PROJECT REPORT

Programming Machine Learning Applications (DSC478)

Sanjana Gowda Hetthur Chandrashekar (2114504)

CONTENTS

0. Dataset Description

Dataset: Heart Disease Dataset

This dataset consists of 14 different attributes that contribute to the detection of heart disease. The target variable is 'target' i.e., the presence of heart disease not otherwise. The columns:

age: Patient's age
sex: Patient's gender
cp: Patient's Constrictive pericarditis
trestbps: Patient's resting blood pressure
chol: Patient's cholesterol levels
fbs: Patient's fasting blood sugar
restecg: Patient's resting ECG levels
thalach: Patient's thalach levels
exang : Patient's Exang levels
oldpeak: Patient's old peak history recorded
slope: Patient's slope levels
ca: Patient's Calcium levels
thal: Patient's thal levels
target: If the patients is healthy or has a heart disease (0 – Healthy, 1 – Heart disease)

1. Data Exploration

The dataset consists of 14 features and 303 observations. It has 13 integer variables and 1 float variable. There were no missing variables found in the dataset and there are at least 2 unique value counts in each variables.

2. Data Visualization and Analysis

We obverse the cholesterol levels as the age goes up and sort of see the cholesterol levels rising as the age goes up with a few outliers. Then, we see the visualizations of the categorical variables in our dataset. We see there are more males than females with rest ecg levels of 1 being the highest and 2 being the least. We can also see there more no exang patients than the ones with exang and same goes for fbs of patients as well. CP levels of 0 are the most observed and 3 are the least. We already know that are more male patients than female patients but out of the total female patients we observe more heart disease positive cases. However, out of the total male patients we less heart disease positive patients but it is still more than that of female positive patients. In the next visualization, we can observe a crosstab between age and target variables. This tells us that people aged 51 have the highest positive heart disease followed by people aged 41 and 59. Whereas, people aged 77 shocking have no positive heart disease observed.

3. Models

After splitting the dataset into train and test datasets on both dependent and independent variables and performing minmax normalization on them,

3.1 K-Nearest Neighbor

We start with our first model KNN, which helps us find nearest neighbors. Here, we start with first instance for Euclidian distance and k value 5. We also check the distance of the neighbors to our first test instance. Next, we continue with the classifier function that calls the search function and returns the majority class among the k-nearest neighbors of the instance to be classified. Here, we will run the classifier on each test instance in our test data and compared the predicted classes to the actual class. Following which, is our evaluation function which can be reused with different parameters of KNN classifier.

Using all these function we have tested for a wide range of K using both Euclidian distance and Cosine similarity and also graphed it to compare them.

3.2 Decision Tree Classifier

Decision Tree Classifier with entropy as the criterion for splitting nodes and a minimum of three samples required to split a node. The model is trained on the normalized training data and is then used to predict labels for the normalized test data. We then evaluate its performance on both testing and training data. Here, the accuracy of training data indicates how well the model is able to correctly predict the target variable for the data it was trained on and accuracy on testing data is able to correctly predict the target variable for the unseen testing instances. The classification report then provides insights into how well the model performs along with visualization of the decision tree.

3.3 Regression

We start with basic linear regression on normalized training data and obtained it coefficients and plotted them. We then compare the performance of the model on the training data to its performance on 10-fold cross validation. The RMSE is low both on training and cross validation telling us that it is a well fitted model. However, the difference between them is not large but the training RMSE is slightly lower than the cross-validation RMSE which can be a possibility of overfitting.

We gave it a try on both Ridge and Lasso regression for a variety of alpha values and also predicting the model using their best alpha values but the RMSE of training data was higher than the RMSE of test data. This could have been worked on more and fixed if we had a little time longer.

3.4 KMeans Clustering

KMeans clustering is applied to the normalized combined dataset with specified number of clusters i.e., 3. The model is trained to identify clusters in the data and is used to predict the cluster assignments for the instances in the training data and the testing data. These predicted cluster assignments were further used in measuring the quality of clustering by computing the silhouette values for each instance in the data.

4. Conclusion

This project aims to present a comprehensive plan for predictive modeling of heart disease using the Kaggle dataset. By employing various classifier and prediction learning techniques we were able to detect the heart disease presence.