Analyzing the Heart Failure Prediction Improving Patient Outcomes through Early Detection

Introduction

Heart failure is a huge medical problem with significant mortality and morbidity globally. Therefore, early diagnosis and accurate prediction of heart failure risk is an essential factor for the promotion and administration of life-saving therapies. Importantly, machine learning approaches are of great value in forecasting heart failure risk based on different patient-related characteristics and parameters of clinical tests.

The primary objective of this study is to establish a model that can predict heart failure using the variables in the dataset, including sex, age, chest pain type, resting blood pressure, cholesterol, fasting blood sugar test results in years, resting electrocardiographic results, maximum heart rate achieved during the exercise test, exercise-induced angina, ST depression induced by exercise relative to rest (Oldpeak), and the slope of the ST segment peak exercise.

By analyzing these variables, we seek to create a robust predictive model that can accurately identify individuals at high risk of heart failure. This model could potentially assist healthcare providers in early intervention and personalized risk management strategies, ultimately leading to improved patient outcomes and reduced burden on healthcare systems.

Data Collection

The dataset used for this project is named "Heart Failure Prediction Dataset." It was obtained from Kaggle and consists of 918 records and 12 variables. These variables include Age, Sex, Chest Pain, Resting Blood Pressure (RestingBP), Cholesterol levels, Fasting Blood Sugar (FastingBS), Resting Electrocardiographic Results (RestingECG), Maximum Heart Rate achieved during exercise (MaxHR), Exercise-induced Angina (ExerciseAngina), ST depression induced by exercise relative to rest (Oldpeak), ST Slope, and the presence of Heart Disease.

The goal of this project was to predict the risk of heart failure using machine learning models. By analyzing these variables, we aimed to develop predictive models to accurately identify individuals at high risk of heart failure. The dataset contained 918 records with information collected on various patient characteristics and clinical measurements.

Age	Age of the patient [years]
Sex	Sex of the patient [M: Male, F: Female]
ChestPainType	Chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
RestingBP	Resting blood pressure [mm Hg]
Cholesterol	Serum cholesterol [mm/dl]
FastingBS	Fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
RestingECG	Resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]

MaxHR	Maximum heart rate achieved [Numeric value between 60 and 202]
ExerciseAngina	Exercise-induced angina [Y: Yes, N: No]
Oldpeak	ST [Numeric value measured in depression]
ST Slope	The slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
HeartDisease	Output class [1: heart disease, 0: Normal]

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

Data Exploration and Visualization

Before creating any models, we explored each of the 12 variables using a variety of graphs and statistical summaries to better understand the data we have and to ensure that there are not any easily seen logical errors that would question the validity of the data from the start. It was also important to take sufficient time to understand what each variable meant in depth, so the models are easier to interpret later on.

Sex

Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	193	21.02	193	21.02
M	725	78.98	918	100.00

In this frequency table we can see There are 193 females (21.02% of the total sample) and 725 males (78.98% of the total sample), making up a total of 918 individuals. Overall Male Heart failure frequency is highest than female.

ChestPainType

ChestPainType	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ASY	392	77.17	392	77.17
ATA	24	4.72	416	81.89
NAP	72	14.17	488	96.06
TA	20	3.94	508	100.00

The "ChestPainType" feature describes the type of chest pain a patient experiences. There are four categories of chest pain type: ASY: Asymptomatic (no chest pain), ATA: Atypical Angina (chest pain that doesn't meet the classic angina description), NAP: Non-Anginal Pain (chest pain not related to the heart), TA: Typical Angina (crushing or squeezing chest pain, often radiating to the arm or jaw)

FastingBS				
FastingBS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	338	66.54	338	66.54
1	170	33.46	508	100.00

Fasting Blood Sugar (FastingBS) Distribution: The distribution of the "FastingBS" variable in the dataset is as follows: FastingBS = 0: 338 records (66.54%), FastingBS = 1: 170 records (33.46%) This indicates that 66.54% of the individuals in the dataset have a fasting blood sugar level below the threshold, while 33.46% have a fasting blood sugar level above the threshold.

Resting ECG: This table showing the frequency of resting (ECG) results in a population with heart failure.

RestingECG				
RestingECG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
LVH	106	20.87	106	20.87
Normal	285	56.10	391	76.97
ST	117	23.03	508	100.00

Left ventricular hypertrophy (LVH) is a common finding in heart failure. It can occur when the heart has to work harder to pump blood. **ST segment abnormalities** can also be a sign of heart failure. They can indicate damage to the heart muscle or problems with blood flow to the heart. The table shows that there were 106 ECG results with LVH, which represents 20.87% of the total ECG results. The cumulative frequency for LVH is 106, and the cumulative percent is 20.87%. This means that LVH was the most common ECG result, and it accounted for 20.87% of all ECG results.

ExerciseAngina: Exercise angina is a common symptom of heart failure.

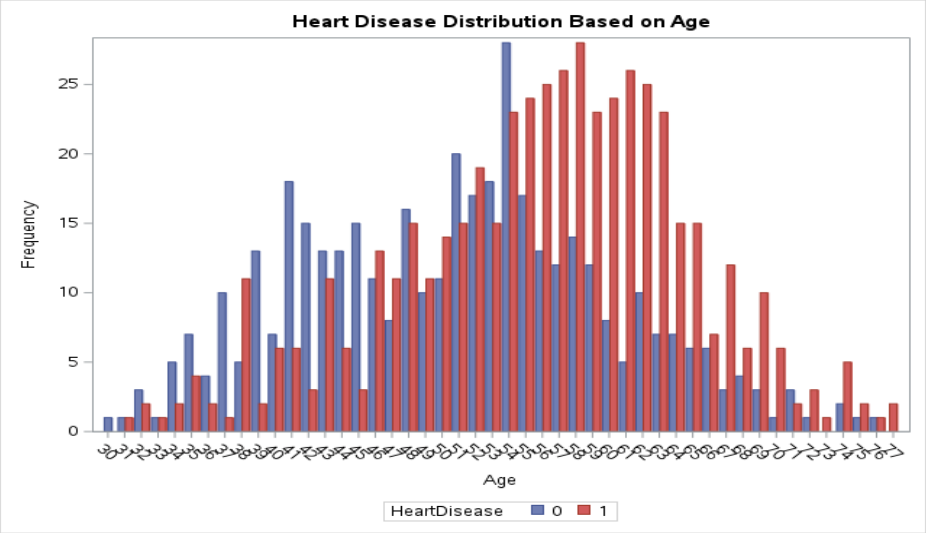
ExerciseAngina				
ExerciseAngina	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	192	37.80	192	37.80
Y	316	62.20	508	100.00

The table shows two categories: **N:** No exercise angina, **Y:** Yes, exercise angina. The table shows that 192 people did not report exercise angina (N), which represents 37.8% of the total people in the study. The cumulative frequency for no exercise angina (N) is 192, and the cumulative percent is 37.8%. This means that more people did not report exercise angina than did (37.8% vs. 62.2%). Overall, the table shows that exercise angina is a common symptom of heart failure in the people studied, affecting over 60% of the participants.

ST_Slope				
ST_Slope	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Down	49	9.65	49	9.65
Flat	381	75.00	430	84.65
Up	78	15.35	508	100.00

The ST_Slope: feature in ECGs assesses heart failure risk. It consists of three main categories: downsloping, flat, and upsloping. Downsloping seen in 9.65% of ECGs it indicates reduced blood flow to the heart muscle, flat segments may be normal or abnormal, it seen in 75% of ECGs and upsloping is the least common category, seen in 15.35% of ECGs, generally considered normal. In this slide we can see frequency table for 4 heart failure features which are chestpaintype, fasting blood sugar, Resting ECG, Exercise Angina.

Bar Charts:



The bar chart shows the distribution of heart disease based on age. The y-axis labeled "frequency" represents the number of people who have heart failure, and the x-axis labeled "age" represents different age groups. The bars on the chart are colored red and blue. Red represents present heart failure, and blue represents absent heart failure. Moreover, it can be observed from the chart that the age group of 58 people has the highest heart failure count.

Correlation Matrix:

Pearson Correlation Coefficients, N = 918 Prob > |r| under H0: Rho=0							
	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
Age Age	1.00000	0.25440 <.0001	-0.09528 0.0039	0.19804 <.0001	-0.38204 <.0001	0.25861 <.0001	0.28204 <.0001
RestingBP RestingBP	0.25440 <.0001	1.00000	0.10089 0.0022	0.07019 0.0335	-0.11213 0.0007	0.16480 <.0001	0.10759 0.0011
Cholesterol Cholesterol	-0.09528 0.0039	0.10089 0.0022	1.00000	-0.26097 <.0001	0.23579 <.0001	0.05015 0.1289	-0.23274 <.0001
FastingBS FastingBS	0.19804 <.0001	0.07019 0.0335	-0.26097 <.0001	1.00000	-0.13144 <.0001	0.05270 0.1106	0.26729 <.0001
MaxHR MaxHR	-0.38204 <.0001	-0.11213 0.0007	0.23579 <.0001	-0.13144 <.0001	1.00000	-0.16069 <.0001	-0.40042 <.0001
Oldpeak Oldpeak	0.25861 <.0001	0.16480 <.0001	0.05015 0.1289	0.05270 0.1106	-0.16069 <.0001	1.00000	0.40395 <.0001
HeartDisease HeartDisease	0.28204 <.0001	0.10759 0.0011	-0.23274 <.0001	0.26729 <.0001	-0.40042 <.0001	0.40395 <.0001	1.00000

This table shows the Pearson correlation coefficients between different health factors measured in a dataset of 918 individuals. Pearson correlation coefficient is a measure of the linear dependence between two variables. It ranges from -1 to 1, with 1 indicating a perfect positive correlation, -1 indicating a perfect negative correlation, and 0 indicating no correlation.

The correlation coefficient of Age has a weak positive linear relationship with Resting Blood Pressure (0.2544), Fasting Blood Sugar (0.1980), Oldpeak (0.2586), and the presence of Heart Disease (0.2820). Age has a moderate negative linear relationship with Maximum Heart Rate achieved (-0.3820). Age has a weak negative linear relationship with Cholesterol (-0.0953). Resting Blood Pressure (RestingBP) is positively correlated with Age (0.2544) and Cholesterol (0.1009), and negatively correlated with MaxHR (-0.1121). Cholesterol is positively correlated with RestingBP (0.1009) and negatively correlated with Age (-0.0953). Fasting Blood Sugar (FastingBS) is positively correlated with Age (0.1980) and negatively correlated with MaxHR (-0.1314) and Cholesterol (-0.2609). Maximum Heart Rate achieved (MaxHR) is negatively correlated with Age (-0.3820), RestingBP (-0.1121), and Oldpeak (-0.1607). Oldpeak is positively correlated with Age (0.2586), RestingBP (0.1648), and Heart Disease (0.4040), and negatively correlated with MaxHR (-0.1607). Presence of Heart Disease (HeartDisease) is positively correlated with Age (0.2820), RestingBP (0.1076), FastingBS (0.2673), and Oldpeak (0.4040), and negatively correlated with MaxHR (-0.4004). This table suggests that as age increases, these values tend to rise slightly. Conversely, there's a weak negative correlation between age and cholesterol (-0.09528) and maximal heart rate (-0.38204), implying that cholesterol levels and heart rate tend to decrease somewhat with advancing age.

High-Performance Regression Model:

The HPREG Procedure

Selection Summary				
Step	Effect Entered	Number Effects In	SBC	Validation ASE
0	Intercept	1	-901.3684	0.2454
1	MaxHR	2	-1025.8658	0.2196
2	Oldpeak	3	-1112.6162	0.1848
3	FastingBS	4	-1136.5391	0.1690
4	Cholesterol	5	-1145.7261*	0.1647*

Root MSE	0.40616
R-Square	0.33977
Adj R-Sq	0.33568
AIC	-515.11861
AICC	-514.98817
SBC	-1145.72606
ASE (Train)	0.16370
ASE (Validate)	0.16472

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	54.84304	13.71076	83.11	<.0001
Error	646	106.57017	0.16497		
Corrected Total	650	161.41321			

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.325641	0.092695	14.30	<.0001
Cholesterol	1	-0.000623	0.000157	-3.97	<.0001
FastingBS	1	0.176309	0.039567	4.46	<.0001
MaxHR	1	-0.006103	0.000642	-9.50	<.0001
Oldpeak	1	0.162921	0.015514	10.50	<.0001

Model Evaluation:

i. The adjusted R square is **33.56%,** which is low model performance. This indicates that **33.56% of the** of the variation in the variable fare is explained by the model. The root mean square error (RMSE) is **0.40616**.
ii. The p-values of the slope coefficients of all the variables except "cholesterol" are less than α, indicating that this variable is a significant predictor.
iii. In this case, the output, HPREG, selected a model with Intercept, Cholesterol, FastingBS, MaxHR, and Oldpeak as predictors; the validation ASE is **0.1647**.

For the high-performance model, selection summaries include Maxhr, Oldpeak, cholesterol, and fasting blood sugar, or BS. The lowest validation average square error is seen from the table. Therefore, the value of ASE (validation) is 0.1647 and the coefficient for cholesterol is essentially statistically significant at the 5% level. It's clear from observing these variables that there is a lot of evidence to support the hypothesis that cholesterol may predict the dependent variable statistically significantly. Furthermore, when the p-value is less than 0.05, it is considered to be statistically significant that the predictor variable has an impact on the response variable.

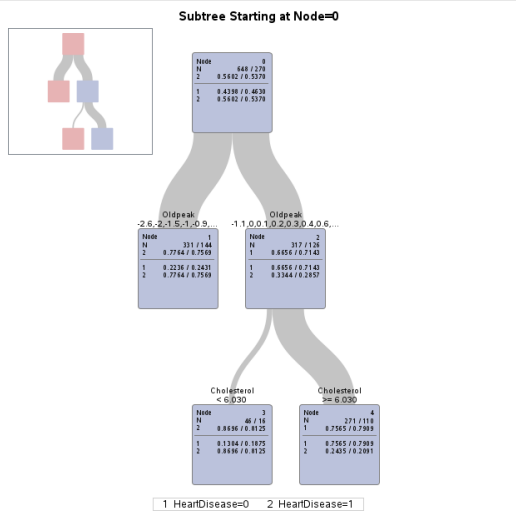
Furthermore, when the p-value is less than 0.0001, it means that the coefficient estimate differs significantly from zero.

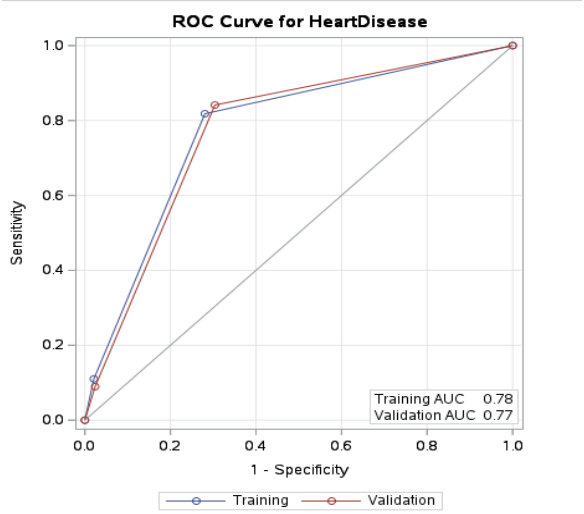
In this case, all predictors in the parameter estimates have p-values ("<.0001") that are incredibly low, indicating that they are very statistically significant. It is possible to conclude that the model has poor performance in predicting heart failures based on the adjusted R square (0.33568) and R square (0.33977) values, which are largely similar but differ slightly in points.

Decision Tree:

Model Information	
Split Criterion Used	Gini
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	10
Maximum Tree Depth Achieved	10
Tree Depth	2
Number of Leaves Before Pruning	100
Number of Leaves After Pruning	3
Model Event Level	1

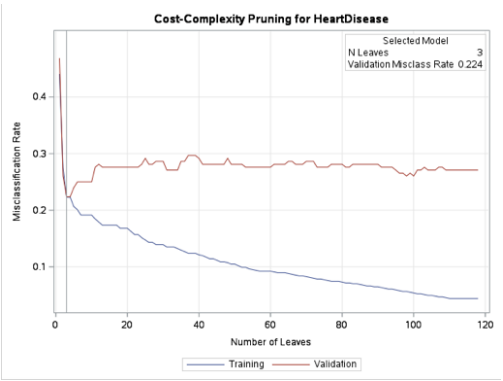
Number of Observations Read	918
Number of Observations Used	918
Number of Training Observations Used	648
Number of Validation Observations Used	270





Fit Statistics for Selected Tree									
	N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
Training	3	0.1722	0.2231	0.7906	0.7594	0.7610	0.3444	250.0	0.7778
Validation	3	0.1746	0.2240	0.8039	0.7444	0.7672	0.3475	67.0643	0.7733

Confusion Matrices				
		Predicted		Error
	Actual	0	1	Rate
Training	0	243	77	0.2406
	1	85	321	0.2094
Validation	0	67	23	0.2556
	1	20	82	0.1961



Node Information									
		Training Data				Validation Data			
ID	Path	Count	0		1		Count	0	1
0	Root Node	726	0.4408		0.5592	*	192	0.4688	0.5313
1	Root Node	726	0.4408		0.5592		192	0.4688	0.5313
	Oldpeak = -2.6,-2,-1.5,-1,-0.9,-0.8,-0.7,0.5,0.7,0.9,1,1.1,1....	342	0.1988		0.8012	*	92	0.2174	0.7826
2	Root Node	726	0.4408		0.5592		192	0.4688	0.5313
	Oldpeak = -1.1,-0.5,-0.1,0,0.1,0.2,0.3,0.4,0.6,0.8,1.6,2.3,3....	384	0.6563	*	0.3438		100	0.7000	0.3000
3	Root Node	726	0.4408		0.5592		192	0.4688	0.5313
	Oldpeak = -1.1,-0.5,-0.1,0,0.1,0.2,0.3,0.4,0.6,0.8,1.6,2.3,3....	384	0.6563		0.3438		100	0.7000	0.3000
	Cholesterol < 6.03	56	0.1607		0.8393	*	13	0.2308	0.7692
4	Root Node	726	0.4408		0.5592		192	0.4688	0.5313
	Oldpeak = -1.1,-0.5,-0.1,0,0.1,0.2,0.3,0.4,0.6,0.8,1.6,2.3,3....	384	0.6563		0.3438		100	0.7000	0.3000
	Cholesterol >= 6.03 or Missing	328	0.7409	*	0.2591		87	0.7701	0.2299
* Selected target level									

Variable Importance							
		Training		Validation			
Variable	Variable Label	Relative	Importance	Relative	Importance	Relative Ratio	Count
Oldpeak	Oldpeak	1.0000	8.7004	1.0000	4.7090	1.0000	1
Cholesterol	Cholesterol	0.6522	5.6743	0.5491	2.5858	0.8420	1

The dataset is first divided by the tree according to the old peak feature. The dataset is further divided by Nodes 1 and 2 according to the oldpeak value. The dataset is further divided by Nodes 3 and 5 according to the level of cholesterol. The tree predicts whether or not a patient has heart failure based on these divisions. There are two significant variables: cholesterol and oldpeak.

i. The specificity is 75.94% in the training set and 74.44% in the validation set, which is also high. These values are close to each other for both training and validation.
ii. The sensitivity is 79.06% in the training set and 80.39% in the validation set, which are high values and reasonably close.
iii. There is no sign of overfitting because the Sensitivity, Specificity and AUC values are close to each other for both the training and the validation data.
iv. The AUC for training set is 77.56% and for validation set is 77.22% which shows moderate performance model.

Logistic Regression:

Model Information		
Data Set	WORK.HEART	
Response Variable	HeartDisease	HeartDisease
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	551
Number of Observations Used	551

Response Profile		
Ordered Value	HeartDisease	Total Frequency
1	0	250
2	1	301

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	4.5155	1.2803	12.4389	0.0004
Age	1	-0.00376	0.0135	0.0772	0.7811
RestingBP	1	0.00393	0.00610	0.4160	0.5189
Cholesterol	1	-0.00335	0.00115	8.4741	0.0036
FastingBS	1	1.2700	0.3040	17.4541	<.0001
MaxHR	1	-0.0369	0.00520	50.4742	<.0001
Oldpeak	1	1.1167	0.1328	70.6760	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	0.996	0.970	1.023
RestingBP	1.004	0.992	1.016
Cholesterol	0.997	0.994	0.999
FastingBS	3.561	1.962	6.461
MaxHR	0.964	0.954	0.974
Oldpeak	3.055	2.355	3.963

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	86.4	Somers' D	0.728
Percent Discordant	13.6	Gamma	0.728
Percent Tied	0.0	Tau-a	0.362
Pairs	75250	c	0.864

The LOGISTIC Procedure

Fit Statistics for SCORE Data											
Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC	Brier Score
WORK.HEART_TRAIN	551	-252.6	0.1996	519.1416	519.3479	549.3238	549.3238	0.369311	0.493833	0.864093	0.145561

The LOGISTIC Procedure

Fit Statistics for SCORE Data											
Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC	Brier Score
WORK.HEART_VALID	367	-196.7	0.2807	407.4112	407.7232	434.7487	434.7487	0.257612	0.345392	0.810568	0.180021

This logistic regression model aims to predict the likelihood of heart failure based on several predictor variables. The model identifies significant predictors such as cholesterol levels, fasting blood sugar, maximum heart rate, and ST depression induced by exercise relative to rest (oldpeak). Cholesterol, fasting blood sugar, maximum heart rate, and ST depression are positively associated with the likelihood of heart disease, while age and resting blood pressure are not significant predictors. The model's ability to discriminate between those with and without heart disease is reasonably good.

The odd ratio estimate of Fasting BS is 3.404. AUC is slightly lower on Validation set compared to training set by 5%, but it is moderate. R-squared values are high in both sets, suggesting that the model explains portion of the variance in the data. AUC is moderate, with AIC and BIC values are higher on training set, which suggest the model could be improved.
In summary, the logistic regression model provides the moderate fit for predicting heart disease, with cholesterol, fasting blood sugar, maximum heart rate, and ST depression being significant predictors.

Neural Network:

Model Information	
Data Source	WORK.HEART
Architecture	MLP
Number of Input Variables	11
Number of Hidden Layers	1
Number of Hidden Neurons	11
Number of Target Variables	1
Number of Weights	221
Optimization Technique	Limited Memory BFGS

Number of Observations Read	918
Number of Observations Used	918
Number Used for Training	726
Number Used for Validation	192

Misclassification Table for HeartDisease		
Class:	1	0
1	96	6
0	15	75

Fit Statistics Table

NAME	Train: Number of Observations	Valid: Number of Observations	L1 Norm of Weights	Train: Average Error Function	Valid: Average Error Function	Train: Average Absolute Error	Valid: Average Absolute Error	Train: Maximum Absolute Error	Valid: Maximum Absolute Error	Train: Number of Wrong Classifications	Valid: Number of Wrong Classifications	Train: Misclassification Rate	Valid: Misclassification Rate
HeartDisease	726	192	136.567898	0.268237	0.295387	0.166556	0.175960	0.994099	0.960645	79	21	0.1088	0.1094

Training Table

Try	Iterations	Avg Training Error	Avg Validation Error	Reason for Stopping	Best?
1	61	0.157194	0.322952	Validation Error	
2	63	0.126134	0.295387	Validation Error	Y
3	66	0.141431	0.317735	Validation Error	
4	52	0.186431	0.333147	Validation Error	
5	65	0.141606	0.307625	Validation Error	

Iteration Table			
Iteration	Avg Training Error	Avg Validation Error	Best?
22	0.268237	0.295387	Y

The HPNEURAL procedure built a Multilayer Perceptron (MLP) neural network model to predict heart disease. Architecture: 1 hidden layer with 11 neurons. Optimization Technique: Limited Memory BFGS.
Performance:
Train Data: Average error = 0.268, Misclassification rate = 10.88%
Validation Data: Average error = 0.295, Misclassification rate = 10.94%
Training:
Best Model: Iteration 22.
Reason for stopping: Validation error minimized.
Iterations: The model was trained over 63 iterations to minimize validation errors.

In this neural network model, the response variable is heart disease. It can be observed from the table that the average error function was computed over all the training observations. It indicated how well the model fits into the training data. A lower value suggests a better fit. In this case, the average error value for training is 0.268237. Similarly, for validation, it is providing insights into how well the model generalizes to unseen data; in this case, the average error value is 0.295387.

Furthermore, the misclassification ratio for the training model is 10.88. Similar to validation, it provides insights into the model misclassification rate on unseen data, which is 10.94.

Overall, based on these factors, we can say that the neural network appears to be a good fit for the data. It demonstrates a low error rate, a low misclassification rate, and reliable performance on both training and validation.

Appendix A

```
proc import out=heart  datafile="/home/u63735896/sasuser.v94/heart.xlsx"
dbms=xlsx replace;
run;

/* Calculate percentages for positive heart disease cases for each categorical feature */
proc freq data=heart;
  tables sex;
run;

proc freq data=heart(where=(HeartDisease=1));
   tables Sex / out=sex_counts;
run;

proc freq data=heart(where=(HeartDisease=1));
   tables ChestPainType / out=cp_counts;
run;

proc freq data=heart(where=(HeartDisease=1));
   tables FastingBS / out=fbs_counts;
run;

proc freq data=heart(where=(HeartDisease=1));
   tables RestingECG / out=restecg_counts;
run;

proc freq data=heart(where=(HeartDisease=1));
   tables ExerciseAngina / out=exang_counts;
run;

proc freq data=heart(where=(HeartDisease=1));
   tables ST_Slope / out=slope_counts;
run;

/* Step 3: Create the plot */
PROC SGPLOT DATA=heart;
   VBAR Sex / GROUP=HeartDisease GROUPDISPLAY=CLUSTER;
   TITLE 'Heart Disease Distribution Based on Gender';
RUN;

PROC SGPLOT DATA=heart;
   VBAR Age / GROUP=HeartDisease GROUPDISPLAY=CLUSTER;
```

```
   xaxis values=(30 to 77 by 1);
   TITLE 'Heart Disease Distribution Based on Age';
RUN;

PROC SGPLOT DATA=heart;
   VBAR ChestPainType / GROUP=HeartDisease GROUPDISPLAY=CLUSTER;
   TITLE 'Heart Disease Distribution Based on Chest Pain Type';
RUN;

PROC SGPLOT DATA=heart;
   VBAR RestingECG / GROUP=HeartDisease GROUPDISPLAY=CLUSTER;
   TITLE 'Heart Disease Distribution Based on Resting ECG';
RUN;

PROC SGPLOT DATA=heart;
   VBAR ExerciseAngina / GROUP=HeartDisease GROUPDISPLAY=CLUSTER;
   yaxis values=(0 to 350 by 50);
   TITLE 'Heart Disease Distribution Based on Exercise Angina';
RUN;

PROC SGPLOT DATA=heart;
   VBAR ST_Slope / GROUP=HeartDisease GROUPDISPLAY=CLUSTER;
   yaxis values=(0 to 400 by 50);
   TITLE 'Heart Disease Distribution Based on ST Slope';
RUN;

PROC SGPLOT DATA=heart;
   VBAR FastingBS / GROUP=HeartDisease GROUPDISPLAY=CLUSTER;
   TITLE 'Heart Disease Distribution Based on Fasting BS';
RUN;

PROC SGPLOT DATA=heart;
   VBAR RestingBP / GROUP=HeartDisease GROUPDISPLAY=CLUSTER;
   xaxis values=(0 to 200 by 10);
   yaxis values=(0 to 100 by 10);
   TITLE 'Heart Disease Distribution Based on Resting BP';
RUN;

PROC SGPLOT DATA=heart;
   VBAR MaxHR / GROUP=HeartDisease GROUPDISPLAY=CLUSTER;
   xaxis values=(80 to 200 by 5);
   TITLE 'Heart Disease Distribution Based on Max HR';
RUN;

PROC SGPLOT DATA=heart;
   VBAR Oldpeak / GROUP=HeartDisease GROUPDISPLAY=CLUSTER;
```

```
   xaxis values=(-0.2 to 6.2 by 0.3);
   yaxis values=(0 to 100 by 5);
   TITLE 'Heart Disease Distribution Based on Old peak';
RUN;

/* Calculate correlation matrix */
proc corr data=heart;
   var Age RestingBP Cholesterol FastingBS MaxHR Oldpeak HeartDisease;
run;

/* Building a regression model  */
proc surveyselect data=heart samprate=0.8 method=srs out=heart_part outall seed=12345;
run;

data heart_train heart_val;
       set heart_part;
       if selected=1 then output heart_train; else output heart_val;
run;

/* Perform variable selection using HPREG */
proc hpreg  data=heart seed=12345;
   partition fraction(validate=0.3);
   model HeartDisease = Age RestingBP Cholesterol FastingBS MaxHR Oldpeak;
   selection method=stepwise(choose=validate);
run;

/* Decision Tree*/
proc hpsplit data=heart nodes=detail;
       class HeartDisease OldPeak FastingBS; /* Includes all the categorical variables */
       model HeartDisease(event="1")= Age RestingBP Cholesterol FastingBS MaxHR Oldpeak;
       partition fraction(validate=.3 seed=12345);
       grow gini;
       prune cc;
run;

/* Neural Network */
proc hpneural data=heart;
   partition fraction(validate=0.2 seed=12345);
   target HeartDisease/ level=nom; /* categorical target variable */
   input Sex ChestPainType RestingECG ExerciseAngina ST_Slope/level=nom;
   input Age RestingBP Cholesterol FastingBS MaxHR Oldpeak/ level=int; /* continuous
predictors */
   hidden 11; /* first hidden layer with 11 neurons */
   train maxiter=1000 numtries=5;
run;
```

```
/* Logistic Rgression */
proc logistic data=heart;
	model HeartDisease(event="1")=Age RestingBP Cholesterol FastingBS MaxHR
Oldpeak;
run;

proc surveyselect data=heart samprate=0.6 method=srs outall out=heart_part seed=12345;
run;

data heart_train heart_valid;
	set heart_part;
	if selected=1 then output heart_train; else output heart_valid;
run;

proc logistic data=heart_train outmodel=heart_model;
	model HeartDisease(event="1")=Age RestingBP Cholesterol FastingBS MaxHR
Oldpeak;
run;

proc logistic inmodel=heart_model;
	score data=heart_train fitstat;
run;

proc logistic inmodel=heart_model;
	score data=heart_valid fitstat;
run;
```