Assignment 4 – sxb180048 – Sai Pratheek Banda

ENERGY DATASET

Clustering Problem:

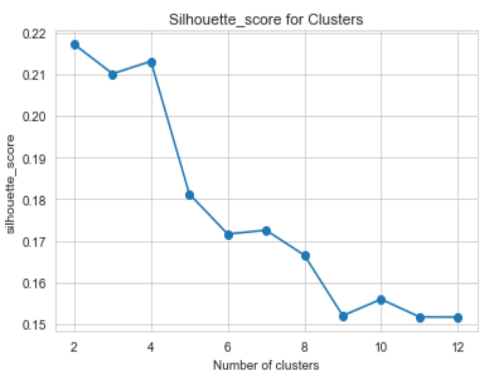
The aim of the problem is to cluster data to predict the classification of the energy consumption into high and low. The interesting part of this classification is that it will help us understand the patterns for energy consumption thus understanding why energy consumption might be low or high. **Class 0 (Low)**: 10357 Observations| **Class 1 (High)**: 8048 Observations

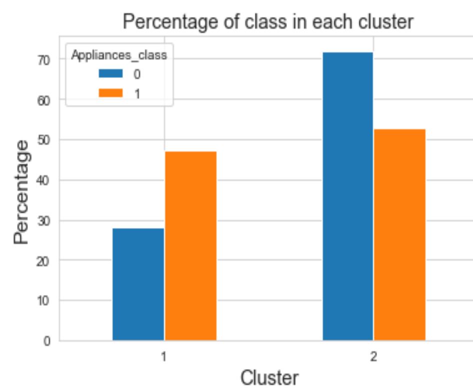
EDA:

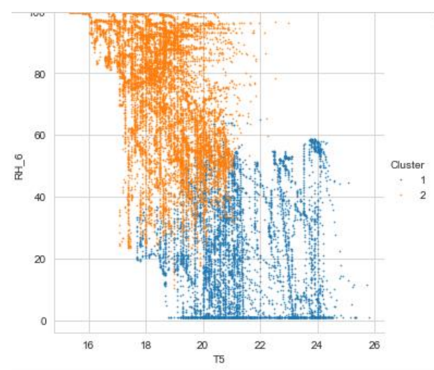
1) Features 'T_out, RH_7, T9, RH_4, T3, rv2, RH_9, T1, T7 ' were dropped due to high correlation.
2) Date, Visibility features are insignificant due to its inability to explain the target variable.
3) Lights Feature was removed due to high null values.
4) The outliers were removed based on the Inter Quartile Range.
5) Data has been Scaled for computational productivity.

CLUSTERING*:*

K Means:

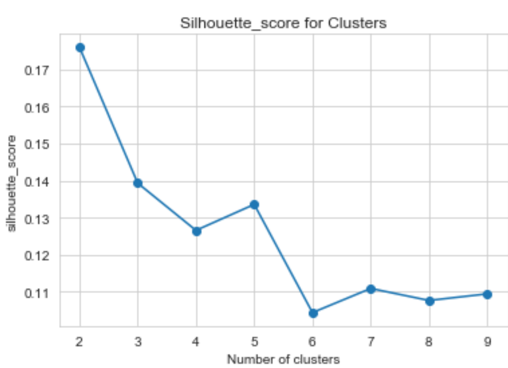


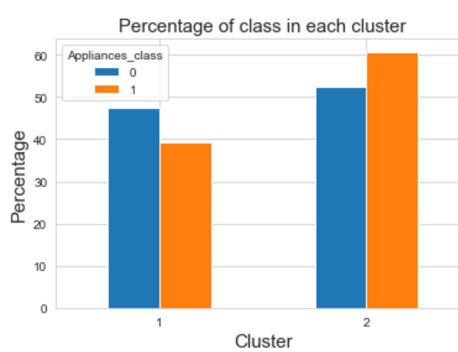


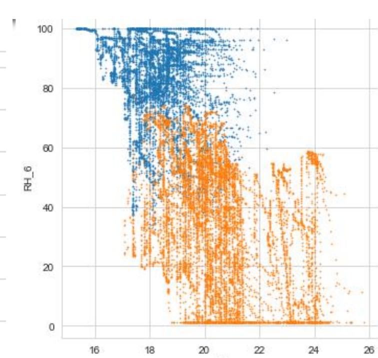


K means clustering done with multiple Clusters along which Silhoutte score was calculated for each Cluster. The optimal number of clusters found **(K)** was two.	The graph shows us the distribution of the observations in each cluster. They did not align with the class labels. The distrubution across the clusters are also not similar.	The seperation between the Clusters is not clear as you see from the graph. The clusters are not compact The 2 Clusters speration is not linear. This is due to multiple features clustering.

Expectation Maximization:



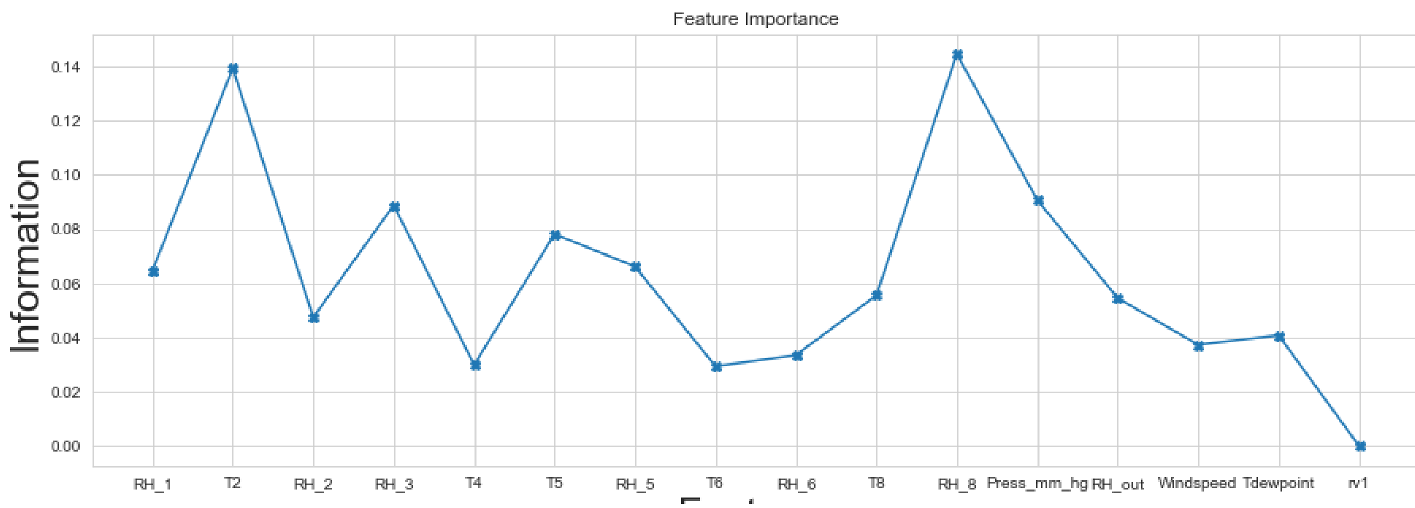




Expectation Maximization was done with Gaussian Mixture model, along with silhouette score was calculated, the **optimum number of clusters to be found was two.**	The Observation distribution follows the trajectory of the k means in the distribution of observations across clusters. They did not align with the Class labels nor within themselves	The Clustering isn't clearly separable. It clusters the **same way K means does. The clusters are not compact but spread out.** This is due to the clusters being dependent on various features.

Feature Selection:

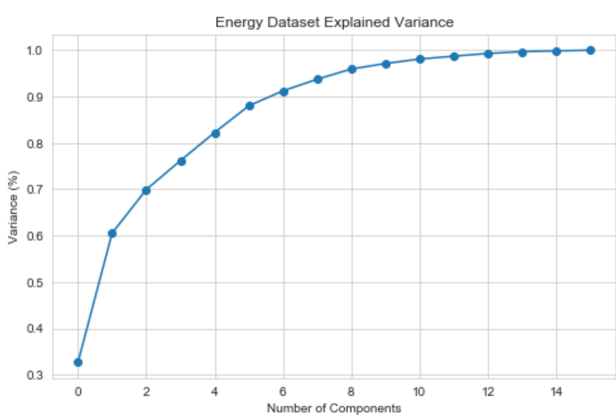
Decision Tree:



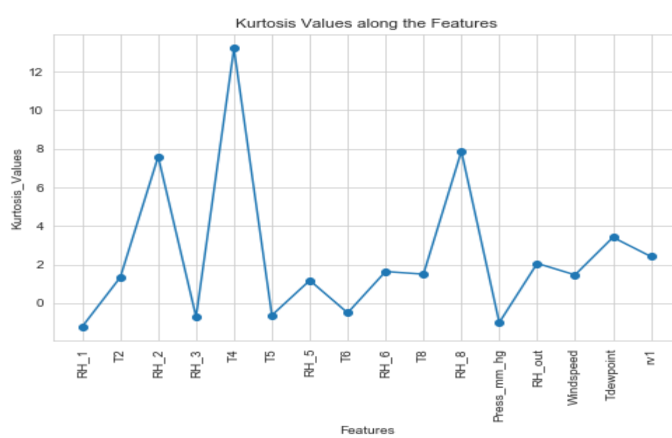
The features picked through feature Selection are 'RH_1','T2','T4','T5','RH_5','RH_6','T8', 'RH_8', 'Press_mm_hg'

Feature Reduction:

Principal Component Analysis:



Independent Component Analysis:



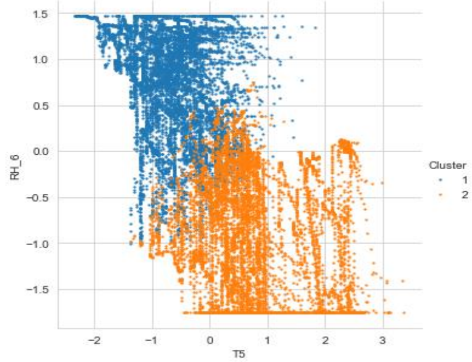
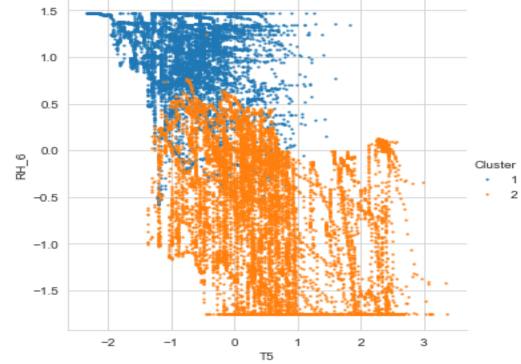
Principal Component Analysis: The graph is **the Explained Variance** across components, we pick 6 Components as it explains **90%** of the data.	Independent Component Analysis To pick the best features for ICA, we use the kurtosis values, and pick the features with highest value which are T2, T4, RH_8

Random Optimization:

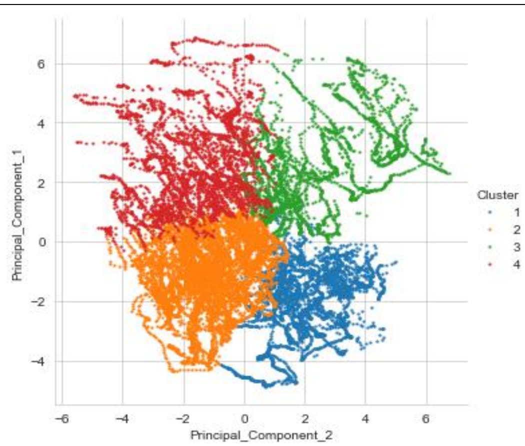
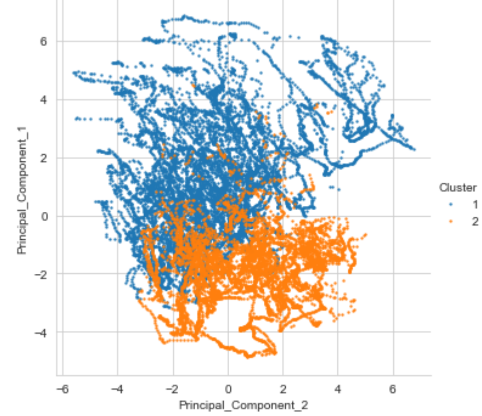
We use Gaussian Random Projection to pick Random Components. We pick 8 components which has previously proven to be the number best suited for understanding the data.

Clustering After Dimension Reduction:

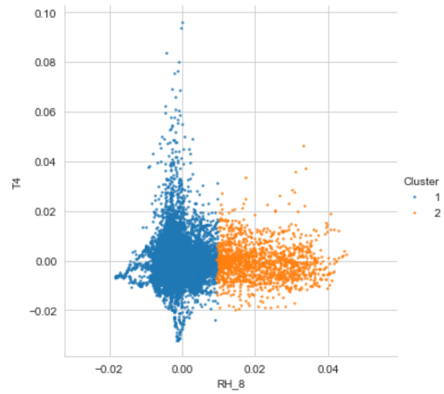
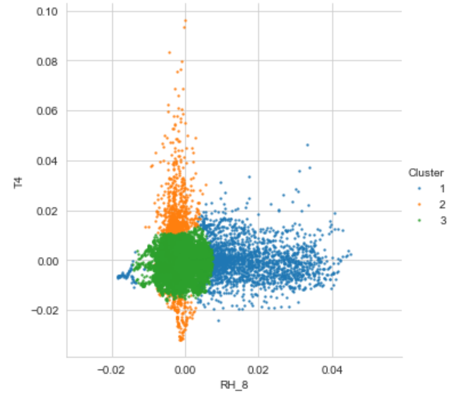
Decision tree:

K Means	Expectation Maximization
	
Explanation: Optimum Clusters using Silhoutte Score is 2 The Class Labels are not aligned with the clusters formed. They did not naturally line up either. This explains that the features could not distingush the dependent variable well enough.	*Explanation:* Optimum Clusters using Silhoutte Score is 2 The clusters are not compact in nature nor did they align themselves in terms of class labels or the clutser itself. The distribution though distinguishable is not clearly seperated.
Difference: K Means on the Feature Selection with descion tree did not change the clustering pattern nor was it effeictive for us to learn anything meanigful from the data.	*Difference:* Expectation maximization on Feature Selection with decision tree was also redundant to classify the observations and into the target classes.

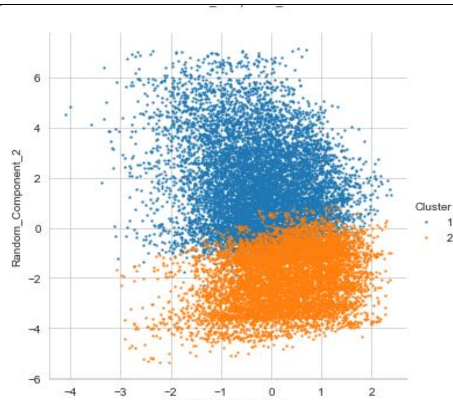
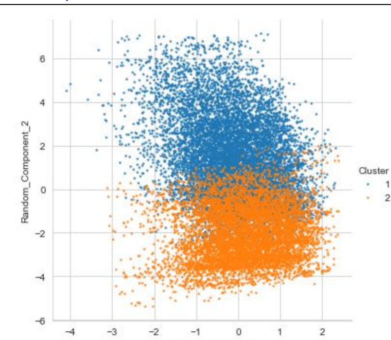
Principal Component Analysis

K Means	Expectation Maximization
	
Explanation: Optimal Clusters using Silhoutte core is 4 The Clustering though not compact does a good job with PC1 and PC2. The compononets help distinguish the Class lables well	Explanation: Optimal Clusters using Silhoutte core is 2 Clusters are not compact in nature nor did they align themselves in terms of class labels or the clutser itself.
Difference: Does a better job when compares to the decision tree but still does not converge to make it distinguishable	Difference: The Clusters to not converge nor do they do a better job than Decision tree to classify the class labels

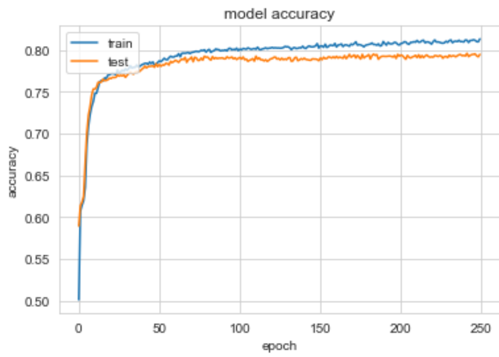
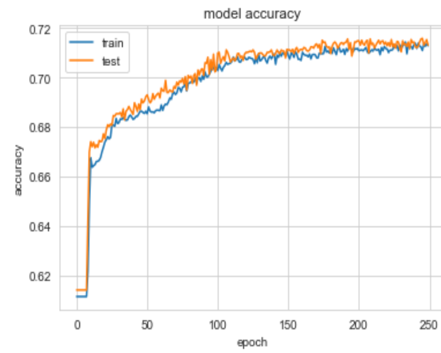
Independent Component Analysis:

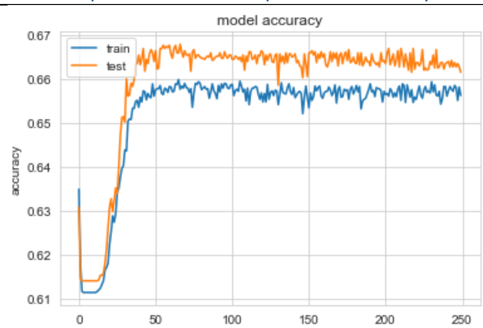
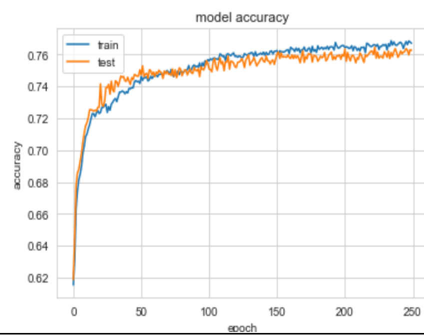
K Means	Expectation Maximization
	
Explanation: Optimal Clusters using Silhoutte core is 2 The Clustering compact, does a good job with being distinguishable . The compononets show a clear clusters.	Explanation: Optimal Clusters using Silhoutte core is 3 Cluster is spread out and is overalapped into each other and is clearly a bad combination to work on.
Difference: ICA does a better job than Decision tree and PCA and the simple K means, ICA does better to understand the data.	Difference: Expectaion maximization does not improve with the original , Decision tree, PCA or ICA.

Random Optimization:

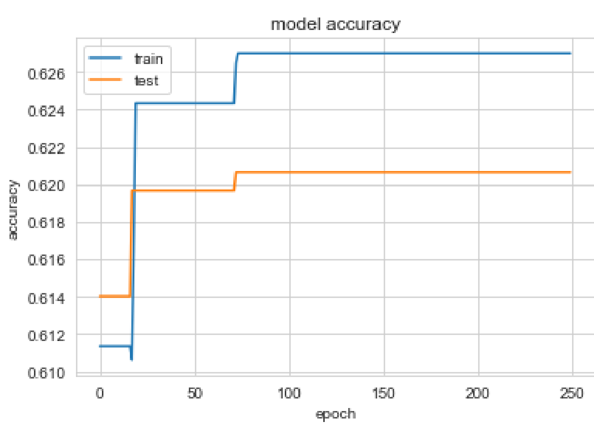
K Means	Expectation Maximization
	
Explanation: Optimal Clusters with Silhouttee Score is 2 Random Optimization does a good job in understanding the target variable. The clusters formed are compact but also have a slight overlap.	Explanation: EM with random optimization gives a better result in understanding the data , but the overlap is still huge for it to be considered a good clustering algorithm for this data.
Difference: Definitely better than K means but is not as good as PCA. Thus can be considered a competent clustering for the random optimization features.	Difference: EM does well when compared with the other Dimension reduction techniques but still isnt operable to consider.

Artificial Neural Networks: It was comparatively Faster than the original ANN ran to optimize the hyper parameters.

Decision Tree	Principal Component Analysis
	
Optimal Neurons and Layers: (9,9,7,9)	Optimal Neurons and layers: (9,9,7,9)
Accuracy: 80.71%	**Accuracy**: 69.87%
Feature Selection works well and does relatively well when compared to the last ANN with 84% Accuracy.	Accuracy dives down with PCA losing information in the dimension reduction.

Independent Component Analysis	Random Optimization
	
Optimal Neurons: (9,9,7,9)	Optimal Neurons: (9,9,7,9)
Accuracy:67%	**Accuracy**:75%
Since the data cannot be separated like signals and the data is dependent on each other the ICA ANN fails to perform well.	Using Random Components worked to find a good accuracy score but was worse than the original ANN.

Artificial Neural Network on Clustered Data




```
Test Measures
        Loss   Accuracy  F1 Score  Precision
0  0.648505    0.62178    0.4808    0.520166
```

- Creating a separate dataset with just the clusters of K means and Expectation Maximization as features.
- These features did not enough information to carry out the classification well
- Thus, we resulted in an Accuracy of 62.17%.
- F-1 Score of 48.08% provides evidence of the target classification worse than a coin flip.
- Thus, running ANN on the just the cluster results is not an encouraging idea for this data.