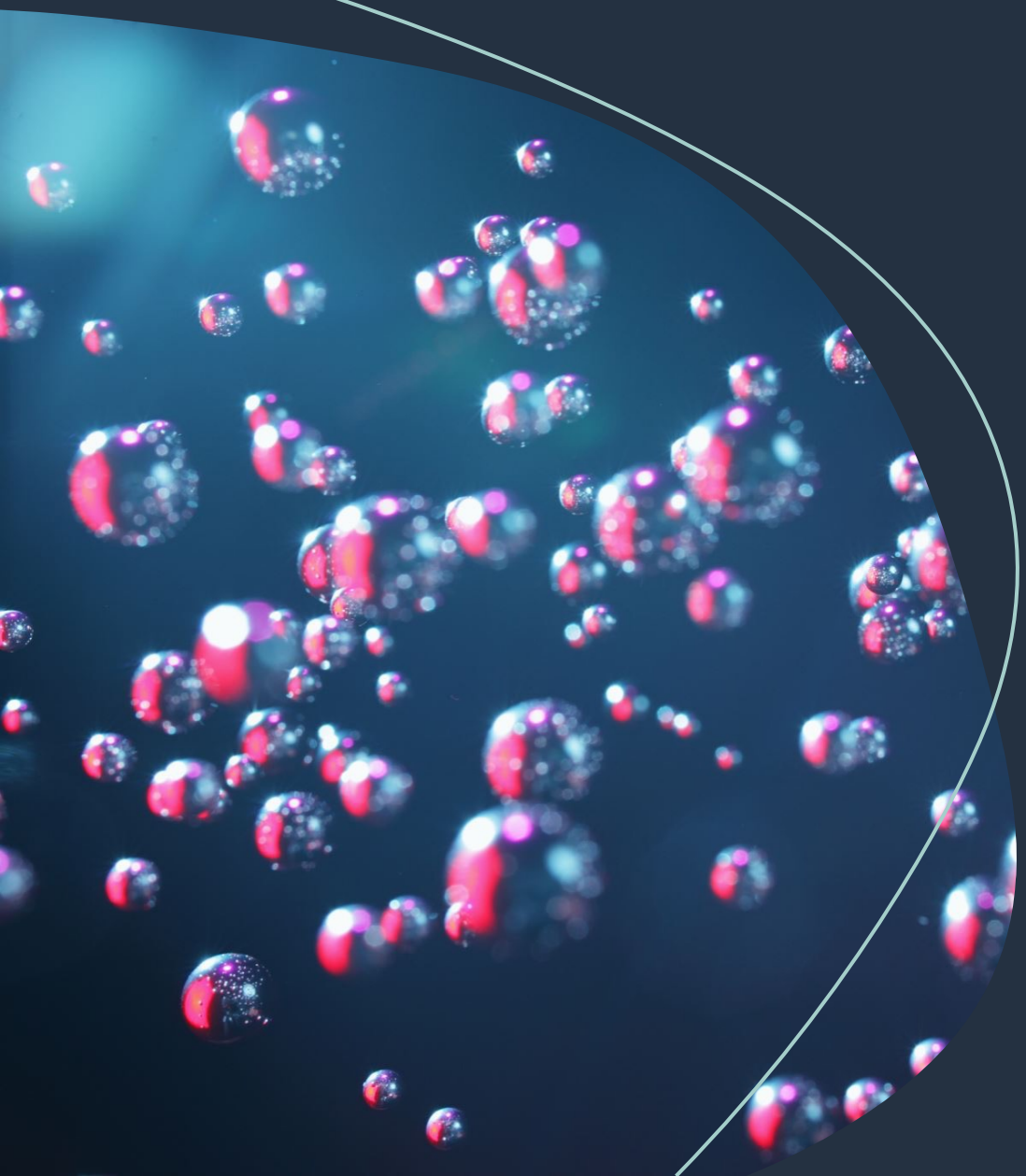


Health Outcomes and Machine Learning: Heart Disease

Clay, Dalit, Patrali

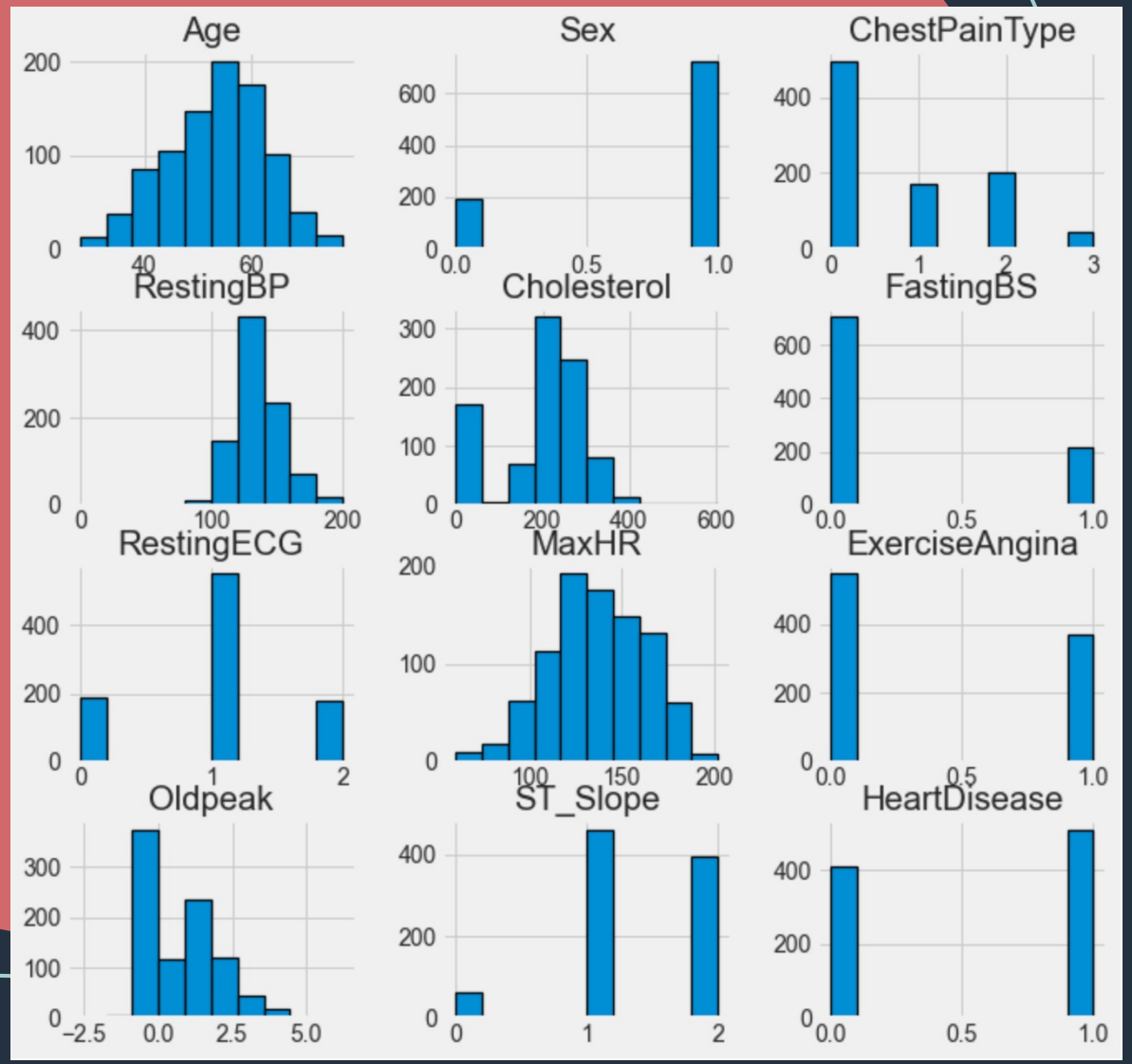
Data 201B Final Project

Introduction:
Heart Disease Background Information

- Heart disease: catchall term for several heart conditions
  - Often reduces blood flow to the heart – which can cause a heart attack
- Risk factors: diabetes, obesity, unhealthy diet, physical inactivity, and excess alcohol usage
- Leading cause of death in the United States - avg. 1 of every 4
  - "One person dies every 36 seconds in the United States from cardiovascular disease"- CDC
- Costs the U.S. roughly half $363 billion a year
  - Due to: health care services, medicines, and lost productivity due to death
- Our question: **What factors/features are the most important when it comes to predicting heart disease?**

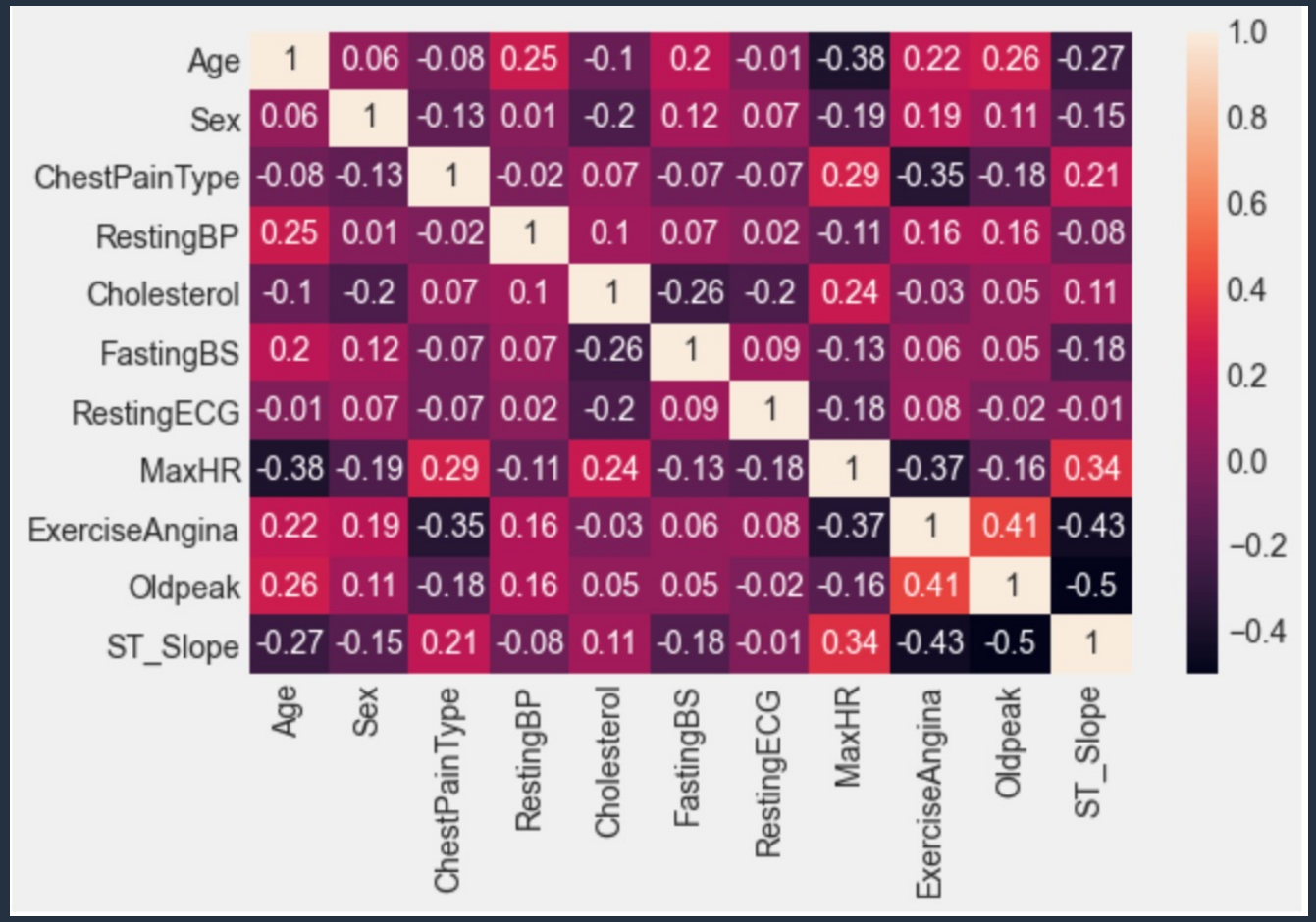
All heart disease facts were taken from the CDC website: https://www.cdc.gov/heartdisease/facts.htm

Data Summary:



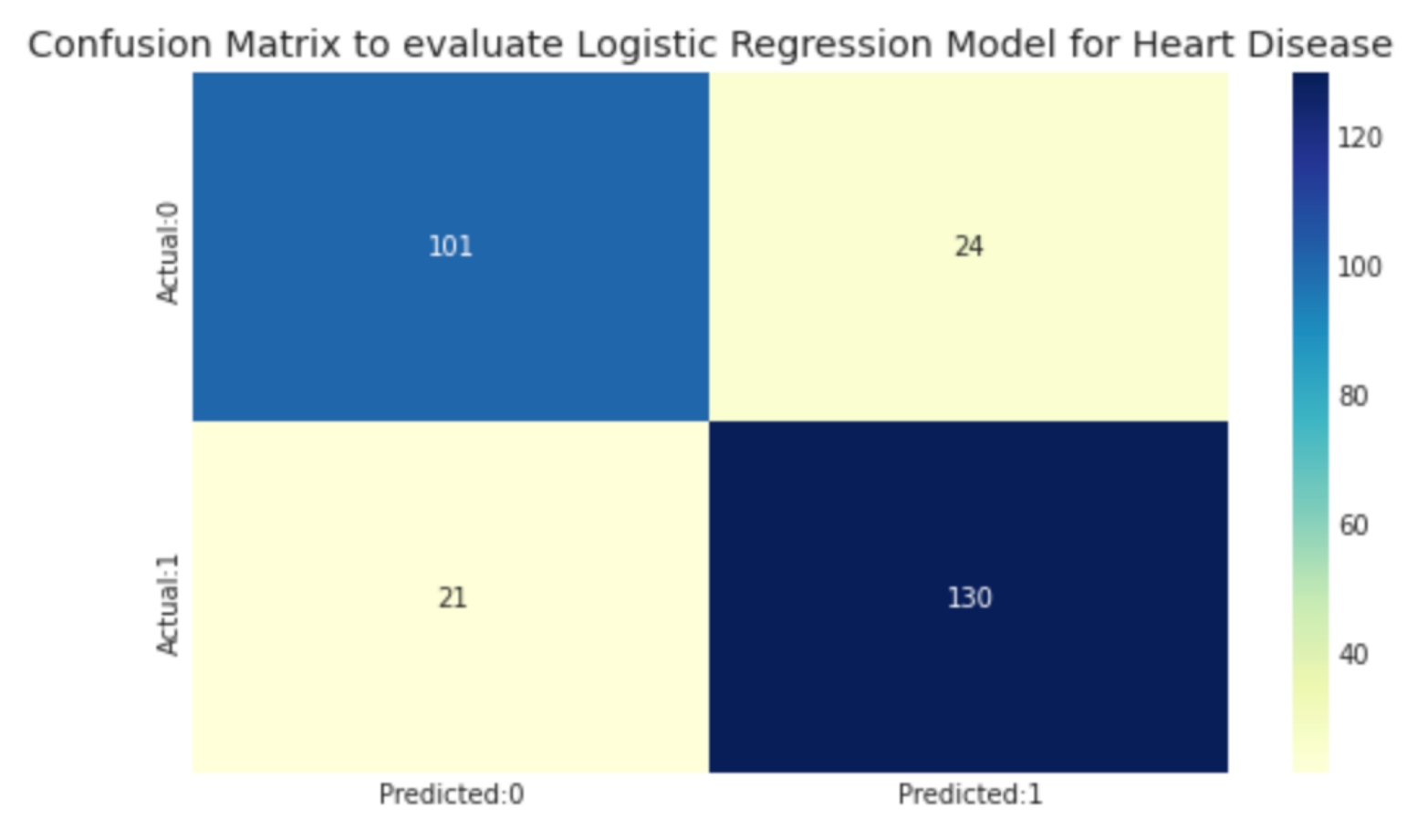
- 12 variables
  - 11 we will use to predict Heart Disease
- 5 variables are categorical were label encoded
  - Sex, ChestPainType, ExerciseAngina, RestingECG,ST_Slope
- 5 datasets from UCI Machine Learning Repository, amalgamated, and uploaded to Kaggle on 2021-09-10
- 918 rows 12 columns
- Heart Disease – categorical variable
  - 1 == heart disease present with 508 values
  - 0 == no heart disease present with 410 values

Heat Map



- Only moderate correlation
  - Oldpeak = ST [Numeric value measured in depression]
  - ST_slope = the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]

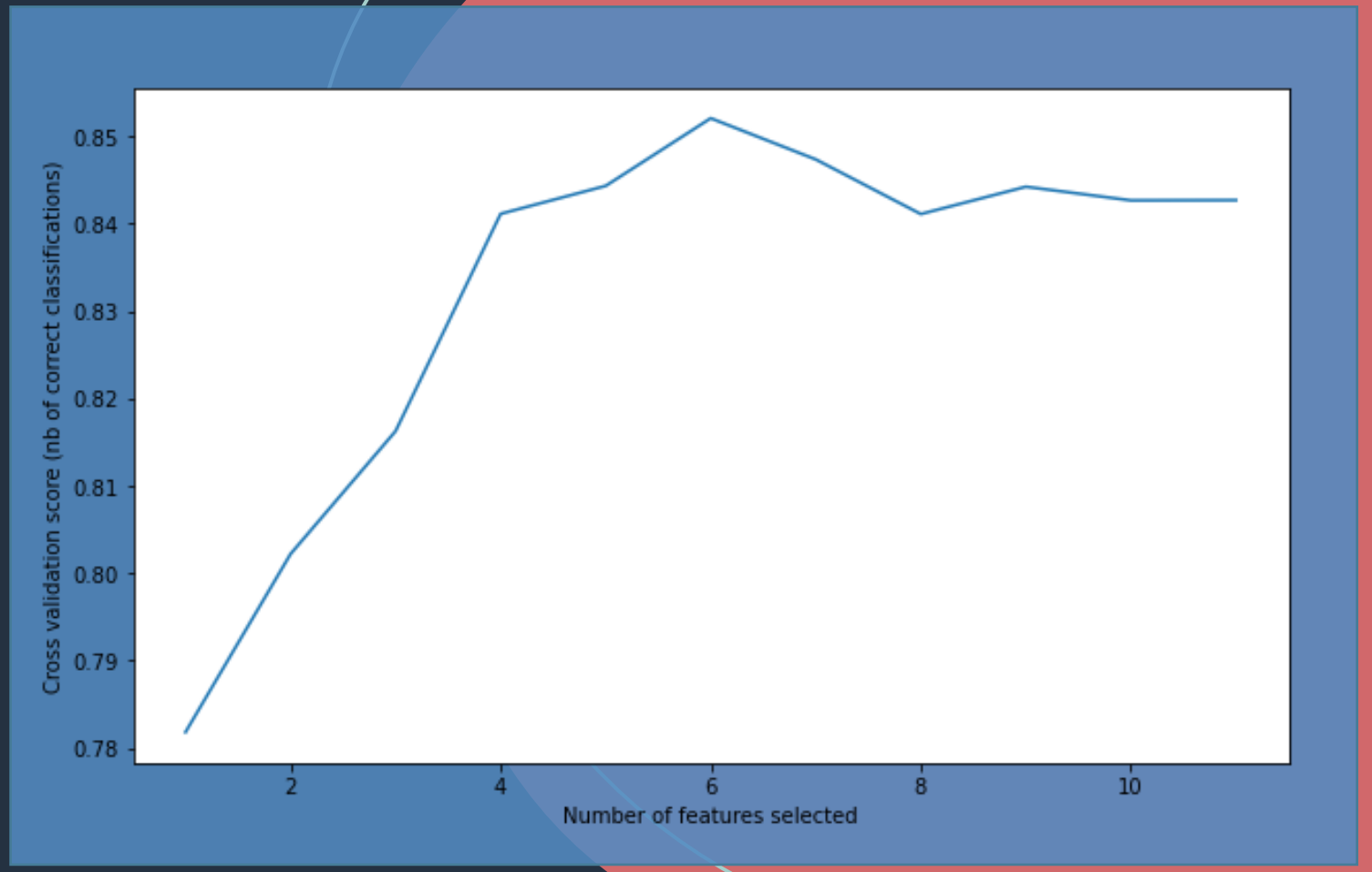
Logistic Regression



- Accuracy is 83.7%

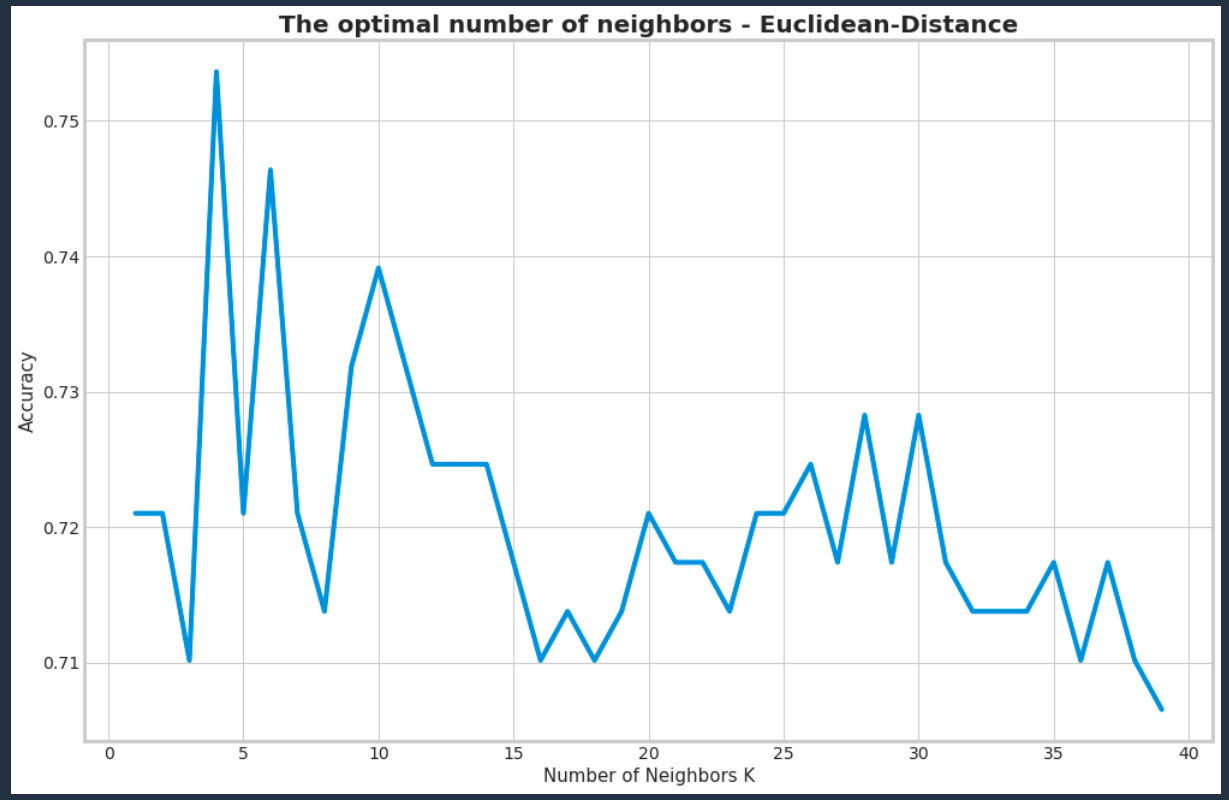
Cross Validation

- Optimal number of features selected are 6 :
- 'Sex', 'ChestPainType', 'FastingBS', 'ExerciseAngina', 'Oldpeak', 'ST_Slope'

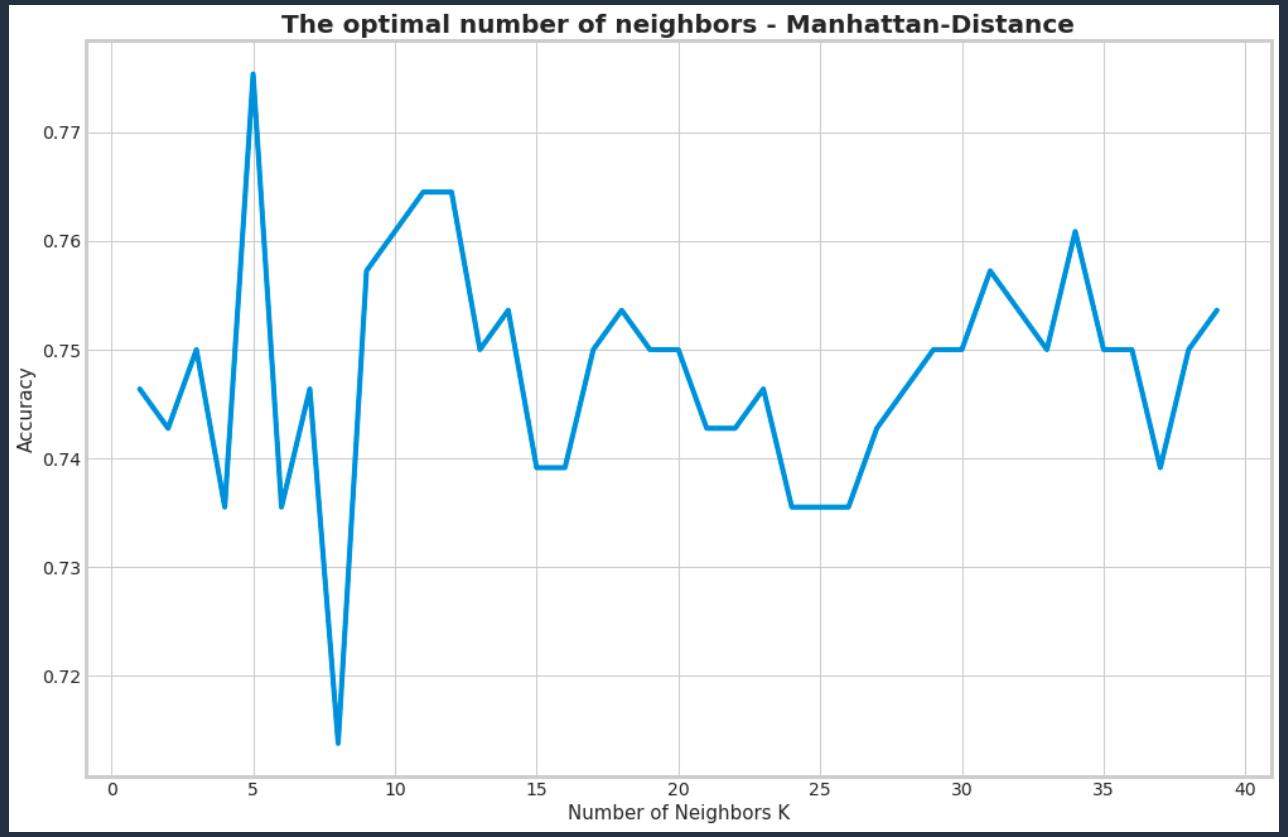


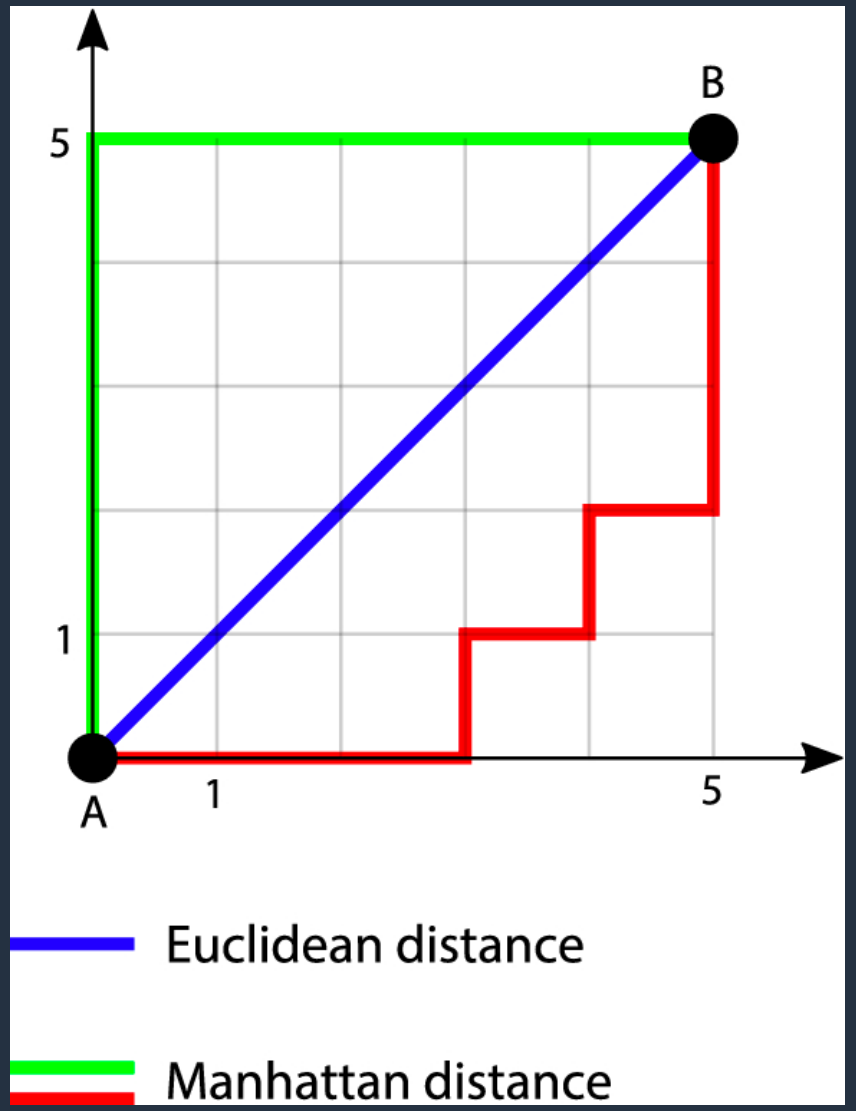
KNN Results

- Best model found with 4 Neighbors (accuracy: 75.3%)
  - Accuracy Results (Num Neighbors 1-39)



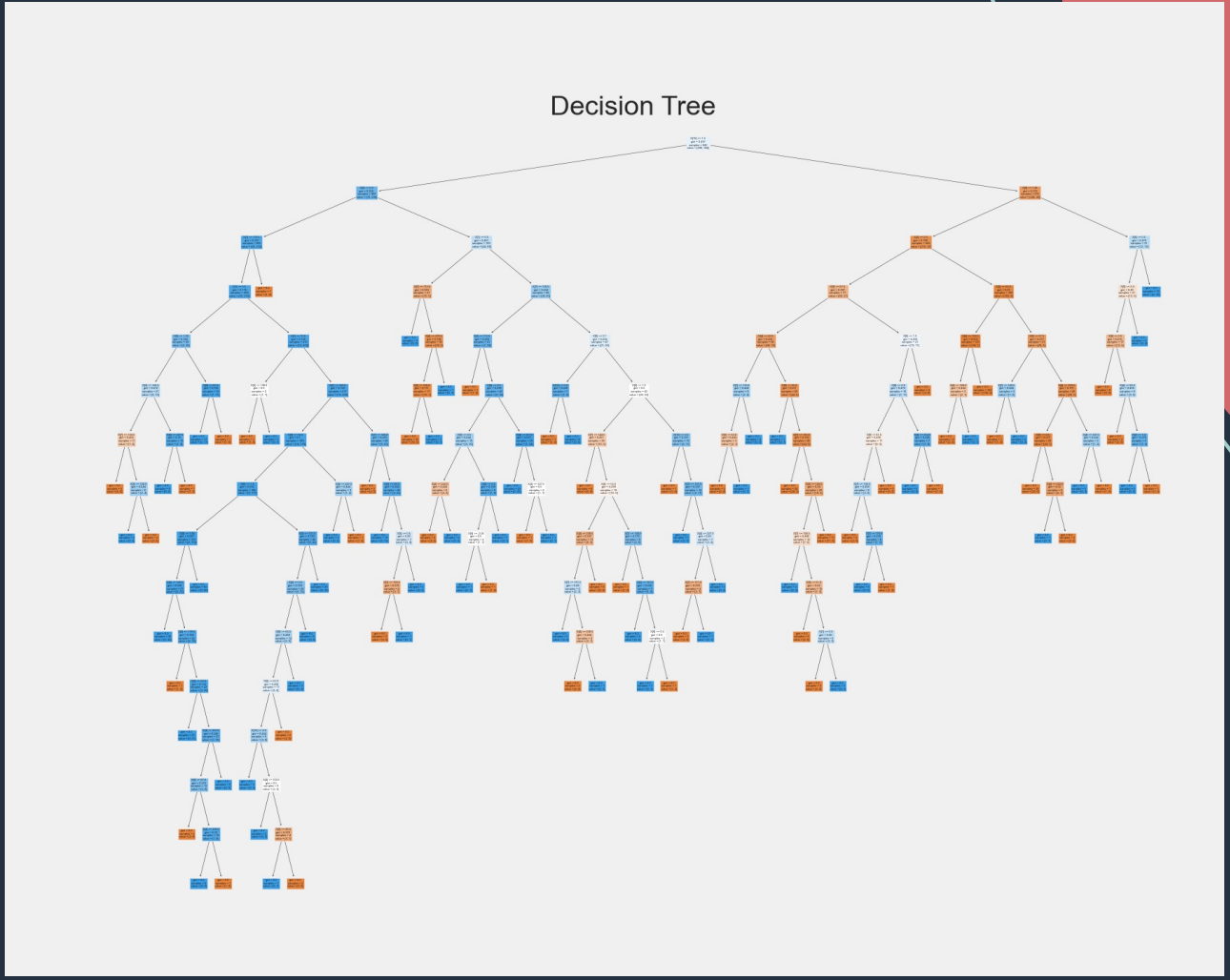
KNN – Different Distance Calculation

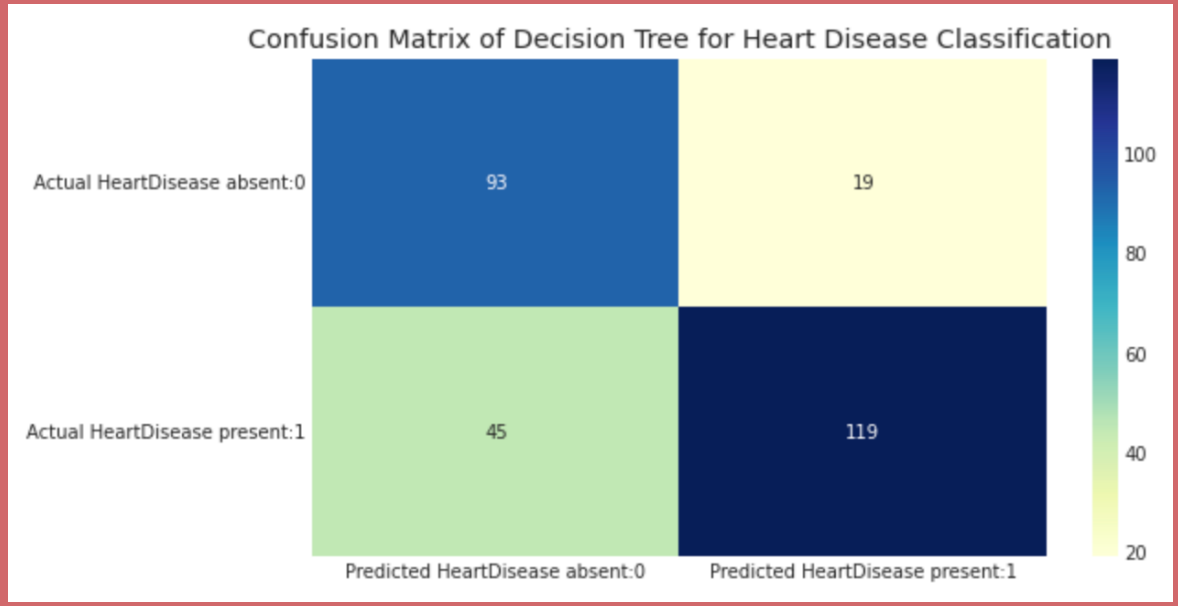




The Manhattan-distance based KNN model performed 2.335% better than Euclidean (77%)

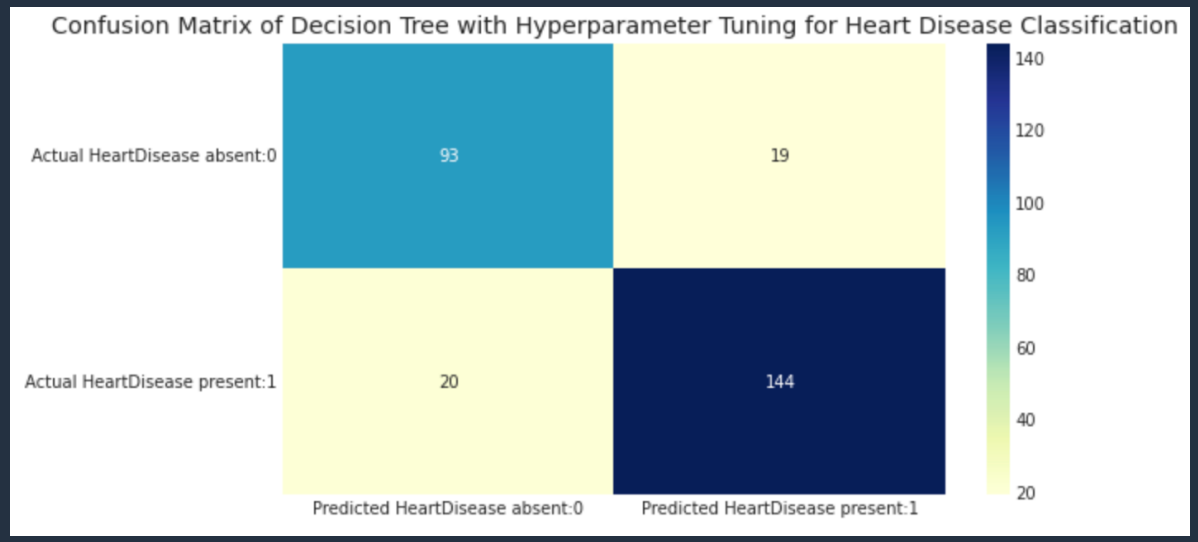
Decision Tree

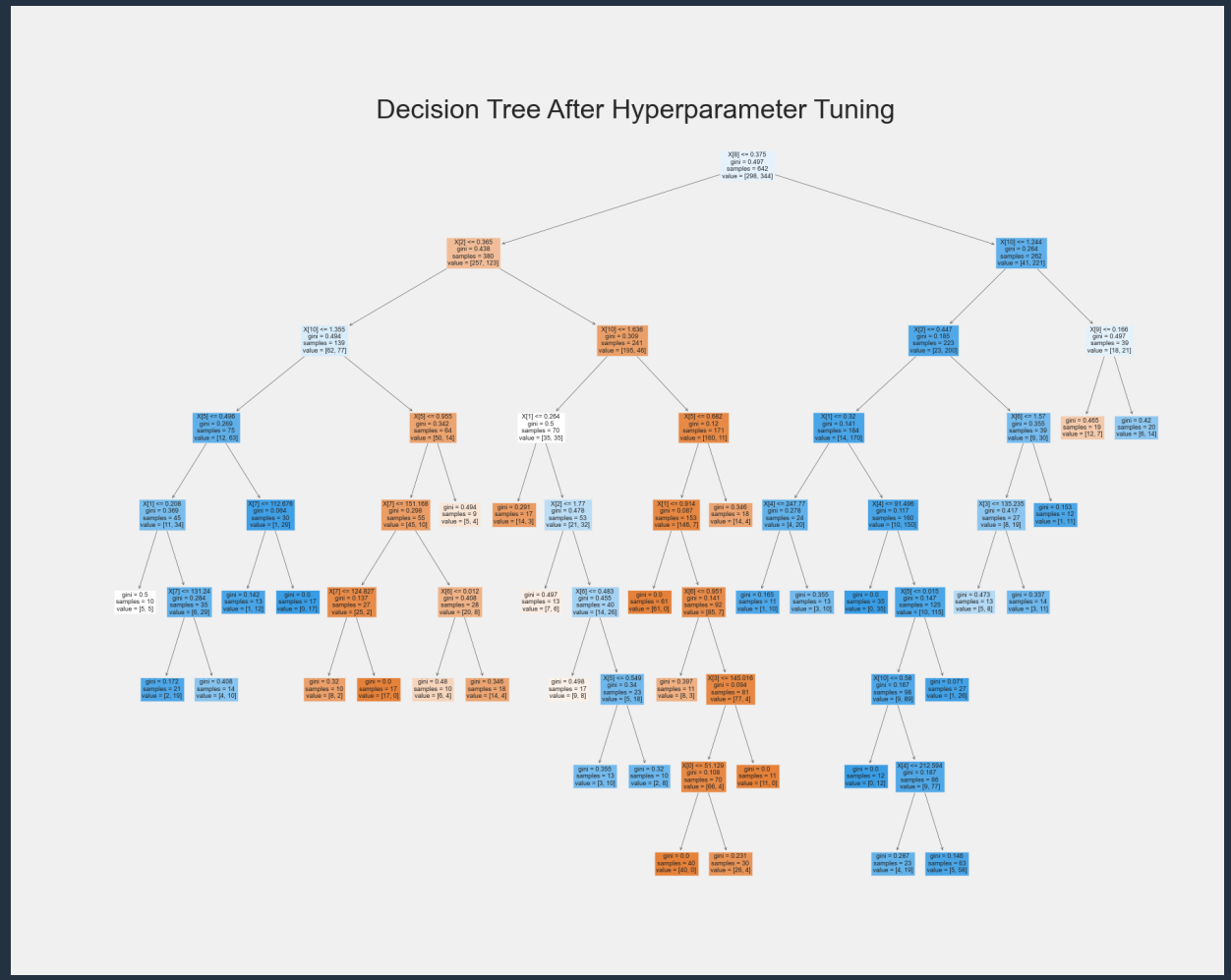




Accuracy score: 76.8%

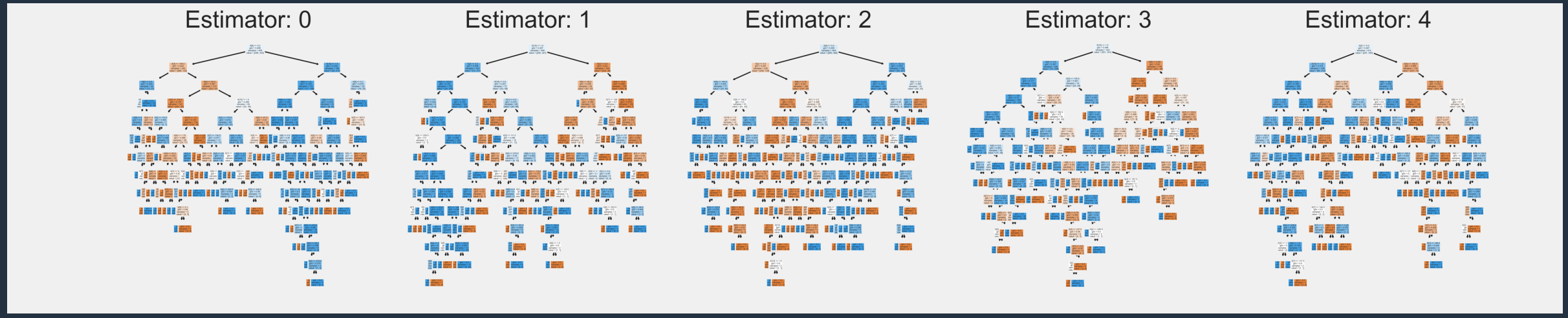
Decision Tree with Hyperparameter Tuning





Accuracy score: 85.8%

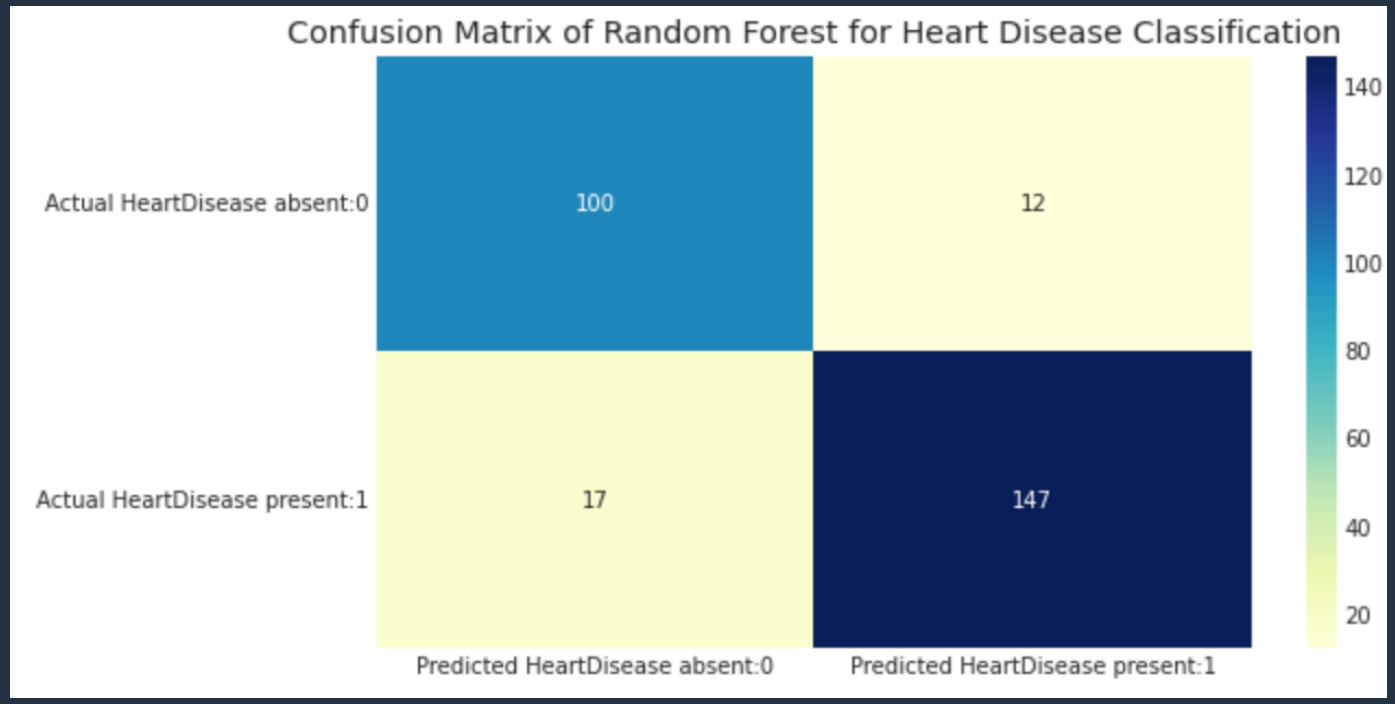
0 less Type 1 error
25 less Type 2 error



Random Forest

Better than DT:
7 less Type 1 error
28 less Type 2 error

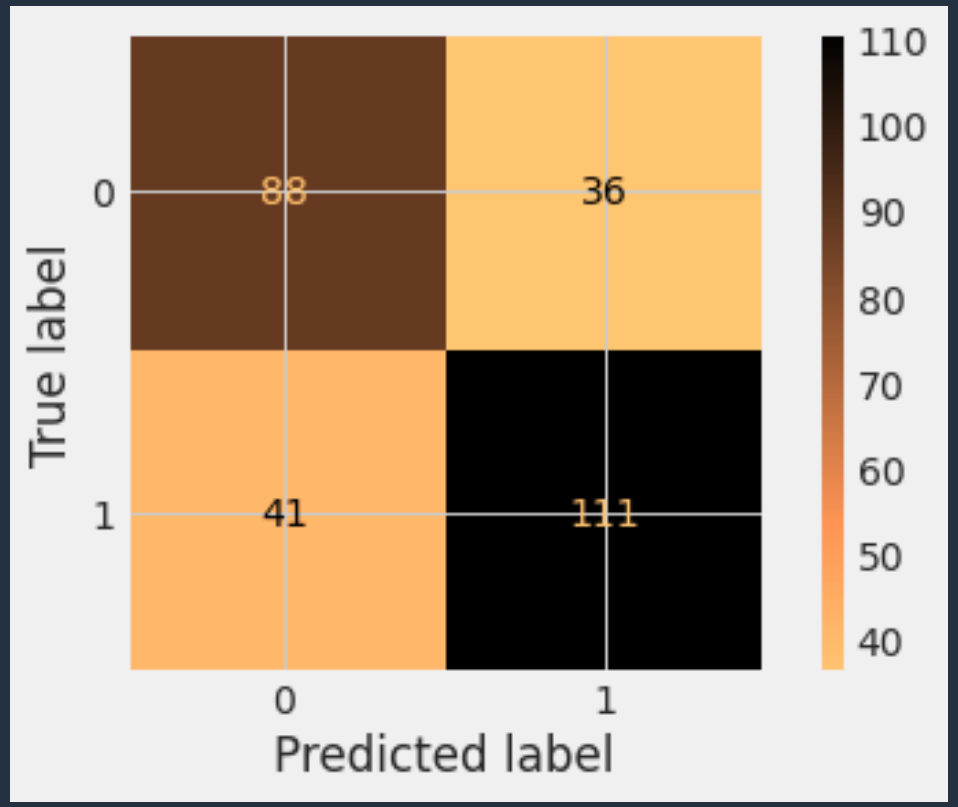
Better than DT w/ HT:
7 less Type 1 error
3 less Type 2 error



Accuracy: 89.4%

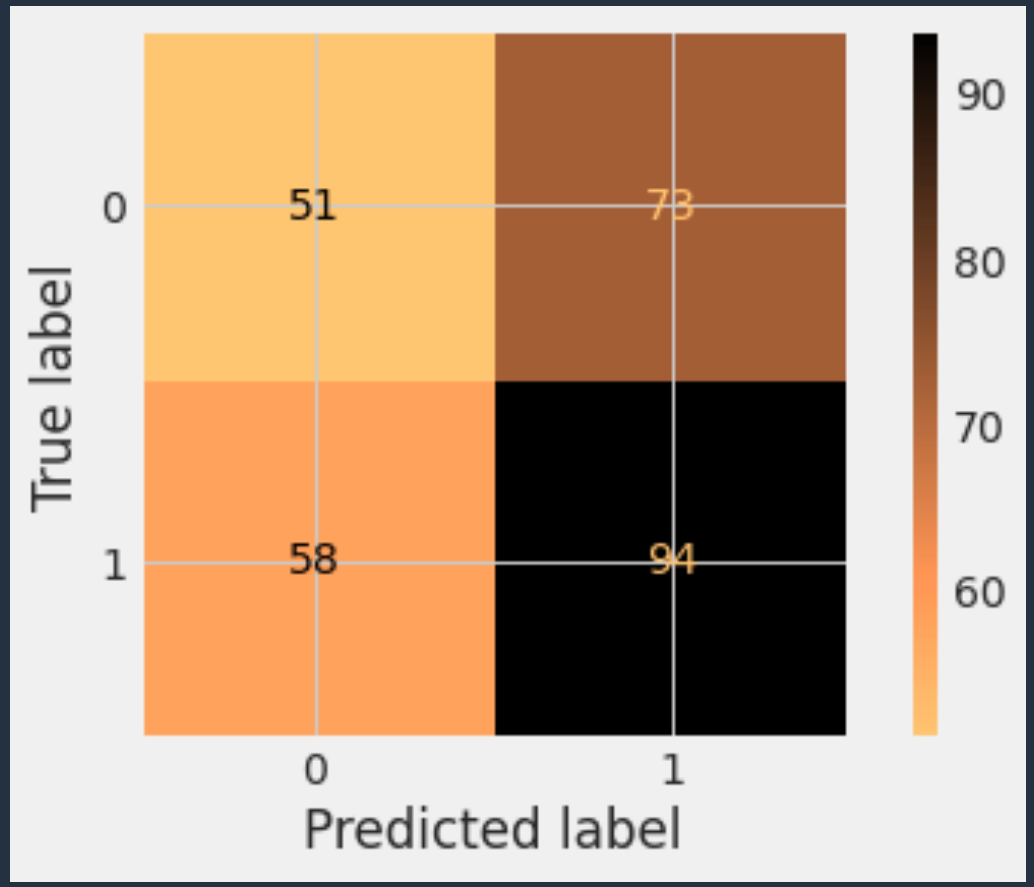
SVM Results

Radial Basis Kernel Function



Accuracy: 72%

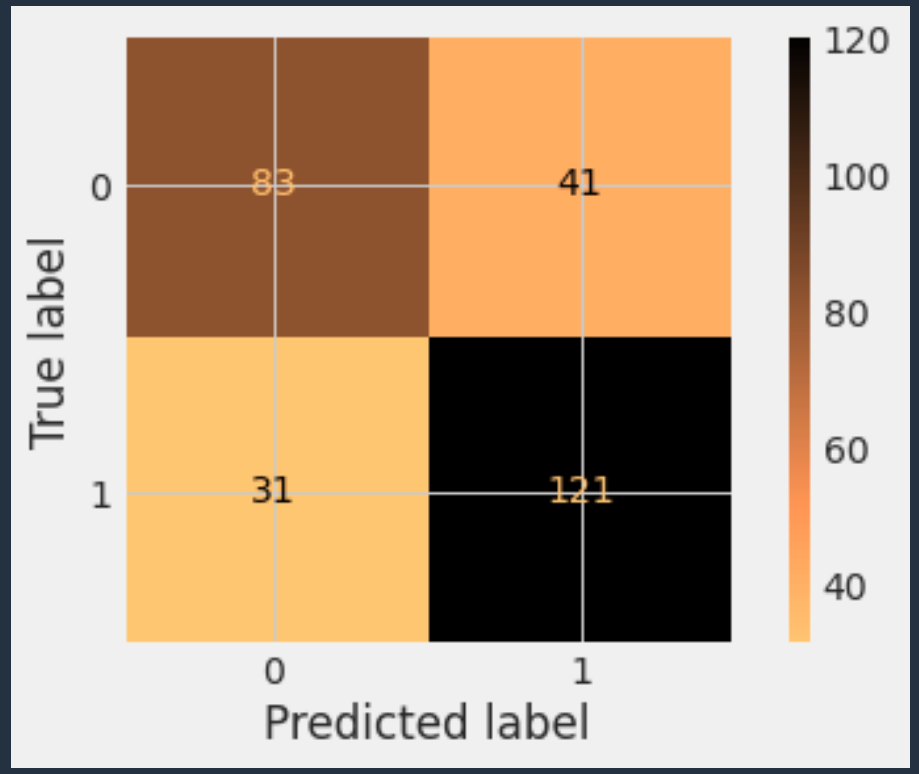
Sigmoid Kernel Function



Accuracy: 53%

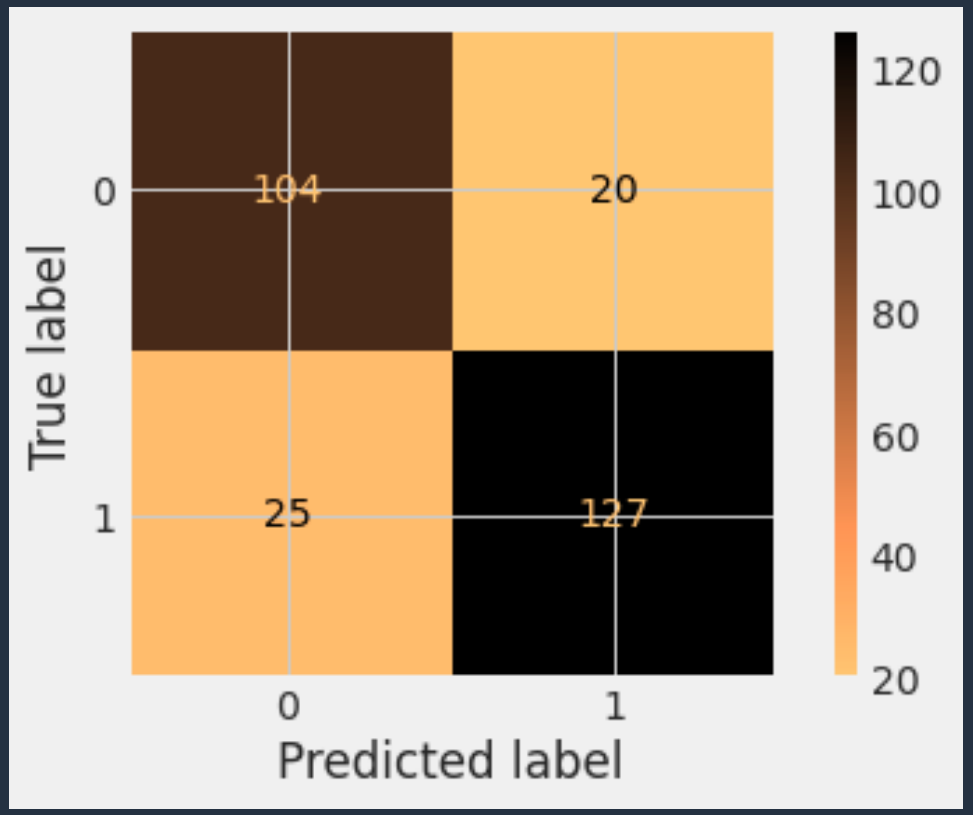
SVM Results (pt. 2)

Polynomial Kernel Function



Accuracy: 74%

Linear Kernel Function



Accuracy: 84%

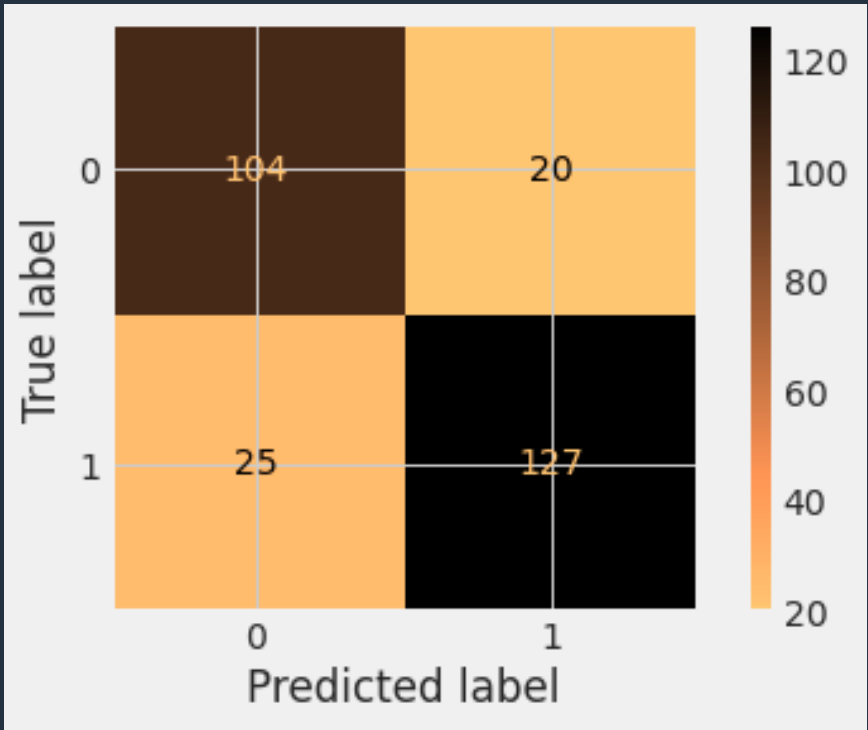
Question: Does it make sense that a linear function would outperform a polynomial one?

Linear > Polynomial

$$h(x) = mx^1 + bx^0$$

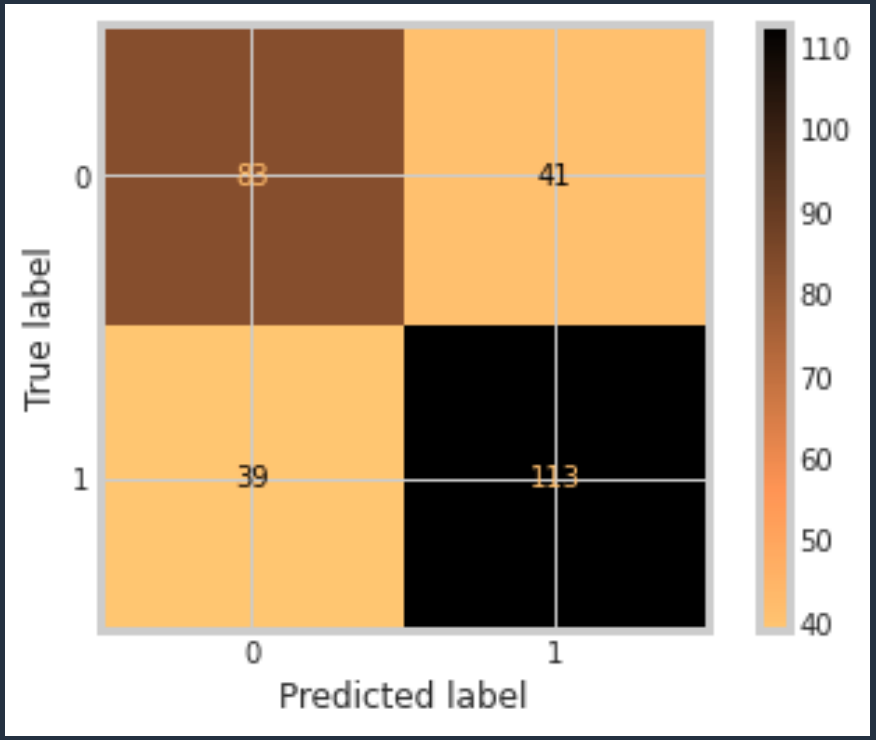
- Linear model (best performing) accuracy almost 10% better than our polynomial

Linear



Accuracy: 84%

svm.SVC(kernel = 'poly', degree=1)

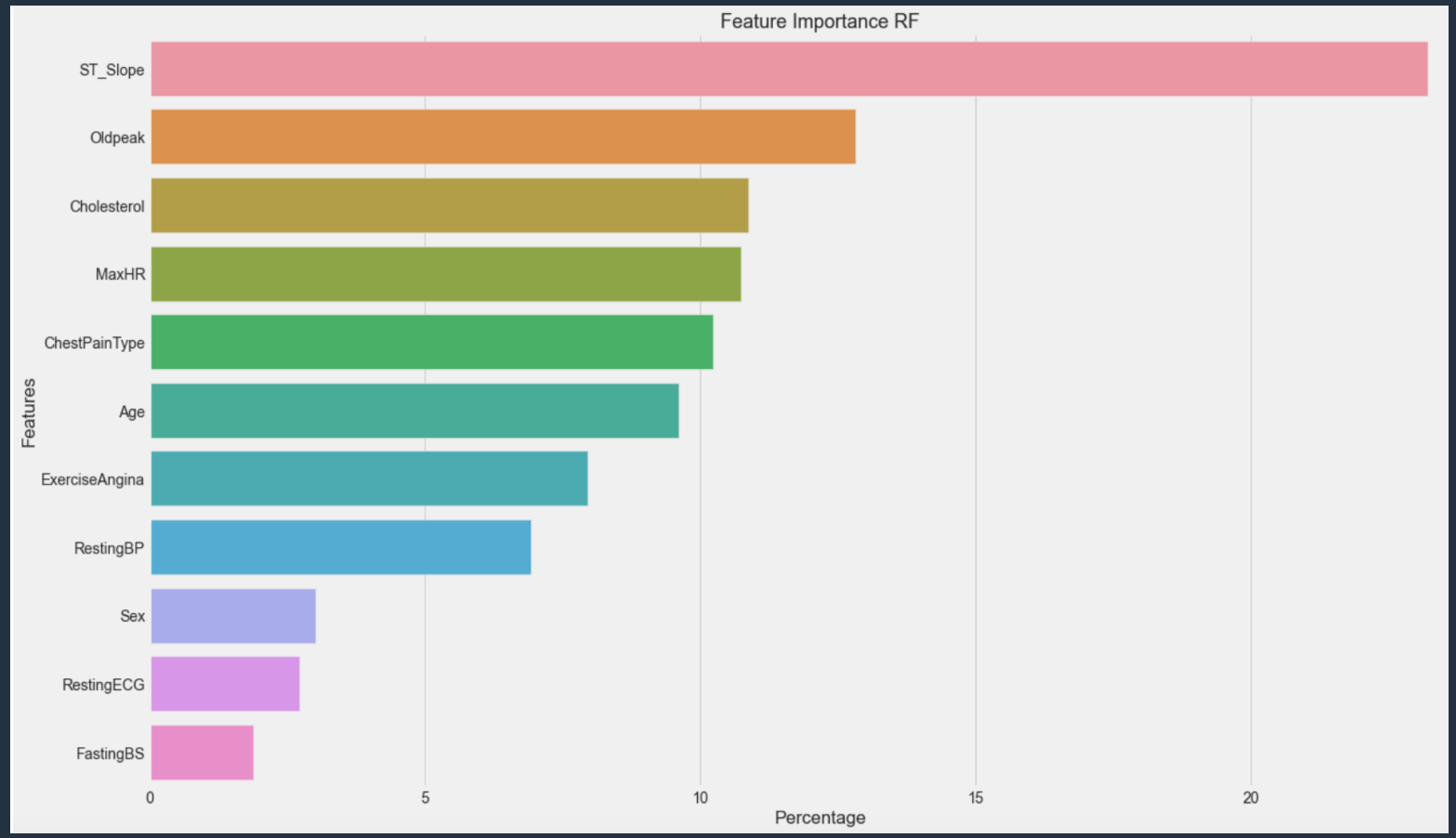


Accuracy: 71%

Which Model Performed the Best?

- Model accuracies:
  - Logistic Regression: 83.7%
  - KNN:
    - Euclidean Distance: 75.3%
    - Manhattan Distance: 77.5%
  - Decision Tree: 76.8%
  - Decision Tree w/ Hyperparameter Tuning: 85.8%
  - Random Forest: 89.4%
  - SVM (Linear Function): 84%

Feature Importance for Best Performing Model: Random Forest



Feature	Percentage
ST_Slope	23.202805
Oldpeak	12.814285
Cholesterol	10.882240
MaxHR	10.737578
ChestPainType	10.237212
Age	9.619771
ExerciseAngina	7.955635
RestingBP	6.930208
Sex	3.015218
RestingECG	2.720309
FastingBS	1.884739

Future Work

- Continue collecting de-identified patient data
- Scaling patient data to more input features
  - Build-up of calcium in a major artery outside of the heart "could **predict future heart attack or stroke" -** Edith Cowan University (2021)
- Important features are crucial to monitor on patients for preventative care
  - Particularly ST_Slope
- General tracking for patients

Questions?

Thank you

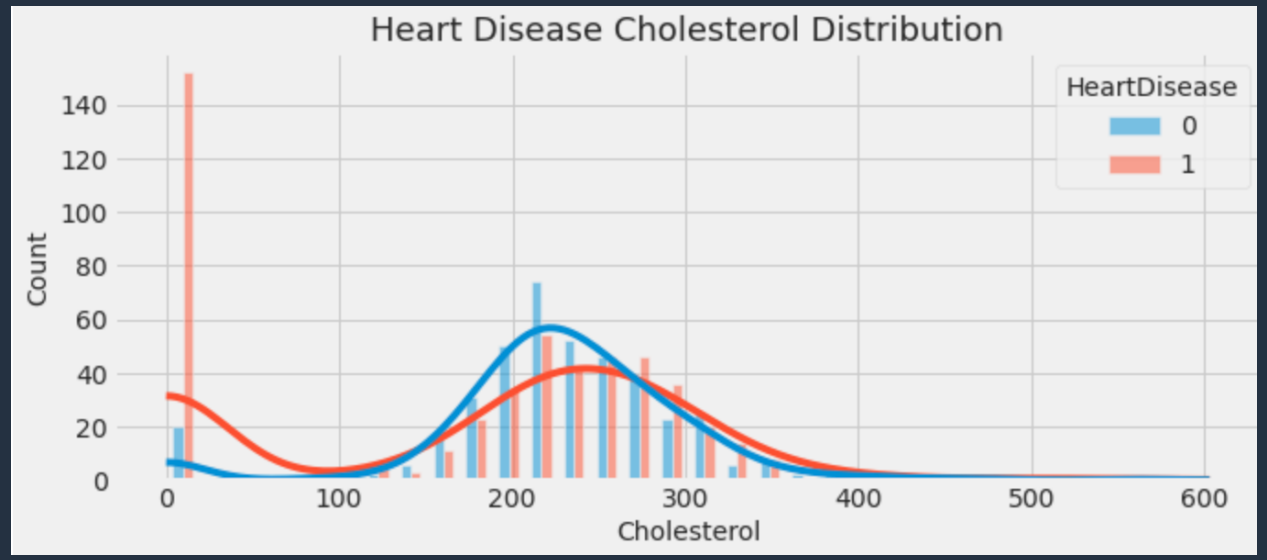
Appendix

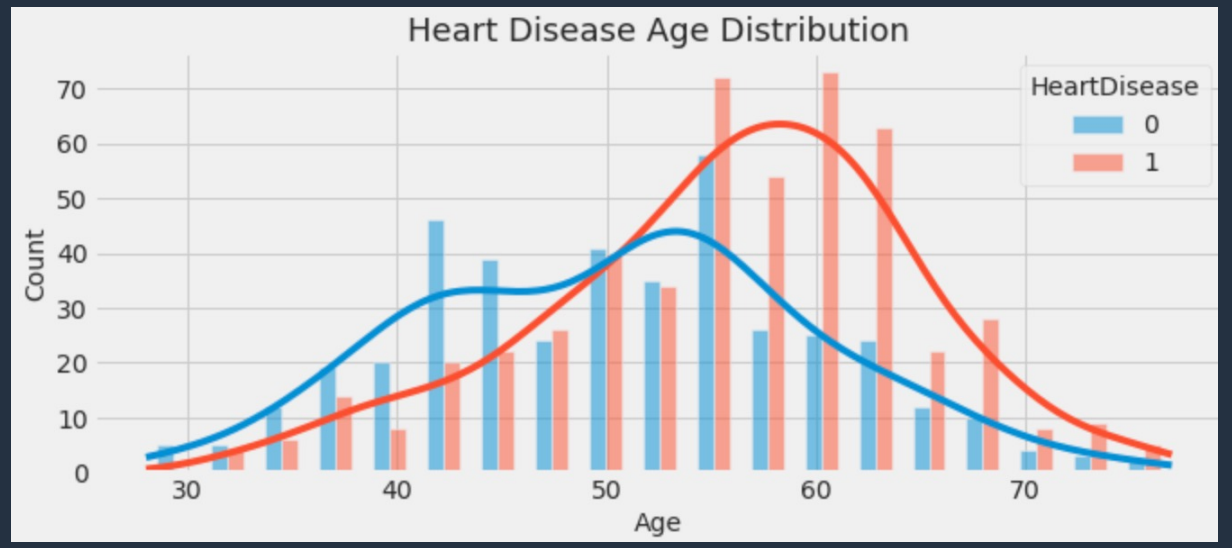
Attributes

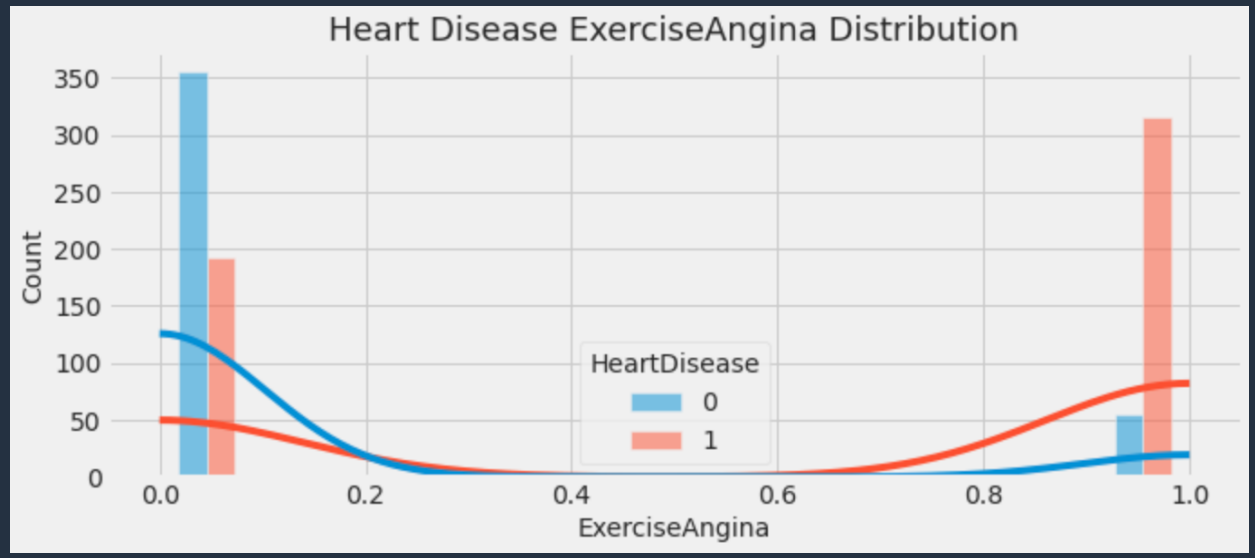
- Age: age of the patient [years]
- Sex: sex of the patient [M: Male, F: Female]
- ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- RestingBP: resting blood pressure [mm Hg]
- Cholesterol: serum cholesterol [mm/dl]
- FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
- ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
- Oldpeak: oldpeak = ST [Numeric value measured in depression]
- ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
- HeartDisease: output class [1: heart disease, 0: Normal]

(Fedesoriano 2021)

A closer look:







Models Used & Why

Image Source: Power point image catalogue

Why Logistic Regression?

- Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis.
- PRO: Simple to understand, easy to implement, and efficient to train
- CON: Sensitive to outliers

Logistic Regression Results

	precision	recall	f1-score	support
0	0.83	0.81	0.82	125
1	0.84	0.86	0.85	151
accuracy			0.84	276
macro avg	0.84	0.83	0.84	276
weighted avg	0.84	0.84	0.84	276

Image Source:
Power point image catalogue

Why a kNN Model?

- Can be used for both binary and multi-class classification
- Simple model with no training step
  - Uses training data to classify test/future data
- Sensitive to Outliers
  - Plethora of causes for Heart Disease (some that we don't factor in our model at all)
  - Differently scaled **future/testing** data impacts model greatly – or null values
- Results on next page!

Why DT & RF Models?

- Decision Tree
  - Intuitive supervised learning classification problem
  - Goal is to infer class labels
    - 1 = heart disease & 0 = normal
  - PRO: Resistant to outliers
  - CON: Prone to overfitting (high variance)
    - Biased if one class if dominating
- Random Forest
  - Ensemble method of many trees improves performance (with random sampling of features) and aggregates results
  - Parameter tuning not necessary
  - Pro: Overfitting is generally not a problem
  - Con: Becomes complex when there are many class variables

Image Source: Power point image catalogue

Image Source: Power point image catalogue

Why a Support-Vector Machine Model?

- We have a relatively small number of features compared to observations
  - Plot training data on an 11-dimensional hyperplane (# features)
- A con of SVM is that it can underperform on highly skewed/imbalanced data sets
  - Num samples in one class outweigh num samples in other
  - Less applicable to our model: heart disease value count (508 T and 410 F)

Works Cited

- Centers for Disease Control and Prevention. (2021, September 27). *About heart disease*. Centers for Disease Control and Prevention. Retrieved December 14, 2021, from https://www.cdc.gov/heartdisease/about.htm
- Fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [2021-11-11] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.
- "New Research Reveals Early Warning Sign for Heart Disease." *EurekAlert!*, https://www.eurekalert.org/news-releases/761185.

- Acknowledgements
  - Creators:
    - Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
    - University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
    - University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
    - V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.