PREDICTION OF INFECTED SAMPLES IN AGRICULTURAL DATABASE USING DATA MINING TECHNIQUES

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT

FOR THE AWARD OF THE DEGREE

OF

MASTER OF TECHNOLOGY

IN

INFORMATION SYSTEMS

Submitted by:

Shreya Khurana

2K16/ISY/14

Under the supervision of

Dr. Kapil Sharma

IT DEPARTMENT

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

May, 2018

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, Shreya Khurana, Roll No 2K16/ISY/14 student of M.Tech (ISY), hereby declare that the project Dissertation titled "Prediction of Infected Samples in Agricultural Database using Data Mining Techniques" which is submitted by me to the Department of Information Technology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma, Associateship, Fellowship or other similar title or recognition.

Place: Delhi **Shreya Khurana**

Date:

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "Prediction of Infected Samples in Agricultural Database using Data Mining Techniques" which is submitted by Shreya Khurana, Roll No 2K16/ISY/14, Department of Information Technology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi
Date:

DR. KAPIL SHARMA
SUPERVISOR
(HOD, IT)

ABSTRACT

Data mining in agriculture has come up very recently as a research topic. It is based on using data mining techniques on agricultural data to derive some analysis and patterns out of the data. Many technologies are available these days which help to obtain a lot of information on agriculture and this can be further analysed to get more important information.

Data mining in agriculture has a lot of applications like predicting crop yield, predicting the optimum environmental parameters for improving crop yield, predicting the infections in crops etc. By obtaining such useful information we can increase our productivity of crops, we can avoid many problems like infestations in the crops and we can sow the crops according to the optimum conditions. We can try to make the soil more suitable for the crops which are to be grown in them. Hence by using data mining techniques we can extract important and useful information from agricultural data sets which can further be used for important applications and in daily life practical problems.

In this project, agricultural data has been analysed which had been provided by the Indian Agricultural Research Institute, New Delhi. The data is regarding various crops imported from various countries over the past 5 years. It focuses on the interceptions (in the form insects, fungi, viruses etc.) present in the crop samples imported. Data mining techniques have been implemented in R Language over the data to discover interesting patterns and relationships in the data.

In particular, clustering algorithms are used to analyse the data and cluster it. From the clusters effort has been made to predict whether a particular crop to be imported from a particular country will be infected with some interception or not and to determine the probability of the samples being infected. On the basis of such result, we can find out whether it will be beneficial for us to import a crop 'A' from a country 'B'.

ACKNOWLEDGEMENT

I am very thankful to Dr. Kapil Sharma (HOD, Department of Information Technology) and all the faculty members of the IT Department of DTU. They all provided us with immense support and guidance for the project.

I would also like to express my gratitude to the university for providing us with the laboratories, infrastructure, testing facilities and environment which allowed us to work without any obstructions.

I would also like to appreciate the support provided to us by our lab assistants, seniors and our peer group who aided us with all the knowledge they had regarding various topics.

SHREYA KHURANA

2K16/ISY/14

CONTENTS

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 General Introduction: Data Analysis

Data analysis is a combination of several processes with the goal of discovering useful information and coming at conclusions which help in decision making. These processes are the inspection of data followed by cleaning and transformation and then modelling it by various algorithms. There are various strategies for data analysis which have various techniques under different names depending on the domains.

Hence data analysis means obtaining raw data and then transforming it into information which is usable in decision making.

Data mining is a particular technique of data analysis. The main focus of data mining is modelling the data and knowledge discovery which is used for prediction rather than just descriptive analysis. it also includes other techniques like predictive analysis which uses statistical models for prediction and classification and text analysis which use statistical techniques to obtain information from textual sources.

1.2 Data Mining

Data mining is a process of data analysis whose aim is to analyse large amounts of data known as big data with the hope of searching patterns and relationships between various attributes in the data and then to validate the results by testing the patterns with the help of another data set. Hence data mining is performed with the goal of prediction and predictive data mining has a lot of applications in business sector. Data mining techniques are also used in other domains like economy etc and their importance is also being realized in agriculture related areas.

The process of Knowledge Discovery in databases has the following stages:

1) selection
2) preprocessing
3) transformation
4) data mining
5) interpretation.

There are other variations also such as the cross industry standard process for data mining which have different phases. In a simplified way a process can be mentioned as preprocessing, data mining and results validation. Hence, data mining is the step in KDD at which analysis has to be carried out.

Data mining techniques can be divided into two categories: descriptive and predictive.

1) Descriptive data mining techniques describe the general properties of data present in the data set with the help of various techniques like classification etc.

2) Predictive data mining techniques are used to predict some values explicitly based on the patterns which are obtained from known results. In most cases predictive data mining approaches are used.Basically prediction means to use some attributes in the data set which are already known in order to predict the unknown or future values of other attributes which are of interest.

Usually the data set for mining is selected from a data mart or a warehouse since it has to be a large data set or big data for it to contain useful patterns to be obtained during mining. Pre-processing includes cleaning the data set to remove any noise or anomalies like missing data values in the data.

Data mining can be done in 6 different ways.

1) Anomaly detection- This includes finding some unusual data observations or errors in the data which might help in giving some interesting information or may require deeper understanding.

2) Association rule mining- This means searching for relationships between the attributes in the data, for example in a supermarket one may be willing to find which products are being purchased together by the customers and then this information can be used for marketing purpose. The idea is to find how different items in a transactional database are related to each other. Elements which occur together repeatedly within a database are found out with the help of association rules. Association rule mining algorithms include Apriori algorithm, partition ,dynamic, hashing and pruning etc.

3) Clustering- In clustering we group the data observations according to their similarity without knowing the structure in the data. It is an unsupervised learning technique. Data records are partitioned into clusters such that the data observations present in each cluster are similar or close to each other. Hence clustering helps to find categories in the data or groups in the data. It can be used to find some information about the groups. In this case we don't know the structure about the data beforehand. Similar data points are grouped together where as different data points are grouped in different groups. Clustering techniques include hierarchical clustering, partitioning clustering, density based, methods grid based methods, soft computing methods, fuzzy, neural network.

4) Classification- it is a supervised learning technique. It can be used on training data set to form a model which predicts class labels for the data and this model can then be used to predict class labels for the test data set. Hence classification algorithms can be used to classify the data into different class labels. This is helpful in predicting future trends. Classification algorithms include rule based classifier, bayesian networks, decision tree, artificial neural network, support vector machine.

5) Regression- in regression we find a function which models the data such that each data item can be mapped to a real valued prediction variable with the least error. Regression is of two kinds: Linear and non - linear.

6) Summarization- this gives a summary of the data set which may include generating reports or visualising the data.

Data Validation- The next step after data mining is validating the results to know whether they actually predict any future behaviour and can be reproduced on a new data set. If too many hypotheses are investigated and proper statistical hypothesis testing is not performed then the results may not be of much use.

Applications of data mining include games, business, science and engineering, medical data mining, spatial data mining, surveillance, pattern mining.

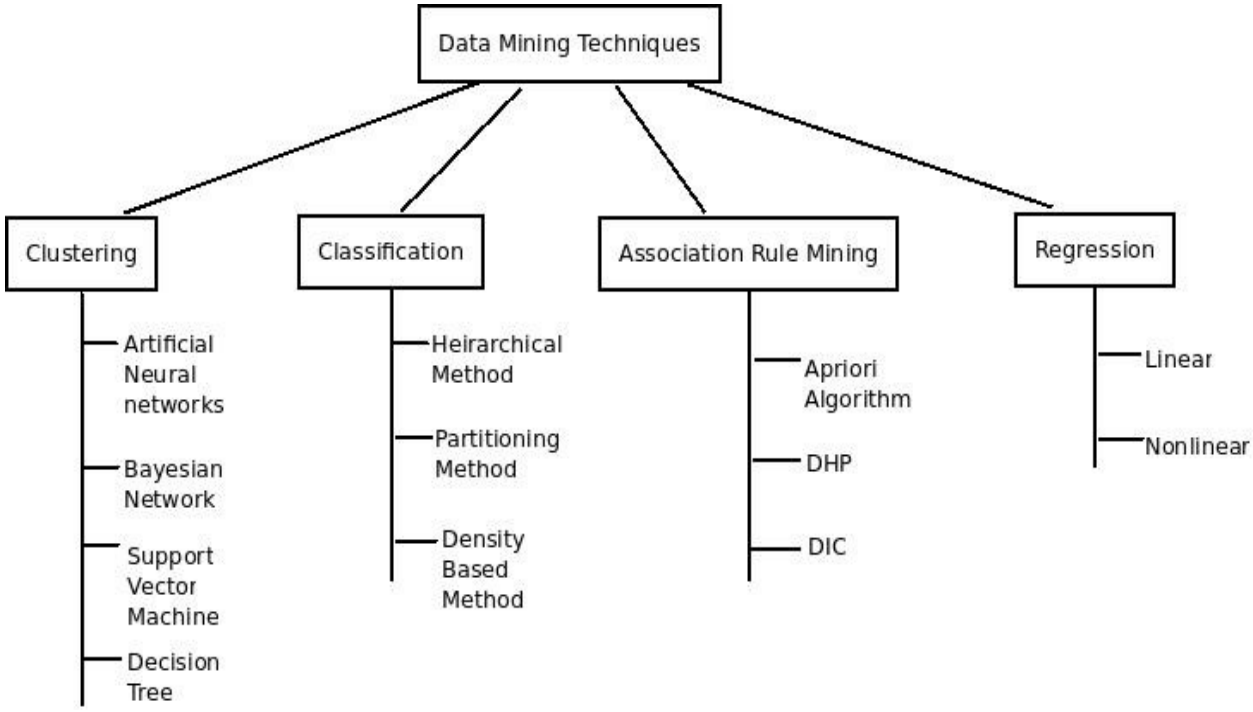

Fig 1.1: Data Mining Techniques

1.3 Data Mining in Agriculture

Data mining techniques can be applied to agricultural data sets with the aim of discovering interesting knowledge and patterns which can further be used for predicting the future patterns of Agricultural processes. In data mining we analyse the data from various perspectives and then summarise it to obtain some useful information which can be used for decision making. Any type of data can be analysed by data mining which can be obtained from a relational database, data warehouse, web server log, or simple text file. To analyse data effectively we need to understand the appropriate techniques of data mining.

Research has been carried out on how data mining techniques can be used for agricultural data sets. For the classification of soils in large soil experiment data sets, Naive Bayes data mining technique can be used.Also in order to predict the fertility of soil decision tree algorithm can be used.
Various data mining techniques can be used for crop yield analysis. These include algorithms like k-means, K nearest neighbour (KNN), artificial neural network (ANN) support vector machines (SVM).

For agricultural data mining predictive techniques are used in order to predict the future crop field, to do weather forecasting, to predict which and how much pesticides and fertilizers are to be used and where to be used, how much revenue can be generated, which are the optimum conditions for crop production and so on.
Clustering data mining techniques are widely used in agricultural applications.

1.4 Cluster Analysis

Cluster analysis in an important part of data analysis as it is used to study the internal structure of data set. Data clustering is unsupervised, descriptive and summarises the data to reveal important information from it; hence it has become a core method of data mining. Cluster analysis can be defined as finding groups which naturally occur in the dataset when little or no prior information is known about the categories present in the data. That is the reason clustering is known as unsupervised classification because of the unknown structure of the data whereas in classification the category structure of data is known beforehand.

In order to carry out the task of cluster analysis successfully we need to consider some elements in a cluster analysis study which are as follows:

1) Data presentation
2) Choosing the objects
3) Choosing variables(data units/variables)
4) What to cluster
5) Normalisation
6) Choice of distance measures
7) Missing data strategy
8) Clustering criterion
9) Clustering algorithm and its implementation
10) Number of clusters
11) Interpretation of results

1.5 Hierarchical Clustering

Hierarchical clustering technique is a method of cluster analysis in which hierarchy of clusters is being built from bottom to top. At the first stage all the observations are taken individually. Then in each successive stage the observations are clustered together into groups of clusters depending on the similarity between them. The similar data observations will be clustered together whereas the dissimilar data objects will be present in different clusters.

So in the initial stage we assign each item to one cluster. Hence if there are N data items we will have N clusters, each containing only one data item. So the distances or similarities between the clusters will be the same as the distance between the items which are present in them. Next we intend to find the most similar or the closest pair of clusters among them and then we will me into merge them into a single cluster. In this way the number of clusters will reduce by one.
Then again we will compute the distances between the new cluster and each of the old clusters and merge the most similar or the closest pair of clusters. In this way we keep combining the closest clusters at each step till all the items are present in one single cluster which has the size N (number of data items).

The computation of distances between each of the cluster pairs can be done in 3 different ways which are: single linkage, complete linkage and average linkage clustering. In single linkage clustering (minimum method) the distance between two clusters is taken as the shortest distance between a particular member of one cluster to a member of the other cluster. Or in other words if similarity between the data is considered then the greatest similarity between a member of one cluster to another cluster is taken as the similarity between the two clusters; because if the distance is the least then the similarity will be the greatest. In complete linkage clustering the distance between two clusters is considered to be as the distance greatest distance or the maximum distance from a particular member of one cluster to a member of another cluster. In average linkage clustering, the distance

between two clusters is taken as the average distance from a particular member of one cluster to a member of another cluster.

The type of hierarchical clustering described above is known as agglomerative hierarchical clustering or bottom-up since it merges the clusters in an iterative manner. Another type of hierarchical clustering is divisive hierarchical clustering or top-down in which all the objects are taken initially into one cluster and then that one cluster is divided into smaller clusters. Generally agglomerative hierarchical clustering is used because divisive hierarchical clustering methods are not generally available.

1.6 Partitioning Clustering

Partitioning based clustering algorithms are iterative relocation algorithms in which the data points are iteratively relocated in order to minimise a given clustering criteria. The relocation is done between clusters such that the optimal solution is attained. In this case we know the number of clusters that have to be formed unlike in hierarchical clustering.
Each cluster group must contain at least one data object and at the same time each object must belong to only one group. Hence the clusters should not be overlapping.
If K clusters are to be formed and the clustering criteria is J then we choose K centres initially. Then depending on the criterion function we compute which data points will belong to which cluster. After this computation we will update the cluster centres depending on the new membership of the data observations. Steps 2 & 3 will be repeated until the data points do not change cluster.

1.6.1 K-Means Clustering Algorithm

K means clustering algorithm is one of the most popular algorithms for clustering in an unsupervised way.
Firstly we need to define the cluster centroids which will be K for K clusters. The position of the centroid needs to be defined such that they are as far away from each other as possible. The next step we do is that we take each data point from the data set and then compute its distance with the centroids and place it near that centroid to with which its distance is the least or with which it is the most similar. In that way with each iteration we form the clusters. Then after each iteration we need to re-calculate the centroid of the clusters which we have found in the previous step. Now we have K new centroids so again we need to check which data points will fall near which centroid. In this way the iterative relocation continues till a locally optimal solution is found.

1.6.2 CLARA Algorithm

Clara is especially designed to handle large data sets. It stands for clustering large applications. Clara doesn't take the whole data set into consideration instead it uses a small portion of the actual data as a representative of the data then PAM is applied on that sample data set or subset and mediods are chosen. PAM is also a partitioning based algorithm. It finds a set of objects which are known as mediods that are centrally located. The mediods are chosen from the data set only then the nearest data points to the mediods are calculated and they are grouped together as clusters. Initially k objects are chosen as the initial mediods then each object is assigned to the cluster with the closest mediod. The quality of the k-medoids is improved by randomly selecting a non-mediod object and swapping a mediod with it. The cost of swapping is calculated and the minimum swapping cost is found.

CLARA applies PAM to a sample of the data set and finds the mediod of the sample. CLARA picks up multiple samples from the data set and gives the best clustering as output. Here the quality of clustering is measured depending on the average dissimilarity of the objects in the data set.

CHAPTER 2: LITERATURE REVIEW

Different researchers all over the world have developed various data mining methodologies which can be used for forecasting and predicting various useful information in the field of agriculture.

Researchers like Majumdar, Naraseeyappa and Ankalaki [1] have used clustering data mining techniques like PAM, CLARA, DBSCAN and multiple linear regression for the analysis of agricultural data and they have tried to obtain optimal parameters in order to maximize the production of crops. They obtained the agricultural data of different districts in Karnataka which was based on various parameters like year, district, crop, area of production, temperature, rainfall, soil, pH value, soil type, fertilizers, etc.
DBSCAN method has been modified and data clustering is done on the basis of district which have same temperature, rainfall and soil type. This modification of DBSCAN method has been done by automatically determining the optimal eps value which was not done in the traditional DBSCAN method. Hence by the modified DBSCAN method they proposed to find the minimum points and Epsilon value automatically. This is done with the help of KNN Plot and Bachelor Wilkins clustering algorithm to find the value of K. Then they have also applied PAM and CLARA in order to cluster the data on the basis of district having the maximum production of crop. On the basis of the results of these two clustering they have tried to obtain the optimal parameters to have the maximum crop production. They have also tried to predict the annual crop production with the help of multiple linear regression technique.

A.Mucherino, Petraq Papajorgji, P.M. Pardalos have presented some of the data mining techniques which are mostly used in the field of agriculture. They have discussed some useful techniques like K-means, K nearest neighbour, artificial neural networks and support vector machines in detail and have also given an insight as to how these can be used in the field of agriculture. They have also suggested directions for future research work in the field of agriculture. They have discussed clustering and classification techniques of data

mining and their uses in agriculture. They have mentioned how data mining techniques are useful to study weather conditions and forecast, eg: k-means method can be used to obtained forecast of atmospheric pollution. Another application of data mining techniques is to study the characteristics of soils. [2]

Ramesh and Vishnu Vardhan have made an effort to predict the crop yield of the future year.[3] They have taken the data from the districts of Andhra Pradesh which consist of eight variables which are year, rainfall, area of sowing, production, etc. They have used multiple linear regression technique and density based clustering technique for this purpose. Multiple linear regression is used with seven predictors or independent variables and one dependant variable that is production. Then they have tried to know the crop yield of particular regions in that district through these two methods.

Ramesh and Ramar have obtained soil database of a large size from Kanchipuram and Tamilnadu. They have tried to classify the soil types based on various soil properties. This classification helps to decide which soil type is best suited for growing which crops, for example the soils of Kanchipuram District are categorised into eight classes and class 1 is the best suited for crop farming in rows as it has many desirable properties like it is not prone to erosion, its drainage is good and it texture is also suitable. In this paper they have proposed the usage of WEKA tool for the classification of soils with the help of Naive Bayes classifier.[4]

Gandhi and Leisa Armstrong have given an extensive study of various data mining techniques like artificial neural network, bayesian networks and support vector machines and their applications in agricultural domain. They have shown how these techniques can be used in order to obtain the relationship of climatic and other factors which affect crop production. They have proposed how these methods can further be used on complex agricultural databases for the purpose of predicting crop yield. This can also be done by integration of seasonal and spatial factors with the help of GIS Technologies.[5]

This paper also provides a survey of how data mining techniques can be used in agricultural domain. They have studied various techniques like K-means, KNN, SVM, Naive Bayesian classifier. They have mentioned that K means can be used for soil classification using GPS based Technology and forecasting atmospheric pollution; KNN can be used for predicting weather and soil water parameters; neural networks can be used for predicting maturity dates of soybean and with water resource variables; and support vector machine can also be used in classification of crops and analysing how the climate changes[6]. They have also mentioned that unsupervised clustering can be used to generate clusters and determine any kind of patterns that exist in the data. WEKA tool can also be used for classification as in for the sorting and grading of Mushrooms.

CHAPTER 3: PROPOSED METHOD

3.1 Tool Survey

R Studio: R Studio is an integrated development environment (IDE) for writing and executing R scripts. R studio has various features like database viewer, code editor, console and data visualisation which are present on a single screen. In R Studio we can easily write and execute R scripts. It also has the feature of running the code line by line. It also enables importing and exporting of R scripts and also draw beautiful plots as results of data mining. It is free and open source.

CSV Database: It is a comma-separator values files. It stores all types of data in an easily readable format.

3.2 Cluster Analysis in R

Traditional algorithms of cluster analysis are available in R through various functions. R allows extensive cluster analysis through the use of many different functions and plots.

3.2.1 Hierarchical clustering in R

Hierarchical agglomerative clustering forms a cluster for each of the observations initially. It then calculates the distance between observations and clusters the observations together which are the closest to each other. It continues like this till one cluster with all the data objects is formed. The distance between the observations can be taken as either the minimum distance between an observation and a particular member of the cluster, or the maximum distance or the average distance or some other method can also be used which will minimise the distance between the observations within the cluster. Agglomerative hierarchical clustering in R is done through the hclust function. The Cluster library in R also provides agnes function which is similar to H class function, the only difference being in its updating method of distance matrix of observations.

R provides dendogram in order for us to analyse the cluster solution. Hclust or agnes function is used to do hierarchical clustering. The result obtained from the clustering is plotted with the help of plot function and the resulting plot formed is a dendrogram or a tree which shows the hierarchy of clusters formed from bottom to top. Hclust function in R uses complete linkage method for hierarchical clustering hence the maximum distance between the members of clusters will be taken as the cluster distance. At every iteration the clusters between whom the distance will be the least will be merged together to form a single cluster. This process is continued till the whole data set is agglomerated into one cluster. clusplot function can also be used in order to plot the clusters formed in a graphical way.

3.2.2. Partitioning clustering in R

K-Means Clustering

The most popular method of partitioning clustering is k-means. In this one needs to mention the number of clusters which have to be formed from the data. R allows k-means algorithm to be implemented with the help of k-means function. Then k observations will be chosen as centres for the clusters and the distance of every other observation is calculated with respect to the K clusters and the observations are put in those clusters to which they will be the nearest. After new clusters are formed, the centres are recalculated and if required the observations might be moved to a different cluster to which they will be closer, so again the distances will be calculated. This continues still no observations change their clusters.

The number of clusters can be formed with the help of obtaining a plot of within groups sum of squares by number of clusters. It forms an elbow plot and the number of clusters at which the bend is obtained determines the appropriate number of clusters which should be taken for the data set.

```
#Obtaining the number of clusters
wss <- (nrow(data)-1)*sum(apply(data,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(data,
centers=i)$withinss)
plot(1:15, wss, type="b", xlab="Number of Clusters",
ylab="Within groups sum of squares")

#Applying k-means algorithm in R
res <- kmeans(data, 3)

# get cluster means
aggregate(data,by=list(res$cluster),FUN=mean)
```

CLARA clustering

R Cluster library also gives an option of using partitioning around medoids or PAM clustering which acts as a better alternative to K-means clustering. Mediod means such an observation inside a cluster for which the sum of distances between itself and every other cluster member is minimum. Like K-means we need to mention the number of clusters beforehand but it is more accurate than K-means because it makes sure that the mediods are true representatives of the clusters. For this it does more extensive computation than K means. In K-means the cluster centres, which may or may not be actual observations, are calculated after all the observations have moved from one cluster to another. Whereas in the case of PAM the sum of distances between objects within a cluster are calculated constantly when the observations are moving between the clusters. Hence it also provides a more reliable cluster solution. Also it identifies the observations that represent the mediods which can be taken as a representative of each cluster.

For large data sets, CLARA applies PAM to few samples of the data set by taking some samples at a time, instead of the entire data set at a time and obtains the best clustering solution.

CLARA is implemented in R in the following way:
clara(x, k, metric = "euclidean")

x= the data set to be clustered; only numeric value variables are allowed
k= number of clusters obtained from wss plot; 0<k<n where n is the number of
observations
metric= the distance measure to be used; current options are:
Euclidean: root sum-of-squares of differences
Manhattan: sum of absolute differences

3.3 Proposed Work

The data for this project has been taken from the Indian Agricultural Research Institute, PUSA road, New Delhi. It consists of the data of crops imported from various countries over the past few years. The attributes present in the data include the name of the crop, the source country, the year of import and number of crops infected out of the total number of samples. With the help of data preprocessing, the percentage of infected samples for each observation is obtained Some of the samples of the crops imported have some interception in the form of insects or viruses and fungi. Sometimes all the samples may be infected or sometimes only a few out of many. When all the samples or majority of the samples turn out to be infected then it is a huge loss for our country. Hence we should be able to predict whether a particular crop should be imported from a particular
country on the basis of previous data of that crop being imported from that country. We try to find out the possibility of a particular crop imported from a particular country being infected and on the basis of the result we can determine if importing a crop from a particular country will be beneficial for us or a bad deal for us.

R language is used to analyse the data using data mining

Both Hierarchical clustering and Partition-based clustering are used to cluster the data points which are similar to each other together in a cluster. The clustering is done on the basis of the attribute
PERCENTAGE OF INFECTED SAMPLES. So those data points or observations which have percentages near to each other will be clustered together.

The appropriate number of clusters is obtained with the help of within sum of squares with the number of clusters plot. This forms an elbow plot and the bending point gives the appropriate number of clusters.

After we have obtained the result of hierarchical clustering algorithm, we need to find the

mean of each of the clusters that is the mean of the percentage of infected samples. Firstly we obtain a dendogram plot and then assign the points to clusters by cutting the dendogram tree at a given height(h) or by specifying the number of clusters to be formed(k). Then we can find the mean of each of the clusters

In the case of partitioning algorithm the result itself gives a set of clusters along with their means in the case of k-means and with their mediods in the case of CLARA.

Then the problem statement is to find out whether a crop 'A' should be imported from a country 'B'. For this we will find the data point corresponding to the crop 'A' and country 'B' and then find which cluster does the data point belong to. Depending on the mean or mediod of the cluster to which the data point belongs to, we can predict the possible percentage of infection to be found when 'A' is imported from 'B'. On the basis of this, it is possible to know whether importing 'A' from 'B' is beneficial or not.

Hence this work is a combination of clustering and prediction. In clusters the data set on the basis of the percent of infected samples. Then given a country name and a crop name we predict the cluster to which it belongs and depending on the cluster mean we decide whether to import the crop from that particular country or not.

We then try to make a comparison of the results obtained from all the three clustering algorithms that is hierarchical agglomerative clustering, K means clustering and CLARA clustering.

CHAPTER 4: RESULTS

4.1 Results of Hierarchical Clustering

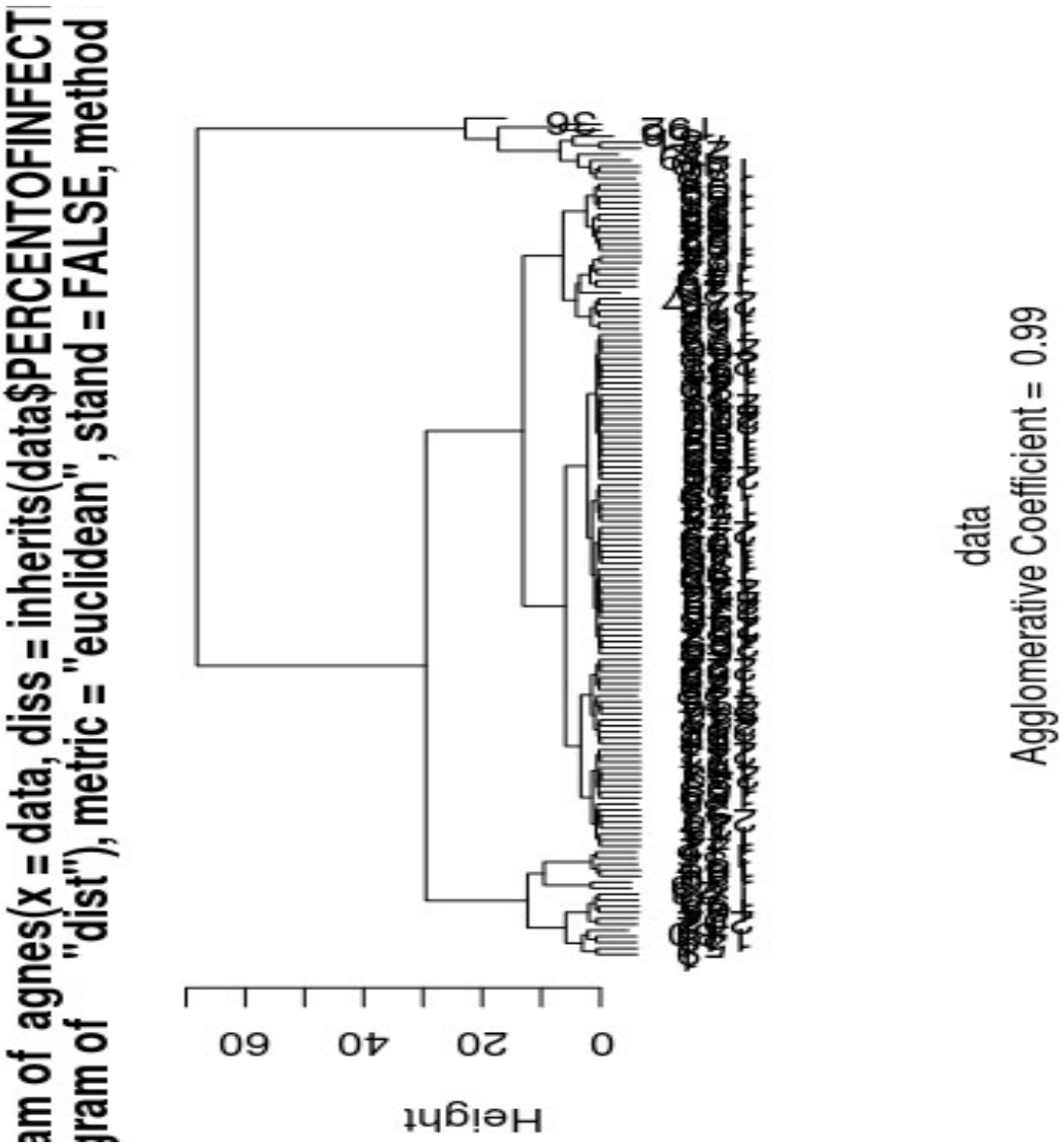

Fig 4.1 Dendogram of Hierarchical Clustering

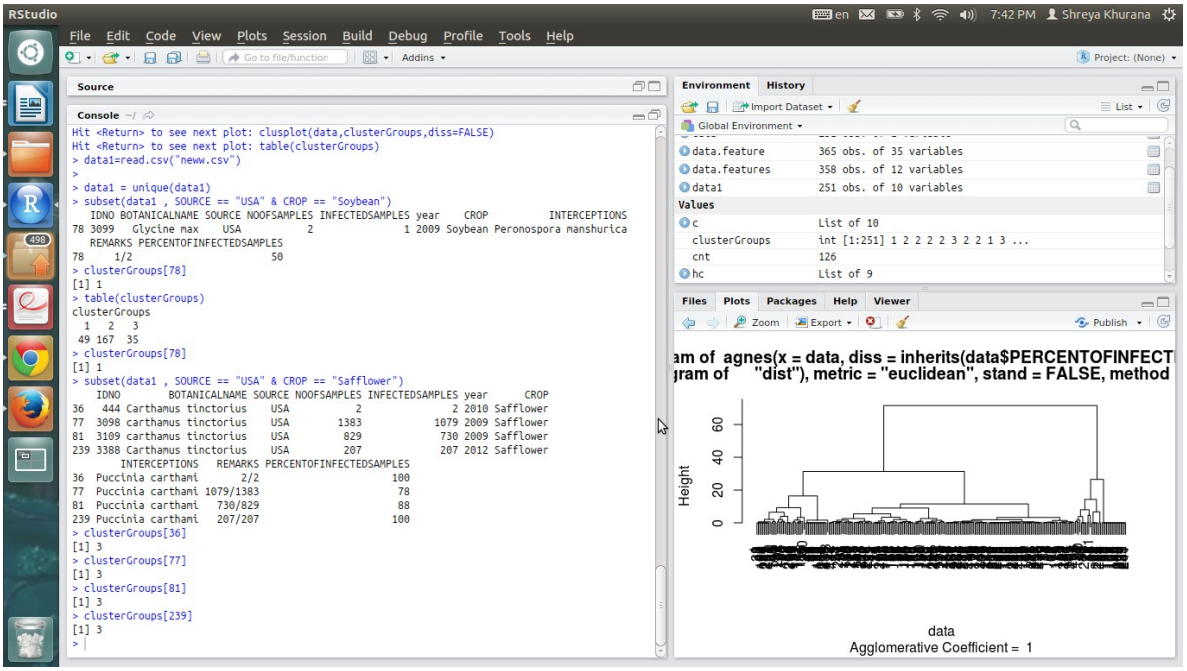

Fig 4.2 Result of Hierarchical Clustering

Mean of PERCENTOFINFECTEDSAMPLES in each cluster:

Cluster 1: 41.8667

Cluster 2: 11.9233

Cluster 3: 77.3134

These are the mean percentage of infected crop samples imported from various countries in each cluster.

So the data points present in cluster 3 indicate a high probability of infected samples imported.

If we want to find out that suppose we should import a crop 'A' from a country 'B' we will write the command in R

subset(data, SOURCE == “A” & CROP == “B”)

[result will be the observation number from our data set say 21]
clusterGroups[21]
[the result will be the cluster number to which the data point belongs and if it is 3 we know that the crop is most likely to be infected so we should not import A from B .
Otherwise if the cluster number is 1 we can still consider but likely not be imported because there is a 41% percent chance that the crop will be infected. For cluster number 2 the crop is less likely to be infected so we can import . This way we can find for any crop and country]

We can take different combinations of country and crop and check which cluster the data point belongs to.

1) Country = USA & Crop = Soybean

#Finding the data point for USA & Soybean

> subset(data1 , SOURCE == "USA" & CROP == "Soybean")

IDNO BOTANICALNAME SOURCE NOOFSAMPLES INFECTEDSAMPLES year CROP INTERCEPTIONS

78 3099 Glycine max USA 2 1 2009 Soybean Peronospora manshurica

```
  REMARKS PERCENTOFINFECTEDSAMPLES
78       1/2                      50
```

#Finding the cluster in which the observation no 78 lies

```
> clusterGroups[78]
[1] 1
```

=> cluster 1 has mean value of percentage of infected samples 41.8 so soybean imported from USA will likely be 41.8 percent infected (on an average). So we see if soybean should be or should not be imported from USA.

2) Country = USA & Crop = Safflower

#Finding the data points for USA & Safflower

```
subset(data , SOURCE == "USA" & CROP == "Safflower")
```

#output in R

```
> subset(data1 , SOURCE == "USA" & CROP == "Safflower")
          IDNO           BOTANICALNAME           SOURCE           NOOFSAMPLES
INFECTEDSAMPLES year  CROP
36   444 Carthamus tinctorius         USA          2              2 2010 Safflower
77  3098 carthamus tinctorius         USA          1383           1079 2009 Safflower
81  3109 carthamus tinctorius         USA          829            730 2009 Safflower
239 3388 Carthamus tinctorius         USA          207            207 2012 Safflower
          INTERCEPTIONS   REMARKS PERCENTOFINFECTEDSAMPLES
36  Puccinia carthami          2/2              100
77  Puccinia carthami 1079/1383                 78
81  Puccinia carthami   730/829              88
239 Puccinia carthami   207/207              100
>
```

#Finding the cluster in which the observations lie

```
> clusterGroups[36]
[1] 3
```

```
> clusterGroups[77]
[1] 3
> clusterGroups[81]
[1] 3
```

All the obervations lie in cluster 4 which has a high mean of 77.3. So we can conclude that safflower imported from USA will most likely be infected and hence should not be imported.

3) Country = Thailand & Crop = Corn

```
> subset(data1 , SOURCE == "Thailand" & CROP == "Corn")
     IDNO BOTANICALNAME    SOURCE NOOFSAMPLES INFECTEDSAMPLES
year CROP     INTERCEPTIONS
25   427     Zea mays Thailand          536              2      2010      Corn      Fusarium
moniliforme
70  3084     Zea mays Thailand          648             11      2009      Corn      Fusarium
moniliforme
121 3171     Zea mays Thailand          98               5      2010      Corn      Fusarium
moniliforme
     REMARKS PERCENTOFINFECTEDSAMPLES
25     2/536                    0.37
70   11/648                1.70
121     5/98                    5.10
> clusterGroups[25]
[1] 2
> clusterGroups[70]
[1] 2
> clusterGroups[121]
[1] 2
```

All the observations lie in cluster 2 which has an average of 11.9 percent of infected samples so we can conclude that we can safely import corn from Thailand.

4.2 Results of Partitioning Clustering

4.2.1 Result of WSS Plot

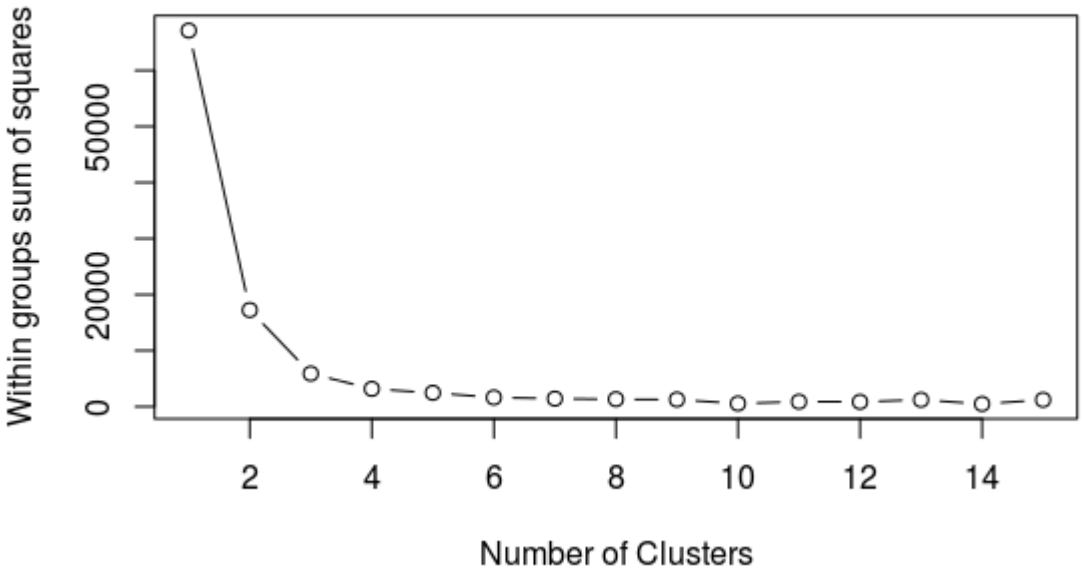

Fig 4.3 WSS Plot

The optimal number of clusters is obtained at the elbow bend point of the graph which comes around 3.

4.2.2 Results of K-Means Clustering Algorithm

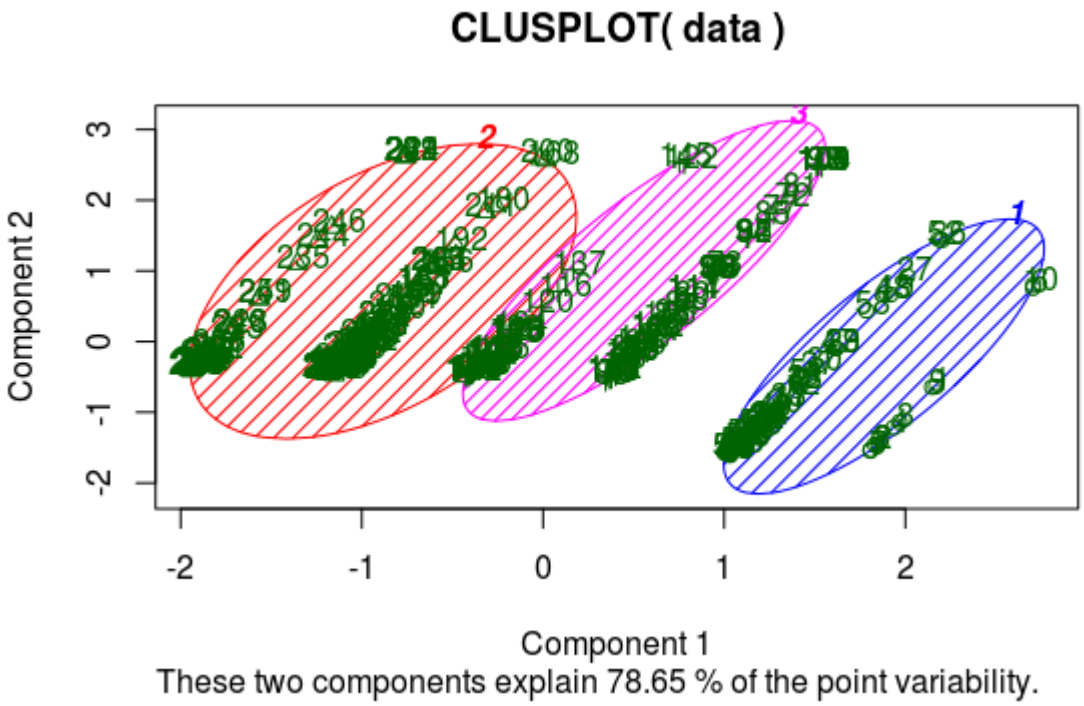

Fig 4.4 Result of K-Means Clustering

#Output in R

K-means clustering with 3 clusters of sizes 59, 105, 87

Cluster means:

	IDNO	year	PERCENTOFINFECTEDSAMPLES
1	432.5593	2009.831	14.55915
2	3320.7619	2011.276	89.03571
3	3145.0115	2009.368	30.69759

“data1” is the original data file with all the attributes present.
“data” is the file which contains the attribute percentofinfectedsamples and the cluster to which the observation
Belongs.

Now again we want to check whether a particular crop should be imported from a particular country by trying to find out that which cluster it will fall in and if the cluster has a high mean value as in the case of cluster 2 in the result then it is beneficial not to import that grow from that particular country.

We can take different combinations of country and crop and check which cluster the data point belongs to.
1) Country = USA & Crop = Soybean
#Finding the data point for USA & Soybean
> subset(data1 , SOURCE == "USA" & CROP == "soybean")
#Output in R
IDNO BOTANICALNAME SOURCE NOOFSAMPLES INFECTEDSAMPLES year CROP
78 3099 Glycine max USA 2 1 2009 Soybean
INTERCEPTIONS REMARKS PERCENTOFINFECTEDSAMPLES
78 Peronospora manshurica ½ 50

#Checking which cluster observation 78 belongs to
> res$cluster[78]
78
3

=> cluster 3 has mean value of percentage of infected samples 30.69 so soybean imported

from USA will likely be 30.69 percent infected (on an average). So we see if soybean should be or should not be imported from USA.

2) Country = USA & Crop = Safflower

#Finding the data points for USA & Safflower

```
> subset(data1 , SOURCE == "USA" & CROP == "Safflower")
     IDNO          BOTANICALNAME      SOURCE          NOOFSAMPLES
INFECTEDSAMPLES year  CROP
36   444 Carthamus tinctorius       USA          2              2 2010 Safflower
77  3098 carthamus tinctorius       USA          1383           1079 2009 Safflower
81  3109 carthamus tinctorius       USA          829            730 2009 Safflower
239 3388 Carthamus tinctorius       USA          207            207 2012 Safflower
     INTERCEPTIONS   REMARKS PERCENTOFINFECTEDSAMPLES
36  Puccinia carthami          2/2              100
77  Puccinia carthami 1079/1383                 78
81  Puccinia carthami   730/829              88
239 Puccinia carthami   207/207              100

> res$cluster[81]
81
 2
> res$cluster[239]
239
  2
> res$cluster[77]
77
 2
> res$cluster[36]
36
```

2

All the obervations lie in cluster 2 which has a high mean of 89. So we can conclude that safflower imported from USA will most likely be infected and hence should not be imported.

3) Country = Thailand & Crop = Corn

```
> subset(data1 , SOURCE == "Thailand" & CROP == "Corn")
     IDNO BOTANICALNAME    SOURCE NOOFSAMPLES INFECTEDSAMPLES
year CROP     INTERCEPTIONS
25   427     Zea mays Thailand          536             2      2010      Corn      Fusarium
moniliforme
70  3084     Zea mays Thailand          648            11      2009      Corn      Fusarium
moniliforme
121 3171     Zea mays Thailand          98              5      2010      Corn      Fusarium
moniliforme
     REMARKS PERCENTOFINFECTEDSAMPLES
25     2/536                    0.37
70   11/648                1.70
121     5/98                    5.10

> res$cluster[25]
[1] 1
> res$cluster[70]
[1] 1
> res$cluster[121]
[1] 1
```

All the observations lie in cluster 1 which has an average of 14.5 percent of infected samples so we can conclude that we can safely import corn from Thailand.

4.2.3 Results of CLARA Clustering

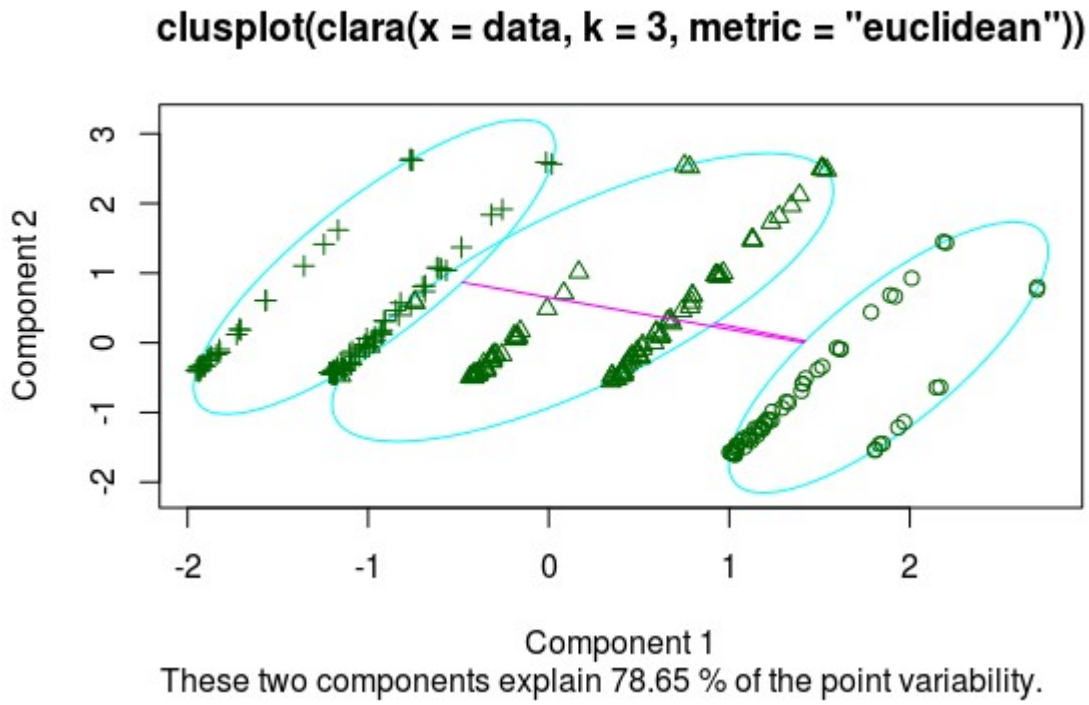

Fig 4.5 Result of CLARA Clustering

#Output in R

```
> c<-clara(data, 3, metric = "euclidean")
> c
Call:     clara(x = data, k = 3, metric = "euclidean")
Medoids:
IDNO year PERCENTOFINFECTEDSAMPLES
33   438 2010                  33.3
110 3151 2009                  3.3
207 3334 2011                  100
```

We again take different combinations of country and crop and check which cluster the data point belongs to.

1) Country = USA & Crop = Soybean

#Finding the data point for USA & Soybean

```
> subset(data1 , SOURCE == "USA" & CROP == "soybean")
```

#Output in R

```
IDNO BOTANICALNAME SOURCE NOOFSAMPLES INFECTEDSAMPLES year CROP
78 3099 Glycine max USA 2 1 2009 Soybean
INTERCEPTIONS REMARKS PERCENTOFINFECTEDSAMPLES
78 Peronospora manshurica ½ 50
```

#Checking which cluster observation 78 belongs to

```
> c$clustering[78]
78
 1
```

=> cluster 1 has mediod value of percentage of infected samples 33.3 so soybean imported from USA will likely be 33.3 percent infected (on an average). So we see if soybean should be or should not be imported from USA.

2) Country = USA & Crop = Safflower

#Finding the data points for USA & Safflower

```
> subset(data1 , SOURCE == "USA" & CROP == "Safflower")
    IDNO BOTANICALNAME SOURCE NOOFSAMPLES INFECTEDSAMPLES year CROP
36   444 Carthamus tinctorius USA 2 2 2010 Safflower
```

```
77  3098 carthamus tinctorius        USA            1383            1079 2009 Safflower
81  3109 carthamus tinctorius        USA            829             730 2009 Safflower
239 3388 Carthamus tinctorius        USA            207             207 2012 Safflower
          INTERCEPTIONS   REMARKS PERCENTOFINFECTEDSAMPLES
36  Puccinia carthami          2/2                100
77  Puccinia carthami 1079/1383                   78
81  Puccinia carthami   730/829                   88
239 Puccinia carthami   207/207                   100

> c$clustering[81]
81
 3
> c$clustering[239]
239
  3
> c$clustering[77]
77
 3
> c$clustering[36]
36
 3
```

All the observations lie in cluster 3 which has a high mediod of 100. So we can conclude that safflower imported from USA will most likely be infected and hence should not be imported.

3) Country = Thailand & Crop = Corn

```
> subset(data1 , SOURCE == "Thailand" & CROP == "Corn")
          IDNO BOTANICALNAME    SOURCE NOOFSAMPLES INFECTEDSAMPLES
year CROP     INTERCEPTIONS
```

25	427	Zea mays	Thailand	536	2	2010	Corn	Fusarium moniliforme
70	3084	Zea mays	Thailand	648	11	2009	Corn	Fusarium moniliforme
121	3171	Zea mays	Thailand	98	5	2010	Corn	Fusarium moniliforme

	REMARKS	PERCENTOFINFECTEDSAMPLES
25	2/536	0.37
70	11/648	1.70
121	5/98	5.10

```
> res$cluster[25]
[1] 2
> res$cluster[70]
[1] 2
> res$cluster[121]
[1] 2
```

All the observations lie in cluster 2 which has an average of 3.3 percent of infected samples so we can conclude that we can safely import corn from Thailand.

4.3 Comparison of Outputs

1) For Soybean to be imported from the USA,

Hierarchical Clustering predicts that it might be 41.8% infected.
K-Means Clustering predicts that it might be 30.69% infected.
CLARA Clustering predicts that it might be 33.3% infected.

Hence, all three algorithms conclude that it might be averagely infected.

2) For Safflower to be imported from the USA,

Hierarchical Clustering predicts that it might be 77.3% infected.
K-Means Clustering predicts that it might be 89% infected.
CLARA Clustering predicts that it might be 100% infected.

Hence, all three algorithms conclude that it might be highly infected and hence should not be imported.

3) For Corn to be imported from Thailand,

Hierarchical Clustering predicts that it might be 11.9% infected.
K-Means Clustering predicts that it might be 14.5% infected.
CLARA Clustering predicts that it might be 3.3% infected.

Hence, all three algorithms conclude that it might be slightly infected and can be imported.

CHAPTER 5: CONCLUSION

With the emerging field of data mining one can easily discover information which is hidden inside data. With the help of advanced technologies surprising information can be obtained from use amounts of data which helps in decision making. Data mining helps best to predict something from large amounts of data or big data and it also helps to discover some interesting patterns and relationships. There are many applications to data mining. One of them being agriculture, in which it is now increasingly being used to help the agricultural sector in improving the productivity of crops, classifying soils, classifying crops, etc. In this project, data mining has been used to analyse the data of crops which is which are imported from various countries and the interceptions in them in the form of worms, fungi, insects or other unwanted mixed crops.

And, effort has been made in order to determine whether a crop should be imported from a particular country or not. R programming has been used which is a very robust language for mining. Rstudio IDE has been used which is a powerful and productive user interface for R. With the help of hierarchical clustering and partitioning clustering in R, effort has been used to cluster similar observations on the basis of the percentage of infected samples. Then the data points which lie in those clusters which have a high average percentage of infected samples can be concluded to have a high infection rate. So, the cluster models have been made and now we can predict whether a crop should be imported from a particular country or not depending on the cluster in which the observation which that crop and country lies. So effort has been made to make a prediction model which can be used to improve the import efficiency of our country. With the help of the prediction model, we will be able to predict the probability of a particular crop being imported from a particular country to be infected. This can be greatly helpful to avoid loss of crops and also benefit us financially. Resulting prediction models of hierarchical clustering and partitioning clustering algorithms have been compared to show similar results.

For developing countries like India, agriculture is one of the most important occupation of a

large section of the population and if data mining can be used effectively in the field of agriculture it will be greatly useful to the farmers who are directly involved in it as well as to the population as a whole. This is because with the help of data mining we can discover hidden information and help the farmers to increase crop productivity by making important decisions with the help of the discovered information.

CHAPTER 6: REFERENCES

[1]Majumdar, J., Naraseeyappa, S. & Ankalaki, "Analysis of agriculture data using data mining techniques: application of big data," S. J Big Data (2017) 4: 20. https://doi.org/10.1186/s40537-017-0077-4

[2]Mucherino, A., Papajorgji, P. & Pardalos, "A survey of data mining techniques applied to agriculture," P.M. Oper Res Int J (2009) 9: 121. https://doi.org/10.1007/s12351-009-0054-6

[3]D. Ramesh, B. Vishnu Vardhan, "Analysis of crop yield production using data mining techniques," International Journal of Research in Engineering and Technology, 4(1), 470(2015).

[4] Vamanan, Ramesh & Ramar, K. (2011). Classification of Agricultural Land Soils: A Data Mining Approach. Agricultural Journal. 6. 82-86. 10.3923/aj.2011.82.86.

[5] N. Gandhi and L. J. Armstrong, "A review of the application of data mining techniques for decision making in agriculture," 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), Noida, 2016, pp. 1-6.
doi: 10.1109/IC3I.2016.7917925

[6] Patel, Hetal & Patel, Dharmendra. (2014). A Brief survey of Data Mining Techniques Applied to Agricultural Data. International Journal of Computer Applications. 95. 6-8. 10.5120/16620-6472.

[7] E. Vintrou, D. Ienco, A. Bégué and M. Teisseire, "Data Mining, A Promising Tool for Large-Area Cropland Mapping," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 6, no. 5, pp. 2132-2138, Oct. 2013.

doi: 10.1109/JSTARS.2013.2238507

[8] Bhargavi, P. and M. Tech. “Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils.” (2009).

[9] Nisha and Puneet Jai Kaur, "A survey of clustering techniques and algorithms," *2*015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2015, pp. 304-307.

[10] Veenadhari S, Misra B, Singh CD. Data mining techniques for predicting crop productivity—A review article. In: IJCST. 2011; 2(1).

[11] Majumdar J, Ankalaki S. Comparison of clustering algorithms using quality metrics with invariant features extracted from plant leaves. In: Paper presented at international conference on computational science and engineering. 2016.

[12] Berkhin P. A survey of clustering data mining technique. In: Kogan J, Nicholas C, Teboulle M, editors. Grouping multidimensional data. Berlin: Springer; 2006. p. 25–72.

[13] Han J, Kamber M. Data mining: concepts and techniques. Massachusetts: Morgan Kaufmann Publishers; 2001.

[14] Ramesh D, Vishnu Vardhan B. Data mining techniques and applications to agricultural yield data. In: International
journal of advanced research in computer and communication engineering. 2013; 2(9).

[15] MotiurRahman M, Haq N, Rahman RM. Application of data mining tools for rice yield prediction on clustered
regions of Bangladesh. IEEE. 2014;2014:8–13.

[16] Veenadhari S, Misra B, Singh D. Machine learning approach for forecasting crop yield

based on climatic parameters.
In: Paper presented at international conference on computer communication and informatics (ICCCI-2014), Coimbatore. 2014.

[17] Rahmah N, Sitanggang IS. Determination of optimal epsilon (Eps) value on DBSCAN algorithm to clustering data on peatland hotspots in Sumatra. IOP conference series: earth and environmental. Science. 2016;31:012012.

[18] Ng RT, Han J. CLARANS: A Method for Clustering Objects for Spatial Data Mining. In: IEEE Transactions on Knowledge and Data Engineering. 2002; 14(5)

[19] Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. Wiley. 1990.
doi:10.1002/9780470316801.

[20] Kirkl O, De La Iglesia B. Experimental evaluation of cluster quality measures. 2013. 978-1-4799-1568-2/13. IEEE.

[21] Meila M (2003) Comparing clustering. In: Proceedings of COLT 2003.