Machine Learning Project

Prediction of the presence of Heart Disease using KNN Algorithm and Support Vector Classifier Technique

Course: Machine Learning

Course Code: UE17EC337

ECE Department

PES University, Bangalore.

Name: Pindi Yashwanth

Department: Electrical and Electronics Engineering.

Class: 6th Semester B Section.

SRN: PES1201700505

Date: 18th April, 2020.

Index

8. References in IEEE format
9. Appendix: Code and sample Data

Abstract

Heart disease is a commonly occurring disease and is the major cause of sudden death nowadays. This disease attacks the person instantly. Most people are not aware of the symptoms of heart disease. Timely attention and proper diagnosis of heart disease will reduce the mortality rate. **Medical data mining** is to explore hidden patterns from the data sets. Supervised algorithms are used for the early prediction of heart disease in machine learning. For this purpose, in this project I have made use of 2 techniques for the prediction of the presence of heart disease in a person based on the sample data of that person:

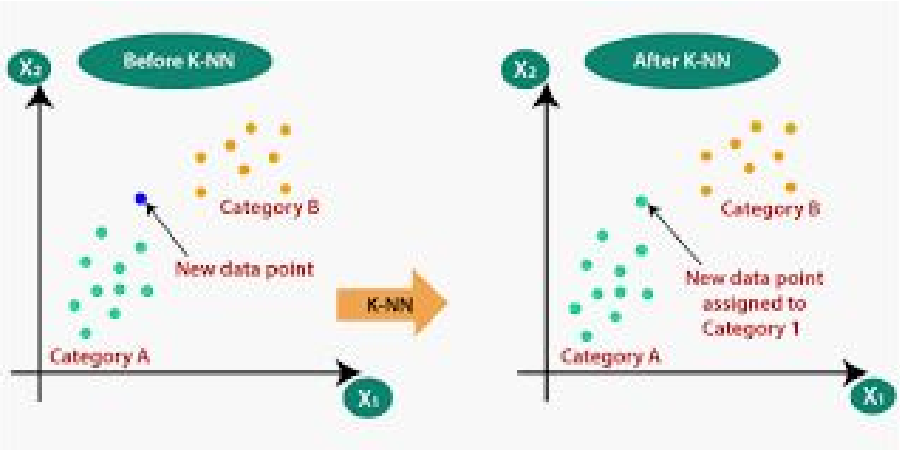
1. K Nearest Neighbor Algorithm (KNN Method)
2. Support Vector Classifier Technique.

Medical data sets contain a large number of features. The Performance of the classifier will be reduced if the data sets contain noisy features. Feature subset selection is proposed to solve this problem. Feature selection will improve accuracy and reduces the running time. For this purpose, we do feature selection. There are a number of feature selection techniques widely known.

Introduction

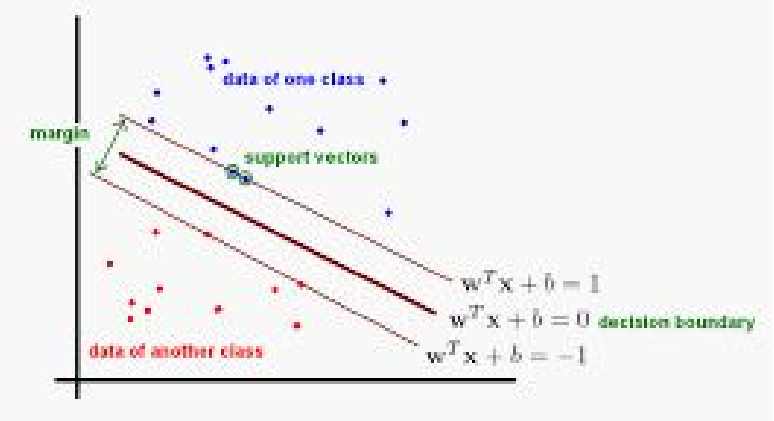
Medical data mining is used to infer diagnostic rules and help physicians to make the diagnosis process more accurate. **K-nearest neighbor** is the most widely used lazy classification algorithm as it reduces misclassification error. K-Nearest neighbor (KNN) is a simple, lazy and nonparametric classifier. KNN is preferred when all the features are

continuous. KNN is also called case-based reasoning and has been used in many applications like pattern recognition, statistical estimation. Classification is obtained by identifying the nearest neighbor to determine the class of an unknown sample. KNN is preferred over other classification algorithms due to its high convergence speed and simplicity.



The objective of the support vector machine algorithm is to find a **hyperplane** in an N-dimensional space (N — the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e. the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Support Vector Classifier/Machine is another simple algorithm that every machine learning expert should have in his/her arsenal. Support Vector Classifier is highly preferred by many as it produces significant accuracy with less computation power. Support Vector Machine (or SVM) can be used for both regression and classification tasks. But, it is widely used in classification objectives. It is used in both linear regression and logistic regression.



Problem Statement and Objective

Coronary Heart Disease (CHD) is obstruction of the coronary arteries with symptoms such as angina, chest pain, and heart attacks. Arteries supply blood to the heart muscle. CHD is a leading cause of death in many countries. In India there are roughly 3 crore heart patients and 2 lakh open heart surgeries are performed every year. CHD is a leading cause of mortality claiming nearly 17.3 million people every year. The reason for this is smoking, high levels of cholesterol, diabetes. Early prediction of heart disease is essential to reduce the mortality rate. **Data mining** provides a user-oriented approach to extract novel and uncovered patterns in the data set. Data mining is to extract useful knowledge within medical data for medical diagnosis.

In this project, I used Machine Learning to predict whether a person is suffering from a heart disease. After importing the data, I analyzed it using plots. Then, I generated dummy variables for categorical features and scaled other features. I then applied four Machine Learning algorithms, K Neighbors Classifier algorithm and Support Vector Classifier Technique.. I varied parameters across each model to improve their scores. In the end, K Neighbors Classifier achieved the highest

score of 87% with 8 nearest neighbors over the support vector classifier which had a score prediction of 81% with linear kernel.

Methodology, Tools and Data set used

In this machine learning project, I have collected the dataset from Kaggle (a website) and I will be using Machine Learning to make predictions on whether a person is suffering from Heart Disease or not. This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. The 'goal field' refers to the presence of heart disease in the patient. It is from a scale of 0 to 4.

Dataset used: https://www.kaggle.com/ronitf/heart-disease-uci

Implementation

Firstly, the dataset has been stored in the file dataset.csv in my Google Drive. I have linked my Google Drive to the **Google Collaborate Website** which I have made use of to implement my code. Next, we import all the necessary libraries. I'll use **numpy** and **pandas** to start with. For visualization, I will use pyplot subpackage of **matplotlib**, use rcParams to add styling to the plots and rainbow for colors.

For implementing Machine Learning models and processing of data, I will use the **sklearn** library. For processing the data, I'll import a few libraries. To split the available dataset for testing and training, I'll use the train_test_split method. To scale the features, I am using **StandardScalar**.

Now that we have all the libraries we will need, I can import the dataset and take a look at it. I'll use the pandas read_csv method to read the dataset. The dataset is now loaded into the variable dataset. I'll just take

a glimpse of the data using the describe() and info() methods before I actually start processing and visualizing it.

The scale of each feature column is different and quite varied as well. While the maximum for age reaches 77, the maximum of chol (serum cholesterol) is 564. Now, we can use visualizations to better understand our data and then look at any processing we might want to do.

From the code, (the link has been provided in references), taking a look at the correlation matrix, it's easy to see that a few features have negative correlation with the target value while some have positive.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
age         303 non-null int64
sex         303 non-null int64
cp          303 non-null int64
trestbps    303 non-null int64
chol        303 non-null int64
fbs         303 non-null int64
restecg     303 non-null int64
thalach     303 non-null int64
exang       303 non-null int64
oldpeak     303 non-null float64
slope       303 non-null int64
ca          303 non-null int64
thal        303 non-null int64
target      303 non-null int64
dtypes: float64(1), int64(13)
memory usage: 33.2 KB
```

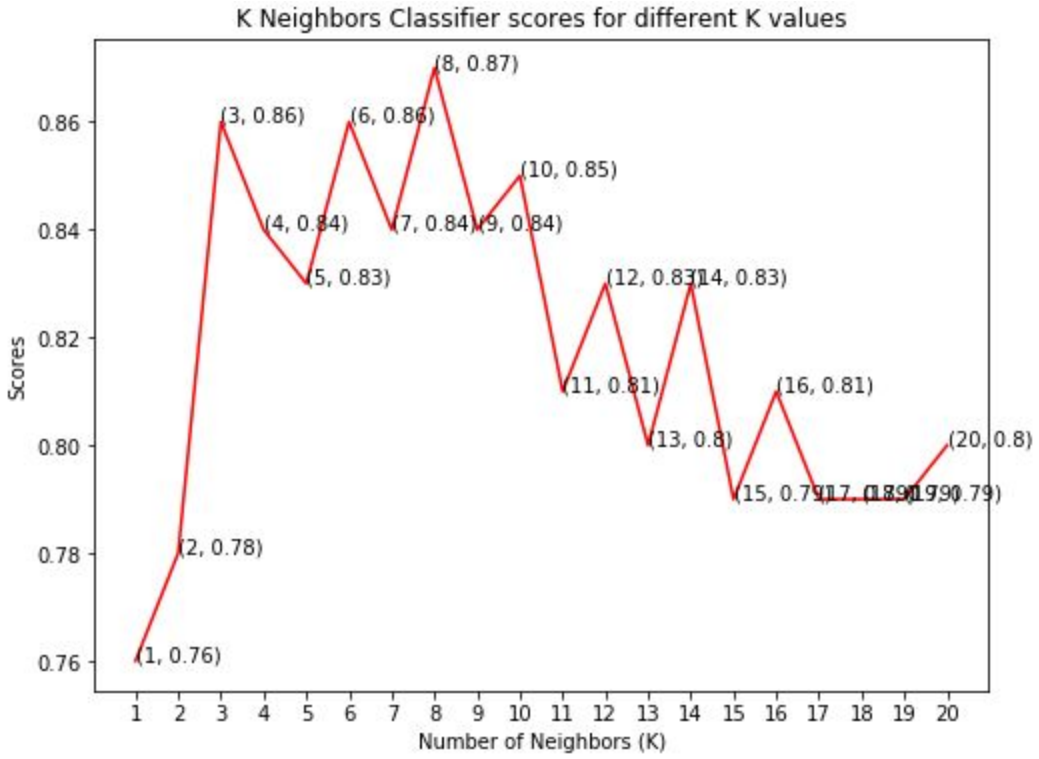
The dataset has a total of 303 rows and there are no missing values. There are a total of 13 features along with one target value which we wish to find.

Next, by taking a look at the **histograms** for each variable, we can see that each feature has a different range of distribution. Thus, using scaling before our predictions should be of great use. Also, the categorical features do stand out. Our target classes are both of approximately equal sizes. I'll now import train_test_split to split our dataset into training and

testing datasets. Then, I'll import all Machine Learning models I'll be using to train and test the data.

1. **K Nearest Neighbour**

The classification score varies based on different values of neighbors that we choose. Thus, I'll plot a score graph for different values of K (neighbors) and check when I will achieve the best score. I have the scores for different neighbor values in the array knn_scores. I'll now plot it and see for which value of K I get the best scores.

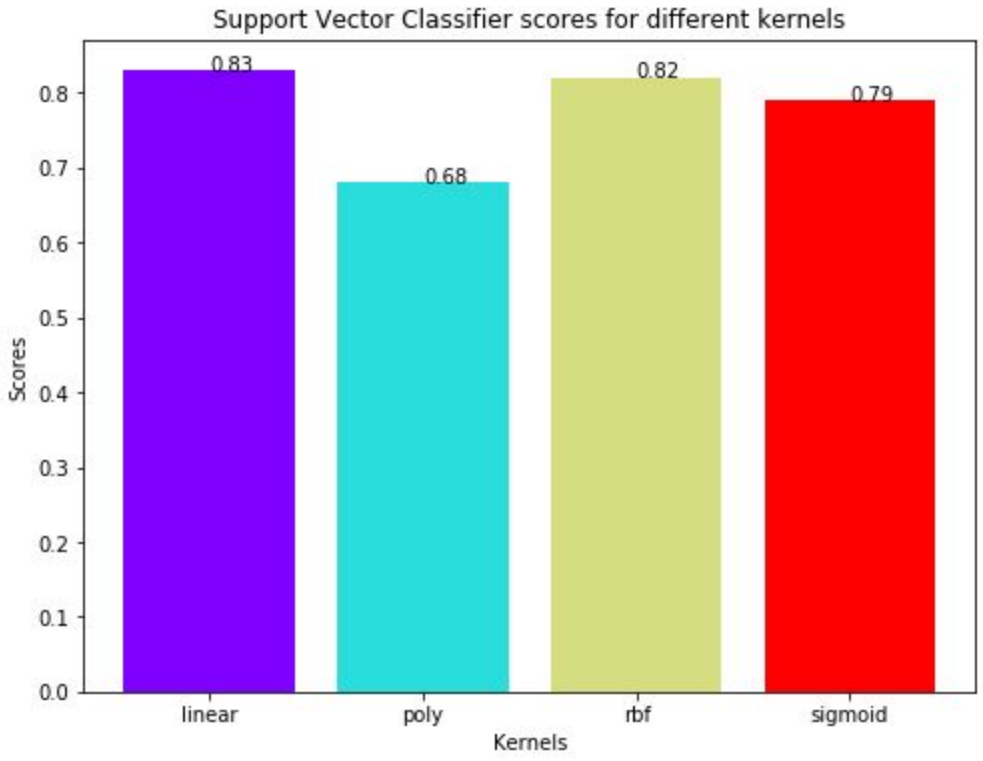


The maximum score achieved was 0.87 for the 8 neighbors

2. Support Vector Classifier

There are several kernels for Support Vector Classifier. I'll test some of them and check which has the best score. I'll now plot a bar plot of

scores for each kernel and see which performed the best. The score for Support Vector Classifier is 83.0% with linear kernel.



Results and Analysis

Therefore, from the above analysis we have predicted the possibility of the presence of heart disease among individuals using their datasets by the methods of K nearest neighbour algorithm (KNN Algorithm) technique and the Support Vector Classifier Technique. We have found out the different predictions from these two models and have found out that in this particular dataset for heart disease prediction, the KNN approach gives us a higher accuracy of 87% for 8 neighbours compared to the support vector classifier technique which gives us an accuracy of 81%.

Conclusion and Future Scope

In this project, I used Machine Learning to predict whether a person is suffering from a heart disease. After importing the data, I analyzed it using plots. Then, I generated dummy variables for categorical features and scaled other features. I then applied four Machine Learning algorithms of K Neighbors Classifier and Support Vector Classifier. I varied parameters across each model to improve their scores. In the end, K Neighbors Classifier achieved the highest score of 87% with 8 nearest neighbors.

The results suggest that the proposed approach can significantly improve the learning accuracy. This model helps the physicians in an efficient prediction of diseases with predominant features. The scope of this particular application is immense. In future, it would be possible to integrate and ensemble classifiers by reducing the number of parameters and variables for better feature selection. For this purpose, algorithms such as **Particle Swarm Optimization** (PSO) can be used to improve the classification performance to develop a decision support system for early diagnosis of heart disease.

References

[1] Code:

https://colab.research.google.com/drive/1PJUIp9GpmFlMA6EdNRoCYw_RjqlvxunQ

[2]

https://www.alliedacademies.org/articles/prediction-of-heart-disease-using-knearest-neighbor-and-particle-swarm-optimization.html

[3] https://www.kaggle.com/ronitf/heart-disease-uci

[4] https://github.com/kb22/Heart-Disease-Prediction/blob/master/Heart%20Disease%20Prediction.ipynb

[5] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[6] Bishop - Pattern Recognition And Machine Learning - Springer 2006

[7] Machine-Learning-Tom-Mitchell

Appendix: Code and sample Data

Medical data mining, Heart disease, KNN, Feature selection, Particle swarm optimization,

Code: https://colab.research.google.com/drive/1PJUIp9GpmFlMA6EdNRoCYw_RjqlvxunQ

Sample data: https://www.kaggle.com/ronitf/heart-disease-uci

Thank You!