CS 760: Machine Learning Spring 2024

Homework 4: Decision Trees

AUTHORS: Jed Pulley

DO NOT POLLUTE! AVOID PRINTING, OR PRINT 2-SIDED MULTIPAGE.

Problem 4.1

Passenger Class: First class is 1 and other classes are 0.
Sex: I left sex alone, since it was already binary.
Age: I split age based on the median. If you were above the median age, you get 0, otherwise 1.
Siblings/Spouse: I chose 1 for no siblings/spouses and 0 for any number of them.
Parents/Children: Similarly, I chose 1 for no parents/children and 0 for any
Fare: I split fare based on the median, if you are above, you get 1, otherwise 0.

Problem 4.2

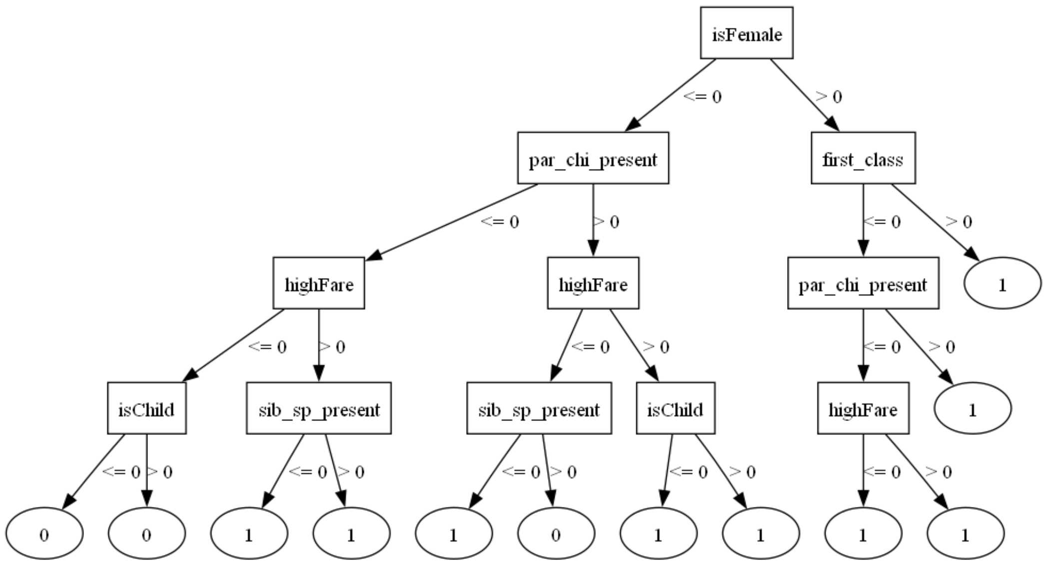
See *decision_trees.ipynb* under section 4.2 for code

Problem 4.3

See *decision_trees.ipynb* under section 4.3 for code

Two variables, *max_depth* and *min_samples_split*, are defined as my stopping conditions. If a node exceeds the max depth, then stop. Otherwise, if a node is about to be split more times than our minimum sample split, then it will stop as well. This prevents the tree from getting too deep or wide.

Problem 4.4



Problem 4.5

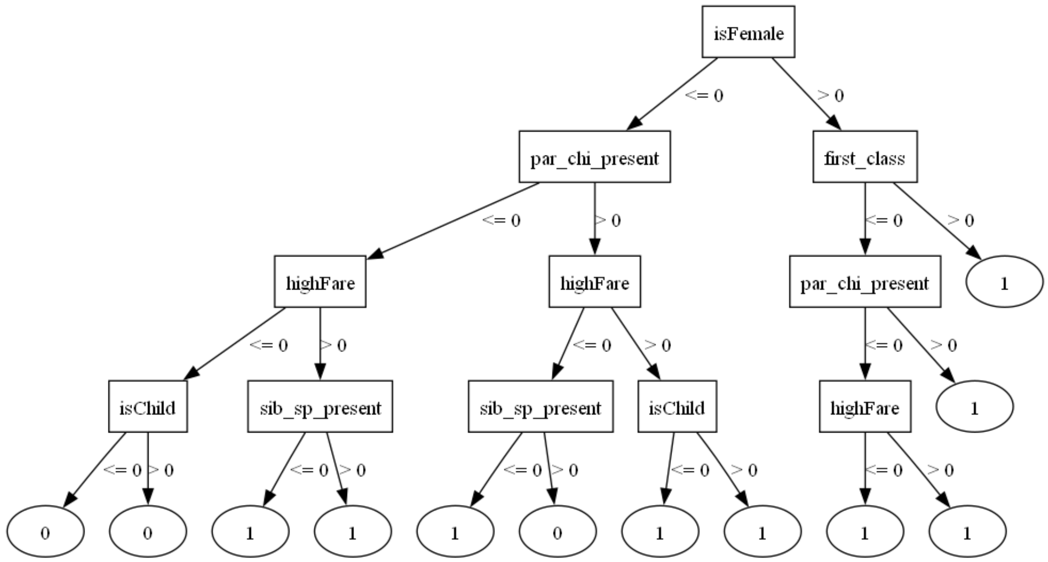
Using *train_test_split* from sklearn, I came up with an accuracy of 0.893

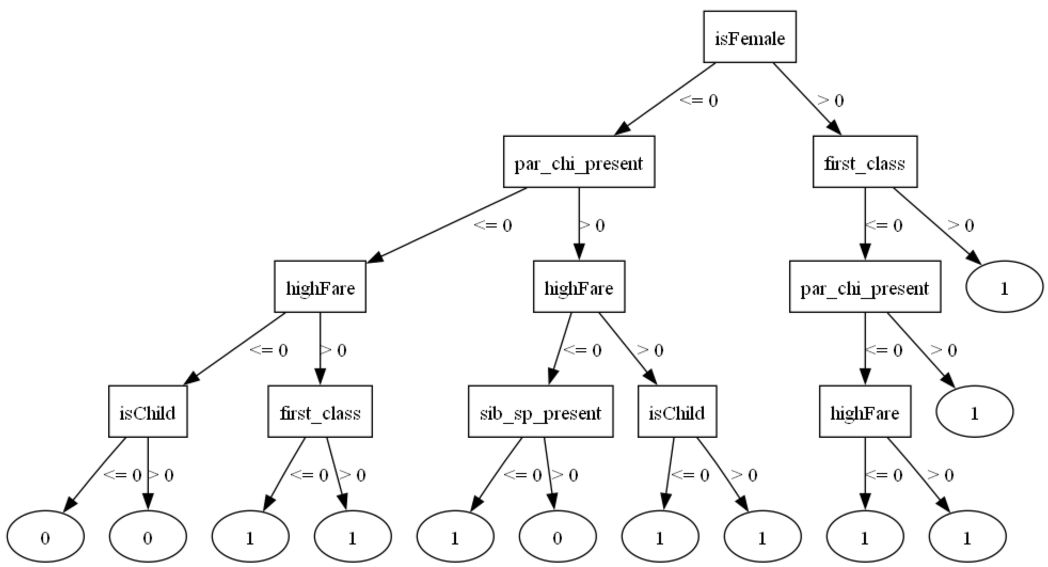
Problem 4.6

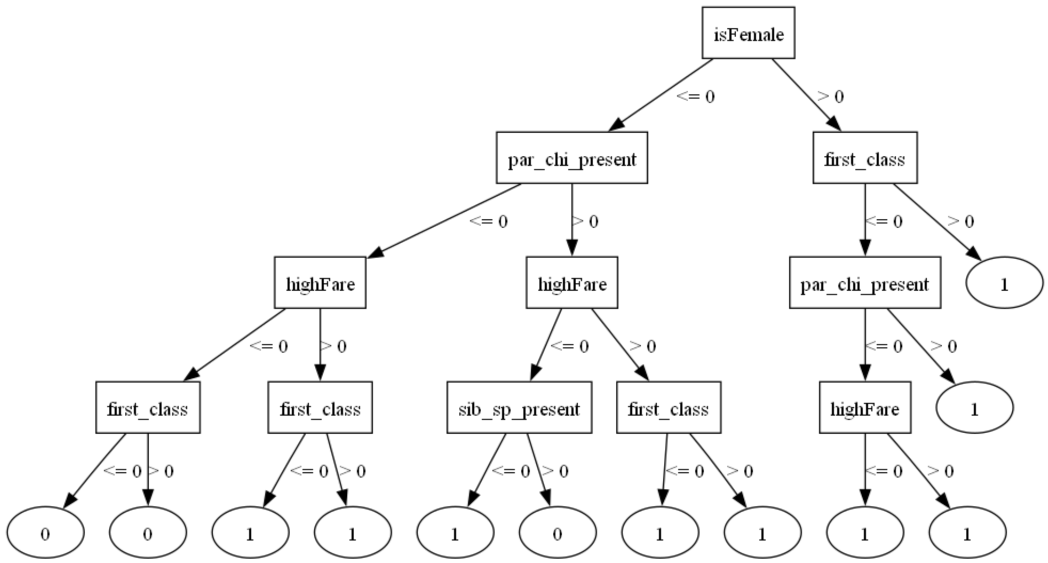
For a man paying a low fare with no children, siblings and parents on board, I would have survived the titanic.

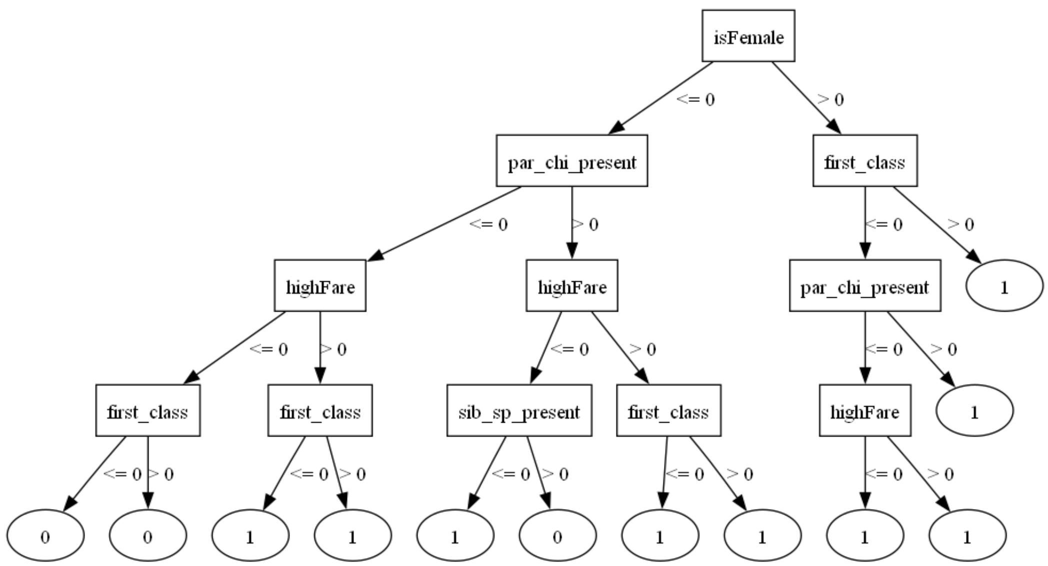
Problem 4.7

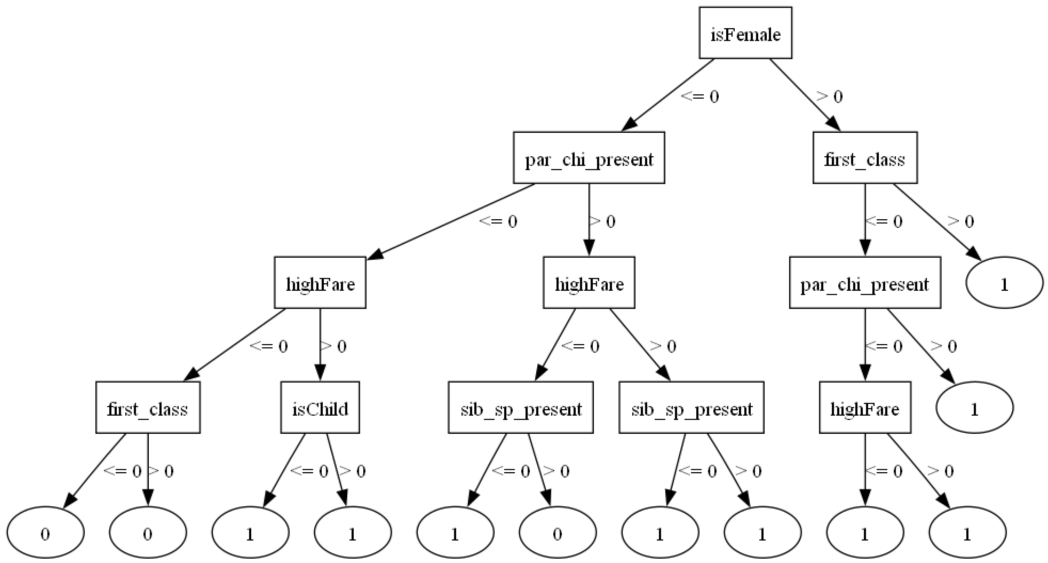
(a)


isFemale
<= 0
> 0
par_chi_present
first_class
highFare
isChild
sib_sp_present
1
0


isFemale
<= 0
> 0
par_chi_present
first_class
highFare
isChild
sib_sp_present
1
0


isFemale
<= 0
> 0
par_chi_present
first_class
highFare
highFare
par_chi_present
1
first_class
first_class
sib_sp_present
first_class
highFare
1
0
0
1
1
1
0
1
1
1
1


isFemale
<= 0
> 0
par_chi_present
first_class
highFare
highFare
par_chi_present
1
first_class
first_class
sib_sp_present
first_class
highFare
1
0
0
1
1
1
0
1
1
1
1

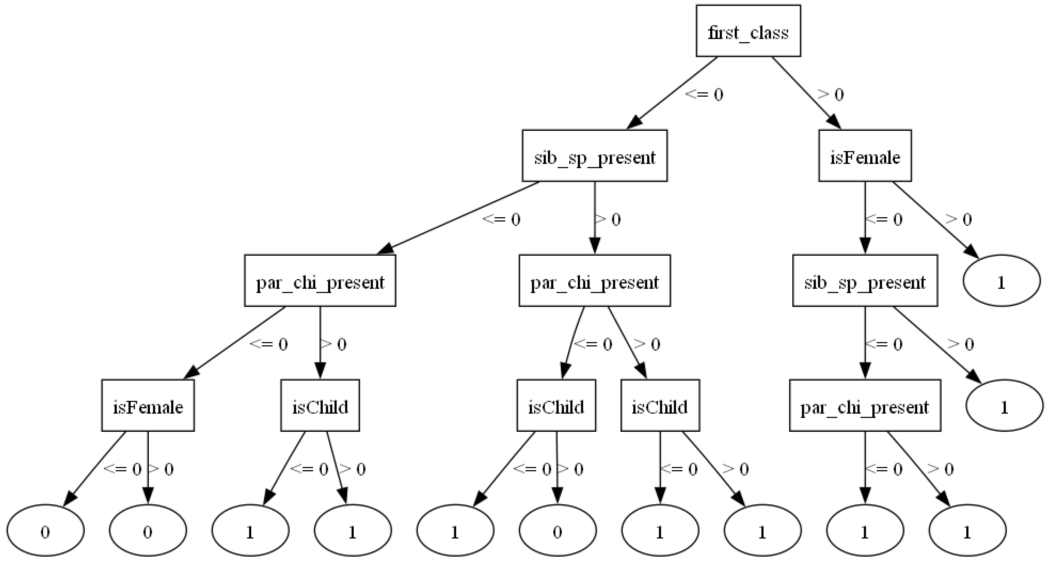


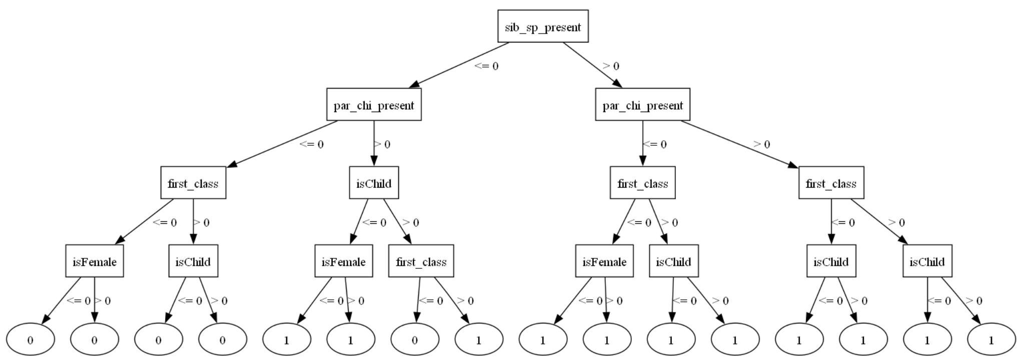
(b) Using 10-fold cross validation, I get an accuracy of 89%.

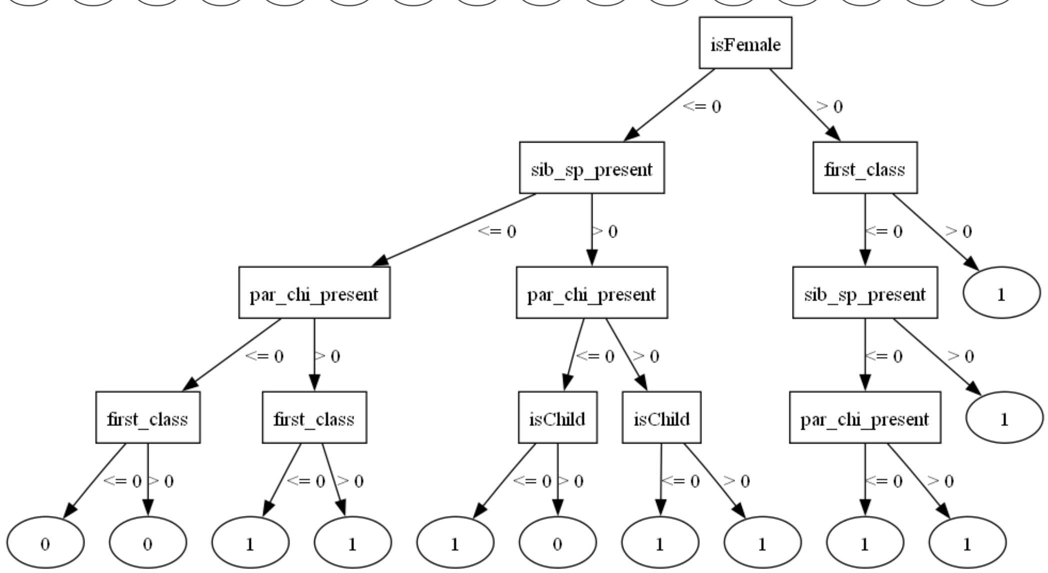
(c) Using the same feature from before, I get the same results.

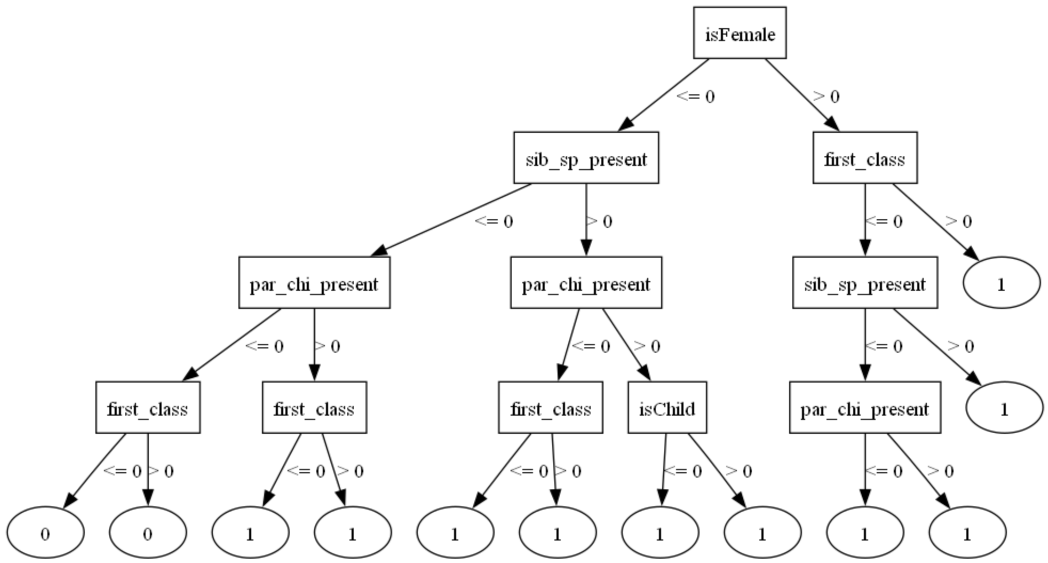
Problem 4.8

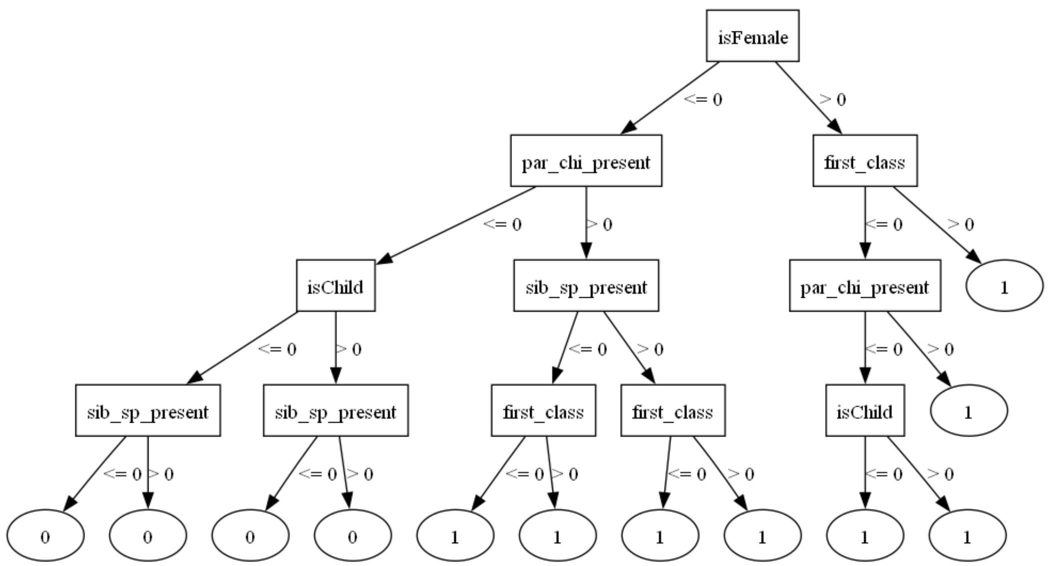
(a)

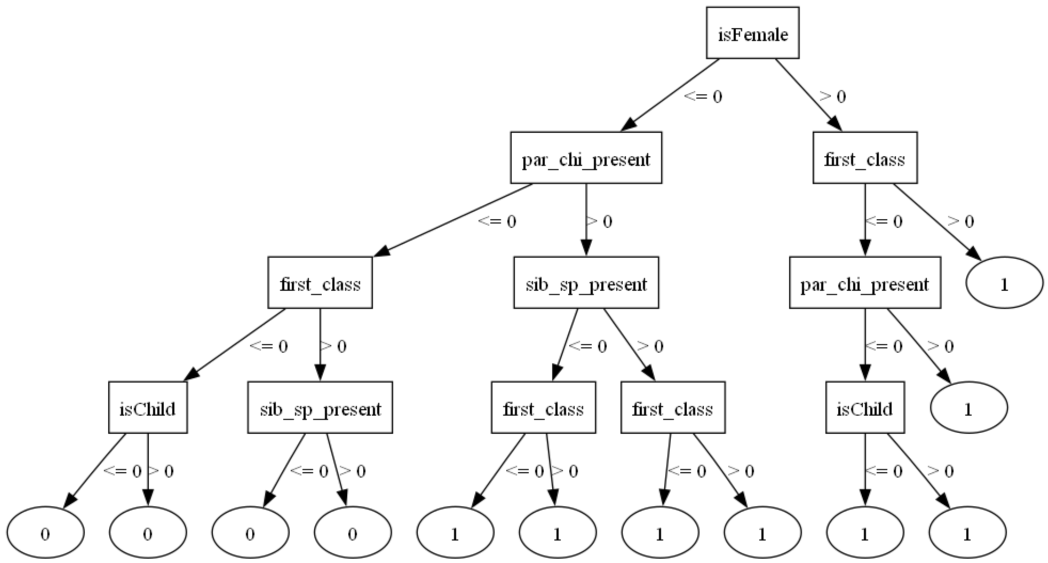

first_class
<= 0
> 0
sib_sp_present
isFemale
par_chi_present
par_chi_present
sib_sp_present
1
isFemale
isChild
isChild
isChild
par_chi_present
1
0
0
1
1
1
0
1
1
1
1


sib_sp_present
<= 0
> 0
par_chi_present
par_chi_present
first_class
isChild
first_class
first_class
isFemale
isChild
isFemale
first_class
isFemale
isChild
isChild
isChild
0
0
0
0
1
1
0
1
1
1
1
1
1
1
1
1


isFemale
<= 0
> 0
sib_sp_present
first_class
par_chi_present
par_chi_present
sib_sp_present
1
first_class
first_class
isChild
isChild
par_chi_present
1
0
0
1
1
1
0
1
1
1
1


isFemale
<= 0
> 0
sib_sp_present
first_class
par_chi_present
par_chi_present
sib_sp_present
1
first_class
first_class
first_class
isChild
par_chi_present
1
0
0
1
1
1
1
1
1
1
1


isFemale
<= 0
> 0
par_chi_present
first_class
isChild
sib_sp_present
par_chi_present
1
sib_sp_present
sib_sp_present
first_class
first_class
isChild
1
0
0
0
0
1
1
1
1
1
1



(b) My accuracy reduced, but not by a lot. It went down to 86%.

(c) Yes, I would have survived. This is using the same feature as the previous two.

Problem 4.9

My decision tree predictions agree, but those disagree with my logistic regression predictions. My assumption as to why is because I'm binarizing my data here and potentially am losing information in the process. I would prefer to use logistic regression, in no small part because it's FAR simpler to implement and understand.

Problem 4.10

From the definition of mutual information, we know that:

$$I(x; y) = H(x) - H(x|y)$$

So, to prove that $I(x; y) = I(y; x)$, we need to show:

$$H(x) - H(x|y) = H(y) - H(y|x)$$

Conditional entropy is defined as:

$$H(x|y) = H(x, y) - H(y) \text{ and, conversely } H(y|x) = H(y, x) - H(x)$$

Substituting these equations into the above, we get:

$$H(x) - (H(x, y) - H(y)) = H(y) - (H(y, x) - H(x))$$

After rearranging, we get:

$$H(x) + H(y) - H(x, y) = H(x) + H(y) - H(y, x)$$

To clean up, we subtract $H(y)$ and $H(x)$ from both sides, multiply by -1, and get:

$$H(x, y) = H(y, x)$$

This still necessitates that we show that joint entropy is symmetric, so we define joint entropy as:

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} \mathbb{P}(x, y) log_2[\mathbb{P}(x, y)]$$

$$H(Y, X) = -\sum_{y \in Y} \sum_{x \in X} \mathbb{P}(y, x) log_2[\mathbb{P}(y, x)]$$

Since the order of summation doesn't affect the result and we know that joint probability is symmetric, we can say prove that $H(x, y) = H(y, x)$