MACHINE LEARNING LAB (CSE 336L)

Predicting Electricity Consumption

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology/Master of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Under the Guidance of

Dr. Neeraj Kumar Sharma

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

05,2024

CERTIFICATE

10-06-2024

This is to certify that the work present in this Project entitled "**Predicting Electricity Consumption**" has been carried out by "D **Sai Dhanush**" under Dr.Neeraj Kumar Sharmasupervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology in the School of Engineering and Sciences.

Supervisor

(Signature)

Dr. Neeraj Kumar Sharma

Assistant professor, CSE Department

SRM UNIVERSITY, AP

Abstract:

Accurate estimation of power usage is necessary for efficient energy management, supporting utilities and legislators in maximizing resource distribution and guaranteeing system stability. The dynamic nature of consumption patterns, which are impacted by variables including population increase, economic activity, and seasonal variations, makes this work difficult.

In this study, we suggest a machine learning-based method for estimating Finland's electricity usage. We employ a K-nearest neighbors (KNN) regression approach in conjunction with principal component analysis (PCA) to reduce dimensionality by using past consumption data. The difficulties in predicting consumption are addressed by this method, which also manages noisy or missing data and handles big, diverse datasets. Non-linear correlations are also captured.

The KNN regression technique was selected due to its ease of use and ability to identify certain patterns within the data. By taking into account nearby data points inside the feature space, KNN regression offers a versatile and comprehensible model for precisely forecasting patterns of consumption.

The dataset that was utilized includes historical data on Finland's electricity usage, along with seasonal indications, date, time, and related consumption values. In order to manage missing data, normalize features, and extract pertinent time-related features for model training, preprocessing processes are used.

Standard regression metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) coefficient are used to assess the predictive model's performance. These metrics provide information about how well the model predicts consumption patterns and how accurate it is.

Our study attempts to create a precise and trustworthy predictive model for Finland's power usage by utilizing machine learning techniques and historical consumption data. A model like this can help stakeholders make well-informed decisions, optimize the distribution of energy, and improve the overall stability and efficiency of the power system.

Literature Survey:

Author: Fazli Wahid

Review of the literature: Examines several methods for predicting energy usage, such as genetic algorithms, time series, neural networks, regression, and hybrid models.

Reference Link: https://gvpress.com/journals/IJSH/vol10_no2/10.pdf

Methodology: makes use of 520 apartment data sets, predicts using the K-Nearest Neighbor classifier, and assesses performance using a variety of metrics and cross-validation.

Author: Nitesh Kushwaha

Related Work: Investigates energy usage prediction methods including time series forecasting and various machine learning techniques.

Reference Link: https://www.ijfmr.com/papers/2023/6/9988.pdf

Methodology: Proposes a supervised machine learning approach encompassing data preprocessing, model selection, training, evaluation, and deployment, while considering normality testing, missing data handling, and algorithm performance comparison.

Author: Mel Keytingan M. Shapi

Work: Energy consumption prediction by using machine learning for smart building.

Reference link: https://www.sciencedirect.com/science/article/pii/S266616592030034X

Methodology: Three methodologies - Support Vector Machine, Artificial Neural Network, and k-Nearest Neighbour - are proposed for the algorithm of the predictive model for energy consumption in a cloud-based machine learning platform.

Author: Goopyo Hong

Work: The work presents an hourly energy consumption prediction method using the K-Nearest Neighbor (KNN) algorithm for community buildings with different types of buildings.

Reference link: https://www.researchgate.net/publication/364295118_The_Hourly_Energy_Consumption_Prediction_by_KNN_for_Buildings_in_Community_Buildings

Methodology: The methodology involves data preparation of hourly energy consumption and weather data, applying the KNN algorithm to cluster similar energy consumption patterns for each season, averaging the clustered patterns, and evaluating the prediction accuracy using RMSE and CVRMSE metrics.

Author: Ch Poojitha

About the Work: The paper presents a machine learning-based approach for predicting electricity consumption.

Reference link: Electricity Consumption Prediction Using Machine Learning | E3S Web of Conferences (e3s-conferences.org)

Methodology: A power utility provider provides historical data on electricity consumption, which is preprocessed to handle outliers and missing numbers. Using feature engineering, pertinent features are extracted. The preprocessed data is used to train machine learning models, such as artificial neural networks, decision trees, random forests, and linear regression. The top-performing model is chosen using evaluation measures like MAE, RMSE, and R2, and it is then utilized to forecast electricity consumption using the features that were extracted.

Author: Ragupathi Chinnaraji

The work: This work proposes an enhanced long short-term memory (E-LSTM) deep learning model for accurate electricity consumption prediction.

Reference Link: https://ietresearch.onlinelibrary.wiley.com/doi/epdf/10.1049/cmu2.12384

The methodology: In order to address the vanishing gradient problem, the methodology incorporates rectified linear activation into the LSTM model architecture and optimizes hyperparameters. On the basis of training and evaluation on the UCI residential building dataset, the suggested E-LSTM model outperforms existing cutting edge methods.

Author: Zhifeng Lin*

Work: This paper proposes an electricity consumption prediction model based on Long Short-Term Memory (LSTM) with an attention mechanism.

Reference Link: https://onlinelibrary.wiley.com/doi/abs/10.1002/tee.23088

Methodology: The attention mechanism uses cell outputs to anticipate electricity usage, while the LSTM computes weight coefficients for the input sequence data. To reduce prediction errors, weights are improved using Back Propagation Through Time (BPTT).

Author: Alexander Agung Santoso Gunawan

Work: Identifying the dominant features in Indonesia smart home dataset by interpreting electrical energy consumption prediction results.

Reference Link: https://ijisae.org/index.php/IJISAE/article/view/5086/3794

Methodology: K-Nearest Neighbors (KNN) with hyperparameter tweaking for k and distance algorithm selection is used to predict electrical energy usage. SHAP and LIME aid in the interpretation of prediction findings by highlighting prominent traits. Using error metrics like RMSE, MSE, and MAE, the prediction model is validated.

Author: Maha Alanbar

Work: Proposing an energy consumption prediction model using deep learning algorithm (LSTM) for a computer college building.

Reference Link: https://doi.org/10.3991/ijim.v14i10.14383

Methodology: Collecting data on temperature, workdays, number of devices, and historical energy consumption for 13 years, and using LSTM to predict medium-term energy consumption with trial-and-error to determine optimal neural network structure.

Author: Metodija Atanasovski

The work described in the PDF focuses on using the k-nearest neighbor (KNN) machine learning model to estimate electricity demand in North Macedonia's power system and evaluating its effectiveness against sinusoidal and polynomial regression models.

Reference Link: https://trinityh2020.eu/wp-content/uploads/2021/06/K-Nearest-Neighbor-Regression-for-Forecasting-Electricity-Demand-1-1.pdf

The methodology: uses the date and air temperature as independent factors and the KNN model to estimate the electricity load, which is the dependent variable. Data from 2014 to 2018 are used to train the model, while data from 2019 are used to test it. Cross-validation methods and performance indicators such as correlation coefficient, mean absolute error (MAE), and root mean squared error (RMSE) are used to assess the model's efficacy.

Author: Oleg Valgaev

Associated Work: examines short-term load forecasting strategies, focusing on building-level approaches including individual load profiles, neural networks, SVMs, and decision trees.

Reference Link:
https://www.researchgate.net/publication/321202217_Building_power_demand_forecasting_using_K-nearest_neighbours_model_-_practical_application_in_Smart_City_Demo_Aspern_project

Methodology: Forecasts the load curve for each building for the following day without requiring manual setup by combining past daily load curves with an automated cross-validation approach based on K-nearest neighbors.

Proposed Work:

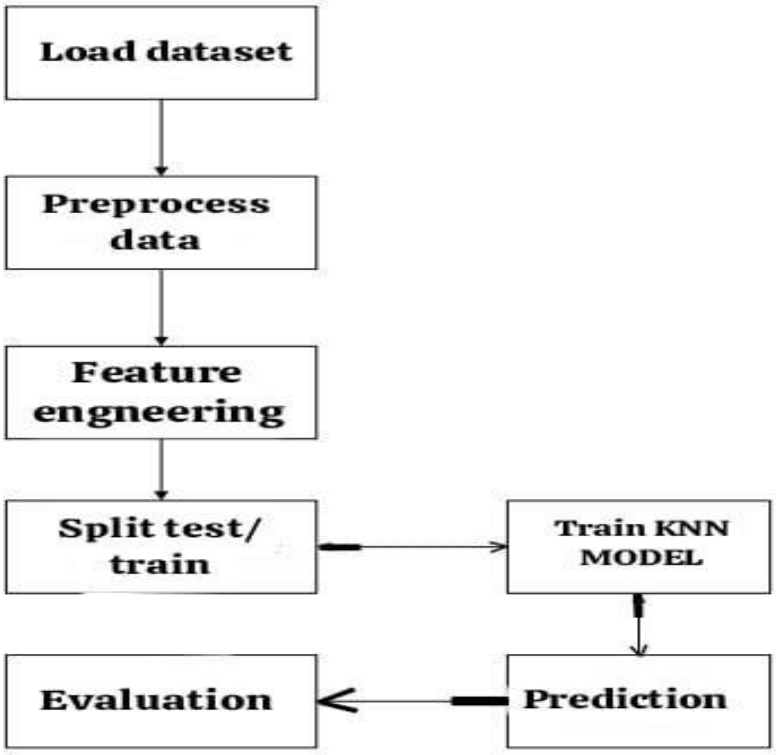

Fig.1 flow chart of data set

Fig 1 shows the data flow Once loaded, the timestamp and consumption data is cleaned and may be enhanced with additional characteristics like month or year. After training on a subset of the data, the model is applied to forecast consumption for fresh data points. The correctness of the model is assessed last.

	Start time UTC	End time UTC	Start time UTC+03:00	End time UTC+03:00	Electricity consumption in Finland
0	2015-12-31 21:00:00	2015-12-31 22:00:00	2016-01-01 00:00:00	2016-01-01 01:00:00	10800.0
1	2015-12-31 22:00:00	2015-12-31 23:00:00	2016-01-01 01:00:00	2016-01-01 02:00:00	10431.0
2	2015-12-31 23:00:00	2016-01-01 00:00:00	2016-01-01 02:00:00	2016-01-01 03:00:00	10005.0
3	2016-01-01 00:00:00	2016-01-01 01:00:00	2016-01-01 03:00:00	2016-01-01 04:00:00	9722.0
4	2016-01-01 01:00:00	2016-01-01 02:00:00	2016-01-01 04:00:00	2016-01-01 05:00:00	9599.0
...	...	...	...	...	...
52961	2021-12-31 16:00:00	2021-12-31 17:00:00	2021-12-31 19:00:00	2021-12-31 20:00:00	11447.0
52962	2021-12-31 17:00:00	2021-12-31 18:00:00	2021-12-31 20:00:00	2021-12-31 21:00:00	11237.0
52963	2021-12-31 18:00:00	2021-12-31 19:00:00	2021-12-31 21:00:00	2021-12-31 22:00:00	10914.0
52964	2021-12-31 19:00:00	2021-12-31 20:00:00	2021-12-31 22:00:00	2021-12-31 23:00:00	10599.0
52965	2021-12-31 20:00:00	2021-12-31 21:00:00	2021-12-31 23:00:00	2022-01-01 00:00:00	10812.0

Fig.02 the first and last 5 data

Fig.02 About the data set:

The dataset provided contains electricity consumption data for Finland. Here's a description of the columns:

- Start time UTC: This column represents the start time of each electricity consumption interval in Coordinated Universal Time (UTC). UTC is a time standard that is commonly used across different time zones.
- End time UTC: This column represents the end time of each electricity consumption interval in UTC
- Start time UTC+03:00: This column represents the start time of each electricity consumption interval adjusted to the UTC+03:00 time zone. This adjustment likely accounts for the time zone difference in Finland, where UTC+03:00 is often used.
- End time UTC+03:00: This column represents the end time of each electricity consumption interval adjusted to the UTC+03:00 time zone.
- Electricity consumption in Finland: This column contains the actual electricity consumption values recorded during each interval. The values are typically measured in kilowatt-hours (kWh) or another appropriate unit of energy.

- The data appears to be collected at hourly intervals, with each row representing one hour of electricity consumption data.
- The dataset covers a period starting from December 31, 2015, and likely continues for a certain duration, possibly up to the present or a specific end date.
- The electricity consumption values vary over time, reflecting fluctuations in energy usage throughout the day and possibly seasonally as well.
- The dataset provides a valuable resource for analyzing and modeling electricity consumption patterns in Finland, which can inform energy management strategies, infrastructure planning, and policy decisions.

KNN Algorithm:

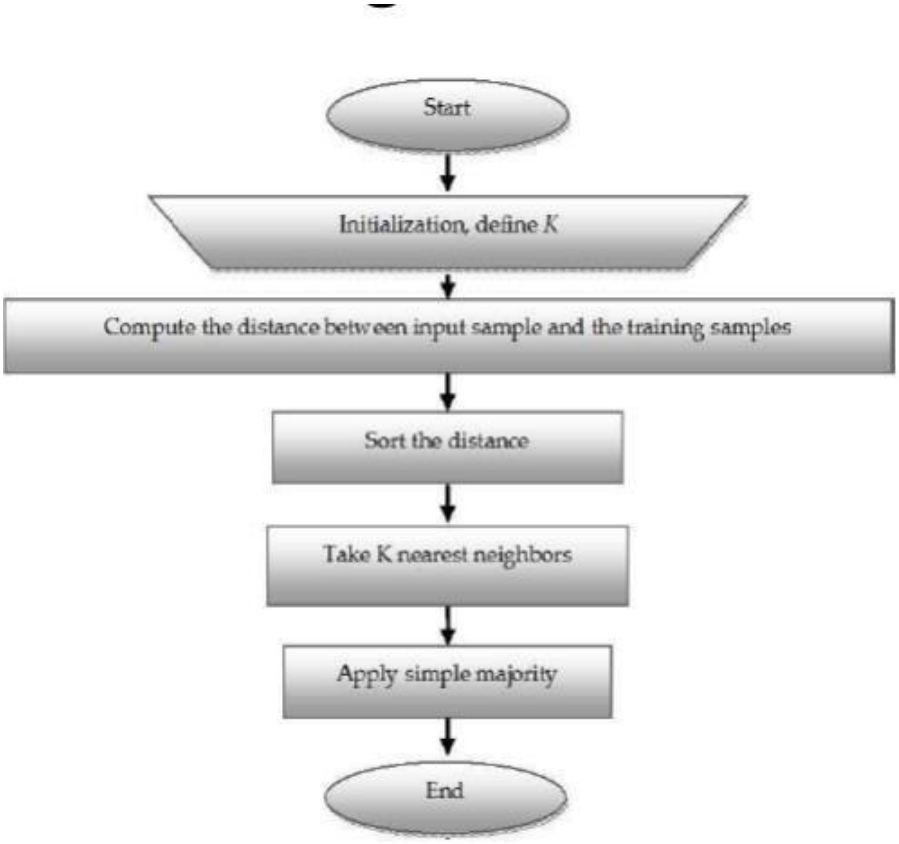

Fig. 03 KNN Algorithm

- Start: The algorithm begins here.
- Set k = number of nearest neighbors: This step determines how many nearby data points to consider when classifying a new data point. The value of k is typically chosen beforehand based on experimentation or domain knowledge.
- Calculate distance between objects, d(x,z): For each new, unseen data point (x), the algorithm calculates its distance to all data points in the training dataset (z). This distance can be measured using various metrics, such as Euclidean distance or Manhattan distance.
- Select class based on majority in k neighborhood: After calculating distances, the algorithm identifies the k nearest neighbors of the new data point (x) from the training dataset. It then predicts the class label for the new data point by selecting the most frequent class among these k neighbors.
- Maximum iteration reached?: This decision point checks if the algorithm has processed all the new data points or if it has reached a predetermined maximum number of iterations. If so, the process ends; otherwise, it repeats from step 3 for the next data point.
- Stop: The algorithm terminates here.

Overall KNN Algorithm:

KNN is a non-parametric and supervised machine learning algorithm.

It predicts the class of a new data point based on the majority class among its nearest neighbors.

The parameter k determines the number of neighbors to consider, and choosing the right value of k is important.

Different distance metrics can be used, such as Euclidean distance, depending on the data and problem at hand.

h

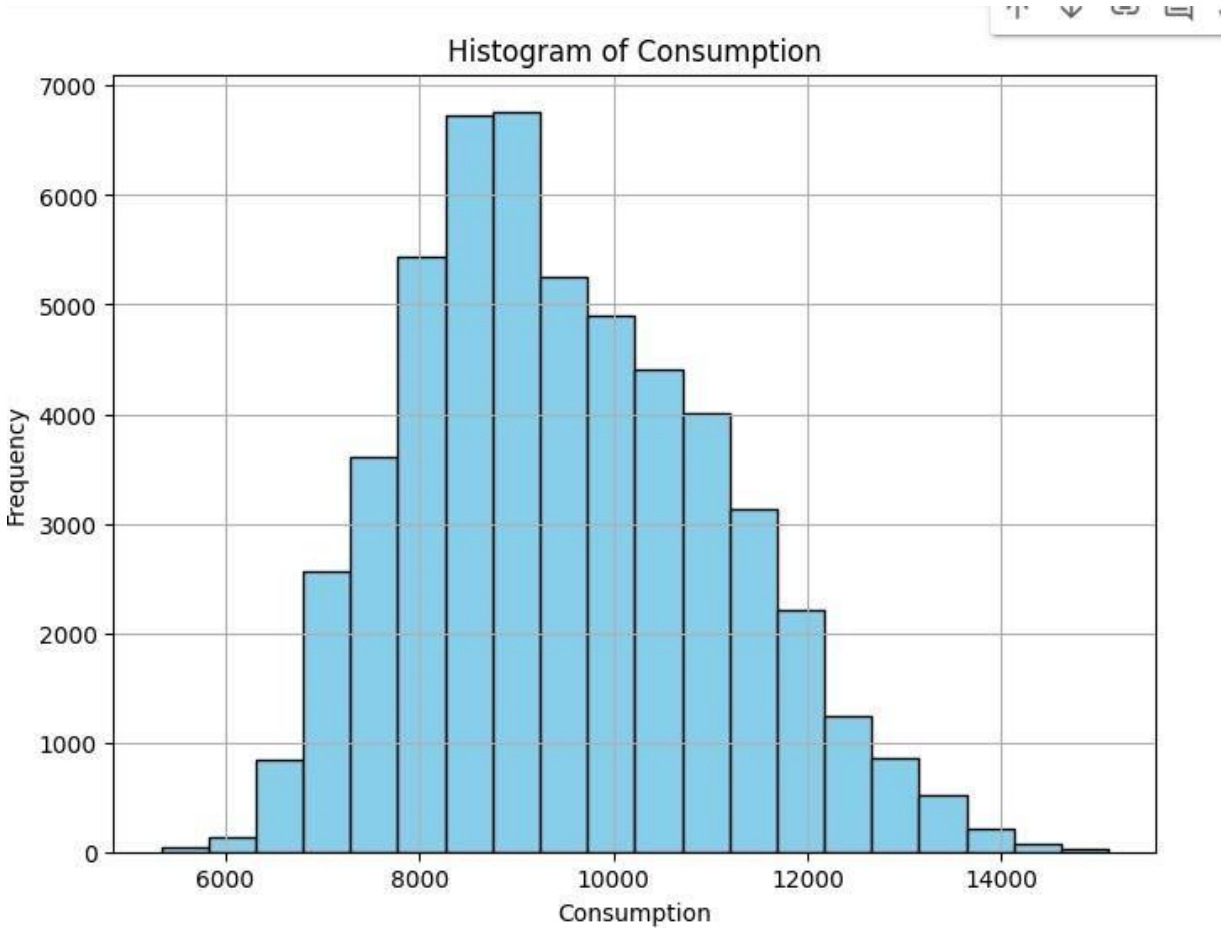

Fig.04

The distribution of consumption in the Finland is shown by the histogram. The horizontal axis displays consumption levels, while the vertical axis displays frequency, or the number of occurrences. The distribution is bell-shaped, with consumption most commonly falling between 8000 and 10,000 units in the middle and frequency falling as consumption rises outside of this range. Specific conclusions regarding consumption patterns are hard to come to without further context, but the histogram gives the distribution of the data a clear visual representation.

Experimental Setup:

Libraries Used:

Pandas: Pandas is an effective library for data analysis and manipulation. It offers data structures including Series and DataFrame, which are widely used for jobs involving data analysis and processing. Pandas is used in your script to read the CSV file into a DataFrame, clean and manipulate the data, and add new features like week, month, and year.

NumPy: NumPy is a core package for Python scientific computing. Mathematical functions, matrices, and arrays are supported. NumPy is utilized throughout your script to do numerical operations, such figuring out distances using the k-nearest neighbors (KNN) technique.

Matplotlib: One popular Python plotting library is Matplotlib. It offers an interface for creating static, interactive, and animated visualizations that is similar to MATLAB. Matplotlib is used to create plots and visualizations in your script, but it appears that you imported it but didn't utilize it directly in the code that was provided.

Seaborn: A Python data visualization library called Seaborn is built on top of Matplotlib. It offers a sophisticated interface for making eye-catching and educational statistical visuals. Seaborn is especially helpful for displaying data contained in DataFrames because it builds upon Matplotlib and connects tightly with Pandas data structures.

Based on Matplotlib, Seaborn is a Python data visualization package. It offers a sophisticated drawing tool for creating eye-catching and educational statistical visuals. Although you imported Seaborn into your script, it doesn't appear like you used it specifically.

scikit-learn (sklearn): Scikit-learn is a Python machine learning package that offers easy-to-use and effective tools for data analysis and mining. It includes tools for model selection and evaluation along with a variety of methods for classification, regression, and clustering. Principal Component Analysis (PCA), and k-nearest neighbors (KNN) regression.

Languages Used: **Python**

Hardware Requirements:

CPU: AMD Ryzen series or the Intel Core i5/i7are appropriated.
RAM, or memory: Ideally, you should have at least 8 GB of RAM, but larger datasets or complicated models may require more.
Storage: SSDs are the recommended option for quicker model training and data access.

Software Requirements:

Python: Python 3.x is the recommended programming language for creating heart disease prediction models since it is used in the majority of machine learning tools and frameworks.

Machine Learning Libraries: To create and train predictive models, make use of well-known machine learning frameworks and libraries.

Integrated Development Environment (IDE): For coding and experimentation, select an IDE or text editor. Jupyter Notebook, PyCharm, Visual Studio Code, and Spyder are used.

Operating System: Windows, macOS, Linux, and other popular operating systems are compatible with the majority of machine learning tools and libraries which we used.

Google Collaboration:

Google cloud-based technology that allows Python code to be executed in a web browser.
allows free computation access to CPUs, GPUs, and TPUs.
integrates with Google Drive to facilitate collaboration and data management.
provides an interactive scripting and visualization experience using a Jupyter Notebook.
has Python libraries installed out of the box and allows you to install more packages.
allows for notebook sharing and real-time collaboration.
offers a permanent runtime environment so that work can be picked up again between sessions.

Result:

Based on the given features, the K-Nearest Neighbors (KNN) regression model's output predicts the amount of power used. The primary components that result from applying PCA to the original features—month, year, and week—are the features that were utilized for prediction in this instance. An estimate of the amount of electricity used for the specified set of features is shown by the expected consumption value. Nonetheless, it's crucial to interpret this conclusion cautiously and take into account a number of factors:

1. Accuracy: A number of parameters, such as the characteristics selected, the number of neighbors (k) selected for the KNN method, and the quality of the dataset, affect how accurate the prediction is.
2. Feature Selection: To lower the dimensionality of the feature space, PCA was applied. PCA may cause information loss even if it can aid in identifying the most important variances in the data. Making ensuring the chosen characteristics accurately depict the underlying trends in the data is crucial.
3. Model Performance: The Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE) are suitable metrics to assess the KNN model's performance. These measures shed light on how well the model applies to previously untested data.
4. Hyperparameter Tuning: The model's performance can be greatly affected by the number of neighbors (k) selected in the KNN method. To get the greatest results, it is imperative to adjust this hyperparameter.
5. Interpretability: The relationship between the input characteristics and the target variable is not automatically revealed by KNN regression. Understanding how each feature contributes to the prediction and learning more about the behavior of the model can be accomplished by utilizing interpretability techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations).

   In conclusion, additional research and assessment are required to determine the robustness and dependability of the prediction model, even though the anticipated consumption number offers an estimate of the amount of power used. This entails assessing the effectiveness of the model, determining the significance of the features, and maybe improving the model through feature engineering and hyperparameter tuning.

Conclusion and Future work:

With this initiative, a thorough analysis of Finland's power consumption trends was initiated. The main machine learning method was K-nearest neighbors (KNN) regression, but the process started with the careful creation of a reliable dataset. A large range of timestamps and associated electricity usage values were recorded in this dataset. This data required a multi-step curation process. First, the collection of data made sure that Finland's patterns of electricity demand were fully represented. Then, using Python libraries, painstaking preprocessing methods were used. During this phase, any anomalies or inconsistencies that would impair the performance of the model were removed from the data. Furthermore, feature engineering was quite important. Key temporal components, such as week, year, and month, were extracted, enriching the dataset to capture the complex temporal trends inherent in power usage. actions. The algorithm was able to detect daily and seasonal fluctuations in addition to overall consumption levels because to these enhanced characteristics. After this thorough preprocessing phase, the study moved on to its main focus: employing KNN regression to model electricity consumption. The foundation for training the model on the preprocessed dataset was given by Scikit-learn's K-Neighbors Regressor. KNN regression is an effective method for finding patterns in similar historical data points because of its proximity-based learning methodology. The model was skilled at identifying complex consumption patterns and projecting them into the future by utilizing this power. The performance of the trained model was clearly promising, indicating that it can accurately forecast future levels of power use. This capacity to predict turned out to be quite useful, providing insight into the temporal dynamics. and the inherent seasonality of Finland's electricity usage pattern. Equipped with these model-derived insights, stakeholders were able to comprehend variations in consumption to a great extent. This gave them the ability to strengthen energy distribution plans, optimize resource allocation tactics, and proactively predict demand surges. Thus, the study provided a powerful example of how machine learning may be used to find hidden patterns in intricate temporal datasets. It emphasizes how much potential this technology has to completely transform energy management techniques not just in Finland but all around the world.

To improve predicted accuracy, future research will examine other regression algorithms such as decision trees, ensemble techniques, and linear regression. Models could be enhanced and greater insights could be obtained by integrating other data sources, such as socioeconomic indicators and meteorological data. By utilizing interactive visualization tools in conjunction with predictive models deployed in real-world settings, stakeholders would be better equipped to make educated decisions and effect significant change in energy management practices. Furthermore, carrying out a more thorough temporal analysis and improving model explainability methods would enhance the project's impact and practical usability by improving the predictability and accuracy of results.