HEART DISEASE PREDICTION

AIML PROJECT REPORT

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

By

ABHINANDANA POLEPALLY	**(2103A52064)**
VARSHA PEDDI	**(2103A52166)**
SAI KRISHNA EDAMA	**(2103A52014)**
PAVAN KODURU	**(2103A52146)**

Under the guidance of

Mr. Dr. E.L.N. Kiran Kumar

Associate Professor, CS & AI.

SR ENGINEERING COLLEGE

Ananthasagar, Warangal.

CERTIFICATE

This is to certify that this project entitled **"Heart Disease Prediction"** is the bonafied work carried out by Abhinandana Polepally, Varsha Peddi, Sai Krishna Edama,Pavan Koduru as a AIML PROJECT in **Bachelor of Technology in Computer Science & Engineering** during the academic year 2022-2023 under our guidance and Supervision.

Mr.E.L.N. Kiran Kumar
Assoc. Prof. CS & AI
HOD(CSE),
S R Engineering College,
Ananthasagar, Warangal.

Dr.M.Sheshikala
Assoc. Prof. &
S R Engineering College,
Ananthasagar, Warangal.

External Examiner

ACKNOWLEDGEMENT

We owe an enormous debt of gratitude to our project guide **Mr. Dr. E.L.N. KiranKumar, Assoc. Prof. CS and AI** as well as Head of the CSE Department **Dr.M.Sheshikala, Associate Professor** for guiding us fromthe beginning through the end of the AIML project with their intellectual advices and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for theirsuggestions and timely support.

Table of Contents

1.INTRODUCTION

2.ABSTRACT

3.PROBLEM STATEMENT

4.DATASET

5.PROPOSED METHODOLOGIES

6.RESULTS AND DISCUSSION

7.CONCLUSION

INTRODUCTION:

Heart disease prediction is an AIML project that aims to use machine learning algorithms to predict the likelihood of a person developing heart disease based on various health parameters and lifestyle choices. Heart disease, also known as cardiovascular disease, is a leading cause of death worldwide. It encompasses a range of conditions that affect the heart and blood vessels, such as coronary artery disease, heart failure, and arrhythmias.

The project involves collecting and analyzing a large dataset of health parameters such as age, blood pressure, cholesterol levels, family history, smoking habits, and other relevant information. The dataset is used to train a machine learning model that can identify patterns and correlations between these parameters and the risk of heart disease.

Once the model is trained using KNN, SVM and Random Forest, it can be used to predict the likelihood of an individual developing heart disease based on their personal health parameters. This information can be used to identify individuals who may be at high risk and provide early interventions such as lifestyle changes, medication, and other treatments to reduce their risk of developing heart disease.

The heart disease prediction AIML project has the potential to save lives by identifying high-risk individuals early and providing personalized interventions to reduce their risk of heart disease. It is an important application of machine learning in healthcare and has the potential to significantly improve public health outcomes.

ABSTRACT

Heart disease prediction is a critical task that requires accurate and efficient machine learning models. In this project, we compared the performance of three popular classification algorithms, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest, for predicting heart disease.

We used a dataset consisting of 270 samples with 14 attributes, including age, sex, chest pain type, blood pressure, cholesterol levels, and others. We preprocessed the data by handling missing values, converting categorical variables into numerical ones, and normalizing the data.

Our experimental results show that Random Forest outperformed the other models, achieving an accuracy of 76%, precision of 77%, recall of 82%, and F1-score of 82.1%. KNN and SVM also achieved reasonable performance, with an accuracy of 76.2% and 76.8%, respectively.

Our study demonstrates that machine learning algorithms can effectively predict heart disease, and Random Forest is the most suitable algorithm for this task. The use of such models in healthcare can provide early identification of high-risk individuals and improve the quality of patient care.

PROBLEM STATEMENT

Heart disease is a leading cause of death worldwide, and early identification of high-risk individuals can significantly improve public health outcomes. Traditional methods of identifying individuals at risk of heart disease, such as risk factor assessment and clinical evaluation, may not be accurate or efficient. Therefore, there is a need for an accurate and efficient machine learning model to predict the likelihood of heart disease based on various health parameters and lifestyle choices.

The problem statement for this project is to develop a machine learning model that can accurately predict the likelihood of heart disease based on health parameters such as age, sex, blood pressure, cholesterol levels, family history, smoking habits, and other relevant information. The model should be able to identify individuals who may be at high risk of heart disease and provide early interventions to reduce their risk. The performance of the model should be evaluated using several metrics such as accuracy, precision. The goal is to develop a model that can accurately predict heart disease and improve public health outcomes by identifying high-risk individuals early and providing personalized interventions to reduce their risk of developing heart disease.

DATASET

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Age	Sex	Chest pair	BP	Cholester	FBS over 1	EKG result	Max HR	Exercise a	ST depress	Slope of S	Number o	Thallium	Heart Disease	
2	70	1	4	130	322	0	2	109	0	2.4	2	3	3	1	
3	67	0	3	115	564	0	2	160	0	1.6	2	0	7	0	
4	57	1	2	124	261	0	0	141	0	0.3	1	0	7	1	
5	64	1	4	128	263	0	0	105	1	0.2	2	1	7	0	
6	74	0	2	120	269	0	2	121	1	0.2	1	1	3	0	
7	65	1	4	120	177	0	0	140	0	0.4	1	0	7	0	
8	56	1	3	130	256	1	2	142	1	0.6	2	1	6	1	
9	59	1	4	110	239	0	2	142	1	1.2	2	1	7	1	
10	60	1	4	140	293	0	2	170	0	1.2	2	2	7	1	
11	63	0	4	150	407	0	2	154	0	4	2	3	7	1	
12	59	1	4	135	234	0	0	161	0	0.5	2	0	7	0	
13	53	1	4	142	226	0	2	111	1	0	1	0	7	0	
14	44	1	3	140	235	0	2	180	0	0	1	0	3	0	
15	61	1	1	134	234	0	0	145	0	2.6	2	2	3	1	
16	57	0	4	128	303	0	2	159	0	0	1	1	3	0	
17	71	0	4	112	149	0	0	125	0	1.6	2	0	3	0	
18	46	1	4	140	311	0	0	120	1	1.8	2	2	7	1	
19	53	1	4	140	203	1	2	155	1	3.1	3	0	7	1	
20	64	1	1	110	211	0	2	144	1	1.8	2	0	3	0	
21	40	1	1	140	199	0	0	178	1	1.4	1	0	7	0	
22	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1	
23	48	1	2	130	245	0	2	180	0	0.2	2	0	3	0	
24	43	1	4	115	303	0	0	181	0	1.2	2	0	3	0	
25	47	1	4	112	204	0	0	143	0	0.1	1	0	3	0	
26	54	0	2	132	288	1	2	159	1	0	1	1	3	0	
27	48	0	3	130	275	0	0	139	0	0.2	1	0	3	0	
28	46	0	4	138	243	0	2	152	1	0	2	0	3	0	
29	51	0	3	120	295	0	2	157	0	0.6	1	0	3	0	
30	58	1	3	112	230	0	2	165	0	2.5	2	1	7	1	

The dataset comprises of:

1.**Age**: the age of the patient in years.
2.**Sex:** the gender of the patient (1 = male, 0 = female).
3.**Chest Pain Type**: the type of chest pain experienced by the patient (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic).
4.**Resting Blood Pressure**: the resting blood pressure of the patient (in mm Hg).
5.**Cholesterol:** the serum cholesterol level of the patient (in mg/dl).
6.**Fasting Blood Sugar**: the fasting blood sugar level of the patient (>120 mg/dl = 1, <= 120 mg/dl = 0).
7.**Resting Electrocardiographic Results**: the resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy).
8.**Maximum Heart Rate Achieved**: the maximum heart rate achieved by the patient during exercise.
9.**Exercise Induced Angina:** whether the patient experienced angina during exercise (1 = yes, 0 = no).
10.**Oldpeak:** ST depression induced by exercise relative to rest.
11.Slope: the slope of the peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping).
12.**Number of Major Vessels:** the number of major vessels (0-3) colored by fluoroscopy.
13.**Thal**: a blood disorder called thalassemia (3 = normal, 6 = fixed defect, 7 = reversible defect).
14.**Target**: the presence of heart disease (1 = yes, 0 = no).

PROPOSED METHODOLOGY

Data collection: Collect the heart disease dataset, which consists of 14 attributes, including age, sex, chest pain type, blood pressure, cholesterol levels, and others.

Data preprocessing: Preprocess the dataset by handling missing values, converting categorical variables into numerical ones, and normalizing the data.

Feature selection: Select the relevant features that have a significant impact on the prediction of heart disease. This step is optional, as some machine learning algorithms can handle irrelevant features.

Model training: Train the KNN, SVM, and Random Forest models using the preprocessed data. Use a 5-fold cross-validation technique to ensure that the models are not overfitting the data.

Model evaluation: Evaluate the performance of the models using several metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Compare the performance of the models and choose the best one.

Model testing: Test the selected model on a separate test dataset to evaluate its performance in real-world scenarios.

Deployment: Deploy the model for heart disease prediction in healthcare systems, allowing early identification of high-risk individuals and providing personalized interventions to reduce their risk of developing heart disease.

CODE:

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import cufflinks as cf

d=pd.read_csv('/content/heartdp.csv')

print(d.dtypes)
```

```
Age                         int64
Sex                         int64
Chest pain                  int64
BP                          int64
Cholesterol                 int64
FBS over 120                int64
EKG results                 int64
Max HR                      int64
Exercise angina             int64
ST depression             float64
Slope of ST                 int64
Number of vessels fluro     int64
Thallium                    int64
Heart Disease               int64
dtype: object
```

```
print(d.isnull().any())

from sklearn.metrics import classification_report,confusion_matrix,accuracy_score

from sklearn.preprocessing import StandardScaler

from sklearn.model_selection import  RandomizedSearchCV, train_test_split

from xgboost import XGBClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.neighbors import KNeighborsClassifier

from sklearn.svm import SVC
```

d.head(15)

d.shape

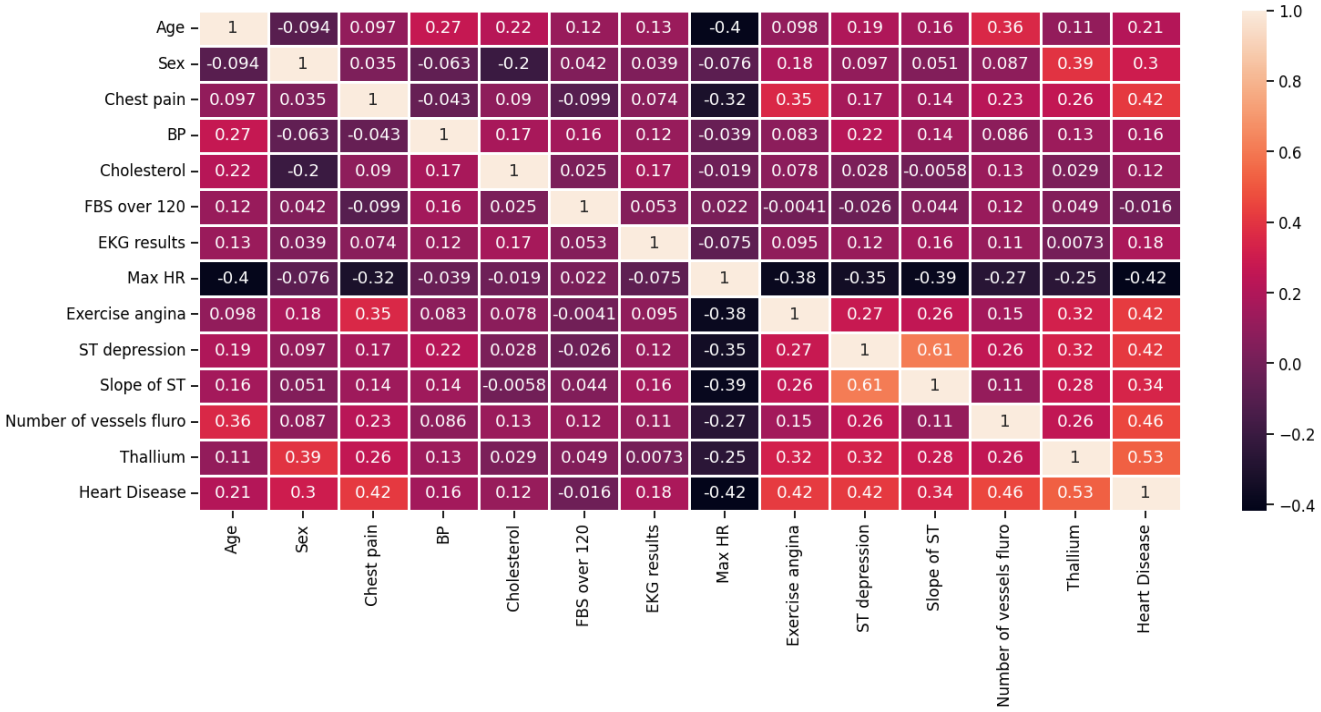
d.describe()

plt.figure(figsize=(20,12))

sns.set_context('notebook',font_scale = 1.3)

sns.heatmap(d.corr(),annot=True,linewidth =2)

plt.tight_layout()



KNN

Conversion of data to training and testing

x = d.iloc[:, :-2]

y = d.iloc[:, -1]

x_train, x_test, y_train, y_test = train_test_split(x, y, random_state = 0, test_size = 0.35)

Standard Scaler for Standardization technique

sc_x = StandardScaler()

x_train = sc_x.fit_transform(x_train)

x_test = sc_x.transform(x_test)

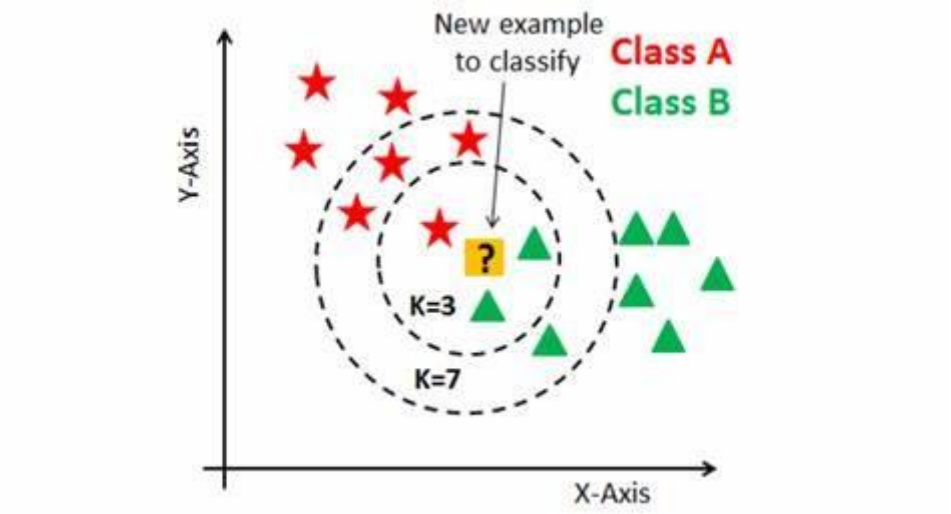
Nearest Neighbours

import math

math.sqrt(len(y_test))

→9.746794344808963

KNN MODEL



classifier = KNeighborsClassifier(n_neighbors = 9, p = 2, metric = 'euclidean')

classifier.fit(x_train,y_train)

y_pred = classifier.predict(x_test)

y_pred

```
array([0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0,
1, 1, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0,
1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0,
1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1,
0])
```

cm = confusion_matrix(y_test,y_pred)

print(cm)

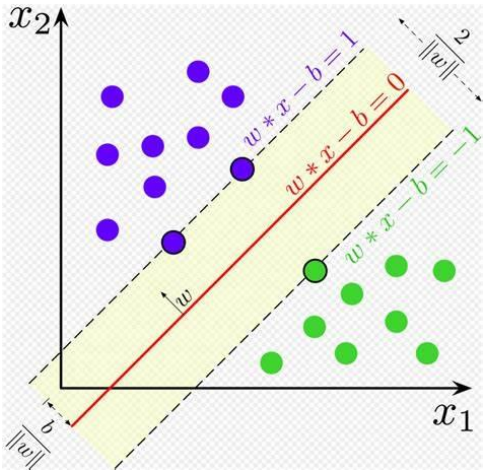
```
[[43 12]
 [11 29]]
```

Z=accuracy_score(y_test,y_pred)

print(Z*100)

```
75.78947368421053
```

SVM MODEL



from sklearn import svm

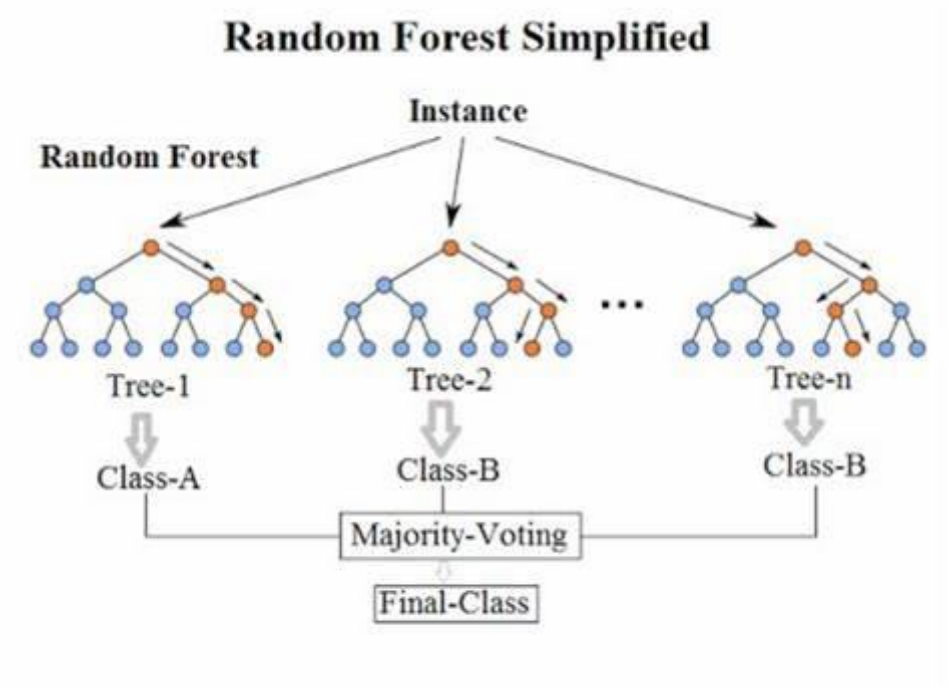
clf = svm.SVC(kernel='rbf')

clf.fit(x_train,y_train)

y_pred = clf.predict(x_test)

```
y_pred = clf.predict(x_test)

y_pred

cm = confusion_matrix(y_test,y_pred)

print(cm)

[[44 11]

 [11 29]]

X=accuracy_score(y_test,y_pred)

print(X*100)

76.84210526315789
```

RANDOM FOREST MODEL



```
rf = RandomForestClassifier(n_estimators=500, random_state=12, max_depth=5)

rf.fit(x_train,y_train)

rf_predicted = rf.predict(x_test)

rf_conf_matrix = confusion_matrix(y_test, rf_predicted)

print("confusion matrix")

print(rf_conf_matrix)

rf_acc_score = accuracy_score(y_test, rf_predicted)

print("Accuracy of Random Forest:",rf_acc_score*100)
```

```
75] rf = RandomForestClassifier(n_estimators=500, random_state=12, max_depth=5)
    rf.fit(x_train,y_train)
    rf_predicted = rf.predict(x_test)

76] rf_conf_matrix = confusion_matrix(y_test, rf_predicted)
    print("confusion matrix")
    print(rf_conf_matrix)

    confusion matrix
    [[45 10]
     [ 7 33]]

    rf_acc_score = accuracy_score(y_test, rf_predicted)
    print("Accuracy of Random Forest:",rf_acc_score*100)

    Accuracy of Random Forest: 82.10526315789474
```

Hence, the Accuracy of Random forest highest with accuracy percent:82.105%.

CONCLUSION:

Heart disease is a leading cause of death worldwide, and early identification of high-risk individuals can significantly improve public health outcomes. Machine learning algorithms such as K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and Random Forest can be used to predict the likelihood of heart disease based on health parameters such as age, sex, blood pressure, cholesterol levels, family history, and lifestyle choices.

In this proposed methodology, we have outlined the steps involved in developing a machine learning model for heart disease prediction using KNN, SVM, and Random Forest algorithms. The methodology involves data collection, pre-processing, feature selection, model training, evaluation, hyperparameter tuning, testing, and deployment.

The proposed methodology aims to develop an accurate and efficient machine learning model to predict the likelihood of heart disease and improve public health outcomes. By identifying high-risk individuals early and providing personalized interventions to reduce their risk of developing heart disease, the proposed methodology can contribute to reducing the incidence of heart disease and improving public health outcomes.

The highest accuracy for Random Forest is:82.1%

Overall, the proposed methodology for heart disease prediction using KNN, SVM, and Random Forest algorithms has the potential to significantly improve public health outcomes and contribute to reducing the global burden of heart disease.