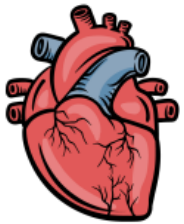


COMP5310

Presentation by Faiyam Islam

Agenda

Introduction and Problem Statement

Understanding the data

Data Preprocessing

Data Summarisation and Analysis

Discussion

Introduction and Problem Statement

- Heart disease is the leading cause of death worldwide, and early detection plays a critical role in preventing adverse health conditions.
- Traditional diagnostic methods often rely on invasive procedures and subjective interpretations, making them time-consuming and costly.
- By leveraging the power of data analysis, we aim to report useful insights to assist healthcare professionals in making accurate and timely diagnosis.

Understanding the Data

Description

- The dataset consists of information related to heart failure prediction, containing various features that could potentially contribute of heart disease occurrences
- The data was collated from different datasets in the UCI Machine Learning Repository:
  - https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

Understanding the Data

Features

Figure 1: Attributes of the heart dataset

Age	Sex	ChestPain	RestingBP	Cholesterc	FastingBS	RestingEC	MaxHR	ExerciseAr	Oldpeak	ST_Slope	HeartDisease
40	M	ATA	140	289	0	Normal	172	N	0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
37	M	ATA	130	283	0	ST	98	N	0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0	Up	0
39	M	NAP	120	339	0	Normal	170	N	0	Up	0
45	F	ATA	130	237	0	Normal	170	N	0	Up	0
54	M	ATA	110	208	0	Normal	142	N	0	Up	0
37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
48	F	ATA	120	284	0	Normal	120	N	0	Up	0
37	F	NAP	130	211	0	Normal	142	N	0	Up	0
58	M	ATA	136	164	0	ST	99	Y	2	Flat	1
39	M	ATA	120	204	0	Normal	145	N	0	Up	0
49	M	ASY	140	234	0	Normal	140	Y	1	Flat	1
42	F	NAP	115	211	0	ST	137	N	0	Up	0
54	F	ATA	120	273	0	Normal	150	N	1.5	Flat	0
38	M	ASY	110	196	0	Normal	166	N	0	Flat	1

Understanding the Data

Key insights

- Categorical predictors: Sex, ChestPainType, RestingECG, ExerciseAngina, ST_Slope
- Numerical predictors: Age, RestingBP, Cholesterol, FastingBS, MaxHR, Old peak

- Desrciptive statistics reveals that FastingBS and HeartDisease are binary variables with minimum and maximum values of 0 and 1
- Categorical predictors aren't shown
- Cholesterol has the highest mean value with FastingBS having the lowest

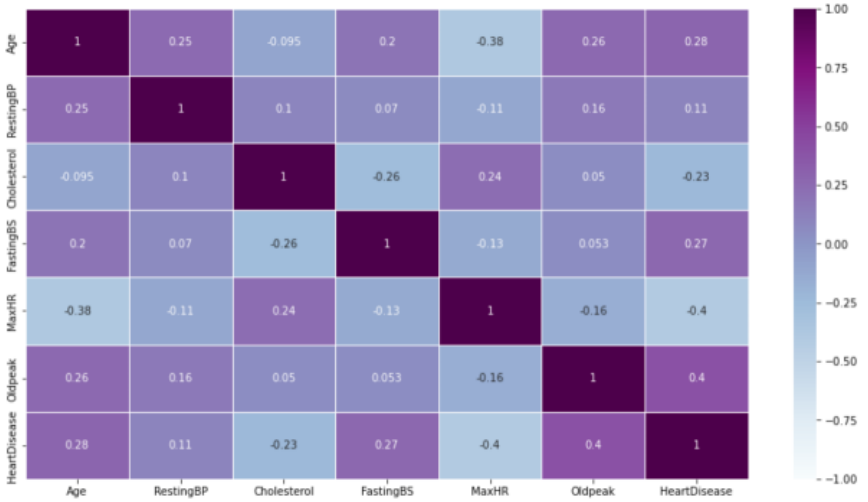
Figure 2: Descriptive statistics of numerical predictors of heart dataset

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	53.510893	132.396514	198.799564	0.233115	136.809368	0.887364	0.553377
std	9.432617	18.514154	109.384145	0.423046	25.460334	1.066570	0.497414
min	28.000000	0.000000	0.000000	0.000000	60.000000	-2.600000	0.000000
25%	47.000000	120.000000	173.250000	0.000000	120.000000	0.000000	0.000000
50%	54.000000	130.000000	223.000000	0.000000	138.000000	0.600000	1.000000
75%	60.000000	140.000000	267.000000	0.000000	156.000000	1.500000	1.000000
max	77.000000	200.000000	603.000000	1.000000	202.000000	6.200000	1.000000

Understanding the Data

- Correlation matrix reveals that most predictors are not highly correlated
- This is evident from R-score values, which did not exceed 0.7
    - No indication of issues with multicollinearity
- MaxHR and HeartDisease exhibit negative correlation with R-score of -0.4

Figure 3: Correlation matrix of numerical predictors in heart dataset



Data Preprocessing

Data Transformation

- No null values found in the dataset using the `isna().sum()` command

Figure 4: Checking null values on heart dataset

```
# To check null values
heart_df.isna().sum()
```

```
Age               0
Sex               0
ChestPainType     0
RestingBP         0
Cholesterol       0
FastingBS         0
RestingECG        0
MaxHR             0
ExerciseAngina    0
Oldpeak           0
ST_Slope          0
HeartDisease      0
dtype: int64
```

Data Preprocessing

- No duplicate values found in the dataset using the `.duplicated()` command

Figure 5: Checking duplicate values on heart dataset

```
# Check duplicate values
duplicates = heart_df.duplicated()
print(duplicates)
```

```
0      False
1      False
2      False
3      False
4      False
       ...
913    False
914    False
915    False
916    False
917    False
Length: 918, dtype: bool
```

```
heart_df.drop_duplicates(inplace = True)
heart_df.shape
```

```
(918, 12)
```

Data Preprocessing

- Categorical predictors were converted to numerical predictors to allow for the use of machine learning models that require numerical inputs.

Figure 6: One-hot encoding on categorical variables in heart dataset

```
# Select categorical variables
categ = heart_df.select_dtypes(include=object).columns

# One hot encoding
heart_df = pd.get_dummies(heart_df, columns=categ, drop_first=True)
heart_df.head()
```

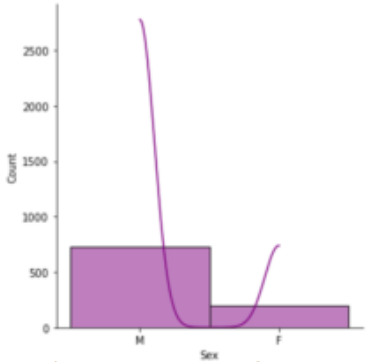
	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease	Sex_M	ChestPainTy
0	40	140	289	0	172	0.0	0	1	
1	49	160	180	0	156	1.0	1	0	
2	37	130	283	0	98	0.0	0	1	
3	48	138	214	0	108	1.5	1	0	
4	54	150	195	0	122	0.0	0	1	

Data Summarisation and Analysis

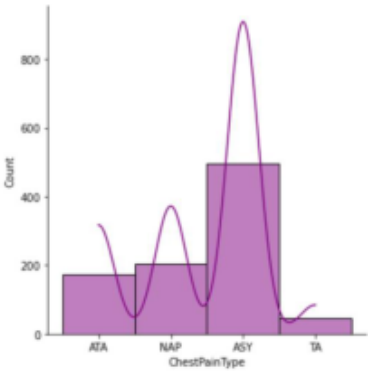
Exploratory Data Analysis

- There are more males than females in the sample
- The most frequent chest pain type experienced by patients is Asymptomatic, whereas Typical Angina is the least common
- Histogram on RestingBP variable has a left-tail distribution with mode around 120-140 resting blood pressure
- Mode for Cholesterol levels seems to be 0, which is not possible
    - Further data transformation has been taken to remove all zeros

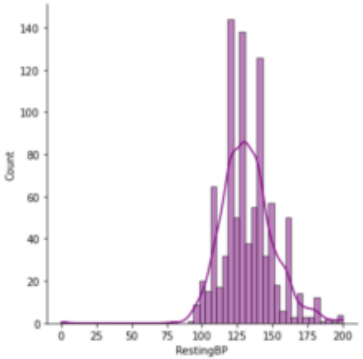
Plot 1: Histogram of patient's sex



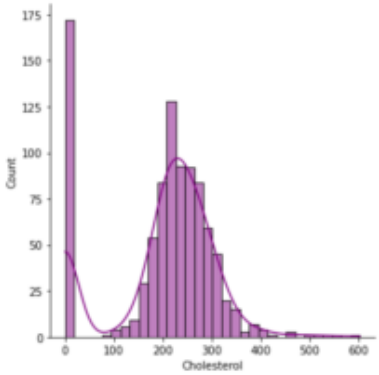
Plot 2: Histogram of patient's ChestPainType



Plot 3: Histogram of patient's RestingBP



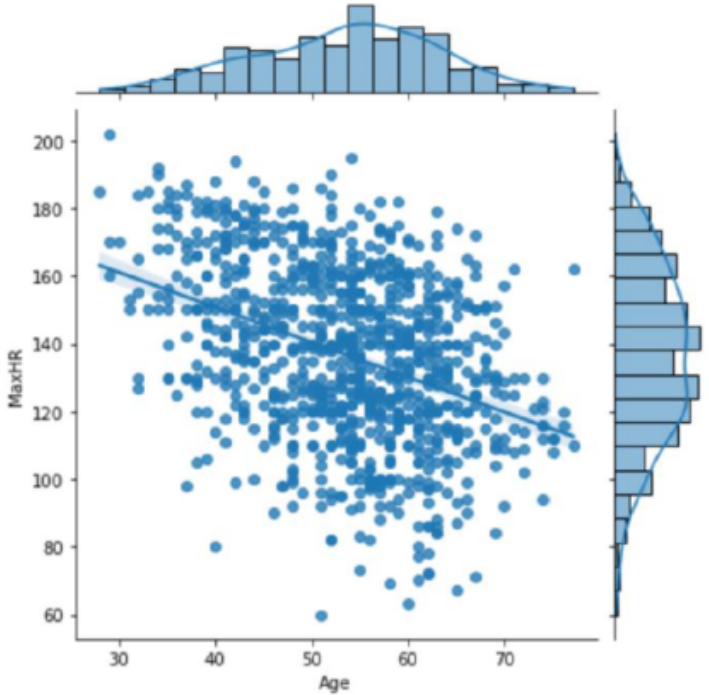
Plot 4: Histogram of patient's Cholesterol



Data Summarisation and Analysis

- Jointplot visualises a negative relationship between MaxHR and Age
- Presence of a few outliers, removal of these data points could potentially strengthen the bivariate relationships between variables
- Histograms on the top and right side allows us to understand where most of the data lies which is around ages 50 and MaxHR of 140.

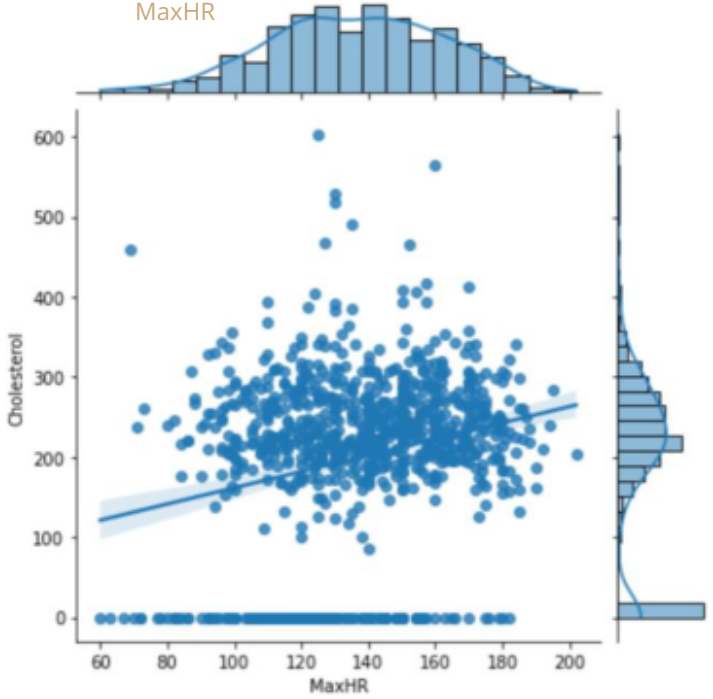
Plot 5: Jointplot of MaxHR against Age



Data Summarisation and Analysis

- Bivariate relationship between MaxHR and Cholesterol is positive
  - Could indicate that patients with higher maximum heart rate generally tend to have higher Cholesterol levels
- We cannot confirm this is a strong relationship due to the various outliers around 500 to 600 Cholesterol levels as shown in this plot

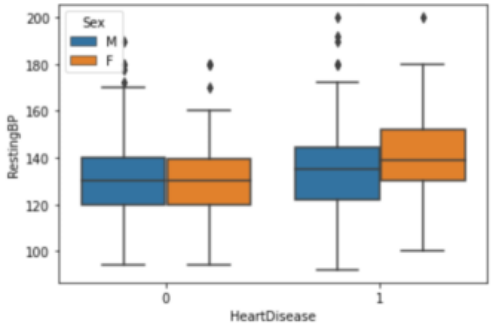
Plot 6: Jointplot of Cholesterol against MaxHR



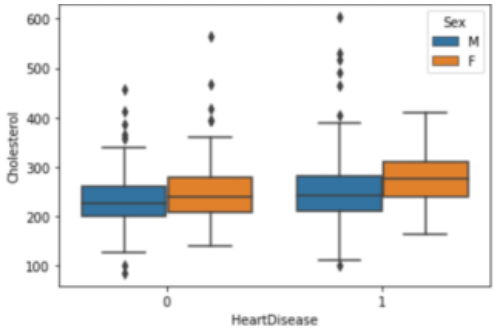
Data Summarisation and Analysis

- Potential outlier detection through box plot visualisations
- Boxplot of HeartDisease over RestingBP and boxplot of HeartDisease over Cholesterol had various zeros, these were removed for better accuracy during model deployment and evaluation metrics are applied

Plot 7: Boxplot of HeartDisease over RestingBP



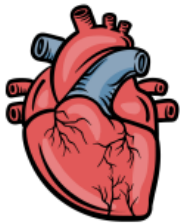
Plot 8: Boxplot of HeartDisease over Cholesterol



Discussion

What's next?

- This is a classification problem, the dataset has a variety of different data types, alluding to various model deployments possible
- Logistic Regression, Multi-class classification, Naive Bayes, Decision Trees can be used as machine learning techniques for modelling
- Evaluation metrics can range from Accuracy, Precision-Recall, F1 score and RMSE can be used when measuring the effectiveness of these models on the above supervised learning methods
- Evidently, our results should indicate which variables are the key indicators and are the most prevalent when determining heart disease in these patients, hence answering the researchquestion



Thank you for listening!