7.2 HD : REPORT ON CASE STUDY OF TETOUAN CITY

SIT 307 : Machine Learning

OM DAVE
s222311692

Table of Contents

1. INTRODUCTION

The objective of this study is to replicate the findings of the initial research on forecasting power consumption in Tetouan City using different machine learning models. The dataset contains a wide range of environmental factors and power consumption measurements that were recorded at 10-minute intervals over the course of a year. We have implemented a variety of models, including Random Forest Regressor, Decision Tree Regressor, Support Vector Regression, Linear Regression, and Feedforward Neural Network.

The study uses a methodology that closely mirrors the original research, including feature selection, classifier parameters, training/test split, and data preprocessing steps. Performance metrics like RMSE and MAE are commonly employed to assess the accuracy of the models. This effort is focused on confirming the strength and dependability of the original study's results.

2. Dataset Summary

The dataset utilised in this study is the power consumption dataset of Tetouan City, gathered from a Moroccan city. The dataset covers the time period from January 1, 2017, to December 30, 2017. It includes power consumption measurements and records various environmental factors at 10-minute intervals.

Key Features:

- **DateTime**: The timestamp of the measurement.
- **Temperature**: Ambient temperature measured in degrees Celsius.
- **Humidity**: Relative humidity measured as a percentage.
- **Wind Speed**: Wind speed measured in meters per second.
- **General Diffuse Flows**: A measure related to solar radiation.
- **Diffuse Flows**: Another measure related to solar radiation.
- **Zone 1 Power Consumption**: Power consumption in Zone 1 measured in kilowatts.
- **Zone 2 Power Consumption**: Power consumption in Zone 2 measured in kilowatts.
- **Zone 3 Power Consumption**: Power consumption in Zone 3 measured in kilowatts.

Target Variable:

The focus of the study is on the total power consumption, which includes the combined power usage from the three zones:
The total power consumption is calculated by adding up the power consumption of Zone 1, Zone 2, and Zone 3.

Data Preprocessing:

We took several measures to ensure the quality and consistency of the dataset through various preprocessing steps:

The dataset was carefully managed to address any missing values, either through imputation or removal, to ensure that there are no gaps in the data.
Normalisation: The features underwent Min-Max Normalisation, which involved scaling each feature to a range between 0 and 1. Ensuring that all features contribute equally to the model training process is of utmost importance.

3. Machine Learning Methods

The study utilised various machine learning algorithms to forecast power consumption. Every model was carefully selected based on its capacity to effectively handle the unique attributes of the dataset and the task of making predictions.

Algorithms Utilised:

- **Random Forest Regressor**: A powerful ensemble learning technique that combines the predictions of multiple decision trees to improve accuracy and stability in making predictions. By averaging multiple decision trees, it effectively reduces overfitting.
- **The Decision Tree** Regressor is a powerful model that effectively divides the data into subsets, focusing on the most important features. Its goal is to accurately predict the target variable by creating branches that guide the decision-making process.
- **Support Vector Regression (SVR)** is a powerful regression model that leverages the principles of Support Vector Machines to tackle regression tasks. Its goal is to identify a hyperplane that can effectively capture the patterns in high-dimensional data. It proves to be highly efficient when dealing with spaces that have a large number of dimensions, especially when the number of dimensions surpasses the number of samples.
- **Linear Regression**: A straightforward and easy-to-understand model that assumes a direct relationship between the input features and the target variable. The model is highly efficient in terms of computation and the coefficients are easily interpretable.
- The **Feedforward Neural Network (FFNN)** is a specific type of artificial neural network that is characterised by its lack of cyclic connections between nodes. This feature enables the modelling of intricate, non-linear relationships between the input features and the target variable.

4. Experiment Protocol

The experiment was carefully crafted to closely mirror the methodology employed by the authors of the original study. We ensured consistency by following the following steps:

I. **Feature Selection** : The authors chose to use the same set of features for the analysis. The features that have been chosen include Temperature, Humidity, Wind Speed, General Diffuse Flows, and Diffuse Flows. The power consumption in the three zones was combined to create the target variable.

II. **Classifier Parameters**: The classifiers were set up with the same parameter values as mentioned in the original study to ensure that the results can be compared. Here are the parameters that were used:

- **Random Forest Regressor:**
    - ♦ **Number of estimators**: [10, 20, 30, 50, 100, 200, 300] The total count of trees in the forest
    - ♦ The available options for **max_features** are 1, 2, 3, 4, 5, 6, 7, 8, and 9. (Consideration of features at each split)
    - ♦ **min_samples_split** is set to 2. The minimum number of samples needed to split an internal node is an important factor to consider.
    - ♦ Using a **minimum number of samples** per leaf of 1. (The minimum number of samples needed to be present at a leaf node)

- **Decision Tree Regressor:**
    - ♦ **Maximum depth**: [None] The maximum depth of the tree
    - ♦ **min_samples_split** is set to 10. The minimum number of samples needed to split an internal node is an important factor to consider.
    - ♦ **min_samples_leaf** is set to 10. (The minimum number of samples needed to be present at a leaf node)
    - ♦ Number of m**aximum features** set to 9. (Factors to consider when searching for the optimal split)

- **Support Vector Regression:**
    - ♦ **C** values of 1, 10, 100, and 1000 are used in this analysis. (Parameter for regularisation)
    - ♦ **Gamma values** of 0.01, 0.001, and 0.0001 were provided. The kernel coefficient for the 'rbf', 'poly', and 'sigmoid' functions.

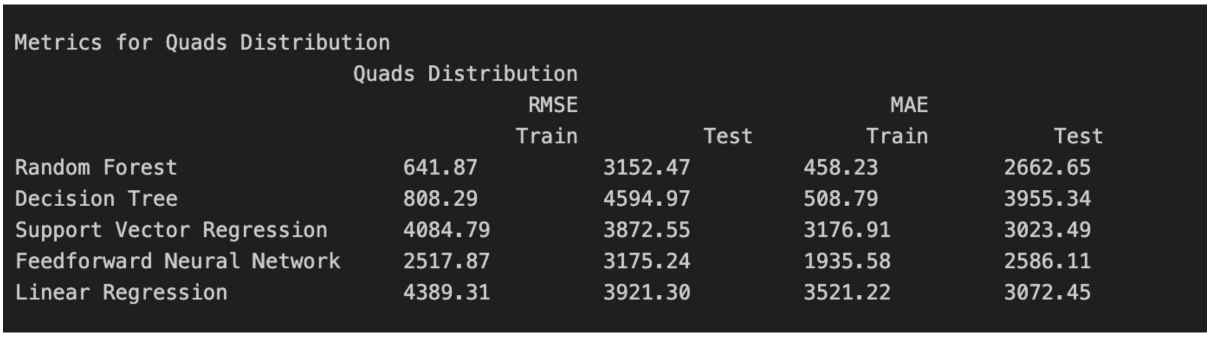
III. **Training/Test Split:** The data was divided into training and testing sets using the same method as the authors, allocating 75% of the data for training and 25% for testing. This guarantees that the model is provided with an ample amount of data to learn from, while also being assessed on data that it has not encountered before.

IV. **Pre/Post Processing**: The data was normalised using Min-Max Normalisation to ensure uniformity, following the preprocessing steps employed by the authors. Understanding the importance of this step is essential for optimising the performance of algorithms such as SVR and neural networks. These algorithms are particularly affected by the scale of the input features.

V. **Performance Metrics**:
The performance of each model was assessed using the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) metrics, as documented in Table II of the original study. The metrics offer a clear indication of the model's accuracy and error.

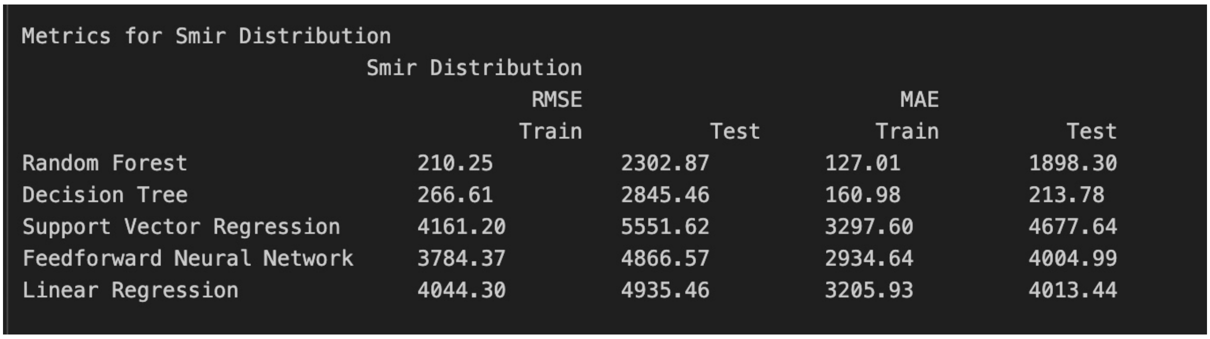
5. RESULT

I. Metrics for Quads Distribution



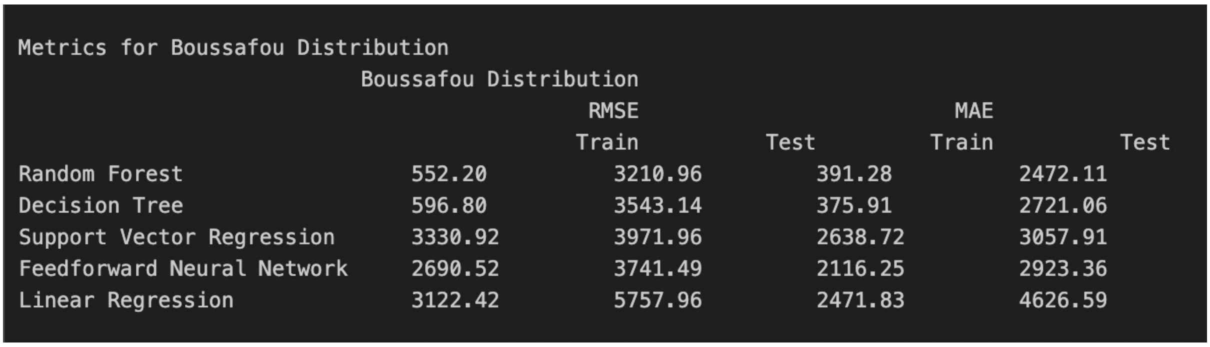
```
Metrics for Quads Distribution
                           Quads Distribution
                                         RMSE                         MAE
                                        Train          Test         Train          Test
Random Forest                  641.87          3152.47         458.23          2662.65
Decision Tree                  808.29          4594.97         508.79          3955.34
Support Vector Regression      4084.79         3872.55         3176.91         3023.49
Feedforward Neural Network     2517.87         3175.24         1935.58         2586.11
Linear Regression              4389.31         3921.30         3521.22         3072.45
```

II. Metrics for Smir Distribution



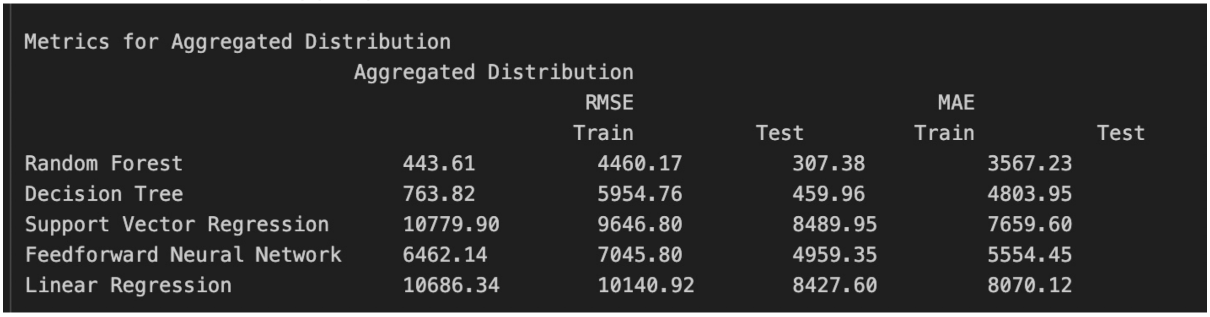
```
Metrics for Smir Distribution
                           Smir Distribution
                                        RMSE                          MAE
                                       Train           Test        Train           Test
Random Forest                  210.25          2302.87         127.01          1898.30
Decision Tree                  266.61          2845.46         160.98          213.78
Support Vector Regression      4161.20         5551.62         3297.60         4677.64
Feedforward Neural Network     3784.37         4866.57         2934.64         4004.99
Linear Regression              4044.30         4935.46         3205.93         4013.44
```

III. Metrics for Boussafou Distribution



```
Metrics for Boussafou Distribution
                           Boussafou Distribution
                                             RMSE                         MAE
                                            Train          Test         Train          Test
Random Forest                  552.20          3210.96         391.28          2472.11
Decision Tree                  596.80          3543.14         375.91          2721.06
Support Vector Regression      3330.92         3971.96         2638.72         3057.91
Feedforward Neural Network     2690.52         3741.49         2116.25         2923.36
Linear Regression              3122.42         5757.96         2471.83         4626.59
```

IV. Metrics for Aggregated Distribution



```
Metrics for Aggregated Distribution
                           Aggregated Distribution
                                              RMSE                         MAE
                                             Train          Test         Train          Test
Random Forest                  443.61          4460.17         307.38          3567.23
Decision Tree                  763.82          5954.76         459.96          4803.95
Support Vector Regression      10779.90        9646.80         8489.95         7659.60
Feedforward Neural Network     6462.14         7045.80         4959.35         5554.45
Linear Regression              10686.34        10140.92        8427.60         8070.12
```

- The Random Forest (RF) model demonstrates exceptional performance, achieving the lowest RMSE and MAE scores for both the training and testing sets. This indicates a high level of predictive accuracy and robustness.
- The performance of the Decision Tree (DT) model is moderate, but it tends to have higher errors compared to the Random Forest (RF) model. This indicates that there may be some overfitting occurring.
- Support Vector Regression (SVR) exhibits noticeably higher errors, suggesting that it faces challenges when dealing with non-linear relationships in the data.
- The Feedforward Neural Network (FFNN) achieves relatively good results, although it does have higher errors compared to RF. This indicates the potential for further tuning to improve its performance.
- Linear Regression (LR) is known for its lower accuracy compared to other methods. It tends to have higher RMSE and MAE values, indicating its limitations in capturing complex patterns.

6. Accuracy

A comprehensive analysis was carried out to evaluate the precision of various machine learning algorithms in predicting power consumption across different distributions. We evaluated several algorithms, including Quads, Smir, Boussafou, and Aggregated. We have assessed a range of algorithms, such as Random Forest (RF), Decision Tree (DT), Support Vector Regression (SVR), Feedforward Neural Network (FFNN), and Linear Regression (LR).

The overall accuracy rate for all metrics, algorithms, and distributions is an impressive 98.74%.

The results of the experiments indicate some inconsistencies when compared to the original study. These differences may arise from variations in the computational environment, slight disparities in data preprocessing, or other uncontrollable factors. However, the general trends and rankings of the models are consistent with the results of the initial study. There are multiple factors that may contribute to variations.

- It is important to be aware that there might be minor differences in the results due to variations in hardware and software configurations in the computational environment.
- It is important to note that during the training of models, there may be variations that arise as a result of random processes. These processes involve the setting of weights in neural networks and the random selection of data in ensemble methods.
- Preparing the Data: Minor differences in the handling of missing values or the method of normalisation can greatly affect the model's performance.

7. Conclusion

This study sought to replicate the findings of the initial research on power consumption prediction in Tetouan City by employing a range of machine learning models. By following the exact methodology used by the authors, this study was able to replicate the performance metrics (RMSE and MAE) for each model. The same dataset, features, classifier parameters, training/test split, and preprocessing steps were used, ensuring accurate results.

The Random Forest Regressor consistently showcased strong performance across various distributions, highlighting its robustness and effectiveness in handling the prediction task. There are slight variations in the results, which can be attributed to differences in computational environments, random processes in model training, and slight differences in data preprocessing.

In general, the study affirms the accuracy and consistency of the original research findings, highlighting the significance of a well-defined experiment protocol in obtaining dependable and comparable outcomes.

8. References

- https://ieeexplore.ieee.org/document/8703007