Heart Disease Identification Tool

using K Nearest Neighbours classifier

Submitted to

Dr. Mohammad Shoyaib

Professor

IIT, University of Dhaka

and

Kishan Kumar Ganguly

Lecturer

IIT, University of Dhaka

Submitted by

Yasin Sazid (BSSE **1006**, Exam Roll: **1027**)

Date of Submission: 5 September, 2021

Table of Contents

1. Introduction

Heart diseases are one of the leading causes of death in the world. Different signs and symptoms are often attributed to this deadly medical condition. However, a correct diagnosis is very difficult. That has prompted researchers to gather data and study medical conditions of patients deeply to make better predictions about the presence of heart disease [1]. In this project, we built a Machine Learning based model using the K Nearest Neighbours (KNN) classifier to predict the risk of heart disease. We have used a dataset for this purpose that has been widely used by researchers working in this field. The dataset dates from 1988 and is recognized as a benchmark dataset in the field of heart disease prediction [1].

2. Scope of the Project

The scope of this project is strictly limited to using K Nearest Neighbours (KNN) algorithm as the one and only classification technique to predict heart disease. We aspired to build the best possible machine learning model using KNN, but did not consider any other classification algorithm.

3. Objectives

The main objective of this project was to build a fully working tool for heart disease identification. Another objective was to improve the machine learning model used in this project to have a higher accuracy in heart disease prediction.

4. Methodology

The machine learning model used in this heart disease identification tool was carefully built using dataset preprocessing, machine learning classifier and evaluation process.

4.1 Description of the Dataset

The dataset used for this project originally contained 76 attributes, but all published experiments refer to using a subset of 14 of them. The "Num" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Other experiments with the dataset used for this project have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0) [2]. We also built our model to predict the presence or absence of heart disease.

The dataset is properly explained below -

Number of Instances: 300

Number of Attributes: 14

Attribute Information:
1. Age: Age in years
2. Sex: Sex (1 = Male; 0 = Female)
3. Cp: Chest pain type
    Value 1: Typical Angina
    Value 2: Atypical Angina
    Value 3: Non Anginal pain
    Value 4: Asymptomatic
4. Trestbps: Resting blood pressure (in mmHg)
5. Chol: Serum cholesterol in mg/dl
6. Fbs: (Fasting blood sugar > 120 mg/dl) (1 = True; 0 = False)
7. Restecg: Resting electrocardiographic results
    Value 0: Normal
    Value 1: Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    Value 2: Showing probable or definite left ventricular hypertrophy by Estes' criteria
8. Thalach: Maximum heart rate achieved
9. Exang: Exercise induced angina (1 = Yes; 0 = No)
10. Oldpeak: ST depression induced by exercise relative to rest
11. Slope: The slope of the peak exercise ST segment
    Value 1: Upsloping
    Value 2: Flat
    Value 3: Downsloping
12. Ca: Number of major vessels (0-3) colored by fluoroscopy
13. Thal : 3 = Normal; 6 = Fixed Defect; 7 = Reversible Defect
14. Num: Angiographic disease status (No Disease = 0 and Disease = 1,2,3,4)

Missing Attribute Values: 3, distinguished with value -1

Class Distribution:

0	1	2	3	4	Total
161	55	36	35	13	300

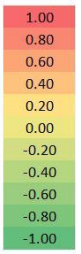
4.2 Preprocessing of the Dataset

At first, we did not do any preprocessing of the dataset and directly applied the data to the machine learning algorithm, and the results were not promising. Then we tried preprocessing the data in order to have better results. We used feature selection and scaling on the dataset and this time the results were quite promising.

4.2.1 Feature Selection

For feature selection, we used two techniques. At first, we created a correlation heatmap -

	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Num
Age	1.00	-0.10	0.11	0.30	0.19	0.12	0.15	-0.39	0.09	0.20	0.16	0.37	0.12	0.22
Sex	-0.10	1.00	0.01	-0.06	-0.18	0.05	0.03	-0.04	0.15	0.12	0.03	0.08	0.37	0.23
Cp	0.11	0.01	1.00	-0.04	0.08	-0.04	0.06	-0.33	0.38	0.18	0.15	0.23	0.26	0.41
Trestbps	0.30	-0.06	-0.04	1.00	0.15	0.16	0.13	-0.05	0.06	0.17	0.11	0.08	0.11	0.14
Chol	0.19	-0.18	0.08	0.15	1.00	0.02	0.16	-0.02	0.08	0.06	-0.02	0.14	0.00	0.09
Fbs	0.12	0.05	-0.04	0.16	0.02	1.00	0.07	0.00	0.02	0.02	0.06	0.13	0.05	0.06
Restecg	0.15	0.03	0.06	0.13	0.16	0.07	1.00	-0.09	0.08	0.12	0.14	0.13	0.03	0.18
Thalach	-0.39	-0.04	-0.33	-0.05	-0.02	0.00	-0.09	1.00	-0.40	-0.36	-0.39	-0.27	-0.28	-0.41
Exang	0.09	0.15	0.38	0.06	0.08	0.02	0.08	-0.40	1.00	0.31	0.26	0.14	0.31	0.39
Oldpeak	0.20	0.12	0.18	0.17	0.06	0.02	0.12	-0.36	0.31	1.00	0.53	0.27	0.32	0.47
Slope	0.16	0.03	0.15	0.11	-0.02	0.06	0.14	-0.39	0.26	0.53	1.00	0.11	0.27	0.38
Ca	0.37	0.08	0.23	0.08	0.14	0.13	0.13	-0.27	0.14	0.27	0.11	1.00	0.25	0.52
Thal	0.12	0.37	0.26	0.11	0.00	0.05	0.03	-0.28	0.31	0.32	0.27	0.25	1.00	0.50
Num	0.22	0.23	0.41	0.14	0.09	0.06	0.18	-0.41	0.39	0.47	0.38	0.52	0.50	1.00



Here, we see that the correlation coefficients of the features “Trestbps”, “Chol” and “Fbs” with respect to the target attribute “Num” are 0.14, 0.09 and 0.06 respectively. This means that they have very little impact on the value of target attribute “Num”. So we removed these features from the dataset. Later, we used forward elimination on the remaining 10 features. We found that the best performing subset is the one where we discarded the features “Sex”, “Thalach” and “Oldpeak”. So at the end, the dataset was left with the features “Age”, “Cp”, “Restecg”, “Exang”, “Slope”, “Ca” and “Thal”.

4.2.2 Data Scaling

Scaling is a technique of constraining the values of all the independent attributes of our dataset within the same scale. Without scaling, the machine learning model might give higher weightage to higher values and lower weightage to lower values. We used normalization to scale our dataset. Normalization is a scaling technique where data values are rescaled between 0 to 1 [3]. The equation to normalize a data value x is -

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

4.3 Machine Learning Classifier Used

K Nearest Neighbours (KNN) classifier was used to build the machine learning model for this project. Different variations of KNN were experimented with to have better results. We tried basic KNN and also weighted KNN. No machine learning library was used to implement the classifier used in this project.

4.4 Evaluation Process Used

We used confusion matrix, accuracy score, precision, recall and F1 score to evaluate the quality of the machine learning model.

		Predicted value: P	Predicted value: N
True value	P	TP	FN
True value	N	FP	TN

Here, P = positive, N = negative, TP = true positive, FP = false positive, FN = false negative and TN = true negative.

We used an accuracy score to evaluate how well the model is performing. The equation for accuracy is -

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The equations used to calculate precision, recall and F1 score are -

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = \frac{2*Precision*Recall}{Precision + Recall}$$

5. Analysis of Results

Some variations of KNN were tried out and we found that the basic KNN shows better results than weighted ones. Also K=5 was chosen to be an optimum value for better results. Then three approaches were used to see what differences come when the KNN classifier is applied on the data. At first, we used the original dataset directly for classification. In the second approach, we used feature selection and got rid of unimportant features. The results in this approach were better than the first one. In the third approach, we used feature selection and we also scaled the dataset using normalization. This yielded the best results among all three approaches.

5.1 Without Using Feature Selection and Scaling

The average accuracy, precision, recall and F1 score when the dataset was used directly -

Metric	Value
Accuracy	64.5%
Precision	62%
Recall	59.5%
F1 Score	61%

5.2 Using Feature Selection and No Scaling

The average accuracy, precision, recall and F1 score after feature selection -

Metric	Value
Accuracy	76.5%
Precision	74%
Recall	76%
F1 Score	75%

5.3 Using Feature Selection and Scaling

The average accuracy, precision, recall and F1 after feature selection and data scaling -

Metric	Value
Accuracy	83%
Precision	83%
Recall	79.5%
F1 Score	81%

6. Conclusion

In this project, we intended to create an usable tool for heart disease identification. We used various techniques to improve the accuracy of the machine learning model used in this project. We used three different approaches and compared the results based on metrics such as accuracy, precision, recall and F1 score. We used the approach that yielded the best results in our final identification tool. It has an accuracy of 83% which is very promising compared to the other approaches.

7. References

1. Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning, https://www.hindawi.com/journals/cin/2021/8387680/

2. Heart Disease Data Set, https://archive.ics.uci.edu/ml/datasets/heart+disease

3. Feature Scaling In Machine Learning! | by SagarDhandare | Jul, 2021, https://medium.datadriveninvestor.com/feature-scaling-in-data-science-5b1e82492727