Inspiring Excellence

HEART DISEASE PREDICTION

CSE422

SUBMITTED TO:

SYED ZAMIL HASAN (SZH)

SUBMITTED BY:

Name: Nahian Hoque Dipto
ID: 20101260
Section: 11

Name: Adhara Labannya
ID: 20101181
Section: 11

Name: Sahadat Hossain
ID: 20101421
Section: 03

Name: Somiya Azadi Tonni
ID: 20101187
Section: 11

ARTIFICIAL INTELLIGENCE

1. Introduction:

Heart diseases mostly refer to conditions that cause blocked or constricted blood vessels, which may lead to angina, followed by a stroke or heart attack. Disruption of the heart's rhythm, valve, or muscle are signs of the early phases of the fatal disease. Machine learning is crucial for predicting whether someone is likely to suffer from heart disease in the near future or if someone is already afflicted with it. A common scenario with a health issue as severe as this is that it requires staying up to date with body conditions and having services ready in case of emergency. Not to mention, not everyone can afford / have access to these facilities. The paper focuses on the construction of an artificial intelligence-based heart disease detection system using machine learning algorithms. We demonstrate how artificial intelligence can be used to forecast if someone would get cardiac disease. A python-based application is created for healthcare research in this study since it is more dependable and aids in tracking and establishing various kinds of health monitoring applications.

2. Dataset Description:

2.1. Source:

Link- https://www.kaggle.com/johnsmith88/heart-disease-dataset/metadata

Reference - https://intellipaat.com/blog/confusion-matrix-python

2.2. Dataset Description:

- Features in our dataset are 13 in number and consist of the following:

2.2.1. **Age:** Age in years.

2.2.2. **Sex:** Two unique results. (male = 1, female = 0)

2.2.3. **Cp:** 4 types of pain in the chest are experienced. (Typical angina = 1, atypical angina = 2, non-anginal pain = 3, asymptomatic = 4)

2.2.4. **Trestbps:** Resting blood pressure.

2.2.5. **Chol:** Serum cholesterol in mg/dl

2.2.6. **Fbs:** Fasting blood sugar > 120 mg/dl

2.2.7. **Restecg:** Resting electrocardiographic results (values 0,1,2).

2.2.8. **Thalach:** Maximum heart rate achieved.

2.2.9. **Exang:** Exercise induced angina.

2.2.10. **Oldpeak :** Oldpeak = ST depression induced by exercise relative to rest.

2.2.11. **Slope:** The slope of the peak exercise ST segment.

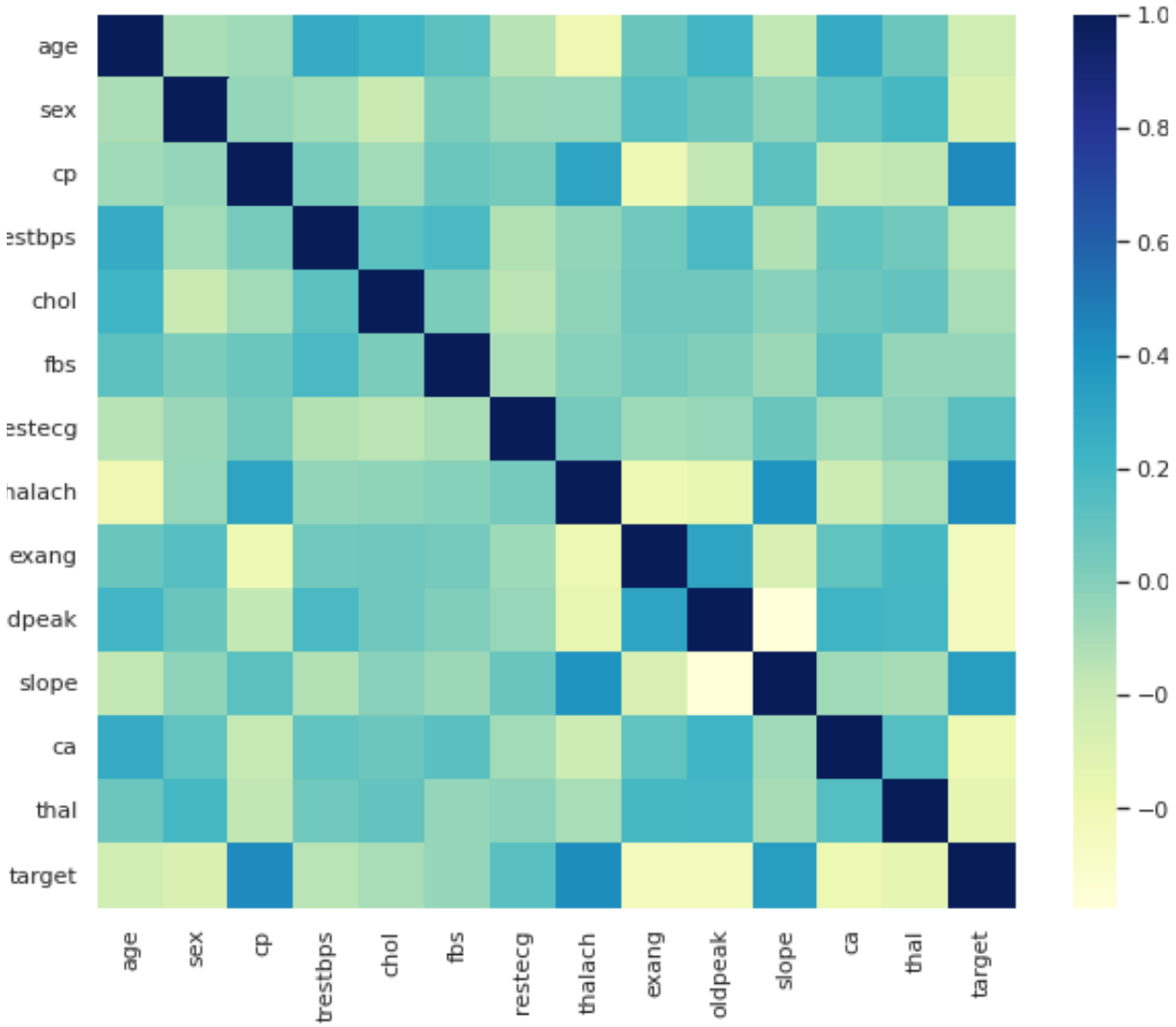
2.2.12. **Ca :** Number of major vessels (0-3) colored by fluoroscopy

2.2.13. **Thal:** Thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

- **Classification or regression?** => It is a classification problem.

- **Data Points:** Total of -1024 Rows and 14 Columns

- **Correlation of all the features (HEATMAP):**



3. Dataset pre-processing:

3.1. Faults: Luckily, we did not have any null values. Therefore, we do not have to change or remove any rows or columns in our dataset. Also, there are total of 5 types of categorical values and these are:

* Sex(m/f)
* Chest Pain (CP)
* Resting ECG (Restecg)
* Number of major vessels (CA)
* Defects (Thal)

3.2. Solution: In case of handling categorical and binary values, we have used binary encoding for this purpose and later we transformed the encoded values.

4. Dataset splitting:

4.1. Train set (70%)

4.2. Test set (30%)

5. Model training & testing:

5.1. Logistic Regression (for classification problem):

A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

5.2. Decision Tree:

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

5.3. Random Forest:

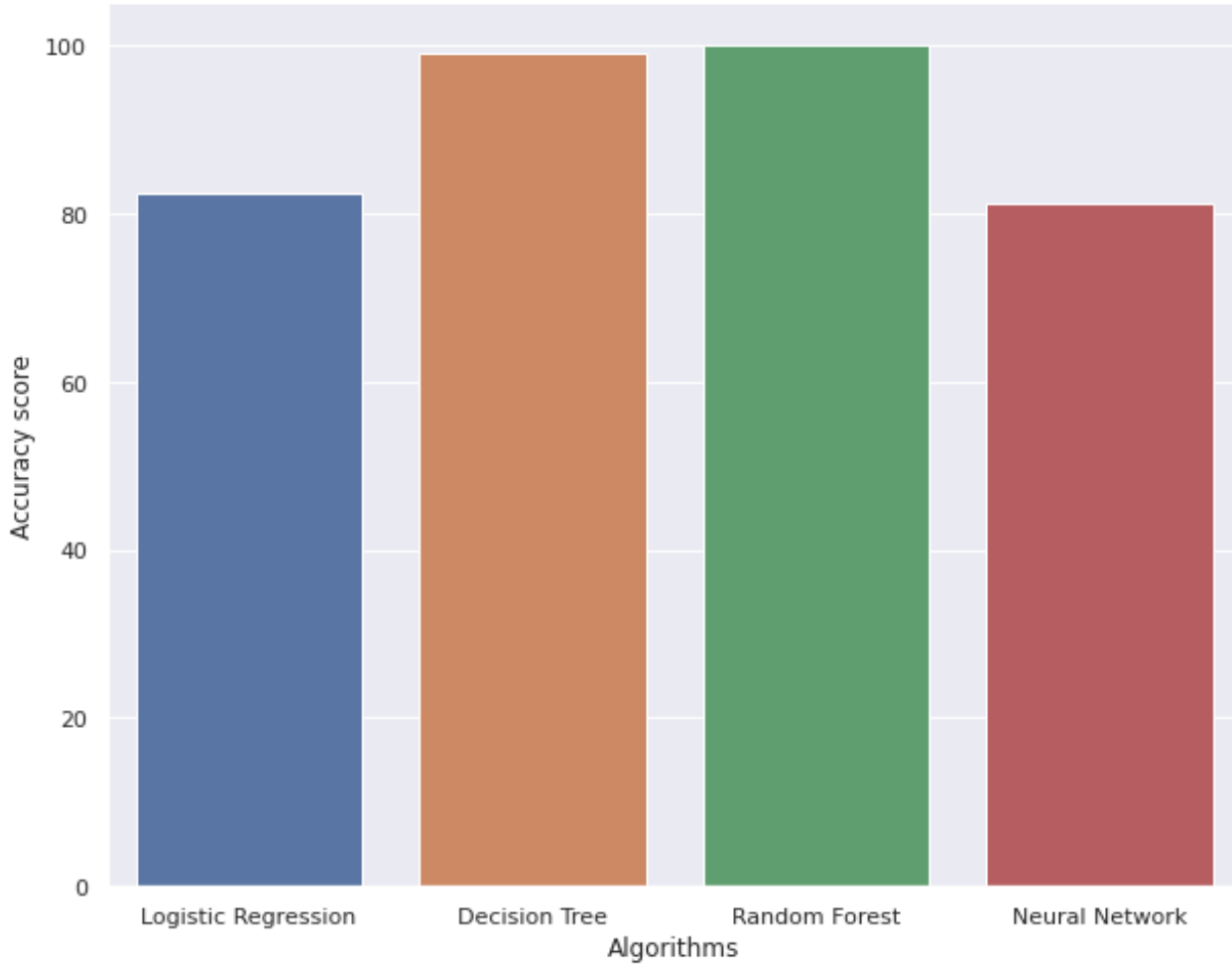
Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

5.4. Neural Network:

Neural Networks are the computing system that is designed to simulate the way the human brain analyzes and processes the information. Artificial Neural Networks have self-learning capabilities that enable it to produce a better result as more data become available.

6. Model selection/Comparison analysis:

6.1. Bar chart showcasing prediction accuracy of all models:



6.2. Precision, recall comparison of each model:

6.2.1. Logistic Regression:

	precision	recall	f1-score	support
0	0.91	0.73	0.81	161
1	0.76	0.93	0.83	147
accuracy			0.82	308
macro avg	0.84	0.83	0.82	308
weighted avg	0.84	0.82	0.82	308

6.2.2. Decision Tree:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	161
1	1.00	0.98	0.99	147
accuracy			0.99	308
macro avg	0.99	0.99	0.99	308
weighted avg	0.99	0.99	0.99	308

6.2.3. Random Forest:

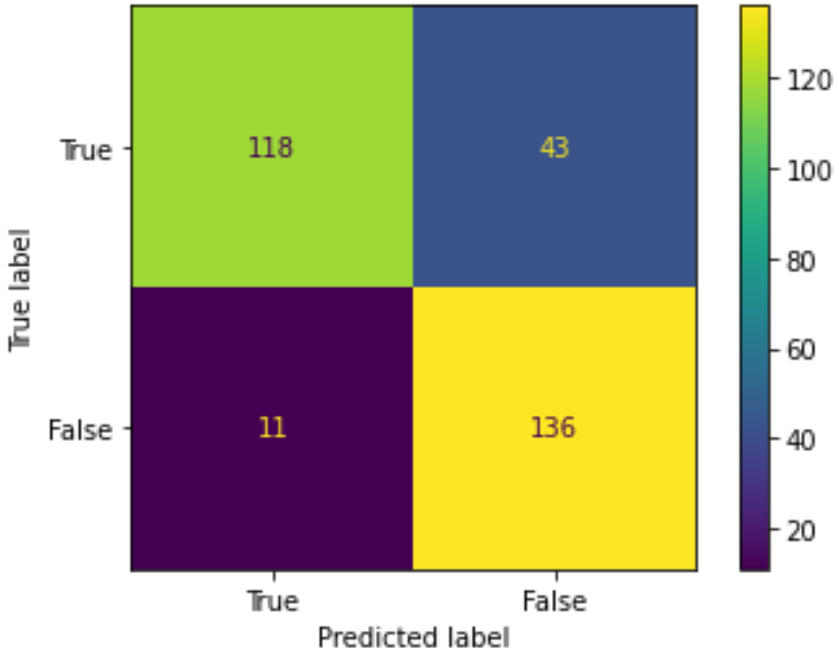
	precision	recall	f1-score	support
0	1.00	1.00	1.00	161
1	1.00	1.00	1.00	147
accuracy			1.00	308
macro avg	1.00	1.00	1.00	308
weighted avg	1.00	1.00	1.00	308

6.2.4. Neural Network:

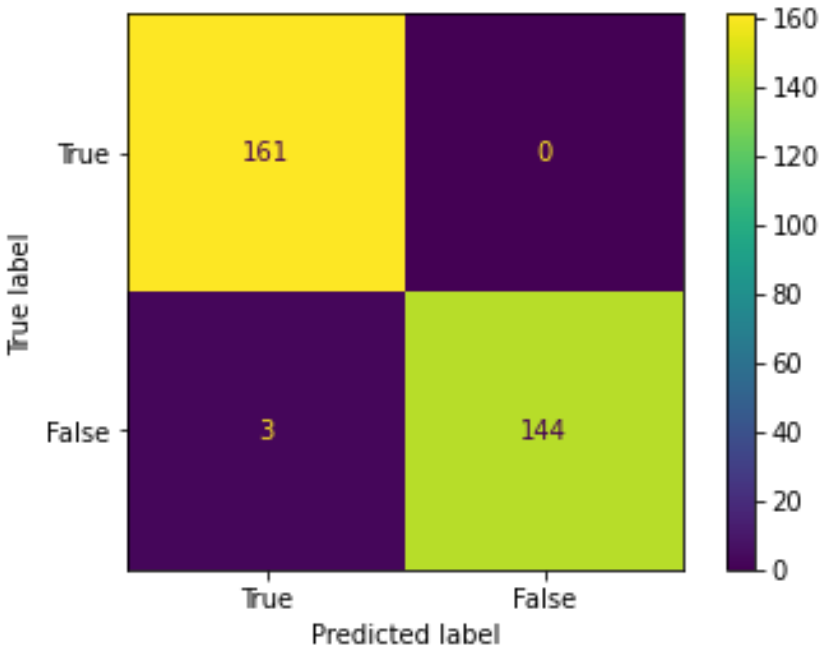
	precision	recall	f1-score	support
0	0.86	0.76	0.81	161
1	0.77	0.86	0.81	147
accuracy			0.81	308
macro avg	0.81	0.81	0.81	308
weighted avg	0.82	0.81	0.81	308

6.3. Confusion Matrix (for each model):

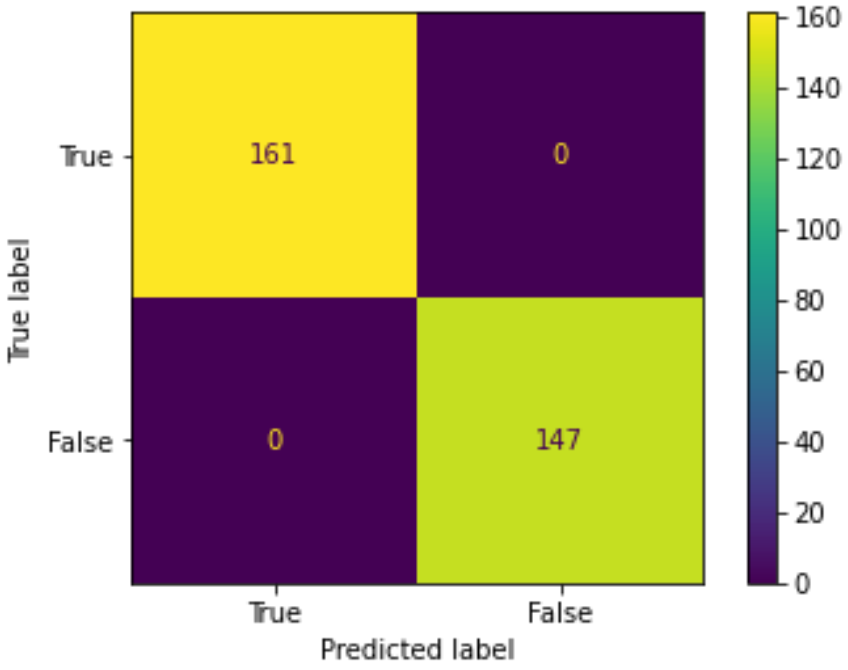
6.3.1. Logistic Regression:



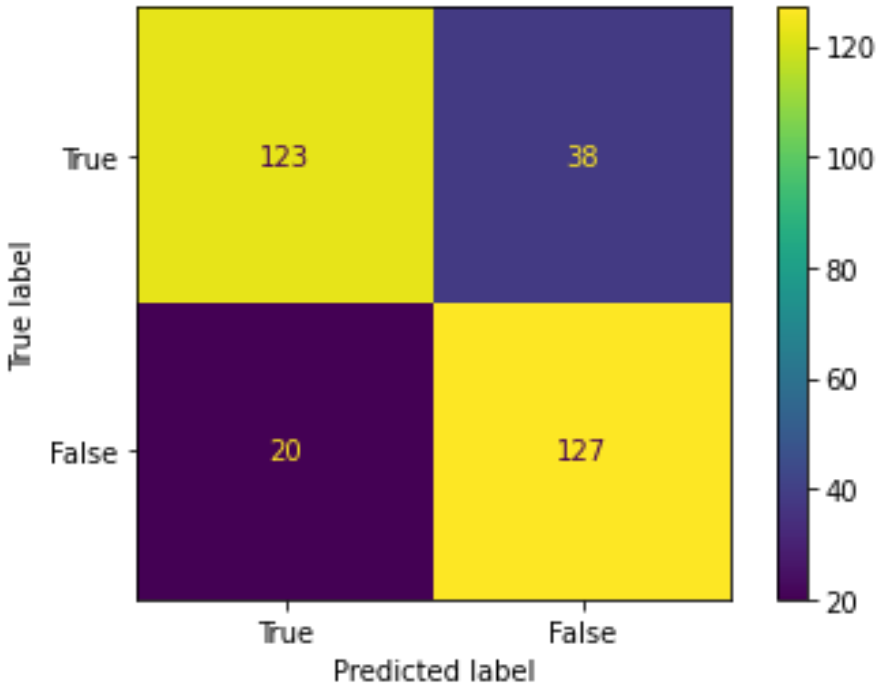
6.3.2. Decision Tree:



6.3.3. Random Forest:



6.3.4. Neural Network:



7. Conclusion:

An online user support and consulting initiative, the Heart Disease Prediction system. Here, we suggest a web application that enables users to receive immediate knowledge on their cardiac problems through a sophisticated online system. The application system with different information, along with the heart illness that is related to that information. Users of the program can share heart-related problems. After that, it examines user-specific information to look for various illnesses that might be connected to it. Here, we employ a few clever data mining algorithms to identify the ailment that is most likely to be connected

to the patient's specifics.