Table of Contents

[Project Overview 1](#_Toc165507590)

[Data Preprocessing 1](#_Toc165507591)

[Data Correlation Logic 2](#_Toc165507592)

[Stage 1 2](#_Toc165507593)

[Stage 2 2](#_Toc165507594)

[Stage 3 2](#_Toc165507595)

[Stage 4 2](#_Toc165507596)

[Stage 5 2](#_Toc165507597)

[Stage 6 3](#_Toc165507598)

[Stage 7 3](#_Toc165507599)

[Load and Temperature Mapping 3](#_Toc165507600)

[Machine Learning 4](#_Toc165507601)

[Model 1 4](#_Toc165507602)

[Code 4](#_Toc165507603)

[Hyperparameters Tuning 5](#_Toc165507604)

[Output 5](#_Toc165507605)

[Observations 5](#_Toc165507606)

[Prediction Errors 5](#_Toc165507607)

[Dedicated Prediction 5](#_Toc165507608)

[Model 2 5](#_Toc165507609)

[Code 5](#_Toc165507610)

[Hyperparameters Tuning 5](#_Toc165507611)

[Output 5](#_Toc165507612)

[Observations 5](#_Toc165507613)

[Prediction Errors 5](#_Toc165507614)

[Dedicated Prediction 5](#_Toc165507615)

[Models Comparison 5](#_Toc165507616)

[Project Conclusion 5](#_Toc165507617)

[Appendix 6](#_Toc165507618)

**List of Figures**

[Figure 1: Logic behind the correlation between the Load Zones and Temperature Stations 1](#_Toc165506197)

**List of Tables**

[Table 1: Zone ID and Station ID mapped matched values. 3](#_Toc165506198)

[Table 2: Correlation between Load Zones and Temperature Stations 4](#_Toc165506199)

# Project Overview

The final project focusses mainly on the entire life cycle of a machine learning task right from data preprocessing, data cleaning, data mapping, data correlation and then to training a couple of machine learning models to understand the correlation between the data files given.

This correlation is ultimately used to train a set of machine learning models for them to predict a week worth of load data using the learning from temperature correlation. The concept here is, that the data being input into the system is an instance from the real-world set of data available.

This adds a challenge to the process since there is a requirement to process the data and provide results within a certain timeframe, however, on the contrary, the size of the data is huge and to deal with big data in a certain limited timeframe by also providing a higher accuracy of prediction needs a machine learning model to be robust.

# Data Preprocessing

The data from both files namely ‘Load\_history\_final.csv’ and ‘Temp\_history\_final.csv’ contain a lot of values that are zero. This would lead to misleading the correlation phase since these zero values add no practical significance to the raw data and the ratio of increase or decrease to the data in relation between the Load and Temperature data.

To avoid this, the zero values are ignored during the correlation phase and not discarded entirely however, to preserve the raw data integrity. The logic used for the preprocessing of the data is pictorially represented below with the explanation of each stage following the image.



Figure 1: Logic behind the correlation between the Load Zones and Temperature Stations.

## Data Correlation Logic

### Stage 1

The correlation first begins by understanding how the values are varying each hour since that is the lowest level of the data available for prediction. The difference calculated for each hour of the data for every row there is an array created. These row wise arrays can be looked at as day wise hourly value differences. This is done for both the Temperature and Load data files.

### Stage 2

The data files however need cleaning since it is found that there are abrupt 0 values in the data which can hinder the model’s understanding of the correlation and mislead the model during the training phase, leading to poor efficiency of model’s prediction capability.

To counter this, the logic is designed to ignore that particular iteration of the difference calculation and to fill the place of that indexed position of the array, the code automatically fills that value of the array with a 0 (NaN).

### Stage 3

For each unique value of “Zone ID” and “Station ID”, the entire list of arrays is averaged to each indexed value of the array, i.e. the length of the individual arrays is the same and thus, the average array consisting of averaged values in each of its indices of values is of the same length of each of the row wise arrays.

During this process, the 0 values from each row wise arrays is ignored to avoid loss of significance in the data. So, the average is calculated for all the non-zero elements only, so the ratio of change remains rich without the interference of manipulated and abrupt data. This will ensure a certain amount of robustness assistance to the model through a cleaned database.

### Stage 4

At this stage, the individual and separate calculations of overall average arrays from Load values are linked with Temperature values. The ratio between each combination is extracted, i.e., The individual values from the overall average temperature array for the particular “Station ID” is divided by each particular indexed element from the overall average load array for the particular “Zone ID”.

The resultant array is stored for that combination of “Zone ID” and “Station ID”. Since we know that there are 20 load zones and 11 temperature stations, there will be a total of 220 arrays.

### Stage 5

The increase or decrease in each hourly consecutive value ensures either a positive or negative value in the initial array. This sign is carried through the entire process as well. If there is a similar increase or decrease in the hourly data across both the Temperature and Load data files, then the ratio array for a set combination would be positive. Otherwise, it would indicate a negative value which means that the data for that set combination of “Zone ID” and “Station ID” is inversely proportional.

### Stage 6

The positive values from this each ratio are copied onto another positive array for that set combination and similarly the negative values are copied onto another array for that set combination. The percentage of positives and negatives for each set combination is calculated based on the length of the positive and negative array to the length of the overall ratio array of that set combination. This will indicate the correlation from each Load Zone to Temperature Station.

### Stage 7

There are 20 Load Zones in total along with 11 Temperature stations. Since we look at ratios of each load zone to each temperature station to find the best correlation. The result is the combination of 220 ratios and 220 sets of positive and negative percentages.

The code, also find the highest correlation positive percentage from a zone ID to station ID and documents that that Load Zone affects that particular Temperature Station. This signifies the mapping between the combination, which will further aid the model training and prediction thereafter.

## Load and Temperature Mapping

The Table 1 below provides direct mapping explaining the mapping of Zone ID to the Station ID.



Table 1: Zone ID and Station ID mapped matched values.



Table 2: Correlation between Load Zones and Temperature Stations.

Table 2 indicates an interactive way to describe the correlation mapping between the Zone ID and Station ID. The Green box indicates that there is a correlation between the two nodes (ex. There is a correlation between Zone ID 6 and Station ID 3), and the red boxes indicate that there is no significant correlation between the two nodes (ex. Zone ID 10 and Station ID 5).

# Machine Learning

## Design Process and Trade Offs

## Model 1

The choice of model 1 for this task is chosen to be performed by the Logistic Regression algorithm. This is to provide a linear model that is better than the basic yet un-complex enough to test its performance upon the real-world large dataset.

### Code

The code for the model 1 learning and prediction is attached in the Model 1-code portion of the appendix of this document.

### Hyperparameters Tuning

### Output

### Observations

### Prediction Errors

### Dedicated Prediction

## Model 2

The choice of model 2 for this task is chosen to be performed by the Random Forest algorithm. This is to provide a dedicated complex approach in enabling the algorithm to approach the big data through multiple depths and thus aiding in prediction to greater accuracies.

### Code

The code for the model 2 learning and prediction is attached in the Model 2-code portion of the appendix of this document.

### Hyperparameters Tuning

### Output

### Observations

### Prediction Errors

### Dedicated Prediction

# Models Comparison

|  |  |
| --- | --- |
| Linear Regression | Random Forest Regressor |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# Project Conclusion

# Appendix