May 4th, 12:00 AM

Cardiovascular Disease Prediction Modelling: A Machine Learning Approach

Usmaan Al-Shehab
Rowan University

Maduka Gunasinghe

Yousuf Elkhoga

Nimay Patel

Juliana Yang

ROWAN-VIRTUA School of Osteopathic Medicine

Cardiovascular Disease Prediction Modelling: A Machine Learning Approach

Usmaan Al-Shehab, Maduka Gunasinghe Yousuf Elkhoga, Nimay Patel, Juliana Yang

Introduction:

- The objective of this project is to utilize UCI's Heart Disease dataset to deduce which physiological biomarkers are most highly correlated with heart disease incidence.
- A predictive model can be developed using these biomarkers to calculate the probability of someone having or potentially developing a heart-related condition.
- The efficacy of predicting cardiovascular disease as an outcome was compared between three different machine learning algorithms.
- Support Vector Machine works by creating hyperplanes between data points to conduct classification.[1]
- Gaussian Naive Bayes works by using the conditional probabilities of events to classify the target.[1]
- In logistic regression, the independent variables included all features in the data set except for "target," which is a categorical variable that indicates whether the patient has cardiovascular disease. The dependent variable included the "target" variable.[1]
- The findings of this study can help healthcare professionals with developing new preventative protocols for assessing and treating cardiovascular disease.

Objectives:

- Identify which physiological factors are most strongly associated to heart disease
- Determine what type of model would be best for predicting one's risk for heart disease

Data Set Description[2]

#	Feature	Label	Description
1	Age	age	in years
2	Sex	sex	0 = female, 1 = male
3	Chest pain type	cp	0 = asymptomatic (i.e., no symptoms of chest pain exhibited) 1 = atypical angina (i.e., non-cardiac chest pain) 2 - non-anginal pain (i.e., esophageal contractions resemblant of angina) 3 - typical angina (i.e., chest pain caused by inadequate blood flow to the heart)
4	Resting blood pressure	trestbps	in mmHg
5	Serum cholesterol	chol	in mg/dl
6	Fasting blood sugar	fbs	>120 mg/dl (0 = false, 1 = true)
7	Resting ECG results	restecg	0 = probable or definite left ventricular hypertrophy or heart enlargement (irregular heartbeat) 1 = normal heartbeat 2 = ST-T wave abnormalities (abnormal heartbeat)
8	Maximum heart rate achieved during thallium stress test	thalach	beats/minute
9	Exercise induced angina	exang	0 = no, 1 = yes
10	ST depression induced by exercise relative to rest (i.e., ST segment below baseline)	oldpeak	ST relates to the positions on the ECG wave
11	Slope of the peak exercise ST segment (or heart rate slope)	slope	0 = downsloping (indicative of unhealthy heart) 1 = flat (typical of healthy heart) 2 = upsloping (indicative of healthy heart)
12	Number of major vessels colored by fluoroscopy (i.e., number of vessels through which blood flow is observed)	ca	0,1,2,3
13	Thallium stress test result (i.e., how well blood flows through one's heart during exercise or at rest)	thal	0 - NULL (dropped already) 1 - fixed defect (zero blood flow in some part of the heart) 2 - normal blood flow 3 - reversible defect (observed blood flow is not normal)
14	Presence of Disease	target	0 = disease, 1 = no disease

Support Vector Machine:

Classification Report:

	0	1	accuracy	macro avg	weighted avg
precision	0.82927	0.84	0.835165	0.834634	0.835047
recall	0.80952	0.857143	0.835165	0.833333	0.835165
f1-score	0.81928	0.848485	0.835165	0.833881	0.835004
support	42	49	0.835165	91	91

Accuracy: 82.2%

Confusion Matrix:

	Actual Positive	Actual Negative
Predicted Positive	34	4
Predicted Negative	6	47

Gaussian Naive Bayes (NB):

Classification Report:

	0	1	accuracy	macro avg	weighted avg
precision	0.86364	0.83871	0.849057	0.851173	0.849997
recall	0.79167	0.896552	0.849057	0.844109	0.849057
f1-score	0.82609	0.866667	0.849057	0.846377	0.848291
support	96	116	0.849057	212	212

Accuracy: 84.9%

Confusion Matrix:

	Actual Positive	Actual Negative
Predicted Positive	38	7
Predicted Negative	6	40

Logistic Regression:

Classification Report:

	0	1	accuracy	macro avg	weighted avg
precision	0.93617	0.898305	0.915094	0.917238	0.916166
recall	0.88	0.946429	0.915094	0.913214	0.915094
f1-score	0.90722	0.921739	0.915094	0.914478	0.914889
support	100	112	0.915094	212	212

Accuracy: 91.2%

Confusion Matrix:

	Actual Positive	Actual Negative
Predicted Positive	34	4
Predicted Negative	4	49

References:

[1] Bruce P, Bruce A, Gedeck P. Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. 2nd ed. O'Reilly Media; 2021.)

[2] Heart Disease UCI. https://archive.ics.uci.edu/ml/datasets/Heart+Disease (accessed February 17, 2021).