomass volume (m³) – 2021
REP9	Average number of rainy days	REP22	Total fertilizer consumption – 2021
REP10	Total precipitation amount	REP23	Crop production value (thousand TL) – 2021
REP11	Total average streamflow (m³/s) – 2021	REP24	Livestock production value (thousand TL) – 2021
REP12	Elevation	REP25	Production of cereals and other crops (tons) – 2021
REP13	Groundwater (hm³/year) – 2021	REP26	Average PM10 station values (µg/m³) – 2022

The dataset used in the study was obtained from publicly accessible databases in Türkiye (Turkish State Meteorological Service, 2024; Ministry of Energy and Natural Resources, 2024c; Ministry of Agriculture and Forestry, 2024; Ministry of Environment, Urbanization and Climate Change, 2024a; 2024b; General Directorate of State Hydraulic Works, 2024).

The dataset covers all 81 provinces of Türkiye. Although the inclusion of many criteria may raise concerns regarding dimensionality, this approach is necessary to accurately reflect the multidimensional structure of SEP. SEP is influenced not only by technical indicators related to solar, wind, hydropower, geothermal, and biomass resources but also by environmental, socioeconomic, and geographical factors.

Solar energy potential is largely determined by radiation levels; regions with higher radiation naturally exhibit stronger solar energy generation capacity. Average temperature, as well as maximum and minimum temperatures, affects the technical efficiency and seasonal performance of photovoltaic systems. Average sunshine duration is a direct determinant of regional solar potential. In the context of wind energy, average wind speed is a critical indicator. Furthermore, wind power density and capacity factor are essential parameters for assessing the technical and economic feasibility of wind energy in each region. Hydropower potential depends on factors such as precipitation levels, river discharge, and elevation. Biomass potential is influenced by agricultural land availability, crop production values, livestock activities, forest resources, and water availability—reflecting

the biophysical and agricultural capacity of a region. Environmental factors such as air quality (PM10) and forest stock serve as indicators both of sustainability performance and the need for renewable energy deployment. Among socioeconomic factors, population and population growth rate provide important insights into future regional energy demand. Geographical variables were incorporated into the model through latitude and longitude, given their direct influence on climatic conditions.

All variables were normalized to the range [0,1] using the min–max scaling method to eliminate differences in measurement units. To avoid redundancy, one variable from each pair with correlation coefficients above 0.95 was removed from the dataset following a correlation analysis.

Figure 7 presents the correlation matrix of REP criteria, and the variable codes shown in the figure correspond to those listed in Table 3.

Figure 8 presents the correlation matrix for the socioeconomic criteria. Here, SE1, SE2, SE3, SE4, SE5, SE6, SE7, and SE8 represent land area, population, annual population growth rate (‰), GDP per capita, total exports, total imports, industrial volume, and billed electricity consumption, respectively.

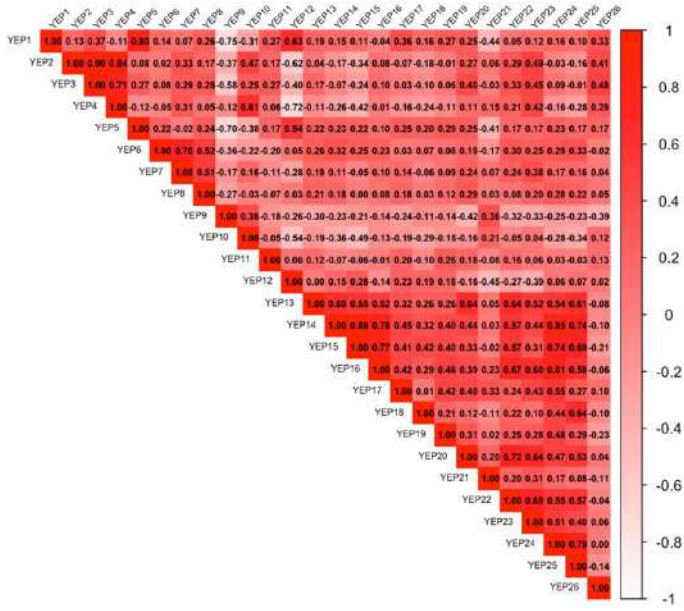

Figure 7: Correlation Matrix for REP Criteria

Figure 8: Correlation Matrix for Socioeconomic Criteria

3.2.2. FCM Findings

The FCM algorithm was implemented using three primary groups of criteria. FCM was chosen for the initial clustering phase because it can accommodate overlapping observations and assign each data point a degree of membership across clusters. This property is especially advantageous when working with complex datasets in which cluster boundaries are not sharply defined. The algorithm's flexibility made it suitable for capturing the nuanced relationships among socioeconomic characteristics, geographical location, and REP. In this context, FCM served to uncover the multidimensional structure underlying SEP.

The algorithm was executed with three different distance metrics—Euclidean, Manhattan, and Minkowski—and with several alternative cluster numbers. Since Türkiye is divided into seven geographical regions, the algorithm was tested using 4, 5, 6, 7, 8, 9, and 10 clusters. The analysis was carried out in RStudio version 2024.12.1+563, an integrated development environment that provides extensive support for statistical programming and visualization in R.

Because FCM is a fuzzy clustering technique, each observation receives a membership value between 0 and 1. This enabled the quantification of each province's degree of association with each cluster. The membership values produced by the algorithm were then consolidated into a single dataset. For instance, when the cluster count is 4, the resulting dataset contains 12 variables (4 clusters × 3 main criteria). Likewise, when 10 clusters are used, the dataset includes 30 variables.

Table 4 illustrates an example of the dataset generated by the FCM procedure, specifically the output obtained with Euclidean distance for k=4.

Table 4: Dataset for Euclidean Distance and k = 4

	Socioeconomic				Location				REP			
Province	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
P1	0.367	0.175	0.346	0.112	0.147	0.161	0.371	0.321	0.609	0.015	0.255	0.120
P2	0.068	0.126	0.041	0.765	0.098	0.340	0.380	0.182	0.057	0.002	0.930	0.011
P3	0.113	0.244	0.064	0.579	0.642	0.067	0.109	0.183	0.171	0.005	0.795	0.029
P4	0.099	0.173	0.060	0.668	0.111	0.509	0.220	0.160	0.142	0.006	0.820	0.032
P5	0.114	0.201	0.070	0.614	0.134	0.089	0.220	0.557	0.047	0.002	0.942	0.009
P6	0.092	0.212	0.050	0.646	0.121	0.152	0.405	0.322	0.096	0.004	0.880	0.020
P7	0.359	0.363	0.124	0.155	0.240	0.089	0.176	0.495	0.116	0.022	0.095	0.768
P8	0.252	0.206	0.368	0.174	0.413	0.119	0.189	0.280	0.634	0.011	0.279	0.077
P9	0.064	0.123	0.037	0.776	0.118	0.482	0.230	0.170	0.162	0.007	0.794	0.037
P10	0.158	0.093	0.683	0.066	0.108	0.506	0.225	0.161	0.128	0.005	0.838	0.028
P11	0.152	0.539	0.072	0.237	0.503	0.113	0.162	0.222	0.419	0.008	0.523	0.050
P12	0.166	0.108	0.647	0.079	0.567	0.097	0.140	0.196	0.697	0.007	0.247	0.049
P13	0.366	0.379	0.109	0.146	0.324	0.120	0.203	0.353	0.110	0.004	0.863	0.024
P14	0.080	0.144	0.048	0.728	0.082	0.599	0.193	0.127	0.046	0.002	0.943	0.009
P15	0.097	0.171	0.059	0.672	0.069	0.641	0.179	0.112	0.163	0.007	0.792	0.038
P16	0.143	0.394	0.073	0.390	0.794	0.041	0.064	0.102	0.272	0.007	0.679	0.042
P17	0.067	0.125	0.040	0.768	0.030	0.845	0.078	0.048	0.141	0.006	0.822	0.032
P18	0.048	0.092	0.028	0.831	0.093	0.567	0.201	0.139	0.147	0.006	0.814	0.033
P19	0.249	0.201	0.380	0.169	0.435	0.096	0.164	0.306	0.187	0.005	0.775	0.032
P20	0.141	0.537	0.066	0.256	0.510	0.097	0.154	0.239	0.061	0.002	0.924	0.012
P21	0.195	0.113	0.612	0.079	0.677	0.068	0.102	0.152	0.073	0.012	0.060	0.855
P22	0.212	0.113	0.599	0.077	0.477	0.125	0.171	0.227	0.220	0.006	0.737	0.037
P23	0.146	0.599	0.065	0.190	0.178	0.098	0.205	0.519	0.081	0.003	0.900	0.016
P24	0.246	0.543	0.080	0.132	0.132	0.121	0.305	0.442	0.278	0.009	0.659	0.054
P25	0.691	0.154	0.084	0.072	0.538	0.099	0.148	0.215	0.745	0.007	0.197	0.051
P26	0.120	0.219	0.072	0.588	0.074	0.611	0.196	0.120	0.228	0.007	0.724	0.041
P27	0.678	0.141	0.107	0.074	0.501	0.090	0.149	0.260	0.257	0.007	0.695	0.041

	Socioeconomic				Location				REP			
Province	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
P28	0.088	0.188	0.050	0.675	0.456	0.131	0.178	0.235	0.045	0.002	0.945	0.009
P29	0.076	0.138	0.046	0.740	0.071	0.566	0.238	0.125	0.071	0.003	0.912	0.014
P30	0.046	0.089	0.027	0.839	0.064	0.630	0.196	0.110	0.108	0.004	0.866	0.023
P31	0.161	0.437	0.081	0.321	0.072	0.650	0.167	0.111	0.097	0.004	0.879	0.021
P32	0.071	0.831	0.029	0.070	0.771	0.043	0.069	0.118	0.670	0.011	0.244	0.076
P33	0.086	0.160	0.052	0.703	0.115	0.252	0.414	0.219	0.396	0.028	0.264	0.312
P34	0.697	0.142	0.090	0.071	0.107	0.355	0.341	0.197	0.067	0.002	0.917	0.014
P35	0.189	0.573	0.075	0.164	0.084	0.531	0.241	0.144	0.145	0.006	0.817	0.033
P36	0.089	0.159	0.054	0.700	0.126	0.459	0.237	0.177	0.153	0.006	0.805	0.035
P37	0.148	0.488	0.072	0.292	0.144	0.208	0.378	0.270	0.520	0.014	0.374	0.092
P38	0.093	0.164	0.057	0.686	0.127	0.462	0.234	0.177	0.156	0.006	0.802	0.036
P39	0.132	0.574	0.062	0.232	0.495	0.097	0.157	0.250	0.034	0.001	0.958	0.007
P40	0.147	0.446	0.072	0.335	0.567	0.095	0.138	0.200	0.007	0.978	0.007	0.008
P41	0.302	0.397	0.123	0.178	0.496	0.117	0.165	0.222	0.136	0.028	0.114	0.723
P42	0.177	0.643	0.063	0.117	0.102	0.199	0.484	0.215	0.679	0.007	0.270	0.044
P43	0.201	0.113	0.607	0.079	0.293	0.113	0.202	0.391	0.131	0.005	0.837	0.028
P44	0.061	0.124	0.035	0.779	0.229	0.127	0.248	0.397	0.048	0.002	0.941	0.009
P45	0.090	0.173	0.053	0.684	0.118	0.486	0.227	0.169	0.150	0.006	0.809	0.034
P46	0.255	0.223	0.324	0.198	0.208	0.126	0.240	0.426	0.045	0.002	0.945	0.009
P47	0.081	0.158	0.047	0.714	0.093	0.106	0.480	0.321	0.783	0.007	0.162	0.048
P48	0.096	0.169	0.059	0.677	0.128	0.066	0.146	0.660	0.095	0.003	0.884	0.019
P49	0.775	0.097	0.078	0.051	0.473	0.125	0.172	0.230	0.361	0.008	0.582	0.049
P50	0.090	0.159	0.055	0.697	0.054	0.037	0.097	0.811	0.090	0.003	0.889	0.018
P51	0.095	0.168	0.058	0.679	0.125	0.241	0.400	0.234	0.137	0.006	0.827	0.031
P52	0.109	0.250	0.059	0.582	0.622	0.077	0.118	0.183	0.138	0.027	0.115	0.720
P53	0.243	0.318	0.157	0.282	0.283	0.109	0.207	0.401	0.619	0.014	0.254	0.113
P54	0.259	0.130	0.525	0.087	0.900	0.020	0.031	0.050	0.296	0.007	0.653	0.044
P55	0.066	0.123	0.039	0.773	0.089	0.359	0.382	0.170	0.159	0.005	0.810	0.027
P56	0.194	0.561	0.080	0.165	0.514	0.112	0.158	0.216	0.501	0.022	0.282	0.196
P57	0.081	0.149	0.049	0.721	0.093	0.537	0.224	0.146	0.159	0.005	0.809	0.028
P58	0.327	0.188	0.359	0.126	0.174	0.153	0.320	0.353	0.644	0.011	0.268	0.077
P59	0.101	0.063	0.790	0.046	0.481	0.117	0.169	0.233	0.311	0.008	0.635	0.046
P60	0.078	0.143	0.047	0.732	0.074	0.645	0.169	0.113	0.140	0.006	0.823	0.031

	Socioeconomic				Location				REP			
Province	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
P61	0.090	0.160	0.055	0.695	0.109	0.094	0.288	0.510	0.112	0.004	0.860	0.024
P62	0.085	0.155	0.051	0.708	0.135	0.117	0.319	0.429	0.073	0.003	0.910	0.015
P63	0.512	0.274	0.102	0.112	0.112	0.301	0.373	0.214	0.067	0.002	0.919	0.012
P64	0.120	0.302	0.064	0.515	0.125	0.186	0.427	0.262	0.291	0.009	0.650	0.050
P65	0.175	0.617	0.067	0.141	0.095	0.533	0.223	0.150	0.079	0.003	0.902	0.016
P66	0.328	0.445	0.094	0.135	0.622	0.074	0.116	0.189	0.686	0.011	0.225	0.078
P67	0.699	0.140	0.091	0.069	0.133	0.194	0.383	0.290	0.434	0.008	0.507	0.051
P68	0.062	0.115	0.037	0.786	0.095	0.555	0.207	0.143	0.140	0.006	0.823	0.031
P69	0.235	0.127	0.552	0.087	0.176	0.171	0.305	0.348	0.137	0.006	0.827	0.031
P70	0.143	0.533	0.068	0.256	0.075	0.158	0.594	0.174	0.087	0.003	0.894	0.017
P71	0.177	0.281	0.112	0.430	0.106	0.388	0.324	0.183	0.372	0.011	0.551	0.066
P72	0.054	0.105	0.031	0.811	0.110	0.505	0.225	0.161	0.135	0.005	0.831	0.029
P73	0.082	0.160	0.048	0.711	0.512	0.113	0.158	0.217	0.383	0.028	0.264	0.324
P74	0.077	0.827	0.030	0.066	0.098	0.165	0.499	0.239	0.081	0.003	0.899	0.017
P75	0.671	0.162	0.090	0.077	0.096	0.496	0.249	0.159	0.116	0.004	0.860	0.021
P76	0.079	0.156	0.045	0.720	0.053	0.693	0.164	0.091	0.162	0.007	0.794	0.037
P77	0.110	0.260	0.059	0.571	0.664	0.069	0.106	0.161	0.156	0.005	0.814	0.027
P78	0.096	0.171	0.059	0.674	0.117	0.491	0.226	0.166	0.127	0.005	0.841	0.027
P79	0.079	0.151	0.046	0.725	0.648	0.074	0.111	0.167	0.327	0.008	0.616	0.049
P80	0.115	0.290	0.061	0.535	0.102	0.090	0.265	0.543	0.112	0.004	0.859	0.024
P81	0.615	0.155	0.144	0.086	0.382	0.112	0.185	0.322	0.212	0.006	0.745	0.037

The remaining FCM results are not included in the book due to page limitations.

3.2.3. K-Means Findings

The new datasets obtained from the FCM results were subsequently subjected to crisp clustering using the K-Means algorithm. This process ensured that the clustering analysis could represent the multidimensional structure of the energy system independently of the number of original variables. In this way, a novel two-stage

hierarchical clustering approach was proposed by integrating fuzzy and hard clustering methods.

As an example, the visualizations of the 4-cluster and 10-cluster solutions generated using the Euclidean distance metric are presented in Figure 9 and Figure 10, respectively.

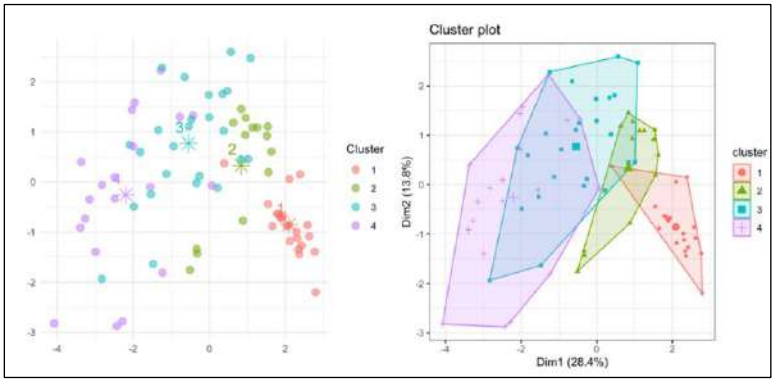

Figure 9: $k = 4$, Metric = Euclidean

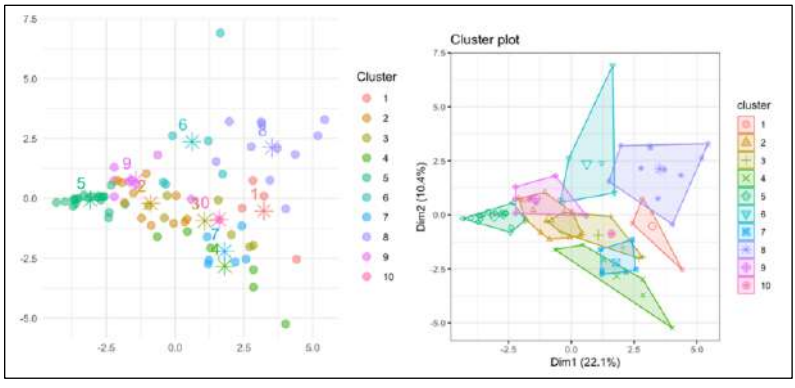

Figure 10: k =10, Metric = Euclidean

For the same number of clusters, the clustering results obtained using the Manhattan metric are presented in Figures 11 and 12, respectively.

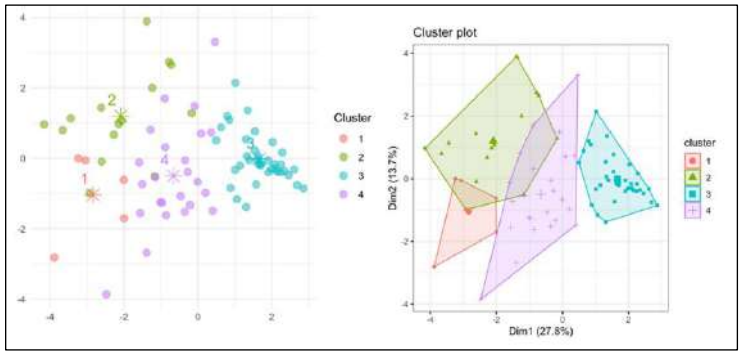

Figure 11: k =4, Metric = Manhattan

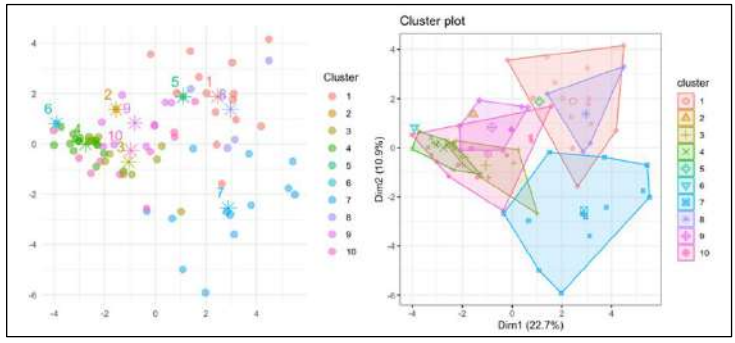

Figure 12: k =10, Metric = Manhattan

Finally, the clustering results obtained using the Minkowski metric are presented in Figures 13 and 14, respectively.

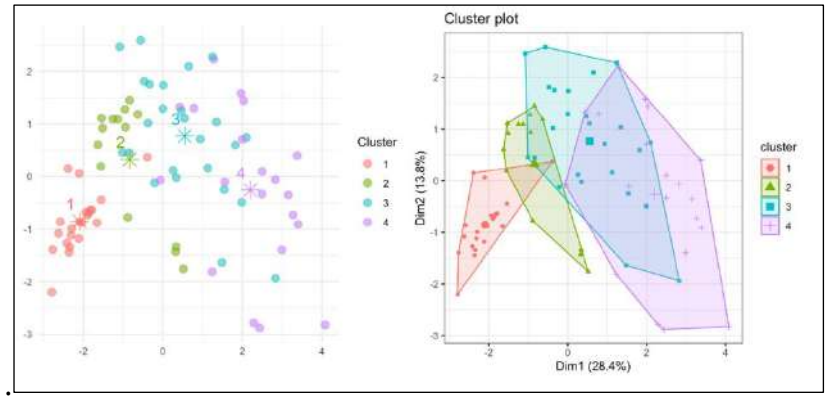

Figure 13: k =4, Metric = Minkowski

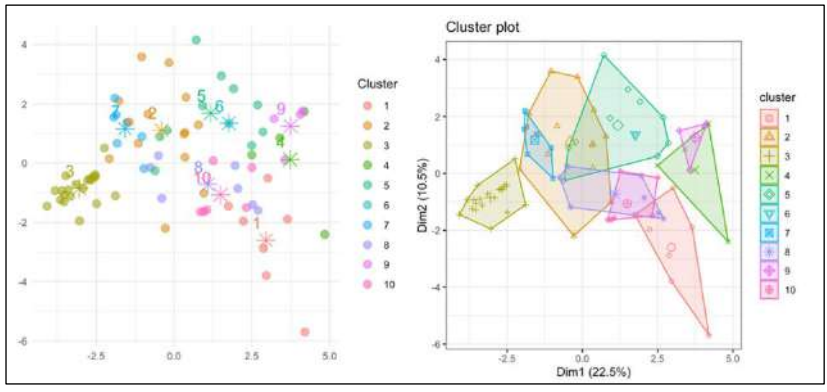

Figure 14: k =10, Metric = Minkowski

All cluster visualizations obtained using the K-Means algorithm are presented in Appendix A (Figures A.1, A.2, A.3, A.4, A.5, A.6, A.7, A.8, A.9, A.10, A.11, A.12, A.13, A.14, A.15, A.16, A.17, A.18, A.19, A.20, and A.21).

3.2.4. Analysis of Clustering Performance

In the second stage of the methodology, the effectiveness of the clustering results was evaluated to determine the final cluster

structure. The class labels obtained from the clustering process were added to the dataset, and the problem was then treated as a classification task. Using the KNN, SVM, RF, and XGBoost algorithms, classification errors were computed. Additionally, voting and stacking ensemble learning techniques were applied to examine whether classification performance could be improved. The results were compared with widely used clustering performance metrics—Silhouette and CHI indices—and the applicability of classification algorithms as a measure of clustering accuracy was critically assessed. The cluster configuration with the lowest classification error was selected as the final clustering solution. Furthermore, the effects of cluster number and distance metric on clustering performance were thoroughly analyzed.

The classification algorithms utilized in this study were chosen based on their distinct strengths. KNN was preferred due to its simplicity and effectiveness. RF was selected for its strong generalization ability and robustness against overfitting, despite being based on decision trees. SVM was employed because of its high performance in cases where data are not linearly separable and its ability to maximize the margin between classes. XGBoost was chosen due to its high predictive accuracy, built-in regularization mechanisms that prevent overfitting, and computational efficiency.

The CHI and Silhouette scores for all cluster configurations are presented in Table 5.

Table 5: CHI ve Silhoutte Scores for all Clusters

Metric	Cluster Number	Average Silhouette Score	CHI	Silhouette Score Ranking	CHI Ranking
Euclidean	4	0.32	31.72	1	1
Euclidean	5	0.27	25.05	4	3
Euclidean	6	0.26	19.22	5	8
Euclidean	7	0.17	13.46	20	15
Euclidean	8	0.21	13.25	16	16
Euclidean	9	0.21	11.85	17	20
Euclidean	10	0.14	9.55	21	21
Manhattan	4	0.26	22.64	10	6
Manhattan	5	0.26	24.15	11	4
Manhattan	6	0.26	18.01	7	11
Manhattan	7	0.26	19.57	8	7
Manhattan	8	0.26	19.22	5	8
Manhattan	9	0.23	14.26	13	14
Manhattan	10	0.20	12.36	19	19
Minkowski	4	0.28	23.13	2	5
Minkowski	5	0.25	18.57	12	10
Minkowski	6	0.28	26.18	3	2
Minkowski	7	0.26	17.87	9	12
Minkowski	8	0.22	14.77	15	13
Minkowski	9	0.21	12.58	18	17
Minkowski	10	0.23	12.56	14	18

According to Table 5, the most effective clustering configuration was achieved when the Euclidean distance metric was used with four clusters. Nonetheless, some clustering outcomes showed only minor performance differences. Additionally, discrepancies between the rankings of the validity indices are noteworthy. For example, while the Silhouette score identified the Minkowski distance with four clusters as the second-best solution, the CHI index ranked the Minkowski metric with six clusters as the second-best alternative. This inconsistency indicates that relying on a single validity measure may not provide a sufficiently robust assessment when comparing various

distance metrics and cluster counts. Therefore, this study incorporates classification algorithms as an additional approach to more accurately determine the optimal number of clusters.

For the classification procedures, 75% of the data was used for model training and the remaining 25% for testing. Stratified sampling ensured that each cluster (class) was proportionally represented in both subsets, helping to reduce potential bias. Multiple precautions were taken to mitigate overfitting during model development. Cross-validation was employed for hyperparameter tuning across all classifiers. In the KNN algorithm, the optimal value of k was selected. For the SVM model, the regularization parameter C and kernel width σ were optimized using grid search. In XGBoost and other ensemble techniques, key hyperparameters such as maximum tree depth (d_{max}) and learning rate (η) were fine-tuned through cross-validation to strike a balance between model complexity and generalization ability. An early-stopping rule was applied to halt the training process once the validation error stopped decreasing, preventing unnecessary complication of the model.

Given the relatively limited sample size, ensemble learning methods were adopted to improve predictive performance. In total, 21 clustering experiments were conducted by combining seven different cluster numbers with three distance metrics; however, not all cluster configurations yielded valid outcomes.

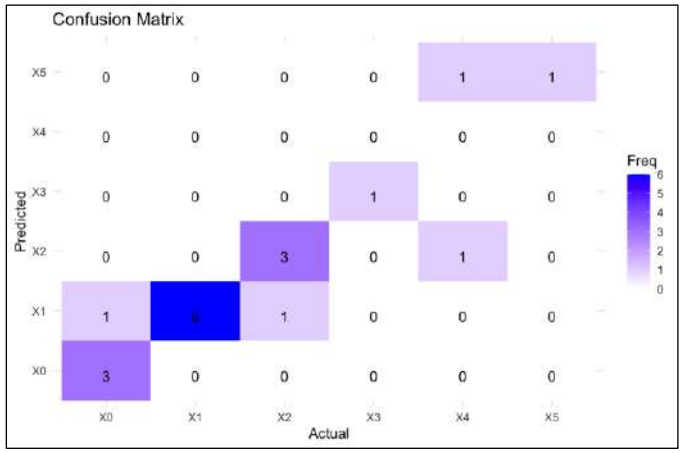

Figure 15: Confusion Matrix for k = 6, Metric = Euclidean

For instance, Figure 15 displays the confusion matrix produced by the KNN classifier when the Euclidean distance metric was used and the number of clusters was set to k=6. As illustrated in the figure, the KNN algorithm identified 5 classes in the dataset, even though the original clustering specified 6 groups. Classification outputs that exhibited such inconsistencies were therefore removed from further evaluation.

The classification errors observed in the analysis were not regarded as a shortcoming of the methodology. Instead, they were interpreted as indicators of clustering quality. When classification errors are low, this suggests that the resulting clusters are clearly separated and readily distinguishable, which in turn supports the appropriateness of the chosen clustering settings and distance metrics. Conversely, high classification error rates may point to overlaps between clusters or challenges in differentiating the underlying variables. The results

obtained using the KNN classifier are summarized in Table 6. For each clustering configuration, cross-validation was performed by adjusting the number of neighbors k from 1 to 20.

Table 6: Result of KNN

Number of clusters	Distance metric	MCC	Kappa	Accuracy	Best k
k=4	Euclidean	0.86	0.86	0.90	2
k=5	Euclidean	0.94	0.93	0.95	2
k=7	Euclidean	0.58	0.58	0.74	5
k=8	Euclidean	0.70	0.74	0.78	4
k=4	Manhattan	0.76	0.75	0.84	4
k=5	Manhattan	0.86	0.86	0.89	3
k=7	Manhattan	0.73	0.72	0.76	2
k=4	Minkowski	0.87	0.87	0.90	4
k=5	Minkowski	0.87	0.86	0.89	2
k=7	Minkowski	0.69	0.68	0.74	2

The results obtained from the SVM method are shown in Table 7. In the SVM approach, the RBF kernel was selected because, when appropriately tuned, it typically yields strong classification performance. The optimal values of the regularization parameter C and the kernel width parameter σ were identified using a grid search procedure, while the number of support vectors was determined automatically by the model. A smaller C value permits more classification errors and thus reduces the likelihood of overfitting, whereas a larger C value forces the model to fit the training data more closely, increasing the risk of overfitting. Similarly, a small σ value produces a more complex and flexible decision boundary (Cássia, 2024).

Table 7: SVM Results

Number of clusters	Distance metric	Regularization parameter (*C*)	σ	Number of support vectors	MCC	Kappa	Accuracy
k=4	Euclidean	3.5	0.02	35	0.87	0.87	0.91
k=5	Euclidean	2.5	0.02	44	0.79	0.76	0.81
k=6	Euclidean	1.5	0.02	51	0.80	0.77	0.82
k=4	Manhattan	2.5	0.02	38	0.92	0.93	0.95
k=5	Manhattan	7	0.01	44	0.68	0.69	0.77
k=6	Manhattan	2	0.02	48	0.80	0.77	0.82
k=7	Manhattan	1.5	0.02	51	0.84	0.83	0.86
k=4	Minkowski	10	0.02	37	0.88	0.87	0.90
k=5	Minkowski	4.5	0.02	41	0.86	0.82	0.86
k=6	Minkowski	5.5	0.01	47	0.67	0.76	0.81

The results obtained using the RF method are presented in Table 8.

Table 8: RF Results

Number of clusters	Distance metric	*mtry*	MCC	Kappa	Accuracy
k=4	Euclidean	4	0.94	0.94	0.95
k=5	Euclidean	2	0.94	0.93	0.95
k=6	Euclidean	4	0.74	0.71	0.77
k=4	Manhattan	3	0.81	0.80	0.86
k=5	Manhattan	4	0.68	0.67	0.77
k=6	Manhattan	2	0.78	0.77	0.82
k=7	Manhattan	4	0.67	0.78	0.83
k=4	Minkowski	3	0.87	0.87	0.90
k=5	Minkowski	4	0.74	0.71	0.77
k=6	Minkowski	6	0.89	0.89	0.91

In Random Forest, the hyperparameter *mtry*—which specifies the number of variables considered at each split—was tuned through cross-validation. A smaller *mtry* increases the diversity among trees and helps mitigate overfitting. Conversely, a larger *mtry* value can improve the accuracy of individual trees but also raises the correlation

between them, potentially diminishing the ensemble's generalization capability.

The XGBoost results and the corresponding hyperparameter configurations are presented in Table 9.

Table 9: XGBoost Results

Number of clusters	Distance metric	MCC	Kappa	Accuracy	Best parameter combination
k=4	Euclidean	0.7559	0.8679	0.9048	d_{max}= 3, w_{min}= 1, s = 1, s_f = 0.75, γ = 0, η = 0.05
k=5	Euclidean	0.9366	0.9333	0.9474	d_{max}= 9, w_{min}= 1, s = 1, s_f = 0.5, γ = 0.2, η = 0.1
k=7	Euclidean	0.7284	0.7224	0.7647	d_{max}= 9, w_{min}= 1, s = 0.75, s_f = 0.5, γ = 0.1, η = 0.1
k=8	Euclidean	0.7169	0.7104	0.7500	d_{max}= 9, w_{min}= 1, s = 1, s_f = 1, γ = 0, η = 0.05
k=4	Manhattan	0.9331	0.9298	0.9500	d_{max}= 9, w_{min}= 1, s = 0.8, s_f = 1, γ = 0.1, η = 0.1
k=6	Manhattan	0.6667	0.7627	0.8095	d_{max} = 6, w_{min}= 1, s = 1, s_f = 1, γ = 0, η = 0.1
k=7	Manhattan	0.8281	0.8779	0.8571	d_{max}= 3, w_{min}= 1, s = 1, s_f = 1, γ = 0.2, η = 0.1
k=4	Minkowski	0.9386	0.9359	0.9545	d_{max}= 9, w_{min}= 1, s = 1, s_f = 1, γ = 0.1, η = 0.1
k=5	Minkowski	0.8070	0.8014	0.8421	d_{max}= 9, w_{min}= 1, s = 1, s_f = 1, γ = 0, η = 0.1
k=6	Minkowski	0.8373	0.8281	0.8636	d_{max}=9, w_{min}= 1, s = 1, s_f = 1, γ = 0, η = 0.1

In this framework, d_{max} specifies the maximum depth of the trees; increasing this parameter creates more complex models but also heightens the likelihood of overfitting. The parameter w_{min} represents the minimum total Hessian (second derivative) weight required for a node to undergo a split. Conceptually, it functions as the model's mathematical mechanism for determining whether a split is justified, thereby regulating tree growth and helping to control overfitting. Lower values make the model more responsive to variations in the

data, while higher values tend to enhance its ability to generalize. The parameter *s* refers to the fraction of data samples randomly selected for training each tree, whereas *sf* indicates the fraction of features randomly chosen for the same purpose. The parameter γ defines the minimum loss reduction needed to allow a leaf node to split; larger values require greater improvement in the objective function before a split is permitted. The learning rate η determines how much each additional tree contributes to the model's final prediction. Although smaller values of η often produce more stable and potentially more accurate models, they can substantially increase the training duration (GitHub, 2024; Chen & Guestrin, 2016).

Additionally, within the proposed methodology, the predictions of the RF and XGBoost algorithms were combined to create a stacked dataset. Then, MLR, SVM, and KNN models were used as meta-models, and accuracy values were calculated. The meta-models that produced the best results in the stacked method and their performance metrics are presented in Table 10.

Table 10: Results of Ensemble Model 1

Number of clusters	Distance metric	Best meta-model	MCC	Kappa	Accuracy
k=4	Euclidean	MLR-KNN	0.76	0.87	0.90
k=5	Euclidean	MLR-KNN-SVM	0.85	0.80	0.86
k=4	Manhattan	MLR-KNN-SVM	0.85	0.93	0.95
k=5	Manhattan	MLR-KNN-SVM	0.78	0.80	0.86
k=6	Manhattan	MLR-KNN-SVM	0.67	0.76	0.81
k=7	Manhattan	MLR-SVN	0.83	0.88	0.86
k=4	Minkowski	MLR-KNN	0.94	0.93	0.95
k=5	Minkowski	SVM	0.80	0.79	0.76
k=6	Minkowski	MLR-KNN-SVM	0.84	0.81	0.86

The cross-validation procedures previously applied to the RF, XGBoost, SVM, and KNN models were repeated, and the optimal hyperparameters identified through these validations were incorporated into the final model. However, the detailed cross-validation outputs were not included in Table 10 due to page layout constraints. In the final stage, predictions generated by the RF and XGBoost algorithms were combined using a majority voting ensemble strategy, and the corresponding results are presented in Table 11. As in earlier steps, the necessary cross-validation processes for tuning the RF and XGBoost parameters were also carried out in this part of the analysis.

Table 11: Results of Ensemble Model 2

Number of clusters	Distance metric	MCC	Kappa	Accuracy
k=4	Euclidean	0.94	0.94	0.95
k=5	Euclidean	0.85	0.93	0.95
k=6	Euclidean	1.00	1.00	1.00
k=7	Euclidean	0.87	0.87	0.89
k=8	Euclidean	0.81	0.80	0.83
k=4	Manhattan	0.78	0.76	0.84
k=5	Manhattan	1.00	1.00	1.00
k=6	Manhattan	1.00	1.00	1.00
k=7	Manhattan	0.88	0.87	0.89
k=8	Manhattan	0.72	0.71	0.75
k=10	Manhattan	0.52	0.50	0.56
k=4	Minkowski	1.00	1.00	1.00
k=5	Minkowski	0.93	0.93	0.95
k=6	Minkowski	1.00	1.00	1.00
k=7	Minkowski	0.43	0.42	0.53

A review of Tables 6 through 11 shows that the KNN, SVM, RF, and XGBoost models each yielded 10 valid classification results, while Ensemble Model 1 produced 9 valid classifications and Ensemble

Model 2 produced 15. With the exception of a single valid outcome obtained from Ensemble Model 2 (k = 10, Manhattan), none of the methods generated acceptable classifications for cluster sizes of 9 or 10. Table 12 summarizes the mean MCC, Kappa, and accuracy values computed for each model.

Table 12: Comparison of Distance Metrics

Method	Number of Predictions	Distance Metric	Average MCC	Average Kappa	Average Accuracy
KNN	4	Euclidean	0.77	0.78	0.84
SVM	3	Euclidean	0.82	0.80	0.85
RF	3	Euclidean	0.87	0.86	0.89
XGBoost	4	Euclidean	0.78	0.81	0.84
Ensemble Model 1	2	Euclidean	0.80	0.84	0.88
Ensemble Model 2	5	Euclidean	0.89	0.91	0.93

Based on these findings, Ensemble Model 2 was selected as the basis for determining the optimal number of clusters, as it outperformed the other approaches. The most successful clustering structures were identified through the results obtained by this ensemble. Initially, when the Euclidean distance metric was used, the 6-cluster solution achieved perfect classification performance, with MCC, Kappa, and accuracy values all equal to 1.00. Under the Manhattan distance metric, both the 5-cluster and 6-cluster solutions likewise produced perfect results across all performance measures. Similarly, the Minkowski metric yielded two fully accurate solutions, with the 4-cluster and 6-cluster configurations each reaching MCC, Kappa, and accuracy values of 1.00.

Table 13 provides a comparison between the clustering structures identified as completely successful by Ensemble Model 2 and the Silhouette and CHI indices reported in Table 5.

Table 13: Evaluation of Ensemble Model 2 in Relation to the Silhouette and CHI

Cluster number	Distance metric	MCC	Kappa	Accuracy	Average Silhouette score ranking	CHI ranking
k=6	Euclidean	1.00	1.00	1.00	5	8
k=5	Manhattan	1.00	1.00	1.00	11	4
k=6	Manhattan	1.00	1.00	1.00	7	11
k=4	Minkowski	1.00	1.00	1.00	2	5
k=6	Minkowski	1.00	1.00	1.00	3	2

As shown in Table 13, the clustering obtained using the Minkowski distance metric with a cluster count of 6 was selected as the final solution, as it consistently appears within the top three ranks across all evaluation criteria.

Table 14: Final Clustering Results

Cluster	Province Group
Cluster a	Adana, Antalya, Burdur, Denizli, Gaziantep, Hatay, Isparta, Kahramanmaraş, Kayseri, Mersin, Şanlıurfa, Osmaniye
Cluster b	Adıyaman, Ağrı, Aksaray, Ardahan, Batman, Bayburt, Bingöl, Çorum, Diyarbakır, Elazığ, Erzincan, Hakkari, Iğdır, Karaman, Kars, Kırıkkale, Kırşehir, Malatya, Mardin, Muş, Nevşehir, Niğde, Tunceli, Van, Şırnak, Bitlis, Siirt
Cluster c	Amasya, Ankara, Artvin, Çankırı, Erzurum, Gümüşhane, Karabük, Konya, Rize, Sivas, Tokat, Yozgat
Cluster d	Bursa, İstanbul, Kocaeli
Cluster e	Afyonkarahisar, Aydın, Balıkesir, Bilecik, Bolu, Çanakkale, Eskişehir, İzmir, Kütahya, Manisa, Muğla, Uşak, Yalova
Cluster f	Bartın, Düzce, Edirne, Giresun, Kastamonu, Kırklareli, Ordu, Sakarya, Samsun, Sinop, Tekirdağ, Trabzon, Zonguldak

The list of provinces assigned to each cluster is given in Table 14, and the spatial representation of the clusters on the map of Türkiye is illustrated in Figure 16.

Figure 16: Visualization of the Final Clustering on the Map of Türkiye

Table 15 presents the cluster means for socioeconomic criteria, while Table 16 reports the cluster means for selected REP criteria.

As presented in Table 15, the average values of the socioeconomic criteria offer important insights into the economic structures, energy demand levels, and regional development dynamics of the clusters. Evaluating each cluster based on indicators such as industrial capacity, population size, foreign trade volume, per capita income, and electricity consumption provides a more accurate understanding of

regional energy demand and supports more effective planning of sustainable energy investments.

Table 15: Cluster Means for Socioeconomic Criteria

Cluster	Area	Population	Annual population growth rate	GDP	Total exports	Total imports	Industrial Volume	Electricity consumption
a	12055	1486881	12	63942	2696762	2626115	5132260	4567433
b	8231	454644	3	48849	248297	184956	741051	776142
c	14882	993522	11	63379	1344522	1186612	4107826	2330135
d	6557	7060581	19	130287	47871234	56412045	52091641	20016168
e	9912	1030761	18	83178	1857114	1371351	6502726	3376702
f	5985	631005	9	68489	928029	707197	3399168	1922170

Cluster d clearly distinguishes itself from all other clusters. This cluster represents the economic and demographic center of Türkiye, characterized by extremely high population density, a very large industrial capacity, a substantial share of the country's total exports and imports, and exceptionally high electricity consumption. These characteristics demonstrate the prevalence of intensive industrial activities, large-scale production facilities, organized industrial zones, and high-technology sectors. Consequently, this cluster is critical in terms of energy supply security and requires both strong electricity generation and transmission infrastructure. Clusters a and e have moderate levels of industrial activity but stand out with strong agricultural production, commercial capacity, and service sector dynamics. The relatively high electricity consumption observed in these clusters results from mixed economic structures in which

industrial and agricultural activities coexist. In these regions, renewable energy investments may be strengthened through distributed energy systems and flexible resources such as solar and wind power, especially to support areas with concentrated demand. Cluster b exhibits lower values in both population and economic indicators. Electricity consumption and industrial activities are relatively limited, which reflects a predominantly rural structure with low-density economic activity. In such areas, small-scale and locally targeted renewable energy projects—such as biomass systems, small hydropower plants, or micro-scale solar installations—are more feasible. This approach can enhance energy accessibility while contributing to regional economic development. Clusters c and f represent regions with moderate economic activity and industrial capacity, accompanied by a more balanced population structure. These clusters contain diversified agricultural and industrial elements, leading to a more balanced overall energy demand profile. As a result, hybrid systems combining different renewable energy technologies—such as wind, solar, and biomass—may be particularly suitable for these areas. Overall, the clustering analysis reveals significant regional variations in Türkiye's energy demand structure. Industrially intensive regions require large-scale generation facilities and robust grid infrastructure, while low-population rural regions benefit more from flexible, localized, and sustainable energy solutions. Therefore, shaping energy policies and investment strategies according to these regional differences is crucial for ensuring both economic efficiency and long-term sustainability.

Table 16: Cluster Averages for Selected REP Criteria

	Average temperature	Average sunshine duration	Average wind speed	Average power density	Total precipitation	Total average discharge	Elevation	Groundwater	Agricultural land	Crop production value	Livestock production value
a	16.45	7.16	4.57	157.23	668.02	157.48	470.17	412.45	334148	8431778	3383013
b	11.81	6.60	4.47	137.55	545.08	170.28	1144.57	173.30	287033	1856390	2613220
c	11.27	6.12	4.15	118.62	617.09	182.03	836.17	172.81	556745	3709095	3906816
d	14.90	4.27	4.83	160.58	727.70	168.70	92.67	192.44	188618	4264319	2226005
e	13.92	6.63	4.79	171.27	638.95	106.75	445.85	283.25	379918	4264451	3926453
f	13.71	5.08	4.26	143.93	829.35	112.40	100.31	119.88	233940	3113323	1561956

The average values presented in Table 16 reveal the distinct renewable energy characteristics of each cluster. When temperature, solar radiation, wind indicators, hydrological parameters, and agricultural production variables are jointly considered, it becomes evident that the clusters exhibit significantly different profiles in terms of renewable energy planning and resource suitability.

Cluster a exhibits relatively high temperatures and long sunshine durations compared to the other clusters, making it highly suitable for solar energy investments. The presence of moderate wind speeds and power density further indicates that certain areas within this cluster may also support wind energy applications. The moderate levels of precipitation and river discharge suggest that hydropower potential is present but not dominant. Additionally, the extensive agricultural land

and high levels of crop production highlight that biomass energy can serve as a meaningful option for this cluster. Cluster b displays lower levels of temperature and sunshine duration, which limits its suitability for solar energy applications. However, the high altitude strengthens the hydrological structure of the region, and the relatively high discharge values suggest considerable hydropower potential. The cluster's extensive agricultural land and significant livestock production also indicate a favorable environment for biomass energy. Wind speeds are moderate, making wind power feasible in selected locations. Cluster c presents medium-level temperature and sunshine duration, indicating balanced potential for both solar and wind energy. With wind speeds above 4 m/s and reasonable power density, certain areas are suitable for wind farm development. Furthermore, high discharge levels coupled with notable elevation differences strengthen the hydropower potential of this cluster. Large agricultural areas and substantial crop production also provide a solid foundation for biomass energy. Overall, Cluster c represents a hybrid renewable energy profile suitable for integrated energy strategies. Cluster d has the lowest average sunshine duration among all clusters, limiting its suitability for solar power. In contrast, its wind speed and power density are notably high, positioning wind energy as the primary renewable resource for this region. Although precipitation and discharge values are relatively high, the low elevation limits hydropower feasibility. Agricultural and livestock indicators are moderate, suggesting that biomass potential exists but is not dominant. Overall, this cluster stands out primarily due to its strong wind energy

characteristics. Cluster e demonstrates a highly balanced structure with strong performance across multiple renewable energy indicators. High sunshine duration, strong wind speeds, and high-power density make this cluster one of the most favorable regions in Türkiye for both solar and wind energy. Additionally, extensive agricultural land and exceptionally high crop production values highlight biomass as another strong alternative. Although hydrological parameters are moderate, hydropower serves as a complementary rather than primary source. This cluster is well-suited for diverse and integrated renewable energy investments. Cluster f exhibits relatively low sunshine duration but possesses wind speeds and power density values that support wind energy development in specific areas. High rainfall and particularly high discharge values suggest that hydropower constitutes the most promising resource for this cluster. The relatively smaller agricultural land area and lower crop production values limit biomass potential. Given its geographical characteristics, hydropower and wind energy emerge as the dominant renewable resources for this cluster.

Overall, the clusters reveal distinct renewable energy profiles, with each cluster demonstrating comparative advantages in different energy sources. Some clusters are dominated by a single resource (e.g., wind in Cluster d; hydropower in Clusters b and f), while others exhibit broad multi-resource suitability (e.g., Cluster e). These findings highlight the necessity of incorporating regional differences into Türkiye's renewable energy transition strategies and suggest that

region-specific energy policies may lead to more efficient and sustainable outcomes.

4. CONCLUSION AND DISCUSSION

Climate change, rising energy demand, and the environmental impacts of fossil fuels have become the key drivers compelling countries to develop sustainable energy policies. In this context, the effective evaluation of renewable energy resources and the integration of scientific methods into energy planning processes are not only environmental necessities but also strategic requirements for enhancing economic competitiveness and ensuring long-term social welfare. Türkiye's diverse geographical, climatic, and socioeconomic characteristics make it difficult to conduct energy planning through a homogeneous structure. Therefore, systematically analyzing regional differences in energy potential provides decision makers with valuable insights into which renewable energy source is most suitable for each region.

This study was designed with this need in mind and aimed to examine Türkiye's provinces based on their sustainable energy potential through a comprehensive machine learning–based clustering framework. Using socioeconomic structure, geographical characteristics, and REP as the primary criteria groups, the provinces were first clustered using the FCM algorithm. FCM was preferred due to its ability to assign membership degrees to multiple clusters, an important feature when analyzing multidimensional problems such as sustainable energy potential, where cluster boundaries are not sharply

defined. The algorithm was executed using three different distance metrics (Euclidean, Manhattan, and Minkowski) and seven different cluster numbers (from 4 to 10). The membership degrees obtained from each clustering configuration were then combined to form a new dataset, which was subsequently subjected to a second-stage crisp clustering process using the K-Means algorithm. In this two-tiered hierarchical structure, the fuzzy nature of the data was first captured, and then more explicit cluster structures were derived from this representation.

To test the reliability of the clustering results, the cluster labels obtained from K-Means were incorporated into the dataset as response variables. Subsequently, several classification algorithms including KNN, SVM, RF, and XGBoost were applied, and cluster validity was assessed based on classification errors. Furthermore, ensemble learning strategies—specifically voting and stacking approaches that combined the predictive outputs of RF and XGBoost—were utilized to enhance classification performance. Supported by cluster validity metrics such as Silhouette and CHI indices, the structure with the lowest classification error was selected as the final clustering solution.

According to these evaluations, the most successful clustering was achieved when using the Minkowski distance metric with six clusters. Among the identified clusters, Cluster d—which includes Bursa, İstanbul, and Kocaeli—stands out as Türkiye's most strategically significant region in terms of energy demand. This cluster exhibits the highest population, the largest industrial capacity, and the most

intensive electricity consumption across the country. These characteristics are directly linked to the region's high concentration of industrial facilities, strong economic activity, and developed urban infrastructure. As a result, this cluster requires prioritized consideration in energy supply planning, including the deployment of large-scale and reliable energy sources to support its substantial and continuously increasing energy demand.

The findings highlight that Türkiye's sustainable energy potential is far from uniform and that provinces with similar characteristics tend to group together in meaningful ways. This reinforces the importance of adopting region-specific strategies rather than relying on a single, national-level renewable energy policy. The results also demonstrate that machine learning–based clustering can provide significant support to policymakers by revealing regional energy profiles, identifying critical demand centers, and offering a scientific foundation for targeted investment planning.

For future research, each cluster's unique energy profile could be analyzed more comprehensively using MCDM techniques. Conducting MCDM analyses separately for each cluster would provide deeper insights into the most suitable renewable energy sources at the regional level and would help develop more precise and actionable energy investment strategies. Additionally, monitoring changes in cluster structures over time and performing long-term dynamic analyses would enable researchers to examine how climate change, demographic shifts, and economic transformations influence

regional energy potentials. Integrating cost, environmental impacts, carbon emissions, technical feasibility, and social acceptability indicators into the regional analyses may also support the development of more holistic and multidimensional energy planning models in future studies.

ACKNOWLEDGEMENTS

This book was developed by expanding and revising a section of Selen AVCI AZKESKİN's doctoral dissertation titled "**Assessment of Renewable Energy Alternatives Based on Regional Energy Potential Through Clustering and Fuzzy Multi-Criteria Decision Making: A Site Selection Model for Wind Power Plants**" (**Thesis No: 970827**).

REFERENCES

Abdullah, L., & Najib, L. (2016). Sustainable energy planning decision using the intuitionistic fuzzy analytic hierarchy process: choosing energy technology in Malaysia. *International Journal of Sustainable Energy*, *35*(4), 360-377.

Afsordegan, A., Sánchez, M., Agell, N., Zahedi, S., & Cremades, L. V. (2016). Decision making under uncertainty using a qualitative topsis method for selecting sustainable energy alternatives. *International Journal of Environmental Science and Technology, 13*, 1419-1432.

Andrade, R., Faria, W. M., Silva, S., Chakraborty, S., & Curi, N. (2020). Prediction of soil fertility via portable x-ray fluorescence (PXRF) spectrometry and soil texture in the brazilian coastal plains. *Geoderma, 357*, 113960.

Avcı Azkeskin, S. & Aladağ, Z. (2025). Evaluating regional sustainable energy potential through hierarchical clustering and machine learning. *Environmental Research Communications*, *7*(1), 015002.

Bezdek, J.C. (1981). Modified objective function algorithms. In: *Pattern recognition with fuzzy objective function algorithms. advanced applications in pattern recognition (p.* 155-201). Boston: Springer.

Breiman, L. (2001). Random forests. *Mach Learn, 45*, 5–32.

Cássia, S. (2024). *Understanding SVM Hyperparameters*. https://stackabuse.com/understanding-svm-hyperparameters/

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods, 3*(1), 1-27.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USD, August 13-17, 785-794.

Coughenour, C., Paz, A., Fuente-Mella, H., & Singh, A. (2015). Multinomial logistic regression to estimate and predict perceptions of bicycle and transportation infrastructure in a sprawling metropolitan area. *Journal of Public Health, 38*(4), E401–E408.

Dall'o', G., Norese, M. F., Galante, A., & Novello, C. (2013). A multi-criteria methodology to support public administration decision making concerning sustainable energy action plans. *Energies, 6*(8), 4308-4330.

Erdoğan, N. (2020). Interaction between the reflections and financial incentives for renewable energy and renewable energy production in Turkey. Sivas Cumhuriyet University, Institute of Social Sciences, Master's Thesis, Sivas, 603142.

Github. (2024). *XGBOOST Parameters*. https://xgboost.readthedocs.io/en/latest/parameter.html/

Gostkowski, M., Rokicki, T., Ochnio, L., Koszela, G., Wojtczuk, K., Ratajczak, M., . . . , & Bełdycka-Bórawska, A. (2021). Clustering analysis of energy consumption in the countries of the Visegrad group. *Energies, 14*(18), 5612.

Grigoras, G., & Scarlatache, F. (2015). An assessment of the renewable energy potential using a clustering based data mining method: case study in Romania. *Energy, 81*, 416-429.

Güleryüz, E. (2022). Improvement of classification performance evaluation criteria using hybrid clustering methods. Bursa Uludağ University, Institute of Science, Doctoral Dissertation, Bursa, 717947.

Gürcün, D., & Petek, A. (2021). The evaluation of geothermal energy potential by SWOT analysis: the case of Aydın. Academic Review of Economics and Administrative Sciences, 14(2), 349-364.

İllez, B. (2020). Türkiye's energy outlook: Biomass energy in Türkiye. *TMMOB Chamber of Mechanical Engineers Report*, MMO/717, 317-346.

Jain, A. K. (2010). Data clustering: 50 years beyond K-Means. *Pattern Recognition Letters, 31*(8), 651-666.

Kacperska, E., Łukasiewicz, K., & Pietrzak, P. (2021). Use of renewable energy sources in the European Union and the Visegrad group countries—results of cluster analysis. *Energies, 14*(18), 5680.

Kaya, F., & Kaya, S. (2024). Increasing using wind energy in Turkey with comparative evaluation according to the European Union. The Journal of Social Sciences, 14(14), 363-380.

Kosowski, P., Kosowska, K., & Janiga, D. (2023). Primary energy consumption patterns in selected european countries from 1990 to 2021: A Cluster analysis approach. *Energies, 16*(19), 6941.

Kotsiantis, S. B. (2007). Supervised machine learning: a review of classification techniques. Maglogiannis I., Karpouzis K., Wallace B. A., Soldatos J. (Ed.), In *Emerging Artificial Intelligence Applications in Computer Engineering* (1st ed.) (3-24). Amsterdam: IOS Press.

Li, S., & Hu, Y. (2022). A multi-criteria framework to evaluate the sustainability of renewable energy: a 2-tuple linguistic grey relation model from the perspective of the prospect theory. *Sustainability, 14*(8), 4419.

Liu, A., Ledwich, G., Miller, W., & Cholette, M. (2020). A new multi-dimension clustering-based method for planning sustainable energy investment. *Asia-Pacific Sustainable Development of Energy Water and Environment Systems*. https://eprints.qut.edu.au/201261/.

Marinakis, V., Doukas, H., Xidonas, P., & Zopounidis, C. (2017). Multicriteria decision support in local energy planning: An evaluation of alternative scenarios for the sustainable energy action plan. *Omega, 69*, 1-16.

Matenga, Z. (2022). Assessment of energy market's progress towards achieving sustainable development goal 7: a clustering approach. *Sustainable Energy Technologies and Assessments, 52*(C), 102224.

Movlyanov, A., & Selçuklu, S. B. (2025). Energy efficiency optimization model for sustainable campus buildings and transportation. *Buildings*, 15(12), 1993.

Özgür, E. (2020). Türkiye's energy outlook: Solar energy in Türkiye. *TMMOB Chamber of Mechanical Engineers Report,* MMO/717, 297-315.

Pelau, C., & Chinie, A. C. (2018). Cluster analysis for the determination of innovative and sustainable oriented regions in Europe. *Studia Universitatis Vasile Goldis Arad–Economics Series, 28*(2), 36-47.

Quatrosi, M. (2022). Clustering environmental performances, energy efficiency and clean energy patterns: a comparative static approach across EU countries. *Sustainability Environmental Economics and Dynamics Studies*. https://www.sustainability-seeds.org/papers/RePec/srt/wpaper/0722.pdf.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65.

Saraswat, S. K., & Digalwar, A. K. (2021). Evaluation of energy alternatives for sustainable development of energy sector in india: an integrated shannon's entropy fuzzy multi-criteria decision approach. *Renewable Energy, 171*, 58-74.

Sasidharan, A. (2021). *Support Vector Machine Algorithm. https://www.geeksforgeeks.org/support-vector-machine-algorithm/*

Schölkopf, B., & Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (1st ed.). Cambridge, Massachusetts: MIT Press.

Selçuklu, Saltuk B., Rodgers, Mark D., Movlyanov, Atabek. (2022). Economically and environmentally sustainable long-term power system expansion, *Computers & Industrial Engineering*, 164, 107892.

Seddiki, M., & Bennadji, A. (2019). Multi-criteria evaluation of renewable energy alternatives for electricity generation in a residential building. *Renewable and Sustainable Energy Reviews, 110*, 101-117.

Serdar, S. (2020). Türkiye's energy outlook: Türkiye's hydroelectric potential and development status. *TMMOB Chamber of Mechanical Engineers Report* MMO/717, 271-282.

Shih, M., Yuan, Y., & Shi, G. (2024). Comparative analysis of LDA, PLS-DA, SVM, RF, and voting ensemble for discrimination origin in greenish white to white nephrites using LIBS. *J. Anal. At. Spectrom, 39*, 1560-1570.

Solangi, Y. A., Tan, Q., Mirjat, N. H., & Ali, S. (2019). Evaluating the strategies for sustainable energy planning in Pakistan: An integrated SWOT-AHP and Fuzzy-TOPSIS approach. *Journal of Cleaner Production, 236*, 117655.

Solomon, D. D., Khan, S., Garg, S., Gupta, G., Almjally, A., Alabduallah, B. I., . . . , & Abdallah, A. A. (2023). Hybrid majority voting: prediction and classification model for obesity. *Diagnostics, 13*, 2610.

Soylu, B. N. (2019). Renewable energy sources and renewable energy potential of Konya province. Selçuk University, Institute of Social Sciences, Master's Thesis, Konya, 567548.

Şahin, M. (2021). A Comprehensive analysis of weighting and multicriteria methods in the context of sustainable energy. *International Journal of Environmental Science and Technology, 18*(6), 1591-1616.

Tahtalı, Y. (2020). Classification of raw milk composition and somatic cell count in water buffaloes with support vector machines. *Kafkas Üniversitesi Veteriner Fakültesi Dergisi, 26*(4), 541-549.

T.C. Ministry of Environment, Urbanization and Climate Change. (2024a). *General Directorate of Meteorology: Official Climate*

Statistics. https://mgm.gov.tr/veridegerlendirme/il-ve-ilceler-istatistik.aspx?

T.C. Ministry of Environment, Urbanization and Climate Change. (2024b). *2021 Provincial Environmental Status Reports*. https://ced.csb.gov.tr/2021-yili-il-cevre-durum-raporlari-i-104268/ (Ziyaret tarihi: 05 Haziran 2024).

T.C. Ministry of Energy and Natural Resources. (2022). *Türkiye Electricity Generation and Consumption Emission Factors,* ETKB-EVÇED-FRM-042 Rev.01.

T.C. Ministry of Energy and Natural Resources. (2023). *2023 Annual Report.* Strategy Development Department.

T.C. Ministry of Energy and Natural Resources. (2024a). S*olar Energy Potential Atlas* (GEPA). https://gepa.enerji.gov.tr/

T.C. Ministry of Energy and Natural Resources. (2024b). *Wind Energy Potential Atlas (REPA).* https://enerji.gov.tr/bilgi-merkezi-enerji-ruzgar

T.C. Ministry of Energy and Natural Resources, General Directorate of Energy Affairs. (2024c). *Türkiye Biomass Energy Potential Atlas (BEPA).* https://bepa.enerji.gov.tr/

T.C. Ministry of Agriculture and Forestry. (2024). *Land Cover Classes.* https://corine.tarimorman.gov.tr/corineportal/araziortususiniflari.html /

TMMOB Chamber of Mechanical Engineers. (2024). *Türkiye's Energy Outlook Report* (MMO/758). Ankara.

Türkiye Electricity Transmission Corporation (2025). 2023 Statistics of Turkey Electricity Generation-Transmission. Electricity Production–Consumption–Losses, 2023 Turkey Electricity Generation by Sources Distribution. https://www.teias.gov.tr/turkiye-elektrik-uretim-iletim-istatistikleri/

Trappey, A. J., Wang, D. Y., Ou, J. J., Trappey, C., & Li, S. J. (2014). Evaluating renewable energy policies using hybrid clustering and analytic hierarchy process modeling. Proceedings of the 2014 *IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD),* Hsinchu, Tayvan, 21-23 May, 716-720.

Tutak, M., Brodny, J., Siwiec, D., Ulewicz, R., & Bindzár, P. (2020). Studying the level of sustainable energy development of the european union countries and their similarity based on the economic and demographic potential. *Energies, 13*(24), 6643.

Wang, Q., & Yang, X. (2020). Investigating the sustainability of renewable energy: an empirical analysis of European Union countries using a hybrid of projection pursuit fuzzy clustering model and accelerated genetic algorithm based on real coding. *Journal of Cleaner Production, 268*, 121940.

Wu, C., Peng, Q., Lee, J., Leibnitz, K., & Xia, Y. (2021). Effective hierarchical clustering based on structural similarities in nearest neighbor graphs. *Knowledge-Based Systems, 228*, 107295.

Zorlutuna, Ş., Erilli, N. A. (2018). Sosyo-ekonomik verilere göre illerin bulanık c-ortalamalar yöntemi ile sınıflandırılması: 2002-2008-2013 dönemleri karşılaştırması. *İktisadi Yenilik Dergisi*, 5(2), 13-31.

APPENDIX

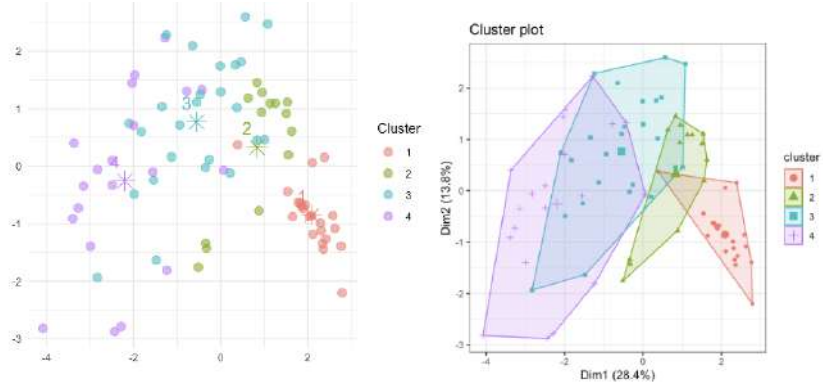

Figure A. 1. K-Means result for k = 4, metric = Euclidean

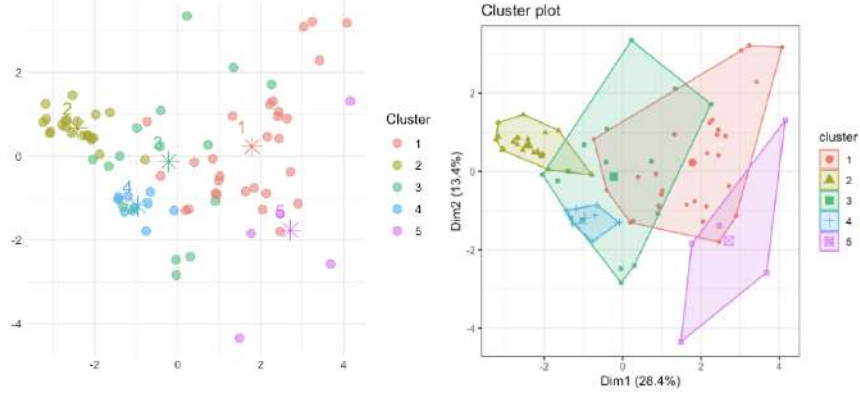

Figure A. 2. K-Means result for k = 5, metric = Euclidean

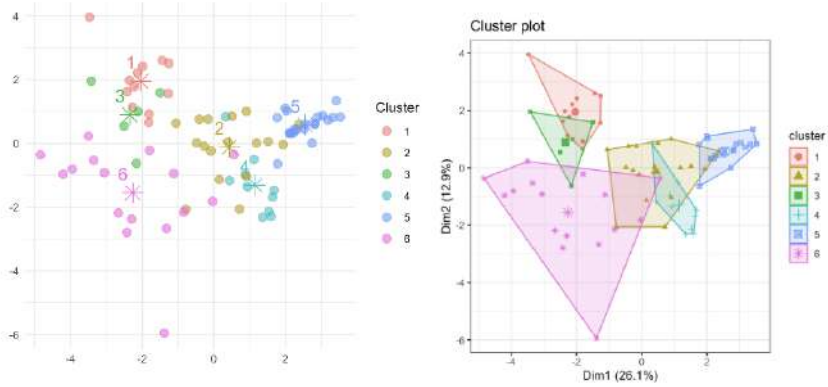

Figure A. 3. K-Means result for k = 6, metric = Euclidean

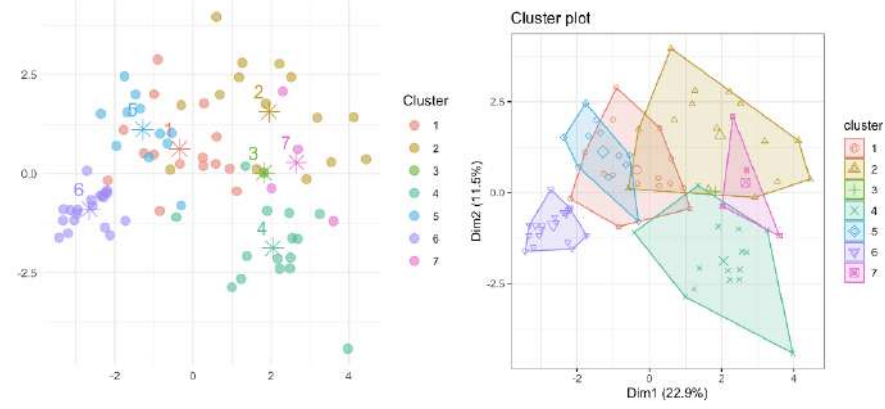

Figure A. 4. K-Means result for k = 7, metric = Euclidean

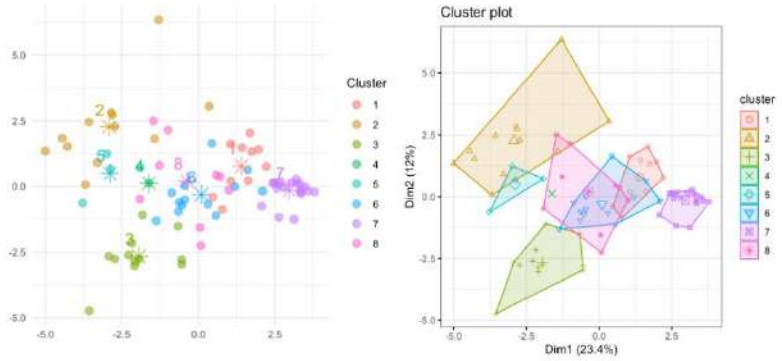

Figure A. 5. K-Means result for k = 8, metric = Euclidean

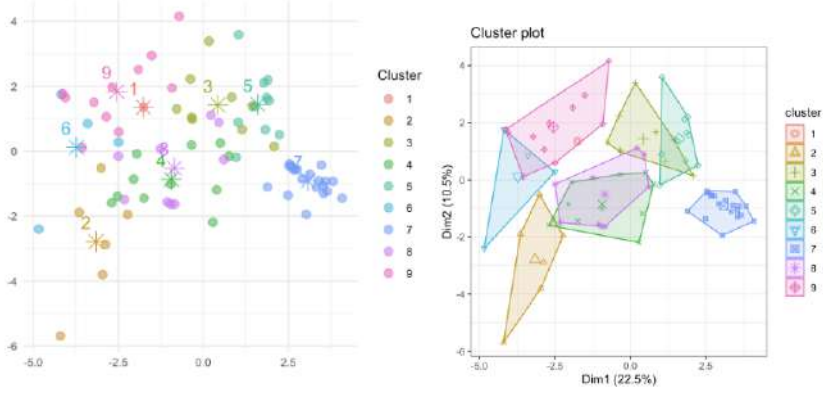

Figure A. 6. K-Means result for k = 9, metric = Euclidean

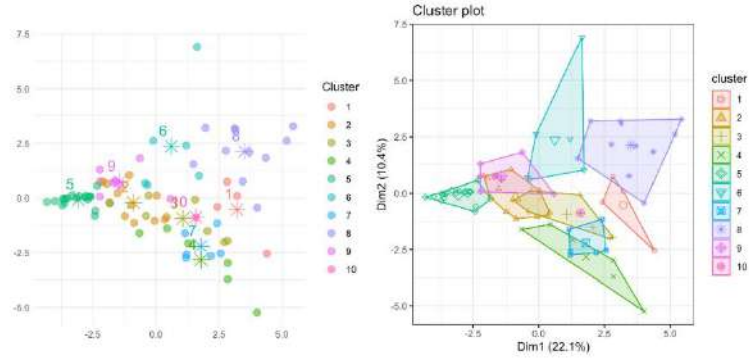

Figure A. 7. K-Means result for k = 10, metric = Euclidean

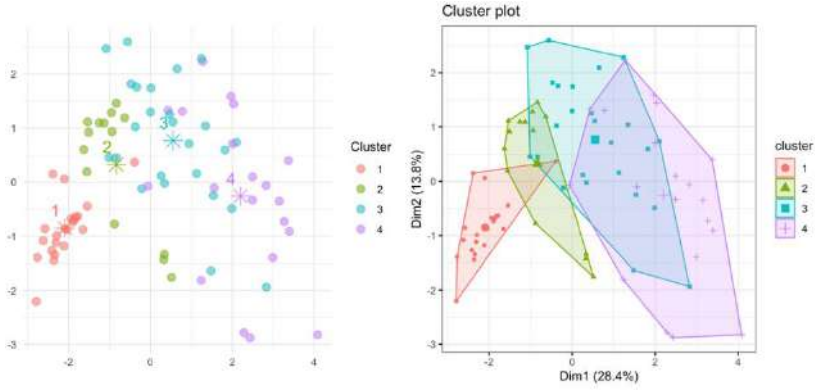

Figure A. 8. K-Means result for k = 4, metric = Manhattan

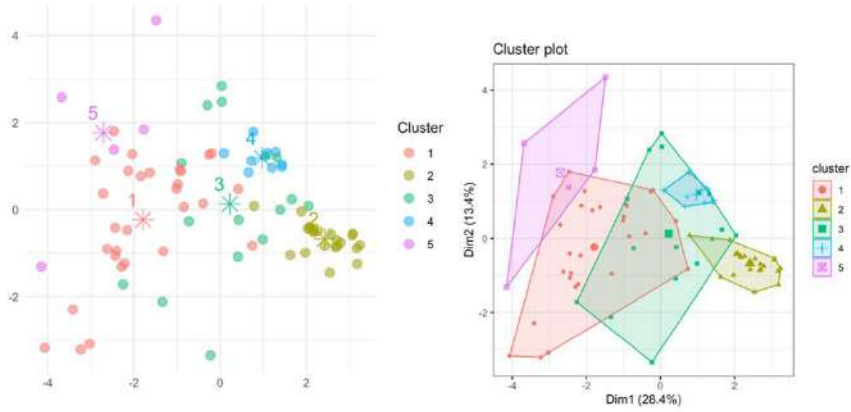

Figure A. 9. K-Means result for *k* = 5, metric = Manhattan

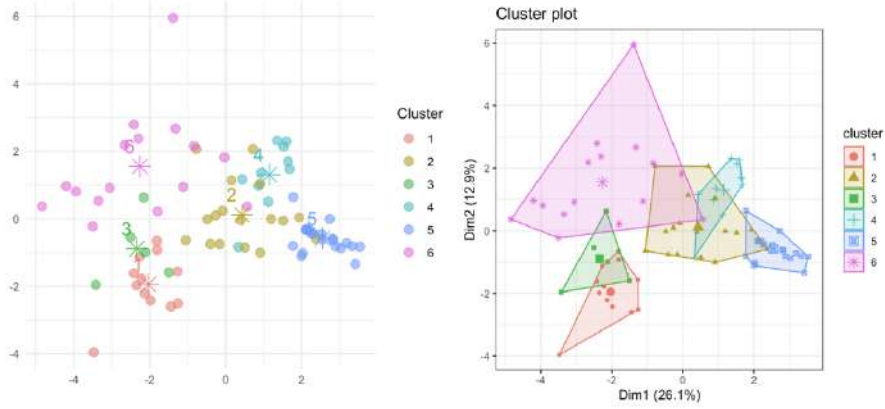

Figure A. 10. K-Means result for *k* = 6, metric = Manhattan

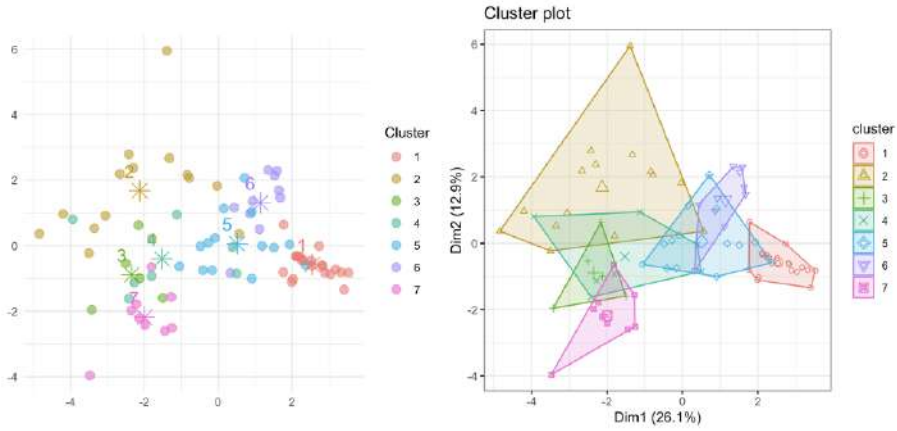

Figure A. 11. K-Means result for *k* = 7, metric = Manhattan

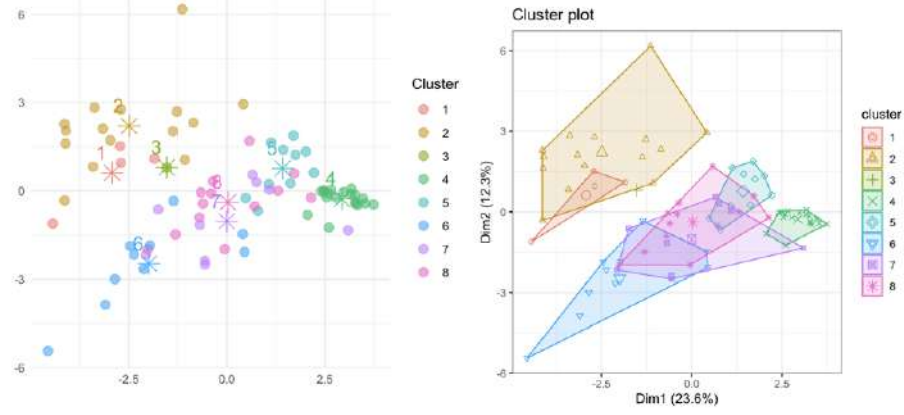

Figure A. 12. K-Means result for *k* = 8, metric = Manhattan

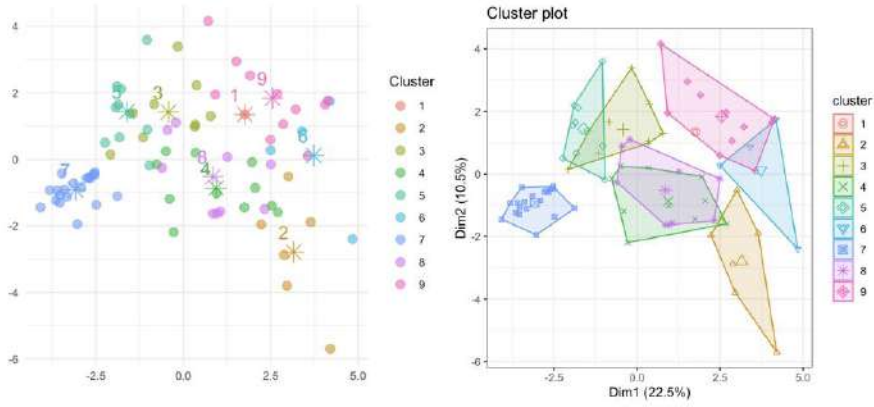

Figure A. 13. K-Means result for k = 9, metric = Manhattan

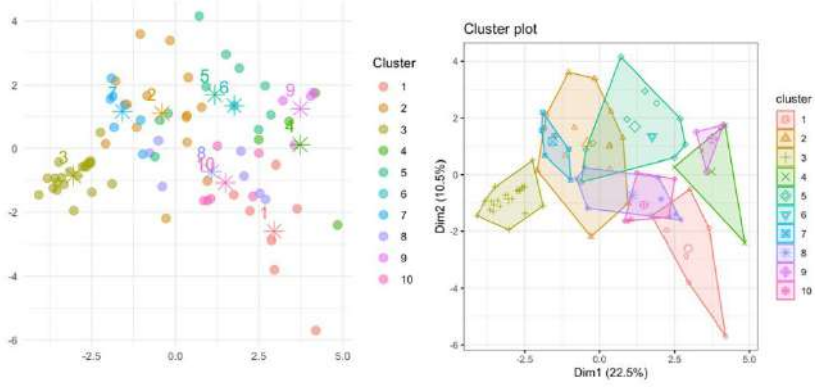

Figure A. 14. K-Means result for k = 10, metric = Manhattan

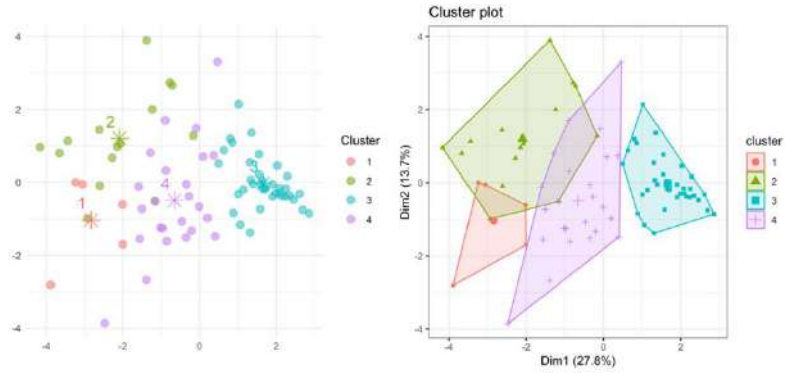

Figure A. 15. K-Means result for k = 4, metric = Minkowski

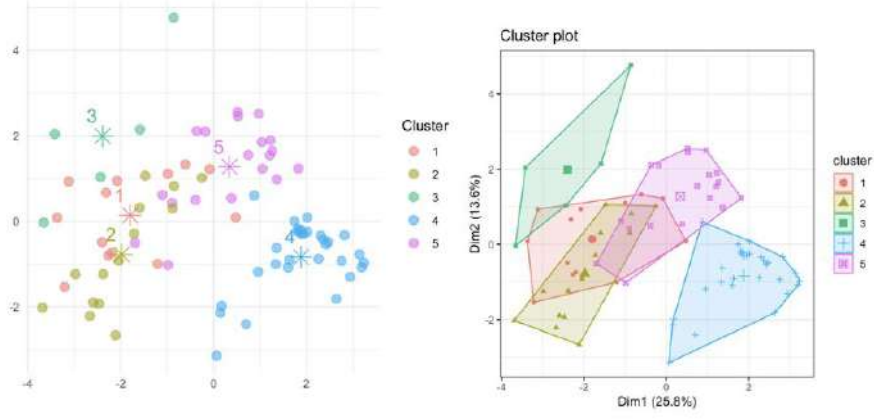

Figure A. 16. K-Means result for k = 5, metric = Minkowski

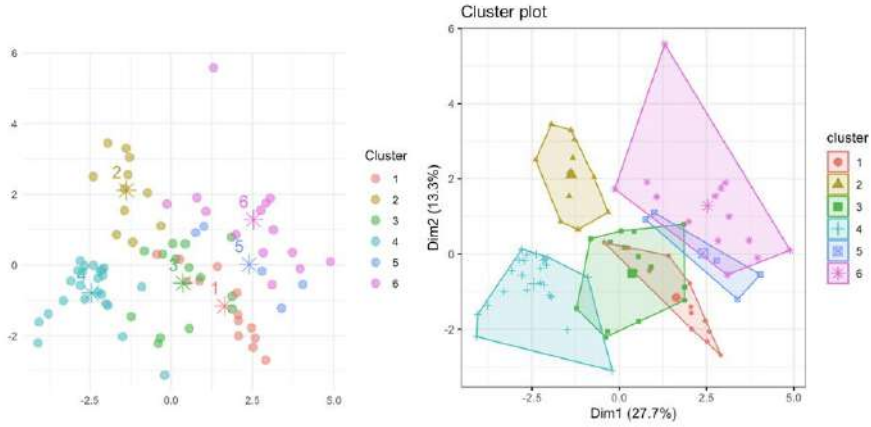

Figure A. 17. K-Means result for k = 6, metric = Minkowski

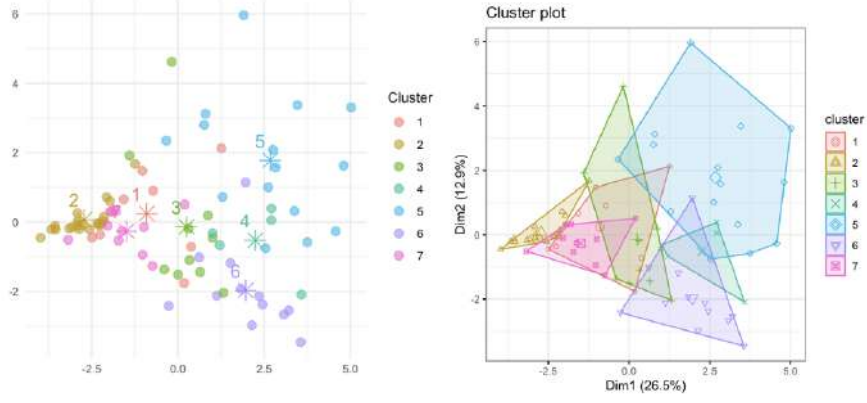

Figure A. 18. K-Means result for k = 7, metric = Minkowski

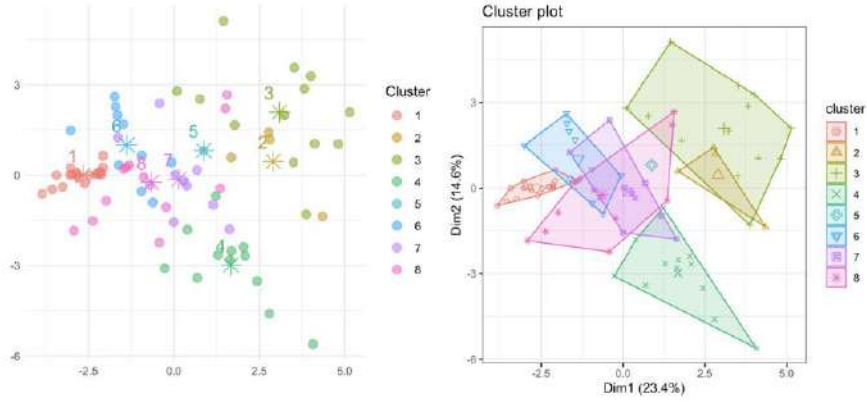

Figure A. 19. K-Means result for k = 8, metric = Minkowski

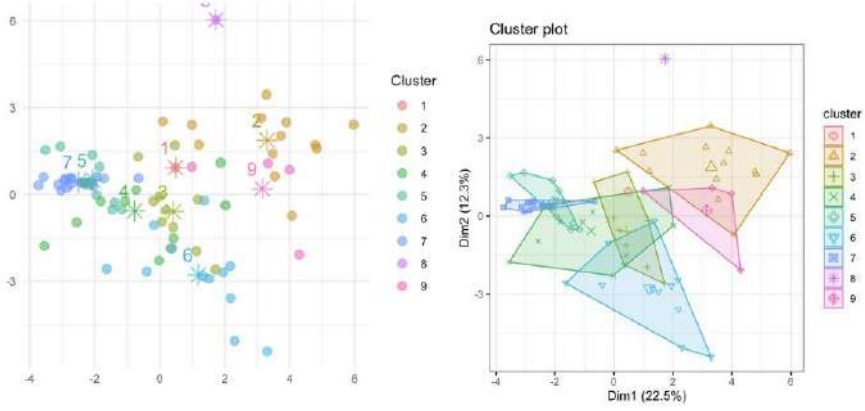

Figure A. 20. K-Means result for k = 9, metric = Minkowski

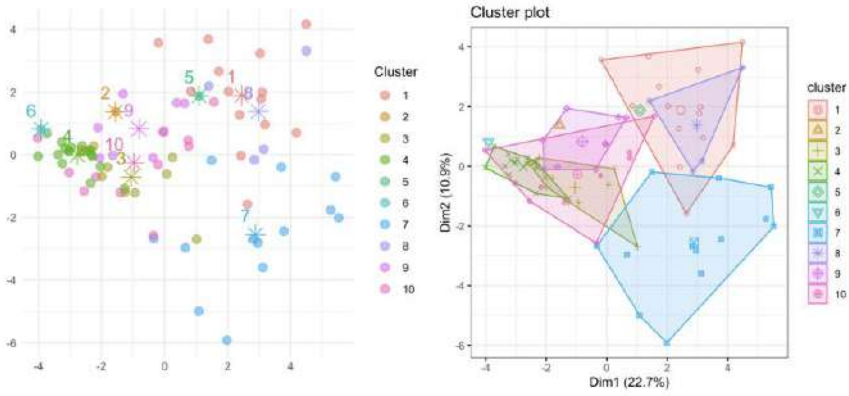

Figure A. 21. K-Means result for k = 10, metric = Minkowski

SHAPING THE FUTURE OF ENERGY: A MACHINE LEARNING–BASED ANALYSIS OF TÜRKİYE'S REGIONAL RENEWABLE ENERGY POTENTIAL