ISM 6136 Data Mining

Project 2

Predicting Mortality from Heart Failure

Group 7

James Long

Andres Zambrano

Samiksha Mhatre

Background of the Problem

Heart failure is form of cardiovascular disease. It is a chronic, progressive condition in which the heart muscle is unable to pump enough blood to meet the body's needs for blood and oxygen. The heart cannot keep up with its workload. Heart failure is a serious condition that is usually incurable, but the condition can be managed (American Heart Association, n.d.). Numerous factors contribute to heart failure, genetics being one of them. This makes one wonder whether heart failure has always been a part of our humanity. Nonetheless, many other controllable factors play a part. In this paper, we attempt to predict whether patients with severe heart failure will die within a specified period. If such predictions can be made reliably, then the leading indicators can be used to better control the condition and extend life expectancy.

In 2009, the University of California at Irvine published a study of Egyptians mummies. They found evidence of cardiovascular diseases (University of California - Irvine, 2009). Descriptions of cardiovascular diseases exist from ancient Greece and India as well. Despite this, the 1900s mark the period when people became interested in the scientific study of cardiovascular disease. In 1915, the Association for the Prevention and Relief of Heart Disease was formed in New York City, which comprised a group of social workers and doctors. In the years that followed, other similar groups formed in other U.S. cities. In 1924, multiple such heart associations merged to become the American Heart Association (American Heart Association, n.d.). As time went by, doctors started experimenting more with cardiovascular diseases.

Studies of the epidemiology of cardiovascular disease have been complicated by the lack of universal agreement on a definition of cardiovascular disease, which is primarily a clinical diagnosis. National and international comparisons have therefore been difficult, and mortality data, postmortem studies, and hospital admission rates are not easily translated into incidence and prevalence. Several different systems have been used in large population studies, with the use of scores for clinical features determined from history and examination, and in most cases chest radiography, to define heart failure (Davis, Hobbs, & Lip, 2000). Heart failure cannot be cured, but according to statistics, proper treatment and early diagnosis have helped people with heart failure live longer (American Heart Association, n.d.). New treatments have been introduced resulting in a reduction of mortality rate due to heart failure because of our better understanding of heart failure.

Motivation for Solving the Problem

The societal and personal costs of cardiovascular disease are extremely high. According to the Centers for Disease Control and Prevention:

> Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States. One person dies every 36 seconds in the United States from cardiovascular disease. About 655,000 Americans die from heart disease each year—that's 1 in every 4 deaths. Heart disease costs the United States about $219 billion each year from 2014 to 2015. This includes the cost of health care services, medicines, and lost productivity due to death (Centers for Disease Control and Prevention, 2020).

According to the AHA, the costs of cardiovascular disease in the United States will rise from $555 billion in 2016 to $1.1 trillion by 2035 (American Heart Association, n.d.). Irrespective of our society's focus on healthy living, the situation seems to be getting worse. Prevention of heart failure has become an urgent public health need globally. Because heart failure is commonly the result of an acute and chronic cardiac injury that can be prevented with aggressive risk factor management, there is a critical need to examine our current approach to this enormous health threat. Much of the knowledge about the epidemiology, risk factors, prognosis, treatment, and prevention of heart failure is based on North American and European studies. The prevalence and burden of heart failure will likely continue to increase in developed countries, where better care has improved survival with cardiovascular conditions such as myocardial infarction and heart failure. Although survival in clinical trials is improving, heart failure remains a lethal condition in the population with an estimated 287,000 deaths per year in the United States (Emory Healthcare, 2019).

The World Health Organization (WHO) estimates that simple measures could advance life expectancy by five to 10 years and would dramatically reduce global disparities in life expectancy. Three of the 10 items on the WHO's list of simple measures would, directly and indirectly, prevent heart failure. These include control programs aimed at blood pressure, cholesterol, and tobacco. In summary, the next decades will offer tremendous opportunities for advancing the prevention of heart failure (Schocken, et al., 2008). To seize the opportunities, healthcare professionals need to know which risk factors are most relevant to prevention and management. That is the central question we seek to answer in this paper, subject to the limitations of our data set.

Solution Methodology and Evaluation Metrics

The nature of the problem is one of prediction. Given a data set containing various patient measurements taken from blood reports and physician examinations, and given a timeframe of interest, can death be reliably predicted to occur within the timeframe? For such a question, the outcome is binary (death or survival). Thus, we employed two-class classification algorithms. We pitted five algorithms against other to see which would perform best. They included Two-Class Support Vector Machine, Two-Class Boosted Decision Tree, Two-Class Decision Forest, Two-Class Neural Network, and Two-Class Logistic Regression. We constructed multiple models with all features and then repeated the exercise with selected features that demonstrated higher Pearson correlations with the outcome event. The selected features were Event, Ejection.Fraction, Creatinine, Time, Age, and Sodium. Our hope was to narrow the list of relevant patient measurements, thereby simplifying the tasks of prevention and management for healthcare professionals. Upon viewing the results, we decided to refilter the features based on the final model parameters of the best performing model, which came from the unfiltered data set. The selected features were Event, Ejection.Fraction, Creatinine, Time, Age, and CPK. Thus, we created three sets of competing models using different features. For all models, we split the data set 80/20 for training and testing with no validation subset.

We considered several metrics. Falsely predicting a negative could lead to patient death because the patient would not be given correct guidance to prevent heart failure or manage their condition. Thus, our highest priority was minimizing false negatives (or maximizing recall). False positives could lead to patients making lifestyle changes that are unnecessary. This is not ideal and could undermine patient confidence in healthcare professionals if the fault in the model were ever discovered and made public. So, our second highest priority was minimizing false positives (or maximizing specificity). The data set was slightly imbalanced with a ratio of negative to positive outcomes of approximately 2:1. Still, we considered accuracy to be a valid metric. However, accuracy is reduced by false positives and false negatives equally, so accuracy did not serve our needs as well as recall. We included accuracy in our results as a convenient summary metric for overall performance albeit lower weighted. Similarly, precision, F1 score, and area under the curve (AUC) were less useful to us due to their focus on false positives rather than false negatives.

Description of Our Dataset

The data set we selected has 12 features describing patients with class III or class IV heart failure. Of the four classes defined by the New York Heart Association, classes III and IV are the most severe. As such, these patients were at elevated risk of death. A 13th feature indicates survival of each patient during the period of study. The data set was originally used to predict the mortality of the patients in the period following diagnosis by using standard biostatistics methods. There are 299 patients in the data set of which 105 are women, and 194 are men. Survivors outnumber the deceased 203 to 96. We discovered the data set via Kaggle, but it was originally published in the Public Library of Science on figshare (Ahmad, Munir, Bhatti, Aftab, & Raza, 2017). The 13 features are described in Table 1.

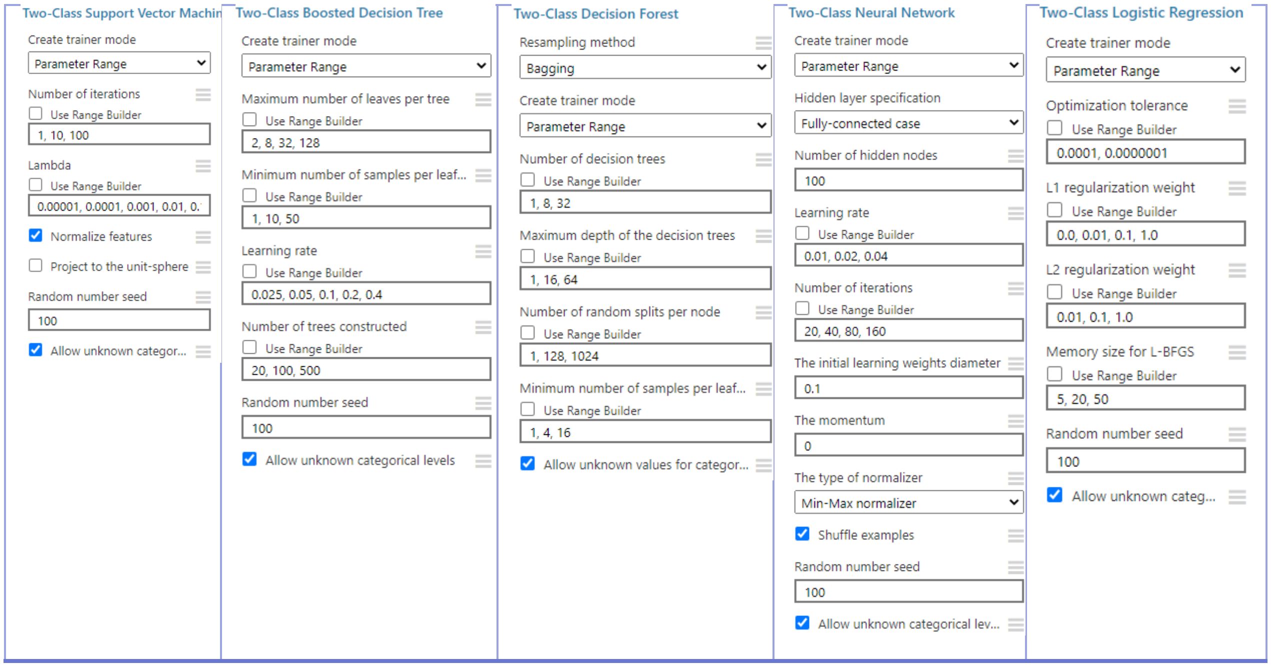
Table 1: Heart Failure Patient Data

Feature Name	Feature Description	Type	Values
Event	Did death occur within the total follow up period of 285 days	Binary	0 = survival 1 = death
Time	How many days after diagnosis did the patient have a medical follow up	Numeric	4-285
Gender	Patient gender	Binary	0 = female 1 = male
Smoking	Is the patient a smoker	Binary	0 = no 1 = yes
Diabetes	Does the patient have diabetes	Binary	0 = no 1 = yes
BP	Does the patient have high blood pressure (hypertension)	Binary	0 = no 1 = yes
Anaemia	Does the patient have low levels of red blood cells or hemoglobin	Binary	0 = no 1 = yes
Age	Age of patient in years	Numeric	40-95
Ejection.Fraction	Percentage of blood leaving the ventricle chamber per contraction	Numeric	14-80
Sodium	Blood sodium level mEg/L	Numeric	114-148
Creatinine	Blood creatinine level mg/dL	Numeric	0.50-9.40
Pletelets	Kiloplatelets per mL of blood	Numeric	25.01-850.00
CPK	Blood Creatinine phosphokinase level mcg/L	Numeric	23-7861

Comparison of Algorithms

For each of the five algorithms we used, we set the trainer mode to Parameter range and used the default settings. We then paired each algorithm with a Tune Model Hyperparameters module. For all tuning modules, we set the parameter sweeping mode to Random Sweep, increased the Maximum Number of Runs from five to 10, set Random Seed to 100, and changed the Metric for Measuring Performance for Classification from Accuracy to Recall. The algorithm parameters are captured in Table 2.

Table 2: Algorithm Parameters



Two-Class Support Vector Machin

Create trainer mode: Parameter Range

Number of iterations
□ Use Range Builder
1, 10, 100

Lambda
□ Use Range Builder
0.00001, 0.0001, 0.001, 0.01, 0.

☑ Normalize features

□ Project to the unit-sphere

Random number seed
100

☑ Allow unknown categor...

Two-Class Boosted Decision Tree

Create trainer mode: Parameter Range

Maximum number of leaves per tree
□ Use Range Builder
2, 8, 32, 128

Minimum number of samples per leaf...
□ Use Range Builder
1, 10, 50

Learning rate
□ Use Range Builder
0.025, 0.05, 0.1, 0.2, 0.4

Number of trees constructed
□ Use Range Builder
20, 100, 500

Random number seed
100

☑ Allow unknown categorical levels

Two-Class Decision Forest

Resampling method: Bagging

Create trainer mode: Parameter Range

Number of decision trees
□ Use Range Builder
1, 8, 32

Maximum depth of the decision trees
□ Use Range Builder
1, 16, 64

Number of random splits per node
□ Use Range Builder
1, 128, 1024

Minimum number of samples per leaf...
□ Use Range Builder
1, 4, 16

☑ Allow unknown values for categor...

Two-Class Neural Network

Create trainer mode: Parameter Range

Hidden layer specification: Fully-connected case

Number of hidden nodes
100

Learning rate
□ Use Range Builder
0.01, 0.02, 0.04

Number of iterations
□ Use Range Builder
20, 40, 80, 160

The initial learning weights diameter
0.1

The momentum
0

The type of normalizer: Min-Max normalizer

☑ Shuffle examples

Random number seed
100

☑ Allow unknown categorical lev...

Two-Class Logistic Regression

Create trainer mode: Parameter Range

Optimization tolerance
□ Use Range Builder
0.0001, 0.0000001

L1 regularization weight
□ Use Range Builder
0.0, 0.01, 0.1, 1.0

L2 regularization weight
□ Use Range Builder
0.01, 0.1, 1.0

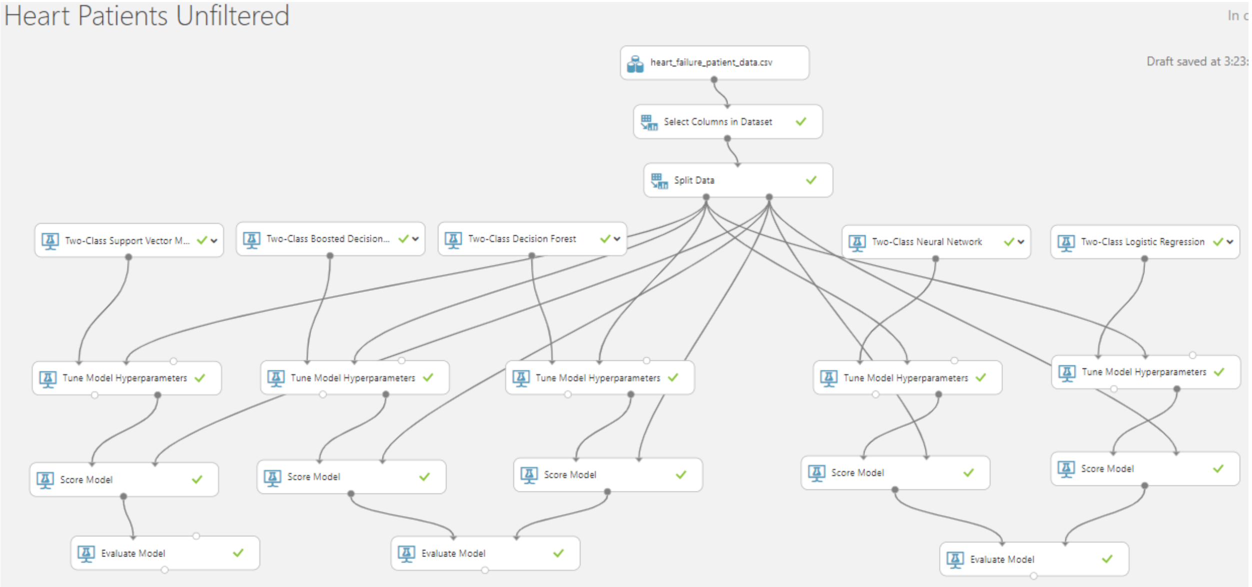
Memory size for L-BFGS
□ Use Range Builder
5, 20, 50

Random number seed
100

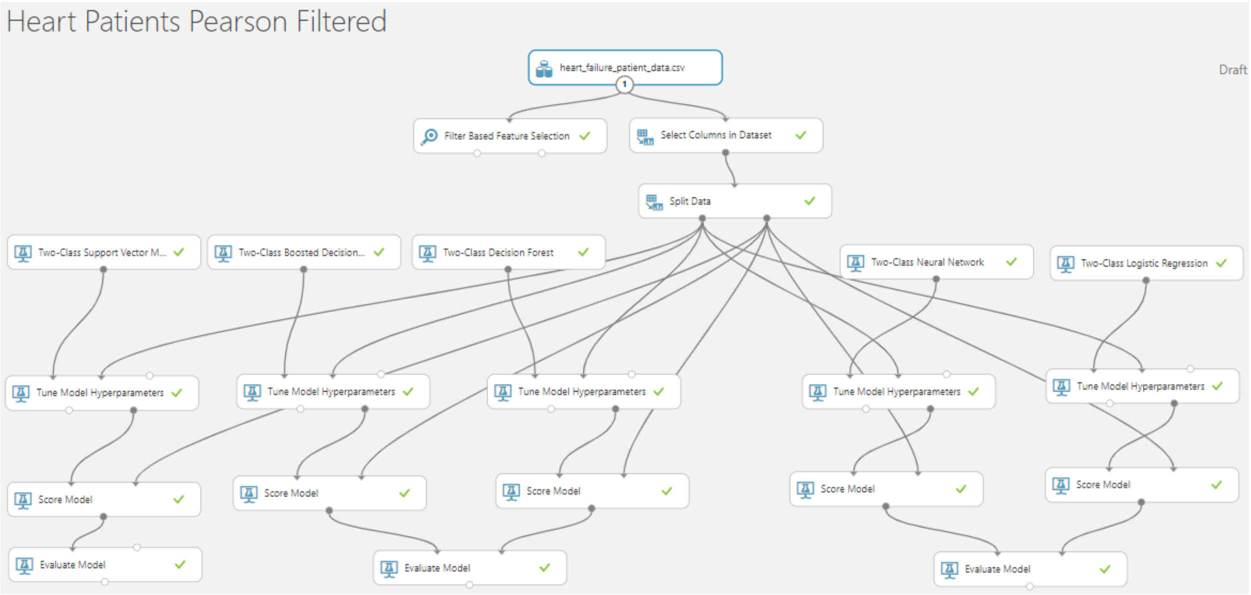
☑ Allow unknown categ...

With this setup, the tuning module selected multiple random combinations of parameter values from the parameter range defined for each algorithm. It then conducted a run with each combination of parameter values up to the maximum number of runs we specified (10). This setup automated the laborious task of determining the optimal combination of values for each algorithm. Screenshots of the experiments follow.

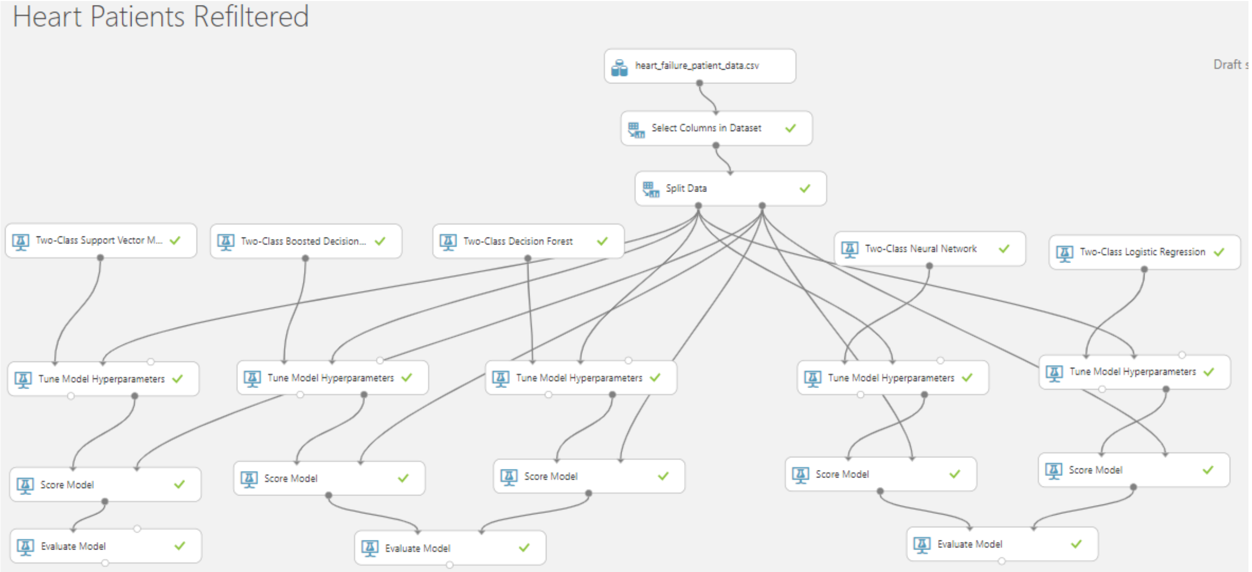
Screenshot 1: Unfiltered Data Set


Heart Patients Unfiltered
In c
Draft saved at 3:23:
heart_failure_patient_data.csv
Select Columns in Dataset
Split Data
Two-Class Support Vector M...
Two-Class Boosted Decision...
Two-Class Decision Forest
Two-Class Neural Network
Two-Class Logistic Regression
Tune Model Hyperparameters
Tune Model Hyperparameters
Tune Model Hyperparameters
Tune Model Hyperparameters
Tune Model Hyperparameters
Score Model
Score Model
Score Model
Score Model
Score Model
Evaluate Model
Evaluate Model
Evaluate Model

Screenshot 2: Pearson Filtered Data Set

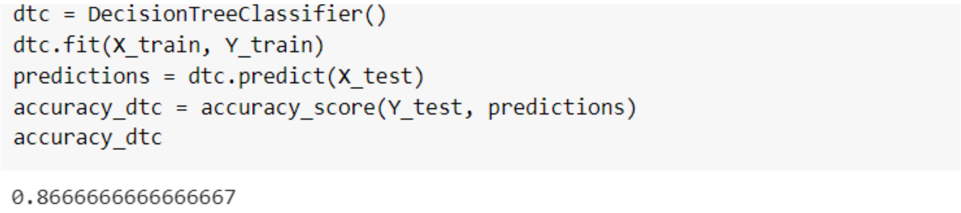

Heart Patients Pearson Filtered
Draft
heart_failure_patient_data.csv
1
Filter Based Feature Selection
Select Columns in Dataset
Split Data
Two-Class Support Vector M...
Two-Class Boosted Decision...
Two-Class Decision Forest
Two-Class Neural Network
Two-Class Logistic Regression
Tune Model Hyperparameters
Tune Model Hyperparameters
Tune Model Hyperparameters
Tune Model Hyperparameters
Tune Model Hyperparameters
Score Model
Score Model
Score Model
Score Model
Score Model
Evaluate Model
Evaluate Model
Evaluate Model

Screenshot 3: Refiltered Data Set


Heart Patients Refiltered
heart_failure_patient_data.csv
Draft s
Select Columns in Dataset
Split Data
Two-Class Support Vector M...
Two-Class Boosted Decision...
Two-Class Decision Forest
Two-Class Neural Network
Two-Class Logistic Regression
Tune Model Hyperparameters
Tune Model Hyperparameters
Tune Model Hyperparameters
Tune Model Hyperparameters
Tune Model Hyperparameters
Score Model
Score Model
Score Model
Score Model
Score Model
Evaluate Model
Evaluate Model
Evaluate Model

For comparison, we also generated a Python script to train four prediction models. The experiment includes a Decision Tree classifier, Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) classifier. For Logistic Regression, we scaled the independent variables to obtain a higher accuracy. We also performed a grid search on the SVM parameters to train the best performing model. Based on the accuracy score for each of the models, we concluded that the best performing model was the Decision Tree with an accuracy of 87% followed by the SVM model with an accuracy of 81%.

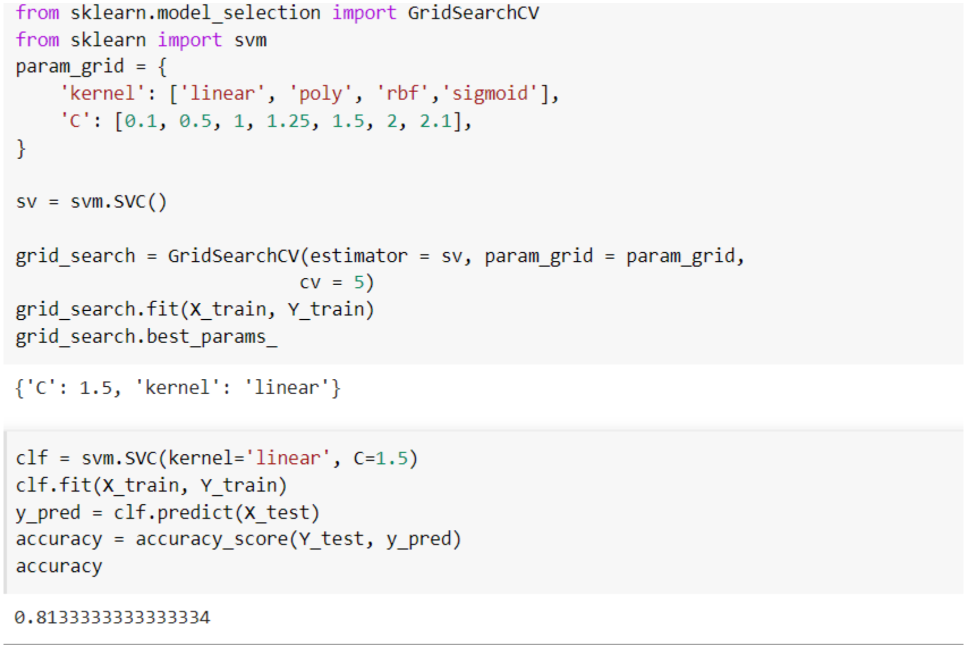
Screenshot 4: Decision Tree Classifier Python Code



```
dtc = DecisionTreeClassifier()
dtc.fit(X_train, Y_train)
predictions = dtc.predict(X_test)
accuracy_dtc = accuracy_score(Y_test, predictions)
accuracy_dtc
```

```
0.8666666666666667
```

Screenshot 5: SVM Python Code



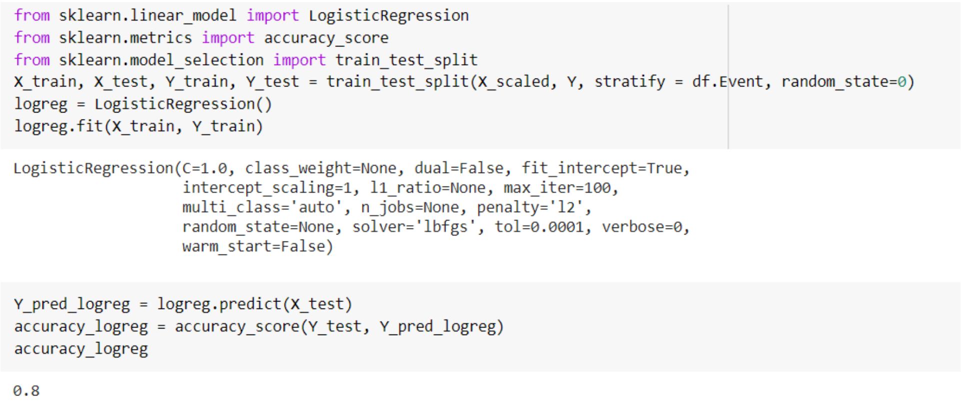
```
from sklearn.model_selection import GridSearchCV
from sklearn import svm
param_grid = {
    'kernel': ['linear', 'poly', 'rbf','sigmoid'],
    'C': [0.1, 0.5, 1, 1.25, 1.5, 2, 2.1],
}

sv = svm.SVC()

grid_search = GridSearchCV(estimator = sv, param_grid = param_grid,
                          cv = 5)
grid_search.fit(X_train, Y_train)
grid_search.best_params_
```

```
{'C': 1.5, 'kernel': 'linear'}
```

```
clf = svm.SVC(kernel='linear', C=1.5)
clf.fit(X_train, Y_train)
y_pred = clf.predict(X_test)
accuracy = accuracy_score(Y_test, y_pred)
accuracy
```

```
0.8133333333333334
```

Screenshot 6: Logistic Regression Python Code



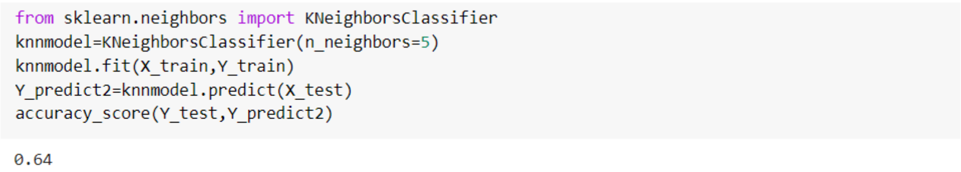
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X_scaled, Y, stratify = df.Event, random_state=0)
logreg = LogisticRegression()
logreg.fit(X_train, Y_train)
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=None, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)
```

```
Y_pred_logreg = logreg.predict(X_test)
accuracy_logreg = accuracy_score(Y_test, Y_pred_logreg)
accuracy_logreg
```

```
0.8
```

Screenshot 7: KNN Classifier Python Code



```
from sklearn.neighbors import KNeighborsClassifier
knnmodel=KNeighborsClassifier(n_neighbors=5)
knnmodel.fit(X_train,Y_train)
Y_predict2=knnmodel.predict(X_test)
accuracy_score(Y_test,Y_predict2)
```

```
0.64
```

Summary Sheet

The best performing model with the unfiltered data set outperformed the best performing model with each of the filtered data sets. This outcome means we were unable to simultaneously improve model performance and reduce model complexity. However, this might be achievable with further experimentation of feature selection. We used a threshold of 0.5 for all models. Table 3 summarizes the evaluation metrics for the unfiltered data set, while Table 4 summarizes the Pearson filtered data set, and Table 5 summarizes the refiltered data set. The best performing model in each table is highlighted.

Table 3: Unfiltered Data Performance Metrics

	True Positive	**False Negative**	**Recall**	**True Negative**	**False Positive**	**Specificity**	**Accuracy**
Two-Class Support Vector Machine	12	2	0.857	44	2	0.957	0.933
Two-Class Boosted Decision Tree	12	2	0.857	43	3	0.935	0.917
Two-Class Decision Forest	12	2	0.857	43	3	0.935	0.917
Two-Class Neural Network	9	5	0.643	46	0	1	0.917
Two-Class Logistic Regression	11	3	0.786	43	3	0.935	0.900

Table 4: Pearson Filtered Data Performance Metrics

	True Positive	**False Negative**	**Recall**	**True Negative**	**False Positive**	**Specificity**	**Accuracy**
Two-Class Support Vector Machine	11	3	0.786	44	2	0.957	0.917
Two-Class Boosted Decision Tree	12	2	0.857	43	3	0.935	0.917
Two-Class Decision Forest	11	3	0.786	42	4	0.913	0.883
Two-Class Neural Network	9	5	0.643	46	0	1	0.917
Two-Class Logistic Regression	11	3	0.786	43	3	0.935	0.900

Table 5: Refiltered Data Performance Metrics

	True Positive	False Negative	Recall	True Negative	False Positive	Specificity	Accuracy
Two-Class Support Vector Machine	11	3	0.786	43	3	0.935	0.900
Two-Class Boosted Decision Tree	12	2	0.857	43	3	0.935	0.917
Two-Class Decision Forest	11	3	0.786	43	3	0.935	0.900
Two-Class Neural Network	10	4	0.714	45	1	0.978	0.917
Two-Class Logistic Regression	11	3	0.786	43	3	0.935	0.900

The final model parameters of the overall best performing model are summarized in Table 6 and Table 7.

Table 6: Two-Class SVM Settings

	Value
Num Iterations	56
Lambda	0.009685557
Normalize Features	True
Perform Projection	False
Allow Unknown Levels	True
Random Number Seed	100

Table 7: Two-Class SVM Feature Weights

	Weight
Time	-2.6444
Ejection.Fraction	-1.68955
Age	1.19534
Creatinine	1.10545
CPK	0.65141
Bias	0.627795
Sodium	-0.358741
Smoking	-0.214102
Anaemia	0.129778
Gender	0.129723
Pletelets	0.12356
Diabetes	0.0993895
BP	-0.00773669

Conclusions

Our models were all limited by the data set. Though we were able to train a model with recall of 85.7% and accuracy of 93.3%, we cannot assert the model would generalize well. The first concern is that all the data were collected from patients in Faisalabad, Pakistan. The local dietary habits, lifestyle, genetic diversity, environmental exposures, etc. may vary significantly from other locations. And each of these factors (among others) could impact patient health. Next, the outcome event showed very low correlation with factors such as diabetes and sodium that are known to be correlated with cardiovascular disease. Finally, the small sample size of just 299 records potentially introduces statistical issues.

Before any healthcare organization or professional uses our model, we recommend further analysis be conducted. Using a larger data set is strongly advised. Likewise, gathering a more diverse data set from different countries with a broader range of ages, ethnic groups, etc. would be wise. On such a data set, clustering techniques might be useful prior to prediction. It may very well be the case that different patient groups have different leading indicators of heart failure induced mortality. Furthermore, other patient measures may need to be included. The best leading indicator might not be present in our limited data set. Once the model is made robust and reliable, guidance could be developed and promulgated by organizations such as the AHA and the WHO. Armed with this information, physicians would be better able to assist their patients in preventing and managing heart failure.

References

Ahmad, T., Munir, A., Bhatti, S., Aftab, M., & Raza, M. (2017). *DATA_MINIMAL*, 1. Retrieved from PLOS ONE: https://doi.org/10.1371/journal.pone.0181001.s001

American Heart Association. (n.d.). *History of the American Heart Association*. Retrieved from American Heart Association: https://www.heart.org/en/about-us/history-of-the-american-heart-association

American Heart Association. (n.d.). *Out of shape and middle-aged? It's not too late to turn it around*. Retrieved from American Heart Association: https://www.heart.org/en/news/2018/09/27/out-of-shape-and-middle-aged-its-not-too-late-to-turn-it-around

American Heart Association. (n.d.). *Treatment Options for Heart Failure*. Retrieved from American Heart Association: https://www.heart.org/en/health-topics/heart-failure/treatment-options-for-heart-failure

American Heart Association. (n.d.). *What is Heart Failure?* Retrieved from American Heart Association: https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure

Centers for Disease Control and Prevention. (2020, September 8). *Heart Disease Facts*. Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/heartdisease/facts.htm

Davis, R. C., Hobbs, F., & Lip, G. (2000). History and epidemiology. *British Medical Journal*.

Emory Healthcare. (2019). *Heart Failure Statistics*. Retrieved from Emory Healthcare: https://www.emoryhealthcare.org/heart-vascular/wellness/heart-failure-statistics.html

Schocken, D., Benjamin, E., Fonarow, G., Krumholz, H., Levy, D., Mensah, G., . . . Hong, Y. (2008, April 7). Prevention of Heart Failure. *Circulation, 117*, 2544–2565. doi:https://doi.org/10.1161/CIRCULATIONAHA.107.188965

University of California - Irvine. (2009, November 17). *Heart disease found in Egyptian mummies*. Retrieved from ScienceDaily: www.sciencedaily.com/releases/2009/11/091117161017.htm