Predicting the Onset of Coronary Atherosclerosis in ICU Patients with MIMIC-IV

CSCI 1420 (Machine Learning) Capstone

Suraj Zaveri (szaveri3), Spring 2023

Introduction

Heart disease remains the leading cause of death in the world, resulting in one out of four deaths in the United States in 2017, and coronary artery disease by atherosclerosis is the most common type of heart disease. Atherosclerosis refers to a condition where the arteries narrow as the fatty deposits called plaque build up in the wall. The risk and the mortality of the disease can be used to diagnose as a preventative care indicator to guide the development of personalized treatment in the critical care setting. Past studies have identified hypercholesterolemia, hypertension, and atherosclerosis as direct predictors of cardiovascular disease, but there is minimal existing literature documenting the prediction of atherosclerosis using hypercholesterolemia and hypertension. To solve this problem, several machine learning models (namely logistic regression and decision tree) to predict the diagnosis of atherosclerosis were developed, with the aim for the model to be implemented within a clinical context to assist medical staff.

Methodology

Link to Github Repository: https://github.com/szaveri3/mlcapstone.git

MIMIC-IV Dataset

MIMIC-IV is a large dataset of deidentified patient data collected between 2001 and 2012 at Beth Israel Deaconess Medical Center in Boston, MA. It includes data containing relevant lab results, demographic information, and diagnoses. This data, when employed to train and test the various machine learning models could help point out a correlation between certain risk factors and atherosclerosis. Such knowledge of the association between atherosclerosis and its risk factors can help clinicians rapidly identify atherosclerosis in a critical care setting, resulting in improved patient outcomes.

Data Used from MIMIC-IV

- admissions.csv - for patient subject IDs (de-identified)
- d_icd_diagnoses.csv - to look up codes for specific diagnoses
- diagnoses_icd.csv - for features of whether a patient was diagnosed with a specific disease
- patients.csv - demographic information such as age or gender

subject_id	hadm_id	seq_num	icd_code	icd_version
10000032	22595853	1	5723	9
10000032	22595853	2	78959	9
10000032	22595853	3	5715	9
10000032	22595853	4	07070	9
10000032	22595853	5	496	9
10000032	22595853	6	29680	9
10000032	22595853	7	30981	9
10000032	22595853	8	V1582	9

Figure 1: Sample Raw Diagnosis Data from diagnoses_icd.csv

Preprocessing (In preprocess.py)

1) preprocessAdmissions() - Sampled 10,000 random patients from the entire dataset and loaded into dataframe (dropping duplicates)
2) preprocessDiagnoses() - Gets the rows of the selected subjects from diagnoses_icd.csv and assigns a binary value, 0 or 1, for three diseases
    a) Atherosclerosis (label)
    b) Hypertension (feature)
    c) Hypercholesterolemia (feature)
3) preprocessPatients() - extracts demographic information from patients.csv and stores gender (M or F) as well as age group in different columns, once again as a binary value
    a) Age groups - <40, 40-59, 60-79, 80+ (based on literature)

master_df

	subject_id	Hypertension	Hypercholesterolemia	Atherosclerosis	Male	Female	Age <40	Age 40-59	Age 60-79	Age 80+
65957	11552378	1	0	0	1	0	0	0	0	1
401046	19295317	1	1	0	1	0	0	1	0	0
80866	11897489	1	1	1	1	0	0	0	0	1
311699	17225920	0	0	0	1	0	1	0	0	0
304154	17052504	0	0	0	1	0	0	1	0	0
211569	14917172	0	0	0	1	0	1	0	0	0
317963	17373113	0	0	0	1	0	0	1	0	0
424888	19855099	1	1	1	1	0	0	1	0	0
343925	17970554	1	1	1	1	0	0	1	0	0
40411	10959728	0	0	0	1	0	1	0	0	0

Figure 2: Preprocessed Master Dataframe (first 10 rows)

The final patient population for the models tested consisted of 9318 patients, 1546 who had a diagnosis of atherosclerosis, 4850 who had a diagnosis of hypertension, and 1254 who had a diagnosis of hypercholesterolemia.

Models Tested

- Logistic Regression - LogisticRegression class from scikit-learn
- Decision Tree - DecisionTreeClassifier class from scikit-learn
- Neural Network - MLPClassifier (Multi-Layer Perceptron) class from scikit-learn
- Majority Classifier of all three models - takes majority vote (2 out of 3) as prediction

Each model was trained and tested using the scikit-learn library's standard functions for fitting the models to the data and making predictions.

Accuracy

The Scikit Learn accuracy function was utilized to calculate accuracy. This function calculates the frequency of the predicted labels matching the binary labels where **0** is negative diagnosis of atherosclerosis and **1** is positive diagnosis of atherosclerosis.

Challenges

Lab Events

Initially, the goal was to incorporate labevents.csv from the MIMIC-IV dataset, which contained useful measurements from the literature, such as C-Reactive Protein, Cholesterol level, Troponin T, or Lipase. These metrics could be useful in predicting atherosclerosis because the labevents.csv file contains a column that flags the measurement if it is deemed as medically abnormal. However, the labevents.csv file was 13 GB large, and it took an extremely large amount of time to read the csv. Thus, due to time constraints, the features for lab events were dropped and replaced with demographics.

Model Accuracy by Type

The models were each initially returning the same accuracy (up to 6 decimal places), which hinted at flaws in either the setup of the models or the preprocessing. In the first iteration of featurization, weights were assigned to each of the age groups all in one column titled “Age” (<40 mapped to 1, 40-59 mapped to 2, 60-79 mapped to 3, and 80+ mapped to 4). This resulted in a lower accuracy for all the models; in order to fix this, separate columns were created for each age group and incorporated with binary values, such that if a patient was 45 years old, they had a 1 in the 40-59 column and 0 for all other age groups. This allowed the neural network to distinguish its accuracy as minimally better than the other models, but it did not completely solve the problem. These observations in accuracy could be due to the dataset itself, but they could also warrant further preprocessing in future studies.

Model Tuning

Initially, model tuning in this project posed a challenge due to the imbalanced distribution of diagnoses in the patient dataset. The low prevalence of atherosclerosis compared to hypertension and hypercholesterolemia created an imbalance in the target variable, making it difficult to achieve optimal performance.

In order to tune the parameters of each of the models, various values were iterated over and accuracy was computed. For example, in the Decision Tree, maximum depths from 5 to 50 in increments of 5 were tested. While each of these values resulted in similar accuracies, the maximum depth of 5 resulted in the highest, which is why it was chosen. A similar approach was followed for each classifier.

Final Results

Model	Accuracy
Logistic Regression	0.83
Decision Tree	0.83
Neural Network	0.84
Majority Classifier	0.83

Model arguments:

- Logistic Regression - C = 0.01 (optimized over C = 0.01, 0.1, 1, 10, 100)
- Decision Tree - Max Depth = 5, Min Samples Leaf = 0.16
- Neural Network (MLP Classifier) - Hidden Layer Size (100, 100)

Discussion

The accuracy results indicate that all three models perform similarly in predicting the diagnosis of atherosclerosis. The Neural Network model achieved slightly higher accuracy compared to the other two models and the Majority Classifier. However, the differences in accuracy among the models are minimal.

These results suggest that the selected features, hypercholesterolemia, and hypertension, have a reasonable predictive power for identifying atherosclerosis. The models' ability to predict atherosclerosis with a relatively high accuracy indicates a potential correlation between these risk factors and the presence of the disease.

Considering the distribution of diagnoses within the patient population, it is evident that atherosclerosis had the lowest prevalence compared to hypertension and hypercholesterolemia. This imbalance in the distribution of diagnoses may have influenced the model's performance, particularly in terms of accuracy.

When interpreting the results of the models, it is important to consider the class distribution in the dataset. In this case, the models were tested on a population where the prevalence of atherosclerosis was relatively low compared to the other two diagnoses. This imbalance may have affected the accuracy results, as the models might have a higher tendency to predict the majority class (negative diagnosis of atherosclerosis) due to the skewed class distribution.

To better evaluate the models' performance, additional metrics such as precision, recall, and F1-score could be considered in the future. These metrics take into account the true positive, false positive, and false negative rates, providing a more comprehensive evaluation of the models' predictive ability, especially when dealing with imbalanced datasets.

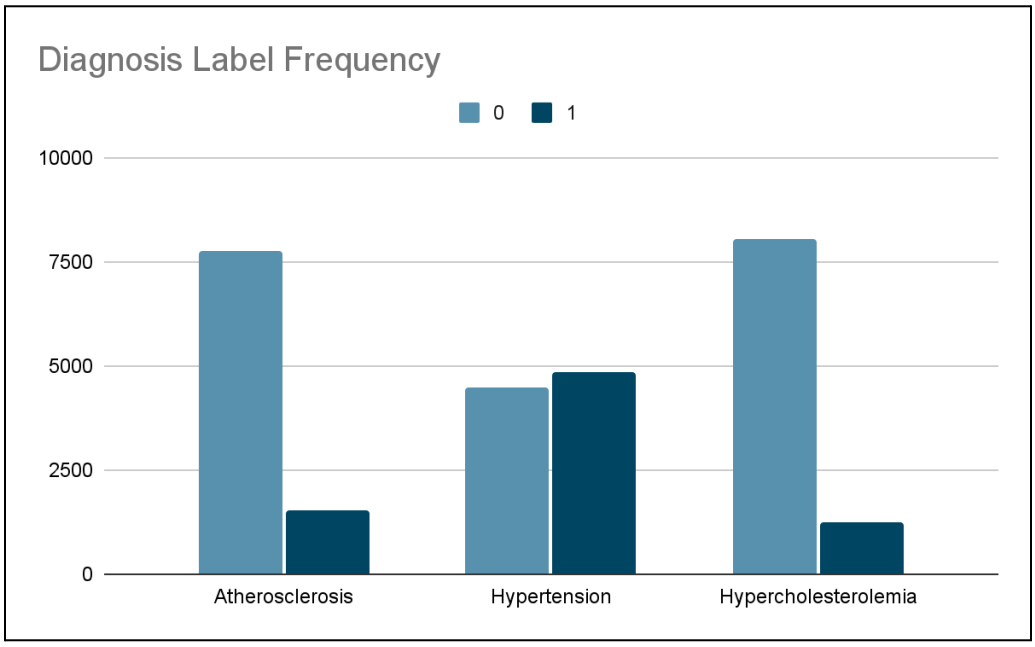

Figure 3: Distribution of Diagnoses Across Data

Ethical Considerations

The above project and the use of the MIMIC-IV dataset raise several ethical considerations that need to be addressed. Firstly, the dataset contains deidentified patient data, but it is essential to ensure the privacy and confidentiality of the individuals involved. Although the data is anonymized, there is still a risk of re-identification, especially with the increasing availability of external data sources and advanced data linkage techniques. Therefore, it is crucial to implement stringent data security measures and adhere to privacy regulations to protect patient confidentiality.

Secondly, bias in the dataset and the models should be carefully addressed. Biases may arise due to factors such as underrepresentation of certain demographic groups, systemic disparities in healthcare access, or variations in data collection practices. If these biases are not properly identified and mitigated, the models can perpetuate existing inequalities and lead to differential healthcare outcomes. It is crucial to conduct thorough data analysis to identify any biases and take appropriate steps to mitigate them, such as using balanced datasets or employing fairness-aware algorithms.

Moreover, the deployment of machine learning models in a clinical context must be done with caution. While these models can assist medical staff in decision-making, they should be viewed as decision support tools rather than replacements for human judgment. The models' predictions should be interpreted in conjunction with the expertise of healthcare professionals, and the final decisions should always consider the unique circumstances and preferences of individual patients. Transparent and interpretable models can help build trust and facilitate collaboration between the models and the medical staff.

Future Directions

Further research could explore additional features and refine the models to enhance their predictive performance. These additional features could include the lab events that were omitted as well as medications, past patient history, and further demographics. The MIMIC-IV dataset is so large and versatile that the only limitations in terms of data are how representative it is of the entire population of those with risk for coronary atherosclerosis.

Moreover, integrating these models into clinical decision support systems could provide real-time assistance to healthcare professionals, facilitating timely interventions and improving patient management. If a high enough accuracy is reached, a model such as the MLPClassifier or Majority Voting Classifier could be utilized as a method of flagging high-risk patients.

In conclusion, this study demonstrates the feasibility of using machine learning models to predict the diagnosis of atherosclerosis. The models achieved promising accuracy rates, suggesting that hypercholesterolemia and hypertension are potential indicators of atherosclerosis. These findings contribute to the growing body of knowledge in the field of cardiovascular disease prediction and highlight the potential for implementing machine learning in critical care settings.

Acknowledgements

This project was made possible through CSCI 1420 (Machine Learning) at Brown University. A special thank you to Professor Stephen Bach for teaching the concepts necessary to complete this project.

References

Ambale-Venkatesh, Bharath, Xiaoying Yang, Colin O. Wu, Kiang Liu, W. Gregory Hundley, Robyn McClelland, Antoinette S. Gomes et al. "Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis." Circulation research 121, no. 9 (2017): 1092-1101.

Boudoulas KD, Triposciadis F, Geleris P, Boudoulas H. Coronary Atherosclerosis: Pathophysiologic Basis for Diagnosis and Management. Prog Cardiovasc Dis. 2016 May-Jun;58(6):676-92. doi: 10.1016/j.pcad.2016.04.003. Epub 2016 Apr 25. PMID: 27091673.

Rodgers JL, Jones J, Bolleddu SI, Vanthenapalli S, Rodgers LE, Shah K, Karia K, Panguluri SK. Cardiovascular Risks Associated with Gender and Aging. J Cardiovasc Dev Dis. 2019 Apr 27;6(2):19. doi: 10.3390/jcdd6020019. PMID: 31035613; PMCID: PMC6616540.