%web\_drop\_table(WORK.IMPORT);  
ods pdf file="\home\u63561731\cardio\_output.pdf";  
  
FILENAME REFFILE '/home/u63561731/cardio\_train.csv';  
  
PROC IMPORT DATAFILE=REFFILE  
 DBMS=CSV  
 OUT=WORK.IMPORT;  
 DELIMITER=';';  
 GETNAMES=YES;  
RUN;  
  
PROC CONTENTS DATA=WORK.IMPORT; RUN;  
  
  
%web\_open\_table(WORK.IMPORT);  
  
DATA CARDIO;  
 set Work.Import;  
  
  
proc surveyselect data=CARDIO  
out=CARDIO  
method=srs  
sampsize=7000  
seed=123;  
Run;  
  
\*Change age variable in years and add BMI variable;  
DATA CARDIO;  
SET CARDIO;  
age = age/365;  
BMI = WEIGHT / (HEIGHT\*0.01) \*\* 2;  
RUN;  
  
  
\*Changing format for the output;  
proc format;  
 value cardiofm  
 0 = "No Disease"  
 1 = "Heart Disease";  
RUN;  
Data CARDIO1;  
 Set CARDIO;  
 format Cardio cardiofm.;  
Run;  
  
  
/\*\*\*\*\*\*\*\* EDA \*\*\*\*\*\*\*\*/  
  
/\* Distribution of heart disease \*/  
proc sgplot data=CARDIO1;  
VBAR cardio;  
TITLE 'Distribution of Heart Disease';  
RUN;  
  
/\* Distribution of heart disease by gender \*/  
proc sgplot data=CARDIO1;  
VBAR cardio / GROUP = gender;  
TITLE 'Distribution of Heart Disease by Gender';  
RUN;  
  
/\* Bar plot of cholesterol \*/  
proc sgplot data=CARDIO1;  
VBAR cholesterol / GROUP = cardio GROUPDISPLAY = CLUSTER;  
TITLE 'Bar Plot of Cholesterol by target variable';  
RUN;  
  
/\* Bar plot of glucose level \*/  
proc sgplot data=CARDIO1;  
VBAR gluc / GROUP = cardio GROUPDISPLAY = CLUSTER;  
TITLE 'Bar Plot of Glucose Level by target variable';  
RUN;  
  
/\* Bar plot of smoking \*/  
proc sgplot data=CARDIO1;  
VBAR smoke / GROUP = cardio GROUPDISPLAY = CLUSTER;  
TITLE 'Bar Plot of Smoking by target variable';  
RUN;  
  
/\* Box plot of age \*/  
proc sgplot data=CARDIO1;  
vbox age/ category=cardio;  
TITLE 'Distribution of Age';  
Run;  
  
/\* Remove outliers in ap\_hi, ap\_lo, and BMI \*/  
DATA CARDIO1;  
SET CARDIO1;  
IF ap\_hi > 750 THEN ap\_hi = 130;  
IF ap\_lo > 750 THEN ap\_lo = 97;  
IF BMI > 80 THEN BMI = 28;  
Run;  
  
/\* Box plot of systolic blood pressure \*/  
proc sgplot data=CARDIO1;  
vbox ap\_hi/ category=cardio;  
TITLE 'Distribution of Systolic Blood Pressure';  
Run;  
  
/\* Box plot of diastolic blood pressure \*/  
proc sgplot data=CARDIO1;  
vbox ap\_lo/ category=cardio;  
TITLE 'Distribution of Diastolic Blood Pressure';  
Run;  
  
/\* Box plot of BMI \*/  
proc sgplot data=CARDIO1;  
vbox bmi / category=cardio;  
TITLE 'Distribution of BMI';  
run;  
  
/\* T-test\*/  
PROC TTEST data=CARDIO;  
CLASS cardio;  
VAR age height weight ap\_hi ap\_lo BMI;  
Run;  
  
\*SRS;  
proc surveyselect data=CARDIO  
 out=CARDIO\_RANDOM  
 method=srs   
 sampsize=7000   
 seed=123;   
Run;  
  
\*Checking the missing values;  
proc means data = cardio Nmiss; \*no missing values;  
  
\*Changing format for the output;  
proc format;  
 value cardiofm  
 0 = "No Disease"  
 1 = "Heart Disease";  
RUN;  
  
  
Data Cardio;  
 Set Cardio\_random;  
 format Cardio cardiofm.;  
run;  
  
  
\*Logistic Regression;  
  
\*The variables have low corralation;  
  
  
\*Dividing the data into train and test;  
\*Sorting the dataset;  
proc sort data = Cardio out= Cardio1;  
by Cardio;  
run;  
  
\*Perform Simple random sampling;  
proc surveyselect data= Cardio1 rate=0.7 outall out=Cardio2 seed=123;  
run;  
  
\*Split the data into train 70% and test30%;  
data train test;   
set Cardio2;   
if selected = 1 then output train;   
else output test;   
drop selected;  
run;  
  
\*Checking the frequency of Cardio in each train and testd dataset;  
title "Train data";  
proc freq data = train;  
table Cardio;  
run;   
title "Test data";  
proc freq data = test;  
table Cardio;  
run;  
  
\*Model Building;\* with gluc and smoke;  
proc logistic descending data = train;  
 class gender cholesterol(ref="1") gluc(ref="1") smoke(ref="0") alco(ref="0") active(ref="0")/ param = ref;  
 model cardio(event = "Heart Disease") = age gender BMI ap\_hi ap\_lo cholesterol gluc smoke alco active/selection=stepwise lackfit;  
 output out = mod\_train p = Probability predprobs = individual;  
 store cardio\_logistic;  
 run;  
   
DATA mod\_train;  
 set mod\_train;  
 Prediction = \_INTO\_;  
run;  
  
\*Confusion Matrix;  
proc freq data = mod\_train;  
table Cardio\*Prediction /out = Counts nocol nopercent;  
run;  
  
proc contents data = cardiocount;  
  
\*Accuracy;  
data cardiocount;  
 set Counts;  
 Cardio1 = put(cardio, cardiofm.);  
 Match = 0;  
 if Cardio1 = Prediction then Match = 1;  
 run;  
  
proc means data = cardiocount mean;  
freq count;  
var MATCH;  
RUN;  
  
  
\*Model Building;\* without gluc and smoke;  
proc logistic descending data = train;  
 class gender cholesterol(ref="1") alco(ref="0") active(ref="0")/ param = ref;  
 model cardio(event = "Heart Disease") = age gender BMI ap\_hi ap\_lo cholesterol alco active/expb selection=stepwise lackfit;  
 output out = mod\_train p = Probability predprobs = individual;  
 store cardio\_logistic;  
 run;  
   
DATA mod\_train;  
 set mod\_train;  
 Prediction = \_INTO\_;  
run;  
  
\*Confusion Matrix;  
proc freq data = mod\_train;  
table Cardio\*Prediction /out = Counts nocol nopercent;  
run;  
  
proc contents data = cardiocount;  
  
\*Accuracy;  
data cardiocount;  
 set Counts;  
 Cardio1 = put(cardio, cardiofm.);  
 Match = 0;  
 if Cardio1 = Prediction then Match = 1;  
 run;  
   
proc means data = cardiocount mean;  
freq count;  
var MATCH;  
RUN;  
  
  
\*Model with interaction;  
proc logistic descending data = train;  
 class gender cholesterol(ref="1") gluc(ref="1") smoke(ref="0") alco(ref="0") active(ref="0")/ param = ref;  
 model cardio(event = "Heart Disease") = age | gender | BMI | ap\_hi |ap\_lo| cholesterol |gluc |smoke |alco |active/selection=stepwise lackfit;  
 output out = mod\_train p = Probability predprobs = individual;  
 store cardio\_logistic;  
 run;  
  
\*Confusion Matrix;  
proc freq data = mod\_train;  
table Cardio\*\_INTO\_ /out = Counts nocol nopercent;  
run;  
  
proc contents data = cardiocount;  
  
\*Accuracy;  
data cardiocount;  
 set Counts;  
 Cardio1 = put(cardio, cardiofm.);  
 Match = 0;  
 if Cardio1 = \_INTO\_ then Match = 1;  
 run;  
   
proc means data = cardiocount mean;  
freq count;  
var MATCH;  
RUN;  
  
\*Testing the dataset;  
proc plm source = cardio\_logistic;  
score data = test out= pred predicted = p lclm = lower uclm = upper /ilink;  
run;  
  
\*Classifies prediction probability;  
data pred;  
 set pred;  
 if p >= 0.5 then prediction = "Heart Disease";  
 else prediction = "No Disease";  
 keep prediction cardio id;  
 run;  
   
\*Confusion matrix;  
proc freq data = pred;  
table cardio\*prediction /out = Counts2 nocol nopercent;  
run;  
  
  
\*Accuracy;  
data cardiocount2;  
 set Counts2;  
 Cardio2 = put(cardio, cardiofm.);  
 Match = 0;  
 if Cardio2 = prediction then Match = 1;  
 run;  
  
proc means data = cardiocount2 mean;  
freq count;  
var Match;  
run;  
  
  
\*Try Predicting a Value;  
\*Variable Importance;  
proc logistic descending data = train;  
 class gender cholesterol(ref="1") alco(ref="0") active(ref="0")/ param = ref;  
 model cardio(event = "Heart Disease") = age gender BMI ap\_hi ap\_lo cholesterol alco active/ stb selection=stepwise lackfit;  
 ods output ParameterEstimates = estimates;  
 run;  
  
Data estimates1;  
 set estimates;  
 keep Variable StandardizedEst ClassVal0;  
 StandardizedEst = abs(StandardizedEst);  
 if \_n\_ >1;  
 run;  
  
data estimates1;  
 set estimates1;  
 Level = ClassVal0;  
 drop ClassVal0;  
 run;  
   
proc rank data= estimates1 out=rank descending;  
ranks rank;  
var StandardizedEst;  
run;  
  
proc sort data = rank;  
by rank;  
run;  
  
title "Variable Importance Rank";  
proc print data = rank;  
run;  
  
   
\*Model Building; \*with gluc and smoke;  
proc logistic descending data = train;  
 class gender cholesterol(ref="1") gluc(ref="1") smoke(ref="0") alco(ref="0") active(ref="0")/ param = ref;  
 model cardio(event = "Heart Disease") = age gender BMI ap\_hi ap\_lo cholesterol gluc smoke alco active/selection=stepwise lackfit;  
 output out = mod\_train p = Probability predprobs = individual;  
 store cardio\_logistic;  
 run;  
  
\*without gluc and smoke and interaction;  
proc logistic descending data = train;  
 class gender cholesterol(ref="1") alco(ref="0") active(ref="0")/ param = ref;  
 model cardio(event = "Heart Disease") = age | gender | BMI | ap\_hi |ap\_lo| cholesterol |alco |active/selection=stepwise lackfit;  
 output out = mod\_train1 p = Probability predprobs = individual;  
 store cardio\_logistic;  
 run;  
  
  
\*with gluc and smoke and interaction;  
proc logistic descending data = train;  
 class gender cholesterol(ref="1") alco(ref="0") active(ref="0")/ param = ref;  
 model cardio(event = "Heart Disease") = age | gender | BMI | ap\_hi |ap\_lo| cholesterol |alco |active/selection=stepwise lackfit;  
 output out = mod\_train1 p = Probability predprobs = individual;  
 store cardio\_logistic;  
 run;  
  
ods pdf close;