A Comparative Study of various Machine Learning Algorithms in Appliance Energy Prediction and Mushroom Classification

ABSTRACT:

Machine Learning is undeniably one among the foremost influential and powerful technologies in today's world. it is a tool for turning information into knowledge. It is a way of knowledge analysis that automates analytical model building and is predicated on the thought that systems can learn from data, identify patterns and make decisions with minimal human intervention.

The subsequent report may be a compilation of the varied methodologies and algorithms that can be used for Supervised and Unsupervised Learning. This includes methodologies for prediction, classification and clustering. It also provides implementation details, results and associated discussions and conclusions from this study. The report also provides details of the work done on a number of these algorithms and provides references for an equivalent.

Keywords:

Machine Learning, Supervised, Unsupervised, Prediction, Classification, Regression, Clustering,
Apriori, Linear, Multivariate, Regularized, Logistic, Decision Trees, K-means, Hierarchical,
FP-Growth, Principal Component Analysis

1. INTRODUCTION AND LITERATURE SURVEY

Introduction

Machine learning involves computers discovering how they will perform tasks without being explicitly programmed to try to do so. It supports the thought that systems can learn from data, identify patterns and make decisions with minimal human intervention.

Surely for advanced tasks like face recognition for instance, it is tough to make the needed algorithms, partly as it isn't easy for humans to exactly define how we recognise faces. Abundant face related data exists however. So far, compared to the problem in directly creating the specified algorithms, it is clothed in practice to be easier to help computers to find out themselves the way to recognise faces from available data. The discipline of machine learning develops various approaches for computers to find out to accomplish tasks that no algorithm exists.

Broadly classified, there are three categories of Machine Learning.

1. Supervised Learning: The computer is presented with example inputs and their desired outputs, and therefore the goal is to find out a general rule that maps inputs to outputs. It is utilized in Classification and Regression.
2. Unsupervised Learning: During this sort of learning no labels are given to the training algorithm, leaving it on its own to seek out structure in its input. Unsupervised learning can be a goal in itself (discovering hidden

patterns in data) or a way towards an end (feature learning). Clustering, Associations and Dimensionality Reduction fall in this category.

3. Reinforcement Learning: A computer program interacts with a dynamic environment in which it must perform a particular goal (such as driving a vehicle or playing a game against an opponent). Because it navigates its problem space, the program is provided feedback that is analogous to rewards, which it tries to maximise.

Literature Survey

1. Chunhui Yuan and Haitao Yang in their paper 'Research on K-Value Selection Method of K-Means Clustering Algorithm' have discussed the K-Means Clustering Algorithm and analyzed four K-value selection algorithms, namely Elbow Method, Gap Statistic, Silhouette Coecient, and Canopy. The results showed that the inadequacy within the Elbow Method algorithm which uses SSE as a performance metric, traverses the K value, finds the inflection point, and features a simple complexity is that the inflection point depends on the connection between the K value and therefore the distance value. If the inflection point isn't obvious, the K value can't be determined. The Gap Statistic algorithm compares the arithmetic mean of the averaged reference data set thereupon of the observed data set in order that the fastest k value decreases. However, for several practical large-scale data sets, this method isn't desirable for both time complexity and space complexity. Because the distance matrix must be calculated, the defect with the Silhouette Coecient algorithm is that the computational complexity is O(n^2). The addition of overlapping subsets

within the Canopy algorithm increases its fault tolerance and noise immunity, thus effectively avoiding the issues caused by large computations. Thus, they concluded from their study that for the clustering of small data sets, the four methods mentioned within the paper can meet the requirements and that, for large and complex data sets, the cover algorithm is the most suitable option.

2. Himani Sharma and Sunil Kumar in their work 'A Survey on Decision Tree Algorithms of Classification in Data Mining' published within the International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 studied the different Decision Tree Algorithms like C4.5, ID3 and CART and discussed the varied applications of decision trees is different areas of knowledge mining. The paper highlighted the differences within the three algorithms and discussed the benefits and drawbacks of using one over the over. For instance , while CART works on continuous and nominal attributes data, C4.5 works on continuous and categorical data and ID3 works on categorical data. Also, the speed of ID3 was found to be low, which of CART was average while C4.5 also performed better than ID3. While ID3 uses information entropy and knowledge gain, C4.5 uses split info and gain ratio and CART also uses Gini index. Thus, each has its own pros and cons.

3. Jiao Yabing in his work 'Research of an Improved Apriori Algorithm in data processing Association Rules' published in International Journal of Computer and Communication Engineering, Vol. 2, No. 1, January 2013 discussed the issues faced by the Apriori algorithm when encountered with

dense data due to which the performance of the algorithm declines dramatically. He proposed an improved algorithm of association rules, the classical Apriori algorithm.The optimized algorithm counts the frequent item set Lk which goes to attach . Consistent with the result it deletes item sets with the amount but k-1 in Lk-1 to decrease the amount of the connecting item set and take away some element that does not satisfy the conditions. This may decrease the likelihood of combination, and decline the amount of candidate itemsets in Ck. For giant databases, this algorithm can save time and increase the efficiency of knowledge mining. This is what the Apriori algorithm does not possess. Although this process can decrease the amount of candidate itemsets in Ck and reduce time cost of knowledge mining, the worth of pruning frequent item sets could cost a particular time. For dense databases (such as, telecom, population census, etc.), as large amounts of long forms occur, the efficiency of this algorithm is above Apriori.

4. Sidharth Prasad Mishra, Uttam Sarkar, Subhash Taraphder, Sanjay Datta, Devi Prasanna Swain1, Reshma Saikhom, Sasmita Panda and Menalsh Laishram in their work ‘Multivariate Statistical Data Analysis- Principal Component Analysis (PCA)’ published within the International Journal of Livestock Research eISSN : 2277-1964 NAAS Score -5.36 discusses the oldest and best known technique of multivariate data analysis, Principal Component Analysis. The paper discusses the goals of PCA and discusses the specified statistics. It further discusses the methodology for implementing PCA.

2. PROBLEM FORMULATION

The objective of this report is to review the varied methodologies, techniques and Machine Learning algorithms that are useful in analysis of knowledge .

The problems considered for the study are:

1. Prediction of energy consumed by appliances in a low energy house using Linear Regression, Multivariate Regression and Regularized Regression.
2. Classification of mushrooms into poisonous and edible using Logistic Regression
3. Classification of mushrooms into poisonous and edible using Decision Tree Algorithm
4. Clustering using K-Means Algorithm
5. Classification of mushrooms using Hierarchical Agglomerative Clustering
6. Association analysis of groceries using Apriori Algorithm
7. Association analysis of groceries using Frequent Pattern Growth Algorithm
8. Principal Component Analysis for dimensionality reduction of mushroom dataset

3. THE METHODOLOGY

3.1 Linear Regression , Multivariate Regression and Regularized Regression

Linear Regression

Linear Regression may be defined as a linear approach used for modeling the connection between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).
The case of 1 explanatory variable is named simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

- Start with a training set with x1,x2, x3 .. and y
- Start with parameters c0,c1,c3 with random values
- Start with a learning rate alpha
- Then repeat the following update

  c0 = c0 - alpha * h(x) - y

  c1 = c1 - alpha * (h(x) - y) * x
- Repeat this process till it converges

Multivariate Regression

The general linear model or Multivariate linear regression on the opposite hand, may be understood as a generalization of multiple linear regression to the case of more than one dependent variable .

Regularized Regression

If the hypothesis has high variance, while it fits the training data with good accuracy, it fails to predict the values for unseen cases with same accuracy due to the high variance within the prediction curve. This is called overfitting.

It is mathematically seen that the high variance of an overfit hypothesis is attributed to higher value of parameters like higher order features i.e. more the dependency is biased on a single feature, greater are the chances of a hypothesis overfitting. This effect can be counteracted by ensuring that the values of parameters are small. This is done by penalizing the algorithm proportional to value of θj which can ensure small values of those parameters and hence would prevent overfitting by attributing small contributions from each feature and hence removing high bias or high variance.
Thus, in regularized regression an additional regularization parameter is used for penalizing the algorithm to stop overfitting.

3.2 Logistic Regression

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which may then be mapped to 2 or more discrete classes.

3.3 Decision Tree

A decision tree is a flowchart -like structure in which each internal node represents a 'test' on an

attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the result of the test, and every leaf node represents a category label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

```
INPUT: S, where S = set of classified instances
OUTPUT: Decision Tree
Require: S ≠ ∅, num_attributes > 0
 1: procedure BUILDTREE
 2:     repeat
 3:         maxGain ← 0
 4:         splitA ← null
 5:         e ← Entropy(Attributes)
 6:         for all Attributes a in S do
 7:             gain ← InformationGain(a, e)
 8:             if gain > maxGain then
 9:                 maxGain ← gain
10:                 splitA ← a
11:             end if
12:         end for
13:         Partition(S, splitA)
14:     until all partitions processed
15: end procedure
```

3.4 K-means Clustering

K-means algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each item belongs to just one group. It tries to form the inter-cluster data points as similar as

possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the information points and therefore the cluster's centroid (arithmetic mean of all the information points that belong thereto cluster) is at the minimum. The less variation we've within clusters, the more homogeneous (similar) the information points are within an equivalent cluster.

K-MEANS($\{\vec{x}_1, \ldots, \vec{x}_N\}, K$)
1 $(\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_K) \leftarrow$ SELECTRANDOMSEEDS($\{\vec{x}_1, \ldots, \vec{x}_N\}, K$)
2 **for** $k \leftarrow 1$ **to** K
3 **do** $\vec{\mu}_k \leftarrow \vec{s}_k$
4 **while** stopping criterion has not been met
5 **do for** $k \leftarrow 1$ **to** K
6 **do** $\omega_k \leftarrow \{\}$
7 **for** $n \leftarrow 1$ **to** N
8 **do** $j \leftarrow \arg\min_{j'} |\vec{\mu}_{j'} - \vec{x}_n|$
9 $\omega_j \leftarrow \omega_j \cup \{\vec{x}_n\}$ *(reassignment of vectors)*
10 **for** $k \leftarrow 1$ **to** K
11 **do** $\vec{\mu}_k \leftarrow \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} \vec{x}$ *(recomputation of centroids)*
12 **return** $\{\vec{\mu}_1, \ldots, \vec{\mu}_K\}$

► **Figure 16.2** The *K*-means algorithm. For most IR applications, the vectors $\vec{x}_n \in \mathbb{R}^M$ should be length-normalized. Alternative methods of seed selection and initialization are discussed on page 16.4 .

3.5 Hierarchical Agglomerative Clustering

Hierarchical Clustering is a method of study which seeks to create a hierarchy of clusters i-e; a tree type structure supported hierarchy. Agglomerative clustering follows a bottom-up approach. It doesn't require us to pre-specify the amount of clusters. Each item is treated as one cluster then it successively agglomerates pairs of clusters until there is one cluster containing all data points. The algorithm is shown as follows:

```
given a dataset (d1, d2, d3, ....dN) of size N
# compute the distance matrix
for i=1 to N:
  # as the distance matrix is symmetric about
  # the primary diagonal so we compute only lower
  # part of the primary diagonal
  for j=1 to i:
    dis_mat[i][j] = distance[di, dj]
each data point is a singleton cluster
repeat
  merge the two cluster having minimum distance
  update the distance matrix
until only a single cluster remains
```

3.6 Apriori Algorithm

The Apriori algorithm is employed for mining frequent itemsets and devising association rules from a transactional database. The parameters "support" and "confidence" are used. Support refers to items' frequency of occurrence; confidence may be a contingent probability. Items during a transaction form an item set. The algorithm begins by identifying frequent, individual items (items with a frequency greater than or adequate to the given support) within the database and continues to increase them to larger, frequent itemsets .

The following are the most steps of the algorithm:

1. Calculate the support of item sets (of size k = 1) within the transactional database (note that support is the frequency of occurrence of an itemset). This is called generating the candidate set.
2. Prune the candidate set by eliminating items with a support but the given threshold.
3. Join the frequent itemsets to make sets of size k + 1, and repeat the above sets until no more itemsets are often formed. This may happen when the set(s) formed have support but the given support.

3.7 Frequent Pattern Growth Algorithm

This algorithm is an improvement to the Apriori method. A frequent pattern is generated without the necessity for candidate generation. FP growth algorithm represents the database within the sort of a tree called a frequent pattern tree or FP tree. This tree structure will maintain the association between the itemsets.

The algorithm is as follows:

1) The primary step is to scan the database to seek out the occurrences of the itemsets within the database.

2) The second step is to construct the FP tree. For this, create the basis of the tree. The basis is represented by null.

3) The subsequent step is to scan the database again and examine the transactions. Examine the primary transaction and determine the itemset in it. The itemset with the max count is taken at the highest, subsequent itemset with lower count then on.

4) Subsequent transactions within the database are examined. The itemsets are ordered in descending order of count. If any itemset of this transaction is already

present in another branch then this transaction branch would share a standard prefix to the basis .

5) Also, the count of the itemset is incremented because it occurs within the transactions. Both the common node and new node count is increased by 1 as they're created and linked consistent with transactions.

6) The subsequent step is to mine the created FP Tree. For this, the rock bottom node is examined first alongside the links of rock bottom nodes. a rock bottom node represents the frequency pattern length 1. From this, traverse the trail within the FP Tree. This path or paths are called a conditional pattern base. Conditional pattern base may be a sub-database consisting of prefix paths within the FP tree occurring with rock bottom node (suffix).

7) Construct a Conditional FP Tree, which is made by a count of itemsets within the path. The itemsets meeting the edge support are considered within the Conditional FP Tree.

8) Frequent Patterns are generated from the Conditional FP Tree.

3.8 Principal Component Analysis for dimensionality reduction

When there are tons of variables aka features n (>10), then we are advised to try to do PCA. PCA may be a statistical technique which reduces the size of the information and helps us understand, plot the information with lesser dimension compared to original data. As the name says, PCA helps us compute the Principal components in data. Principal components are basically vectors that are linearly uncorrelated and have a variance within data. From the principal components top p is picked which have the foremost variance. This is how PCA works -

1. Calculate the covariance matrix X of knowledge points.

2. Calculate eigenvectors and corresponding eigenvalues.
3. Sort the eigenvectors consistent with their eigenvalues in decreasing order.
4. Choose first k eigenvectors which are going to be the new k dimensions.
5. Transform the first n dimensional data points into k dimensions.

4. RESULTS AND DISCUSSIONS

Appliances Energy Prediction

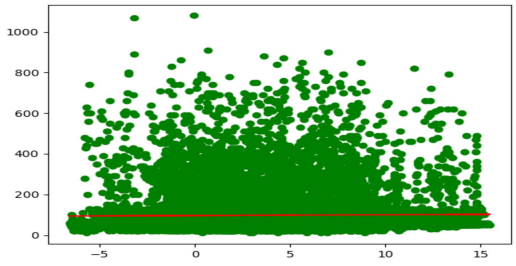
The dataset used to create regression models is that of appliances' energy use during a low energy building. The information was recorded for a period of 4.5 months every 10 minutes employing a Zigbee.It consists of 19735 instances with 29 attributes including temperature and humidity conditions. We perform feature engineering and selection on the information , then train on the specified features. We perform univariate, multivariate and regularized (ridge) regression. The results are shown below.

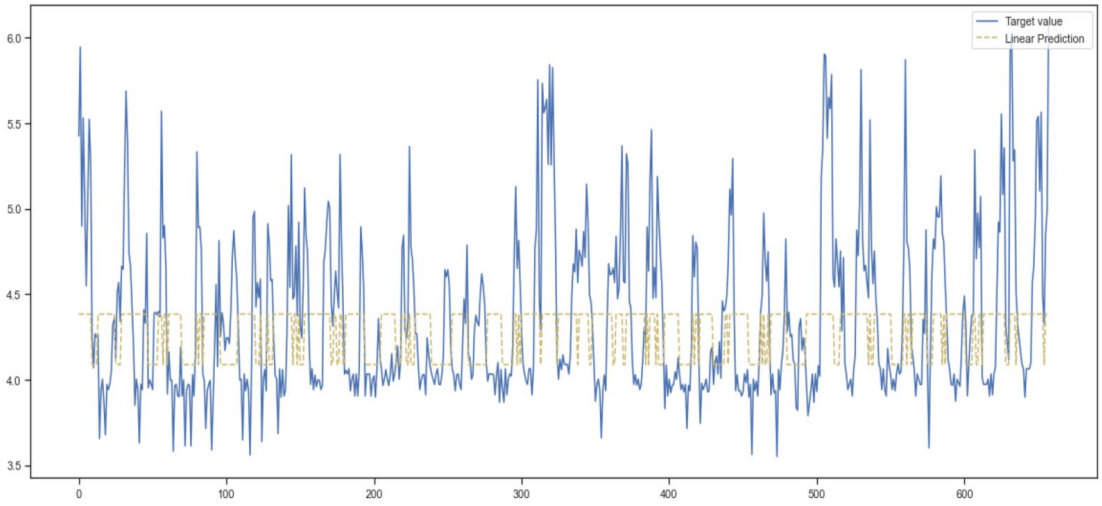
4.1 Univariate, Multivariate, Regularized Linear Regression

We evaluate the performance results calculating the negative mean absolute error, R^2 (coefficient of determination) regression score function. The results and therefore the visualization of an equivalent are displayed below.

Univariate: (Feature used : T6)

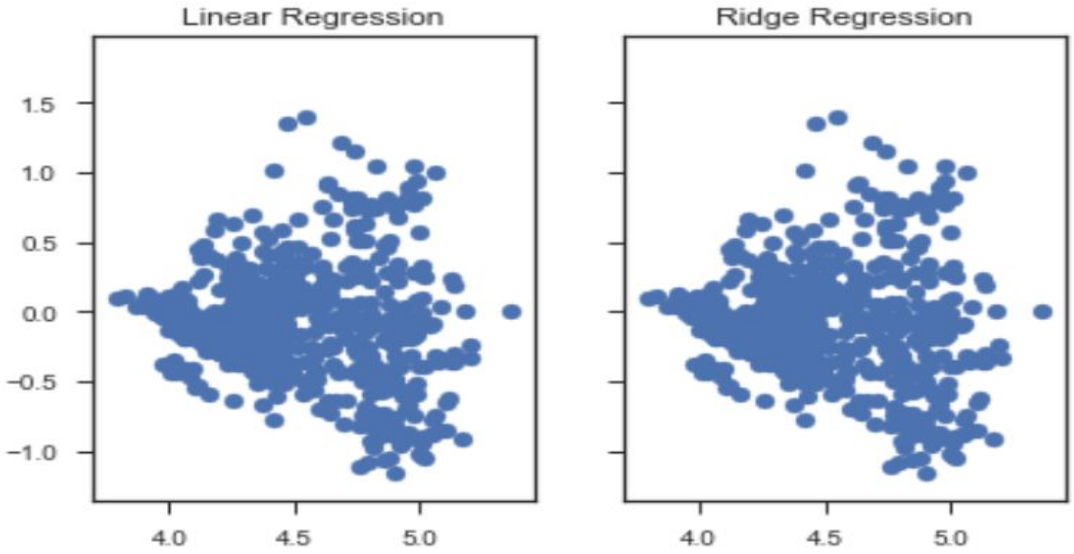
```
Variance score R^2  : -1.48%
```

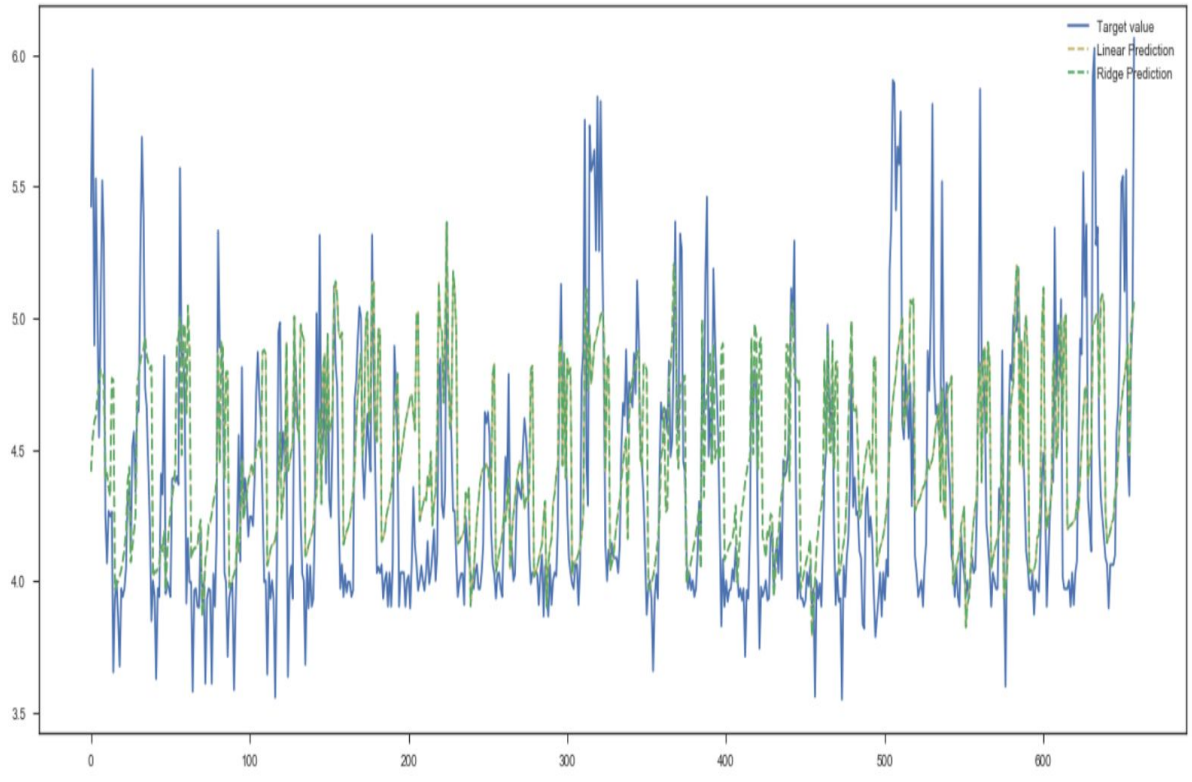




Multivariate and Regularized:

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)

Average Error       : 0.3093 degrees
Variance score R^2  : 28.15%
Accuracy            : 92.92%

Ridge(alpha=1.0, copy_X=True, fit_intercept=True, max_iter=None,
      normalize=False, random_state=None, solver='auto', tol=0.001)

Average Error       : 0.3094 degrees
Variance score R^2  : 28.16%
Accuracy            : 92.91%
```


Linear Regression
Ridge Regression
1.5
1.0
0.5
0.0
−0.5
−1.0
4.0
4.5
5.0
4.0
4.5
5.0


Target value
Linear Prediction
Ridge Prediction
6.0
5.5
5.0
4.5
4.0
3.5
0
100
200
300
400
500
600

Mushroom Classification

The mushroom dataset includes descriptions of hypothetical samples like 23 species of gilled mushrooms within the Agaricus and Lepiota Family. Each species is identified as definitely edible or definitely poisonous. We perform the specified feature engineering on the information and perform classification using the subsequent algorithms:

- ❑ Logistic Regression
- ❑ Decision Tree
- ❑ KMeans Clustering
- ❑ Hierarchical Agglomerative Clustering

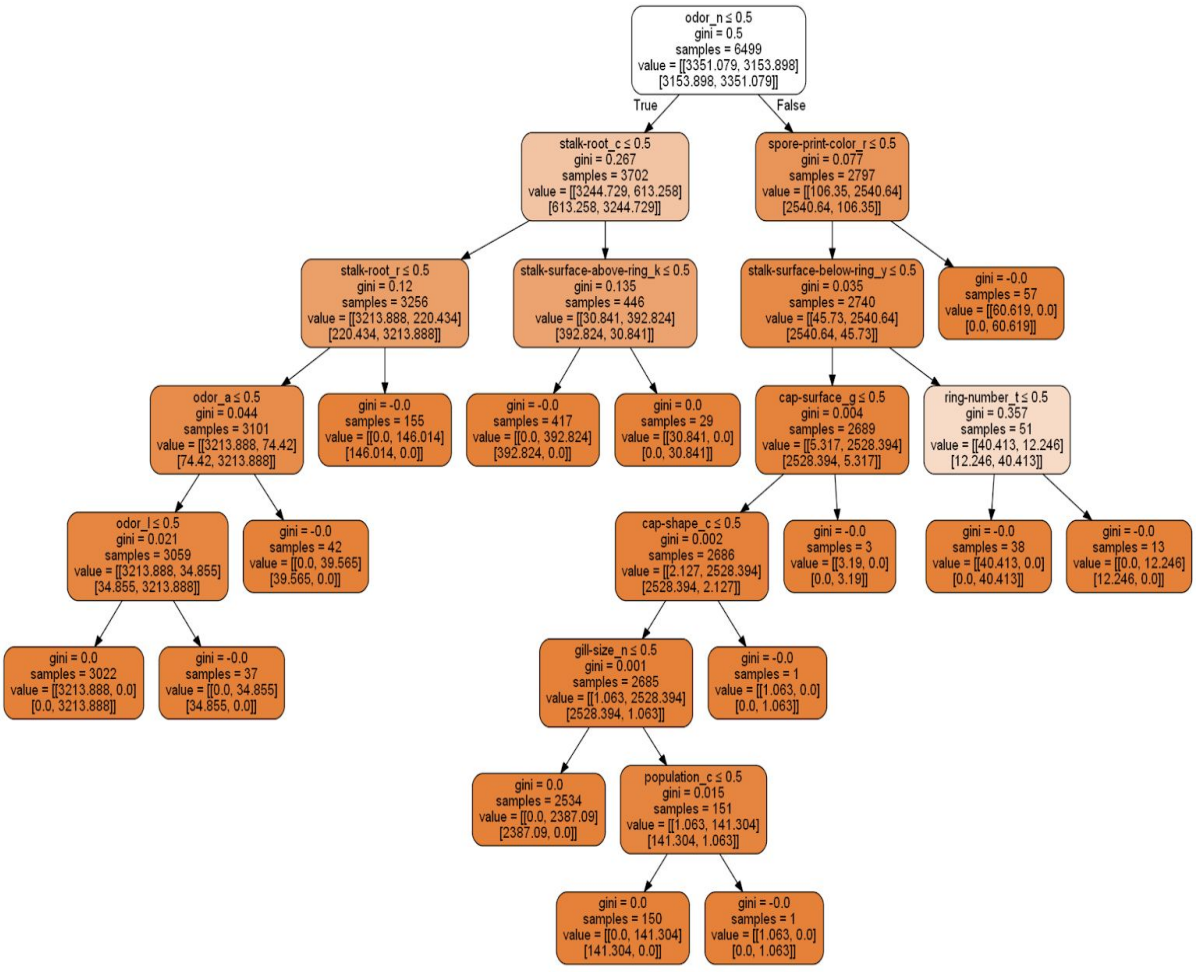
4.2 Logistic Regression

We perform feature engineering and evaluate the result. We calculate the AUC score and also plot the confusion matrix, which indicates the right and incorrect classifications. The results calculated are displayed below.

```
Cross-validation score of feature engineered variables [Logistic Regression]:0.99
Accuracy score of feature engineered variables [Logistic Regression] : 98.77 %
Feature engineered variables confusion matrix [Logistic Regression]:
 [[713  20]
 [  0 892]]
AUC score for feature engineered variables [Logistic Regression]: 0.99
[Logistic Regression] classification report:
              precision    recall  f1-score   support

           0       1.00      0.97      0.99       733
           1       0.98      1.00      0.99       892

    accuracy                           0.99      1625
   macro avg       0.99      0.99      0.99      1625
weighted avg       0.99      0.99      0.99      1625
```

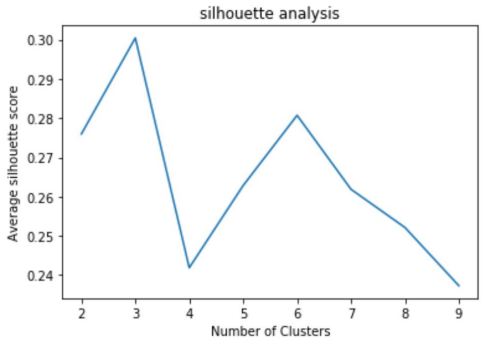
4.3 Decision Tree

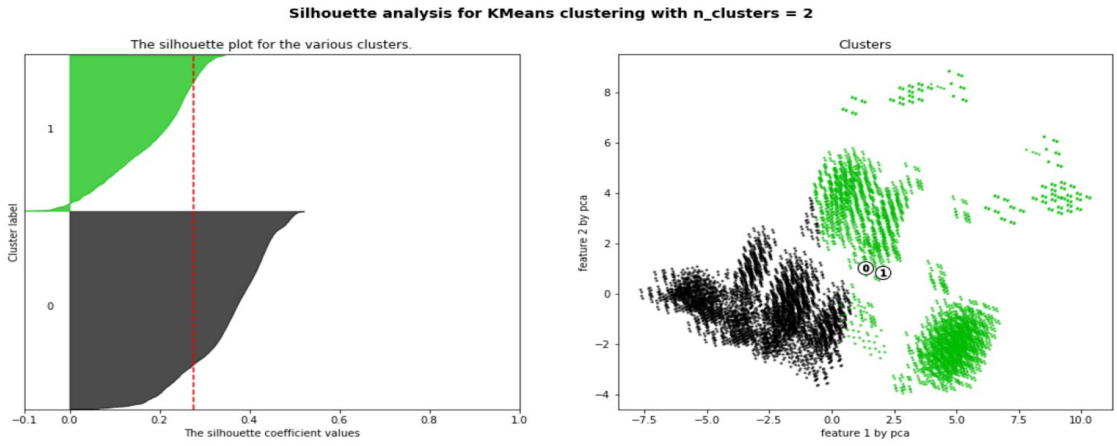
We use a score for evaluation which returns the mean accuracy on the given test data and labels. The visualization of the tree is depicted below. The standards used here is that the Gini index which measures the degree or probability of a specific variable being wrongly classified when it is randomly chosen.



4.4 KMeans

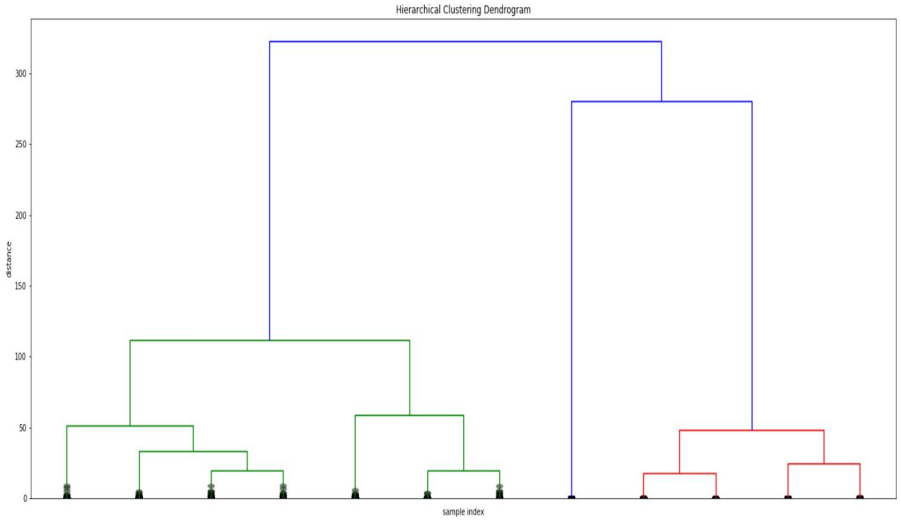
For selection of the optimal number of clusters, we perform silhouette analysis.The optimal number of clusters chosen is 2 despite the score being the very best for 3 because the samples are more uniform for k=2. Thus we use KMeans to cluster the information accordingly. The results and visualization are depicted below.

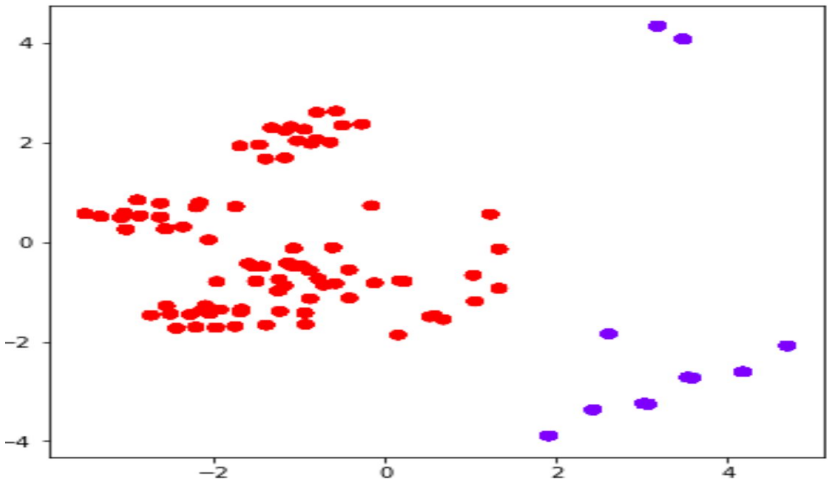




4.5 Hierarchical Agglomerative clustering

For selection of the optimal clusters, we use a dendrogram to work out the utmost distance at which we split then we elect the specified clusters. The specified results and visualization is shown below.





Grocery Transaction Analysis

The groceries dataset contains 9835 transactions by customers buying groceries. The information contains 169 unique items. The csv file was read transaction by transaction and every transaction was saved as an inventory . A mapping was created from the unique items within the dataset to integers in order that each item corresponded to a singular integer. the whole data was mapped to integers to scale back the storage and computational requirement. A reverse mapping was created from the integers to the item, in order that the item names might be written within the final computer file. We obtain the simplest association rules which are displayed below by using the following:

- ❏ Apriori
- ❏ FP Growth

4.6 Apriori Algorithm

The apriori algorithm is employed to get the specified association rules and thus we all know the related items . Association rules with the very best confidence are shown below.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
10	(whipped/sour cream)	(whole milk)	0.071683	0.255516	0.032232	0.449645	1.759754	0.013916	1.352735
0	(root vegetables)	(whole milk)	0.108998	0.255516	0.048907	0.448694	1.756031	0.021056	1.350401
15	(root vegetables)	(other vegetables)	0.108998	0.193493	0.047382	0.434701	2.246605	0.026291	1.426693
1	(tropical fruit)	(whole milk)	0.104931	0.255516	0.042298	0.403101	1.577595	0.015486	1.247252
11	(yogurt)	(whole milk)	0.139502	0.255516	0.056024	0.401603	1.571735	0.020379	1.244132

4.7 FP Growth

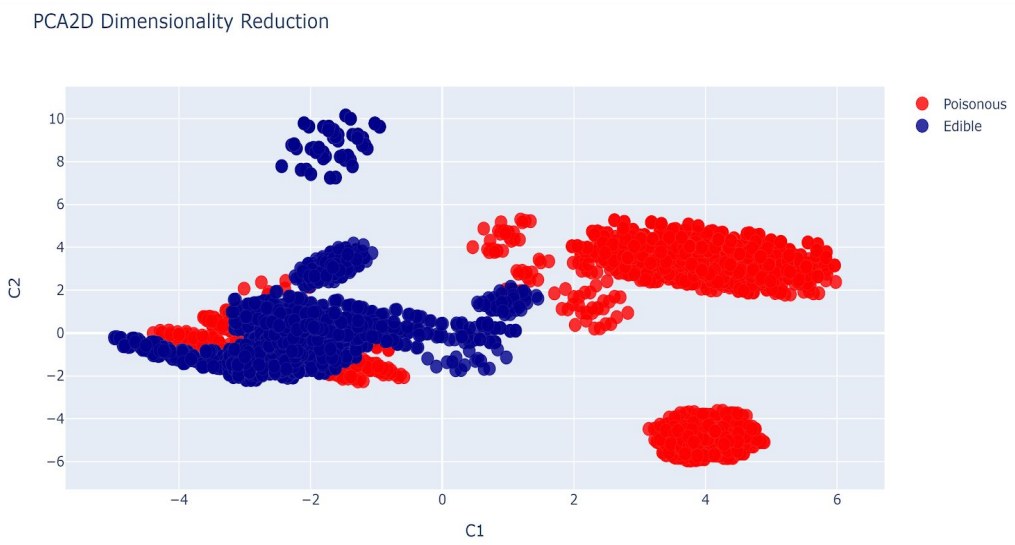
The FP Growth is employed to get the specified association rules. It uses an FP-Tree to get the foremost associated items. Association rules with the very best confidence are shown below.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
15	(other vegetables, yogurt)	(whole milk)	0.043416	0.255516	0.022267	0.512881	2.007235	0.011174	1.528340
91	(other vegetables, root vegetables)	(whole milk)	0.047382	0.255516	0.023183	0.489270	1.914833	0.011076	1.457687
93	(root vegetables, whole milk)	(other vegetables)	0.048907	0.193493	0.023183	0.474012	2.449770	0.013719	1.533320
17	(yogurt, whole milk)	(other vegetables)	0.056024	0.193493	0.022267	0.397459	2.054131	0.011427	1.338511
92	(other vegetables, whole milk)	(root vegetables)	0.074835	0.108998	0.023183	0.309783	2.842082	0.015026	1.290900
16	(other vegetables, whole milk)	(yogurt)	0.074835	0.139502	0.022267	0.297554	2.132979	0.011828	1.225003

4.8 Principal Component Analysis for Dimensionality Reduction

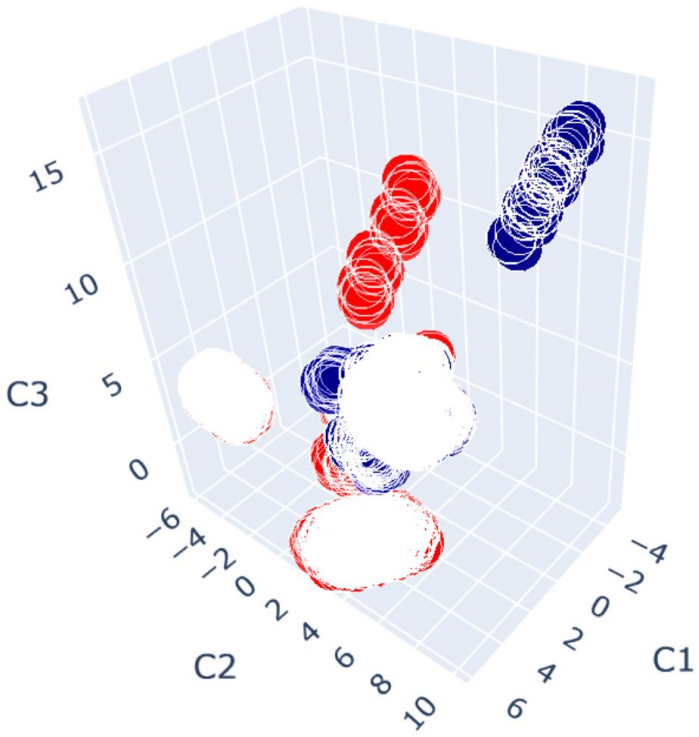
We also perform Dimensionality reduction on the mushroom dataset using Principal Component Analysis. This is then used as input to a Random Forest Classifier. The results and the visualization of the extracted features is depicted below.

```
[[1242   32]
 [  93 1071]]
              precision    recall  f1-score   support

           0       0.93      0.97      0.95      1274
           1       0.97      0.92      0.94      1164

    accuracy                           0.95      2438
   macro avg       0.95      0.95      0.95      2438
weighted avg       0.95      0.95      0.95      2438
```




```
[[1260   14]
 [  41 1123]]
```

	precision	recall	f1-score	support
0	0.97	0.99	0.98	1274
1	0.99	0.96	0.98	1164
accuracy			0.98	2438
macro avg	0.98	0.98	0.98	2438
weighted avg	0.98	0.98	0.98	2438


15
10
C3
5
0
−6
−4
−2
0
2
4
6
8
10
C2
6
4
2
0
−2
−4
C1

5. CONCLUSION

From studying the research work of others on various Machine Learning algorithms and implementing a few ourselves we have come to a conclusion that there can be various algorithms that can be applied for a specific type of analysis. The feature selection method, the parameters used, and the ways to select these parameters can differ in each implementation of any algorithm. And the selection of a different technique can give a better or worse result. While an algorithm may perform better in certain cases, it might perform poorly on others. Thus, selection of methodology needs to be done carefully. Also, pre-processing data can result in better scores on various evaluation metrics.

REFERENCES (Journals, Books, Book Chapter or Conference can be referred)

1. Luis M. Candanedo, Veronique Feldheim, Dominique Deramaix, Data driven prediction models of energy use of appliances in a low-energy house, Energy and Buildings, Volume 140, 1 April 2017, Pages 81-97, ISSN 0378-7788

2. Chunhui Yuan and Haitao Yang, Research on K-Value Selection Method of K-Means Clustering Algorithm, Multidisciplinary Scientific Journal, Published: 18 June 2019

3. A Survey on Decision Tree Algorithms of Classification in Data Mining, International Journal of Science and Research (IJSR) 5(4) · April 2016

4. Himani Sharma and Sunil Kumar, A Survey on Decision Tree Algorithms of Classification in Data Mining, International Journal of Science and Research (IJSR)
   ISSN (Online): 2319-7064

SOME USEFUL WEB RESOURCES

1. Appliance Energy Dataset: http://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction
2. Mushroom Dataset: http://archive.ics.uci.edu/ml/datasets/Mushroom