/\*INFILE method to get the dataset into the SAS application\*/

**data** heart;

title "Heart Failure Prediction dataset";

infile "heart.csv" delimiter = ',' missover firstobs=**2**;

input Age Sex $ ChestPainType $ RestingBP Cholesterol FastingBS RestingECG $ MaxHR ExerciseAngina $ Oldpeak ST\_Slope $

HeartDisease;

/\*Sex\*/

Sex\_Male=(Sex='M');

/\*ChestPainType\*/

CP\_ASY = (ChestPainType = 'ASY');

CP\_ATA = (ChestPainType = 'ATA');

CP\_NAP = (ChestPainType = 'NAP');

/\*RestingECG\*/

ECG\_LVH = (RestingECG = 'LVH');

ECG\_Normal = (RestingECG = 'Normal');

/\*ExerciseAngina\*/

Angina\_Yes = (ExerciseAngina = 'Y');

/\*ST\_Slope\*/

Slope\_Down = (ST\_Slope = 'Down');

Slope\_Flat = (ST\_Slope = 'Flat');

**run**;

**proc** **print** data=heart;

**run**;

/\*checking for variable data types\*/

**proc** **contents** data=heart;

**run**;

/\*checking for frequency data types\*/

**proc** **freq** data = heart;

tables HeartDisease;

**run**;

/\*add labels to variable names\*/

**data** heart;

set heart;

label Age="age of the patient"

Sex\_Male="sex of the patient [1: male]"

CP\_ASY = "chest pain type [1: Asymptomatic]"

CP\_ATA = "chest pain type [1: Atypical Angina]"

CP\_NAP = "chest pain type [1: Non-Anginal Pain]"

RestingBP= "resting blood pressure"

Cholesterol= "serum cholesterol"

FastingBS= "fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]"

ECG\_LVH= "esting electrocardiogram results [1: LVH]"

ECG\_Normal= "esting electrocardiogram results [1: Normal]"

MaxHR= "maximum heart rate achieved"

Angina\_Yes= "exercise-induced angina [1: Yes]"

Oldpeak= "oldpeak = ST [Numeric value measured in depression]"

Slope\_Down= "slope of the peak exercise ST segment [1: Down]"

Slope\_Flat= "slope of the peak exercise ST segment [1: flat]"

HeartDisease= "output class [1: heart disease, 0: Normal]";

/\*Histogram for age \*/

title "Histogram";

**proc** **univariate** normal;

var age;

histogram / normal(mu=est sigma=est); \*mu-> sample mean, sigma-> std;

**run**;

/\*Scatter Plot\*/

title "Scatterplot";

**proc** **sgscatter**;

matrix Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal MaxHR Angina\_Yes Oldpeak

Slope\_Down Slope\_Flat HeartDisease;

**run**;

/\*Correlation Matrix\*/

title "Pearson Correlation matrix";

**proc** **corr**;

var Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal MaxHR Angina\_Yes Oldpeak

Slope\_Down Slope\_Flat HeartDisease;

**run**;

/\*sort HeartDisease type Variable\*/

title "boxplot";

**proc** **sort**;

by HeartDisease;

**run**;

/\*Boxplot of HeartDisease vs Age,RestingBP,Cholesterol,MaxHR,Oldpeak variables\*/

**proc** **boxplot**;

plot Age\*HeartDisease; \*y-var\*x-var;

plot RestingBP\*HeartDisease; \*y-var\*x-var;

plot Cholesterol\*HeartDisease; \*y-var\*x-var;

plot MaxHR\*HeartDisease; \*y-var\*x-var;

plot Oldpeak\*HeartDisease; \*y-var\*x-var;

**run**;

/\*Full model - checking for standardized coefficients, collinearity, influential points and outliers\*/

**proc** **logistic** data=heart;

title "Full Model - Standardized coefficients(stb), Checking collinearity(corrb), influential points(influential obs.)

and outliers(iplots)";

model HeartDisease(event='1')= Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal

MaxHR Angina\_Yes Oldpeak Slope\_Down Slope\_Flat/stb corrb influence iplots rsquare;

**run**;

/\*delete observations from Pearson residual graph which are above +3 and below -3\*/

**data** newheart;

set heart;

if \_n\_ in (**18**,**57**,**115**,**191**,**197**,**199**,**221**,**232**,**235**,**236**,**242**,**248**,**290**,**295**,**389**,**591**,**781**,**809**,**810**,**817**,**830**,**845**,**895**,**903**) then delete;

**run**;

/\*Analyzing data after removing above outliers\*/

**proc** **logistic** data=newheart;

title "Analyzing outliers and influential points - 1";

model HeartDisease(event='1')= Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal

MaxHR Angina\_Yes Oldpeak Slope\_Down Slope\_Flat/stb corrb influence iplots;

**run**;

/\*delete observations from Pearson residual graph which are above +3 and below -3\*/

**data** newheart1;

set newheart;

if \_n\_ in (**36**,**199**,**215**,**244**,**349**,**381**,**451**,**779**,**782**,**879**,**894**) then delete;

**run**;

/\*Analyzing data after removing above outliers\*/

**proc** **logistic** data=newheart1;

title "Analyzing outliers and influential points - 2";

model HeartDisease(event='1')= Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal

MaxHR Angina\_Yes Oldpeak Slope\_Down Slope\_Flat/stb corrb influence iplots;

**run**;

/\*delete observations from Pearson residual graph which are above +3 and below -3\*/

**data** newheart2;

set newheart1;

if \_n\_ in (**81**,**121**,**189**,**195**,**198**,**203**,**236**,**243**,**288**,**866**) then delete;

**run**;

/\*Analyzing data after removing above outliers\*/

**proc** **logistic** data=newheart2;

title "Analyzing outliers and influential points -3";

model HeartDisease(event='1')= Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal

MaxHR Angina\_Yes Oldpeak Slope\_Down Slope\_Flat/stb corrb influence iplots;

**run**;

/\*delete observations from Pearson residual graph which are above +3 and below -3\*/

**data** newheart3;

set newheart2;

if \_n\_ in (**30**,**275**,**334**,**426**,**829**,**853**) then delete;

**run**;

/\*Analyzing data after removing above outliers\*/

**proc** **logistic** data=newheart3;

title "Analyzing outliers and influential points -4";

model HeartDisease(event='1')= Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal

MaxHR Angina\_Yes Oldpeak Slope\_Down Slope\_Flat/stb corrb influence iplots;

**run**;

/\*delete observations from Pearson residual graph which are above +3 and below -3\*/

**data** newheart4;

set newheart3;

if \_n\_ in (**81**,**146**,**186**,**232**,**330**,**346**,**378**,**759**) then delete;

**run**;

/\*Analyzing data after removing above outliers\*/

**proc** **logistic** data=newheart4;

title "Analyzing outliers and influential points -5";

model HeartDisease(event='1')= Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal

MaxHR Angina\_Yes Oldpeak Slope\_Down Slope\_Flat/stb corrb influence iplots;

**run**;

/\*delete observations from Pearson residual graph which are above +3 and below -3\*/

**data** newheart5;

set newheart4;

if \_n\_ in (**26**,**111**,**264**,**301**,**753**,**811**) then delete;

**run**;

/\*Analyzing data after removing above outliers\*/

**proc** **logistic** data=newheart5;

title "Analyzing outliers and influential points -6";

model HeartDisease(event='1')= Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal

MaxHR Angina\_Yes Oldpeak Slope\_Down Slope\_Flat/stb corrb influence iplots;

**run**;

/\*delete observations from Pearson residual graph which are above +3 and below -3\*/

**data** newheart6;

set newheart5;

if \_n\_ in (**294**,**304**,**775**,**842**) then delete;

**run**;

/\*Analyzing data after removing above outliers\*/

**proc** **logistic** data=newheart6;

title "Analyzing outliers and influential points -7";

model HeartDisease(event='1')= Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal

MaxHR Angina\_Yes Oldpeak Slope\_Down Slope\_Flat/stb corrb influence iplots;

**run**;

/\*delete observations from Pearson residual graph which are above +3 and below -3\*/

**data** newheart7;

set newheart6;

if \_n\_ in (**787**,**806**,**825**) then delete;

**run**;

/\*Analyzing data after removing above outliers\*/

**proc** **logistic** data=newheart7;

title "Analyzing outliers and influential points -8";

model HeartDisease(event='1')= Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal

MaxHR Angina\_Yes Oldpeak Slope\_Down Slope\_Flat/stb corrb influence iplots;

**run**;

/\*delete observations from Pearson residual graph which are above +3 and below -3\*/

**data** newheart8;

set newheart7;

if \_n\_ in (**68**,**824**) then delete;

**run**;

/\*Analyzing data after removing above outliers\*/

**proc** **logistic** data=newheart8;

title "Analyzing outliers and influential points -9";

model HeartDisease(event='1')= Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal

MaxHR Angina\_Yes Oldpeak Slope\_Down Slope\_Flat/stb corrb influence iplots;

**run**;

/\*delete observations from Pearson residual graph which are above +3 and below -3\*/

**data** newheart9;

set newheart8;

if \_n\_ in (**185**,**764**,**787**) then delete;

**run**;

/\*Analyzing data after removing above outliers\*/

**proc** **logistic** data=newheart9;

title "Analyzing outliers and influential points -10";

model HeartDisease(event='1')= Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal

MaxHR Angina\_Yes Oldpeak Slope\_Down Slope\_Flat/stb corrb influence iplots;

**run**;

/\*delete observations from Pearson residual graph which are above +3 and below -3\*/

**data** newheart10;

set newheart9;

if \_n\_ in (**162**,**761**) then delete;

**run**;

/\*Analyzing data after removing above outliers\*/

**proc** **logistic** data=newheart10;

title "Analyzing outliers and influential points -11";

model HeartDisease(event='1')= Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal

MaxHR Angina\_Yes Oldpeak Slope\_Down Slope\_Flat/stb corrb influence iplots;

**run**;

/\*delete observations from Pearson residual graph which are above +3 and below -3\*/

**data** newheart11;

set newheart10;

if \_n\_ = **433** then delete;

**run**;

/\*Analyzing data after removing above outliers\*/

**proc** **logistic** data=newheart11;

title "Analyzing outliers and influential points -12";

model HeartDisease(event='1')= Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal

MaxHR Angina\_Yes Oldpeak Slope\_Down Slope\_Flat/stb corrb influence iplots;

**run**;

/\*delete observations from Pearson residual graph which are above +3 and below -3\*/

**data** newheart12;

set newheart11;

if \_n\_ in (**31**,**145**,**177**,**185**,**194**,**210**,**214**,**229**,**254**,**284**,**299**,**307**,**313**,**351**,**354**,**371**,**388**,**398**,**405**,**406**,**435**,**466**,**483**,**504**,**652**,**658**,**673**,

**678**,**765**,**775**,**798**,**808**) then delete;

**run**;

/\*Analyzing data after removing above outliers\*/

**proc** **logistic** data=newheart12;

title "Analyzing outliers and influential points -13";

model HeartDisease(event='1')= Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal

MaxHR Angina\_Yes Oldpeak Slope\_Down Slope\_Flat/stb corrb influence iplots;

**run**;

/\*using backward selection method for full model\*/

**proc** **logistic** data=newheart12;

title "Selection Method - Backward";

model HeartDisease(event='1')= Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal

MaxHR Angina\_Yes Oldpeak Slope\_Down Slope\_Flat/selection=backward stb rsquare;

**run**;

/\*Final model - checking for collinearity, influential points and outliers based on backward selection method\*/

**proc** **logistic** data=newheart12;

title "Final Model - Checking collinearity(corrb), influential points(influential obs.) and outliers(iplots)";

model HeartDisease(event='1')= Sex\_Male CP\_ASY CP\_ATA Cholesterol FastingBS ECG\_LVH ECG\_Normal Angina\_Yes Oldpeak

Slope\_Flat/stb corrb influence iplots rsquare;

**run**;

/\*Create new dataset with new value\*/

title "Predictions";

**data** pred;

input Sex\_Male CP\_ASY CP\_ATA Cholesterol FastingBS ECG\_LVH ECG\_Normal Angina\_Yes Oldpeak Slope\_Flat;

datalines;

1 0 1 120 1 1 0 0 0.6 0

0 1 0 80 0 0 1 1 -0.8 1

;

**proc** **print**;

**run**;

/\*add blank and new dataset\*/

**data** prediction;

set pred newheart12;

**run**;

**proc** **print**;

**run**;

/\*updated logistic regression model\*/

**PROC** **logistic** data=prediction;

title "Final prediction model";

model HeartDisease(event='1')= Sex\_Male CP\_ASY CP\_ATA Cholesterol FastingBS ECG\_LVH ECG\_Normal Angina\_Yes Oldpeak

Slope\_Flat;

output out = prediction p = phat lower = lcl upper = ucl;

**run**;

**proc** **print** data = prediction;

**run**;

/\* Model validation \*/

title "Test and Train Sets for Heart Stroke Prediction dataset";

**proc** **surveyselect** data=newheart12 out=heart\_all seed=**353678**

samprate=**0.75** outall; \*outall - show all the data selected (1) and not selected (0) for training;

**run**;

**proc** **print** data=heart\_all;

**run**;

/\* create new variable new\_y = HeartDisease for training set, and = NA for testing set\*/

**data** heart\_all;

set heart\_all;

if selected then new\_y=HeartDisease;

**run**;

**proc** **print** data=heart\_all;

**run**;

/\* using backward selection method of full model\*/

title "Backward Selection";

**proc** **logistic** data=heart\_all;

model new\_y (event='1')= Age Sex\_Male CP\_ASY CP\_ATA CP\_NAP RestingBP Cholesterol FastingBS ECG\_LVH ECG\_Normal

MaxHR Angina\_Yes Oldpeak Slope\_Down Slope\_Flat/selection =backward stb rsquare;

**run**;

/\*Significant predictors of train dataset\*/

title "Significant predictors of train dataset";

**proc** **logistic** data=heart\_all;

model new\_y(event='1')=Sex\_Male CP\_ASY CP\_ATA Cholesterol FastingBS ECG\_LVH ECG\_Normal Angina\_Yes Oldpeak

Slope\_Flat/ rsquare;

**run**;

/\*Test Set: Final prediction model\*/

title "Test Set: Final prediction model";

**proc** **logistic** data=heart\_all;

model new\_y(event='1')=Sex\_Male CP\_ASY CP\_ATA Cholesterol FastingBS ECG\_LVH ECG\_Normal Angina\_Yes Oldpeak Slope\_Flat;

output out=pred (where=(new\_y=**.**)) p=phat lower=lcl upper=ucl;

**run**;

**proc** **print** data=pred;

**run**;

/\*Frequency table for the final model of the test dataset\*/

title "Frequency table for the final model of the test dataset";

**data** final;

set pred;

pred\_y=**0**;

threshold=**0.5**;

if phat>threshold then pred\_y=**1**;

**run**;

**proc** **print** data=final;

**run**;

**proc** **freq** data=final;

tables HeartDisease\*pred\_y/norow nocol nopercent;

**run**;