Tampere University
Tampere University of Applied Sciences

BBT.MJS.101 Small Sample Data Analysis

PROJECT WORK REPORT

Name: Fizra Khan

Student number: 152177548

1. Descriptive Statistics

1.1 Key statistics

In the table below, we calculated the key statistics for each variable listed below. The important thing to note is that, for data that is not normally distributed (normality indicated in section 1.2 for each variable), we take only median into account instead of mean and standard deviation.

CAD patients:

Numeric distributions

	Age	maxHR	ST depression	ST/HR index	ΔRWA
Mean	53.98	122.74	0.1369	3.241	0.192
Median	55	120.5	0.110	2.945	0.190
Standard Deviation	8.39	21.67	0.11	2.51	0.145
95% CI	52.31 – 55.64	118.43 – 127	0.114 – 0.159	2.742 – 3.740	0.163 – 0.220

Categorical distributions

Variable	Categories	Count
Sex	Male	76
	Female	24
Myocardial Infarction	0 (Never)	57
	1 (Once)	20
	2 (Twice)	23
Beta	0 (No)	18
	1 (Yes)	82
Calsi	0 (No)	65
	1 (Yes)	35
Digit	0 (No)	98

	1 (Yes)	2
Nitro	0 (No)	26
	1 (Yes)	74

Non-CAD patients

Numeric distributions

	Age	maxHR	ST depression	ST/HR index	ΔRWA
Mean	48.38	159.07	0.0537	0.9105	0.117
Median	49.50	161	0.03	0.70	0.07
Standard Deviation	12.13	21.20	0.061	0.859	0.147
95% CI	45.97 – 50.78	154.86 - 163.27	0.041 – 0.065	0.739 – 1.081	0.088 – 0.146

Categorical distributions

Variable	Categories	Count
Sex	Male	50
	Female	50
Myocardial Infarction	0 (Never)	100
	1 (Once)	0
	2 (Twice)	0
Beta	0 (No)	95
	1 (Yes)	5
Calsi	0 (No)	98
	1 (Yes)	2
Digit	0 (No)	99

	1 (Yes)	1
Nitro	0 (No)	97
	1 (Yes)	3

1.2. Normality Tests

The Shapiro-Wilk Test and visual inspection using distribution plots were used to test the normality of the data in patients with CAD and non-CAD to make the decision about parametric and non-parametric statistical tests. The normality is tested for the given variables including maxHR because it will be required in the upcoming sections of this project. Following hypothesis are used in Shapiro-Wilk test:

Null hypothesis (Ho): The data is normally distributed.

Alternative hypothesis (H1): The data is not normally distributed.

If p-value is < 0.05, we rejected null hypothesis concluding that the data is not normally distributed. After applying the test for different variables that are used in upcoming tasks, we get the following p-values:

Variable	p-value	Normality
Patients having CAD		
MaxHR	0.737 > 0.05	Ho accepted, normally distributed
ST depression	3.86e-06 < 0.05	Ho rejected, not normally distributed
RWA	0.002 < 0.05	Ho rejected, not normally distributed
ST/HR	1.77e-06 < 0.05	Ho rejected, not normally distributed
Age	0.060 > 0.05	Ho accepted, normally distributed
Patients not having CAD		
MaxHR	8.03e-06 < 0.05	Ho rejected, not normally distributed
ST depression	2.72e-08 < 0.05	Ho rejected, not normally distributed
RWA	3.69e-1 < 0.05	Ho rejected, not normally distributed
ST/HR	1.55e-06 < 0.05	Ho rejected, not normally distributed
Age	0.0017 < 0.05	Ho rejected, not normally distributed
In male population of patients having CAD		

Age	0.260 > 0.05	Ho accepted, normally distributed
MaxHR	0.289 > 0.05	Ho accepted, normally distributed
In female population of patients having CAD		
Age	0.046 < 0.05	Ho rejected, not normally distributed
MaxHR	0.963 > 0.05	Ho accepted, normally distributed
In total population		
Age	0.00604 < 0.05	Ho rejected, not normally distributed
RWA	2.208e-10 < 0.05	Ho rejected, not normally distributed

The risk level to be wrong if we make the wrong decision on distribution of maxHR to be non-normal would be 5%.

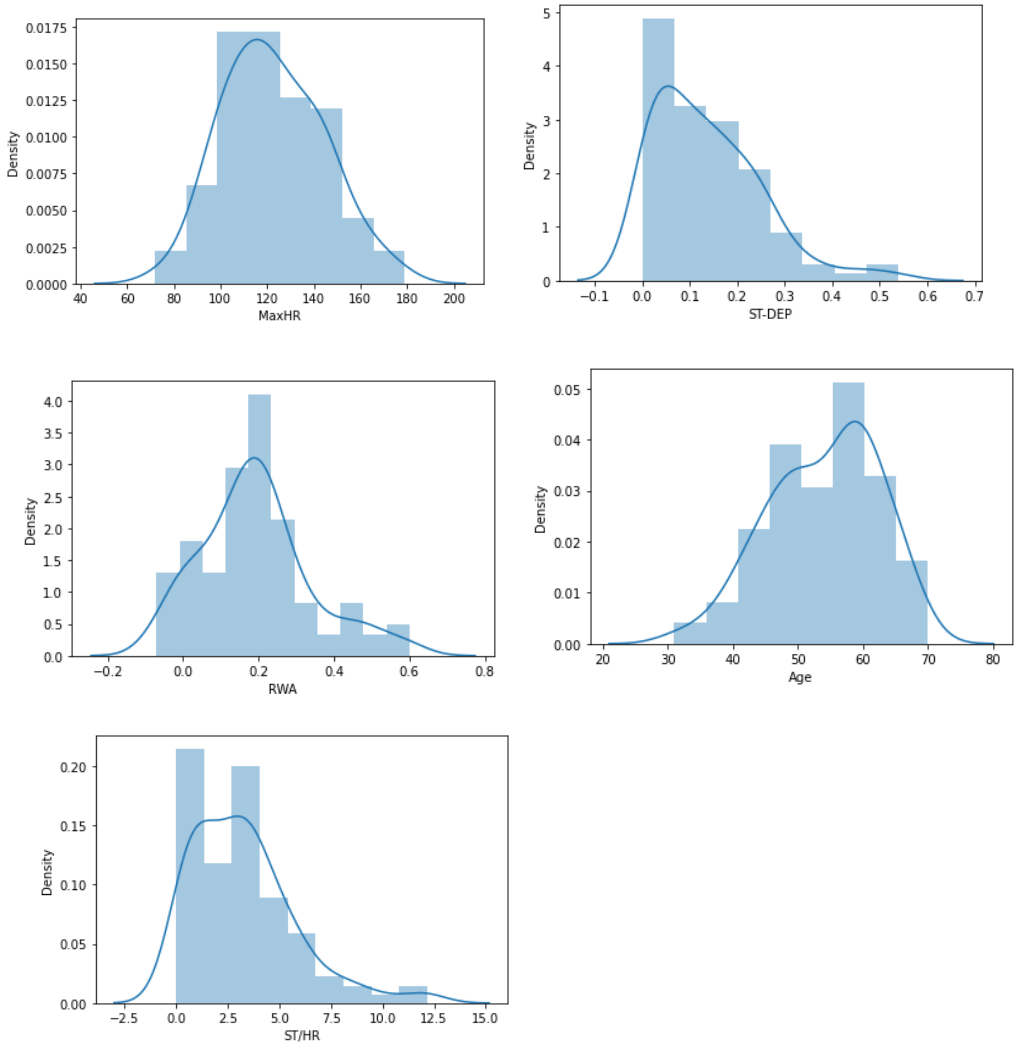


Figure 1: Distribution of different variables in patients having CAD

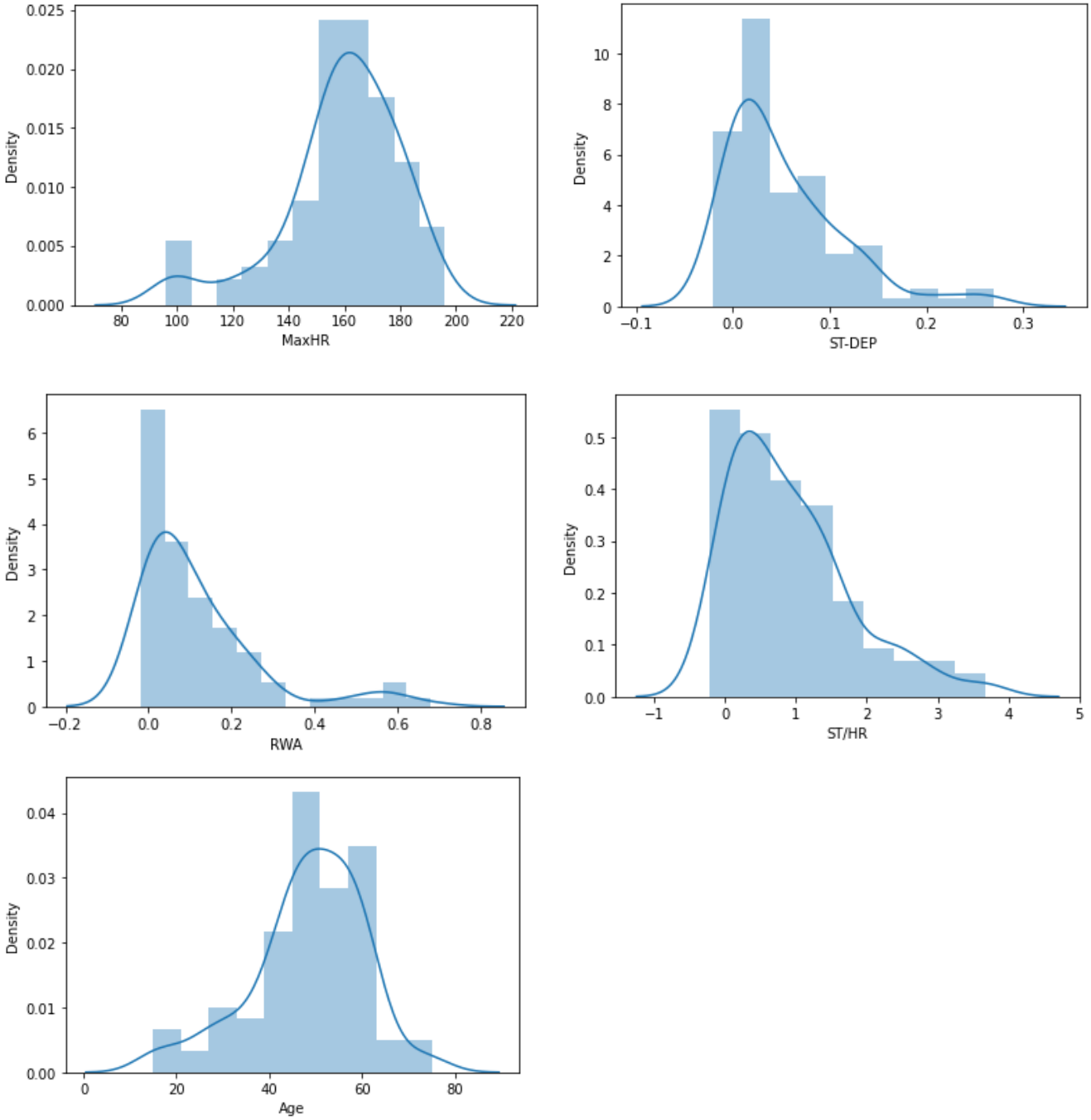


Figure 2: Distribution of different variables in patients not having CAD

2. Comparing the distributions of two groups

2.1. ST depression, ST/HR index, ΔRWA and maxHR.

The statistical tests for finding the difference is commonly done to make sure that the observed difference is not occurring merely by chance. This helps us to make validate any conclusions made regarding the dataset. To test the statistical difference, we first inspect the distribution of the dataset and select the parametric or non-parametric test. In case of maxHR, the data is normally distributed in CAD population but not normally distributed in non-CAD population (normality is evaluated in section 1.2). In that case, we assume that both groups are not normally distributed and select a non-parametric test. We use Mann Whitney U test which is used to test difference between two samples by checking the difference between their medians. The assumption for this test is that the data is ordinal and non-normal. Therefore, when applying the test, the software will first rank the continuous values of the dataset and then evaluate the statistics.

For ST Depression, ST/HR index and RWA, the data in both CAD and non-CAD is not normally distributed, therefore, Mann Whitney U can be used for them as well. We can use following hypothesis for our problem:

Null hypothesis (H0): There is no statistical difference

Alternative hypothesis (H1): There is a statistical difference.

After applying the test, the following statistics were obtained.

Group1	Group2	Test applied	Test statistic	P-value
MaxHR in patients who don't have CAD	MaxHR in patients who have CAD	Mann Whitney U Test	8812.0	1.229e-20 < 0.05 H0 is rejected
ST Depression in patients who don't have CAD	ST Depression in patients who have CAD	Mann Whitney U Test	2532.5	1.564e-09 < 0.05 H0 is rejected
ST/HR index in patients who don't have CAD	ST/HR index in patients who have CAD	Mann Whitney U Test	1758.5	2.378e-15 < 0.05 H0 is rejected
ΔRWA in patients who don't have CAD	ΔRWA in patients who have CAD	Mann Whitney U Test	3167.5	7.507e-06 < 0.05 H0 is rejected

Based on the obtained results, we can conclude there is a significant statistical difference between the two groups of maxHR in CAD and maxHR in non-CAD with a risk level of under 5% of being wrong. The same is also true for ST Depression, ST/HR index and RWA in CAD and non-CAD patients.

2.2. Age and Sex

a)

To select the parametric or non-parametric test, we look at the distributions of age, described in 1.2, in CAD and non-CAD patients. We can see that the data related to age in CAD and non-patients is not normally distributed, so we use Mann Whitney U test. The hypothesis are as follows:

Null hypothesis (H0): There is no statistical difference between ages of CAD and non-CAD patients.

Alternative hypothesis (H1): There is a statistical difference between ages of CAD and non-CAD patients

The test statistics are calculated as follows:

Group1	Group2	Test applied	Test statistic	P-value
Age in patients who don't have CAD	Age in patients who have CAD	Mann Whitney U Test	3590.0	0.00056 < 0.05 H0 is rejected

By looking at the p-value and rejecting null hypothesis, we can conclude that the two groups are statistically different from each. The median age of CAD is greater than median age of non-CAD patients by 5 months as indicated in section 1.1, we can say that the patients in former group are older than patients in latter group with under 5% risk level of being wrong.

b)

The non-parametric test for categorical data is Chi2 test. It tests the independence of two variables and calculates the expected frequency for each category and compares it with the observed frequency. Since, the categorical data do not form ordinal scale and is not normally distributed, it is considered non-parametric. We want to test if male or female have different disease classification in our population. If the classification in them is different, we can say that disease can depend on being male or female but if the classification is same, it would be

indicative of disease not depending on sex. The hypothesis to test the statistical difference between two groups is:

Null hypothesis (H0): There is no difference between sex (male or female) and disease (CAD or No CAD)

Alternative hypothesis (H1): There is a difference between sex (male and female) and disease (CAD or No CAD)

Disease	Male	Female	Test applied	Test statistic	P-value
Non-CAD	50	50	Chi2	13.406	0.00025 < 0.05 H0 rejected.
CAD	76	24			

Since, we were able to reject null hypothesis based on the obtained statistics we can say that there is a statistically significant difference between gender and disease with a risk level of under 5% to be wrong. It indicates that disease classification can depend on being male or female. The population of male in CAD patients is 76% which is higher than females. Based on this, we can conclude that disease can depend on being male however, the sample size is relatively small to make a population wide conclusion. Moreover, there might be some other confounding factors for example, genetic factors associated with males and females that lead to developing the disease.

2.3. Age and MaxHR differences in CAD patients

The normality of age distributions and maxHR in females and males having CAD has been tested (in section 1.2). For age, the distribution is assumed non-normal for both male and female groups (however, one is normal, and one is non-normal, according to the test). This satisfies the condition of Mann Whitney U test and based on the rank of each value in the dataset, statistic values are evaluated.

For maxHR, the distribution in both groups is normal (in section 1.2). The data is interval scale and normally distributed, which satisfy the condition of parametric two-tailed independent t-test. Our hypothesis is following:

Null hypothesis (H0): The age of female and male are not different.

Alternative hypothesis (H1): The age of female and male are different.

Group1	Group2	Test applied	Test statistic	P-value
Age of females in patients who have CAD	Age of males in patients who have CAD	Mann Whitney U Test	1052.5	0.258 > 0.05 H0 accepted
maxHR of females in patients who have CAD	maxHR in patients who have CAD	Independent t-test	-0.546	0.586 > 0.05 H0 accepted

Based on the p-value, we can say that the ages of female and male are statistically same. This can be interpreted as any observed differences in their means is merely by chance and do not hold any significant value statistically.

3. Correlation and regression between two variables

3.1 a)

To study the linear correlation between ST depression, ST/HR index and RWA, normality test is conducted (in section 1.2). All the three variables turned out to be non-normal therefore, spearman correlation is applied. The hypothesis for the test is:

Null hypothesis (H0): The correlation between two groups does not exist.

Alternative hypothesis (H1): The correlation between two groups exists.

The test statistics is then calculated:

Group1	Group2	Test applied	Correlation coefficient	P-value
ST depression in patients who have CAD	ST/HR in patients who have CAD	Spearman Correlation	0.783	5.86e-22 < 0.05 H0 is rejected.
ST depression in patients who have CAD	RWA in patients who have CAD	Spearman Correlation	0.078	0.439 > 0.05 H0 is accepted
ST/HR in patients who have CAD	RWA in patients who have CAD	Spearman Correlation	0.011	0.909 > 0.05 H0 is accepted

ST depression in patients who don't have CAD	ST/HR index in patients who don't have CAD	Spearman Correlation	0.724	1.65e-17 < 0.05 H0 is rejected
ST depression in patients who don't have CAD	RWA in patients who don't have CAD	Spearman Correlation	0.0991	0.326 > 0.05 H0 accepted
ST/HR in patients who don't have CAD	RWA in patients who don't have CAD	Spearman Correlation	0.0318	0.753 > 0.05 H0 accepted

The linear correlations were found in ST Depression and ST/HR in CAD patients and ST/HR and RWA in non-CAD patients. It is important to note that the sample size is small so only strong correlations should be considered. Therefore, ST Depression and ST/HR in CAD are 78.3% correlated, ST Depression and ST/HR in non-CAD patients are 72.4% correlated which is quite strong.

3.2

Linear regression is used to evaluate the relationship between two variables. The equation of linear regression is linear function of slope, baseline constant and dependent variable. For maxHR and age in CAD patients, the slope was found to be -0.883 and constant was 201.8. The equation is given in the table below. For maxHR and age in non-CAD patients, the slope was -0.88 and constant was 170.2 and the equation is given below. To test the significance of correlation, we use the following hypothesis:

Null hypothesis (H0): There is no correlation between independent and dependent variables (slope = 0)

Alternative hypothesis (H1): There is a correlation between independent and dependent variables (slope ≠ 0)

The computed statistics and regression plots are given in the table below:

Independent Variable	Dependent Variable	Equation	R-value	R^2-value	P-value
Age in patients who don't have CAD	maxHR in patients who don't have CAD	maxHR = 201.8 – 0.883 x age	-0.505	0.255	7.942e-08 < 0.05 H0 rejected
Age in patients who have CAD	maxHR in patients who have CAD	maxHR = 170.2 – 0.88 x age	-0.341	0.116	0.000514 < 0.05 H0 rejected

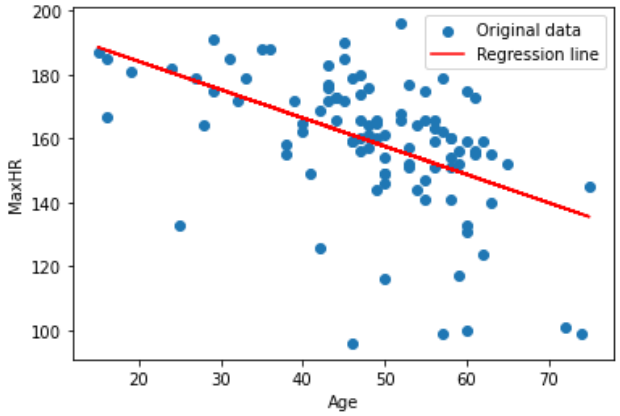


Fig: Regression plot (No CAD patients)

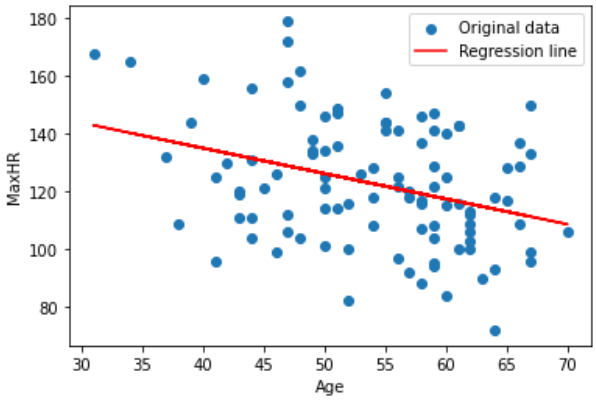


Fig: Regression plot (CAD patients)

According to the obtained results, we found that there is a significant correlation between maxHR and age in both CAD and non-CAD patients. The equation for no-CAD is approximately equal to normal heart in our population which shows that our model represents the general population very well. While the equation in CAD patients deviates more from normal heart rate which indicates some abnormality in the heart rate of those patients. Based on the coefficient of determination (R^2), we can say that the 25.5% and 11.6% variance in maxHR in non-CAD and CAD patients respectively can be explained by age. The strength of the correlation is relatively stronger in no-CAD patients than in CAD patients as the R (0.505) is closer to 1 and R (0.34) is closer to zero, in each case. The final decision should be made on sample size and the requirement of robustness. Since, we have relatively small size, we can conclude that CAD patients have weak correlation between MaxHR and Age while non-CAD has acceptable correlation but not very strong.

3.3

The regression analysis was conducted between maxHR and ST depression in total population, and between maxHR and ST/HR index. The hypothesis is formulated as following:

Null hypothesis (H0): There is no relationship between two variables.

Alternative hypothesis (H1): There is a relationship between two variables.

The statistics computed as given as follows:

Independent variable	Dependent variable	R value	R^2 value	P value
maxHR in total population	ST depression in total population	-0.293	0.0858	2.414e-05 < 0.05 H0 rejected
maxHR in total population	ST/HR index in total population	-0.551	0.3036	2.536e-17 < 0.05 H0 rejected

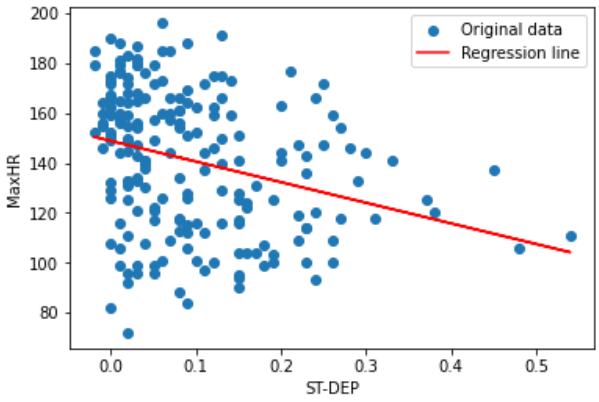


Fig: Regression plot for ST Depression

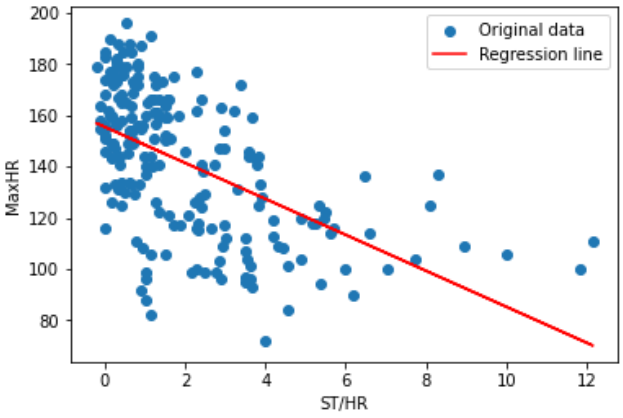


Fig: Regression plot for ST/HR index

The p-values for both tell us that there is a relationship between maxHR and ST depression, and between maxHR and ST/HR index. The former has weak correlation (0.29 close to zero) and the latter has acceptable but not very strong (0.551 close to 1), however, both of them are negative indicating that with the decrease in ST depression or ST/HR, maxHR also decreases. The correlation of determination (R^2) tells us that 8.5% variation in maxHR can be explained by

ST depression while 30.3% variation in maxHR can be explained by ST/HR index. This shows that ST/HR might be a good indicator for predicting maxHR of a patient.

3.4

The relationship between age and RWA is tested in total population using Spearman correlation. The spearman correlation is suited for data that is not normally distributed and the normality for age and RWA is evaluated in section 1.2. The hypothesis used are given as:

Null hypothesis (H0): There is no correlation between age and RWA in total population

Alternative hypothesis (H1): There is a correlation between age and RWA in total population

The statistics are computed as follows:

Group1	Group2	Test applied	Test statistic	P-value
Age in total population	RWA in total population	Spearman Correlation	0.193	0.00604 < 0.05 H0 is rejected

Based on the results, we can conclude that there is a positive correlation between age and RWA in total population. However, the strength of correlation is quite weak (0.193 close to 0). This might occur due to differences in sample size as the sample size in each category is 50% less than the total population. Other possible reasons may include, existence of confounding factors that are affecting the results.

4. Estimating the accuracy of diagnostic classifiers

4.1.

To move forward with the evaluation of accuracy with different parameters, we first need to find true positives (all positive diseased cases that were detected positive by the classifier), true negative (all negative diseased cases that were detected negative by the classifier), false positive (all negative diseased cases that were detected positive by the classifier) and false negative (all positive diseased cases that were detected negative by the classifier). Contingency table was created for ST Depression, ST/HR index and RWA at given partition values. The required parameters are evaluated as:

Sensitivity: TP / (TP + FN)

Specificity: TN / (TN + FP)

Diagnostic accuracy: (TP + TN) / Total diseased cases (N)

Diagnostic performance: (Sensitivity + Specificity) / 2

Positive Prediction value (PV+): TP / (TP + FP)

Negative prediction value (PV-): TN / (TN + FN)

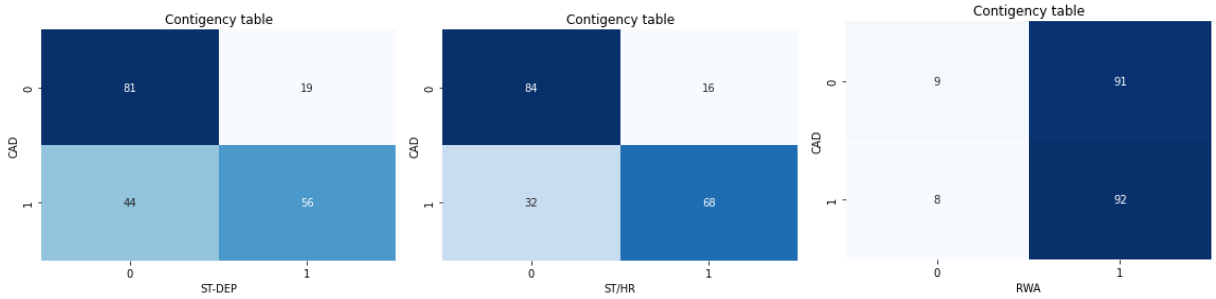


Fig: Contingency table for variables

Based on the values of contingency table, all parameters are evaluated as follow in total and in female population:

Classifier	True Negative	True Positive	False Negative	False positive	Sensitivity	Specificity
In total population (N=200)						
ST Depression	81	56	44	19	56/(56+44) =0.56	81/(19+81) =0.81
ST/HR index	84	68	32	16	68/(68+32) = 0.68	84/(16+84) = 0.84
RWA	9	92	8	91	92/(92+8) = 0.92	9/(91+9)= 0.09
In female population (N=74)						
ST Depression	36	13	11	14	13/(13+11) = 0.54	36/(36+14) =0.72

ST/HR index	38	18	6	12	18/(18+6) = 0.75	38/(38+12) = 0.76
RWA	7	19	5	43	19/(19+5) = 0.79	7/(7+43) = 0.14

Classifier	Diagnostic accuracy	Diagnostic performance	Positive predictive value	Negative predictive value
In total population				
ST Depression	(81+56)/200 = 0.685	(0.56+0.81)/2 = 0.685	56/(56+19) = 0.746	81/(81+44) = 0.648
ST/HR index	(68+84)/200 = 0.76	(0.68+0.84)/2 = 0.76	68/(68+16) =0.809	84/(84+32) = 0.724
RWA	(92+9)/200 = 0.505	(0.92+0.09)/2 =0.505	92/(92+91) = 0.502	9/(9+8) = 0.529
In female population				
ST Depression	36+13/74 = 0.66	0.54+0.72/2 = 0.63	13/13+14 = 0.48	36/36+11 = 0.76
ST/HR index	38+18/74 = 0.75	0.75+0.76/2 = 0.75	18/18+12 = 0.6	38/38+6 = 0.86
RWA	7+19/74 = 0.35	0.14+0.79/2 = 0.46	19/19+43 = 0.306	7/7+5 = 0.58

After the analysis for each, we found ST/HR index to have a better trade-off between sensitivity (68% for total, 75% for female population) and specificity (84% for total, 76% for female population) as compared to other two classifiers. The diagnostic accuracy and diagnostic performance are also better in ST/HR index in both total and female population. PV+ is 80.9% in total population and 60% in female population while PV- is 72.4% in total and 86% in female population which is also higher than other two classifiers.

4.2

a) The contingency table is as follows for each partition value:

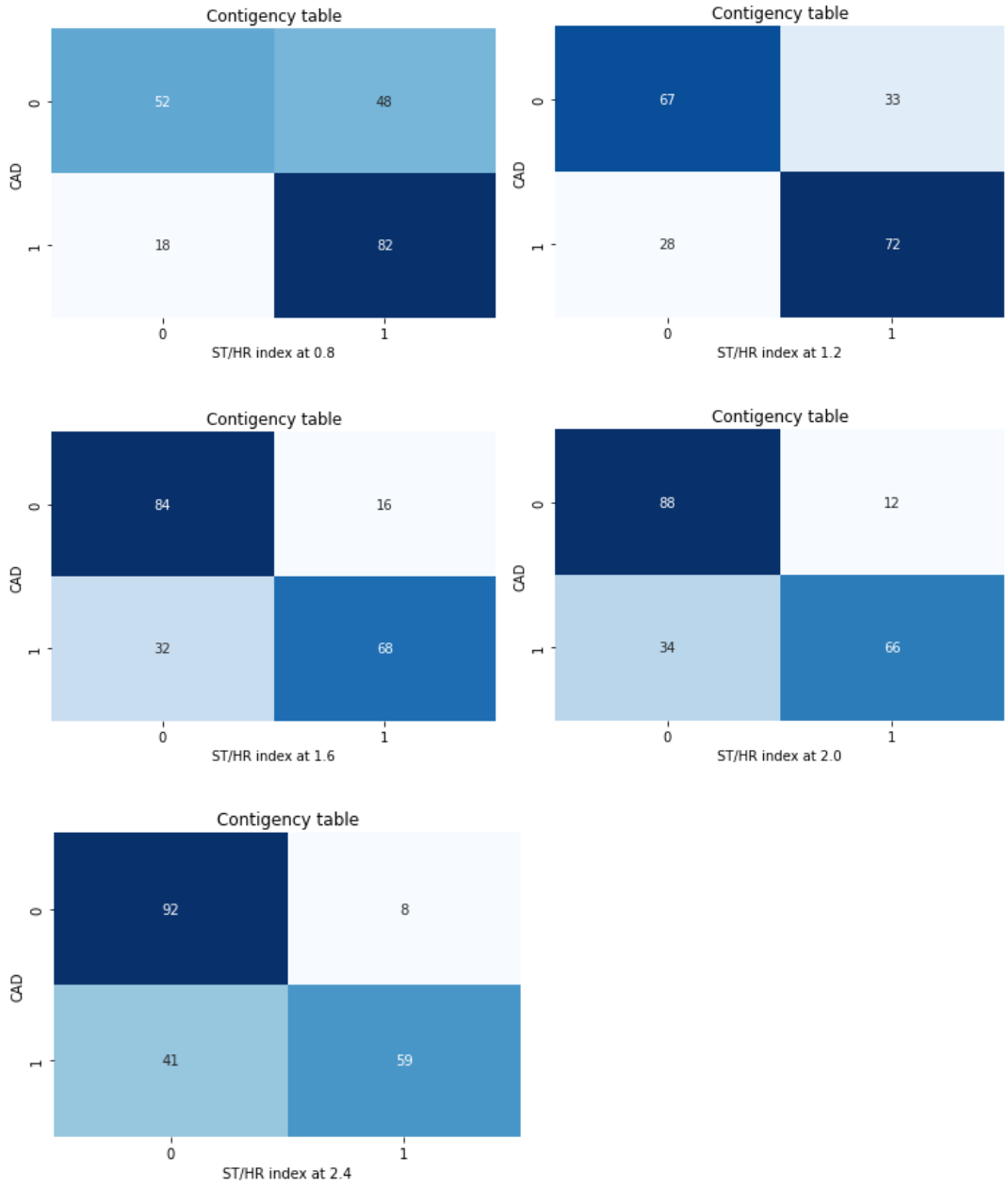


Fig: Contingency table at each partition value

The sensitivity and specificity are then calculated based on the formulae mentioned in section 4.1 The results are as follows:

Partition values µV/bpm	Sensitivity	Specificity
0.8	0.82	0.52
1.2	0.72	0.67
1.6	0.68	0.84
2.0	0.66	0.88
2.1	0.59	0.92

b) If the prevalence is 20% in 200 population, this indicates that 40 people have CAD disease and 160 people do not have CAD disease. The sensitivity and specificity of ST/HR at 2.0 µV/bpm is 0.66 and 0.88 respectively. With the given information, we can calculate true positives and true positives that are required for calculating diagnostic accuracy.

True Positives (TP) = Diseased cases x Sensitivity

TP = 40 x 0.66 = 26.4 ≈ 24

True Negatives (TN) = Non-diseased cases x Specificity

TN = 160 x 0.88 = 141

Diagnostic accuracy = (TP + TN) / Total population

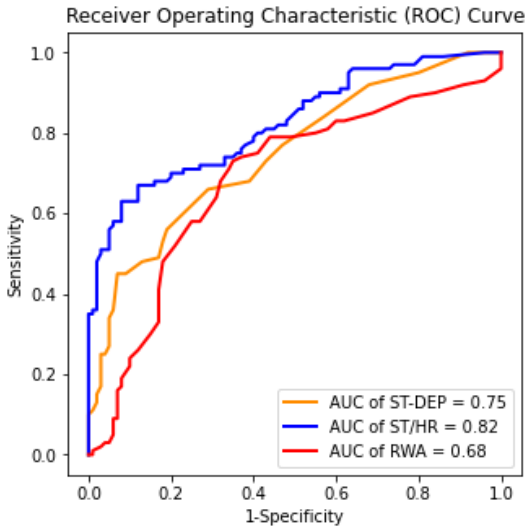
Diagnostic accuracy = (24+141) / 200 = 0.825 or 82.5%

The same evaluation is done for each partition value and the results are:

Partition values (µV/bpm)	Diagnostic accuracy
0.8	0.58 or 58%
1.2	0.68 or 68%
1.6	0.705 or 70.5%
2.0	0.825 or 82.5%
2.1	0.855 or 85.5%

The diagnostic accuracy of ST/HR is highest (85.5%) for partition value of 2.1 µV/bpm.

c) The ROC curve for all diagnostic classifiers is illustrated below:



Based on the ROC curve, area under the curve is highest for ST/HR index with the value of 0.82.

4.3

a) To determine at sensitivity at 85% specificity, we look at the coordinates of ROC curve and find the sensitivity value. In the ROC curve, the coordinates are available for 1 - Specificity, therefore, to find sensitivity coordinate at 85% or 0.85 specificity, 1 – Specificity coordinate is 0.15. The coordinates are shown in the figure below:

Index	Sensitivty	1-Specificity
10	0.27	0.05
11	0.3	0.05
12	0.34	0.05
13	0.36	0.06
14	0.45	0.07
15	0.45	0.09
16	0.48	0.13
17	0.49	0.17
18	0.53	0.18
19	0.56	0.19

Fig: ST Depression

Index	Sensitivty	1-Specificity
14	0.56	0.05
15	0.57	0.06
16	0.58	0.06
17	0.58	0.08
18	0.6	0.08
19	0.62	0.08
20	0.63	0.08
21	0.63	0.12
22	0.67	0.12
23	0.67	0.16
24	0.68	0.16
25	0.68	0.19

Fig: ST/HR

Index	Sensitivty	1-Specificity
13	0.16	0.07
14	0.17	0.08
15	0.19	0.08
16	0.22	0.1
17	0.24	0.1
18	0.26	0.12
19	0.3	0.15
20	0.33	0.17
21	0.35	0.17
22	0.41	0.17
23	0.48	0.18
24	0.52	0.21

Fig: RWA

The sensitivity for ST Depression was 0.49 or 49% at 0.17 (close to 0.15) 1 – Specificity. The sensitivity for ST/HR was 0.67 or 67% at 0.16 1 – Specificity. The sensitivity for RWA was 0.3 or 30% at 0.15 1 – Specificity. ST/HR was found to have the highest sensitivity at 85% specificity.

b) Since we now know the sensitivity values at 85% specificity, we can now calculate true positives and false positives for both ST Depression and ST/HR index. McNemar test is chosen to evaluate the statistical significance between ST depression and ST/HR index. This test has two assumptions; data is non-normal (normality is evaluated in section 1.2) and nominal, the test groups are matched pairs which means that both data should overlap with each other. Assuming that observed data overlaps with each other i.e. 67 data points in ST/HR index are the same for exact 49 data points in CAD and 33 data points in ST/HR index are the same for 51 data points in non-CAD patients, we can use the following hypothesis:

Null hypothesis (H0): There is no significant difference between two groups.

Alternative hypothesis (H1): There is a significant difference between two groups.

	ST Depression	ST/HR index	Total	P-value for McNemar
CAD	100 x 0.49 = 49	100 x 0.67 = 67	100	0.0008 < 0.05 H0 is rejected
No-CAD	100 – 49 = 51	100 – 67 = 33	100	
Total	100	100	200	

Based on the results, we can say there is a significant difference between sensitivity of ST depression and ST/HR index at 85% specificity with a risk level of 5% to be wrong. This indicates that sensitivity of 67% in ST/HR is statistically better than sensitivity of 49% in ST depression and the observed higher value of in the former is not observed merely by chance.

EXTRA: 4.4

a) If the prevalence is 15% in 6000,000 population, this indicates that 900000 people have a disease and 5100000 people do not have a disease. The sensitivity and specificity of ST depression is 0.56 and 0.81 respectively. With the given information, we can calculate true positives and true positives that can help us in identifying false positives and false negatives.

True Positives (TP) = Diseased cases x Sensitivity

TP = 900000 x 0.56 = 504000

False Negatives (FN) = 900000 – 504000 = 396000

True Negatives (TN) = Non-diseased cases x Specificity

TN = 5100000 x 0.81 = 4131000

False Positives (FP) = 5100000 – 4131000 = 969000

Therefore, the total misdiagnosed patients are the sum of FP and FN which is 1365000 approx. 1.37 million people.

b) Cost spent on unncecessary treatment when cost per patient is 600 EUR:

Cost for misdiagnosed positive patients = 969000 x 600 = 581,400,000 EUR

c) To find the improvement by ST/HR, we need to first find FP and TP cases in the population. The sensitivity is 0.68 and specificity is 0.84 using which we can calculate TP and TN.

True Positives (TP) = Diseased cases x Sensitivity

TP = 900000 x 0.68 = 612000

False Negatives (FN) = 900000 – 612000 = 288000

Improvement in False Negatives = 396000 – 288000 = 108000 or 12%

True Negatives (TN) = Non-diseased cases x Specificity

TN = 5100000 x 0.84 = 4284000

False Positives (FP) = 5100000 – 4284000 = 816000

Improvement is FP = 969000 – 816000 = 153000 or 15.7%

When assessing the obtained values, we can say that total *1104000* (TN+TP) or approx. 1.10 million cases are incorrectly diagnosed by ST/HT index classifier which is lower than ST depression.

d) The total cost for unnecessary treatments caused by ST/HR index when 600 EUR per patient is spent can be calculated as: *FP x cost per patient = 816000 x 600 = 489600000 EUR*

The money saved by ST/HR index as a classifier when compared to ST Depression can now be found: *581,400,000 – 489600000 = 91800000 EUR.* This shows that ST/HR index as a classifier will save 91.8 million EUR by reducing any unnecessary treatments.

Feedback:

The project was very lucrative to my learning and I had a good time doing it. I had some prior knowledge of statistics, so I did not find it intermediate and not extremely hard. However, it took quite a bit of my work time. It provided good basic understanding of statistics especially medical statistics. I wanted to learn this aspect because I am doing my research work in the same area which is also why I conducted the analysis on Python. I added most of results in a tabular format obtained from Python because results were not organized well on Python terminal.

During the whole project, I was able to clarify some of my doubts about why some tests are suitable for certain conditions with the help of lectures and exercises. Moreover, I would suggest that project work is given earlier in the course duration and some assignments (related to the project or even some parts of the project) can be made a requirement during the course. Other than that, I had a good learning experience throughout the course.