**LIVER DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS**

A main project submitted in partial fulfillment of the requirements of the degree of

**MASTER OF COMPUTER APPLICATION**

**In**

**COMPUTER SCIENCE**

to the

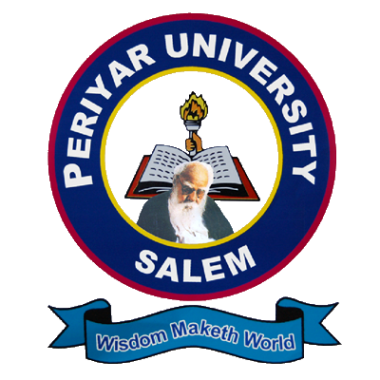
**Periyar University, Salem**

by

**RABIYA A**

**(Reg. No: P17CSC1009)**

s



**DEPARTMENT OF COMPUTER SCIENCE**

**PERIYAR UNIVERSITY**

**SALEM - 636 011**

Dr. Chandrasekar,Phd,

Professor,

Department of Computer Science,

Periyar University,

Salem-636011.

CERTIFICATE

This is to certify that the mini project report entitled as “**Liver Disease Prediction using Machine Learning Algorithms**” submitted in partial fulfillment of the requirements for the award of the degree of Master of Science in Computer Science to Periyar University, Salem is a bonafide work carried out by RABIYA.A **(Reg.No: P17CSC1009)** under my supervision and guidance and that no part of the project has been submitted for the award of any degree or diploma.

Place: Salem-11

Date:

Signature of the Guide Head of the Department

Submitted of the Viva-Voice Examination held on \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

Internal Examiner External Examiner

**DECLARATION**

Iherebydeclarethatthe miniprojectworkentitled“Liver Disease Prediction Using Machine Learning Algorithms”submittedto PeriyarUniversityinpartialfulfillmentof therequirementfortheawardoftheDegreeofMasterofScienceinComputer Scienceistherecordworkcarriedoutbyme,underthesupervisionDr.K.Savitha,AssistantProfessor,Departmentof Computer Science, Periyar University,Salem, duringtheyear2019-2020.

Tothebest ofmyknowledge,theworkreportedhereisnotapartofanyotherworkonthebasis of whichadegreeorawardwasconferredonanearliertooneoranyothercandidate.

Place: Salem -11 Signature of the Student

Date: [RABIYA A.]

**ACKNOWLEDGEMENT**

First, I would like to thank The Almighty for providing me with everything that I required in completing this project.

I extend my heart-felt gratitude to Prof.**Dr. K. Thangavel,**Head of Department, Department of Computer Science, PeriyarUniversity, for his timely triggers.

I acknowledge with thanks the kind patronage, loving inspiration and timelyguidance which I have received from my guide **Dr.K.SAVITHA**, Assistant Professor, Department of Computer Science, PeriyarUniversity.

I extend my thanks to my parents and well-wisher for their constant support and encouragement.

CONTENTS

|  |  |  |
| --- | --- | --- |
| S.NO | TITLE | PAGE NO |
| 1 | **INTRODUCTION**  1.1 Objective  1.2 Scope  1.3 Machine Learning |  |
| 2 | **SYSTEM CONFIGURATION**  2.1 Hardware Specification  2.2 Software Specification |  |
| 3 | **SYSTEM STUDY**  3.1 Existing System  3.2 Proposed System |  |
| 4 | **REVIEW OF LITERATURE** |  |
| 5 | **METHODOLOGY**  5.1 Data Pre-processing |  |
| 6 | **MATERIALS AND METHODS**  6.1 Sampling  6.2 Model Development |  |
| 7 | **SYSTEM DESIGN**  7.1 Input Design |  |
| 8 | **EXPERIMENTAL**  8.1 Screen Shot  8.2 Accuracy Calculation  8.3 Results |  |
| 9 | **APPENDIX**  9.1 Source Code |  |
| 10 | **CONCLUSION** |  |
| 11 | **REFERENCES** |  |

**ABSTRACT**

Data Mining is one of the most critical aspects of automated disease diagnosis and disease prediction. It involves data mining algorithms and techniques to analyze medical data. In recent years, liver disorders have excessively increased

and liver diseases are becoming one of the most fatal diseases in several countries. In this thesis, liver patient datasets are investigate for building classification models in order to predict liver disease. This is implemented a feature model

construction and comparative analysis for improving prediction accuracy of Indian liver patients in three phases. In first phase, min max normalization algorithm is applied on the original liver patient datasets collected from UCI repository. In

liver dataset prediction second phase, by the use of PSO feature selection, subset (data) of liver patient dataset from whole normalized liver patient datasets is obtained which comprises only significant attributes. Third phase, classification algorithms are applied on the data set. In the fourth phase, the accuracy will be calculated Finally, the evaluation is done based on accuracy values.

The main objective of this project is to use classification algorithms to identify the liver patients from healthy individuals. This project also aims to compare the classification algorithms based on their performance factors.

1. **INTRODUCTION**

Problems with liver patients are not easily discovered in an early stage as it will be functioning normally even when it is partially damaged. An early diagnosis of liver problems will increase patient’s survival rate. Liver failures are at high rate of risk among Indians. It is expected that by 2025 India may become the World Capital for Liver Diseases. The widespread occurrence of liver infection in India is contributed due to deskbound lifestyle, increased alcohol consumption and smoking. There are about 100 types of liver infections. Therefore, developing a machine that will enhance in the diagnosis of the disease will be of a great advantage in the medical field. These systems will help the physicians in making accurate decisions on patients and also with the help of Automatic classification tools for liver diseases,one can reduce the patient queue at the liver experts such as endocrinologists. The main objective of this project is to use classification algorithms to identify the liver patients from healthy individuals. In this study, classification algorithms, Support Vector Machines (SVM), Random forest, Logistic regression,KNN have been considered for comparing their performance based on the liver patient data.The dataset used is The Indian Liver Patient Dataset (ILPD) which was selected from UCI Machine learning repository for this study. It is a sample of the entire Indian population collected from Andhra Pradesh region and comprises of 583 patient data.

**1.1 objective**

The number and size of medical databases are increasing rapidly but most of these data are not analyzed for finding the valuable and hidden knowledge. Advanced data mining techniques can be used to discover hidden patterns and relationships. Models developed from these techniques are useful for medical practitioners to make right decision.Data mining techniques also ensemble of these methodsfor the diagnosis of this leading liver disease.

**1.2 scope**

The main of the liver disease data using data mining classification algorithms using KNN and SVM to analyze the accuracy of both models. Then to compare the models which model produce the high accuracy of performance.

**1.3 Machine Learning**

Machine learning is an application of artificial Intelligent (AT) that provides system the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

Machine learning is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence.

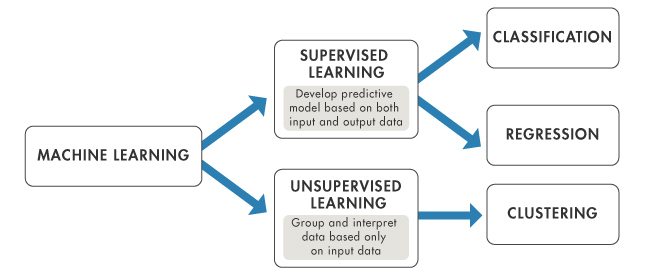
**Some machine learning methods**,

Machine learning algorithms are often categorized as supervised or unsupervised.

**Supervised machinelearning** algorithms can apply what has been learned in the part to new data using labeled examples to predict future events. Starting from the analysis of known training dataset, the learning algorithm produces an inferred function to make predictions about the outputvalues.

**Unsupervised machine learning**algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how the systems can infer a function to describe a hidden structure from unlabeled data.

**Semi-supervised machine learningalgorithms fall somewhere in between supervised and unsupervised learning.Since they use both labeled and unlabeled data for training –typically a small amount of labeled data and a large amount of unlabeled data.**



**Figure 1.2: Machine Learning Methods**

**Regression**:

It predicts continuous valued output. The Regression analysis is the statistical model which is used to predict the numeric data instead of labels. It can also identify the distribution trends based on the available data or historic data them.

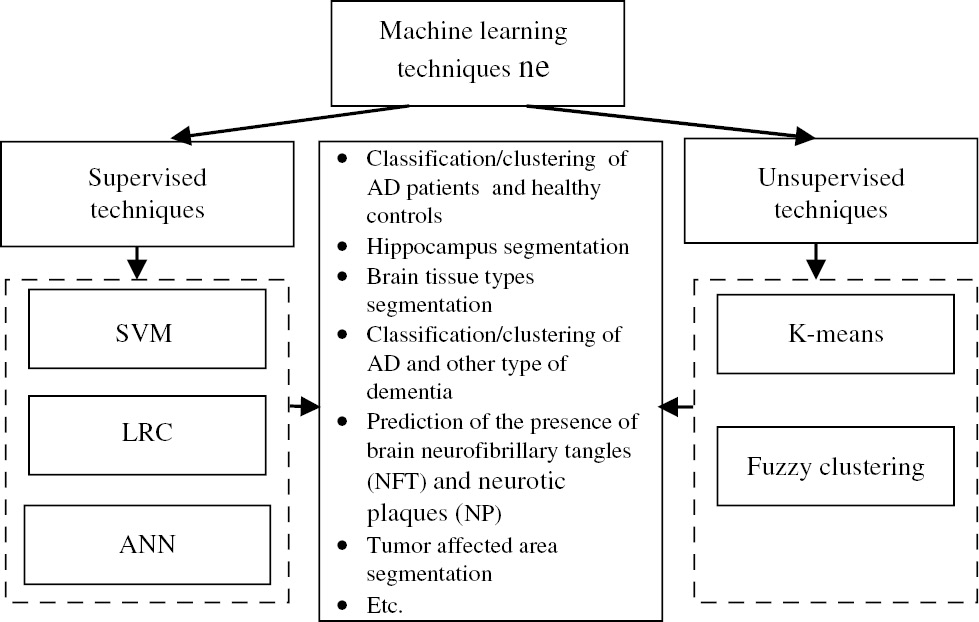
**Classification**:

It predicts discrete number of values. In classification the data is categorized under different labels according to some parameters and then the labels are predicted for the data. Classifying emails as either spam or not spam is example of classification problem.

**Clustering**:

Clustering is the task of partitioning the dataset into groups, called clusters. The goal is to split up the data in such a way that points within single cluster are very similar and points in different clusters are different. It determines grouping among unlabeled data.

**Machine Learning Techniques are,**



**Figure 1.3: Machine Learning Techniques**

**2. SYSTEM CONFIGURATION**

2.1   HARDWARE SPECIFICATION**:**

**Minimum RAM : 2.00 GB**

**Hard Disk : 320 GB**

**Processor : Intel, X86 Compatible processor with 1.7 GHz**

**Clock speed.**

**2.2   SOFTWARE SPECIFICATION:**

**Language : R programming**

**Operating System : Windows8**

**System type: 32 bit**

**Software tool: R studio**

**3. SYSTEM STUDY**

**3.1 EXISTING SYSTEM**

The data mining techniques for predictive data mining task includes, KNN, Random forest, Logistic Regression, SVM .These methods are used for generating knowledge to make it useful for decision making. Each method will produce different result to classify the available variable in dataset.

**3.2 PROPOSED SYSTEM**

The proposed system study focuses on predicting liver disease dataset with different algorithms such as Support Vector Machine, Random forest, Logistic Regression,KNN. The data is collected from UCI Repository .The dataset consist of 583 instances,11 attributes and 2 classes.

# 4. REVIEW OF LITERATURE

**Paul Mangiameli** et al. Proposed a model selection affects the decision support systems accurately. In their model selection, how to affects the accuracy of decision support system hydrides by single model and ensembles. They proposed single model is not more accurate than ensembles. Ahmed M. Hashem proposed to predict Liver Cirrhosis or fibrosis single stage classification model and multistage classification model. In their model based on K-Nearest Neighborhood clustering and Logistic Regression.

8

**Bendi Venkata Ramana** et al. Proposed the different types of liver datasets that is AP liver dataset and UCLA dataset and then he evaluated the performance of the classification techniques from precision, accuracy, specificity and sensitivity. The author said, AP liver dataset is better than the UCLA liver dataset. Using classification algorithm, they are support vector machine, Random forest, Logistic Regression and KNN classifier.

**Bendi Venkata Ramana**et al. Proposed a Modified Rotation Forest algorithm to calculate the accuracy of the liver classification techniques in UCIliver dataset using the combo of feature selection technique and selected classification technique algorithm.

**Ratnamala Kiruba.H**et al. Proposed intelligent agent based system to hike a precise and accurate of diagnosis system. Random forest algorithm are used to predict. Two different types of liver disease disorder dataset are combined and predict the accuracy of the disease. And then conclude these both algorithms gives very good accuracy for diagnosing liver disease disorder.

**CK Ghosh** et al., In his study, the liver abscess was the commonest cause of

hepatomegaly and it was due to amoebiasis, followed by fatty liver, congestive cardic failure, hepatocellular carcinoma, and viral hepatitis seen only in few patients. The most common disease was hepatic steatosis, followed by cirrhosis, portal traditis and chronic hepatitis.

**Rosalina** et al. Predicted a hepatitis prognosis disease using Support Vector machine (SVM) and Wrapper Method. Before classification process they used wrapper methods to remove the noise features. Firstly SVM carried out feature selection to get better accuracy.

5. METHODOLOGY

**5.1 DATA PRE-PROCESSING:**

**Missing data**:

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data pre- processing is a proven method of resolving such issues. Data pre -processing prepares raw data for further processing.

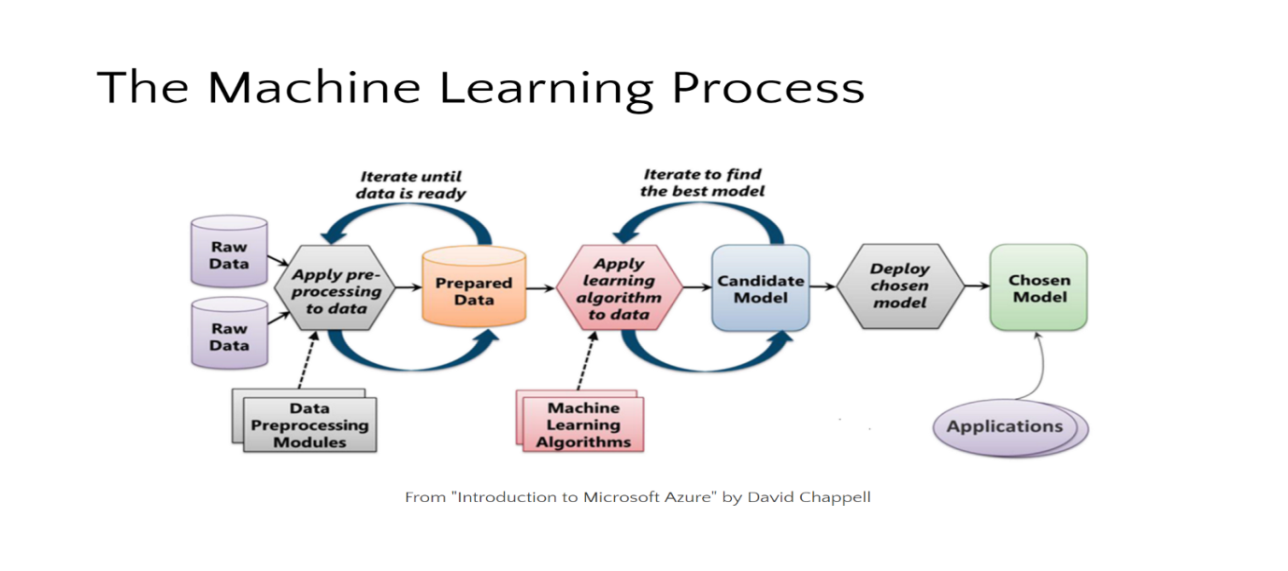
Steps are,

**Data Cleaning**: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.

**Data Transformation**: Data is normalized, aggregated and generalized.

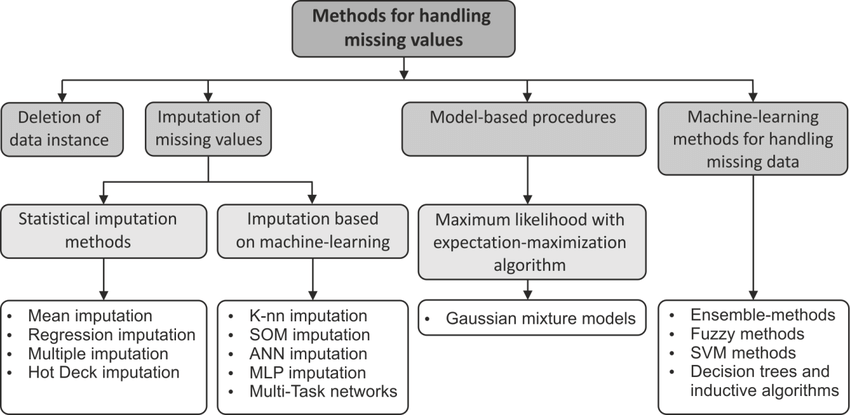
**Data Reduction**: This step aims to present a reduced representation of the data in a data warehouse.

**Data Discretization**: Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

****

**Figure 5.1: Process of machine learning**

**Methods for handling missing values**:



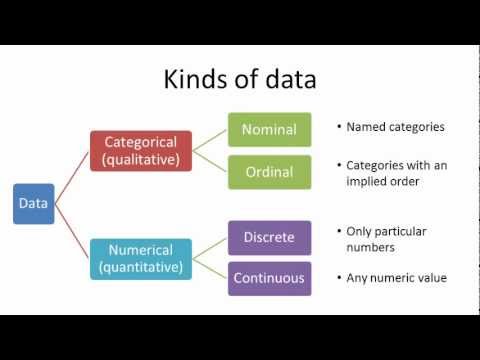
**Figure 5.2: Methods for handling missing values**

**Replace the missing values with mean imputation:**

In our dataset, the missing values occur in attribute name of radius. In this method missing values are replaced by using mean.

**Categorical variable to numerical data:**

In the liver disease dataset one of the attribute names is gender. It contains categorical variable. Then this replaces data from categorical to numerical data.



**Figure 5.3: Kinds of data**

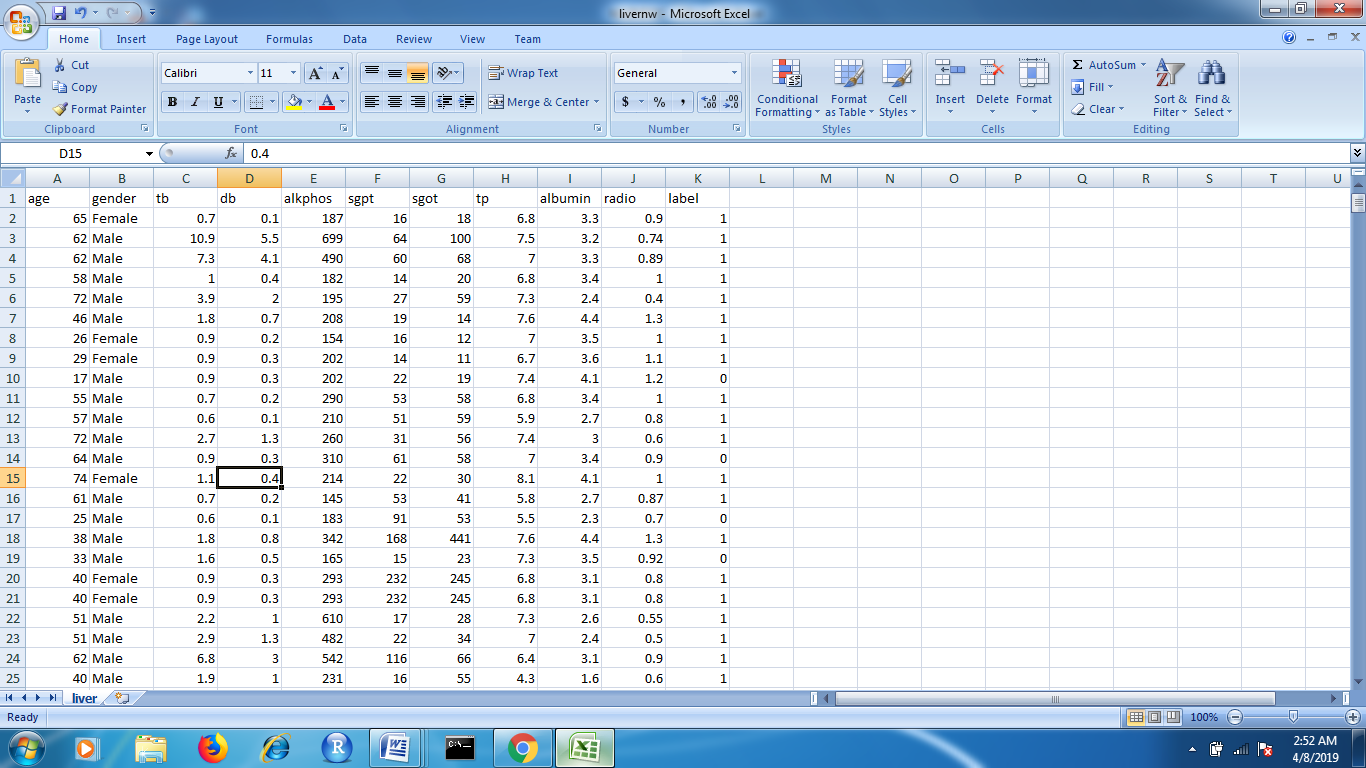
## 6. MATERIALS AND METHODS

**6.1SAMPLING**

The Data Sets are used to evaluate the performance of classification techniques using liver disease. The classification techniques are KNN and Random forest. The datasets are publicly available from the UCI repository.

### 6.1.1 DATASET DESCRIPTION

This data set contains 416 liver patient records and 167non liver patient records. The data set was collected from north east of Andhra Pradesh, India. Selector is a class label set to divide into groups (liver patient or not).This dataset contains 441 male patient records and 142 female patient records. These ten real values are the following: radius, texture, perimeter, concavity, smoothness, compactness, area, concave points, symmetry, and fractaldimension.



**Figure 6.1: Attributes of liver dataset**

**AttributeInformation:**

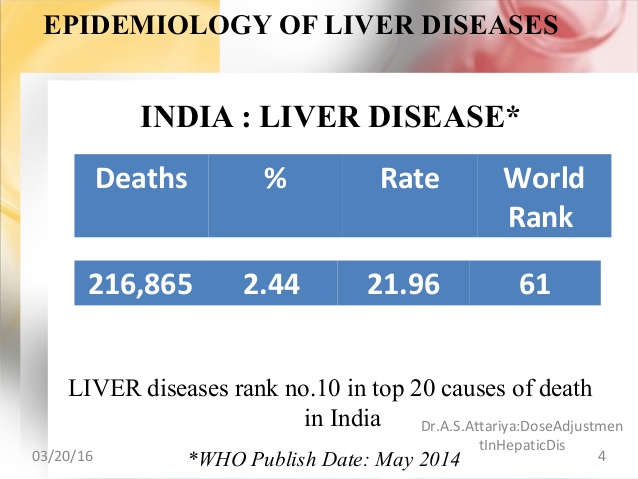
|  |  |
| --- | --- |
| **Attribute name** | **Type** |
| Age | Integer |
| Gender | Number |
| Total\_bilirubin (tb) | number |
| Direct\_ bilirubin (db) | Number |
| Alkaline Phosphotase  (Alkphos) | Integer |
| Alamine Aminotransferase   (Sgpt) | Integer |
| Aspartate Aminotransferase  ( Sgot) | integer |
| Total Protiens  (TP) | Number |
| Albumin | Number |
| Albumin and Globulin Ratio (radio) | Number |
| Label | Integer 1-liver patient  0-Non liver patient |

**Table 6.1: Attributes and its types**

**Liver disease in India**:

According to the latest WHO data published in 2017 Liver Disease Deaths in India reached 257,749 or 2.95% of total deaths. The age adjusted Death Rate is 22.93 per 100,000of population ranks India #63 in the world.

2, 00,000 Indians die of liver failure every year. 25,000 liver transplants need to be done every year in India. Only 1,100 transplants performed in India every year.



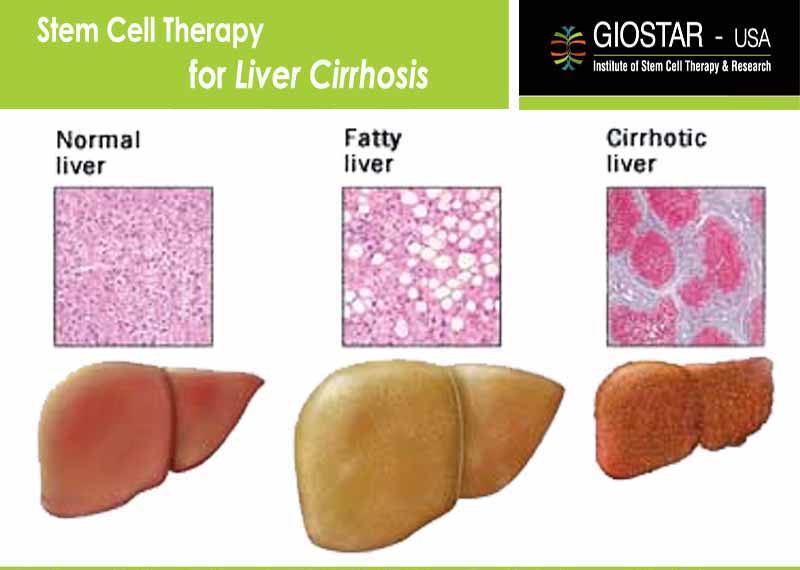
**Figure 6.2: Death rate in India**

Parasites and viruses can infect the liver, causing inflammation and that reduces liver function. The viruses that cause liver damage can be spread through blood or semen, contaminated food or water, or close contact with a person who is infected. The most common types of liver infection are hepatitis viruses, including:

* Hepatitis A
* Hepatitis B
* Hepatitis C

Cancer and other growths:

* Liver cancer
* Bile duct cancer
* Liver adenoma



**Figure 6.3: Types of liver cirrhosis**

**6.2MODEL DEVELOPMENT**

The classification is a data mining function that assigns items in collection to target categories or classes. The goal of classification is to accurately target class for each case in the data. The classification is also known as supervised learning or predictive modeling which is based on the nature of the information being extracted.

* **Types of classification models**

Random Forest

K-Nearest Neighbors algorithm

***Logistic Regression***

* **Algorithms used:**

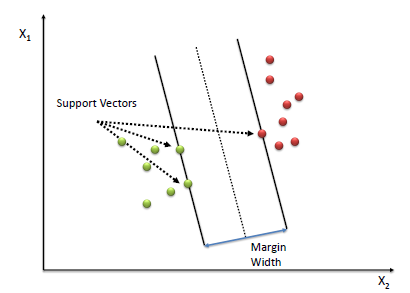
Support Vector Machine(SVM)

**6.2.2Support Vector Machine(SVM)**

The original **SVM algorithm** was invented by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963. In 1992, Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik suggested a way to create nonlinear classifiers by applying the kernel trick to maximum-margin hyperplanes.

According to the SVM algorithm we find the points closest to the line from both the classes.These points are called support vectors. Now, we compute the distance between the line and the support vectors. This distance is called the margin. Our goal is to maximize the margin. The hyperplane for which the margin is maximum is the optimal hyperplane.

Thus SVM tries to make a decision boundary in such a way that the separation between the two classes(that street) is as wide as possible. A hyperplane in an n-dimensional Euclidean space is a flat, n-1 dimensional subset of that space that divides the space into two disconnected parts.



**Figure 6.7: Support Vector Machine**

**Kernel:**

The learning of the hyperplane in linear SVM is done by transforming the problem using some linear algebra. This is where the kernel plays role.

For **linear kernel** the equation for prediction for a new input using the dot product between the input (x) and each support vector (xi) is calculated as

follows:

f(x) = B(0) + sum(ai \* (x,xi))

This is an equation that involves calculating the inner products of a new input vector (x) with all support vectors in training data. The coefficients B0 and ai (for each input) must be estimated from the training data by the learning algorithm.

The **polynomial kernel** can be written as K(x,xi)=1+sum(x\*xi)^d and exponential as K(x,xi)=exp(-gamma\*sum((x-xi²)).

Polynomial and exponential kernels calculates separation line in higher dimension. This is called ***kernel trick.***

7. SYSTEM DESIGN

**7.1 INPUT DESIGN:**

**7.1.1 NORMALIZATION:**

Normalization is scaling technique or a mapping technique or a processing stage. The technique which provides linear transformation on original range of data is called Min-Max Normalization. Where, we can find new range from an existing one range. It can be helpful for the prediction or forecasting purpose a lot. As we know there are so many ways to predict or forecast but all can vary with each other a lot. So to maintain the large variation of prediction and forecasting the Normalization technique is required to make them closer.

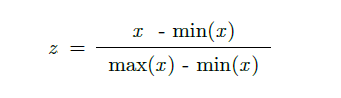
**NORMALIZATION TECHNIQES ARE,**

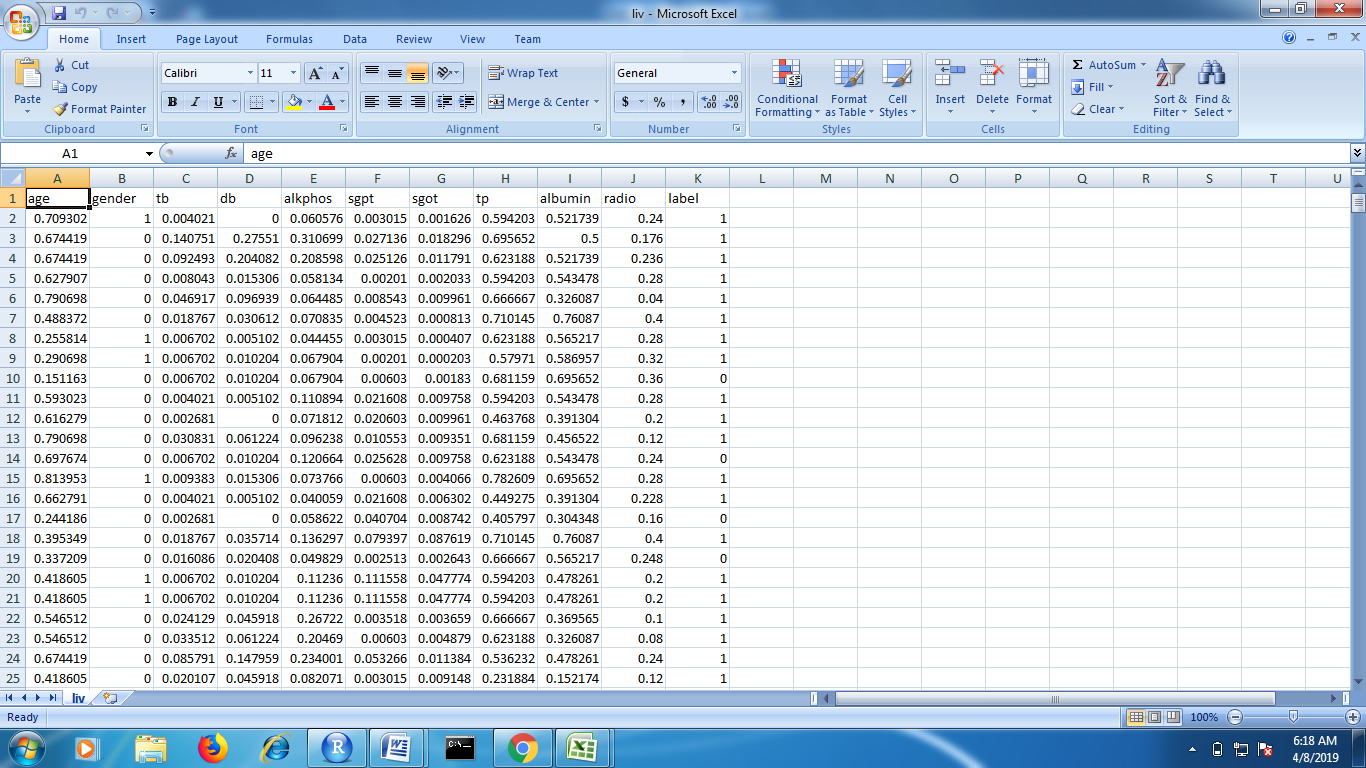
* Min Max normalization
* Z score normalization
* Decimal Scaling

**MIN MAX NORMALIZATION**

The min-max normalize linearly rescales every feature to the [0,1] interval.Rescaling to the [0, 1] interval is done by shifting the values of each feature so that the minimal value is 0, and then dividing by the new maximal value (which is the difference between the original maximal and minimal values).

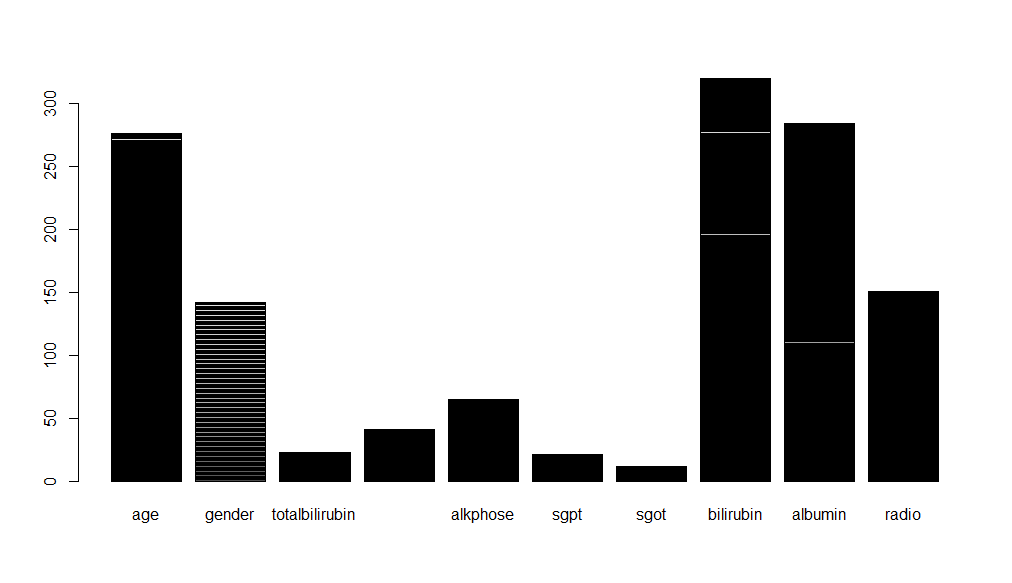
Min Max normalization technique is a technique that helps to normalize the data. It will scale the data between the 0 and 1.**Formula as**:





**Figure 7.1: Normalized the data**

* After normalized the all attributes are shown in figure :



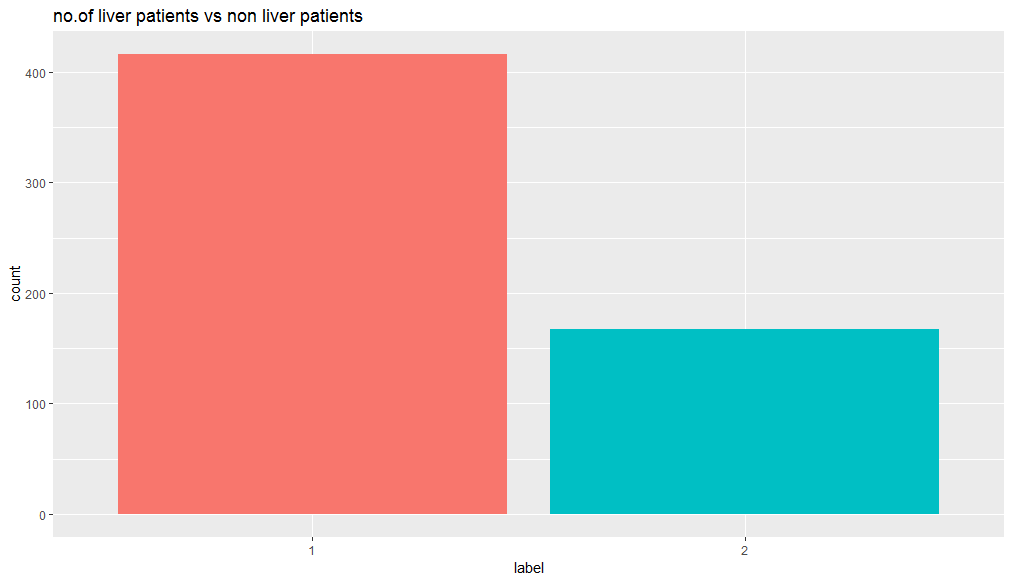
**Figure 7.2: After Normalizing the data**

**8. EXPERIMENTAL ANALYSIS**

**8.1 SCREEN SHOT**:

**8.1.1 LIVER PATIENTS VS NON LIVER PATIENTS:**

The plot represents the count of liver patients and non liver patients. The class label 1 represents liver disease patients and class label 0 represents non liver patients.



**Figure 8.1: Number of liver patients and non liver patients**

**8.2:ACCURACY CALCULATION**

Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

**PREDICT**

|  |  |  |
| --- | --- | --- |
| **ACTUAL** | **POSITIVE** | **NEGATIVE** |
| **POSITIVE** | TP | FP |
| **NEGATIVE** | FN | TN |

**Table 8.1: Confusion Matrix**

**Accuracy** = (true positive + true negative) / (true positive + true negative

+false positive + false negative)

**F measurement:**

The F measure (F1 score or F score) is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test.

F measure= 2 \* ((precision \* recall) / (precision + recall))

**Precision**

Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances.

**Recall**

Recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.

**8.3 RESULTS**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | False Negatives | Cross validation |
| SVM | 0.6348 | 19 | 0.713 |
| Logistic Regression | 0.7043 | 10 | 0.6957 |
| K-Nearest Neibourhood | 0.7043 | 3 | 0.6435 |
| Random forest | 0.713 | 11 | 0.7217 |

**9.APPENDIX**

**9.1 Source Code**

**#Read the datafiles in csv formate**

**liver\_data<- read.csv("file:///F:\\indian\_liver\_patient.csv")**

**#LODING ALL THE REQUIRED LIBRARIES**

**library(psych)**

**library(ggplot2)**

**library(caret)**

**library(psych)**

**library(VIM)**

**library(mice)**

**library(kernlab)**

**library(randomForest)**

**library(caretEnsemble)**

**library(shiny)**

**library(shinydashboard)**

**II.DATA**

**#BASIC UNDERSTANDING OF LIVER\_DATA DATAFRAME**

**str(liver\_data)**

**#Using str() i analyze the basic structure of data. Theconsist of 11 variables and 583 observation. Response variableis dataset is an int with liver disease and where 1 indicate a patient with liver disease and 0 indicate patient without liverdisease. Gender variable is a factor with 2 levels male and female, while other variable are int and num type.**

**III.EXPLORATORY PLOTS**

**temp<-liver\_data**

**temp$Dataset<-as.factor(temp$Dataset)**

**temp$Gender<-as.numeric(temp$Gender)**

**#Plotting a kpie chart of atrribute dataset**

**mytable<-table(temp$Dataset)**

**lbls<-c("patients without liver disease", "patients with disease")**

**lbls<-paste(lbls, "\n", mytable)**

**pie(mytable, labels =lbls, main="pie Chart of Dataset\n (with sample sizes)", col = rainbow(length(lbls)))**

**#Plotting a histogram of attribute gender according to response attribute dataset**

**ggplot(data = temp, aes(x=Gender))+geom\_histogram(binwidth = 0.2, color="black", aes(fill=Dataset))+xlab("Gender")+ ylab("Dataset")+ggtitle("Histogram of Gender")**

**par(mfrow=c(3,3))**

**#Further explored the skewness in the data by plotting a histogram**

**col\_hist <- c("#56B4E9", "#000000", "#009E73", "#E69F00", "#F0E442", "#0072B2", "#D55E00", "#CC79A7")**

**hist(liver\_data$Age,col ="#999999", border="black",las=1, xlab= "Age", main= "Histogram of Age")**

**#plotting histogram for all the feature.**

**for(i in 3:10)**

**{**

**hist(liver\_data[, i], cex.axis=.5, col=col\_hist[i-2], las=1, xlab = names(liver\_data)[i], main = paste("Histogram of", names(liver\_data)[i]))**

**}**

**x <- temp[,1:5]**

**y <- temp[,11]**

**scales <- list(x=list(relation="free"), y=list(relation="free"))**

**featurePlot(x=x, y=y, plot="density", scales=scales)**

**x <- temp[,5:10]**

**y <- temp[,11]**

**scales <- list(x=list(relation="free"), y=list(relation="free"))**

**featurePlot(x=x, y=y, plot="density", scales=scales)**

**ggplot(temp, aes(x=factor(1), y = Age))+geom\_boxplot(width=0.4, fill="white")+geom\_jitter(aes(color=Dataset, shape=Dataset),width=0.1, size=1)+scale\_color\_manual(values = c("#00AF88", "#E78500"))+labs(x=NULL)**

**par(mfrow=c(2,2))**

**col\_boxplot=c("#56B4E9", "#009E73", "#E69F00", "#F0E442", "#56B4E9", "#D55E00", "#CC79A7", "#999999")**

**boxplot(liver\_data$Age, cex.axis=.5, col="#999999", main="Age")**

**for(i in 3:10)**

**{**

**boxplot(liver\_data[,i], cex.axis=.5, col=col\_boxplot[i-2], main=names(liver\_data)[i])**

**}**

**mod<-glm(Dataset~., data=liver\_data)**

**cooksd<-cooks.distance(mod)**

**plot(cooksd, pch="\*", cex=2, main="Influential obs by cooks distance")**

**abline(h=3\*mean(cooksd, na.rm = T), col="red")**

**text(x=1:length(cooksd)+1, y=cooksd, labels = ifelse(cooksd>3\*mean(cooksd, na.rm = T),names(cooksd),""),col="red")**

**pairs.panels(liver\_data, pch = 10)**

**pairs.panels(liver\_data[,c(3,4,6,7,8,9)])**

**scatter<-ggplot(data=temp, aes(x= temp$Total\_Bilirubin, y=temp$Direct\_Bilirubin))**

**scatter+geom\_point(aes(color=Dataset, shape=Dataset))+**

**xlab("Total\_Bilirubin")+ ylab("Direct\_Bilirubin")+**

**ggtitle("Total\_Bilirubin")**

**table(is.na(liver\_data))**

**sapply(liver\_data,function(x) sum(is.na (x)))**

**missing\_plot <- aggr(liver\_data, col=c('navyblue','yellow'), numbers=TRUE, sortVars=TRUE,**

**labels=names(liver\_data),cex.axis=.5, gap=3, ylab=c("Missing data", "Pattern"))**

**liver\_data\_dummy<-liver\_data**

**dmy <- dummyVars("~.", data=liver\_data\_dummy, fullRank=T)**

**liver\_data\_dummy <-data.frame(predict(dmy,newdata=liver\_data\_dummy))**

**str(liver\_data\_dummy)**

**imputed\_liver\_dt\_dummy<-mice(liver\_data\_dummy, m=4, maxit = 20, method= 'pmm', seed= 500)**

**summary(imputed\_liver\_dt\_dummy)**

**liver\_data\_dummy\_imputed <- complete(imputed\_liver\_dt\_dummy,2)**

**table(is.na(liver\_data\_dummy\_imputed))**

**influential <- as.numeric(names(cooksd)[(cooksd > 3\*mean(cooksd, na.rm = T))])**

**influential**

**liver\_data\_dummy\_imputed\_noOut <- liver\_data\_dummy\_imputed[-influential,]**

**normalize<-function(x)**

**{**

**return((x-mean(x))/sd(x))**

**}**

**liver\_data\_dummy\_imputed\_noOut\_norm<-**

**as.data.frame(lapply(liver\_data\_dummy\_imputed\_noOut[,c(1,3:10)],normalize))**

**Dataset<-liver\_data\_dummy\_imputed\_noOut$Dataset**

**Gender<-liver\_data\_dummy\_imputed\_noOut$Gender.Male**

**liver\_data\_dummy\_imputed\_noOut\_norm$Gender<-Gender**

**liver\_data\_dummy\_imputed\_noOut\_norm$Dataset<-Dataset**

**summary(liver\_data\_dummy\_imputed\_noOut\_norm)**

**liver\_clean\_data<-liver\_data\_dummy\_imputed\_noOut\_norm**

**str(liver\_clean\_data)**

**set.seed(80)**

**liver\_clean\_data\_lm <- liver\_clean\_data**

**liver\_clean\_data$Dataset <- as.factor(liver\_clean\_data$Dataset)**

**Index <-createDataPartition(liver\_clean\_data$Dataset, p=0.8, list = FALSE)**

**train\_data <- liver\_clean\_data[Index,]**

**test\_data<-liver\_clean\_data[-Index,]**

**train\_data\_lm <-liver\_clean\_data\_lm[Index,]**

**test\_data\_lm <-liver\_clean\_data\_lm[-Index,]**

**#vector machine.**

**set.seed(200)**

**model\_SVM <- ksvm(Dataset ~., data = train\_data, kernel= "tanhdot")**

**pred\_SVM<-predict(model\_SVM,test\_data)**

**confusionMatrix(pred\_SVM,test\_data$Dataset)**

**#logestic regression.**

**model\_glm <- glm(formula = Dataset ~. , family = binomial, data = liver\_clean\_data\_lm)**

**summary(model\_glm)**

**model\_glm2<-glm(formula = Dataset~., family = binomial, data = subset(liver\_clean\_data\_lm,**

**select = c(-Gender, -Total\_Bilirubin, -Direct\_Bilirubin, -Aspartate\_Aminotransferase)))**

**summary(model\_glm2)**

**pred\_lm<-predict(model\_glm2,test\_data\_lm[,c(1,4,5,7,8,9)],type="response")**

**pred\_lm<-ifelse(pred\_lm>=0.5,1,0)**

**confusionMatrix(as.factor(pred\_lm),as.factor(test\_data\_lm$Dataset))**

**#k-nearest neighboor.**

**ctrl<-trainControl(method = "repeatedcv", repeats = 3)**

**model\_knn <- train(Dataset ~ ., data = train\_data, method="knn", trControl=ctrl,preProcess=**

**c("center","scale"), tuneLength=20)**

**pred\_knn<- predict(model\_knn, newdata=test\_data)**

**confusionMatrix(pred\_knn,test\_data$Dataset)**

**#random forest**

**set.seed(200)**

**model\_randomForest<-randomForest(Dataset~.,data=train\_data,importance=TRUE)**

**pred\_randomForest<-predict(model\_randomForest,test\_data)**

**confusionMatrix(pred\_randomForest,test\_data$Dataset)**

**SVM <- confusionMatrix(pred\_SVM,test\_data$Dataset)$overall['Accuracy']**

**GLM<-confusionMatrix(as.factor(pred\_lm),as.factor(test\_data$Dataset))$overall['Accuracy']**

**KNN<-confusionMatrix(pred\_knn,test\_data$Dataset)$overall['Accuracy']**

**RandomForest<-confusionMatrix(pred\_randomForest,test\_data$Dataset)$overall['Accuracy']**

**accuracy<-data.frame(model=c("Support Vector Machine", "Logistic Regression","knn","randomforest"),Accuracy=c(SVM,GLM,KNN,RandomForest))**

**ggplot(accuracy,aes(x=model, y=Accuracy))+ geom\_bar(stat='identity')+ theme\_bw()+ggtitle('Comparison of model Accuracy')**

**levels(liver\_clean\_data$Dataset)<-make.names(levels(liver\_clean\_data$Dataset))**

**control<-trainControl(method="repeatedcv", number=10,repeats=3, savePredictions=TRUE, classProbs=TRUE)**

**algorithmList<-c('svmRadial','glm','knn','rf')**

**set.seed(200)**

**models<-caretList(Dataset~., data=liver\_clean\_data,trControl = control, methodList = algorithmList)**

**output<-resamples(models)**

**summary(output)**

**dotplot(output)**

**#MODEL IMPROVEMENT**

**set.seed(200)**

**train\_control\_SVM<-trainControl(method = "repeatedcv", number=10)**

**Cross\_model\_SVM<-train(Dataset~., data=train\_data, trControl=train\_control\_SVM, method="svmRadial")**

**print(Cross\_model\_SVM)**

**pred\_svm\_Cross<-predict(Cross\_model\_SVM, test\_data)**

**confusionMatrix(pred\_svm\_Cross,test\_data$Dataset)**

**set.seed(200)**

**train\_control\_GLM<-trainControl(method="repeatedcv", number = 10)**

**Cross\_model\_GLM<-train(Dataset~., data = train\_data\_lm, trControl=train\_control\_GLM, method="glm")**

**print(Cross\_model\_GLM)**

**pred\_glm\_cross<-predict(Cross\_model\_GLM,test\_data\_lm)**

**pred\_glm\_cross<-ifelse(pred\_glm\_cross >=0.5,1,0)**

**confusionMatrix(as.factor(pred\_glm\_cross),as.factor(test\_data\_lm$Dataset))**

**set.seed(200)**

**train\_control\_knn<-trainControl(method = "repeatedcv", number=10)**

**Cross\_model\_knn<-train(Dataset~., data=train\_data,trControl=train\_control\_knn, method="knn")**

**print(Cross\_model\_knn)**

**pred\_knn\_cross<-predict(Cross\_model\_knn,test\_data)**

**confusionMatrix(pred\_knn\_cross,test\_data$Dataset)**

**set.seed(200)**

**train\_control\_RF<-trainControl(method="repeatedcv", number = 10)**

**model\_rf <- train(Dataset~., data=train\_data, trControl=train\_control\_RF, method="rf")**

**print(model\_rf)**

**pred\_rf\_cross<-predict(model\_rf,test\_data)**

**confusionMatrix(pred\_rf\_cross,test\_data$Dataset)**

**10. CONCLUSION**

The algorithm of Back Propagation Neural Network has shown 80% accuracy with short duration when compared with Support Vector Machine in classification. This comparison among algorithms is to provide proper utilization of best algorithm to give correct solution to problems .This comparison should be used in further proceedings in scientific research and prediction of liver disease.

**11. REFERENCES**

Anju Gulia, Dr. Rajan Vohra, Praveen Rani (2014). “ Liver patient classification using intelligent techniques”. International Journal of Computer Science andInformation Technologies.

Bendi Venkata Ramana, Prof. M. S. Prasad Babu and Prof. N. B. Venkateswarlu, “Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis”, International Journal of Computer Science Issues, ISSN :1694-0784, May 2012.

Cortes C, Vapnik V (1995). “Support-vector networks using Machine Learning.

Esraa M Hashem, Mai S Mabrouk (2014). “A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis”. American Journal of Intelligent Systems.

Gish, H. 1990. “A probabilistic approach to the understanding and training of neural network classifiers”. In Proceedings of l€E€ Conference on Acoustics Speech and Signal Processing, pp. 1361-1364, April.

Jankishran Pahariyavohra, Jagdeesh makhijani and sanjay patsariya. “Liver patient classification using intelligence techniques”. International journal of advanced research in computer science and software engineering.

Rajeswari, G.Sophia Reena (2010). “Analysis of Liver Disorder using Data Mining Algorithm”.