

20CS3020AA – Machine Learning

A Skill based Project Report

on

**To predict the onset of heart disease in a patient**

Submitted in partial fulfilment of the Requirements for the award of the Degree of Bachelor of Technologyin Computer Science and Engineering

Under the Guidance of

Dr. K.Swathi

designation

by

I.D NUMBER NAME

2000030200 Ch.Meenakshi

2000030231 S.Deepika

2000030278 G.Deva Ram

2000030290 G.Akhilesh

Department of Computer Science and Engineering

(DST-FIST Sponsored Department)

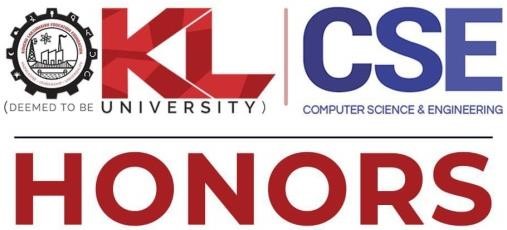
KL Education Foundation

Green Fields, Vaddeswaram, Guntur District-522 502

2022-2023

KONERU LAKSHMAIAH EDUCATION FOUNDATION

DEPARTMENT OF COMPUTER SCIENCE AND ENIGNEERING (DST-FIST Sponsored Department)



**Declaration**

We here by declare that this Skill Based Project report entitled **To predict the onset of heart disease in a patient.** has been prepared by us in the course 20CS3020AA Machine Learning in partial fulfilment of the requirement for the award of Degree Bachelor of Technology in COMPUTER SCIENCE AND ENGINEERING during the Odd Semester of the academic year 2022-2023. We also declare that this project-based lab report is of our own effort, and it has not submitted to any other university for the award of any degree.

Date: 26-10-22

Place: Vaddeswaram

Signature of the Student

**Student Name Id Number**

Ch.Meenakshi 2000030200

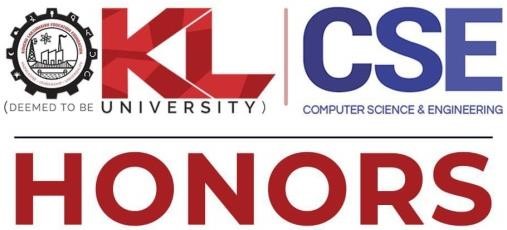
S.Deepika 2000030231

G.Deva Ram 2000030278

G.Akhilesh 2000030290

KONERU LAKSHMAIAH EDUCATION FOUNDATION

DEPARTMENT OF COMPUTER SCIENCE AND ENIGNEERING (DST-FIST Sponsored Department)



CERTIFICATE

This is to certify that the project-based Skill report entitled “**To predict the onset of heart disease in a patient.**” is a bonafide work done Ch.Meenakshi, S.Deepika, G.Deva Ram, G.Akhilesh bearing Regd. No.2000030200,2000030231, 2000030278, 2000030290 to the course 20CS3020AA Machine Learning in partial fulfillment of the requirements for the award of Degree in Bachelor of Technology in COMPUTER SCIENCE AND ENGINEERING during the Even Semester of Academic year 2022-2023.

Dr.K.Swathi Dr.Hari Kiran Vege

**FACULTY INCHARGE HEAD OF THE DEPARTMENT**

**ACKNOWLEDGEMENTS**

Our sincere thanks to Dr. K.Swathi in the skill sessions for her/his outstanding support throughout the project for the successful completion of the work

We express our gratitude to Dr. K.Swathi, Course Co-Ordinator for the course 20CS3020AA Machine Learning in the Department of Computer Science and Engineering for providing us with adequate planning and support and means by which we can complete this project.

We express our gratitude to Prof. V. HARI KIRAN, Head of the Department for Computer Science and Engineering for providing us with adequate facilities, ways and means by which we can complete this project.

We would like to place on record the deep sense of gratitude to the Vice Chancellor, K L University for providing the necessary facilities to carry out the project.

Last but not the least, we thank all Teaching and Non-Teaching Staff of our department and especially our classmates and our friends for their support.

**Name Student id**

Ch.Meenakshi 2000030200

S.Deepika 2000030231

G.Deva Ram 2000030278

G.Akhilesh 2000030290

# **INDEX**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **TITLE** | **PAGE NO** |
| 1 | Abstract | 6 |
| 2 | Introduction | 7 |
| 3 | System Requirement Specification | 8 |
| 4 | Methodology | 9 |
| 5 | Coding and implementation | 11 |
| 6 | Result Analysis | 14 |
| 7 | Conclusion | 19 |
| 8 | Future Enhancement | 20 |

**ABSTRACT**

The main objective of this project is to predict the onset of heart disease in a patient, where we are given a dataset that comprises of predictors such as cholesterol, age, diabetes, and family history.

Python can be used to analyse the relationships present in this dataset by which we will get to know few things such as whether there are patients with diabetes more likely to develop heart disease at an early age, whether there is a certain demographic group that is at higher risk of heart disease than others, whether frequent exercise lower the risk of developing heart disease and finally whether there are smokers more likely to develop heart disease than non-smokers.

**INTRODUCTION**

**Machine learning is a subset of AI, which enables the machine to automatically learn from data, improve performance from past experiences, and make predictions.** Machine learning contains a set of algorithms that work on a huge amount of data. Data is fed to these algorithms to train them, and based on training, they build the model & perform a specific task.

Machine learning (ML) proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry.

This makes heart disease a major concern to be dealt with. But it is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and many other factors. Due to such constraints, scientists have turned towards modern approaches like Data Mining and Machine Learning for predicting the disease.

Here we will be applying Machine Learning approaches (and eventually comparing them) for classifying whether a person is suffering from heart disease or not, using one of the most used datasets— heart disease dataset.

Thus, preventing heart diseases has become more than necessary. Good data-driven systems for predicting heart diseases can improve the entire research and prevention process, making sure that more people can live healthy lives. This is where Machine Learning comes into play. Machine Learning helps in predicting the heart diseases, and the predictions made are quite accurate.

**SYSTEM REQUIREMENT SPECIVICATION**

* SOFTWARE REQUIREMENTS:

The major software requirements of the project are as follows:

Language : Python, Python Libraries

Tools : Microsoft Word and Jupiter Notebook

* HARDWARE REQUIREMENTS:

The hardware requirements that map towards the software are as follows:

• Intel (or AMD equivalent) i5 or better processor, 7th generation or newer

• Windows 10 Operating System

• 1920 x 1080 or greater screen resolution

• 500 GB or larger SSD

• Minimum 8 GB of RAM (12GB -16GB RAM recommended)

• Access to High-Speed Internet

# **METHODOLOGY**

Our research explores data-driven approaches which utilize supervised machine learning models to identify patients with such diseases.

Using the Heart dataset, we conduct an exhaustive search of all available feature variables within the data to develop models for cardiovascular, prediabetes, and diabetes detection.

Using different timeframes and feature sets for the data (based on laboratory data), multiple machine learning models (logistic regression, support vector machines, random forest, and gradient boosting) were evaluated on their classification performance.

**DATA ANALYTICS**

Let us look at the people’s age who are suffering from the disease or not.  
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. Next, let us look at the distribution of age and gender for each target class.

We see that for females who are suffering from the disease are older than males.

**CODING**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.impute import SimpleImputer

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LogisticRegression

from sklearn.linear\_model import LinearRegression

from sklearn.tree import DecisionTreeClassifier

from sklearn.metrics import accuracy\_score

dataset=pd.read\_csv("heartDisease.csv")

dataset

dataset.shape

dataset.head(5)

dataset.tail(5)

dataset.describe()

dataset.info()

dataset.isnull().sum()

correlation=dataset.corr()

correlation

x=dataset[['Age','Sex','ChestPainType','RestingBP','Cholesterol','FastingBS','RestingECG','MaxHR','ExerciseAngina','Oldpeak','ST\_Slope']].values

x

y=dataset[['HeartDisease']].values

y

Imputer=SimpleImputer(missing\_values=np.nan,strategy='most\_frequent')

Imputer=Imputer.fit(x[:,1:12])

x[:,1:12]=Imputer.transform(x[:,1:12])

x

from sklearn.preprocessing import LabelEncoder

l\_encode=LabelEncoder()

x[:,2]=l\_encode.fit\_transform(x[:,2])

x[:,1]=l\_encode.fit\_transform(x[:,1])

x[:,8]=l\_encode.fit\_transform(x[:,8])

x[:,10]=l\_encode.fit\_transform(x[:,10])

x[:,6]=l\_encode.fit\_transform(x[:,6])

x

from sklearn.model\_selection import train\_test\_split

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.2,random\_state=0)

x\_train

x\_test

y\_train

y\_test

print(x.shape,x\_train.shape,x\_test.shape,y\_train.shape,y\_test.shape)

model=LogisticRegression()

model.fit(x\_train,y\_train)

x\_train\_pred=model.predict(x\_train)

training\_data\_accuracy=accuracy\_score(x\_train\_pred,y\_train)

print("accuracy on training data is : ",training\_data\_accuracy)

x\_test\_pred=model.predict(x\_test)

test\_data\_accuracy=accuracy\_score(x\_test\_pred,y\_test)

print("Acuraccy on testing data is : ",test\_data\_accuracy)

input\_data=(37,1,1,130,283,0,1,98,0,0.0,2)

input\_data\_as\_numpy\_array=np.array(input\_data)

input\_data\_reshaped=input\_data\_as\_numpy\_array.reshape(1,-1)

prediction=model.predict(input\_data\_reshaped)

print( prediction)

if(prediction==0):

print("Person not having heart disease")

else:

print("Person is having heart disease")

sns.heatmap(correlation, annot=True)

sns.displot(dataset)

sns.pairplot(dataset)

plt.show()

from sklearn.tree import DecisionTreeClassifier

decision\_tree = DecisionTreeClassifier(random\_state=0, max\_depth=2)

model = decision\_tree.fit(x\_train, y\_train)

y\_pred = model.predict(x\_test)

from sklearn.metrics import accuracy\_score

acc\_score=accuracy\_score(y\_test, y\_pred)

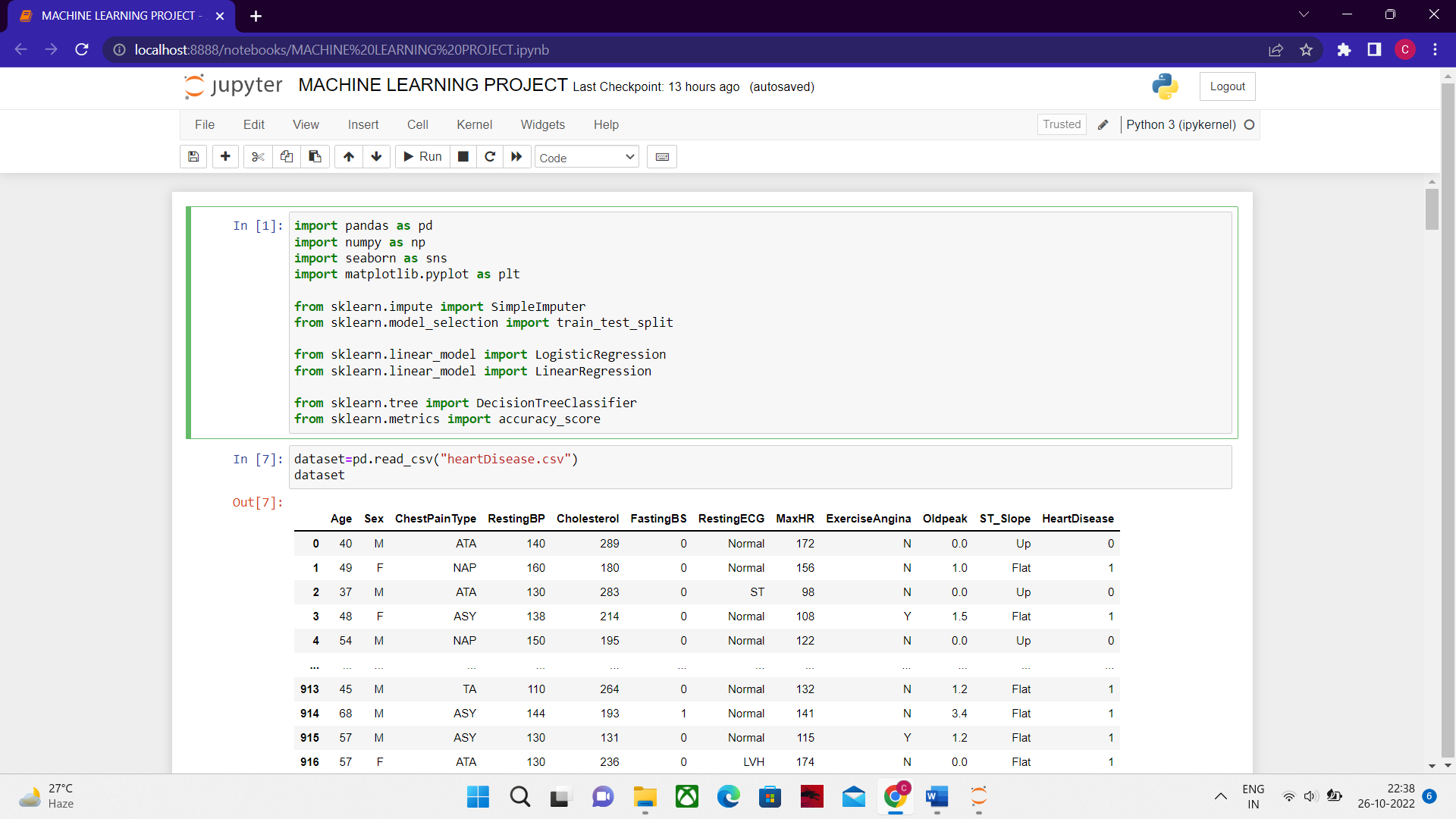
print("accuracy is :" ,acc\_score)

from sklearn import tree

tree.plot\_tree(model)

**RESULT ANALYSIS**

Screen Shots:



A screenshot of a computer

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application, email

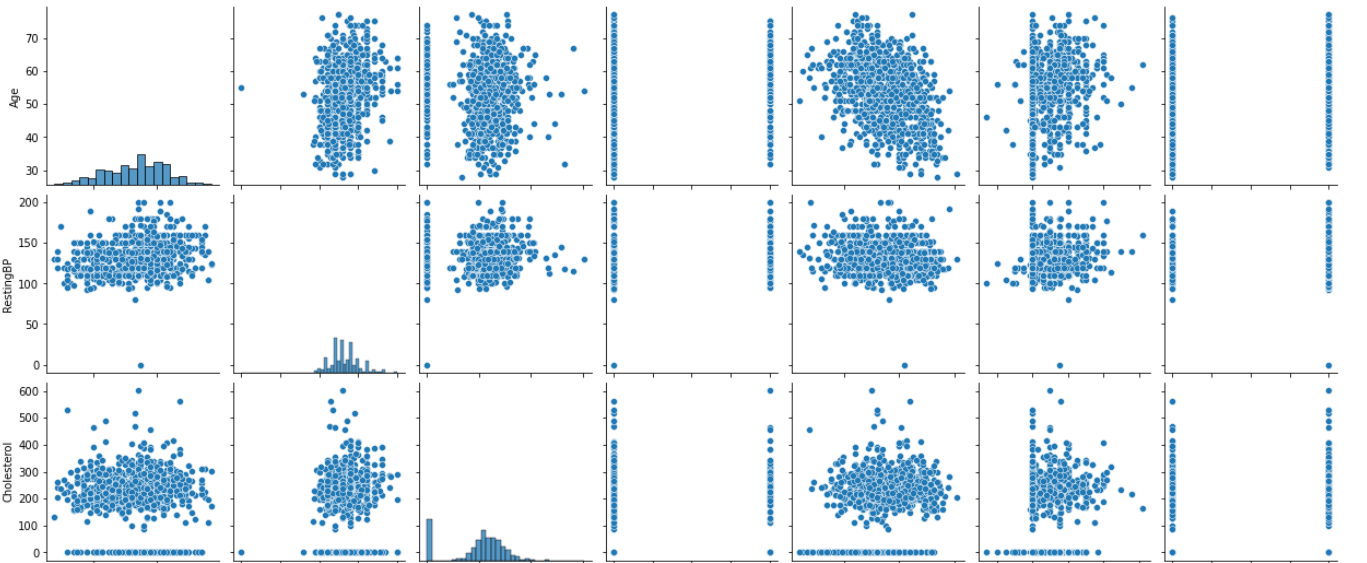
Description automatically generated

Graphical user interface

Description automatically generated

A picture containing graphical user interface

Description automatically generated

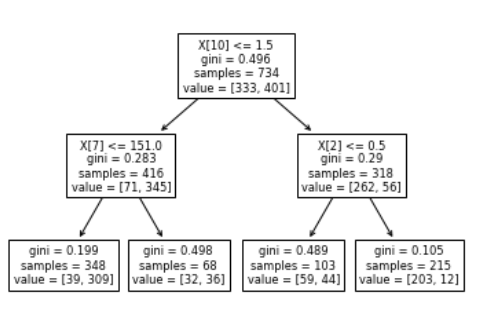


A picture containing chart

Description automatically generated

Text

Description automatically generated with medium confidence



# **CONCLUSION**

Heart Disease is one of the major concerns for society today. It is difficult to manually determine the odds of getting heart disease based on risk factors. However, machine learning techniques are useful to predict the output from existing data.

**FUTURE ENHANCEMENT**

Machine learned models based on survey questionnaire can provide an automated identification mechanism for patients at risk of diabetes and cardiovascular diseases.

We also identify key contributors to the prediction, which can be further explored for their implications on electronic health records.