**CHAPTER 1**

**INTRODUCTION**

* 1. **MACHINE LEARNING**

Machine learning is a fast-growing trend in the health care industry, thanks to the advent of wearable devices and sensors that can use data to assess a patient's health in real time. The technology can also help medical experts analyze data to identify trends or red flags that may lead to improved diagnoses and treatment.

Diagnosis of the Parkinson disease through machine learning approach provides better understanding from PD dataset in the present decade. Orange v2. 0b and weka v3.4.10 has been used in the present experimentation for the statistical analysis, classification, Evaluation and unsupervised learning methods. Voice dataset for Parkinson disease has been retrieved from UCI Machine learning repository from Center for Machine Learning and Intelligent Systems. The two machine learning models used in this proposed model are:

* **Decision trees.** These models use observations about certain actions and identify an optimal path for arriving at a desired outcome.
* **K-means clustering.** This model groups a specified number of data points into a specific number of groupings based on like characteristics.

**1.2 SYSTEM OVERVIEW**

Parkinson’s disease (PD) is a progressive, neurodegenerative disease that belongs to the group of conditions called motor system disorders. Parkinson’s disease sufferers get worse over time as the normal bodily functions, including breathing, balance, movement, and heart function worsen. Other neurodegenerative disorders include Alzheimer’s disease, Huntington’s disease, and amyotrophic lateral sclerosis or Lou Gehrig’s disease. An estimated seven to 10 million people worldwide are suffering from Parkinson’s disease. Occurrence of Parkinson’s increases with age, but an estimated four percent of people with PD are diagnosed before the age of 50. There is no cure or prevention for PD. However, the disease can be controlled in early stage. Hence data mining techniques can play effective role in early detection and diagnosis.

Data mining techniques in medicine is a research area that combines sophisticated representational and computing techniques with the insights of expert physicians to produce tools for improving healthcare. Data mining is a computational process to find hidden patterns in datasets by building predictive or classification models that can be learnt from past experience and applied to future cases. With the vast amount of medical data available to hospitals, medical centers, and medical research organizations, the field of medicine supported by data mining techniques can increase healthcare quality and can help physicians make decisions about their patients’ care. Parkinson’s disease is a neurological disorder that causes memory loss and dementia. It is mainly observed in elderly individuals over the age of 60 but can also be caused by concussions or traumatic brain injuries. It causes brain cells to die and spread the damage across the brain, in some severe cases rendering an individual unable to perform daily necessary tasks. It is also considered as neuro-degenerative type of dementia. There are many ways to diagnose and detect consisting of MRI scans, MMSE (mini-mental state exams) both expressed in terms of CDR standards. But identifying distinctions between Alzheimer brain and normal brain in elderly individuals over the age of 75 is difficult because they share similar brain patterns and image intensities. PD causes shrinkage in hippocampus and cerebral cortex and it enlarges the ventricles in the brain.

**1.2 SYSTEM SCOPE**

It is planned to develop a machine learning model that caters to the issue of handling the huge voluminous patient data for exhaustive and complete analysis profile. This model will be designed further with the best collection of classifiers for enhancing the accuracy of the classifiers. In addition, this model tries to integrate the real time PD patients’ data with the existing voice data set for better prediction with good sensitivity and specificity. Sensitivity and specificity are important for disease data analysis. The predictive model will use the required symptoms that will be important in identifying the PD and they are appropriately accommodated in the data set for better treatment and care.

**CHAPTER 2**

**LITERATURE SURVEY**

**[1] Xeuyun Sharon Wang,”Infrastructure for a clinical-decision-intelligence system”, IBM Systems Journal Volume 46 Issue 1, January 2007.**

Clinical decision intelligence (CDI) was an emerging area in health care, covering a broad range of subjects, from clinical data integration and data analysis to knowledge management and application development. The goal of CDI systems was to improve health-care quality and reduce costs through the discovery, management, and application of clinical intelligence from heterogeneous and rapidly expanding data sources. These sources included data from clinical practice, nursing, health-care management, health-care administration, and medical research. In this model, the functional requirements and reference architecture for CDI systems and their clinical applications have been discussed. This architecture included an integrated framework for managing the entire CDI process, a standardized enterprise ontology management system, and a clinical knowledge representation platform. The CDI approach had the potential to transform the health-care management process through the integration of business intelligence, business rule management, and business process management in a clinical setting, and to help health-care organizations move closer to an on demand model.

**[2] Jin David Gill A, Magnus Johnson B,” Diagnosing Parkinson by using Artificial Neural Networks and Support Vector Machines”, Global journal of Computer science and Technology 9.4:63-71, 2009.**

Parkinson's Disease (PD) is the second most common neurodegenerative action only surpassed by Alzheimer's Disease (AD). Moreover, it is expected to increase in the next decade with accelerating treatment costs as a consequence. This situation had lead towards the need to develop a Decision Support System for PD. In this model methods based on ANNs and SVMs to aid the specialist in the diagnosis of PD were developed. Data recorded during 195 examinations carried out on 31 patients was used to verify the capacity of the developed system. The results had showed a high accuracy of around 90%.

**[3] Dr.R.Geetha Ramani,”Reference analysis and classification of Parkinson’s disease tele monitoring Data through Data Mining Techniques”, IEEE on smart grid, 2012.**

The developed model places emphasis on classifying the severity of Parkinson’s disease. Symptomatic identification includes loss of specific brain cells and decline in dopamine concentration. Unified Parkinson Disease Rating Scale (UPDRS), captures multiple aspects of Parkinson Disease that includes Mutation, Behavior and mood, Activities of Daily Life (ADL), Motor Examination and complications of therapy. In the Parkinson Disease Tele-Monitoring Dataset, the data consisted of 16 biomedical voice measures with test subject information, motor UPDRS and total UPDRS scores. The main goal of the work was to predict the motor and total UPDRS scores from the voice measures. Predicting the scores was done by extracting useful knowledge and thereby providing the scientific decision- making classification rules necessary for the diagnosis of disease severity. This was done by precisely classifying the given dataset and relegating them into people with high scores and low scores.

**[4] N.Sairam,“New machine-learning algorithms for prediction of Parkinson's disease”, International journal of system science Pages 647-666, 2012.**

This model presented an enhanced prediction accuracy of diagnosis of Parkinson’s disease (PD) to prevent the delay and misdiagnosis of patients using the robust inference system. New-machine learning methods were developed and performance comparison were based on specificity, sensitivity accuracy and other measurable parameters. The robust methods of treating Parkinson’s disease included sparse multinomial logistics regression, rotation forest ensemble with support vector machines and principal components analysis, artificial neural networks, boosting methods.

**[5] Farhad Soleimanian Gharehchopogh,”A Case Study of Parkinson’s disease Diagnosis using Artificial Neural”, IEEE Computer Application (0975-8887) Volume 73 –NO. 19, July 2013.**

Artificial Neural Network based diagnosis of medical diseases had been taken into great consideration. Two types of ANNs were used to classify effective diagnosis of Parkinson’s disease. Multilayer Perceptron with back-propagation learning algorithm and Radial Basis Function (RBF) ANNs were used to differentiate between clinical variables of samples (N= 195) who were suffering from Parkinson's disease and who were not. For this purpose, Parkinson's disease data set was taken from UCI machine learning database was used. Mean squared normalized error function was used to measure the usefulness of our networks during trainings and direct performance calculations. It was observed that MLP is the best classification with 93.22% accuracy for the data set. This technique was used to assist neurologists to make their ultimate decisions without hesitation and more astutely.

**[6] Peyman Mohammadi and Mohammad Masdari,”International Journal in Foundations of computer science and technology Vol 3 No 6”, International Research Journal of Engineering and Technology (IRJET) April-2018.**

In recent years applications of data mining methods have become more popular in many fields of medical diagnosis and evaluations. The data mining methods are appropriate tools for discovering and extracting of available knowledge in medical databases. In this model, they have divided 11 patients clinical variables data with Parkinson’s Disease (PD) to study the disease progression. The data set included 22 properties of 42 people that all of their algorithms were applied to this data set. The decision table stump with 0.7919 correlation coefficients had the lowest accuracy.

**[7] Chen, Hui-Ling, etal,”An efficient diagnosis system for detection of Parkinson’s disease using fuzzy K-nearest neighbor approach”, Volume 40, Issue 1, Pages 263-271 January 2013.**

An effective and efficient diagnosis system using fuzzy K-nearest (FKNN)for Parkinson’s disease(PD) diagnosis had been developed. The developed FKNN-based system was compared with the support vector machines(SVM)based approaches. In order to further improve the diagnosis accuracy for detection of PD, the principle component analysis was employed to construct the most discriminative new feature sets on which the optimal FKNN model was constructed. The effectiveness of the developed model has been rigorously estimated on a PD data set in terms of classification accuracy, sensitivity, specificity. The best classification accuracy obtained by the FKNN-based system using a 10-fold cross validation method could ensure a reliable diagnostic model for detection of PD.

**[8] Patawala Amatulla.H, Bansode Navnath.P, Bhong Yogesh.P, Prof. Zadbuke,”Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms”, I Volume 3 issue 3 September 2013.**

Diagnosis of the Parkinson disease through machine learning approach provides better understanding from PD dataset in the present decade. Orange v2.0b and weka v3.4.10 had been used in the present experimentation for the statistical analysis, classification, Evaluation and unsupervised learning methods. Voice dataset for Parkinson disease had been retrieved from UCI Machine learning repository from Center for Machine Learning and Intelligent Systems. The parallel coordinates showed higher variation in Parkinson disease dataset. SVM had shown good accuracy (88.9%) compared to Majority and k-NN algorithms. Classification algorithm like Random Forest had shown good accuracy (90.26) and Naïve Bayes had shown least accuracy 69.23% Higher number of clusters in healthy dataset in Fo and less number in diseased data has been predicted by Hierarchical clustering and SOM.

**[9] P.Suganya and C.P Sumathi,”Novel Metaheuristic Data Mining Algorithm for the Detection and Classification of Parkinson Disease”, Volume 8, Issue 14,2015.**

The model adopted a novel meta heuristic data mining algorithm for the detection and classification of Parkinson Disease where about 195 instances were selected for the investigation. In the initial phase the data underwent five phases, which included training dataset, data pre-process, feature selection, classification and evaluation. This included the confusion matrix, precision, recall and error rate. The confusion matrix was evaluated with various attributes like Specificity, Sensitivity, Accuracy and Positive and Negative predictive values. Findings: The model also performed a comparative study on five classification algorithms i.e. ABO, SCFW with KELM, FCM, ACO and PSO algorithms. This comparison results from confusion matrix of the selected algorithms which supported in identifying the specificity, sensitivity and accuracy of performance measures index showed that ABO algorithm is found to have best specificity, sensitivity and accuracy compared to all other algorithms, i.e. SCFW with KELM, FCM, PSO and ACO. In addition, the classifiers comparison results of the selected algorithms indicated that ‘ABO’ had the highest accuracy. This model was intended to estimate the efficiency and efficacy of the selected algorithm to best detect the Parkinson Dataset using various classifiers, as early detection of any kind of disease is an essential factor.

**[10] Musaed Alhussein,” Monitoring Parkinson’s Disease in Smart Cities “Volume 5 Page(s): 19835 - 19841, September 2017.**

Parkinson's Disease (PD) is one of the most severe neurological diseases prevalent in the world. A neurodegenerative disease, it impairs the body's balance, damages motor skills, and leads to disorder in speech production. These problems also affect decision-making processes and the expression of emotions. In this model, a PD monitoring framework for use in smart cities was proposed. Using this framework, city residents will have their health constantly monitored and get feedback on their PD situation. Early PD symptoms can, therefore, be detected and the proper medication provided. In this framework, they used speech signals from clients captured from various sensors and transmitted to the cloud for processing. In the cloud, decisions were made using a support vector machine-based classifier. Decisions, along with the signal features, are sent to registered doctors, who then prescribe certain medications to the client.

**CHAPTER 3**

**SYSTEM ANALYSIS**

**3.1 EXISTING SYSTEM**

PD is a neurological disease. It affects certain brain cells that help in controlling the movement and coordination. Dopamine is a hormone and neurotransmitter, a chemical that is generated by brain cell. It is used to send signals to other brain cells to control the muscle activity. PD causes, degeneration of dopamine in the brain cell which is unable to control the movement and activity of muscles. The symptoms may vary from person to person. In existing system, PD is detected at the secondary stage only (Dopamine deficiency) which leads to medical challenges. Also doctor has to manually examine and suggest medical diagnosis in which the symptoms might vary from person to person so suggesting medicine is also a challenge.

**3.1.1 Issues of the existing system**

* Possibility of Uncertainty: Due to the presence of uncertain errors, no model can predict hundred percent accurate results in terms of detecting all the stages of Parkinson’s disease.
* Due to the exponential growth of data, maintaining the data warehouse is difficult.
* Among various stages of Parkinson’s disease only primary and secondary stages are predicted in existing system.
* The main drawback of the used algorithms in existing systems are privacy and security which arises from patient’s data.

**3.2 PROPOSED SYSTEM**

In healthcare industries, the demand for maintaining large amount of patients’ data is steadily growing due to rising population which has resulted in the increase of details about clinical and laboratory tests, imaging, prescription and medication.

The proposed predictive analytics framework is a combination of K-meansclustering and Decision Tree which is used to gain insights from patients. By using machine learning techniques, the problem can be solved with minimal error rate. Parkinson’s disease voice dataset from UCI Machine learning repository is used as input. Thus our experimental results will show early detection of disease will facilitate clinical monitoring of elderly people and increase the chances of their life span and improved lifestyle to lead peaceful life.

**3.2.1 ADVANTAGES**

* The proposed system has a greater accuracy than the other existing applications.
* The proposed system can predict all the stages of Parkinson’s disease including Hereditary-Genetic Origin and Multiple system atrophy- Degeneration of parts other than mid brain.
* The proposed system identifies the people affected by PD more accurately.
* The performance is based on accuracy, sensitivity and specificity.
* Only authorized people can have access to the application.

**3.3 REQUIREMENTS SPECIFICATIONS**

**3.3.1 Hardware Specifications**

* Hard Disk: 40GB and above
* RAM: 512MB and above
* Processor: Pentium IV and above

**3.3.2 Software Specifications**

## Windows OS

## Shiny application

## R Studio

**3.4 LANGUAGE SPECIFICATION**

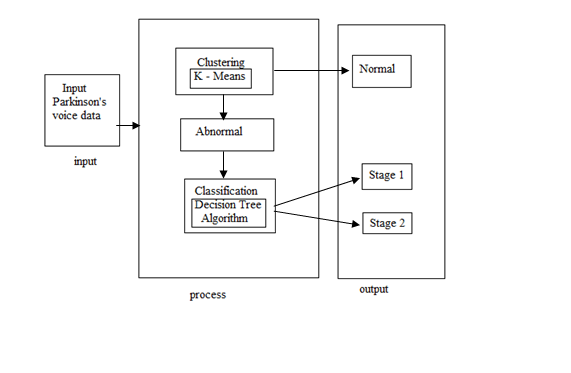
**3.4.1 R PROGRAMMING**

R is a programming language and software environment for statistical analysis, graphics representation and reporting. A huge amounts of multidimensional data have been collected in various fields such as marketing, bio-medical and geo-spatial fields. Mining knowledge from these big data becomes a highly demanding field. However, it far exceeded human’s ability to analyze these huge data. Unsupervised Machine Learning or clustering is one of the important data mining methods for discovering knowledge in multidimensional data.

* RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.
* RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) or in a browser connected to RStudio Server or RStudio Server Pro (Debian/Ubuntu, RedHat/CentOS, and SUSE Linux).
* RStudio is written in the C++ programming language and uses the Qt framework for its graphical user interface.
* R is a powerful language and environment for statistical computing and graphics. It is a public domain (a so called “GNU”) project which is similar to the commercial S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a diﬀerent implementation of S, and is much used in as an educational language and research tool. The main advantages of R are the fact that R is freeware and that there is a lot of help available online. It is quite similar to other programming packages such as MatLab (not freeware), but more user-friendly than programming languages such as C++ or Fortran. You can use R as it is, but for educational purposes we prefer to use R in combination with the RStudio interface (also freeware), which has an organized layout and several extra options. This document contains explanations, examples and exercises, which can also be understood (hopefully) by people without any programming experience. Going through all text and exercises takes about 1 or 2 hours. Examples of frequently used commands and error messages are listed on the last two pages of this document and can be used as a reference while programming.

**CHAPTER 4**

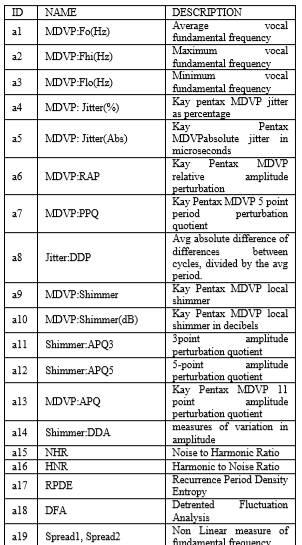
**4.1 SYSTEM ARCHITECTURE**



**Figure 1.1 Architecture of proposed system**

In the Figure1.1, the input Parkinson’s voice data fetched from UCI machine learning repository. The gathered information may not be in a particular format. It can produce petabytes of data. This kind of mixed data of big data need to be converted into structured form which will allow to retrieve actionable insights without missing the required information. The process consists of two major steps. They are Clustering and Classification. The k-Means clustering is used to segment the datasets into training and testing datasets. In this proposed framework the training datasets are clustered into Normal and Abnormal categories based on threshold values of Detrented Fluctuation Analysis feature of datasets. The Normal dataset is the dataset of people who are not affected by PD. Whereas Abnormal ones are for people who suffer due to PD.

The K-means clustering performs the clustering on the basis of the Euclidean distance formula. This segregates the patients into Normal and Abnormal. The next step of the process is to classify the patients who fall in the abnormal category. The people who are abnormal fall under the Stage 1 or Stage 2. These stages are predicted with the help of classification using Decision Tree. Decision tree helps in classifying the dataset into various parts with some constraints given to it. The constraint here is that the person having the Detrented Fluctuation Analysis below 0.77 fall into the Stage 1 i.e., Hereditary-Genetic origin and the person having the Detrented Fluctuation Analysis above 0.77 fall into the Stage 2 Multiple system atrophy.

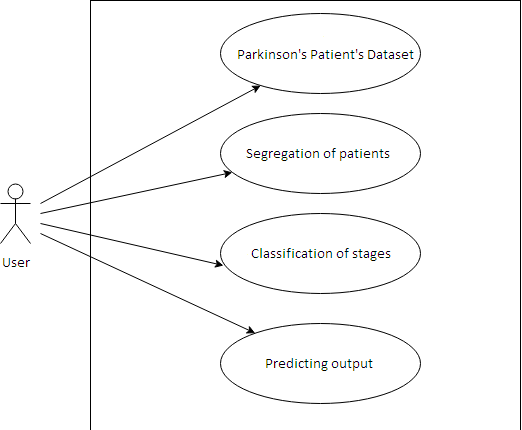
****

**Parkinson Dataset**

**Obtained From UCI**

**Figure 4.1.1 Parkinson’s disease input data set**

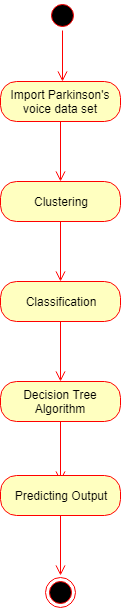
**4.2 USE CASE DIAGRAM**



**Figure 4.2 Use case diagram**

The use case diagram is a representation of a user’s interaction with the system that shows the relationship between the user and the different use cases in which the user is involved.

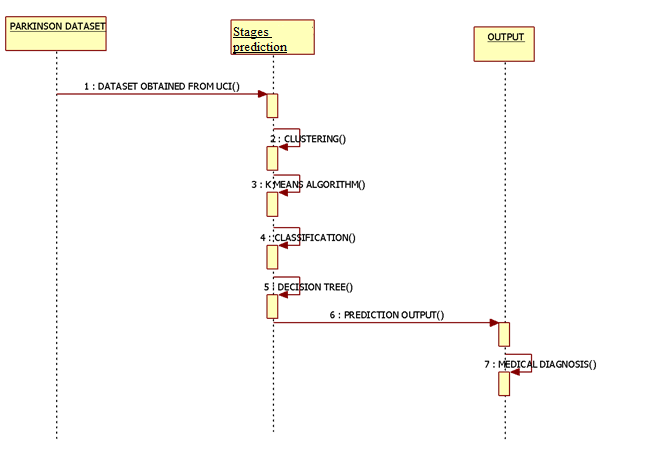
**4.3 ACTIVITY DIAGRAM**

****

**Figure 4.3 Activity diagram**

In this Figure 4.3 the activity carried out is that Dataset of the user is first imported and then the clustering is done on it. The result produced by the k-means clustering are the two clusters, Normal and the Abnormal which segregates the people. After which the dataset of the normal people is given to the decision tree. This classification results into the two stages of the Parkinson’s disease.

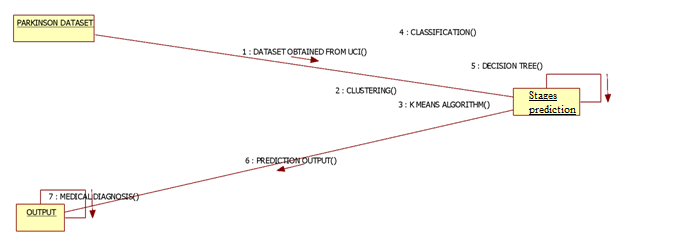
**4.4 SEQUENCE DIAGRAM**

****

**Figure 4.4 Sequence diagram**

In the Figure 4.4 the dataset is first imported after which the stages of the Parkinson’s disease is detected. The stages are predicted from the results of the k-means clustering. Clustering gives the two clusters as Normal and Abnormal. The Abnormal people’s dataset is then given to the decision tree which gives the final output for medical diagnosis.

**4.5. COLLABORATION**

****

**Figure 4.5 COLLABORATION**

In the Figure 4.5 the Parkinson’s Disease Dataset is given to the k-means clustering algorithm which clusters the given dataset into two which is normal and abnormal. The stages of PD are predicted by applying the decision tree.

**CHAPTER 5**

**MODULE DESCRIPTION**

#### 5.1 MODULES:

The modules used in the system are:

1. Parkinson’s Patient’s dataset.
2. Segregation of patients.
3. Classification of stages.
4. Predicting output.

**5.1.1 Parkinson’s Patient’s dataset**

Organize your data in an Excel worksheet, such that the first row (Row 1) contains the column names and each subsequent row contains all the necessary information for each data point in the experiment [i.e. classification levels and measurement(s)]. Save your worksheet as a comma separated values (.csv) file type Save your Excel spreadsheet as normal (default file type: Excel Workbook); this will be your master file that you can always return to in order to modify things, add new data, etc. Next, to create a version of your data to input into R, click "Save As…" A window will appear in which you can specify your desired filename as well as your desired file type. In Windows, for example, when you click "Save As…" you will see an option for "Choose other formats." Click that option, and then you will be presented with a pull-down menu of available file types. Select "Comma Separated Values (.csv)" When you click to save the worksheet as a comma separated file (.csv), Excel will present you with a couple of warnings. One warns you that the .csv format cannot accommodate multiple worksheets. That's fine; just click OK/Continue to save the active worksheet. The second warning tells you that the .csv format cannot accommodate some of the fancy features of an .xls or .xlsx file. That's fine, just click OK/Continue. When you're done, you will see that a new file has been created (FILENAME.csv). If you were to open this file in a text editor, you would see that it looks like this:

Block, Genotype, Location, Replication, Subsample, Yield, Quality, Disease

1, A, 1, 1, 1, 23, 143, 2

1, A, 1, 1, 2, 25, 135, 2

1, A, 1, 2, 1, 24, 152, 3

1, A, 1, 2, 2, 27, 160, 2

1, B, 1, 1, 1, 32, 123, 1

1, B,1,1,2,34,112,1

1, B,1,2,1,32,134,1

1, B,1,2,2,38,118,1 And so on…

Refer to Lab 1 for instructions on how to import data into RStudio, using the "Import data..." feature. Remember, to learn R, you need to play with R.

**5.1.2 SEGREGATION OF PATIENT**

K Means Clustering is an unsupervised learning algorithm that tries to cluster data based on their similarity. In k means clustering, we have to specify the number of clusters we want the data to be grouped into. The algorithm randomly assigns each observation to a cluster, and finds the centroid of each cluster. Then, the algorithm iterates through two steps:

* Reassign data points to the cluster whose centroid is closest.
* Calculate new centroid of each cluster.

These two steps are repeated till the within cluster variation cannot be reduced any further. The within cluster variation is calculated as the sum of the Euclidean distance between the data points and their respective cluster centroids.

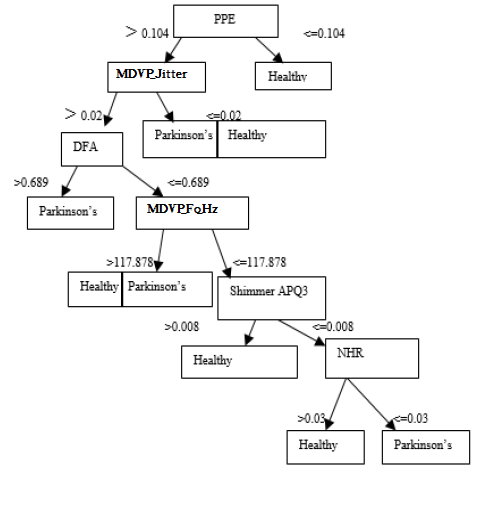
**5.1.3 CLASSIFICATION OF STAGES**

**5.1.3.1 DECISION TREE**

A decision tree is also called a prediction tree. A decision tree uses a structure to specify sequences of decisions and consequences. Given input X= {X1, X2, .., Xn}, the goal is to predict a response or output variable Y. Each member of the set {X1, X2, …, Xn} is called an input variable.

The prediction can be achieved by constructing a decision tree with test points and branches. A decision is made at each test point, to pick a specific branch and traverse down the tree Decision trees can be used in a variety of disciplines, such as: On the basis of individual characteristics deciding whether or not to offer a loan to an individual, predicting the rate of return of various investment strategies, predict whether or not send a direct mail to a potential customer, etc.

A decision tree consists of nodes, and thus form a rooted tree, this means that it is a directed tree with a node called root. There are no incoming edges on root node, all other nodes in a decision tree have exactly one incoming edge. An internal node is a node with an incoming edge and outgoing edges, internal node is also known as test node. Nodes with no outgoing edges are known as leaves or terminal nodes.



Generally model performance is measures based on accuracy, specificity and accuracy. These results are shown in confusion matrix. The accuracy achieved to predict the disease after processing the dataset in all three algorithms i.e., ANN, Decision tree and Naïve Bayes algorithm. ANN provides good accuracy in PD and Naïve Bayes provides best accuracy in tumors. Naïve Bayes classifies the result based on the confusion matrix. Even though ANN gives good accuracy in PD it takes more time. Hence we have chosen Naïve Bayes algorithm.

**5.1.4 PREDICTING OUTPUT**

Each dataset is divided into training and testing sets in the ratio of 70:30. The task of dividing the dataset is to test the effectiveness of proposed model. The dataset for testing is given as input for the k-means algorithm. The output of k-means are the two clusters which are used to identify the normal and abnormal patients. The abnormal patient’s dataset is further classified using the decision tree. Decision tree has the lowest error probability. Decision tree distinguishes abnormal result obtained from clustering into stage 1 and stage 2. Hence there are some overall error probabilities. The vital and critical part of the proposed model is built on complete case sample as it has to be deployed in the health care domain which only provides estimated target score for individuals with complete data. This machine learning approach is to build a better data model.

Formulae used

Sensitivity=TP/(TP+FN) \*100

Specificity=TN/(TN+FP) \*100

Accuracy=(TP+TN)/(TP+TN+FP+FN) \*100

Where,

TP – True Positive,

FP – False Positive,

TN – True Negative,

FN – False Negative

**5.4 METHODOLOGY**

The voice dataset for Parkinson disease has been retrieved from UCI Machine learning repository from Centre for Machine Learning and Intelligent Systems [11]. The dataset range of biomedical voice measurements from 31 people, where 23 people are showing Parkinson's disease. The class column represents "status” which is set to 0 for healthy and 1 for PD. The data is in ASCII CSV format. The following are the Attribute Information:

name,MDVP:Fo(Hz),MDVP:Fhi(Hz),MDVP:Flo(Hz),MDVP:Jitter(%),MDVP:Jitter(Abs),MDVP:RAP,MDVP:PPQ,Jitter:DDP,MDVP:Shimmer,MDVP:Shimmer,Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shimmer:DDA,NHR,HNR,status, RPDE,DFA,spread1,spread2,D2,PPE Matrix column entries (attributes) show the following description:

Name - ASCII subject name and recording number

MDVP : Fo(Hz) - Average vocal fundamental frequency

MDVP : Fhi(Hz) - Maximum vocal fundamental frequency

MDVP : Flo(Hz) - Minimum vocal fundamental frequency

MDVP : Jitter(%),MDVP:Jitter(Abs),MDVP:RAP,MD

VP : PPQ,Jitter : DDP - Several measures of variation in fundamental frequency

**CHAPTER 6**

**CONCLUSION AND FUTURE ENHANCEMENT**

**6.1 CONCLUSION**

Parkinson’s disease diagnosis has become a vital one in medical field because they are caused by abnormal and uncontrolled growth of cells inside the brain. The aim of this project is to recognize decision tree classifiers performance and examine the effectiveness of attribute selection, discretization, and test mode on the selected classifier performance when implemented on the PD dataset.

**6.2 FUTURE ENHANCEMENT**

As future work, we are trying to include the image dataset along with the voice dataset to improve the accuracy of the result obtained. Future improvement can be with the curvatures in the brain which will also use 3D images for prediction of Parkinson’s disease.

**APPENDIX 1**

**SAMPLE CODING**

data\_s<-Datasets

data\_s$HNR

data\_s$status

data\_s$name<-NULL

omit<-na.omit(data\_s)

New\_dataset1<-data.frame(omit)

New\_dataset1$DFA

New\_dataset1$parameter[New\_dataset1$DFA<=0.77]<-"Stage1"

New\_dataset1$parameter[New\_dataset1$DFA>=0.77]<-"Stage2"

New\_dataset1$parameter

Stage1<-subset(New\_dataset1,parameter=="Stage1")

Stage2<-subset(New\_dataset1,parameter=="Stage2")

print(New\_dataset1)

#kmeans#

New\_dataset\_s<- New\_dataset1

New\_dataset\_s$parameter<-NULL

f<-New\_dataset1

result<-data.frame(f$HNR,f$spread2)

print(result)

result->k

(s<- kmeans(k,6))

s$cluster

plot(k[c("f.HNR","f.spread2")],

col = s$cluster)

points(s$centers[,c("f.HNR","f.spread2")],

col = 1:6,pch = 8, cex=2)

s$betweenss

#naive Bayes#

library(e1071)

library(caTools)

library(caret)

train <- New\_dataset1[1:16,]

test <-New\_dataset1[17:24,]

train$parameter<- as.factor(unlist(train$parameter))

test$parameter <- as.factor(unlist(test$parameter ))

train\_class <- train[,-24]

model <- naiveBayes(train\_class,train$parameter)

summary(model)

pred <- predict(model,test[,-24] )

summary(pred)

confusionMatrix(pred, test$parameter)

confusion\_f<-confusionMatrix(pred, test$parameter)

plot(pred)

#support vector machine##

library(caTools)

library(e1071)

library(caret)

train1 <- New\_dataset1[1:15,]

test1 <-New\_dataset1[16:24,]

train1$parameter<- as.factor(unlist(train1$parameter))

test1$parameter<- as.factor(unlist(test1$parameter))

train\_class1 <- train1$parameter

New\_dataset1$parameter<-as.factor(New\_dataset1$parameter)

data<- function(New\_dataset1){

New\_dataset1<-New\_dataset1

return(New\_dataset1)}

data(New\_dataset1)

model1<-svm(New\_dataset1$parameter~., data = New\_dataset1)

summary(model1)

pred1<-predict(model1,New\_dataset1)

summary(pred1)

plot(pred1)

confusionMatrix(pred1, New\_dataset1$parameter)

##decision tree#

library(rpart)

library(rpart.plot)

h<-New\_dataset1

New\_set<- h[ sample( nrow( h ) ), ]

train\_tree<-h[1:14,]

test\_tree<-h[15:24,]

train\_tree$parameter<- as.factor(unlist(train\_tree$parameter))

test\_tree$parameter<- as.factor(unlist(test\_tree$parameter))

h$parameter<-as.factor(h$parameter)

model\_fit <- rpart(parameter~., data = h)

print(model\_fit)

rpart.plot(model\_fit, extra = 106)

predict\_seen <-predict(model\_fit, h, type = 'class')

summary(predict\_seen)

table\_pred <- table(h$parameter, predict\_seen)

table\_pred

accuracy\_Test <- sum(diag(table\_pred)) / sum(table\_pred)

print(accuracy\_Test)

**APPENDIX 2**

**SCREENSHOTS**

**REFERENCES**

[1] Michael Minelli, Michele Chambers, Ambiga Dhiraj. Big Data Big analytics: emerging business intelligence and analytics trend for today’s businesses, Feb 2013.

[2] Parkinson’s disease, challenges, progress and promise:National Institute Of Neurological Disorder And Stroke, National Institute of Health, November 2004.

[3] Jiawai Han and Micheline Kamber. Data Mining Concepts and Techniques: second edition.

[4] Pravin Kumar and Vijay Singh Rathore. Efficient Capabilities of Processing of Big Data using HadoopMapReduce: International Journal of Advanced Research

in Computer and Communication Engineering June 2014;Vol: 3, Issue 6.

[5] Wei Dai and Wei Ji. A MapReduce Implementation of C4.5 Decision Tree Algorithm: International Journal of Database Theory and Application; Vol: 7, No.1 (2014),pp.49-60.

[6] TawseefAyoub Shaikh. A Prototype of Parkinson’s and Primary Tumor Diseases Prediction Using Data Mining Techniques: International Journal of Engineering Science Invention April 2014, Vol: 3 Issue 4, pp. 23-28.

[7] Anil Radhakrishnan and kirankalmadi.Big Data Medical Engine in the cloud (BDMEiC): your new Health Doctor: vol: 11 Nov 1, 2013.

[8] DivyaTomar and Sonali Agarwal. A survey on Data Mining approaches for Healthcare: International Journal of Bio-Science and Bio-Technology 2013,vol: 5.

[9] GeethaRamani R, Sivagami G, ShomanaGracia Jacob.Feature Relevance Analysis and Classification of Parkinson’s disease Tele-Monitoring data Through Data Mining: International Journal of Advanced Research in Computer Science and Software Engineering March 2012,Vol: 2.

[10] Tarigoppula et al. Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms: IJEIT September 2013, Vol: 3, Issue 3.

[11] Dr.Hariganesh and Gracyannamary. Comparative study of Data Mining Approaches for Parkinson’s Disease: IJARCET September 2014, Vol: 3, Issue 9.

[12] Chandrasekhar Azad, Sanjay Jain, Vijay Kumar Jha. Design and Analysis of Data Mining Based Prediction Model for Parkinson’s disease: IJCSE.

[13] Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig. Suitability of dysphonia measurements for tele monitoring of Parkinson's disease',IEEE Transactions on Biomedical Engineering 2008.

[14] Hoglinger GU, Rizk P, Muriel MP, Duyckaerts C, Oertel WH, Caille I, et al. Dopamine depletion impairs precursor cell proliferation in Parkinson disease: National Neuroscience 2004,7:726–35.

[15] David gila, magnus Johnson b. Diagnosing Parkinson by using Artificial Neural Networks and Support Vector Machines: Global Journal of Computer Science and Tchnology, pp. 63-71.