Project Report

Heart Disease Prediction

December 2023

ISSUED BY

Dr. Uzair Iqbal

REPRESENTATIVE

IBRAHIM ABID (21I-0298)

Table of Contents

Introduction and Domain Background

Introduction

Heart disease remains a leading cause of mortality worldwide, emphasizing the critical need for accurate prediction and early intervention. In this report, we are diving into the application of machine learning for heart disease prediction, aiming to contribute to the advancements in cardiovascular healthcare.

Domain Background

Cardiovascular diseases encompass a range of conditions affecting the heart and blood vessels, including coronary artery disease, heart failure, and arrhythmias. Early detection of heart disease risk factors is paramount for preventive measures and personalized patient care. Machine learning offers a promising avenue for developing predictive models that can assist healthcare professionals in identifying individuals at high risk.

Dataset Description:

Dataset Overview:

My analysis is based on a comprehensive dataset containing various attributes related to heart health. The dataset comprises numerical features, including age, sex, blood pressure, cholesterol levels, and exercise-induced angina, among others. These attributes provide a rich set of information for training and evaluating machine learning models.

Attribute Details:

- ***Age:*** The age of the individual.
- ***Sex:*** Gender of the individual (0 for female, 1 for male).
- ***CP:*** Chest pain type (0 to 3).
- ***Trestbps:*** Resting blood pressure.
- ***Chol:*** Serum cholesterol.
- ***Fbs:*** Fasting blood sugar > 120 mg/dl (1 for true, 0 for false).
- ***Restecg:*** Resting electrocardiographic results (0 to 2).
- ***Thalach:*** Maximum heart rate achieved.
- ***Exang:*** Exercise-induced angina (1 for yes, 0 for no).
- ***Oldpeak:*** ST depression induced by exercise.
- ***Slope:*** Slope of the peak exercise ST segment (0 to 2).
- ***Ca:*** Number of major vessels colored by fluoroscopy (0 to 3).
- ***Thal:*** Thalassemia type (0 to 3).
- ***Target:*** Presence of heart disease (1 for yes, 0 for no).

Descriptive Statistics:

The table below provides descriptive statistics for the dataset:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	1052.000000	1052.000000	1052.000000	1052.000000	1052.000000	1052.000000	1052.000000	1052.000000	1052.000000	1052.000000	1052.000000	1052.000000	1052.000000	1052.000000
mean	54.449620	0.701521	0.933460	131.629278	246.143536	0.147338	0.531369	148.660646	0.339354	1.086217	1.377376	0.766160	2.327947	0.500000
std	9.058134	0.457808	1.028306	17.462463	51.290792	0.354612	0.528867	23.245471	0.473715	1.175888	0.617830	1.033444	0.629030	0.500238
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	48.000000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	132.000000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	56.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	152.000000	0.000000	0.800000	1.000000	0.000000	2.000000	0.500000
75%	61.000000	1.000000	2.000000	140.000000	276.000000	0.000000	1.000000	165.250000	1.000000	1.800000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

My Contribution:

In this project there were certain flaws which I fixed such as:

- *Data Imbalance*
- *Applied Standard Scaler.*
- *Printed Classification Report.*
- *Removed Outliers from dataset.*
- *Checked for any NaN Values in Dataset.*
- *Used Random OverSampler to fix imbalanced dataset.*
- *Used Neural Network with Sigmoid Activation Function at its last Layer.*

Used Standard libraries:

- *NumPy:*

np.asarray():
Converts input to an array.
np.reshape():
Gives a new shape to an array without changing its data.

- *Pandas:*

pd.read_csv():
Reads a comma-separated values (csv) file into a DataFrame.
DataFrame.head():
Returns the first n rows of a DataFrame.
DataFrame.isna():
Detects missing values.
DataFrame.info():
Prints a concise summary of a DataFrame.
DataFrame.isnull().sum():
Counts the number of missing values in each column.
DataFrame.select_dtypes():
Selects columns based on their data types.

- *Seaborn:*

sns.countplot():
Shows the counts of observations in each categorical bin using bars.
sns.boxplot():
Draws a box plot to show distributions with respect to categories.

- *Matplotlib:*

plt.figure():
Creates a new figure.

plt.show():
Displays the figure.

- *Imbalanced-learn*

RandomOverSampler():
Randomly over-samples the minority class.

- *Scikit-learn:*

train_test_split():
Splits arrays or matrices into random train and test subsets.
accuracy_score():
Accuracy classification score.
StandardScaler():
Standardizes features by removing the mean and scaling to unit variance.

- *TensorFlow-Keras:*

keras.Sequential():
Linear stack of layers for building the neural network model.
model.compile():
Configures the model for training.
model.fit():
Trains the model for a fixed number of epochs.
model.predict():
Generates predictions for input samples.

- *Other Python built-in libraries:*

os (for file path operations)