Heart Disease Prediction

Group 2: Cassidy Kratwell, Caleb O’kane, Amanda Garza

Heart Disease

Heart disease is a term used for conditions that affect the heart's overall structure and function. Also referred to as Cardiovascular disease, it is currently the leading cause of death throughout the United States and globally. There are several problems that are related to heart disease, the most common of these being heart attacks, strokes, heart failure, or arrhythmia to name a few. There are several different types of treatments for these issues such as medication, surgery, and lifestyle changes. For our project we will look to analyze a set of data that was collected to see if we can predict if they have heart disease or not based on the various health conditions presented.

Factors that affect the Heart

- High blood pressure
- High cholesterol
- Smoking
- Unhealthy diet
- Physical inactivity

Data

From Kaggle:

https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data

Variables:

- Age
- Sex
- Cp-chest pain type
- Trestbps- resting blood pressure
- Chol-cholesterol level
- Fbs-fasting blood sugar

Variables Defined

Avg Chol - Average Cholesterol

Avg Cp - Average Chest Pain. Chest pains range from 0 to 3 with 0 being low chest pain and 3 being severe chest pain.

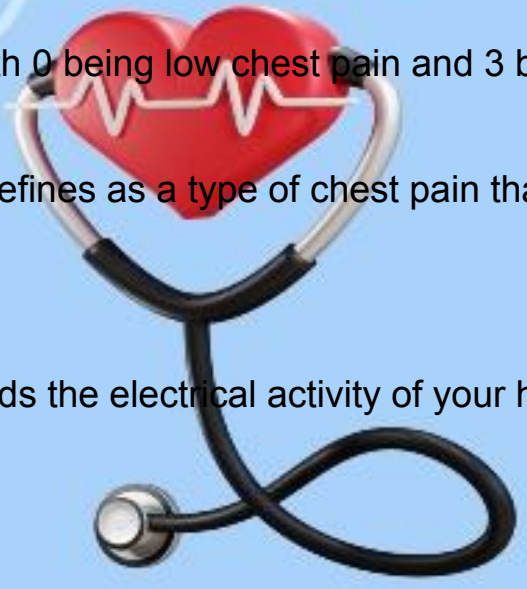
Avg Exang - Average Exercise Induced Angina. Which further defines as a type of chest pain that occurs when the heart needs more oxygen than usual during exercise.

Avg Fbs - Fasting Blood Sugar

Avg Restecg - resting electrocardiographic results. Which records the electrical activity of your heart when you are at rest.

Avg Thalach - Maximum Heart Rate Achieved.

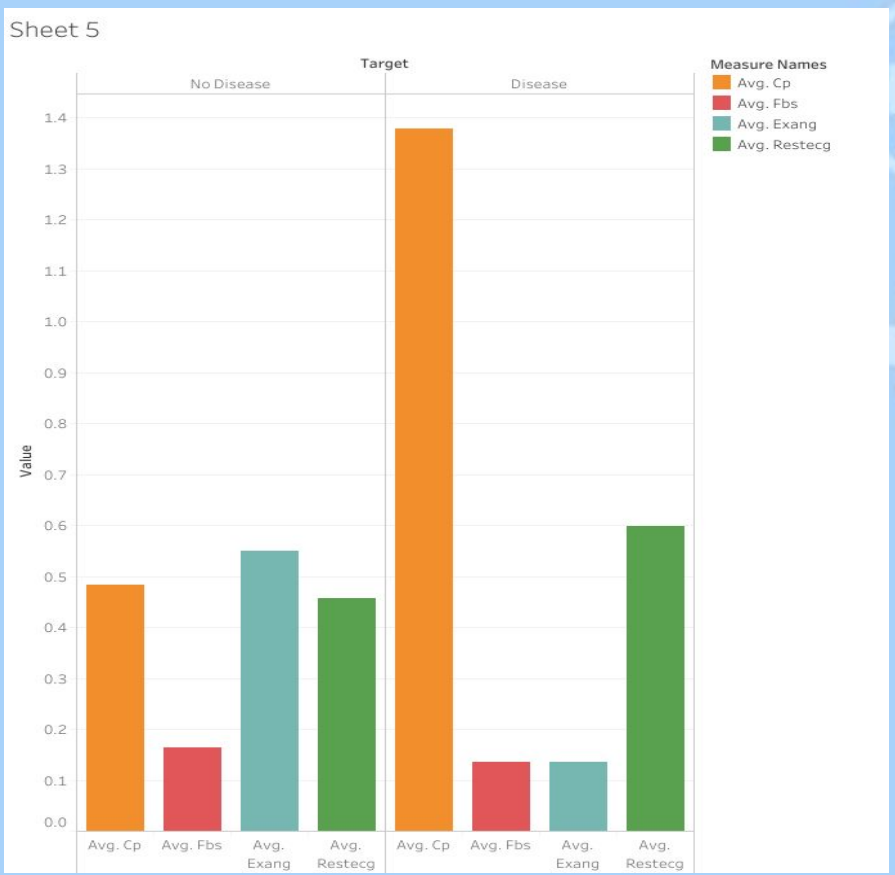
Avg Trestbps - Resting Blood Pressure.

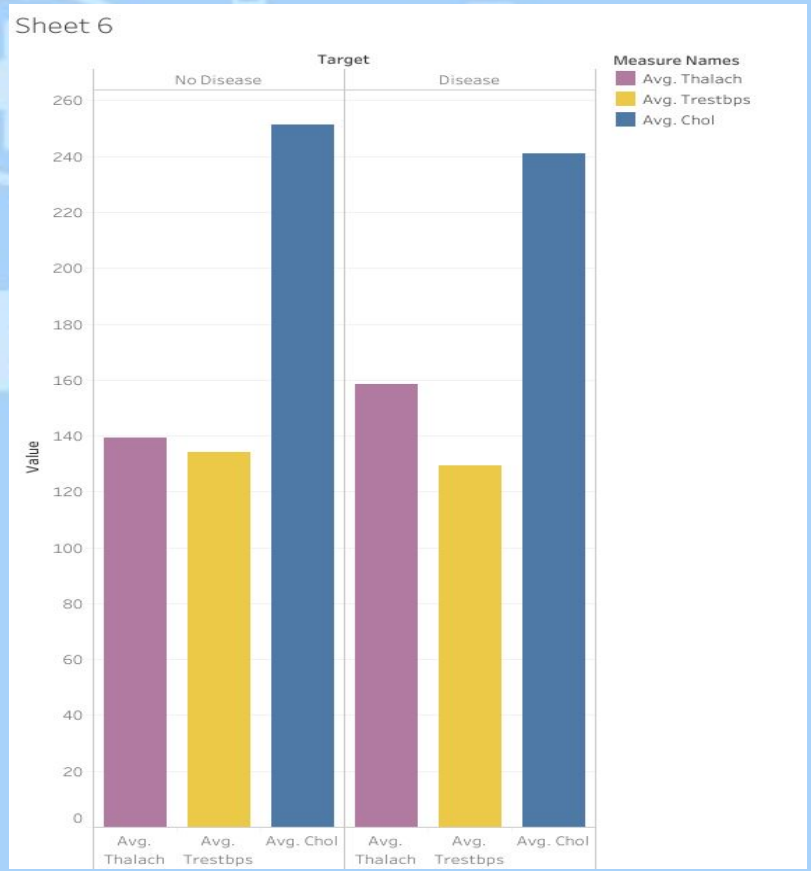


Age/Sex Comparison of People With and Without Heart Disease

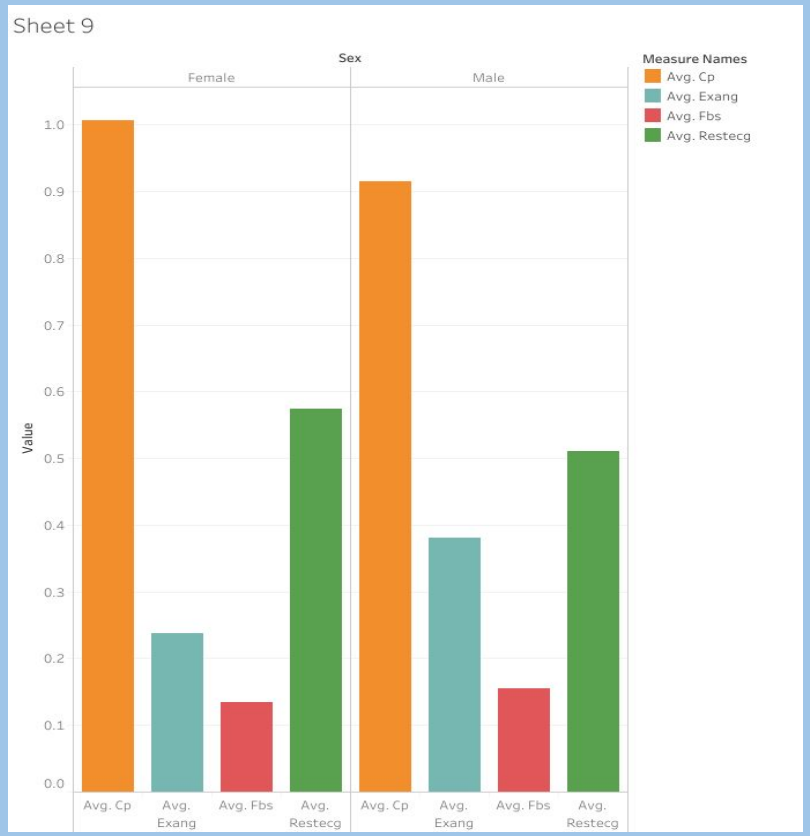
Sex	Target	Avg. Chol	Avg. Cp	Avg. Exang	Avg. Fbs	Avg. Restecg	Avg. Thalach	Avg. Trestbps
Female	No Disease	276.7	0.2	0.6	0.3	0.5	142.3	146.5
	Disease	255.6	1.3	0.1	0.1	0.6	154.1	128.8
Male	No Disease	246.0	0.5	0.5	0.1	0.4	138.5	131.5
	Disease	229.9	1.4	0.2	0.2	0.6	162.0	129.6

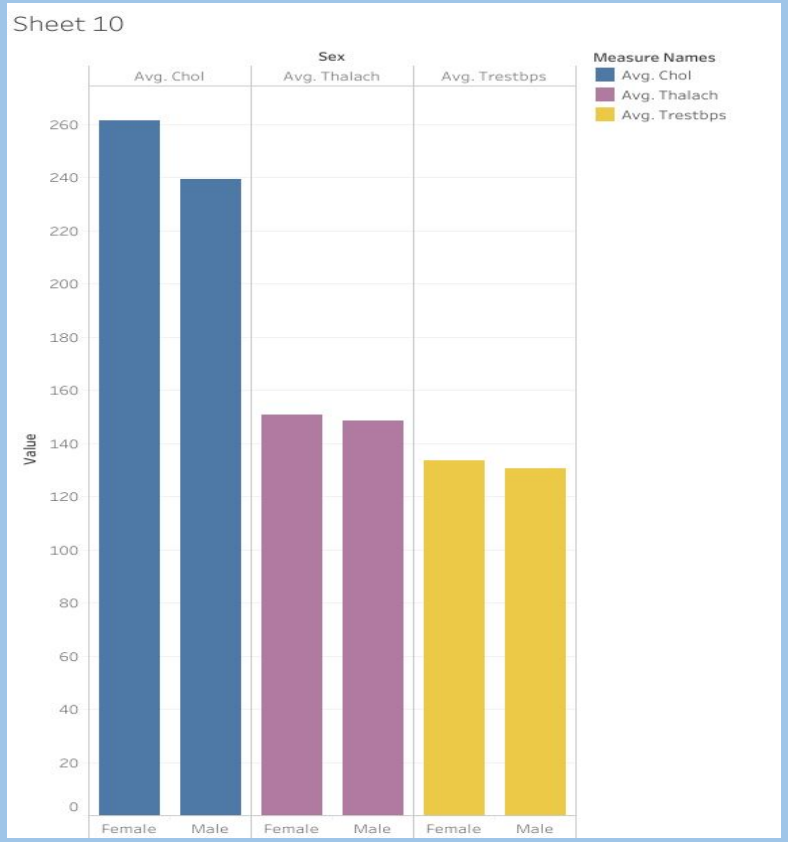
Comparison of Someone with/without Heart Disease



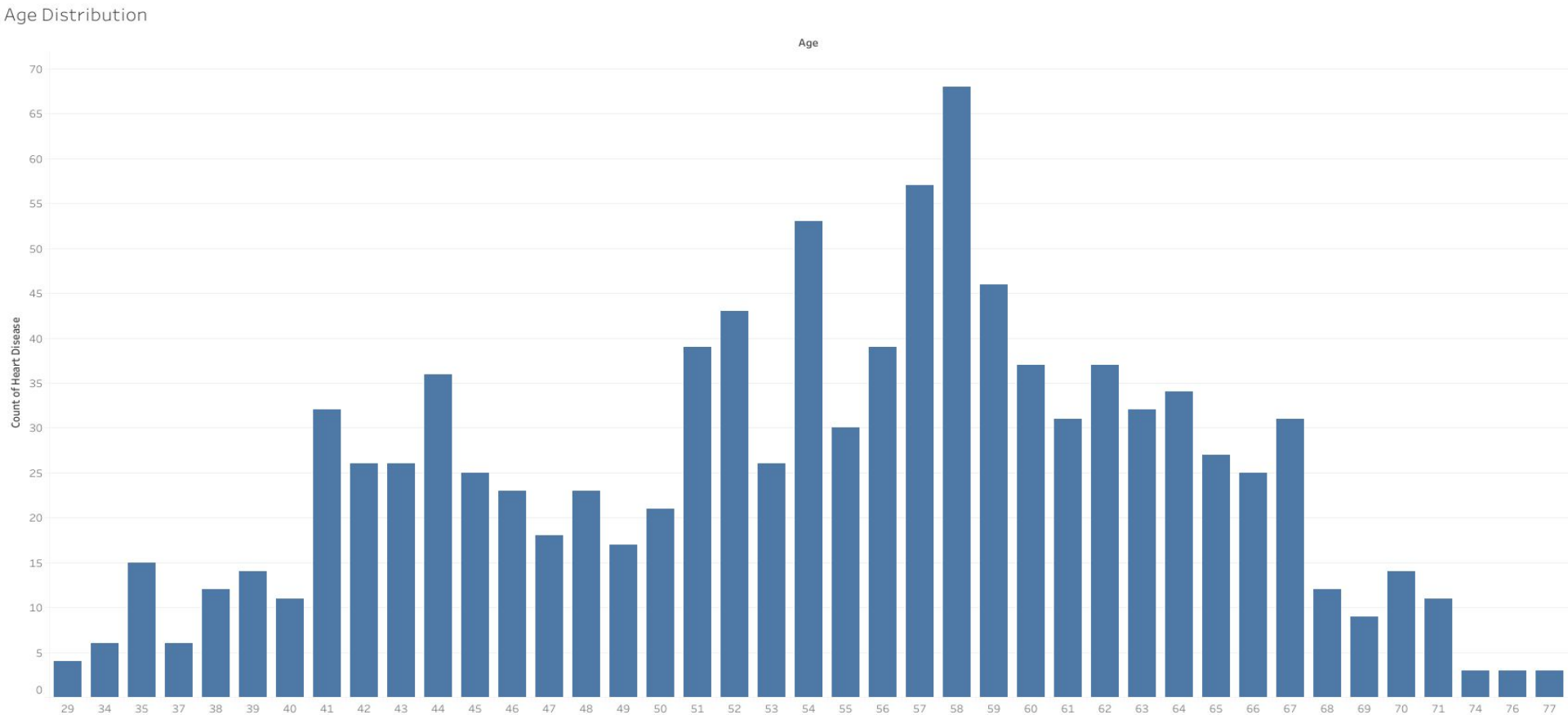


Gender Comparison





Comparison of The Age With The Highest Number of Heart Disease



Machine Learning

Separate the x and y values

```
y = heart_df['target']

X = heart_df.drop(columns=['target'])
```

Split the data into training and testing datasets using train_test_split

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1)
```

```
from sklearn.linear_model import LogisticRegression
logistic_regression_model = LogisticRegression(solver='lbfgs', random_state=1)
lr_model = logistic_regression_model.fit(X_train, y_train)
```

Machine Learning

- Fit a logistic regression model using the testing data
- Saved the predictions on the testing data labels using the testing feature and the fitted model

	Predictions	Actual
49	0	0
525	1	1
119	1	1
629	1	0
186	0	0
...	...	...
1014	1	1
429	1	0
828	1	1
99	1	1
233	1	1

257 rows × 2 columns

Machine Learning

Generated a confusion matrix for the model

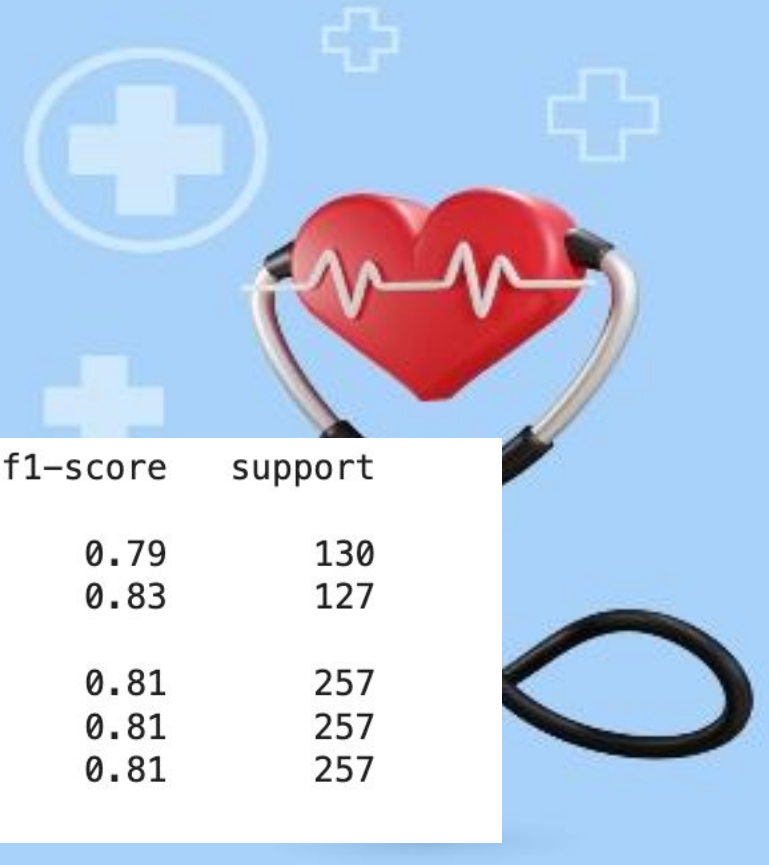
```
cf_test_matrix = confusion_matrix(y_test, test_predictions)
cf_test_matrix
```

```
array([[ 91,  39],
       [ 10, 117]])
```

Machine Learning

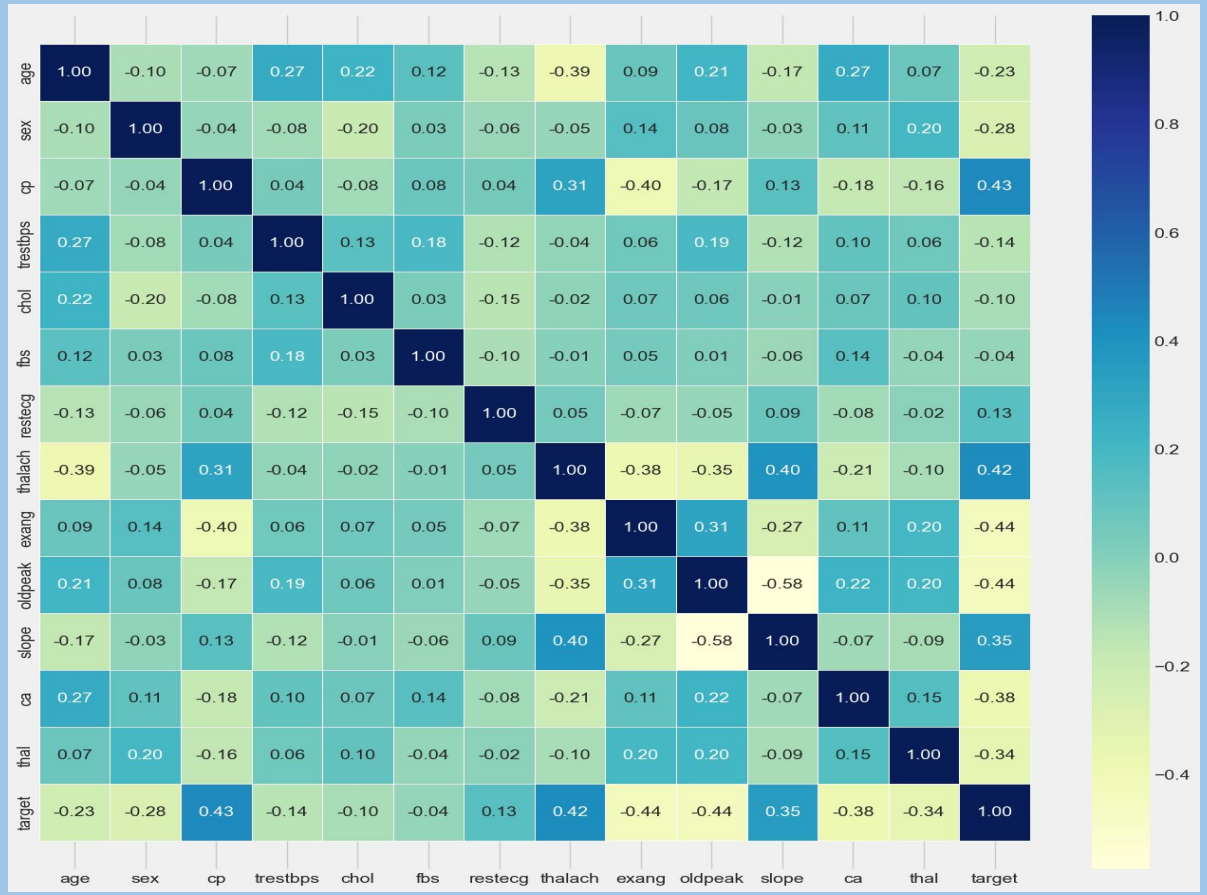
Logistic Regression:

Model demonstrated a predictive power of 83% classification accuracy



	precision	recall	f1-score	support
0	0.90	0.70	0.79	130
1	0.75	0.92	0.83	127
accuracy			0.81	257
macro avg	0.83	0.81	0.81	257
weighted avg	0.83	0.81	0.81	257

Correlation Heatmap



Neural Network

- Preprocessed our data and split into testing and training datasets
- Dropped the exercised induced angina column (exang)
- Used whether they had heart disease or not (target) as our target variable

```
#drop excang column
heart_df=heart_df.drop(columns=['exang'])
```

```
# Split our preprocessed data into our features and target arrays
x =heart_df.drop(columns='target')
y = heart_df['target']

# Split the preprocessed data into a training and testing dataset
x_train, x_test, y_train, y_test = train_test_split(x, y, random_state=15)
```

Neural Network

- Defined the model
- 3 Hidden Layers
- Use Sigmoid as our activation function

```
# Define the model - deep neural net, i.e., the number of input features and hidden nodes for each layer.
number_input_features = len(x_train_scaled[0])
hidden_nodes_layer1 = 15
hidden_nodes_layer2 = 10
hidden_nodes_layer3 = 5
nn = tf.keras.models.Sequential()

# First hidden layer
nn.add(tf.keras.layers.Dense(units=hidden_nodes_layer1,
             input_dim=number_input_features, activation="sigmoid"))

# Second hidden layer
nn.add(tf.keras.layers.Dense(
    units=hidden_nodes_layer2, activation="sigmoid"))

#Third hidden layer
nn.add(tf.keras.layers.Dense(
    units=hidden_nodes_layer3, activation="sigmoid"))

# Output layer
nn.add(tf.keras.layers.Dense(units=1, activation="sigmoid"))

# Check the structure of the model
nn.summary()
```

Neural Network

Compiled the model
Trained the model using 50 epochs
Achieved 87.5% accuracy with a loss of 0.3599

```
# Compile the model
nn.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

```
# Train the model
fit_model = nn.fit(x_train_scaled, y_train, epochs=50)
```

```
9/9 - 0s - 15ms/step - accuracy: 0.8755 - loss: 0.3599
Loss: 0.3598513603210449, Accuracy: 0.8754863739013672
```

Questions?