Statistical Learning
Report
On
Home Assignment 2

HÖGSKOLAN
DALARNA

By
Manthri Bala Kiran
H19baman@du.se

Introduction:

We have been given the Nuclear data set which is collected from a nuclear fraction of cortex on Mice.

Problem Description:

1. We have been given different classification task to perform at the first step. They are:
   a. Using all the 77 proteins in the dataset as predictors for Decision Trees and Support Vector Machine to make Binary Classification and Multiple Class Classification.
   b. Perform Principal Component Analysis on the 77 Numerical Features. Using Appropriate number of Principal Components as predictors and performing the same classification task.
   c. Using Bagging, Random Forest, and Boosting performing the same classification task. Comparing the results of the three methods.
2. In the second task we have been performing task on clustering. We tried using K-means Clustering and Hierarchical Clustering. In every case, we need to find the number of clusters that make sense and try to explain what each cluster describes.

Dataset:

The Data Cortex Nuclear data set concerns the expression levels of 77 proteins, detectable in the nuclear fraction of cortex of mice. There are 38 control mice and 34 mice with Down syndrome in the dataset, which are further divided into 4 categories each (8 categories total, 4 for the control mice and 4 for the Down Syndrome mice). Therefore, for the control mice, we have 38*15 which equals 570 measurements, and for trisomic mice, we have 34*15 which equals 510 measurements. In total we have 1080 measurements per protein and 77 different proteins, Mouse ID, Genotype of Mice, Treatment type, Behaviour and Class. There are numerical and categorical variables on the dataset.

Initial Steps:

Firstly, I have loaded the Excel file “Data_Cortex_Nuclear.xls” to the R Studio as a dataframe. In the dataset we have 1080 entries as the rows distributes between 82 columns. In total we have 1396 of null values in the data set we have replaced the null values with the minimum value of the column. I have deleted the column mouse id from our dataset. We have created two datasets for Binary class and Multiple class. For Binary class, I have considered Genotype feature and Class feature for the Multiple class. We have created train and test data set. For training we have used 750 observations which make nearly 70% of the data and rest 330 observations of the data for testing.

Methods Task 1:

For the Classification task, we need to perform Binary and Multiple class classification. For the Binary class I have taken Genotype as the response variable and for the Multiple class I have taken class feature as the response variable. For both binary and multiple class we have performed Decision tree and SVM classification. At the second part I have performed PCA and using the PCA components I have reperformed Decision Tree and SVM classification. At the third part I have performed Random Forest, Bagging and Boosting. For the Clustering task, I have performed K-means and Hierarchical clustering.

Decision Tree:

A Decision Tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. A decision tree is a flowchart-like structure in which each internal node represents as “test” on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label [1].

a. Firstly, I have applied decision tree for the Multiclass classifier while considering Class as the response variable for the multiclass training dataset and we have converted the Class variable as categorial variable.
b. Now, we tried to predict() using the fitted tree and the test data of the multiclass. After drawing the confusion matrix, we can see that we have obtained the accuracy of 74%.
c. Next, I tried to apply the decision tree for Binary Classifier while using Genotype as my response variable for the binary class training and testing dataset. And I converted the Genotype variable as categorical variable.
d. We tried to predict() using the fitted tree and the test data of the binary class. Using the confusion matrix, we can say that I had an accuracy of 85%.
e. Now, I tried to fit a cross validation to the multiclass decision tree and the binary class decision tree and tried to prune the tree which helps to improve the accuracy of the predication.
f. I have predicated, using the pruned tree and test data for the multiclass and converted the Class variable from the test data to categorical and drawn the confusion matrix. I had the accuracy of 74%.
g. Now, I tried predicting the binary class using the pruned tree and the test data for the binary tree and converted the Genotype variable from the test data to categorical and drawn the confusion matrix. I had the accuracy of 85%.

Support Vector Machines:

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N – the number of features) that distinctly classifies the data points [2].

1. Linear Kernel is used when the data is Linearly separable, that is it can be separated using a single line.
2. The polynomial kernel is a kernel function commonly used with support vector machines(SVM), that represents the vector machines, that represents the vector in a feature space over polynomials of the original variable, allowing learning of non-linear models.
3. Radial kernel is another popular kernel method used n SVM models. Radial Kernel is a function whose values depends on the distance from the origin or from some point.

a. For SVM, I have applied multiple predictions on Multiclass and Binary class classifier converting Class and Genotype Variable as the categorical variable and opted the accuracy.
b. For the better accuracy, we have tuned our model for maximizing the model performance without overfitting the model. And from the tuning, I have applied multiple Multiclass and Binary class classifier converting Class and Genotype Variable as categorical variable and opted the accuracy.

PCA:

Principal Component Analysis (PCA) is an unsupervised, non-parametric statistical technique primarily used for dimensionality reduction in machine learning. High dimensionality means that the dataset has many features. The primary problem associated with high-dimensionality in the machine learning field is model overfitting, which reduces the ability to generalized beyond the examples of the training set [3].

a. Using the PCA components, I have created the dataframe and divided the data frame into train and test data.
b. Now performed Decision Tree and Support Vector Machine for both Binary and Multiclass classifiers.

Random Forest:

The Random Forest is a model made up of many decision trees. Rather than just simply averaging the prediction of trees(which would be calling as 'forest') [4].

a. I have applied the Random Forest for the multiclass train data and predicted the data using the test data for the multiclass while converting the Class variable as categorical variable. I have obtained 98% of accuracy for the same.
b. For the binary classifier, I have predicated using Genotype as my response variable and obtained the accuracy of 98% while converting Genotype as categorical variable.

Bagging:

Bootstrap aggregating, is also called Bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression [5].

a. Applied bagging for the multiclass train data and predicated the data using the test data for the multiclass using categorical class feature. And, I have opted an accuracy of 98%.
b. For the binary classifier, using Genotype as my response variable and converting it to categorical variable, I have opted an accuracy of 0.95.

Boosting:

Boosting is one of the ensemble learning techniques in machine learning and it is widely used in regression and classification problems. The main concept of this method is to improve the week learners sequentially and increase the model accuracy with a combined model [6].

1. After applying boosting for the binary class, I have opted an accuracy of 99%.
2. After applying boosting for the multiclass, I have opted an accuracy of 99%.

Results Task 1- A,B:

	Binary Class		Multiclass	
	Original	**PCA**	**Original**	**PCA**
Decision Tree	**85%**	**78%**	**74%**	**53%**
Decision Tree.cv	**85%**	**80%**	**74%**	**80%**
SVM-Linear	**61%(cost=0.01)**	**73% (cost = 0.01)**	**76%(cost=0.1)**	**68%(cost=0.01)**
SVM-Linear – post tuning	**79%(cost=0.1)**	**79% (cost= 5)**	**91% (cost =1)**	**73% (cost = 5)**
SVM-Radial	**54% (gamma = 1, cost = 1)**	**98% (gamma= 1, cost =1)**	**48% (gamma = 0.1, cost =0.1)**	**66% (gamma = 0.1, cost = 0.1)**
SVM-Radial-post tuning	**100% (gamma = 0.01, cost = 5)**	**98% (gamma = 1, cost= 5)**	**99% (gamma = 0.001, cost = 100)**	**95% (gamma = 0.1, cost =10)**
SVM-Polynomial	**98% (degree= 2, gamma = 1, cost = 0.01)**	**81% (degree = 2, gamma =1, cost = 0.01)**	**94% (degree = 2, gamma = 0.1, cost = 0.01)**	**30% (degree = 2, gamma = 0.1, cost = 0.01)**
SVM-Polynomial-post tuning	**99% (degree = 3, gamma = 1, cost = 5)**	**93% (degree = 3, gamma = 1, cost = 100)**	**100% (degree = 3, cost = 5)**	**94% (degree = 3, cost = 10)**

For the original data, we can see that SVM performs better than the decision tree. For the best SVM accuracy values I have performed tuning for the best fit of the model for choosing best cost and gamma values. Based on all SVM fits we can see that Radial and Polynomial gave the best results and at one point they have an 100% accuracy.

I have performed the SVM and Decision Tree for the PCA transformed data. From the fits we can see at some points the original fits have the best accuracy than the PCA fits and SVM performed better than the decision tree. I have performed tuning for choosing the best cost and gamma values. SVM Radial and Polynomial gave the best accuracy for the models.

Task 1- C:

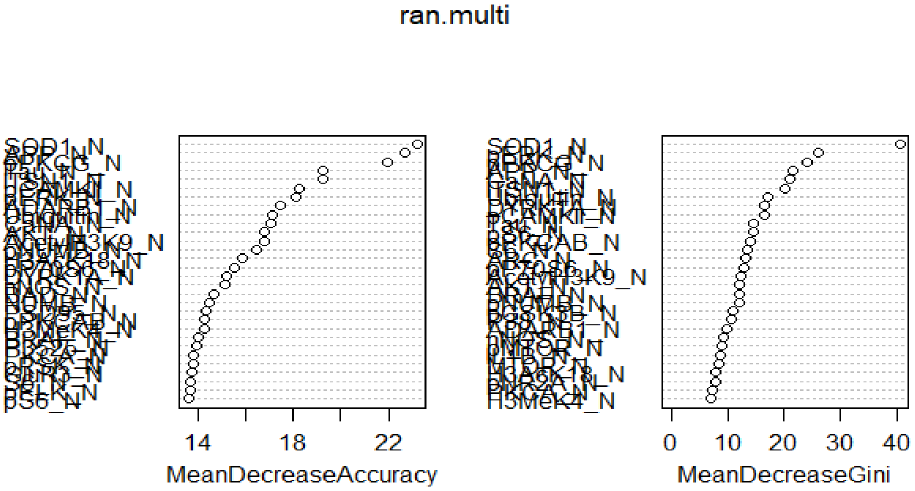


Fig: Random Forest-Multi

From the above figure we can say that SOD1_N is considered as a most important variable considering MeanDecreaseAccuracy and MeanDecreaseGini.

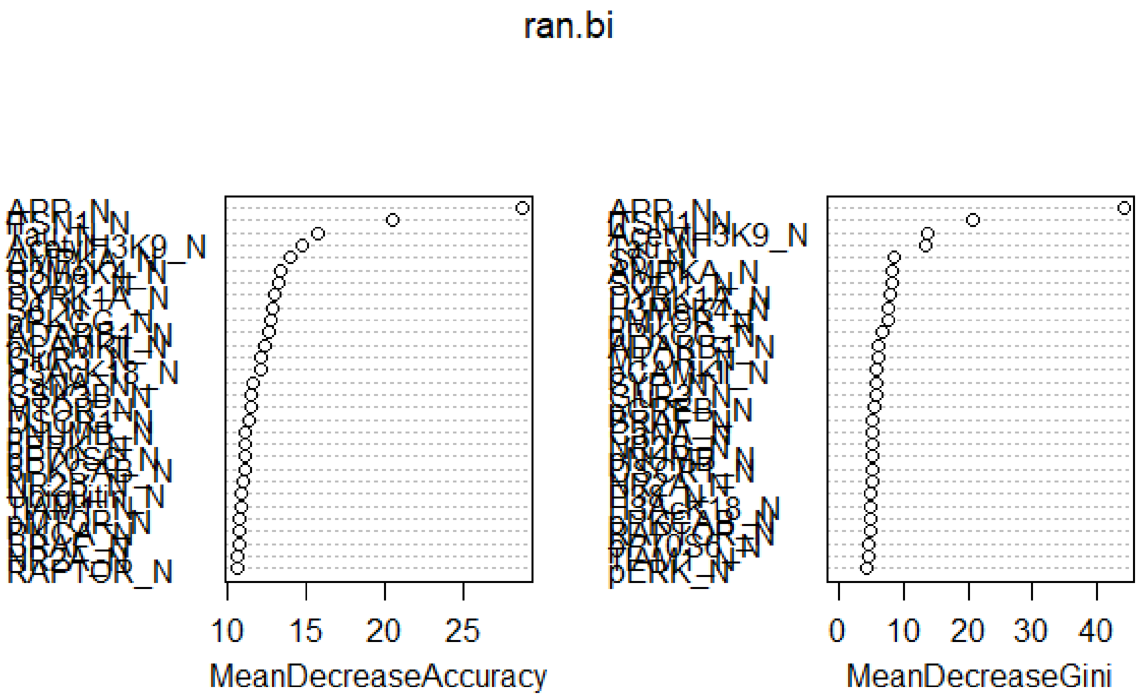


Fig: Random Forest Binary

From the above figure we can say that APP_N is considered as a most important variable considering MeanDecreaseAccuracy and MeanDecreaseGini.

	Binary Class	**MultiClass**
Random Forest	**98%**	**98%**
Bagging	**95%**	**98%**
Boosting	**99%**	**99%**

For the fits between Random Forest, Bagging and Boosting we have the best results for Multiclass and Binary class for the Boosting fit.

Discussion:

First, I have performed Fitting Decision Tree and Support Vector Machine keeping Class as the response variable for the multiclass and Genotype as the response variable for the binary class using 70% of the data for the training and 30% of the data for testing for both. I have also performed tuning for obtaining the best accuracy results. Next, I have tried implementing PCA and performed all the steps for Decision Tree and Support Vector Machine for Multiclass and Binary class and tuning for obtaining the best results. SVM gave the best results compare to the decision tree and SVM for the original fit gave the best results than the PCA fit.

For the third part in the task 1, performed Random Forest, Bagging and Boosting for Multiclass using Class as the response variable and Binary Class using Genotype as the response variable.

Task 2:

Clustering:

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them to clusters [7].

Kmeans Clustering:

Kmeans algorithm is an iterative algorithm that tries to partition the dataset into kpre-defined distinct non-overlapping subgroups which are called as clusters. Where each data point belongs to only one group.

Hierarchical Clustering:

Hierachical clustering involves creating clusters that have a predetermined ordering from top to bottom.

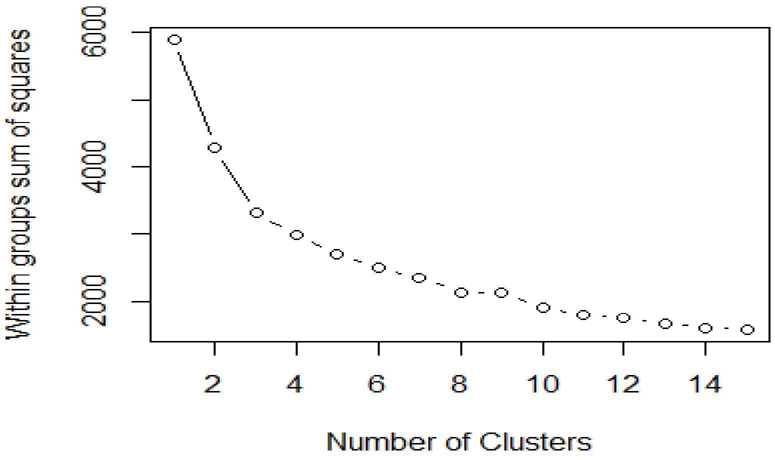
Methods:

Kmeans Clustering:

Kmeans is performed on the 77 proteins features of the nuclear fraction of cortex of the mice. For the clustering, I have removed all the categorical variables from the data set and performed the clustering.

Clusters:

The number of the clusters are decided based on the elbow plot. We also know that our class variable is divided into 8 classes.



From the above plot, I have decided to choose the number of cluster K as 8 but k= 6,7 are also a best choice to choose. We can choose K = 2 for the variables Genotype, Behaviour and Treatment.

Hierarchical Clustering:

Hierarchical clustering analysis is performed for the data which consists of nuclear fraction of cortex of mice. In Hierarchical clustering each data point is considered as a cluster and then are they are grouped as a single cluster based on the similarities of the features. The results are displayers in the form of dendrogram.

Results:

Kmeans Clustering:

When K = 8:

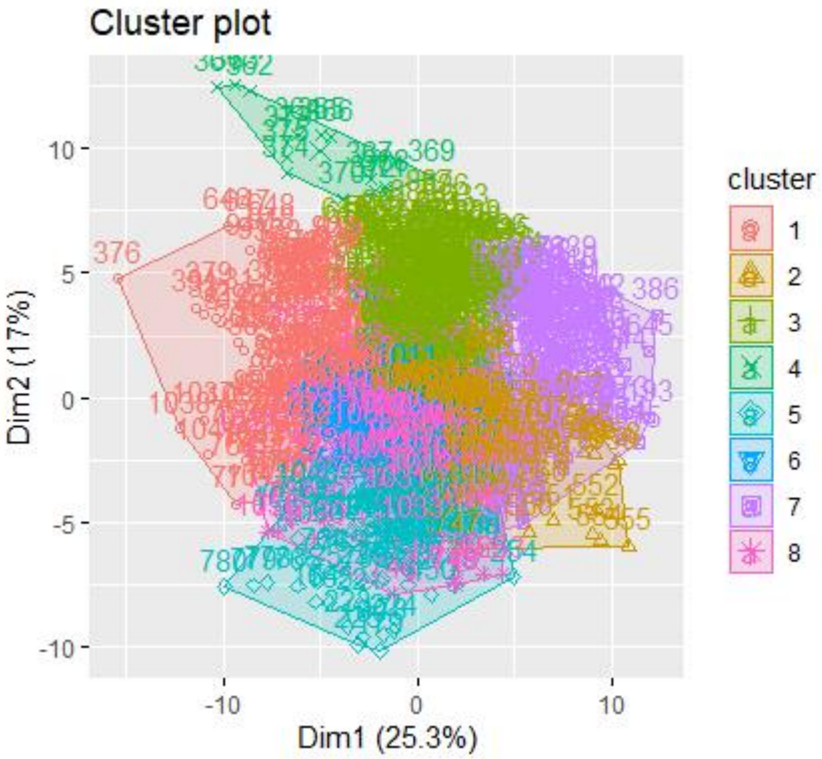
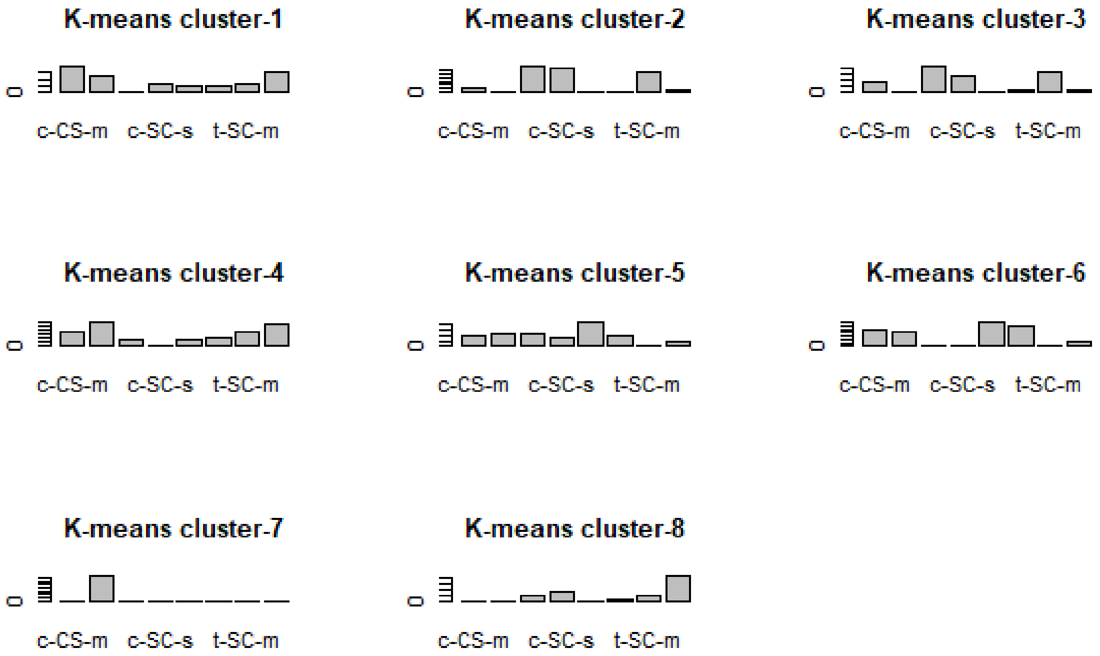


Table:

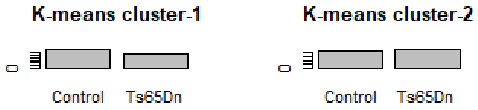
	c-CS-m	c-CS-s	c-SC-m	c-SC-s	t-CS-m	t-CS-s	t-SC-m	t-SC-s
1	39	24	0	12	11	11	14	31
2	11	0	69	65	0	0	56	7
3	18	0	43	27	0	3	34	3
4	17	31	8	2	9	11	17	29
5	19	24	22	14	44	20	3	9
6	46	41	0	0	71	57	0	12
7	0	15	0	0	0	0	0	0
8	0	0	8	15	0	3	11	44



For K = 8, from the above figure we can say that only cluster 7 has clearly classified compared to other clusters from that we can conclude that the clusters have not been efficiently clustered.

For K = 2 and cluster for Genotype

	Control	**Ts65Dn**
1	**345**	**275**
2	**225**	**235**



From the above figure for Genotype we can see that the clusters have not been correctly classified.

For K = 2 and cluster for Behaviour:

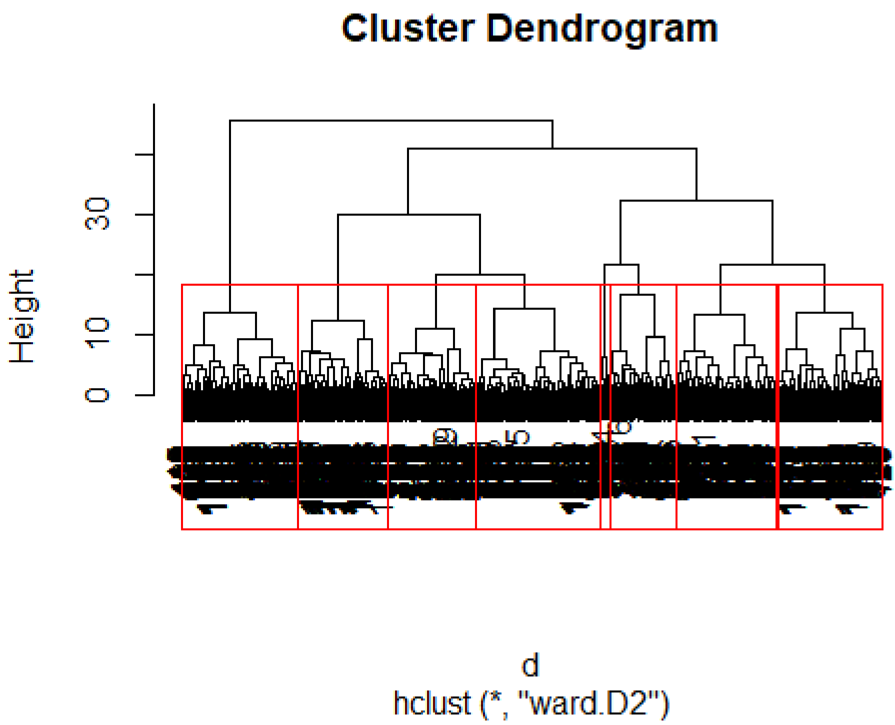
	C/S	**S/C**
1	**255**	**365**
2	**270**	**190**

For K = 2 and cluster for Treatment:

	Memantine	**Saline**
1	**336**	**284**
2	**234**	**226**

Kmeans clustering analysis works very fast and it is efficient with huge amounts of the datasets. But from the above tables we can see that the cluster data has been mis-classified.

Hierarchical:



I have decided with the optimal number of clusters by visualizing the dendrogram. The threshold is picked based up on the highest number of clusters after cutting the dendrogram. The distance metric used is “Euclidean” and the method is “ward.D” through hclust.

Discussion:

I have performed Kmeans and the Hierarchical clustering and chosen the best cluster numbers to fit to the model. I felt that the cluster have not been classified accurately. I choose to have 8 cluster for the multi class as we have 8 categorical variables in the class feature and 2 cluster for Binary class as we have 2 different categorical variables in the columns in the Genotype, Behaviour and Treatment features.

References

[1] R. S. Bird, "Medium," Medium, [Online]. Available: https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb. [Accessed 01 06 2020].

[2] R. Gandhi, "towarddatascience," towarddatascience, [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47. [Accessed 06 01 2020].

[3] H. Goonewardana, "Medium," Medium, [Online]. Available: https://medium.com/apprentice-journal/pca-application-in-machine-learning-4827c07a61db. [Accessed 01 06 2020].

[4] W. Koehrsen, "towards data science," towards data science, [Online]. Available: https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76. [Accessed 01 06 2020].

[5] Kangrinboqe, "rpubs," rpubs, [Online]. Available: https://rpubs.com/kangrinboqe/268745. [Accessed 01 06 2020].

[6] Datatechnotes, "datatechnotes," datatechnotes, [Online]. Available: https://www.datatechnotes.com/2018/03/classification-with-gradient-boosting.html. [Accessed 01 06 2020].

[7] S. Kaushik, "Analytics Vidhya," Analytics Vidhya, [Online]. Available: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/. [Accessed 03 06 2020].