Heart Disease Prediction

Eric Arthur
28.10.2021

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

The objective of this capstone project is to detect whether patients will suffer a heart disease or not, given a number of features from the patients medical history.

The motivation of the project is to use several algorithms to help improve accuracy of diagnosis. This project focused on the use of different algorithms such as Support Vector Machines, Random Forest and Decision Tree to make accurate predictions.

During the analytical stage, it was evident in the dataset that, heart disease is imminent as one gets older in life. The results of the models deployed gave an 85% accuracy rate by the Random Forest algorithm.

Introduction

- **Project background and context**

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

- **Problems you want to find answers**

The problem this project seeks to solve is, how to predict whether patients have heart disease given some features of their diagnosis. Granted that the algorithms employed give accurate predictions, we can not only avoid wrong diagnosis but also save human lives. Certain group of patients without a heart disease are often times diagnosed with heart disease, which sometimes put them into a panic situations. This project will attempt to curb the rate at which erroneous diagnosis are used by qualified medical practitioners. With accurate predictions, one can be assured of the result given by the models. The dataset used for this project contains 12 different features. The algorithms used are Logistic Regression, SVM, Random Forest and Decision Tree.

Section 1

Methodology

Methodology

In this project, 3 different algorithms were implemented, and these are::

- Support Vector Machines
- Random Forest, and
- Decision Tree

- **Data wrangling:** the features and the completeness of the datasets were checked with missingno python library. Several strings variables were also changed into categorical variables.
- **EDAs:** pandas_profiling EDA was introduced to understand some of the features of the dataset as well as other pandas methods, such as, describe, correlations.
- **Visualization:** Seaborn and Matplotlib library were largely used on almost all the features of the dataset to derive insights
- **Predictive Analysis:** the algorithms mentioned above were used to test the accuracy of the predictions

Data Collection

The data used for this project was sourced from kaggle via the link below:

https://www.kaggle.com/fedesoriano/heart-failure-prediction?select=heart.csv

"This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

Cleveland: 303 observations

Hungarian: 294 observations

Switzerland: 123 observations

Long Beach VA: 200 observations

Stalog (Heart) Data Set: 270 observations

Total: 1190 observations
Duplicated: 272 observations

Final dataset: 918 observations"

Data Collection – Data Preview

```
In [223]: #Let's load the data into a panda dataframe

          df = pd.read_csv('heart.csv')
          df
```

Out[223]:

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
...	...	...	...	...	...	...	...	...	...	...	...	...
913	45	M	TA	110	264	0	Normal	132	N	1.2	Flat	1
914	68	M	ASY	144	193	1	Normal	141	N	3.4	Flat	1
915	57	M	ASY	130	131	0	Normal	115	Y	1.2	Flat	1
916	57	F	ATA	130	236	0	LVH	174	N	0.0	Flat	1
917	38	M	NAP	138	175	0	Normal	173	N	0.0	Up	0

918 rows × 12 columns

Data Wrangling

- After previewing the data in pandas and running several EDAs to pick insights from the data, it was observed that some of the EDAs were not coming out right, hence there was a need to modify some of the data in the dataset so we can generate the right EDAs.

- For this reason, some of the strings variables were converted/mapped into categorical variables for the algorithms to understand; sample shown below:

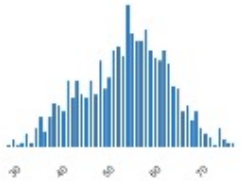
```
# Map the values of the Sex column into binary

df['Sex'] = df['Sex'].map({'M': 0,'F': 1})
```

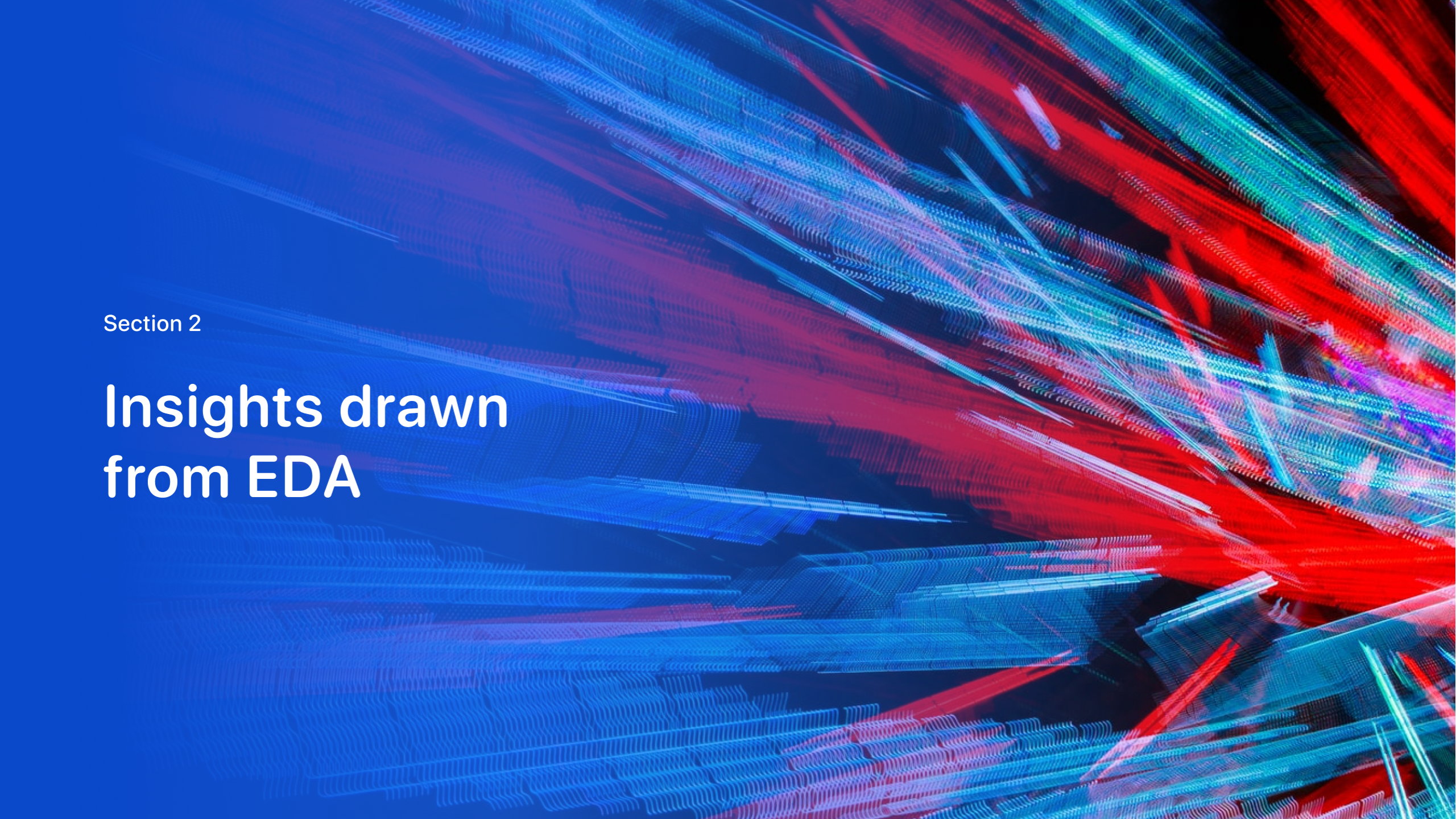
EDAs

- The EDAs adopted for this project basically were pandas correlation and describe methods. Pandas Profiling (which gives an overview of every variable) was also introduced in the notebook for this project to give a general overview of the dataset. The example below show in glance all the details about the Age column using the pandas profiling

Age
Real number ($\mathbb{R}_{\geq 0}$)

Distinct	50	**Minimum**	28
Distinct (%)	5.4%	**Maximum**	77
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	53.51089325	**Memory size**	7.3 KiB



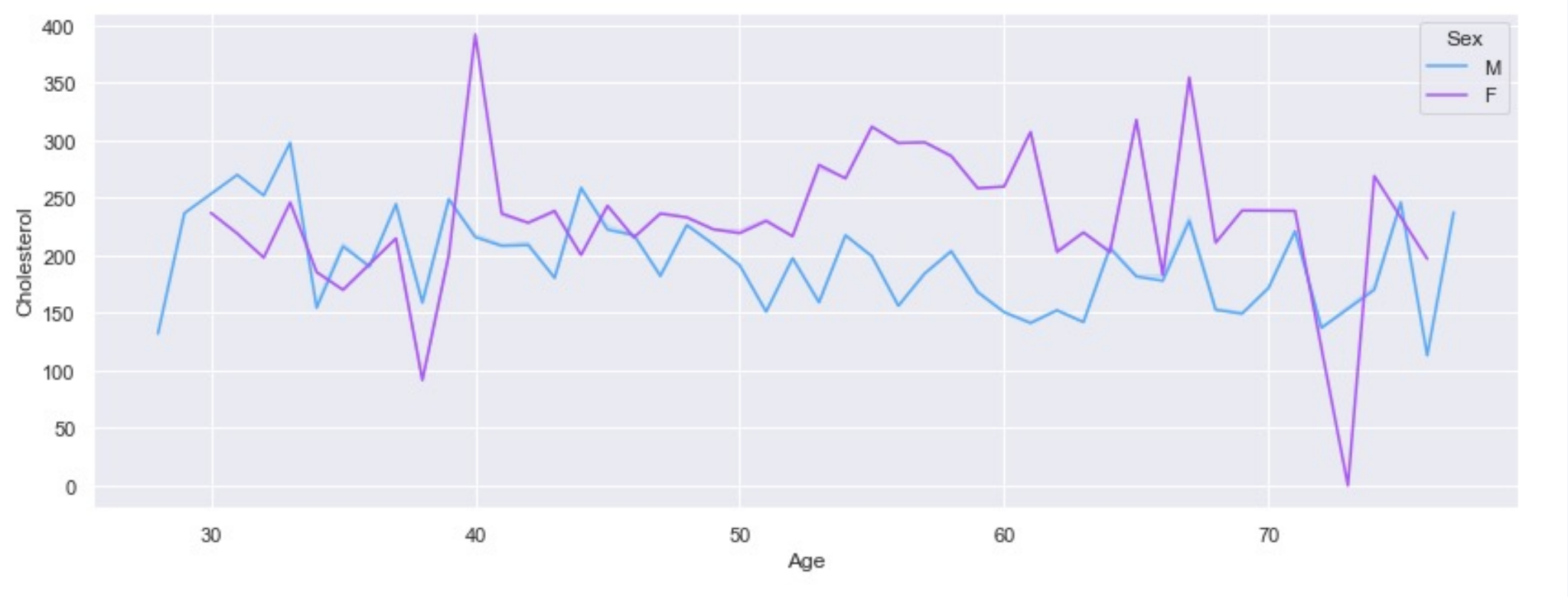


Section 2

Insights drawn from EDA

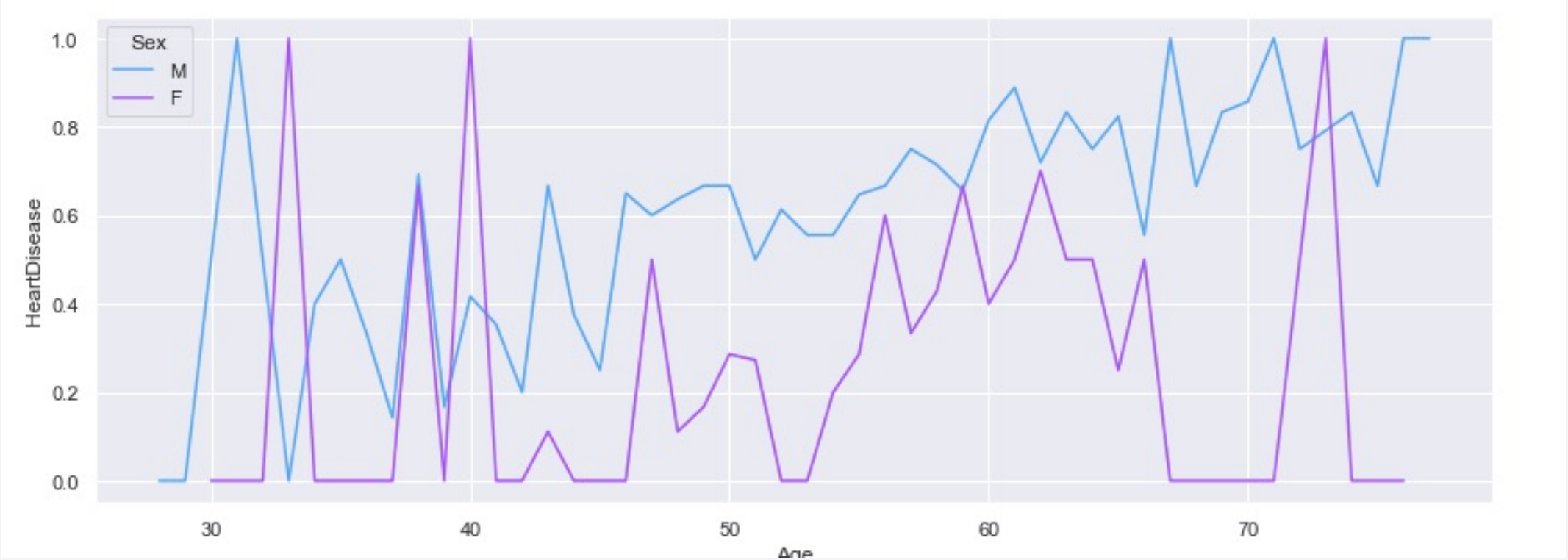
Cholesterol level among Males & Females

- The first thing noticed in during the EDAs showed that, Females tend to develop higher cholesterol than their Male counterparts.



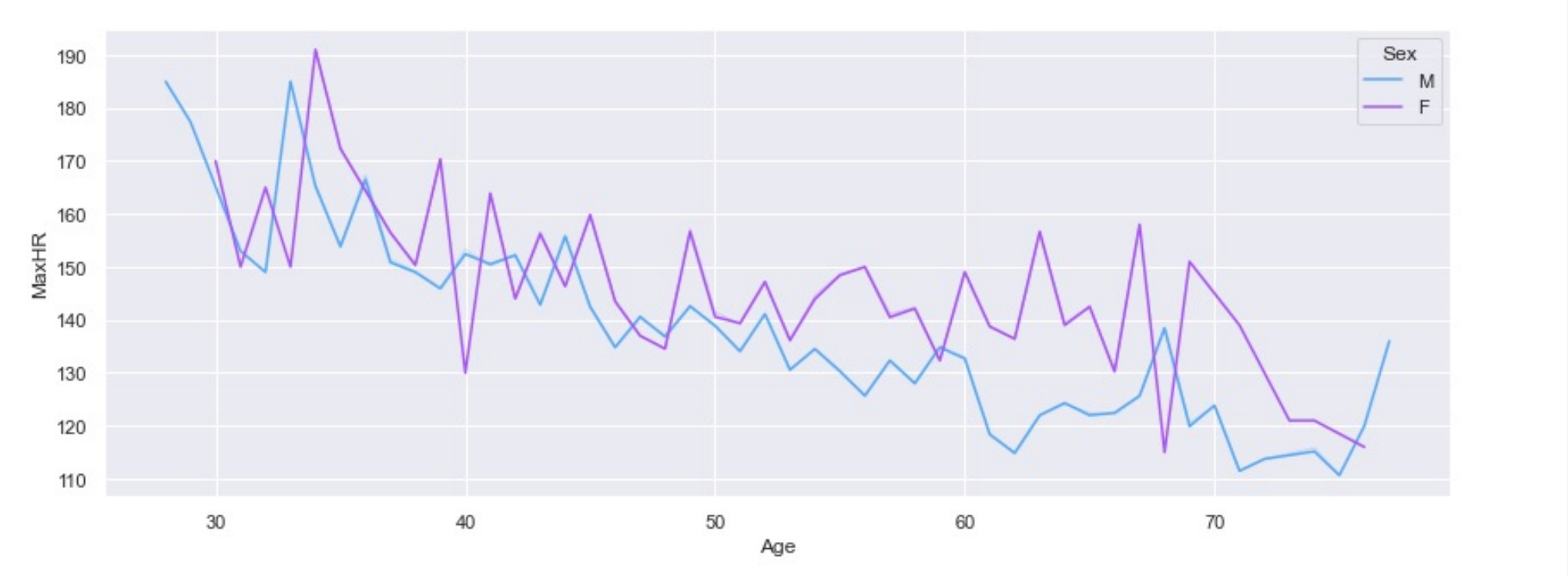
Heart Disease among Males & Females

- Secondly, it was also observed from the data set that, Men develop more heart disease than Females.



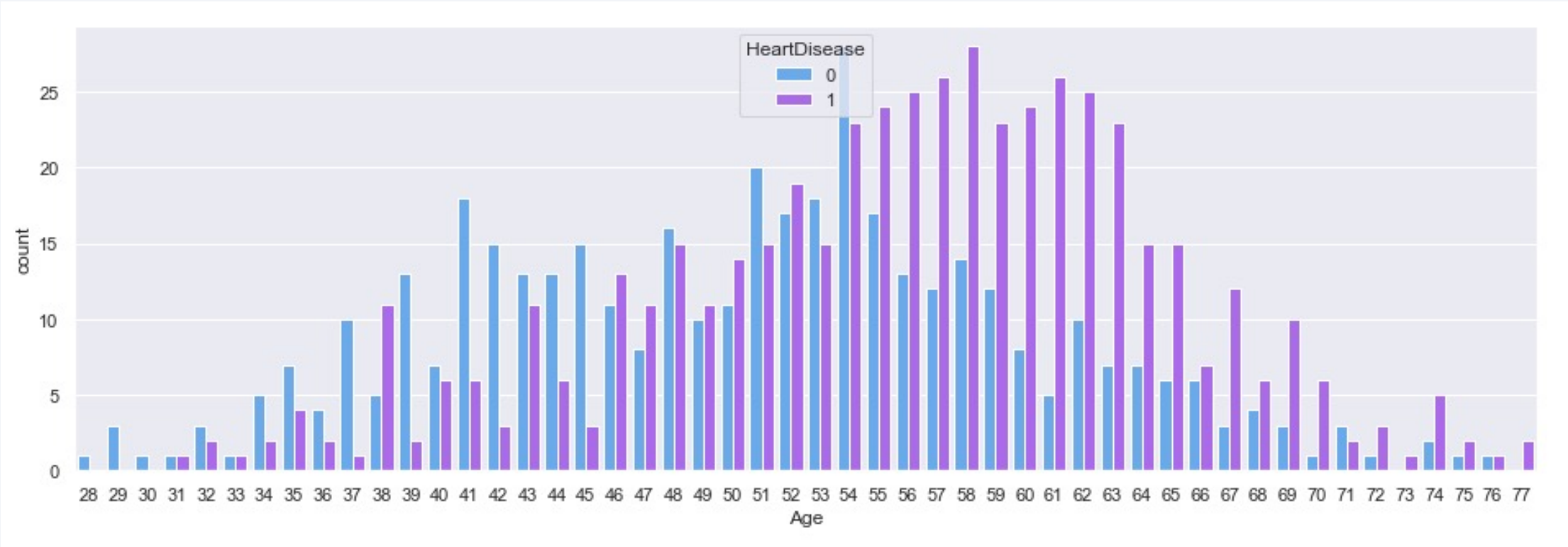
Maximum Heart Rate - Males vs Females

- Again, based on the dataset analysed it was seen that Females have high heart rates than Men. Perhaps due to jealousness



Heart Disease

- Overall, it was noticed that, heart disease is generally associated with the age of a person.



EDA with pandas

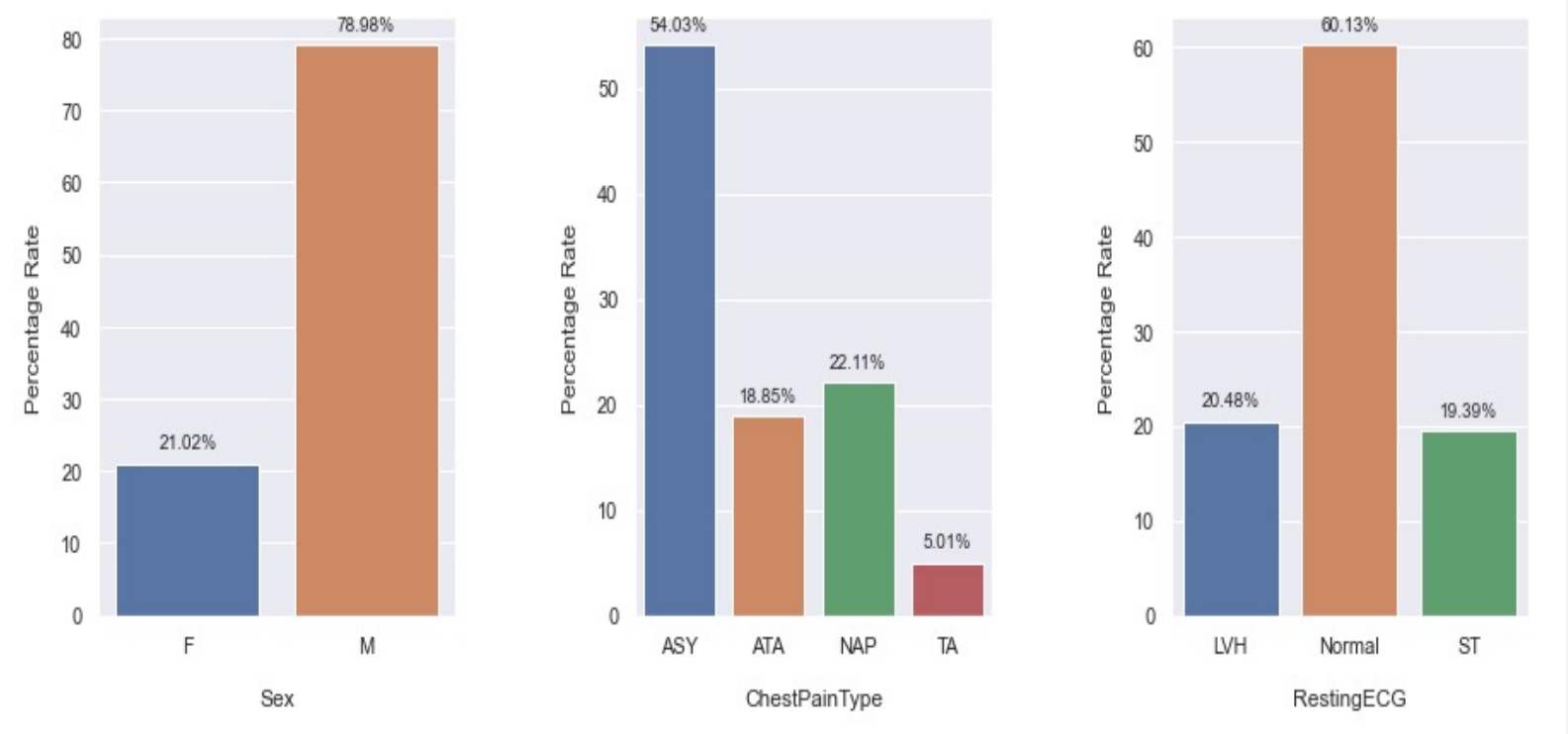
- The initial EDA used to understand the dataset was the pandas describe method which gave a mathematical understanding of each calculated column.

```
df.corr()
```

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
Age	1.000000	0.254399	-0.095282	0.198039	-0.382045	0.258612	0.282039
RestingBP	0.254399	1.000000	0.100893	0.070193	-0.112135	0.164803	0.107589
Cholesterol	-0.095282	0.100893	1.000000	-0.260974	0.235792	0.050148	-0.232741
FastingBS	0.198039	0.070193	-0.260974	1.000000	-0.131438	0.052698	0.267291
MaxHR	-0.382045	-0.112135	0.235792	-0.131438	1.000000	-0.160691	-0.400421
Oldpeak	0.258612	0.164803	0.050148	0.052698	-0.160691	1.000000	0.403951
HeartDisease	0.282039	0.107589	-0.232741	0.267291	-0.400421	0.403951	1.000000

Visualizations

- Several visualizations were used in this projects to generate insights from the datasets. Below are some of the initial visuals used to understand the dataset.



Predictive Analysis (Classification)

- Prior to testing our predictions, the UCI data had to be split into 80% for training, and 20% for testing all them algorithms deployed.

Results

The Random Forest algorithm gave the best test accuracy of 85%. This is because, the random forest is expands to all the features of the dataset, hence the best model in this case.

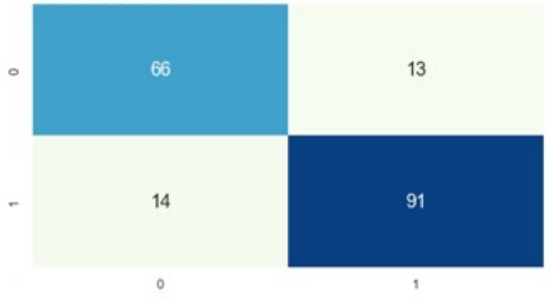
Irrespective of the good result achieved, it is believed that there are more room for improvement and to increase the accuracy of the model. This could be achieved by having a dataset with more instances to experiment and better train the model

Consolidated Model Test Report

Model Name: Default Random Forest

* 1 - No Heart Disease
* MCC - Matthews Correlation Coefficient
* ROC - Receiver Operating Characteristics
* 0 - Heart Disease
* AUC - Area Under The Curve
* AUROC - Area Under the Receiver Operating Characteristics

1. Confusion Matrix



2. Classification Metrics

Classification Metric	Score Value
F1	0.8708
MCC	0.7011
Precision	0.8750
Recall	0.8667
Accuracy	0.8533
AUC_ROC	0.8511
AUC_PR	0.9393

3. Classification Report

	precision	recall	f1-score	support
0	0.82	0.84	0.83	79
1	0.88	0.87	0.87	105
accuracy			0.85	184
macro avg	0.85	0.85	0.85	184
weighted avg	0.85	0.85	0.85	184

Appendix

- https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3018605/

- Insights from:
- http://noiselab.ucsd.edu/ECE228-2020/projects/Report/72Report.pdf

- Code snippet from:
- https://github.com/sauravmishra1710/Heart-Failure-Prediction/blob/main/Heart_Failure_Prediction.ipynb

Thank you!