Rory Maguire

Dr. David Allee

EEE 591

10/3/2024

Machine Learning Project 1

13 different biometrics were recorded from 277 different individuals to create and analyze models used to predict heart disease, were “a1p2” is the presence of heart disease (Table 1). A collection of machine learning algorithms were implemented to compose these models. For each algorithm, 70% of the dataset was used to train the model, and the remaining 30% were used to test the accuracy of the model. Each variable was standardized using a standard scalar transformation to equalize the weights among each variable despite their unique ranges.

Table 1. Variable Names and Value Ranges

Name	Num	Description
age	0	age
sex	1	sex
cpt	2	chest pain type (4 values)
rbp	3	resting blood pressure
sc	4	serum cholestoral in mg/dl
fbs	5	fasting blood sugar > 120 mg/dl
rer	6	resting electrocardiographic results (values 0,1,2)
mhr	7	maximum heart rate achieved
eia	8	exercise induced angina
opst	9	oldpeak = ST depression induced by exercise relative to rest
dests	10	the slope of the peak exercise ST segment
nmvcf	11	number of major vessels (0-3) colored by flourosopy
thal	12	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
a1p2	13	absence of heart disease = 1, presence = 2

Of all the metrics recorded, the most correlated are found in Table 2. Variables correlated to a1p2 accounted for 6 of the top 7 correlated values. Although there are several variables with a relatively high correlation to a1p2 in relation to the other variables, the a1p2 correlation values are relatively low, ranging from 40% to 50%. The covariances to a1p2 were similarly unimpressive and were visualized using a pair plot (Figure 1). From the pair plot, it can be seen that there are no variables that are easily identified as highly covarying with a1p2. a1p2 was not among the most covarying values (Table 3). Having more than one variable, however, allows for

correlations in more than just two dimensions. Although these cannot be plotted, the combination of correlations of a few of the highest correlated variables to a1p2 will be helpful in predicting heart disease. It is interesting to note that the highest correlated variables are not the same as the highest covarying variables. Based on these results, it can be concluded that thal, nmvcf, eia, mhr, opst, and cpt will play a significant role in predicting the value of a1p2.

Table 2. Most Correlated Variables

Variable 1	Variable 2	Correlation (%)
dests	opst	61.0
a1p2	thal	52.5
a1p2	nmvcf	45.5
a1p2	eia	41.9
a1p2	mhr	41.9
a1p2	opst	41.8
a1p2	cpt	41.7

Table 3. Most Covarying Variables

Variable 1	Variable 2	Covariance
sc	rbp	159.7
sc	age	103.6
mhr	age	84.9
rbp	age	44.4
mhr	sc	22.4
mhr	rbp	16.2
thal	mhr	11.4

6 different machine learning methods were implemented on the given biometric dataset and scored for accuracy (Table 4). The linear kernel was used for Support Vector Machine. For K-Nearest Neighbors, 11 was chosen for the value of K. For Logistic Regression the ‘sag’ solver was used. In order to establish which of these is the best model to use for this dataset, the Test Accuracy and Training Accuracy were compared. In a 2-dimensional dataset, the data and model can be visualized and investigated to determine if the model is overfitting or underfitting. With a larger dataset such as this one, the fit characteristics can be studied by comparing the Test Accuracy to the Training Accuracy. Models with a Training Accuracy that greatly exceeds the Test Accuracy are overfit. For example, using a depth of 10 for Decision Tree Learning results in a Training Accuracy of 100% but a Test Accuracy of 74.1%. The goal was to create models where the Training Accuracy and Test Accuracy were relatively close to one another. For Random Forest, it was quite hard to minimize the difference between the two values. It seems that for this dataset, Random Forest has the tendency to overfit. Of each of the models being

analyzed, Logistic Regression had the best combination of accuracies. The Test Accuracy and Training Accuracy were 85.2% and 87.8% respectively. Because both the Training Accuracy and Test Accuracy were high and the difference between the two were low, Logistic Regression is the best machine learning method to use for this dataset based on this analysis.

Figure 1. Covariance Pair Plot

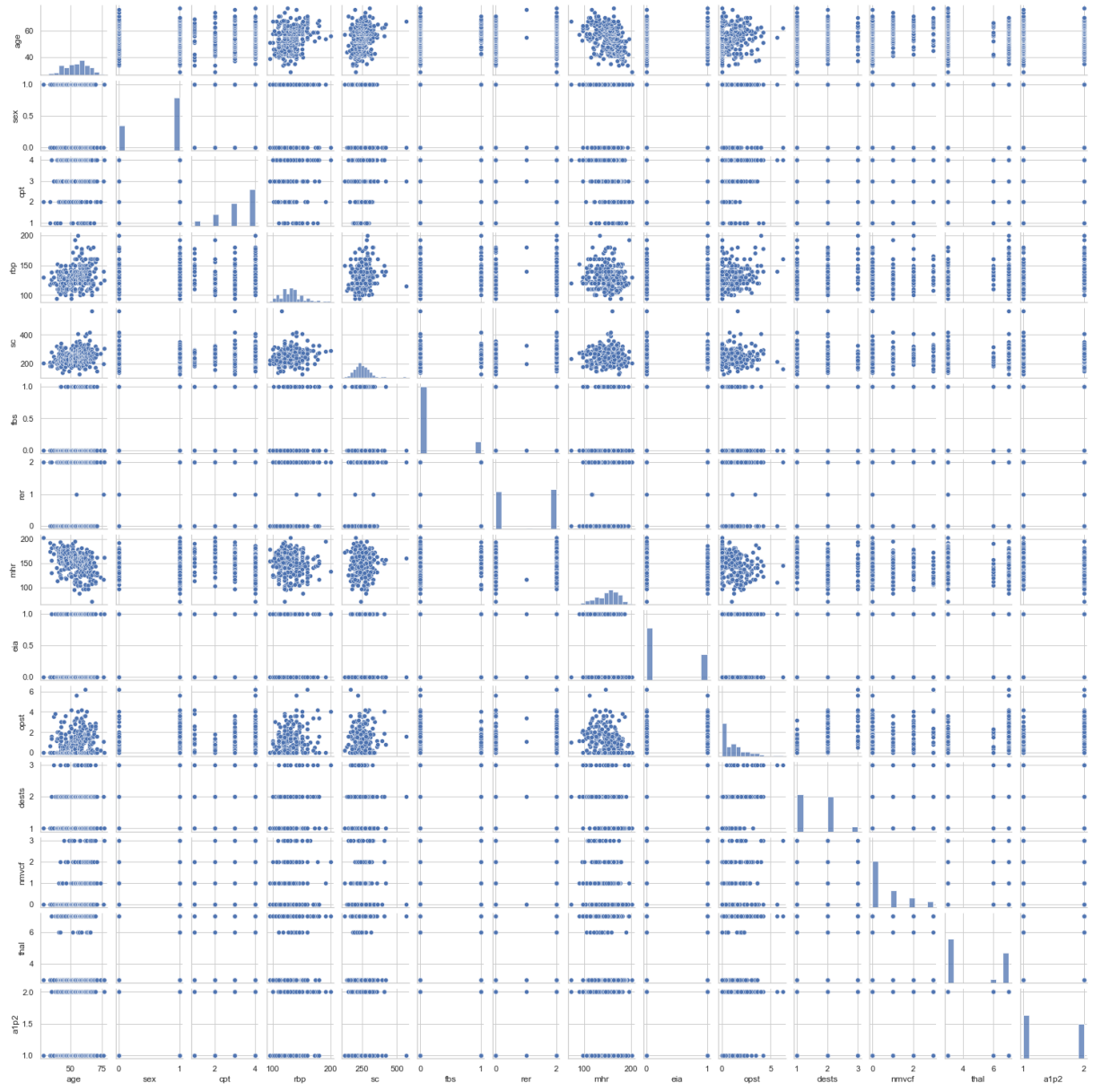

Table 4. Machine Learning Method Accuracy

Machine Learning Method	Test Accuracy (%)	Training Accuracy
Perceptron	84.0	86.2
Logistic Regression	85.2	87.8

Support Vector Machine (Linear)	84.0	86.8
Decision Tree Learning	75.3	76.7
Random Forest	80.2	98.9
K-Nearest Neighbors	82.7	87.8