DOI: https://doi.org/10.33103/uot.ijccce.23.3.4

Big Data Aggregation, Visualization and Clustering for Smart Grid in Smart City using Machine Learning

Hussain Jassim Fahad[1], Haider M. Al-Mashhadi[2]

[1,2]Department of Computer Information Systems, College of Computer Science and Information Technology
University of Basrah, Basrah, Iraq
[1]hussein0992@gmail.com, [2]mashhad01@gmail.com

Abstract— *Electrical energy is one of the most important components of life today where different fields depend on it. The field of electrical energy distribution (electricity network), which transmits electrical energy from sources to consumers, is one of the most important areas that need to be developed and improved. In addition to analyzing electrical energy consumption, it needs to forecast consumption and determine consumer behavior in terms of consumption and how to balance supply and demand. The research aims to analyze weather data and find the relation between the weather factors and energy consumption in order to prepare data to use as a suitable data in machine learning model for future use. This model analyzes the building consumption rate for a particular area and takes into account the weather factors that affect electrical energy consumption, where (temperature, dew point, ultraviolet index) are selects based on the correlation confidence and then divided these factors into a set of categories using the K-Means algorithm to show the effect of each factors on the other.*

Index Terms— *Big data, IoT, smart city, smart grid, smart meter, clustering, k-means clustering.*

I. INTRODUCTION

Big data has attracted attention in the world of technology, where big data represents one of the most profound and pervasive developments in the digital world. Examples of big data come from Internet of Things (IoT) devices, as well as smart cities, but also from the use of social networks, industries, etc. Data sources are numerous and ever-increasing, therefore, what distinguishes big data is not only the size but also the complexity due to the heterogeneity of the information that can be obtained [1][2]. So big data is data that exceeds the processing power of traditional database systems. The data is too large, moves too fast, or doesn't fit within the constraints of your database structures. To get a value from this data, the research must choose an alternative way to process it. The first definition of big data comes from Merv Adrian: “Big data goes beyond the reach of commonly used hardware environments and software tools to capture, manage, and process it within an acceptable elapsed time for its users.” Another good definition given by the McKinsey Global Institute is: “Big data refers to data sets whose size exceeds the ability of typical database software tools to capture, store, manage, and analyze” [3]. However, big data can be described as the volume of data grows along with the growth of information and communications technology, there is more data than technology can effectively manage, store and process [4]. Energy and how to predict it and give the optimum use of it is one of the most important applications in the world of electrical energy. The concept of a smart city has been developed is a city that uses various technological methods to enhance operational efficiency, make services, control and management more aware, interactive and effective, share information with the public, and improve service quality. A smart city manages technologies such as transportation, health, education, energy, housing, buildings and the environment in a smart way [5][6].

Information and Communication Technology (ICT) is at the heart of building smart cities (ICT). It assists countries around the world in developing, disseminating and improving sustainable

Received 11/November/2022; Accepted 05/January/2023

ISSN (Print) 1811-9212 *ISSN (Online) 2617-3352*

DOI: https://doi.org/10.33103/uot.ijccce.23.3.4

development practices in order to address the problems of increasing urbanization [7]. Governments in many major cities are adopting the concept of smart cities, and have started collecting large data sets in order to obtain valuable information from them. This information helps governments improve the standards of living and sustainability required for their residents. In order to increase the comfort and quality of life of citizens, it is necessary to reduce costs and optimize the consumption of various energy resources. This cost reduction, for example, could improve performance in areas such as education, healthcare, transportation, security, and emergency services [8]. In this regard, massive data storage using smart grid technologies is spreading [9]. For example, the energy consumption of water or electricity in public institutions is constantly monitored [10]. Where the smart grid can be defined is one of the concepts in which the Internet of things has been used, where energy is a valuable resource that the state depends on, and the research also need an energy network capable of providing the best consumption for citizens, but the problems of the network are due to increased energy consumption due to population growth or the use of equipment new, resulting in significant challenges to energy security [11]. Modernizing electricity grids has been suggested using smart grids. Power generation, transmission, distribution, and load components are all interconnected with sensors, computers, and communication networks in the smart grid. This offers methods for acquiring data, managing supply and demand, and predicting energy usage. Smart meters, for instance, can be used to collect real-time demand data at a highly accurate level from loads (such as residences and buildings) and assist system operators in forecasting future demand along spatial or temporal dimensions [12]. Where smart energy meters are considered one of the most important instant energy meters where the meter uses a two-way communication scheme, which is an improved energy of the meter that reveals the consumer's energy consumption and gives the facility additional information [13]. In situations of high demand, demand response strategies can be implemented by combining knowledge of load behavior with knowledge of generation. Another illustration is the usage of network equipment sensors, which can be used to identify issues and stop outages by enabling staff to pinpoint the precise position of malfunctioning devices. In this way, the smart grid improves the current electrical system, allowing for the reuse of existing infrastructure and lowering rollout costs. This kind of network creates a new host of issues since big data technologies are needed to handle the enormous number of data that needs to be collected from the power grid. Across industries, big data analytics are influencing how decisions are made [14].

In this paper, the research presents a district-level analysis of energy consumption calculation of energy usage data from electricity meters for each building per day, to provide customers and businesses with useful information as well as more accurate forecasts of energy production and consumption required in each part of the network. The data set of the region is collected with the weather data and looking at the relationship generated between them and giving a general perception of the amount of consumption where the weather factors (temperature, dew point, UV index) were chosen for their clear impact on the rate of consumption and the establishment of five groups of these features using K-Means Algorithm. It is necessary to divide the data into a set of clusters because in the next steps it needs the weather condition and it is better to define the weather using clusters. When conducting our analysis, our goal is to measure the rate of energy consumption and the extent of the impact of the weather on it in terms of consumption and analysis, which is necessary in order to overcome the challenges of energy networks.

II. RELATED WORK

Muhajiri et al. [15] proposed a polar projection-based approach for reducing the computational complexity of dimensionality reduction in unsupervised learning. K-mean clustering is used in this approach. To assess peak consumption in cluster consumer load profiles, a new distance scale has been proposed. It's used to categorize load profiles based on total and peak consumption. A randomization-

Received 11/November/2022; Accepted 05/January/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.3.4

based strategy has been devised to limit the search space for cluster centers in order to speed up the aggregation process. The numerical findings reveal a significant reduction in computational complexity, and the cross-sectional investigation clearly shows that the suggested polarity-dependent technique cuts complexity significantly over the classic means algorithm. Furthermore, random clustering can reduce computational complexity while keeping an acceptable estimation error.

Timothy Evely et al. [16] four distinct machine learning algorithms have been proposed to estimate consumer eligibility for Demand Tesponse (DR) programs using real energy data from users obtained over time by smart meters. Communication must be two-way (between facilities and users). Instead of off-peak power plants. Random forests, with an accuracy of 95.1 percent, have proven to be a successful strategy for evaluating large data sets using smart meter data, followed by the K-Nearest Neighbor (k-NN) rating, which has an accuracy of 75.6 percent. Artificial Neural Network outperforms decision tree by a large margin (66.5%). against 34.9%). Utility companies will have an optimal selection of customers using the proposed methodology, with an emphasis on consumers well suited to DR programs. Study results show that customer participation in a DR program can be expected at more than 90% that typically includes consumers' habits and lifestyles, and target marketing is more likely to be successful if it targets a specific group.

Ning Lu et al. [17] they analyzed data sets from residential meters that were read every 15 minutes in order to find potential value propositions for smart meter initiatives. Whereas, 50-house countermeasures were used to identify some critical data signatures for a variety of applications, such as calculating demand response capabilities, detecting trends in abnormal loads, and troubleshooting. According to the results of the study, the communication requirements for different applications, ranging from meters to data storage and management capabilities, as well as the difficulty of processing intelligence data, vary greatly. Despite the fact that the methodology can be used with different metrics, including Supervisory Control and Data Acquisition (SCADA) and Phasor Measurement Units (PMU), it has been Create the study for smart meter measurements. Our next research will focus on data interconnection between different data sources from electricity transmission and distribution networks. To reduce overburdened communication lines and to quickly identify issues, such as anomalies in network performance, it has been stated that it is critical to develop a dynamic database for data signing and to enhance the allocation of data processing power across nearby workstations and command centers.

J Jeyaranjani et al. [18] proposed a residential load prediction system and residential load estimation, which is one of the major problems of smart meter data. Due to the fluctuation of the loading pattern, this data has a highly variable load. This system is based on deep learning that uses smart scale and demographic data sets. When compared to a shallow network that performs better in error computation, a Deep Neural Network (DNN) performs better. It was discovered through an experiment that demonstrated the effect of demographic data on pregnancy prediction. The results show that deep learning has a lot of potential in the following areas: Deep learning uses multiple layers of deep learning to solve the prediction problem at scale. 2) Deep learning is capable of handling large amounts of data. DNN has been found to be useful in other carry issues, including demand response, pricing forecasting, and error determination.

According to how similar the users' typical electricity usage patterns are to each other, Zigui Jiang et.al. [19] proposed a hybrid machine learning approach that includes unsupervised assembly and supervised classification. A flexible demand management strategy and efficient energy control may be made possible through the three phases of the methodology, which include classification of electricity consumers, characterization of their consumption, and classification of new consumers. A non-obvious consumer classification is made after common electricity usage patterns have been captured using an unsupervised aggregation technique. Several consumer classes and their consumption characteristics are then proposed to be identified using a new method. New customers are scored using a supervised

Received 11/November/2022; Accepted 05/January/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.3.4

rating algorithm, which also assesses the accuracy of the selected categories. The proposed model is checked using real data. The model was tested on an actual data set of non-US consumers obtained over a year by smart meters. The results show that large or private companies usually have separate consumption characteristics.

Wilson Rivera et al. [20] explored big data technologies and difficulties in the context of smart grids, which have been recommended as a way to update energy infrastructure. The electricity grid is integrated with sensors, computers, and communication networks in a smart grid. Big data analysis techniques are necessary to enable data analysis and decision making due to the rise in sensor density and the capacity to gather vast volumes of data. It displays a real-world scenario using numerous data sources. Through a dashboard created with Python Flask and D3, information about the weather, Twitter activity, and energy usage is interactively shown. The Apache Spark framework is used for data analysis.

By Alexander et al. [21] show ways to improve clustering by harvesting inherent structure from the smart meter data. They collected local electricity consumption using smart meter data from the Danish city of Esbjerg. Methods from time-series and wave analysis have been applied to enable the K-Means clustering method to calculate autocorrelation in the data and thus improve clustering performance. The results showed that there is an autocorrelation in specific electricity data of the smart meter. It is not general evidence of autocorrelation in all smart meter datasets, but it is an indication that smart meter data should be examined for autocorrelation before starting the analysis to identify consumption subgroups within housing types. K-Means were able to produce these aggregates by taking into account time. This method succeeded in extracting large autocorrelation coefficients and merging them into subsequent groups using K-Means. The Wavelet transformation of the input data to K-Means succeeded in compressing the data and removing the multiple linear relationships, but was unsuccessful in determining the optimal number of groups. The results show that intelligent data transformation before compiling K-Means can improve performance and enable K-Means to handle data and information of types they were not originally intended for. This result enables the production of sets of smart meter data that are better defined by smaller groups with less intragroup variance.

III. DATASET COLLECTION AND DESCRIPTION

A smart grid system uses many components to collect data, and a smart meter is the ultimate device that devices use to collect data from homes, buildings, and any other places where the smart meter is built. A large amount of smart meter data must be provided for analysis, but the research faced with the problem that there is not enough data set collected from smart meters, so electrical energy consumption data for one of the projects in London was adopted. The demand for energy increases in the future, so the research only take the total energy consumed per day in a particular house. The data, which are readings for the energy consumption of a sample of 5,567 London households who participated in the London Low Carbon Electricity Networks project dataset that this research adopt, for the period between November 2011 and February 2014, were collected and contain different readings available for each half. Read the average daily consumption by the hour and by the hour. However, this data alone cannot be relied upon. Other features should be added to help us analyze and give the best results. Weather data has been added, and daily weather information has been taken using the Dark Sky API in the dataset which has a clear energy impact and, in addition to being date-based, is a platform for data analysis. So, the data on which the research is based is the meter data as well as the weather [22].

IV. THE PROPOSED SYSTEM

The system is based on a set of sequential and interdependent steps that depend on each other. The actual instantaneous consumption of electrical energy is read and then transmitted from the reading

Received 11/November/2022; Accepted 05/January/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.3.4

devices (electricity meters) via the Internet to the storage location. After the data is stored, the real-time weather is read from this data and combined with it. After that, the processing of this data, which includes (organizing, analyzing and extracting the influencing characteristic), is then done to extract the results. The results are discussed based on the steps of the pretreatment stage. Theoretical basis of the system steps mentioned and shown in *Fig. 1*, below:

A. Data Preprocessing

Preprocessing techniques: The process of converting raw data into an understandable format is known as data preprocessing. Data preprocessing is one of the most common data mining activities that involves cleaning and transforming data into a form that is acceptable for mining operations. Reducing the quantity of the data, identifying relationships between it, normalizing it, removing outliers, and extracting features from it are the goals of data preparation.

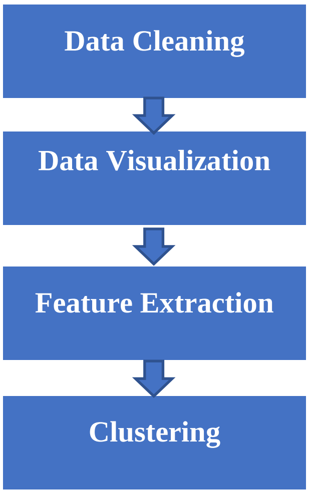

FIG. 1. DATA PREPROCESSING.

It includes several techniques like data cleaning, integration, transformation and reduction [23], so this is an essential stage before using machine learning or data mining methods, to ensure that the data is good quality and is formatted in a way that is suitable for the required work.

i. Data Cleaning

Data with the aim of removing duplicate, damaged, inaccurate, badly structured, inaccurate, or incomplete data from a data set. When combining multiple data sources, it is possible for data to be duplicated or given inaccurate labels in a number of different ways. There is no exact way to categorize the specific processes in the data cleaning process because the procedures will vary from dataset to dataset. But it's crucial to model your data cleaning process so you know you're doing it the right way every time [24]. While the methods used to clean data may vary according to the types of data you store, there are a number of steps the research follow which are removing redundant or irrelevant observations, fixing structural errors, filtering out unwanted outliers, and handling missing data, Validation and quality assurance.

ii. Data Visualization

Visualization, which turns abstract data into concrete insights (such as length, location, shape, colors, etc.), is essential in today's data-driven corporate environment. It has been extensively utilized to support the making of decisions that are closely related. Naturally, data visualization is a wonderful fit for providing a thorough overview of huge data, and facilitating interpretation of data analysis results for data scientists using visual elements such as charts, graphs, and maps [25]. The type of data visualization to use is part of the data visualization strategy. Data visualization

Received 11/November/2022; Accepted 05/January/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.3.4

can take many forms. Scatter charts, line graphs, pie charts, bar charts, heat maps, area charts, corrective maps, and graphs are the most popular as shown in *Fig. 2*.

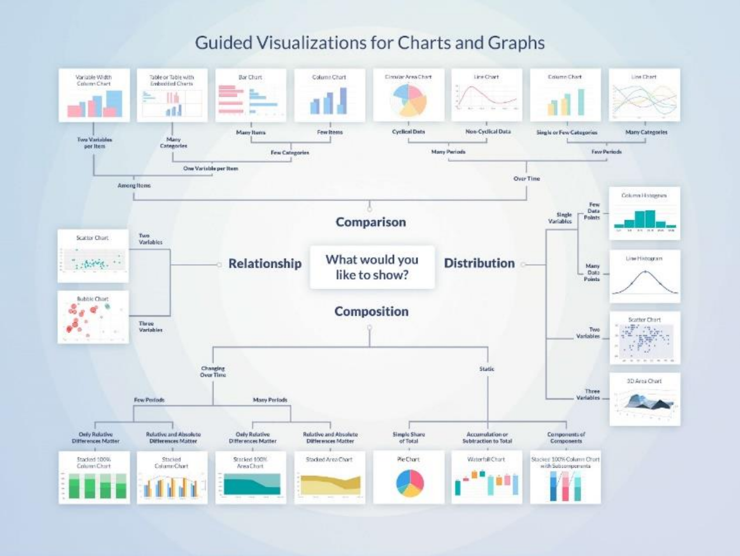

FIG. 2. VISUALIZATION GRAPHS TYPE.

Time-series graphics were used because the energy consumption data is time-related. A time series is a set of data points that are collected with respect to their time. This technique is used to diagnose past behavior and predict future behavior, using a wide range of techniques, including statistical methods. It is based on assumptions; therefore, successive values are observed at regular intervals [26]. Time series is a set of observations, each of which is recorded at a specific time that describes a very wide range of phenomena. There are two main objectives of time series analysis: determining the nature of the series and predicting future values from current and past data. Both of the above objectives require a sequential model to be formally defined and described [27].

iii. Feature Selection

When the data set received, many features are often revealed in a dataset. Not every feature that discover in the data set will be helpful in creating a machine learning model for the required prediction job. Utilizing any functions might worsen the situation. As a result, choosing the right features is crucial when creating a machine learning model. There is always a desire to have a model with the best predictive power that can be explained. However, in many cases there is a trade-off between interpretability and predictive performance and it is generally difficult to maximize both simultaneously. Feature selection increases the model's ability to interpret and at the same time simplifies the model. The complexity of the model is reduced because only most of the relevant features are included that make the model easy to interpret and permit the construction of practical models for the phenomena under study [28]. The following categories essentially describe how feature selection methods in machine learning are used [29]:

- Supervised methods: These techniques can be used to categorical data and are used to find pertinent characteristics that improve the performance of supervised models in data with labels.
- Unsupervised Technologies: These technologies can be used for unlabeled data.

For classification approaches, most feature selection algorithms have been put out for consideration. The bulk of feature selection algorithms employ statistical metrics such

Received 11/November/2022; Accepted 05/January/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.3.4

mutual information, correlation, and information gain measure [30]. Based on assessment metrics, there are three general strategies for feature selection Filter, Wrapper, Embedded methods.

a. Correlation Coefficient

Formulas for the correlation coefficient are used to calculate the degree of relationship between two features in the sets of data *Fig. 3*. The computations get a result that ranges from -1 to +1 [31] where:

- A score of +1 denotes a highly favorable association.
- A score of -1 denotes a seriously adverse association.
- A 0-result means there is absolutely no relationship

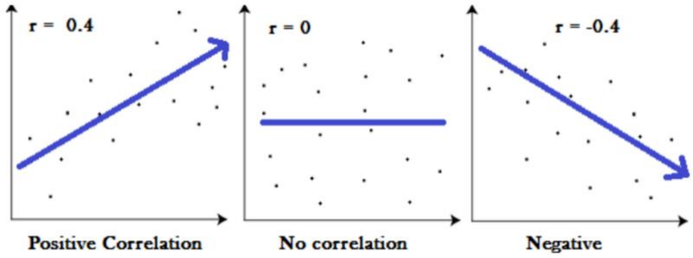

FIG. 3. CORRELATION COEFFICIENT RELATIONSHIPS.

The absolute value of the correlation coefficient appears to indicate the strength of the association. (1) refere to the correlation coefficient for two features [32].

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

r= correlation coefficient

x_i = values of x- variable in a sample

$\bar{x}$ = mean of the values of x- variable

y_i= values of the y-variable in a sample

y = mean of the values of the y-variable

b. Correlation matrix

Correlation matrix is a table with the correlation coefficients for mutable variables. A matrix in a table represents the correlation between all feasible pairings of values. It is considered a helpful tool for fast condensing a big dataset and finding and displaying data trends. In a correlation matrix, these variables are represented by both columns and rows. Numerous types of statistical analysis are typically used with the correlation matrix. It might be useful for examining several linear regression models, for instance. the models contain a lot of independent variables. The correlation matrix in multivariate linear regression establishes the correlation coefficients between the independent variables in a model. The correlation matrix indicates to the symmetric array of numbers as shown in the below matrix in which each cells represents the correlation coefficient between the intersected features and each element in matrix calculated by Eq (2):

Received 11/November/2022; Accepted 05/January/2023

$$\mathbf{R}=\begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ r_{31} & r_{32} & 1 & \cdots & r_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{pmatrix} \qquad (2)$$

iv. Data clustering

The goal of data clustering, also known as cluster analysis, is to discover the natural set(s) of a set of patterns, points, or objects It can be defined as the task of identifying sub-clusters in the data so that data points in the same sub-set (cluster) are very similar while points the data is in completely different sets [33]. Cluster analysis is "a statistical categorization approach for determining whether people from a population fall into various groups by making quantitative comparisons of several features," according to Webster (Merriam-Webster Online Dictionary, 2008) [34]. Any discipline that requires the study of multivariate data frequently uses cluster analysis. The research found a reason for the large number of variables used. The variables must be grouped, as the K-Means algorithm was used to group blocks into groups [35]. There are various types of clustering methods that solve one or more of these problems and of course many statistical clustering algorithms and machine learning algorithms that implement the methodology, in this paper K-Means algorithm is chosen.

a. K-Means

One unsupervised learning approach that addresses the well-known clustering issue is k-means. The process uses a fixed number of clusters (let's assume k clusters) to classify a given data set in an uncomplicated and straightforward manner. To define k centers, one for each cluster, is the main notion. These centers should be strategically positioned because different locations yield various effects. The preferable option is to situate them as far apart from one another as you can. The following phase is connecting each point from a given data set to the closest center. The first step is ended and an early group age is finished when there are no points still open. As the barycenter of the clusters formed in the previous phase, the new centroids must be recalculated k. The same data set points must now be bound to the closest new center once have these k new centroids. There has been created a loop. This loop may cause the k centers to gradually shift positions until no more modifications are made, or, to put it another way, the centers stop moving altogether [36].

The pseudocode of K-means algorithm [37]**:**

Enter: the datasets x1,..., xP, the centroids' initializations c1,..., cK, and the maximum number of iterations J.

Algorithm 1: pseudocode of K-means algorithm

```
for j=1,…,J
#Update cluster assignments
        for  p=1,…,P
            ap=argmink=1,…,K‖$c_k$−$x_p$‖2
        end for
#Update centroid location
        for  k=1,…,K
            denote  $S_k$  the index set of points  $X_p$  currently assigned to the  kth  cluster
            update  $c_k$  via  $c_k$=1|Sk|$\sum_{p\in Sk}$ $X_p$
        end for
end for
#Update cluster assignments using final centroids
 for p=1,…,P
        $a_p$=argmink=1,…,K‖$c_k$−$x_p$$\|_2$
end for
output: optimal centroids and assignments
```

Received 11/November/2022; Accepted 05/January/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.3.4

V. IMPLEMENTATION AND RESULTS

A. Cleaning

Data cleaning includes removing or replacing erroneous data i.e., removing blanks, redundant data and noise, in addition to filling in the empty fields using the average daily rate of consumption. Then, calculated the daily average building electricity consumption by summation the buildings consumption and dividing it by buildings number. *Fig. 4* shows the meters per day from which the data was collected, the x-coordinate shows the days, while the y-coordinate shows the number of meters on that day.

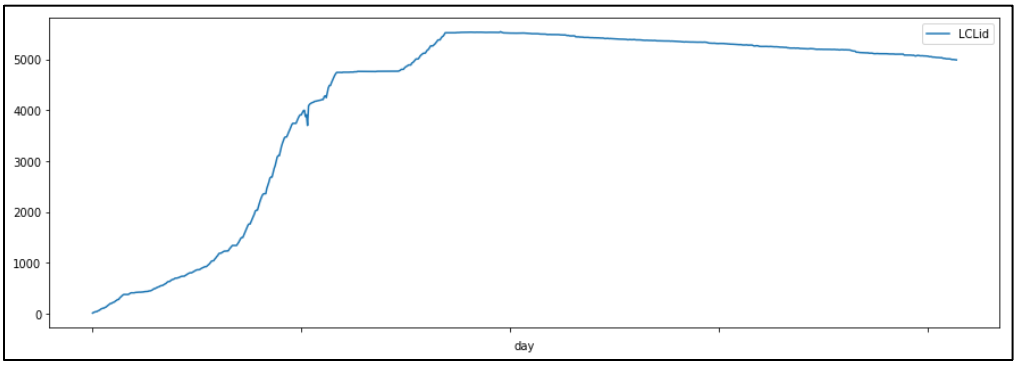

FIG. 4. COUNT OF METERS PER DAY.

B. Visualization

Data visualization, which transforms abstract data into physical insights, is naturally suitable for giving a good overview of data, and facilitating interpretation of data analytics results using visual elements such as charts, graphs, and maps. Visualization helps to finding the relationship between weather conditions and consumption energy by visual way for same period of time. The relationship of weather factors with electricity consumption is through the effect of the factor on the percentage of consumption, and will explain each factor and the extent of its impact as follows:

i. Electricity Consumption VS Temperature

Temperature one of the weather factors that have a significant impact on the energy consumed as shown in the *Fig*. 5 below.

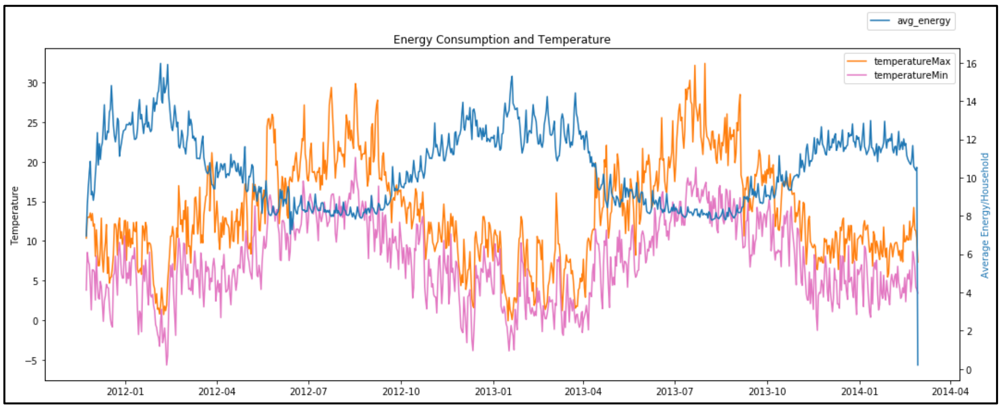

FIG. 5. RELATIONSHIP BETWEEN ENERGY AND TEMPERATURE.

The relationship is inverse between energy and temperature through what see from the height of the blue peaks, which shows the consumption of electricity with a cavity in the other orange

Received 11/November/2022; Accepted 05/January/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.3.4

(TemperatureMax) and pink (TemperatureMin). This confirms the business intuition that during low temperatures, the energy consumption through the use of heaters is likely to increase due to the often-cold weather.

ii. Humidity

Another weather factors that do not have a significant impact on the energy consumed. humidity and average energy consumption have the same direction, meaning there is no clear relationship between them. As the energy consumption in a humid atmosphere does not make a difference if the humidity is high or low through what we see from the colors that show the same direction. as shown in *Fig. 6* below:

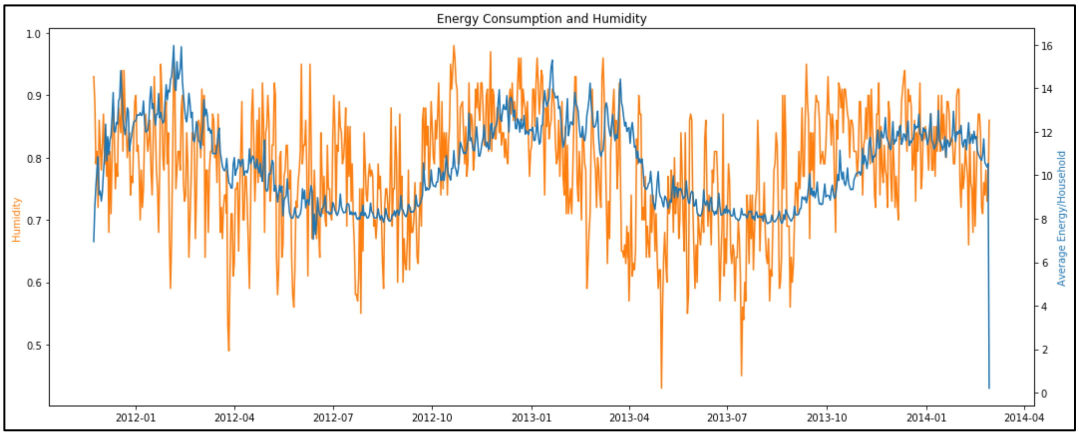

FIG. 6. RELATIONSHIP BETWEEN ENERGY AND HUMIDITY.

iii. Cloud Cover

One of the weather factors that do not have a significant impact on the energy consumed, as shown in *Fig. 7* below:

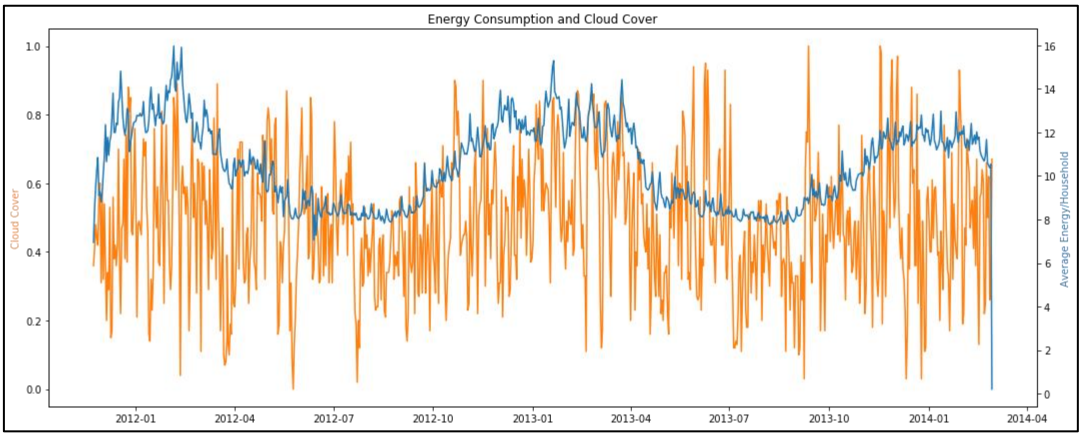

FIG.7. RELATIONSHIP BETWEEN ENERGY AND CLOUD COVER.

As in the case of humidity, it appears that the value of cloud cover (clouds) follows the same pattern of energy consumption, meaning that rarefactions in it do not have an effect on the level of energy consumption.

iv. Visibility

Visibility is a weather factor that does not have a significant impact on the energy consumed, as shown in *Fig. 8* below.

Received 11/November/2022; Accepted 05/January/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.3.4

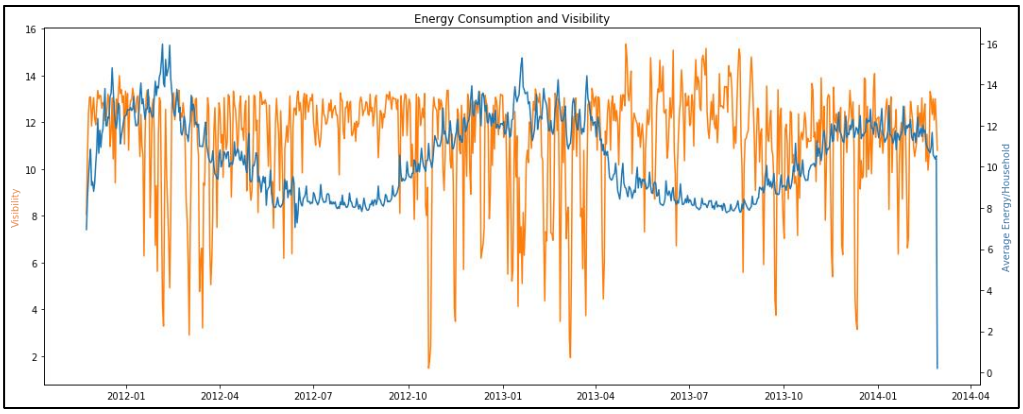

FIG. 8. RELATIONSHIP BETWEEN ENERGY AND VISIBILITY.

Take note that the visibility element has no impact whatsoever on energy usage through what that seen from the inverse proportion of the rise and fall of the blue color with the other two colors, so the increase or decrease does not affect energy consumption.

v. Wind speed

It is similar to the factor of wind speed and visibility, as both do not affect energy consumption, as it is also an external factor as shown in *Fig. 9* below.

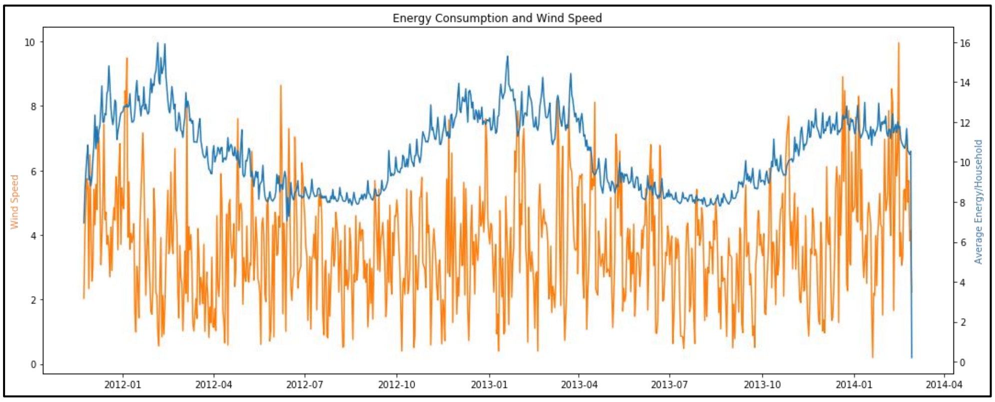

FIG. 9. RELATIONSHIP BETWEEN ENERGY AND WIND SPEED.

vi. Ultraviolet (UV) Index

Energy use and the UV index are inversely correlated, as noted the clear effect of the sun's ultraviolet rays on energy consumption, that is, the higher the rays, the lower the energy consumption and vice versa as shown in the *Fig. 10* below:

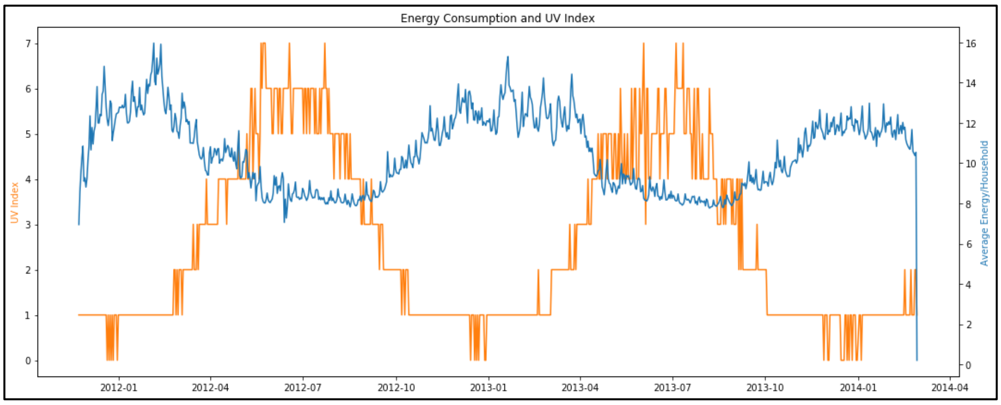

FIG. 10. RELATIONSHIP BETWEEN ENERGY AND UV INDEX.

Received 11/November/2022; Accepted 05/January/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.3.4

vii. Dewpoint

Dew point: It depends on temperature and relative humidity. The graph below shows that there is a strong negative correlation between energy consumption and dew point and that this relationship affects energy consumption significantly, as shown in *Fig. 11*.

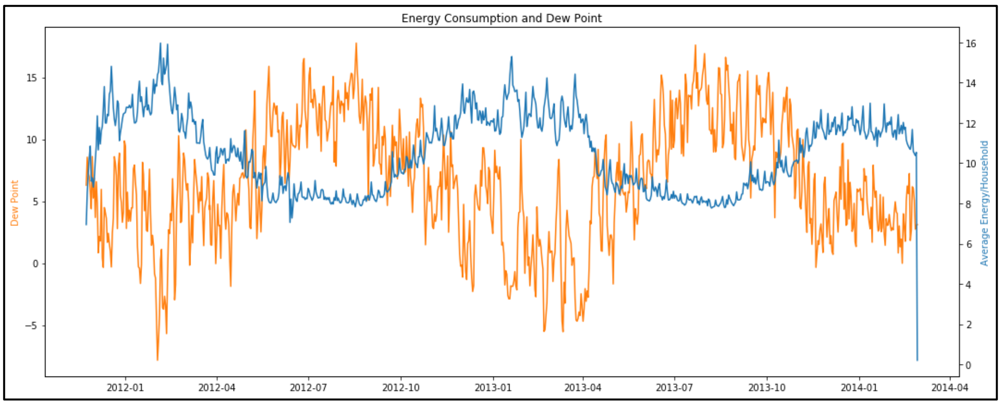

FIG. 11. RELATIONSHIP BETWEEN ENERGY AND DEW POINT.

From the visual perception of atmospheric factors initially, conclude that their relationship to energy consumption is as follows:

- Energy consumption has a high negative (inverse) relationship with temperature, UV index and Dew point.
- Wind speed, Visibility, Cloud Cover and Humidity have low relationship with Energy consumption therefore eliminated

C. Feature Selection

After the data visualization step, a large vision was formed of the weather features affecting electrical consumption, to choose the most important features that have the highest impact ratios, a correlation matrix was used. Correlation matrix implements between weather features and electrical consumption the result is showing in *Fig. 12* below:

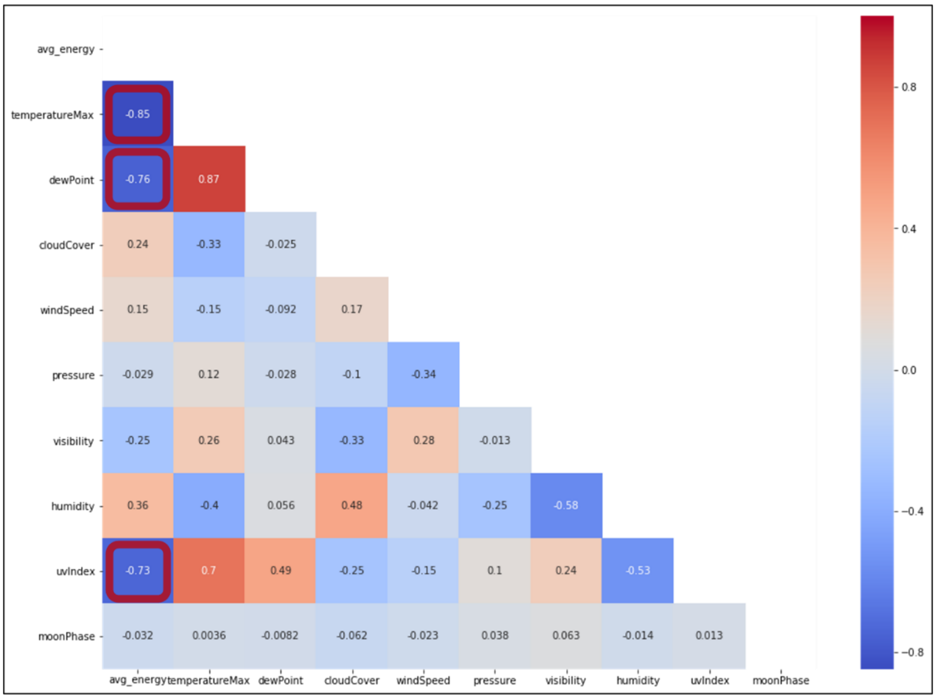

FIG. 12. CORRELATION MATRIX PRESENTATION.

Received 11/November/2022; Accepted 05/January/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.3.4

The correlation coefficients for the various variables are displayed in the correlation matrix figure. The matrix that incorporates the meteorological variables (energy consumption rate, maximum temperature, dew point, cloud cover, wind speed, pressure, humidity, UV index, visibility, moon phase) with rows and columns in a matrix represents the relationship between all possible pairings of values in the figure, the values of the cells within the matrix are between (+1, -1) and this value represents the strength of the relationship between the pairs of intersecting factors. The absolute value represents the amount of the relationship, while the sign represents the type of relationship (negative sign represents inverse relationship, positive sign represents positive relationship). On this basis, the features that have absolute values greater than 0.6 are selected, i.e. (temperature, Dew point, UV index) are selected.

D. Weather data clustering

After select strong weather features, make five clasters of this feature using K-Means. it is necessary to divide the data into a group of clusters because in the following steps needs the weather condition and it is better to determine the weather condition using clusters, also in the following steps some artificial intelligence techniques are applied, it is better to use summerized data when needs AI work quickly.The following *Fig. 13* shows relations between weather factures (temperature, Dew point, UV index) that selected in previous step and the result weather cluster.

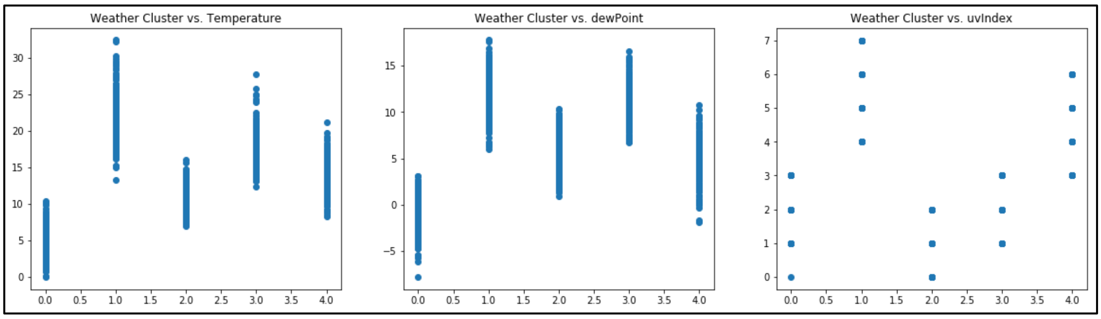

FIG. 13. WEATHER CLUSTER.

Fig.14 represents another way to visualize the relation between weather clusters and selected feature, each dote in figure refire to a day and there is four parameters that a dote represents (X-Access refire to temperature, Y-Access refire to Dewpoint, dote size refire to UV-Index, dote color refire to the class).

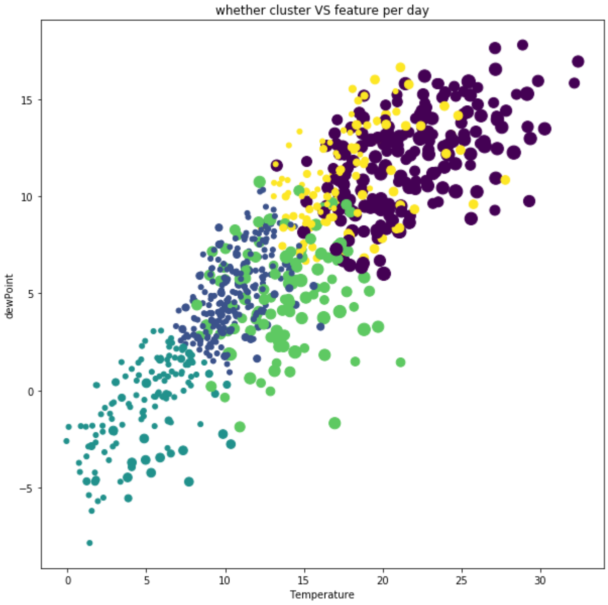

FIG. 14. ELECTRICITY CONSUMPTION AND WEATHER DATA WITH CLUSTERING PER DAY.

Received 11/November/2022; Accepted 05/January/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.3.4

VI. CONCLUSIONS

This research presents a method to collect, analysis and visualize the relation between weather factors (temperature, dew point, UV-Index) that selects based on the correlation confidence and energy consumption values that can collects by electrical meter based on the Internet of Things. The research focus on collects the data of energy consumption for buildings in an entire area and calculate the total consumption rate for this area. After this step the data are inputs to preprocessing stage that consists of multi steps (cleaning in which removing duplicate, damaged, inaccurate, badly structured, inaccurate, or incomplete data from a data set, visualize in which turns abstract data into concrete insights, feature extraction, select the most important features to process the data, clustering in which dived the data into clusters depends on these features using k-means model). The relationship between the selected weather factors that significantly affected energy consumption and from the results of the implementation of K-Means on weather factors, the results show that the data are grouped into five groups, each group representing a close relationship between (temperature, dew point, UV index). The result of weather clusters and electrical consumption per day will be using in the next research to predict the future electrical consumption per day for a region.

REFERENCES

[1] F. Arena, G. Pau,” An overview of big data analysis,” Bulletin of Electrical Engineering and Informatics, vol. 9, no. 4, pp. 1646-1653, ISSN: 2302-9285, August 2020; doi: 10.11591/eei.v9i4.2359.

[2] I. Mohsin ", H. Mohammed", K. Radhi ", H. Mohammed,” IoT Based Multitasking Games And Entertainment Arcade Station Using Raspberry-Pi,” Journal of Southwest Jiaotong University, vol. 54, no. 3, ISSN: 0258-2724, June 2019; doi, 10.35741/issn.0258-2724.543A.

[3] A. Trnka,” Big Data Analysis,” European Journal of Science and Theology, vol.10, Suppl.1, 143-148 October 2014.

[4] S. Kaisler, F. Amnour, J. Alberto, “Big Data: Issues And Challenges Moving Forward,”46th IEEE international conference on system science, Wailea, Maui, HI, USA, 7-10 Jan. 2013;doi:10.1109/HICSS.2013.645.

[5] NA. Jasim , Haider TH. Sal. ALRikabi, “Design and Implementation of Smart City Applications Based on the Internet of Things,” International Journal of Interactive Mobile Technologies (iJIM), vol. 15, no. 13, 2021; doi: https://doi.org/10.3991/ijim.v15i13.22331.

[6] H. Mohammed, K. Radhi,” Design And Implementation of A Smart Integrated Framework to Monitor And Control The Smart City Using the Internet of Things,” Journal of Southwest Jiaotong University, vol. 54, no. 6, ISSN: 0258-2724, December 2019; doi : 10.35741/issn.0258-2724.54.6.61.

[7] A. Ojo, E. Curry, T. Janowski, and Z. Dzhusupova, "Designing next generation smart city initiatives: The SCID framework," in Transforming city governments for successful smart cities: Springer, pp. 43-67, 2015; doi: https://doi.org/10.1007/978-3-319-03167-5_4.

[8] E. Al Nuaimi, H. Al Neyadi, N. Mohamed, J. Al-Jaroodi, “Applications of big data to smart cities,” Journal of Internet Services and Applications, Springer ,vol.6, 2015; doi:10.1186/s13174-015-0041-5.

[9] VC. Gungor, D. Sahin, T. Kocak, S. Ergüt, C. Buccella, C. Cecati, Gerhard P. Hancke,” Smart Grid Technologies: Communication Technologies and Standards,” IEEE Transactions on Industrial Informatics .vol.7, issue 4,2011; doi: 10.1109/TII.2011.2166794.

[10] R. Pérez-Chacón, JM. Luna-Romera , A. Troncoso,F. Martínez-Álvarez and José C. Riquelme,"Big Data Analytics for Discovering Electricity Consumption Patterns in Smart Cities," Energies, 2018, vol.11, issue 3, pp. 683; doi: 10.3390/en11030683.

[11] P. Pritzker, W. May,“NIST Framework and Roadmap for Smart Grid Interoperability Standards, Release 3.0,” National Institute of Standards and Technology Special Publication 1108r3,2014 ;doi:http://dx.doi.org/10.6028/NIST.SP.1108r3.

[12] A. Sanchez and W. Rivera,"Alejandro Sanchez and Wilson Rivera,"IEEE 6th International Congress on Big Data, Honolulu, HI, USA, 2017; doi: 10.1109/BigDataCongress.2017.59.

[13] S. Depuru, W. Lingfeng, V. Devabhaktuni and N. Gudi ,"Smart meters for power grid Challenges, issues, advantages and status," in Proc. Power Systems Conjerence and Exposition (PSCE), pp. i-7, Phoenix, AZ, USA, 20-23 March 2011; doi: 10.1109/PSCE.2011.5772451.

[14] A.S. Chhabra, T. Choudhury, A.V. Srivastava, A. Aggarwal,“Prediction for big data and IoT in 2017,” International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), Dubai, pp. 181-187 ,2017;doi: 10.1109/ICTUS.2017.8286001.

Received 11/November/2022; Accepted 05/January/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.3.4

[15] M. Mohajeri, A. Ghassemi, and T. A. Gulliver, "Fast Big Data Analytics for Smart Meter Data," IEEE Open Journal of the Communications Society, vol.1, 2020; doi: 10.1109/OJCOMS.2020.3038590.

[16] M. Martinez-Pabon, T. Eveleigh and B. Tanju,"Smart meter data analytics for optimal customer selection in demand response programs," 3rd International Conference on Energy and Environment Research, ICEER 2016, 7-11 September 2016, Barcelona, Spain, vol.107, pp 49 – 59,2017; doi:https://doi.org/10.1016/j.egypro.2016.12.128.

[17] N. Lu, P. Du, X. Guo, and FL. Greitzer," Smart Meter Data Analysis," PES T&D, Orlando, FL, USA,2012;doi: 10.1109/TDC.2012.6281612.

[18] J. Jeyaranjani and D. Devaraj,"Deep Learning based Smart Meter Data Analytics for Electricity Load Prediction," IEEE International Conference on Clean Energy and Energy Efficient Electronics Circuit for Sustainable Development (INCCES), 2019; doi: 10.1109/INCCES47820.2019.9167704.

[19] Z. Jiang, R. Lin and F. Yang," A Hybrid Machine Learning Model for Electricity Consumer Categorization Using Smart Meter Data," Energies, vol.11, issue 9, 2018; doi: 10.3390/en11092235.

[20] A. Sanchez and W. Rivera," Big Data Analysis and Visualization for the Smart Grid, "IEEE 6th International Congress on Big Data,Honolulu, HI, USA2017; doi: 10.1109/BigDataCongress.2017.59.

[21] A. Tureczek, P. Sieverts Nielsen and H. Madsen," Electricity Consumption Clustering Using Smart Meter Data, "Energies, vol.1, issue 4, 2018; doi:https://doi.org/10.3390/en11040859.

[22] DB. Avancini, J. J.P.C. Rodrigues, S.G.B. Martins, R. A.L. Rabelo, J. Al-Muhtadi, P. Solic," Energy meters evolution in smart grids: A review," Journal of Cleaner Production, vol.217, pp. 702-715, 2019; doi:https://doi.org/10.1016/j.jclepro.2019.01.229.

[23] M. Lorena Tuballa, M. Lochinvar Abundo," A review of the development of Smart Grid technologies," Renewable and Sustainable Energy Reviews Science Direct, vol. 59, pp. 710-725, June 2016; doi: http://dx.doi.org/10.1016/j.rser.2016.01.01.

[24] S. A. Alasadi and W. S. Bhaya," Review of Data Preprocessing Techniques in Data Mining," Journal of Engineering and Applied Sciences ,vol.12 ,issue 16,pp. 4102-4107, 2017.

[25] H. Liu, Ashwin Kumar TK, J. P Thomas, X. Hou," Cleaning Framework for Big Data," IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), Oxford, UK,2016;doi: 10.1109/BigDataService.2016.41.

[26] X. Qin, Y. Luo, N. Tang, G. Li, "Making data visualization more efficient and effective: a survey," Springer-Verlag GmbH Germany, part of Springer Nature, vol.29, 2019; doi: https://doi.org/10.1007/s00778-019-00588-3.

[27] B. Vishwas and A. Patel," Hands-on time series analysis with Python. Apress," Springer, 2020. https://link.springer.com/book/10.1007/978-1-4842-5992-4.

[28] R. Krispin," Hands-On time series analysis with R. Birmingham: Packt," Packet, 2019.

[29] M. Kuhn and K. Johnson," Feature engineering and selection: A practical approach for predictive models," CRC Press Tajlor & Francis Group, 2019.

[30] S. Alelyani, J. Tang and H. Liu," Data Clustering book, chapter Feature Selection for Clustering: A Review," Taylor & Francis Group, 2018.

[31] D. Aksu1, S. Üstebay, M. Ali Aydin and T. Atmaca,"Intrusion Detection with Comparative Analysis of Supervised Learning Techniques and Fisher Score Feature Selection Algorithm," Springer Nature Switzerland, pp. 141–149, 2018; doi: 10.1007/978-3-030-00840-6_16.

[32] U. M. Khaire, R. Dhanalakshmi," Stability of feature selection algorithm: A review," Journal of King Saud University -Computer and Information Sciences 65. vol. 34, Issue 4, pp. 1060-10732019, April 2022; doi:https://doi.org/10.1016/j.jksuci.2019.06.012;

[33] A. G. Asuero, A. Sayago, and A. G. Gonz´alez," The Correlation Coefficient: An Overview," Critical Reviews in Analytical Chemistry, vol.36, issue 1, pp. 41–59, 2006; doi: 10.1080/10408340500526766.

[34] R. Ahmadi, G. Ekbatanifard and P. Bayat," A Modified Grey Wolf Optimizer Based Data Clustering Algorithm," Applied Artificial Intelligence, vol. 35,issue 1, pp. 63-79,2021; doi: 10.1080/08839514.2020.1842109.

[35] J. Anil K., "Data clustering: 50 years beyond K-means," In Pattern Recognition Letters, vol.31, Issue 8,p.p 651-666, ISSN 0167-8655, 2010; doi: https://doi.org/10.1016/j.patrec.2009.09.011.

[36] M. Sarstedt, E. Mooi," Cluster Analysis. In: A Concise Guide to Market Research. Springer Texts in Business and Economics," Springer, Berlin, Heidelberg, 2014.

[37] J. Pérez-Ortega, N. N. Almanza-Ortega, A. Vega-Villalobos, R. Pazos-Rangel, C. Zavala-Díaz and A. Martínez-Rebollar, Clustering book, "chapter The K-Means Algorithm Evolution" Introduction to Data Science and Machine Learning,2019; doi: 10.5772/intechopen.85447.

Received 11/November/2022; Accepted 05/January/2023