

M.A.M.

Medical Analysis and Modeling

Angelina Martineau

Rensselaer Polytechnic Institute

May 04, 2020

Table of Contents

[1. Introduction 3](#_Toc39522387)

[2. Data Mining in Medicine 3](#_Toc39522388)

[3. Legality of Data Acquisition 5](#_Toc39522389)

[4. Heart Disease 5](#_Toc39522390)

[5. Models for Kaggle Heart Disease Data 6](#_Toc39522391)

[5.1 Multi-Linear Regression 7](#_Toc39522392)

[5.2 Linear Regression 1 VS. 1 8](#_Toc39522393)

[5.3 Multi-Logistic Regression 9](#_Toc39522394)

[5.4 Logistic-Regression 1 VS. 1 10](#_Toc39522395)

[5.5 Naïve Bayes 11](#_Toc39522396)

[5.6 Deep Learning 12](#_Toc39522397)

[6. Resources 14](#_Toc39522398)

# 1. Introduction

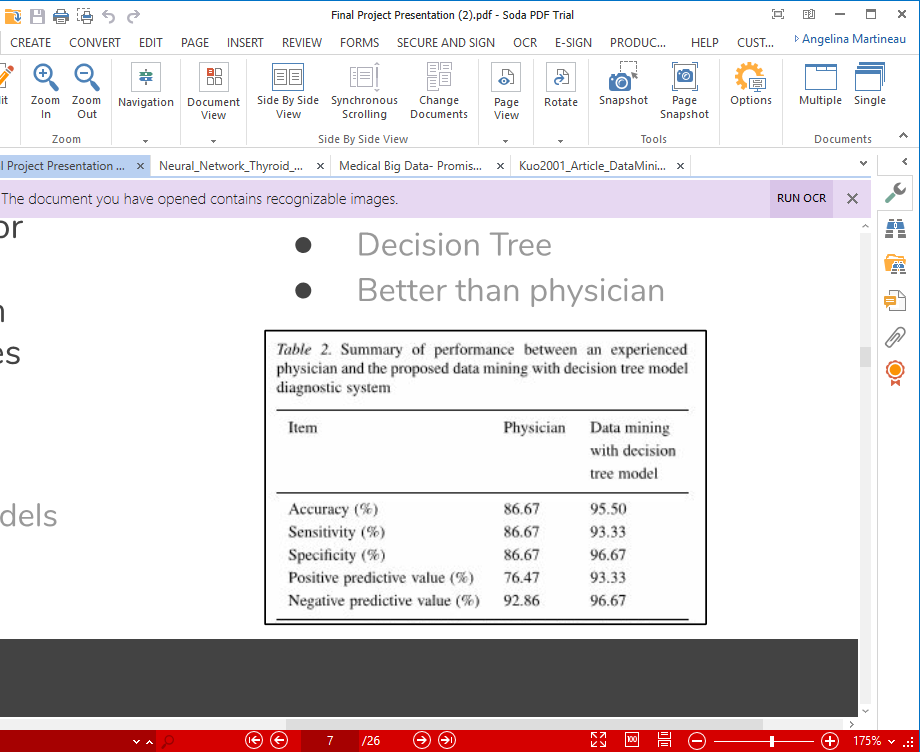
Data mining in medical science is crucial to accurate and timely diagnosis. This projects serves to illuminate the effects of furthering the study in the field, and aims to convince field leaders to push for greater analysis of medical data. Medical data mining has proven to be incredibly useful, whether it is for prediction of disease or improving the efficiency of the healthcare system. Models have been developed for various afflictions allowing medical professionals to make quicker, more accurate predictions, as will be shown in later sections.

Consider taking a subset of characteristics about a person related to their health. Given that small amount of information, learning model can determine (to a certain degree of accuracy) whether a person is afflicted with a specified medical condition. These are the extreme advancements in the medicine world that save lives. This paper aims to demonstrate the need for data mining in medicine. It will begin with an overall view of data mining in medicine. It will then go on to explore the trouble with obtaining medical data. Next, the effects of data mining on heart disease diagnosis will be reviewed. Finally, an analysis of several models created for a heart disease dataset from Kaggle will be shown.

# 2. Data Mining in Medicine

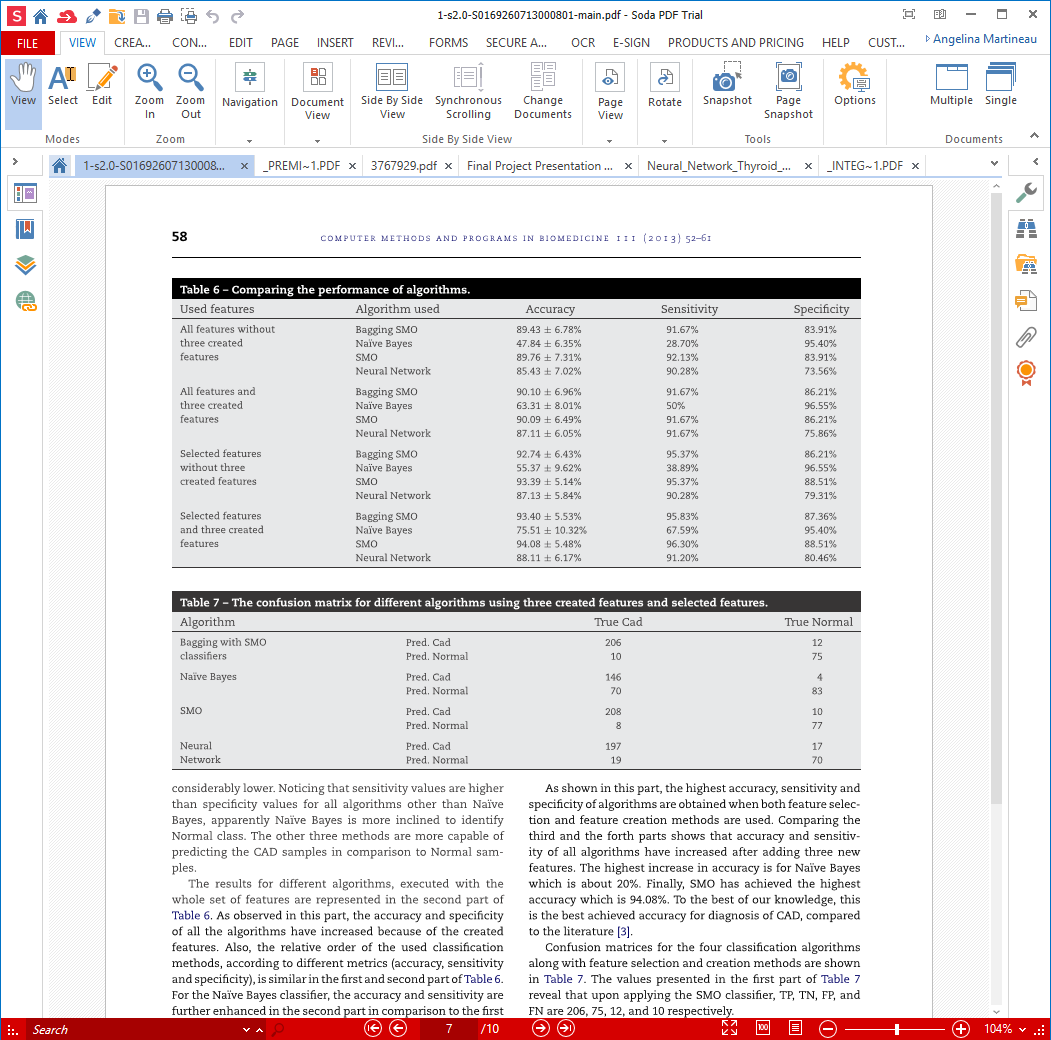
Data Mining has had a great impact on medicine, and continues to advance the field. According to Lee and Hoon, “There are various sources of medical big data, such as administrative claim record, clinical registries, electronic health records, biometric data, patient-reported data, the internet, medical imaging, biomarker data, prospective cohort studies, and large clinical traits,” (Medical big data: promise and challenges). There are many applications to data mining and modeling in medicine, and the need for work with medical data is great.

One application of data mining in medicine by Kuo et. al. demonstrates the use of breast tumor images for better diagnosis (Data mining with decisions trees for diagnosis of breast tumor in medical ultrasonic images). This team obtained 243 instances of image data for patients ages 17-64. The tumors in the images varied in size from 0.8cm to 4.2cm. They decided to use a decision tree to develop a model that could discover tumors during breast cancer imaging. The model proved to be momre accurate and effective at diagnosing breast cancer than a physician. The accuracy, sensitivity, specificity, positive predictive value, and negative predictive value for both the physician and the decision tree model are shown below in a table developed by the research team.



*Table One: Comparison of physician and data model for breast cancer study. Obtained from “Data mining with decisions trees for diagnosis of breast tumor in medical ultrasonic images”.*

Another study focused on coronary artery disease. This project was developed by Alizadehsani et. al. in 2012, and focused on the effects of adding features to the data to improve modeling. The three features that were added related to blocked arteries, and were able to be derived from other features in the data. This expansion let to more accurate models that could perform better diagnosis. Four different models were developed: Bagging SMO (a combination of the sequential minimal operation algorithm and the bagging algorithm), Naïve Bayes, a Neural Network, and SMO. All of the models demonstrated that they were able to perform better when the added features were calculated and accounted for. The table below demonstrates the accuracy, sensitivity, and specificity of each model with different combinations of features:



*Table Two: Metrics for the various models created in the study on coronary artery disease data.*

This table demonstrates that the models that performed the best on the data only considered selected features and the three created features. This study demonstrates the further need for research projects in medical data mining, as there are many ways to develop models, and as many as possible should be attempted. A newer model may end up being more accurate than an older one in use, and it may have the potential to save lives via early diagnosis.

# 3. Data Acquisition in Medicine

A large obstacle in the way of data mining in medicine is the need for data. Medical data can be difficult to obtain for several reasons, especially because it is hard data to maintain and clean. Many institutions have been upgrading to electronic systems to manage patient data, but there are still many places that haven’t, and many that have found some middle ground between paper-based data and electronic data. As Kuilboer et. al. explain, “Although many decision-support systems were built, most failed to become incorporated into daily clinical practice. An important reason for this failure was that these systems tended to be stand-alone, requiring separate data entry.” (The availability of unavailable information). Kuilboer et. al. had worked to develop a system that would not only take data input for patient records, but would also advise doctors about how to proceed given the data it has stored. Systems like these can be developed with data mining practices, but there needs to be a greater deal of organization; essentially, all the data must be stored together in a clean, well designed system.

Health data is also difficult to obtain due to privacy laws. For example, in the United States in 1996, the Health Insurance Portability and Accountability Act as passed, which prevents medical information from being shared without consent from the patient (Center for Disease Control and Prevention). Careful consideration must be taken to avoid legal infractions based on improper usage of health care data. HIPAA requires that patient consent be needed (in most cases) for research purposes, and patients are often hesitant to give their personal health data up for use, even when their identity is removed from the data. Electronic Health Reporter clarifies the situation, explaining that “Under HIPAA rules, data mining is a secondary, future use of health data, and thus requires the explicit permission of the patient before being used,” (Electronic Health Reporter). For effective data mining, large amounts of data need to be gathered with equal representation among the target values for the data (e.g., the heart disease data in this project has approximately 50% positive instances and 50% negative instances). This becomes especially hard, though, when data acquisition becomes reliant on patient consent.

# 4. Heart Disease

The term ‘Heart Disease’, also referred to as ‘Cardiovascular Disease,’ represents all diseases that affect the heart, arteries, and blood vessels of humans. It is one of the leading causes of death worldwide, and is the leading cause of death in the United States. While there are markers for the disease, it is not always detected. In fact, some people have mild enough heart attacks that they don’t even realize it’s happening. The damage is still done, though, and an early diagnosis or prediction for high risk of heart disease could have helped to prevent such instances. The study discussed in *Section 2* involving coronary artery disease demonstrates how useful data mining is to heart disease, as it can improve prediction and diagnosis so patients are aware of the risks they have.

Most importantly, application of data mining to heart disease data will save lives, but it will also reduce the overall heart disease cost. Each year, the United States spends over $200 billion dollars on heart disease. Early prediction and diagnosis can aid in reducing that cost, as fewer patient afflicted with heart disease will be ignorant to their conditions. With more accurate diagnosis, they will be able to adjust their lifestyle before it is too late, saving not only themselves but the cost of medical care they would need to receive.

# 5. Models for Kaggle Heart Disease Data

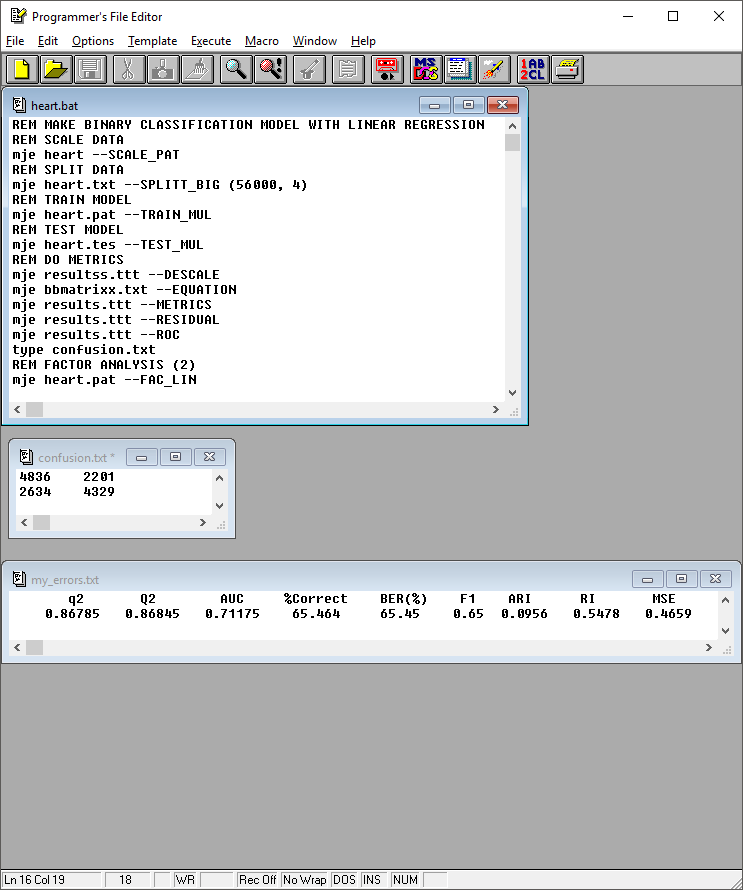
For this project, several models were developed for a heart disease dataset that was obtained from Kaggle. The data was posted by user ‘Svetlana Ulinova,’ who could not be contacted to inquire about the origin of the data. The only other information about data collection stated that for each instance, the data was collected at the time of the individual’s exam.

There are three types of characteristics that were collected for each instance. First, there are objective features, which are physical measures that cannot be disputed. Objective features include age (integer) in days, height (integer) in centimeters, weight (float) in kilograms, and gender (category). The second type of features is an examination feature, which is determined during the exam. This includes systolic blood pressure (integer), diastolic blood pressure (integer), cholesterol (1-normal, 2-above normal, 3-well above normal), and glucose (1-normal, 2-above normal, 3-well above normal). Finally, there are subjective features, which are based on information supplied by the patient. These features are all binary (1-yes, 0-no). They include whether the patient smokes, consumes alcohol, and exercises. The target variable for this data is the presence of cardiovascular disease, which is indicated by a binary value (1-yes, 0-no). It is important to note that approximately half of the instances in this dataset have cardiovascular disease while the other half do not. This helps gain a better prediction model, as there is less bias.

For all models created, the data was split into a training set and a testing set. The training set contained 56,000 (80%) instances, while the test set contained 14,000 instances (20%). All confusion matrices and metrics are based on the testing sets after the model was applied to them. Specificity is calculated as TP / (TP + FN), and sensitivity is calculated as TN / (TN + FP). The confusion matrices shown have the top left box representing true negatives, the bottom right box representing true positives, the bottom left box representing false negatives, and the top right box representing false positives. All models were developed using the DMaK software developed by Mark J. Embrechts.

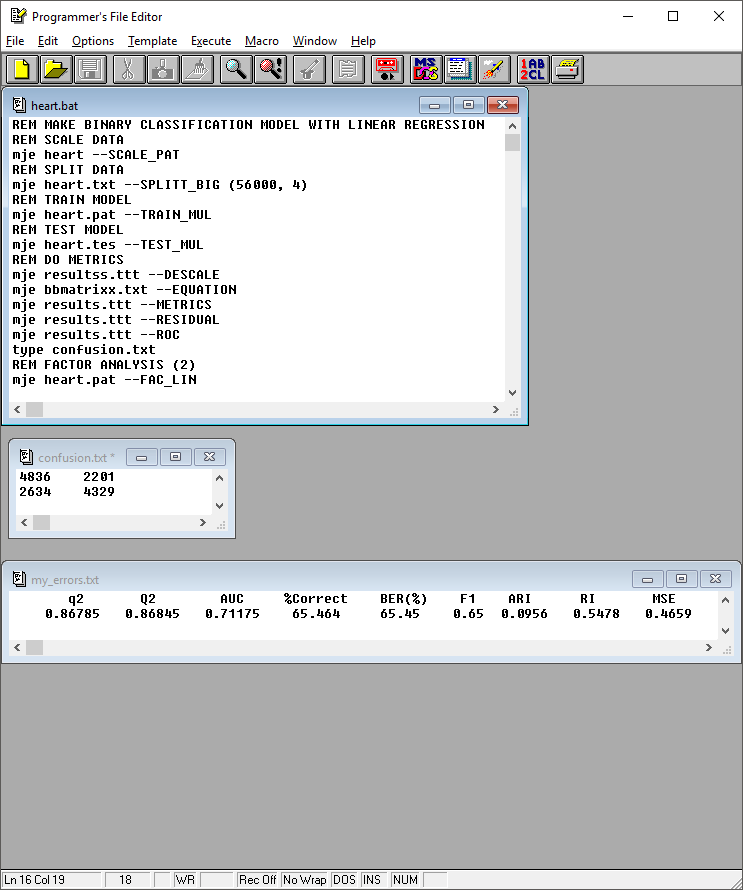
## 5.1 Multi-Class Linear Regression

The first model developed was a multi-linear regression model. This was done using the *--TRAIN\_MUL* and *--TEST\_MUL* commands. The script that was generated for the multi-class linear regression model is shown below in *Image One.*



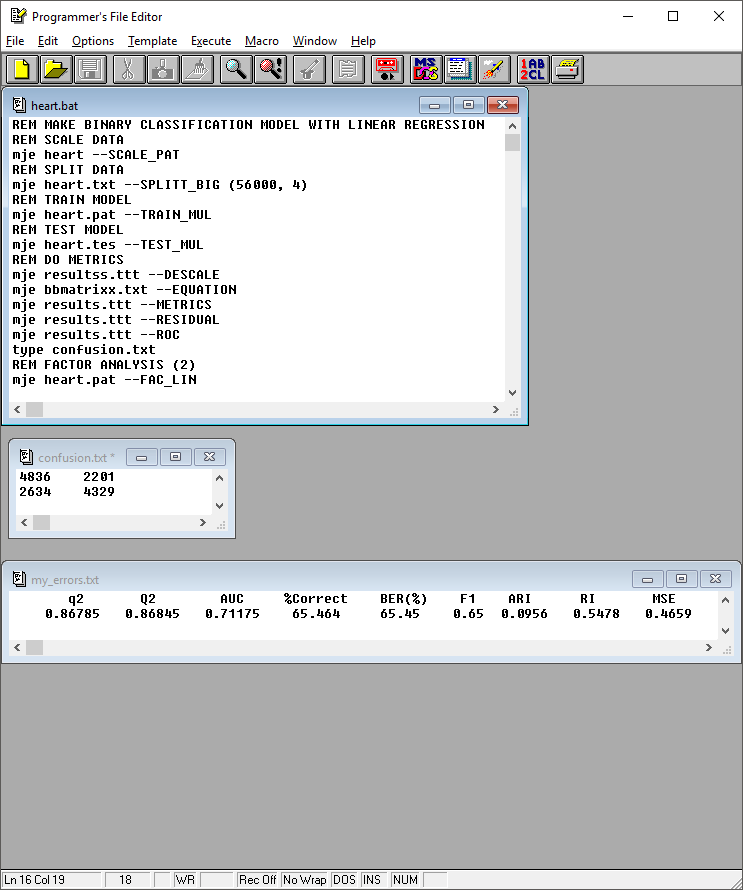
*Image One: DMak Script for the multi-class linear regression model.*

The script involves scaling the data, splitting it into training and testing sets, creating the model, testing the model, descaling the data, and developing error metrics. The error metrics are shown below in *Image Two*.



*Image Two: Error metrics for the multi-class linear regression model.*

The model was only able to predict about two-thirds of the test data correctly, and only had an F1 metric of 0.65. Because this is medical data, it is important to look at the sensitivity and specificity. The confusion matrix is displayed below.

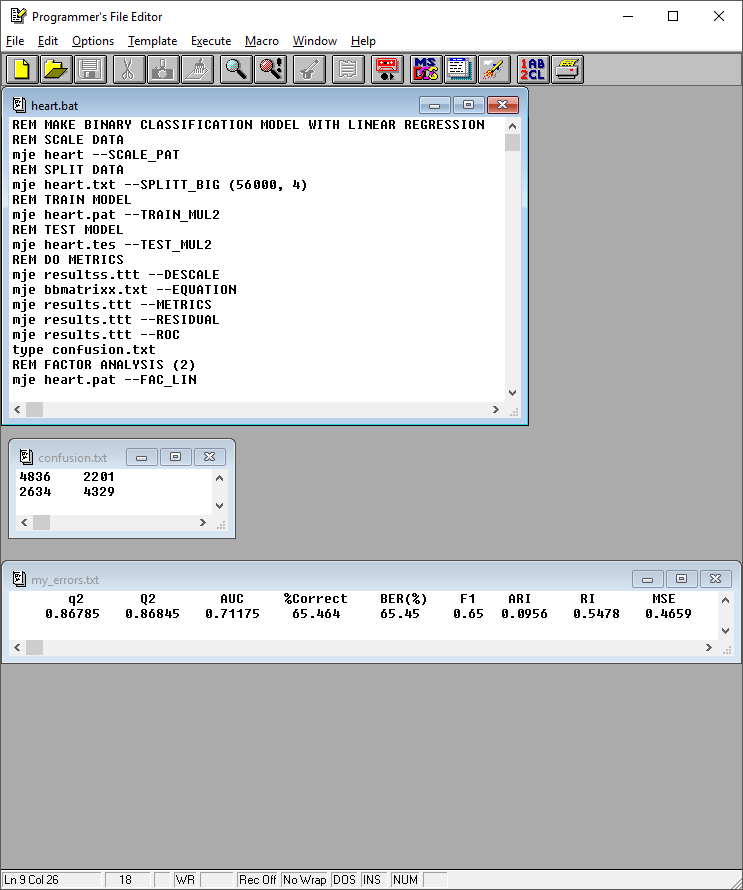
**

*Image Three: Confusion matrix for the test data that was predicted by the multi-class linear regression model.*

Using this confusion matrix, the specificity was calculated to be .6217 and the sensitivity was found to be .6872. As the number of false negatives increase, the specificity will decrease. It is especially important to have a high specificity because false negatives can lead to incorrect diagnosis. Neither the specificity nor the sensitivity are very high, suggesting that other models would be better for this data.

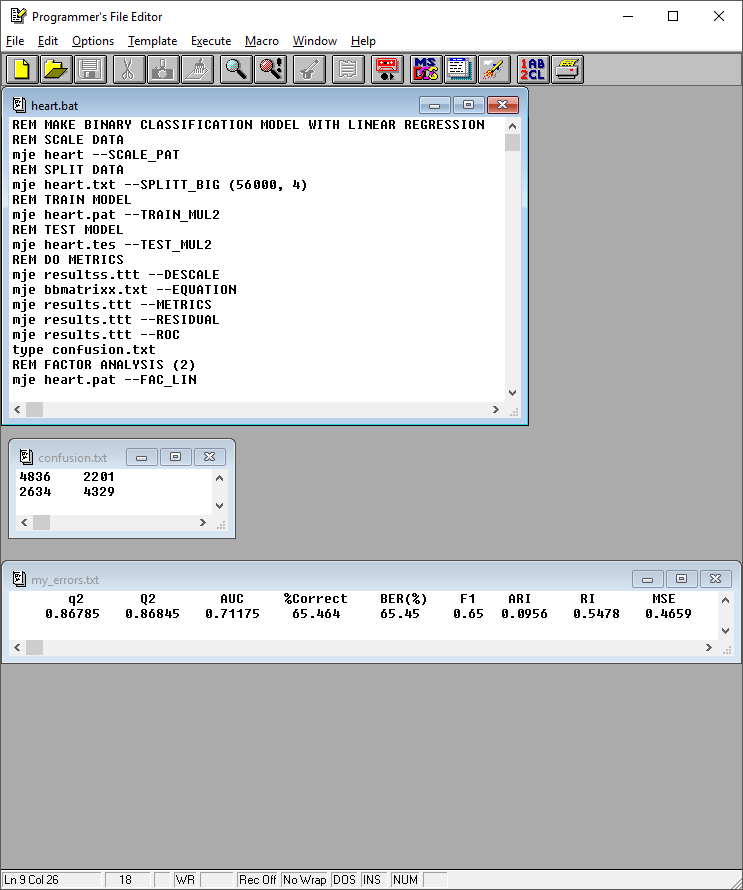
## 5.2 Linear Regression 1 VS. 1

The second model developed was also a linear regression model, except this one was a 1 vs. 1 linear regression model. The *–TRAIN\_MUL2* and *–TEST\_MUL2* options were used. The script that was generated is shown below.

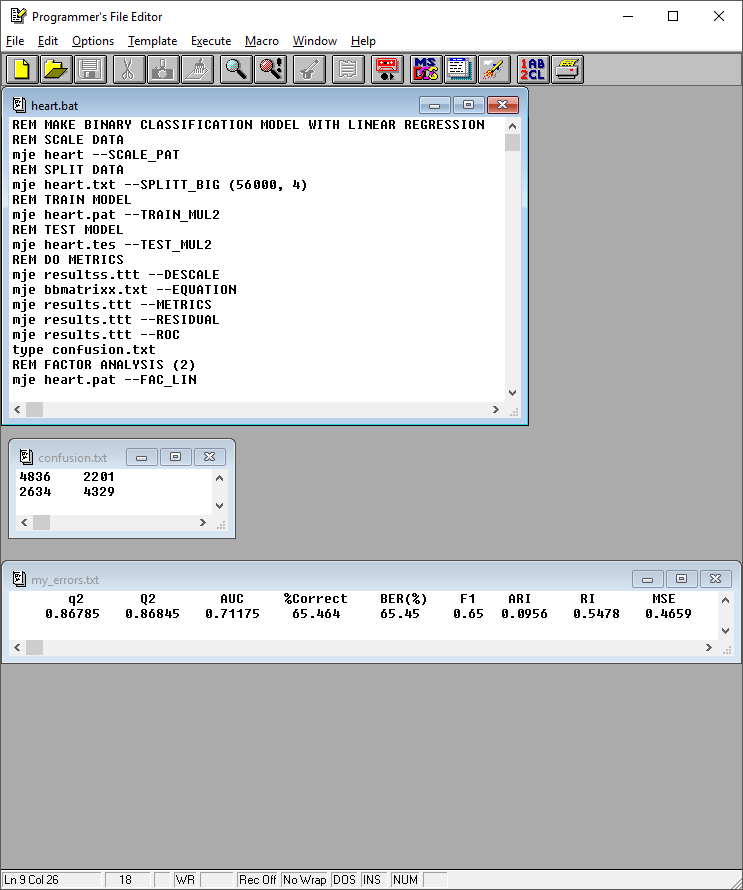


*Image Four: DMak script generated for the 1 vs. 1 linear regression model.*

This script is very similar to the first model, and this model performed with exactly the same metrics as the first linear regression model. The error metrics and confusion matrix are shown below.



*Image Five: Error Metrics for the 1 vs. 1 linear regression model.*

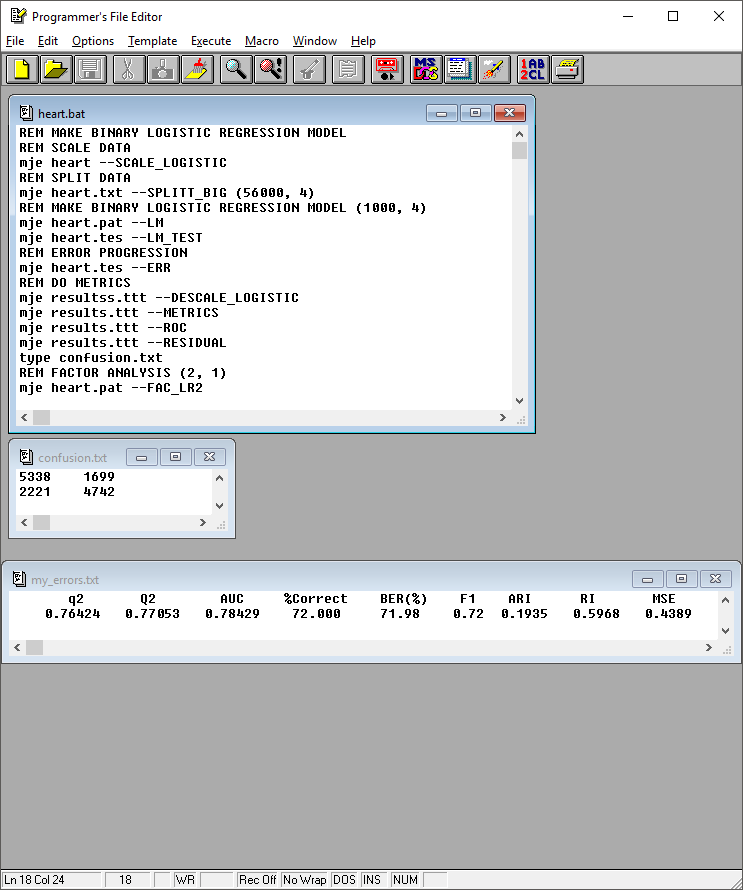


*Image Six: Confusion matrix for the 1 vs. 1 linear regression model.*

Since the confusion matrix for this model is the same as the confusion matrix for the first linear model, the specificity and sensitivity are also the same. The fact that there is no improvement from the other linear model further suggests that the best model for this data is not linear.

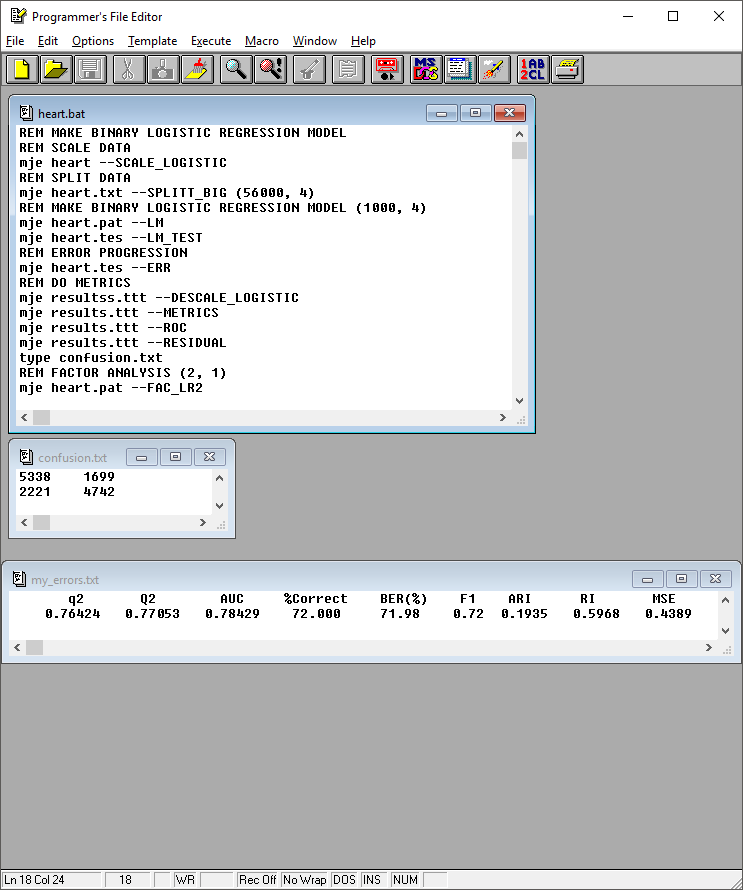
## 5.3 Multi-Class Logistic Regression

The third model developed was a multi-class logistic regression model. The script generated for this model made use of the *–LM\_TRAIN* and *–LM\_TEST* commands. The script for this model is shown below in *Image Seven*.

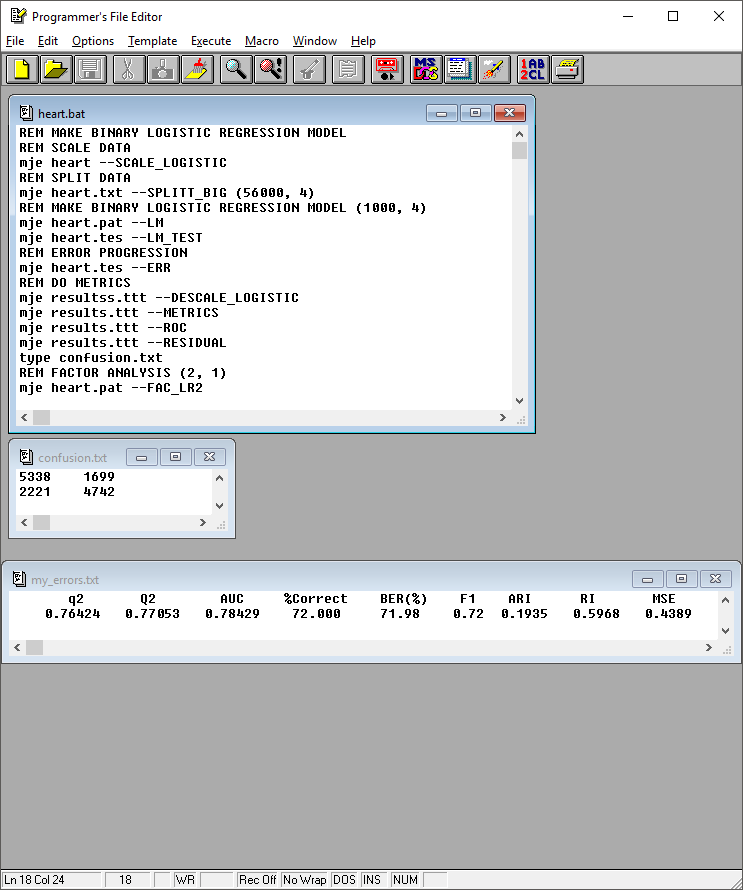


*Image Seven: DMak script for the multi-class logistic model.*

The multi-class logistic regression model performed much more strongly than the linear models did, as is shown by the error metrics and the confusion matrix below.



*Image Eight: Error metrics for the multi-class logistic regression model.*

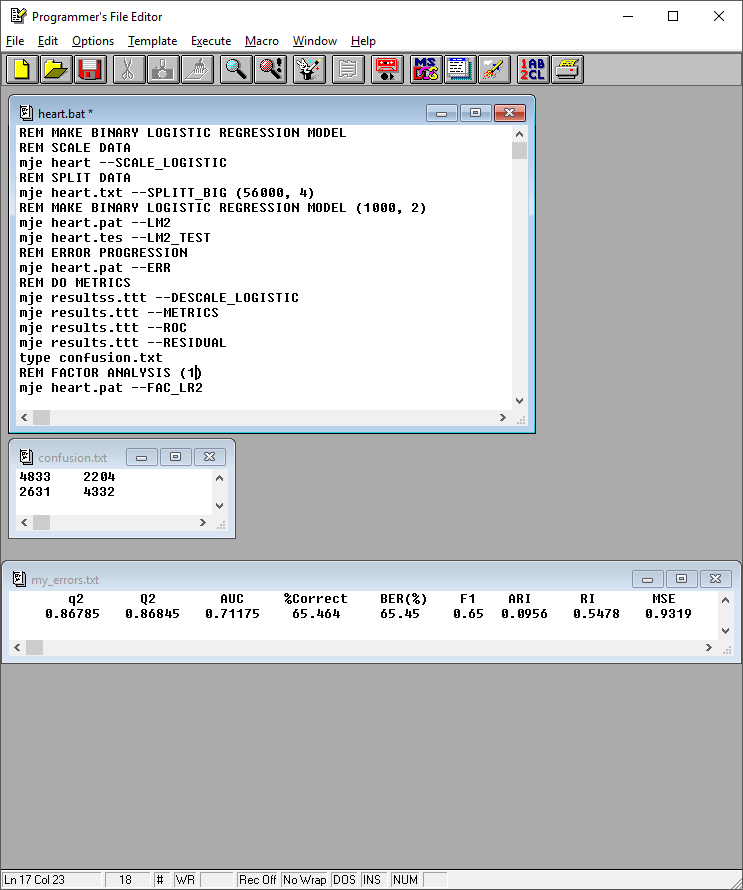


*Image Nine: Confusion matrix for the multi-class logistic regression model.*

This model is able to correctly classify 72% of the test data, a definite improvement over the linear models. It has an F1 metric of 0.72, and has improved specificity and sensitivity as well, which were calculated using the information form the above confusion matrix. The specificity was found to be .6810, and the sensitivity was .7586. The specificity is, again, very important, as it factors in the number of false negatives. So, the improvement in specificity with this model is important.

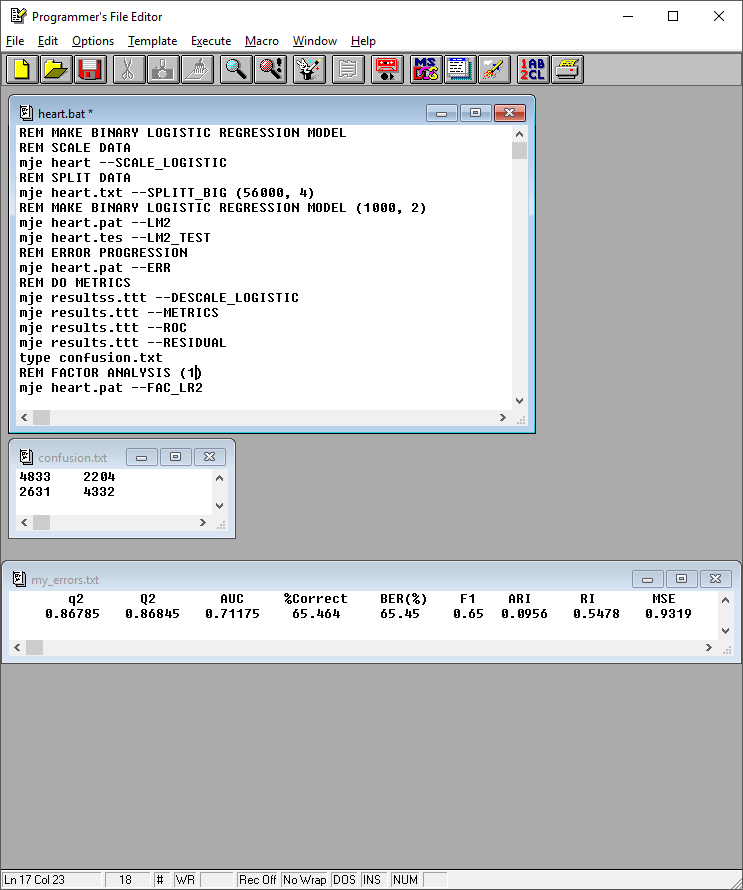
## 5.4 Logistic-Regression 1 VS. 1

The fourth model generated overall was a logistic regression 1 vs. 1 model. This script involved the –LM2 command. The full script is shown in *Image Ten*.

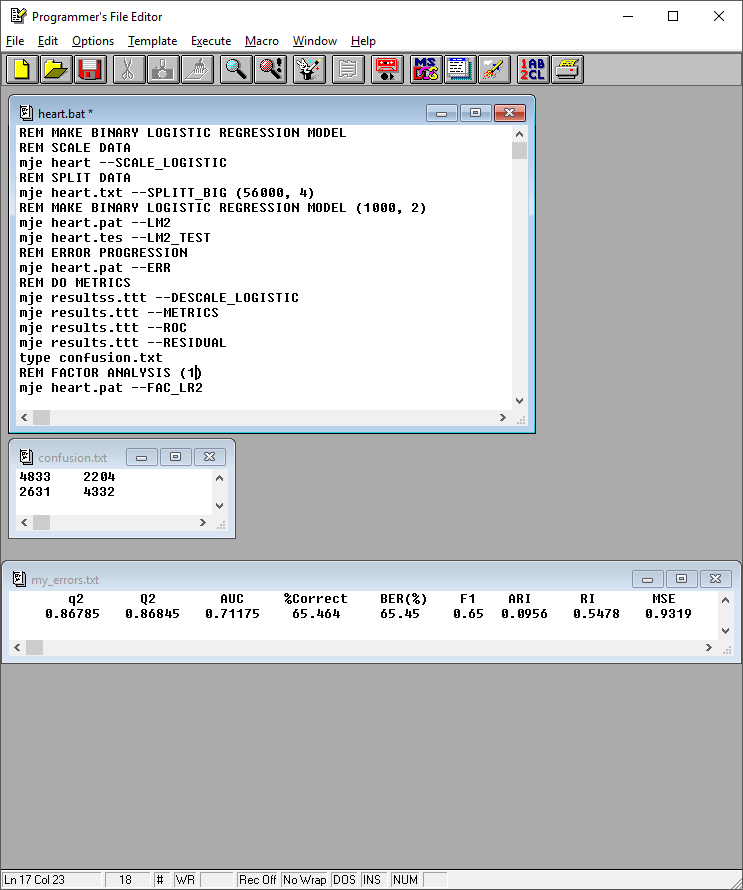


*Image Ten: DMak Script for the 1 vs. 1 logistic regression model.*

This model ran with the same accuracy as the linear models, but it produced a slightly different confusion matrix, leading to slightly different specificity and sensitivity. The error metrics and the confusion matrix can be seen below:



*Image Eleven: Error metrics for the 1 vs. 1 logistic regression model.*

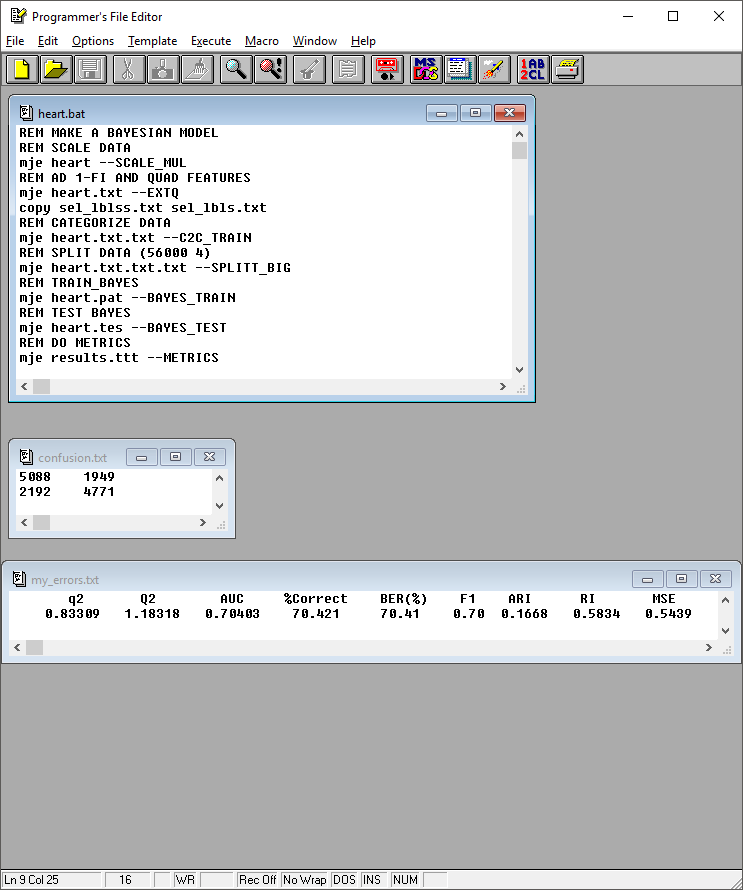
**

*Image Twelve: Confusion matrix for the 1 vs. 1 logistic regression model.*

The model was only correct for about two thirds of the test data, like the linear models. It had an F1 metric of 0.65, which is a decrease from the other logistic model. The specificity was calculated to be .6221, which represents only a slight improvement over the linear models. The sensitivity was .6868, which is slightly less than the linear models. Overall, this model can be considered as effective at classifying the data as the linear models. The multi-class logistic regression model demonstrates that this model is clearly not the best model for the heart disease data.

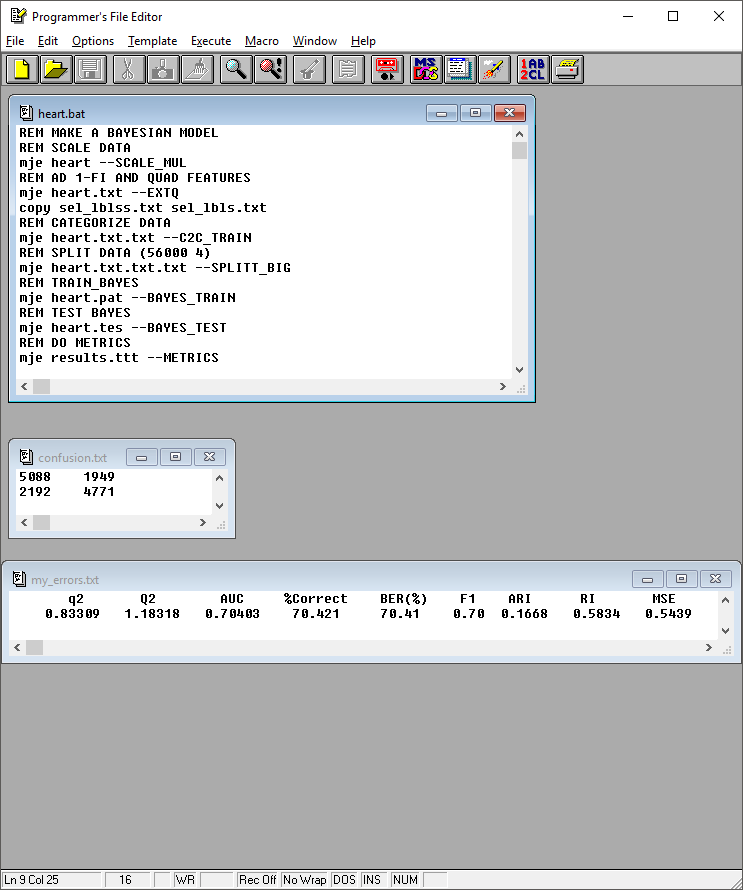
## 5.5 Naïve Bayes

The fifth model developed for the heart disease data was a Naïve Bayes model. The script was generated using the *–MS\_BAYES* command, and can be seen below in *Image Thirteen.*

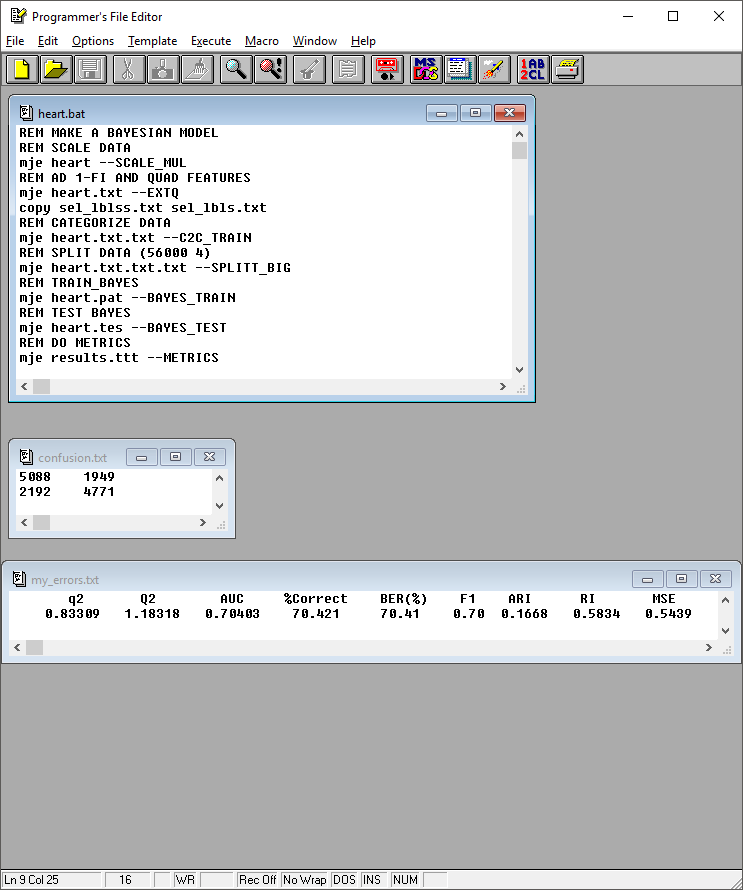


*Image Thirteen: DMak script generated for the Naïve Bayes model.*

Like the previous scripts, this one split the data into training and testing sets, with 80% of the data remaining in the training set. The model was then developed based on the training data and applied to the test data. The error metrics and the confusion matrix are shown below.



*Image Fourteen: Error metrics for the Naïve Bayes model.*

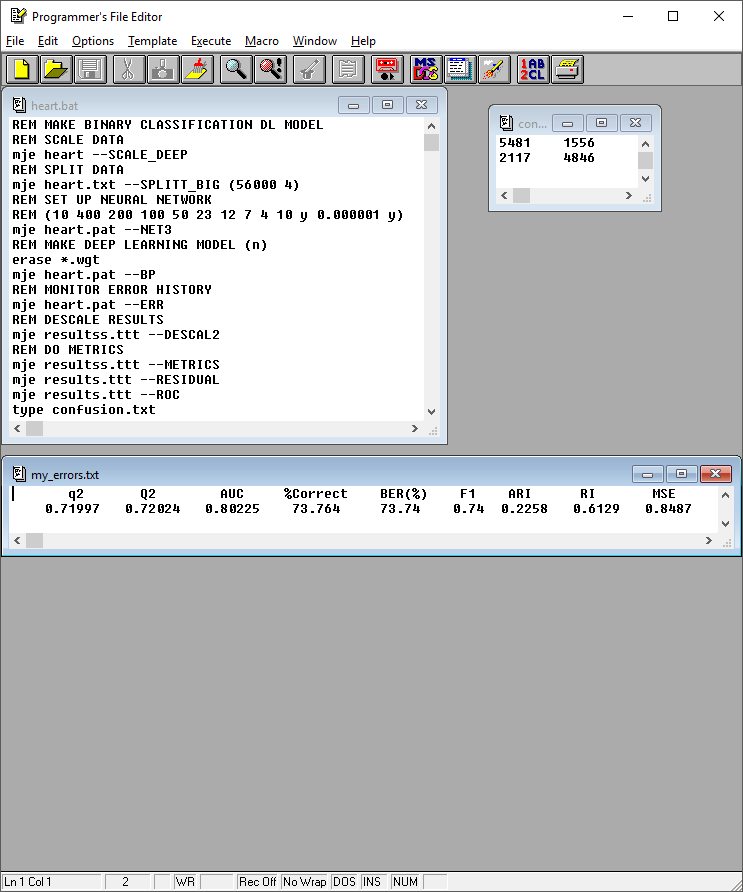


*Image Fifteen: Confusion matrix for the Naïve Bayes model.*

This model performed better than the linear models and the 1 vs. 1 logistic regression model, but it was slightly worse than the multi-class logistic regression model. It was able to accurately predict 70.421% of the test data, and had an F1 metric of 0.70. Based on the confusion matrix, the specificity was found to be .6852, which is a slight improvement over the multi-class logistic regression model. Improvements in the specificity are notable, as this means a reduction in false negatives and an increase in true positives. The sensitivity, however, was not as good as the sensitivity for the multi-class logistic model, as it was only .7230.

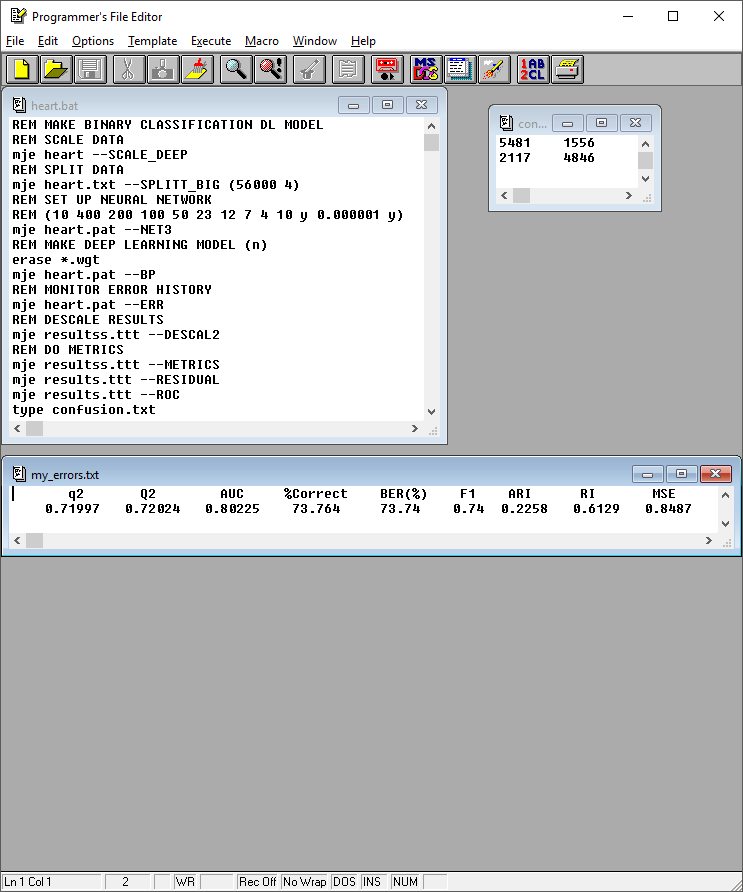
## 5.6 Deep Learning

The final model developed for the cardiovascular dataset was a deep learning model. For this model, the following script was created by using the –MS command and selecting ‘y’ for deep learning model and manual set-up of the model.

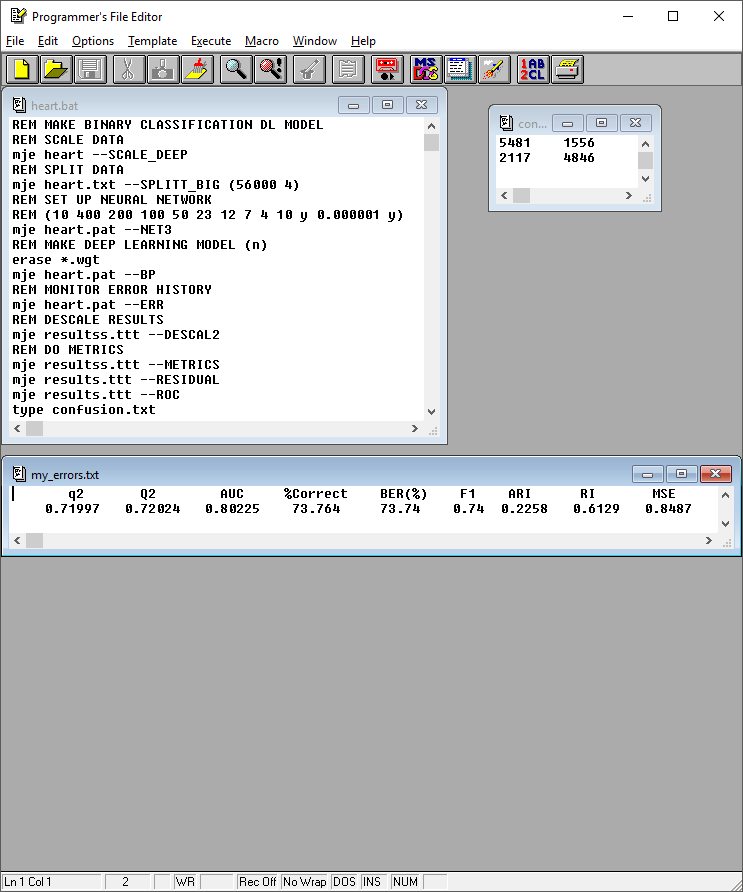


*Image Sixteen: DMak script generated for the deep learning model.*

The network had 10 layers. There were 11 inputs and 1 output, and the middle eight layes had 400, 200, 100, 50, 23, 12, 7, and 4 nodes. The model used adam and RELUs, and the stopping error was set to be 0.000001. Dropout was not used for this model. After manually setting up the model, it was run on the data. The error metrics and the confusion matrix for the test data are shown below.

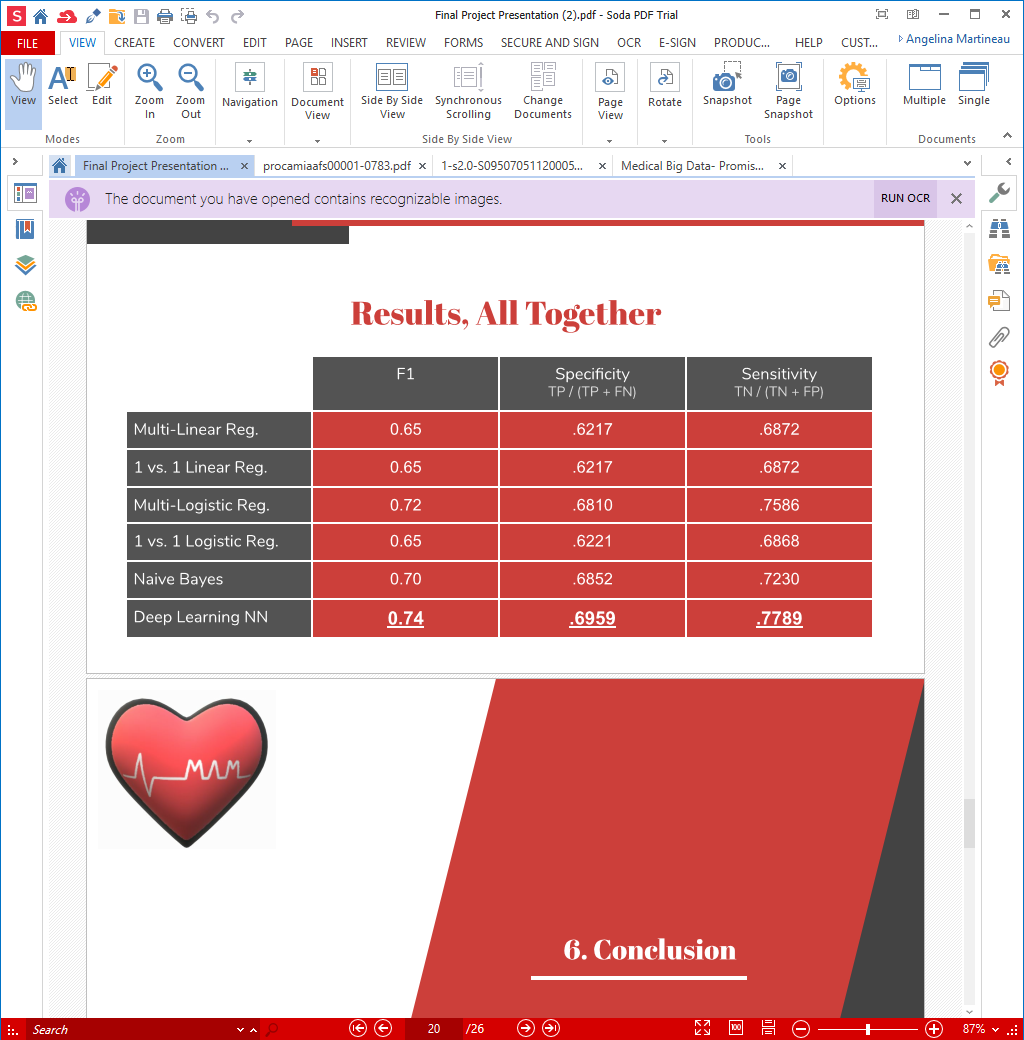


*Image Seventeen: Error metrics generated for the deep learning model.*



*Image Eighteen: Confusion matrix for the deep learning model.*

The deep learning model performed the best of the models developed for this project. It was able to predict 73.764% of the cardiovascular test set correctly, and it had an F1 metric of 0.74, which is the best of all the models. This model also had the best sensitivity and specificity as well. The sensitivity was .7789 and the specificity was .6959. It is important to maximize all of these metrics, and for all four of them, the deep learning model performed the best. This indicates that the deep learning is the most effective model for the cardiovascular disease data, though there could be a different deep learning model that would better predict the data than the model presented here. A final table displaying the difference in performance for all the models is shown below in *Image Nineteen.* This table clearly demonstrates how the deep learning model out-performed the other models when considering the F1 metric, the specificity, and the sensitivity.

**

*Image Nineteen: Metrics for all models developed for the heart disease data.*

# 6. Resources

Alizadehsani, R., et al. (2013). A data mining approach for diagnosis of coronary artery disease. Computer Methods and Programs in Biomedicine, Vol. 111, pp. 52-61. Retrieved from <https://www.sciencedirect.com/science/article/abs/pii/S0169260713000801>

Baldwin, L. M., et. al. (Aug., 2002). Linking Physician Characteristics and Medicare Claims Data: Issues in Data Availability, Quality, and Measurement. Medical Care, Vol. 40, No. 8, pp. IV82 - IV95. Retrieved from <https://www.jstor.org/stable/3767929>

Berka, P., et. al. (2009). Data Mining and Medical Knowledge Management Cases and Applications. Retrieved from <http://gen.lib.rus.ec/book/index.php?md5=82be13d2bd187b56e4ea3e3819d01b24>

Chen, H., Fuller, S. S., Friedman, C., Hershi, W. (2005). Medical Informatics: Knowledge Management and Data Mining in Biomedicine. Integrated Series in Information Systems, edition 1. Retrieved from <http://gen.lib.rus.ec/book/index.php?md5=b53f2c5b462ff86977c7c8495b6ed354>

*Data Mining: How to Protect Patient Privacy and Security When Using Data.* (2020). Electronic Health Reporter. Retrieved May 02, 2020 from <https://electronichealthreporter.com/data-mining-protect-patient-privacy-security-using-data/>

Gupta, S., Sharma, A. (April 2011). Data mining classification techniques applied for breast cancer diagnosis and prognosis. Indian Journal of Computer Science and Engineering, Vol. 2. Retrieved from <https://www.researchgate.net/publication/268054343_Data_mining_classification_techniques_applied_for_breast_cancer_diagnosis_and_prognosis>

Health Insurance Portability and Accountability Act of 1996 (HIPAA). (2018, September 14). Center for Disease Control and Prevention. Retrieved May 02, 2020, from <https://www.cdc.gov/phlp/publications/topic/hipaa.html>

Huang, M., Chen, M., Lee, S. (April 2007). Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis.  Expert Systems with Applications, Vol. 32, Issue 2, pp. 856-867. Retrieved from <https://www.sciencedirect.com/science/article/abs/pii/S095741740600042X>

Kuilboer, M. M., et. al. (1997). The availability of unavailable information. Proc AMIA Annu Fall Symp 749-753. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2233582/>

Kuo, W., Chang, R., Chen, D. et al. (2001). Data mining with decision trees for diagnosis of breast tumor in medical ultrasoundic images. Breast Cancer Res Treat, Vol 66, pp. 51-57. Retrieved from <https://link.springer.com/article/10.1023/A:1010676701382>

Mookiah M.R.K., Rajendra Acharya U., Lim C.M., Petznick A., Suri J.S. (2001) Data mining technique for automated diagnosis of glaucoma using higher order spectra and wavelet energy features. Knowledge-Based Systems, Vol. 33, pp. 73-82. Retrieved from <https://www.sciencedirect.com/science/article/abs/pii/S0950705112000500>

Prerana, Sehgal, P., Taneja, K. (April 2015). Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network. International Journal of Research in Management, Science & Technology, Vol. 3, No. 2. Retrieved from <http://papersim.com/wp-content/uploads/Neural_Network_Thyroid_Disease_2015.pdf>