Final Report

Coronary Artery Disease Detection Using Data Mining

Hanxiao Chen

Data Science Initiative, Brown University

https://github.com/hanxiao-chen/Data1030-project.git

Introduction

1. Problem description

Coronary artery disease (CAD), one of the main causes of heart attacks, may lead to life-threatening illness by reducing the blood and oxygen supplying, which induces approximately 7 million deaths worldwide every year.

To effectively lower the risk of CAD, an early diagnosis is of great importance. Machine learning, as an effective tool for prediction, may serve as a good way of accurate medical evaluation.

The application of machine learning in disease diagnosis issues has been an emerging research field. For instance, past research utilizes gene expression data to predict cancer, use MRI data to assess level of Alzheimer's disease, analyze search query and social media to track epidemics. When it comes to CAD evaluation, given that we have enough indicators highly associated with CAD such as a person's blood pressure, BMI, lifestyle, etc., it is reasonable to hypothesize that we can predict whether a person gets CAD by applying machine learning. Therefore, in this study, I focus on processing Z-Alizadeh Sani Data Set retrieved from UCI Machine Learning Repository to predict the occurrence of CAD.

Early research regarding this dataset mainly explored data collection and disease prediction methods. The result of Alizadehsani et al. (2013) indicated that Bagging SMO is the best model for prediction, with an accuracy of 90%. Yadav et al. (2014) found that Association Rule Data Mining is a reliable method when predicting CAD by comparing the overall performance of several models. Alizadehsani et al. (2012) examined stenosis of each vessel and checked the capacity of different algorithms. Their study showed that the accuracy of diagnosis of stenosis of each vessel could achieve 74.20%.

Inspired by previous scholars, the major objective of this study is to use machine learning algorithms like logistic regression, random forest, and AdaBoost with Z-Alizadeh Sani dataset (http://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani) to predict patients who are most likely to get CAD and compare the CAD prediction accuracy of different classification algorithms.

2. Data description

Z-Alizadeh Sani Data Set is 303*56 table, which has the medical records of 303 patients and their corresponding 55 features collected from Tehrans Shaheed Rajaei Cardiovascular, Medical and Research Centre. The last column of the data set contains the label of each person, which can either be CAD or Normal. Note that a person is categorized as CAD only if his/her diameter narrowing is greater than or equal to 50%. All the features, a brief description and the range of them are listed in Table 1.

Table 1 Data and descriptive statistics

Name of features	Brief description	Range	Name of features	Brief description	Range
Age	-	30-86	**Nonanginal CP**	-	Yes, No
Weight	-	48-120	**Exertional CP**	Exertional Chest Pain	Yes, No
Length	-	140-188	**Low Th Ang**	Low Threshold angina	Yes, No
Sex	-	Male, Female	**Rhythm**	-	Sin, AF
BMI	Body Mass Index (kg/m2)	18-41	**Q Wave**	-	Yes, No
DM	Diabetes Mellitus	Yes, No	**ST Elevation**	-	Yes, No
HTN	Hyper Tension	Yes, No	**ST Depression**	-	Yes, No
Current Smoker	-	Yes, No	**T inversion**	-	Yes, No
Ex-Smoker	-	Yes, No	**LVH**	Left Ventricular Hypertrophy	Yes, No
FH	Family History	Yes, No	**Poor R Progression**	Poor R Wave Progression	Yes, No
Obesity	Yes if MBI>25, No otherwise	Yes, No	**FBS**	Fasting Blood Sugar(mg/dl)	62-400
CRF	Chronic Renal Failure	Yes, No	**Cr**	Creatine (mg/dl)	0.5-2.2
CVA	Cerebrovascular Accident	Yes, No	**TG**	Triglyceride (mg/dl)	37-1050
Airway Disease	-	Yes, No	**LDL**	Low density lipoprotein (mg/dl)	18-232
Thyroid Disease	-	Yes, No	**HDL**	High density lipoprotein (mg/dl)	15-111
CHF	Congestive Heart Failure	Yes, No	**BUN**	Blood Urea Nitrogen (mg/dl)	6-52
DLP	Dyslipidemia	Yes, No	**ESR**	Erythrocyte Sedimentation rate (mm/h)	1-90
BP	Blood Pressure (mmHg)	90-190	**Hb**	Hemoglobin (g/dl)	8.9-17.6
PR	Pulse Rate (ppm)	50-110	**K**	Potassium (mEq/lit)	3-6.6
Ede ma	-	Yes, No	**Na**	Sodium (mEq/lit)	128-156
Weak peripheral pulse	-	Yes, No	**WBC**	White Blood Cell (cells/ml)	3700-18000
Lung Rales	-	Yes, No	**Lymph**	Lymphocyte (%)	7-60
Systolic murmur	-	Yes, No	**Neut**	Neutrophil (%)	32-89
Diastolic murmur	-	Yes, No	**PLT**	Platelet (1000/ml)	25-742
Typical Chest Pain	-	Yes, No	**EF**	Ejection Fraction (%)	15-60
Dyspnea	-	Yes, No	**Region with RWMA**	Region with Regional Wall Motion Abnormality	0,1,2,3,4
Function Class	-	1,2,3,4	**VHD**	Valvular Heart Disease	Normal, Mild, Moderate, Severe
Atypical	-	Yes, No			

Exploratory data analysis

1. Data balance

As shown in Fig.1, the ratio between CAD patients and Normal people is 7 to 3. This indicates that this dataset is a little bit imbalanced and thus stratified k-fold cross validation is necessary and the baseline accuracy is 70%.

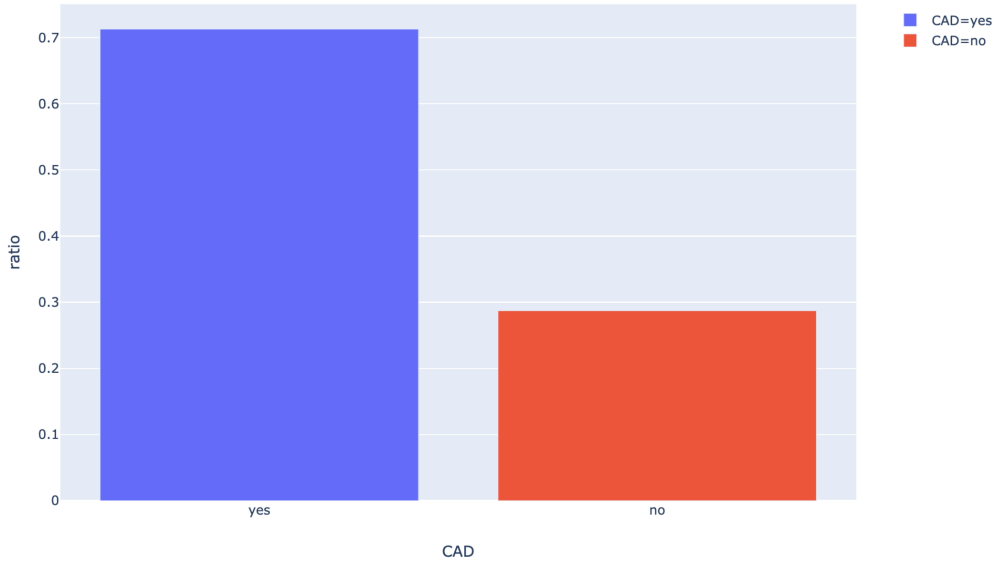

Fig.1 Ratio between the two labels. *The distribution of each class is represented by different colors: red represents normal people and blue represents CAD patients. This graph shows the imbalance of the dataset.*

2. Age and CAD

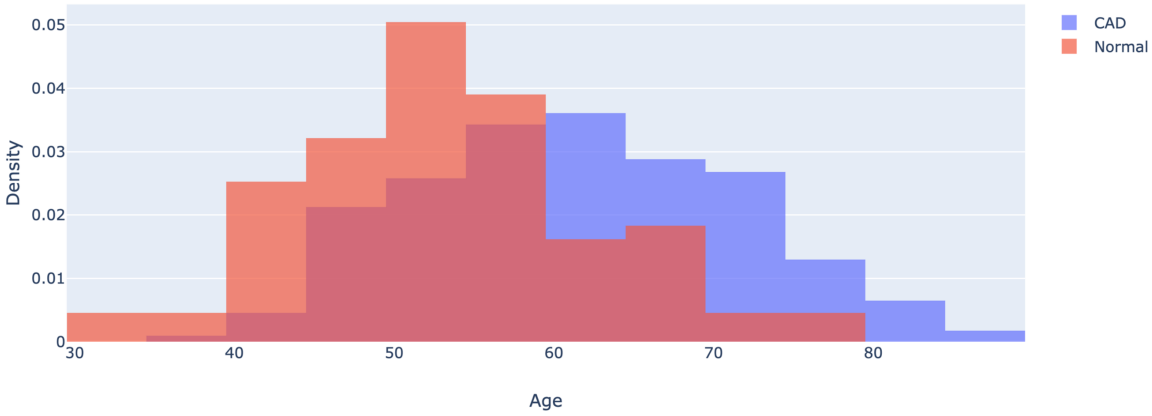

Fig.2: Age distribution of of CAD patients and normal people*. The distribution of each class is represented by different colors: red represents normal people and blue represents CAD patients. A positive correlation can be observed.*

This histogram in Fig.2 shows that people's age is approximately normally distributed and that the average age of CAD patients is greater than normal people's. Thus, it is reasonable to hypothesize that age is an effective indicator of the likelihood of being ill.

3. Chest pain and CAD

The bar plot in Fig.3 presents that nearly all patients that have chest pain get CAD, and half of the patients without chest pain have CAD. Whether a person has chest pain is a strong indicator of his/her health condition.

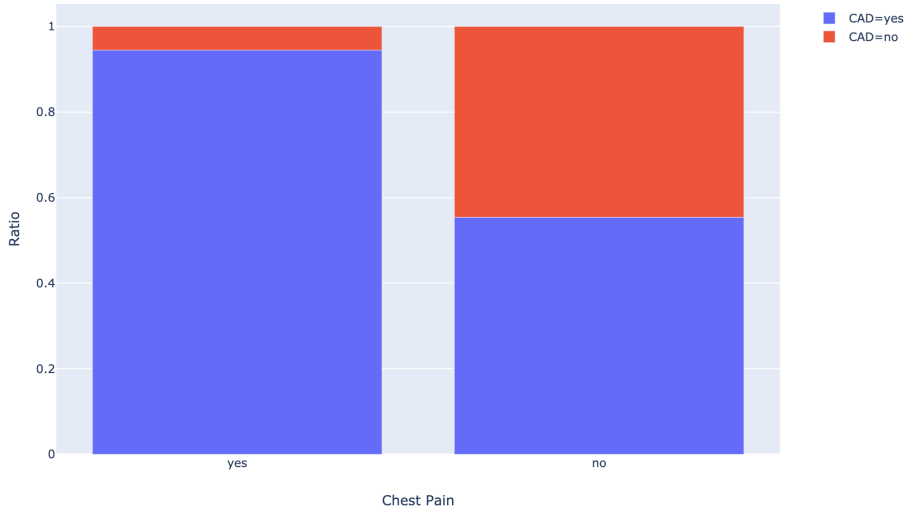

Fig.3: Relationship between chest pain and CAD*. The ratio of each class is represented by different colors: red represents normal people and blue represents CAD patients. A negative relationship can be observed.*

4. EF-TTE and CAD

EF-TTE is a measurement of how much blood the left ventricle pumps out with each contraction. The box plot presented in Fig.4 indicates that CAD patients tend to have a lower ejection fraction than normal people since a higher ejection fraction usually means a well-conditioned heart.

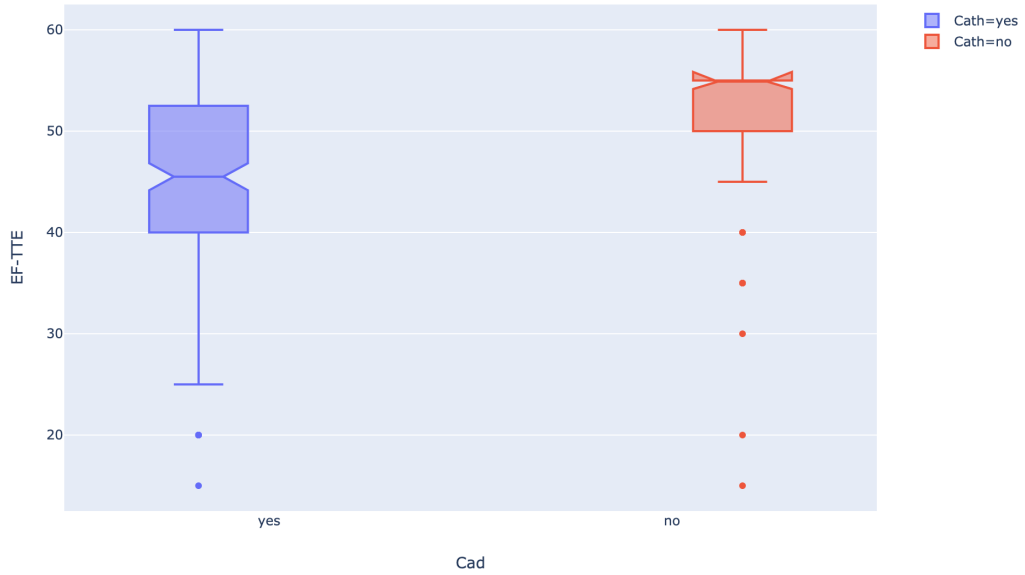

Fig.4 Relationship between EF-TTE and CAD. *The ratio of each class is represented by different colors: red represents normal people and blue represents CAD patients. A negative relationship can be observed.*

Methods

1. Data preprocessing

I apply OneHotEncoder to the features below because they are categorical features without order.

Sex, Obesity, CRF, CVA, Airway Disease, Thyroid Disease, CHF, DLP, Weak Peripheral Pulse, Lung rales, Systolic Murmur, Diastolic Murmur, Dyspnea, Function Class, Atypical, Nonanginal, Exertional CP, LowTH Ang, LVH, Poor R Progression, BBB, Region RWMA.

I apply OrdinalEncoder to feature “VHD” since it is classified into ordered categories: Normal, Mild, Moderate, and Severe.

I apply StandardScaler to the features below because they are continuous variables and I do not find any concrete evidence to transform them with MinMaxScaler.

Age, Weight, Length, BMI, BP, PR, HB, K, Na, Neut, FBS, CR, TG, LDL, HDL, BUN, ESR, WBC, Lymph, PLT, EF-TTE.

The following features are binary, and I do not need to transfer them:

DM, HTN, Current Smoker, EX-Smoker, FH, Edema, Typical Chest Pain, Q Wave, St Elevation, St Depression, Tinversion.

After preprocessing, there are 91 features in the preprocessed data.

2. Data split and uncertainty control

The whole dataset is divided into a training set and a test set and the size is the test set is 30%. Then I utilize 10-fold cross validation to find the optimal parameters. To control the uncertainty brought by randomly splitting and model specification (in random forest), I specify a set of random seeds each time I train a model, and compute the mean and standard deviation of every metric for each set of seeds. Ten sets of seeds are specified for each model to get a better approximation of the performance of the model.

3. Models and parameters

In this study, six classification models are considered as candidates to predict CAD. A variety of models are taken into consideration, which include probability-based classifier (logistic regression), distance-based classifiers (support vector machine, KNN), and ensemble models (random forest, AdaBoost, XGBoost). The diversity guarantees the extensiveness of this comparison study. The tuning parameters of each model and their corresponding range are listed in Table 2.

Table 2 Models and tuning parameters

Model	Tuning parameter(s)	Range of parameter(s)
Logistic regression	C	np.logspace(1,-3,num=50)
Random forest	max_depth min_samples_split	range(2,11) range(2,11)
Support vector machine	C gamma	np.logspace(6,-2,num=55) np.logspace(6,-2,num=55)
KNN	n_neighbors p (power parameter for the Minkowski metric)	range(3,15) range(3,15)
Adaptive Boosting	max_depth learning_rate	range(1,15) np.logspace(1,-4,num=25)
Gradient Boost	n_estimators learning_rate	range(10,201,10) np.logspace(1,-4,num=25)

4. Evaluation metrics

To measure the correctness of classification, five metrics (accuracy, precision, recall, F_β, confusion matrix) are computed for each of the models. Note that we hope to find as many patients as we could because false negative cases would cause more serious consequences during disease diagnosis. For this reason, I set β to be 1.3 instead of 1 when calculating F_β since I assign more weight to recall.

Results

1. Model evaluation metrics

The evaluation metrics of each model and their corresponding parameters are summarized in Table 3.

Table 3 Models and evaluation metrics

Model (parameters)	Accuracy (baseline: 0.7)	Precision	Recall	$F_{1.3}$	Confusion Matrix		
					0	1	
Logistic regression (C=0.719)	0.871 ± 0.030	0.908 ± 0.021	0.915 ± 0.033	0.912 ± 0.025	0.773	0.090	0
					0.227	0.909	1
Random forest (max_depth=6, min_samples_split=2)	0.852 ± 0.037	0.863 ± 0.033	0.948 ± 0.024	0.914 ± 0.026	0.807	0.134	0
					0.193	0.866	1
Support vector machine (C=1362, gamma=0.56)	0.847 ± 0.041	0.879 ± 0.036	0.918 ± 0.030	0.902 ± 0.028	0.748	0.100	0
					0.252	0.899	1
K-nearest neighbors (n_neighbors=3, p=13)	0.754 ± 0.031	0.803 ± 0.047	0.879 ± 0.015	0.848 ± 0.016	0.401	0.265	0
					0.598	0.734	1
Adaptive Boosting (max_depth=6, learning_rate=0.019)	0.820 ± 0.028	0.877 ± 0.017	0.876 ± 0.036	0.875 ± 0.025	0.699	0.151	0
					0.300	0.849	1
Gradient Boost (n_neighbots=160, learning_rate=0.063)	0.824 ± 0.030	0.883 ± 0.018	0.876 ± 0.041	0.877 ± 0.025	0.783	0.152	0
					0.217	0.848	1

Among all the six models, logistic regression is better than the rest of the models in terms of accuracy, and precision. Random forest is also an acceptable model because it has the highest recall and $\boldsymbol{F_{1.3}}$, which is necessary when applied to medical examination. The confusion matrix of them looks very similar. K-nearest neighbors model has the worst performance in nearly all

the metrics. The accuracy of all the models are greater than baseline accuracy (0.7).

2. Feature importance

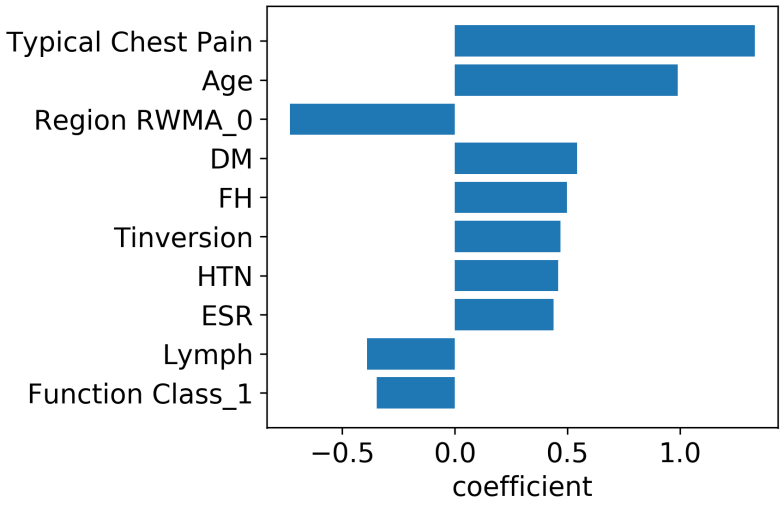

Fig.5 Coefficients of logistic regression. *Typical Chest Pain, Age, DM, FH, Tinversion, HTN, ESR are positively related to CAD condition, and Rrgion RWMA_0, Lymph, and Function Class_1 are negatively associated with CAD.*

The coefficients of logistic regression can be interpreted as the significance of each feature. The only two features whose coefficients are greater than 1 are Typical Chest Pain and Age. In accordance with EDA, Typical Chest Pain is the most important feature to diagnose CAD and the age of a person is also strongly positively related to the likelihood of getting the disease. Other features like Region RWMA_0, DM, FH, and Tinversion also play key roles in CAD detection.

3. Real-life application

Results from EDA and feature importance both indicate that Chest Pain and Age are a key feature that greatly affects the prediction outcomes. This is useful information because we can use this fact to do a rough physical examination by ourselves. In real life, we need to be aware of CAD and go to see a doctor if we are old enough and if we feel that we have chest pain.

The classifier I previously obtained helps potential patients to make an early diagnosis of CAD, since in the current medical situation, doctors highly rely on angiography, which is a very costly technique. The application of machine learning provides a much cheaper way in that it is more accessible to most people.

Outlook

Sample size is the first weakness I have to mention in this analysis. A small sample size could result in high bias and variance since accuracy of a classifier greatly depends on sufficient training and test data. Moreover, other algorithms such as deep Learning can only be applied on a dataset made of millions of data points, and a small sample eliminates the possibility of employing such an effective classification technique.

The way that this dataset is collected may be another factor that affects the performance of classification. All the records in Z-Alizadeh Sani dataset are collected from a hospital. People who care more about their health are more likely to regularly participate in a physical examination and people who pay less attention to their physical conditions would not always show up in a hospital, which indicates the possibility that is dataset could be biased. My suggestion would be expanding the coverage of the survey and take people outside the hospital into consideration.

In order to improve the accuracy of the classifier, feature engineering could be a good choice and thus a conversation with domain experts is necessary.

Finally, although the speed of training is pretty fast in this project, this efficiency does not benefit from the algorithm itself, but the size of the dataset. In future work, dimension reduction techniques such as t-SNE will be used to exclude the redundancy in the data to reduce training cost.

Reference

[1] C. Yadav, S. Lade, and M. K. Suman, Predictive Analysis for the Diagnosis of Coronary Artery Disease using Association Rule Mining, International Journal of Computer Applications, vol. 87, 2014.

[2] R. Alizadehsani, J. Habibi, M. J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, et al., A data mining approach for diagnosis of coronary artery disease, Computer Methods and Programs in Biomedicine, vol. 111, pp. 52-61, 2013/07/01/ 2013.

[3] Lohita, Kodali et al. Performance Analysis of Various Data Mining Techniques in the Prediction of Heart Disease. Indian Journal of Science and Technology, [S.l.], dec. 2015. ISSN 0974 -5645.

[4] R. Alizadehsani, J. Habibi, B. Bahadorian, H. Mashayekhi, A. Ghandeharioun, R. Boghrati, et al., Diagnosis of Coronary Arteries Stenosis Using Data Mining, Journal of Medical Signals and Sensors, vol. 2, pp. 153-159, Jul-Sep

[5] R. Alizadehsani, J. Habibi, Z. Alizadeh Sani, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, et al., Diagnosing Coronary Artery Disease via Data Mining Algorithms by Considering Laboratory and Echocardiography Features, Research in Cardiovascular Medicine, vol. 2, pp. 133-139, 07/31