Final Group Project on Full Data Mining Analysis

PREDICTING PATIENTS WHO ARE AT A HIGH RISK OF DYING DUE TO cardiovascular heart disease (CVD'S)

Class: ISM6136.002.F21

Group No. 6

Sashanth Embakula USFid: U36956282

Syed Omar USFid: U05608232

Claudio Moncada USFid: U37768007

BACKGROUND OF THE PROBLEM:

Cardiovascular diseases (CVDs) are the **number 1 cause of death globally**, taking an estimated **17.9 million lives each year**, which accounts for **31% of all deaths worldwide**. Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure.

Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes or already established disease) need **early detection** and management wherein a machine learning model can be of great help.

ONE PERSON DIES ABOUT EVERY 37 SECONDS FROM CVD IN THE US.
IT IS ESTIMATED BY 2030, OVER 1 MILLION PEOPLE IN THE US COULD DIE EACH YEAR FROM CVD.

MOTIVATION TO SOLVE OUR PROBLEM:

Data mining is the process of examining data from various angles and synthesizing it into useful information. The information shown here could be used to increase revenue or reduce expenses. Data mining is the process of locating specific links or models among dozens of domains in massive relational databases.
Data mining and machine learning together are a powerful tool for solving a wide range of prob lems. Medical data mining is tough to handle manually due to the large number of data source. Artificial intelligence advancements have also resulted in more precise and accurate systems for medical applications that handle sensitive medical data. Heart disease is the top cause of death in even the wealthiest countries. In early stages of cardiac disease, machine learning algorithms are frequently employed to identify hazards.
According to a comprehensive literature assessment, existing approaches performed well in the prediction of heart disease on various datasets. On the other hand, several optimization algorit hms have been used to improve a variety of parameters, including accuracy, precision, and reca ll. The primary goal of this research is to analyze several machine learning algorithms in order to determine the best method for predicting heart disease survival.

Cardiovascular Diseases are a major problem in many populations and with this study we want to help those that might be at risk. This is an unfortunate disease that millions suffer, and we planto apply machine learning models to assess all features of the dataset to predict heart patie nt survival.

DESCRIPTION OF OUR DATASET:

- RISK (Output variable): encodes whether the patient was diagnosed positive with cvd (1) or whether they were tested negative (0).
- TIME: number of days the patient was kept under observation.
- Age: age of the patient.
- Anaemia Decrease of red blood cells or hemoglobin (boolean)
- creatinine_phosphokinase Level of the CPK enzyme in the blood (mcg/L)
- diabetes: If the patient has diabetes or not
- ejection_fraction Percentage of blood leaving the heart at each contraction (percentage)
- high_blood_pressure If the patient has hypertension (boolean)
- platelets: Platelets in the blood (kiloplatelets/mL)
- serum_creatinine: Level of serum creatinine in the blood (mg/dL)
- serum_sodium: Level of serum sodium in the blood (mEq/L)
- sex: Woman or man (binary)

age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets
75	0	582	0	20	1	265000
55	0	7861	0	38	0	263358
65	0	146	0	20	0	162000

serum_creatinine	serum_sodium	sex	smoking	time	Risk
1.9	130	1	0	4	1
1.1	136	1	0	6	1
1.3	129	1	1	7	1

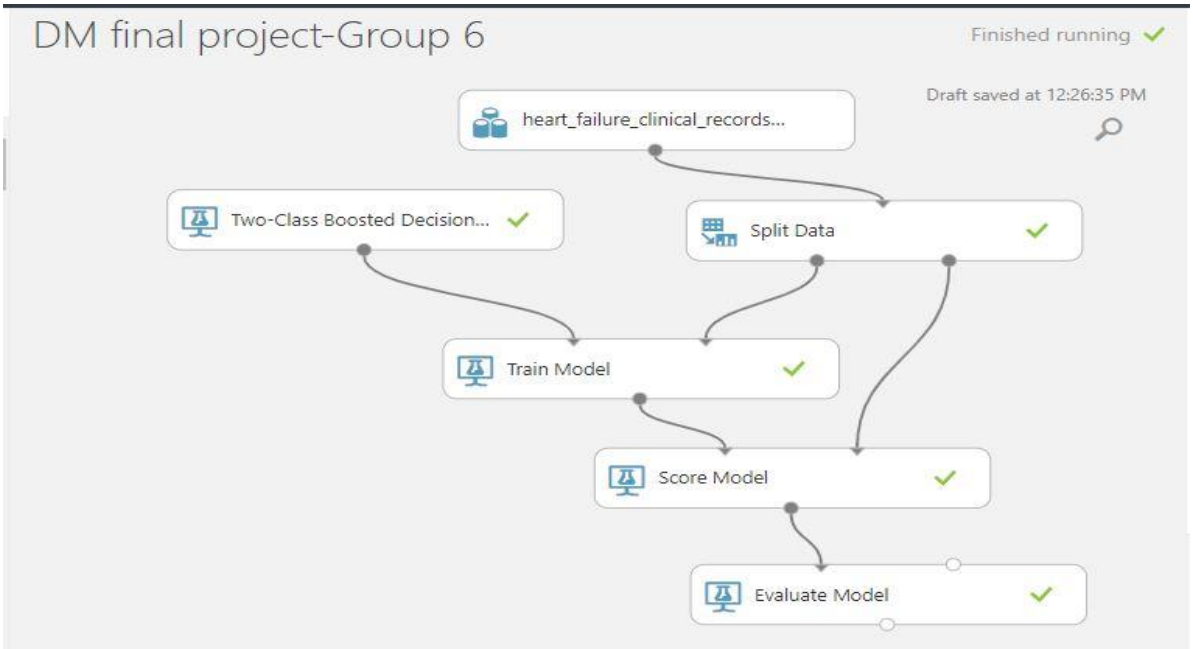
SOLUTION METHODOLOGY AND EVALUATION METRICS:

In evaluating the High risk patients, we have to identify if the patient is likely to succumb to the heart disease based on the above factors or not. For that we use the two class boosted decision tree and neural networks algorithm. The output column **RISK** would be used as a Y or a dependent variable. This is the prediction if the patient was tested positive (1) for CVD or not. The variables would be used to understand how factors like Age, sex, diabetes, creatine and platelets levels etc affect the decision of heart disease risk. The dataset is split into sets, 80% for training and the rest 20% on which the algorithm is tested. We have 300 rows of records in our system of customers who have over the years been observed and the records show if he or she was diagnosed with CVD's. These records would be used as a training data set to our model to understand the accuracy of our model.
Based on the accuracy, F1 score and AUC of our model, we would re-assess the factors affecting our dependent variable. If there are factors that need to be added or removed, the model would be readjusted and retrained.

We would then draw a lift curve to understand how accurate our model is:

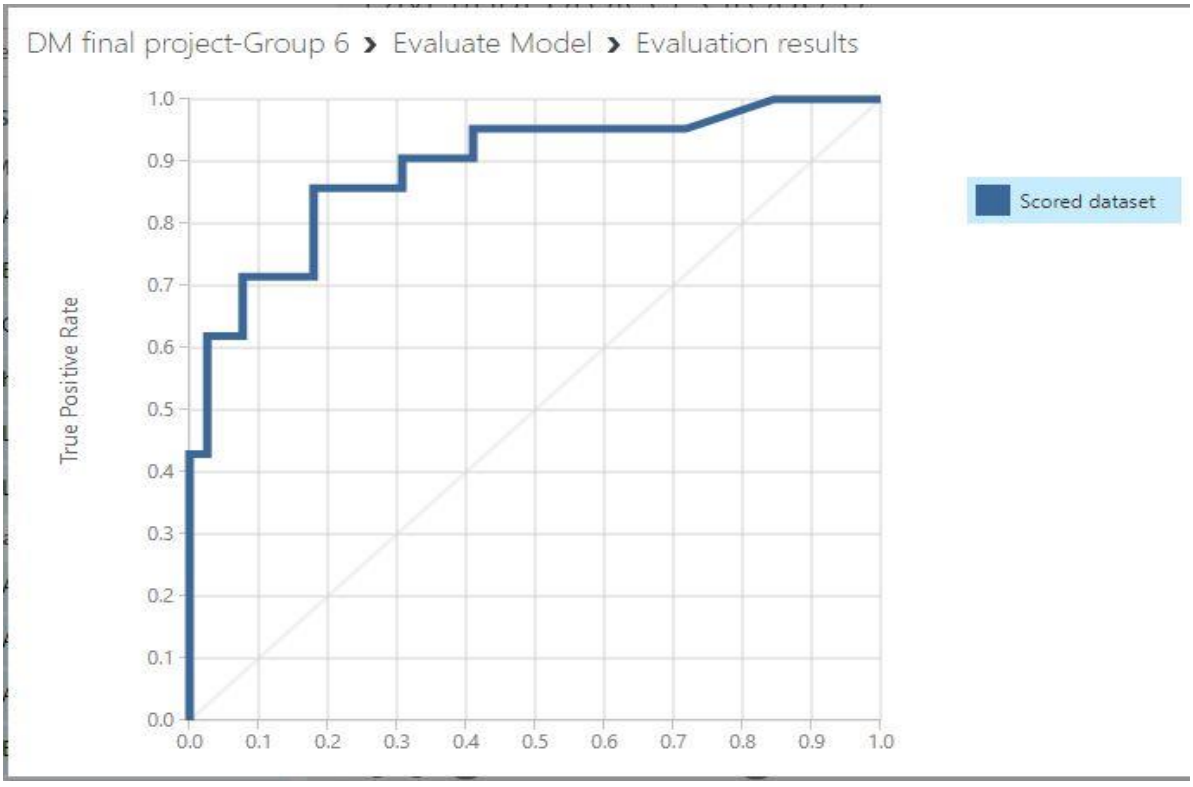
LIFT = Percentage of respondents in some sample/ percentage of respondents in a random sample of same size.

Two class Boosted Decision tree:



Parameters:

Default parameters: No. of leaves 20, no. of trees: 100



DM final project-Group 6 › Evaluate Model › Evaluation results

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
15	6	0.833	0.789	0.5	0.893

False Positive	True Negative	Recall	F1 Score
4	35	0.714	0.750

Positive Label	Negative Label
1	0

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	10	1	0.183	0.800	0.625	0.909	0.476	0.776	0.974	0.011
(0.800,0.900]	3	0	0.233	0.850	0.743	0.929	0.619	0.826	0.974	0.011
(0.700,0.800]	0	2	0.267	0.817	0.703	0.813	0.619	0.818	0.923	0.043
(0.600,0.700]	0	0	0.267	0.817	0.703	0.813	0.619	0.818	0.923	0.043
(0.500,0.600]	2	1	0.317	0.833	0.750	0.789	0.714	0.854	0.897	0.061
(0.400,0.500]	0	2	0.350	0.800	0.714	0.714	0.714	0.846	0.846	0.098
(0.300,0.400]	2	1	0.400	0.817	0.756	0.708	0.810	0.889	0.821	0.116
(0.200,0.300]	1	0	0.417	0.833	0.783	0.720	0.857	0.914	0.821	0.116

Changed threshold to 0.8 and observed the results:

We can see a slight improvement in the true negative results when the threshold was changed to 0.8. So the model predicted it right for those who are at no risk for CVD's, saving the patients an approximately $20,000 each, for the treatment.

Moreover the parameters like No. of leaves, tress, learning rate were changed and the results were observed, However there was no significant change in the confusion matrix or the AUC and F1 score.

DM final project-Group 6 › Evaluate Model › Evaluation results

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
13	8	0.850	0.929	0.8	0.893

False Positive	True Negative	Recall	F1 Score
1	38	0.619	0.743

Positive Label	Negative Label
1	0

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	10	1	0.183	0.800	0.625	0.909	0.476	0.776	0.974	0.011
(0.800,0.900]	3	0	0.233	0.850	0.743	0.929	0.619	0.826	0.974	0.011
(0.700,0.800]	0	2	0.267	0.817	0.703	0.813	0.619	0.818	0.923	0.043
(0.600,0.700]	0	0	0.267	0.817	0.703	0.813	0.619	0.818	0.923	0.043
(0.500,0.600]	2	1	0.317	0.833	0.750	0.789	0.714	0.854	0.897	0.061
(0.400,0.500]	0	2	0.350	0.800	0.714	0.714	0.714	0.846	0.846	0.098
(0.300,0.400]	2	1	0.400	0.817	0.756	0.708	0.810	0.889	0.821	0.116
(0.200,0.300]	1	0	0.417	0.833	0.783	0.720	0.857	0.914	0.821	0.116

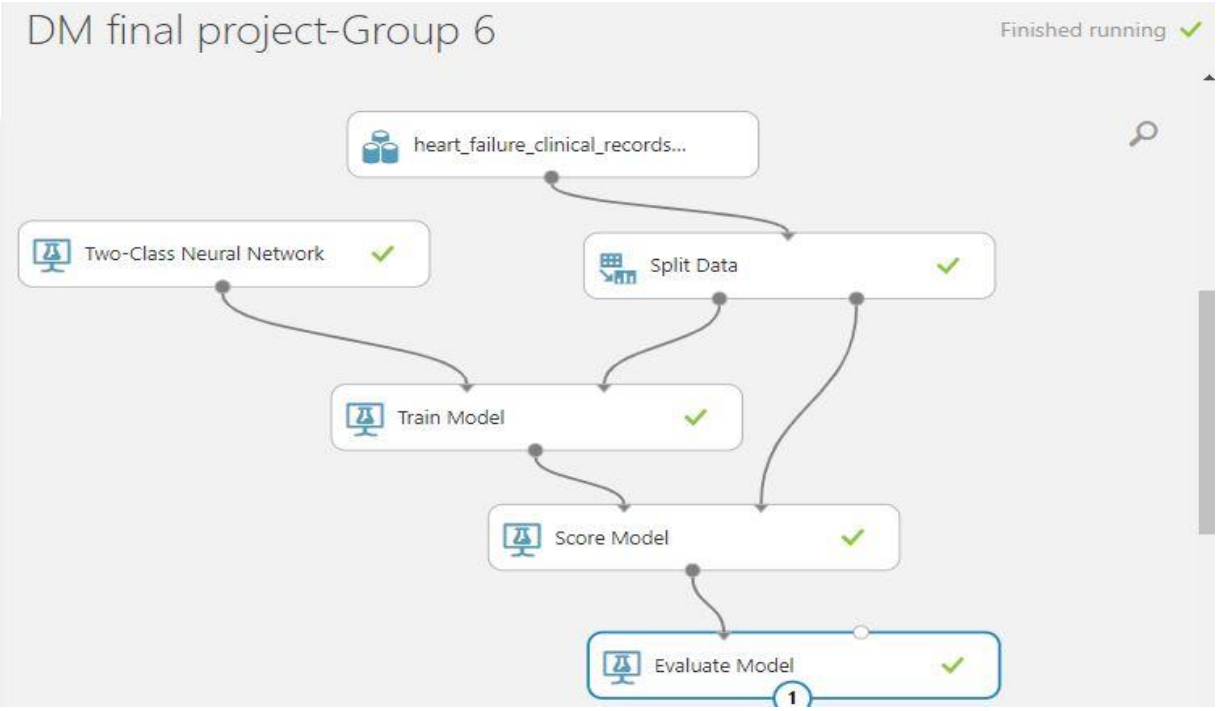
Threshold Analysis:

Threshold	Accuracy	Precision	Recall
0.5	0.833	0.789	0.714
0.8	0.850	0.929	0.619

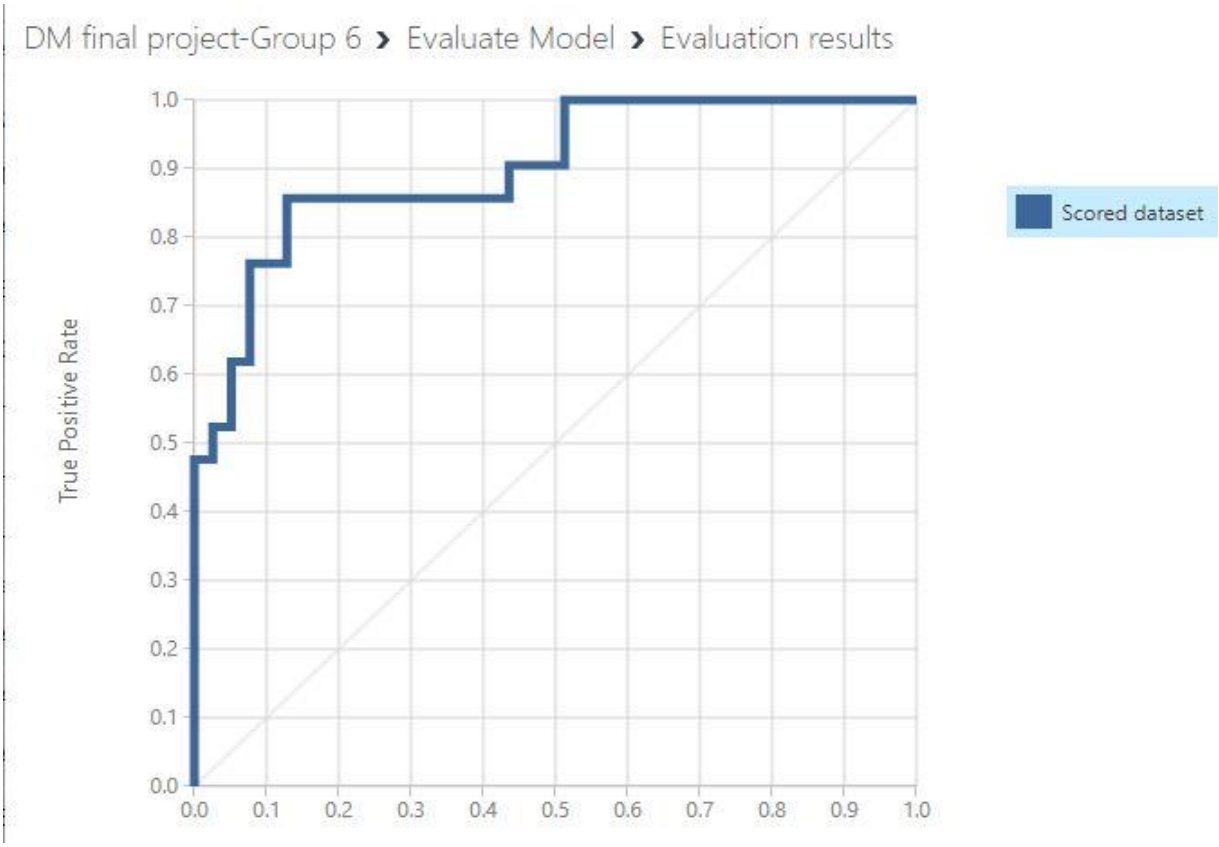
Threshold	FP	FN	TP	TN	TP+FP total
0.5	4	6	15	35	**19**
0.8	1	8	13	38	**14**

Comparing the models on 2 different thresholds and speaking from the Hospital's business perspective. They would be treating a total of 19 patients who are at risk of getting CVD's based on the 0.5 threshold model's positive prediction. The business would be making a total of approximately $380,000 yearly which is around $100,000 more than what the hospital would make if they followed the model that ran on 0.8 threshold treating only 14 patients. Moreover the Hospital is also losing 2 of its extra patients as the model predicted them as false negatives. However the tradeoff is that from a patient's perspective, 3 patients would unnecessarily spend a total of approximately $60,000 an year on their medical treatments if the 0.5 threshold model is followed.

Two class neural network:



Default Parameters: no of hidden nodes= 100, No. of learning iterations: 100



DM final project-Group 6 › Evaluate Model › Evaluation results

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
15	6	0.850	0.833	0.5	0.901

False Positive	True Negative	Recall	F1 Score
3	36	0.714	0.769

Positive Label	Negative Label
1	0

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	6	0	0.100	0.750	0.444	1.000	0.286	0.722	1.000	0.000
(0.800,0.900]	4	1	0.183	0.800	0.625	0.909	0.476	0.776	0.974	0.012
(0.700,0.800]	1	0	0.200	0.817	0.667	0.917	0.524	0.792	0.974	0.012
(0.600,0.700]	2	1	0.250	0.833	0.722	0.867	0.619	0.822	0.949	0.026
(0.500,0.600]	2	1	0.300	0.850	0.769	0.833	0.714	0.857	0.923	0.042
(0.400,0.500]	1	2	0.350	0.833	0.762	0.762	0.762	0.872	0.872	0.081
(0.300,0.400]	2	0	0.383	0.867	0.818	0.783	0.857	0.919	0.872	0.081

Changing parameters and comparing the results:

We can see a slight improvement in the True positive results when the No. of hidden layers where increased to 200. So, the model predicted it right for the patients who were at risk of having CVD's and the patients could start their treatment at early stage only. Moreover the model also provides with better F1 score and AUC.

DM final project-Group 6 › Evaluate Model › Evaluation results

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
16	5	0.867	0.842	0.5	0.905

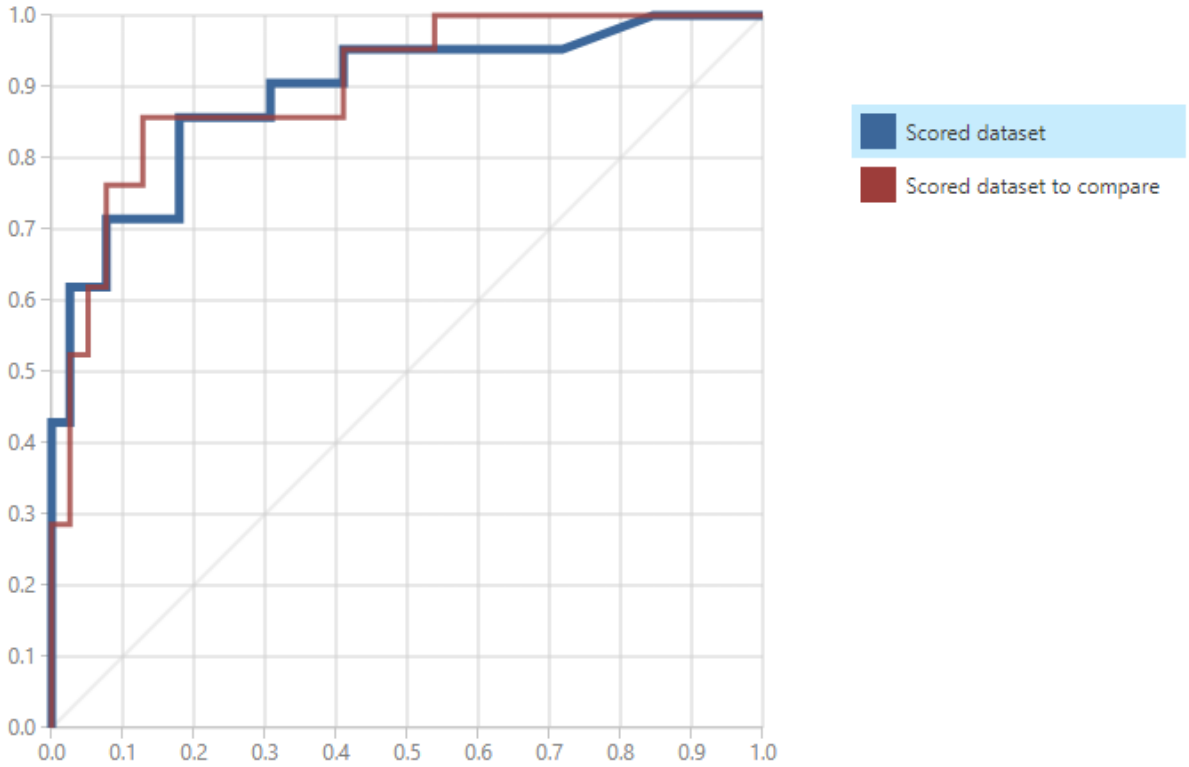
False Positive	True Negative	Recall	F1 Score
3	36	0.762	0.800

Positive Label	Negative Label
1	0

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	6	0	0.100	0.750	0.444	1.000	0.286	0.722	1.000	0.000
(0.800,0.900]	4	1	0.183	0.800	0.625	0.909	0.476	0.776	0.974	0.012
(0.700,0.800]	1	0	0.200	0.817	0.667	0.917	0.524	0.792	0.974	0.012
(0.600,0.700]	2	2	0.267	0.817	0.703	0.813	0.619	0.818	0.923	0.040
(0.500,0.600]	3	0	0.317	0.867	0.800	0.842	0.762	0.878	0.923	0.040
(0.400,0.500]	0	1	0.333	0.850	0.780	0.800	0.762	0.875	0.897	0.060
(0.300,0.400]	2	2	0.400	0.850	0.800	0.750	0.857	0.917	0.846	0.103
(0.200,0.300]	0	5	0.483	0.767	0.720	0.621	0.857	0.903	0.718	0.212

Comparing the 2 algorithms:

Two class boosted decision tree (Blue) vs Two class neural network (Red):



Conclusion:

The comparison between both algorithms with default parameters and a threshold of 0.5 shows the following results:
It shows that the Two-Class Neural Network is comparatively better to predict if a patient is at risk of a heart disease. There's a change in the No. of false positives. This change can look as something miniscule but, since this is a serious matter of a person's health condition. It means that one less person is being viewed as "At RISK" even though in reality is diagnosed negative of getting a heart disease. Moreover Area under the curve and F1 score are comparatively better in the Two-Class Neural Network algorithm.

Monetary Value:

Scenario	Description	Expected		
True Positive (TP)	At Risk, Model Predicted Risk	Approx. $20,000 per patient for the treatment of CVD.	**Positive**	**At Risk of CVD**
True Negative (TN)	No Risk, Model Predicted negative	$1,000 per patient for tests and diagnostics.	**Negative**	**No Risk of CVD**
False Positive (FP)	No Risk, Model Predicted they were at Risk	$1,000 for tests and diagnostics. Patient might waste $20,000 on treatment due to model's wrong prediction.		
False Negative (FN)	At Risk, Model Predicted they were at not Risk	Patient should be treated as they are at risk of getting CVD. Loss of $20,000 per patient for the hospital.		

For this data the monetary value varies with each patient and facility where the patient and tests might be involved. An average patient spends almost about $20,000 yearly on cardiovascular disease treatments and this price might increase depending on the severity of the disease. There are many factors that can make this amount increase or decrease but the most common factors involved are drugs, testing and appointments with medical professionals. In case of an emergency, additional costs might be added such as ambulances and other hospital charges. In case a patient is falsely diagnosed as not at risk, and they are at risk this can cause a major cost in terms for both the patient and medical facility.

References:

(DATASET): https://www.kaggle.com/andrewmvd/heart-failure-clinical-data

https://www.ajmc.com/view/ajmc_10marnicholswebx_e86to93

https://www.webmd.com/healthy-aging/features/heart-disease-medical-costs#1