*Iraqi National Journal of Earth Science, Vol. 26, No. 03, 2026 (207-224)*

**Iraqi National Journal of Earth Science**

www.earth.uomosul.edu.iq

# Advancing Solar Energy Planning Using Machine Learning and GIS: A Suitability Analysis in Nineveh Governorate Iraq

**Salah J Ibrahim** [1*], **Ali Z.A. Al-Ozeer** [2], **Qusay K. Al-Ahmady** [3]

salah.23evp15@student.uomosul.edu.iq aalozeer@uomosul.edu.iq prof.kossayalahmady@uomosul.edu.iq

[1] *Department of Environmental Science, College of Environmental Science, University of Mosul, Duhok, Iraq.*
[2] *Department of Climate Change, College of Environmental Science, University of Mosul, Mosul, Iraq.*
[3] *Department of Environmental Engineering, College of Engineering, University of Mosul, Mosul, Iraq.*

Received: 14 April 2025 Received in revised form: 29 May 2025 Accepted: 07 July 2025

Available online: 01 July 2026

## Abstract

*Solar energy is growing rapidly due to its abundance, availability, cleanliness, cost-effectiveness, and ease of installation. The site selection process plays a crucial role in maximizing the efficiency of solar projects while minimizing the environmental impact. This study focuses on Nineveh Governorate, Iraq, by introducing an advanced solar project site selection methodology that utilizes the Geographic Information System (GIS) and Machine Learning (ML). For this aim, seven machine learning (ML) algorithms, namely Logistic Regression, Averaged Perceptron, Boosted Decision Tree, Decision Forest, Support Vector Machine, Neural Network, and K-Means, are applied on Microsoft Azure ML to identify a suitable site for solar farms. Thirteen factors are identified that influence the solar farm site, including solar radiation, sandstorms, land surface temperature, main road, population, power lines, substation, land ownership, land cover, water resources, slope, and aspect. These factors are categorized into four groups: environmental, climatic, topological, and socio-economic. Moreover, the dataset is analyzed in Azure Machine Learning using unsupervised and supervised ML. As a result, the Boosted Decision Tree model has achieved the highest accuracy at 94.2%. The output results indicate that 27% of the study area is highly suitable for solar farm development. This research underlines the immense potential of the Nineveh Governorate for renewable energy projects. It creates a replicable framework linking GIS and ML to planning energy needs with minimal environmental impacts.*

## Keywords:

*Solar Power, Geographic Information System (GIS), Machine Learning (ML), Optimal Site Selection, Iraq.*

*DOI: 10.33899/injes.v26i3.60958,* 

.==========

## 1. Introduction

Electricity is one of the most essential infrastructures in modern industry as it drives productivity and economic growth (Bayounis et al., 2022). Renewable energy is a source of energy replenished naturally over short periods (Shrestha et al., 2022). There are five renewable energy sources: solar, wind, water, geothermal, and biomass (Rahman et al., 2022). Solar energy is considered one of the most important renewable sources because it is widely available, environmentally friendly, affordable, and relatively simple to harness compared to other options (Ghosh and Halder, 2022). For solar energy projects, selecting the best location is essential to optimizing effectiveness and financial feasibility. Important variables include things like solar radiation, land availability, grid proximity, and minimum environmental restrictions. Choosing a suitable location guarantees increased energy production, lowers expenses, and lessens environmental disturbance (Spyridonidou and Vagiona, 2023). In Nineveh Governorate, Iraq, selecting appropriate sites for solar energy projects presents challenges due to geographic and terrain conditions. To determine the best sites for solar farms, factors including land ownership, slope, aspect, land use patterns, and environmental restrictions must be carefully considered (Hassan et al., 2024). According to the Renewable Energy Agency (2024), the global renewable power capacity hit 3,870 gigawatts (GW) by 2023. Several countries made a serious contribution to the global rise of renewable energy capacity.

Notably, China added approximately 374 GW, and the United States added 38.3 GW of solar capacity, up 54% from last year. This was a record-breaking growth, adding 473 GW in 2023. Geographic information systems (GIS) software and spatial analysis methodologies have evolved as popular tools for selecting suitable sites by combining multiple geographic data sources (Mahmood Faisal and Abdaki, 2021). Machine learning models have been successfully used in mapping site selection for various purposes, including the suitability of renewable energy (Ashraf et al., 2025). Iraq has historically depended on fossil fuels to fulfill its energy needs, but in recent years, there have been efforts to expand its energy portfolio with renewable initiatives. The Renewable Energy Agency (2024) reported that renewable energy accounted for about 2% of Iraq's electricity capacity in 2022, marking a decrease from 9% in 2014. As reported, the total renewable energy capacity in 2014 was 2311 MW. This energy was generated mainly by hydroelectric dams such as Mosul Dam. But this capacity decreased year by year to reach 1599 MW in 2023. The decline in Iraq's hydroelectric capacity is largely due to water scarcity, drought, and reduced river inflows exacerbated by upstream damming in Turkey and Iran, which have significantly lowered water availability, directly impacting hydroelectric generation (Jiang et al., 2024). As a previous study, the only research that specifically focused on solar energy site selection in Nineveh Governorate is a master's thesis by Al-Sabawi (2023), applying GIS and the Analytic Hierarchy Process (AHP) to identify suitable locations for solar photovoltaic energy in Nineveh Governorate, Iraq. The findings revealed that approximately 24% of the governorate holds high suitability for solar development, with these areas predominantly consisting of unused or low-productivity agricultural land, while 7% is very highly suitable. The study incorporated several key factors, including solar radiation, frequency of sandstorms, slope, aspect, DEM, soil type, and proximity to infrastructure such as power lines, roads, and water sources. Based on the spatial analysis, the study recommended the northern parts of the governorate, particularly Zummar, Tall Kayf, Tal Afar, and Sinjar districts, as the most promising locations for future photovoltaic projects. Research by Sachit et al. (2022) applied machine learning (ML) and GIS techniques for the site selection of both solar and wind energy projects. By utilizing 13 factors such as solar radiation, air temperature, and distance to cities. Applying machine learning techniques, such as random forest, support vector machine, and multi-layer perceptron. The obtained result confirms that the Random Forest achieved a maximum accuracy of 89% for solar energy. Furthermore, the result reveals that 19.4% of the land on our planet is very well adapted to solar farms. The research highlights the use of ML and GIS in renewable energy planning. Another relevant study is the research conducted by Ashraf et al. (2025). Machine learning algorithms, Random Forest, XGBoost, and Multilayer Perceptron were applied to solar PV site suitability in Pakistan's Cholistan Desert. The study utilized 14 factors like solar radiation, slope, DEM, land cover, and proximity to roads, power lines, and water bodies. Random Forest had the highest accuracy (92%) among the algorithms used. The study determined 10–11% of the area, Bahawalnagar and Bahawalpur, as highly suitable for solar development. Such studies point to the growing use of machine learning and GIS methods in assessing site suitability for solar farms

This research aims to identify the key factors that influence solar farm site suitability and select the optimal site for solar energy projects in Nineveh Governorate. This research supports the United Nations Sustainable Development Goals (SDGs), which aim to address global challenges such as poverty, inequality, and climate change. In particular, the study contributes to SDG 7 (Affordable and Clean Energy), SDG 8 (Decent Work and Economic Growth), SDG 9 (Industry, Innovation, and Infrastructure), SDG 11 (Sustainable Cities and Communities), and SDG 13 (Climate Action).

## 2. Location of The Study Area

Nineveh Governorate is located in the north and northwest of Iraq, lying within the coordinates between latitudes (35°00'00'' N to 37°00'00'' N) and longitudes (41°00'00'' E to 44°00'00'' E). It shares the border with Syria, Dohuk, Erbil, Salah al-Din, and Anbar. The Governorate consists of nine districts: Tel Kayf, Sheikhan, Makhmoor, Tal Afar, Sinjar, Ba'aj, Al-Hatra, Hamdaniya, and Mosul. In 2021, its population was estimated at 4030006 inhabitants. Its total area is about 39632

Square Kilometers (Abed, 2024). Figure (1) shows the map of the study area.

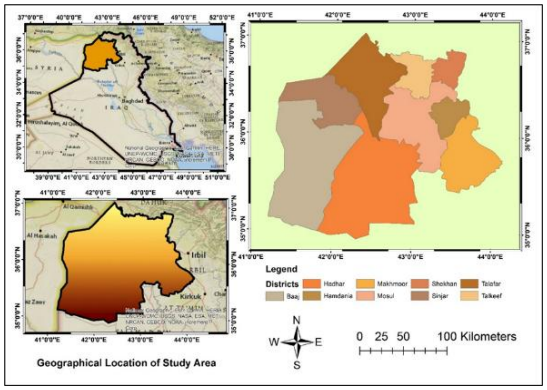



**Fig. 1: Geographical location of the study area.**

Based on the Iraqi Ministry of Electricity Annual Statistical Report (2023), electricity in Nineveh is generated from two main sources (Government Sector). Electricity is produced through electrical units that primarily use fossil fuels such as oil and some natural gas. The government also generates electricity through the Mosul Dam via hydropower. The second sector is (Civil Generators). There is also a significant reliance on privately owned generators distributed throughout the region. These generators often use gasoline as their fuel source. Both sectors contribute to environmental pollution. According to this report, Nineveh requires a load of 1785 MW. Government sectors generate 1144 MW in total, which means there is an electricity Deficiency of 641 MW. This deficiency is generated by Civil generators. That means the capacity of civil generators is 641 MW. In general, the actual supplied load is 64%, and the power deficiency is 36%. Figure (2) represents the power deficiency.

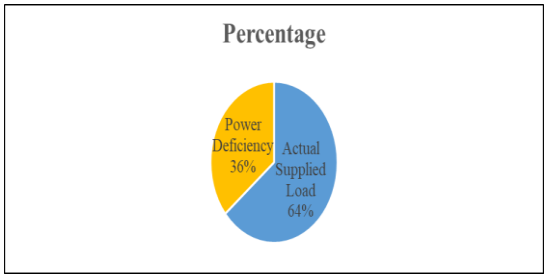



**Fig. 2: Energy supply and demand in Nineveh Governorate.**

## 3. Materials and Methods

In this research, the methodology for evaluating solar project site suitability comprised several sequential stages. Initially, relevant factors influencing site suitability are identified. Corresponding data for each factor are acquired and subsequently analyzed using ArcMap. Appropriate analytical techniques are applied according to the nature of each factor, and all layers are reclassified into five suitability categories: Very Low, Low, Moderate, High, and Very High. In the following stage, the classified maps were pre-processed for machine learning by transforming the study area into a dataset containing 3222 spatial points, with each point assigned the respective factor values. This dataset was then utilized in the machine learning phase by implementing both unsupervised and supervised learning algorithms. Fig. 3 illustrates the methodology stages, and Fig. 4 shows the flowchart of the methodology.

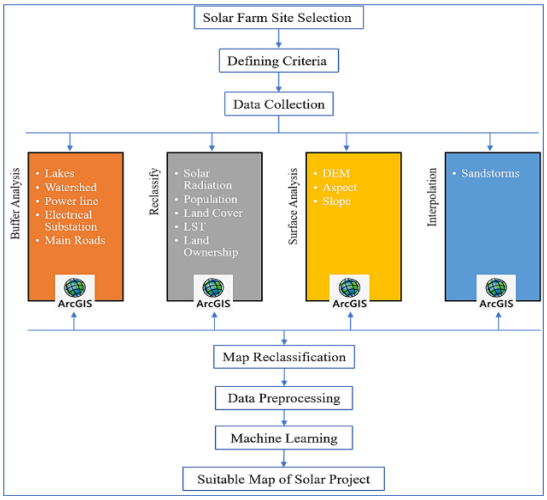



**Fig. 3: Methodology phases.**

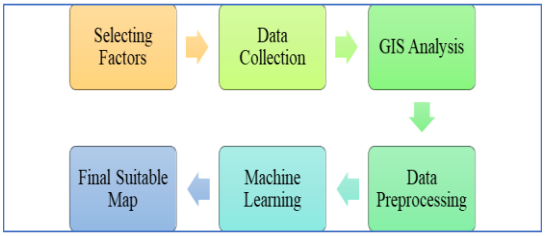



**Fig.4: Solar farm site selection framework. It begins with selecting criteria and spatial data collection. GIS methods (buffer, reclassification, surface analysis, interpolation) are applied, and outputs are standardized into five suitability classes. Data are converted to points and analyzed using machine learning to produce the final suitability map.**

### *Maps Preparation*

In this research, the literature review and international regulations are studied. After that, comprehensive sets of factors are selected, and based on the availability of data in the study area,

factors were defined and classified into four groups: environmental (Land Cover, Distance to lakes, Distance to watersheds), socioeconomic (Main Road, Population, Power Lines, Electrical Substation, Land Ownership) topological (Slope and Aspect) and climatic (Solar Radiation, Sandstorms, LST). Fig. 5 represents the selected criteria and their categories. The defined criteria and layers are collected and acquired from various sources, including satellite images such as the Space Shuttle Radar Topography Mission (SRTM) used for DEM data, government departments of the governorate, scientific centers, research, and some of them are created and derived from software such as QGIS and ArcGIS. Table (1) shows the details of the data sources.

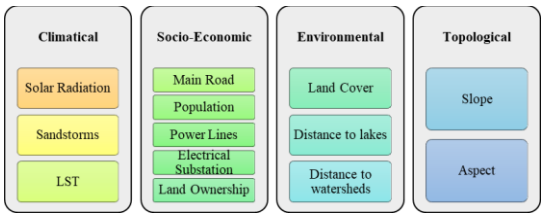



**Fig.5: Selected criteria and their categories.**

**Table 1: Sources of selected factors.**

| Factors | Source |
|---|---|
| Solar irradiation | Global Solar Atlas ((HTTPs://globalsolaratlas.info/). The solar radiation data were a raster file and were clipped to the boundary of Nineveh Governorate for use in the analysis. |
| Power line | The power line data were created using QGIS, specifically utilizing the QuickOSM plugin. |
| Electrical Substation | The Substation data were originally obtained from the Ministry of Electricity General Company for Northern Electricity Distribution, Training and Development Department, as cited in a master's thesis. After extracting the coordinates and capacities from an Excel sheet, the data were displayed in ArcMap. |
| Land ownership | Land ownership information was obtained from the Agriculture Directorate of Nineveh. |
| DEM | NASA ((https://data.nasa.gov/). The DEM raster was clipped to the extent of Nineveh Governorate to extract relevant DEM data for the study. |
| Slope | The slope was derived from the DEM data using the Surface Analysis Slope Tool in ArcMap. |
| Roads | Humanitarian Data Exchange (https://data.humdata.org/dataset/iraq-roads) |
| Water Resources | Water resources were derived from the DEM in ArcGIS using flow direction, stream identification, and watershed delineation |
| Land cover | The land cover data for Nineveh were analyzed using Google Earth Engine (2019-2023 averages) and reclassified in ArcMap into five suitability categories for solar farms. |
| Sandstorm | The sandstorm data used in this study were sourced from the Ministry of Transport General Authority for Meteorology and Seismology, as reported in a master's thesis, and were then analyzed using IDW in ArcMap. |
| Population | Humanitarian Data Exchange (https://data.humdata.org/dataset/iraq-populated-places-2021). ). Population data, also in raster format, was clipped to the Nineveh Governorate boundary to ensure spatial consistency with other layers. |
| Aspect | The aspect was derived from the DEM data using the Surface Analysis aspect Tool in ArcMap. |
| LST | The LST data for Nineveh was analyzed using Google Earth Engine (2019-2023 averages) and reclassified in ArcMap into five suitability categories for solar farms. |

### *Influencing factors for solar farms*

#### A. Solar Radiation

It's the most important factor for solar energy projects. Solar radiation levels at a site directly impact the efficiency and total energy production of solar panels (Abdo and El-Shimy, 2013). The solar radiation data were obtained from the Global Solar Atlas website, specifically, the GHI (Global horizontal irradiation) annual average for 24 years from 1999 to 2023. According to the data, the south of the governorate has the highest solar radiation, and the Mosul dam lake has the lowest value.

#### B. Sandstorms

Sandstorms can block the sunlight and minimize the solar panel's efficiency. Sandstorms may cover the solar panel surface with sand particles, which increases the maintenance cost (Alanzi et al., 2024). The sandstorm data were sourced from a master's thesis (Al-Sabawi, 2023). The data represent the annual average of sandstorms and suspended dust in Nineveh Governorate for 30 years from 1990 to 2020, taken from five weather stations. Table (2) shows the data.

**Table 2: Sandstorms Data.**

| Station | Latitude | longitude | Average number of sandstorms over 30 years |
|---|---|---|---|
| Baaj | 36.03 | 41.73 | 0.425 |
| Mosul | 36.33 | 43.16 | 0.38 |
| Talabta | 35.94 | 42.56 | 0.233 |
| Talafar | 36.36 | 42.38 | 0.211 |
| Rabeeaa | 36.74 | 42.23 | 0.268 |
| Sinjar | 36.33 | 41.87 | 0.401 |

#### C. Land Surface Temperature

Higher temperatures reduce solar panel efficiency, requiring cooling solutions that add to overall system costs (Xu et al., 2024). LST ranges

from 22 °C to 52 °C, and Mosul Dam Lake has the lowest LST value based on the map data.

D. Land Cover

Land Cover, built-up area, and tree cover can cause shading and reduce the productivity of solar panels, while sparse vegetation and grassland provide a large scale of open spaces (Hernandez et al., 2015).

E. Distance to Water Resources

Distance to the lake, river, and watersheds can prevent the solar project from flooding; also, areas near water resources are unstable, which will increase construction costs. Construction near water resources can damage the ecosystem and break environmental regulations (Al-Saidi and Lahham, 2019).

F. Distance to the Roads

Accessibility to the roads is an economic factor that ensures easy and cost-effective transport of solar project materials in the construction phase, such as solar panels and inverters (Kırcalı and Selim, 2021).

G. Distance to Power lines and substations

Solar farms need to be located near existing high-voltage power lines or substations to minimize the cost and complexity of connecting to the grid (Khazael and Al-Bakri, 2021).

H. Population Density

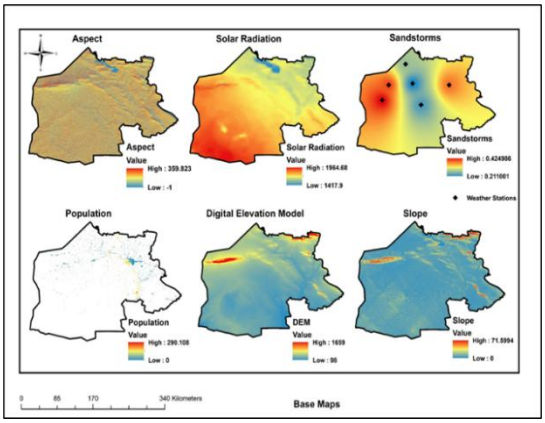
Population is a social factor. Highly populated areas need more energy, so the site should be a populated area to minimize energy loss (Scovell et al., 2024).

I. Land Ownership

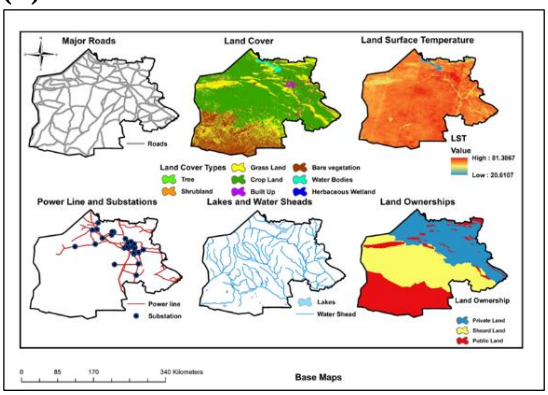
Land Ownership refers to the control of land and is divided into three types: first, public land, which is owned by the local government; private land owned by individuals and mainly used for agriculture; and finally, shared land, which is a mix of public and private land (Abashidze and Taylor, 2023).

J. Digital Elevation Model

The DEM dataset, featuring a 30-meter resolution, is derived from NASA Earth Data using the SRTM satellite. The slope, a flat slope, is ideal because it minimizes construction costs (Keshavarz et al., 2012). Aspect represents the direction of the slope face. It is a key factor for solar project site planning as it affects the amount of sunlight received by the site (Fan et al., 2021). Figure (6) shows all base maps.




**(A)**




**(B)**

**Fig. 6: (A) Base maps of Aspect, Solar Radiation, Sandstorms, Population Density, DEM, and Slope. (B) Base maps of Roads, Land Cover, Land Surface Temperature (LST), Power Lines and Substations, Lakes and Watersheds, and Land Ownership.**

### *Map Reclassification*

Map reclassification is a critical step in site selection analysis because, in this step, factors are classified into suitable classes, which are five classes (Very low, Low, Moderate, High, and Very high). These classes reflect the suitability level and simplify complex datasets; also, reclassification helps in identifying optimal sites while excluding low-suitability areas. Moreover, it offers a clear foundation for further analysis using machine learning. In this study, an innovative methodology has been used for map reclassification by utilizing an AI Model, DeepSeek, to classify and arrange criteria for solar project development in Nineveh Governorate. First, the dataset for the study area, Nineveh, was provided to the AI model. The dataset includes the maximum and minimum values for key factors such as solar radiation, temperature, slope, etc. These values determine the local conditions of the study area. Second, 23 articles are provided as sources of world research

data. These articles contain the criteria ranges employed in similar solar farm site selection studies around the world. The AI then analyzes these articles to establish the best range of criteria that matches the conditions in the Nineveh Governorate. The criteria range of the articles in the study area dataset is compared by the AI to identify the most appropriate ranges for the Nineveh Governorate. This new method provides a robust basis for integrating local environmental information with results from global studies. A total of 23 articles have been applied as data sources to determine appropriate ranges for factor reclassification. The articles are (Georgiou and Skarlatos, 2016; Aldrin Wiguna et al., 2017; Kereush and Geomatics, 2017; Mierzwiak and Calka, 2017; Sadeghi and Karimi, 2017; Hussain and Mahdi, 2018; Ibrahim et al., 2019; Koc et al., 2019; Nebey et al., 2020; Badi et al., 2021; Khazael and Al-Bakri, 2021; Suprova et al., 2021; Bandira et al., 2022; Elboshy et al., 2022; Goshem and Hailu, 2022; Sachit et al., 2022; Uyan and Dogmus, 2023; Farahani et al., 2024; Kamali Saraji et al., 2024; Li et al., 2024; Rane et al., 2024; Rusol and Al-Timimi, 2024; Salama, 2024). When given ranges from a specific dataset (article criteria), they are collected and refined. The system compares, analyzes, and evaluates these values against global benchmarks and prior studies. The system retrieves the pertinent data and arranges the factors into five classes, ensuring that the class structure fits the context of the project and the empirical data from the region. This is done automatically and optimized with the help of special algorithms that process large volumes of data. Thus, an optimized classification of factors is produced that fits the requirements of the study area. Table 3 shows the ideal ranges of each factor, while Fig. 7 presents the reclassified maps.

### *Reclassification Process*

In this study, after identifying the suitable ranges for each factor, the maps are reclassified. Various analyses are performed based on the spatial data of each factor. Before performing the reclassification process, all spatial data layers were standardized to a 30-meter spatial resolution (Fatah et al., 2022). This resampling step is essential to ensure consistency in cell size across all input rasters, enabling accurate analysis and comparison using a uniform resolution that minimizes spatial discrepancies (Qin et al., 2025). The 30m resolution is selected as it balances data detail with computational efficiency, making it suitable for regional-scale environmental analysis and decision-making processes (Westgate et al., 2025).

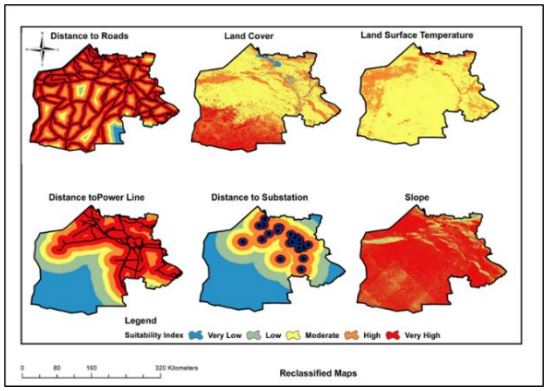
The set of spatial data of raster layers, which are Solar Radiation, Sandstorms, LST, Population, Land Cover, Slope, and Aspect raster maps, is grouped into five suitability classes except Land Ownership. There are three classes: Class 1 (Low), Class 2 (Moderate), and Class 3 (High). Reclassification is carried out using ArcMap 10.8 with the Reclassify tool under Spatial Analyst Tools > Reclass > Reclassify.

The set of vector layers, such as Main Roads, Power Line, Substation, Distance to Lake, and Distance to Watershed, underwent spatial analysis using buffering methods to calculate proximity-based suitability. More specifically, the Multiple Ring Buffer function in ArcMap Analysis Tools is utilized to generate concentric buffer areas around each feature.

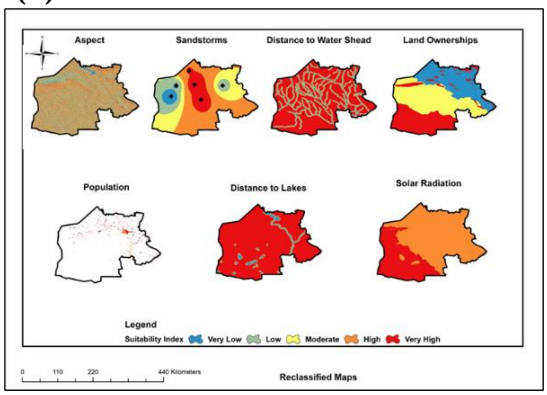
**Table 3: Criteria ranges based on the AI Model**

| Factors | Unit | Class 1 | Class 2 | Class 3 | Class 4 | Class 5 |
|---|---|---|---|---|---|---|
| | | Very Low | Low | Moderate | High | Very high |
| Solar Irradiation | kWh/m²/year | 1417-1543 | 1543-1669 | 1669-1795 | 1795-1921 | 1921-1965 |
| Power line and Substation | m | > 50,000 | 30000-50000 | 20000-30000 | 10000-20000 | < 10000 |
| Land ownership | - | Private Land | | Shared Land | Public Land | |
| DEM | m | > 1600 | 1200 – 1600 | 800 – 1200 | 400 – 800 | < 400 |
| Slope | Degree | >30° | 15 – 30° | 5 – 15° | 3 – 5° | 0 – 3° |
| Roads | m | > 20,000 | 15,000 – 20,000 | 10,000 – 15,000 | 5,000 – 10,000 | < 5,000 |
| water Resources | m | < 500 | 500 – 1000 | 1000 – 1500 | 1500 – 2000 | > 2000 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Land Cover | - | Tree cover, Water Body | Built-up Areas | Cropland | Grassland | Sparse Vegetation |
| Sandstorm | Day/year | > 7 | 5 – 7 | 3 – 5 | 2 – 3 | < 2 |
| Population | Inhabitants/km² | 200-290 | 150-200 | 100 – 150 | 50 – 100 | < 50 |
| Surface Temperature | °C | > 60 | 50 – 60 | 40 – 50 | 30 – 40 | 20.6 – 30 |
| Aspect | Degree | North-facing | Northwest/North east | West/East | South/Southeast | Flat/Southwest |




**(A)**




**(B)**

**Fig.7: (A) Classified maps of roads, land cover, LST, power lines, and substations, and slope; (B) Classified maps of aspect, sandstorms, watersheds and lakes, population, solar radiation, and land ownership.**

### *Data Preprocessing*

Data preprocessing is a vital step for data analysis using machine learning. In this process, all vector layers are converted into raster format using standard conversion tools, including the transformation of polygon and polyline features. To ensure consistency across the dataset, all layers are in a uniform spatial resolution of 30 meters (Fatah et al., 2024). The data preprocessing stage ensures the data are in a prepared format for advanced analysis in ML by converting spatial raster data into numerical values. In this study, the study area is divided initially into 3,222 evenly distributed spatial grid points using the Fishnet tool in ArcMap, creating a systematic framework for sampling and spatial analysis. Then, each point is assigned spatial coordinates by using the Add XY Coordinates tool, ensuring precise geolocation. To assign values from all selected criteria to each point, the Extract Multi Values to Points tool is applied. This ensured that each point contained complete data for all selected factors, such as solar radiation, slope, and land cover. The resulting dataset is then exported as a CSV file, creating a tabular format suitable for machine learning algorithms. This structured dataset has provided a comprehensive and precise foundation for advanced analysis and modeling. Fig. 8 represents the spatial grid points row dataset in the appendices.

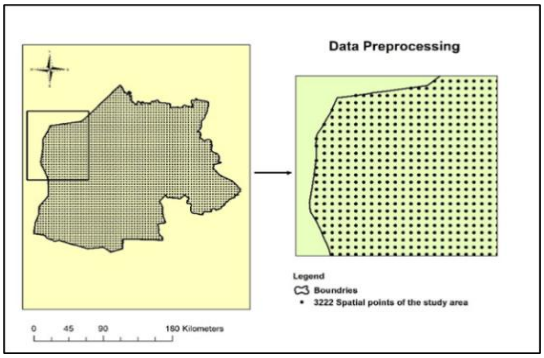



**Fig. 8: Spatial points of the study area.**

### *Machine Learning*

The scientific study of computer algorithms that automatically improve their performance with exposure and experience is known as machine learning (ML), which is a branch of artificial intelligence that uses data to learn and make predictions (Abdaki et al., 2025). Two of the most widely adopted ML techniques are supervised and unsupervised. Unsupervised ML generates target data, which means it works with unlabeled data. Its function is to group the data based on similarity. Supervised ML deals with labeled data, which involves training the algorithm on a labeled dataset. Labeled data means each input has a known output (Almuqati et al., 2024). This type of ML is useful for classification or regression. This study's methodology involves both unsupervised and supervised machine learning techniques. Microsoft Azure Machine Learning Studio has been used, which is available at (Azure Data Studio | Microsoft Azure).

### *Unsupervised Machine Learning*

Data: refers to the dataset, which contains the values of all factors, with 3222 rows and 13 columns. Columns are linked to all 13 factors, such as solar radiation, slope, etc. Clean Missing Data: This step is important because it cleans the missing data from the dataset, ensuring the data is complete and the clustering process will run accurately (Hua and Pei, 2007). Normalize Data: This process arranges all factor values into the same scale, which prevents the range between the factor values and ensures fair treatment of clustering (Li et al., 2021). Algorithm (K-Means Clustering): This algorithm groups the data into K clusters. It assigns each data point to the cluster with the nearest centroid based on the normalized values. Train Clustering Model: The cleaned and normalized dataset is used to train the model. Initial centroids are computed, data points are iteratively assigned to the closest cluster, and centroids are updated until convergence (Niu et al. 2022). Assign Data to Cluster: Each data point in the dataset is assigned to the cluster it belongs to based on proximity to the cluster's centroid. This results in clear groupings of data based on similarity. Convert to CSV: After clustering, the result of the process is converted to CSV format for further analysis with supervised ML. Figure (9) exhibits a flowchart of unsupervised ML.

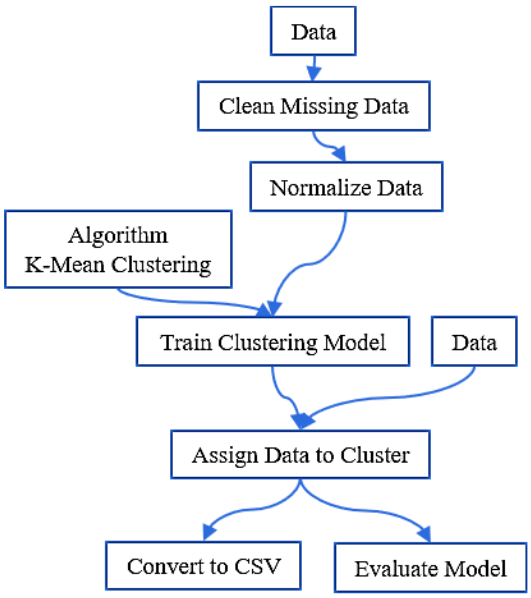



**Fig.9: Pipeline flowchart of unsupervised ML in Microsoft Azure Platform**

### *Supervised Machine Learning*

In supervised ML, six algorithms have been applied, which are (Two-Class Logistic Regression, Two-Class Averaged Perceptron, Two-Class Boosted Decision Tree, Two-Class Decision Forest, Two-Class Support Vector Machine, and Two-Class Neural Network). Split the data: the dataset is divided into two parts: 70% for training, used to train the supervised machine learning model, and 30% for testing, used to evaluate the performance of the trained model (Joseph and Vakayil, 2022). Model Training algorithms are trained using the labeled dataset derived from the unsupervised ML process (Chen et al., 2024). Model Scoring is used to test the trained model, which is 30% of the dataset (Mizumoto and Eguchi, 2023). Evaluate Model calculates the performance, accuracy, and efficiency of the model (Chang et al., 2024). Fig.

10 represents the flowchart of supervised ML. Various binary classification algorithms are used in this study. Logistic Regression predicts event probabilities using a logistic function (Omar et al., 2024). The Averaged Perceptron, a simple neural model, classifies inputs using weighted linear functions (Noori and Qasim, 2023). Boosted Decision Trees build successive trees to correct previous errors (Coadou, 2022). Support Vector Machines handle both linear and non-linear classification (Al-Thanoon et al., 2018). Decision Forests aggregate multiple decision trees for final predictions (Ahmed Salman et al., 2024). Neural Networks use interconnected neurons to detect complex patterns for binary outcomes (Tian and Tang, 2025).

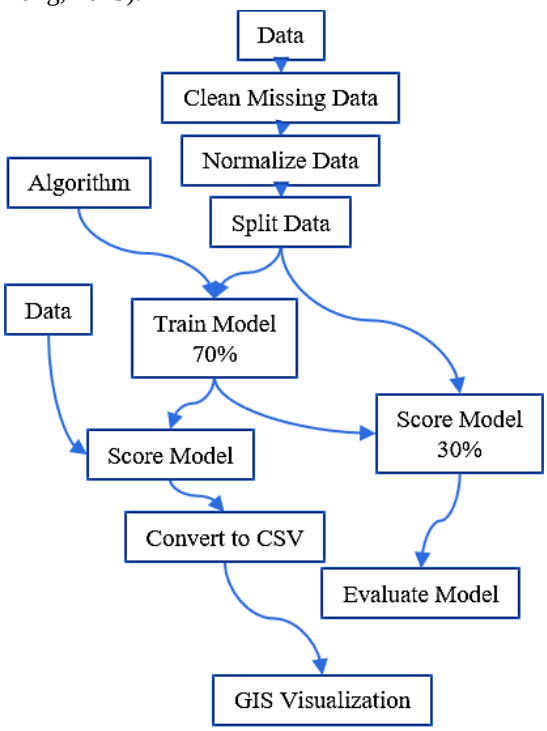



**Fig.10: Pipeline flowchart of supervised ML in Microsoft Azure Platform.**

### Error Statistics Used to Evaluate the Model's Performance

Once the data are cleaned and preprocessed, the model produces results in the form of probabilities. To measure how well the model performs, a confusion matrix is used. This tool breaks down the results into true and false predictions, helping to calculate important metrics like accuracy, precision, recall, F-score, and AUC. These values show how reliable and effective the model is at making correct classifications. Figure (11) shows the Confusion Matrix.

| | Actually Positive (1) | Actually Negative (0) |
|---|---|---|
| Predicted Positive (1) | True Positives (TPs) | False Positives (FPs) |
| Predicted Negative (0) | False Negatives (FNs) | True Negatives (TNs) |

**Fig. 11: Confusion Matrix.**

Azure offers numerous metrics to distinguish between the best model performance. Five assessment metrics have been used in this study for result validation:

Accuracy (ACC): The proportion of true results to total cases is used to assess the quality of a classification model (Naidu et al., 2023) and can be calculated as follows:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + FP + FN + TN}}$$

Precision: the value of real outcomes over all positive outcomes (Berrar, 2024). It can be calculated as follows:

$$\boldsymbol{Precision\ (P)} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}}$$

Recall: the model's percentage of all correct results recovered (Maxwell et al., 2021). It can be calculated as follows:

$$\boldsymbol{Recall\ (R)} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}}$$

F-Score: The mean value of precision and recall between 0 and 1 is calculated, with 1 being the ideal F-score value (Li et al., 2022). It can be calculated as follows:

$$\boldsymbol{F} = \frac{\boldsymbol{2PR}}{\boldsymbol{P + R}}$$

Area under the curve (AUC): AUC evaluates the region under the curve of true y positives and false x positives. This is helpful because it gives a single value that allows us to compare different types of models (Al-Ozeer et al., 2021). It can be calculated as follows:

$$\boldsymbol{False\ Positive\ Rate\ (FPR)} = \frac{\boldsymbol{FP}}{\boldsymbol{FP + TN}}$$

$$\boldsymbol{True\ Positive\ Rate\ (TPR)} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}}$$

$$\boldsymbol{AUC} = \sum_{\boldsymbol{i \in (TP+TP+FN+TN)}} \frac{\boldsymbol{(TPR_i + TPR_{i-1}).(FPR_i + FPR_{i-1})}}{\boldsymbol{2}}$$

## 4. Results and Discussion

In unsupervised ML, the K-Means algorithm is applied to the dataset to divide it into two clusters. Table 4 shows the results.

**Table 4: Results of Unsupervised ML.**

| Result Description | Average distance to other centers | Average distance to the Cluster center | Number of points | Maximal distance to the Cluster center |
|---|---|---|---|---|
| Evaluation for Cluster 0 | 1.652429 | 0.826701 | 2145 | 1.780448 |
| Evaluation for Cluster 1 | 1.616859 | 0.71384 | 1077 | 1.786231 |
| Combined evaluation | 1.640539 | 0.788975 | 3222 | 1.786231 |

The results indicate that the average distance to other centers for Cluster 0 is 1.652429, and for Cluster 1, 1.616859. This value is larger than 0.826701 and 0.71384, which represent the Average distance to the Cluster Center. This means that the cluster is well separated and distinct. Cluster 0 has 2145 points, and Cluster 1 has 1077 points. The percentage of each data point can be calculated using the following equation (Chitralekha and Roogi, 2021).

$$Percentage = \left(\frac{Number\ of\ points\ in\ Cluster\ 0}{Total\ number\ of\ points}\right) \times 100$$

$$Percentage = \left(\frac{2145}{3222}\right) \times 100 = 66.56\%$$

$$Percentage = \left(\frac{Number\ of\ points\ in\ Cluster\ 1}{Total\ number\ of\ points}\right) \times 100$$

$$Percentage = \left(\frac{1077}{3222}\right) \times 100 = 33.44\%$$

In supervised Machine Learning Analysis, the performance of six Supervised ML algorithms is evaluated, and accuracy is calculated for each algorithm to identify which algorithm has the highest performance and accuracy. The results are shown in Table 5.

**Table 5: Performance and accuracy evaluation of supervised ML algorithms**

| Algorithm | Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|
| Two-Class Support Vector Machine | 0.921 | 0.963 | 0.864 | 0.91 | 0.959 |
| Two-Class Averaged Perceptron | 0.927 | 0.97 | 0.868 | 0.916 | 0.957 |
| Two-Class Decision Forest | 0.928 | 0.926 | 0.917 | 0.921 | 0.969 |
| Two-Class Logistic Regression | 0.933 | 0.95 | 0.902 | 0.925 | 0.977 |
| Two-Class Neural Network | 0.94 | 0.964 | 0.904 | 0.933 | 0.962 |
| Two-Class Boosted Decision Tree | 0.942 | 0.969 | 0.904 | 0.935 | 0.962 |

The Two-Class Boosted Decision Tree algorithm has achieved the highest performance across multiple metrics, including accuracy (0.942), precision (0.969), recall (0.904), F1 score (0.935), and AUC (0.62). Its superior performance can be attributed to several factors related to the nature of the dataset. First, the dataset includes numerical variables (e.g., slope, solar radiation, land ownership), which BDT handles efficiently due to its ability to process mixed data types without requiring normalization or encoding. Second, the classification task involves capturing complex, non-linear relationships between environmental and infrastructural variables, for which BDT models are effective through ensemble learning. Third, the algorithm's robustness to outliers and missing values made it particularly suitable, given that some inputs (such as population density) varied widely across the study area. These characteristics made BDT more effective than simpler models, which struggled to generalize well across the diverse conditions of Nineveh Governorate (Imani et al., 2025). Figure (12) represents the results validation; the original outputs are mentioned in the appendices

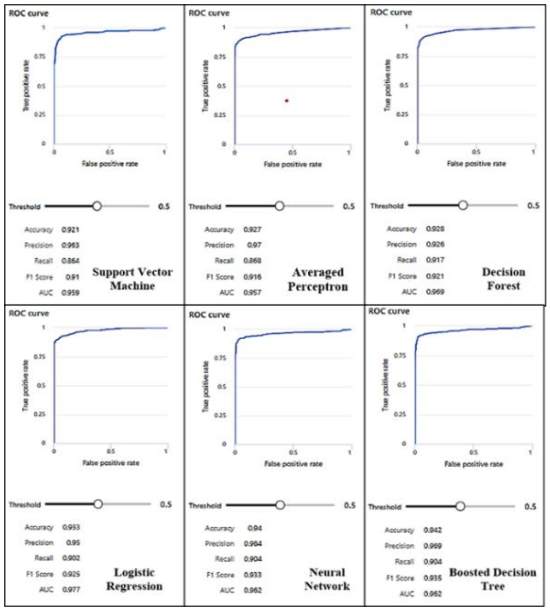



**Fig.12: Result validation (Accuracy, Precision, Recall, F1 Score, and AUC).**

### *Map of optimal site for solar farm*

The results produced by the Boosted Decision Tree algorithm are imported into ArcMap 10.8 and visualized using Empirical Bayesian Kriging interpolation. The final map highlights five levels of land suitability: very low, low, moderate, high, and very high, classified using the natural breaks method. Figure (13) shows the suitability distribution, and Table (6) provides the area for each category.

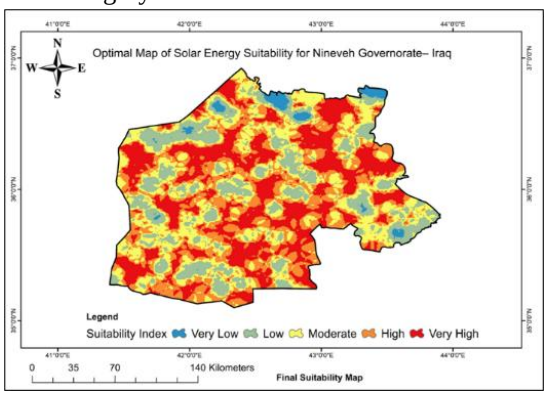



**Fig 13: Final Suitable Map of Solar Farm in Nineveh Governorate.**

**Table 6: Suitability analysis results by area.**

| Suitability | Area km$^2$ | Area Percentage |
|---|---|---|
| Very Low Suitable | 753.9507 | 1.90% |
| Low Suitable | 7039.3757 | 17.76% |
| Moderate Suitable | 10407.0165 | 26.26% |
| High Suitable | 10712.1845 | 27.03% |
| Very High Suitable | 10716.4476 | 27.04% |

Very High Suitable: This category covered 27.04% of the total study area. These sites offer favorable conditions for solar energy development such as high levels of solar radiation, low slopes, and fewer sandstorms. These areas are considered optimal for solar energy project development. High and Moderate Suitability: These areas are distributed across the study area and cover 53.29%; they are optimal for solar energy project development, but they have higher costs and require additional infrastructure investment such as power lines or road access. Low and Very Low Suitability: These classes represent areas that face significant challenges for the construction and development of solar energy because some factors negatively affect site suitability, such as steep slopes and land cover types such as tree cover or water bodies, and a high frequency of sandstorms, and it covers 19.66% of the study area. In the final map, the very high suitability zones are spread across the entire study area rather than being concentrated in a specific region, like the north or south. This even distribution presents a significant opportunity for solar energy development in Nineveh, offering greater flexibility in site selection by allowing planners to prioritize locations with better infrastructure access and potentially lower development costs. Dispersing solar farms throughout the governorate reduces the risk of overloading infrastructure in one area, minimizes transmission losses and enhances grid reliability and stability. Moreover, this spatial pattern supports balanced regional development by improving access to renewable energy in both urban and rural areas.

This research applies ML algorithms that analyze the database based on factor values such as solar radiation, slope, and sandstorms. It does not depend on the weight of factors, such as the AHP method, where some factors are more important than others. ML uses the values of the factors to determine the suitability of each point in the dataset, which is 3222. If a point has a high value of several factors, this point is considered highly suitable, even if some factors have low values. This is the reason why all classes of suitability are scattered throughout the study area (Ohmer et al., 2025).

Compared to the previous study by Al-Sabawi (2023), which used AHP combined with GIS in Nineveh, this research employs machine learning (ML) alongside GIS, resulting in notable

differences in site suitability outcomes and spatial patterns. Al-Sabawi's AHP approach depends heavily on expert judgment to assign weights to factors like solar radiation, slope, and proximity to infrastructure. This led to identifying very high suitability areas as only 7% of the governorate, clustered around existing roads and power lines, indicative of the infrastructure's dominant impact in AHP weighting. In contrast, the ML model examines intricate relationships between multiple factors based directly on data without depending on subjective weights. Consequently, it identified a larger portion, 27.04% of the area, as very highly suitable and distributed more widely across the province, including some remote zones. This broader detection highlights ML's ability to recognize favorable conditions based on actual data patterns such as optimal solar radiation and low sandstorm frequency, even when infrastructure proximity is less favorable.

## 5. Conclusion

This study represents the methodology of combining ML algorithms with GIS to determine a suitable location for a solar farm in Nineveh Governorate, Iraq. 13 factors have been selected affecting site suitability, which are grouped into environmental, climatic, topological, and socio-economic. By using both supervised and unsupervised ML, seven algorithms are applied, one of which is K-Means, used to generate target data, while six supervised algorithms train the data. The Boosted Decision Tree achieves the highest accuracy at 94.2%. As a result, the final suitable map, which is generated using EBK interpolation, indicates that 27.04% of the total study area is very high for solar farm development, and the suitability is distributed across the study area. Beyond identifying suitable locations, this research demonstrates how ML can improve decision-making in energy planning, making the process faster, more accurate, and less reliant on traditional manual assessments.

## 6. Recommendation

- Additional factors can be used, such as distance to faults, distance to oil pipeline, wind speed, and soil texture.
- This methodology can be applied to other governorates in Iraq with similar climatic conditions to determine optimal sites for solar energy projects and increase Iraq's renewable energy capacity.
- This framework can be used to identify suitable sites for wind, geothermal, hydro, and biomass energy.
- Further study can be applied for real-world validation of very suitable classes such as solar radiation level, actual slope, security of the region, and LST.

## 7. References

Abashidze, N. and Taylor, L.O., 2023. Utility-Scale Solar Farms and Agricultural Land Values. Land Economics, 99(3), pp. 327–342. https://doi.org/10.3368/LE.99.3.102920-0165R

Abdaki, M., Sanchez-Azofeifa, A., and Hamann, H.F., 2025. A Machine Learning Approach for Filling Long Gaps in Eddy Covariance Time Series Data in a Tropical Dry Forest. Journal of Geophysical Research: Biogeosciences, 130(1), e2024JG008375. https://doi.org/10.1029/2024JG008375

Abdo, T. and El-Shimy, M., 2013. Estimating the global solar radiation for solar energy projects – Egypt case study. International Journal of Sustainable Energy, 32(6), pp. 682–712. https://doi.org/10.1080/14786451.2013.822872

Abed, A.N., 2024. Framing Iraqi Refugees in The Guardian and Deutsche Welle: A Critical Discourse and Multimodal Analysis Study. https://doi.org/10.4995/THESIS/10251/207879

Ahmed Salman, H., Kalakech, A., Steiti, A., and History, A., 2024. Random Forest Algorithm Overview. Babylonian Journal of Machine Learning, 2024, pp. 69–79. https://doi.org/10.58496/BJML/2024/007

Alanzi, S. S., Aldalali, B. and Kamel, R. M., 2024. Effects of sandstorms on hybrid renewable energy sources and load demand in arid desert climates: A case study. Energy for Sustainable Development, 81, 101473. https://doi.org/10.1016/J.ESD.2024.101473

Aldrin Wiguna, K., Sarno, R. and Ariyani, N. F., 2017. Optimization Solar Farm site selection using Multi-Criteria Decision Making Fuzzy AHP and PROMETHEE: Case study in Bali. Proceedings of 2016 International Conference on Information and Communication Technology and Systems, ICTS 2016, 237–243. https://doi.org/10.1109/ICTS.2016.7910305

Almuqati, M. T., Sidi, F., Rum, S. N. M., Zolkepli, M. and Ishak, I., 2024. Challenges in Supervised and Unsupervised Learning: A Comprehensive Overview. International Journal on Advanced Science, Engineering and Information Technology, 14(4), 1449–1455. https://doi.org/10.18517/IJASEIT.14.4.20191

Al-Ozeer, A. ZA., Al-Abadi, A. M., Hussain, T. A., Fryar, A. E., Pradhan, B., Alamri, A. and Maulud, K. N. A., 2021. Modeling of Groundwater Potential Using Cloud Computing Platform: A Case Study from Nineveh Plain, Northern Iraq. Earth and Environmental Sciences

Faculty Publications, 13(23), 3330. https://doi.org/https://doi.org/10.3390/w13233330

Al-Saidi, M. and Lahham, N., 2019. Solar energy farming as a development innovation for vulnerable water basins. Development in Practice, 29(5), 619–634. https://doi.org/10.1080/09614524.2019.1600659

Al-Sabawi, W. W., 2023. Selection of sites for the production of solar photovoltaic energy in Nineveh Governorate using geographic information systems. University of Mosul.

Al-Thanoon, N.A., Qasim, O.S. and Algamal, Z.Y., 2018. Tuning parameter estimation in SCAD-support vector machine using firefly algorithm with application in gene selection and cancer classification. Computers in Biology and Medicine, 103, 262–268. https://doi.org/10.1016/j.compbiomed.2018.10.03

Ashraf, H. A., Li, J., Li, Z., Sohail, A., Ahmed, R., Butt, M. H. and Ullah, H., 2025. Geographic Information System and Machine Learning Approach for Solar Photovoltaic Site Selection: A Case Study in Pakistan. Processes 2025, Vol. 13, Page 981, 13(4), 981. https://doi.org/10.3390/PR13040981

Azure Data Studio | Microsoft Azure. (n.d.). Retrieved 21 January 2025, from https://azure.microsoft.com/en-us/products/data-studio

Badi, I., Pamucar, D., Gigović, L. and Tatomirović, S., 2021. Optimal site selection for sitting a solar park using a novel GIS- SWA'TEL model: A case study in Libya. International Journal of Green Energy, 18(4), 336–350. https://doi.org/10.1080/15435075.2020.1854264

Bandira, P. N. A., Tan, M. L., Teh, S. Y., Samat, N., Shaharudin, S. M., Mahamud, M. A., Tangang, F., Juneng, L., Chung, J. X. and Samsudin, M. S., 2022. Optimal Solar Farm Site Selection in the George Town Conurbation Using GIS-Based Multi-Criteria Decision Making (MCDM) and NASA POWER Data. Atmosphere 2022, Vol. 13, Page 2105, 13(12), 2105. https://doi.org/10.3390/ATMOS13122105

Berrar, D., 2024. Performance Measures for Binary Classification 1.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. (2024). A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology, 15(3), 39. https://doi.org/10.1145/3641289

Chen, J., Liu, B., Liao, X., Gao, J., Zheng, H., and Li, Y. (2024). Adaptive Optimization for Enhanced Efficiency in Large-Scale Language Model Training. https://arxiv.org/abs/2412.04718v1

Chitralekha, G. and Roogi, J. M., 2021. A Quick Review of ML Algorithms. Proceedings of the 6th International Conference on Communication and Electronics Systems, ICCES 2021. https://doi.org/10.1109/ICCES51350.2021.9488982

Coadou, Y., 2022. Boosted decision trees. Artificial Intelligence for High Energy Physics, 9–58. https://doi.org/10.1142/9789811234033_0002

Elboshy, B., Alwetaishi, M., M. H. Aly, R. and Zalhaf, A. S., 2022. A suitability mapping for the PV solar farms in Egypt based on GIS-AHP to optimize multi-criteria feasibility. Ain Shams Engineering Journal, 13(3), 101618. https://doi.org/10.1016/J.ASEJ.2021.10.013

Farahani, M., Kazemi, A., Hedayati Aghmashadi, A., Masteri Farahani, F., Kazemi, A. and Hedayati Aghmashadi, A., 2024. Optimal Site Selection of Solar Power Plant Stations Using GIS-ANP and Genetic Optimization Algorithm in Markazi Province, Iran. Journal of Green Energy Research and Innovation, 1(4), 47–63. https://doi.org/10.61186/JGERI.1.4.47

Fatah, K. K., Mustafa, Y. T., and Hassan, I. O., 2022. Flood Susceptibility Mapping Using an Analytic Hierarchy Process Model Based on Remote Sensing and GIS Approaches in Akre District, Kurdistan Region, Iraq. The Iraqi Geological Journal, 55(2), 123–151. https://doi.org/10.46717/IGJ.55.2C.10MS-2022-08-23

Fatah, K. K., Mustafa, Y. T., and Hassan, I. O., 2024. Geoinformatics-based frequency ratio, analytic hierarchy process and hybrid models for landslide susceptibility zonation in Kurdistan Region, Northern Iraq. Environment, Development and Sustainability, 26(3), 6977–7014. https://doi.org/10.1007/S10668-023-02995-7/METRICS

Georgiou, A., and Skarlatos, D., 2016. Optimal site selection for sitting a solar park using multi-criteria decision analysis and geographical information systems. Geoscientific Instrumentation, Methods and Data Systems, 5(2), 321–332. https://doi.org/10.5194/GI-5-321-2016

Ghosh, N., and Halder, G., 2022. Current progress and perspective of heterogeneous nanocatalytic transesterification towards biodiesel production from edible and inedible feedstock: A review. Energy Conversion and Management, 270, 116292. https://doi.org/10.1016/J.ENCONMAN.2022.116292

Goshem, G. K., and Hailu, B. T., 2022. Solar Farm Site Selection Using Spatial Multi-Criteria Evaluation Method: A Case Study of Kewet Wereda, North-Center Ethiopia. https://doi.org/10.21203/RS.3.RS-1007080/V1

Hassan, Q., Algburi, S., Al-Musawi, T. J., Viktor, P., Jaszczur, M., Barakat, M., Sameen, A. Z., and Hussein, A. A. H., 2024. GIS-based multi-criteria analysis for solar, wind, and biomass energy potential: A case study of Iraq with implications for climate goals. Results in Engineering, 22, 102212. https://doi.org/10.1016/J.RINENG.2024.102212

Hernandez, R. R., Hoffacker, M. K., Murphy-Mariscal, M. L., Wu, G. C., and Allen, M. F., 2015. Solar energy development impacts on land cover change and protected areas. Proceedings of the National Academy of Sciences of the United States of America, 112(44), 13579–13584. https://doi.org/10.1073/PNAS.1517656112/SUPPL_FILE/PNAS.1517656112.ST05.DOCX

Hua, M., and Pei, J., 2007. Cleaning disguised missing data: A heuristic approach. Proceedings of the ACM SIGKDD International Conference on Knowledge

Discovery and Data Mining, 950–958. https://doi.org/10.1145/1281192.1281294

Hussain, M. T., and Mahdi, E. J., 2018. Assessment of Solar Photovoltaic Potential in Iraq. Journal of Physics: Conference Series, 1032(1), 012007. https://doi.org/10.1088/1742-6596/1032/1/012007

Ibrahim, G. R. F., Rasul, A., Hamid, A. A., Ali, Z. F., and Dewana, A. A., 2019. Suitable site selection for rainwater harvesting and storage case study using Dohuk governorate. Water (Switzerland), 11(4). https://doi.org/10.3390/W11040864

Imani, M., Beikmohammadi, A., and Arabnia, H. R., 2025. Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels. Technologies 2025, Vol. 13, Page 88, 13(3), 88. https://doi.org/10.3390/TECHNOLOGIES13030088

Jiang, D., Jones, I., Liu, X., Simis, S. G. H., Cretaux, J. F., Albergel, C., Tyler, A., and Spyrakos, E., 2024. Impacts of droughts and human activities on water quantity and quality: Remote sensing observations of Lake Qadisiyah, Iraq. International Journal of Applied Earth Observation and Geoinformation, 132, 104021. https://doi.org/10.1016/J.JAG.2024.104021

Joseph, V. R., and Vakayil, A., 2022. SPlit: An Optimal Method for Data Splitting. Technometrics, 64(2), 166–176. https://doi.org/10.1080/00401706.2021.1921037

Kamali Saraji, M., Streimikiene, D., and Suresh, V., 2024. A novel two-stage multicriteria decision-making approach for selecting solar farm sites: A case study. Journal of Cleaner Production, 444, 141198. https://doi.org/10.1016/J.JCLEPRO.2024.141198

Kereush, D., and Geomatics, I. P., 2017. Determining criteria for optimal site selection for solar power plants. Geomatics, Landmanagement and Landscape, 2017(4), 39–54. https://doi.org/10.15576/GLL/2017.4.39

Khazael, S. M., and Al-Bakri, M., 2021. The Optimum Site Selection for Solar Energy Farms using AHP in GIS Environment, A Case Study of Iraq. Iraqi Journal of Science, 62(11), 4571–4587. https://doi.org/10.24996/IJS.2021.62.11(SI).36

Kırcalı, Ş., and Selim, S., 2021. Site suitability analysis for solar farms using the geographic information system and multi-criteria decision analysis: the case of Antalya, Turkey. Clean Technologies and Environmental Policy, 23(4), 1233–1250. https://doi.org/10.1007/S10098-020-02018-3/METRICS

Koc, A., Turk, S., and Şahin, G., 2019. Multi-criteria of wind-solar site selection problem using a GIS-AHP-based approach with an application in Igdir Province/Turkey. Environmental Science and Pollution Research, 26(31), 32298–32310. https://doi.org/10.1007/S11356-019-06260-1/METRICS

Li, B., Wu, F., Lim, S.-N., Belongie, S., and Weinberger, K. Q., 2021. On Feature Normalization and Data Augmentation (pp. 12383–12392). https://github.com/Boyiliee/MoEx.

Li, X. Y., Dong, X. Y., Chen, S., and Ye, Y. M., 2024. The promising future of developing large-scale PV solar farms in China: A three-stage framework for site selection. Renewable Energy, 220, 119638. https://doi.org/10.1016/J.RENENE.2023.119638

Li, Z., Yoon, J., Zhang, R., Rajabipour, F., Srubar, W. V., Dabo, I., and Radlińska, A., 2022. Machine learning in concrete science: applications, challenges, and best practices. Npj Computational Materials 2022 8:1, 8(1), 1–17. https://doi.org/10.1038/s41524-022-00810-x

Mahmood Faisal, R., and Abdaki, M., 2021. multi-criteria analysis for selecting suitable sites of water harvesting in northern al tharthar watershed. Penerbit UMT Journal of Sustainability Science and Management, 16, 218–236. https://doi.org/10.46754/jssm.2021.10.017

Maxwell, A. E., Warner, T. A., and Guillén, L. A., 2021. Accuracy Assessment in Convolutional Neural Network-Based Deep Learning Remote Sensing Studies—Part 1: Literature Review. Remote Sensing 2021, Vol. 13, Page 2450, 13(13), 2450. https://doi.org/10.3390/RS13132450

Mierzwiak, M., and Calka, B., 2017. Multi-Criteria Analysis for Solar Farm Location Suitability. Reports on Geodesy and Geoinformatics, 104(1), 20–32. https://doi.org/10.1515/RGG-2017-0012

Mizumoto, A., and Eguchi, M., 2023. Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2(2), 100050. https://doi.org/10.1016/J.RMAL.2023.100050

Naidu, G., Zuva, T., and Sibanda, E. M., 2023. A Review of Evaluation Metrics in Machine Learning Algorithms. Lecture Notes in Networks and Systems, 724 LNNS, 15–25. https://doi.org/10.1007/978-3-031-35314-7_2

Nebey, A. H., Taye, B. Z., and Workineh, T. G., 2020. Site Suitability Analysis of Solar PV Power Generation in South Gondar, Amhara Region. Journal of Energy, 2020(1), 3519257. https://doi.org/10.1155/2020/3519257

Niu, C., Shan, H., and Wang, G., 2022. SPICE: Semantic Pseudo-Labeling for Image Clustering. IEEE Transactions on Image Processing, 31, 7264–7278. https://doi.org/10.1109/TIP.2022.3221290

Noori, N. M., and Qasim, O. S., 2023. Deep Features Selections with Binary Marine Predators Algorithm for Effective Classification of Image Datasets. Fusion: Practice and Applications, 10(2), 86–94. https://doi.org/10.54216/FPA.100208

Ohmer, M., Doll, F., and Liesch, T., 2025. Incorporating Spatial Information for Regionalization of Environmental Parameters in Machine Learning Models. Mathematical Geosciences, 57(2), 251–273. https://doi.org/10.1007/S11004-024-10163-4/FIGURES/5

Omar, E. D., Mat, H., Karim, A. Z. A., Sanaudi, R., Ibrahim, F. H., Omar, M. A., Ismail, M. Z. H., Jayaraj, V. J., and Goh, B. L., 2024. Comparative Analysis of Logistic Regression, Gradient Boosted Trees, SVM, and

Random Forest Algorithms for Prediction of Acute Kidney Injury Requiring Dialysis After Cardiac Surgery. International Journal of Nephrology and Renovascular Disease, 17, 197. https://doi.org/10.2147/IJNRD.S461028

Qin, S., Zhang, Y., Sun, K., and Chen, F., 2025. Aileron fault detection with dynamic resampling based on fuzzy entropy and multi-branch neural network. Measurement, 241, 115773. https://doi.org/10.1016/J.MEASUREMENT.2024.115773

Rahman, A., Farrok, O., and Haque, M. M., 2022. Environmental impact of renewable energy source based electrical power plants: Solar, wind, hydroelectric, biomass, geothermal, tidal, ocean, and osmotic. Renewable and Sustainable Energy Reviews, 161, 112279. https://doi.org/10.1016/J.RSER.2022.112279

Rane, N. L., Günen, M. A., Mallick, S. K., Rane, J., Pande, C. B., Giduturi, M., Bhutto, J. K., Yadav, K. K., Tolche, A. D., and Alreshidi, M. A., 2024. GIS-based multi-influencing factor (MIF) application for optimal site selection of solar photovoltaic power plant in Nashik, India. Environmental Sciences Europe, 36(1), 1–25. https://doi.org/10.1186/S12302-023-00832-2/TABLES/7

Renewable Energy Agency, 2024. World Energy Transitions Outlook 2024: 1.5°C pathway. www.irena.org

Rusol, I.A., and Al-Timimi, Y.K., 2024. Determination of optimum sites for solar energy harvesting in Iraq using multi-criteria. iraqi journal of agricultural sciences, 55(Special), 25–33. https://doi.org/10.36103/IJAS.V55ISPECIAL.1882

Sachit, M. S., Shafri, H. Z. M., Abdullah, A. F., Rafie, A. S. M., and Gibril, M. B. A., 2022. Global Spatial Suitability Mapping of Wind and Solar Systems Using an Explainable AI-Based Approach. ISPRS International Journal of Geo-Information, 11(8), 422. https://doi.org/10.3390/IJGI11080422/S1

Sadeghi, M., and Karimi, M., 2017. GIS-based solar and wind turbine site selection using multi-criteria analysis: case study tehran, iran. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-4-W4(4W4), 469–476. https://doi.org/10.5194/ISPRS-ARCHIVES-XLII-4-W4-469-2017

Salama, S. W., 2024. Towards advanced sustainable criteria for choosing the best site for collecting solar energy in cities using multi-criteria GIS. City, Territory and Architecture, 11(1), 1–14. https://doi.org/10.1186/S40410-024-00243-7/FIGURES/22

Scovell, M., McCrea, R., Walton, A., and Poruschi, L, .2024. Local acceptance of solar farms: The impact of energy narratives. Renewable and Sustainable Energy Reviews, 189, 114029. https://doi.org/10.1016/J.RSER.2023.114029

Shrestha, A., Mustafa, A. A., Htike, M. M., You, V., and Kakinaka, M., 2022. Evolution of energy mix in emerging countries: Modern renewable energy, traditional renewable energy, and non-renewable energy. Renewable Energy, 199, 419–432. https://doi.org/10.1016/J.RENENE.2022.09.018

Spyridonidou, S., and Vagiona, D. G., 2023. A systematic review of site-selection procedures of PV and CSP technologies. Energy Reports, 9, 2947–2979. https://doi.org/10.1016/J.EGYR.2023.01.132

Suprova, N. T., Zidan, A. R., and Rashid, A., 2021. Optimal Site Selection for Solar Farms Using GIS and AHP: A Literature Review.

Tian, Y., and Tang, X., 2025. The use of artificial neural network algorithms to enhance tourism economic efficiency under information and communication technology. Scientific Reports 2025 15:1, 15(1), 1–13. https://doi.org/10.1038/s41598-025-94268-8

Uyan, M., and Dogmus, O. L., 2023. An Integrated GIS-Based ANP Analysis for Selecting Solar Farm Installation Locations: Case Study in Cumra Region, Turkey. Environmental Modeling and Assessment, 28(1), 105–119. https://doi.org/10.1007/S10666-022-09870-1/METRICS

Westgate, M., Kucinskaite, K., Konstantinidis, E., Malehmir, A., Papadopoulou, M., Gregersen, U., Keiding, M., and Bjerager, M., 2025. Seismic Imaging of Halokinetic Sequences and Structures With High-Resolution, Dual-Element Acquisition, and Processing: Applications to the Gassum Structure in Eastern Jutland, Denmark. Earth and Space Science, 12(1), e2024EA004014.https://doi.org/10.1029/2024EA004014

Xu, Z., Li, Y., Qin, Y., and Bach, E., 2024. A global assessment of the effects of solar farms on albedo, vegetation, and land surface temperature using remote sensing. Solar Energy, 268, 112198. https://doi.org/10.1016/J.SOLENER.2023.112198

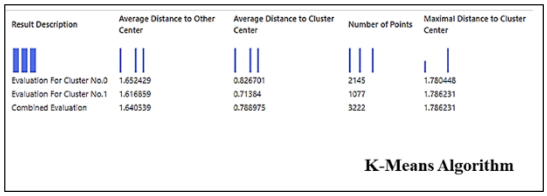
# 8. APPENDICES

***Dataset***

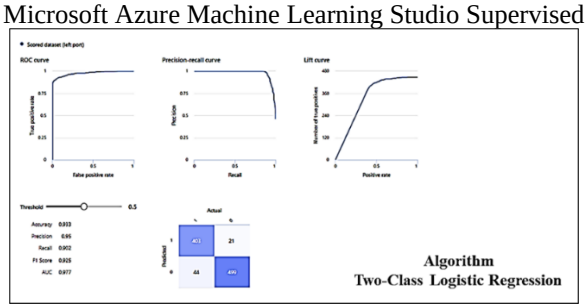
The dataset consists of 3222 rows, which will take about 240 pages to display. For this reason, the file is attached to the QR code below or directly through this link.

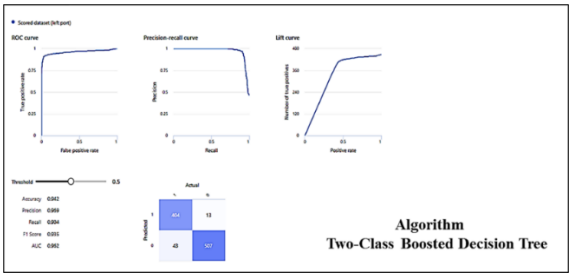
| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Aspect | Land Cover | Population | Sandstorm | Solar Radiation | LST | Land Ownership | Roads | Power Line | Substation | Slop | water Shead | Lakes | Target |
| 2 | 3 | 5 | 5 | 3 | 5 | 3 | 3 | 2 | 1 | 5 | 5 | 5 | 5 | 1 |
| 3 | 3 | 5 | 5 | 3 | 5 | 3 | 3 | 3 | 1 | 5 | 5 | 5 | 5 | 1 |
| 4 | 3 | 5 | 5 | 3 | 5 | 3 | 3 | 3 | 1 | 5 | 4 | 5 | 5 | 1 |
| 5 | 5 | 5 | 5 | 3 | 5 | 3 | 3 | 3 | 1 | 5 | 3 | 5 | 5 | 1 |
| 6 | 5 | 5 | 5 | 3 | 5 | 3 | 3 | 3 | 1 | 5 | 5 | 5 | 5 | 1 |
| 7 | 4 | 5 | 5 | 3 | 5 | 3 | 3 | 3 | 1 | 5 | 3 | 5 | 5 | 1 |
| 8 | 1 | 4 | 5 | 3 | 5 | 3 | 3 | 3 | 1 | 5 | 3 | 5 | 5 | 1 |
| 9 | 4 | 5 | 5 | 3 | 5 | 3 | 3 | 4 | 1 | 5 | 4 | 5 | 5 | 1 |
| 10 | 1 | 5 | 5 | 3 | 5 | 3 | 3 | 4 | 1 | 5 | 5 | 5 | 5 | 1 |
| 11 | 4 | 4 | 5 | 3 | 5 | 3 | 3 | 3 | 1 | 5 | 5 | 5 | 5 | 1 |
| 12 | 5 | 5 | 5 | 3 | 5 | 3 | 3 | 4 | 1 | 5 | 4 | 5 | 5 | 1 |
| 13 | 4 | 4 | 5 | 3 | 5 | 3 | 3 | 4 | 1 | 5 | 4 | 5 | 5 | 1 |
| 14 | 5 | 5 | 5 | 3 | 5 | 3 | 3 | 4 | 1 | 5 | 5 | 5 | 5 | 1 |
| 15 | 4 | 5 | 5 | 3 | 5 | 3 | 3 | 4 | 1 | 5 | 5 | 5 | 5 | 1 |
| 16 | 1 | 5 | 5 | 3 | 5 | 3 | 3 | 3 | 1 | 5 | 5 | 5 | 5 | 1 |
| 17 | 4 | 5 | 5 | 3 | 5 | 3 | 3 | 3 | 1 | 5 | 5 | 5 | 5 | 1 |
| 18 | 3 | 4 | 5 | 3 | 5 | 3 | 3 | 3 | 1 | 5 | 5 | 5 | 5 | 1 |
| 19 | 3 | 5 | 5 | 3 | 5 | 3 | 3 | 5 | 1 | 5 | 5 | 5 | 5 | 1 |
| 20 | 2 | 5 | 5 | 3 | 5 | 3 | 3 | 5 | 1 | 5 | 5 | 5 | 5 | 1 |
| 21 | 4 | 5 | 5 | 3 | 5 | 4 | 3 | 5 | 1 | 5 | 4 | 5 | 5 | 1 |
| 22 | 5 | 5 | 5 | 3 | 5 | 3 | 3 | 4 | 1 | 5 | 5 | 5 | 5 | 1 |

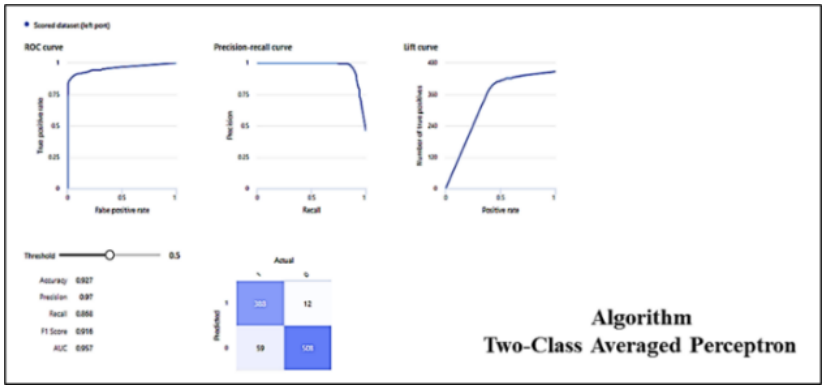
Microsoft Azure Machine Learning Studio Unsupervised

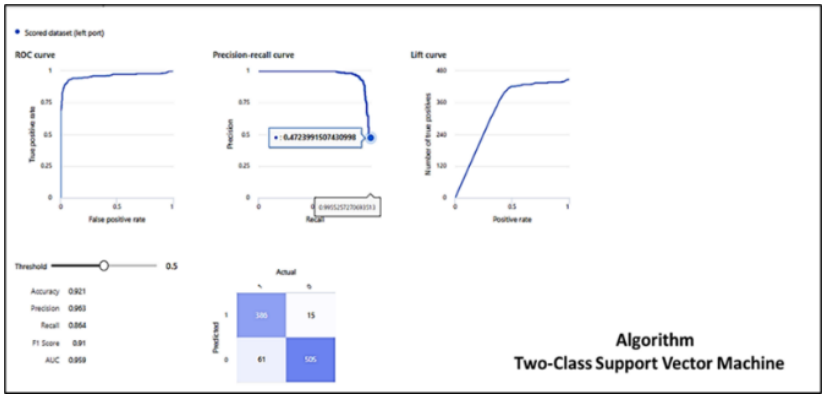


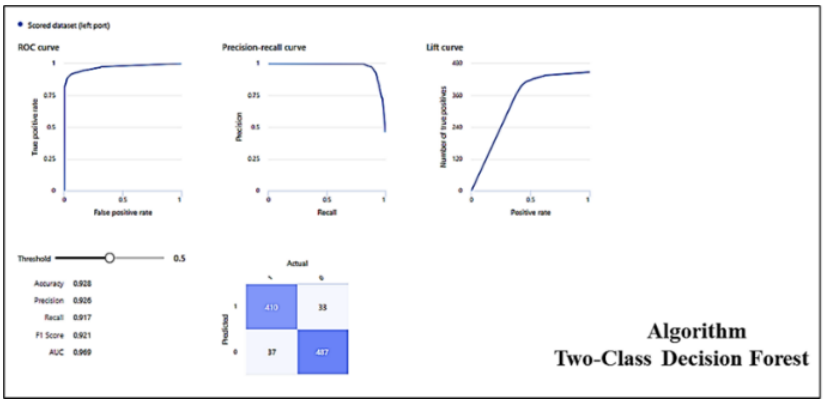

Microsoft Azure Machine Learning Studio Supervised

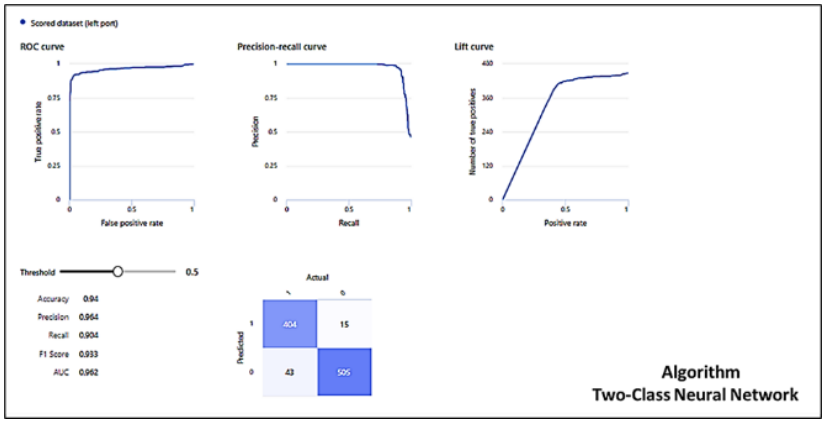



Scored dataset (left port)
ROC curve
Precision-recall curve
Lift curve
Threshold 0.5
Accuracy 0.927
Precision 0.97
Recall 0.866
F1 Score 0.916
AUC 0.957
Actual
Predicted
Algorithm
Two-Class Averaged Perceptron


Scored dataset (left port)
ROC curve
Precision-recall curve
Lift curve
Threshold 0.5
Accuracy 0.921
Precision 0.963
Recall 0.864
F1 Score 0.91
AUC 0.959
Actual
Predicted
Algorithm
Two-Class Support Vector Machine


Scored dataset (left port)
ROC curve
Precision-recall curve
Lift curve
Threshold 0.5
Accuracy 0.928
Precision 0.926
Recall 0.917
F1 Score 0.921
AUC 0.969
Actual
Predicted
Algorithm
Two-Class Decision Forest


Scored dataset (left port)
ROC curve
Precision-recall curve
Lift curve
Threshold 0.5
Accuracy 0.94
Precision 0.964
Recall 0.904
F1 Score 0.933
AUC 0.962
Actual
Predicted
Algorithm
Two-Class Neural Network

*Iraqi National Journal of Earth Science, Vol. 26, No. 03, 2026 (207-224)*

**Iraqi National Journal of Earth Science**

www.earth.uomosul.edu.iq

# تطوير تخطيط الطاقة الشمسية باستخدام التعلم الآلي ونظم المعلومات الجغرافية: تحليل مدى الملاءمة في محافظة نينوى العراق

**صلاح جانكير ابراهيم[1]*** salah.23evp15@student.uomosul.edu.iq ، **علي زين العابدين حيدر [2]** aalozeer@uomosul.edu.iq ، **قصي كمال الدين الاحمدي[3]** prof.kossayalahmady@uomosul.edu.iq

[1] *قسم علوم البيئية، كلية علوم البيئية، جامعة الموصل، دهوك، العراق.*

[2] *قسم التغيرات المناخية، كلية علوم البيئية، جامعة الموصل، موصل، العراق.*

[3] *قسم الهندسة البيئية، كلية الهندسة، جامعة الموصل، موصل، العراق.*

تاريخ الاستلام: 14 نيسان 2025 تاريخ المراجعة: 29 ايار 2025 تاريخ القبول: 07 تموز 2025

تاريخ النشر الالكتروني: 01 تموز 2026

**الملخص**

*تشهد الطاقة الشمسية نموًا سريعًا نظرًا لوفرتها ونظافتها وفعاليتها من حيث التكلفة وسهولة تركيبها. تلعب عملية اختيار الموقع دورًا حاسمًا في تحقيق أقصى كفاءة لمشاريع الطاقة الشمسية مع تقليل التأثير البيئي. تركز هذه الدراسة على محافظة نينوى، العراق. تم تقديم منهجية متقدمة لاختيار مواقع المشاريع الشمسية باستخدام التعلم الآلي (ML) ونظام المعلومات الجغرافية (GIS) لتحقيق هذا الهدف، تم تطبيق سبعة خوارزميات وهي Logistic Regression, Averaged Perceptron, Boosted Decision Tree, Decision Forest, Support Vector Machine, Neural Network و K-Means لتحديد موقع مناسب للمزارع الشمسية. تم تحديد ثلاثة عشر عاملاً مؤثرًا على اختيار مواقع المزارع الشمسية، بما في ذلك الإشعاع الشمسي، العواصف الرملية، درجة حرارة سطح الأرض، الطرق الرئيسية، السكان، خطوط الكهرباء، المحطات الفرعية، ملكية الأراضي، الغطاء الأرضي، الموارد المائية، الانحدار، والاتجاه. ثم تصنيف هذه العوامل إلى أربع مجموعات: بيئية، مناخية، طوبوغرافية، واجتماعية–اقتصادية. بعد ذلك، تم تحليل مجموعة البيانات في منصة Azure Machine Learning باستخدام تقنيات unsupervised و supervised ML. وكنتيجة لذلك، حقق نموذج Boosted Decision Tree أعلى دقة بلغت 94.2% ، حيث تم تصنيف 27% من منطقة الدراسة كمنطقة عالية جدًا من حيث الملاءمة لتطوير المزارع الشمسية. تؤكد هذه الدراسة على الإمكانات الهائلة لمحافظة نينوى لمشاريع الطاقة المتجددة، وتُنشئ إطار عمل يمكن تكراره يربط مابين GIS و ML في تخطيط احتياجات الطاقة بأقل تأثيرات بيئية ممكنة.*

**الكلمات المفتاحية:**

*الطاقة الشمسية، نظام المعلومات الجغرافية، التعلم الآلي، اختيار الموقع الأمثل، العراق.*

*DOI: 10.33899/injes.v26i3.60958, ©Authors, 2026, College of Science, University of Mosul.*



==========