



Spring 2021-2022

Project Report

Project Name : Tic-Tac-Toe Games

Programming Language : Python

Course: Programming In Python

Section : A

Faculty: Akinul Islam Jony

Group Members:

|  |  |  |
| --- | --- | --- |
| Name | Id | Dept |
| Shah, Faysal Mohammed | 19-39367-1 | CSE |
| Chowdhury, Shah Fahim | 17-33640-1 | CSE |
| Israt Jahan | 18-38280-2 | CSE |
| Anindya Datta | 18-37634-1 | CSE |
|  |  |  |

**Heart Disease Detection Project Report**

**I. INTRODUCTION:**

Our problem is that we want to predict whether patients have heart disease by given some features of users. This is important to medical fields. If such a prediction is accurate enough, we can not only avoid wrong diagnosis but also save human resources. When a patient without a heart disease is diagnosed with heart disease, he will fall into unnecessary panic and when a patient with heart disease is not diagnosed with heart disease, he will miss the best chance to cure his disease. Such wrong diagnosis is painful to both patients and hospitals. With accurate predictions, we can solve the unnecessary trouble. Besides, if we can apply our machine learning tool into medical prediction, we will save human resource because we do not need complicated diagnosis process in hospitals. (though it is a very long way to go.) The input to our algorithm is 13 features with number values. We use several algorithms such as Logistic Regression, SVM, Naïve Bayes, Random Forest, Artificial Neural Network to output a binary number 1 or 0. 1 indicates the patient has heart disease and vice versa.

**II. RELATED WORK :**

Before we did the experiments, we did research on how people explored heart disease prediction so that we can broaden our horizons and learn from them. In 2011, Ujma Ansari [1] made use of Decision Tree model to predict heart disease and get a high accuracy of 99%, which inspires us to use a better version of Decision Tree and it is Random Forest. Unfortunately, the paper uses a dataset with 3000 instances but dose not provide a reference of how they get the data. The UCI website only provides 303 instances of dataset so we doubt where the author gets 3000 instances of dataset. In 2012, Chaitrali S. Dangare [2] made the prediction by using three models and such models are Naïve Bayes, Decision Trees and Neural Network. We are using the same dataset as he did. The difference between his work and ours is that he added 2 more features into the dataset, which means there are 15 features of his work while there are 13 features in our dataset. Though there is no big difference between 13 features and 15 features in his work, what he did on dataset inspires us to make useful change to our dataset (Try normalization on dataset) to

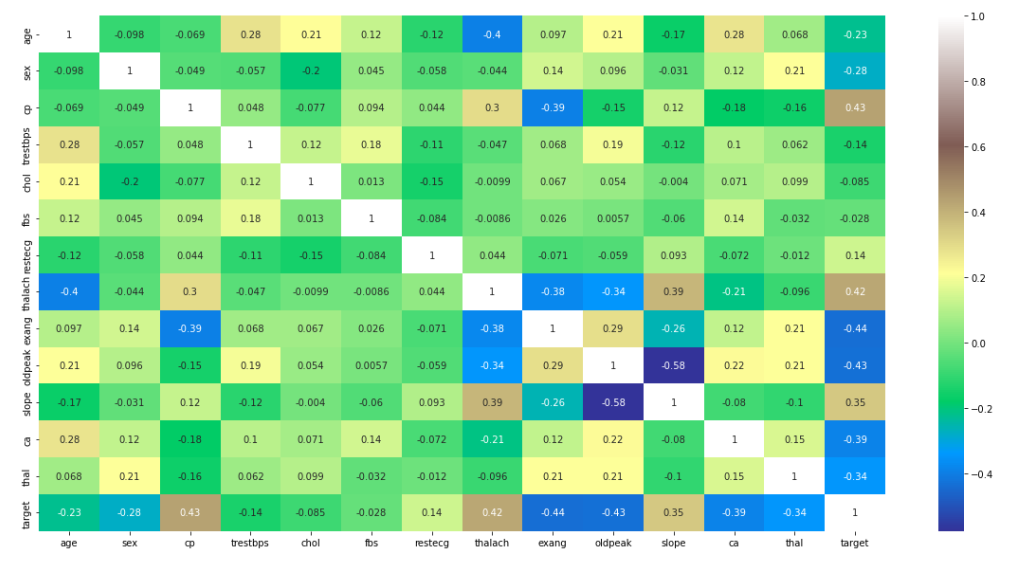
overfitting. What we did to our dataset is to change 1s to 0s in target column and vice versa in order to make value 1 indicate the presence of heart disease and make value 0 indicate the absence of heart disease. Given such dataset we can do many interesting predicative tasks. For example, we can use these features to predict chest pain type. But the most important thing is that given the 13 attributes from a patient, we want to predict whether he has the heart disease or not because keeping healthy is very import to people. make our results comprehensive. However, during this paper there are only 3 models. More models need to be considered so that the results are comprehensive. In 2017, Kaan Uyar and Ahmet İlhan[3] did the same experiment and used the same dataset as we did for projects. During their analysis, “Class distributions are interpreted as 54% absence and 46% presence of a heart disease”. The dataset we download from Kaggle has 54% 1s and 46% 0s in the target column. From their analysis, we realize 1 indicates absence of heart disease and vice versa. To make it easily understood, we switched 1s and 0s in the target column so that 1 indicates presence of heart disease to show our confusion matrix[10] in our results. After reviewing paper [4] and [5], we have learned that neural network has advantage of fault tolerance and it has the ability to work with inadequate knowledge as human beings. Therefore, in our project we decide to spend some time working on neural network to detect heart diseases.

**III. DATASET AND FEATURES :**

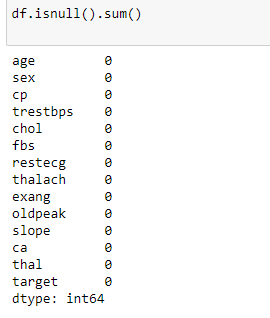
Our dataset is based on Kaggle heart Disease Data Set and we have 303 instances. According to Kaggle, “This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them.”We guess too many features will bring too much noise so people has done feature extraction and reduce 76 features to 14 features. To better understand the meaning of the features, we have the responsibility to explain some of the attributes of original dataset from UCI as follows: • age: age in years • sex: sex (1 = male; 0 = female) • cp: chest pain type -- Value 0: typical angina -- Value 1: atypical angina -- Value 2: non-anginal pain -- Value 3: asymptomatic • trestbps: resting blood pressure (in mm Hg on admission to the hospital) • chol: serum cholestoral in mg/dl • fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false) • target: Heart disease (0 = no, 1 = yes) Since the original dataset has missing values, we just downloaded a clean dataset from Kaggle. We have split the dataset into 80% (242 instances) for training and 20%(61 instances) for test. We did normalization on our dataset to avoid

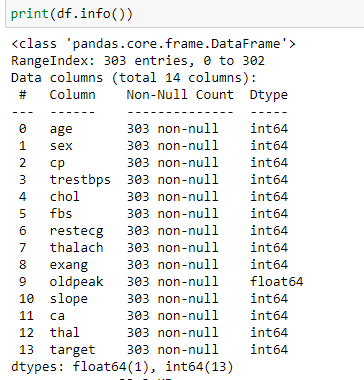
overfitting. What we did to our dataset is to change 1s to 0s in target column and vice versa in order to make value 1 indicate the presence of heart disease and make value 0 indicate the absence of heart disease. Given such dataset we can do many interesting predicative tasks. For example, we can use these features to predict chest pain type. But the most important thing is that given the 13 attributes from a patient, we want to predict whether he has the heart disease or not because keeping healthy is very import to people.

**IV. Data Preprocessing , Data Visualization :**





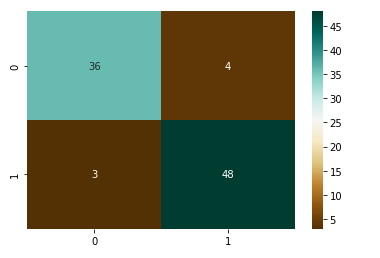




**V. EDA (EXPLOSITORY DATA ANAYSIS) :**

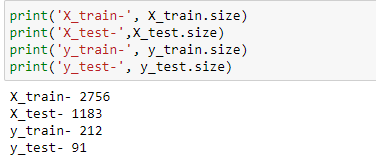
Before building any model, it is essential to understand the data set. EDA (exploratory data analysis) is a crucial step in building any successful model. The dataset used for this model is straightforward. We are trying to predict the weight of the individual based on their height. So let us first see if there is any correlation between height and weight of the individual.

This Project EDA is Given Below:

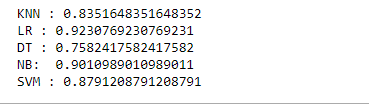


**VI. Model Development:**

We Used Logistic Regression , Naive Bayes , K Nearest Neighbours (KNN), Decision Tree , Support Vector Machine (SVM) model in this entire project.



Accuracy Test of this project in (LR,NB,KNN,DT,SVM) :



Best Accuracy: Logistic Regression : 92

Same Accurancy NB and Decision tree : 90

**VII. CONCLUSION :**

We use some libraries provided by Python to implement this project. After the experiments, the algorithm of Logistic Regression us the best test accuracy, which is 92%. The reason why it outperforms others is that it is not limited to the property of the dataset. Naive Bayes requires the features to be mutually independent. Logistic Regress requires the features to be linearly separable. SVM requires the parameters to be appropriately set. Though we get a good result of 92% accuracy, that is not enough because it cannot guarantee that no wrong diagnosis happens. To improve accuracy, we hope to require more dataset because 300 instances of dataset are not sufficient to do an excellent job. In the future, to predict disease we want to try different diseases such as lung cancer by using image detection. In this way, the dataset becomes complicated and we can apply convolutional neural network to make accuracy predictions.