Heart Disease Prediction System

A COURSE PROJECT REPORT

Submitted by

SHIVAM PAHARIYA(RA2011027010007)

PRATYUSH VATS(RA2011027010018)

SHIVANSH SHARMA(RA2011027010039)

Under the guidance of

Dr. A.Shanthini

In partial fulfilment for the Course

of

Data science (18CSE396T)

in

DEPARTMENT OF DATA SCIENCE AND BUSINESS SYSTEMS

SCHOOL OF COMPUTING

COLLEGE OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Deemed to be University u/s 3 of UGC Act, 1956)

KATTANKULATHUR - 603 203

November, 2022

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Under Section 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that this mini project titled **HEART DISEASE PREDICTION MODEL** is the bonafide work of **SHIVAM PAHARIYA(RA2011027010007), PRATYUSH VATS(RA2011027010018), SHIVANSH SHARMA(RA2011027010039)** who carried out the project work under my supervision.

SUPERVISOR

Dr. A.Shanthini

Associate Professor

Department of Data Science and Business Systems

SRM Institute of Science and Technology

Kattankulathur – 603 203

HEAD OF THE DEPARTMENT

Dr. M. LAKSHMI

Professor & Head

Department of Data Science and Business Systems

SRM Institute of Science and Technology

Kattankulathur – 603 203

ABSTRACT

Day by day the cases of heart diseases are increasing at a rapid rate and it's very Important and concerning to predict any such diseases beforehand. This diagnosis is a difficult task i.e. it should be performed precisely and efficiently. The research paper mainly focuses on which patient is more likely to have a heart disease based on various medical attributes. We prepared a heart disease prediction system to predict whether the patient is likely to be diagnosed with a heart disease or not using the medical history of the patient. We used different algorithms of machine learning such as logistic regression and KNN to predict and classify the patient with heart disease. A quite Helpful approach was used to regulate how the model can be used to improve the accuracy of prediction of Heart Attack in any individual. The strength of the proposed model was quiet satisfying and was able to predict evidence of having a heart disease in a particular individual by using KNN and Logistic Regression which showed a good accuracy in comparison to the previously used classifier such as naive bayes etc. So a quiet significant amount of pressure has been lift off by using the given model in finding the probability of the classifier to correctly and accurately identify the heart disease. The Given heart disease prediction system enhances medical care and reduces the cost. This project gives us significant knowledge that can help us predict the patients with heart disease It is implemented on the .pynb format.

ACKNOWLEDGEMENT

We express our heartfelt thanks to our honorable **Vice Chancellor Dr. C. MUTHAMIZHCHELVAN**, for being the beacon in all our endeavors.

We would like to express my warmth of gratitude to our **Registrar Dr. S. Ponnusamy,** for his encouragement

We express our profound gratitude to our **Dean (College of Engineering and Technology) Dr. T. V.Gopal,** for bringing out novelty in all executions.

We would like to express my heartfelt thanks to Chairperson, School of Computing **Dr. Revathi Venkataraman,** for imparting confidence to complete my course project

We wish to express my sincere thanks to **Course Audit Professor** and **Course Coordinator** for their constant encouragement and support.

We are highly thankful to my Course project Faculty **Dr. A.Shanthini, Associate Professor, Department of Data Science and Business Systems,** for his assistance, timely suggestion and guidance throughout the duration of this course project.

We extend our gratitude to our **HoD, Dr. M. Lakshmi, Professor, Department of Data Science and Business Systems,** and my Departmental colleagues for their Support.

Finally, we thank our parents and friends near and dear ones who directly and indirectly contributed to the successful completion of our project. Above all, I thank the almighty for showering his blessings on me to complete my Course project.

TABLE OF CONTENTS

1. INTRODUCTION

"Machine Learning is a way of Manipulating and extraction of implicit, previously unknown/known and potential useful information about data" . Machine Learning is a very vast and diverse field and its scope and implementation is increasing day by day. Machine learning Incorporates various classifiers of Supervised, Unsupervised and Ensemble Learning which are used to predict and Find the Accuracy of the given dataset. We can use that knowledge in our project of HDPS as it will help a lot of people.

Cardiovascular diseases are very common these days, they describe a range of conditions that could affect your heart. World health organization estimates that 17.9 million global deaths from (Cardiovascular diseases) CVDs . It is the primary reason of deaths in adults. Our project can help predict the people who are likely to diagnose with a heart disease by help of their medical history. It recognizes who all are having any symptoms of heart disease such as chest pain or high blood pressure and can help in diagnosing disease with less medical tests and effective treatments, so that they can be cured accordingly. This project focuses on mainly two data mining techniques namely: (1) Logistic regression, (2) KNN . The accuracy of our project is 80.48% for which is better than previous system where only one data mining technique is used. So, using more data mining techniques increased the HDPS accuracy and efficiency. Logistic regression falls under the category of supervised learning. Only discrete values are used in logistic regression. The objective of this project is to check whether the patient is likely to be diagnosed with any cardiovascular heart diseases based on their medical attributes such as gender, age, chest pain, fasting sugar level, etc. A dataset is selected from the UCI repository with patient's medical history and attributes. By using this dataset, we predict whether the patient can have a heart disease or not. To predict this, we use 14 medical attributes of a patient and classify him if the patient is likely to have a heart disease. These medical attributes are trained under two algorithms: Logistic regression, KNN . Most efficient of these algorithms is Logistic Regression which gives us the accuracy of 80.48%. And, finally we classify patients that are at risk of getting a heart disease or not and also this method is totally cost efficient.

2. LITERATURE SURVEY

2.1 Heart Disease Prediction

Author: Nayab Akhtar(Fatima Jinnah Women University)
Published on Research Gate

Heart disease is the major cause of deaths worldwide. To give treatment for heart disease, a lot of advanced technologies are used. In medical center it is the most common problem because many medical persons do not have equal knowledge and expertise to treat their patient so they deduce their own decision and as a result it shows poor outcome and sometimes lead to death. To overcome these problems, prediction of heart disease is being done by using machine learning algorithms and data mining techniques, it has become easy to perform automatic diagnosis in hospitals as they are playing vital role in this regard. Heart disease can be predicted by performing analysis on patient's different health parameters. There are different algorithm to predict heart disease like naïve Bayes, k Nearest Neighbor (KNN), Decision tree, Artificial Neural Network (ANN). We have used different parameters to predict heart disease. Those parameters are Age, Gender, Cerebral palsey (CP), Gender, Cerebral palsey (CP), Blood Pressure (bp), Fasting blood sugar test (fbs) etc. In our research paper, we have used built in dataset. we have implemented the five different techniques with same dataset to predict heart disease These implemented algorithm are Naive Bayes, k Nearest Neighbor (KNN), Decision tree, Artificial Neural Network (ANN), Random Forest. This paper investigates that which technique gives more accuracy in predicting heart disease based on health parameters. Experiment shows that Naïve Bayes has the highest accuracy of 88%.

2.2 A novel approach for heart disease prediction using strength score with significant predictors

Author: Armin Yazdani, Kasturi Dewi Varathan, Yin Kia Chiam, Asad Malik & Wan Azman Wan Ahmad
Published on BMC Medical Informatics and Decision Making

Cardiovascular disease is the leading cause of death in many countries. Physicians often diagnose cardiovascular disease based on current clinical tests and previous experience of diagnosing patients with similar symptoms. Patients who suffer from heart disease require quick diagnosis, early treatment and constant observations. To address their needs, many data mining approaches have been used in the past in diagnosing and predicting heart diseases. Previous research was also focused on identifying the significant contributing features to heart disease prediction, however, less importance was given to identifying the strength of these features

2.3 An Introduction to Logistic Regression Analysis and Reporting

Author: Joanne Peng (National Taiwan University)
Published on Research Gate

The purpose of this article is to provide researchers, editors, and readers with a set of guidelines for what to expect in an article using logistic regression techniques. Tables, figures, and charts that should be included to comprehensively assess the results and assumptions to be verified are discussed. This article demonstrates the preferred pattern for the application of logistic methods with an illustration of logistic regression applied to a data set in testing a research hypothesis. Recommendations are also offered for appropriate reporting formats of logistic regression results and the minimum observation-to-predictor ratio. The authors evaluated the use and interpretation of logistic regression presented in 8 articles published in The Journal of Educational Research between 1990 and 2000. They found that all 8 studies met or exceeded recommended criteria.

2.4 KNN Model-Based Approach in Classification

Author: Gongde Guo (Fuijan Normal University)
Published on ResearchGate

The k-Nearest-Neighbors (kNN) is a simple but effective method for classification. The major drawbacks with respect to kNN are its low efficiency - being a lazy learning method prohibits it in many applications such as dynamic web mining for a large repository, and its dependency on the selection of a "good value" for k . In this paper, we propose a novel kNN type method for classification that is aimed at overcoming these shortcomings. Our method constructs a kNN model for the data, which replaces the data to serve as the basis of classification. The value of k is automatically determined, is varied for different data, and is optimal in terms of classification accuracy. The construction of the model reduces the dependency on k and makes classification faster. Experiments were carried out on some public datasets collected from the UCI machine learning repository in order to test our method.

3. REQUIREMENTS ANALYSIS

- **Numpy**

  NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It is open-source software.

- **Pandas**

  Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.

- **Sklearn**

  Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering, and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

- **Sklearn.metrics**

  The sklearn. metrics module implements several loss, score, and utility functions to measure classification performance. Some metrics might require probability estimates of the positive class, confidence values, or binary decisions values

- **Sklearn.model_selection**

  Split arrays or matrices into random train and test subsets.Quick utility that wraps input validation and next(ShuffleSplit().split(X, y)) and application to input data into a single call for splitting (and optionally subsampling) data in a oneliner.

- **Pickle**

  Python pickle module is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk. What pickle does is that it “serializes” the object first before writing it to file. Pickling is a way to convert a python object (list, dict, etc.) into a character stream. The idea is that this character stream contains all the information necessary to reconstruct the object in another python script.

- **Streamlit**

  Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

4. DATA SET DESCRIPTION

The dataset used in this project is the Cleveland Heart Disease dataset taken from the UCI repository.

```
]: heart_data.head()
```

]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

```
]: heart_data.tail()
```

]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

The dataset consists of 1024 individuals data. There are 14 columns in the dataset, which are described below.

1. ***Age***: displays the age of the individual.

2. ***Sex***: displays the gender of the individual using the following format :
   1 = male
   0 = female

3. ***Chest-pain type***: displays the type of chest-pain experienced by the individual using the following format :
   1 = typical angina
   2 = atypical angina
   3 = non — anginal pain
   4 = asymptotic

4. ***Resting Blood Pressure***: displays the resting blood pressure value of an individual in mmHg (unit)

5. ***Serum Cholestrol***: displays the serum cholesterol in mg/dl (unit)

6. ***Fasting Blood Sugar***: compares the fasting blood sugar value of an individual with 120mg/dl.
   If fasting blood sugar > 120mg/dl then : 1 (true)
   else : 0 (false)

7. ***Resting ECG*** : displays resting electrocardiographic results
   0 = normal
   1 = having ST-T wave abnormality
   2 = left ventricular hyperthrophy

8. ***Max heart rate achieved*** : displays the max heart rate achieved by an individual.

9. ***Exercise induced angina*** :
   1 = yes
   0 = no

10. ***ST depression induced by exercise relative to rest***: displays the value which is an integer or float.

11. ***Peak exercise ST segment*** :

    1 = upsloping

    2 = flat

    3 = downsloping

12. ***Number of major vessels (0–3) colored by flourosopy*** : displays the value as integer or float.

13. ***Thal*** : displays the thalassemia :

    3 = normal

    6 = fixed defect

    7 = reversible defect

14. ***Diagnosis of heart disease*** : Displays whether the individual is suffering from heart disease or not :

    0 = absence

    1, 2, 3, 4 = present.

In the actual dataset, we had 76 features but for our study, we chose only the above 14 because :

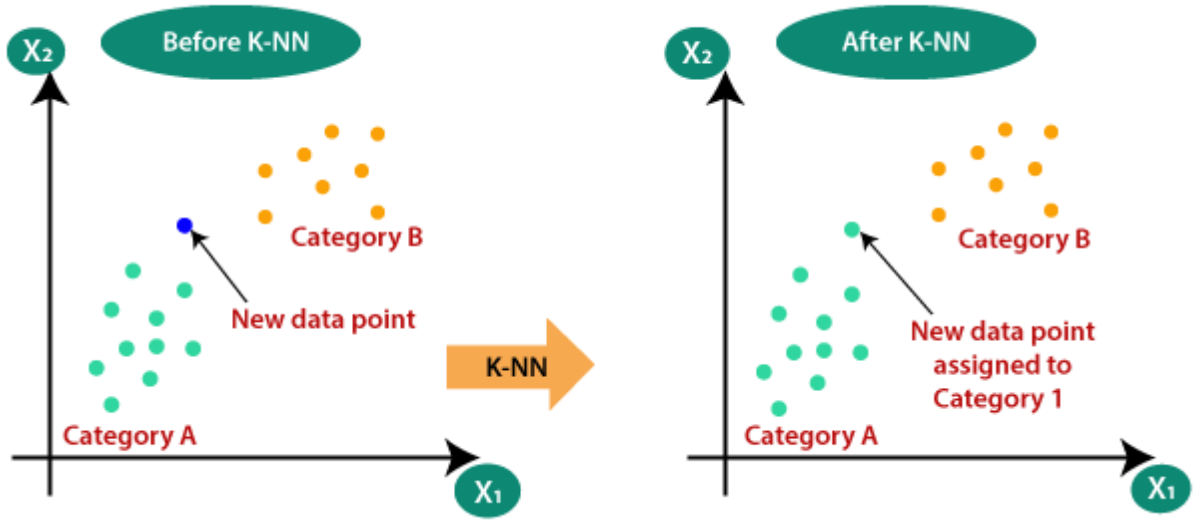
1. **Age**: Age is the most important risk factor in developing cardiovascular or heart diseases, with approximately a tripling of risk with each decade of life. Coronary fatty streaks can begin to form in adolescence. It is estimated that 82 percent of people who die of coronary heart disease are 65 and older. Simultaneously, the risk of stroke doubles every decade after age 55.
2. **Sex**: Men are at greater risk of heart disease than pre-menopausal women. Once past menopause, it has been argued that a woman's risk is similar to a man's although more recent data from the WHO and UN disputes this. If a female has diabetes, she is more likely to develop heart disease than a male with diabetes.
3. **Angina (Chest Pain)**: Angina is chest pain or discomfort caused when your heart muscle doesn't get enough oxygen-rich blood. It may feel like pressure or squeezing in your chest. The discomfort also can occur in your shoulders, arms, neck, jaw, or back. Angina pain may even feel like indigestion.
4. **Resting Blood Pressure**: Over time, high blood pressure can damage arteries that feed your heart. High blood pressure that occurs with other conditions, such as obesity, high cholesterol or diabetes, increases your risk even more.

5. **Serum Cholesterol**: A high level of low-density lipoprotein (LDL) cholesterol (the "bad" cholesterol) is most likely to narrow arteries. A high level of triglycerides, a type of blood fat related to your diet, also ups your risk of a heart attack. However, a high level of high-density lipoprotein (HDL) cholesterol (the "good" cholesterol) lowers your risk of a heart attack.
6. **Fasting Blood Sugar**: Not producing enough of a hormone secreted by your pancreas (insulin) or not responding to insulin properly causes your body's blood sugar levels to rise, increasing your risk of a heart attack.
7. **Resting ECG**: For people at low risk of cardiovascular disease, the USPSTF concludes with moderate certainty that the potential harms of screening with resting or exercise ECG equal or exceed the potential benefits. For people at intermediate to high risk, current evidence is insufficient to assess the balance of benefits and harms of screening.
8. **Max heart rate achieved**: The increase in cardiovascular risk, associated with the acceleration of heart rate, was comparable to the increase in risk observed with high blood pressure. It has been shown that an increase in heart rate by 10 beats per minute was associated with an increase in the risk of cardiac death by at least 20%, and this increase in the risk is similar to the one observed with an increase in systolic blood pressure by 10 mm Hg.
9. **Exercise induced angina**: The pain or discomfort associated with angina usually feels tight, gripping or squeezing, and can vary from mild to severe. Angina is usually felt in the center of your chest but may spread to either or both of your shoulders, or your back, neck, jaw or arm. It can even be felt in your hands. o Types of Angina a. Stable Angina / Angina Pectoris b. Unstable Angina c. Variant (Prinzmetal) Angina d. Microvascular Angina.
10. **Peak exercise ST segment**: A treadmill ECG stress test is considered abnormal when there is a horizontal or down-sloping ST-segment depression ≥ 1 mm at 60–80 ms after the J point. Exercise ECGs with up-sloping ST-segment depressions are typically reported as an 'equivocal' test. In general, the occurrence of horizontal or down-sloping ST-segment depression at a lower workload (calculated in METs) or heart rate indicates a worse prognosis and higher likelihood of multi-vessel disease. The duration of ST-segment depression is also important, as prolonged recovery after peak stress is consistent with a positive treadmill ECG stress test. Another finding that is highly indicative of significant CAD is the occurrence of ST-segment elevation > 1 mm (often suggesting transmural ischemia); these patients are frequently referred urgently for coronary angiography.

5. METHOD/ALGORITHM USED

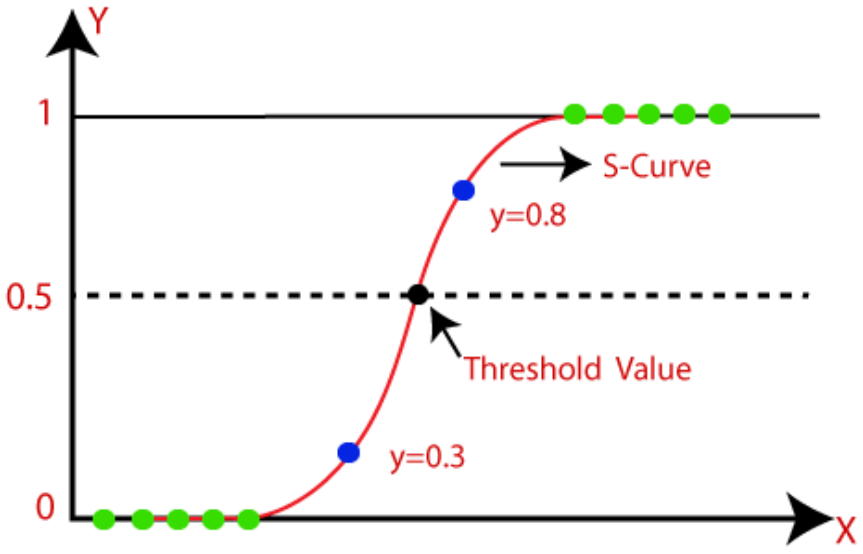
K-Nearest Neighbor

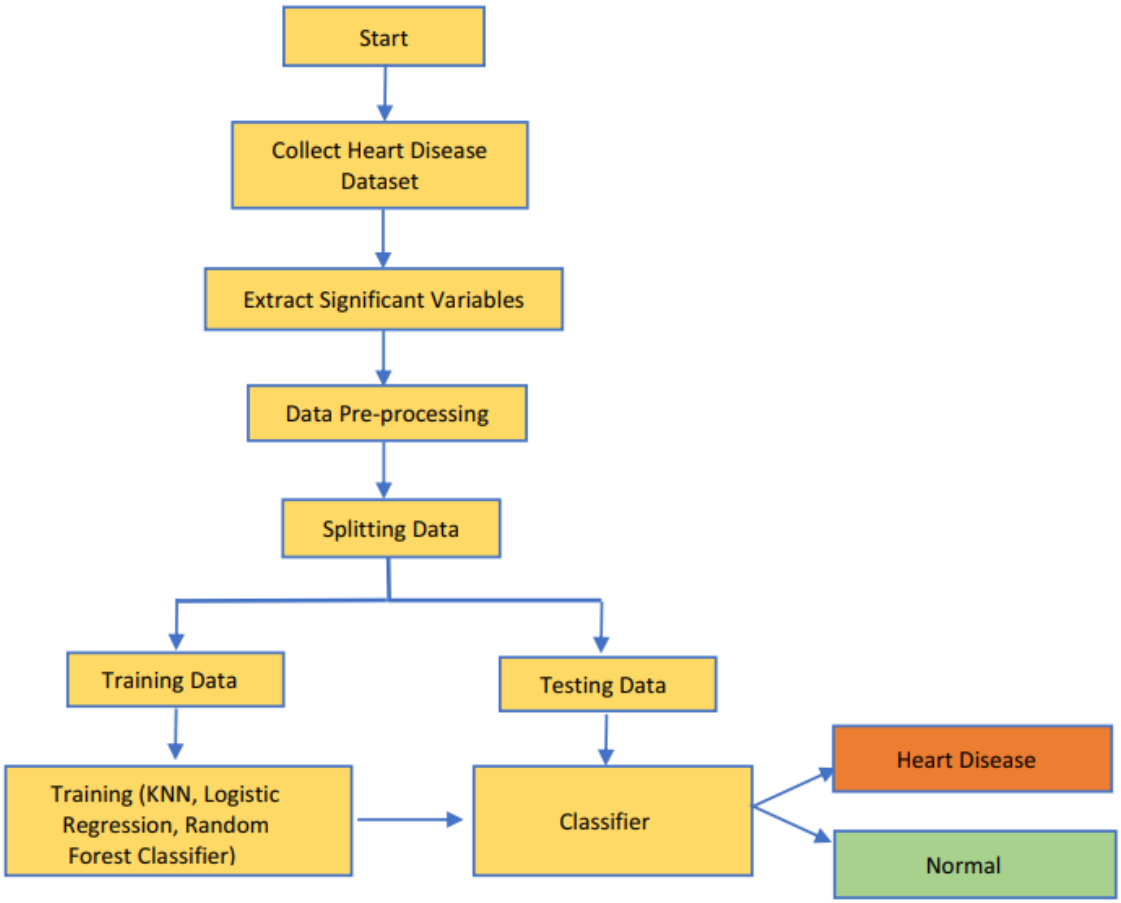
- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.



Logistic Regression

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.




Start
Collect Heart Disease Dataset
Extract Significant Variables
Data Pre-processing
Splitting Data
Training Data
Testing Data
Training (KNN, Logistic Regression, Random Forest Classifier)
Classifier
Heart Disease
Normal

6. RESULTS AND DISCUSSION

6.1 Accuracy of the model

```
# accuracy on training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

```
print('Accuracy on Training data : ', training_data_accuracy)
```

```
Accuracy on Training data :  0.8524390243902439
```

```
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

```
print('Accuracy on Test data : ', test_data_accuracy)
```

```
Accuracy on Test data :  0.8048780487804879
```

6.2 Testing the model on Google Collab

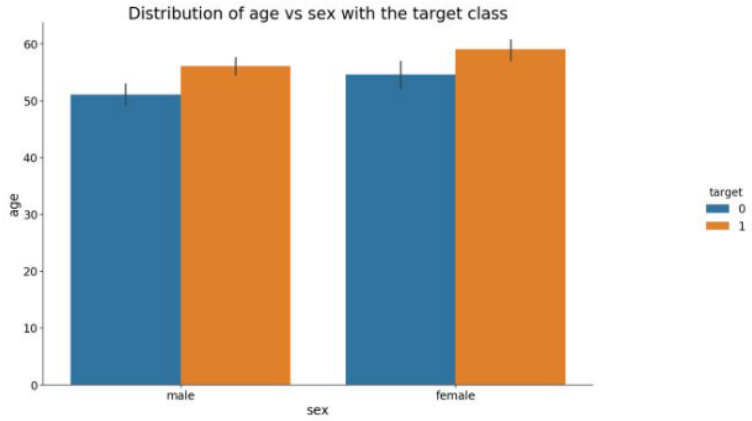
```
input_data = (62,0,0,140,268,0,0,160,0,3.6,0,2,2)

# change the input data to a numpy array
input_data_as_numpy_array= np.asarray(input_data)

# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0]== 0):
  print('The Person does not have a Heart Disease')
else:
  print('The Person has Heart Disease')
```

```
[0]
The Person does not have a Heart Disease
```

Here, target = 1 implies that the person is suffering from heart disease and target = 0 implies the person is not suffering.

We see that most people who are suffering are of the age of 58, followed by 57.

Majorly, people belonging to the age group 50+ are suffering from the disease.

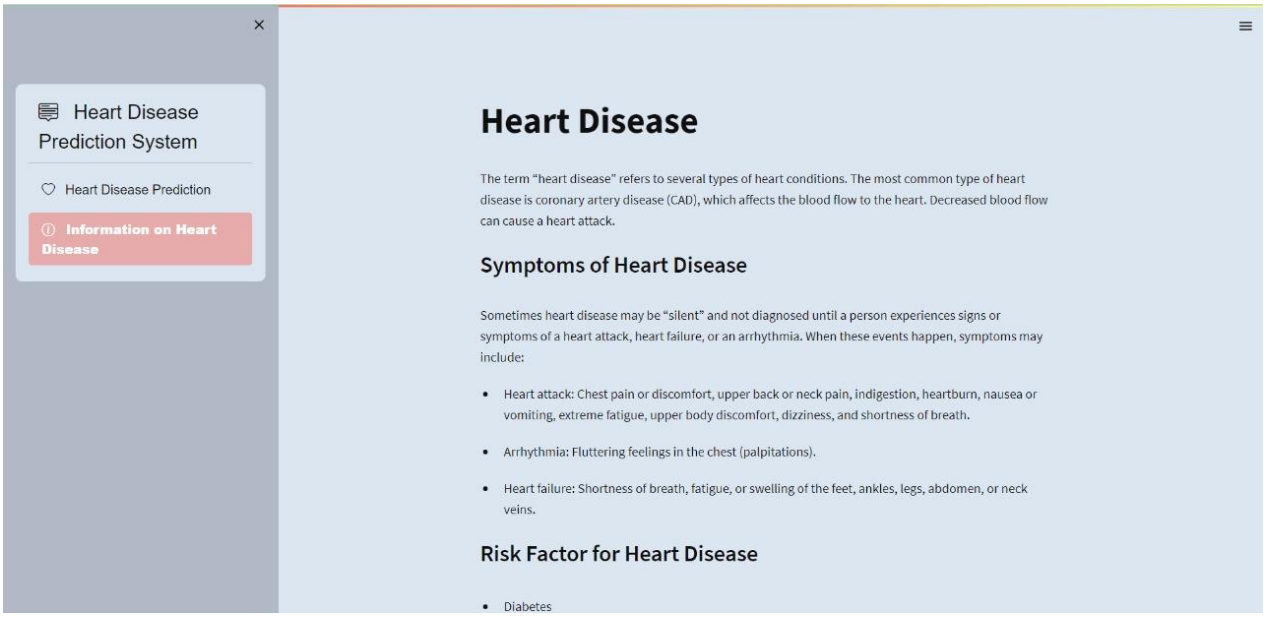


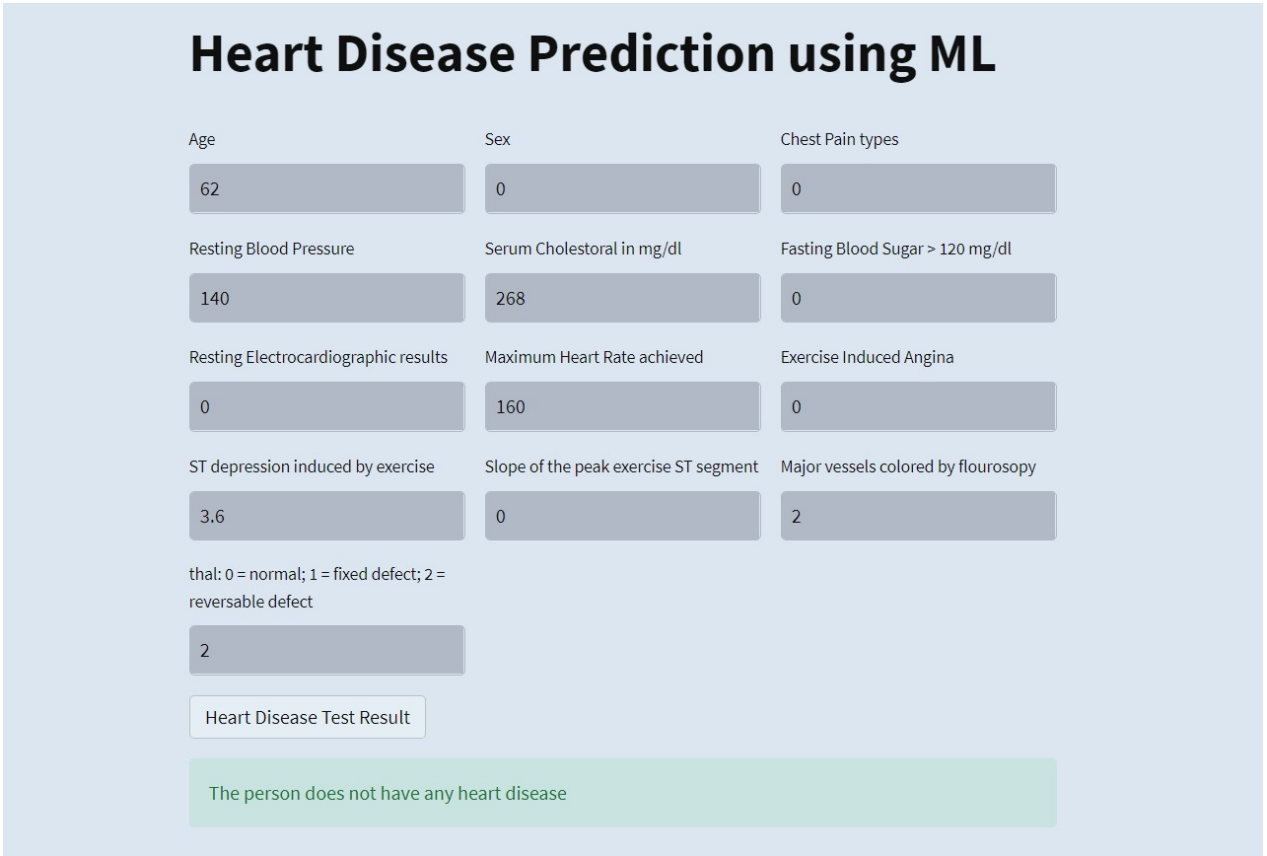
In the second graph, we can see that females who are suffering from the disease are older than males.

From these results, we noticed that. our accuracy has improved due to the increased medical attributes that we used from the dataset we took. As we analyzed the results given through both algorithms, we were able to conclude that the results given while using logistic regression was more accurate than the results given using KNN. The maximum accuracy obtained by logistic regression is equal to **80.4%**.

We were able to conclude that Logistic Regression outperforms KNN. This proves that Logistic Regression is better in diagnosis of a heart disease.

6.3 Using Pickle and Streamlit package of python we created the web-app for the model.





7. CONCLUSION AND FUTURE ENHANCEMENT

Heart Disease Prediction System was developed using two ML classification modelling techniques. This project predicts people with cardiovascular disease by extracting the patient medical history that leads to a fatal heart disease from a dataset that includes patients' medical history such as chest pain, sugar level, blood pressure, etc. This Heart Disease detection system assists a patient based on his/her clinical information of them been diagnosed with a previous heart disease. The algorithms used in building the given model are Logistic regression and KNN

There are a couple of things that we wish to add to the project. That way, the model will provide better predictions and be much easier to use.

Future Enhancements:

- We wish to add more parameters in the data set. That way, we can get more accurate results and better predictions.
- We will use more effective classification algorithms to give better performance and accurate results.
- Use of more training data ensures the higher chances of the model to accurately predict whether the given person has a heart disease or not.

8.REFERENCES

- Numpy documentation: https://numpy.org/doc/stable/
- Pandas documentation: https://pandas.pydata.org/
- Sklearn documentation: https://scikit-learn.org/stable/
- Stream lit: https://streamlit.io/
- Pickle: https://docs.python.org/3/library/pickle.html
- Heart Disease Prediction: https://www.researchgate.net
- A novel approach for heart disease prediction using strength score with significant predictors: https://bmcmedinformdecismak.biomedcentral.com