الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي

Université Badji Mokhtar - Annaba
Badji Mokhtar – Annaba University

جامعة باجي مختار – عنابـــة

Faculty : Technologie

Department : Informatique

Field : Mathématique-Informatique

Sector : Informatique

Specialization : systèmes informatiques

Thesis

Presented for the purpose of obtaining the Bachelor's Degree

Theme

Initiation to the creation of a system to help forecast the demand for electrical load

Presented by: Labar Mohamed Yaniss

Supervisor : Farah Nadir Grade : Professeur

University**: Badji Mokhtar Annaba**

College year: 2022/2023

Table of contents:

Lists of figures:

Lists of tables:

Greetings:

I would like to express my heartfelt gratitude to my research supervisor, Professor Farah Nadir from Badji-Mokhtar Annaba University, for the assistance he provided me with, his patience, and his encouragement. I would like to thank all the members of the jury for their interest in my work.

I would also like to thank the entire staff of the computer science department, the teachers, the administrative staff, and all my colleagues without exception.

Lastly, a special thank you goes to my parents from the bottom of my heart. They have been an invaluable source of strength and support. I would also like to extend my final words of gratitude to my family, who have always helped me achieve my goals.

Abstract:

This project provides a thorough exploration of K-means clustering, linear regression, and artificial neural networks and their applications in electricity consumption forecasting. It demonstrates the practical implementation of these techniques, presents the obtained results, and conducts a comparative analysis to assess their accuracy in forecasting. By offering a comprehensive overview and evaluating the effectiveness of each method, this project sheds light on the potential of K-means clustering, linear regression, and artificial neural networks for electricity consumption forecasting.

Résumé:

Ce projet propose une exploration approfondie du clustering K-means, de la régression linéaire et des réseaux de neurones artificiels et de leurs applications dans la prévision de la consommation d'électricité. Il démontre la mise en œuvre pratique de ces techniques, présente les résultats obtenus et procède à une analyse comparative pour évaluer leur exactitude dans les prévisions. En offrant un aperçu complet et en évaluant l'efficacité de chaque méthode, ce projet met en lumière le potentiel du clustering K-means, de la régression linéaire et des réseaux de neurones artificiels pour la prévision de la consommation d'électricité.

ملخص:

يوفر هذا المشروع استكشافًا شاملاً لتجمعات mean-K ، والانحدار الخطي، والشبكات العصبية الاصطناعية وتطبيقاتها في التنبؤ باستهلاك الكهرباء.

يوضح التطبيق العملي لهذه التقنيات، ويعرض النتائج التي تم الحصول عليها، ويقوم بإجراء تحليل مقارن لتقييم دقتها في التنبؤ.

من خلال تقديم نظرة عامة شاملة وتقييم فعالية كل طريقة، يسلط هذا المشروع الضوء على إمكانات mean clustering-K ، والانحدار الخطي، والشبكات العصبية الاصطناعية للتنبؤ باستهلاك الكهرباء

1-Introduction:

Forecasting time series poses a notable challenge in various domains, such as finance where predicting stock exchange rates or market indices is essential, and in data processing where experts forecast information flow on networks. The global surge in human population, the desire for improved living standards, industrialization in developing nations, and the necessity for positive economic growth rates have led to a rapid escalation in energy consumption. Accurate forecasting plays a crucial role in investment planning for energy production, generation, and distribution. Developing reliable forecasts can be challenging due to the difficulty in determining the necessary and sufficient information for accurate predictions. Inadequate or redundant information can result in flawed modeling or skewed predictions. While complex models may yield accurate predictions, their management can be arduous. In certain cases, a simpler model may be preferred, especially if the forecasting component is part of a larger planning tool [10]. Numerous studies have explored the factors driving electricity consumption, such as Jannuzzi and Shipper's analysis of residential electricity usage in Brazil, Harris and Lon-Mu's investigation of the dynamic relationships between electricity consumption and variables like weather, price, and consumer income, and Ranjan and Jain's analysis of electricity consumption patterns in different seasons in Delhi. Over the past 15 years, several papers have focused on forecasting electricity demand using various techniques. For example, Abdel-Aal et al. applied an AIM model to domestic consumption in Saudi Arabia, Yan presented residential consumption models for Hong Kong utilizing climatic variables, and Egelioglu et al. explored the impact of economic variables on annual electricity consumption in Northern Cyprus. Mohamed and

Bodger developed a model for electricity forecasting in New Zealand based on multiple linear regression, incorporating economic and demographic variables. Al-Ghandoor et al. introduced a model for forecasting electricity consumption in the Jordanian industrial sector using multivariate linear regression of time series, Erdogdu proposed an electricity demand estimation and forecast model for Turkey based on ARIMA, and Amarawickrama and Hunt conducted a time series analysis of electricity demand in Sri Lanka, studying the performance of various time series estimation methods in modeling past electricity demand and forecasting future consumption. Algeria has made significant efforts to develop its electricity and gas infrastructure, resulting in universal electricity access for its citizens in 2018, as reported by the World Bank. The country has experienced a substantial increase in electricity consumption in recent years, with a growth rate of 65 percent between 2010 and 2016. The surge in electricity demand is primarily driven by population growth, increased household and transport sector demand, and subsidized electricity prices. Natural gas serves as the primary source for electricity generation in Algeria, accounting for over 96 percent of the electricity generated in 2016. So with this came the idea of this work, trying different methods and concepts to predict the electricity consumption of day a month or even a year at least being close to the real value but what is close enough that can be accepted, and are these methods effective and in what they differ and what are exactly those concepts and how they operate.

Chapter 1

The principles of clustering and linear regression

1-Machine Learning:

In recent years, machine learning has risen to prominence as a leading artificial intelligence technique, capturing significant attention. Its primary goal is to improve the accuracy of software applications without relying on explicit programming. Nevertheless, the question persists: how do machines acquire knowledge? At its core, machine learning involves developing algorithms that can ingest large volumes of data and utilize statistical analysis to generate reasonably precise outcomes. Machine learning algorithms are commonly classified into two primary categories: supervised and unsupervised.

2-Unsupervized machine learning:

Unsupervised learning involves training a machine using unclassified and unlabeled information, enabling the algorithm to operate independently based solely on that data. The objective of the machine in this context, is to autonomously organize, unstructured information into clusters based on similarities, patterns, and distinctions, without any pre-existing data training. Unlike supervised learning, unsupervised learning does not rely on a teacher or explicit training for the machine. Consequently, the machine is intrinsically responsible for independently unraveling the hidden structure within unlabeled data. One of its most used and known implementations or techniques is called clustering.

3-Clustering:

Clustering refers to the process of categorizing unlabeled data into groups, also known as clusters, based on similarities between the data items. [1] A cluster comprises data items that share common characteristics with each other, while being different from the data items in other clusters. An example of clustering in action is the work of John Snow, a London physician who used a map to plot

cholera deaths during an outbreak in the 1850s. By identifying clusters of cases around specific intersections with polluted wells, Snow was able to reveal the problem and the solution to the outbreak.

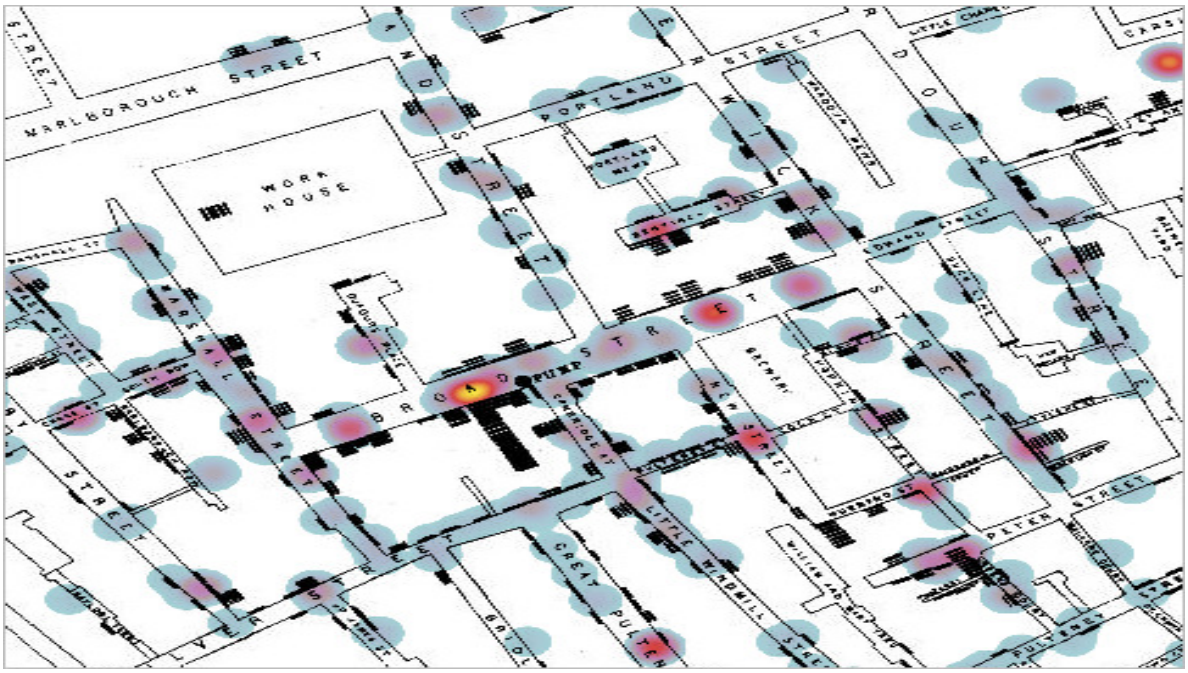


Fig 1: Cholera map 1850 in London

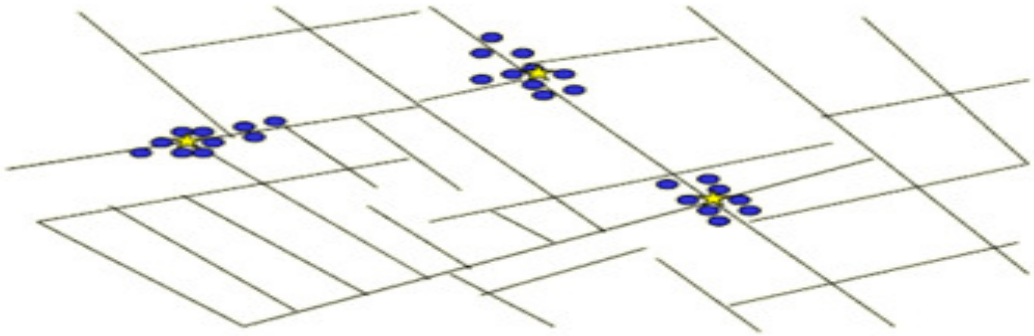


Fig 2: Cholera map divided by clusters from Nina mishra HP labs

And this was one of the first Historical applications of clustering.

Clustering is an unsupervised learning method that involves grouping data points into classes or clusters of similar objects without any predefined labels assigned by a human expert. Unlike classification, where labels are assigned to

data points, clustering relies on algorithms to group similar data points together based on their properties and features. The goal of clustering is to identify similarities and dissimilarities between data points and to group them accordingly. For instance, if we have a set of data points, we can use clustering algorithms to classify each data point into a specific group, where data points within the same group share similar characteristics or features. Conversely, data points in different groups have highly dissimilar properties or features. For instance, a simple example of clustering is shown in the next figure, where we can visually see that there are three distinct clusters of points that share similar properties or features.

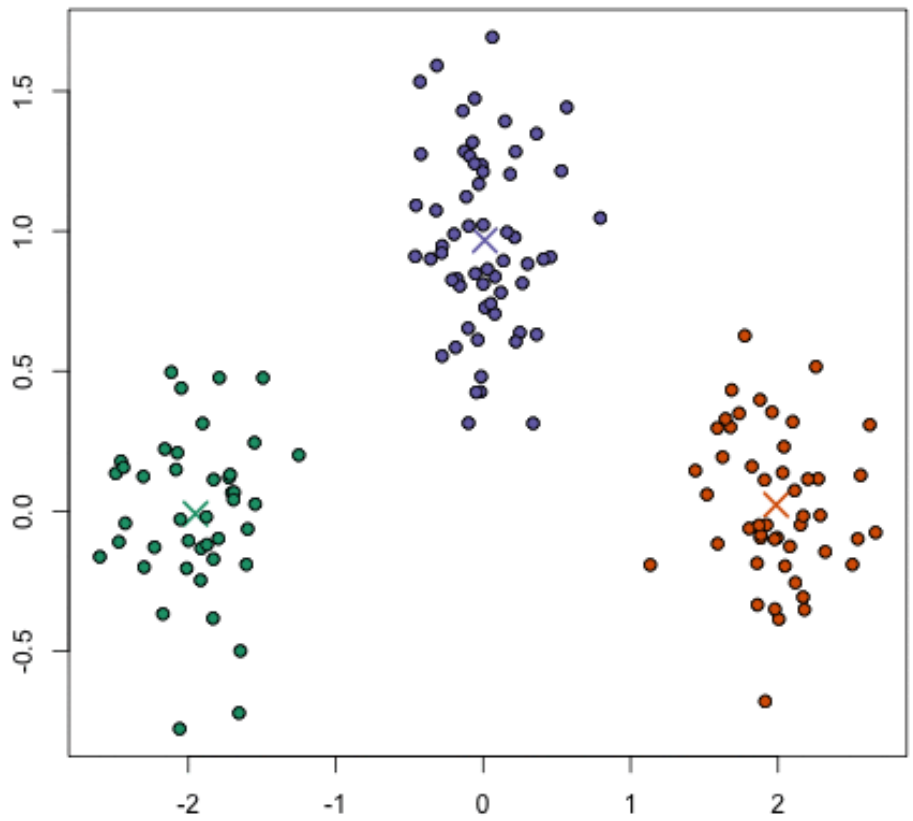


Fig 3: Group of points divided into 3 clusters

In the field of Data Science, we can utilize clustering analysis to extract valuable insights from our data. This involves observing the groups that data points are classified into when subjected to a clustering algorithm. A crucial factor in

clustering algorithms is the distance measure used. Additionally, the determination of the optimal number of clusters is a significant consideration in clustering analysis.

From example, how many clusters are they here?

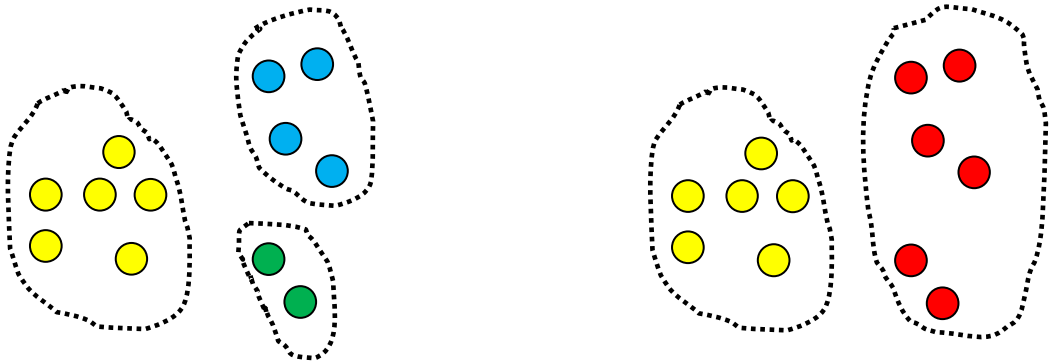


Fig 4:clusters number

Is it 3 or 2 clusters?

Here there is 2 possible approaches either taking a random number n of clusters or find the optimal n and the optimal number of elements that is within each cluster by using for example the k-means method that will be detailed in the next section [2].

With the understanding of what is a cluster the next question could be is there only one approach to define clusters for a given set of data, actually no there is quite a few techniques hierarchical, partitional and Bayesian.

Now being more specific Clustering techniques involve various methods for grouping data points into clusters based on their similarities. Hierarchical algorithms are one type of clustering method that can be either agglomerative or divisive. Agglomerative algorithms start with individual elements as separate clusters and merge them into larger clusters, while divisive algorithms begin with the entire dataset and then divide it into smaller clusters [3].

Another type of clustering method is partitional algorithms, which determine all clusters at once but can also be used in a hierarchical clustering approach as divisive algorithms.

Finally, Bayesian algorithms attempt to generate a posteriori distribution over the possible partitions of the data, allowing for more flexibility in the clustering process.

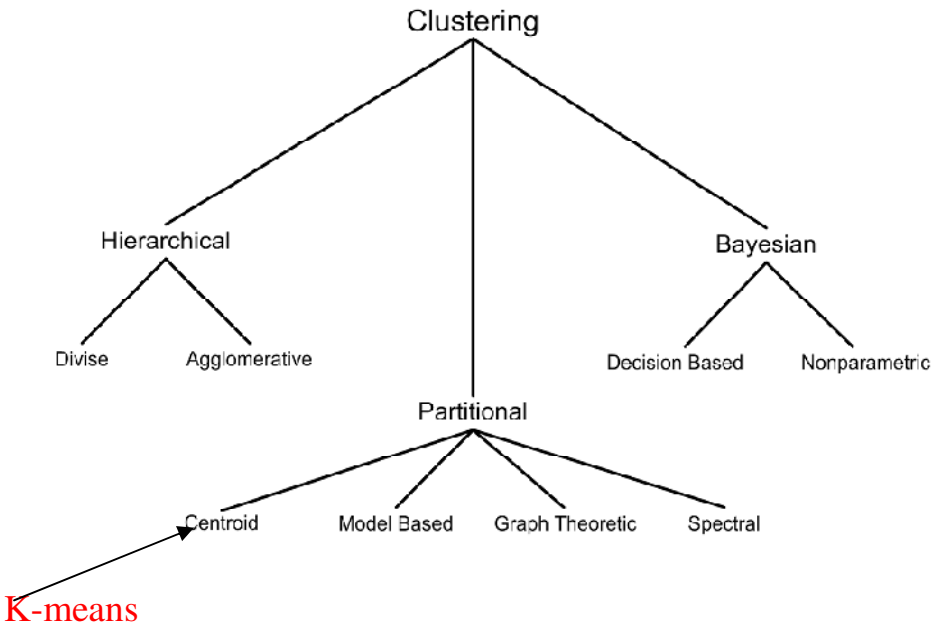
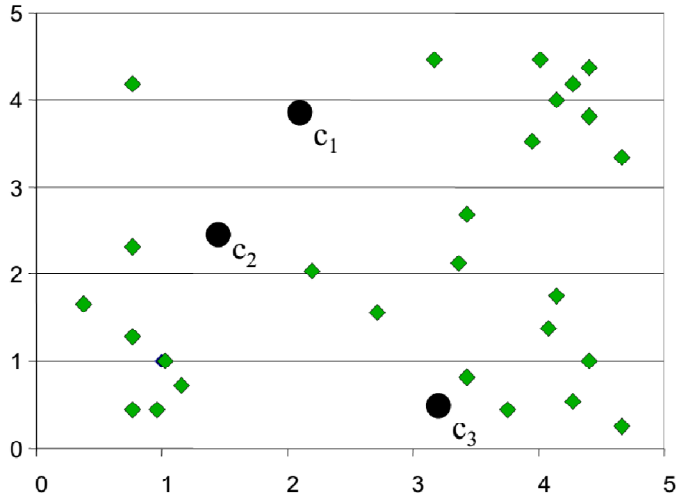


Fig 5: Clustering techniques

Seeing all these clustering techniques, the more suitable one is partitional more precisely using cancroids usually called K-means, because it is simple to understand and apply, the results are shown fast due to its algorithm that requires a single pass through the dataset witch make it very good for large number of elements or data, when rapid results are needed and it is by far the most popular one among the other techniques.

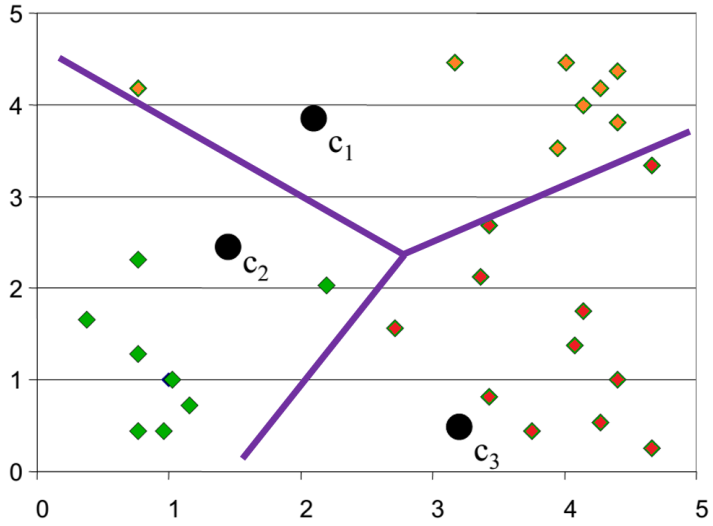
Here's how the algorithm works:

Step 1: Randomly initialize the cluster centers (synaptic weights):



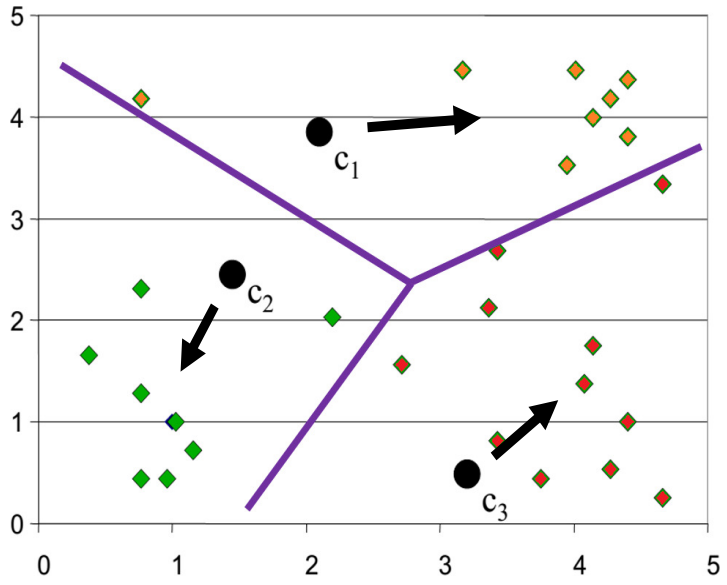
The algorithm begins by randomly selecting initial cluster centers. These centers act as the central points around which clusters will form. The number of cluster centers is determined in advance based on the desired number of clusters.

Step2: Determine cluster membership for each input (“winner-takes-all” inhibitory circuit):



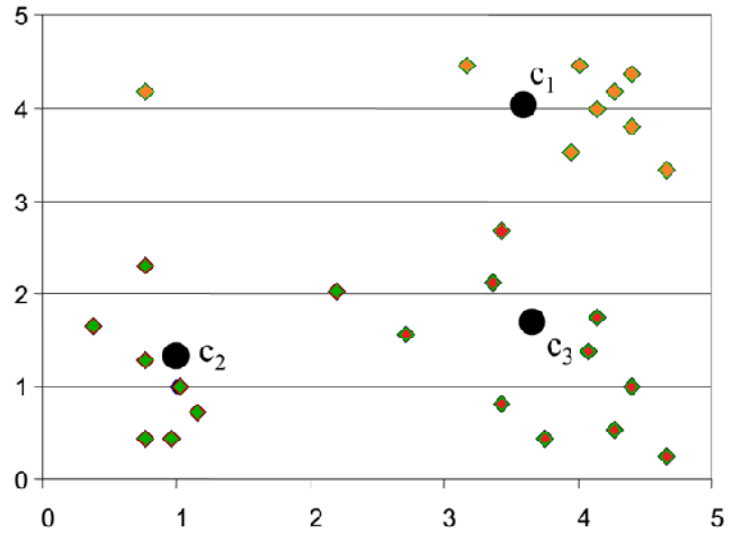
Each input data point is assigned to the nearest cluster center using a distance metric, usually the Euclidean distance. The algorithm calculates the distance between each data point and all the cluster centers. The data point is then assigned to the cluster with the closest center. This process is often referred to as the "winner-takes-all" mechanism.

Step3:Re-estimate cluster centers (adapt synaptic weights):



After all data points are assigned to clusters, the algorithm recalculates the cluster centers. The new centers are determined by taking the average position of all the data points within each cluster. This step adjusts the cluster centers to represent the data points in each cluster.

Step4: Result of the first iteration:



Once the first iteration is complete, the algorithm provides the initial result, which includes the cluster assignments for each data point and the updated

cluster centers. This result represents the initial clustering based on the randomly initialized cluster centers.

More mathematically the main formulas used in k-means clustering are:

Euclidean distance: This is used to calculate the distance between each data point and the cluster centroids. The Euclidean distance between two points $A(x_1, y_1, z_1)$ and $B(x_2, y_2, z_2)$ is:

$\text{distance}(A, B) = \text{sqrt}((x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2)$

Cluster centroid: This is the mean of all the data points in a cluster. For a cluster C with n data points, the centroid (c_x, c_y, c_z) is:

$C_x = (x_1 + x_2 + ... + x_n) / n$

$C_y = (y_1 + y_2 + ... + y_n) / n$

$C_z = (z_1 + z_2 + ... + z_n) / n$

Sum of squared distances: This is the objective function that k-means tries to minimize. For a cluster C with centroid (cx, cy, cz) and n data points $(x_i, y_i z_i)$, the sum of squared distances is:

$SSD(C) = (x_1 - c_x)^2 + (y_1 - c_y)^2 + (z_1 - c_z)^2 + ... + (x_n - c_x)^2 + (y_n - c_y)^2 + (z_n - c_z)^2$

By minimizing the sum of squared distances between each data point and its nearest cluster centroid, k-means tries to find the optimal partition of the data into k clusters.

The Flowchart of K-means is:

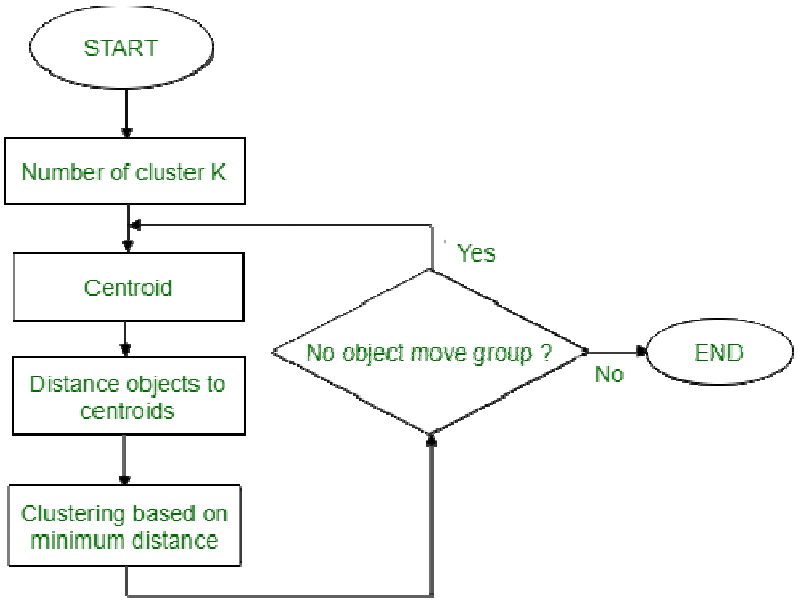


Fig 6: Flowchart of k-means

4-Supervised machine learning:

Supervised learning revolves around the involvement of a supervisor or teacher to guide the learning process. In this approach, the machine is trained using labeled data, where each data point is already associated with its correct answer. The primary objective is to enable the machine to establish a connection between input data and the desired output [11]. The process commences by furnishing the machine with a set of well-labeled training examples. The supervised learning algorithm then scrutinizes this data to identify patterns and relationships. Once the learning phase concludes, the machine becomes capable of making accurate predictions or providing correct answers when confronted with new, unlabeled data.

5-Linear regression:

Linear regression is a statistical approach that establishes a connection between a dependent variable and one or more independent variables. In the context of

forecasting, it can be used to anticipate future values of the dependent variable based on previous data for the independent variable(s).

There are two types of Linear Regression: Simple and Multiple [4]. Simple Linear Regression involves one independent variable, and the model establishes a linear relationship between it and the dependent variable. On the other hand, Multiple Linear Regression involves more than one independent variable, and the model has to determine the relationship between them and the dependent variable.

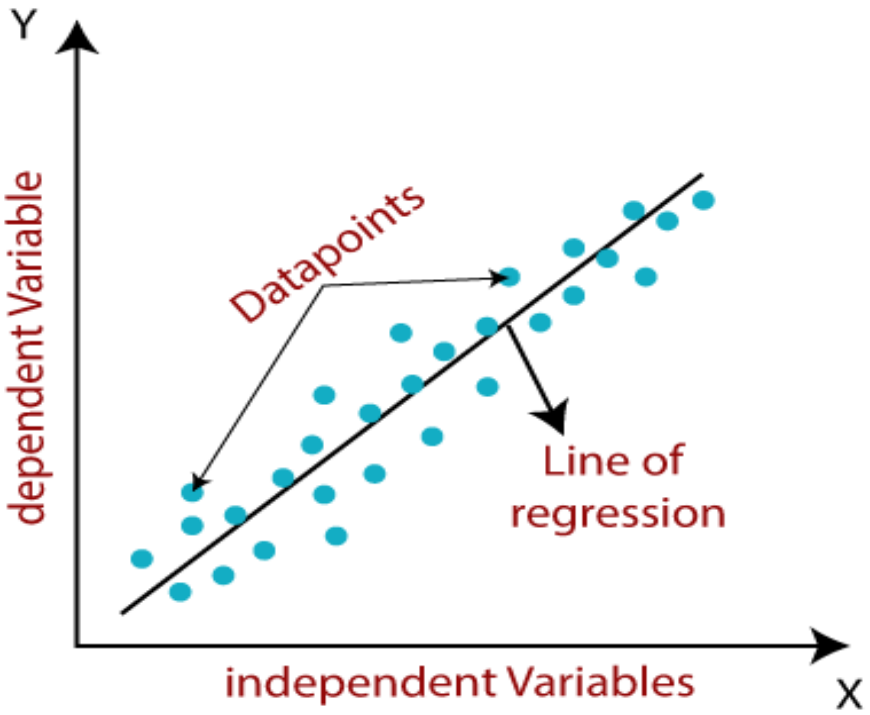


Fig7: scatter plot of a linear regression

In Simple Linear Regression, the equation includes an intercept (b_o) and a coefficient or slope (b_1) that describe the linear relationship between the independent variable (x) and dependent variable (y).

$y = b_0 + b_1 x$

[5] In Multiple Linear Regression, the equation also includes an intercept (bo) and multiple coefficients or slopes (b_1, b_2, b_3, ..., b_n) that correspond to each independent variable (x_1, x_2, x_3, ..., x_n) and their relationship with the dependent variable (y).

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots\ldots\ldots + b_nx_n$$

The primary objective of a Linear Regression model is to determine the optimal values of the intercept and coefficients that result in the best fit linear line, which minimizes the error between the actual and predicted values. The error represents the difference between the actual value and the predicted value, and the ultimate goal is to minimize this difference.

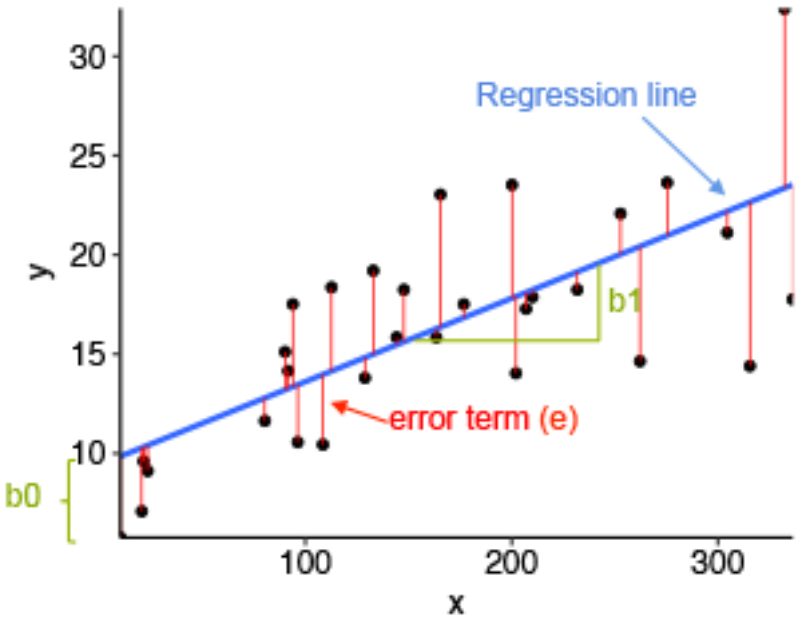


Fig8:A detailed scatter plot of a linear regression

The above diagram displays a scatter plot where the independent variable (x) is shown on the x-axis, and the dependent variable (y) is shown on the y-axis. The black dots represent the actual data points.

The intercept of the linear regression model is represented by bo, which has a value of 10, and the coefficient or slope of the x variable is represented by b1.

The blue line represents the best fit line predicted by the model, and the predicted values are expected to fall on this line.

With the understanding of what is exactly a linear regression here are the general steps to perform forecasting using linear regression:

1*Collect data: The first step is to gather historical data for the independent variable(s) and the dependent variable.

2*Split the data: Split the data into two parts, one for training the model and another for testing the model. Typically, a ratio of 70:30 is used, where 70% of the data is used for training and 30% is used for testing.

3*Choose the model: Linear regression models can be simple or multiple. A simple linear regression model has only one independent variable, while a multiple linear regression model has two or more independent variables.

4*Fit the model: Using the training data, fit the linear regression model to establish the relationship between the independent variable(s) and the dependent variable.

4*Predict the values: Once the model is fitted, use it to predict future values of the dependent variable based on the values of the independent variable(s) in the test dataset.

5*Evaluate the model: Evaluate the accuracy of the model by comparing the predicted values to the actual values in the test dataset. Common metrics used to evaluate the model include mean squared error (MSE), mean absolute error (MAE), and R-squared.

6*Refine the model: If the model's performance is not satisfactory, refine the model by tweaking its parameters and repeating steps 4-6 until the model produces satisfactory results.

The vertical distance between each data point and the regression line is referred to as the error or residual. Each data point has its residual, and the sum of all the residuals is known as the Sum of Residuals/Errors.

But there is still one point to fully comprehend the linear regression and that being the metrics that are used to evaluate the model which are: MAE,MSE, RMSE and R squared at least these are the most popular ones [6]:

1-The Mean Absolute Error (MAE) is a metric used to measure the average size of the errors in a set of predictions. An error represents the absolute difference between the actual or true values and the predicted values. This means that negative signs are ignored, and only the magnitude of the difference is considered.

The formula for MAE is calculated by taking the absolute difference between the true values and the predicted values and then averaging this error over all the samples in a dataset.

In summary, MAE provides an estimate of the average magnitude of the errors between the actual and predicted values in a dataset.

2-The Mean Squared Error (MSE) is a measure of how close a regression line is to a set of points. It is calculated as the squared mean of the difference between the actual values and the predicted values.

To calculate MSE, these steps need to be followed:

1*Find the regression line that best fits the data.

2*Use the regression line to predict the Y values (Y') for each X value in the dataset.

3*Calculate the difference between each actual Y value and its predicted Y' value to get the error for each point.

4*square each of the error values obtained in step 3.

5*Add up all the squared errors from step 4.

Divide the total sum of squared errors by the number of observations in the dataset to get the mean squared error.

The formula for MSE can be written as:

MSE = 1/N * Σ(i=1 to N) (Yi - Y'i)^2

Where N is the total number of observations in the dataset, Yi is the actual Y value, and Y'i is the predicted Y value.

In summary, MSE is a measure of how well a regression line fits a set of data points, and it provides a way to quantify the errors between the actual and predicted values in a dataset.

3- RMSE stands for Root Mean Squared Error, which is a measure of the differences between actual and predicted values in a dataset. RMSE is calculated by taking the square root of the average of the squared differences between the predicted and actual values. It is similar to the MSE (Mean Squared Error) but the root of the value is taken, which makes it more interpretable in the same units as the dependent variable.

RMSE is commonly used in regression analysis to evaluate the performance of a model. A lower RMSE indicates a better fit of the model to the data.

4- Finally,the R-squared metric, also called the coefficient of determination, evaluates how well a linear regression model fits a given dataset. It measures the proportion of the variance in the dependent variable that can be explained by the independent variable(s) in the model.

The R-squared value ranges from 0 to 1, where 0 indicates that the model does not fit the data at all, and 1 indicates that the model fits the data perfectly. Essentially, R-squared indicates how close the regression line (i.e., the line of predicted values) is to the actual data values.

6-Conclusion:

In summary, K-means and linear regression serve as invaluable tools in the arsenal of a data scientist. K-means excels in uncovering patterns and clusters in datasets, whereas linear regression enables the prediction and understanding of

relationships between variables. Acquiring a deep understanding of these techniques and applying them effectively can greatly enhance the processes of data analysis and decision-making, as will be evident in the forthcoming results chapter.

Chapter 2:

Artificial Neural Network

1-Introduction:

To understand the next method used to forecast, that is artificial neural network, the understanding of IA and deep learning is quite important to comprehend, in what the artificial neural network defer from the other techniques.Therefore, in general, AI refers to the ability of a machine to replicate cognitive functions typically associated with human intelligence, such as learning and problem solving. However, AI can also involve simple programmed rules that dictate a machine's behavior in specific situations. Thus, AI can range from complex algorithms to basic if-else statements, which are simple rules programmed by humans, in the other hand Deep learning is a specialized branch of AI that aims to create artificial neural networks capable of learning from data. By training on large volumes of labeled data, deep learning algorithms are designed to improve automatically over time. This approach has paved the way for remarkable advancements in several areas, including speech and image recognition, natural language processing, and self-driving cars.

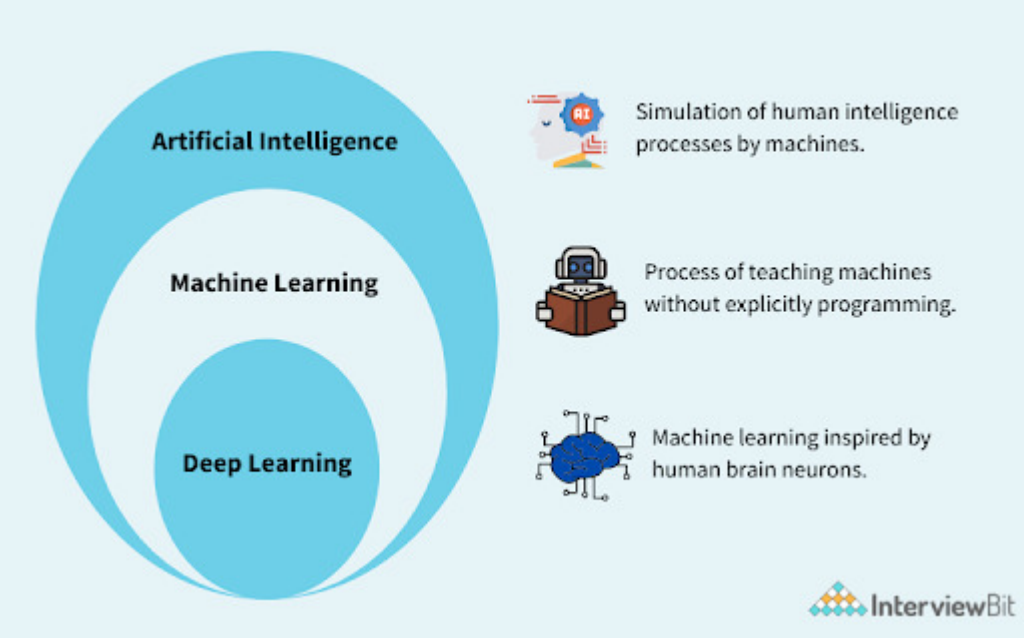


Fig9: comparison between AI, machine learning and Deep learning

With figure10 things become clearer and make a good initiation to artificial neural networks.

2- Artificial Neural Network:

Artificial Neural Networks consist of artificial neurons called units, which are organized into layers to create a comprehensive system. The number of units in each layer can vary greatly, depending on the complexity of the patterns that the neural network needs to learn from the dataset.[7] Typically, an Artificial Neural Network is composed of an input layer, one or more hidden layers, and an output layer. The input layer receives data from external sources, and the hidden layers process this information to generate valuable input for the output layer, which then produces a response to the input data.

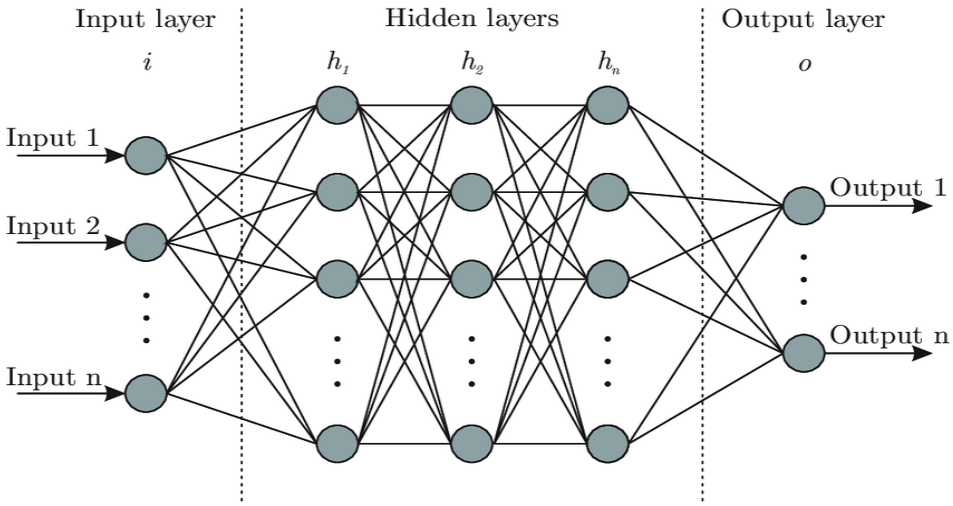


Fig 10: Artificial Neural Networks

In general, the units in an Artificial Neural Network are linked between layers, and the weight of each connection determines how much one unit influences another. As data passes through the network, it learns more about it, leading to an output from the output layer. Artificial Neural Networks are based on the structure and function of human neurons, and they are also known as neural networks or neural nets. During training, the connection weights are adjusted to

optimize the effects of the inputs from the previous layer, resulting in an enhanced model performance.

Artificial neural networks are modeled after biological neurons found in animal and even human brains, sharing similarities in both structure and function. Biological neurons consist of dendrites, a cell body or soma, and an axon that transfers impulses to other neurons. Similarly, artificial neural networks have input nodes that receive signals, hidden layer nodes that process these signals, and output layer nodes that compute the final output using activation functions. In both biological and artificial neurons, synapses facilitate the transmission of impulses and are represented by weights that connect nodes. Learning in biological neurons occurs in the cell body nucleus or soma through synaptic plasticity, while in artificial neural networks; back propagation adjusts weights between nodes based on prediction errors. Activation in biological neurons is the firing rate of the neuron, while in artificial neural networks, an activation function maps the input to the output.

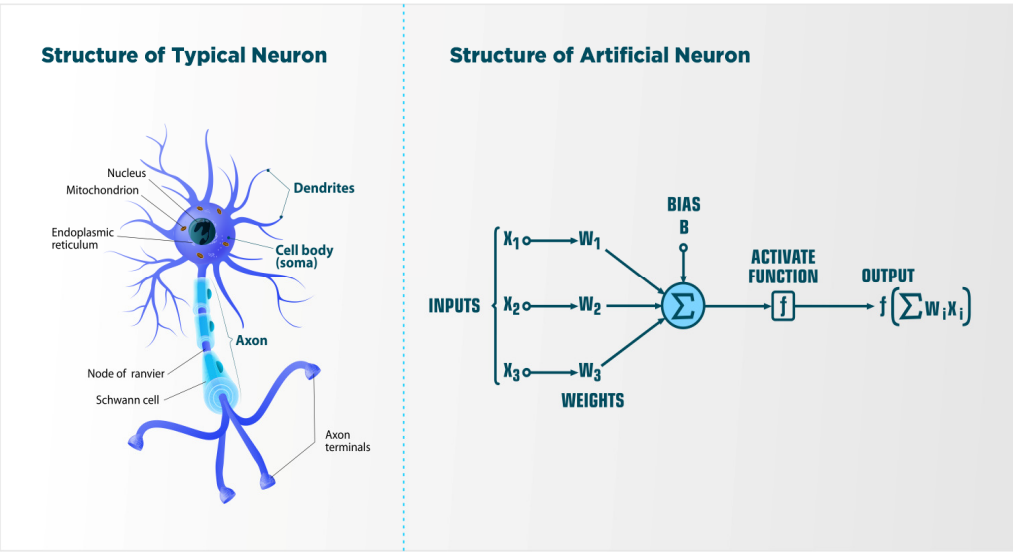


Fig11: neural network in human brain next to an artificial neural network

The introduction of the most important basics in artificial neural network starts with weights, in artificial neural networks, weights refer to the values assigned to the connections between nodes [8]. Every connection in the network has a

corresponding weight that indicates the strength of the link between two nodes. These weights serve as adjustable parameters that are fine-tuned during the network's training process to produce the intended output. As seen in the fig12 there is a term that has not been mentioned in the paragraphs above and that is the bias, which is not very hard to comprehend because the concept of bias involves adding a fixed value to the input of an activation function. The primary objective of bias is to shift the function, much like how a constant in a linear function adjusts the position of the line. This allows for precise control over the output of the function by manipulating the input. The use of bias in Neural Networks allows for more nuanced and accurate modeling of complex relationships, providing greater flexibility and adaptability in the learning process.

How the output is actually generated what technique does go throw the neural network to give the result and here comes the role of the forward propagation,Forward propagation is a crucial process in neural networks that helps determine the effectiveness of the assigned weights in solving a given problem. This process involves two key steps: first, the weighted vector is multiplied by the input vector to obtain a product, then sum that is computed in each layer until the final decision is made. Second, the product sum is passed through an activation function in each layer, which generates the output for that layer. This output becomes the input for the next layer, and the process repeats until the final output layer is reached. Then there are activation functions that refers to a mathematical function applied to the output of each neuron or node in the network[9].The main objective of the activation function is to introduce nonlinearity into the output of a neuron, which enables the neural network to learn and model complex relationships between inputs and outputs.

Typically, activation functions are applied to the weighted sum of the inputs to a neuron, also known as the activation. By applying a transformation to this

activation value, the activation function produces the output of the neuron, which is then passed to the next layer in the network.

Several types of activation functions are commonly used in neural networks, including the sigmoid function, ReLU (Rectified Linear Unit) function, tanh (hyperbolic tangent) function, and sigmoid function among others. The choice of activation function depends on the specific problem being addressed and the properties of the data involved.

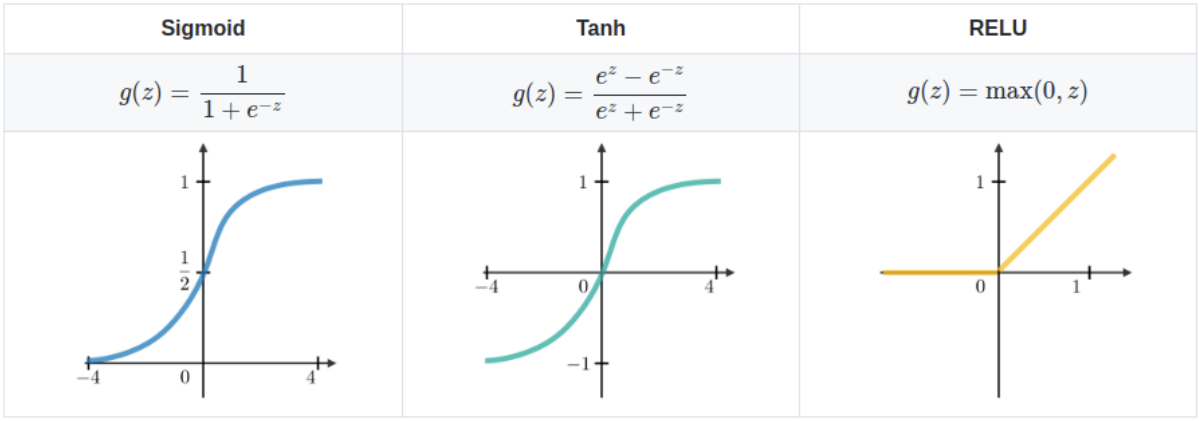


Fig12 : Activation functions

Finally comes the role of backward propagation, Backpropagation plays a key role in training neural networks by adjusting their weights based on the error rate computed in the previous iteration which is the difference between the predicted and the wanted output. By tweaking the weights appropriately, the error rate can be lowered, leading to a more dependable model that can generalize better, and this can be repeated until the model reaches a satisfactory result.

To resume it:

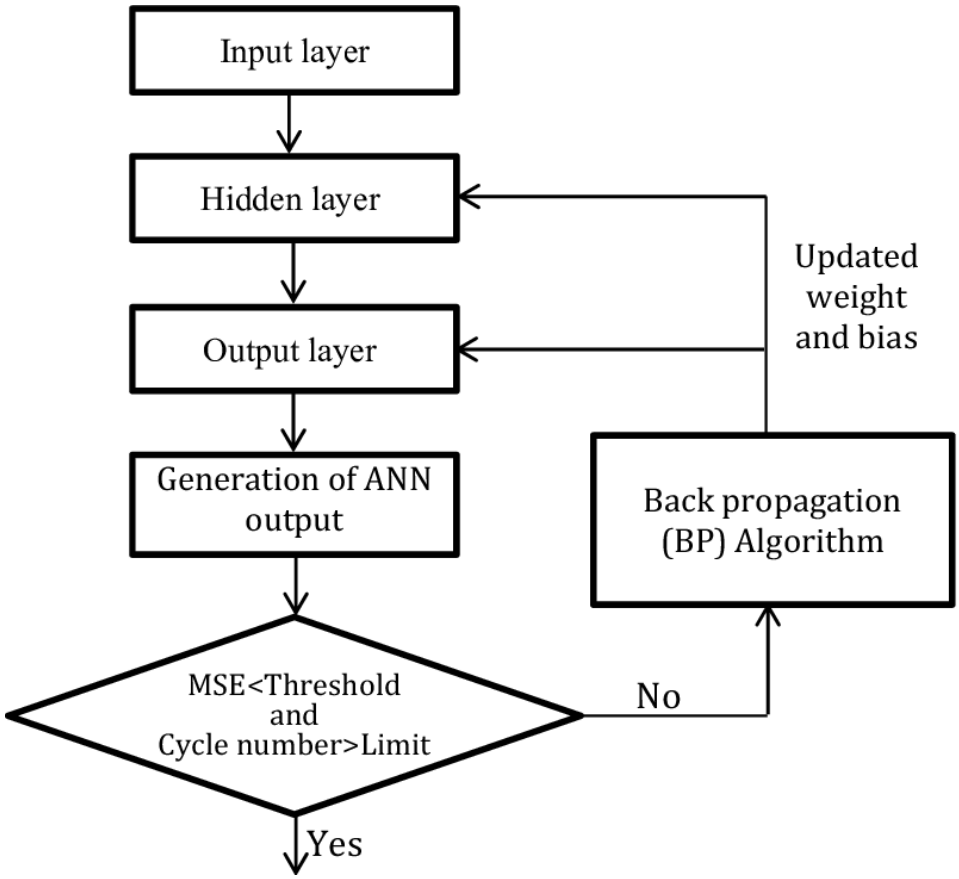


Fig13: Flowchart of artificial neural network

3-Conclusion:

The chapter provided a comprehensive examination of the essential elements comprising artificial neural networks. These include the foundational components such as input layers, hidden layers, and output layers, as well as the critical role of activation functions in introducing non-linearity to the model. Furthermore, the chapter delved into the crucial training process, emphasizing the significance of techniques like backpropagation in effectively optimizing the network's weights and biases.All this, will be applied and tested in the next chapter.

Chapter3:
Applications
And
Results

1-Introduction:

This chapter will show the way that the methods shown in the previous chapters were used in real life case scenario alongside their respective results.

The data set:

The data set used here is the global electricity consumption in Algeria between 01-01-2008 and 01-02-2020 in KW contained in an excel file.

Clustering:

In this case study, the k-means clustering algorithm is employed to identify similarities among data points (days in this case). The goal is to group together the days that exhibit similar behaviors, enabling the selection of a representative data point (day) to represent each cluster of data.

At first this set of data were separated by year and each year is divided into different clusters, the number of clusters is defined using the elbow method, taking on consideration the size of the data only the years 2017, 2018 and 2019 since they are the most recent ones.

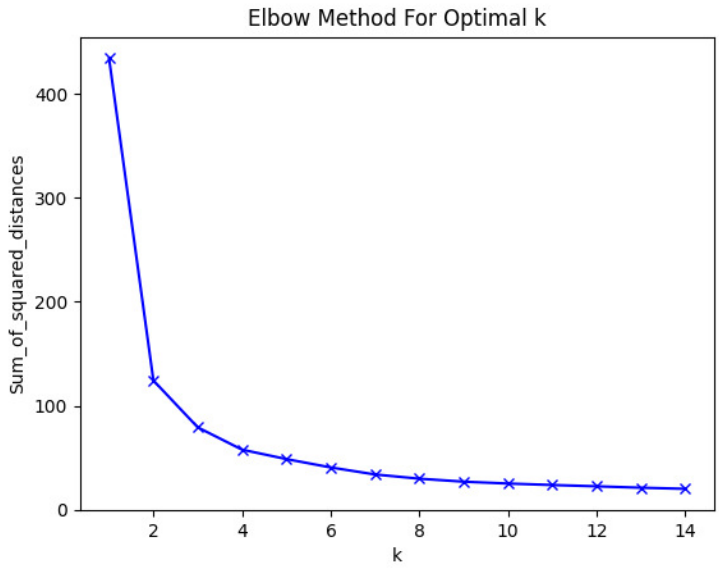


Fig14: Elbow method used in the year 2018

This plot represents the result of the elbow method in the year 2018. The results are the same for the other years and from it can be seen that the optimal number

of clusters is around 3 to 8 and then it stabilize. Thus 4 clusters seems to be the right choice because from there the difference between the number of clusters decreases and because this number mustn't be too small or too high cause it negate the whole point of clustering making the data hard to analyze.

Now that number of clusters is set to 4 the k-means algorithm can be applied to the years. The used code is shown below.

```
importpandasaspd
importnumpyasnp
importmatplotlib.pyplotasplt
fromsklearn.clusterimportKMeans
fromsklearn.preprocessingimportMinMaxScaler
fromsklearn.preprocessingimportStandardScaler
df = pd.read_excel(r'C:\Users\PC\Desktop\BDD.xlsx')
wmax = df.set_index('Date').resample('1D').max()
df2 = wmax.loc['2018-01-01':'2018-12-31', :]
df3 = wmax.loc[:, '1h':'24h']
kmeans = KMeans(n_clusters=4).fit(df2)
centroids = kmeans.cluster_centers_
df2['kmeans_4'] = kmeans.labels_

nbjour = df2.pivot_table(index = ['kmeans_4'], aggfunc
='size')

print(nbjour)

plt.scatter(df2.index, df2['kmeans_4'], s=1)

plt.show()
```

According to this program, results are as follow:

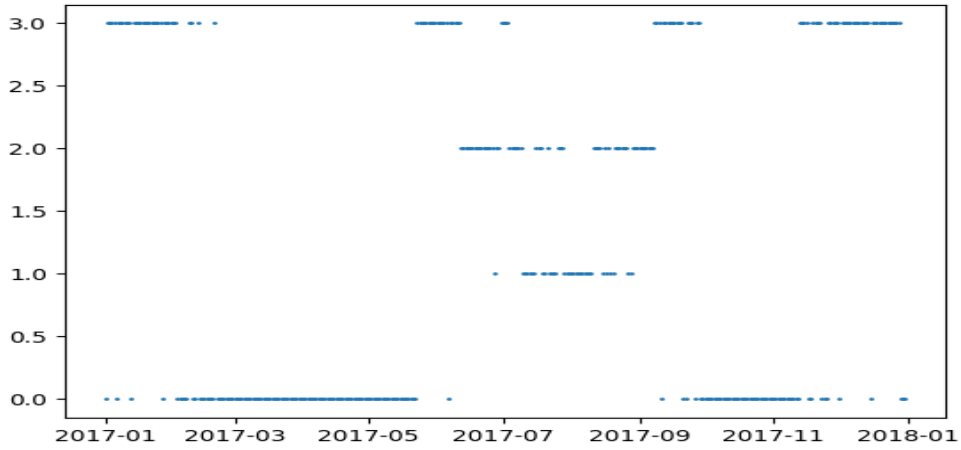


Fig15:2017 into 4 clusters

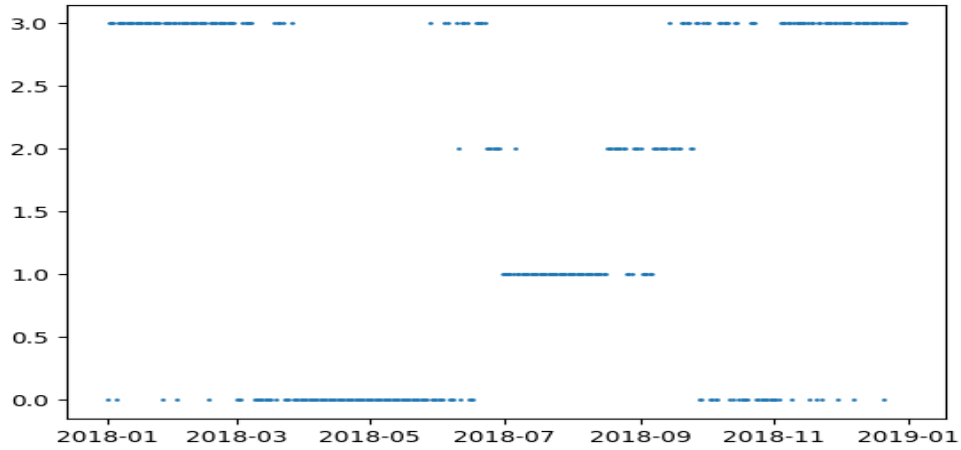


Fig16:2018 into 4 clusters

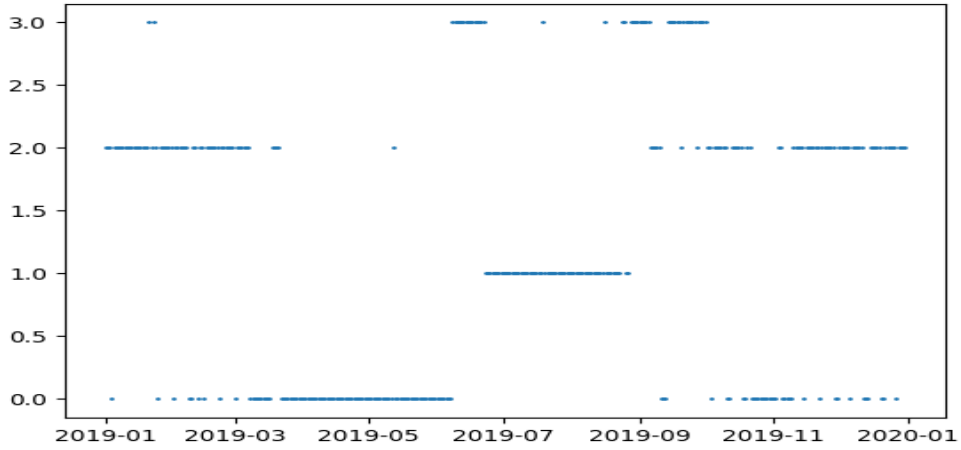


Fig17:2019 into 4 clusters

At first glance without seeing the plots, it can be expected that results will represent seasons with each cluster representing a season since the number of clusters is 4.But after a depth analyze of the plots above, 2 clusters can be each assigned to summer and winter respectively and the other two clusters contain separate fragments of spring and fall kind of a mix and sometimes the clusters contain more than one season but the summer can always be distinguished.

In theory, the days in within the same cluster follow the same pattern (model), these days could be any day of week and a week is constituted of 7 days. Is it possible to regroup some similar days together so that it would not take a model for each cluster and then a model for each day of the week that isa bit much.

As the number of clusters in a year was determined by the elbow method as being 4, the same process was applied for the week.

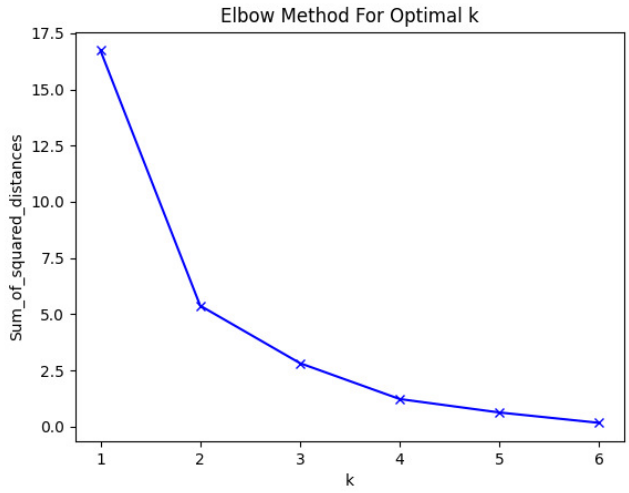
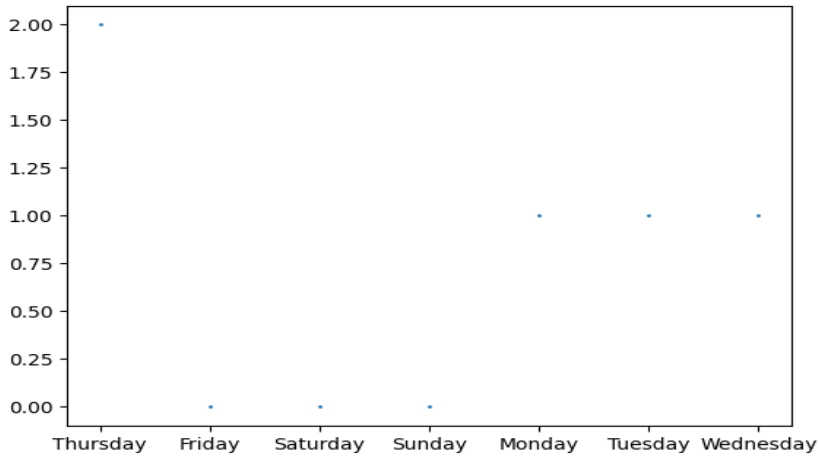
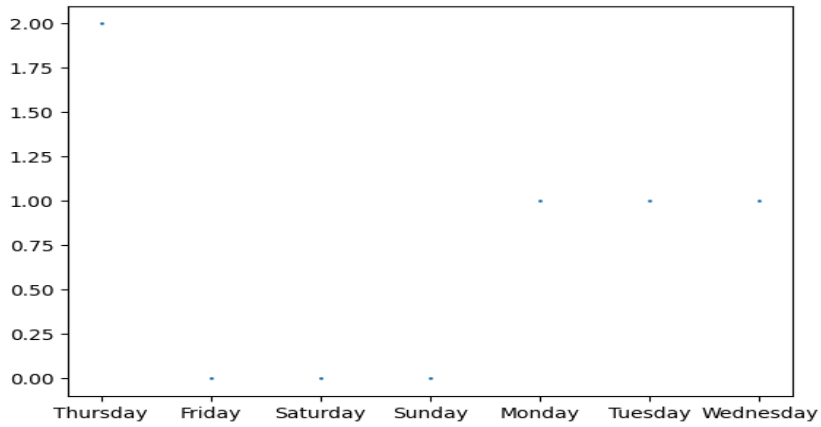


Fig18: Elbow method used on a week

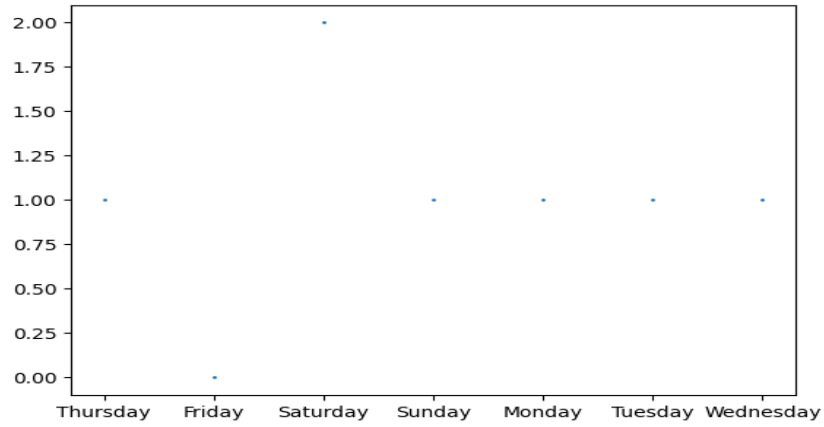
This plot is valid for almost if not every week; based on it the number 3 seems to be the best option with that the results of clustering were as follow:



a)-



b)-



C)-

Fig19: Plots a, b and c represent the result of clustering in 3 different weeks

The plots above are different, but in all of them, the days in the middle of the week are in the same cluster. Being more precise 80 percent of the time one cluster will have Saturday another one for Friday and the last one for the remaining days which will be referred to them by working days like the plot C.
To summaries clustering a year into 4 clusters and the week having 3 clusters the tables are as follow:

Table 1: Data clustering by 4 (2018)

Cluster number	Each Friday		Each Saturday		Working Days		All The Days	
	Number of elements per cluster	Percent %	Number of elements per cluster	percent	Number of elements per cluster	Percent %	Number of elements per cluster	Percent %
1	27	52	21	41	105	40	147	40
2	6	12	7	14	43	16	55	15
3	6	12	17	33	89	34	127	34
4	13	24	6	12	24	10	36	11

Table 2: Data clustering by 4 (2017)

Cluster number	Each Friday		Each Saturday		Working Days		All The Days	
	Number of elements per cluster	Percent %	Number of elements per cluster	percent	Number of elements per cluster	Percent %	Number of elements per cluster	Percent %
1	6	12	26	50	144	55	111	30
2	26	50	6	12	39	15	32	9
3	6	12	15	30	17	7	171	47
4	14	26	5	8	61	23	51	14

Table 3: Data clustering by 4 (2019)

Cluster number	Each Friday		Each Saturday		Working Days		All The Days	
	Number of elements per	Percent	Number of elements per	percent	Number of elements per	Percent	Number of elements	Percent

	cluster	%	cluster		cluster	%	per cluster	%
1	8	15	8	15	47	18	132	36
2	2	4	23	44	98	38	62	17
3	7	13	8	15	33	13	125	34
4	35	68	13	26	83	31	46	13

Linear regression:

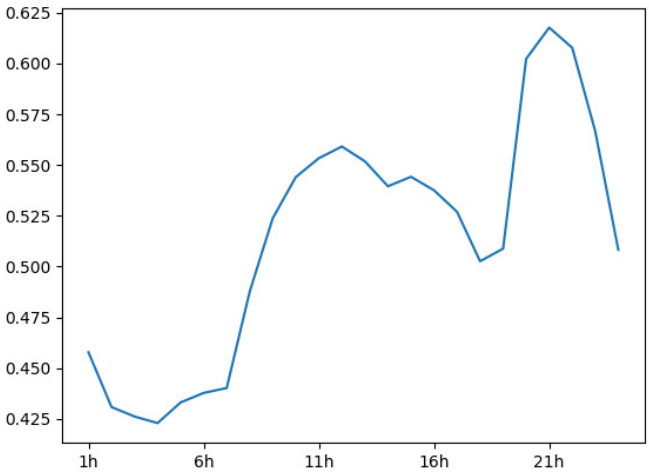
Linear regression is based on a training set, usually the training set is 70%, and the test set 30% of the whole data set. There is a little problem, the used data set is too big to have comprehensive result over 4000 days and for that instead of taking the 365 days of every year simply take the max out of every year, that will be 12 values without counting 2020 because it hasonly 3 months. In the data set, the power consumption profile of each year is divided by the maximum power of the corresponding year in order to get reorganized data set in pair unit values, with that each year will have a model.

The linear regression will have 12 values one for each year, 8 for training and 4 for test and the code is as follow:

```
importpandasaspd
importnumpyasnp
fromsklearn.model_selectionimporttrain_test_split
fromsklearn.linear_modelimportLinearRegression
importmatplotlib.pyplotasplt
importseabornassns
fromscipyimportstats
df = pd.read_excel(r'C:\Users\PC\Desktop\BDD.xlsx')
df2 = df.set_index('Date').resample('1Y').max()
df3 = df2.max(axis=1)
df4 = pd.DataFrame(df3).reset_index()
df4.columns = ['Date', 'Max Consomation']

df4['Year'] = df4['Date'].dt.strftime('%Y')
deldf4["Date"]

df4 = df4.drop(df4.index[12])

df4['Year'] = df4.Year.astype(int)
```

```
X = df4.iloc[:,1].values
y = df4.iloc[:,:-1].values
X = X.reshape(12, 1)
print(X.shape)
print(y.shape)
print(X)
print(y)

X_train,            X_test,            y_train,            y_test            =
train_test_split(X,y,test_size=1/3,random_state=0)
regressor = LinearRegression()
regressor.fit(X_train,y_train)
y_pred = regressor.predict(X_test)
print(X_test)
print(y_test)

print(y_pred)
plt.scatter(X_train, y_train, color='red') # plotting the observation line

plt.plot(X_train,  regressor.predict(X_train),  color='blue')  #  plotting  the
regression line

plt.title("(Training set)") # stating the title of the graph

plt.xlabel("Years") # adding the name of x-axis
plt.ylabel("Consomation max") # adding the name of y-axis
plt.show()

plt.scatter(X_test, y_test, color='red')
plt.plot(X_train,  regressor.predict(X_train),  color='blue')  #  plotting  the
regression line

plt.title("(Testing set)")

plt.xlabel("Years ")
plt.ylabel("Consomation max")
plt.show()
```

This is the model of a random day, with x axis being the hours and the y axis being the consumption divided by the max of that year.



To find the real values simply the model will be multiplied by the max of the year.

And as for the results of the previous given code:

Fig20: Train set for linear regression

In this representation, the coordinates of each point indicate the year on the x-axis and the corresponding maximum consumption (kw) for that specific year on the y-axis.

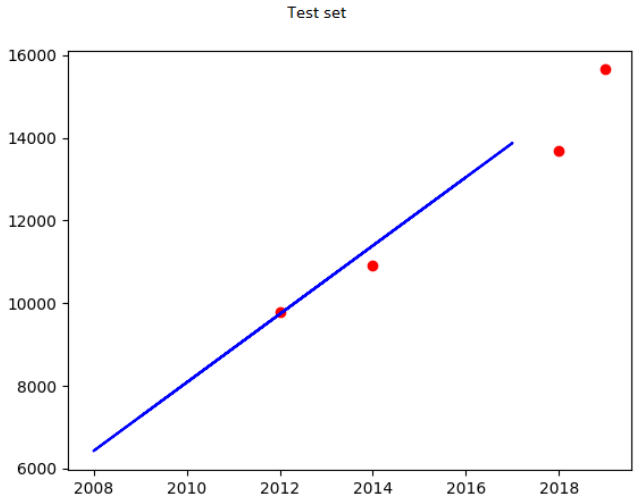


Fig21: Test set for linear regression

Using the function that resulted from the linear regression, the max of 2020 can be predicted, then the profile of 2019 can be used to find the eclectic consumption of the whole year of 2020 by just multiplying the predicted max of 2020 by 2019 profile, and some of the results are represented in the plots bellow: note that the x axis are hours and the y axis electricity consumption Kw.

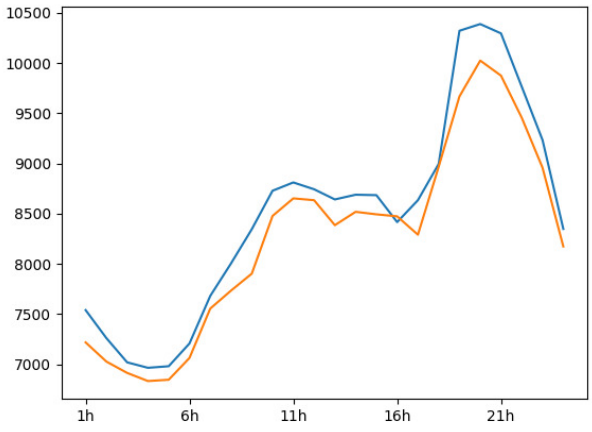


Fig22: Saturday real (blue)/ predicted (orange) in 2020

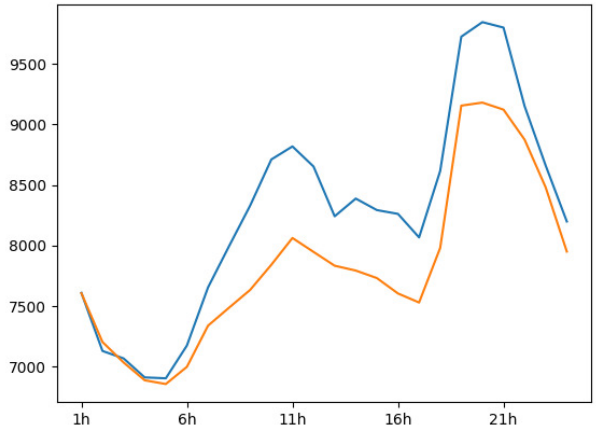


Fig23: Wednesday real (blue)/ predicted (orange) in 2020

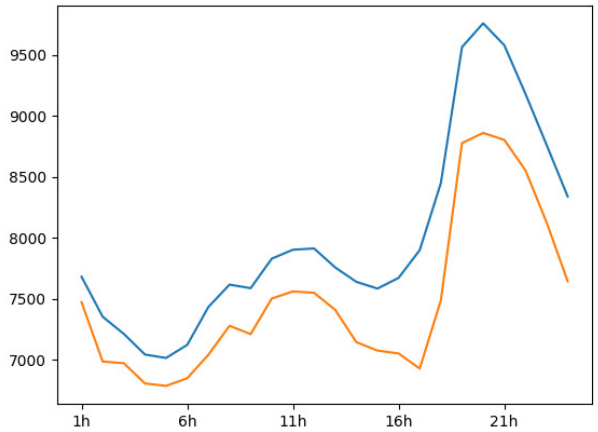


Fig24: Friday real (blue)/ predicted (orange) in 2020

Since the data set only has 3 months of 2020, only the winter could be tested with its real values.So showing the predicted plots without comparing them with the real ones doesn't give a step ahead into evaluating this linear regression, even though the previous plots makes the function acceptable, but to really determine if the results are good enough, the mean between all the predicted maxes needs to be calculated which in this case its equal to 2.9 percent which is not bad, but if the years were seen individually the error between the real consumption and the predicted can be very low as it can reach a difference of a whole power plant or two which is huge.

This is probably because both the training and the test sets were chosen randomly, and it makes sense. Predicting a year like 2019, the training set would rather be the recent years to 2019 than 2008 and 2009 for example because the further the years are the more likely the differences between them will grow and in linear regression the training set is the key of the whole program.It basically decide the linear regression formula so by applying the training set to be the most recent resent years to 2019 and the test set being farthest including 2019.

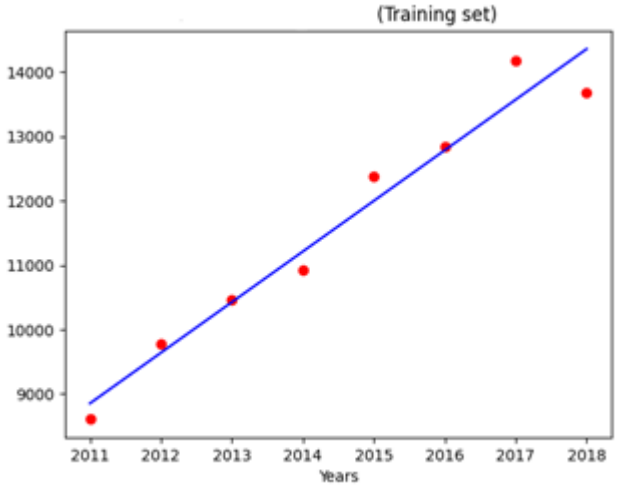


Fig25: Training set for linear regression

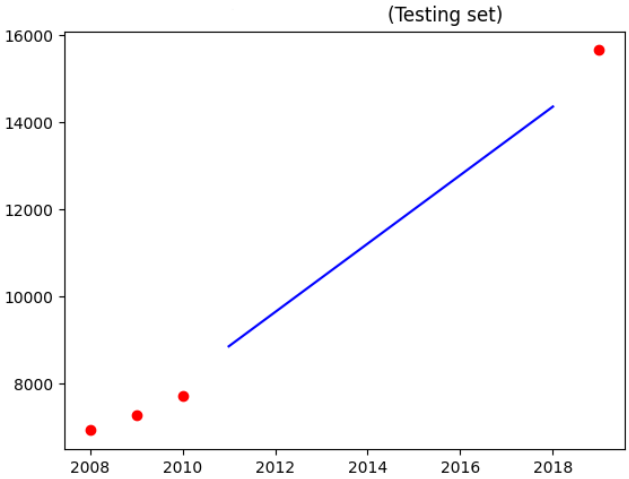


Fig26: Test set for linear regression

With the results of the difference between the real and the predicted values in each year in this table:

	Year	Regression result [MW]	Actual power [MW]	Relative error %	Absolute error [MW]
Test	2008	6492.04	6925	-6.25	432.96
	2009	7278.86	7280	-0.02	1.14
	2010	8065.68	7718	4.50	347.68
Training	2011	8852.50	8606	2.86	246.5
	2012	9639.32	9777	-1.41	137.68
	2013	10426.14	10464	-0.36	37.86
	2014	11212.96	10927	2.62	285.96
	2015	11999.79	12380	-3.07	380.21
	2016	12786.61	12839	-0.41	52.39
	2017	13573.43	14182	-4.29	608.57
	2018	14360.25	13676	5	684.25
Predict	2019	15147.07	15656	-3.25	508.93

Table4: Result of linear regression

Linear regression function: P = 786.82*T -1573445.40.

As seen above in this version of the linear regression model the results are better than the random affiliation and the mean error drops to 2.8 percent, which is not that big of a gap compared to the previous one, but taking the error in each year separately.It is better now even though it is not perfect but still good and acceptable.

Artificial neural network:

For the ANN (artificial neural network) model there are quite a few parameters that need to be set or changed depending on the needs like the number of layers, the loss function and others but first here is the most important part of its code:

```
model = Sequential()
# input layer
model.add(Dense(10, input_dim=26,activation="relu"))

# hidden layers

model.add(Dense(20,activation="relu"))
#model.add(Dense(20,activation="relu"))

# output layer
model.add(Dense(24))
model.compile(optimizer='adam',loss='mean_squared_error',metrics=["mape"])
model.fit(X_train,Y_train,batch_size=364,epochs=200)
performance = model.evaluate(X_test,Y_test,verbose=0)[1]
print("La performance du modèle est : {} % MAPE".format(performance))
```

The model = sequential() Is just for ordering the layers within the model, the dense() is to specify that each neuron is connected to all the neurons in the previous layer. In other words a FC(fully connected) model, the number next to it define how many neurons that layer contains with obviously an activation function here in the code it set to be relu. The last 2 important parts that need to be explained are the batch size and epochs basically. The batch size determines how many data samples are grouped together, and processed before updating the model. In contrast, the number of epochs represents the total number of complete iterations made through the entire training dataset. The verbose is just to display or not the processing information.

Now results of the ANN are show in this table:

Training set	Test set	Input layer	Hidden layer	Output layer	Error
Hours+max	Max and	24	20	2	3,06

and min	Min				
Hours+max and min	Max and Min	24	18	2	3,7
Hours	Hours	24	20	24	3,2
Max and Min	Max and Min	2	20	2	0,3
Hours	Max and Min	24	7	2	2,5
Max and Min	Max and Min	2	4	2	2,6

Table5: Results of the artificial neural network

The result seem to be good the more neurons there are in the hidden layer the more the model becomes more accurate. With that, it over fit the model and it may achieve high accuracy during training. It often fails to generalize well and performs poorly when evaluated on validation or test data, and if that number is low then the network will simply not learn.The real problem is the results in the table above, are not the ones same with each run of the program the difference could be high or low, and by a lot sometimes and that's because each time the weights are randomly generated resulting in a different network with each run.

Conclusion: The data mining was started in the first time by searching independent clusters defining seasons it was seen that winterand summer are almost independent, unfortunately spring and autumn are highly depending. Kmeans allow to define three specific consumption pattern. As a second step, the regression method was used successfully to forecast maximum power consumption for each year this is proved by using an appropriate data for training and another for testing. As a final step in this work, ANN algorithm was

used to forecast the pattern consumption of each hour separately.As seen in the results the linear regression proved to be more efficient and that is due to the linearity of the data. With each yearwith each year the electrical consumption increases making its progression more like straight line rather than an irregular function(ANN) more predictable in a certain way.Right now, the results are promising but can be improved if data is much more important.

Chapitre4:

Conclusion

And

Perspectives

Conclusion and perspectives:

Conclusion: Forecasting is still an uncharted field, optimize current models, suggest different approaches, and add more variables to the equation.In the future the economy of every country will rely on the efficiency of its forecasts like Paul Saffo once said:" the goal of forecasting is not to predict the future but to tell you what you need to know to take meaningful action in the present".
The dataset containing electricity consumption data from the years 2008 to 2019, along with the first two months of 2020, was provided by Sonelgaz. The data has undergone normalization, treatments, and clustering into different groups using the K-means algorithm. Subsequently, linear regression was applied to the dataset, resulting in a high accuracy rate of 98% when predicting the maximum consumption for the year.

On the other hand, when using the artificial neural network algorithm, the error rate ranged between 2.9% and 3.7%, sometimes even higher. This further supports the idea that the data exhibits a linear relationship.

Perspectives:the results formed in this work seems to be acceptable but it doesn't reach the accuracy that we are looking for so as a perspective In order to improve the forecasting, hybrid program will be combined between the three methods previously discussed (kmeans, linear regression, ANN), include deep learning . Other considerations should be highlighted such as Covid period feast days, Ramadan…. Etc . Other parameters could be included such as temperature correlation, technology evolution (smart devices, electric vehicles,….),dry and wet weather.

Bibliography:

[1]. Abdulnassar, Latha R. Nair, Performance analysis of Kmeans with modified initial centroid selection algorithms and developed Kmeans9+ model, Measurement: Sensors, 14 January 2023

[2]. Zexian Sun, Mingyu Zhao, Guohong Zhao, Hybrid model based on VMD decomposition, clustering analysis, long short memory network, ensemble learning and error complementation for short-term wind speed forecasting assisted by Flink platform, Energy, 26 August 2022

[3]. Junzeng He,Dong Jiang,Qingguo FeiInterval, model validation for rotor support system using Kmeans Bayesian method,Probabilistic Engineering Mechanics, 15 September 2022...

[4]. Rasyidah,Riswan Efendi,S. M. Aqil Burney, Cleansing of inconsistent sample in linear regression model based on rough sets theory, Systems and Soft Computing, 13 December 2022...

[5]. Hai Zhang,Huang Qin,Weidong Fan,Revealing the influence of oxygen-containing functional groups on mercury adsorption via density functional theory and multiple linear regression analysis,Fuel7 December 2022...

[6]. Sijia Huang,Linear regression analysis, International Encyclopedia of Education (Fourth Edition)18 November 2022...

[7]. Claudia Borredon,Luis A. Miccio,Gustavo A. Schwartz, Estimating glass transition temperature and related dynamics of molecular glass formers combining artificial neural networks and disordered systems theory, Journal of Non-Crystalline Solids: X17 June 2022...

[8]. Alexander E. Mayer, Vasiliy S. Krasnikov, Victor V. Pogorelko, Homogeneous nucleation of dislocations in copper: Theory and

approximate description based on molecular dynamics and artificial neural networks, Computational Materials Science, 15 February 2022...

[9]. Zengrui Yuan, Mu-Qing Niu, Li-Qun Chen, Predicting mechanical behaviors of rubber materials with artificial neural networks, International Journal of Mechanical Sciences, 28 February 2023...

[10]. Vincenzo Bianco, Oronzio Manca, Sergio Nardini ,Electricity consumption forcasting Italy using linear regression models. Septemner 2021

[11]. Supervised vs unsupervised learning, seldom, September 16 2022

Key Words: Forcasting, electricity consumption, k-means, linear regression, machine learning, deep learning, artificial neural networks.