INTRODUCTION

This purpose of this report is to perform classification using kNN and Decision Trees algorithm on the given Heart Disease dataset. The dataset was obtained on this webpage hosted by the University of California, Irvine (UCI).

BUSINESS UNDERSTANDING

Business Objective: It is to identify which medical and general attributes contributed to diagnosis of heart disease. Therefore, helping with predicting patients with high risk.

The dataset was published in July 1988. The locations where it was gathered and the researchers responsible are as follows:

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

- V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

DATA UNDERSTANDING

➢ COLLECT INITIAL DATA

The initial dataset has been collected beforehand and additional datasets are outside the scope of this report. Relevant domain knowledge concerning heart disease will be gathered and used when it is applicable.

➢ DESCRIBE DATA

The dataset consists of data concerning heart disease diagnosis from 4 different locations. This report will only use the processed data from Cleveland Clinic Foundation (processed.cleveland.data file). Compared to the raw data, the processed file only has 14 attributes instead of 76 attributes. This report follows the existing precedence set

by previous Machine Learning researchers by only using data already processed and coming from Cleveland.

The data is in CSV (comma-separated values) format and consists of 303 rows/instances. The data file does not have headers.

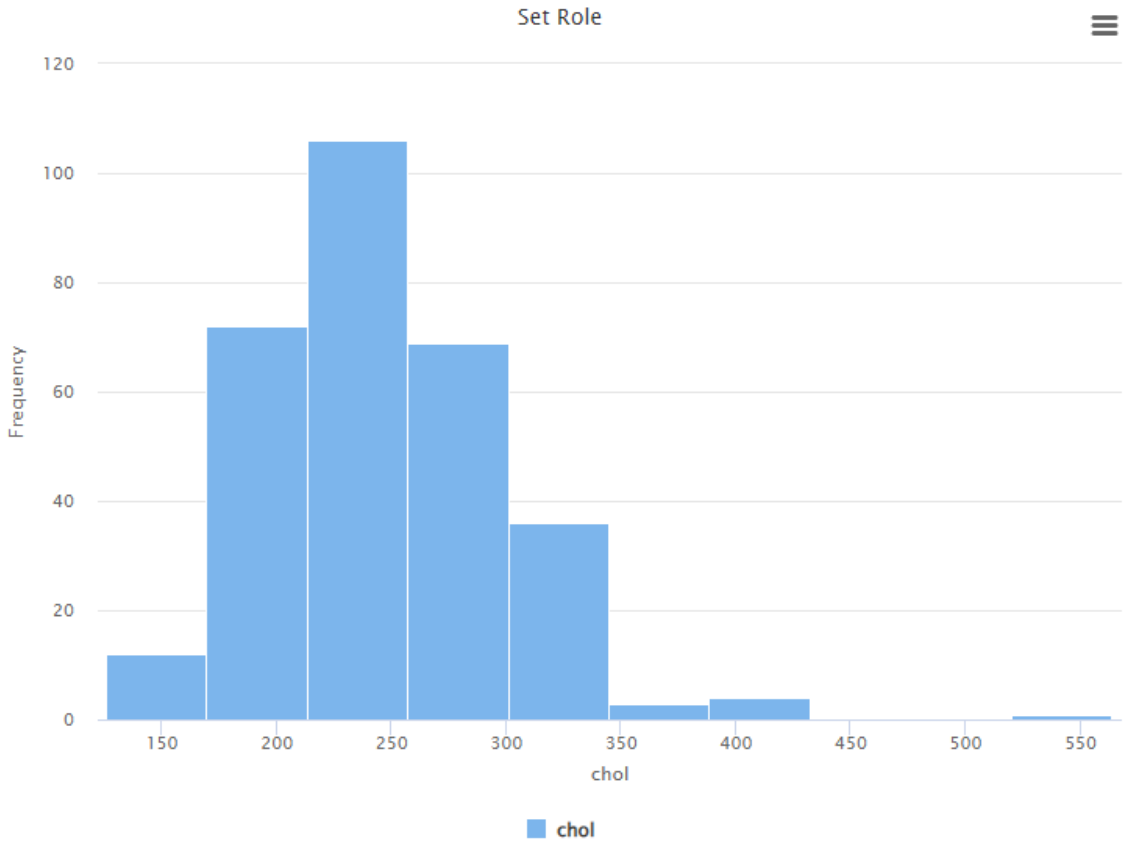
- **EXPLORE DATA**

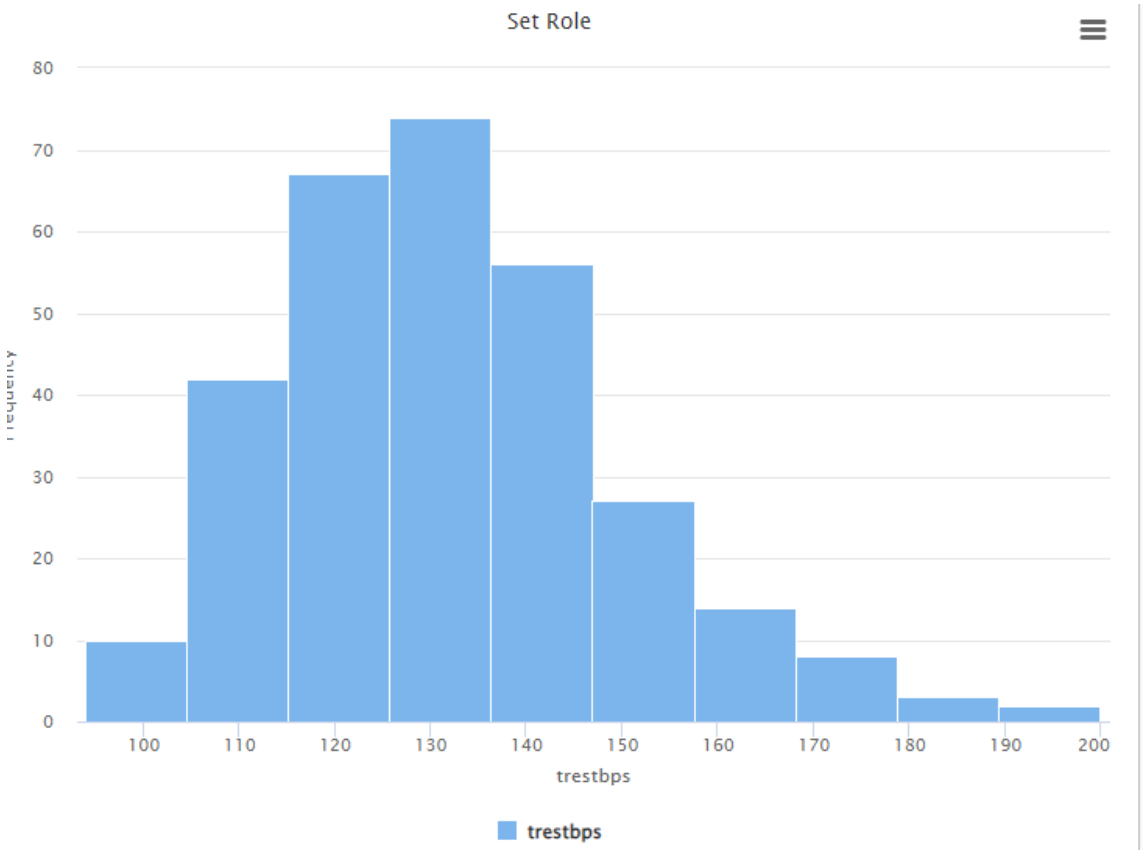
There are 14 attributes in the dataset and the UCI webpage explains what they are.

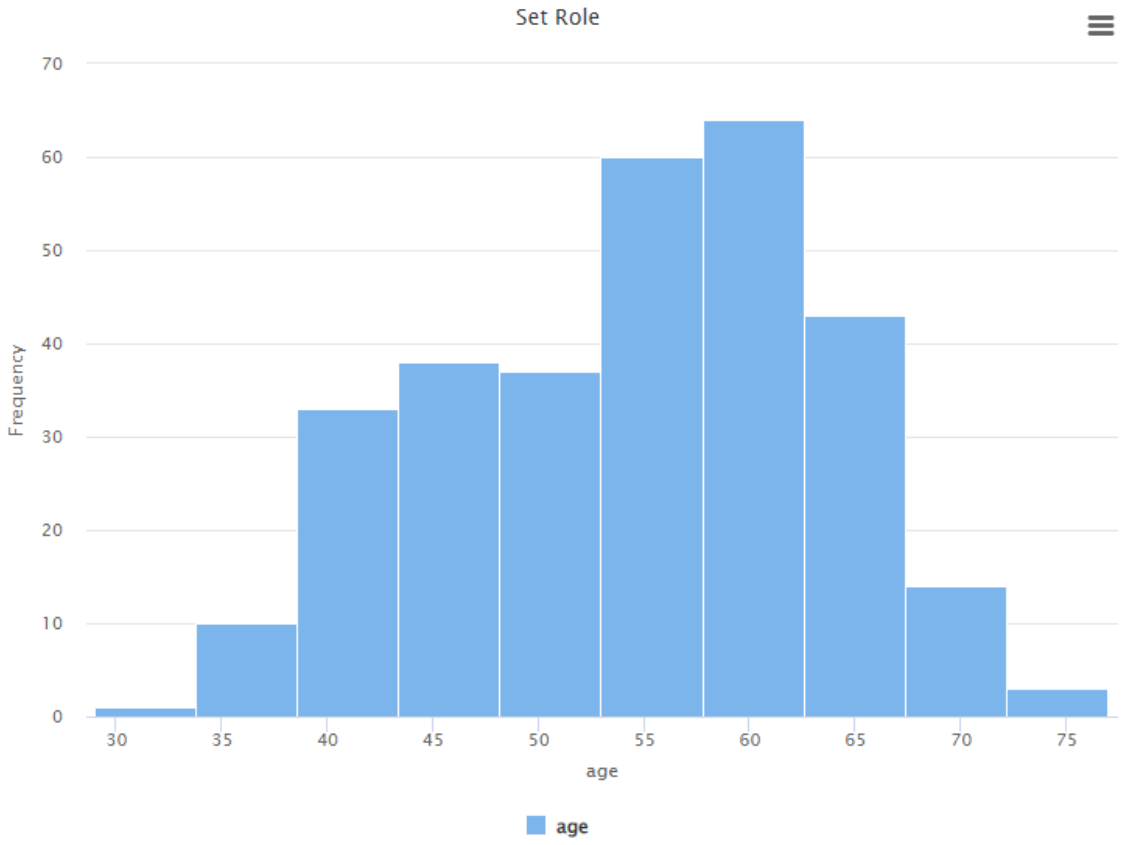
NAME	DESCRIPTION	DATA TYPE (RapidMiner)	DATA QUALITY ISSUES
age	age of patient in years	integer	
sex	(patient's sex) 0 for female, 1 for male	binomial	
cp	(chest pain) 1 for typical angina, 2 for atypical angina, 3 for non-anginal pain, 4 for asymptomatic	polynomial	
trestbps	resting blood pressure in mm Hg on admission to the hospital	integer	
chol	serum cholesterol in mg/dl	integer	outlier present
fbs	(fasting blood sugar > 120 mg/dl) 1 for true, 0 for false	binomial	
restecg	(resting electrocardiographic results) 0 for normal, 1 for having ST-T wave abnormality, 2 for showing probable or definite left ventricular hypertrophy	polynomial	
thalach	maximum heart rate achieved	integer	
exang	(exercise induced angina) 1 for yes, 0 for no	binomial	
oldpeak	ST depression induced by exercise relative to rest	real	

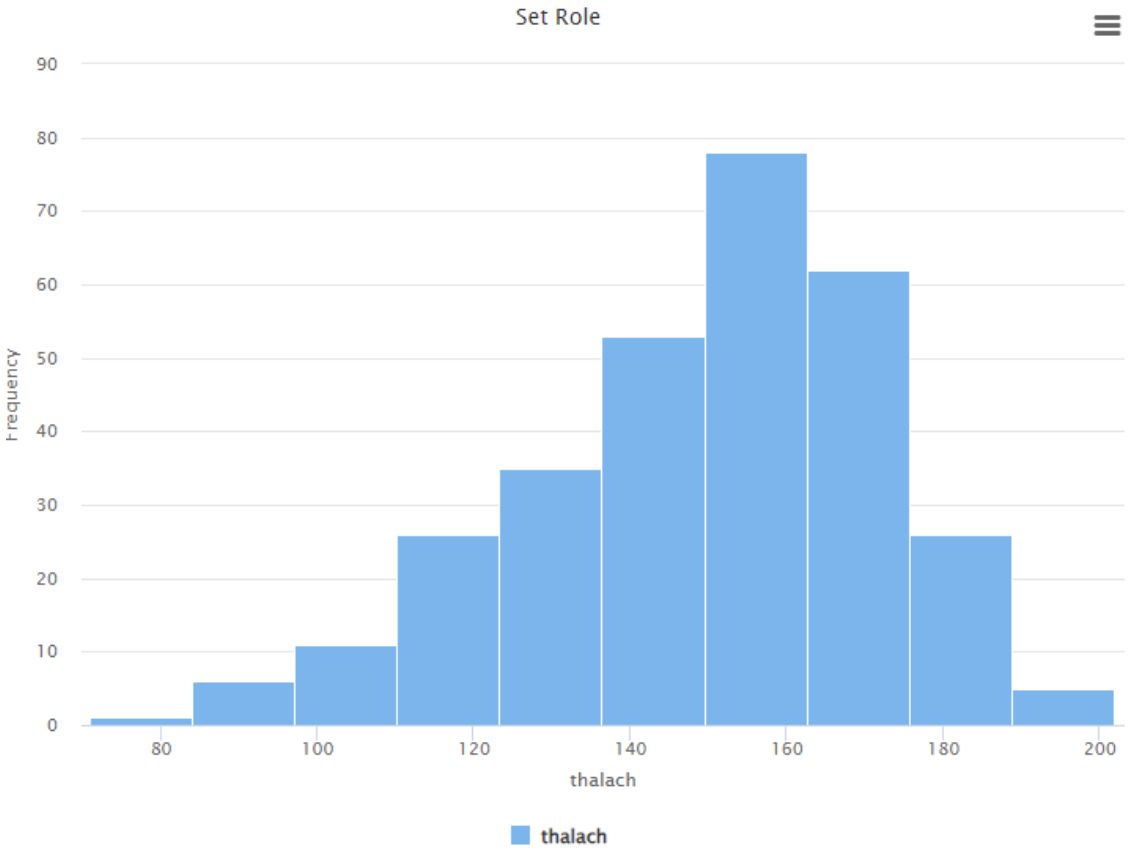
slope	(the slope of the peak exercise ST segment) 1 for upslope, 2 for flat, 3 for downslope	polynomial	
ca	number of major vessels colored by fluoroscopy	integer	missing values
thal	(thallium stress tests results) 3 for normal, 6 for fixed defect, 7 for reversable defect	polynomial	missing values
num	(diagnosis of heart disease) 0 for absence, 1-4 for presence	polynomial	

From the 6 columns with numerical data, 4 attributes (age, chol, trestbps, thalach) follow a normal distribution.

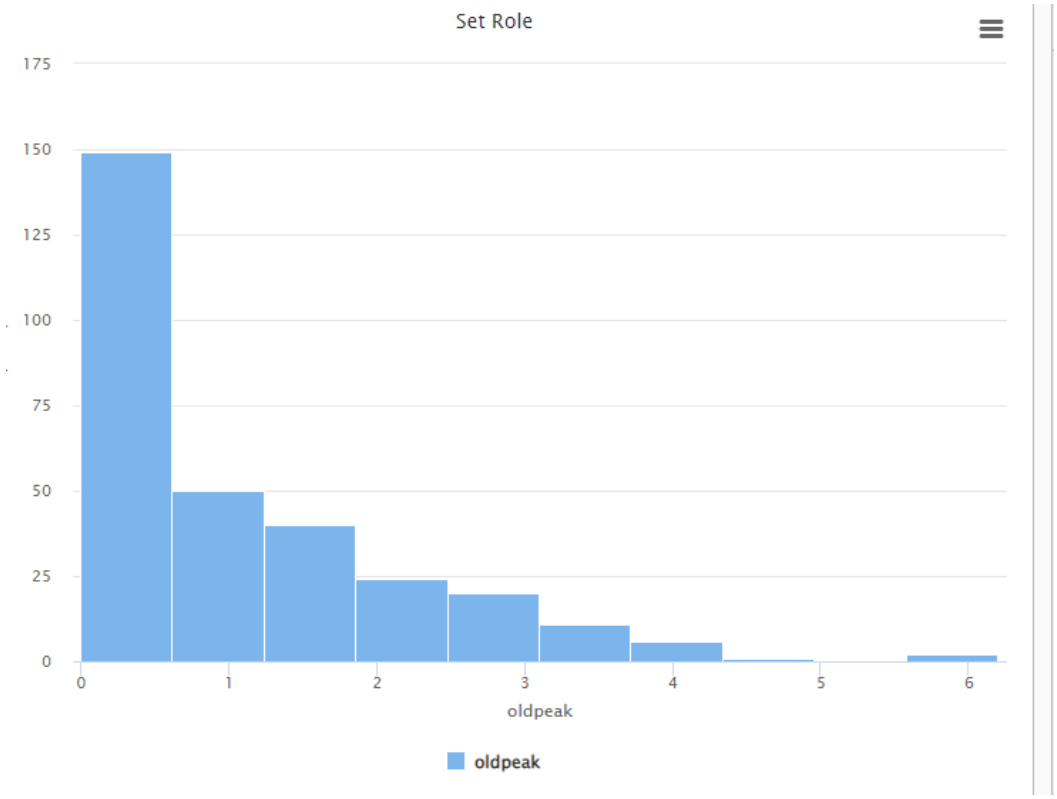


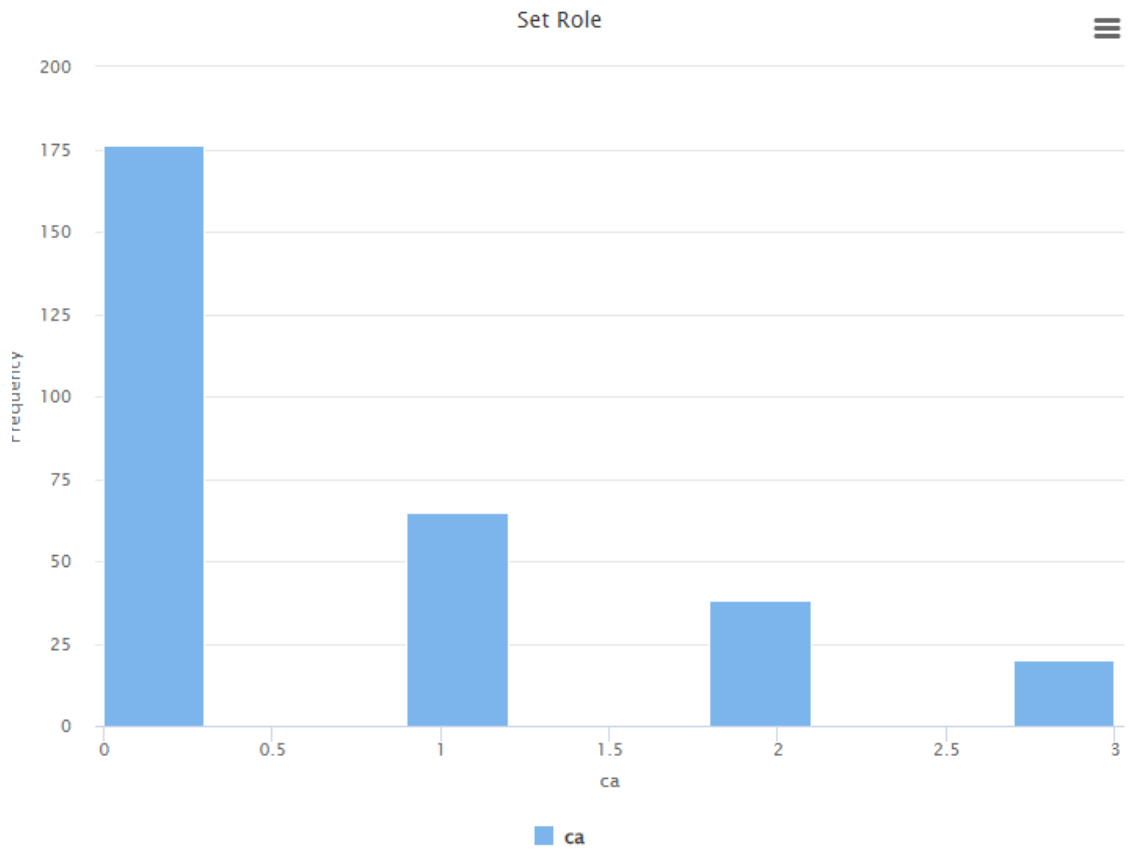

Set Role
80
70
60
50
40
30
20
10
0
100
110
120
130
140
150
160
170
180
190
200
trestbps
trestbps


Set Role
70
60
50
40
30
20
10
0
Frequency
30
35
40
45
50
55
60
65
70
75
age
age



2 attributes (oldpeak, ca) are left-skewed.





- **VERIFY DATA QUALITY**

There are some missing values in the “ca” and “thal” columns.

There is an outlier present in the “chol” column that have much higher values than the rest of the data.

The entries in several columns containing numerical values have decimal places but those decimal places are all zeroes. For this report, those columns will be set to “integer” instead of “real” datatype in RapidMiner.

DATA PREPARATION

- **SELECT DATA**

The various data attributes had already been processed and selected by previous researchers to exclude irrelevant attributes. This report will use all attributes in the processed data file.

- **CLEAN DATA**

The missing values in the “ca” column will be replaced by median value.

The missing values in the “thal” column will be replaced by the modal value.

Row No.	median(ca)	mode(thal)
1	0	3.0

- **CONSTRUCT DATA**

A new instance based on student number 41118874 has been inserted into the data as follows:

- age: 37
- sex: 1
- cp: 4
- trestbps:111
- chol: 118
- fbs: 1
- restecg:7
- thalach: 874
- exang: 1
- oldpeak: 1.1
- slope: 1
- ca: 8
- thal: 4
- num: 0

- **INTEGRATE DATA**

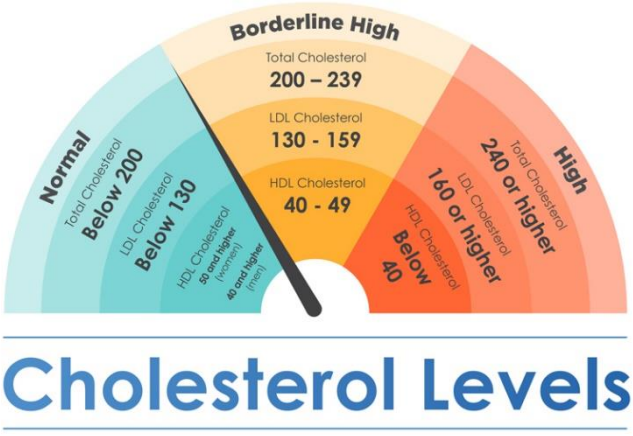
No additional data sources are integrated.

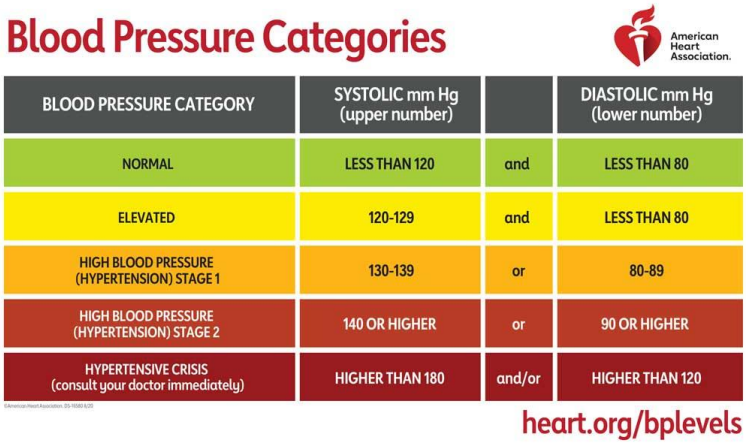
➢ FORMAT DATA

The column “num” indicates the diagnosis of heart disease with 0 as absence and 1-4 as presence with larger numbers having more severity. It will be binned and replaced with a binomial column “disease”. The new value of 0 representing absence will replace the old value of 0. The new value of 1 representing presence will replace the old value of 1-4. This report is more interested in the presence/absence of disease rather than its severity. The binning may also improve the accuracy of the classification models. The “disease” column will be used as the class attribute.

In order to improve the k-NN model accuracy, normalization is performed on the 2 left-skewed columns (oldpeak, ca). Standardization is also performed on the 4 columns with normal distribution (age, chol, trestbps, thalach).

In order to improve the Decision Tree model, the numerical columns will be binned according to relevant domain knowledge if possible. Age will be binned by decades into values of ‘thirties’, forties’, ‘fifties’, ‘sixties’, and ‘seventies’. Trestbps (resting blood pressure) will be binned into values of ‘normal’, ‘elevated’, ‘hypertension 1’, ‘hypertension 2’, and ‘hypertensive crisis’. Chol (cholesterol level) will be binned into ‘normal’, ‘borderline high’, ‘high’.





Oldpeak (ST depression induced by exercise) will be binned by frequency into ‘range1’, ’range2’, ’range3’, ’range4’, ’range5’. Ca will be binned into ‘zero’, ‘one’, ‘two’, ‘three’, ‘above three’. Thalach (maximum heart rate achieved) will be binned by frequency into ‘range1’, ‘range2’, ’range3’, ’range4’, and ’range5’.

MODELING & EVALUATION

- **TEST DESIGN**

This report will use 10-fold cross validation to split the data for both k-NN and Decision Tree.

- **BUILD MODEL**
  - **K-NN**

The k-NN model will use 4 different values for k (3,5,7,9). Results will be evaluated to determine the best value for k.

  - **DECISION TREE**

The decision Tree will use the Optimize Parameters operator to try improving its accuracy. The following parameters in the Decision Tree operator will be optimized: maximal depth, minimal leaf size, minimal size for split, apply pruning, and apply pre-pruning.

- **ASSESS MODEL**
  - **K-NN**

K = 3:

accuracy: 75.60% +/- 10.34% (micro average: 75.66%)

	true false	true true	class precision
pred. false	131	40	76.61%
pred. true	34	99	74.44%
class recall	79.39%	71.22%	

K = 5:

accuracy: 79.23% +/- 8.22% (micro average: 79.28%)

	true false	true true	class precision
pred. false	131	29	81.88%
pred. true	34	110	76.39%
class recall	79.39%	79.14%	

K = 7:

accuracy: 80.88% +/- 8.12% (micro average: 80.92%)

	true false	true true	class precision
pred. false	134	27	83.23%
pred. true	31	112	78.32%
class recall	81.21%	80.58%	

K = 9:

accuracy: 80.27% +/- 7.27% (micro average: 80.26%)

	true false	true true	class precision
pred. false	136	31	81.44%
pred. true	29	108	78.83%
class recall	82.42%	77.70%	

- **DECISION TREE**

The optimal results and optimal parameters are as follows:

```
PerformanceVector [
-----accuracy: 81.19% +/- 8.20% (micro average: 81.25%)
ConfusionMatrix:
True:   false   true
false:  145     37
true:   20      102
]
Decision Tree.maximal_depth     = 50
Decision Tree.minimal_leaf_size = 0
Decision Tree.minimal_size_for_split    = 11
Decision Tree.apply_pruning     = false
Decision Tree.apply_prepruning  = true
```

DISCUSSION OF RESULTS

- **K-NN RESULTS**

The accuracy peaks when the number of k increases from 3 to 7 and increasing it to 9 actually lowers it.

true positive	112
true negative	134
false positive	31
false negative	27
accuracy	80.27%
precision	78.32%
sensitivity	80.58%
specificity	81.21%
F1 measure	79.43%

- **DECISION TREE RESULTS**

Using the Optimize Parameters operator, the best accuracy can be obtained.

true positive	102
true negative	145
false positive	20
false negative	37
accuracy	81.25%
precision	83.61%
sensitivity	73.38%
specificity	87.88%
F1 measure	78.16%

Some rules from the tree:

thal = 3.0

| ca = one

| | cp = 1.0: false {false=2, true=1}

| | cp = 2.0: false {false=4, true=2}

| | cp = 3.0: false {false=10, true=0}

| | cp = 4.0: true {false=1, true=9}

| ca = three: true {false=1, true=5}

| ca = two

| | age = fifties: true {false=0, true=3}

| | age = forties: false {false=2, true=0}

| | age = seventies: false {false=1, true=0}

| | age = sixties: false {false=4, true=4}

The tree considers ca (number of major vessels colored by fluoroscopy) to be the most important parameter.

- **COMBINED RESULTS**

Both results are about the same accuracy in the end which is hovering around 80% accuracy.

CONCLUSION

The accuracy of the models may be further improved in the future if the other dataset from other locations are included.