Myocardial Infarction Complications

By Natalie Blanco, Alexandria Garcia, Melani Rodriguez, Susana Chavez, and Dylan Wegmann

Introduction

Myocardial infarction, also commonly known as a heart attack. is caused by a blood clot in the arteries of the heart. Clots cause a disturbance in the blood supply to the heart, leading to a loss of oxygen to the heart muscle causing severe damage and tissue death. Blockage is caused by a buildup of plaque in the arteries, which is formed by cholesterol and other deposits.

Factors that put people at risk of having a heart attack include: hypertension, obesity, smoking, alcohol abuse, Type 1 and Type 2 diabetes, high levels of low-density lipoprotein (LDL) cholesterol in blood, family history of heart disease, old age, excessive stress, and a diet high in saturated fats.

Heart attacks are the number one cause of death globally. Individuals who suffer from a heart attack can present different complications, such as atrial fibrillation, ventricular tachycardia, and pulmonary edema after suffering a heart attack. Those that DO present complications can have a worsening of the condition that may be fatal, while others can have complications but suffer no adverse effects on their long term health.

The intrinsic properties of heart attacks make it difficult to predict the onset of complications to where even the most experienced specialist is unable to anticipate their development. Foreseeing the development of complications following a heart attack is key to taking the necessary preventative measures for the individual to avoid the worsening of the condition or possible death.

Data

The MI data consists of 124 variables. Columns 2-112 are input data for prediction and columns 113-124 are possible complications, with four possible times a complication prediction could be made.

1. Initial admission time to hospital: all input columns used for prediction except 93, 94, 95, 100, 101, 102, 103, 104, 105
2. 24 hours after admission into the hospital: all input columns used for prediction except 93, 94, 95, 101, 102, 104, 105
3. 48 hours after admission to the hospital: all input columns used for prediction except 95, 102, 105
4. 72 hours after admission to the hospital: all input columns can be used for prediction

The data was cleaned of any “?” characters and missing values for each column imputed by the median. The following list includes attribute variables and their descriptions that had 15% or more missing values.

- Systolic blood pressure according to Emergency Cardiology Team (S_AD_KBRIG) (mmHg)

- Diastolic blood pressure according to Emergency Cardiology Team (D_AD_KBRIG) (mmHg)
- Systolic blood pressure according to intensive care unit (S_AD_ORIT) (mmHg)
- Diastolic blood pressure according to intensive care unit (D_AD_ORIT) (mmHg)
- Hypokalemia (< 4 mmol/L) (GIPO_K)
    - Extremely low potassium levels in blood, yes or no
- Serum potassium content (K_BLOOD) (mmol/L)
    - Specific blood potassium concentrations
- Increase of sodium in serum (more than 150 mmol/L) (GIPER_Na)
    - High sodium levels, either yes or no
- Serum sodium content (Na_BLOOD) (mmol/L)
    - Specific blood sodium concentrations
- Serum AlAT content (ALT_BLOOD) (IU/L)
    - Alanine aminotransferase - enzyme found in liver, high levels indicate possible liver disease
- Serum AsAT content (AST_BLOOD) (IU/L)
    - Aspartate aminotransferase - enzyme found in liver, high levels indicate possible liver disease
- Serum CPK content (KFK_BLOOD) (IU/L)
    - Creatine phosphokinase - enzyme in heart muscle, high levels indicate stress/injury to heart
- Use of opioid drugs by the Emergency Cardiology Team (NA_KB)
    - Medical opioids can be prescribed as pain medications
- Use of NSAIDs by the Emergency Cardiology Team (NOT_NA_KB)
    - Nonsteroidal anti-inflammatory drugs to treat pain and prevent blood clots
- Use of lidocaine by the Emergency Cardiology Team (LID_KB)
    - Anesthetic and can treat irregular heartbeats (arrhythmias)

Exploratory Analysis

Within our exploratory analysis, we outlined two main objectives; those being to 1) to use correlation plot to generate some hypotheses for numeric variables; and 2) to explore variables who appear to be good indicators of MI complication. We began the analysis by generating a heatmap to identify variables that are positively or negatively correlated with each other. As seen in figure 1, none are shown to be significant since there are no strong or negative correlations.

Not only was our data significantly multivariate, but the variables themselves posed a potential issue. That being that the data only had 12 variables which were numeric and of those 12, only 3 had enough observations to be analyzed. In an attempt to clean our data, we set a cutoff where variables with 15% or more of the data were missing. We did not analyze 9 of the 12 numeric variables that fit this cutoff, as well as 5 categorical variables that fit the cutoff, to avoid biases since they did not provide enough information. After completing the analysis, no specific feature variables stood out as being significant predictor variables of MI complications. Thus, based on previous research, the ages, gender and incidence of diabetes variables were explored. These three variables have been shown to be good indicators of whether or not a patient would experience an MI complication.

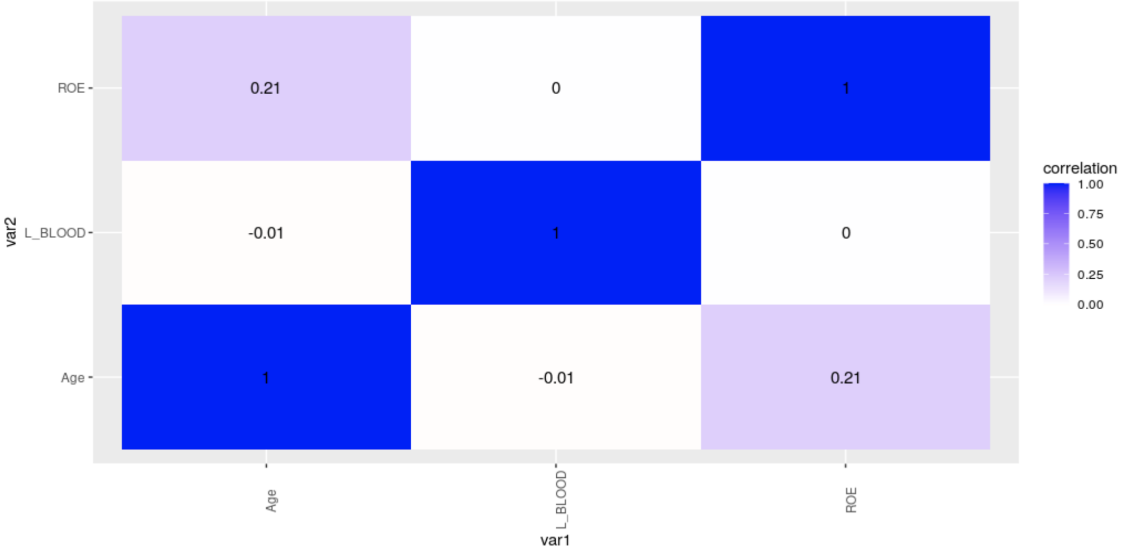

Figure 1: Correlation heatmap of numeric variables: age, white blood cell count, and red blood cell deposition, shows no strong correlation to one another.

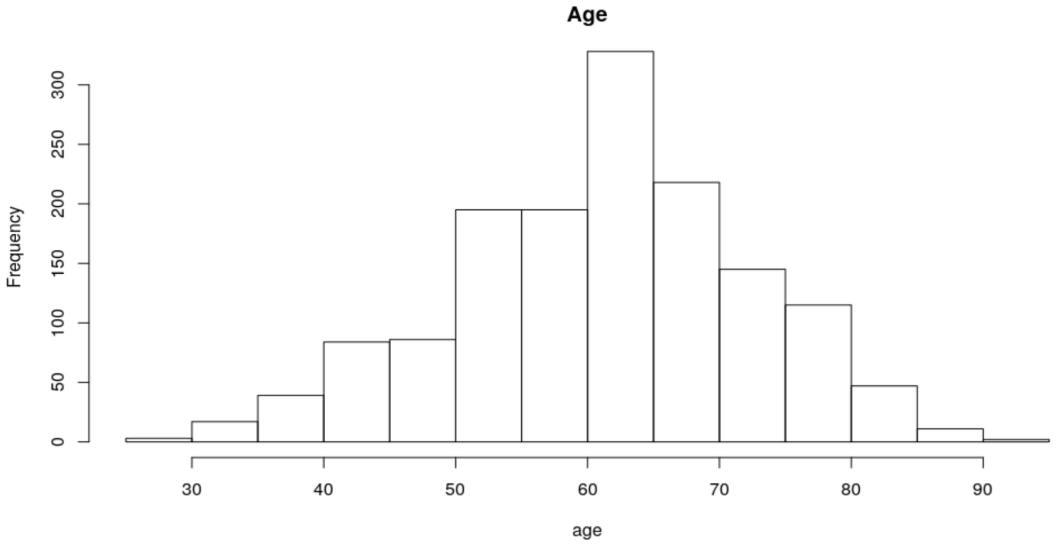

Figure 2: Histogram of patients displaying age, with the mean age approximately being 60 years old.

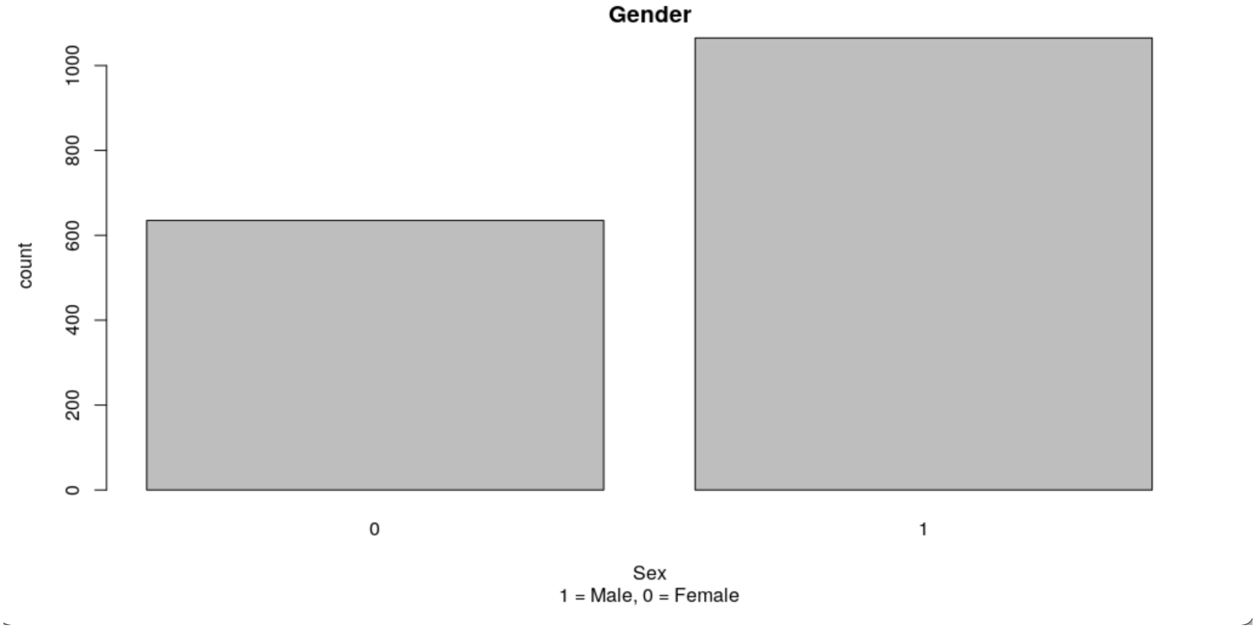

Figure 3: barplot of the patient's gender. Most individuals who experienced MI appear to be male.

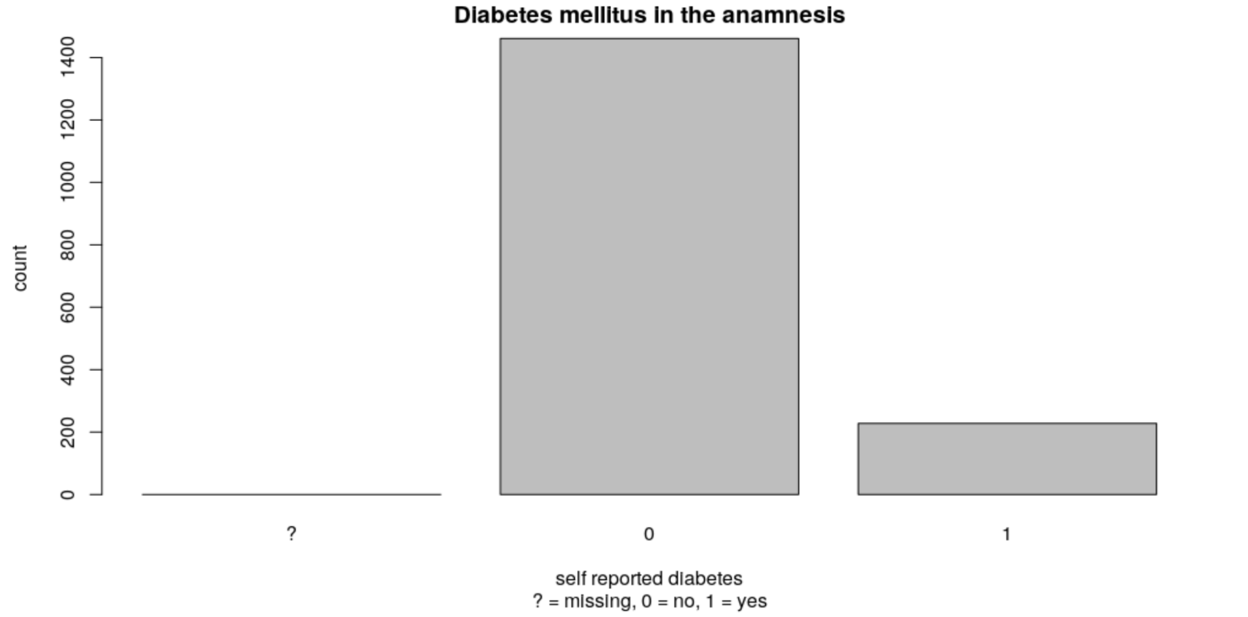

Figure 4: barplot of individuals who reported having diabetes. Most individuals who experienced MI did not report having diabetes.

Modeling

After exploring several different classification models (notably kNN, decision trees, and neural networks) we decided to model our data with the SVM classification model. Most other models explored were ultimately decided not to work well as they cannot handle multiclass-multi output datasets like we worked with. The Support Vector Machine model is a binary classification model that attempts to separate a dataset into two major classes. Then, the objective is to find the hyperplane ("an n-dimensional Euclidean space is a flat, n-1 dimensional subset of that space that divides the space into two disconnected parts" (Pupale, 2018)) that separates the two classes as much as possible in space. This is known as the margin distance. Determining the max margin distance ensures more confidence for future data point classification.

The SVM classification model was selected for our dataset, as the set was quite robust. More specifically, the dataset contained 97 independent variables (after cleaning the data) and 12 dependent variables. This immense number of variables in itself makes it easy to understand why selecting a good model was difficult. In our initial investigation, the main problem we faced was that our exploratory analysis did not point to any one feature variable being a strong predictor of any of the 12 outcome variables. Thus, we instead chose to individually examine each outcome variable in comparison to the combined effects of every independent variable using SVM. This decision was ultimately made since SVM works well with binary inputs, like those that we chose.

Using this model, we were able to obtain the accuracy scores of SVM for each dependent variable. Then, confusion matrices were constructed to assess the validity of such accuracy scores. With this combined information, some marginally significant conclusions can be made from this classification. Most notably, it was determined that the combined effect of each feature variable could most accurately predict myocardial infarction complications associated with tachycardia. The accuracy scores for supraventricular tachycardia and ventricular tachycardia were 0.99 and 0.98, respectively. Further, the associated confusion matrices contained the highest amount of true negatives. Upon interpretation, this means that for patients who have suffered a heart attack, we can accurately predict when they will not have the associated complication of either supraventricular tachycardia or ventricular tachycardia. Another notable conclusion was for the complication of chronic heart failure. Not only was its accuracy score the lowest of all outcome variables, but its confusion matrix also contained the highest amount of false negatives. This means that for many patients in which the complication of chronic heart failure is predicted to not happen, actually does happen. From this, we can determine that this particular complication is rather hard to predict and another form of assessment should be used to better study this complication. The previous conclusions discussed were the only ones deemed significant enough to share due to high or low accuracy scores, but the following table and figure shows the accuracy scores and confusion matrices for each dependent variable so as not to leave any data assessment out of the discussion.

Complication	**Accuracy Value**
Atrial fibrillation	0.9
Supraventricular tachycardia	0.99
Ventricular tachycardia	0.98
Ventricular fibrillation	0.96
Third-degree AV block	0.97
Pulmonary edema	0.91
Myocardial rupture	0.97
Dressler syndrome	0.96
Chronic heart failure	0.77
Relapse of the myocardial infarction	0.91
Post-infarction angina	0.91
Lethal outcome	0.84

Figure 5: Accuracy values obtained using the SVM classification model with every combined independent variable v each outcome variable.

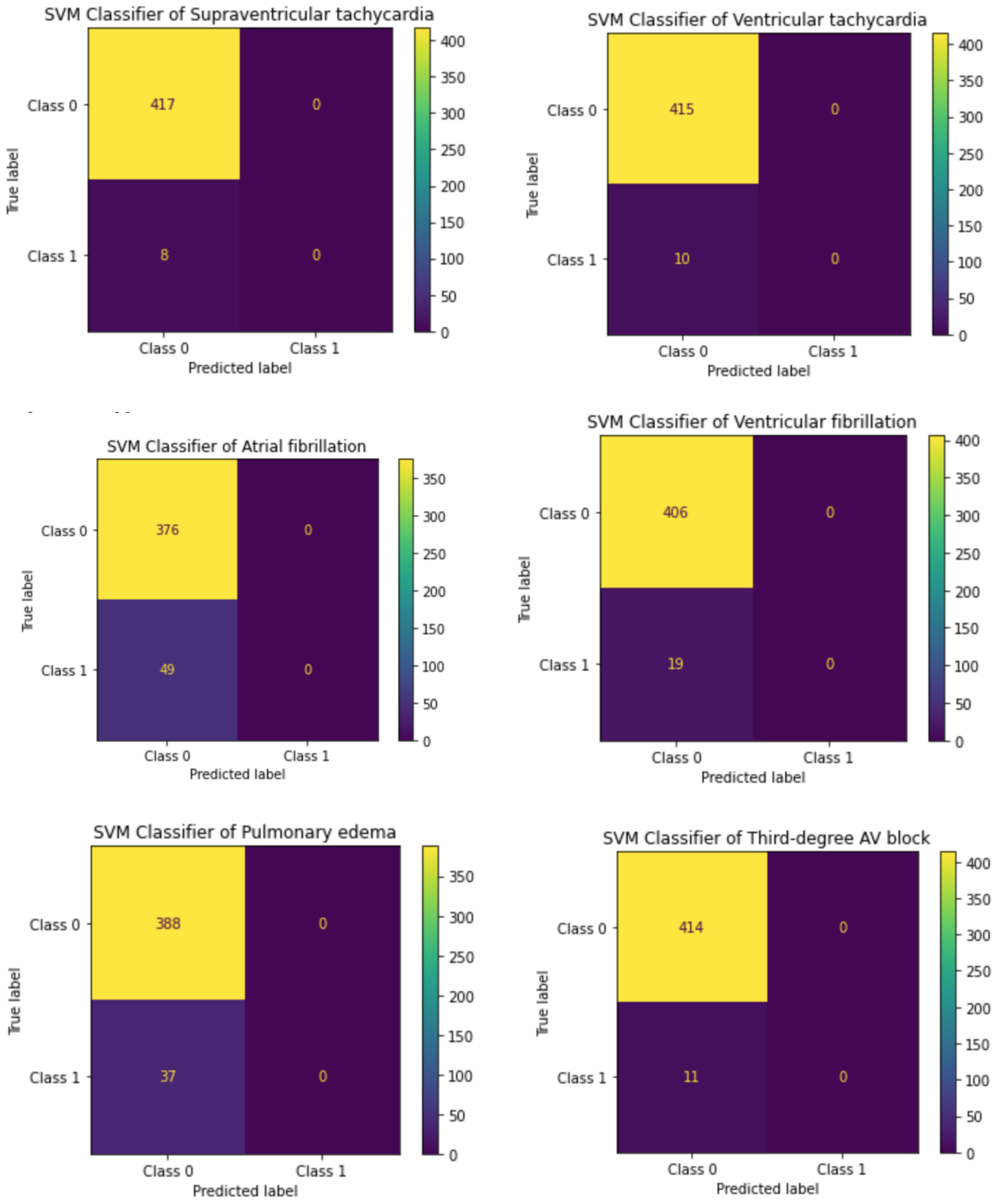

Figure 6: Confusion matrices for outcome variables Atrial fibrillation, Supraventricular tachycardia, Ventricular tachycardia, Ventricular fibrillation, Third-degree AV block, Pulmonary edema.

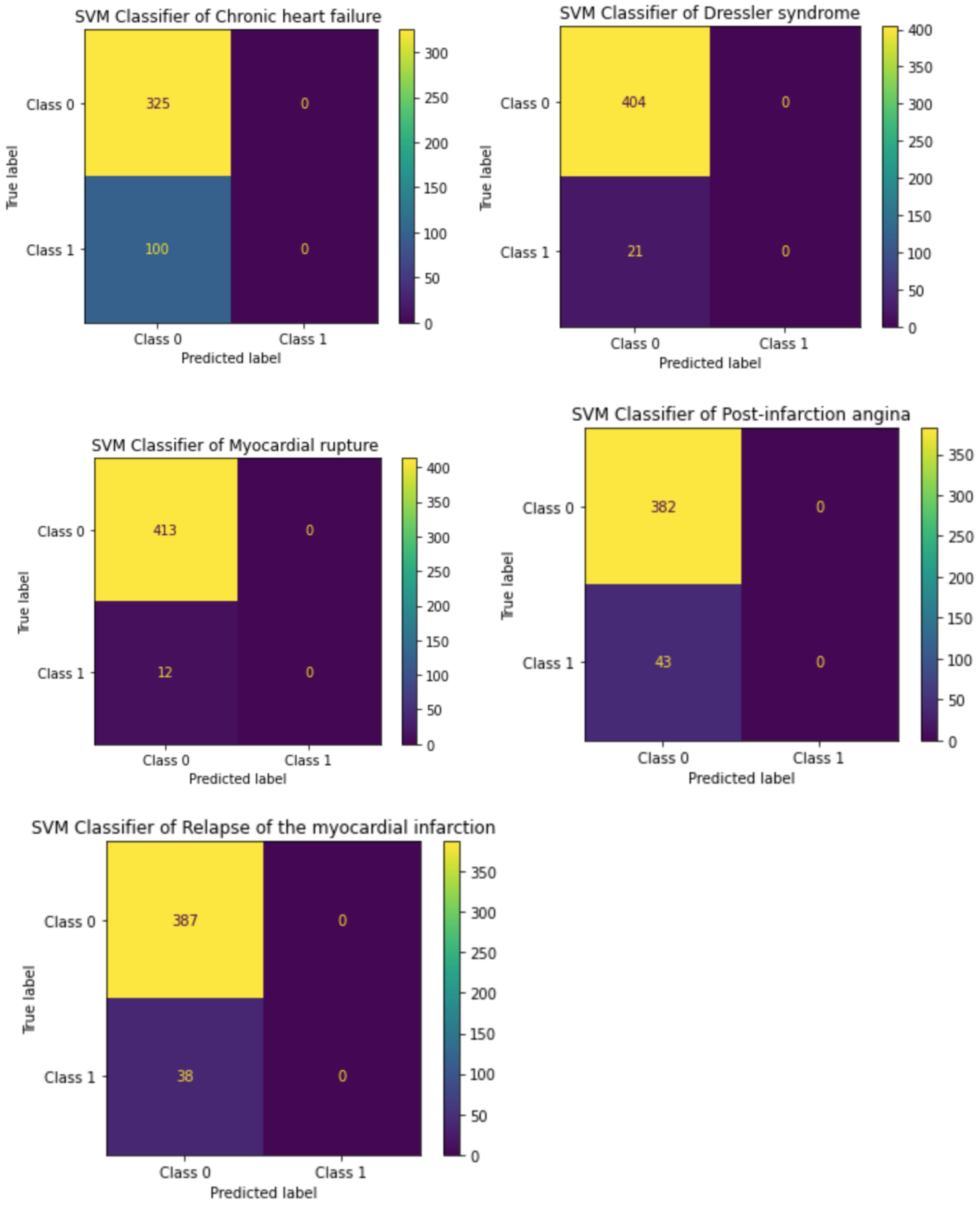

Figure 7: Confusion matrices for outcome variables Chronic heart failure, Dressler syndrome, Myocardial rupture, Post-infarction angina, and Relapse of myocardial infarction.

From the information collected, it is obvious that a better model could have been chosen. (Other models that may have been a better selection for this dataset are discussed in more depth in the ‘Limitations’ section of this report.) One major reason for this assessment is because as seen in figures 6 and 7, there is not a single time in which the model predicted that someone would have a MI complication and they actually did. In simpler terms, no true positives were found. This is a major problem with the model since a major objective for this data collection was to be able to predict when MI complications may occur and to do so accurately. As mentioned, the exploratory analysis did not point us in a clear direction and as such we chose the SVM model and this problem resulted. Yet, another thing that could cause this issue is that it is very likely that a combination of feature variables could better predict MI complications, rather than just a single feature variable or the entire set of feature variables together. Yet, in order to determine which combination would best do this would require much more, extensive exploratory analysis.

Limitations

There are a few limitations to SVM that may explain why it did not perform well with the large, myocardial infarction dataset. For one, the dataset was enormous with an escalated amount of feature variables. SVM is known to perform poorly with large datasets because the “training complexity is highly dependent on the size of the dataset” (Cervantes, 2008), and as such should have been an indicator that another model would have been better. Another limitation of SVM is that there is no probabilistic explanation for classifying datasets. This is because the model simply works by placing data points either above or below the distinguished hyperplane. Thus, with our objective of attempting to predict the likelihood of an event happening, SVM does not provide sufficient information to help answer this goal.

A better model to run in the future would be a random forest model because it is better equipped to handle data that has a large number of features like our data which has 112 variables that are predictors of MI complications. Random forest handles datasets with a large number of features because it is an ensemble of decision trees. Thus, it can run multiple variables against each other in order to determine which are the most significant in prediction and this model would answer which feature variable was selected by most trees to have the greatest impact on MI complications. Further, our data had more categorical variables than numeric and random forest works well with categorical variables for the same reasons as just explained. Finally, random forest is good to work with as the skip of normalizing data can be avoided since it uses a rule-based approach. Finally, running this model with an interpreter like SHAP would better help us understand the features that are the most important predictors. SHAP provides information on the degree to which feature variables have an outcome which is exactly the type of information we are after. Based on which features SHAP predicted the highest rate for, would be the variables we could single in on for predicting complications.

Finally, our confidence matrices show that a significant limitation from our modeling strategy is that it found zero true positives and zero false positives. Essentially, this means that our model didn’t classify any complications for any of the inputs. This is likely explained by the rarity of these complications and further shows how difficult it is to predict complications, as mentioned in the introduction of our project.

Conclusion

We used a Support Vector Machine (SVM), to find a hyperplane which classifies the data points. We looked at 110 independent variables such as age, gender, and stage of Ventricular fibrillation at which the patient was admitted to the intensive care unit to determine if there is plausible predictability for doctors to make from independent patient variables. As discussed earlier, it is difficult for even the most experienced doctors to know which complications a patient may have after a myocardial infarction. We first found accuracy scores to see how well the model holds up. The accuracy scores pointed to the complications dealing with tachycardia being best predicted by the model. Thereby testing the independent variables available to us against each of the 12 complications we have data for, the model can predict complications and we can test its accuracy. The model best predicts for the complication Supraventricular tachycardia, with varying degrees of reliability for the other 11 possible complications. Given the results already explained, it would also be wise to not use the model for the complication, chronic heart failure, and the model does not suit the compilation lethal cause well. Essentially, the model helps us classify external, independent patient variables into the different myocardial infarction complications.

Acknowledgements

A. Alexandria Garcia - Extensive research on myocardial infarctions helped us have a strong understanding of the dataset we were working with and communicate it in our presentation. Worked on the website.

B. Natalie Blanco - Extensive wrangling with python code to clean the data. Created SVM model and confidence matrices, analyzed the model and accuracy scores, and derived conclusions from the findings.

C. Susana Chavez - Extensive wrangling with python code to clean the data. Tested different classifier models and accuracy models to help decide and created a decision tree using age and gender shown in the presentation.

D. Melani Rodriguez - Extensive wrangling to import and clean the data in R. Created exploratory analysis visuals to help visualize and understand the data and generated hypotheses.

E. Dylan Wegmann - Extensive wrangling to import and clean the data in python. Worked closely with exploratory analysis and modeling to get the correct code for what we wanted.

Bibliography

Cervantes, J., Li, X., Yu, W., & Li, K. (2008). Support vector machine classification for large data sets via minimum enclosing ball clustering. *Neurocomputing*, *71*(4–6), 611–619. https://doi.org/10.1016/j.neucom.2007.07.028

D. A. Rossiev, S. E. Golovenkin, V. A. Shulman and G. V. Matjushin, 'Neural networks for

forecasting of myocardial infarction complications,' The Second International Symposium on Neuroinformatics and Neurocomputers, Rostov on Don, Russia, 1995, pp. 292-298, DOI: 10.1109/ISNINC.1995.480871https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=480871

Golovenkin, S.E., Bac, J., Chervov, A., Mirkes, E.M., Orlova, Y.V., Barillot, E., Gorban, A.N.

and Zinovyev, A., 2020. Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data. GigaScience, 9(11), p.giaa128., DOI: 10.1093/gigascience/giaa128 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7688475/

Gorban, A.N., Rossiyev, D.A. and Dorrer, M.G., 2004. MultiNeuron - Neural Networks

Simulator For Medical, Physiological, and Psychological Applications. arXiv preprint q-bio/0411034 (Available at [Web Link])

Gorban, A.N., Rossiev, D.A., Butakova, E.V., Gilev, S.E., Golovenkin, S.E., Dogadin,

S.A.,Dorrer, M.G., Kochenov, D.A., Kopytov, A.G., Maslennikova, E.V., Matyushin, G.V., Mirkes, Ye.M., Nazarov, B.V., Nozdrachev, K.G., Savchenko, A.A., Smirnova, S.V., Shulman, V.A. and Zenkin, V.I., Medical, psychological and physiological applications of MultiNeuron neural simulator. The Second International Symposium on Neuroinformatics and Neurocomputers. DOI: 10.1109/isninc.1995.480831 https://ieeexplore.ieee.org/document/480831

Great Learning Team. (2021, December 1). *Random forest Algorithm in Machine learning: An Overview*. GreatLearning Blog: Free Resources What Matters to Shape Your Career! Retrieved December 9, 2021, from https://www.mygreatlearning.com/blog/random-forest-algorithm/

“Heart Attack.” *Johns Hopkins Medicine*, https://www.hopkinsmedicine.org/health/conditions-and-diseases/heart-attack#:~:text=A%20heart%20attack%20(myocardial%20infarction,the%20heart%20muscle%20is%20blocked.

K, D. (2020, December 26). *Top 4 advantages and disadvantages of Support Vector Machine or SVM*. Medium. Retrieved December 9, 2021, from https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-veCtor-machine-or-svm-a3c06a2b107

Pupale, R. (2019, February 11). *Support Vector Machines(SVM) — An Overview - Towards Data Science*. Medium. Retrieved December 9, 2021, from https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989#:%7E:text=Now%20that%20we%20understand%20the,space%20into%20two%20disconnected%20parts.&text=So%20a%20point%20is%20a%20hyperplane%20of%20the%20line.