**Maters Research Project**

**Prof Srikanth Mudigonda**

**Sathwik Nallamada**

**Introduction:**

I had finalized the topic “**Use of analytics in analyzing the data collected from health trackers connected by the Internet of Medical Things (IoMT)**”. The problem that I going to address is, specifically focused on measuring the heartbeat and beats per minute on a real-time basis by using health trackers connected via the Internet of Things (IoT). Throughout the MRP I will be focusing on building probability models and decision tree models that have a higher correlation with the MAX.HR(This parameter gives the maximum heart rate measured in a certain amount of time) to address the problem statement “How accurate is the prediction of human body complications, by measuring the real-time heartbeat and beats per minute data collected from the health trackers “ (or) “Can the diseases be predicted by analyzing the real-time heartbeat, beats per minute collected from the health data trackers connected via IoT”

The core idea of my project is measuring the heartbeat and beats per minute for the detection of diseases the research questions will be the same but there might be a change accordingly in analyzing the data generated from the health data trackers because there can new diseases predicted by using different analytical techniques with the use of the same predictor variables.

The data set I would be using is the transformative dataset consisting of various parameters, some are predictor variables while the rest are independent variables and categorical variables, the dataset will be transformed based on the metrics provided by the user like age, gender, weight, height, complications in the body and presence of the heart disease. The paraments it would be measuring are chest pain type (Categorical variable), Blood pressure (Numerical variable), Blood cholesterol level (Numerical variable), maximum reported heart rate (Numerical variable), number of vessels blocked (Categorical variable), and Presence of heart disease (Categorical and dichotomous).

The agenda for the MRP based on weekly is jotted down below:

|  |  |  |
| --- | --- | --- |
| **S No** | **Week** | **Plan of action** |
| 1 | Week-1 | The plan for week-1 is to analyze the heart prediction dataset using various functions like summary and str functions. |
| 2 | Week-2 | The plan for week-2 is to categorize the variables i.e predictor variables and outcome variables based on the core idea of the project and finalize the methodologies that aligns with my project idea. |
| 3 | Week-3 | For week-3 I would be focusing on the model building. |
| 4 | Week-4 | For week-4 I will be implementing the models based on the week-3 analysis and parallelly deliver the insights based on the outcomes. |
| 5 | Week-5 | The plan for week -5 is to analyze the results and next set plan of action based on the week-4 conclusion, |
| 6 | Week-6 | For week-6 I will be continuing the analysis based on the week-4 and week-5 insights and writing reflection based on the previous week analysis. |
| 7 | Week-7 | For week -7 I will be performing analysis based on the decision tree models and the pervious weeks analysis. |
| 8 | Week-8 | For the last week of the project,I will be performing the final steps of analysis based on the previous weeks results In addition to that, I will be summarizing the results and deliver conclusions. |

**WEEK-1(Dataset analysis)**

For the first week of the project, I will be analyzing the dataset using summary and str function.

The dataset I finalized for my master’s research project is sourced from KAGGLE. This dataset consists of 272 observations spread across 12 different parameters. The dataset is collected by the health tracking devices that are installed on 272 different people who are facing health related complications.

Here is the breakdown of the dataset I selected:

|  |  |  |  |
| --- | --- | --- | --- |
| SNO | Name | Description | Type |
| 1 | Age | Age of the participant (in years) | Numeric |
| 2 | Sex | Gender of the participant | Categorical |
| 3 | Blood Pressure | The value of the resting blood pressure | Numeric |
| 4 | Cholesterol | The serum cholesterol level in the body. Measured by mg/dl | Numeric |
| 5 | Chest Pain Type | 1-typical angina  2-Atypical angina  3-non-anginal pain  4-asymptotic | Categorical |
| 6 | FBS over 120 | Blood sugar level | Numeric |
| 7 | ST depression | The level of internal depression of the participant | Numeric |
| 8 | Number of vessels | The number of vessels in the heart narrowing down | Numeric |
| 9 | The allium | The allium heart related disease categorized by  3-normal  4-fined effect  7-reversible effect | Categorical |
| 10 | Max.HR | The maximum human heart rate | Numeric |
| 11 | Heart disease | The presence or absence of heart disease. | Categorical |

Here is the summary of the dataset:

Table

Description automatically generated

**WEEK-2(Categorization)**

For Week-2 I will be focusing on categorizing the variables that aligns to my area of interest:

Here is the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SNO | **Name** | **Description** | Type | Role |
| 1 | Age | Age of the participant (in years) | Numeric | Predictor |
| 2 | Sex | Gender of the participant | Categorical | Predictor |
| 3 | Blood Pressure | The value of the resting blood pressure | Numeric | Predictor |
| 4 | Cholesterol | The serum cholesterol level in the body. Measured by mg/dl | Numeric | Predictor |
| 5 | Chest Pain Type | 1-typical angina  2-Atypical angina  3-non-anginal pain  4-asymptotic | Categorical | Outcome |
| 6 | FBS over 120 | Blood sugar level | Numeric | Outcome |
| 7 | ST depression | The level of internal depression of the participant | Numeric | Outcome |
| 8 | Number of vessels | The number of vessels in the heart narrowing down | Numeric | Outcome |
| 9 | The allium | The allium heart related disease categorized by  3-normal  4-fined effect  7-reversible effect | Categorical | Outcome |
| 10 | Max.HR | The maximum human heart rate | Numeric | Predictor |
| 11 | Heart disease | The presence or absence of heart disease. | Categorical | Outcome |

**WEEK-3 (Model Building)**

For week-3 I will be mainly focusing on the model buildings. Since the core idea of the project is to find the presence of heart disease based on MAX.HR. I will be building models to analyze the different sets of parameters that have more statistical significance with the MAX.HR.

In the process of moving forward I will be analyzing three different models:

1. The first model contains all the parameters present in the table.
2. The second model contains the elements that I am interested in examining with the heart rate.
3. The third model contains the most accurate parameters by examining the above two models.

The prior probability density functions for all the three models are the range of the heart rate. The range of the Max.H.R can be determined by using the summary function on the dataset I selected .

* Beta (Max.HR ~all\_predictor variables) ∼ N(m0,s.d^2)
* Beta (Max.HR~Age+BP+FBS.over.120+Cholesterol+Sex) ∼ N (m1, s.d^2)
* Beta (Max.HR~Chest.pain.type+FBS.over.120+Slope.of.ST+Number.of. vessels.fluro) ∼ N(m3, s.d^2)

I will be examining the data model that contains all the parameters present in the table. For the prior intercept, I am considering the range of the heartbeat. In order to gather data on the range of the heartbeat, I analyzed the summary function. Based on this analysis, I selected a prior intercept of (71,1, Autoscale=True) because the heartbeat ranges from 71 to 202. After considering the summary values of the MAX.HR, I plugged in the prior intercept as (71,65.5), with a mean of 71 and a standard deviation of 65.5. Due to the fact that the heart rate cannot be negative, the standard deviation is set to a value of 1, and the auto scale is set to TRUE.

**WEEK-4 (Analysis)**

For week-4 I had examined 3 different probability models that are stated below:

Model -1

model\_1 <- stan\_glm(

Max.HR~.,

data = dataset1, family = gaussian,

prior\_intercept = normal(71,1,autoscale = TRUE),

prior = normal(0, 1, autoscale = TRUE),

prior\_aux = exponential(1, autoscale = TRUE),

chains = 4, iter = 5000\*2, seed = 84735,prior\_PD = TRUE)

Model-2

model\_2 <- stan\_glm(

Max.HR~Age+BP+FBS.over.120+Chest.pain.type+Sex,

data = dataset1, family = gaussian,

prior\_intercept = normal(71,1,autoscale = TRUE),

prior = normal(0, 1, autoscale = TRUE),

prior\_aux = exponential(1, autoscale = TRUE),

chains = 4, iter = 5000\*2, seed = 84735,prior\_PD = TRUE)

Model-3

dataset3=dataset%>%select(Chest.pain.type,FBS.over.120,Slope.of.ST,Number.of.vessels.fluro,Max.HR)

model\_3 <- stan\_glm(

Max.HR~Chest.pain.type+FBS.over.120+Slope.of.ST+Number.of.vessels.fluro,

data = dataset3, family = gaussian,

prior\_intercept = normal(71,65.5),

prior = normal(0, 1, autoscale = TRUE),

prior\_aux = exponential(1, autoscale = TRUE),

chains = 4, iter = 5000\*2, seed = 84735,prior\_PD = TRUE)

**Week-5(Reflections)**

Reflections of the analysis based on the week 4:

By performing the above analysis, we can conclude that that human heart rate solely be accountable for predicting the heart complications. But by analyzing the heart rate there will be some useful insight that can be obtained. According to my data set that I considered. There are only limited parameter values that can be analyzed by real time human heartbeat analysis. The predictor values that can be analyzed and having the most statistical significance for the above solution are:

1)Chest Pain type

2)FBS over 120

3)Slope of ST

4)Number of vessels of Fluro.

cv\_procedure$cv

## mae mae\_scaled within\_50 within\_95

## 1 82.19276 0.9512328 0.2 1

set.seed(84735)

prediction\_summary(model = model\_1, data = dataset)

## mae mae\_scaled within\_50 within\_95

## 1 82.53469 1.057655 0.1925926 1

set.seed(84735)

prediction\_summary(model = model\_2, data = dataset)

## mae mae\_scaled within\_50 within\_95

## 1 82.11212 1.276942 0.08888889 0.962963

set.seed(84735)

prediction\_summary(model = model\_3, data = dataset)

## mae mae\_scaled within\_50 within\_95

## 1 81.95622 0.9551879 0.2037037 1

Based on the prediction summary of all the three models within 50% of the confidence interval the model 3 is considered most accurate followed by the model -1 and model -2.

***For further analysis, I will be using decision tree models to analyze the type of chest pain that is closely associated with heart disease and MAX.HR. By identifying the chest pain type that is linked to these parameters, we can focus on it for deeper analysis and filter out the others. I will calculate the var\_imp across three different decision tree models including bagged tree, random forest, and regression tree models. This analysis will help us gain insights into the types of chest pain that are crucial in detecting heart disease and eliminating other types****.*

**WEEK-6(Decision tree models Implementation)**

For week-6 I am going to address

The problem statement that I am going to analyze is “Can the chest pain type be predicted with the help of different predictor variables that are closely associated.”

The questions that I am trying to address through this MRP are jotted down below:

1)What are the predictor variables closely associated with the chest pain type predictor variable or is there any way can the type of chest pain type be predicted by the use of other different predictor variables that are closely associated?

2)Can the chest pain type be predicted with the help of different predictor variables which are closely associated?

*The explanation of the predictor variables is given in the table below:*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SNO** | **Names** | **Description** | **Measurement type** | **Role** |
| 1 | Age | Age of the participant (in years) | Numeric | Predictor |
| 2 | Sex | Gender of the participant | Categorical | Predictor |
| 3 | Blood Pressure | The value of the resting blood pressure | Numeric | Predictor |
| 4 | Cholesterol | The serum cholesterol level in the body. Measured by mg/dl | Numeric | Predictor |
| 5 | Chest Pain Type | 1-typical angina  2-Atypical angina  3-non-anginal pain  4-asymptotic | Categorical | Outcome |
| 6 | FBS over 120 | Blood sugar level | Numeric | Predictor |
| 7 | ST depression | The level of internal depression of the participant | Numeric | Predictor |
| 8 | Number of Vessels | The number of vessels in the heart narrowing down | Categorical | Predictor |
| 9 | Thallium | The allium heart related disease categorized by  3-normal  4-fined effect  7-reversible effect | Categorical | Predictor |
| 10 | MAX.HR | The maximum human heart rate | Numeric | Outcome |
| 11 | Exercise Angina | The chest discomfort or pain that occurs when the muscles don't get oxygen rich blood | Categorical | Predictor |
| 12 | EKG Results | The Irregularity in heart beat | Numeric | Predictor |
| 13 | Slope of ST | The value of change in ST Segments. | Numeric | Predictor |
| 14 | Heart Disease | The presence or absence of heart disease. | Categorical | Outcome |

Along this project I will be performing the different sets of analysis that align with my MRP:

|  |  |  |  |
| --- | --- | --- | --- |
| **SNO** | **METHOD** | **EXPLANATION** | **RATIONALE OF USING THE METHOD** |
| 1 | LASSO AND RIDGE | LASSO REGRESSION-Lasso regression is used for future selection and regularization. Lasso regression is used to make the least important predictors to zero thus helps in reducing over fitting the model  RIDGE REGRESSION-Ridge regression is form of linear regression, but the difference is ridge regression sets all the predictor variables to zero and performs the operation. | I would be using lasso regression to determine the most important predictor variables in predicting that are aligned to the chest pain type, parallelly shrinking the coefficients of least important variables towards zero this can prevent from over fitting the model.  I would be using ridge regression to access the importance of the variables with the chest pain type, while taking into account of the multicollinearity. |
| 2 | BOOSTED TREE | The boosted tree model is the regression model of the bagged tree model. It helps to improve accuracy of the of the prediction with the combination of the multiple decision tress. In addition of that it also helps to capture nonlinear relationships between the predictor variables which lacks in bagged tree model. | I will be using boosted tree model to access the nonlinear relationships between the other predictor variables with the chest pain type variable. In the process I would be using the it to access the feature importance scores of the other predictor variables. |
| 3 | SVM REGRESSION | SVM regression is used to predict the continuous variables and categorical variables. It works by finding the distance between the predictor variables by using hyper plane and the support vectors. | In context for finding the chest pain type I would be using the SVM regression for finding the non-Linear relationship with the chest pain score. |
| 4 | K-means clustering | K-means clustering is partitions the data set into K-clusters based on the similarity of the data points. K means clustering identify the group of patients associated with the different types of chest pain type. It also helps in iteratively assigning data points to the nearest cluster and updating the cluster based on the average of the points of each cluster. | I will be using K means clustering to gain insights about the patterns of the predictor variables that are associated with the different types of chest pain type. K Means clustering can be also used to identify the data about the group of patients with the similar predictor variables. |

**Week-7(Analysis of the decision tree models)**

|  |  |  |  |
| --- | --- | --- | --- |
| **MODEL NAME** | **CROSS VALIDATION** | **MODEL FORMULA** | **EXPLANATION** |
| LOGISTIC REGRESSION | fitControl <- trainControl(  method = "repeatedcv", ## perform repeated k-fold CV  number = 5,  repeats = 1,  classProbs = TRUE) | glmmodel.fit <- glm(Chest.pain.type ~ ., family = binomial, data = train.data) | Logistic regression is mainly works well when the outcome variable is categorical in nature .As we can see in my case the outcome variable is chest pain type and logistic regression precisely works for my data set.As we can see above it correctly predicts the variable importance that allies with the chest pain type . |
| LASSO REGRESSION | fitControl <- trainControl(  method = "repeatedcv", ## perform repeated k-fold CV  number = 5,  repeats = 1,  classProbs = TRUE) |  | LASSO regression helps the model to not over fit by making the least important predictors to zero.In the above scenario ST depression,EKG results and age is set to zero. |
| RIDGE REGRESSION | fitControl <- trainControl(  method = "repeatedcv", ## perform repeated k-fold CV  number = 5,  repeats = 1,  classProbs = TRUE) |  | Ridge regression is form of linear regression but the difference is ridge regression sets all the predictor variables to zero and perform the operation. For my example data set number of vessels,thallium,heart disease presence are positively related with the chest pain type. |
| BOOSTED TREE MODEL | fitControl <- trainControl(  method = "repeatedcv", ## perform repeated k-fold CV  number = 5,  repeats = 1,  classProbs = TRUE) | boostedfit <- train(Chest.pain.type ~ .,  data = trainTransformed,  method = "gbm",  trControl = fitControl,  verbose = FALSE,  tuneGrid = grid) | Boosted tree model is the regression model of the bagged tree model .It helps to improve accuracy of the of the prediction with the combination of the multiple decision tress.In addition of that it also helps to capture non linear relationships between the predictor variables which lacks in bagged tree model.In the above scenario since it is related to medical based data sensitivity is considered pivotal among all and the boosted tree model performed the best among the rest. |
| K MEANS CLUSTERING | fitControl <- trainControl(  method = "repeatedcv", ## perform repeated k-fold CV  number = 5,  repeats = 1,  classProbs = TRUE) | |  | | --- | | wss <- (nrow(scaled\_data)-1)\*sum(apply(scaled\_data,2,var))  for (i in 2:10) wss[i] <- sum(kmeans(scaled\_data, centers=i)$withinss)  plot(1:10, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares") | | K-means clustering is partitions the data set into K-clusters based on the similarity of the data points.K means clustering identify the group of patients associated with the different types of chest pain type.It also helps in iteratively assigning data points to the nearest cluster and updating the cluster based on the average of the points of each cluster. |
| SVM LINEAR FIT | fitControl <- trainControl(  method = "repeatedcv", ## perform repeated k-fold CV  number = 5,  repeats = 1,  classProbs = TRUE) | svmlinearfit <- train(Chest.pain.type ~ .,  data = trainTransformed,  method = "svmLinear",  trControl = fitControl,  verbose = FALSE,  tuneGrid = grid) | SVM regression is used to predict the continuous variables and categorical variables. It works by finding the distance between the predictor variables by using hyper plane and the support vectors. |

**Results:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **BAGGED TREE** | **RANDOM FOREST** | **BOOSTED TREE** | **K MEANS CLUSTERING** | **SVM LINEAR REGRESSION** |
| ACCURACY | 0.3462 | 0.4038 | 0.4038 | - | 0.5192 |
| CONFIDENCE INTERVAL | (0.2197,0.4909) | (0.2701,0.549) | (0.2701,0.549) | - | (0.3763,0.6599) |
| NO INFORMATION RATE | 0.4808 | 0.4038 | 0.4808 | - | 0.4808 |
| P VALUE (Acc>NIR) | 0.9821 | 0.8946 | 0.8946 | - | 0.3382 |
| SENSITIVITY | TA-0.000  ATA-0.000  NAP-0.333  ASY-0.5200 | |  | | --- | | TA-0.000  ATA-0.000  NAP-0.26667  ASY-0.6880 | | |  | | --- | | TA-0.000  ATA-0.125  NAP-0.0266  ASY-0.6400 | | - | |  | | --- | | TA-0.000  ATA-0.000  NAP-0.5333  ASY-0.7600 | |
| SPECIFICITY | |  | | --- | | TA-0.9710  ATA-0.8636  NAP-0.51351  ASY-0.6667 | | |  | | --- | | TA-1.00  ATA-0.8636  NAP-0.648  ASY-0.556 | | |  | | --- | | TA-0.916  ATA-0.79545  NAP-0.783  ASY-0.6296 | | - | |  | | --- | | TA-1.000  ATA-1.000  NAP-0.6757  ASY-0.5185 | |
| MCNEMAR, s P VALUE | 0.3462 | NA | 0.4074 | - | NA |
| kappa | -0.0063 | 0.0382 | 0.0934 | - | 0.1895 |

Based on the results generated we can conclude that none of the models generated are ideal based on the P-Value which is calculated with respect to an accuracy greater than no information have significantly higher values than for all the three models i.e for the bagged tree(0.9821), random forest (0.8946) and boosted tree (0.8946) and SVM linear fit(0.3382). Among the four models examined the SVM linear fit has a higher accuracy than the rest. But by the results, we can glean that the ASY chest pain type has higher sensitivity values among the rest of the chest pain types.

Since my data set is mainly related to the medical field, Sensitivity is considered the most accurate among the remaining predictor variables (Quave, C. L., Pardo-de-Santayana, M., & Pieroni, A. (2012). The model which can detect three different types of chest pains but with less amount of accuracy is boosted tree model followed by the SVM linear regression.

Based on the above values of sensitivity ASY -Chest pain type can be predicted more accurately among the rest of the chest pain types. The goal of my project is to find the type of chest pain that is closely associated with the remaining paraments and primarily with MAX.HR and heart disease. So, I will be looking at the sensitivity values among all the four types of chest pain. By analyzing the above results, we can conclude that ASY chest pain type is closely associated with the remaining paraments and MAX.HR and Heart disease. In addition to that I will be assuring the variable importance across all the three models to access the association with the chest pain type:

Below are the tables that are generated by var\_imp function across three models:

Bagged tree:

varImp(bagfit)

## rf variable importance

##

## Overall

## Max.HR 100.000

## Cholesterol 99.612

## BP 98.372

## Age 95.841

## Exercise.anginayes 73.283

## ST.depression 65.385

## EKG.resultsventricular.hypertrophy 19.520

## Number.of.vessels.fluromore.than.50.pct 15.266

## Thalliumreversible.defect 14.555

## SexM 14.030

## Slope.of.STflat 13.777

## FBS.over.120true 11.559

## Thalliumfixed.defect 6.401

## Slope.of.STdownsloping 3.370

## EKG.resultswave.abnormality 0.000

Random Forest:

varImp(forestfit)

## rf variable importance

##

## Overall

## Max.HR 100.000

## Cholesterol 93.205

## Age 89.834

## BP 84.069

## ST.depression 67.901

## Exercise.anginayes 52.075

## EKG.resultsventricular.hypertrophy 21.903

## Thalliumreversible.defect 21.354

## Number.of.vessels.fluromore.than.50.pct 20.074

## SexM 17.847

## Slope.of.STflat 17.637

## FBS.over.120true 14.622

## Thalliumfixed.defect 5.719

## Slope.of.STdownsloping 4.508

## EKG.resultswave.abnormality 0.000

Boosted Tree:

*## Rank the variables in terms of their importance*

varImp(boostedfit)

## gbm variable importance

##

## Overall

## Max.HR 100.00000

## Cholesterol 91.61395

## Age 75.59574

## BP 70.73654

## ST.depression 37.43820

## Exercise.anginayes 18.18513

## Thalliumreversible.defect 9.01262

## SexM 7.14636

## EKG.resultsventricular.hypertrophy 6.67734

## FBS.over.120true 6.26255

## Number.of.vessels.fluromore.than.50.pct 4.57975

## Slope.of.STflat 3.30890

## Thalliumfixed.defect 0.38169

## Slope.of.STdownsloping 0.04615

## EKG.resultswave.abnormality 0.00000

Based on the three models, MAX.HR and Cholesterol levels are closely associated with the chest pain type.

To conclude the MRP in next week I will be determining the type of association with MAX.HR and filtering ASY chest pain type with the rest and performing logistic regression with the ASY chest pain type and MAX.HR to determine the heart disease.

**Week-8(Logistic Regression Analysis)**

I performed the logistic regression analysis based on the outcomes of the week-4 where I performed Bayesian analysis and week-7 where I performed decision tree analysis based on the outcome of week-4.

I had performed logistic regression analysis across 4 different models:

**Model-1:**

* **HeartDisease\_Presence ~ Max.HR+ASY\_ChestPain**

Heart disease is the dependent variable and the MAX.HR (Maximum heart rate) and ASY of the chest pain type are the independent variables. By performing the logistic regression for the first model the results generated are:

Call:

glm(formula = HeartDisease\_Presence ~ Max.HR + ASY\_ChestPain,

family = binomial, data = dsobject)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.1548 -0.7057 -0.4659 0.8019 2.3674

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 3.840407 1.116991 3.438 0.000586 \*\*\*

Max.HR -0.033744 0.007261 -4.647 3.36e-06 \*\*\*

ASY\_ChestPain 1.920963 0.298847 6.428 1.29e-10 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 370.96 on 269 degrees of freedom

Residual deviance: 275.30 on 267 degrees of freedom

AIC: 281.3

Number of Fisher Scoring iterations: 4

From the above analysis we can glean that for every one unit increase of the maximum heart rate the predictability decreases by 0.033. But on the contrary side the person with the 'ASY' chest pain type predicted probability of the of the heart disease is 1.920.

**Model-2:**

* **HeartDisease\_Presence ~ Max.HR**

Heart disease is the dependent variable and the MAX.HR (Maximum heart rate) is the independent variable. By performing the logistic regression for the second model the results generated are:

Call:

glm(formula = HeartDisease\_Presence ~ Max.HR, family = binomial,

data = dsobject)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.0858 -0.9208 -0.6243 1.0265 2.1378

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 6.159089 1.020518 6.035 1.59e-09 \*\*\*

Max.HR -0.042753 0.006767 -6.318 2.64e-10 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 370.96 on 269 degrees of freedom

Residual deviance: 320.04 on 268 degrees of freedom

AIC: 324.04

Number of Fisher Scoring iterations: 4

From the above analysis we can glean that for every one unit increase of the maximum heart rate the predictability decreases by 0.04. But the model-1 the Max.HR predictability decreased by the value of 0.03 which is 0.1 lower than the second model. On the contrary side AIC has increased when compared to the prior model which denotes that the 2nd model is not the better fit model when compared to the previous model.

**Model -3**

* **HeartDisease\_Presence ~ ASY\_ChestPain**

Heart disease is the dependent variable and ASy\_chest\_pain i is the independent variable. By performing the logistic regression for the third model the results generated are:

Call:

glm(formula = HeartDisease\_Presence ~ ASY\_ChestPain, family = binomial,

data = dsobject)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.5635 -0.6786 -0.6786 0.8354 1.7785

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.3512 0.2084 -6.485 8.87e-11 \*\*\*

ASY\_ChestPain 2.2245 0.2841 7.830 4.89e-15 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 370.96 on 269 degrees of freedom

Residual deviance: 299.70 on 268 degrees of freedom

AIC: 303.7

Number of Fisher Scoring iterations: 4

The Asy\_type of chest pain type shows the predicted probability of heart disease is 2. 2245.The AIC value is marginally higher than the model-1 which suggests that model -1 is better fit than model-2 and model -3.

* **Model-4**

**HeartDisease\_Presence~ .(All the predictor variables)**

Heart disease is the dependent variable and all the remaining parameters in the dataset are independent variables. By performing the logistic regression for the first model the results generated are:

Call:

glm(formula = HeartDisease\_Presence ~ ., family = binomial, data = dsobject2)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.6123 -0.5120 -0.1739 0.4023 2.4210

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.446417 3.088097 -2.735 0.00624 \*\*

Age -0.017477 0.025720 -0.680 0.49681

Sex 1.542109 0.540762 2.852 0.00435 \*\*

Chest.pain.type 0.700895 0.215280 3.256 0.00113 \*\*

BP 0.025216 0.011450 2.202 0.02765 \*

Cholesterol 0.007228 0.004077 1.773 0.07628 .

FBS.over.120 -0.794811 0.574662 -1.383 0.16664

EKG.results 0.301668 0.197838 1.525 0.12730

Max.HR -0.021045 0.010579 -1.989 0.04666 \*

Exercise.angina 0.829386 0.431091 1.924 0.05436 .

ST.depression 0.343690 0.227068 1.514 0.13013

Slope.of.ST 0.442276 0.391077 1.131 0.25809

Number.of.vessels.fluro 1.165271 0.269283 4.327 1.51e-05 \*\*\*

Thallium 0.341384 0.106066 3.219 0.00129 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 370.96 on 269 degrees of freedom

Residual deviance: 179.60 on 256 degrees of freedom

AIC: 207.6

Number of Fisher Scoring iterations: 6

The logistic regression is performed across all the parameters present in the dataset. Heart disease is the dependent variable, and the remaining predictors are the independent variables. The AIC value is significantly lower than the other three models. By this we can glean that model -4 is the better fit model when compared with the other three models.

**Future analysis:**

* In future I will be analyzing different models to increase the predictability
* I will be implementing a visualization representation of the results for better understanding.
* Parallelly I will be focusing on the detection of the type of heart disease and suggest a diagnosis based on the results generated.

References:

Quave, C. L., Pardo-de-Santayana, M., & Pieroni, A. (2012). Medical ethnobotany in Europe: from field ethnography to a more culturally sensitive evidence-based cam?. *Evidence-based complementary and alternative medicine*, *2012*.

https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction