Problem description

Cardiovascular diseases are the first cause of death globally. According to the estimation of World Health Organization, in 2016, 17.9 million people died from cardiovascular diseases, making up 31% of all global deaths. 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.

The major objective of this study is to employ machine learning algorithms like logistic regression, random forest, and Neural Network with Z-Alizadeh Sani Data Set (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. The project is available at GitHub (https://github.com/hanxiao-chen/Data1030-project).

Data and descriptive statistics

Z-Alizadeh Sani Data Set is 303*56 table, which has the medical records of 303 patients of the data set complied by collecting 303 patients from Tehran's Shaheed Rajaei Cardiovascular, Medical and Research Centre and their corresponding 55 features. The features can be described from four dimensions: demographic, symptom and examination, ECG, and laboratory and echo features. 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, their brief description, and range are listed in the table below:

Name of features	Brief description	Range
Age		30-86
Weight		48-120
Length		140-188
Sex		Male, Female
BMI	Body Mass Index (kg/m2)	18-41

DM	Diabetes Mellitus	Yes, No
HTN	Hyper Tension	Yes, No
Current Smoker		Yes, No
Ex-Smoker		Yes, No
FH	Family History	Yes, No
Obesity	Yes if MBI>25, No otherwise	Yes, No
CRF	Chronic Renal Failure	Yes, No
CVA	Cerebrovascular Accident	Yes, No
Airway Disease		Yes, No
Thyroid Disease		Yes, No
CHF	Congestive Heart Failure	Yes, No
DLP	Dyslipidemia	Yes, No
BP	Blood Pressure (mmHg)	90-190
PR	Pulse Rate (ppm)	50-110
Ede ma		Yes, No
Weak peripheral pulse		Yes, No
Lung Rales		Yes, No
Systolic murmur		Yes, No
Diastolic murmur		Yes, No
Typical Chest Pain		Yes, No
Dyspnea		Yes, No
Function Class		1,2,3,4
Atypical		Yes, No
Nonanginal CP		Yes, No
Exertional CP	Exertional Chest Pain	Yes, No
Low Th Ang	Low Threshold angina	Yes, No
Rhythm		Sin,AF
Q Wave		Yes, No
ST Elevation		Yes, No
ST Depression		Yes, No
T inversion		Yes, No
LVH	Left Ventricular Hypertrophy	Yes, No

Poor R Progression	Poor R Wave Progression	Yes, No
FBS	Fasting Blood Sugar(mg/dl)	62-400
Cr	Creatine (mg/dl)	0.5-2.2
TG	Triglyceride (mg/dl)	37-1050
LDL	Low density lipoprotein (mg/dl)	18-232
HDL	High density lipoprotein (mg/dl)	15-111
BUN	Blood Urea Nitrogen (mg/dl)	6-52
ESR	Erythrocyte Sedimentation rate (mm/h)	1-90
Hb	Hemoglobin (g/dl)	8.9-17.6
K	Potassium (mEq/lit)	3-6.6
Na	Sodium (mEq/lit)	128-156
WBC	White Blood Cell (cells/ml)	3700-18000
Lymph	Lymphocyte (%)	7-60
Neut	Neutrophil (%)	32-89
PLT	Platelet (1000/ml)	25-742
EF	Ejection Fraction (%)	15-60
Region with RWMA	Region with Regional Wall Motion Abnormality	0,1,2,3,4
VHD	Valvular Heart Disease	Normal, Mild, Moderate, Severe

Note that features of the same category (demographic, symptom and examination, ECG, and laboratory and echo features) are of the same color.

Related work

Alizadehsani et al. (2013) applied different classification techniques to predict CAD. Their result indicated that SMO and Bagging SMO are the best models for prediction, with the accuracy of 94.08% and 93.40%, respectively. With the same dataset, Lohita et al. (2015) compared the heart disease prediction accuracy of different data mining classification methods and concluded that the Bagging algorithm achieved highest accuracy of 97.39%. Yadav et al. (2014) found that Association Rule Data Mining is a reliable method when predicting CAD by comparing the overall performance of accuracy, sensitivity, and specificity of several models. Alizadehsani et al. (2012) examined stenosis of each vessel and checked the capacity of algorithms such as Naïve Bayes, and k-nearest. Their study indicated that the accuracy of diagnosis of stenosis of each vessel could achieve 74.20% for Left Anterior Descending, 63.76% for Left Circumflex and 68.33% for Right Coronary Artery.

Data preprocessing

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

Sex, DM, HTN, Current Smoker, EX-Smoker, FH, Obesity, CRF, CVA, Airway disease, Thyroid Disease, CHF, DLP, Edema, Weak Peripheral Pulse, Lung rales, Systolic Murmur, Diastolic Murmur, Typical Chest Pain, Dyspnea, Function Class, Atypical, Nonanginal, Exertional CP, LowTH Ang, Q Wave, St Elevation, St Depression, Tinversion, 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.

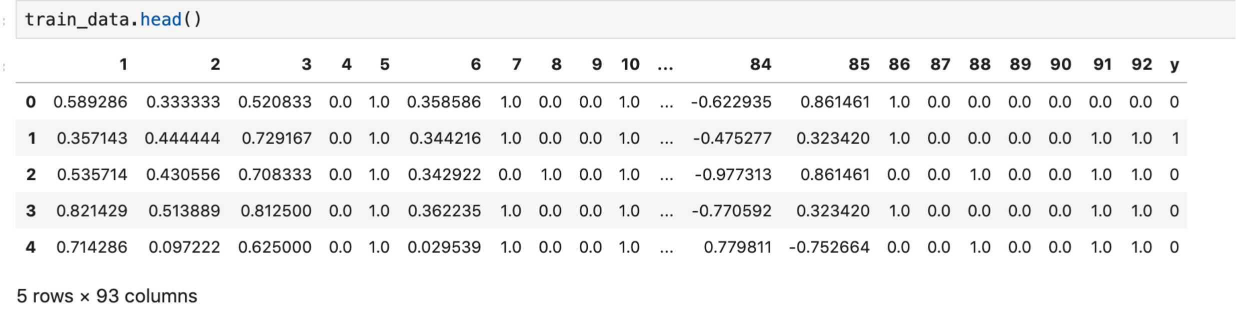
I apply StandardScaler to the features below because they are continuous variables bounded by a number.

Age, Weight, Length, BMI, BP, PR, HB, K, Na, Neut

I apply StandardScaler to the features below because they are continuous variables and sometimes may appear extreme values.

FBS, CR, TG, LDL, HDL, BUN, ESR, WBC, Lymph, PLT, EF-TTE

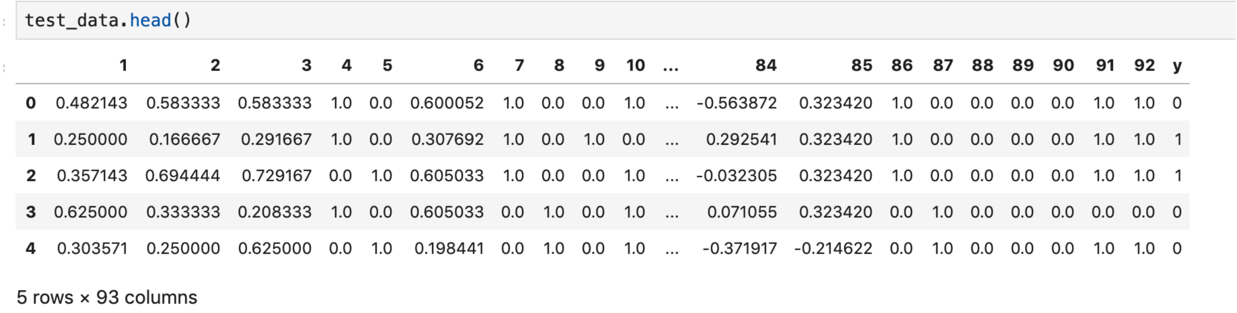
After preprocessing, there are 93 features in the data and a screen shot after preprocessing of training data and testing data are in the below.



```
train_data.head()
```

	1	2	3	4	5	6	7	8	9	10	...	84	85	86	87	88	89	90	91	92	y
0	0.589286	0.333333	0.520833	0.0	1.0	0.358586	1.0	0.0	0.0	1.0	...	-0.622935	0.861461	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0
1	0.357143	0.444444	0.729167	0.0	1.0	0.344216	1.0	0.0	0.0	1.0	...	-0.475277	0.323420	1.0	0.0	0.0	0.0	0.0	1.0	1.0	1
2	0.535714	0.430556	0.708333	0.0	1.0	0.342922	0.0	1.0	0.0	1.0	...	-0.977313	0.861461	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0
3	0.821429	0.513889	0.812500	0.0	1.0	0.362235	1.0	0.0	0.0	1.0	...	-0.770592	0.323420	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0
4	0.714286	0.097222	0.625000	0.0	1.0	0.029539	1.0	0.0	0.0	1.0	...	0.779811	-0.752664	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0

5 rows × 93 columns



```
test_data.head()
```

	1	2	3	4	5	6	7	8	9	10	...	84	85	86	87	88	89	90	91	92	y
0	0.482143	0.583333	0.583333	1.0	0.0	0.600052	1.0	0.0	0.0	1.0	...	-0.563872	0.323420	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0
1	0.250000	0.166667	0.291667	1.0	0.0	0.307692	1.0	0.0	1.0	0.0	...	0.292541	0.323420	1.0	0.0	0.0	0.0	0.0	1.0	1.0	1
2	0.357143	0.694444	0.729167	0.0	1.0	0.605033	1.0	0.0	0.0	1.0	...	-0.032305	0.323420	1.0	0.0	0.0	0.0	0.0	1.0	1.0	1
3	0.625000	0.333333	0.208333	1.0	0.0	0.605033	0.0	1.0	0.0	1.0	...	0.071055	0.323420	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0
4	0.303571	0.250000	0.625000	0.0	1.0	0.198441	0.0	1.0	0.0	1.0	...	-0.371917	-0.214622	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0

5 rows × 93 columns

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