Capstone Project - 2

Appliances Energy Prediction

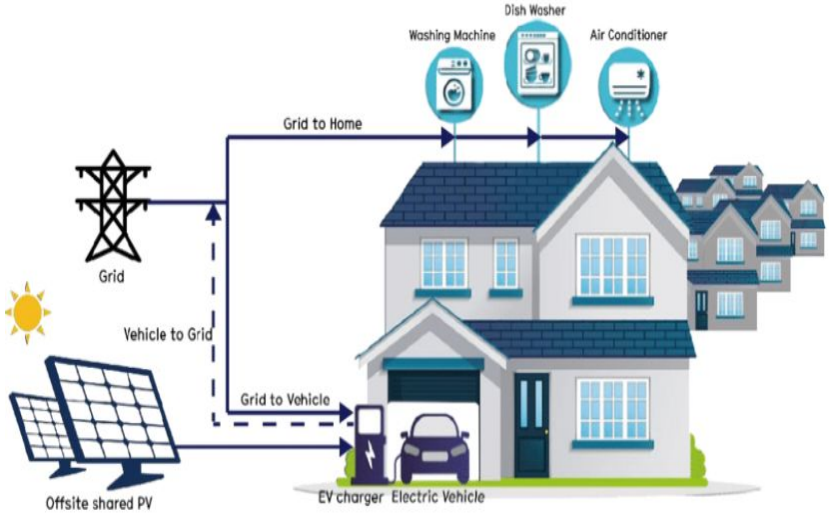
Team Members

Akshada

Nikita

Introduction

We'll be working on appliance energy prediction, where we'll predict appliance energy consumption for a house based on factors like temperature, humidity & pressure. This is a regression problem. Regression analysis is a method of predictive modelling that investigates the relationship between a dependent (target) and independent variables (predictor).



Introduction(Continued)

Power prediction has been a major concern in power systems for effective energy utilization to reduce demand. In order to support economic and social progress and build a better quality of life, the world needs energy in large quantities.

Although, there are still many places in the world where there are outages due to excess load consumed by appliances at home.

In this project, data-driven predictive models will be presented and discussed for predicting the energy utilization of a house's electric appliances. The performances of different models using their algorithms will be compared to find the best algorithm or model for the given dataset of appliance energy prediction.

Objective

The data set is at 10 min for about 4.5 months. The house temperature and humidity conditions were monitored with a ZigBee wireless sensor network. Each wireless node transmitted the temperature and humidity conditions around 3.3 min. Then, the wireless data was averaged for 10 minutes periods. The energy data was logged every 10 minutes with m-bus energy meters. Weather from the nearest airport weather station (Chievres Airport, Belgium) was downloaded from a public data set from Reliable Prognosis (rp5.ru) and merged together with the experimental data sets using the date and time column. Two random variables have been included in the data set for testing the regression models and to filter out non-predictive attributes(parameters).

Objective (Continued)

We need to predict appliance energy consumption for a house based on factors like temperature, humidity, and pressure.

In order to achieve this, the supervised machine learning model was developed using regression algorithms. Regression algorithms are used because data consist of continuous features in the dataset.

The main objective is to understand which algorithm provides a better result in predicting energy consumption.

Variable Breakdown

There are 29 features to describe the energy use of appliances:

- Appliances : Total energy used by appliances, in Wh
- date : time year-month-day hour:minute:second
- lights : energy use of light fixtures in the house in Wh
- T1 : Temperature in kitchen area, in Celsius
- T2 : Temperature in living room area, in Celsius
- T3 : Temperature in laundry room area, in Celsius
- T4 : Temperature in office room, in Celsius

Variable Breakdown (Continued)

- T5 : Temperature in bathroom, in Celsius
- T6 : Temperature outside the building (north side), in Celsius
- T7 : Temperature in ironing room, in Celsius
- T8 : Temperature in teenager room 2, in Celsius
- T9 : Temperature in parents' room, in Celsius
- T_out : Temperature outside (from Chievres weather station), in Celsius
- Tdewpoint : (from Chievres weather station), Â°C

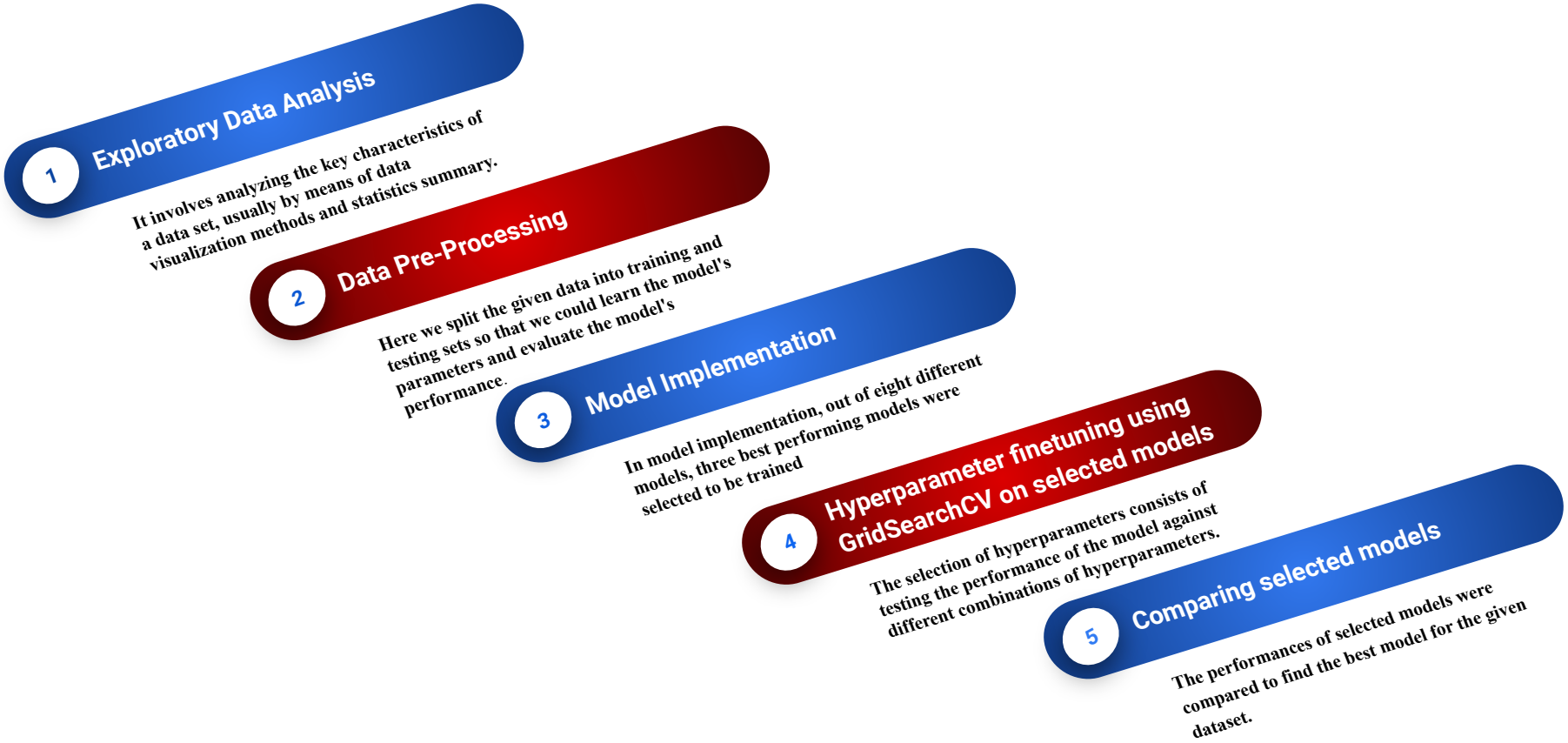
Variable Breakdown (Continued)

- RH_1 : Humidity in kitchen area, in %
- RH_2 : Humidity in living room area, in %
- RH_3 : Humidity in laundry room area, in %
- RH_4 : Humidity in office room, in %
- RH_5 : Humidity in bathroom, in %
- RH_6 : Humidity outside the building (north side), in %
- RH_7 : Humidity in ironing room, in %

Variable Breakdown (Continued)

- RH_8 : Humidity in teenager room 2, in %
- RH_9 : Humidity in parents' room, in %
- RH_out : Humidity outside (from Chievres weather station), in %
- Pressure : (from Chievres weather station), in mm Hg
- Wind speed : (from Chievres weather station), in m/s
- Visibility : (from Chievres weather station), in km
- Rv1 : Random variable 1, non-dimensional
- Rv2 : Random variable 2, non-dimensional

Steps involved in ML



Descriptive statistics

	Appliances	lights	T1	RH_1	T2	RH_2	T3	RH_3	T4	RH_4	T5	RH_5	T6	RH_6
count	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000
mean	97.694958	3.801875	21.686571	40.259739	20.341219	40.420420	22.267611	39.242500	20.855335	39.026904	19.592106	50.949283	7.910939	54.609083
std	102.524891	7.935988	1.606066	3.979299	2.192974	4.069813	2.006111	3.254576	2.042884	4.341321	1.844623	9.022034	6.090347	31.149806
min	10.000000	0.000000	16.790000	27.023333	16.100000	20.463333	17.200000	28.766667	15.100000	27.660000	15.330000	29.815000	-6.065000	1.000000
25%	50.000000	0.000000	20.760000	37.333333	18.790000	37.900000	20.790000	36.900000	19.530000	35.530000	18.277500	45.400000	3.626667	30.025000
50%	60.000000	0.000000	21.600000	39.656667	20.000000	40.500000	22.100000	38.530000	20.666667	38.400000	19.390000	49.090000	7.300000	55.290000
75%	100.000000	0.000000	22.600000	43.066667	21.500000	43.260000	23.290000	41.760000	22.100000	42.156667	20.619643	53.663333	11.256000	83.226667
max	1080.000000	70.000000	26.260000	63.360000	29.856667	56.026667	29.236000	50.163333	26.200000	51.090000	25.795000	96.321667	28.290000	99.900000

T7	RH_7	T8	RH_8	T9	RH_9	T_out	Press_mm_hg	RH_out	Windspeed	Visibility	Tdewpoint	rv1	rv2
19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000
20.267106	35.388200	22.029107	42.936165	19.485828	41.552401	7.411665	755.522602	79.750418	4.039752	38.330834	3.760707	24.988033	24.988033
2.109993	5.114208	1.956162	5.224361	2.014712	4.151497	5.317409	7.399441	14.901088	2.451221	11.794719	4.194648	14.496634	14.496634
15.390000	23.200000	16.306667	29.600000	14.890000	29.166667	-5.000000	729.300000	24.000000	0.000000	1.000000	-6.600000	0.005322	0.005322
18.700000	31.500000	20.790000	39.066667	18.000000	38.500000	3.666667	750.933333	70.333333	2.000000	29.000000	0.900000	12.497889	12.497889
20.033333	34.863333	22.100000	42.375000	19.390000	40.900000	6.916667	756.100000	83.666667	3.666667	40.000000	3.433333	24.897653	24.897653
21.600000	39.000000	23.390000	46.536000	20.600000	44.338095	10.408333	760.933333	91.666667	5.500000	40.000000	6.566667	37.583769	37.583769
26.000000	51.400000	27.230000	58.780000	24.500000	53.326667	26.100000	772.300000	100.000000	14.000000	66.000000	15.500000	49.996530	49.996530

Descriptive statistics (Continued)

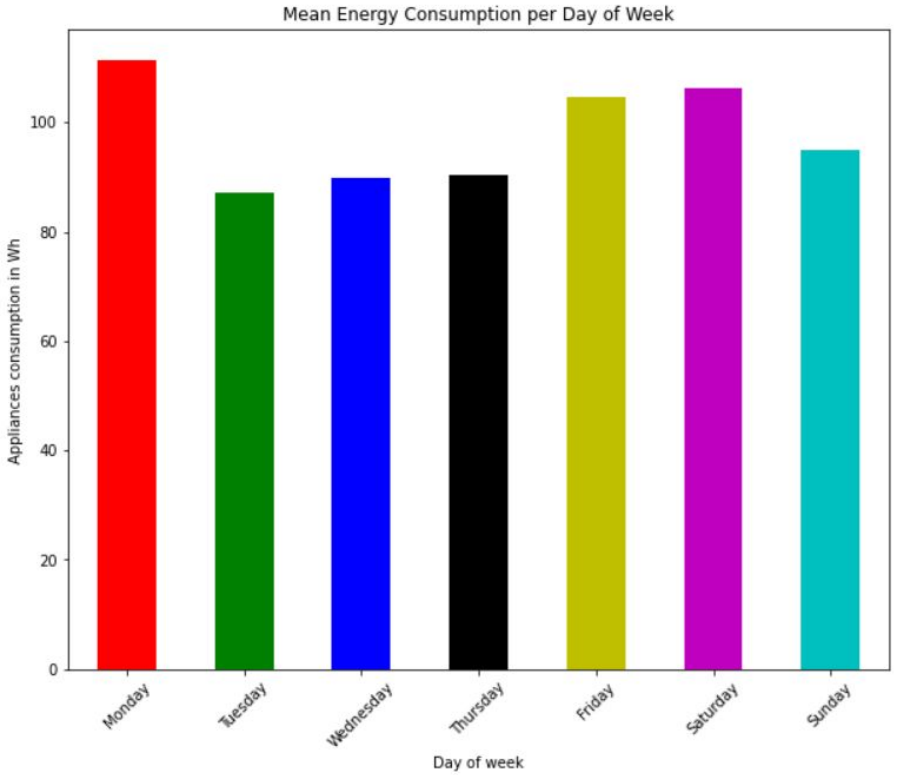
- The dataset consists of 28 independent variables (i.e., 11 temperature, 10 humidity, 1 pressure, 1 wind speed, 1 visibility, 1 light, 1 date, and 2 randoms) and 1 dependent variable (i.e., appliances).
- Temperature ranges from -6°C to 30°C, with the temperature inside the house varying between 14.89°C & 29.85°C and the temperature outside (T6) varying between -6.06°C to 28.29°C.
- Humidity ranges from 1% to 100%, with humidity inside the house ranging from 20.46% to 63.36%, except for the RH_5 (Bathroom) and RH_6 (Outside house), which range from 29.82% to 96.32% and 1% to 99.9%, respectively.

Descriptive statistics (Continued)

- Wind speed ranges from 0 to 14 m/s.
- Visibility ranges from 1 to 66 km.
- Pressure ranges from 729 to 772 mm Hg.
- Appliance energy usage ranges from 10 to 1080 Wh where 75% of appliance consumption is less than 100 Wh. With the maximum consumption of 1080 Wh, there will be outliers in this column and there are few cases where consumption is very high.

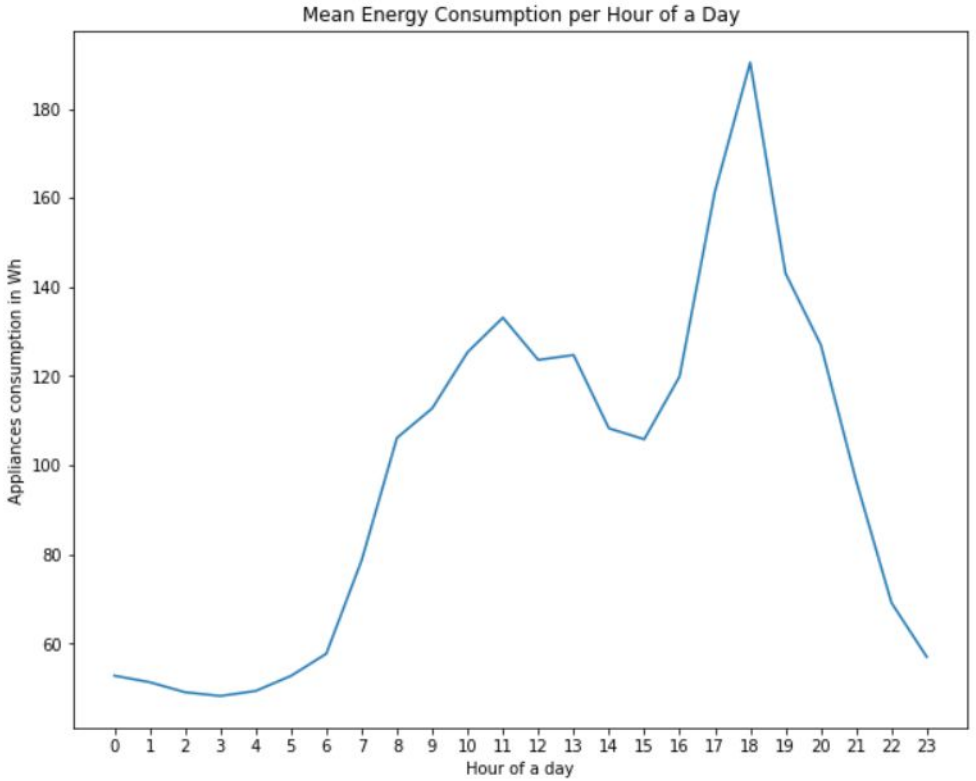
Mean appliance energy consumption per Day of Week

- The maximum mean appliance energy consumption is on Monday and the minimum mean appliance energy consumption is on Tuesday.
- Mean appliance energy consumption is at least 85 Wh (approx.) per day.



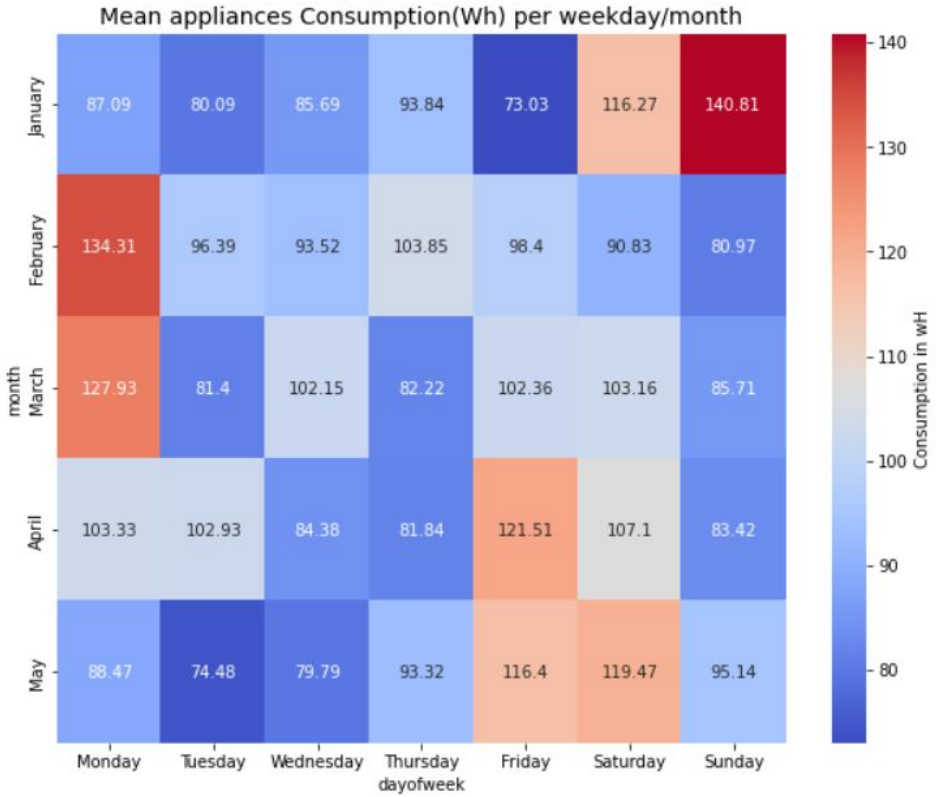
Mean appliance energy Consumption per Hour of a Day

- The mean energy consumption of appliances is increasing in the morning hours (i.e., from 3-11 hrs), after which there is a drop in energy consumption in the afternoon (i.e., 11-15 hrs), then again it rises in the evening hours(i.e., 15-18 hrs) and drops at night (i.e., 18-23 hrs).

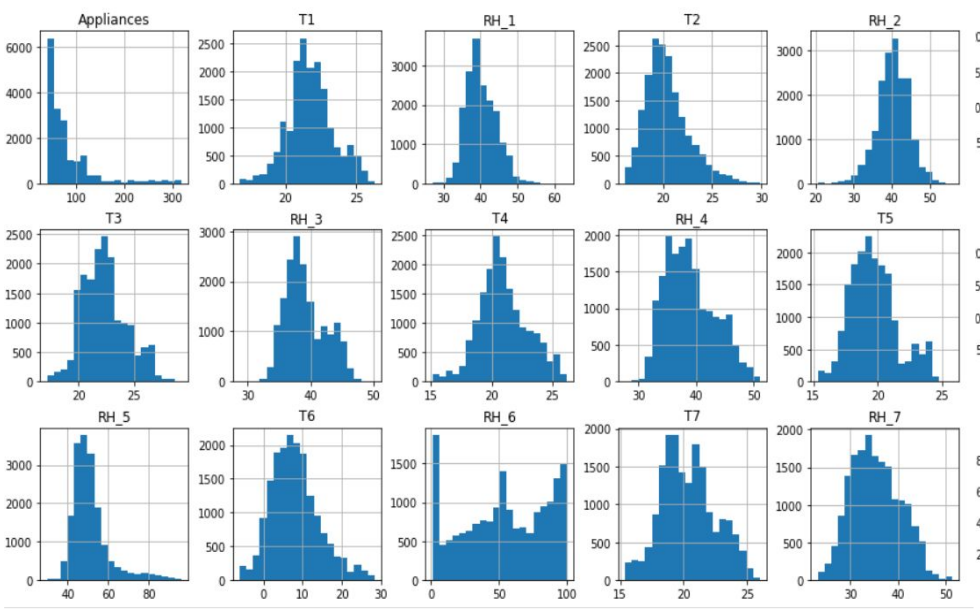


Mean appliance consumption (Wh) per weekday/month

- The highest and lowest mean appliances consumption is on the weekend (i.e. Sunday, with approx. value of 141 (Wh)) and weekday (i.e. Friday, with approx. value of 73 (Wh)) in January.
- Maximum mean appliances consumption is in the months of March and April, with approx. value of 685 (Wh).



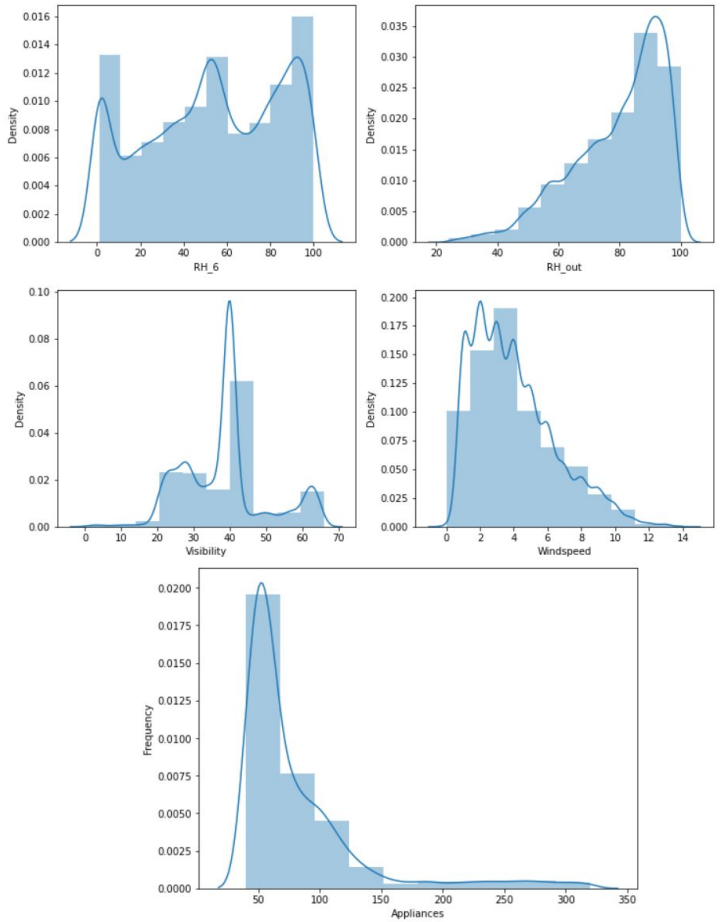
Checking frequency distribution



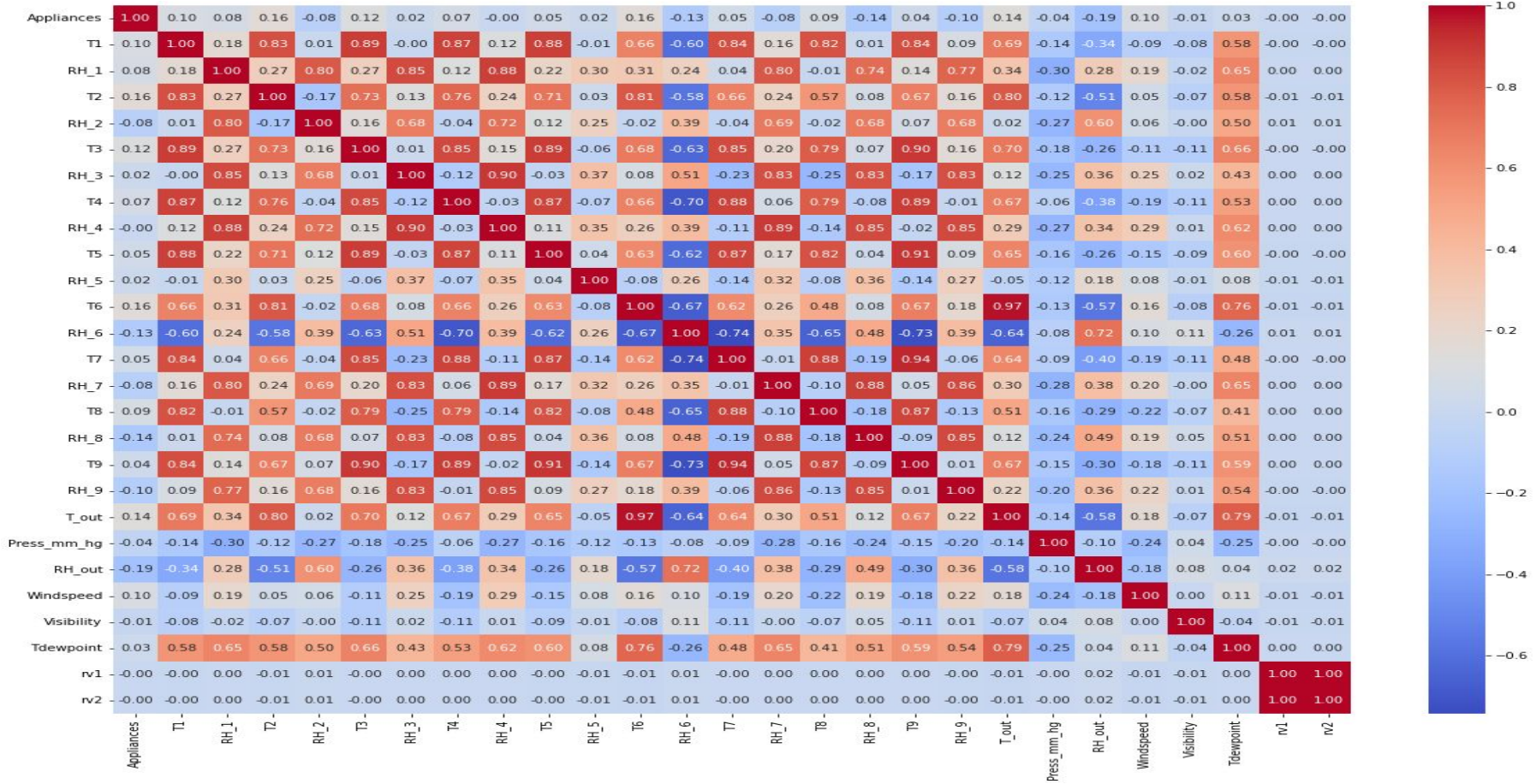
- Almost all temperature variable follows "Normal Distribution" except for T9.
- Similarly, all humidity variable follows "Normal Distribution" except RH_Out and RH_6.
- The random variables rv1 and rv2 have more or less the same values for all the recordings.

Checking frequency distribution (Continued)

- In independent variables, the Windspeed column is positively skewed & the RH_out column is negatively skewed.
- RH_6 and Visibility show an irregular distribution.
- The dependent variable (i.e., Appliance) is positively skewed.
- The dependent variable ‘Appliances’ have most of the values less than 200Wh, showing that high energy consumption cases are very low.



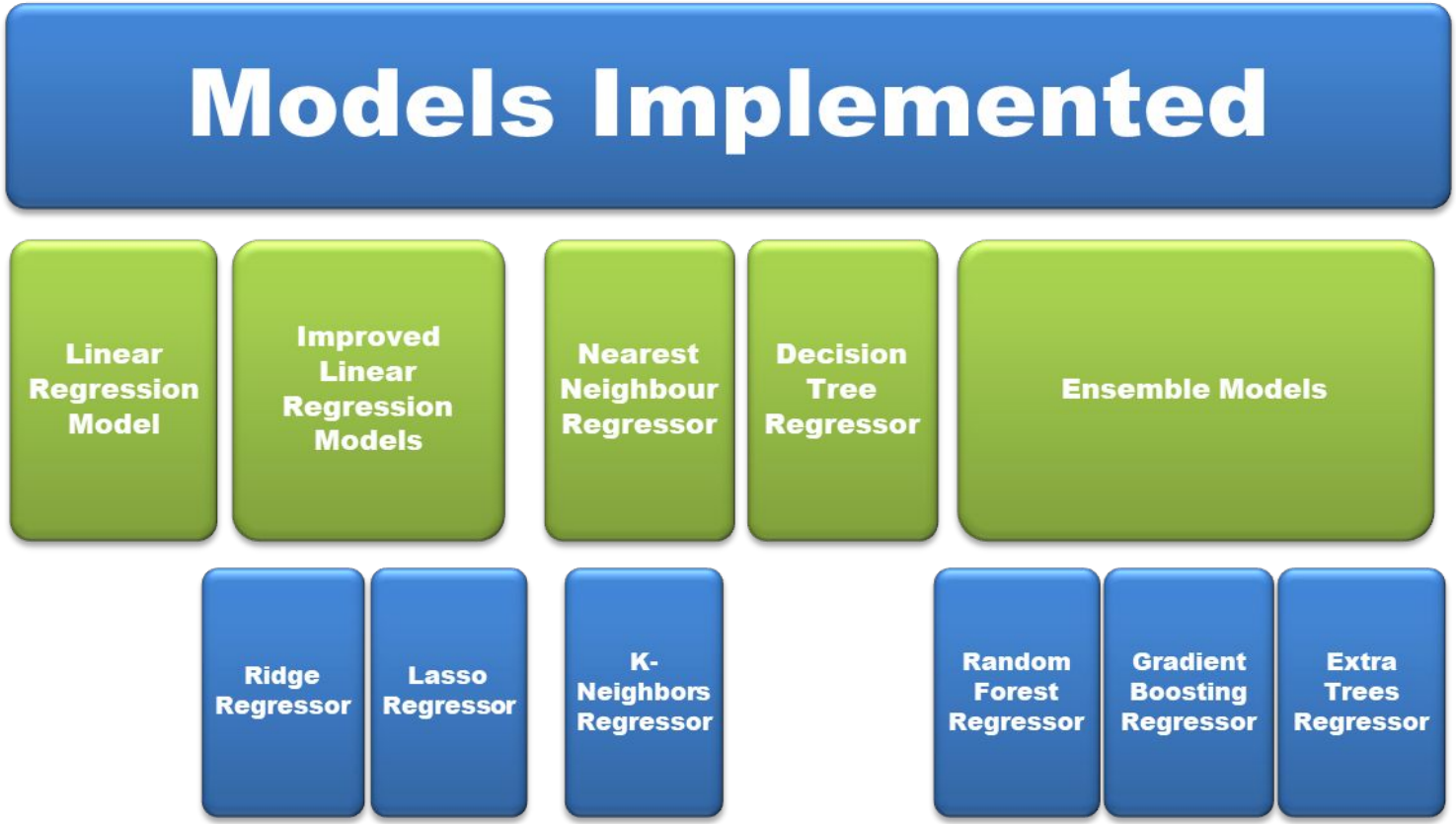
Heat map


Appliances 1.00 0.10 0.08 0.16 -0.08 0.12 0.02 0.07 -0.00 0.05 0.02 0.16 -0.13 0.05 -0.08 0.09 -0.14 0.04 -0.10 0.14 -0.04 -0.19 0.10 -0.01 0.03 -0.00 -0.00
T1 0.10 1.00 0.18 0.83 0.01 0.89 -0.00 0.87 0.12 0.88 -0.01 0.66 -0.60 0.84 0.16 0.82 0.01 0.84 0.09 0.69 -0.14 -0.34 -0.09 -0.08 0.58 -0.00 -0.00
RH_1 0.08 0.18 1.00 0.27 0.80 0.27 0.85 0.12 0.88 0.22 0.30 0.31 0.24 0.04 0.80 -0.01 0.74 0.14 0.77 0.34 -0.30 0.28 0.19 -0.02 0.65 0.00 0.00
T2 0.16 0.83 0.27 1.00 -0.17 0.73 0.13 0.76 0.24 0.71 0.03 0.81 -0.58 0.66 0.24 0.57 0.08 0.67 0.16 0.80 -0.12 -0.51 0.05 -0.07 0.58 -0.01 -0.01
RH_2 -0.08 0.01 0.80 -0.17 1.00 0.16 0.68 -0.04 0.72 0.12 0.25 -0.02 0.39 -0.04 0.69 -0.02 0.68 0.07 0.68 0.02 -0.27 0.60 0.06 -0.00 0.50 0.01 0.01
T3 0.12 0.89 0.27 0.73 0.16 1.00 0.01 0.85 0.15 0.89 -0.06 0.68 -0.63 0.85 0.20 0.79 0.07 0.90 0.16 0.70 -0.18 -0.26 -0.11 -0.11 0.66 -0.00 -0.00
RH_3 0.02 -0.00 0.85 0.13 0.68 0.01 1.00 -0.12 0.90 -0.03 0.37 0.08 0.51 -0.23 0.83 -0.25 0.83 -0.17 0.83 0.12 -0.25 0.36 0.25 0.02 0.43 0.00 0.00
T4 0.07 0.87 0.12 0.76 -0.04 0.85 -0.12 1.00 -0.03 0.87 -0.07 0.66 -0.70 0.88 0.06 0.79 -0.08 0.89 -0.01 0.67 -0.06 -0.38 -0.19 -0.11 0.53 0.00 0.00
RH_4 -0.00 0.12 0.88 0.24 0.72 0.15 0.90 -0.03 1.00 0.11 0.35 0.26 0.39 -0.11 0.89 -0.14 0.85 -0.02 0.85 0.29 -0.27 0.34 0.29 0.01 0.62 0.00 0.00
T5 0.05 0.88 0.22 0.71 0.12 0.89 -0.03 0.87 0.11 1.00 0.04 0.63 -0.62 0.87 0.17 0.82 0.04 0.91 0.09 0.65 -0.16 -0.26 -0.15 -0.09 0.60 -0.00 -0.00
RH_5 0.02 -0.01 0.30 0.03 0.25 -0.06 0.37 -0.07 0.35 0.04 1.00 -0.08 0.26 -0.14 0.32 -0.08 0.36 -0.14 0.27 -0.05 -0.12 0.18 0.08 -0.01 0.08 -0.01 -0.01
T6 0.16 0.66 0.31 0.81 -0.02 0.68 0.08 0.66 0.26 0.63 -0.08 1.00 -0.67 0.62 0.26 0.48 0.08 0.67 0.18 0.97 -0.13 -0.57 0.16 -0.08 0.76 -0.01 -0.01
RH_6 -0.13 -0.60 0.24 -0.58 0.39 -0.63 0.51 -0.70 0.39 -0.62 0.26 -0.67 1.00 -0.74 0.35 -0.65 0.48 -0.73 0.39 -0.64 -0.08 0.72 0.10 0.11 -0.26 0.01 0.01
T7 0.05 0.84 0.04 0.66 -0.04 0.85 -0.23 0.88 -0.11 0.87 -0.14 0.62 -0.74 1.00 -0.01 0.88 -0.19 0.94 -0.06 0.64 -0.09 -0.40 -0.19 -0.11 0.48 -0.00 -0.00
RH_7 -0.08 0.16 0.80 0.24 0.69 0.20 0.83 0.06 0.89 0.17 0.32 0.26 0.35 -0.01 1.00 -0.10 0.88 0.05 0.86 0.30 -0.28 0.38 0.20 -0.00 0.65 0.00 0.00
T8 0.09 0.82 -0.01 0.57 -0.02 0.79 -0.25 0.79 -0.14 0.82 -0.08 0.48 -0.65 0.88 -0.10 1.00 -0.18 0.87 -0.13 0.51 -0.16 -0.29 -0.22 -0.07 0.41 0.00 0.00
RH_8 -0.14 0.01 0.74 0.08 0.68 0.07 0.83 -0.08 0.85 0.04 0.36 0.08 0.48 -0.19 0.88 -0.18 1.00 -0.09 0.85 0.12 -0.24 0.49 0.19 0.05 0.51 0.00 0.00
T9 0.04 0.84 0.14 0.67 0.07 0.90 -0.17 0.89 -0.02 0.91 -0.14 0.67 -0.73 0.94 0.05 0.87 -0.09 1.00 0.01 0.67 -0.15 -0.30 -0.18 -0.11 0.59 0.00 0.00
RH_9 -0.10 0.09 0.77 0.16 0.68 0.16 0.83 -0.01 0.85 0.09 0.27 0.18 0.39 -0.06 0.86 -0.13 0.85 0.01 1.00 0.22 -0.20 0.36 0.22 0.01 0.54 -0.00 -0.00
T_out 0.14 0.69 0.34 0.80 0.02 0.70 0.12 0.67 0.29 0.65 -0.05 0.97 -0.64 0.64 0.30 0.51 0.12 0.67 0.22 1.00 -0.14 -0.58 0.18 -0.07 0.79 -0.01 -0.01
Press_mm_hg -0.04 -0.14 -0.30 -0.12 -0.27 -0.18 -0.25 -0.06 -0.27 -0.16 -0.12 -0.13 -0.08 -0.09 -0.28 -0.16 -0.24 -0.15 -0.20 -0.14 1.00 -0.10 -0.24 0.04 -0.25 -0.00 -0.00
RH_out -0.19 -0.34 0.28 -0.51 0.60 -0.26 0.36 -0.38 0.34 -0.26 0.18 -0.57 0.72 -0.40 0.38 -0.29 0.49 -0.30 0.36 -0.58 -0.10 1.00 -0.18 0.08 0.04 0.02 0.02
Windspeed 0.10 -0.09 0.19 0.05 0.06 -0.11 0.25 -0.19 0.29 -0.15 0.08 0.16 0.10 -0.19 0.20 -0.22 0.19 -0.18 0.22 0.18 -0.24 -0.18 1.00 0.00 0.11 -0.01 -0.01
Visibility -0.01 -0.08 -0.02 -0.07 -0.00 -0.11 0.02 -0.11 0.01 -0.09 -0.01 -0.08 0.11 -0.11 -0.00 -0.07 0.05 -0.11 0.01 -0.07 0.04 0.08 0.00 1.00 -0.04 -0.01 -0.01
Tdewpoint 0.03 0.58 0.65 0.58 0.50 0.66 0.43 0.53 0.62 0.60 0.08 0.76 -0.26 0.48 0.65 0.41 0.51 0.59 0.54 0.79 -0.25 0.04 0.11 -0.04 1.00 0.00 0.00
rv1 -0.00 -0.00 0.00 -0.01 0.01 -0.00 0.00 0.00 0.00 -0.00 -0.01 -0.01 0.01 -0.00 0.00 0.00 0.00 0.00 -0.00 -0.01 -0.00 0.02 -0.01 -0.01 0.00 1.00 1.00
rv2 -0.00 -0.00 0.00 -0.01 0.01 -0.00 0.00 0.00 0.00 -0.00 -0.01 -0.01 0.01 -0.00 0.00 0.00 0.00 0.00 -0.00 -0.01 -0.00 0.02 -0.01 -0.01 0.00 1.00 1.00
Appliances T1 RH_1 T2 RH_2 T3 RH_3 T4 RH_4 T5 RH_5 T6 RH_6 T7 RH_7 T8 RH_8 T9 RH_9 T_out Press_mm_hg RH_out Windspeed Visibility Tdewpoint rv1 rv2
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4 −0.6

Heat map (Continued)

- All the temperature variables ranging from T1 to T9 and T_out have a positive correlation with the target variable (Appliances).
- T3, T4, T5, and T7 have a high degree of correlation with T9.
- T6 & T_Out are highly correlated (both temperatures from outside).
- RH_3 and RH_4 are highly correlated with each other.
- Visibility, Windspeed, and Press_mm_hg have low correlation values.
- For humidity sensors, there are no significantly high correlation cases (>0.9).
- Visibility, rv1, rv2, and Press_mm_hg have low correlations with the target variable(Appliances).

Model Implementation



Model Implementation(Continued)

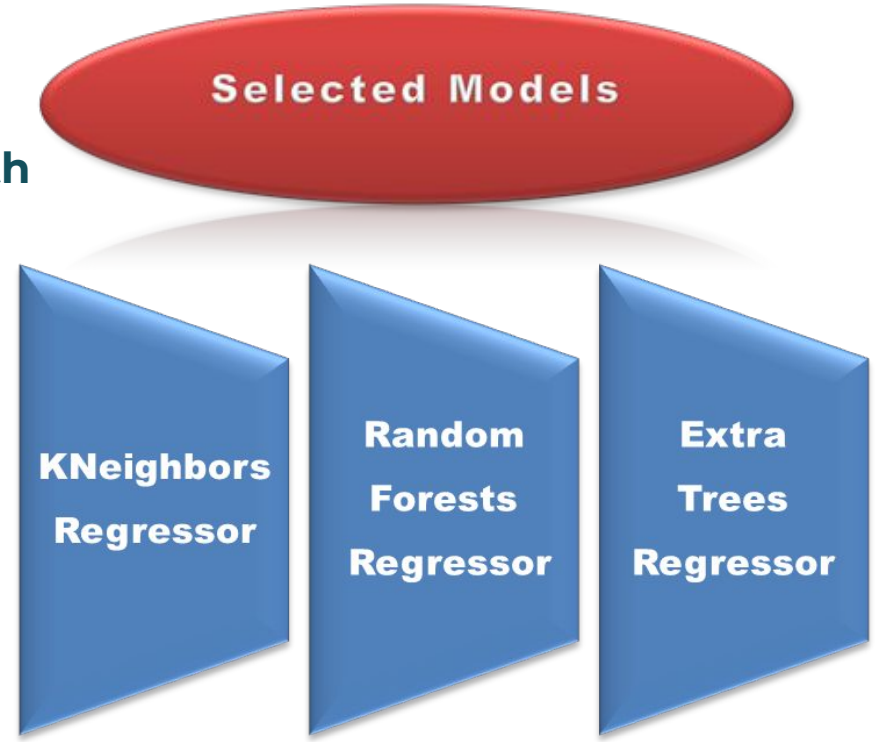
	Name	Train Time	Train R2 Score	Test R2 Score	Train Adjusted R2	Test Adjusted R2	Train MAE	Test MAE	Train RMSE Score	Test RMSE Score
0	Linear Regressor:	0.022841	0.197905	0.187610	0.196825	0.183215	30.090130	29.861799	47.040498	46.906342
1	Lasso:	0.032948	0.164987	0.155447	0.163863	0.150878	30.416715	30.188132	47.996052	47.825835
2	Ridge:	0.008121	0.197905	0.187596	0.196824	0.183201	30.088641	29.860601	47.040506	46.906728
3	KNeighbors Regressor:	0.003511	0.717213	0.552062	0.716832	0.549639	14.488002	18.093998	27.931131	34.830377
4	Decision Tree Regressor:	0.378880	1.000000	0.286622	1.000000	0.282762	0.000000	21.155153	0.000000	43.955089
5	Random Forest:	23.811413	0.942368	0.599894	0.942290	0.597729	6.632859	17.152520	12.609291	32.918259
6	Gradient Boosting Regressor:	6.116991	0.366799	0.297778	0.365946	0.293979	25.699464	26.645732	41.795523	43.610037
7	Extra Trees Regressor:	6.467139	1.000000	0.628284	1.000000	0.626273	0.000354	16.254898	0.009216	31.728870

- **KNeighbors regressor is taking the least training time.**
- **Random forest regressor is taking maximum training time but performing well on the dataset.**
- **The decision tree regressor has zero training error indicating overfitting and performed poorly on the test set.**
- **Lasso, ridge, linear, and gradient boosting regressors are performing poorly on both train and test sets.**
- **Extra trees regressor has the lowest test error.**

Best performing models

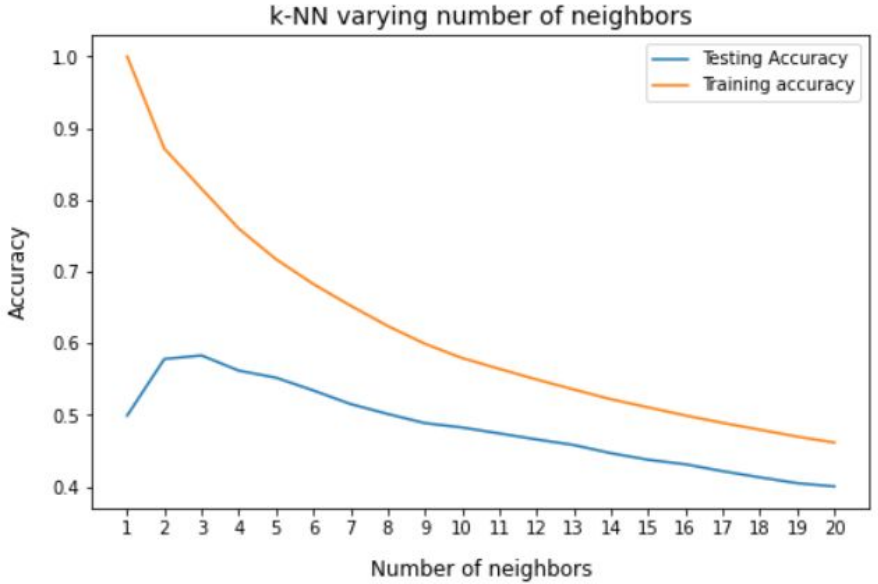
Out of 8 different models, three best performing models were selected to be trained using hyperparameter tuning with GridSearchCV.

- **KNeighbors Regressor**
- **Random Forests Regressor**
- **Extra Trees Regressor**



KNeighbors Regressor

- It is a regression based on k-nearest neighbors. The target is predicted using local interpolation of the targets associated with the nearest neighbors in the training set.
- From the line chart, it is observed that test accuracy is highest at k = 3 (i.e., in the range 1-20) after tuning for n_neighbors.



KNeighbors Regressor (Continued)

After hyperparameter finetuning using GridSearchCV, we observed:

1. **Best tuned parameter are leaf_size = 10, n_neighbors = 4, p = 1 (i.e. metric is 'manhattan'), and weights = 'distance'.**
2. **Improved test accuracy:**
   - **The test set R2 score is 0.614 over 0.552, which was achieved using the untuned model.**
   - **The test set Adjusted R2 score is 0.612 over 0.549, which was achieved using the untuned model.**
3. **Improved test error:**
   - **The test set MAE score is 16.32 over 18.09, which was achieved using the untuned model.**
   - **The test set RMSE score is 32.31 over 34.83, which was achieved using the untuned model.**

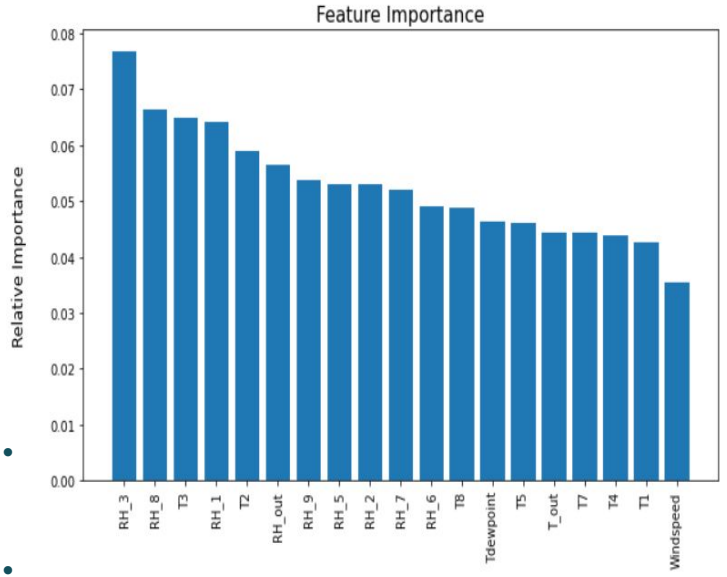
Random Forest Regressor

A random forest is an ensemble method capable of performing both classification and regression tasks with the use of multiple decision trees and a technique called bagging. After hyperparameter finetuning using GridSearchCV, we observed:

1. Best tuned parameters are 'max_depth' = 30, 'max_features' = 'sqrt', 'n_estimators' = 350.
2. Improved test accuracy:
   - The test set R2 score is 0.606 over 0.599, which was achieved using the untuned model.
   - The test set Adjusted R2 score is 0.604 over 0.597, which was achieved using the untuned model.
3. Improved test error:
   - The test set MAE score is 16.93 over 17.15, which was achieved using the untuned model.
   - The test set RMSE score is 32.64 over 32.91, which was achieved using the untuned model.

Feature Importance

1. No improvement in test accuracy over the tuned model:
   - ❑ The test set R2 score of the tuned model 0.606 has come down to 0.594.
   - ❑ The test set Adjusted R2 score of the tuned model 0.604 has come down to 0.592.
2. No improvement in test error over the tuned model:
   - ❑ The test set MAE score is 17.20 over 16.93, which was achieved using the tuned model.
   - ❑ The test set RMSE score is 33.14 over 32.64, which was achieved using the tuned model.
3. Feature reduction was not able to add to better scores.



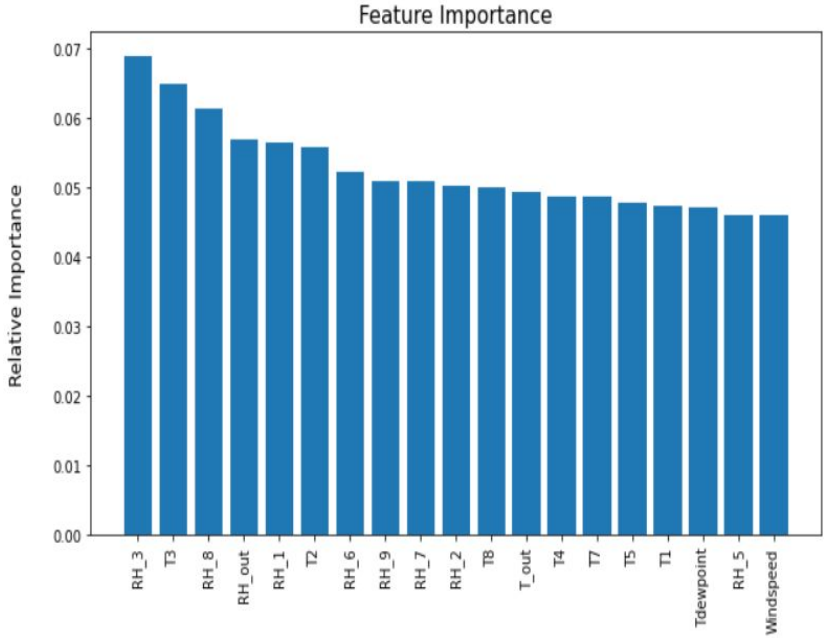
Extra-Trees Regressor

The extra tree can often achieve a better performance than the random forest. It builds an ensemble of the regression trees according to the classical top-down procedure. After hyperparameter finetuning using GridSearchCV, we observed:

1. Best tuned parameter are max_depth = 80, max_features = 'sqrt', and n_estimators = 300.
2. Improved test accuracy:
   - ❑ The test set R2 score is 0.63 over 0.62, which was achieved using the untuned model.
   - ❑ The test set Adjusted R2 score is 0.63 over 0.62, which was achieved using the untuned model.
3. Improved test error:
   - ❑ The test set MAE score is 16.12 over 16.25, which was achieved using the untuned model.
   - ❑ The test set RMSE score is 31.56 over 31.72, which was achieved using the untuned model.

Feature Importance

1. Improved test accuracy over the tuned model:
   - ❑ The test set R2 score is 0.6322 over 0.6321, which was achieved using the tuned model.
   - ❑ The test set Adjusted R2 score is 0.6307 over 0.6301, which was achieved using the tuned model.
2. Improved test error over the tuned model:
   - ❑ The test set RMSE score is 31.55 over 31.56, which was achieved using the tuned model.
3. Feature reduction was able to add to better scores.



Comparing models performance

	Regressor	Train R2 Score	Test R2 Score	Train Adjusted R2	Test Adjusted R2	Train MAE	Test MAE	Train RMSE Score	Test RMSE Score	Best Hyperparameters
0	KNeighbors Regressor	1.000000	0.614368	1.000000	0.612282	0.000000	16.323268	0.000000	32.317335	{'leaf_size': 10, 'n_neighbors': 4, 'p': 1, 'w...
1	Random Forest Regressor	0.945578	0.606429	0.945505	0.604300	6.507231	16.936468	12.253098	32.648306	{'max_depth': 30, 'max_features': 'sqrt', 'n_e...
2	Extra Trees Regressor	0.999998	0.632132	0.999998	0.630142	0.009419	16.125771	0.076480	31.564212	{'max_depth': 80, 'max_features': 'sqrt', 'n_e...

- **It is evident from the above compiled dataframe that the extra trees regressor is performing better on the test set with the highest accuracy and least error as compared to other models.**

Conclusion

- **Performance of selected models:**

  (i) KNeighbors Regressor

  - ❑ **The untuned model was able to explain 55% of the variance on the test set.**
  - ❑ **The tuned model was able to explain 61% of the variance on the test set, which is an improvement of 6%.**

  (ii) Random Forest Regressor

  - ❑ **The untuned model was able to explain 59% of the variance on the test set.**
  - ❑ **The tuned model was able to explain 60% of the variance on the test set, which is an improvement of 1%.**

Conclusion (Continued)

(iii) Extra trees Regressor

- ❑ **The untuned model was able to explain 62% of the variance on the test set.**
- ❑ **The tuned model was able to explain 63% of the variance on the test set, which is an improvement of 1%.**

- **Feature reduction was able to improve the R2 score, adjusted R2 score, and RMSE score in the extra trees regressor.**
- **The final model had 20 features.**
- **The best algorithm to use for this dataset is the extra trees regressor, as the best results for the test set are given by this regressor with the R2 score of 0.632, Adjusted R2 score of 0.63, least MAE score of 16.12, and least RMSE score of 31.56.**


THANK YOU