Analysis and Prediction of Energy Consumption Using Supervised Machine Learning Techniques: A Study of Libyan Electricity Company Data

Ashraf Mohammed Abusida[1] and Aybaba Hançerlioğullari[2]
ashrafbouseda@gmail.com aybaba@kastamonu.edu.tr
[1]Department of Computer Engineering, Kastamonu University, Kastamonu, Turkey,
[2]Art & science faculty, Physics Department, Kastamonu University, Kastamonu, Turkey,

Abstract
The ever-increasing amount of data generated by various industries and systems has led to the development of data mining techniques as a means to extract valuable insights and knowledge from such data. The electrical energy industry is no exception, with the large amounts of data generated by SCADA systems. This study focuses on the analysis of historical data recorded in the SCADA database of the Libyan Electricity Company. The database, spanned from January 1st, 2013, to December 31st, 2022, contains records of daily date and hour, energy production, temperature, humidity, wind speed, and energy consumption levels. The data was pre-processed and analyzed using the WEKA tool and the Apriori algorithm, a supervised machine learning technique. The aim of the study was to extract association rules that would assist decision-makers in making informed decisions with greater efficiency and reduced costs. The results obtained from the study were evaluated in terms of accuracy and production time, and the conclusion of the study shows that the results are promising and encouraging for future use in the Libyan Electricity Company. The study highlights the importance of data mining and the benefits of utilizing machine learning technology in decision-making processes.

Keywords:
GECOL, Energy consumption levels, SCADA, Association Rules.

1. Introduction

In today's fast-paced and highly competitive business environment, decision-making plays a critical role in the success of organizations. Data mining, as a subset of artificial intelligence, has emerged as a powerful tool for uncovering hidden patterns and trends in large datasets, providing valuable insights to support effective decision-making [1]. The use of machine learning algorithms in data mining can enhance the efficiency and accuracy of these insights.

The energy sector is one of the most important sectors in any economy, and the Libyan Electricity Company (GECOL) is no exception [2]. The company's historical data, recorded in the SCADA database, contains information about the daily date and hour, energy production (in MW), temperature, humidity, wind speed, and energy consumption levels (low, medium, high) spanning from January 2013 to December 2022. This vast amount of data presents a valuable opportunity to extract insights that can inform future decision-making.

The overall electric organization of Libya GECOL was set up under Act No. 17 of 1984 AD and is liable for the completion of the activities for working and adjusting the electric organizations, stations for energy creation and their appropriation, and change stations. Likewise, the organization is responsible for the energy transmission lines and their appropriation, the power control centers, and the administration of the activity and adjustment of desalination stations in the entire nation [5].

To address these challenges, the Association Rules and Apriori Algorithm have been proposed to support the automated prediction of energy consumed by the electric network and reduce or prevent load shedding. This technique can predict the expected amount of energy consumed based on peak hours of the day, specific periods of the year, and weather conditions, thus maintaining the stability of the electrical network and reducing costs. Additionally, machine learning techniques can assist decision-makers in taking more informed and quicker decisions, thereby helping to maintain the quality of service provided by the company. [5].

The company's industrial computer system, which monitors and controls the transmission and distribution elements of electrical utilities, including substations, transformers, and other electrical assets, has been applied to SCADA data for the first time [6].

Predicting the hourly energy consumed and managing electric load shedding status is crucial for the stability of the electrical network and the improvement of the quality of services provided by the company. Conventional forecasting methods, however, are time-consuming, costly, and require expertise in the production, distribution, and transmission of electrical energy, as well as prior knowledge of the loads to be provided in the electrical network [5]. With the Association Rules and Apriori Algorithm, the prediction of energy consumed can be automated, reducing the need for manual forecasting and the associated costs.

Manuscript received March 5, 2023
Manuscript revised March 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.3.2

In Zhou and Wang (2010) provides a method for diagnosing power transformers using association rules based on rough sets. The rough sets are used to identify the relationships between different factors and the power transformer performance. The method involves reducing the rough sets to eliminate features that may negatively impact performance. The results of this study show that the approach is effective [7].

The purpose of this study is to investigate the use of machine learning techniques in analyzing the SCADA database of GECOL. The study utilizes the WEKA tool and the Apriori algorithm to extract association rules that can assist decision-makers in making informed decisions with increased efficiency and cost-effectiveness. The focus of the study is to assess the accuracy and processing time of the results obtained, with the ultimate aim of highlighting the benefits of machine learning in analyzing large datasets and its potential impact on decision-making within the energy sector.

A comprehensive review of academic studies on the topic of machine learning and data mining techniques has been conducted. Over the past two decades, numerous research has been conducted on data mining techniques and rule extraction from datasets. These studies have provided valuable insights on the implementation of machine learning algorithms, but none have specifically focused on analyzing and utilizing the database of GECOL to implement data preparation and division techniques and identify the factors that affect the results of machine learning algorithms. The goal is to develop accurate prediction models [3].

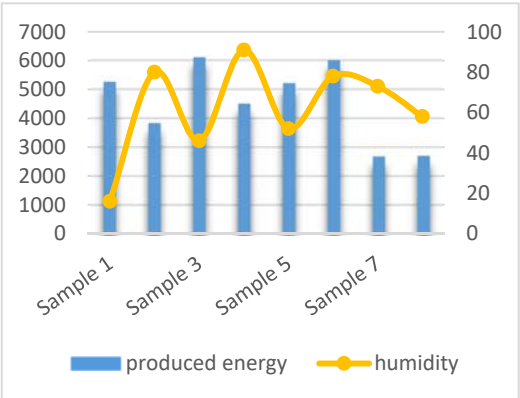

Figure 1. A sample of data showing the value of (energy produced, Temperature)

2. Material

The dataset used in this study consists of historical data on hourly energy production in MW, temperature in Celsius, humidity as a percentage, and wind speed in kilometers per hour, covering the period from January 2013 to December 2022, as shown in Table 1.

Table 1. GECOL Dataset Properties

Variables	Unit	Description
Date	Date time	Represent of Date of day
Time	Time	Represent the time in the day
Energy Generation	Numeric	Represent The amount of electrical energy produced by (MW)
Temperature	Numeric	Represent of Temperature
Humidity	Numeric	Represent a percentage of Humidity
Wind Speed	Numeric	Represent The wind speed by km/h
Energy consumation Level	charcter	{High, Medium, Low}

This study uses collected data to conduct experiments. Figure 1 represents the data in a simplified manner, including the hourly energy production, consumption level, and the impact of weather conditions on energy consumption throughout the day and year.

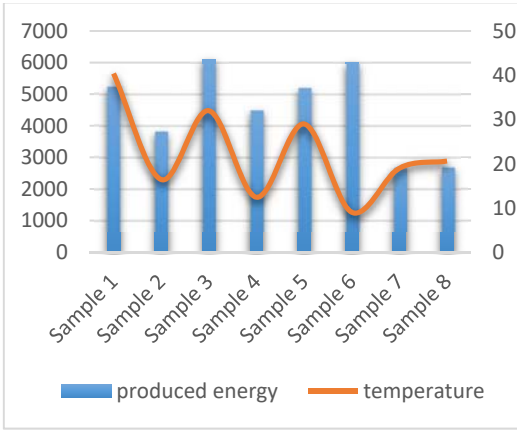

Figure 2. A sample of data showing the value of (energy produced, Humidity)

The figures above show a sample of random data about the production of electrical energy, including a range of time periods, peak times of electrical energy use, and different seasons (like summer and winter), all of which happened in different weather conditions with different humidity, temperature, and wind speed. The results of this study will give important information to decision-makers

that can be used to make plans and make sure the quality and continuity of service.

2.1. Data mining

With the advancement of technologies such as computers and satellites, the information age has seen a tremendous increase in the collection and storage of vast amounts of data. However, the storage of this massive amount of information across various structures has resulted in an overwhelming amount of data that is not utilized effectively [13]. The utilization of data mining techniques can provide a solution to this issue by enabling the extraction of valuable insights from this collected data. With the increasing amounts of data being collected and stored in today's digital age, the extraction of meaningful and useful information from these large databases has become a crucial task. Data Mining techniques aim to identify patterns and relationships within the data, providing companies with the necessary information to make informed decisions. The study of Data Mining involves the development of algorithms and computational paradigms that enable computers to uncover the structure of databases, make predictions and forecasts, and improve their overall performance through interaction with data. The field of Machine Learning is concerned with the creation of computer systems that can improve their performance in a specific domain through the process of learning and experience. The utilization of data mining techniques for the extraction of valuable information from large databases has become an increasingly prevalent practice in recent years. One such software package that has gained popularity in this field is WEKA (Waikato Environment for Knowledge Analysis). Developed at the University of Waikato in New Zealand in 1993, WEKA is a comprehensive suite of machine learning algorithms and tools, written in Java, that provides a user-friendly graphical interface for data mining tasks [11].

The WEKA toolkit includes a diverse range of functions, including basic statistical analysis and visualization tools, as well as pre-processing, classification, and clustering algorithms. WEKA is different from many commercial data mining systems because it can do so much more and has a wider range of algorithms and tools. As the fields of machine learning and data mining continue to grow in importance and applicability, they have been applied to numerous domains, including science and engineering, with great success. Given its comprehensive set of tools and user-friendly interface, WEKA has been widely adopted and is well-suited for use in various data mining projects.

2.1.1. Association Rules

Weka's Apriori algorithm is a data mining method used for finding associations between items in a large dataset [8]. It is one of the most commonly used association rule mining algorithms in the field of data mining. The Apriori algorithm works by first identifying the items that occur frequently in the dataset and then using these frequent itemsets to generate association rules [9].

The algorithm operates in two phases. In the first phase, the algorithm calculates the support of each item in the dataset, which represents the number of transactions in which the item appears. The algorithm then selects the items that have support above a specified minimum support threshold and forms frequent itemsets from these items. In the second phase, the algorithm generates association rules from the frequent itemsets by calculating the confidence of each rule. Confidence represents the likelihood of the rule being true and is calculated as the support of the antecedent and the consequent of the rule divided by the support of the antecedent [10].

The Apriori algorithm is efficient for finding frequent itemsets in large datasets, as it uses a bottom-up approach, where it starts from individual items and forms larger itemsets from them. The algorithm also uses the Apriori property, which states that if an itemset is not frequent, then its subsets cannot be frequent. This property helps in reducing the search space for frequent itemsets, and thus makes the algorithm efficient. The WEKA program was utilized in this study to analyze the prepared, filtered, and preprocessed data files. The data consisted of various attributes, including date, time, electrical power produced, temperature, humidity, and wind speed, as illustrated in Figure 3. This data was deemed appropriate for executing the algorithm through the WEKA program.

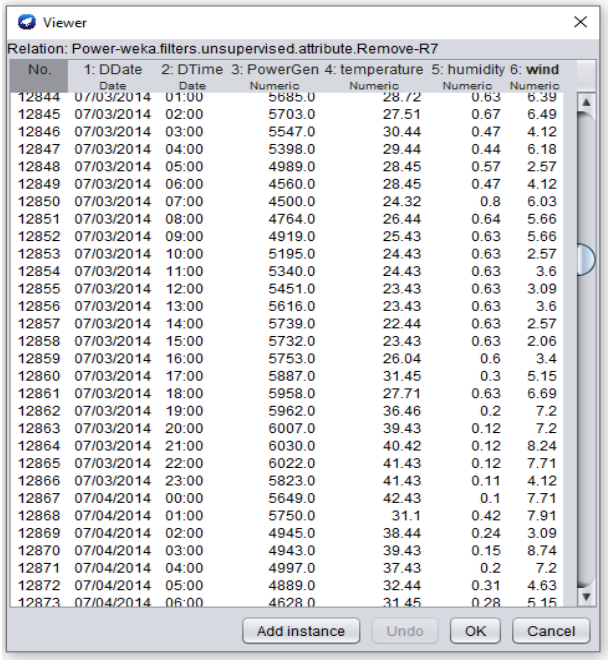

Figure 3. Data file Power

3. Experimental Analysıs

Prior to applying the Apriori algorithm to the Power.Arff dataset, the Discretize filter in Weka was utilized to convert the continuous attributes into a nominal attribute with a fixed number of bins. This was done in order to reduce the complexity of the dataset and enable the use of the Apriori algorithm, which requires nominal data. The Discretize filter effectively partitioned the range of attributes of Power.Arff values into discrete intervals and assigned a unique label to each interval, resulting in a new nominal attribute. This pre-processing step improved the overall accuracy of the Apriori algorithm in generating association rules from the Power.Arff dataset. The Apriori algorithm has been applied to the data file "Power.Arff" to uncover the relationship between energy consumption and the various factors that influence its rate. The process entailed determining the optimal minimum support and minimum confidence to produce meaningful results.

The results of the algorithm implementation were divided into three groups based on the level of electrical energy consumption, which is represented by the attribute class: low, medium, and high. This division was done to enhance the understanding and benefits of the results obtained. Table 2 presents a sampling of the findings in the form of the best-obtained high-level rules. The rules are listed in sequential order and represent the items that surpass the specified minimum support value while also belonging to a minimum confidence level.

Table 2. Best Rules for High level

No	Rules	Confidence
1	PowerGen='(5800.5-inf)' wind='(3.195-8.45]' ==> Level=High	100%
2	DTime='(59400000-70200000]' PowerGen='(5800.5-inf)' ==> Level=High	100%
3	DDate='(1659344400000-1661245200000]' ==> Level=High	100%
4	PowerGen='(5800.5-inf)' wind='(0.485-2.565]' ==> Level=High	100%
5	DDate='(1546074000000-1548925200000]' DTime='(34200000-55800000]' ==> Level=High	100%
6	DTime='(34200000-55800000]' PowerGen='(5800.5-inf)' ==> Level=High	100%
7	PowerGen='(5800.5-inf)' wind='(0.485-2.565]' ==> Level=High	100%
8	DDate='(1661590800000-1663059600000]' ==> Level=High	99%
9	PowerGen='(5713.5-5800.5]' wind='(3.195-8.45]' ==> Level=High	97%
10	DDate='(1561194000000-1567587600000]' DTime='(34200000-55800000]' ==> Level=High	96%

In the context of Association Rule Mining, Support and Confidence are two important measures that are used to evaluate the quality of the discovered rules [14].

Support represents the frequency of occurrence of a particular item or a combination of items in the data set. It is expressed as a ratio of the number of transactions containing the item(s) to the total number of transactions in the data set. The formula for Support can be described in formula 1:

$$\text{Support} = \frac{(X+Y)}{Total} \quad (1)$$

Confidence, on the other hand, is a measure of the degree of certainty that the consequent of a rule will occur given that the antecedent has already occurred. It is expressed as a ratio of the number of transactions containing both the antecedent and consequent to the number of transactions containing only the antecedent. The formula for Confidence can be described in formula 2:

$$Confidence = \frac{(X+Y)}{X} \quad (2)$$

The establishment of a partnership rule requires the determination of minimum support and minimum confidence values. These values are used to find the rules through the application of the confidence formula[14].

The results obtained for the Medium class are presented in Table 3, which lists the top 10 rules in sequential order. Analysis of the results revealed that the highest confidence values are found in the following four rules: (PowerGen = '(4000.5-4269.5]' temperature = '(16.325-18.055]'), (DTime = '(5400000-16200000]' PowerGen ='(4000.5-4269.5]'), (PowerGen = '(4000.5-4148.5]' temperature = '(18.055-21.635]'), and (PowerGen = '(4000.5-4148.5]' temperature ='(18.055-21.635]' wind = '(3.195-8.45]'), with a confidence of (100%). The lowest confidence value among the data was observed in Rule 10 (PowerGen = '(4000.5-4148.5]' humidity = '(0.305-0.515]' wind = '(3.195-8.45]') with a confidence of (96%).

Table 3. Best Rules for Medium level

No	Rules	Confidence
1	PowerGen='(4000.5-4269.5]' temperature='(16.325-18.055]'==> Level=Medium	100%
2	DTime='(5400000-16200000]' PowerGen='(4000.5-4269.5]' ==> Level=Medium	100%
3	PowerGen='(4000.5-4148.5]' temperature='(18.055-21.635]' ==> Level=Medium	100%
4	PowerGen='(4000.5-4148.5]' temperature='(18.055-21.635]' wind='(3.195-8.45]' ==> Level=Medium	100%
5	PowerGen='(4269.5-4513.5]' temperature='(18.055-21.635]' wind='(3.195-8.45]'==> Level=Medium	99%
6	PowerGen='(4148.5-4269.5]' temperature='(18.055-21.635]' ==> Level=Medium	99%
7	PowerGen='(4269.5-4513.5]' temperature='(13.515-14.015]' ==> Level=Medium	99%
8	PowerGen='(4269.5-4513.5]' temperature='(16.325-18.055]' ==> Level=Medium	99%
9	DTime='(5400000-16200000]' PowerGen='(4269.5-4513.5]'==> Level=Medium	98%
10	PowerGen='(4000.5-4148.5]' humidity='(0.305-0.515]' wind='(3.195-8.45]'==> Level=Medium	96%

The results for the class "Low" are displayed in Table 4, which lists the top 10 rules in sequential order. The minimum support and minimum confidence were chosen for optimal results. The results show that rule 1 has the highest confidence value (100%) for the conditions (DTime = '(5400000-16200000]' PowerGen = '(2505.5-3080.5]' temperature = '(16.325-21.635]'). The lowest confidence value (97%) was found for the conditions (PowerGen = '(3080.5-3366.5]' temperature = '(18.055-21.635]') in rule 10.

Table 4. Best Rules for Low level

No	Rules	Confidence
1	DTime='(5400000-16200000]' PowerGen='(2505.5-3080.5]' temperature='(16.325-21.635]' ==> Level=Low	100%
2	DTime='(5400000-16200000]' PowerGen='(2505.5-3080.5]' wind='(0.485-2.565]' ==> Level=Low	99%
3	PowerGen='(2505.5-3080.5]' temperature='(16.325-21.635]' ==> Level=Low	99%
4	DTime='(-1800000-5400000]' PowerGen='(2505.5-3080.5]' ==> Level=Low	99%
5	PowerGen='(2505.5-3080.5]' temperature='(18.055-21.635]' wind='(0.485-2.565]'==> Level=Low	99%
6	DTime='(5400000-16200000]' PowerGen='(2505.5-3080.5]' wind='(2.575-3.195]' ==> Level=Low	98%
7	PowerGen='(2505.5-3366.5]' temperature='(18.055-21.635]' wind='(3.195-8.45]' ==> Level=Low	98%
8	PowerGen='(2505.5-3080.5]' humidity='(0.305-0.515]' ==> Level=Low	97%
9	PowerGen='(2505.5-3080.5]' temperature='(21.635-23.045]' '==> Level=Low	97%
10	PowerGen='(3080.5-3366.5]' temperature='(18.055-21.635]' ==> Level=Low	97%

4. Conclusion

The study analyzed energy consumption levels using the Apriori algorithm and Weka tool and found several rules with high confidence factors. The results showed that when energy production exceeded 5800 MW and wind speed was between 3.19 and 8.45 km/hr, the rate of electricity consumption was high (100% confidence), as demonstrated in High-Level Rule 1 and High-Level Rule 2 (which also had a 100% confidence factor at the specified time of 19:00–21:00). High-Level Rule 10 had a 96% confidence factor in the time period of June 23 to September 4, during the hours of 12:00 to 17:00. For medium-level energy consumption, the study found that a power production of 4000.5–4269.5 MW and a temperature range of 16.325–18.055 °C corresponded to a 100% confidence factor (as in Medium-Level Rule 1), and that energy production between 4000.5–4269.5 MW during 4:00–6:00 am also corresponded to a 100% confidence factor (as in Medium-Level Rule 2). Medium-Level Rule 10 had a 96% confidence factor with power production between 4000.5 and 4148.5 MW, a humidity range of 0.305% to 0.515%, and a wind speed between 3.1985 and 8.45 km/hr. For low-level energy consumption, the study found that the combination of power production between 2505.5 and 3080.5 MW, a temperature ranges of 16.325-21.635 °C, and the time period of 4:00–6:00 am corresponded to a 100% confidence factor (as in Low-Level Rule 1). Low-Level Rule 3 had a 99% confidence factor with similar parameters. Low-Level Rule 10 had a 97% confidence factor with power production between 3080.5 and 3366.5 MW and a temperature range of 18.052 to 21.635 °C. The results of association rules analysis using the Apriori algorithm and Weka tool were as expected. The aim was to set new rules for each level of energy consumption based on factors such as peak times, seasonal variations, and weather conditions to successfully achieve the main objectives. In this paper, the results obtained are very interesting, useful, and significant.

REFERENCES

[1] T. Slimani and A. Lazzez, "Efficient Analysis of Pattern and Association Rule Mining Approaches," International Journal of Information Technology and Computer Science, vol. 6, no. 3, pp. 70-81, 2014.

[2] D. M. Bahssas, A. M. AlBar, and M. R. Hoque, "Enterprise resource planning (ERP) systems: design, trends and deployment," Int. Technol. Manag. Rev., vol. 5, no. 2, pp. 72–81, 2015.

[3] A. M. Abusida and Y. Gültepe, "An Association Prediction Model: GECOL as a Case Study," unpublished.

[4] W. Alsuessi, "General Electricity Company of Libya (GECOL)," Eur. Int. J. Sci. Technol., vol. 4, no. 1, pp. 1–9, 2015.

[5] A. M. Abusida and A. Hançerlioğullari, "The Power Load Prediction in GECOL using Artificial Neural Network," in press.

[6] A. M. Abusida and A. Hançerlioğullari, "A New Approach to Load Shedding Prediction in GECOL Using Deep Learning Neural Network," IJCSNS International Journal of Computer Science and Network Security, vol. 22, no. 3, pp. 220-228.

[7] M. Zhou and T. Wang, "Fault diagnosis of power transformer based on association rules gained by rough set," in Proceedings of the 2nd International Conference on Computer and Automation Engineering (ICCAE), 2010, vol. 3, pp. 123-126.

[8] P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining," 2nd edition, Pearson Education, 2006.

[9] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," in Proceedings of the 20th International Conference on Very Large Data Bases, 1994, pp. 487-499.

[10] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," 3rd edition, Morgan Kaufmann, 2011.

[11] R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald, and D. Scuse, "WEKA Manual," Version 3-6-10, University of Waikato, 2013.

[12] I. H. Witten, E. Frank, and M. A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques," 2nd edition, Elsevier, 2005.

[13] K. Mani and R. Akila, "Enhancing the Performance in Generating Association Rules using Singleton Apriori," International Journal of Information Technology and Computer Science, vol. 9, no. 1, pp. 58-64, 2017.

[14] W. Nisar, A. Khan, and F. Ahmad, "Analysis of Apriori algorithm and improvement of association rule mining," Journal of Emerging Technologies and Innovative Research, vol. 4, no. 3, pp. 100-106, 2017.

Authors' Profiles

Ashraf Mohammed Abusida A native of Libya received his Bachelor degree in Computer Science in 1996 from the Faculty of Science, University of Tripoli, and Tripoli, Libya. From 1998 to 2018, he worked as a software developer for General Electricity Company of Libya (GECOL). Until he became head of the company's developers team. In 2010, he started a Master study of Computer Science in Libyan Academy, Tripoli, Libya. In 2019, he enrolled in a PhD program in Computer Engineering, at the Department of Computer Engineering, University of Kastamonu, Turkey.

Aybaba Hançerlioğullari Graduated BSc in physics Engineering department from Hacettepe University/Turkey in 1992.Msc from Gazi University Institute of science Nuclear Research in 1996. PhD from Gazi University institute of science Nuclear Research/Turkey in 2003. He is currently working as Professor at Institute of Science Physics department, Kastamonu University..