Classification of Heart Disease

Introduction

According to the Centers for Disease Control and Prevention, approximately $\frac{1}{4}$ of all deaths in the US are caused by heart disease (source). Many risk factors for heart disease are well documented, including the influence of diet and lack of physical activity. These features could be used to build a model to predict the presence or probability of an individual being diagnosed with heart disease, leading to early treatment and an improved quality of life.

The Data

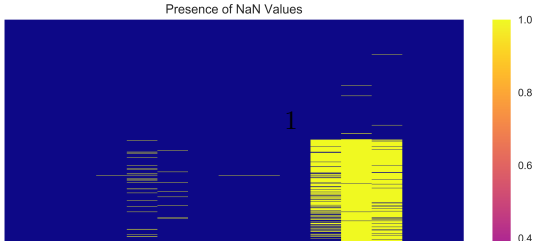
The processed heart disease data used in this report were downloaded from the University of California Irvine machine learning repository. The data sets for Cleveland, Hungary, Long Beach, and Switzerland, each contain a subset of 14 of the original 76 features, and are intended for classification. The outcome is a multiclass label ranging from 0 (no heart disease) to 4, and the values of 1,2,3, and 4 reflect the presence of heart disease.

EDA

Feature	Feature Description	Missing Values Count	Percentage of Missing Values in Feature
age	Age	0	0.00
sex	Sex	0	0.00
cp	Chest Pain Type	0	0.00
trestbps	Resting Blood Pressure	59	6.41
chol	Cholestoral	30	3.26
fbs	Fasting Blood Sugar > 120mg/dl?	90	9.78
restecg	Resting ECG Results	2	0.22
thalach	Maximum Heart Rate	55	5.98
exang	Did exercise induce the heart attack?	55	5.98
oldpeak	ST depression	62	6.74
label	Diagnosis of Heart Disease	0	0.00
Location	Location of Data Collection	0	0.00

feature_selection	model	error_test	train_err	val_err	test_err
Select From Model - LogisticRegression	LinearSVC	Logistic Regression	0.6373391	0.5850000	0.5810811
LinearSVC	LinearSVC	LinearSVC	0.6030043	0.5650000	0.5675676
RFE LinearSVC	KNN	KNN	0.5964126	0.6090909	0.5135135

model	train_err	val_err	test_err
LinearSVC	0.5858369	0.585	0.1351351
KNN	0.6545064	0.680	0.4864865



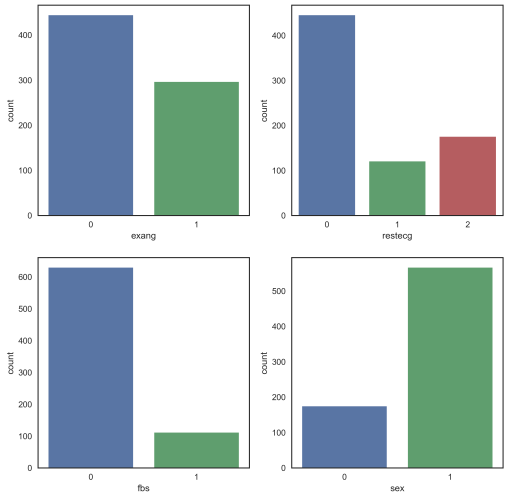


Figure 2:

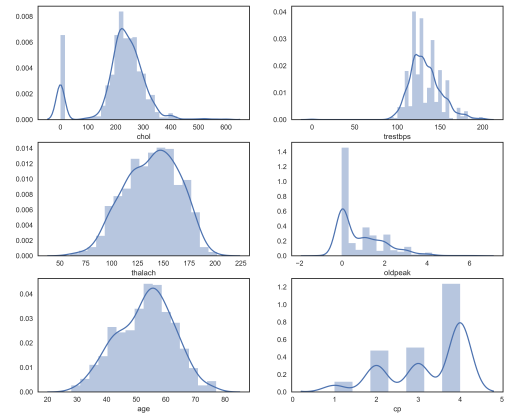


Figure 3: