EEE 419/591 Project 1 Report

Taman Truong

October 6th, 2023

Introduction

The Acme Medical Analysis and Prediction Enterprises (AMAPE) wishes to develop an app that would help doctors predict whether or not patients have heart disease. To examine the effectiveness of an app like this, we performed data analysis and conducted six machine-learning model simulations on the given data of 270 patients based on 14 parameters, as shown in the table below. The last parameter is particularly important because this parameter determines whether or not the patient has heart disease based on the other 13 parameters. We begin by defining statistical terms and machine learning algorithms to aid the understanding of the reader in our analysis.

Abbreviated Name	**Description**
age	Age
sex	Sex
cpt	Chest Pain Type (Values 1, 2, 3, 4)
rbp	Resting Blood Pressure
sc	Serum Cholesterol in mg/dl
fbs	Fasting Blood Sugar > 120 mg/dl
rer	Resting Electrocardiographic Results (Values 0, 1, 2)
mhr	Maximum Heart Rate Achieved
eia	Exercise Induced Angina
opst	Oldpeak - ST Depression Induced by Exercise Relative to Rest
dests	Slope of Peak Exercise ST Segment
nmvcf	Number of Colored Vessels Colored by Fluoroscopy (nmvcf)
thal	Thalassemia (3 = Normal, 6 = Fixed Defect, 7 = Reversible Defect)
a1p2	Absence of Heart Disease (1 = absent, 2 = present)

Correlation is a statistical measure that expresses the strength in which two variables are (linearly) related to each other. This quantity is assigned a number from -1 to 1. Positive values for correlation indicate that the two variables tend to increase together, whereas negative values for correlation indicate that the two variables tend to decrease together. A value of zero for correlation indicates a near-negligible (linear) relationship between the two variables. **Covariance** is a statistical measure that expresses how much two variables vary together. This quantity is assigned a number ranging across all real numbers $(-\infty, \infty)$. Positive values for covariance indicate that the two variables tend to move together, whereas negative values for covariance indicate that the two variables tend to move in opposite directions to each other. A value of zero for covariance indicates a near-negligible relationship between the two variables.

Six supervised machine learning classifier algorithms were used in the analysis of the heart database: Perceptron, Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbors. **Perceptron** contains a single-layer neural network designed to classify binary classifiers. **Logistic Regression** is designed to predict the probability that an object belongs to a given class. **Support Vector Machines** are designed to find a N-dimensional hyperplane that splits up objects into N classes. **Decision Trees** classifies or regresses data based on answers to true-false questions, forming a tree graph. **Random Forests** generalizes decision trees, as forests consist of many trees that are processed together via ensemble learning algorithms, algorithms that combine one or more distinct or non-distinct existing classifiers. **K-Nearest Neighbors** uses proximities to classify data points in certain clusters or groups.

Analysis of Heart Database Data

For both correlation and covariance, we consider the magnitude of these values, since we are looking for the existence of a meaningful relationship between two variables, not necessarily the direction of the the relationship between the two variables.

Variable 1	Variable 2	Correlation Between Two Variables
dests	opst	0.609712
a1p2	thal	0.525020
a1p2	nmvcf	0.455336
a1p2	eia	0.419303
a1p2	mhr	0.418514
a1p2	opst	0.417967
a1p2	cpt	0.417436
mhr	age	0.402215
thal	sex	0.391046
dests	mhr	0.386847

Variable	Correlation to a1p2
thal	0.525020
nmvcf	0.455336
eia	0.419303
mhr	0.418514
opst	0.417967
cpt	0.417436
dests	0.337616
sex	0.297721
age	0.212322
rer	0.182091
rbp	0.155383
sc	0.118021
fbs	0.016319

Variable 1	Variable 2	Covariance Between Two Variables
sc	rbp	159.731185
sc	age	103.605452
mhr	age	84.874721
rbp	age	44.426394
mhr	sc	22.437340
mhr	rbp	16.193432
thal	mhr	11.391904
opst	mhr	9.260037
rer	sc	8.647005
mhr	cpt	6.992028

Variable	Covariance to a1p2
mhr	4.826518
sc	3.036762
age	0.962825
thal	0.507228
opst	0.238290
nmvcf	0.213961
cpt	0.197439
dests	0.103263
eia	0.098306
rer	0.090458
sex	0.069393
fbs	0.002891

We see that the slope of the peak exercise ST segment (dests) and the oldpeak - ST depression induced by exercise relative to rest are the two more highly correlated variables out of all distinct pairs of variables. This makes sense because these two variables are both related to the characteristics of the ST segment. We see that the six most highly correlated variables that influence the absence or presence of heart disease are thalassemia (thal), the number of colored vessels colored by fluoroscopy (nmvef), exercise-induced angina (eia), maximum heart rate achieved (mhr), oldpeak - ST depression induced by exercise relative to rest (opst), and chest pain type (cpt). Even those these variables influence the absence or presence of heart disease that aligns with the top two tables, the magnitude of all of the correlation variables display a weak to moderate relationship with respect to the absence or presence of heart disease.

We see that the serum cholesterol (sc) and the resting blood pressure (rbp) have the highest covariance out of all distinct pairs of variables. This makes sense because the higher or lower the cholesterol levels are, the higher or lower the blood pressure is. We see that the six highest covariance values that influence the absence or presence of heart disease are maximum heart rate achieved (mhr), serum cholesterol (sc), age (age), thalassemia (thal), oldpeak - ST depression induced by exercise relative to rest (opst), and the number of colored vessels colored by fluoroscopy (nmvcf). The magnitude of all of the covariance values displays a weak relationship with respect to the absence or presence of heart disease.

Analysis of Machine Learning Models on Heart Database

For all of the machine learning models used, all of the features from 270 patients in the heart database were used. To set up the training and testing datasets for machine learning simulations, the database was split so that 70% of the data served as the training dataset that the model would use to train for patterns and 30% of the data served as the testing dataset to test whether or not the model can detect patterns seem in the training data to the testing data to a sufficient accuracy. Each model had optimized parameters that generated the accuracies in predicting whether patients have heart disease, as shown in the table below.

Machine Learning Model	**Accuracy**
Perceptron	84%
Logistic Regression	85%
Support Vector Machine	85%
Decision Tree	79%
Random Forest	83%
K-Nearest Neighbors	84%

We see that two machine learning models generate the highest accuracy (85%) with optimized parameters: logistic regression and support vector machines. Only one machine learning model generated the lowest accuracy (79%) with optimized parameters: decision trees. The top five machine learning models consistently perform well, with a discrepancy at most 2% between machine learning model accuracies. However, to truly test how robust these machine-learning models can get, more data from more patients is required to potentially improve the performance of these machine-learning models. Even though logistic regression and support vector machines generated the highest accuracy with optimized parameters, logistic regression should be used with this data because the output parameter, the absence of presence of heart disease, is a binary classifier, and logistic regression is well-suited to deal with binary classifiers because of the natural shape of the graph of the logistic (sigmoid) function.