A MINOR PROJECT REPORT ON

SMART ELECTRONIC HEALTH RECORD

SUBMITTED BY:

Amit Singh	15803006	B 13
Siddharth Agrawal	15104027	B 11
Raghav Maheshwari	15803014	B 13

SUPERVISOR:

Prof. Adwitya Sinha

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY

(A-10, SECTOR 62, NOIDA)

TABLE OF CONTENTS

CERTIFICATE

This is to certify that the work titled "Smart Electronics Health Record " submitted by Amit Singh, Siddharth Agarwal and Raghav Maheshwari in partial fulfilment for the awarding of Bachelor of Technology degree in Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

…………………………..

Signature of Supervisor

Name of Supervisor - Prof. Adwitya Sinha

Date -

ACKNOWLEDGEMENT

It gives us immense pleasure to express our deepest sense of gratitude and sincere thanks to my highly respected and esteemed guide *Prof Adwitya Sinha* for their valuable guidance, encouragement and help for completing this work. Their useful suggestions for this whole work and co-operative behaviour are sincerely acknowledged. We would like to express our sincere thanks to him for giving us this opportunity to undertake this project. We also wish to express my indebtedness to our parents as well as our family members whose blessings and support always helped me to face the challenges ahead. At the end we would like to express my sincere thanks to all our friends and others who helped ous directly or indirectly during this project work.

CANDIDATE’S DECLARATION

We **(Amit Singh, Siddharth Agarwal and Raghav Maheshwari)** hereby declare that the work presented in this report entitled "**Smart Electronic Health Record**", was in fulfilment of the requirement for the awarding of the Bachelor of Technology degree in Computer Science and was submitted to the Computer Science department, affiliated to the Jaypee Institute Of Information Technology, Noida. This report is an authentic record of our own work carried out during my project. The work reported in this report has not been submitted by us for the award of another degree or diploma.

Date : ………………….   Place : Noida

Work Division

The project wouldn’t have been possible if the team members weren’t so devoted and hardworking. Thus the work distribution is as follow:

<table>
<tr><td colspan="4"><b>Project Title: Smart EHR</b></td></tr>
<tr><td><b>S. NO.</b></td><td><b>Enrolment Number</b></td><td><b>Student name</b></td><td><b>Work Done (in percentage) out of 100 contribution of each member in group</b></td></tr>
<tr><td><b>1.</b></td><td><b>15803006</b></td><td><b>Amit Singh</b></td><td><b>33</b></td></tr>
<tr><td><b>2.</b></td><td><b>15104027</b></td><td><b>Siddharth Agarwal</b></td><td><b>33</b></td></tr>
<tr><td><b>3.</b></td><td><b>15803014</b></td><td><b>Raghav Maheshwari</b></td><td><b>33</b></td></tr>
</table>

Introduction

Smart Electronics Health Record was made keeping in mind to ease the medical processes of detecting severe diseases and their treatment. Covering major human diseases like heart, kidney and cancer, we tried to provide an easy, user friendly and near perfect prediction for general people. Through the web based application, a user just by sitting in their homes could know if his/her symptoms will lead to anything or are just temporary.

We have provided an amiable website for users to enter their symptoms and medical values to predict their disease. Just by clicking, one will get to know what is wrong with their heart or lungs or diabetes level. Moreover it’s a age free application i.e. it could be applicable for people of all age groups.

The author of [1] illustrate how healthiness and technology integration and benchmarking have become a challenge for researchers. Being a major factor in strengthening the present medical systems, it also provides a medium to expand the objectives and strategies of organizations.

In [2], it provides a systematic review of the efficacy of patient-oriented care interventions for people with chronic conditions. Around thirty randomized controlled trials were identified from health-related databases. The findings indicated that most interventions were based on the notion of empowering care and included attempts to educate consumers or prompt them about how to manage a health consultation.

The main objective in [3] was to report on the application and accuracy of the popular mining algorithms (artificial neural networks and decision trees) along with logistic regression on a large data set of breast cancer cases, taking advantage of those available technological advancements to develop prediction models for breast cancer survivability.

Research paper [4] intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today’s medical research particularly in Heart Disease Prediction. It could seen that Decision Tree was in close competition with Bayesian, while algorithms like KNN, Neural Networks, etc were not so efficient. Reduction of actual data size by genetics further improved the efficiency.

In [5], in order to determine how data mining techniques (DMT) and their applications have developed in the past decade, the paper reviews data mining techniques and their applications and development from 2000 to 2011. Keywords were used to identify 216 articles concerning DMT applications, from 159 academic journals. The ability to continually change and acquire new understanding is a driving force for the application of DMT and this will allow many new future applications.

Authors of [6] have provided with the review of recent ML approaches employed in the modeling of cancer progression. The predictive models discussed here are based on various supervised ML techniques as well as on different input features and data samples.

Paper [7] explores the use of machine-learning based alternatives to standard statistical data completion (data imputation) methods, for dealing with missing data. It has used two techniques, mainly, unsupervised clustering strategy which uses a Bayesian approach to cluster the data into classes and modeling missing variables by supervised induction of a decision tree-based classifier.

Technology

Smart Electronics Health Record was made keeping in mind to ease the medical processes of detecting severe diseases and their treatment. Covering major human diseases like heart, kidney and lungs, we tried to provide an easy, user friendly and near perfect prediction for general people. Through the web based application merged with R Programming, a user just by sitting in their homes could know if his/her symptoms will lead to anything or are just temporary.

We have provided an amiable website for users to enter their symptoms and medical values to predict their disease. The website is made with the help of a new technology, namely R Shiny. With a little bit help of HTML, we were able to integrate various diseases under one url.

Our purpose was to find the best algorithm which could work for all the diseases. To do so, we have applied eight different machine learning and deep learning algorithms. These are : Linear, Bayesian, Random Forest, Extreme GB Boosting, Decision Trees, Genetic Algorithms and Neural Networks with Tensor Flow. At the end, we found that Extreme GB Boosting gives the most promising results in all the disease data sets.

Besides Classifications, we have also provided data visualization to show relationship between different attributes for a disease. We have used various packages of R, like ggplots, pie charts, time analysis etc, for a better experience.

Description

Smart EHR covers majorly five diseases: Heart, Diabetes, Parkinson’s Kidney and Breast Cancer. Before going further deep into the technical aspects, it is important to understand what each data sets look like and their attributes.

Heart

The heart dataset has 10 attributes which contribute to the prediction of the disease. Each attribute its own level of importance for the cause of heart disease.

Table 1. Attributes of Heart dataset

S.No	Name	Description
1.	Cp	Chest pain type
2.	Trestbps	Resting blood pressure (in mm Hg on admission to the hospital)
3.	Chol	Serum cholestoral in mg/dl
4.	Fbs	Fasting blood sugar
5.	Restecg	Resting electrocardiographic results
6.	Thalach	Maximum heart rate achieved
7.	Exang	Exercise induced angina
8.	Op	Observered pressure
8.	Slope	The slope of the peak exercise ST segment
9.	Ca	Number of major vessels (0-3) colored by flourosopy
10.	Thal	Defect in Thal level
11.	Heart	Is there a heart disease or no

Diabetes

Not only insulin levels in the body affects Diabetes, but there are many other factors like body mass index, age etc also impacts on the level of Diabetes.

Table 2. Attributes of Diabetes dataset

S.No	Name	Description
1.	Pregnant	Number of times pregnant
2.	Plasma_Glucose	Plasma glucose concentration in the body
3.	Dias_BP	Blood pressure
4.	Triceps_Skin	Triceps skin fold thickness
5.	Serum_Insulin	Serum insulin
6.	BMI	Body mass index
7.	DPF	Pedigree function
8.	Age	Age of the patient
9.	Diabetes	Is there Diabetes or not

Cancer

The breast cancer dataset has 11 attributes which contribute to the prediction of the disease. Each attribute contributes into the likeliness of the disease.

Table 3. Attributes of Cancer dataset

S.No	Name	Description
1.	diagnosis	Diagnosis (M = malignant, B = benign)
2.	radius	Distance from centre to points on the perimeter
3.	texture	Standard deviation of grey-scale values
4.	perimeter	Perimeter
5.	area	Area
6.	smoothness	Local variation in radius lengths
7.	compactness	Perimeter^2 / area - 1.0
8.	concavity	Severity of concave portions of the contour
9.	concave points	Number of concave portions of the contour
10.	symmetry	Symmetry
11.	fractal dimension	“Coastline approximation" - 1

Parkinson’s

The Parkinson’s dataset has 25 attributes which contribute to the prediction of the disease. Each attribute has its unique importance and fundamental frequency.

Table 4. Attributes of Parkinson’s dataset

S.No	Name	Description
1.	MDVP:Fo(Hz)	Average vocal fundamental frequency
2.	MDVP:Fhi(Hz)	Maximum vocal fundamental frequency
3.	MDVP:Flo(Hz)	Minimum vocal fundamental frequency
4.	MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP	Several measures of variation in fundamental frequency
5.	MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA	Several measures of variation in amplitude
6.	NHR, HNR	Two measures of ratio of noise to tonal components in the voice
7.	Status	Health status of the subject (one) - Parkinson's, (zero) - healthy
8.	RPDE, D2	Two nonlinear dynamical complexity measures
9.	DFA	Signal fractal scaling exponent
10.	spread1, spread2, PPE	Three nonlinear measures of fundamental frequency variation

Kidney

The kidney dataset has 25 attributes which contribute to the prediction of the disease, which includes parameters like Red Blood Cells, Pus cells clumps, Appetite and so on.

Table 5. Attributes of Kidney Dataset

S.No	Name	Description
1.	age	Age
2.	bp	Blood pressure
3.	sg	Specific gravity
4.	al	Albumin
5.	su	Sugar
6.	rbc	Red blood cells
7.	pc	Pus cell
8.	pcc	Pus cell clumps
9.	ba	Bacteria
10.	bgr	Blood glucose random
11.	bu	Blood urea
12.	sc	Serum creatinine
13.	sod	Sodium
14.	pot	Potassium
15.	Hemo	Hemoglobin
16.	pcv	Packed cell volume
17.	wc	White blood cell count
18.	rc	Red blood cell count
19.	Htn	Hypertension
20.	Dm	Diabetes mellitus
21.	cad	Coronary artery disease
22.	appet	Appetite
23.	Pe	Pedal edema
24.	Ane	Anaemia
25.	Class	Class variable 0 – not CKD 1- CKD

Visualization

It is very important to visualize data before actually applying different algorithm. Visualization helps in seeing different kinds of relationship between different attributes of the data sets.

Heart

We saw that the attribute 'Cp' affects the most in predicting the heart diseases, 'Chol' following it. Many other attributes also play a prominent role in predicting heart disease.

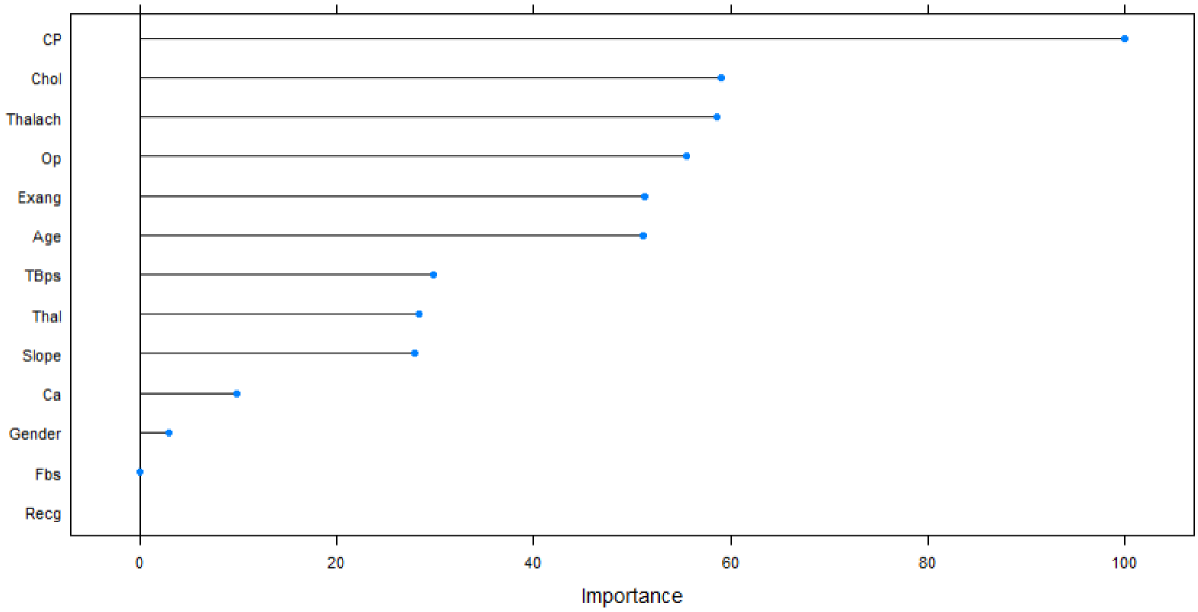


Fig 1.1 Each Attribute importance for predicting Heart Disease

The relationship between 'Thalach' and 'Op' was also quite interesting. As we can see most of the occurrences are between are between 0.3 to 0.8 of OP and above 0.25 of thal which implies that maximum heart rate is almost above 0.25 .

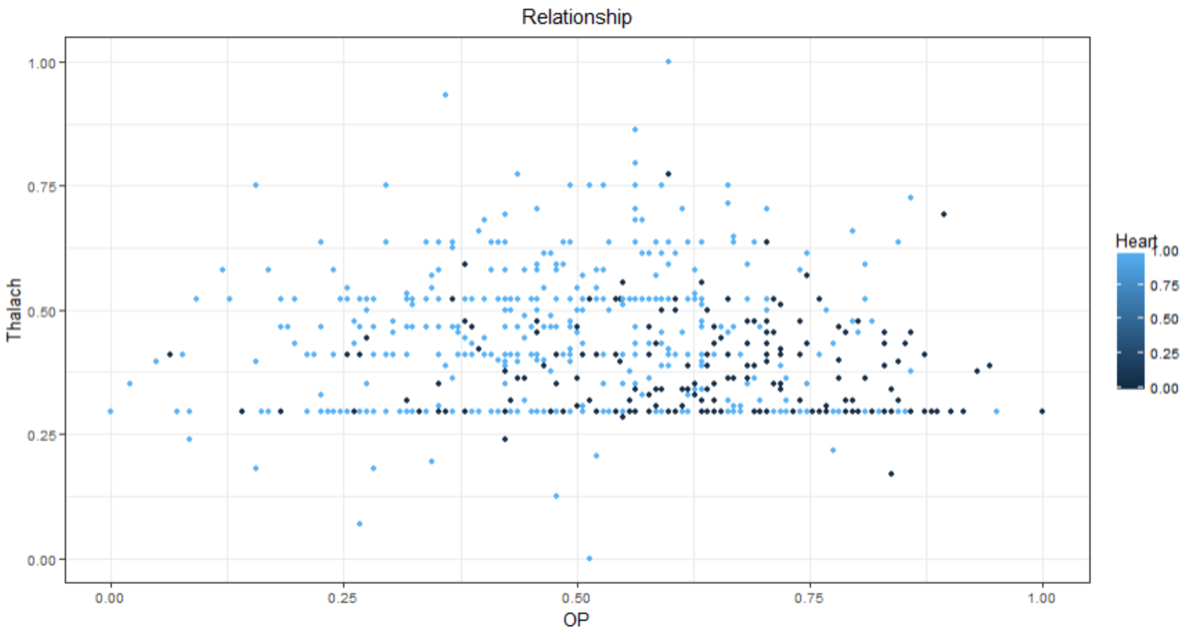


Fig 1.2 Thalach and OP Parameters Relationship for Heart Disease

The working of the neural networks algorithm helped us to understand more clearly about various attributes.

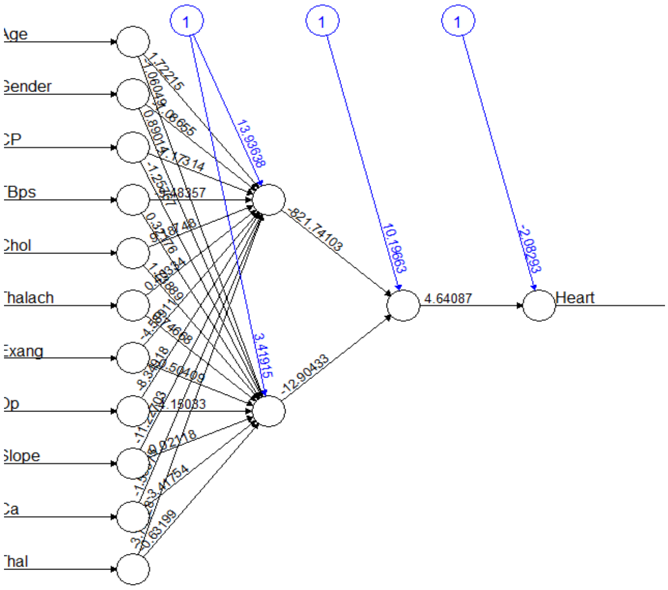


Fig 1.3 Neural net Graph for Heart dataset

Diabetes

Diabetes could attack anyone, but most importantly pregnant women risks chances of catching diabetes. We can clearly see that as the rate of pregnancy increase the risk of diabetes increases.

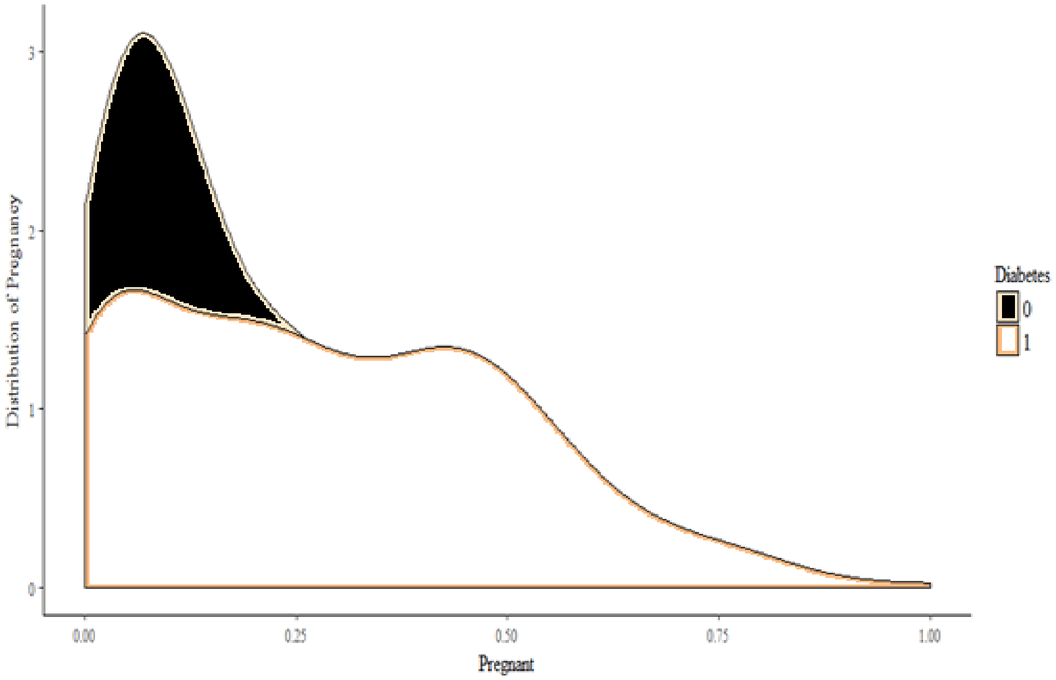


Fig 1.4 Pregnancy and the risk of Diabetes

The correlation plot gives all the 1s in the final boxes, which clearly indicates that all the attributes depend on each other for the correct prediction of diabetes.

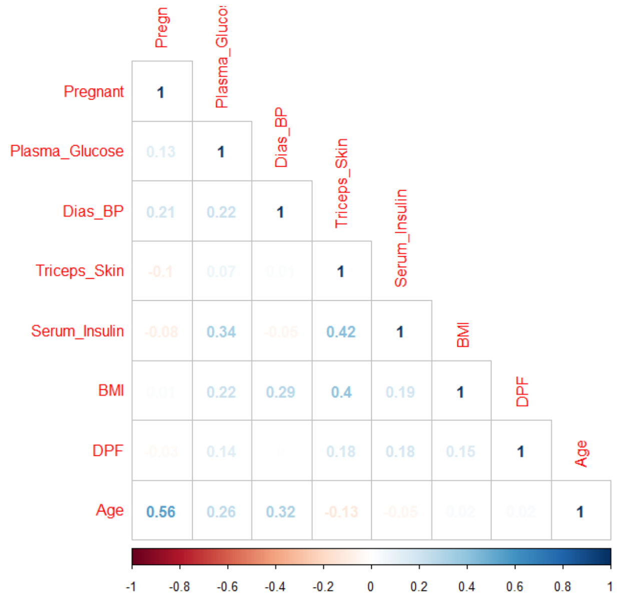


Fig 1.5 Correlation table for Diabetes dataset : All Attributes are Important

Also, it was interesting to see the box plot of serum insulin level and its effects on Diabetes, though there were some outliners too in the dataset. It can be seen that more people had low serum insulin levels than the ones who were diabetic.

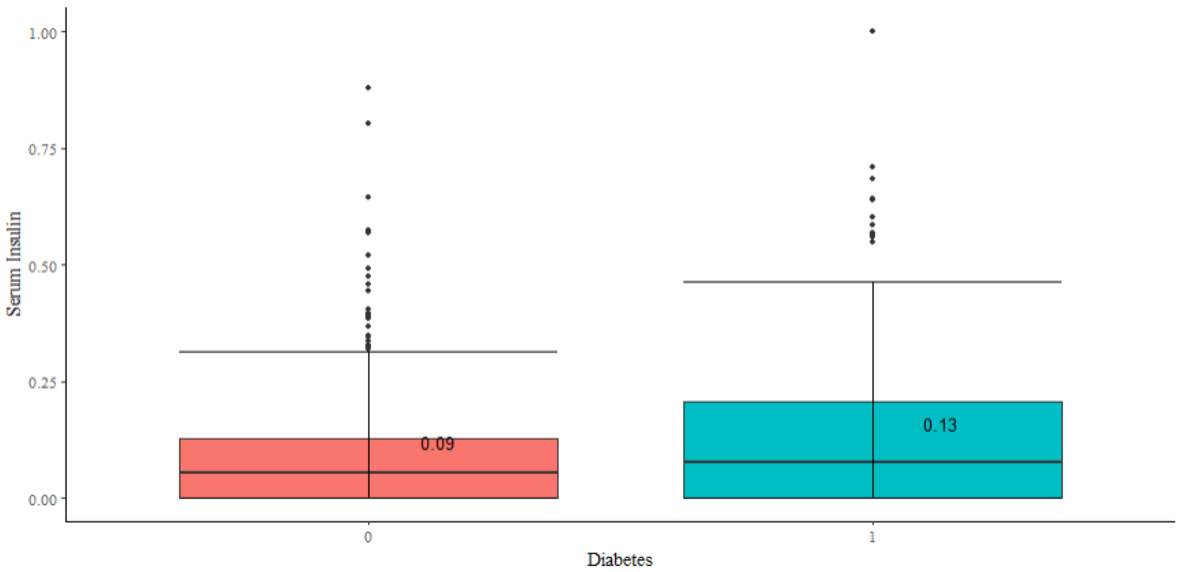


Fig 1.6 Serum insulin level Boxplot for Diabetes

Cancer

The Graph between the Perimeter and Diagnosis of Breast cancer is pretty simple as women who have breast perimeter of 100-150 which is normal have shown no signs of breast cancer whereas the women with perimeters ranging from 50-140 have higher probability of breast cancer.

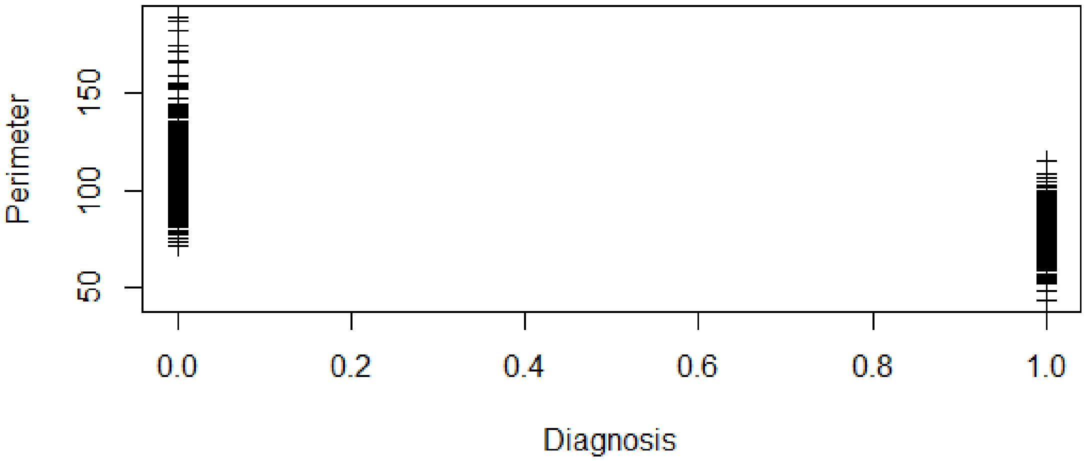


Fig 1.7 Breast perimeter and the likeliness of Cancer

Smoothness of breasts ranging from 0.075 to 0.125 is an indication of having normal breasts whereas women with extremely low or high smoothness can be an indication of having a breast cancer.

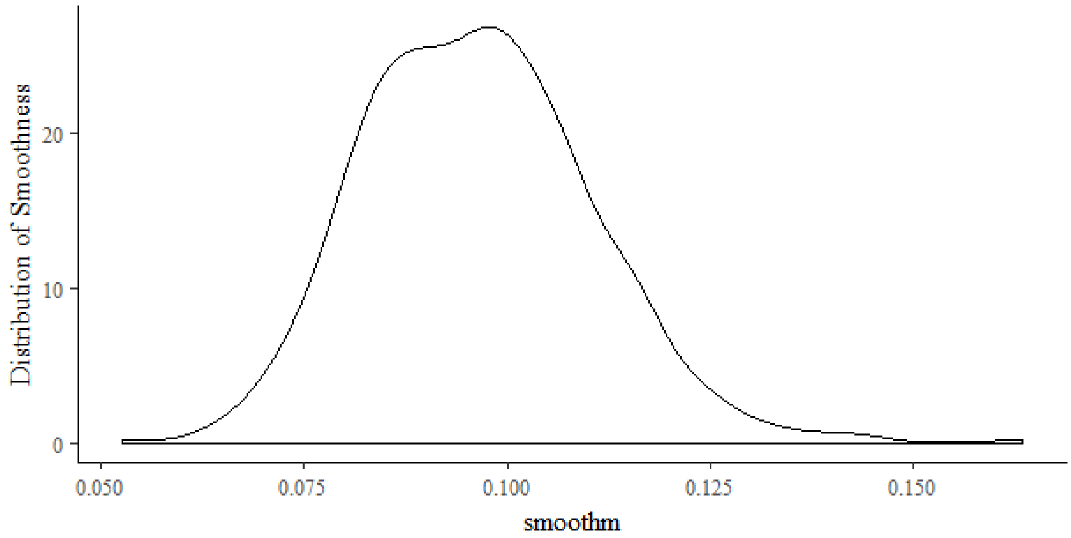


Fig 1.8 Smoothness to Cancer Relationship

Parkinson

The Decision Tree below predicts the outcome of the disease with considerable accuracy. The ppe attribute becomes the root of the tree, followed by apq and so on.

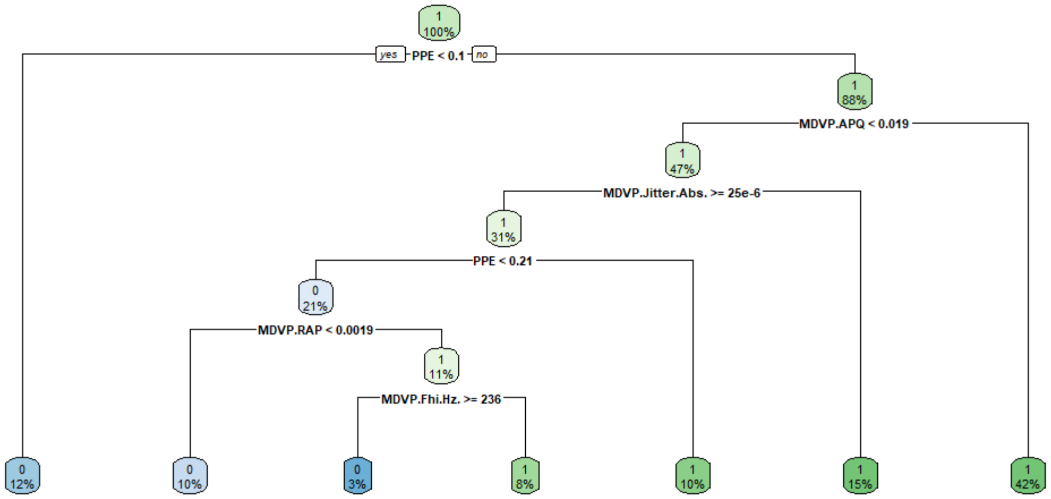


Fig 1.9 Decision Tree for Parkinson Disease with PPE attribute as root of the tree.

We saw that the attribute 'PPE' affects the most in predicting the Parkinson's disease with 'spread1' following it. Other attributes has also played an important role in predicting Parkinson's.

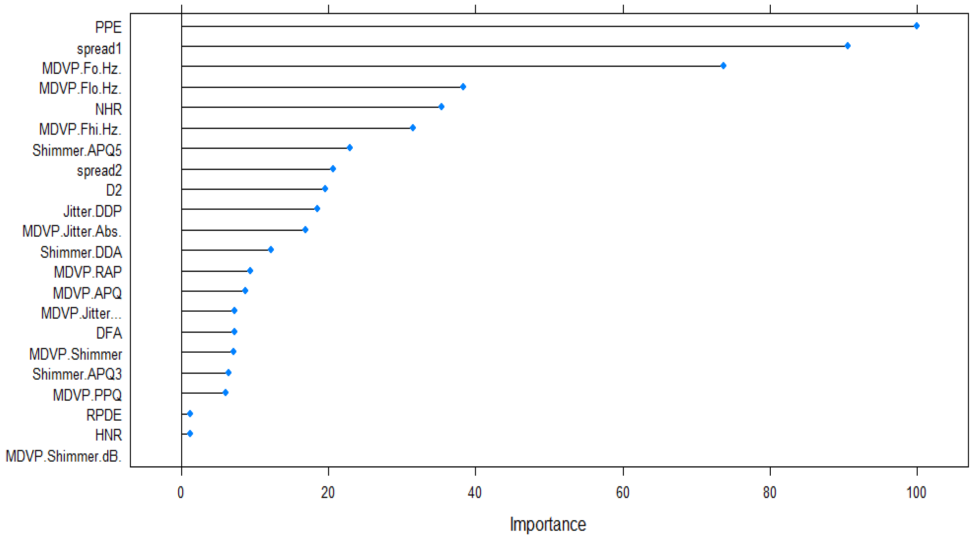


Fig 1.10 Attribute importance Graph for Parkinson Disease

Kidney

The interaction between Blood Glucose Random and Blood Urea is quite interesting as we can say that people who do not have kidney disease have low levels of urea and glucose compared to people suffering from kidney disease.

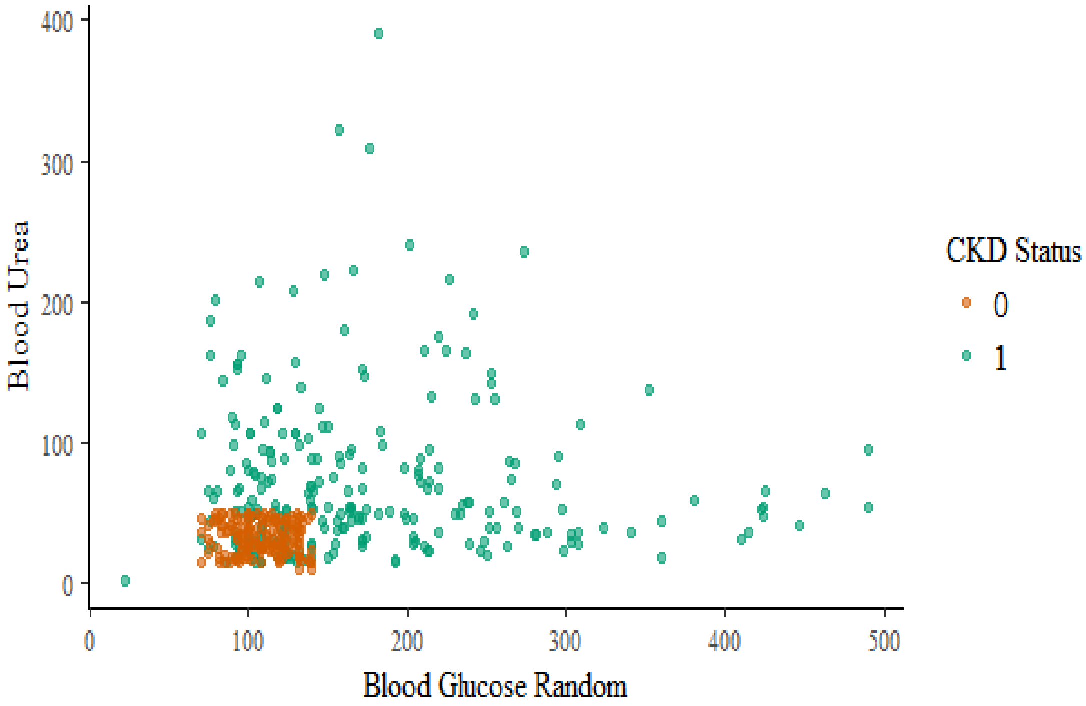


Fig 1.11 Blood Urea and Glucose relationship of Kidney

Graph of hypertension and threat of CKD implies that people with low hypertension levels has a higher rate of Kidney disease then with moderate levels of Hypertension shows low rate of CKD and then again higher levels of Hypertension shows higher rate of CKD.

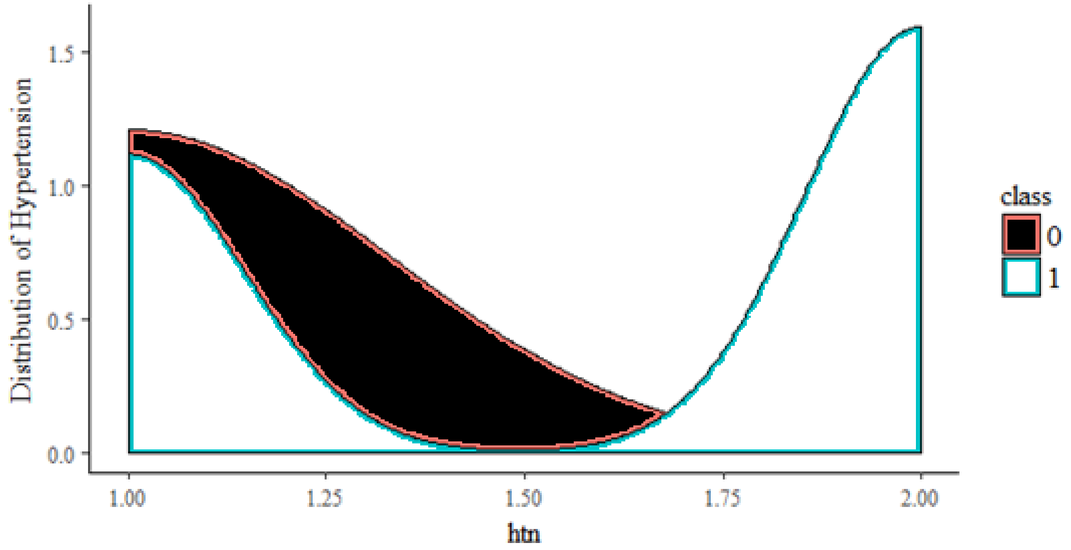


Fig 1.12 Hypertension with the likeliness of occurrence of Kidney Disease

A simple plot between class variable and blood pressure shows that there are equal number of people who have CKD and who do not have CKD at lows levels of blood pressure but at higher levels of blood pressure the chances of CKD decreases but there have been some instances of CKD.

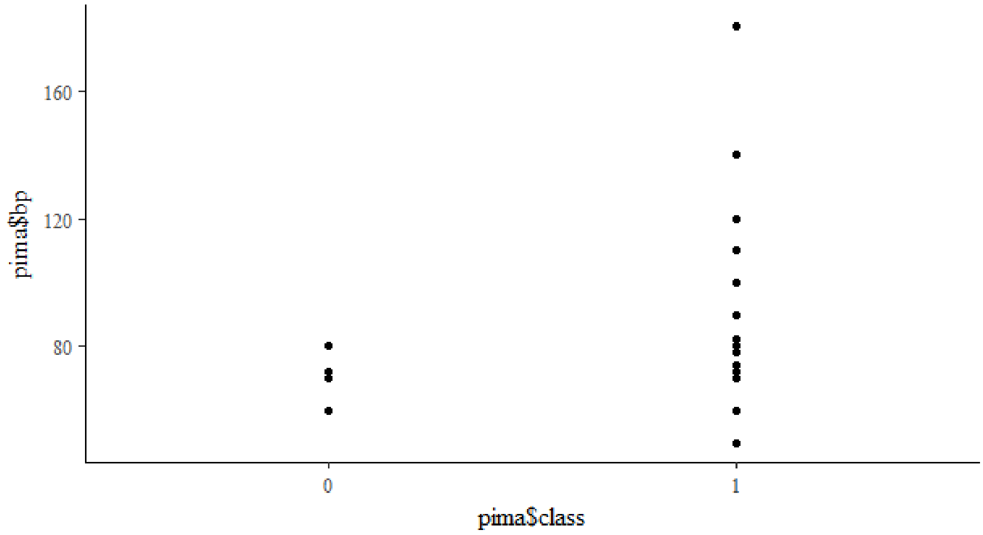


Fig 1.13 Blood pressure and Kidney Disease Box Plot

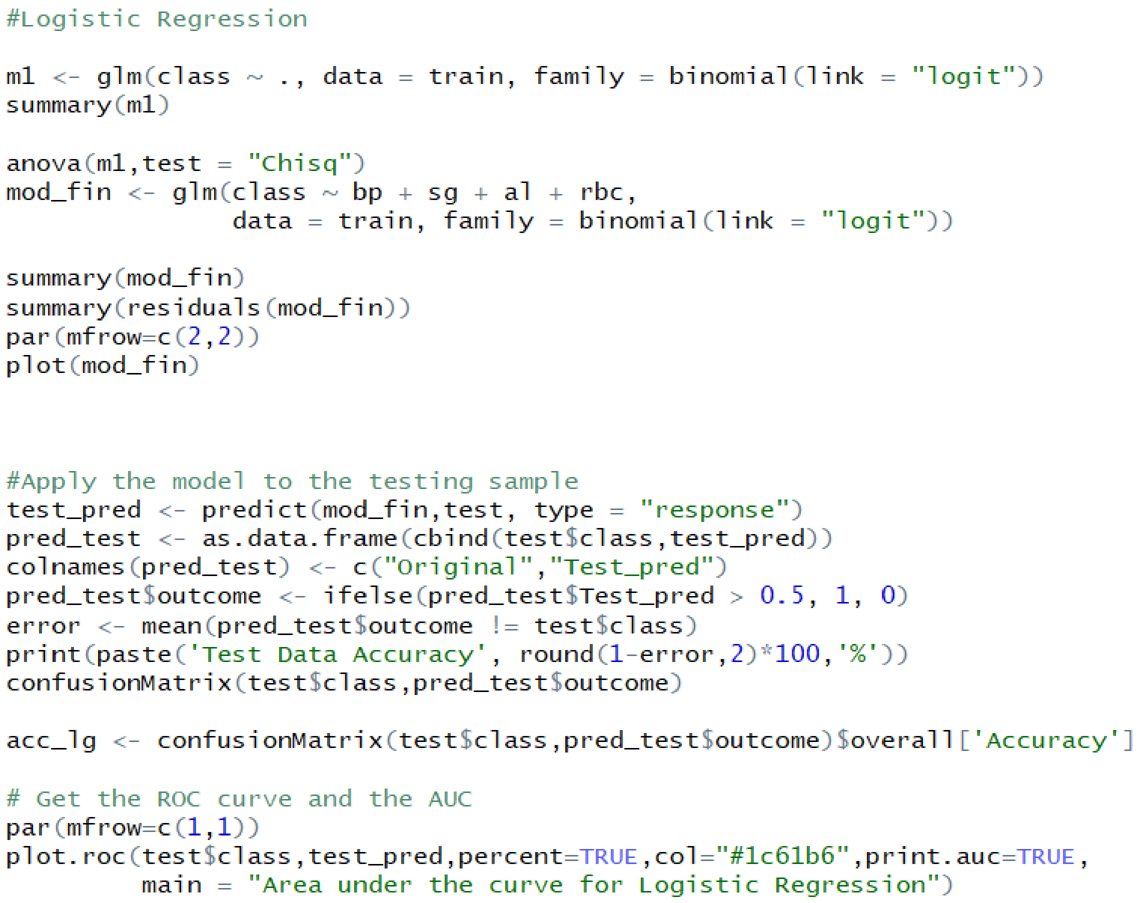
Classification

Classification is the technique which classifies the whole data set according to the class variable. To maintain the credibility of the prediction, we incorporated seven machine learning algorithms for classifying the data set.

Logistic Regression

Logistic regression is a classification regression model where the dependent variable is categorical and is needed to be classified in binary classes. The case of a binary dependent variable that is, where the output can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. Logistic regression is similar to linear regression but unlike it, logistic regression is a classification algorithm. In the project we used the caret module to implement the logistic regression on the given data to predict and classify the tests.

Logistic regression simply uses the technique of reducing the error function and hence deducing the equation for the line of predicted values as in linear regression but it uses a damping function to contain the values between 1 and 0. In general it proved to be a very accurate algorithm as it gave accuracy of 75% in all the datasets that we used. Its average accuracy was 78.3%.



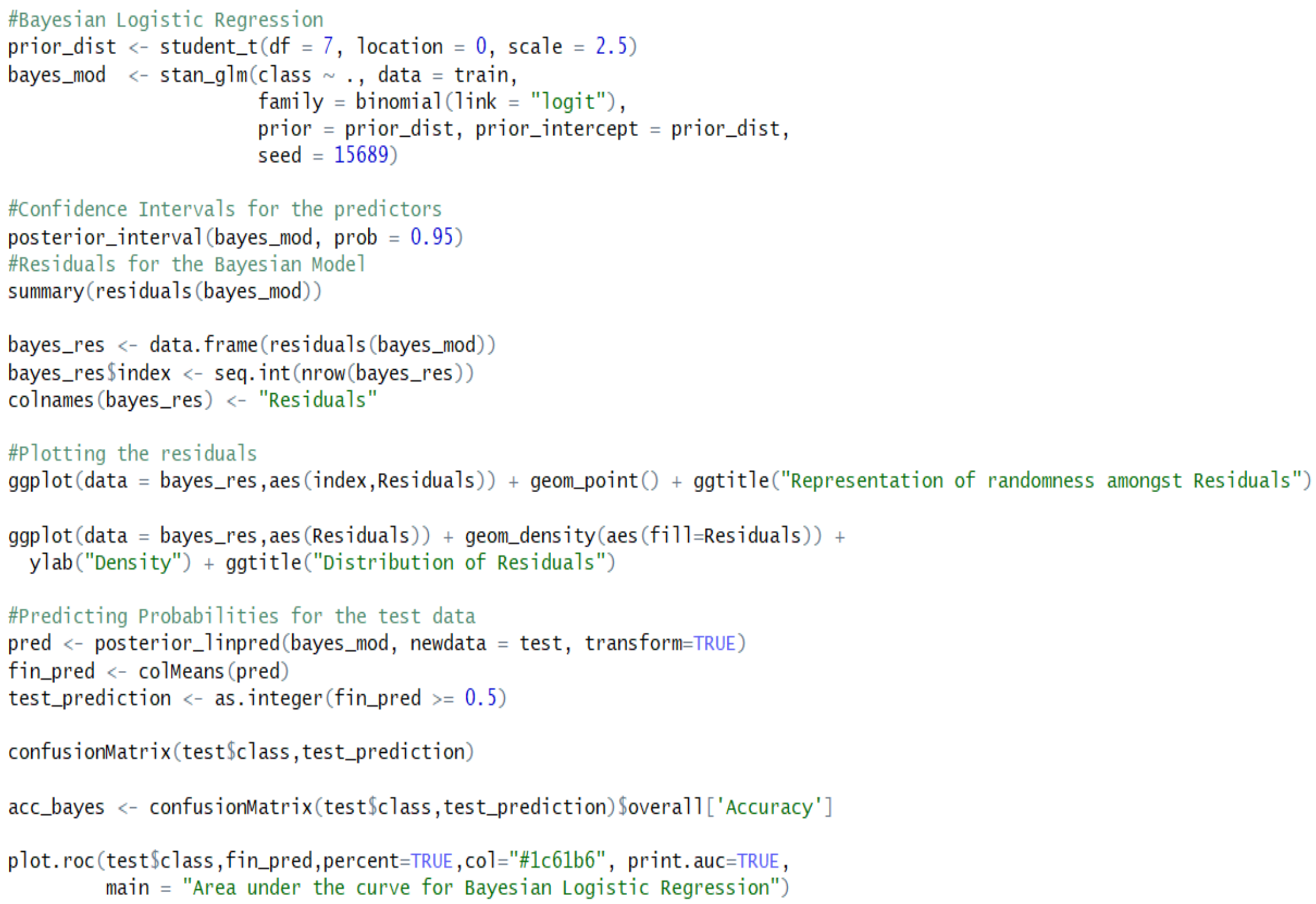
```
#Logistic Regression

m1 <- glm(class ~ ., data = train, family = binomial(link = "logit"))
summary(m1)

anova(m1,test = "Chisq")
mod_fin <- glm(class ~ bp + sg + al + rbc,
               data = train, family = binomial(link = "logit"))

summary(mod_fin)
summary(residuals(mod_fin))
par(mfrow=c(2,2))
plot(mod_fin)



#Apply the model to the testing sample
test_pred <- predict(mod_fin,test, type = "response")
pred_test <- as.data.frame(cbind(test$class,test_pred))
colnames(pred_test) <- c("Original","Test_pred")
pred_test$outcome <- ifelse(pred_test$Test_pred > 0.5, 1, 0)
error <- mean(pred_test$outcome != test$class)
print(paste('Test Data Accuracy', round(1-error,2)*100,'%'))
confusionMatrix(test$class,pred_test$outcome)

acc_lg <- confusionMatrix(test$class,pred_test$outcome)$overall['Accuracy']

# Get the ROC curve and the AUC
par(mfrow=c(1,1))
plot.roc(test$class,test_pred,percent=TRUE,col="#1c61b6",print.auc=TRUE,
         main = "Area under the curve for Logistic Regression")
```

Fig 1.14 Logistic Regression Code Snipet

Bayesian Logistic Regression

Bayesian analyses of binary or categorical outcomes typically rely on probability or mixed effects logistic regression models which do not have a marginal logistic structure for the individual outcomes. The BLR model for individual outcomes has a marginal logistic structure, simplifying interpretation and follow a Bayesian approach to estimation and inference, developing an efficient data augmentation algorithm for computation.

Bayesian logistic regression provide with accurate results of around 80%. It is a very efficient algorithm especially in the parkinson’s dataset in which it gave an accuracy of 87%.



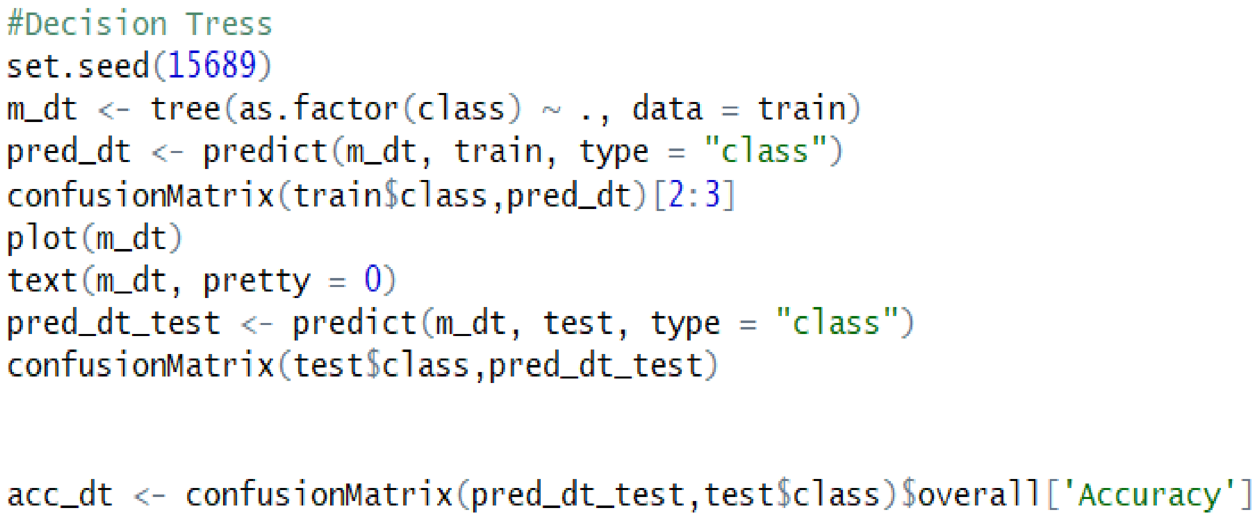
```r
#Bayesian Logistic Regression
prior_dist <- student_t(df = 7, location = 0, scale = 2.5)
bayes_mod  <- stan_glm(class ~ ., data = train,
                       family = binomial(link = "logit"),
                       prior = prior_dist, prior_intercept = prior_dist,
                       seed = 15689)

#Confidence Intervals for the predictors
posterior_interval(bayes_mod, prob = 0.95)
#Residuals for the Bayesian Model
summary(residuals(bayes_mod))

bayes_res <- data.frame(residuals(bayes_mod))
bayes_res$index <- seq.int(nrow(bayes_res))
colnames(bayes_res) <- "Residuals"

#Plotting the residuals
ggplot(data = bayes_res,aes(index,Residuals)) + geom_point() + ggtitle("Representation of randomness amongst Residuals")

ggplot(data = bayes_res,aes(Residuals)) + geom_density(aes(fill=Residuals)) +
  ylab("Density") + ggtitle("Distribution of Residuals")

#Predicting Probabilities for the test data
pred <- posterior_linpred(bayes_mod, newdata = test, transform=TRUE)
fin_pred <- colMeans(pred)
test_prediction <- as.integer(fin_pred >= 0.5)

confusionMatrix(test$class,test_prediction)

acc_bayes <- confusionMatrix(test$class,test_prediction)$overall['Accuracy']

plot.roc(test$class,fin_pred,percent=TRUE,col="#1c61b6", print.auc=TRUE,
         main = "Area under the curve for Bayesian Logistic Regression")
```

Fig 1.15 Bayesian Logistic Regression Code Snipet

Decision Trees

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute , each branch represents the outcome of the test, and each leaf node represents a class label. The paths from root to leaf represent classification rules.

Decision tree is a classification technique that is easy to interpret and understand and hence easy to use. It can be used in both normal and hard data and is pretty accurate if the data is not overfitting to the classifiers.



```
#Decision Tress
set.seed(15689)
m_dt <- tree(as.factor(class) ~ ., data = train)
pred_dt <- predict(m_dt, train, type = "class")
confusionMatrix(train$class,pred_dt)[2:3]
plot(m_dt)
text(m_dt, pretty = 0)
pred_dt_test <- predict(m_dt, test, type = "class")
confusionMatrix(test$class,pred_dt_test)


acc_dt <- confusionMatrix(pred_dt_test,test$class)$overall['Accuracy']
```

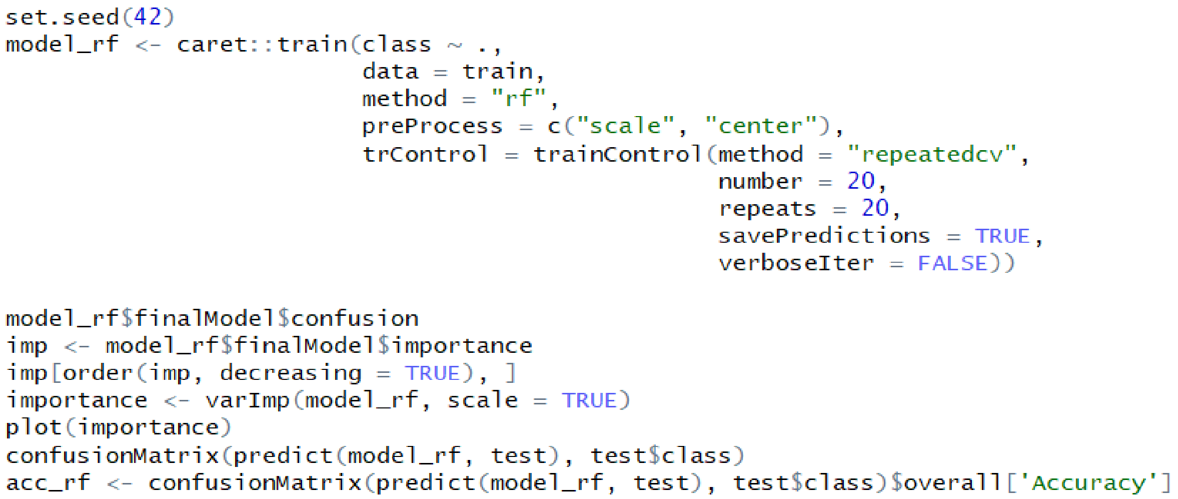
Fig 1.16 Decision Tree Code Snipet

Random Forest

Random forest is a tree-based machine learning algorithm which involves building several trees and combining their output to predict outcomes.

Firstly, the dataset is split into random samples by using bootstrap aggregating algorithm. A new dataset is created by sampling n random cases from original data leaving about one third of data.

Then the new data obtained is trained using the given model in which very small columns of data are selected at random. At the end, each selected data is allowed to grow fully and final prediction is obtained by averaging.



```r
set.seed(42)
model_rf <- caret::train(class ~ .,
                         data = train,
                         method = "rf",
                         preProcess = c("scale", "center"),
                         trControl = trainControl(method = "repeatedcv",
                                                  number = 20,
                                                  repeats = 20,
                                                  savePredictions = TRUE,
                                                  verboseIter = FALSE))

model_rf$finalModel$confusion
imp <- model_rf$finalModel$importance
imp[order(imp, decreasing = TRUE), ]
importance <- varImp(model_rf, scale = TRUE)
plot(importance)
confusionMatrix(predict(model_rf, test), test$class)
acc_rf <- confusionMatrix(predict(model_rf, test), test$class)$overall['Accuracy']
```

Fig 1.17 Random Forest Code Snipet

Extreme Gradient Boosting

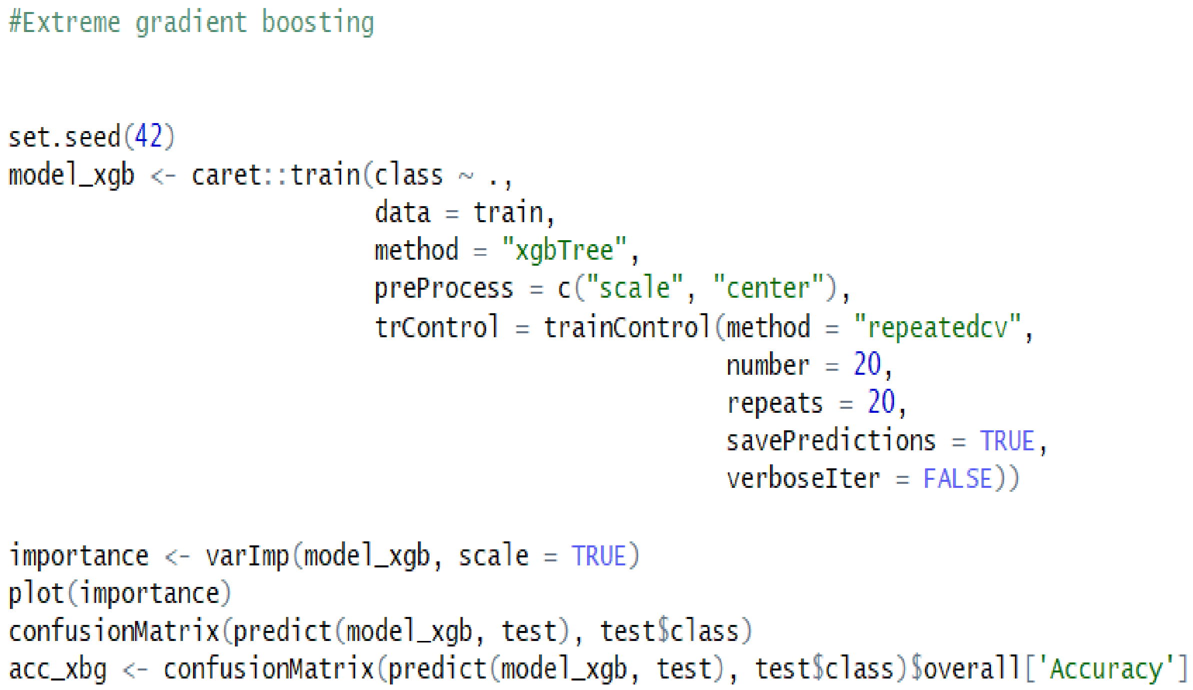
Gradient boosting is a machine learning algorithm which uses weak prediction models to create a final prediction model.

At first data is modelled onto simple models like decision trees and check for error residuals.

For further models we fit the data onto error residuals as target with same input variables.

Now it add the new error residuals with the original ones and fit the data onto updated error residuals.

Repetition of the steps are done until the sum of residuals become constant. In the end all predictors are added to predict outcome.



```
#Extreme gradient boosting


set.seed(42)
model_xgb <- caret::train(class ~ .,
                          data = train,
                          method = "xgbTree",
                          preProcess = c("scale", "center"),
                          trControl = trainControl(method = "repeatedcv",
                                                   number = 20,
                                                   repeats = 20,
                                                   savePredictions = TRUE,
                                                   verboseIter = FALSE))

importance <- varImp(model_xgb, scale = TRUE)
plot(importance)
confusionMatrix(predict(model_xgb, test), test$class)
acc_xbg <- confusionMatrix(predict(model_xgb, test), test$class)$overall['Accuracy']
```

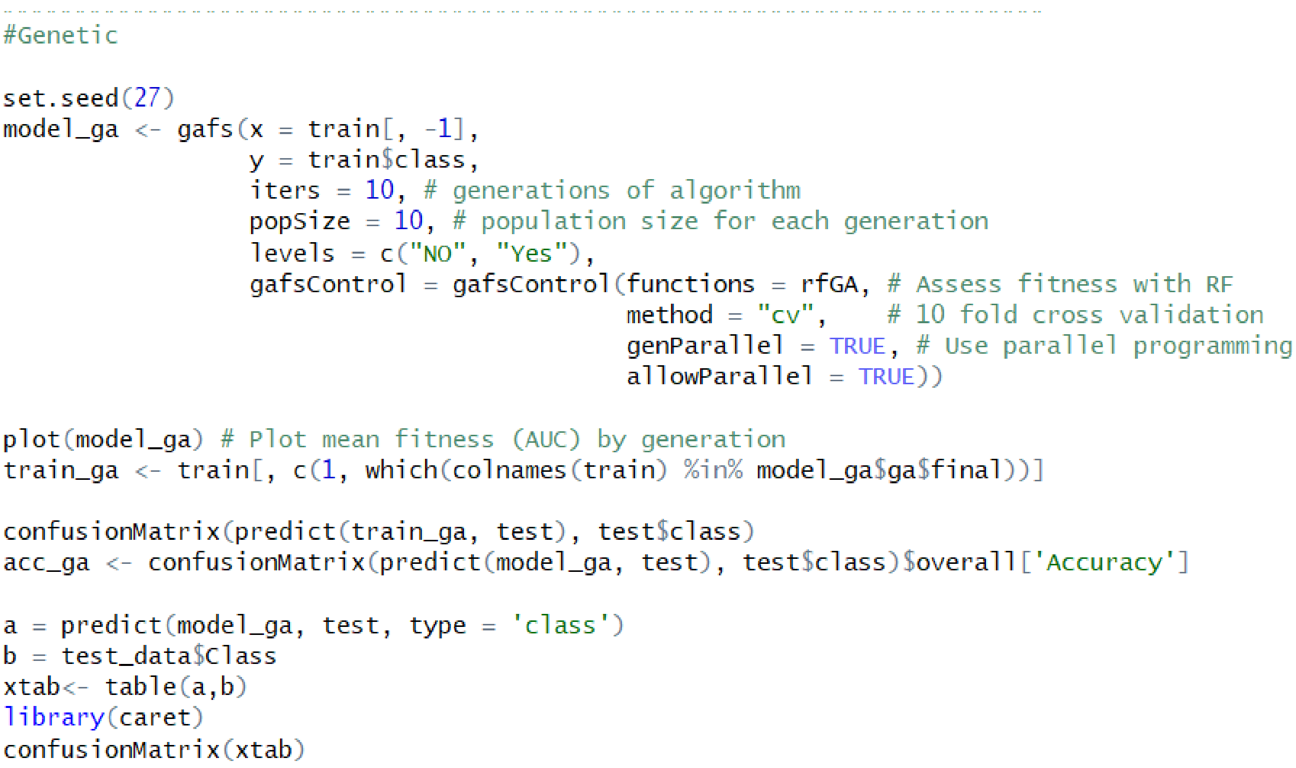
Fig 1.18 Extreme Gradient Boosting Code Snipet

Genetic Algorithm

Genetic algorithms are search based algorithms based on natural selection, recombination and mutation to evolve solutions to a problem.

First, we have a pool of different solutions in hand. These solutions undergo recombination and mutation, and in process produces new solutions. Each solution is given a fitness score accordingly.

Crossover is performed by randomly selecting solutions to produce offspring on the basis of fitness score. The fitter the solutions are, the fitter is their offspring. The process is repeated till a new batch of fitter solutions are created.



```
#Genetic

set.seed(27)
model_ga <- gafs(x = train[, -1],
                 y = train$class,
                 iters = 10, # generations of algorithm
                 popSize = 10, # population size for each generation
                 levels = c("NO", "Yes"),
                 gafsControl = gafsControl(functions = rfGA, # Assess fitness with RF
                                          method = "cv",    # 10 fold cross validation
                                          genParallel = TRUE, # Use parallel programming
                                          allowParallel = TRUE))

plot(model_ga) # Plot mean fitness (AUC) by generation
train_ga <- train[, c(1, which(colnames(train) %in% model_ga$ga$final))]

confusionMatrix(predict(train_ga, test), test$class)
acc_ga <- confusionMatrix(predict(model_ga, test), test$class)$overall['Accuracy']

a = predict(model_ga, test, type = 'class')
b = test_data$Class
xtab<- table(a,b)
library(caret)
confusionMatrix(xtab)
```

Fig 1.19 Genetics Algorithm Code Snipet

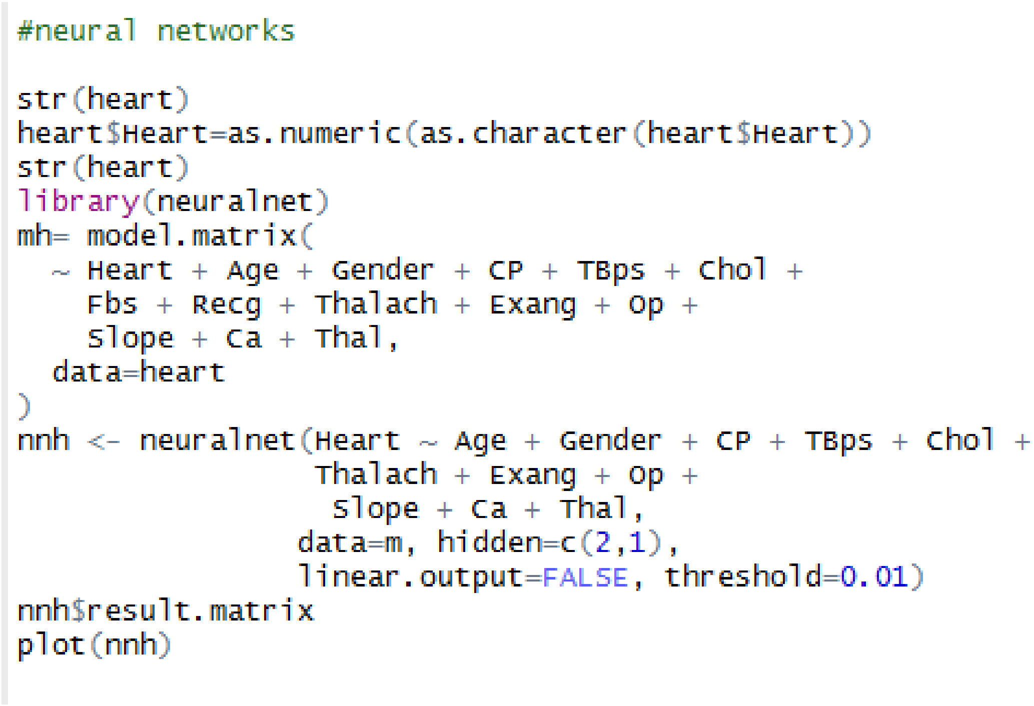
Neural Networks

Artificial Neural Networks work similar to that of a Human Brain. It has the ability to learn on its own from the dataset, so as to make accurate and precise prediction and classification.

Neural Networks have hidden layer in order to increase the productivity and its learning rate. In the first level, input is taken from the parameters itself to compute prediction. This level forms the first hidden layer for Neural Networks.

It then learns from the output about the error rate and recursively compute other predictions to give us a final answer, which has minimum error rate or high accuracy percentage.

The more the number of hidden layer, the more efficiently the model could learn on its own. Thus Neural Network is one of the finest and commonly used Machine Learning algorithms, used for unfeigned classification and prediction.



```
#neural networks

str(heart)
heart$Heart=as.numeric(as.character(heart$Heart))
str(heart)
library(neuralnet)
mh= model.matrix(
  ~ Heart + Age + Gender + CP + TBps + Chol +
    Fbs + Recg + Thalach + Exang + Op +
    Slope + Ca + Thal,
  data=heart
)
nnh <- neuralnet(Heart ~ Age + Gender + CP + TBps + Chol +
                 Thalach + Exang + Op +
                  Slope + Ca + Thal,
                data=m, hidden=c(2,1),
                linear.output=FALSE, threshold=0.01)
nnh$result.matrix
plot(nnh)
```

Fig 1.20 Neural Network Code Snipet

Comparison Models

After applying seven algorithms i.e. Linear Regression, Baeysian Regression, Decision Tree, Random Forest, Extreme Gradient Boosting, Genetic Algorithm and Neural Networks. on all the 5 datasets and comparing their accuracies we came to a conclusion that Extreme Gradient Boosting is giving the best accuracy on all datasets. Hence we have proceeded using by extreme gradient boosting algorithm for predicting the outcomes. Other algorithms were giving varying results in different datasets, thus convincing us not to rely on them.

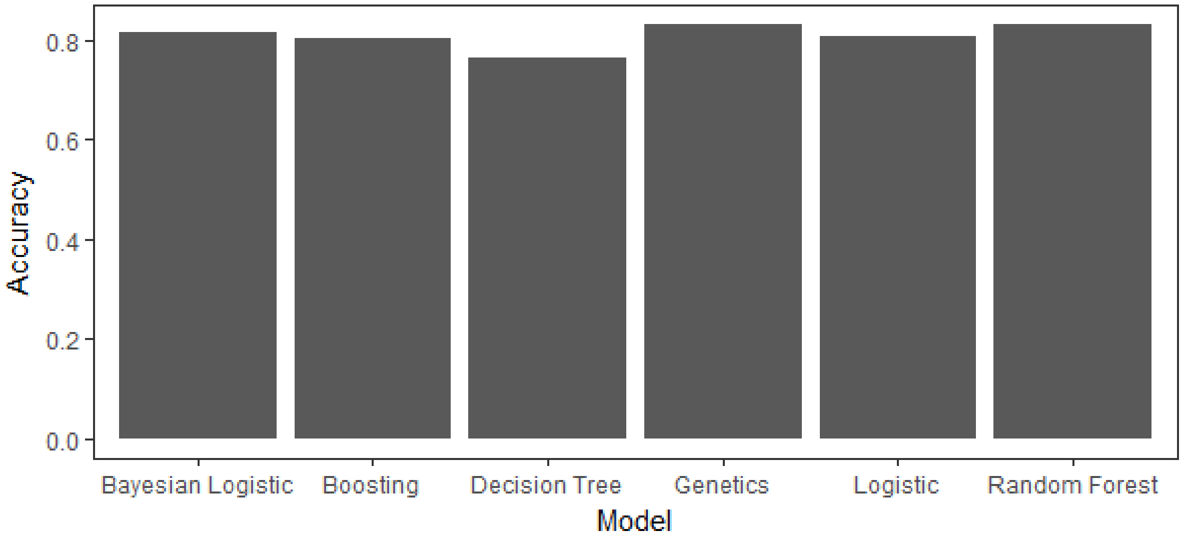


Fig 1.21 Accuracy Graph among different models for Heart disease

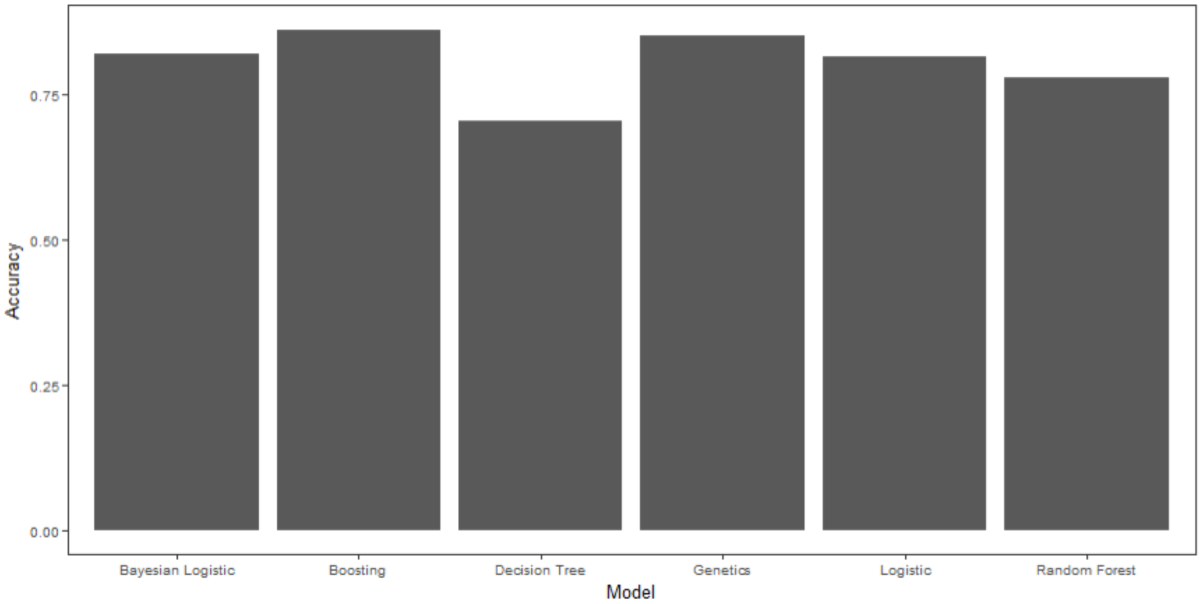


Fig 1.22 Accuracy Graph among different models for Diabetes

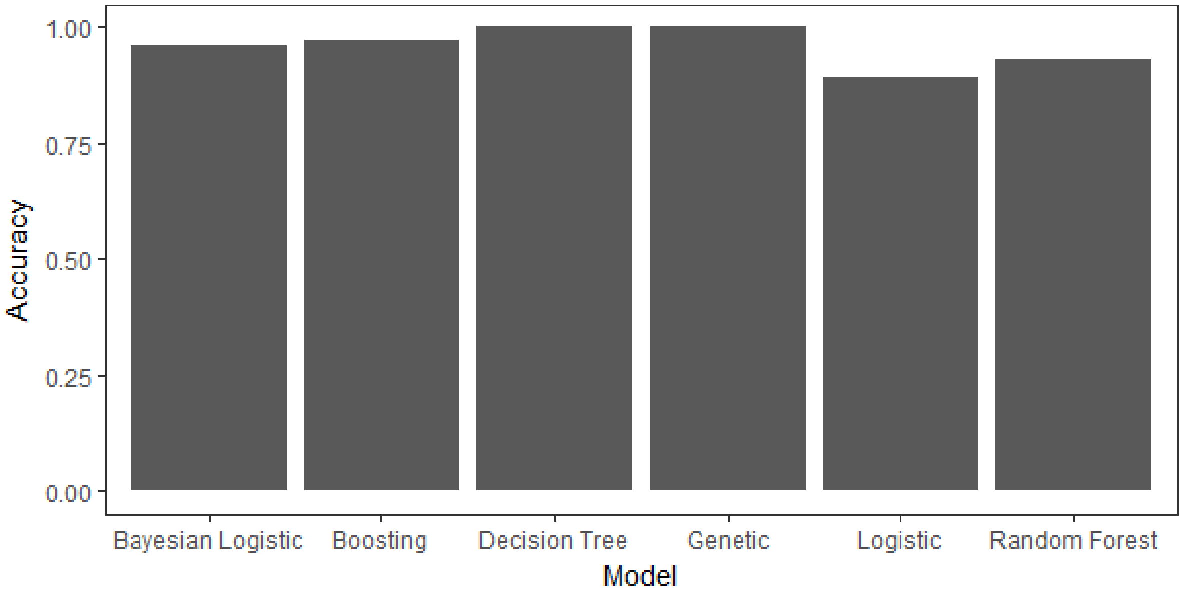


Fig 1.23 Accuracy Graph among different models for Breast cancer

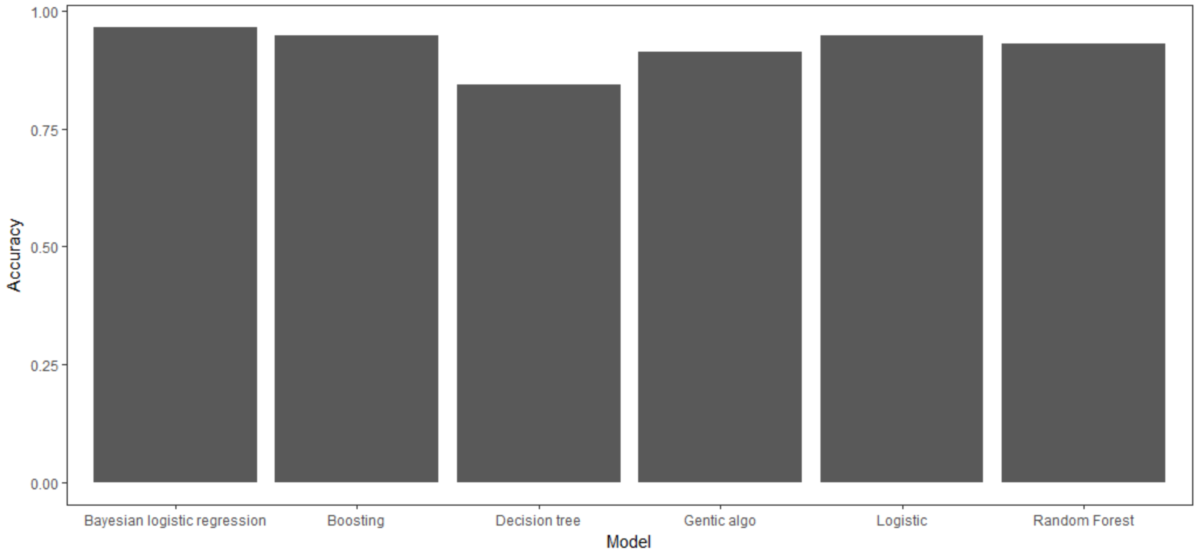


Fig 1.24 Accuracy Graph among different models for Parkinsons

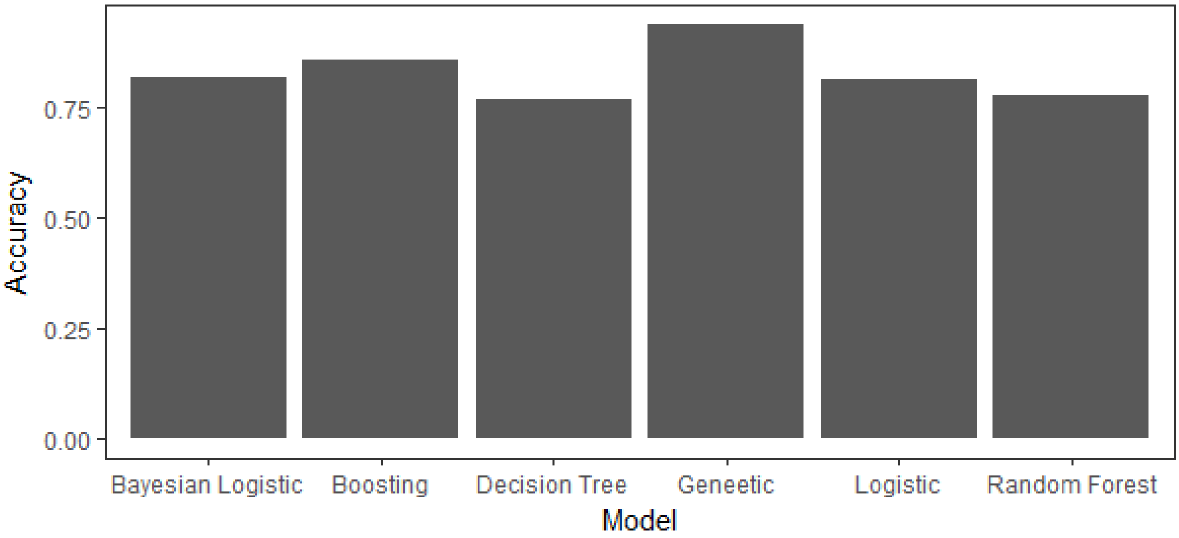
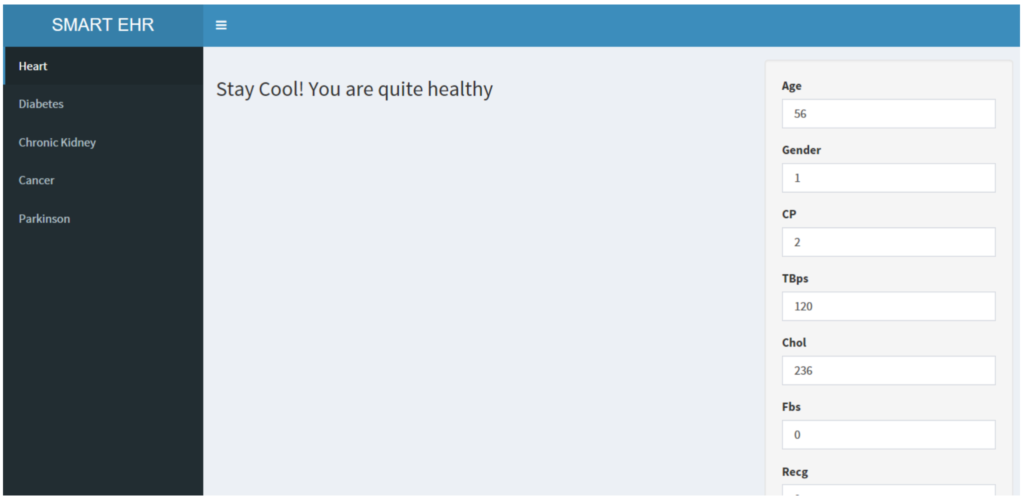


Fig 1.25 Accuracy Graph among different models for Kidney Disease

Extreme Gradient Boosting, though not giving hundred percent accuracy, but was reliable. It gave considerable results with around ninety percent accuracy for each disease dataset. Hence we chose Extreme Gradient Booster as the optimal algorithm for Smart Electronics Health Record.

WEB BASED APPLICATION

We incorporated a responsive web based frontend for easier and better experience for the users. They could easily enter the values of various parameters and see the output, thus taking steps accordingly.



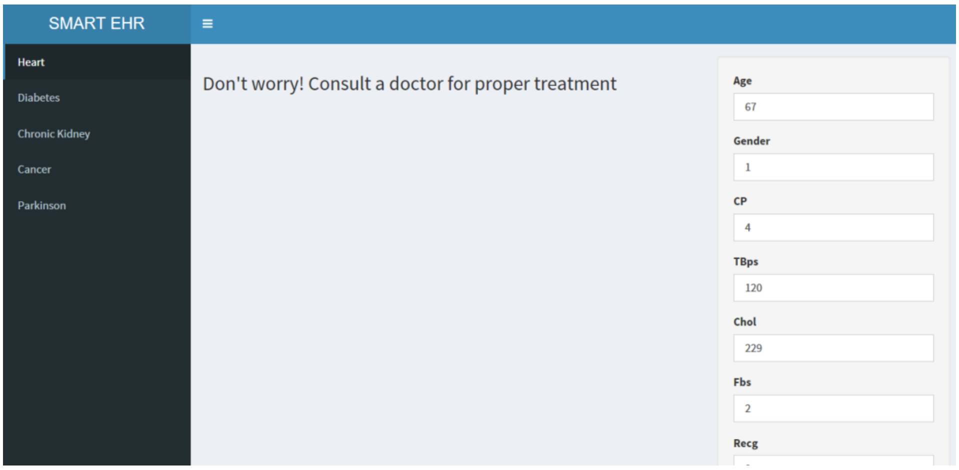
Stay Cool! You are quite healthy

Age: 56
Gender: 1
CP: 2
TBps: 120
Chol: 236
Fbs: 0
Recg

Fig 1.27 Responsive Web Application

The web based application was made with R Shiny. The UI is highly responsive i.e. changing value of the single parameter will enable the computation again.

Users have the choice to test any of the disease provided that are : Heart, Diabetes, Kidney, Parkinson and Cancer. They could easily shuffle through tabs to enter their parameters value, thus giving them an user friendly web experience.



Don't worry! Consult a doctor for proper treatment

Age: 67
Gender: 1
CP: 4
TBps: 120
Chol: 229
Fbs: 2
Recg

Fig 1.28 User Friendly and Easy to Use Website

REFERENCES

With the help and guidance of many people and online websites, we were able to achieve what we wished. Some of the references are:

[1] Health Information Systems. Understanding Health Care IT Alignment, (PMID:20938567), Bréant C, Yearbook of medical informatics [2010;:30-3]

[2] McMillan, S. S., Kendall, E., Sav, A., King, M. A., Whitty, J. A., Kelly, F., & Wheeler, A. J. (2013). Patient-centered approaches to health care: A systematic review of randomized controlled trials.

[3] Predicting breast cancer survivability: a comparison of three data mining methods, Author Dursun Delen, Glenn Walker, Amit Kadam, Department of Management Science and Information Systems, Oklahoma State University, 700 North Greenwood Venue, Tulsa, OK 74106, USA

[4] Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction, Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni, International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011

[5] Data mining techniques and applications – A decade review from 2000 to 2011, Sh HsienLiao, Pei-HuiChu, Pei-YuanHsiao, Department of Management Sciences, Tamkang University, No. 151, Yingzhuan Rd., Tamsui Dist., New Taipei City 25137, Taiwan, ROC

[6] Machine learning applications in cancer prognosis and prediction, Konstantina Kourou, Themis P.Exarchos′ , Konstantinos P.Exarchos, , Dimitrios I.Fotiadis, Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, Ioannina, Greece, IMBB — FORTH, Dept. of Biomedical Research, Ioannina, Greece, Molecular Oncology Unit, Department of Biological Chemistry, Medical School, University of Athens, Athens, Greece

[7] Imputation of missing data using machine learning techniques Kamakshi Lakshminarayan, Steven A. Harp, Robert Goldman and Tariq Samad 3660 Technology Drive Honeywell Technology Center Minneapolis, MN55418, USA

[8] Nikunj Agarwal, M.P.Sebastian, “Use of Cloud Computing and Smart Devices in Healthcare”, WASET, Intl. Jr. Computer, Electrical, Automation, Control and Information Engineering, Vol.10(1): 156-159, 2016.

[9] "Obesity In Australia Modi", Modi.monash.edu.au. N.p 2015, June 2015.

[10] Ya-Li Zheng, Xiao-Rong Ding, Carmen Chung Yan Poon, Unobtrusive Sensing and Wearable Devices for Health Informatics.

[11] Asha Rajkumar, G.Sophia Reena, Diagnosis Of Heart Disease Using Datamining Algorithm, Global Journal of Computer Science and Technology 38 Vol. 10 Issue 10 Ver. 1.0 September 2010.

[12] Sunita Soni, O.P.Vyas, Using Associative Classifiers for Predictive Analysis in Health Care Data Mining, International Journal of Computer Application (IJCA, 0975 – 8887) Volume 4– No.5, July 2010, pages 33-34.

[13]] W.Li, J. Han, J.Pei , CMAR- Classification based on Multiple Association Rules, ICDM‟01, , San Jose, CA, Nov. 2001. pp. 369-376

[14] Carloz Ordonez, Association Rule Discovery with Train and Test approach for heart disease prediction, IEEE Transactions on Information Technology in Biomedicine, Volume 10, No. 2, April 2006.pp 334-343.

[15] M.Y.C. Polley, B. Freidlin, E.L. Korn, B.A. Conley, J.S. Abrams, L.M. McShaneStatistical and practical considerations for clinical evaluation of predictive biomarkers, J Natl Cancer Inst, 105 (2013), pp. 1677-1683

[16] O. Fortunato, M. Boeri, C. Verri, D. Conte, M. Mensah, P. Suatoni, *et al.*Assessment of circulating microRNAs in plasma of lung cancer patients, Molecules, 19 (2014), pp. 3038-3054

[17] Rothman, Brian; Joan. C. Leonard; Michael. M. Vigoda (2012). "Future of electronic health records: implications for decision support". Mount Sinai Journal of Medicine 79 (6): 757-768.

[18] J. Sun, C. D. McNaughton, P. Zhang, A. Perer, A. Gkoulalas-Divanis,J. C. Denny, J. Kirby, T. Lasko, A. Saip, and B. A. Malin, “Predicting changes in hypertension control using electronic health records from a chronic disease management program,” J. Amer. Med. Informat. Assoc., vol. 21, pp. 337–344, 2014.

[19] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, “Big data in health care: Using analytics to identify and manage high risk and high-cost patients,” Health Affairs, vol. 33, pp. 1123–1131, 2014

[20] D. Yach, C. Hawkes, C. L. Gould, K. J. Hofman, "The global burden of chronic diseases: Overcoming impediments to prevention and control", JAMA, vol. 291, no. 21, pp. 2616-2622, 2004.

[21] B. Kayyali, D. Knott, S. Van Kuiken, "The big-data revolution in US health care: Accelerating value and innovation", McKinsey Company, pp. 1-13, 2013.

[22] Heart Disease Facts, 2015, [online] Available: http://www.cdc.gov/heartdisease/facts.htm

[23] Diabetes Globally, 2017, [online] Available: https://www.diabetesaustralia.com.au/diabetes-globally.

[24] D. A. Ludwick, J. Doucette, "Adopting electronic medical records in primary care: Lessons learned from health information systems implementation experience in seven countries", Int. J. Med. Inf., vol. 78, no. 1, pp. 22-31, 2009.

[25] J.Archenaa and E.A. Mary Anita, “A Survey Of Big Data Analytics in Healthcare and Government”, 2nd International Symposium on Big Data and Cloud Computing (ISBCC !"#$% &'()*+,-% ,. Computer Science 50, 408 – 413, 2015.

[26] MimohOjha, KirtiMathur, “Proposed Application of Big Data Analytics in Healthcare at Maharaja Yeshwantrao Hospital”, 2016 3rd MEC International Conference on Big Data and Smart City, IEEE, 2016.