**PJM Hourly Energy Consumption Analysis using Machine Learning Report**

[Over 10 years of hourly energy consumption data from PJM in Megawatts]

**Names:**

Sagnik Chatterjee [19BAI1153]

Sarvesh Sridhar [19BAI1057]

Shubhadip Chakrabarti [19BAI1113]

**Emails:**

[sagnik.chatterjee2019@vitstudent.ac.in](mailto:sagnik.chatterjee2019@vitstudent.ac.in)

[sarvesh.sridhar2019@vitstudent.ac.in](mailto:sarvesh.sridhar2019@vitstudent.ac.in)

[subhadip.chakrabarti2019@vitstudent.ac.in](mailto:subhadip.chakrabarti2019@vitstudent.ac.in)

**Phone numbers:**

8411945801

9361036826

9840379862

**Abstract**

Our project essentially revolves around analysing different patterns and ongoing trends that the energy consumption has on hourly basis, day basis, season basis and few other things. The project is aimed to contribute to the concept of sustainable development. In this project we try to develop a machine learning model that can help the user learn the possible amount of energy consumption in the future. This project is initiated by identifying an appropriate dataset having necessary field information. The project work is then carried out by creating some hypothesis, and then determining if the hypothesis is right or wrong using graphical visualizations and classifications. The project then moves on by analysing the dataset. After performing necessary pre-processing on the dataset, the instances are then passed into the two models namely “Linear Regression” and “Random Forest Regressor”. The project then evaluates the performance of each model designed in order to understand the usability of the model created. Comparative Analysis is performed on the models designed to successfully determine the better model for this project.

**Keywords:** Energy Consumption Data, Hypotheses, Preprocessing, Extracting new attributes, Linear Regression, Random Forest Regressor, Error Analysis

**CHAPTER 1**

**INTRODUCTION**

The world today witnesses an era, where energy has emerged as a topic of prime interest. The human population in the present age of developing world witnesses an ever-increasing demand of energy resources. This demand for energy resources is expected to grow exponentially in the upcoming generations too. Electrical Energy, a vital energy-based resource is an example of resource having a huge demand in the present human society.

The electrical energy is generated in a thermal power plant mainly using heat engines fuelled primarily by combustion of fossil fuels. In this context, it is worth noting that the fossil fuels, being conventional energy sources, are found in limited amounts in the world.  The increasing demand and utilization of fossil fuel sources such as coal, petroleum, natural gas has posed a major challenge to the sustainability of the human civilization. This naturally leads to identifying a sustainable way of utilizing the energy resource. Although the search for alternate sources of energy sources are being carried out actively around the world but there is an urgent need to keep track on the energy utilization in the developing world, so that appropriate steps can be taken to ensure consumption of energy in a sustainable way.

In this project we aim to contribute to the sustainable development concept, rational utilization of electrical energy. In this project we try to develop a machine learning model that can help the user learn the possible amount of energy consumption in the future. We also try to explore the times when the consumption of electrical energy has been the maximum and minimum for better analysis of the energy consumption.

To successfully implement the project, we would need a reliable dataset that can used to predict the energy consumption successfully. The dataset used in this project is obtained from Kaggle. The dataset consists of data records from 12 different electricity generating companies. The dataset on further exploration was found to have features that were enough to meet the basic objectives of the project.

The machine learning model built up in this project can be used to successfully help the user of the model learn the amount of electrical energy consumption in the future. This data can facilitate the target audience to take measures to utilize the electrical energy in a sustainable manner. The model can also help the energy generating companies to understand the estimated surge in demand of electrical energy in the future. This would enable the electrical energy generating companies to take necessary steps to upgrade their equipment’s to match the expected user demands in the future. This model can also support the government in taking effective steps in the field of sustainable energy.

The document hereafter moves on with the literature study carried out for having a better understanding about the project domain. It is followed by a detailed explanation on the methodology followed to successfully implement the project. It also includes the designs of the engineering techniques used in this project. In the later section, a detailed explanation on the implementation technique is scripted. Results obtained as a result of the approved methodology are then explained in the document. This is followed by the Conclusion to the project.

**CHAPTER 2**

**BACKGORUND STUDY**

We first decided about working on a disease predicting model which would predict a disease based on the symptoms, given the current situation it is one the ideas that comes. But then we realized that after getting our dataset that there is not much to work on it, if we want to expand, we would have to move to image detection and identification and then predicting a disease which would be quite complex, so we were just thinking about what to do, then it hit us that why not work on something that was powering the thing we were working on, yeah electricity.

So we saw many articles related to electricity and power, related to energy consumption and we decided to work on this topic, to analyze different patterns and ongoing trends that the energy consumption has on hourly basis, day basis, season basis and few other things. We also created some hypotheses and then proved them whether they right or wrong based on graphs and classifications. Luckily, we found a group of datasets consisting data of a particular company’s energy consumption in different times (date, time).

**Some Research Papers references:**

**Research Paper-1:**

**Household Energy Consumption Segmentation Using Hourly Data**

**Authors:**

***Jungsuk Kwac, Student Member, IEEE, June Flora, and Ram Rajagopal, Member, IEEE***

**About-**The increasing US deployment of residential advanced metering infrastructure (AMI) has made hourly energy consumption data widely available. Using CA smart meter data, the paper investigates a household electricity segmentation methodology that uses an encoding system with a pre-processed load shape dictionary. Structured approaches using features derived from the encoded data drive five sample program and policy relevant energy lifestyle segmentation strategies. They also make sure that the methodologies developed scale to large data sets.

**Index Terms**—Clustering, demand response, segmentation, smart meter data, variability.

**Research Paper-2:**

**Prediction of hourly energy consumption in buildings based on a feedback artificial neural network**

**Authors:**

**Pedro A. Gonza´lez, Jesu´s M. Zamarren˜o\***

**Received 12 January 2004; received in revised form 10 August 2004; accepted 15 September 2004**

**About—**In this paper a replacement approach for short-term load prediction in buildings is shown. the tactic is predicated on a special quite artificial neural network (ANN), which feeds back a neighbourhood of its outputs. This ANN is trained by means of a hybrid algorithm. The new system uses current and forecasted values of temperature, the present load and therefore the reform the hour and the day as inputs. The performance of this predictor was evaluated using real data and results from international contests. The achieved results demonstrate the high precision reached with this technique.

**Keywords—** Intelligent buildings; Energy demands and consumption; Neural networks; Load forecasting; Optimization method

**Research Paper-3:**

**Day-ahead prediction of hourly subentry energy consumption in the building sector using pattern recognition algorithms**

**About:** The accurate day-ahead prediction of subentry electric energy consumption (SEEC) may be a critical basis for elaborative building energy management. However, most of the present studies mainly specialise in modeling overall energy consumption without distinguishing its patterns of various temporal features. At an equivalent time, advances in metering technologies and machine learning methods provide new opportunities for detailed predictions. during this paper, a day-ahead prediction model supported the improved recognized patterns via fuzzy C-means clustering and nonlinear regression is proposed and discussed. The proposed indirect pattern recognition is administered by taking advantage of the connotative incidence relation between fluctuation features and influencing factors. Considering the various temporal characteristics of hourly SEEC, this proposed model is applied in an office block with the scope to manage the day-ahead prediction of hourly HVAC subentry and hourly socket subentry. These are taken because the typical non-stationary sequence and typical stationary sequence respectively. Results show that the proposed pattern recognition is applicable for the non-stationary HVAC subentry, and a stable energy pattern can contribute to accurate predictions. Furthermore, the introduction of additional hourly meteorological parameters improves the accuracy via rolling prediction instead. Finally, the modelling adaptability and applicable implications are summarized for references of optimal building energy operation.

**Research Paper-4:**

**The daily and hourly energy consumption and load forecasting using artificial neural network method: a case study using a set of 93 households in Portugal**

**Authors:**

**Filipe Rodrigues , Carlos Cardeira, J.M.F.Calado**

**About:** It is important to know and forecast a typical or a very household daily consumption so as to style and size suitable renewable energy systems and energy storage. during this research for brief Term Load Forecasting (STLF) it's been used Artificial Neural Networks (ANN) and, despite the consumption unpredictability, it's been shown the likelihood to forecast the electricity consumption of a household with certainty. The ANNs are recognized to be a possible methodology for modelling hourly and daily energy consumption and cargo forecasting. Input variables like apartment area, numbers of occupants, electrical appliance consumption and Boolean inputs as hourly meter system were considered. Furthermore, the investigation administered aims to define an ANN architecture and a training algorithm so as to realize a strong model to be utilized in forecasting energy consumption during a typical household. it had been observed that a feed-forward ANN and therefore the Levenberg-Marquardt algorithm provided an honest performance. For this research it had been used a database with consumption records, logged in 93 real households, in Lisbon, Portugal, between February 2000 and July 2001, including both weekdays and weekend. The results show that the ANN approach provides a reliable model for forecasting household electric energy consumption and cargo profile. Selection and peer-review under responsibility of KES International

**Keywords:** Artificial Neural Networks; Levenberg-Marquardt; Energy forecasting; Hourly and daily energy; Boolean application

**CHAPTER 3**

**PROPOSED METHODOLOGY**

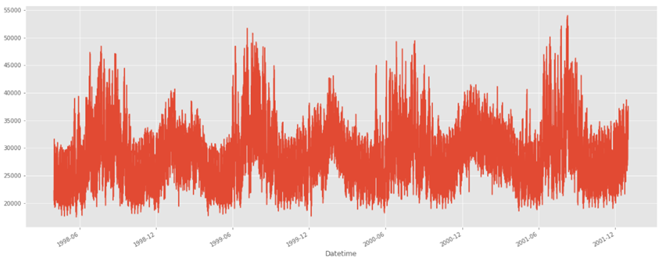
As we discussed earlier our project focuses on hourly energy consumption of 12 different companies based on the data they have provided to their respective datasets. Let’s see the description of our datasets.

**PJM Hourly Energy Consumption Data**

PJM Interconnection LLC (PJM) is a regional transmission organization (RTO) in the United States. It is part of the Eastern Interconnection grid operating an electric transmission system serving all or parts of Delaware, Illinois, Indiana, Kentucky, Maryland, Michigan, New Jersey, North Carolina, Ohio, Pennsylvania, Tennessee, Virginia, West Virginia, and the District of Columbia.

The hourly power consumption data comes from PJM's website and are in megawatts (MW).

The regions have changed over the years so data may only appear for certain dates per region.



Ideas of what we could do with this dataset:

* Split the last year into a test set- can you build a model to predict energy consumption?
* Find trends in energy consumption around hours of the day, holidays, or long-term trends?
* Understand how daily trends change depending of the time of year. Summer trends are very different than winter trends.

**Datasets that we used:**

1. AEP\_hourly.csv:

* American Electric Power(AEP)
* estimated energy consumption in Megawatts (MW)

1. COMED\_hourly.csv:

* Commonwealth Edison (COMED)
* estimated energy consumption in Megawatts (MW)

1. DAYTON\_hourly.csv:

* The Dayton Power and Light Company
* Estimated energy consumption in Megawatts (MW)

1. DEOK\_hourly.csv:

* Duke Energy Ohio/Kentucky (DEOK)
* Estimated energy consumption in Megawatts (MW)

1. DOM\_hourly.csv:

* Dominion Virginia Power (DOM)
* Estimated energy consumption in Megawatts (MW)

1. DUQ\_hourly.csv:

* Duquesne Light Co. (DUQ)
* Estimated energy consumption in Megawatts (MW)

1. EKPC\_hourly.csv:

* East Kentucky Power Cooperative (EKPC)
* Estimated energy consumption in Megawatts (MW)

1. FE\_hourly.csv:

* First Energy (FE)
* Estimated energy consumption in Megawatts (MW)

1. NI\_hourly.csv:

* Northern Illinois Hub (NI)
* Estimated energy consumption in Megawatts (MW)

1. PJME\_hourly.csv:

* PJM East Region: 2001-2018 (PJME)
* Estimated energy consumption in Megawatts (MW)

1. PJMW\_hourly.csv:

* PJM West Region
* Estimated energy consumption in megawatts (MW)

1. PJM\_Load\_hourly.csv:

* PJM Load Combined: 1998-2001
* Estimated energy consumption in Megawatts (MW)

**Let’s see how our methodology is compared to some exiting methodology:**

We’ll see a case study that was conducted in Malaysia in March 2021 to have a better understanding of the methodology.

**Energy consumption prediction by using machine learning for smart building: Case study in Malaysia**

Mel Keytingan M.Shapi; Nor Azuana Ramlib; Lilik J.Awalin

**Some Highlights**

•Cloud based prediction model development does not depend on the performance of the hardware it’s running on.

•This analysis would provide an insight on how reliable and capable AzureML studio for developing a prediction model.

•The consequence of the model training and testing shows that each method performed differently in every cases.

**Abstract**

Building Energy Management System (BEMS) has been a considerable topic nowadays thanks to its importance in reducing energy wastage. However, the performance of 1 of BEMS applications which is energy consumption prediction has been stagnant thanks to problems like low prediction accuracy. Thus, this research aims to deal with the issues by developing a predictive model for energy consumption in Microsoft Azure cloud-based machine learning platform. Three methodologies which are Support Vector Machine, Artificial Neural Network, and k-Nearest Neighbour are proposed for the algorithm of the predictive model. that specialize in real-life application in Malaysia, two tenants from a billboard building are taken as a case study. the info collected is analysed and pre-processed before it's used for model training and testing. The performance of every of the methods is compared supported RMSE, NRMSE, and MAPE metrics. The experimentation shows that every tenant’s energy consumption has different distribution characteristics.

**Keywords**

Building energy management system

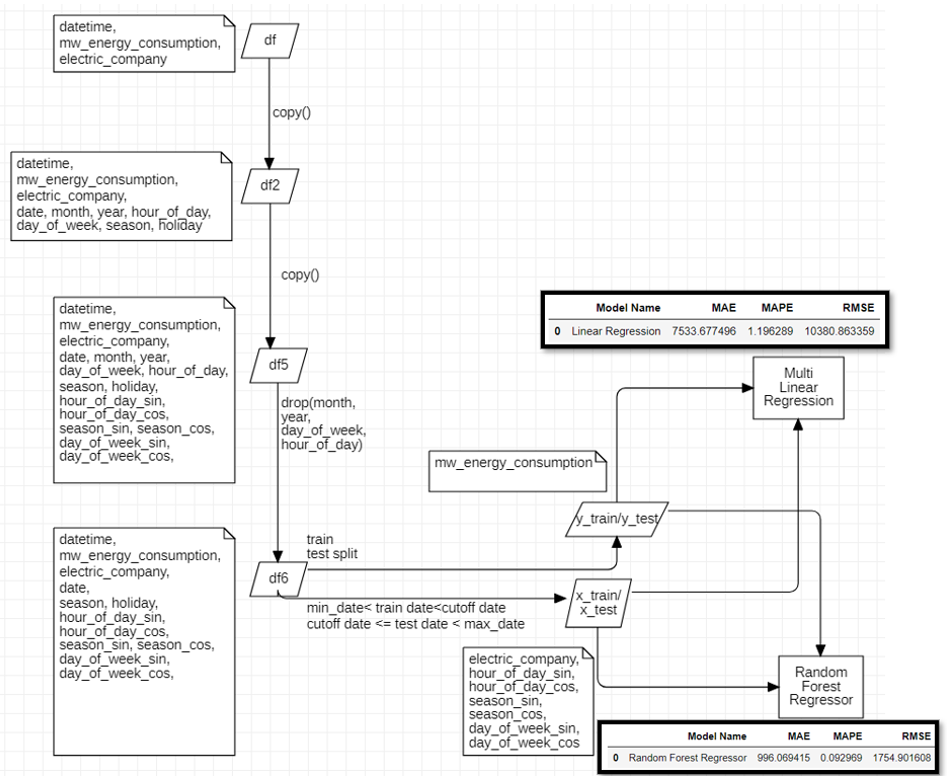
Machine learning

Microsoft Azure

Energy consumption

Prediction

**Design Model:**

****

**Fig 1.** Detailed processing involved in proposed system

**CHAPTER 4**

**IMPLEMENTATION**

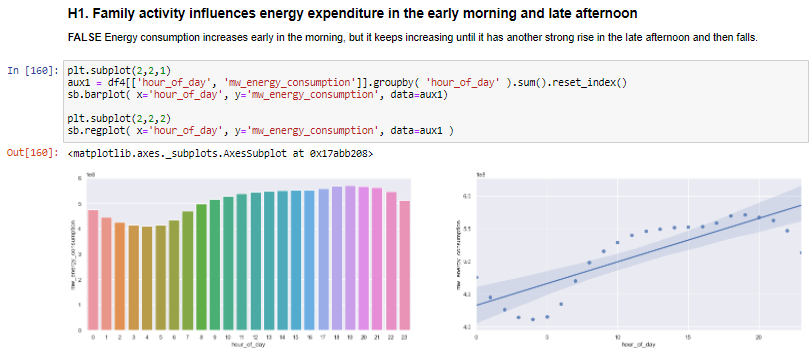
We combine all 12 electric company datasets into one data frame. The features the data frame has are “*datetime”*, *“electric\_company”*,” mw*\_energy\_consumption”.* “*datetime”* feature has data in a condensed format “YYYY-MM-DD HH:MM: SS” and with the interval of one hour. We observe that the total records in the combined data frame (all electric companies) have above the count of 10 million. We then derive new features from the “*datetime”* such as *“date”, ” month”, ” year”,” hour\_of\_day”, ” season”,” day\_of\_week”,” holiday”*. We then observe the features and prove some hypothesis based on observation of the data. In *“hour\_of\_day”, “season”, “day\_of\_week”,* the records are all cyclic (like in “*day\_of\_week*” the data ranges from 0 to 6 then again it comes back to 0 and so it continues). As they are cyclic in nature, we change the nature of its data into “*sine*” and “*cosine*” form. We make the data in “*mw\_energy\_consumption”* into log1p form as it has to be passed into the models and errors while processing maybe large, so we try to process the data and take the inverse (exp1p) in the end. After processing the data, we use “*Boruta”* module to find the most dependent features to the target feature (“*mw\_energy\_consumption”*) and use it to prove our results after comparing them with graphs. We then take cutoff date (last 1 year) and split the data into testing and training. For training data, we have records from the start to the cutoff and for testing data we have from cutoff to the last data of the respective companies. We then pass them into “*Linear Regression*” model and “*Random Forest Regressor”* and do the error analysis for them. We find that “*Random Forest Regressor”* has less error than “*Linear Regression”* model and use it to create graphs for the analysis of the output.

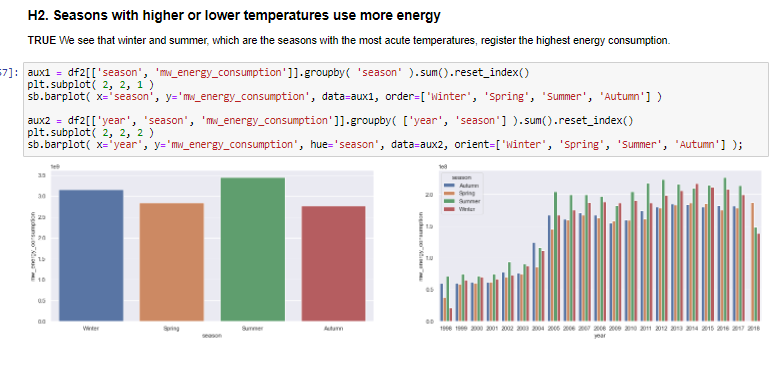
**Fig 1.** Detailed processing involved in proposed system

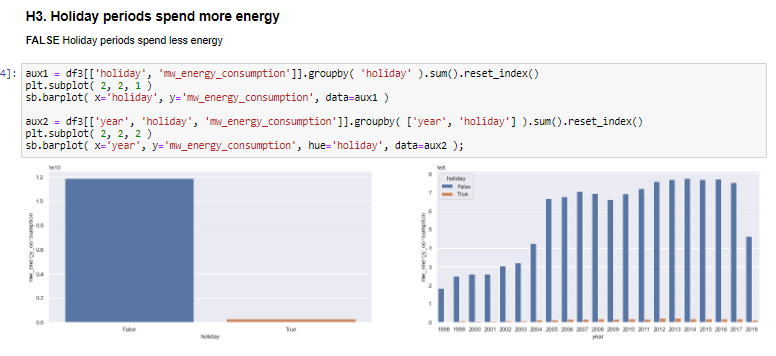
**CHAPTER 5**

**RESULTS AND DISCUSSION**

**The Hypotheses results that we got from our model:** [We assumed some statements related to our domain and then proved them with our data]

****

****

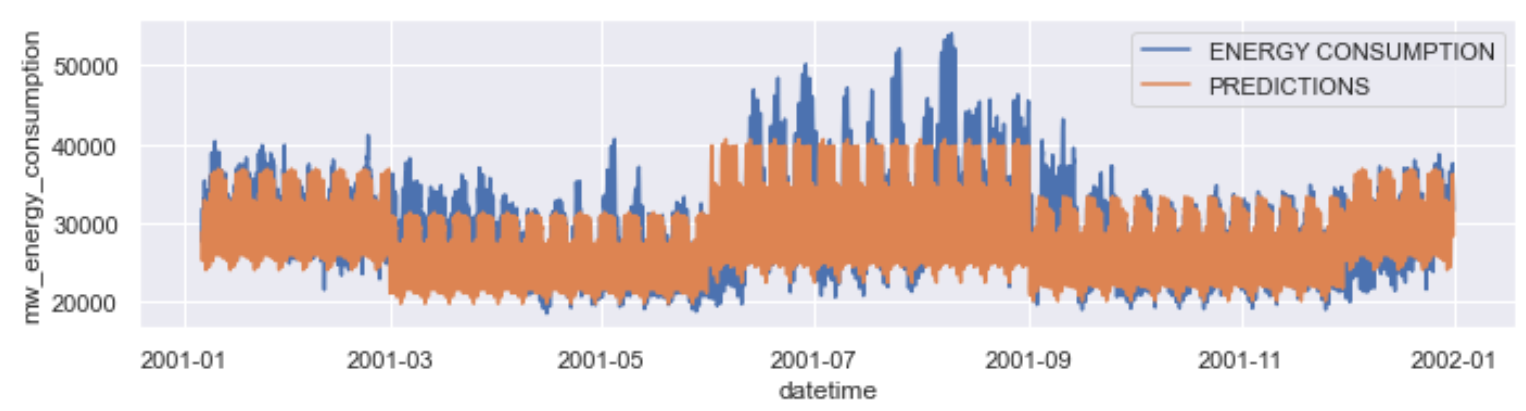
****

****

**Now the results for our models:**

After implementing “*Linear Regression*” and “*Random Forest Regressor”* model, we try calculating metrics for them namely, Most Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE). We observe that although the model is not tuned, we can see big difference while calculating the errors. For example, if we try to compare the RMSE of both models and compare, we see that “*Random Forest Regressor”* has very less value than *“Linear Regression*”. Hence, we can take the “*Random Forest Regressor*” and tune them to make the model more efficient.

We plot the predicted energy consumption which we get from regressor model and try to compare it with all companies and total consumption of all companies. From the graphs plotted, we could observe that predicted results mostly lies as inliers doesn’t shoot out of original data. We could see many overlapping of original and predicted data and very few disconnected points in it.



**Fig.2.** Time series graph between Prediction-Original energy\_consumption

Outcome of this particular program will predict the optimized amount of consumption of power data from a set of power grids located within a region of a geographical/province/country. This result as an output can be used as a statistical optimal demand required to generate power by the power generating companies.

Because the result is based on machine learning model, the result would be very near to the exact demand thereby providing better visibility for the power generating companies to plan for their demand.

**CHAPTER 6**

**CONCLUSION AND FUTURE WORKS**

In this course of project work, we have developed a Machine Learning Model that can help the users learn the possible amount of energy consumption in the future. We have also explored the times when the consumption of electrical energy has been the maximum and minimum for better analysis of the energy consumption. The model is designed to facilitate the users with the information on the prediction of possible energy consumption so that the target departments can take steps accordingly to meet the ambition of having a sustainable energy.

As a way forward, the machine leaning model developed can now be designed to have an application interface so that the users can access the Machine Learning model easily through a friendly user interface. Deployment of this application interface would enable the users to easily have an access to the machine learning model.

Architecture for this API can have the following components in it

Handler API - is the part that receives the requests and plays for the other parts so that the data is processed and then brings everything together, returning the final answer.

Data Preparation - all the treatments and modifications we made to the data will be kept inside. When the Handler receives the raw data it will throw it here within this list of treatment codes so that they are prepared so that they can be ready to be used within the Machine Learning model.

Model Training - this is our trained model that has been saved and will be placed inside this folder in our production architecture. The Handler will take the data processed within Data Preparation and play it inside the model so that it provides the prediction.

**CHAPTER 7**

**REFERENCES**

1) <https://ieeexplore.ieee.org/abstract/document/6693793>

2) <https://www.sciencedirect.com/science/article/abs/pii/S0378778804003032>

3) <https://www.sciencedirect.com/science/article/abs/pii/S0360544220316388>

4) <https://www.sciencedirect.com/science/article/pii/S1876610214034146>

5) Datasets:

<https://www.kaggle.com/robikscube/hourly-energy-consumption>

6) <https://www.eia.gov/todayinenergy/detail.php?id=42915>

7) <https://towardsdatascience.com/part-1-time-series-analysis-predicting-hourly-energy-consumption-of-san-diego-short-term-long-3a1dd1a589c9>