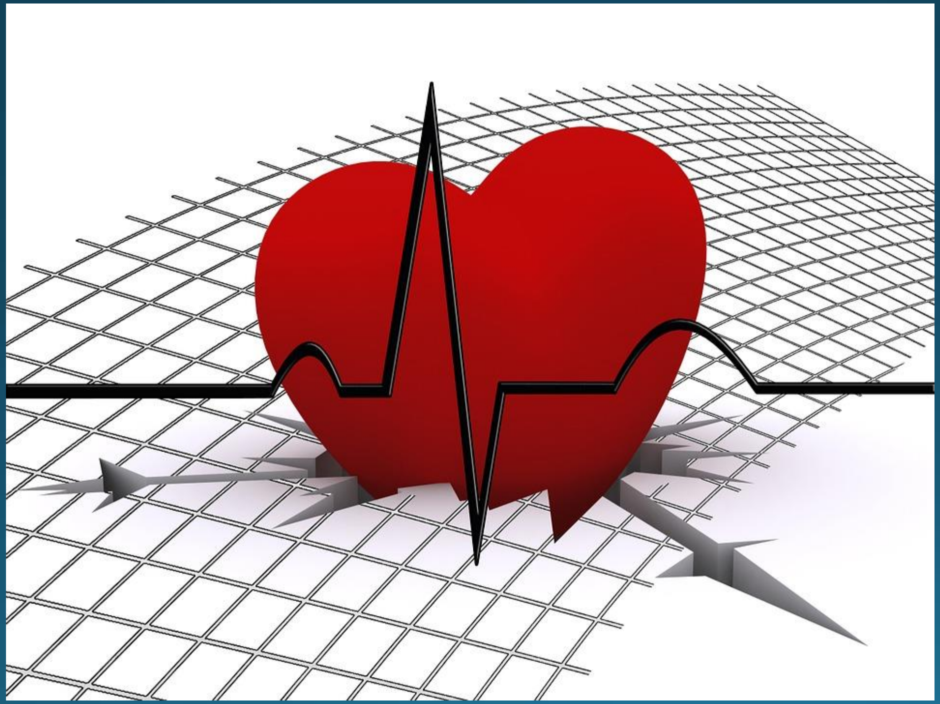
Heart Disease Project

Dr. Nawana Coyle

PROBLEM

Looking at given clinical data, can we predict who will develop heart disease?



WHY IT'S IMPORTANT PREDICT HEART DISEASE

According to CDC (2022),

- ❖ Heart disease is the leading cause of death for men and women in the US.

- ❖ One person dies every 36 seconds in the US from cardiovascular disease.

- ❖ About 659,000 people in the US die from heart disease every year- that's 1 in every 4 deaths.

- ❖ Heart disease costs the US about $363 billion each year from 2016-2017.

https://www.cdc.gov/heartdisease/facts.htm

Data for The Project

- Original data came from the Cleavland data from the UCI Machine Learning Repository

- Data is available on Kaggle. https://www.kaggle.com/ronitf/heart-disease-uci

Data Cleaning

- 14 Columns and 303 rows
- All numerical values
- No missing data
- No duplicated values
- Shuffled data to minimize variance and create a model
- Performing exploratory data analysis (EDA) was straight forward


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       303 non-null    int64  
 1   sex       303 non-null    int64  
 2   cp        303 non-null    int64  
 3   trestbps  303 non-null    int64  
 4   chol      303 non-null    int64  
 5   fbs       303 non-null    int64  
 6   restecg   303 non-null    int64  
 7   thalach   303 non-null    int64  
 8   exang     303 non-null    int64  
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64  
 11  ca        303 non-null    int64  
 12  thal      303 non-null    int64  
 13  target    303 non-null    int64  
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Exploratory Data Analysis (EDA)

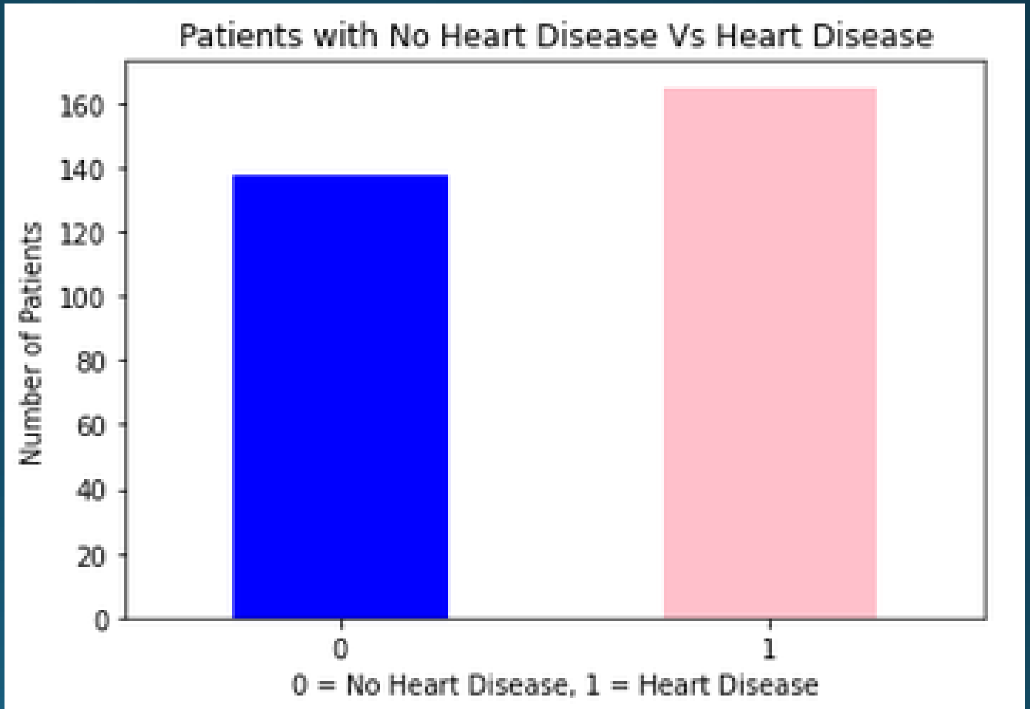
```
file.head(10)
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

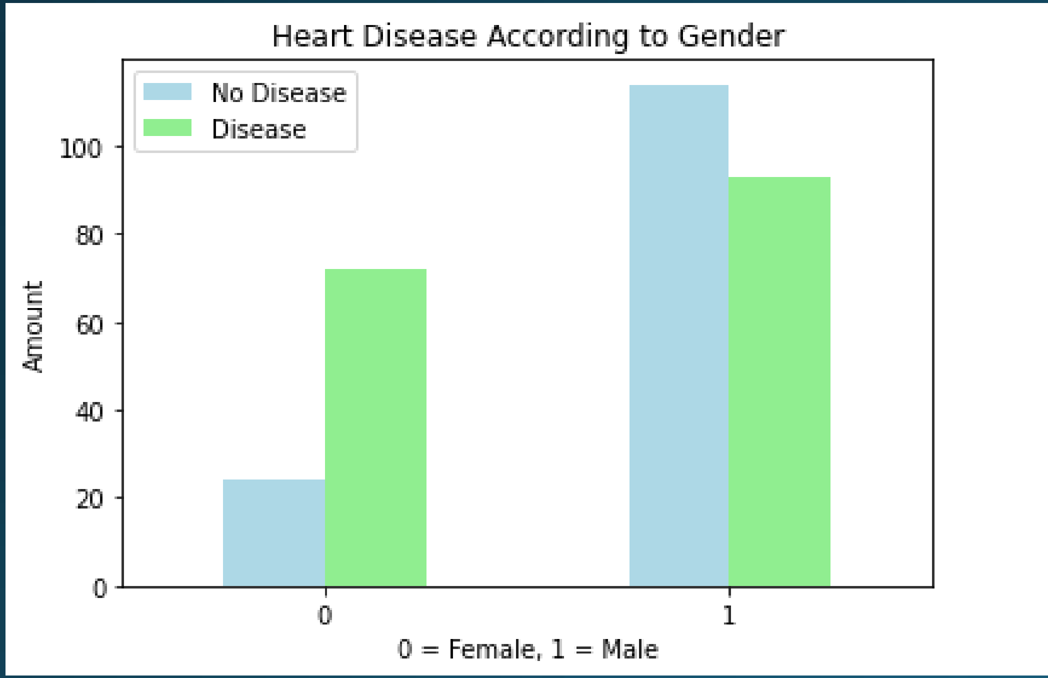
Exploratory Data Analysis (EDA)

❖ Ratio of patients,
Heart disease: No heart disease =~ 1:1

❖ Even distribution of data for the project.



Gender Vs Prevalence of Heart Disease
EDA Continued…

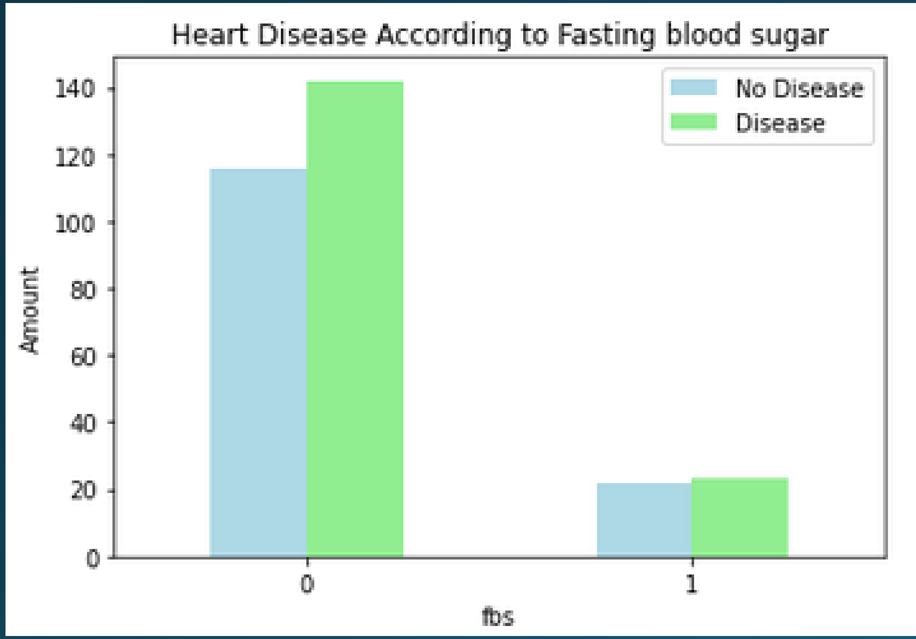


- Significantly more Women with Heart Disease compared to men

Selection Bias?
Or
Reflection of The Real World?

- Further research is required.

Fasting Blood Sugar vs Heart Disease
EDA Continued…



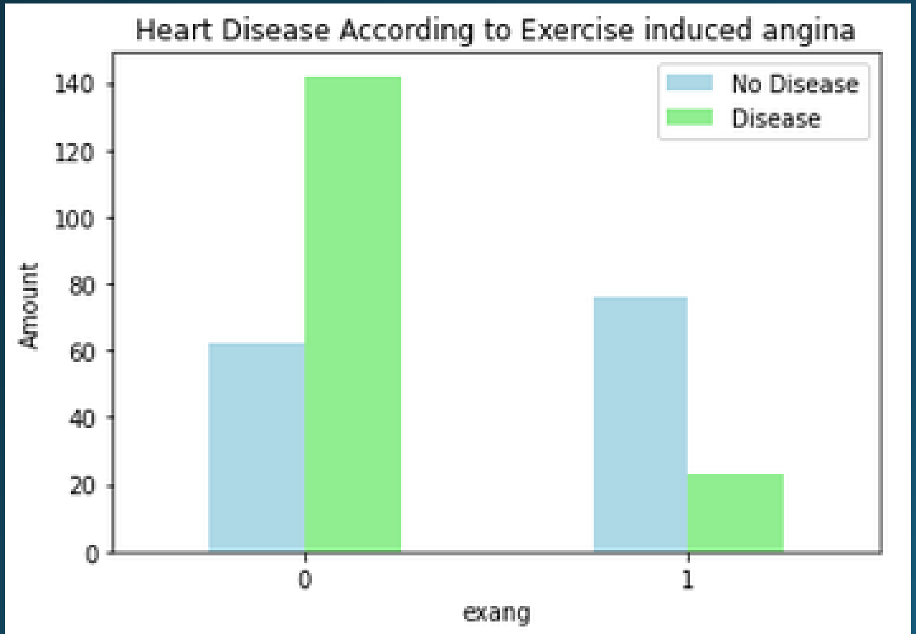
- Observation: Fasting blood sugar doesn’t have a significant effect on developing heart disease.

  Selection Bias?
  Or
  Reflection of The Real World?

- According to CDC, high glucose levels can damage blood vessels and nerves that control the heart.
- Findings are contradictory.
- Further research is required.

Exercise Induced Angina vs Heart Disease

EDA Continued…



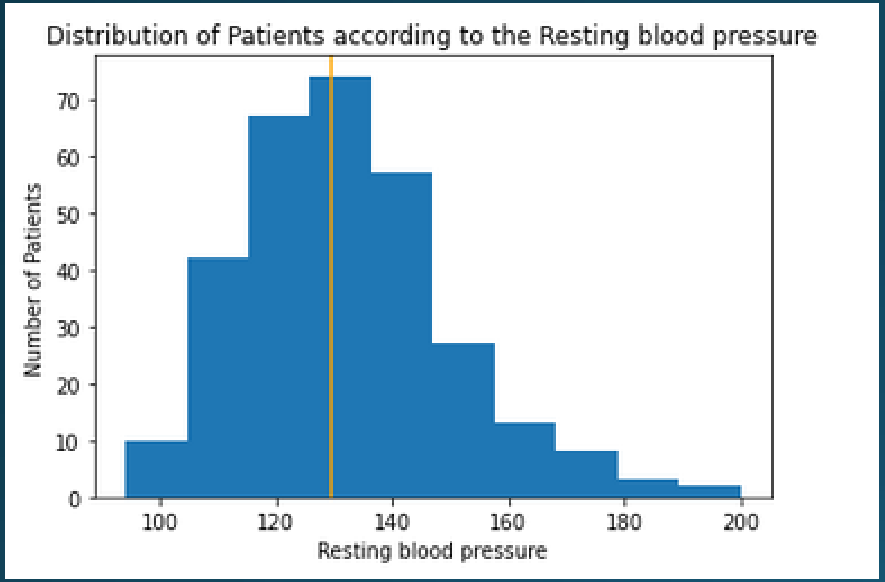
- ❖ Observation: Many people with exercise induced angina do not develop heart disease.
- ❖ According to Clevelandclinic.org, chest pain during exercise is a warning sign for heart disease.
- ❖ Contradictory results.

Selection Bias?

- ❖ More data and research is required.

Resting Blood Pressure Vs Heart Disease

Exploratory Data Analysis (EDA) Continued...



	With No Heart Disease	With Heart Disease
Patients with Resting BP over 170	6	3

- A normal distribution.
- Outliers: 90 and 170 mmHg and 9 data records out of 303 in outliers = 3% of total data
- BP over 170, yet number of patients without heart disease > number of patients with heart disease.
- Needing more data evaluations, misdiagnosis or something else?

Cholesterol Levels vs Heart Disease

Exploratory Data Analysis (EDA) Continued …

❖ General norm

Higher BP

Higher cholesterol

Higher potential for heart disease.

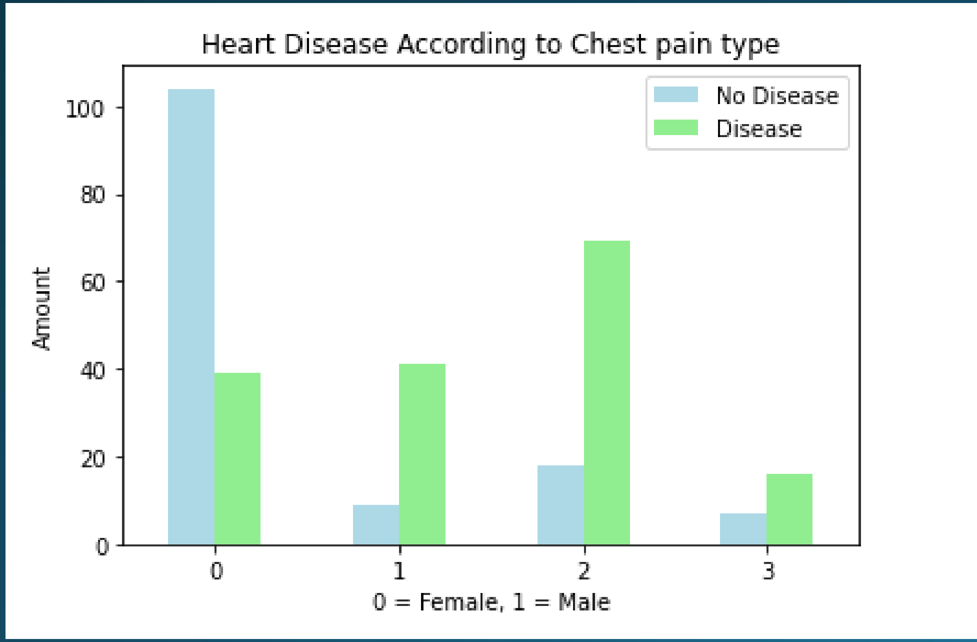
❖ These results contradict the general norm.

Where is the disconnect?

Chest Pain Vs Prevalence of Heart Disease

Exploratory Data Analysis (EDA) Continued …

cp	No Heart Disease	Heart Disease
0	104	39
1	9	41
2	18	69
3	7	16



cp - chest pain type

0 :Typical angina: chest pain related decrease blood supply to the heart
1. Atypical angina: chest pain not related to heart
2. Non-anginal pain: typical esophageal spasms
3. Asymptomatic: chest pain not showing signs of disease

According to the data observations,

Percentage of cp2 patients with Heart Disease < Percentage of cp 0 patients with Heart Disease

How is this possible???
Selection bias, error in diagnosis or something else?

Feature Correlation

Exploratory Data Analysis (EDA) Continued …

+1 or -1 → strong correlation between features.

- Positive correlation:

Feature 1 increase → feature 2 also increase

- Negative correlation:

Feature 1 increase → feature 2 decrease

- Close to 0 → Poor correlation between features

Data Modeling

- Defining success for this project: Reaching accuracy over 95%

- Baseline Models:
  RandomForestClassifier
  KNeighborsClassifier
  LogisticRegression
  XGBClassifier

Why Choose RandomForestClassifer

- Easy to use and generates quick results
- provides high level of accuracy
- Easy to cross validate
- Robust to outliers
- Handles non balanced data
- Does not over fit
- Great for large datasets

Why Choose KNeighborsClassifier?

- Easy to implement
- Fewer parameters to tune: k and distance metric
- No training required to make predictions
- New data can be added when predicting without impacting the outcome

Why Choose Logistic Regression?

- Easier to implement, interpret, and efficient to train without requiring high computational power.
- The feature importance of features can be identified with negative or positive direction.
- Very fast at classifying unknown records
- Good accuracy

Why Choose XGBClassifier?

- Works well in small to medium datasets
- supports regularization to avoid overfitting
- Faster because it uses parallel processing
- Allows to run cross-validation on each iteration

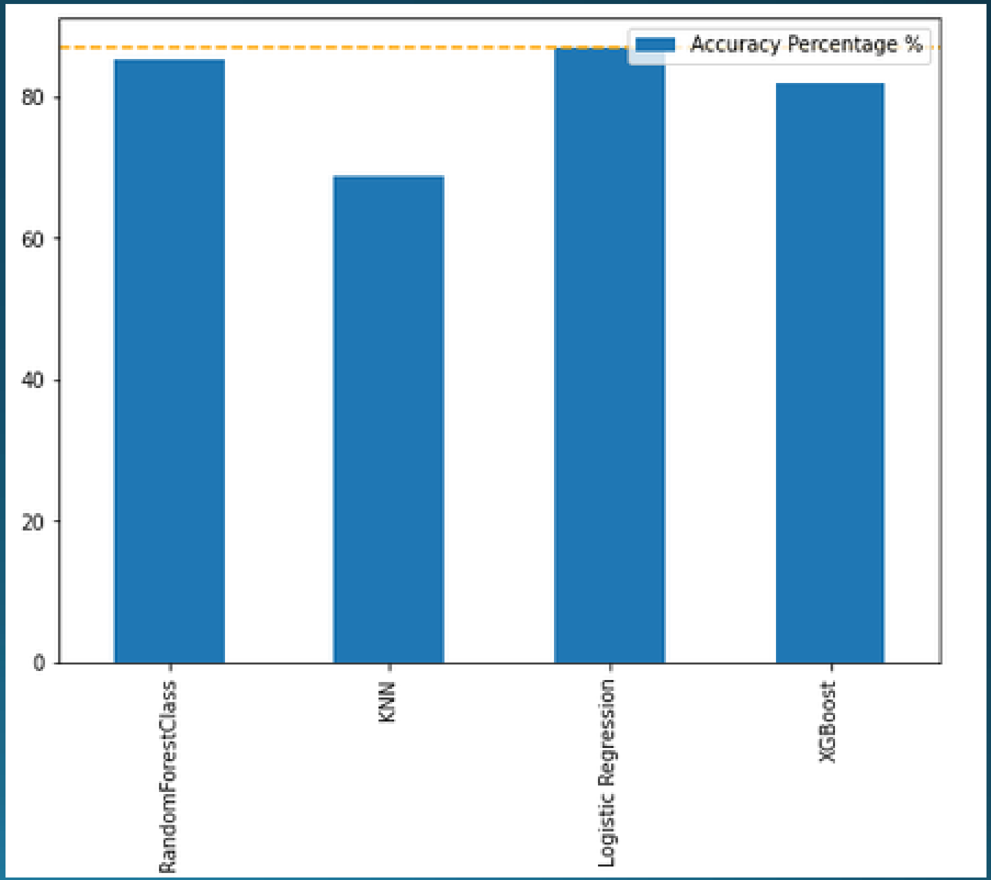
Modeling

1. **Splitting data** : Randomly chosen data, 80% for training, 20% for testing

4 models were evaluated according to their accuracy scores

```
{'RandomForestClass': 85.24590163934425,
 'KNN': 68.85245901639344,
 'Logistic Regression': 86.88524590163934,
 'XGBoost': 81.9672131147541}
```

2. **Visualize** the accuracy scores as percentages



Modeling Continues…

3. Hyperparameter Tuning

a) Since RandomForestClassifier and LogisticRegression models have the highest accuracy score, grid were created to tune them

b) Best parameters were identified

RandomForestClassifier:

```
best_rfc_grid = {'n_estimators': [300],
        'min_samples_split' :  [4],
        'min_samples_leaf': [6],
        'max_features' :['sqrt'],
        'max_depth': [3]}
```

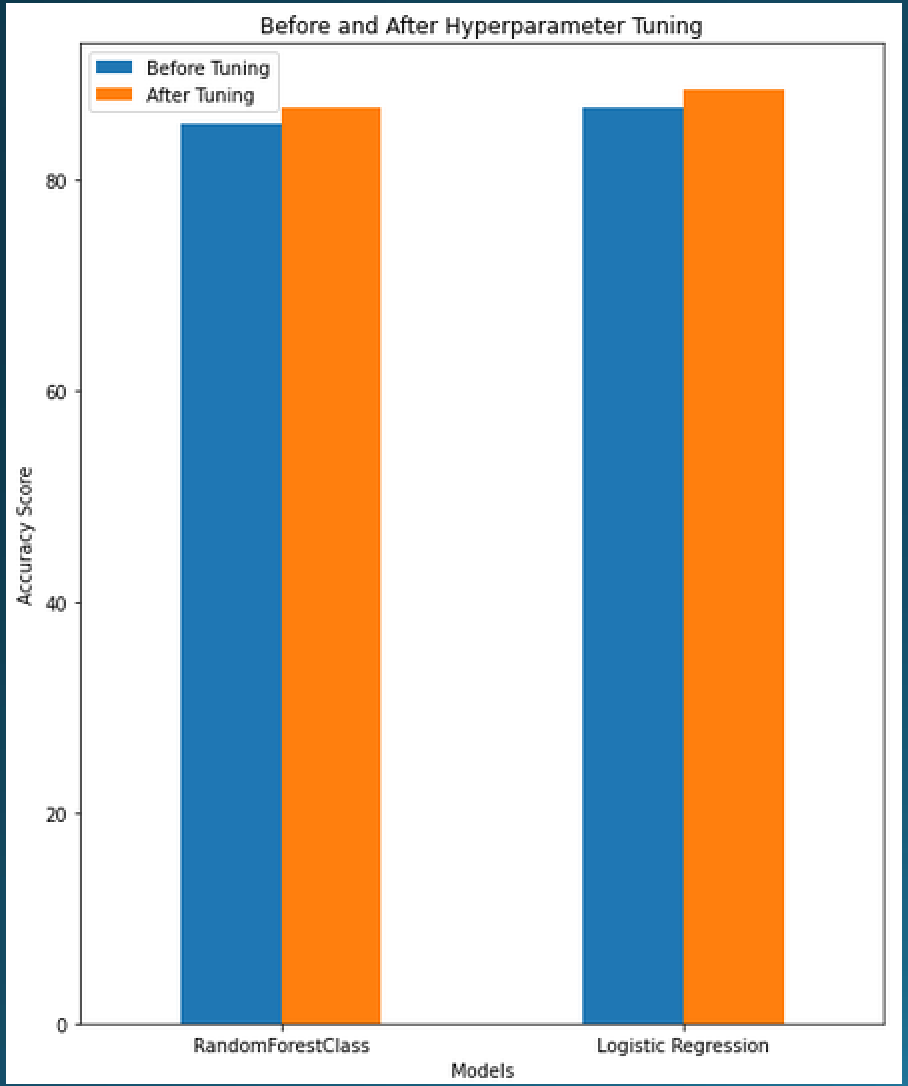
LogisticRegression:

```
best_lg_grid = {'solver': ['liblinear'],
    'penalty':['l2'],
    'C': [0.20433597178569418]}
```

c) Data were retrained on tuned models with best parameters

d) Accuracy scores with best parameters were recalculated for the models.

Modeling Continues…



4. Accuracy score before and after hyperparameter tuning.

Observation:

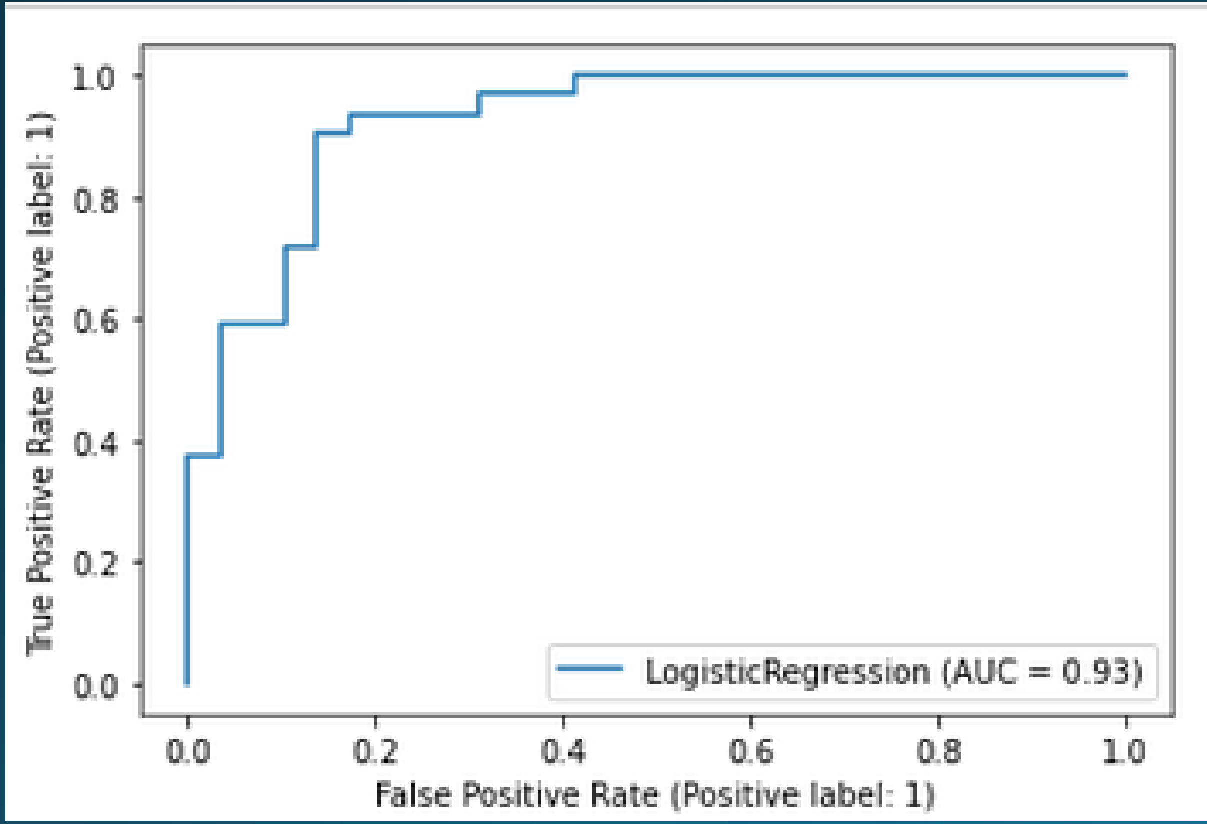
- Accuracy scores have improved after parameter tuning in both models.
- LogisticRegression has achieved the highest accuracy score.

Evaluating The Models

- ROC curve
- Confusion Matrix
- Classification Report
- Precision
- Recall
- F1-score

Evaluating The Model Continues...

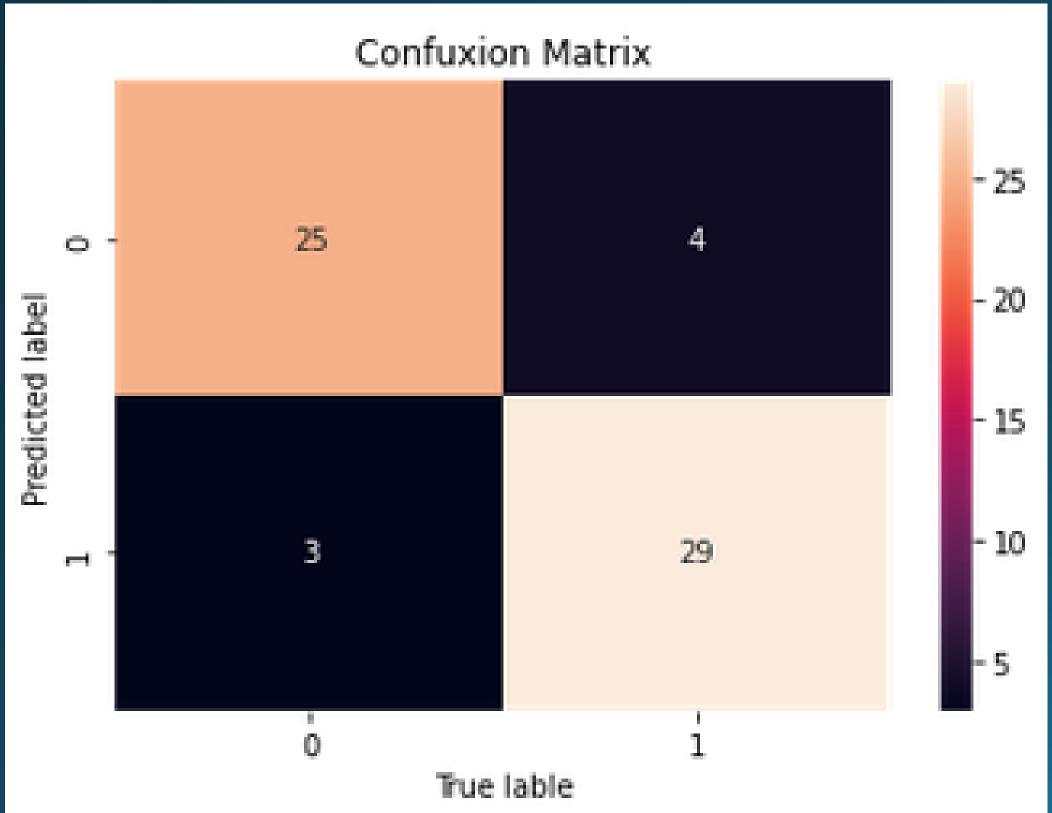
ROC Curve



- The area under the curve (AUC) in LogisticRegression model is 0.93 which is great.
- This indicates that there's only little chance for a patient to be falsely positive.
- Certainly there's room for improvement.

Evaluating The Model Continues...

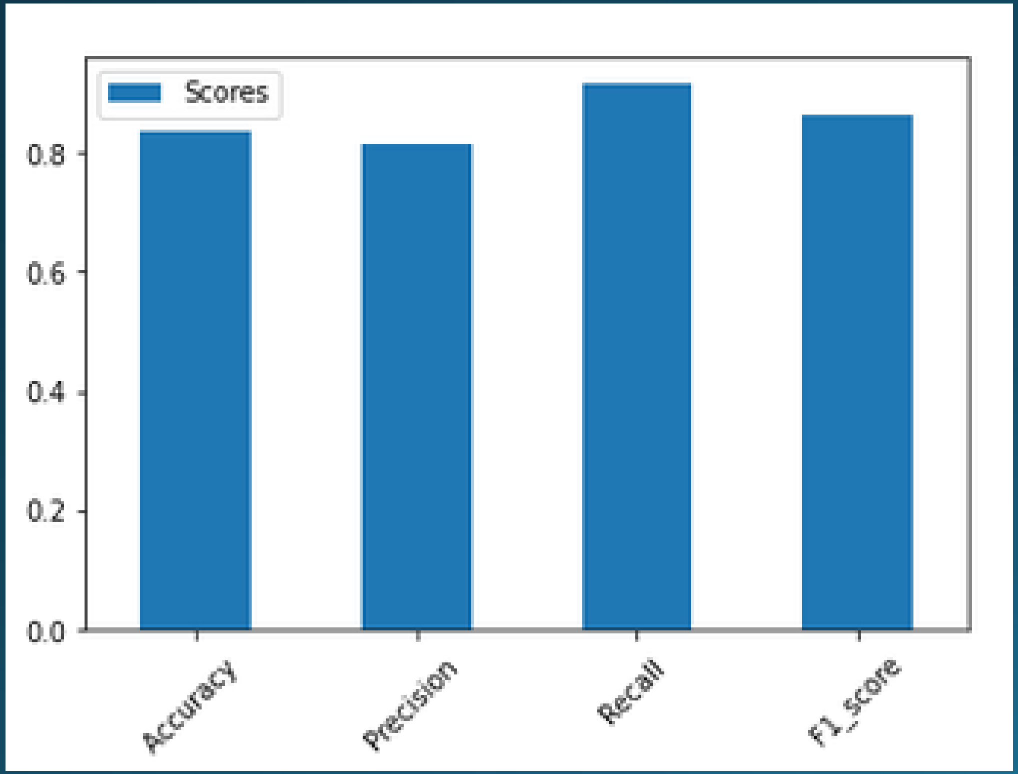
- Confusion Matrix



- The focus of this project is to identifying patients with potential for developing heart disease.
- Improving true positives and reducing false negative values (truly have a higher risk, yet predicts as not) is more important than reducing false positives (model predicts are high risk when they're not).
- Recall is a important feature to focus for this heart disease project.
- Recall = TP / TP + FN = 29 / (29 + 4) = 0.8787

Evaluating The Model Continues…

- Classification Report



Evaluating The Model Continues...

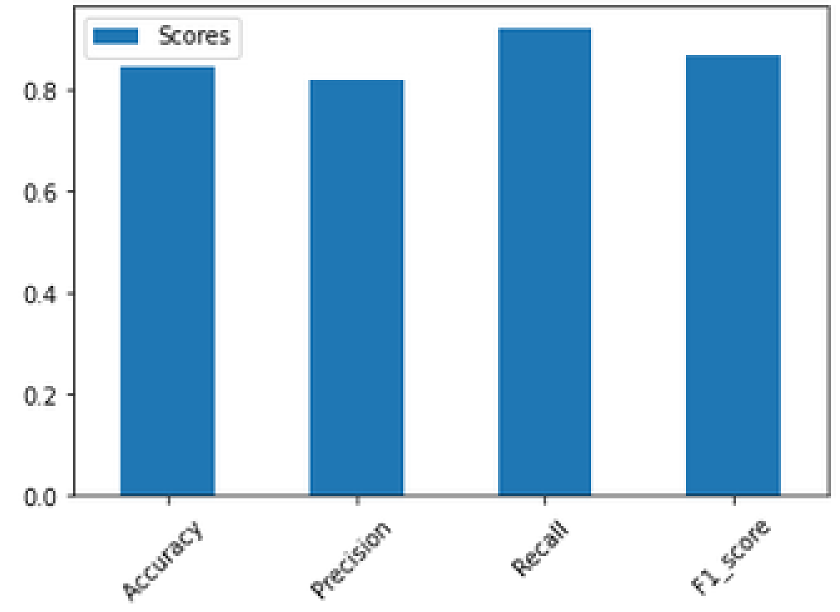
- Classification Report

	precision	recall	f1-score	support
0	0.89	0.86	0.88	29
1	0.88	0.91	0.89	32
accuracy			0.89	61
macro avg	0.89	0.88	0.88	61
weighted avg	0.89	0.89	0.89	61

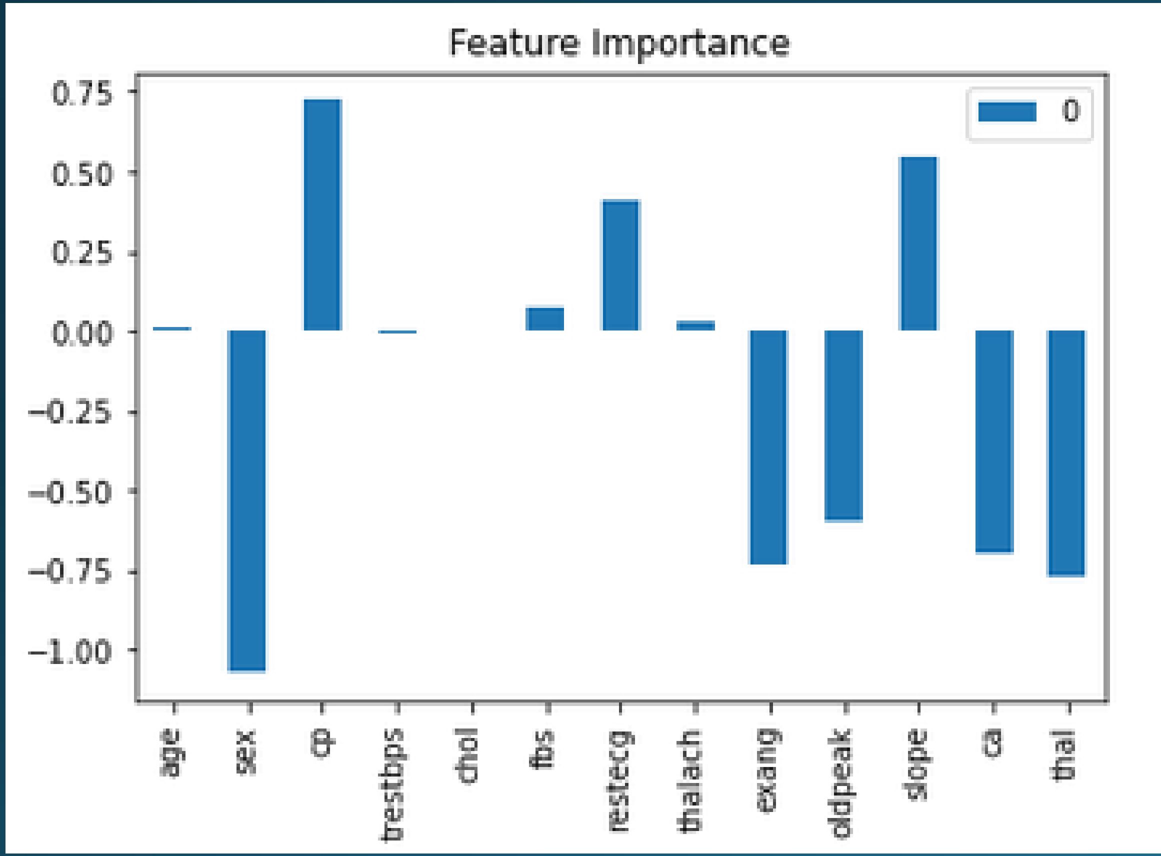
Note: Cross validation provided better scores.

- Cross Validated Accuracy, Precision, Recall, F1-Score

	Accuracy	Precision	Recall	F1_score
Scores	0.844699	0.820794	0.921212	0.867301

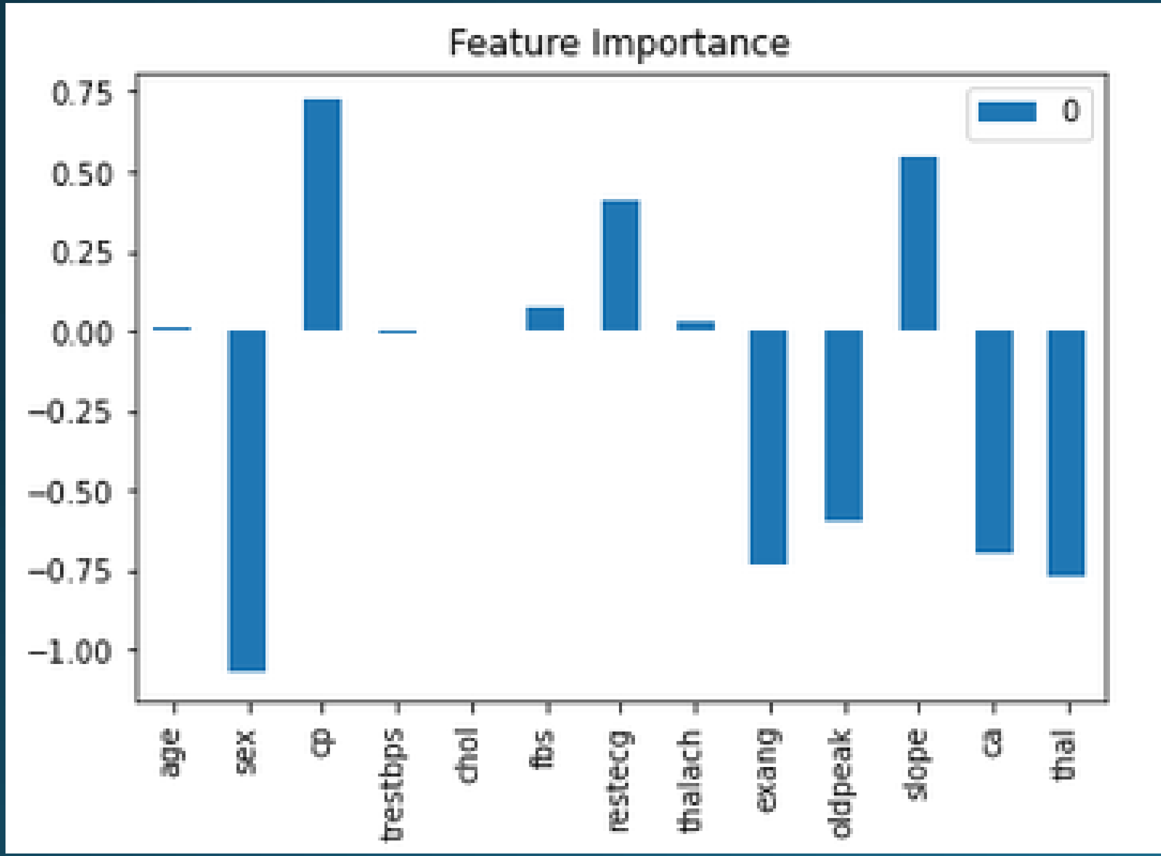


Feature Importance



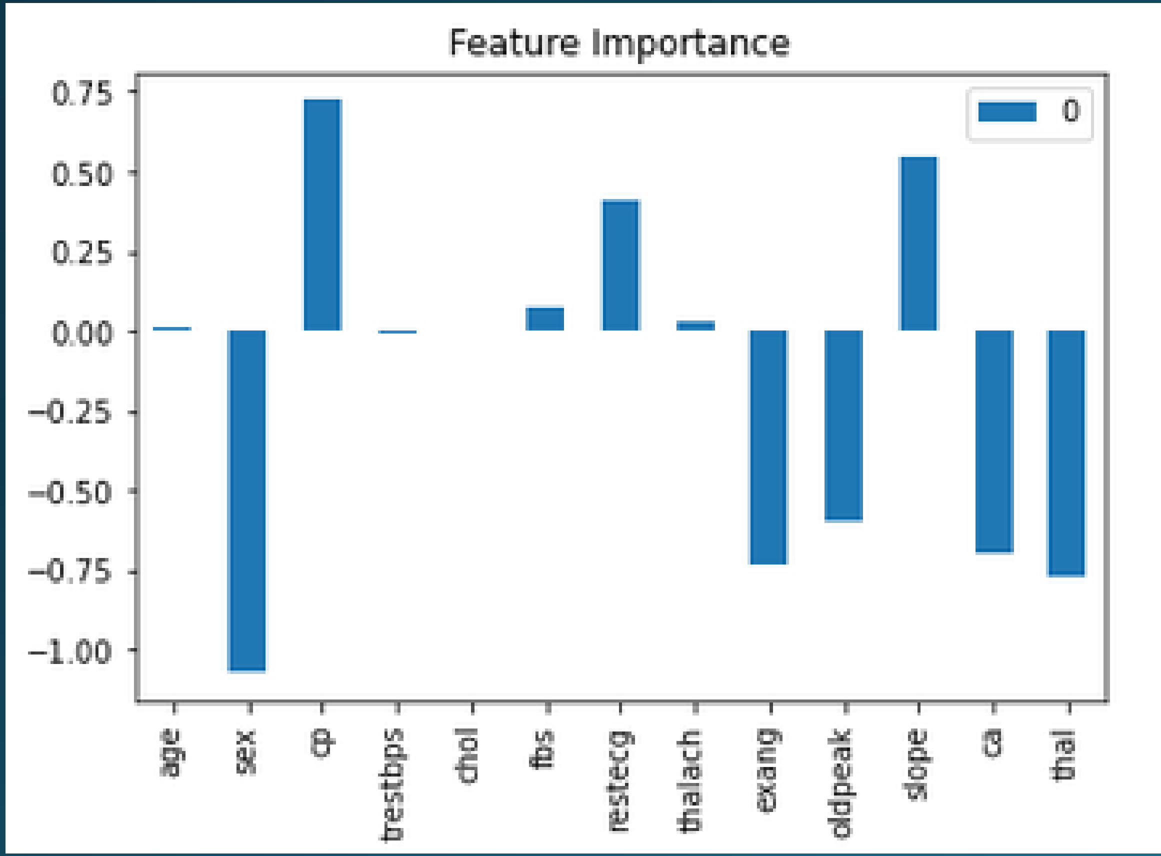
- It answers the question, "which features contributing most to the outcomes of the model?"
- Understanding features that help make predictions is import
- It helps make better predictions and in data gathering process.

Feature Importance



- Features with positive attributes: has a positive correlation with developing heart disease.
  - Higher the value of the attribute, more likely heart disease is developed.
  - E.g.: higher the degree of chest pain, or slope of ST segment, higher likelihood of developing heart disease.

Feature Importance Continues…



- Features with negative values: : has a negative correlation with developing heart disease.
  - Higher the absolute value of the negative feature, it is less likely that heart disease is developed.
  - E.g.: According to EDA, more women (0) has heart disease compared to men (1). So higher the value of sex go (0 to 1) less likely heart disease will be developed.
  - Similarly, more blood vessels visible by flourosopy, better blood supply there is to the heart, and less likely heart disease will be developed.

Summary of the project

- **Problem**
  - Looking at given clinical data, can we predict who will develop heart disease?

- **Findings**

- According to data, significantly more women with heart disease compared to men – Could this be true in the real world?
- According to data, there's no significant effect of fasting blood sugar on developing heart disease. – Could this be true in the real world?
- According to data, many people with exercise induced angina do not develop heart disease – Could this be true in the real world?
- High blood pressure and high cholesterol levels have no significant effect on developing heart disease - – Could this be true in the real world?

- **Results**
  - Logistic regression made the best predictions, compared to other models.
  - The highest accuracy score reached was 88.52% even though the target was 95%.
  - Cross validation provided better recall results, which was indicative that number of false negative values were reduced.
  - Feature importance showed that certain attributes had a positive correlation while the others had a negative correlation.

Areas for improvements

- Consulting with a subject matter experts to understand the disconnect between the findings of the project and what the general public know about heart disease.
- Obtain more random data to represent the general public.
- Try other predictive models
- Try tuning different hyperparameters to obtain better results.

Thank you!