HW 2

- Zaheen E Muktadi Syed
- 5320369
- Date 20 Feb 2022
- CAP 5610

Question 1:

The code is attached in the HW1_task1.ipynb, where I have commented and written step by step process of the data processing I did.

1. The Summary of data processing is that I loaded the data set and investigated the data types and data quality like null values. Then tried to understand what type of data feature can be useful for survival prediction, ie. Name and ticket number are not useful. Basically after loading the data, I started my feature selection.
2. Features Selection:

Variable	Definition	Key	My decision
survival	Survival	0 = No, 1 = Yes	Kept
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd	Kept
sex	Sex	Male 0 female 1	Kept
Age	Age in years		Kept
sibsp	# of siblings / spouses aboard the Titanic		Kept
parch	# of parents / children aboard the Titanic		Kept
ticket	Ticket number		Not Useful, so dropped
fare	Passenger fare		Kept
cabin	Cabin number		Dropped
embarked	Port of Embarkation	C,Q,S=1,2,3	Kept

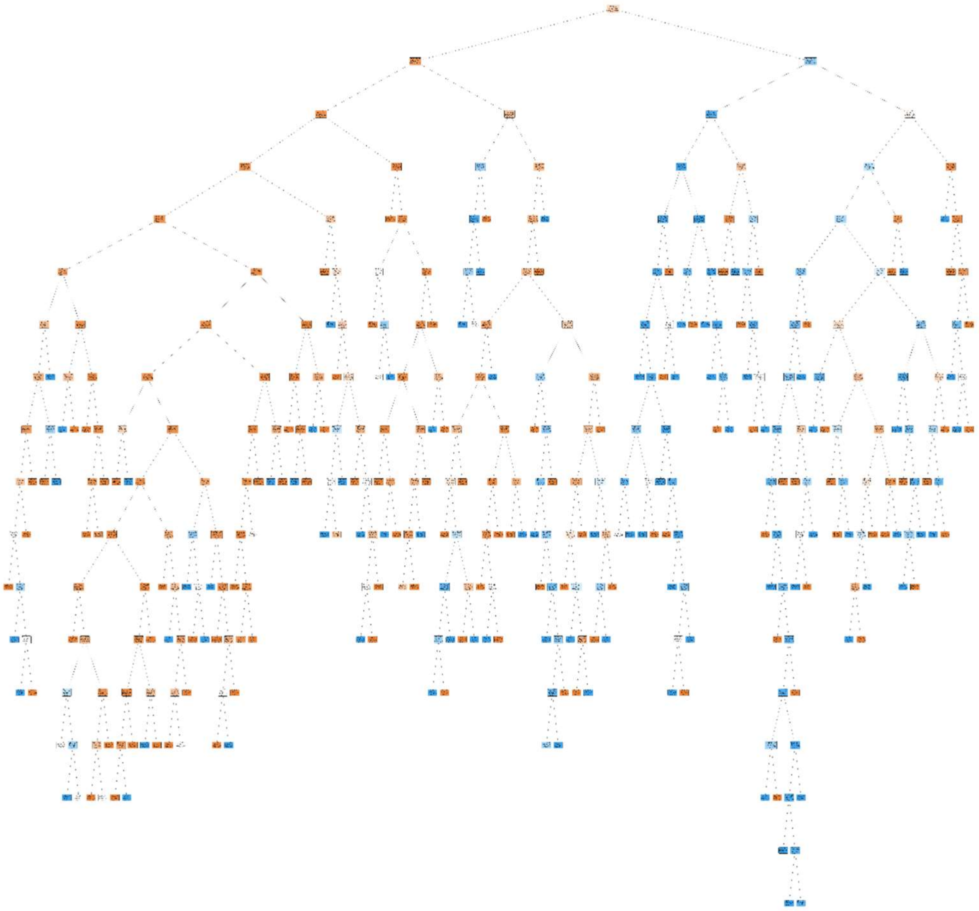
Title	Title of the persion	"Mr": 1, "Miss": 2, "Mrs": 3, "Master": 4, "Rare": 5	

The dropped features were incomplete and simple to identify that they have to useful relation to prediction

However a new feature was used for identification.

3. The decision tree accuracy on training data was 98.43 %

The Decision Tree is plotted in the next page.



4. The average accuracy after 5 fold cross validation with decision tree model is 0.7733036218693113
5. The average accuracy after 5 fold cross validation with random forest model is 0.829408072311845
6. From the accuracy result I see that random forest is the better performer.
7. Conclusion: It is my conclusion that both the models are powerful however sue to the presence of continuous data like fare and age groups I see decision tree performance is weaker. Because removing them enables better performance for decision tree but less performance for RF.

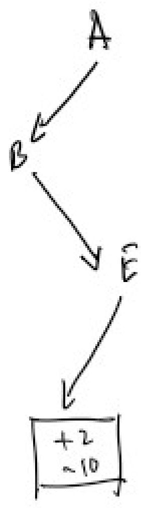
Task 2: Understanding Training Error and Testing

(a) Traing error $= \frac{5+6+2+6+5+5}{100} \times 100$

$= 29\%$

For each node, the majority of the count is considered as the true class and the other as false (errors)

(b) $T = \{ A=0, B=1, C=1, D=1, F=0$



As shown, I have followed the DT chronologically and reach the leaf node wer '–' is majority

So, T will be '–'

Task 3: Understand Splitting Process

①

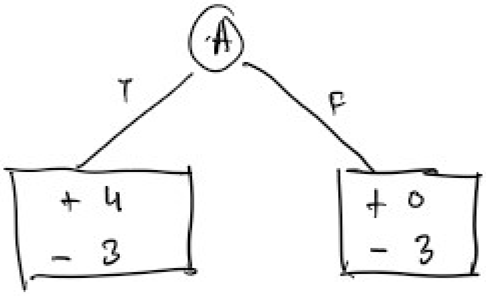
class lebel	
−	6
+	4

① Overall gini before splitting -

$$1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2$$

$= 0.48$

②



$$G_T = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2$$

$$= 0.4898$$

$$G_F = 1 - \left(\frac{3}{3}\right)^2$$

$$= 0$$

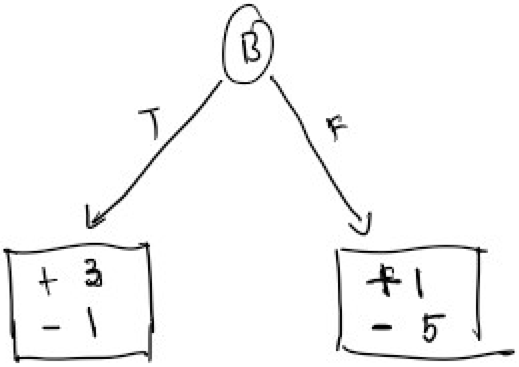
$$w_T = \left(\frac{7}{10}\right) \quad \Big| \quad \left(\frac{3}{10}\right)$$

combined gini $= 0.48 \times \frac{7}{10} + 0 \times \frac{3}{10}$

$= 0.34286$

gain $= 0.48 - 0.34 = 0.14$

③



So,

$G_T = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2$

$= 0.375$

$G_F = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2$

$= 0.278$

combined $= 0.375 \times \left(\frac{4}{10}\right) + 0.278\left(\frac{6}{10}\right).$

$= 0.317.$

gain $= 0.48 - 0.317$

$= 0.163$

Q4. Since the gain after splitting B is higher I will choose B as my attribute.

Task 4: Please answer and explain.

Q1. DT is NOT a linear classifier because it does not split or classify based on the linear combination of the features.

Q2. The DT weakness are :

- Not stable
- Reduced Uncertainty

Because the tree structure varies on the same data set.

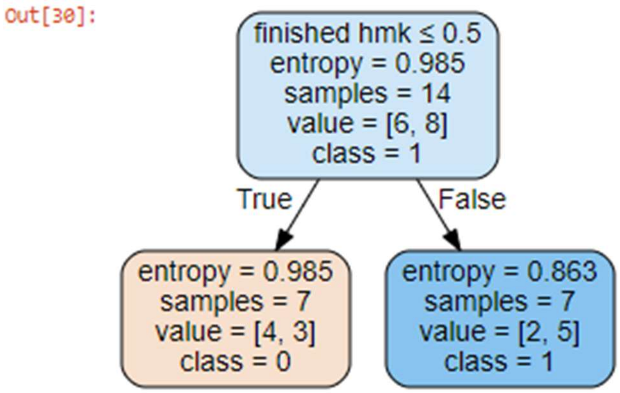
Q3. No Misclassification is not better. It is the least among three. Misclassification have a linear relation with impurity vs probability and less sensitive than gini.

Task 5: Create Decision Trees

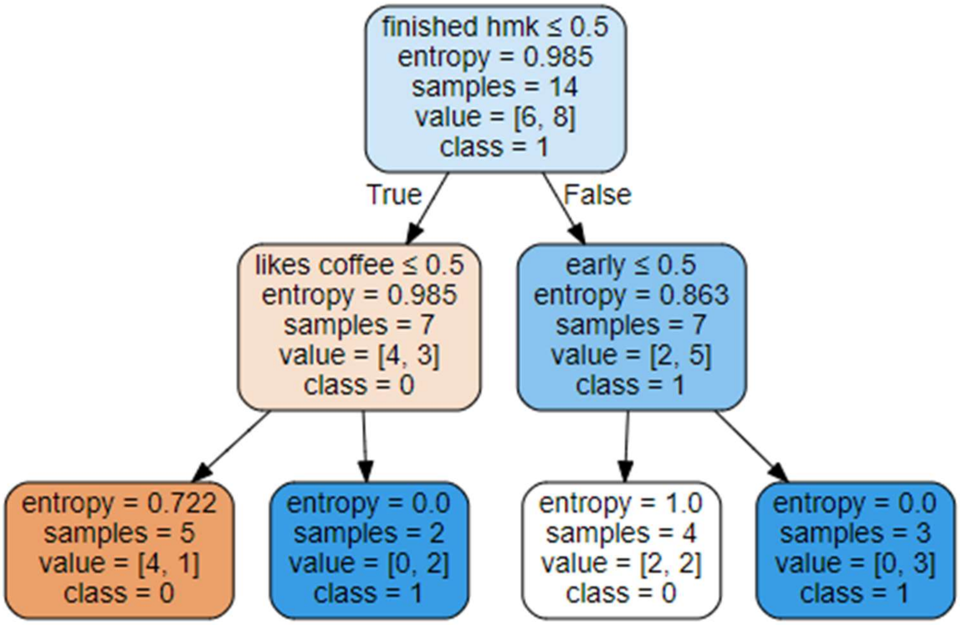
The two decision trees are displayed below.

The Code is uploaded and named “hw2_task5_Visualize Decision Tree.ipynb”. I have manually stored the data as csv file named as data1.csv

Max depth 1, DT 1:



Max depth 2: DT 2



Observation: I would have chosen the second decision tree because the with second depth I see that the with higher we have lower entropy at the lower level meaning total gain is higher. Also, this mean there is a good correlation between the other features.