Detection Of Cardiovascular Disease Through The Use Of Machine Learning Models

Dr. Christina Schweikert
St. John's University
Division of Math, Science And Computer Science
schweikc@stjohns.edu

Christopher Singh
St. John's University
Division of Math, Science And Computer Science
christopher.singh15@stjohns.edu

Abstract

The main goal of implementing a cardiovascular predictive analytics model is to determine the presence of this disease amongst patients of all medical backgrounds and demographics. As it currently stands, cardiovascular disease is the leading cause of death globally. Patient medical data can be analyzed for effective decision making for heart disease. The need for a predictive analytical model is crucial because its usage will help determine the presence or the absence of cardiovascular disease. The prior use of machine learning techniques to identify cardiovascular disease based on previously seen data is constantly changing because new data points are being generated as more patients are diagnosed. This historical data has proven to be quite useful in medical data analysis. This paper documents the results and performance of six machine learning classifiers and their respective ability to predict the presence of heart disease. The machine learning models that are explored include: decision trees, gaussian naïve bayes, k-nearest neighbors, logistic and linear regressions, and an artificial neural network. Moreover, the results revealed that the artificial neural network was able to outperform the other classifiers in terms of accuracy, recall, precision, and f1-measure.

Related Work

The most recent attempt to implement a predictive analytics model mainly focused on using sensor data to create an accurate classifier (Muniasamy, Bhatnagar 2020). Their methodology included gathering data from actual hospital equipment and then performing their analysis against multiple machine learning classifiers; namely decision trees, k-means and random forests. Their results showed that the random forest provided the highest accuracy rate at 84%. Another attempt at cardiac data mining included a big-data approach through the use of PySpark to perform basic machine learning algorithms (Malik, Bilal 2013). Their main advantage was that they were able to perform their analysis while leveraging additional hardware resources. This allowed them to fine-tune their neural network. A further approach at predicting whether or not a person will develop heart disease over a time period of time was made through the use of past family data (Rumsfeld JS, 2016). Their model examined data points from the ancestors of the given patient and then forecasted if they would develop heart disease over a period of 15 years. Their research mainly focuses on identifying symptoms of heart disease at an early age and then identifying the appropriate treatment. A further study was conducted at Tufts University which estimates the probability of contracting heart disease while participating in a clinical trial to help lower blood pressure (Wessler, B 2015). Their main analysis revolved around using k-means clustering to determine low and high-risk participants in the trial. They concluded that the lower a patient's blood pressure is, the less lightly they are to get heart disease. A deeper dive into predicting various types of heart problems such as artery disease, heart failure and the possibility of having a stroke are explored (Krittanawong, C. 2020). Their implementation included building a neural network to provide a label to a patient as having heart disease, artery disease, heart failure or a stroke. Their neural net outperformed support vector machines, decision trees and k-means. The difference in performance amongst the models were very minimal.

A different application of machine learning in cardiovascular disease is the use of ML to determine the best medication to prescribe to the patient. The use of past cardiovascular patient prescription history can be analyzed in an attempt to suggest the best medication for a new patient (Shameer K, 2018). Their methodology involves using dimensionality reduction to identify the top medication to treat heart disease. A further implementation is through the use of calcium and potassium levels as the main factors in determining heart disease. Low levels of both factors typically point towards cardiovascular disease (Sniderman AD, 2015). Another domain of heart disease is the prediction of sudden cardiac attacks through the use of naïve bayes classifiers (Bhatt Anurag, 2017). Their research focused on using Naïve Bayes to determine whether or not a patient is susceptible to a heart attack. Their results were fairly decent at 76%. Outside of tabular data, computer vision and natural language processing has already been applied to detect heart diseases (Kilic Arman 2019). Computer vision techniques determine the overall health of the heart based on images and multiple convolutions while NLP was used to interpret handwritten notes that were made by doctors and nurses. Lastly, a recurrent neural network was used for predictions based on gender and overall health (Paulus Jessica 2016).

The approach of this paper differs such that multiple ma-

chine learning algorithms are examined in terms of accuracy and instead of using a recurrent neural network, an artificial neural network is implemented. Our method uses 10-fold cross validation to measure the performance of the classifier. The proposed implemented model is able to take in unseen data and classify it as either positive or negative for cardiovascular disease.

Data

The initial dataset used in our analysis consists of 70,000 records of patient data with a total number of 11 features and 1 target variable. According to the data source, all of the preliminary features can be categorized into three group types which are either objective, subjective or examinational. An objective feature is defined as factual information about the patient such as age, height, weight and gender. A subjective feature is explained as information that is given by the patient such as smoking habits, alcohol intake and physical activity. The examinational feature is defined as the results of an actual medical examination performed by either a doctor or a nurse such as systolic blood pressure, diastolic blood pressure, cholesterol, and glucose levels. The target variable in the dataset was a binary cardiovascular disease feature which determines the presence of the absence of the disease. Aside from the initial features in the dataset, a calculated feature named body mass index (BMI) was added by dividing the patient's weight by their height squared. In order to perform this calculation, the height column needed to be converted from centimeters to meters by dividing the column values by 100. Another column that was added to the initial dataset was the obesity level column. This column was created by examining the patient's body mass index and then classifying each patient as being under weight, normal weight, overweight, stage 1 obesity, stage 2 obesity, or stage 3 obesity. The thresholds that were used for determining this column were defined by the Centers for Disease Control and Prevention (CDC).

The data pre-processing steps that were used includes the following: checking for null or missing values, identifying possible outliers, and then removing them, and finally min-max normalization to a better feel for the data. The first step in the data preparation process was to remove the patient id column since it provided no useful insights. The following step was to represent the obesity column numerically since it contained textual data. This was done by using value labels from 0 to 5 which represented the entire categorical values from under weight all the way to stage 3 obesity. A simple dataframe null check revealed that there were no missing values. The process that was used to determine possible outliers mainly revolved around blood pressure extreme values. Based on a published study that was performed by multiple doctors, the maximum attainable blood pressure is 270/260 (Brzezinski WA 1990). Given this information, the systolic and diastolic blood pressure columns were checked against those thresholds and the records were appropriately removed. The final data pre-processing step was to normalize the entire dataset using minimum and maximum values such that column value for each row gets transformed into a decimal format between 0 and 1 inclusive. The dataset had to be normalized before any data mining techniques could be applied because each feature variable was measured at different scales. Without performing normalization, the applied models will be skewed and potentially introduce a certain degree of bias.

Methodology

The machine learning models that were implemented were decision trees, naïve bayes classifiers, k-means clustering, logistical and linear regression and an artificial neural network. For predictive analytics purposes, the cardiovascular disease variable was removed from the X – variable. The dataset was split with a 70:30 percent ratio of training and testing. The decision tree application used all features and was evaluated in terms of its Gini index to determine impurity. The training and testing sets were fit on the model and then predictions were made using both the testing sets along with unseen patient data that was not initially in the original dataset. The actual decision tree itself, resulted in multiple children nodes in which all features are considered. The naïve bayes classifier was implemented in a similar manner using its default parameters.

The implementation of k-means relied on first determining the optimal number of clusters that should be used. The procedure used to find the ideal k-value was to create an elbow graph which showed the inertia values from a range of 1 through 14 and then identify the value which most likely resembles the bend in the shape of an elbow. The value that was determined from the graph in our case was k=6. In order to confirm the appropriate k-value, the silhouette score was calculated using a range of potential k-values and the results showed that 6 was the best choice since it had the highest silhouette score. The clustering scatter plots were generated for each feature in the dataset against each column. Even though that there are 6 clusters with very few outliers, the actual data points along with the centroids are still close together as seen in the following graph:

Figure 1: k-means cluster

A more concise and clearer graph can be generated from an un-normalized dataset because all of the datapoints are within the range of 0 to 1.

The two-regression analysis that were performed were logistical and linear. A logistical regression was performed since it is a special form of regression analysis in which the target variable is binary. The ROC curve shows the false positives contrary to the true positives.

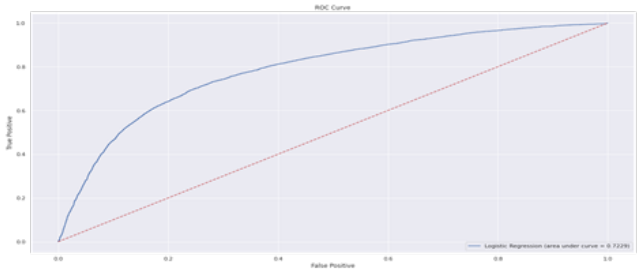

Figure 2: ROC Curve

The linear regression model used all of the features but was not quite as effective when compared to the logistic regression. Below is one of the results of the linear regression model:

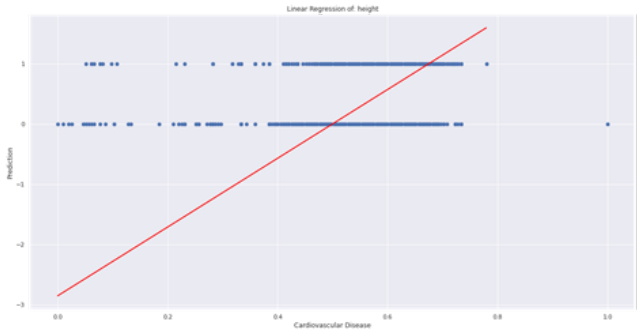

Figure 3: Linear Regression

It is possible that performing this linear regression on an unnormalized dataset will result in a more accurate result. It will also provide a better slope and intercept value.

The neural network architecture that was used is rather simple in that it only has 4 Dense layers. A sequential TensorFlow model was used where the loss function was binary cross entropy because the target variable was either positive or negative for cardiovascular disease. The input layer takes 13 dimensions which is the total number of the feature space and outputs 7 neurons. Each dense layer afterwards sees a reduction in the number of neurons. The activation function for the first three layers are relu which uses the max function to return the value of neurons. If the input is negative, the output is always 0. The activation function for the last layer is sigmoid because it is used to predict probability as an output between the range of 0 and 1. The model was compiled and validated using training and testing sets with a total of 400 epochs. The following graph shows the change in the accuracy of the neural train over all of the epochs.

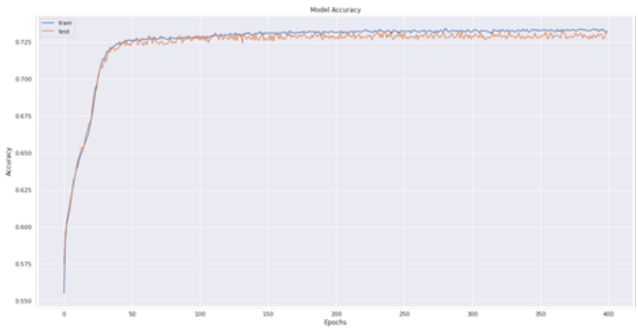

Figure 4: Neural Net Accuracy

The model accuracy seemed to stabilize after about 50 epochs. The main reason that the model stabilized was that we set the number of neurons in the hidden layer to be the median value of the total number of features in the dataset. Predictions were then made against the same testing sets that were used in the other machine learning models.

Results

The performances of all of these models were evaluated in terms of their accuracy, precision, recall and f1-score. For each model, a classification report was generated with the prediction variable and the same y-test variable. The results are displayed in the chart below and show how close each classifier performed in relation to the others:

	Accuracy	Precision	Recall	F1
Decision Tree	0.63	0.64	0.64	0.64
Naive Bayes	0.69	0.67	0.79	0.72
Neural Net	0.73	0.73	0.74	0.73
K-Means	0.68	0.56	0.47	0.51
Logistic Reg	0.72	0.70	0.77	0.74
Linear Reg	0.72	0.72	0.69	0.71

Table 1: Evaluation Metrics Of All Machine Learning Models

The key conclusion from this chart is that the neural network was slightly able to perform better than the other models in terms of its accuracy, precision, recall and f1-score. The main reason for the neural network to outperform the other is that it had multiple layers in the architecture. The decision tree model was the worst out of all the tested models. The results of my models were somewhat close to results of previously implemented models with a tolerance of about 5 percent.

In order to thoroughly determine the effectiveness of our model solution, we examined the p-value of the results with respect to the null hypothesis. In this analysis, we assumed the level of significance to be less than or equal to 0.05. Any p-value which is less than this threshold should be accepted and the null hypothesis is ultimately rejected.

The results of the p-value calculations show that only half of the implemented models turned out to be significant:

	t-value	p-value	result
Decision Tree	0.05	0.35	reject
Naive Bayes	0.05	0.08	reject
Neural Net	0.05	0.039	accept
K-Means	0.05	0.31	reject
Logistic Reg	0.05	0.022	accept
Linear Reg	0.05	0.017	accept

Table 2: p-value results of all predictive models

Since the p-value is less than the level of significance, the null hypothesis can be rejected and the alternative hypothesis can be accepted for the neural network, logistical regression and linear regression models. Unfortunately, the null hypothesis must be accepted for the decision tree, naive bayes classifier and k-means since the p-value was greater than the level of significance.

Final Conclusion

The need for a predictive analytics cardiovascular disease model is trending now than ever before. There is an overwhelming demand by doctors to be able to accurately classify whether or not their patient has heart disease. Our experimentation seems to be able to handle new and unseen data quite well when compared to other implementations. Both neural networks and logistical regressions provide almost identical results, and both can be improved with further training. In final analysis, this is how the application of machine learning affects the medical industry.

References

Muniasamy A., Muniasamy V., Bhatnagar R. (2020) Predictive Analytics for Cardiovascular Disease Diagnosis Using Machine Learning Techniques. In: Hassanien A., Bhatnagar R., Darwish A. (eds) Advanced Machine Learning Technologies and Applications. AMLTA 2020. Advances in Intelligent Systems and Computing, vol 1141. Springer, Singapore.

Malik, Bilal Hussain, Masood Basharat, Iqra Fatima, Mamuna. (2013). Cardiac Data Mining (CDM); Organization and Predictive Analytics on Biomedical (Cardiac) Data. AIP Conference Proceedings. 1559. 10.1063/1.4825018.

Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. Nat Rev Cardiol. 2016 Jun;13(6):350-9. doi: 10.1038/nrcardio.2016.42. Epub 2016 Mar 24. PMID: 27009423.

Wessler, B. et al. "Clinical Prediction Models for Cardiovascular Disease: Tufts Predictive Analytics and Comparative Effectiveness Clinical Prediction Model Database." Circulation: Cardiovascular Quality and Outcomes 8 (2015): 368–375.

Krittanawong, C., Virk, H.U.H., Bangalore, S. et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. Sci Rep 10, 16057 (2020).

Shameer K, Johnson KW, Glicksberg BS, et al Machine learning in cardiovascular medicine: are we there yet? Heart 2018; 104:1156-1164.

Sniderman AD, D'Agostino Sr RB, Pencina MJ. The Role of Physicians in the Era of Predictive Analytics. JAMA. 2015;314(1):25–26. doi:10.1001/jama.2015.6177.

Bhatt Anurag, Dubey Sanjay Kumar, Bhatt A., Sudden Cardiac Arrest Prediction Using Predictive Analytics, International Journal of Intelligent Engineering Systems, 2017.

Kilic Arman, Artificial Intelligence and Machine Learning in Cardiovascular Health Care, Science Direct Journals, 7 November 2019.

Paulus Jessica K., Wessler Benjamin S., Lundquist Christine, Lai Lana L.Y., Raman Gowri, Lutz Jennifer S., Kent David M., Field Synopsis of Sex in Clinical Prediction Models for Cardiovascular Disease, American Heart Association, Circulation: Cardiovascular Quality and Outcomes. 2016;9: S8–S15

CDC. "Defining Adult Overweight and Obesity." Centers for Disease Control and Prevention, 2020, https://www.cdc.gov/obesity/adult/defining.html.

Brzezinski WA. Blood Pressure. In: Walker HK, Hall WD, Hurst JW, editors. Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition. Boston: Butterworths; 1990. Chapter 16.