Project Report

Title: Energy Consumption Prediction

Overview

In a world where using energy wisely is really important, predicting how much energy will be needed is key. That's what the "Energy Consumption Prediction" project is all about. Machine learning models like feed-forward Artificial Neural Networks (ANN) have been used to predict how much electricity will be used based on the weather.

Knowing how much energy will be needed is super important for businesses and homes alike. It helps in planning ahead, using resources wisely, and making smart decisions. By focusing on the Alto Paraná region in Paraguay, understanding how factors like weather impact energy use in that specific area becomes possible. A lot of reading has been done about different ways to predict energy use, like using past data or looking at similar places nearby. This project combines the best of these methods to make really good predictions. Using a big set of data from the region, covering four years from 2017 to 2020 of electricity use and weather conditions ensures the data is clean and ready for analysis.

The plan is to build and train a feed-forward ANN model using the data. The ANN model will learn from the patterns in the data and make predictions about future energy use. Checking how well these models work using simple measures like how close they are to the real numbers will be essential. The project will produce a useful tool: a trained model that can predict energy use accurately.

By doing all this, this project aims to make an efficient model that will accurately predict future data and help to make smarter decisions about energy use, ultimately helping them save resources and money.

Background Studies:

Hong et al. tried to predict how much energy community buildings would use each hour using a method called K-nearest neighbor (KNN). KNN is chosen because it's fast [1], but it's not great at

forecasting future values. Instead, it's good at sorting data into categories. To make KNN better at predicting future values, they suggest adding information about time. [12]

Rodrigues et al. explore the application of Artificial Neural Networks (ANN) for Short-Term Load Forecasting (STLF) of daily and hourly energy consumption in households. The study uses a dataset from 93 real households in Portugal and focuses on key variables such as apartment area, occupants, appliance consumption, and a Boolean hourly meter system. The use of a multilayer perceptron with an ANN feed-forward structure allows the model to capture complex relationships and patterns in the data. Adding a hidden layer in a neural network allows it to understand complex patterns, which is useful for predicting things. Though it's generally believed that one hidden layer can handle most situations, some datasets might need more layers for accurate predictions. [2]

Pîrjan et al. tackle the important task of accurately predicting the hourly energy usage of large businesses, which is crucial for promoting energy efficiency. They introduce dynamic neural networks that are based on non-linear autoregressive (NAR) and non-linear autoregressive with exogenous inputs (NARX) models. Using these models helps capture complex relationships in the data and could enhance the accuracy of energy consumption forecasts. However, it's worth noting that while the model assumes there are non-linear connections within the energy usage patterns, this might not always be the case for every type of energy consumption behavior. [3]

Newsham et al. evaluated an ARIMAX model, which is essentially an ARIMA that includes outside factors that influence the random changes in data. The aim was to predict the electricity consumption of a non-traditional three-story building with a floor area of 5800 square meters, housing laboratories and 81 individual workspaces. The predictors in their model included historical consumption data, weather data, and occupancy data obtained through sensor monitoring of logins. The addition of login data marginally enhanced the model's accuracy. However, the authors suggest that including logoff data or utilizing more advanced sensors, such as cameras, could further improve accuracy in a conventional office building setting. [4]

Different ways of using machine learning have been studied, and Tso et al. looked at decision trees in particular. Decision trees are good because they're easy to understand, unlike Artificial Neural Networks and Support Vector Machines. [5]

However, a promising idea could be to use a mix of different methods like the hybrid method, known as a hybrid approach. This could be even better at predicting things than just using a Support Vector Machine or an Artificial Neural Network by themselves, as demonstrated in the reference. [7]

As per the research conducted by Xiong and Yao, proposed supervised classification techniques include Artificial Neural Networks (ANN), Support Vector Machine (SVM), K-nearest neighbor (KNN), and others [8]. Among these methods, the Artificial Neural Network (ANN) is notable as a regression technique. It's popular because it works similar to the human brain, with interconnected units. But, it needs lots of settings to work well and requires large amounts of data for training. [9]. On the other hand, the SVM algorithm, widely utilized for pattern classification, calculates the linear regression function [10, 11].

Dataset:

The dataset that I will use for my project is **"Electric current consumption and meteorological data of Alto Paraná, Paraguay"**

Dataset Link: https://data.mendeley.com/datasets/hzfwzzsk8f/3

The dataset includes information on electric current consumption and meteorological conditions in the Alto Paraná region of Paraguay. The electric current dataset consists of hourly records with details such as datetime (ISO 8601 without timezone), substation, feeder, and consumption (amperage). Meanwhile, the meteorological dataset includes datetime, temperature (Celsius), relative humidity (percentage), wind speed (km/h), and atmospheric pressure (hPa) at a station level, with readings taken every three hours.

Covering the period from January 2017 to December 2020, the meteorological dataset has 22,445 records from a single weather station. In contrast, the electric consumption dataset involves 55 feeders distributed across 14 substations, resulting in a total of 1,848,947 records.

The data set provides information on electricity consumption and meteorological data of the region of Alto Paraná, Paraguay. It is presented in four files:

- electricity-consumption-raw.csv
- electricity-consumption-processed.csv
- meteorological-raw.csv
- meteorological-processed.csv

After combining the two datasets it looks like this:

	datetime	substation	feeder	consumption	temperature	humidity	wind_speed	pressure
0	2017-01-01T00:00:00	A	A1	64.671363	26.0	85.0	9.3	982.5
1	2017-01-01T00:00:00	A	A2	59.584091	26.0	85.0	9.3	982.5
2	2017-01-01T00:00:00	B	B1	125.500000	26.0	85.0	9.3	982.5
3	2017-01-01T00:00:00	B	B2	138.166667	26.0	85.0	9.3	982.5
4	2017-01-01T00:00:00	B	B3	25.833333	26.0	85.0	9.3	982.5
...	...	...	...	...	...	...	...	...
1928515	2020-12-31T23:00:00	M	M7	NaN	24.6	85.0	14.0	984.0
1928516	2020-12-31T23:00:00	N	N1	NaN	24.6	85.0	14.0	984.0

After combining two datasets, there are now a total of 8 columns and 1928520 rows.

```
Total Rows after combining the datasets: 1928520
Total Columns after combining the datasets: 8
```

Data Imputation:

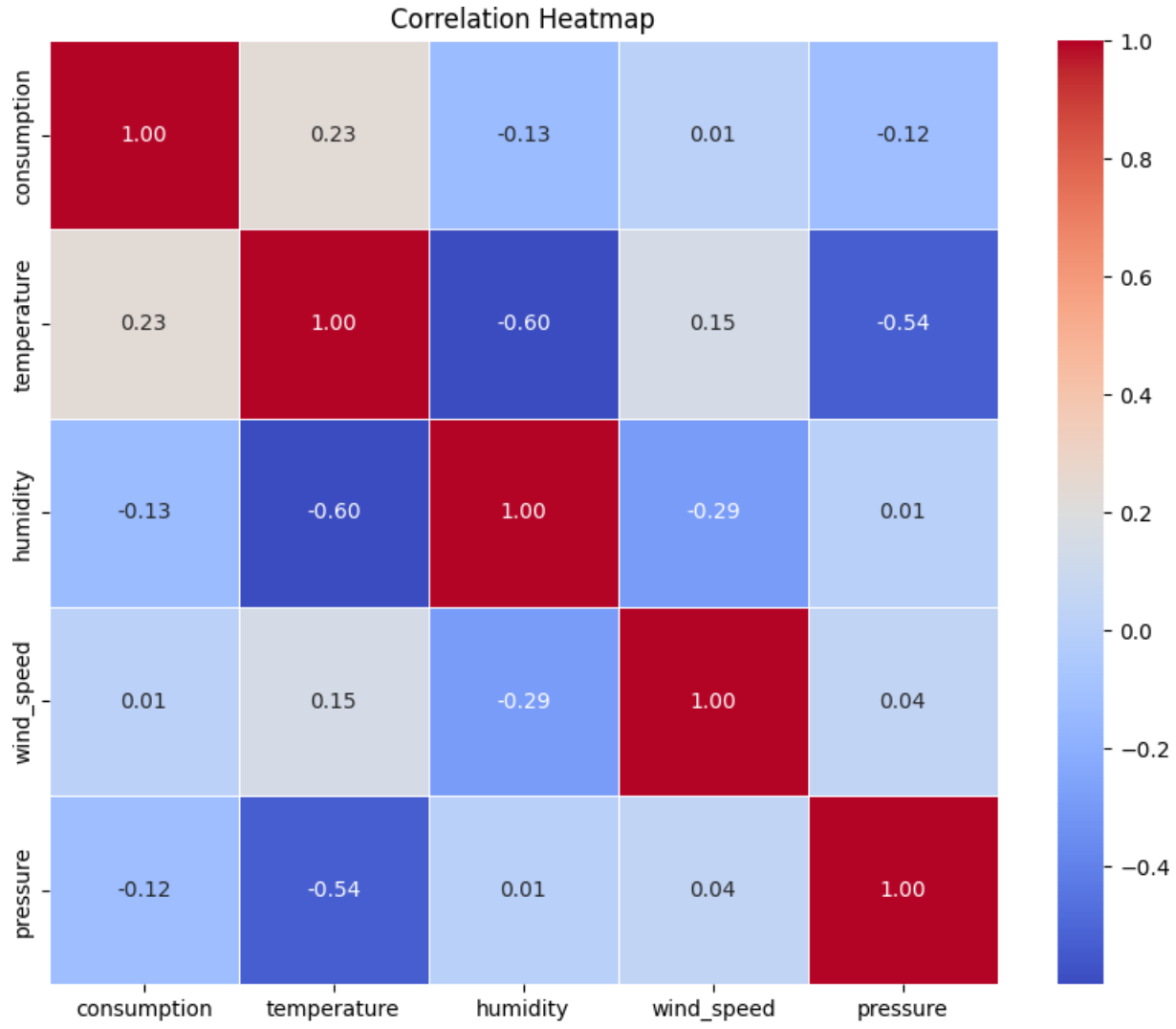
After calculating the null values there are in total of 43560 null values in the consumption column and the other columns do not have any null values. As the null values are only 2.25% of the total value which is very small I have deleted the null value rows.

```
column Name    Null values
datetime             0
substation           0
feeder               0
consumption      43560
temperature          0
humidity             0
wind_speed           0
pressure             0
dtype: int64
```

After eliminating the null values there are now have 1884960 rows and 8 columns.

```
Total Rows after combining the datasets: 1884960
Total Columns after combining the datasets: 8
```

Data Correlation:



In this heatmap, each cell represents how strongly two measurements are related. The values range from -1 to 1. A value of 1 means that when one measurement increases, the other also increases proportionally like when temperature rises, humidity also rises. On the other hand, a

value of -1 indicates a perfect negative relationship when one measurement goes up, the other goes down, and vice versa. For example, if wind speed increases, temperature might decrease.

If the value is around 0, it means there's no clear relationship between the two measurements. They fluctuate independently of each other. So, the correlation heatmap provides a quick way to see which measurements tend to move together or move in opposite directions, helping anyone to understand patterns and relationships in data more easily.

Approach:

Because the dataset is really complicated, using something called an Artificial Neural Network (ANN) is the best way to figure out the patterns in it. These networks are like computer versions of how our brains work, so they're great at handling lots of different pieces of information and figuring out how they're connected. This is perfect for dealing with data about things like energy use and climate, which can be really complicated and have lots of different factors influencing each other. Regular methods might struggle to understand these relationships, but ANNs are really good at it, so they're the top choice for studying this kind of data.

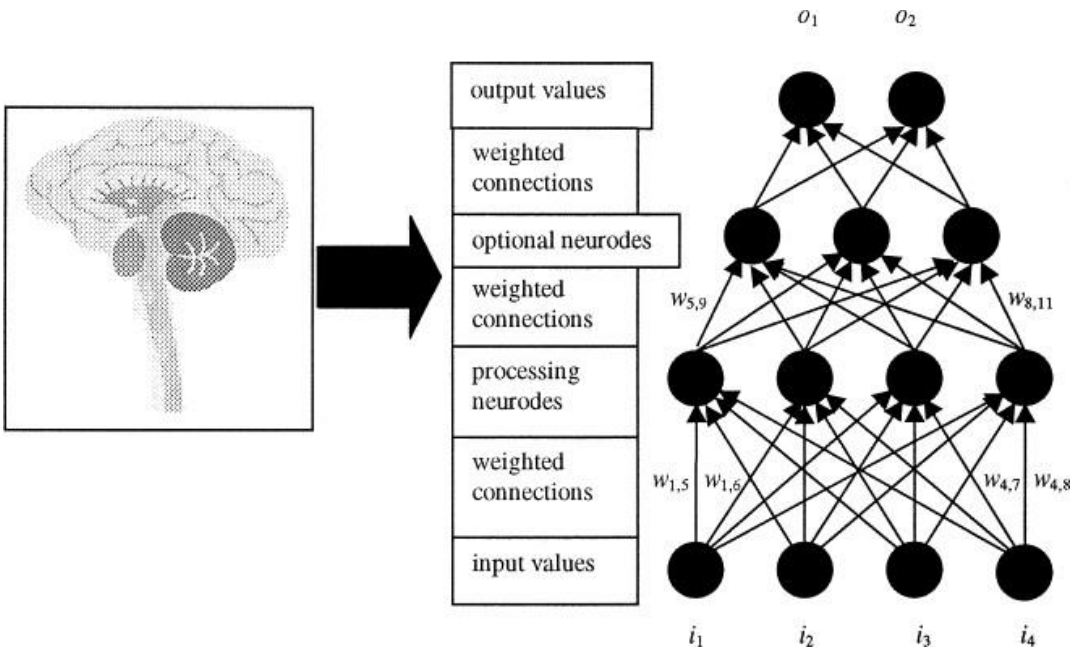


Figure: Sample Artificial Neural Network Architecture [13]

In a neural network, the values coming into a processing part are combined using a function like adding them up, finding the average, picking the highest one, or choosing the most common one. Then, this calculation is used with another function, like a sigmoid or hyperbolic tangent, to get the output. This output then becomes the input for the next layer. This keeps happening until the network gives a final result. While mimicking how our brains work would need parts to activate at different times, most artificial neural networks make each part activate once for every set of inputs.

What makes ANNs special is their knack for understanding complex patterns in data. They can break down these patterns into smaller parts, making it easier to understand them. This is really useful when the connections between different things in the data are complicated and can't be explained by simple models. So, using an ANN for this dataset is a smart choice because it can handle this complexity well. It can also adapt to changes in the data over time and keep getting better.

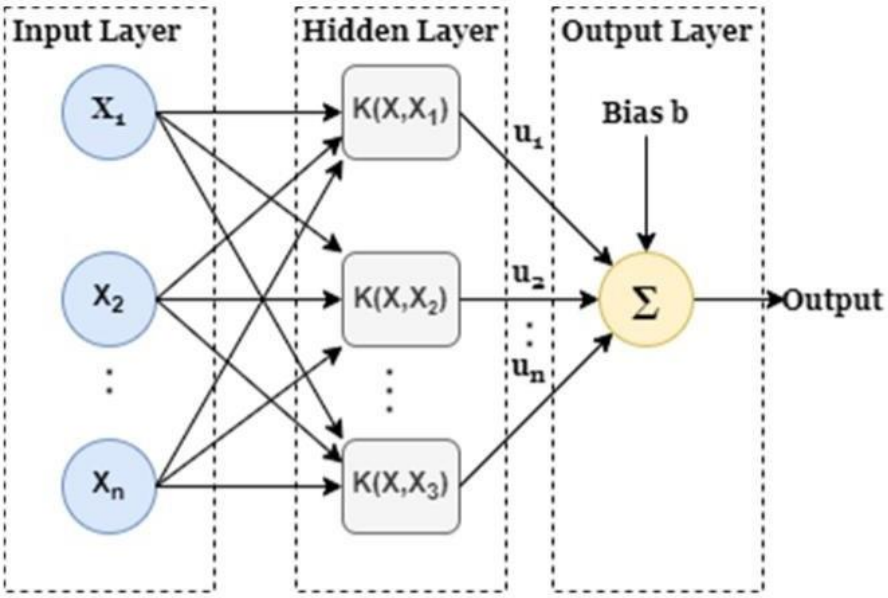


Figure: Architecture of SVM model [14]

When choosing to use Support Vector Machines (SVM) for an energy consumption prediction project, several important factors are considered. Firstly, energy data can be really complicated, with lots of different factors like the time of day, weather, and past usage. SVMs are good at handling this complexity and figuring out how these factors relate to each other. This is crucial

because a method that can understand both simple and complicated patterns in the data is needed to make accurate predictions about energy use.

Another reason for selecting SVMs is their capability to make reliable predictions, even when the data is messy or unpredictable. In energy prediction, where planning and managing resources effectively is essential, having a model that can provide dependable forecasts is crucial. SVMs act like steady guides that can help navigate through uncertain situations, ensuring that resource management stays on track. Additionally, many people in the research and industry world use SVMs for energy forecasting. There have been numerous studies and real-world applications demonstrating how effective SVMs can be in predicting energy usage. Seeing this widespread use and success lends confidence that SVMs are a suitable choice for the project, as there's a wealth of knowledge and experience to draw from.

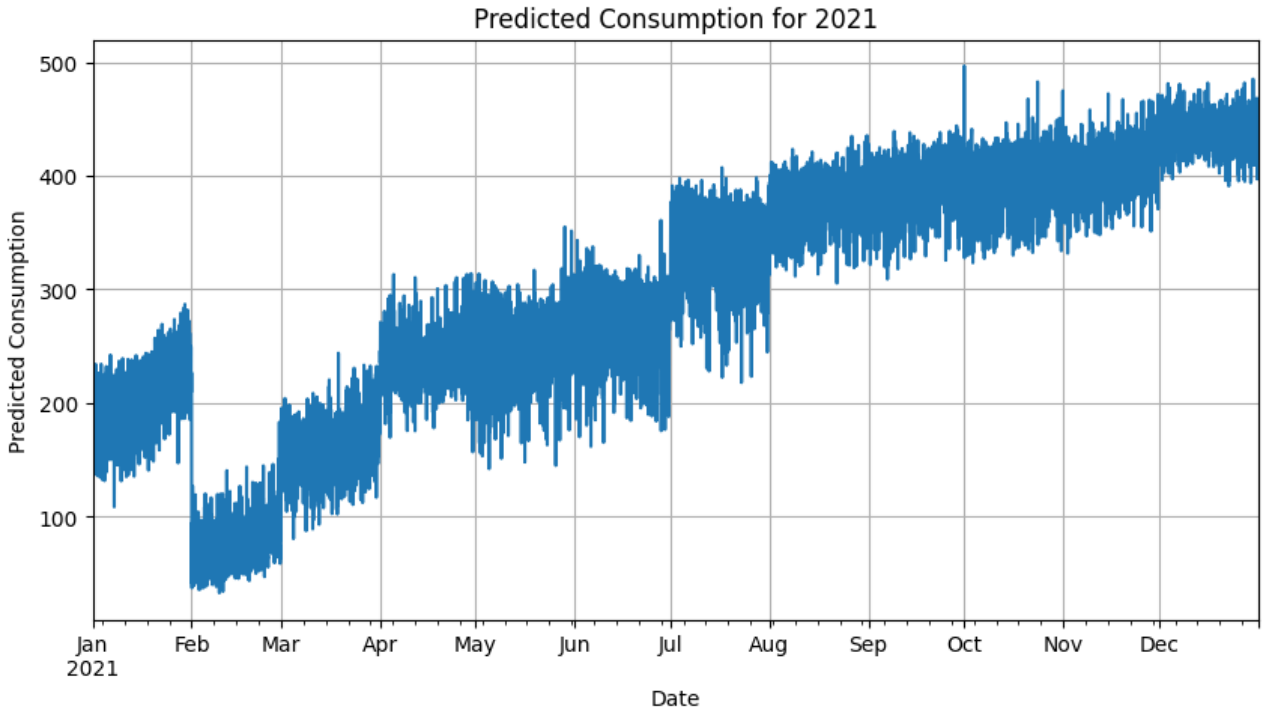
Why ultimately Feed-forward ANN has been chosen for this project:

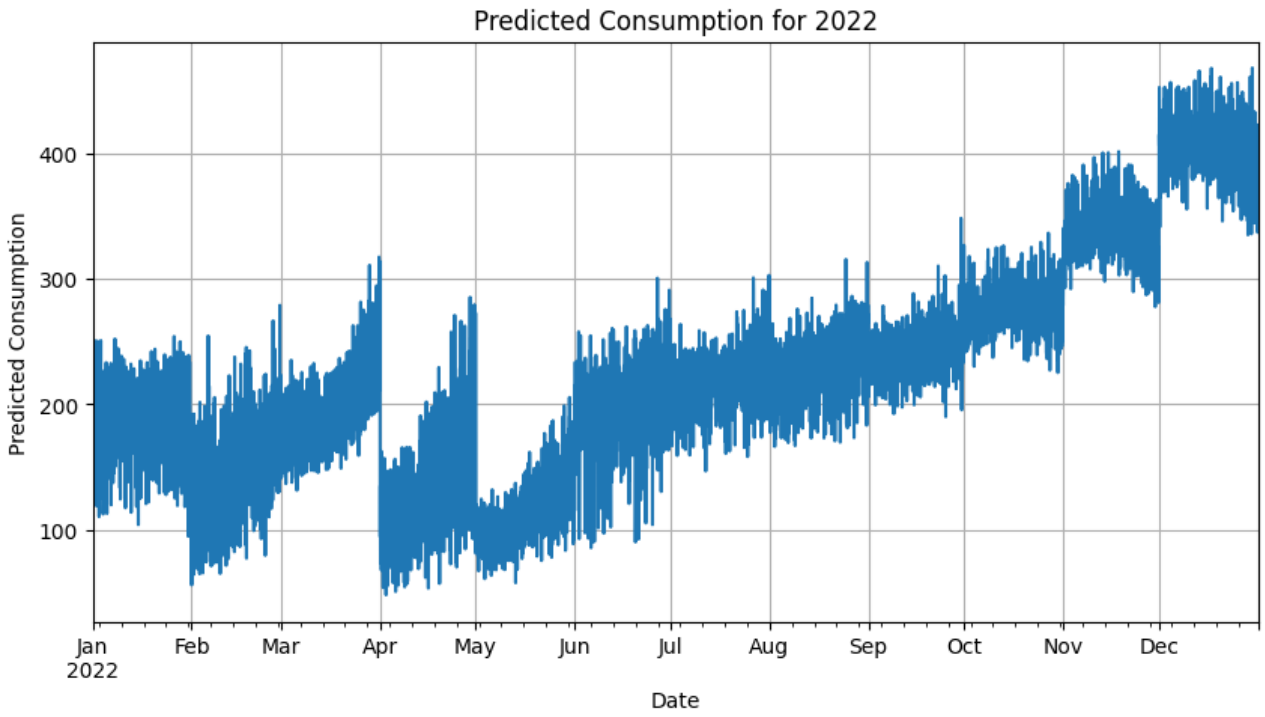
- **Capability to Model Non-linear Relationships**: The use of ReLU (Rectified Linear Unit) activation functions in hidden layers allows the network to capture non-linear relationships in the data. This is important for most real-world datasets which are rarely linearly separable or have simple patterns.
- **Layered Architecture**: The model has several layers, which is why it's sometimes called "deep learning." Each layer can learn different aspects of the data, starting from simple patterns in the early layers to more complex ones in deeper layers.
- **Batch Normalization**: This technique helps in stabilizing the learning process and dramatically reducing the number of training epochs required to train deep networks. It helps in normalizing the input layer by adjusting and scaling activations.
- **Adaptive Learning Rate (Adam Optimizer)**: The Adam optimizer adjusts the learning rate as it learns, which can lead to better convergence in training. This is particularly useful in complex networks that are prone to getting stuck during training or not converging fast enough.
- **Flexibility in Architecture**: Neural networks are highly modular and can be customized and extended to suit specific needs. For example, if the project involves more complex or high-dimensional data, more layers or change layer sizes could be added to capture more detailed features.
- **Availability of Computational Resources**: Training deep neural networks requires significant computational power, especially as the size of the model and data increases. If there is access to GPUs or other efficient hardware, it could facilitate using such models.

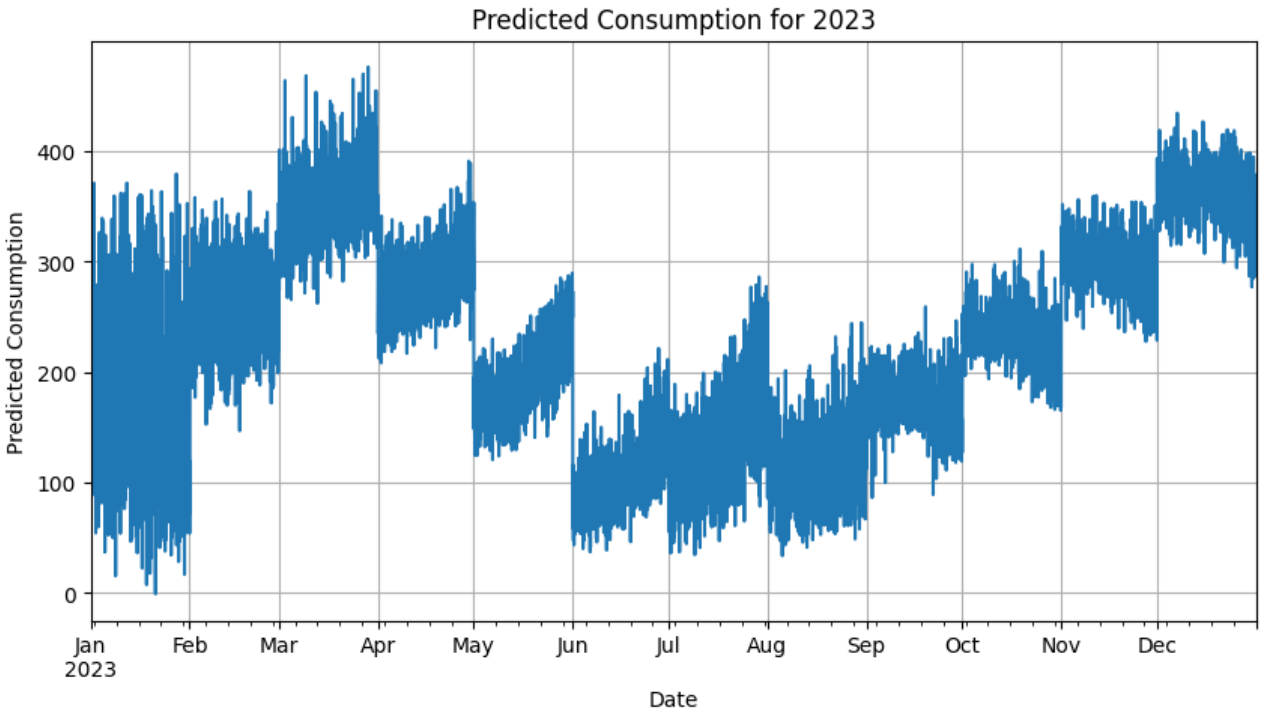
Overall, ANN seemed like the best fit for my energy prediction project because it can handle complexity, give reliable forecasts, is widely used and understood, and helps us understand the

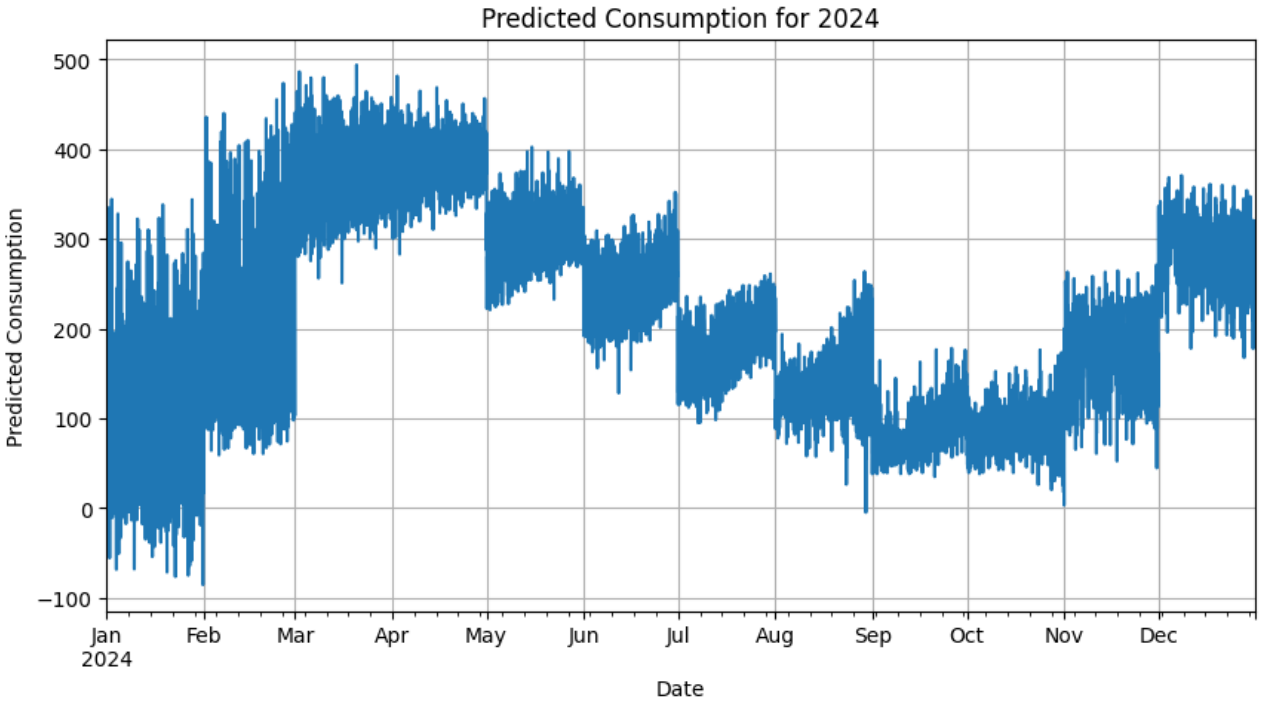
reasons behind the predictions. All of these qualities are really important for making accurate predictions and smart decisions in energy management.

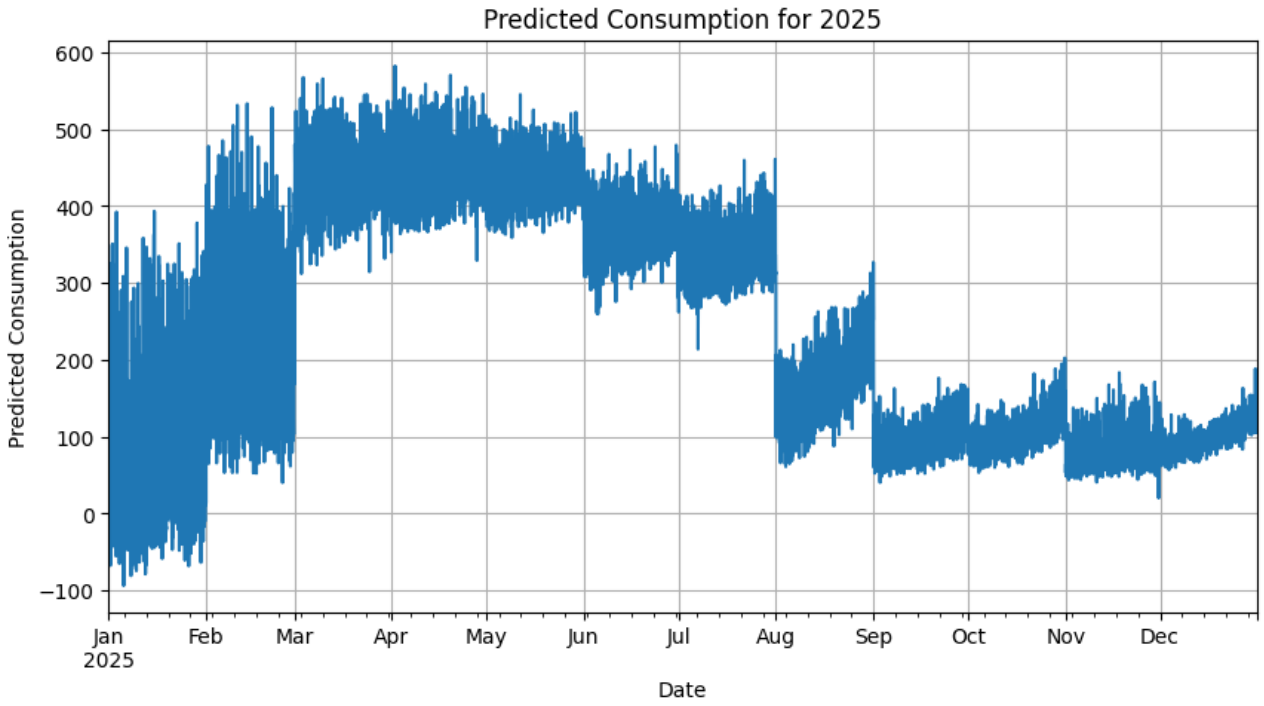
Visualization:






Predicted Consumption for 2023
Predicted Consumption
400
300
200
100
0
Jan
2023
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Date


Predicted Consumption for 2024
Predicted Consumption
500
400
300
200
100
0
−100
Jan
2024
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Date



Performance Evaluation:

The evaluation of the model performance would be based on the following criterion: Mean Absolute Percent Error (MAE) and serial correlation (linear correlation and linear regression R2).

Mean Absolute Error (MAE):

The MAE is a widely used metric for assessing the accuracy of forecasting models, particularly in time series analysis. It calculates the average percentage difference between predicted and actual values using the formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\gamma_i - \widehat{y_i}|$$

Where:

- n is the total number of observations.
- y_i is the actual (observed) value for the i th observation.
- $|\gamma_i - \widehat{y_i}|$ represents the absolute difference between the predicted and actual values for each observation.

A lower MAE indicates better accuracy. However, it's important to be aware of MAE's limitations, such as sensitivity to outliers and its inability to handle cases where actual values are zero. It is advisable to interpret MAE in the specific context of the application and to complement it with other evaluation metrics for a more comprehensive evaluation of the forecasting model's performance.

The Mean Absolute Error (MAE) for the model is 12.087. Unlike MSE, MAE looks at the average error in a simpler way. It just considers how much the predictions differ from the actual values, without squaring them. A lower MAE is better as it means the predictions are closer to the real figures. An MAE of 12.087 means that, on average, the model's predictions are about 12 units off from the actual electricity consumption. This shows the model is quite accurate, making it very useful for helping to plan how much electricity might be needed.

R-squared (R2):

The R-squared (R2) represents the proportion of the variance in the dependent variable (target) that is explained by the independent variables (predictions) in a regression model.

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$$

Sum of Squared Residuals (SSR): This is the sum of the squared differences between the actual values (observed) and the predicted values (estimated) by the regression model.

Total Sum of Squares (SST): This is the sum of the squared differences between the actual values and the mean of the dependent variable.

R2 ranges from 0 to 1, where:

- R^2=1 indicates that the regression model perfectly predicts the dependent variable.
- R^2=0 indicates that the regression model does not explain any variability in the dependent variable.

Interpretation of the Result:

An R^2 value of 0.9218 means that the generated model explains about 92.18% of the variations in electricity usage. This is a high score, which indicates that the model does a great job in understanding and predicting how much electricity will be used based on the data it has.

For predictive modeling, especially in contexts like electricity consumption where external factors such as weather can significantly influence the outcome, achieving an R^2 close to 1 is indicative of a robust model.

Deliverables:

1. **Trained Model:** The main point will be the development of a model, capturing essential weights and biases during the upcoming training phase. This model will serve as a pivotal tool for making accurate predictions on energy consumption.
2. **Prediction Results:** Following the training phase, predictions will be conducted on a designated test dataset or real-time data. The output will encompass predicted energy consumption values, facilitating a comprehensive evaluation against actual consumption.
3. **Performance Metrics:** The model's accuracy will be assessed using various performance metrics, including Mean Absolute Error (MAE) and Mean Squared Error (MSE). These metrics will offer quantitative insights into themodel's predictive capabilities.
4. **Visualization:** Visualizations will be created to aid interpretation, including comparative plots of predicted vs. actual values over time and visualizations of the loss function during training.
5. **Model Evaluation:** An in-depth analysis of the model's strengths and weaknesses will be conducted, providing insights into patterns captured well by the model and areas for potential improvement.
6. **Documentation:** Methodology, encompassing data preprocessing steps, chosen hyperparameters, and other pertinent details, will be thoroughly documented to ensure transparency, reproducibility, and a comprehensive understanding of the project's context.

Project Timeline:

Week 1-2: Project Planning and Research

- Defined project scope and objectives.
- Conducted background research on energy consumption prediction.
- Identified and collected relevant datasets.

Week 3-4: Data Preprocessing and Cleaning

- Cleaning and preprocess collected data.
- Exploring and visualize data to identify patterns.

Week 5-6: Model Selection and Development

- Choose machine learning algorithms.
- Develop and train prediction models.
- Evaluate and fine-tune models for optimal performance.

Week 7-8: Model Validation and Optimization

- Perform cross-validation and overfitting.
- Optimize models for improved performance.

Week 9-10: Documentation and Reporting

- Documentation, preprocessing, and model details.
- Generate a detailed project report summarizing the findings.

Week 11-12: Final Review

- Make final adjustments based on feedback.

Reference:

[1] Hong G, Choi GS, Eum J, Lee HS, Kim DD. The hourly Energy Consumption Prediction by KNN for buildings in community buildings. *Buildings*. 2022;12(10):1636. doi:10.3390/buildings12101636

[2] Rodrigues FM, Cardeira C, Calado JMF. 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. *Energy Procedia*. 2014;62:220-229. doi:10.1016/j.egypro.2014.12.383

[3] Pîrjan A, Oprea S, Căruţaşu G, Petroşanu DM, Bârã A, Coculescu C. Devising hourly forecasting solutions regarding electricity consumption in the case of commercial center type consumers. *Energies*. 2017;10(11):1727. doi:10.3390/en10111727

[4]Newsham, G.R., Birt, B.J.: Building-level occupancy data to improve ARIMA-based electricity use forecasts. In: Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building, pp. 13–18. ACM (2010)

[5]Tso, G.K.F., Yau, K.K.W.: Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks. Energy **32**(9), 1761–1768 (2007)

[6]Yang, X., Yuan, J., Yuan, J., Mao, H.: An improved WM method based on PSO for electric load forecasting. Expert Syst. Appl. **37**(12), 8036–8041 (2010)

[7]Kaytez, F., Cengiz Taplamacioglu, M., Cam, E., Hardalac, F.: Forecasting electricity consumption: a comparison of regression analysis, neural networks and least squares support vector machines. Int. J. Electr. Power Energy Syst. **67**, 431–438 (2015)

[8]Xiong, L.; Yao, Y. Study on an adaptive thermal comfort model with k-nearest-neighbors (knn) algorithm. Build. Environ. **2021**, 202, 108026.

[9] Mohandes, S.R.; Zhang, X.; Mahdiyar, A. A comprehensive review on the application of artificial neural networks in building energy analysis. Neurocomputing **2019**, 340, 55–75.

[10] Ring, M.; Eskofier, B.M. An approximation of the gaussian rbf kernel for efficient classification with svms. Pattern Recognit. Lett. **2016**, 84, 107–113.

[11] Santolamazza, A.; Cesarotti, V.; Introna, V. Anomaly detection in energy consumption for condition-based maintenance of compressed air generation systems: An approach based on artificial neural networks. IFAC-Pap. **2018**, 51, 1131–1136.

[12] Shapi MKM, Ramli NA, Awalin LJ. Energy consumption prediction by using machine learning for smart building: Case study in Malaysia. *Developments in the Built Environment*. 2021;5:100037. doi:10.1016/j.dibe.2020.100037

[13] Alberton KPF, Lima ADM, Nogueira WS, et al. Neural Networks Modeling of Dearomatization of Distillate Cuts with Furfural to Produce Lubricants. In: *Computer-Aided Chemical Engineering*. ; 2016:247-252. doi:10.1016/b978-0-444-63428-3.50046-1

[14] Kadam V, Kumar S, Bongale A, Wazarkar S, Kamat P, Patil S. Enhancing surface fault detection using machine learning for 3D printed products. *Applied System Innovation*. 2021;4(2):34. doi:10.3390/asi4020034