DATA UNDERSTANDING

- **EXPLORE DATA**

There are 12 attribute and 1309 entries in the initial dataset.

NAME	DATA TYPE (RapidMiner)	DATA QUALITY ISSUES
Age	real	missing values
Sex	binomial	
Passenger Class	polynomial	
Name	polynomial	
No of Siblings or Spouses on Board	integer	
No of Parents or Children on Board	integer	
Ticket Number	polynomial	
Passenger Fare	numeric	missing values
Cabin	polynomial	missing values
Port of Embarkation	polynomial	missing values
Life Boat	polynomial	missing values
Survived	binomial	

From the 4 columns with numerical data, 3 attributes (siblings/spouses, parents/children, fare) are right-skewed and age is normally distributed.

- **VERIFY DATA QUALITY**

There are some missing values in 5 columns. Cabin and Life Boat columns each have more than half of instances missing values. Age has about 20% missing values. Fare and Embarked columns each have less than 3 missing values.

DATA PREPARATION

- **SELECT DATA**

Cabin and Life Boat columns will be removed as they have too much data missing. Name and Ticket Number will be removed as they are irrelevant attributes. No of Siblings or Spouses on Board and No of Parents or Children on Board columns will be combined into Relatives column and then removed.

The Passenger Class, Passenger Fare and Port of Embarkation columns will be shortened to PClass, Fare and Embarked respectively.

- **CLEAN DATA**

The missing values in the Age column will be replaced by mean value. The missing values in the Fare column will be replaced by median value. The missing values in the Embarked column will be replaced by the modal value.

Row No.	average(Age)	median(Passenger F... ↑	mode(Port of Embarkation)
1	29.881	14.454	Southampton

- **CONSTRUCT DATA**

A new instance based on student number 41118874 has been inserted into the data. The new Relatives column is made by combining the No of Siblings or Spouses on Board and No of Parents or Children on Board columns.

- **FORMAT DATA**

For tree-based algorithms, the numeric attributes (Fare, Relatives, Age) will be converted to nominal (binning).

For the distance-based kNN algorithm, the nominal attributes will be converted to binary and numeric attributes will be normalized or standardized.

MODELING & EVALUATION

- **TEST DESIGN**

10-fold cross validation will be used to split the data for both Random Forest and Decision Tree. A percentage split of 75/25 will be used for the k-NN algorithm.

- **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: confidence, minimal leaf size, minimal size for split, apply pruning, and apply pre-pruning.

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

K = 3:

accuracy: 75.54%

	true Yes	true No	class precision
pred. Yes	88	43	67.18%
pred. No	37	159	81.12%
class recall	70.40%	78.71%	

K = 5:

accuracy: 77.68%

	true Yes	true No	class precision
pred. Yes	88	36	70.97%
pred. No	37	166	81.77%
class recall	70.40%	82.18%	

K = 7:

accuracy: 77.37%

	true Yes	true No	class precision
pred. Yes	84	33	71.79%
pred. No	41	169	80.48%
class recall	67.20%	83.66%	

K = 9:

accuracy: 77.98%

	true Yes	true No	class precision
pred. Yes	82	29	73.87%
pred. No	43	173	80.09%
class recall	65.60%	85.64%	

- **DECISION TREE**

The optimal results and optimal parameters are as follows:

PerformanceVector [
-----accuracy: 80.92% +/- 2.86% (micro average: 80.92%)
ConfusionMatrix:
True: Yes No
Yes: 319 68
No: 182 741

Decision Tree.confidence = 0.35000002999999996
Decision Tree.minimal_leaf_size = 51
Decision Tree.minimal_size_for_split = 51
Decision Tree.apply_pruning = true
Decision Tree.apply_prepruning = false

DISCUSSION OF RESULTS

- **K-NN RESULTS**

The accuracy does not increase much when going from 5 to 9. Peak accuracy is at 9.

true positive	82
true negative	173
false positive	29
false negative	43
accuracy	77.98%
precision	73.87%
sensitivity	65.60%
specificity	85.64%
F1 measure	69.49%

- **DECISION TREE RESULTS**

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

true positive	319
true negative	741
false positive	68
false negative	182
accuracy	80.92%
precision	82.43%
sensitivity	63.67%
specificity	91.59%
F1 measure	71.85%

The best 5 rules from the tree:

Sex = Female

| PClass = First: Yes {Yes=139, No=5}

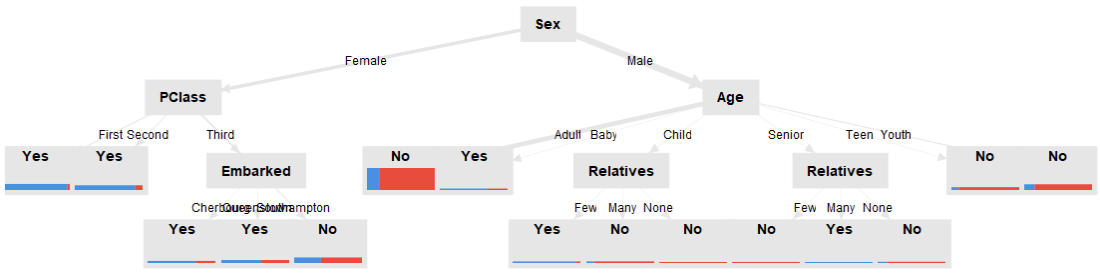
| PClass = Second: Yes {Yes=94, No=12}

| PClass = Third

| | Embarked = Cherbourg: Yes {Yes=22, No=9}

| | Embarked = Queenstown: Yes {Yes=33, No=23}

| | Embarked = Southampton: No {Yes=51, No=78}



The tree considers sex to be the most important parameter.

- **COMBINED RESULTS**

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

DECISION TREE

No of Y = 5
No of sunny / Y = 1
No of overcast / Y = 2
No of rainy / Y = 2
No of hot / Y = 1
No of mild / Y = 1
No of cold / Y = 3
No of high / Y = 2
No of normal / Y = 3
No of false / Y = 4
No of true / Y = 1

No of N = 4
No of sunny / N = 3
No of overcast / N = 0
No of rainy / N = 1
No of hot / N = 2
No of mild / N = 1
No of cold / N = 1
No of high / N = 3
No of normal / N = 1
No of false / N = 2
No of true / N = 2

$P(Yes) = -\frac{5}{9} \times \log_2 \frac{5}{9} = 0.47$
$P(No) = -\frac{4}{9} \times \log_2 \frac{4}{9} = 0.52$
$H(S) = P(yes) + P(no) = 0.99$

$H(sunny) = -\frac{1}{4} \times \log_2 \frac{1}{4} - \frac{3}{4} \times \log_2 \frac{3}{4} = 0.81$
$H(overcast) = -\frac{2}{2} \log_2 \frac{2}{2} - \frac{0}{2} \times \log_2 \frac{0}{2} = 0$
$H(rainy) = -\frac{2}{3} \times \log_2 \frac{2}{3} - \frac{1}{3} \times \log_2 \frac{1}{3} = 0.92$
$M(outlook) = \frac{4}{9} \times 0.81 + \frac{2}{9} \times 0 + \frac{3}{9} \times 0.92 = 0.67$
Gain(outlook) = ~~M outlook~~ $H(S) - M(outlook) = 0.99 - 0.67 = 0.32$

$H(hot) = -\frac{1}{3} \times \log_2 \frac{1}{3} - \frac{2}{3} \times \log_2 \frac{2}{3} = 0.92$
$H(mild) = -\frac{1}{2} \times \log_2 \frac{1}{2} - \frac{1}{2} \times \log_2 \frac{1}{2} = 1$
$H(cool) = -\frac{3}{4} \times \log_2 \frac{3}{4} - \frac{1}{4} \times \log_2 \frac{1}{4} = 0.81$
$M(temperature) = \frac{3}{9} \times 0.92 + \frac{2}{9} \times 1 + \frac{4}{9} \times 0.81 = 0.91$
Gain(temperature) = $H(S) - M(temperature) = 0.99 - 0.89 = 0.10$

$H(high) = -\frac{2}{5} \times \log_2 \frac{2}{5} - \frac{3}{5} \times \log_2 \frac{3}{5} = 0.97$
$H(normal) = -\frac{3}{4} \times \log_2 \frac{3}{4} - \frac{1}{4} \times \log_2 \frac{1}{4} = 0.81$
$M(humidity) = \frac{5}{9} \times 0.97 + \frac{4}{9} \times 0.81 = 0.90$
Gain(humidity) = $H(S) - M(humidity) = 0.99 - 0.90 = 0.09$

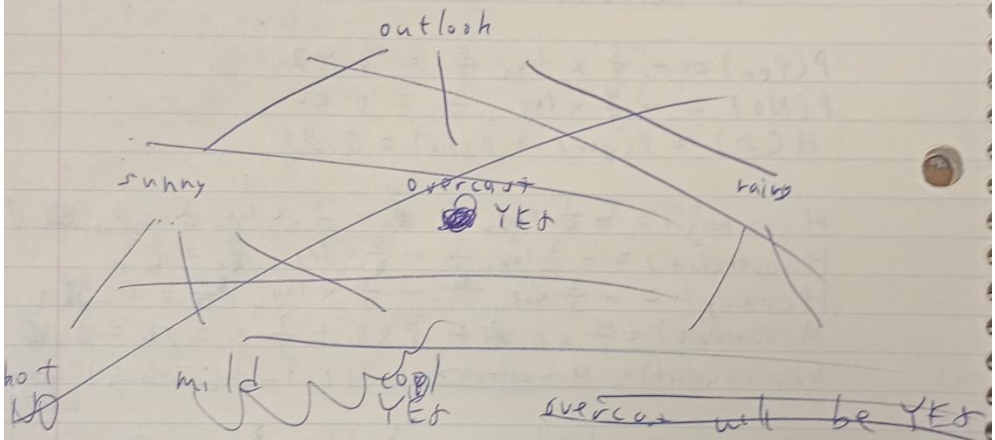
$H(false) = -\frac{4}{6} \times \log_2 \frac{4}{6} - \frac{2}{6} \times \log_2 \frac{2}{6} = 0.$ [illegible]

$H(true) = -\frac{1}{3} \times \log_2 \frac{1}{3} - \frac{2}{3} \times \log_2 \frac{2}{3} = 0.92$

$M(windy) = \frac{6}{9} \times 0.92 + \frac{3}{9} \times 0.92 = 0.$ [illegible]

$Gain(windy) =$ [illegible] $H(S) - M(windy) = 0.93 - 0.$ [illegible] $= 0.07$

	outlook	temperature	humidity	windy
entropy	0.[illegible]67	0.[illegible]89	0.[illegible]90	0.[illegible]92
information gain	0.[illegible]32	0.[illegible]10	0.[illegible]09	0.[illegible]07
ranking	1	2	3	4



overcast will be YES

Outlook = sunny

temp	humid	wind	
hot	high	false	no
hot	high	true	no
mild	high	false	no
cool	normal	false	yes

P(yes) $= -\frac{1}{4} \times \log \frac{1}{4} = 0.5$

P(no) $= -\frac{3}{4} \times \log_2 \frac{3}{4} = 0.31$

H(T) = P(yes) + P(no) = 0.81

H(hot) $= -\frac{2}{2} \times \log_2 \frac{2}{2} - \frac{0}{2} \log_2 \frac{0}{2} = 0$

H(mild) $= -\frac{1}{1} \times \log_2 \frac{1}{1} - \frac{0}{1} \log_2 \frac{0}{1} = 0$

H(cool) $= -\frac{1}{1} \times \log_2 \frac{1}{1} - \frac{0}{1} \log_2 \frac{0}{1} = 0$

M(~~temperature~~ outlook) = 0

G(temperature) = 0.81 − 0 = 0.81

H(high) $= -\frac{3}{3} \times \log_2 \frac{3}{3} - \frac{0}{3} \log_2 \frac{0}{3} = 0$

H(normal) $= -\frac{1}{1} \times \log_2 \frac{1}{1} - \frac{0}{1} \log_2 \frac{0}{1} = 0$

M(humidity) = 0

G(humidity) = 0.81 − 0 = 0.81

H(~~rainy~~ true) $= -\frac{0}{1} \times \log_2 \frac{0}{1} - \frac{1}{1} \log_2 \frac{1}{1} = 0$

H(false) $= -\frac{1}{3} \times \log_2 \frac{1}{3} - \frac{2}{3} \log \frac{2}{3} = 0.92$

M(~~rainy~~ windy) $= \frac{3}{4} \times 0.92 = 0.67$

G(~~humidity~~ windy) = 0.81 − 0.69 = 0.12

Overcast only have YES so it will be 0

Outlook = rainy

	temp	humd	windy	
~~1~~	mild	high	F	Y
	cool	normal	F	Y
	cool	normal	T	N

H(T) $= -\frac{2}{3} \times \log_2 \frac{2}{3} - \frac{1}{3} \times \log_2 \frac{1}{3} = 0.92$

H(mild) $= -\frac{1}{1} \times \log \frac{1}{1} - \frac{0}{1} \log \frac{0}{1} = 0$

H(cool) $= -\frac{1}{2} \times \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} = 1$

M(temp) $= \frac{1}{3} \times 0 + \frac{2}{3} \times 1 = 0.67$

gain(temp) = 0.92 − 0.67 = 0.25

H(high) = $-\frac{1}{1}\log\frac{1}{1} - \frac{0}{1}\log\frac{0}{1} = 0$

H(normal) = $-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$

M(humid) = $\frac{2}{3} \times 1 = 0.67$

Gain(humid) = $0.92 - 0.67 = 0.25$

H(true) = $-\frac{2}{2}\log\frac{2}{2} - \frac{0}{2}\log\frac{0}{2} = 0$

H(false) = $-\frac{1}{1}\log\frac{1}{1} - \frac{0}{1}\log\frac{0}{1} = 0$

M(humid) = 0

Gain(windy) = $0.92 - 0 = 0.92$

outlook=sunny can either split on temperature or humidity as both have equal gain

outlook=rainy can split on windy as it has highest gain

I will choose temperature randomly.

All sunny/hot is no. All sunny/mild is no.

All sunny/cool is yes. All rainy/false is yes. All rainy/true is no.