Heart Disease Analysis

RU DATA SCIENCE PROJECT 4
Group members:
Steph Chestnut, Christina Gabriel, Medha Mallampati, Anna Lewis

Goal

The main goal is to identify the diagnosis of heart disease by building classification models on Heart Disease data, given the provided feature set. By analyzing:

age, sex, type of chest pain, patient history, etc.,

Using both VA Long Beach and Cleveland databases, we preprocessed the datasets for model prediction. After getting a good accuracy score and prediction rate, visualize the prediction model and additional analysis tables using Tableau.

Data Source

The dataset was retrieved from UC Irvine Machine Learning Repository at the following link:

https://archive.ics.uci.edu/dataset/45/heart+disease.

DataFrame stats:

Original: three unique data sets; combined 503 rows and 76 columns

After cleaning: Used the combined data set for analysis and a combined data set with 503 rows and 12 columns to call out disease symptoms.

Ethics:

We made sure that the The UCI Machine Learning Repository website allows complete access to their data

Data Visualization

The data was visualized using Tableau, comparing different factors of the dataset,

Here is the link to our story:

https://public.tableau.com/app/profile/anna.lewis2284/viz/Project4HeartDisease_17222974244730/Story1?publish=yes

Models Used:

- Decision Tree
- Neural Network
- Logistic Regression
- Support Vector Machine (SVM)

Feature Importances

Target Variable: ‘Diagnosis’

0 - absence of Heart Disease

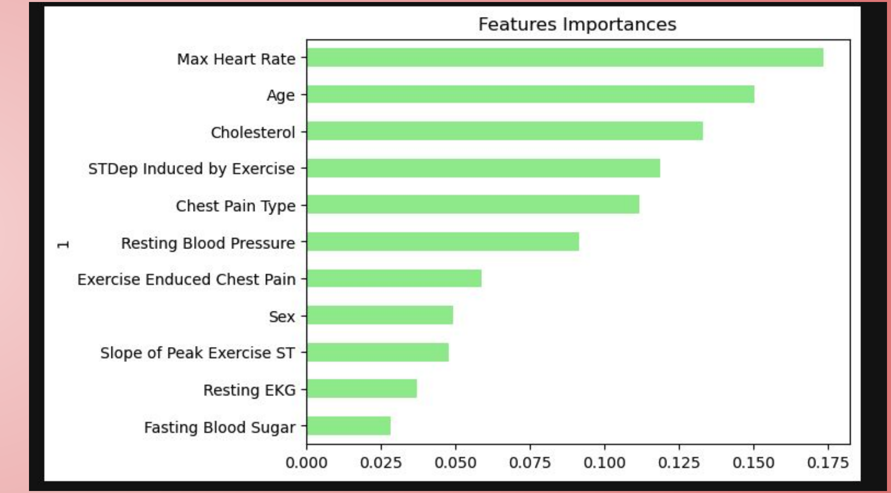
1 - presence of Heart Disease (1,2,3,4)

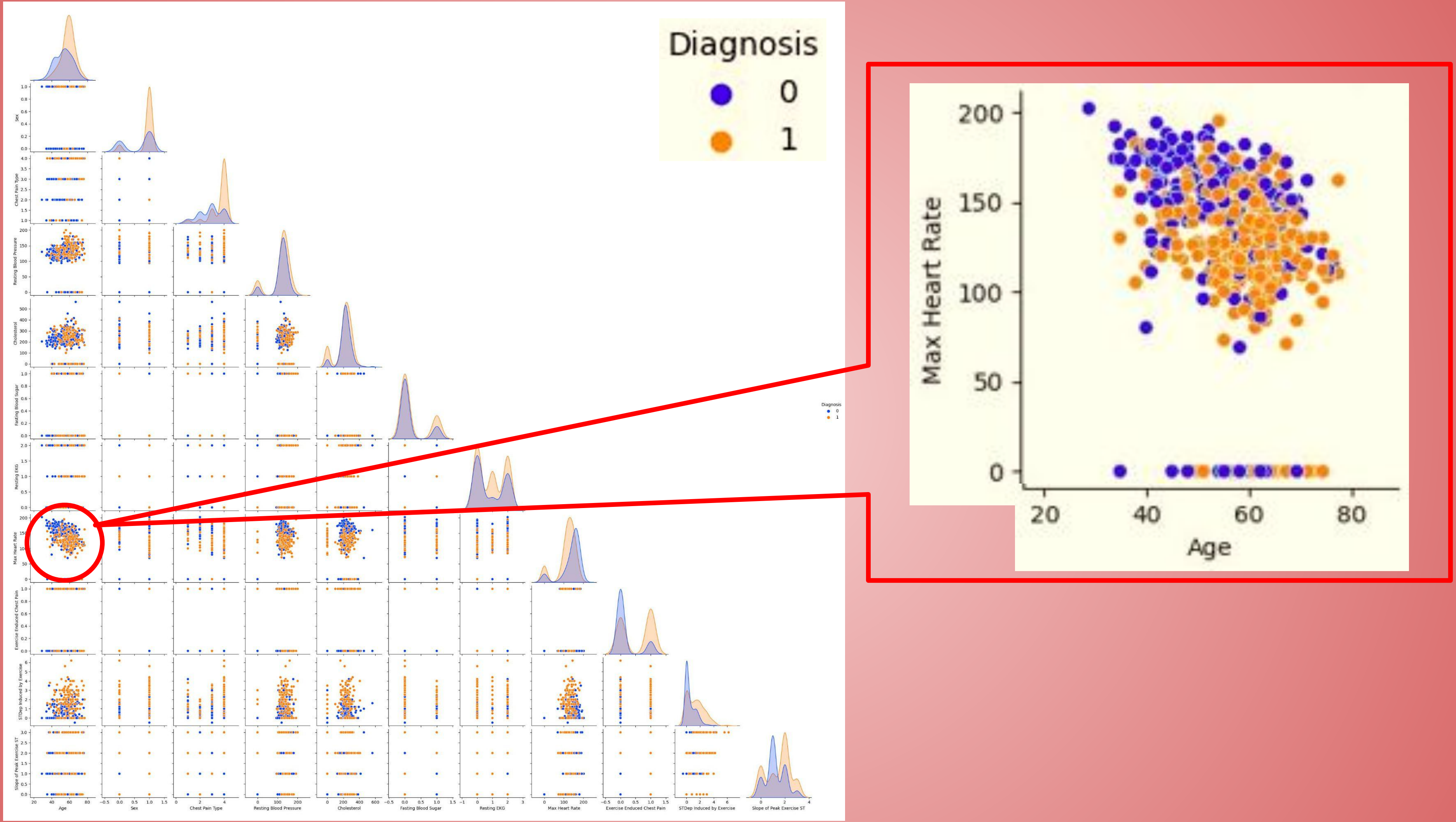
Features: Used all features for some models.

Selected Features for other models were:

- ‘Cholesterol’ and ‘Resting Blood Pressure’
- ‘Max Heart Rate’ and ‘Age’

Used to see if there were specific features that can be used to predict heart disease.

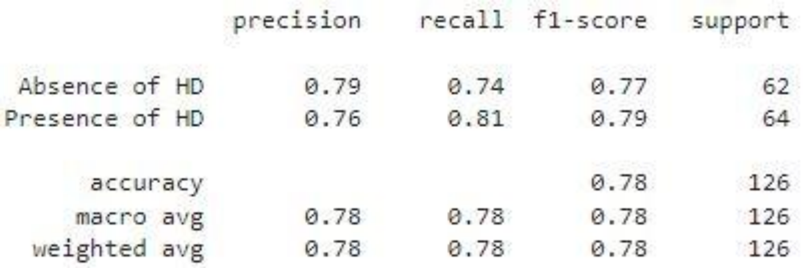




Diagnosis
0
1
Age
Sex
Chest Pain Type
Resting Blood Pressure
Cholesterol
Fasting Blood Sugar
Resting EKG
Max Heart Rate
Exercise Enduced Chest Pain
STDep Induced by Exercise
Slope of Peak Exercise ST
Max Heart Rate
200
150
100
50
0
20
40
60
80
Age

Logistic Regression & SVM

Logistic Regression (solver='lbfgs')

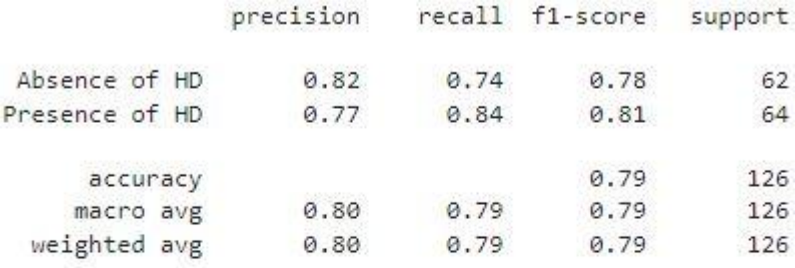


```
                precision    recall  f1-score   support

 Absence of HD       0.79      0.74      0.77        62
Presence of HD       0.76      0.81      0.79        64

      accuracy                           0.78       126
     macro avg       0.78      0.78      0.78       126
  weighted avg       0.78      0.78      0.78       126
```

SVM (kernel='linear')

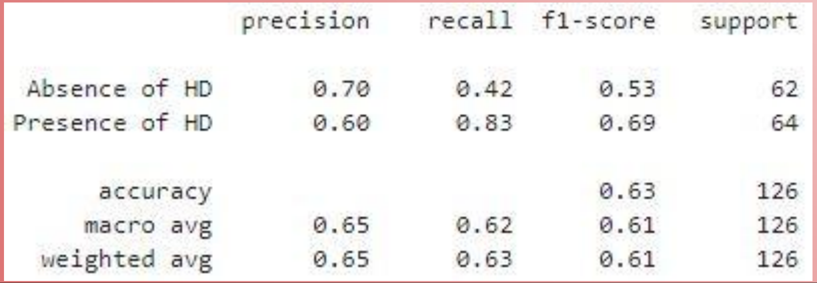


```
                precision    recall  f1-score   support

 Absence of HD       0.82      0.74      0.78        62
Presence of HD       0.77      0.84      0.81        64

      accuracy                           0.79       126
     macro avg       0.80      0.79      0.79       126
  weighted avg       0.80      0.79      0.79       126
```

Max Heart Rate and Age

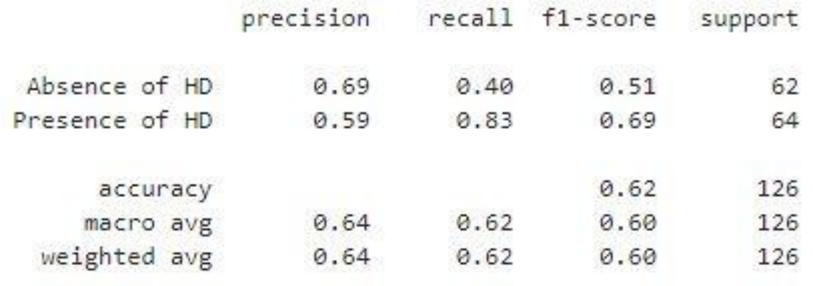


```
                precision    recall  f1-score   support

 Absence of HD       0.70      0.42      0.53        62
Presence of HD       0.60      0.83      0.69        64

      accuracy                           0.63       126
     macro avg       0.65      0.62      0.61       126
  weighted avg       0.65      0.63      0.61       126
```

Max Heart Rate and Age



```
                precision    recall  f1-score   support

 Absence of HD       0.69      0.40      0.51        62
Presence of HD       0.59      0.83      0.69        64

      accuracy                           0.62       126
     macro avg       0.64      0.62      0.60       126
  weighted avg       0.64      0.62      0.60       126
```

NN Data Model & Optimization

Target Variable: ‘Diagnosis’

0 - absence of Heart Disease

1 - presence of Heart Disease (1,2,3,4)

Features: Used all features for some models.

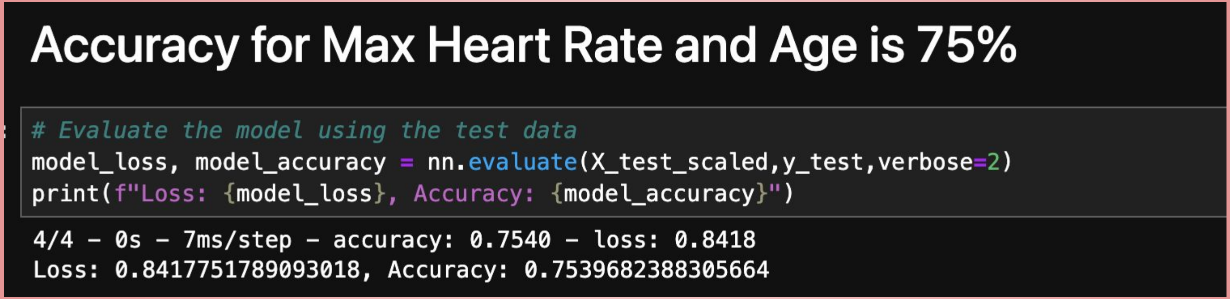
Selected Features for other models were:

- ‘Cholesterol’
- ‘Resting Blood Pressure’
- ‘Max Heart Rate’
- ‘Age’

Used to see if there were specific features that can be used to predict heart disease.

NN Data Model & Optimization

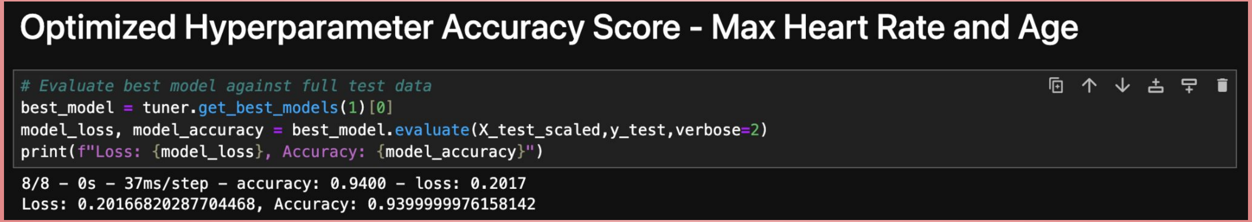
Three layer model using relu, relu and sigmoid activation functions



Accuracy for Max Heart Rate and Age is 75%

```
# Evaluate the model using the test data
model_loss, model_accuracy = nn.evaluate(X_test_scaled,y_test,verbose=2)
print(f"Loss: {model_loss}, Accuracy: {model_accuracy}")
```

```
4/4 - 0s - 7ms/step - accuracy: 0.7540 - loss: 0.8418
Loss: 0.8417751789093018, Accuracy: 0.7539682388305664
```



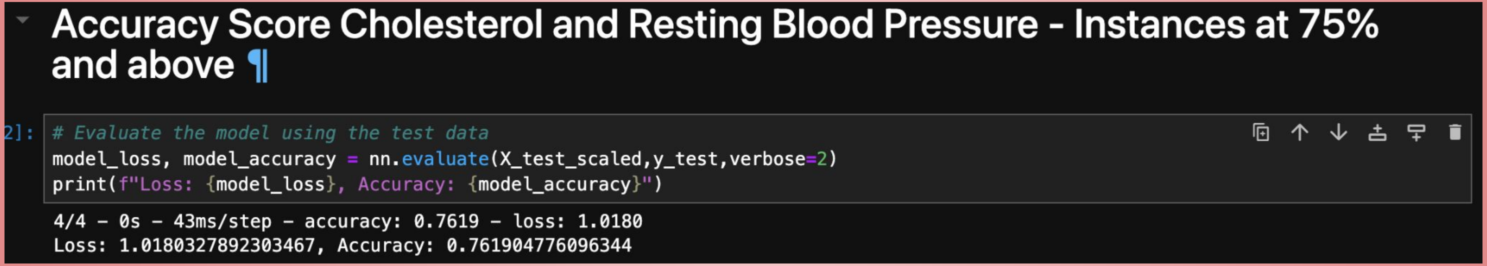
Optimized Hyperparameter Accuracy Score - Max Heart Rate and Age

```
# Evaluate best model against full test data
best_model = tuner.get_best_models(1)[0]
model_loss, model_accuracy = best_model.evaluate(X_test_scaled,y_test,verbose=2)
print(f"Loss: {model_loss}, Accuracy: {model_accuracy}")
```

```
8/8 - 0s - 37ms/step - accuracy: 0.9400 - loss: 0.2017
Loss: 0.20166820287704468, Accuracy: 0.9399999976158142
```

NN Data Model & Optimization

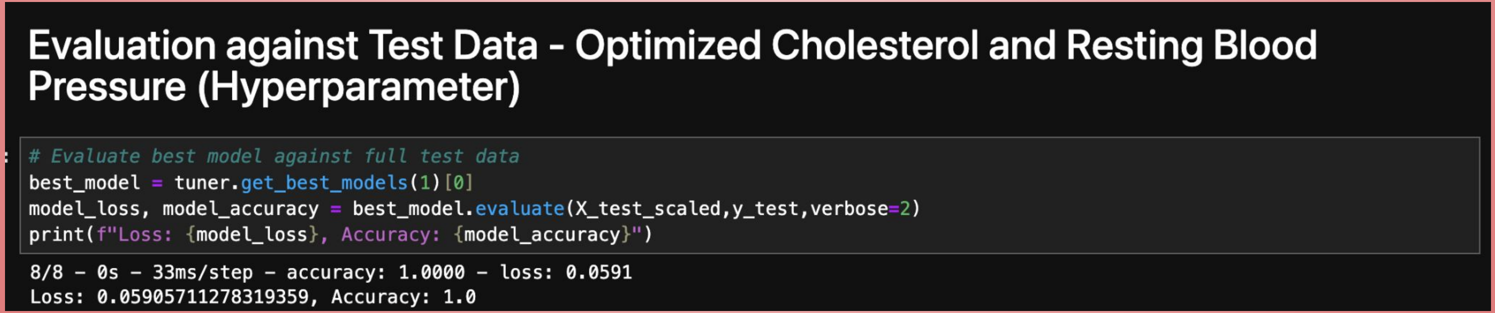
Two layer model using relu and sigmoid activation functions



Accuracy Score Cholesterol and Resting Blood Pressure - Instances at 75% and above ¶

```
# Evaluate the model using the test data
model_loss, model_accuracy = nn.evaluate(X_test_scaled,y_test,verbose=2)
print(f"Loss: {model_loss}, Accuracy: {model_accuracy}")
```

```
4/4 - 0s - 43ms/step - accuracy: 0.7619 - loss: 1.0180
Loss: 1.0180327892303467, Accuracy: 0.761904776096344
```



Evaluation against Test Data - Optimized Cholesterol and Resting Blood Pressure (Hyperparameter)

```
# Evaluate best model against full test data
best_model = tuner.get_best_models(1)[0]
model_loss, model_accuracy = best_model.evaluate(X_test_scaled,y_test,verbose=2)
print(f"Loss: {model_loss}, Accuracy: {model_accuracy}")
```

```
8/8 - 0s - 33ms/step - accuracy: 1.0000 - loss: 0.0591
Loss: 0.05905711278319359, Accuracy: 1.0
```

Comparing Best Model Scores

The top classification models for Heart Disease diagnosis were the following:

Model	Accuracy Score
Logistic Regression (solver=’lbfgs’)	78%
SVM (kernel=’linear’)	79%
NN Optimization Model (Max Heart Rate & Age)	94%
NN Optimization Model (Cholesterol & Resting Blood Pressure)	100%

Conclusion

The model that provided the best accuracy in diagnosing Heart Disease is the Neural Networking Optimized model using features Cholesterol and Resting Blood Pressure. Using those two features can determine the presence and absence of heart disease the best based on the overall scores.

The model that did the best using the full dataset features would be the Support Vector Machine model, where the data is balanced and the based on the precision and recall.

Based on the model results, linear classification models do best in diagnosing heart disease.

References

https://stackoverflow.com/questions/38640109/logistic-regression-python-solvers-definitions

https://github.com/christianversloot/machine-learning-articles/blob/main/how-to-visualize-support-vectors-of-your-svm-classifier.md

https://seaborn.pydata.org/generated/seaborn.pairplot.html