ASC61013 Data Modelling and Machine Intelligence

Course Work – Solar Power Consumption and Generation

Ankur Singh Gulia

Registration Number: - 220181437

Content

1. Introduction to Coursework

The data set provided consists of 50,392 instances with 29 features and 1 meta – attribute in the raw CSV data file named CW_dataset.

The features in the raw CSV datasets are related to the effects on solar power generation and consumption of the energy generated. The meta-attribute present in the raw dataset is time which is spread over 7 days starting from 1st Jan,2016 to 7th Jan 2016. It depicts the energy generation and consumption varying over 7 days with dependency on several weather features like temperature, dewpoint, wind bearing, etc.

This coursework aims to analyse the data and predict the future usage of energy depending on the given data using various regression models. With the understanding of the domain, define a correlation between the features with the target feature. With the help of a PCA or hierarchical clustering analysis predicting which weather feature to keep within the study. Give out a machine learning model to predict the value of energy used from at least two features.

Also, the use of feature engineering to classify the usage of energy based on three classes: Low, Medium, and High. Also, showing which appliance uses most energy in the above categories.

Once the predictions are confirmed, cross-validation of the machine learning pipeline is supposed to be the next step. The application of learning curves and classification evaluation metrics, prove that the pipeline is effective at preventing underfitting as well as overfitting.

Once the learning curve is complete, the understanding of the mathematical peculiarities for different models used can be described along with the reason which regression and classification is best suited for this machine learning model.

2. Domain Analysis – Energy Consumption and Generation

Solar energy is the energy that is generated by the sun through the process of nuclear fusion. It is a renewable and clean source of energy that can be used to generate electricity or provide heat. Solar panels, which are made up of photovoltaic cells, are used to capture the sun's energy and convert it into usable electricity. Solar energy has the potential to provide a significant portion of the world's energy needs, and its use is growing rapidly as technology advances and costs decrease.

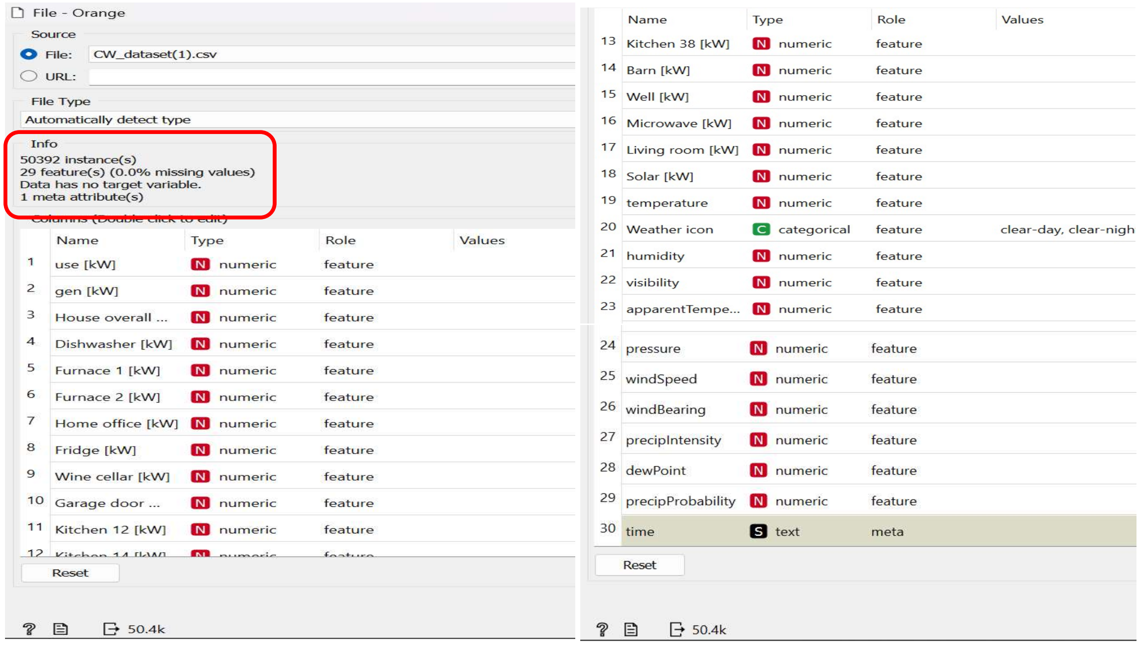
It has been more common to use solar energy now. There have been quite a few prominent factors that affect the generation for the energy. These factors that affect the generation include:

- Temperature - Measure of the thermal energy of a system or substance.
- Apparent Temperature - Measure of how hot or cold it feels to the human body based on the combination of the air temperature, relative humidity, and wind speed.
- Wind Bearing - also known as wind direction or wind direction angle, is the direction from which the wind is blowing.

- Wind Speed – A measure of how fast the wind is blowing.
- Humidity - Humidity is a measure of the amount of water vapor in the air.
- Precipitation intensity - Measure of the rate at which precipitation (e.g., rain, snow, sleet, or hail) is falling at a given moment.
- Precipitation probability - Measure of the likelihood that precipitation will occur at a given location over a specified period of time.
- Visibility - The distance that one can see clearly.
- Pressure - Measure of the force applied to a given area.
- Dew point - The dew point is the temperature at which moisture in the air will condense into liquid.

The above-mentioned factors have been used as the features to predict the model. These factors affect the generation of energy. As for the consumption, it can be relative to various things but for a house, usage can be categorised easily like furnace, kitchen, barn, well, garage, living room, etc. These categories have been taken as the features for consumption of energy which sum up to be 15 features.

According to the data, the weather features that affect the most can be figured out after the data pre-processing is complete with the help of correlation. At present the data is given out as :



Source
● File: CW_dataset(1).csv
○ URL:

File Type
Automatically detect type

Info
50392 instance(s)
29 feature(s) (0.0% missing values)
Data has no target variable.
1 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
1	use [kW]	N numeric	feature	
2	gen [kW]	N numeric	feature	
3	House overall ...	N numeric	feature	
4	Dishwasher [kW]	N numeric	feature	
5	Furnace 1 [kW]	N numeric	feature	
6	Furnace 2 [kW]	N numeric	feature	
7	Home office [kW]	N numeric	feature	
8	Fridge [kW]	N numeric	feature	
9	Wine cellar [kW]	N numeric	feature	
10	Garage door ...	N numeric	feature	
11	Kitchen 12 [kW]	N numeric	feature	
12	Kitchen 14 [kW]	N numeric	feature	
13	Kitchen 38 [kW]	N numeric	feature	
14	Barn [kW]	N numeric	feature	
15	Well [kW]	N numeric	feature	
16	Microwave [kW]	N numeric	feature	
17	Living room [kW]	N numeric	feature	
18	Solar [kW]	N numeric	feature	
19	temperature	N numeric	feature	
20	Weather icon	C categorical	feature	clear-day, clear-nigh
21	humidity	N numeric	feature	
22	visibility	N numeric	feature	
23	apparentTempe...	N numeric	feature	
24	pressure	N numeric	feature	
25	windSpeed	N numeric	feature	
26	windBearing	N numeric	feature	
27	precipIntensity	N numeric	feature	
28	dewPoint	N numeric	feature	
29	precipProbability	N numeric	feature	
30	time	S text	meta	

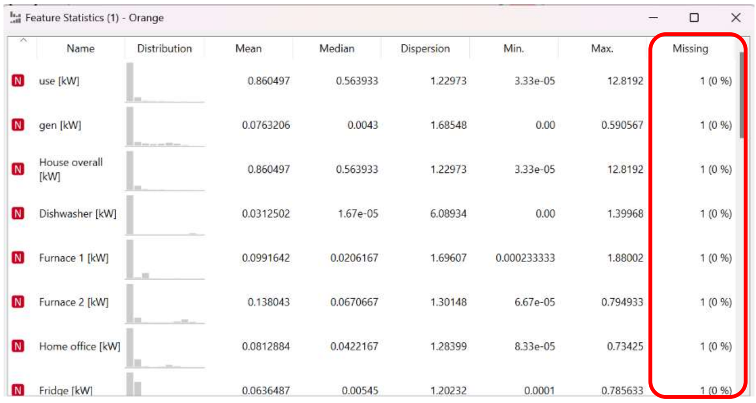
Figure 1 Data Screening

From the quick pre analysis and understanding of the domain, it can be stated that the weather features affecting the generation of the energy would be Pressure and temperature whereas for the usage, the features to be considered are Pressure, dewpoint and apparent temperature.

But this is only initial layout of the analysis based on the domain analysis. It may be very well that some of the features are cleared to be not of use in the in depth analysis.

3. Data Pre-processing

The dataset generally has certain unwanted or non-interactive data which does not allow the pipeline to work in accordance with the model which is required to work the dataset and features.



Feature Statistics (1) - Orange

	Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
N	use [kW]		0.860497	0.563933	1.22973	3.33e-05	12.8192	1 (0 %)
N	gen [kW]		0.0763206	0.0043	1.68548	0.00	0.590567	1 (0 %)
N	House overall [kW]		0.860497	0.563933	1.22973	3.33e-05	12.8192	1 (0 %)
N	Dishwasher [kW]		0.0312502	1.67e-05	6.08934	0.00	1.39968	1 (0 %)
N	Furnace 1 [kW]		0.0991642	0.0206167	1.69607	0.000233333	1.88002	1 (0 %)
N	Furnace 2 [kW]		0.138043	0.0670667	1.30148	6.67e-05	0.794933	1 (0 %)
N	Home office [kW]		0.0812884	0.0422167	1.28399	8.33e-05	0.73425	1 (0 %)
N	Fridge [kW]		0.0636487	0.00545	1.20232	0.0001	0.785633	1 (0 %)

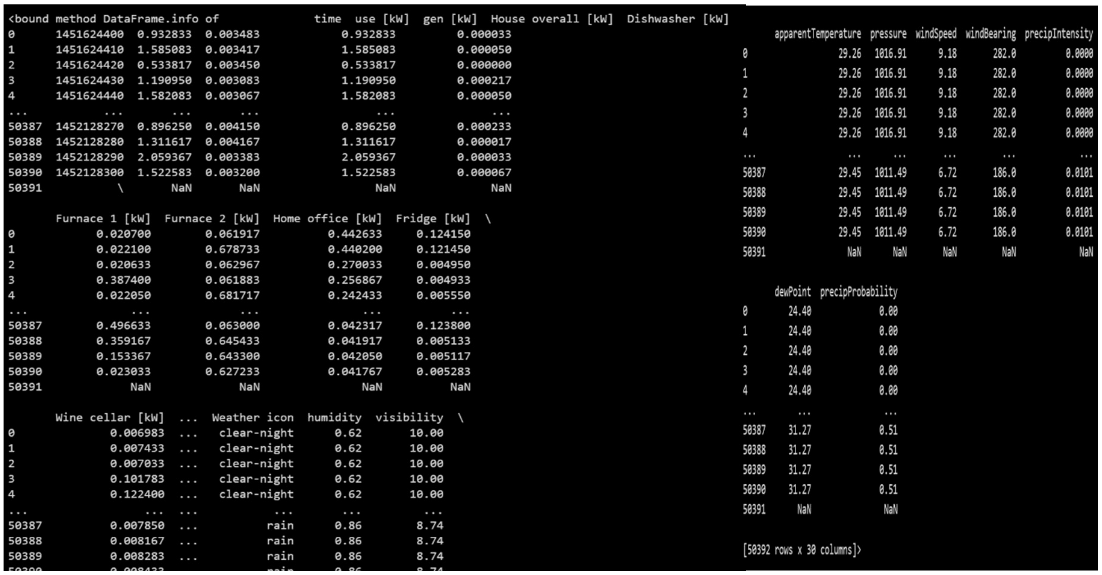
Figure 2 Feature Statistics

For the dataset provided for this course work, by understanding the dataset through feature statistics, It was clear that there were certain unknown values as shown in the fig 2.

With the help of the impute widget, we can clear out the missing values so that it may not affect the pipeline late on while the regression has to be performed.

Also, the meta-attribute Time is irrelevant to the energy generation. It is not necessary to edit the domain to convert this attribute into a feature.

The current data description is given Figure 3.



```
<bound method DataFrame.info of             time  use [kW]  gen [kW]  House overall [kW]  Dishwasher [kW]
0      1451624400  0.932833  0.003483            0.932833         0.000033
1      1451624410  1.585083  0.003417            1.585083         0.000050
2      1451624420  0.533817  0.003450            0.533817         0.000000
3      1451624430  1.190950  0.003083            1.190950         0.000217
4      1451624440  1.582083  0.003067            1.582083         0.000050
...           ...       ...       ...                 ...              ...
50387  1452128270  0.896250  0.004150            0.896250         0.000233
50388  1452128280  1.311617  0.004167            1.311617         0.000017
50389  1452128290  2.059367  0.003383            2.059367         0.000033
50390  1452128300  1.522583  0.003200            1.522583         0.000067
50391           \       NaN       NaN                 NaN              NaN

       Furnace 1 [kW]  Furnace 2 [kW]  Home office [kW]  Fridge [kW]  \
0            0.020700        0.061917          0.442633     0.124150
1            0.022100        0.678733          0.440200     0.121450
2            0.020633        0.062967          0.270033     0.004950
3            0.387400        0.061883          0.256867     0.004933
4            0.022050        0.681717          0.242433     0.005550
...               ...             ...               ...          ...
50387        0.496633        0.063000          0.042317     0.123800
50388        0.359167        0.645433          0.041917     0.005133
50389        0.153367        0.643300          0.042050     0.005117
50390        0.023033        0.627233          0.041767     0.005283
50391             NaN             NaN               NaN          NaN

       Wine cellar [kW]  ...  Weather icon  humidity  visibility  \
0              0.006983  ...   clear-night      0.62       10.00
1              0.007433  ...   clear-night      0.62       10.00
2              0.007033  ...   clear-night      0.62       10.00
3              0.101783  ...   clear-night      0.62       10.00
4              0.122400  ...   clear-night      0.62       10.00
...                 ...  ...           ...       ...         ...
50387          0.007850  ...          rain      0.86        8.74
50388          0.008167  ...          rain      0.86        8.74
50389          0.008283  ...          rain      0.86        8.74

       apparentTemperature  pressure  windSpeed  windBearing  precipIntensity
0                    29.26   1016.91       9.18        282.0           0.0000
1                    29.26   1016.91       9.18        282.0           0.0000
2                    29.26   1016.91       9.18        282.0           0.0000
3                    29.26   1016.91       9.18        282.0           0.0000
4                    29.26   1016.91       9.18        282.0           0.0000
...                    ...       ...        ...          ...              ...
50387                29.45   1011.49       6.72        186.0           0.0101
50388                29.45   1011.49       6.72        186.0           0.0101
50389                29.45   1011.49       6.72        186.0           0.0101
50390                29.45   1011.49       6.72        186.0           0.0101
50391                  NaN       NaN        NaN          NaN              NaN

       dewPoint  precipProbability
0         24.40               0.00
1         24.40               0.00
2         24.40               0.00
3         24.40               0.00
4         24.40               0.00
...         ...                ...
50387     31.27               0.51
50388     31.27               0.51
50389     31.27               0.51
50390     31.27               0.51
50391       NaN                NaN

[50392 rows x 30 columns]>
```

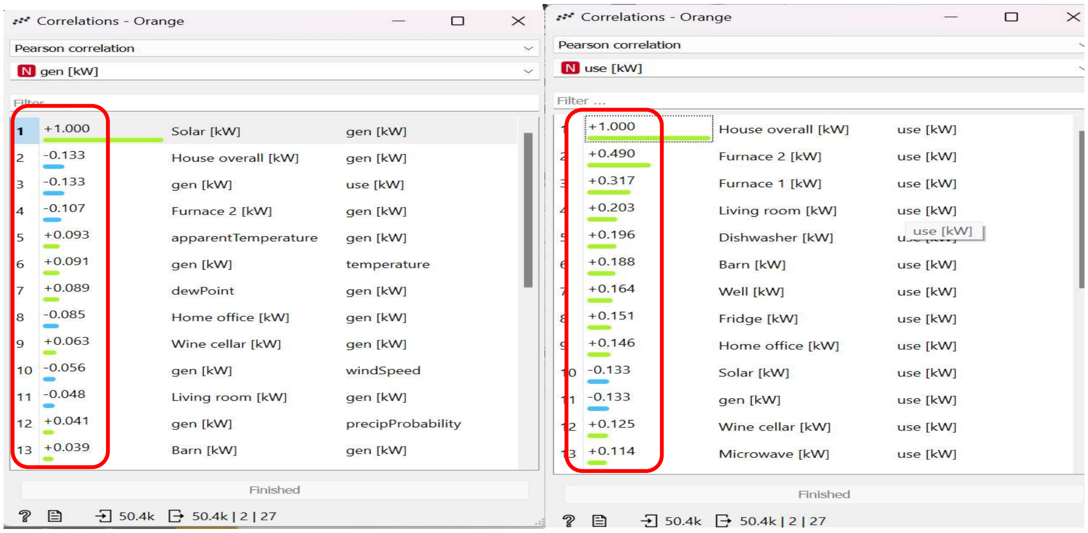
`

Figure 3 Data Description

Once the data pre – processing is complete. The next step is to figure out the correlation between target feature and the rest of features.

4. Correlation

Machine learning models can be classified as good or based on the data that has been provided. The selection features contribute most to the quality of resulting model. The selection of such features that helps in predicting the variable more accurately is known as Feature selection.



Correlations - Orange

Pearson correlation

N gen [kW]

Filter

1	+1.000	Solar [kW]	gen [kW]
2	-0.133	House overall [kW]	gen [kW]
3	-0.133	gen [kW]	use [kW]
4	-0.107	Furnace 2 [kW]	gen [kW]
5	+0.093	apparentTemperature	gen [kW]
6	+0.091	gen [kW]	temperature
7	+0.089	dewPoint	gen [kW]
8	-0.085	Home office [kW]	gen [kW]
9	+0.063	Wine cellar [kW]	gen [kW]
10	-0.056	gen [kW]	windSpeed
11	-0.048	Living room [kW]	gen [kW]
12	+0.041	gen [kW]	precipProbability
13	+0.039	Barn [kW]	gen [kW]

Finished

50.4k 50.4k | 2 | 27

Correlations - Orange

Pearson correlation

N use [kW]

Filter ...

1	+1.000	House overall [kW]	use [kW]
2	+0.490	Furnace 2 [kW]	use [kW]
3	+0.317	Furnace 1 [kW]	use [kW]
4	+0.203	Living room [kW]	use [kW]
5	+0.196	Dishwasher [kW]	use [kW]
6	+0.188	Barn [kW]	use [kW]
7	+0.164	Well [kW]	use [kW]
8	+0.151	Fridge [kW]	use [kW]
9	+0.146	Home office [kW]	use [kW]
10	-0.133	Solar [kW]	use [kW]
11	-0.133	gen [kW]	use [kW]
12	+0.125	Wine cellar [kW]	use [kW]
13	+0.114	Microwave [kW]	use [kW]

Finished

50.4k 50.4k | 2 | 27

Figure 4 Correlation

Feature selection is crucial task which is based on configuring the relation between the available features. With help of such relationships, one can identify the features important based on the target model.

In case of Energy dataset, the correlations between features are given out in Fig (4). The marked values display the effect of the features on the target value that the use [kW]. But simply based on the correlation, it cannot be figured out, which features to drop and which to keep.

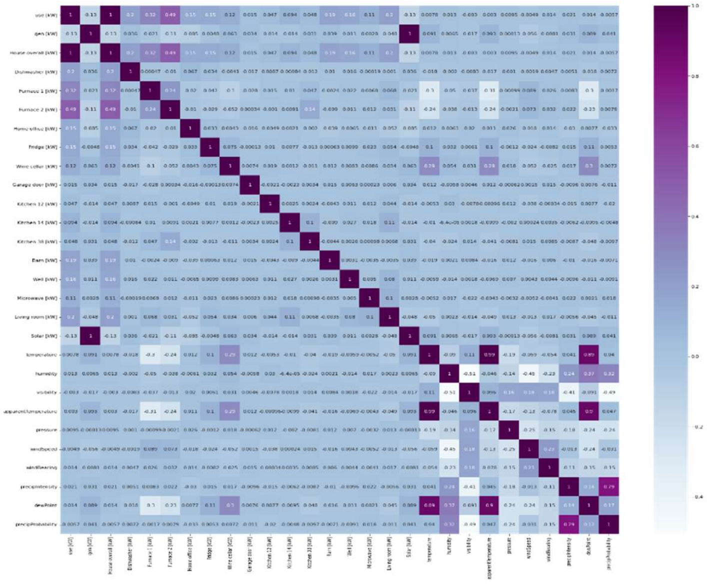


Figure 5 Heat Map for Correlation

The relational features that are to discussed here would the appliance as the weather features are to be further analysed by PCA and HCA. So the most affected features are:

- Furnace 1 and 2 [kW] - 0.489519 and0.316722 respectively.
- Living room - 0.202683.
- Dishwasher - 0.195984
- Barn - 0.187581
- Well - 0.164139

And the rest are correlations can be checked through the heat map.

For the consideration of the weather features , Principal component analysis and hierarchical clustering analysis has to taken into account. The pipeline for the has been given in the Figure 6.

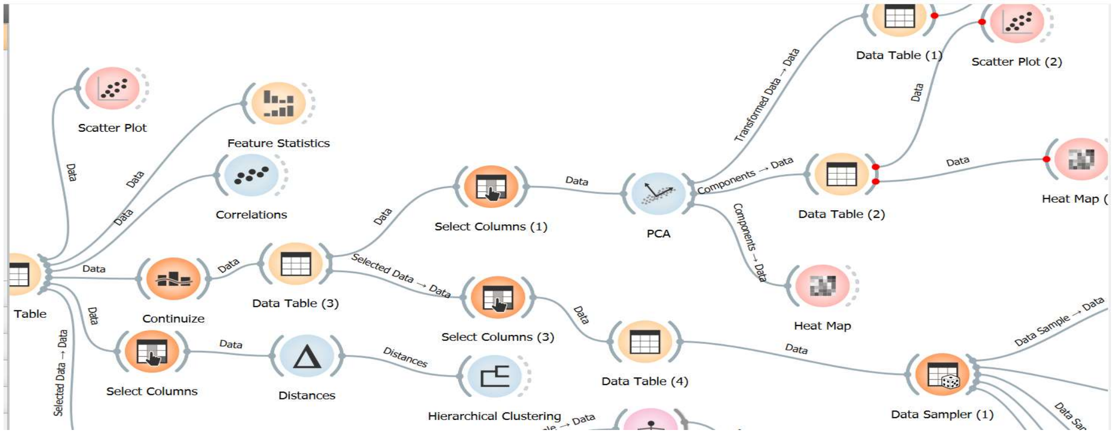
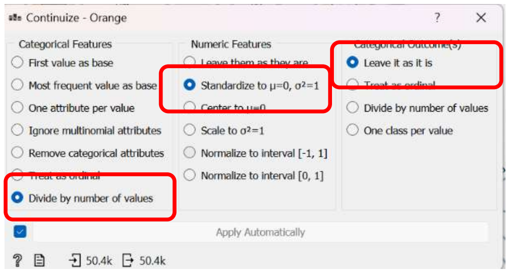


Figure 6 Pipeline for Principal Component analysis and Hierarchical Clustering analysis

From the pipeline, it is clearly understandable that to conduct to analysis, two branches are to be made. For the PCA, Continuize widget is specifically relevant as it determines the basic underlying conditions for the analysis. The condition for the same can be viewed in Figure 7.

Figure 7 Continuize


Continuize - Orange

Categorical Features
○ First value as base
○ Most frequent value as base
○ One attribute per value
○ Ignore multinomial attributes
○ Remove categorical attributes
○ Treat as ordinal
● Divide by number of values

Numeric Features
○ Leave them as they are
● Standardize to μ=0, σ²=1
○ Center to μ=0
○ Scale to σ²=1
○ Normalize to interval [-1, 1]
○ Normalize to interval [0, 1]

Categorical Outcome(s)
● Leave it as it is
○ Treat as ordinal
○ Divide by number of values
○ One class per value

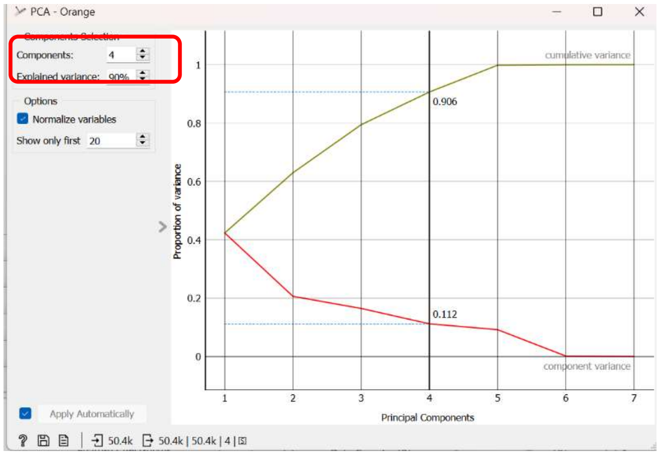
☑ Apply Automatically

50.4k 50.4k

Once the underlying condition have been determined, the relevant column, i.e., the weather features have been selected with the help of the select column widget. By the help of select column widget, only weather features have been selected to confirm the variance amongst them.

Next is to apply the PCA widget and check for the maximum variance with minimum number of components.

The variance can seen in the Figure 8. It can be seen that with only 4 components the variance is 90% which is sufficient in terms of dimensionality reduction. While reducing the PCs, the data presented with is shown in Figure 9.


PCA - Orange

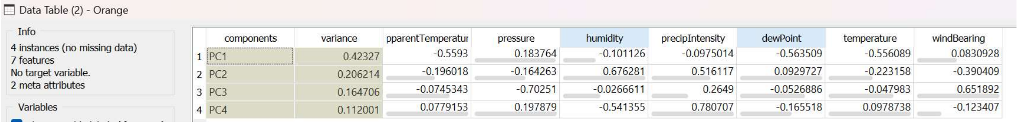
Components Selection
Components: 4
Explained variance: 90%

Options
☑ Normalize variables
Show only first 20

☑ Apply Automatically

50.4k 50.4k | 50.4k | 4

Figure 8 Principal Component Analysis


Data Table (2) - Orange

Info
4 instances (no missing data)
7 features
No target variable.
2 meta attributes

Variables

	components	variance	pparentTemperatur	pressure	humidity	precipIntensity	dewPoint	temperature	windBearing
1	PC1	0.42327	-0.5593	0.183764	-0.101126	-0.0975014	-0.563509	-0.556089	0.0830928
2	PC2	0.206214	-0.196018	-0.164263	0.676281	0.516117	0.0929727	-0.223158	-0.390409
3	PC3	0.164706	-0.0745343	-0.70251	-0.0266611	0.2649	-0.0526886	-0.047983	0.651892
4	PC4	0.112001	0.0779153	0.197879	-0.541355	0.780707	-0.165518	0.0978738	-0.123407

Figure 9 Analysed Data for PCA

The above features have been reduced as the **Wind speed, Visibility and the precipitation probability** were inversely correlated to the to the target feature. Hence, they were kept out the analysis.

Also, the heat map of the data configured by the system should give out the additional data that is required to confirm the features to be added to make impact on the regression model of the system. This heat map is shown in Figure 9.

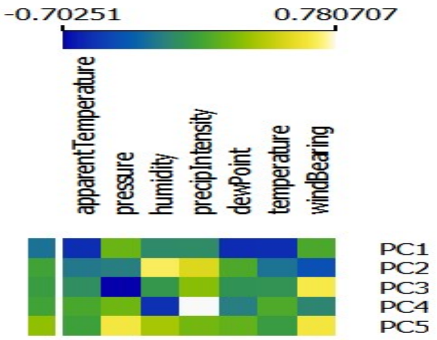
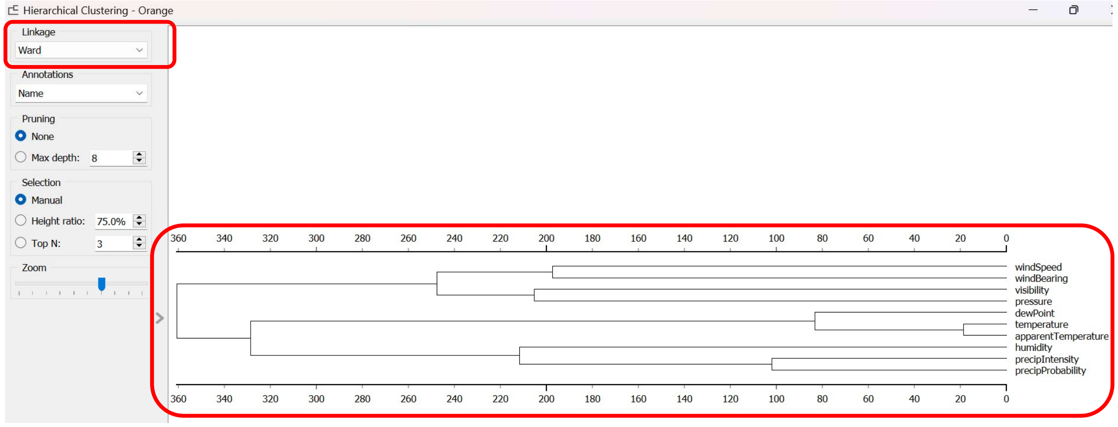


Figure 10 Heat Map for PCA

As for the Hierarchical Clustering, It is important to calculate the distances so as to confirm the relevance so that the features could be decided. Hence, the distances are calculated with the help of Distance widget wherein the metric unit is Euclidean. Once the distances are calculated, the clustering analysis is the next step. The output for the analysis is shown in Figure 10.


Hierarchical Clustering - Orange

Linkage: Ward

Annotations: Name

Pruning
● None
○ Max depth: 8

Selection
● Manual
○ Height ratio: 75.0%
○ Top N: 3

Zoom

Figure 10 Hierarchical Clustering Analysis

With the help of the PCA and Hierarchical Clustering analysis, it is clear that the features with most weightage are given as:

- Pressure
- Wind Bearing
- Dewpoint
- Precipitation Intensity

Since, these features have been selected through above analysis, they will be included in the regression model further ahead.

5. Regression and Classification Model

- Regression model

Based on the domain analysis and the dataset's correlational study, it would be preferable to use multiple regression and cross validate them to obtain an accurate result for the system/pipeline. However, using the feature engineering done on the pipeline is not required.

For the regression model to be applicable, the pipeline would look like:

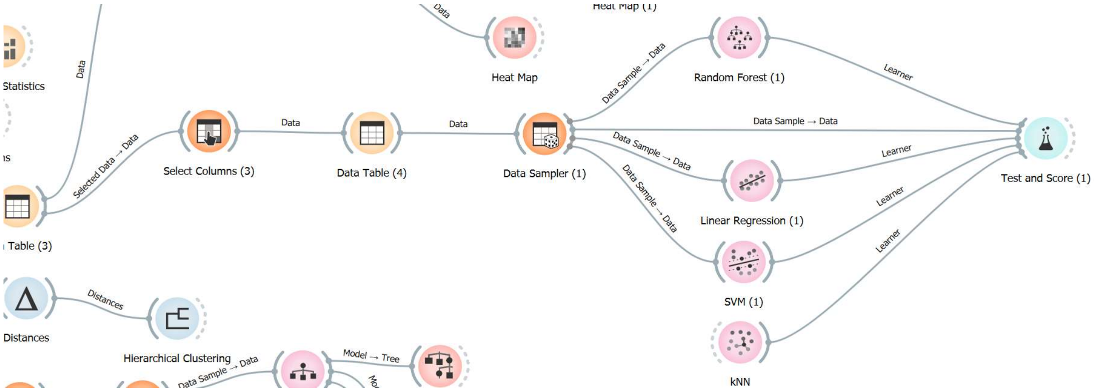


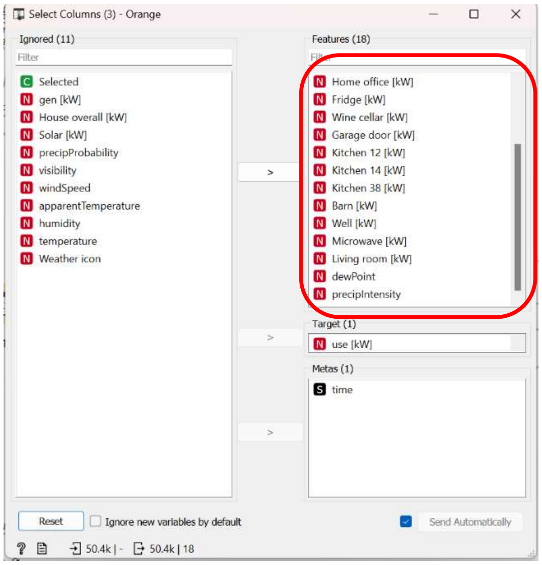
Figure 11 Regression model pipeline

As per the above pipeline, selecting the column that have an impact on the regression model based domain analysis and correlational study is the first step.

Through the PCA and HCA, the major impacting weather features have been added that can be viewed in the Figure 12.

The features that have been ignored are the result of the PCA , HCA and the correlational analsyis done earlier.

After that, all that remains is to divide the data into test and train datasets and to run the regression models through the Test and Score to determine which one is a better fit for the machine learning model.. The accuracy are visible in the Test and Score widget under the R2 column.


Select Columns (3) - Orange

Ignored (11)

Filter

- C Selected
- N gen [kW]
- N House overall [kW]
- N Solar [kW]
- N precipProbability
- N visibility
- N windSpeed
- N apparentTemperature
- N humidity
- N temperature
- N Weather icon

Features (18)

- N Home office [kW]
- N Fridge [kW]
- N Wine cellar [kW]
- N Garage door [kW]
- N Kitchen 12 [kW]
- N Kitchen 14 [kW]
- N Kitchen 38 [kW]
- N Barn [kW]
- N Well [kW]
- N Microwave [kW]
- N Living room [kW]
- N dewPoint
- N precipIntensity

Target (1)

- N use [kW]

Metas (1)

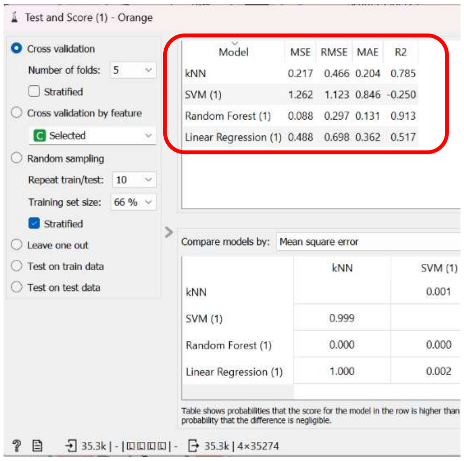
- S time

Figure 12 Select column

The Test and score stated that the random forest regression model is best fit based on the learning curve while linear regression and SVM seems to be a under fit. Although KNN seems to be a good fit as well but Random Forest regression show better accuracy. The values form orange tools are shown in Figure 13.

The above regression model results show that the Random Forest regression model is significantly more effective for the dataset based on Energy Consumption.

Next Step would be to classify the system to figure out which data lie under High, Medium and Low usage.


Test and Score (1) - Orange

● Cross validation
Number of folds: 5
☐ Stratified
○ Cross validation by feature
Selected
○ Random sampling
Repeat train/test: 10
Training set size: 66 %
☑ Stratified
○ Leave one out
○ Test on train data
○ Test on test data

Model	MSE	RMSE	MAE	R2
kNN	0.217	0.466	0.204	0.785
SVM (1)	1.262	1.123	0.846	-0.250
Random Forest (1)	0.088	0.297	0.131	0.913
Linear Regression (1)	0.488	0.698	0.362	0.517

Compare models by: Mean square error

	kNN	SVM (1)
kNN		0.001
SVM (1)	0.999	
Random Forest (1)	0.000	0.000
Linear Regression (1)	1.000	0.002

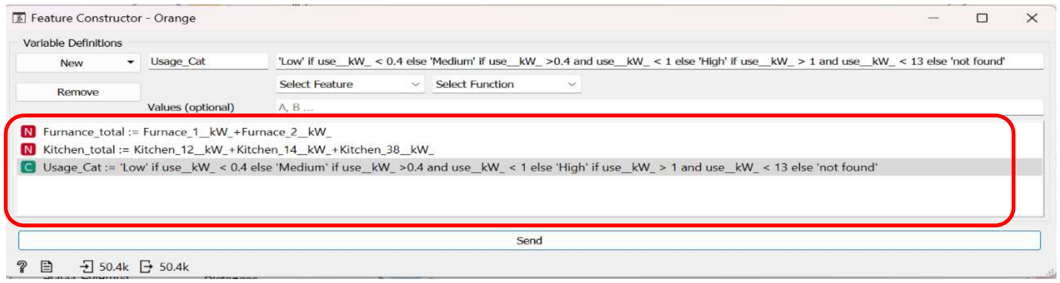
Table shows probabilities that the score for the model in the row is higher than probability that the difference is negligible.

Figure 13 Regression Model result

- Classification model

Classification models decide whether to categorise usage based on prior analysis performed through PCA, HCA, and domain analysis. However, in order to complete the classification model, feature engineering is required so that data can be studied easily.

In this case, the target feature must be converted to classification of 3: Low, Medium and High. Also, from the correlational analysis, it was observed that features Furnace 1 [kW] and Furnace 2 [kW] as well as Kitchen 12, Kitchen 14 and Kitchen 34 can be combine to a simplify the data classification. Hence, the new creations are done with the help of feature constructor widget.



Feature Constructor - Orange

Variable Definitions

New ▾ | Usage_Cat | 'Low' if use__kW_ < 0.4 else 'Medium' if use__kW_ >0.4 and use__kW_ < 1 else 'High' if use__kW_ > 1 and use__kW_ < 13 else 'not found'

Remove | Select Feature | Select Function

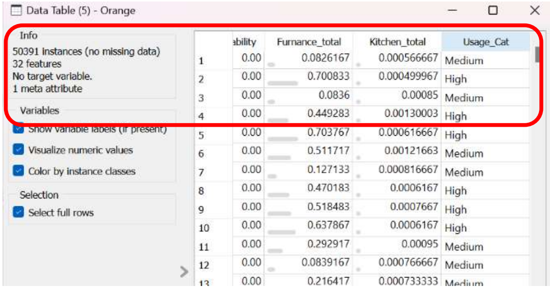
Values (optional): A, B ...

N Furnance_total := Furnace_1__kW_+Furnace_2__kW_
N Kitchen_total := Kitchen_12__kW_+Kitchen_14__kW_+Kitchen_38__kW_
C Usage_Cat := 'Low' if use__kW_ < 0.4 else 'Medium' if use__kW_ >0.4 and use__kW_ < 1 else 'High' if use__kW_ > 1 and use__kW_ < 13 else 'not found'

Send

50.4k 50.4k

Figure 14 Feature Construction



Data Table (5) - Orange

Info
50391 instances (no missing data)
32 features
No target variable.
1 meta attribute

Variables
☑ Show variable labels (if present)
☑ Visualize numeric values
☑ Color by instance classes

Selection
☑ Select full rows

	ability	Furnance_total	Kitchen_total	Usage_Cat
1	0.00	0.0826167	0.000566667	Medium
2	0.00	0.700833	0.000499967	High
3	0.00	0.0836	0.00085	Medium
4	0.00	0.449283	0.00130003	High
5	0.00	0.703767	0.000616667	High
6	0.00	0.511717	0.00121663	Medium
7	0.00	0.127133	0.000816667	Medium
8	0.00	0.470183	0.0006167	High
9	0.00	0.518483	0.0007667	High
10	0.00	0.637867	0.0006167	High
11	0.00	0.292917	0.00095	Medium
12	0.00	0.0839167	0.000766667	Medium
13	0.00	0.216417	0.000733333	Medium

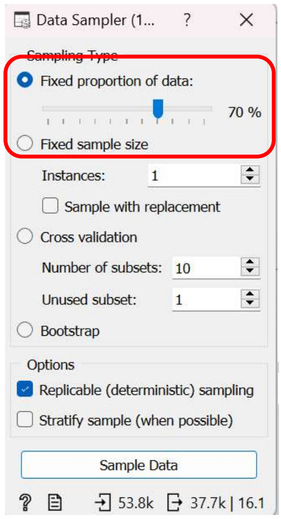
Figure 15 Data for Classification

As presented in the Figure (14), new features have been created. All that is left to do is to use a sampler on the required data and take up the Decision Tree Classifier and Neural network.

Once the required columns have been chosen, sample and test data are generated to set up the system's classifier. In this case, a 70:30 ratio is used for the current pipeline.

When the pipeline's data sampler is prepared, the next step is to use this sample data for classifiers, with the remaining data being used to test the prediction. The classifier pipeline is as follows:

Figure 17 Data Sampling



Data Sampler (1... ? ✕

Sampling Type
● Fixed proportion of data: 70 %
○ Fixed sample size
Instances: 1
□ Sample with replacement
○ Cross validation
Number of subsets: 10
Unused subset: 1
○ Bootstrap

Options
☑ Replicable (deterministic) sampling
□ Stratify sample (when possible)

Sample Data

53.8k 37.7k | 16.1

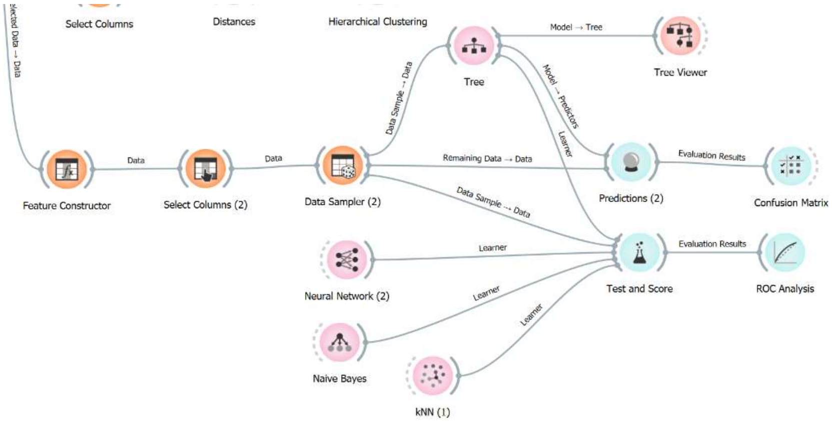
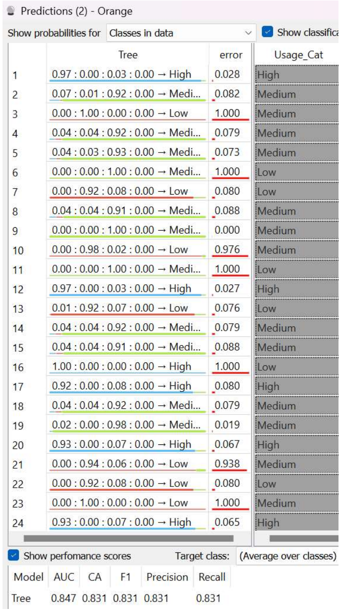


Figure 16 Classification Model Pipeline

From the above pipeline, the predictions are passed through the confusion matrix to make the sure of underfitting and over fitting for the classification.

The confusion matrix provides an accuracy to the predictions on the system so as to validate the pipeline.



Predictions (2) - Orange

Show probabilities for: Classes in data ☑ Show classifica

	Tree	error	Usage_Cat
1	0.97 : 0.00 : 0.03 : 0.00 → High	0.028	High
2	0.07 : 0.01 : 0.92 : 0.00 → Medi...	0.082	Medium
3	0.00 : 1.00 : 0.00 : 0.00 → Low	1.000	Medium
4	0.04 : 0.04 : 0.92 : 0.00 → Medi...	0.079	Medium
5	0.04 : 0.03 : 0.93 : 0.00 → Medi...	0.073	Medium
6	0.00 : 0.00 : 1.00 : 0.00 → Medi...	1.000	Low
7	0.00 : 0.92 : 0.08 : 0.00 → Low	0.080	Low
8	0.04 : 0.04 : 0.91 : 0.00 → Medi...	0.088	Medium
9	0.00 : 0.00 : 1.00 : 0.00 → Medi...	0.000	Medium
10	0.00 : 0.98 : 0.02 : 0.00 → Low	0.976	Medium
11	0.00 : 0.00 : 1.00 : 0.00 → Medi...	1.000	Low
12	0.97 : 0.00 : 0.03 : 0.00 → High	0.027	High
13	0.01 : 0.92 : 0.07 : 0.00 → Low	0.076	Low
14	0.04 : 0.04 : 0.92 : 0.00 → Medi...	0.079	Medium
15	0.04 : 0.04 : 0.91 : 0.00 → Medi...	0.088	Medium
16	1.00 : 0.00 : 0.00 : 0.00 → High	1.000	Low
17	0.92 : 0.00 : 0.08 : 0.00 → High	0.080	High
18	0.04 : 0.04 : 0.92 : 0.00 → Medi...	0.079	Medium
19	0.02 : 0.00 : 0.98 : 0.00 → Medi...	0.019	Medium
20	0.93 : 0.00 : 0.07 : 0.00 → High	0.067	High
21	0.00 : 0.94 : 0.06 : 0.00 → Low	0.938	Medium
22	0.00 : 0.92 : 0.08 : 0.00 → Low	0.080	Low
23	0.00 : 1.00 : 0.00 : 0.00 → Low	1.000	Medium
24	0.93 : 0.00 : 0.07 : 0.00 → High	0.065	High

☑ Show perfomance scores Target class: (Average over classes)

Model	AUC	CA	F1	Precision	Recall
Tree	0.847	0.831	0.831	0.831	0.831

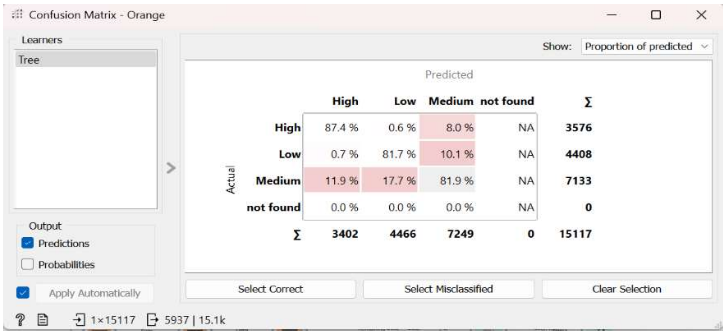
Figure 18 Predictions

The Prediction shows the individual instance's predicted classification based on the pipeline into the three categories created under the feature Usage_Cat with presumptuous error.

Also, the precision for the classifier can also be seen in the same widget's window. As per the pipeline created, the system precision is given as 0.831

The prediction only classifies data, it is not capable of visualising it. As a result, a confusion matrix is used to provide the exact instances for the classification in terms of numbers and percentages.

The study of confusion matrix can be easily suggest out that accuracy of the pipeline and gives a better visualization of the dataset classification.



Confusion Matrix - Orange

Learners: Tree

Show: Proportion of predicted

Output: ☑ Predictions ☐ Probabilities

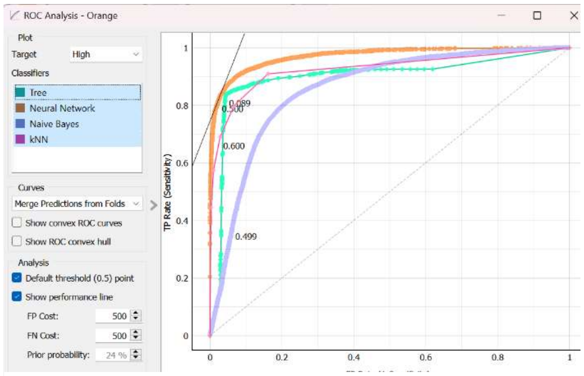
☑ Apply Automatically

Predicted

Actual	High	Low	Medium	not found	Σ
High	87.4 %	0.6 %	8.0 %	NA	3576
Low	0.7 %	81.7 %	10.1 %	NA	4408
Medium	11.9 %	17.7 %	81.9 %	NA	7133
not found	0.0 %	0.0 %	0.0 %	NA	0
Σ	3402	4466	7249	0	15117

Figure 19 Confusion Matrix

A graphial representation for all the classifiers applied to the machine learning model to deduce which is more effective for the system. The ROC analysis shows the result in Figure 20.



ROC Analysis - Orange

Plot
Target: High
Classifiers: Tree, Neural Network, Naive Bayes, kNN

Curves: Merge Predictions from Folds
☐ Show convex ROC curves
☐ Show ROC convex hull

Analysis
☑ Default threshold (0.5) point
☑ Show performance line
FP Cost: 500
FN Cost: 500
Prior probability: 24 %

Figure 20 ROC Analysis`

It is evident from the analysis that the Neural network is better option for the system as the output of the analysis shows that the curve for the neural network to be better.

- Model Complexity

Model complexity refers to the amount of detail or "wiggle room" that is built into a machine learning model. A model with high complexity will have many parameters or variables that can be adjusted, allowing it to capture a wide range of patterns and relationships in the data. For the dataset present and the models used the complexity given are below:

Neural networks, decision trees, and naive bayes are all machine learning algorithms that can be used for classification tasks. The complexity of each of these models can vary depending on a number of factors.

Neural networks are complex machine learning models that are composed of many interconnected nodes, or "neurons." The complexity of a neural network can be measured by the number of neurons and the number of connections, or weights, between them. As the number of neurons and connections increases, the complexity of the model also increases.

Decision trees are relatively simple models that consist of a series of branching decision rules As the number of decision rules and the depth of the tree increase, the complexity of the model also increases.

Naive bayes is a relatively simple probabilistic model that is based on the assumption of conditional independence among the features in the data. As the number of features and the number of possible values increases, the complexity of the model also increases.

K-nearest neighbors (kNN) is a relatively simple non-parametric model that is based on the concept of similarity between examples in the data. As the number of examples and features increases, the complexity of the model also increases.

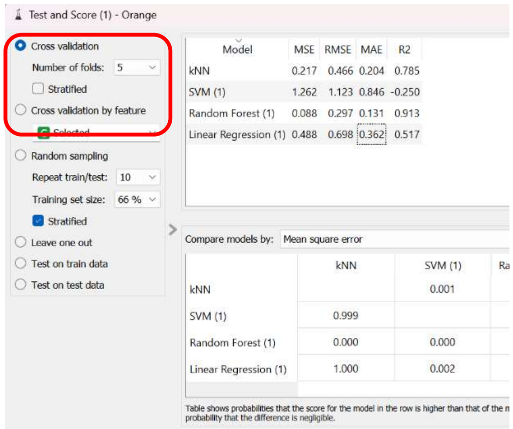
6. Cross Validation

For a Machine learning model, a pipeline is created based on the requirement. But to actually verify the accuracy and the legitimacy of the model, cross validation is done. Essentially, it is method that evaluates and compares the learning algorithm by dividing data in two segments: one to train the model and the other one is used to test the predictive capabilities if the model such as accuracy, error, etc.

For this case, within the test and score result, it can be cross validated with help of K – fold cross validation where in the training data and testing data are split into multiple sets. Then for each set, the regression will take place and the average for all the regression will be shown as the final output.

The second method is by import the cross validation score from python. Which show the result for linear regression and random forest regression, SVM and kNN as:

As shown in Figure 21, cross vaildation for number of folds is given as 5 and the result shows that the random forest regressor is the best fit for the system.



Test and Score (1) - Orange

● Cross validation
Number of folds: 5
☐ Stratified
○ Cross validation by feature
Selected
○ Random sampling
Repeat train/test: 10
Training set size: 66 %
☑ Stratified
○ Leave one out
○ Test on train data
○ Test on test data

Model	MSE	RMSE	MAE	R2
kNN	0.217	0.466	0.204	0.785
SVM (1)	1.262	1.123	0.846	-0.250
Random Forest (1)	0.088	0.297	0.131	0.913
Linear Regression (1)	0.488	0.698	0.362	0.517

Compare models by: Mean square error

	kNN	SVM (1)	Ra
kNN		0.001	
SVM (1)	0.999		
Random Forest (1)	0.000	0.000	
Linear Regression (1)	1.000	0.002	

Table shows probabilities that the score for the model in the row is higher than that of the m
probability that the difference is negligible.

Figure 21 K fold Validation

The validation prevents bias and is critical in maintaining the dataset required for the system to function.

It has an effect on the learning model because the test and train datasets must always be kept separ ate and have the exact ratio specified during time sampling.

7. Learning Curves – Underfitting and Overfitting

Learning curves are the plots that represent the performance and the impact of the model at the training data set increases. The two major causes for error in the pipeline presented are:

Bias: it describes the model so that it makes simpler assumptions such that, for the model, it is easier to approximate.

Variance: It describes the variability of the model's prediction with the change in dataset used for training the model.

What the learning curves helps in representing the exact regressor to be used for prediction as the system requires a good fit.

For this case, 3 regressors were used and trained on the same data sets. The regressors used were Linear regression, Random Forest regression and Neural Network. The plots are shown below:

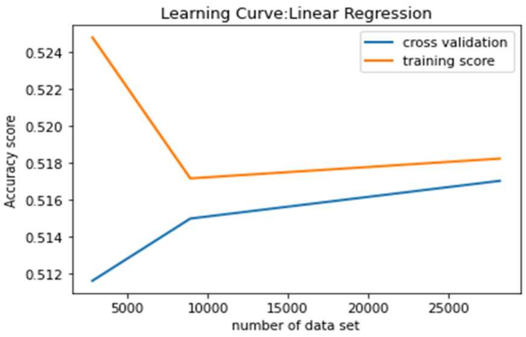


Figure 22 Learning curve for Linear regression

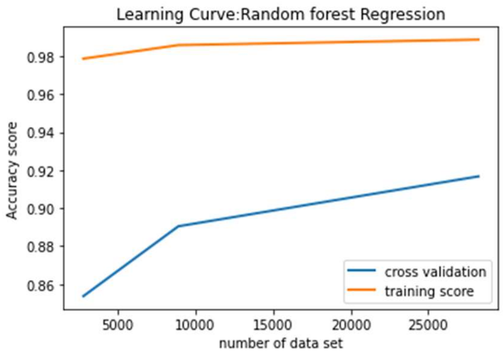


Figure 23 Learning curve for Random Forest regression

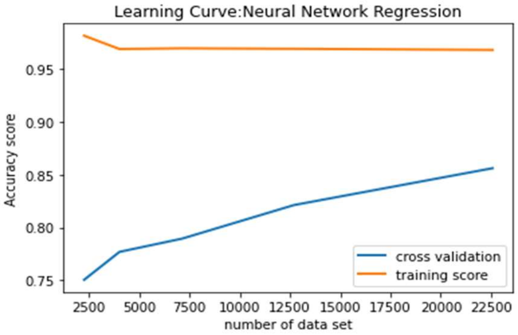


Figure 24 Learning Curve for Neural network

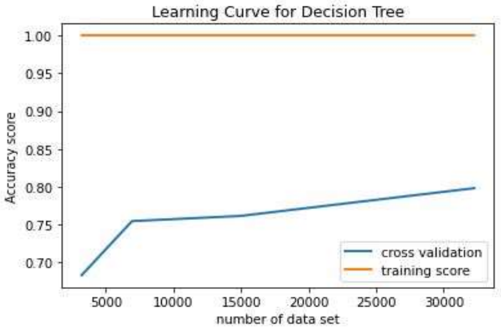


Figure 25 Learning Curve for Decision tree

The reason for choosing the Random - forest regression as fit for the machine learning model can be easily deduced from the above plots for all the regression models.

For the learning curve of linear regression, the result output can be easily justified as the accuracy of the system not at par with the required model accuracy for machine learning model.

The neural network and random forest both have a good fit or the machine learning model, but the random forest seems to have the upper hand for cross validation curve. Hence, the better fit regressor for the system can be concluded to be the random forest.

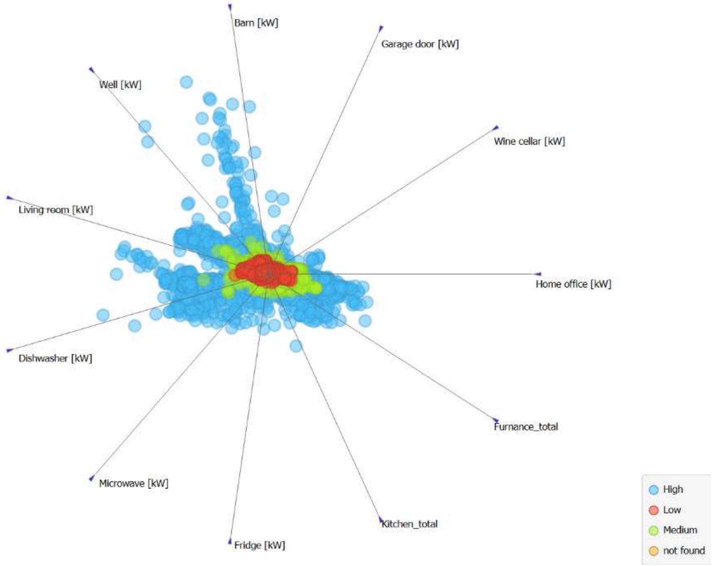


Figure 25 Linear projection

As per the data set used, once the classification is complete the dominance with highest used appliance in the three categories used can be studied with the help of the Linear projection widget. Applying the widget and using the dataset given that only includes the appliances. The projection is given in Figure 26.

The linear projection is a division of all the appliance based on the usage categories classified earlier by the feature engineering and along with that the study of the tree suggested that following appliance were being used during the prediction of the following categories:

- High

1. Barn [kW]

2. Well [kW]

- Medium

1. Living room [kW]

2. Home Office [kW]

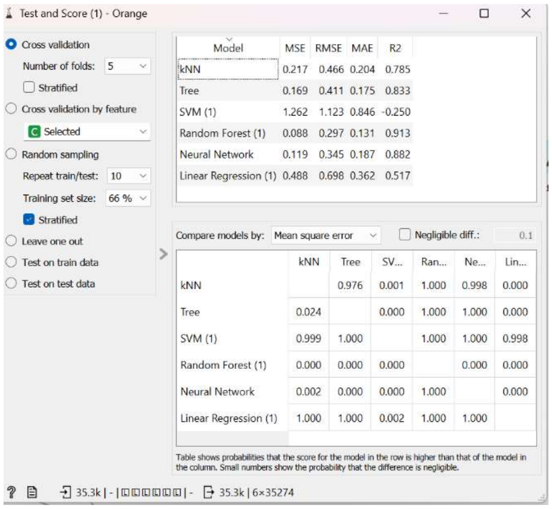
- Low

1. Dishwasher [kW]

8. Mathematical Peculiarity for Machine learning Model

Mathematical Peculiarity for the system can be studied with the help of the output of the system and can be judged based on the accuracies as well the errors that the model present.

To study the peculiarity for the regression and classification models, we need a test and score with all the results of the process taken under single widget.



● Cross validation
Number of folds: 5
☐ Stratified
○ Cross validation by feature
Selected
○ Random sampling
Repeat train/test: 10
Training set size: 66 %
☑ Stratified
○ Leave one out
○ Test on train data
○ Test on test data

Model	MSE	RMSE	MAE	R2
kNN	0.217	0.466	0.204	0.785
Tree	0.169	0.411	0.175	0.833
SVM (1)	1.262	1.123	0.846	-0.250
Random Forest (1)	0.088	0.297	0.131	0.913
Neural Network	0.119	0.345	0.187	0.882
Linear Regression (1)	0.488	0.698	0.362	0.517

Compare models by: Mean square error ☐ Negligible diff.: 0.1

	kNN	Tree	SV...	Ran...	Ne...	Lin...
kNN		0.976	0.001	1.000	0.998	0.000
Tree	0.024		0.000	1.000	1.000	0.000
SVM (1)	0.999	1.000		1.000	1.000	0.998
Random Forest (1)	0.000	0.000	0.000		0.000	0.000
Neural Network	0.002	0.000	0.000	1.000		0.000
Linear Regression (1)	1.000	1.000	0.002	1.000	1.000	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

35.3k | - | □□□□□□ | - 35.3k | 6×35274

Figure 26 Test and Score for Mathematical Peculiarity

It is clear that for linear regression the regression line always passes through the mean of the data. Also finding that the error present in the system is 0.698 which is too large for the model to be considered as a fit machine learning model.

As for the support vector machine regression, they use a kernel function to map the data from the original space into a higher-dimensional space. This allows the algorithm to find nonlinear decision boundaries that are more accurate than those found by other algorithms, such as logistic regression.

But the problem with SVM that the accuracy for the model is in negative which can never be used as a ML model.

One mathematical peculiarity of kNN is that it uses a distance metric to measure the similarity between data points. This distance metric can be Euclidean distance, Manhattan distance, or any other metric that quantifies the "distance" between two points in a given space.

For decision trees, the peculiarity is that they use a metric called entropy to measure the impurity or uncertainty of a node in the tree. Entropy is a measure of the amount of randomness or disorder in a system, and it is used to determine how well a node can be split into sub-nodes based on the values of the independent variables.

The mathematical peculiarity of random forests is that they use a technique called bootstrapping to train multiple decision trees on different subsets of the data. Bootstrapping is a sampling method that involves randomly selecting data points with replacement from the original dataset to create multiple smaller datasets.

- Evaluation matrix

For the given datasets, we have used 6 models in the pipeline. The resulting data given out is discussed in the report. Out of the six of them, but for the evaluation to be done. The matrix suggested that the SVM model could take place as the data set for Energy consumption and generation was too large and the SVM works perfectly for smaller data set.

Whereas the linear regression was not selected because of the accuracy it presented. It states that the prediction for the future use of the energy would only of 51% correct and that model, if used would have caused a lot of trouble for the pipeline.

The evaluation would have been between the rest of the models within the pipeline. kNN, Tree, Random Forest and neural network were taken into account as they were not underfit nor were they any close to being overfit. But the pipeline suggested that the better output was given by random forest of all the rest of them as it has the lowest RMSE and also the cross-validation curve is a rather better fit for random forest only.

The two models that were chosen out side the curriculum were kNN and Naive Bayes.

9. Conclusion

This course work consisted of creation of a complete pipeline for a data set based on study of 29 features of Energy consumption and generation. The process started with the domain analysis of the dataset. Once the domain analysis was complete. It gave enough information to construct a pipeline based on the features for prediction of prices of diamond and classifying them in three categories: low, Medium and High.

This process started with data cleaning and then moved forward to the system where it was required to understand which feature were to be kept and which were to be dropped. After the domain was complete. With the hep of the correlation study, PCA and HCA while keeping the target as use[kW], it was noticed that wind speed, visibility and precipitation probability had negative correlation. From the PCA and HCA, it was analysed which weather features to keep in the pipeline.

From here on, model was divided into two directions: First was creation a model that predict the usage of the energy. Second was the creation of a predictor that classified the usage of energy into three categories.

For the first part of the pipeline, a sampler was taken and the data was divided into a ratio of 70:30. The bigger set was used to train the model and the smaller set to test the regression models in place. It was through cross validation and learning curve found that the Random Forest regression model was a good fit for the machine learning model.

For the second part, feature engineer was done to categorise the class of the price and then through the sampler, a decision tree classification was done to achieve the confusion matrix that explain the prediction with error as well. Also, the pipeline showed a ROC analysis to confirm that the Tree classifier was the correct choice.

With the completion of a machine learning model, the price was predicted and the classification was done which helped in creating an effecting ML Model.