Name – Sparsh Palkhiwala

HW2 CSE 572 Data Mining

ASU ID – 1228656470

Question 1

Task 1 (20 points) For the Titanic challenge (https://www.kaggle.com/c/titanic), we need to guess whether the individuals from the test dataset had survived or not. Please:

1) Preprocess your Titanic training data;

2) (5 points) Learn and fine-tune a decision tree model with the Titanic training data, plot your decision tree;

3) (5 points) Apply the five-fold cross validation of your fine-tuned decision tree learning model to the Titanic training data to extract average classification accuracy;

4) (5 points) Apply the five-fold cross validation of your fine-tuned random forest learning model to the Titanic training data to extract average classification accuracy;

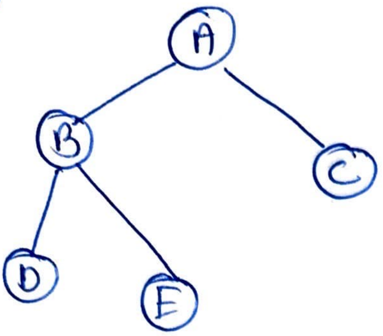
5) (5 points) Which algorithm is better, Decision Tree or Random Forest? What are your observations and conclusions from the algorithm comparison and analysis?

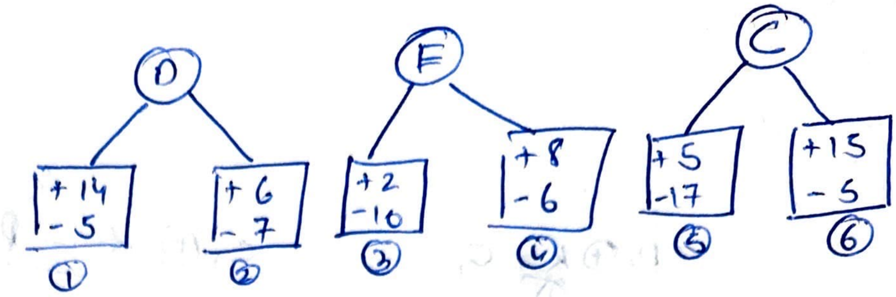
Ans

https://github.com/Sparsh-Palkhiwala/ASU/tree/7165ca8b6a539c9963fe2b91b6cbd5b56bf17808/CSE%20572%20-%20DM/HW_2

⊛ Task - 2

1)





① → Missclassification = 5
Total error = 5

② → Missclassification = 6
Total error = 11

③ → Missclassification = 2
Total error → 13

④ → Miss classification = 6
Total error = 19

⑤ → Miss classification = 5
Total error = 24

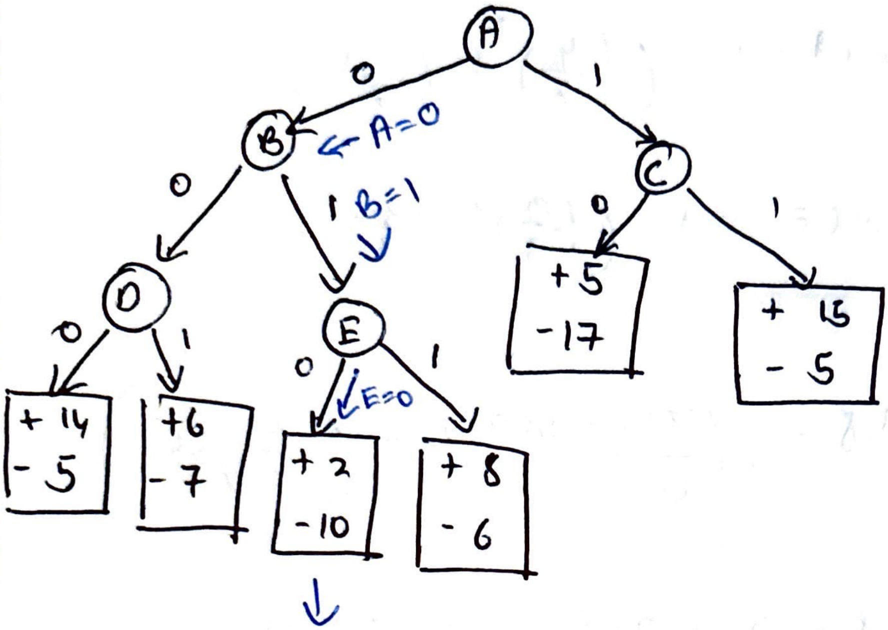
⑥ → Miss classification = 5
Total error = 29

# of missclassification = 29

$$\text{Total error Rate} = \frac{\text{\# of miss classification}}{\text{Total no. of observations}}$$

$$= \frac{29}{100} = 0.29$$

b) For instance   T= {A=0, B=1, C=1, D=1, E=0}



Test T goes to this instance, where it will be classified as '-'.

④ Task-3

1/

Gini Before splitting

'+' Class labels → 4

'-' Class labels → 6

$$\text{Gini} \rightarrow \left(1 - \left(\left(\frac{4}{10}\right)^2 + \left(\frac{6}{10}\right)^2\right)\right.$$

$$= \left(1 - \left(\frac{16}{100} + \frac{36}{100}\right)\right) = \underline{0.48}$$

① For splitting on A :

For A

	+	-
T	4	3
F	0	3

$$\text{Gini (T)} = 1 - \left(\left(\frac{4}{7}\right)^2 + \left(\frac{3}{7}\right)^2\right) = 0.4898$$

$$\text{Gini (F)} = 1 - \left(\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right) = 0$$

Gini Impurity after splitting on A $= 0.48 - \left(\left(\frac{7}{10}\right)(0.4898) - \left(\frac{3}{10}\right)(0)\right)$

Gain $= 0.48 - 0.34286 = 0.13714$

— × — × — × — × — × — × — × — × —

② For splitting on B

For B

	+	-
T	3	1
F	1	5

$$\text{Gini (T)} = 1 - \left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right)$$

$$= 1 - \left(\frac{9}{16} + \frac{1}{16}\right) = 0.375$$

$$\text{Gini (F)} = 1 - \left(\left(\frac{1}{6}\right)^2 + \left(\frac{5}{6}\right)^2\right)$$

$$= 1 - \frac{26}{36} = 0.277$$

Gini Impurity after splitting on B $= 0.48 - \left(\left(\frac{4}{10}\right)(0.375) - \left(\frac{6}{10}\right)(0.277)\right)$

Gain ≈ 0.1633

4) we choose the attribute that gives us more information gain, so splitting over attribute B would be more beneficial for our tree.

Task -4

Q1

Decision trees are non-linear in nature. Unlike linear classifiers that create linear boundaries, decision tree partitions the feature space into an inverse tree like structure. These splits are based on values of indivual features at one node, allowing to find non linear relations in data.

Q2

Missclassification error and Gini both have their own benefits

→ we can use missclassification error to reduce the overall classification errors especially for both balanced classes

→ Gini Index can be used to create balanced trees to handle imbalanced datasets more effeciently.

→ But we prefer Gini due to it's lower sensitivity to noise

④ Task -5

Bagging - we randomly form DT using multiple predictors, it helps reduce variance, thus prevents overfitting. but it leads to

① Lack of interpretability

② Focuses on variance reduction but not on bias.

Random forest - we do sampling on bootstrap we start developing more DTs, using random feature selection.

It addresses bagging weaknesses by

1) Incorporating random feature selection

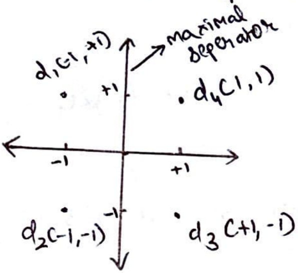
2) We are able to make decorrelated trees through random feature selection.

The difference can improve model interpretability and help reduce variance and bias.

*) Task - 6

Inputs → [-1, -1]
[-1, +1]
[+1, -1]
[+1, +1]

Now we re map the datapoints as : $[\ x_1\ ;\ x_1 x_2]$



$d_1 \rightarrow [-1, +1]$

$d_2 \rightarrow [-1, -1]$

$d_3 \rightarrow [+1, -1]$

$d_4 \rightarrow [+1, +1]$

Now we can have 2 classes → Class (A)(+1) → d_3 and d_4
Class (B) (-1) → d_1 and d_2

Now the seperator would the line in between the points, the midpoint between two of the nearest new generated data points.

The maximal margin seperator is a vertical line at ~~x_2~~ $x_1 = 0$

The margin is perpendicular distance from this line to the nearest data points on either side.
(1, 1) (1, -1)

Task - 7

Equation = $(x_1 - a)^2 + (x_2 - b)^2 - r^2 = 0$

$$(x_1^2 + a^2 - 2x_1 a) + (x_2^{+} + b^2 - 2bx_2) - r^2 = 0$$

$$x_1^2 + a^2 - 2x_1 a + x_2^2 + b^2 - 2bx_2 - r^2 = 0$$

$$x_1^2 + x_2^2 + a^2 + b^2 - 2ax_1 - 2bx_2 - r^2 = 0$$

$$x_1^2 + x_2^2 + a^2 + b^2 - 2ax_1 - 2bx_2 = r^2$$

This equation of circle is in the feature space (x_1, x_2, x_1^2, x_2^2)

→ Every circle area is linear seperable in this feature space

$$x_1^2 + x_2^2 - 2ax_1 - 2bx_2 + (a^2 + b^2 + r^2) = 0$$

weights → $(2a, 2b, 1, 1)$ and intercept → $(a^2 + b^2 + r^2)$

circular equation in this feature space is linear seperable

⊛ Task - 8

$c(x_1-a)^2 + d(x_2-b)^2 -1=0$

$K(u,v) = (1 + u.v)^2$ in the feature space

$(1, x_1, x_2, x_1^2, x_2^2, x_1x_2)$

we expand equation

$$c(x_1^2+a^2-2ax_1) + d(x_2^2+b^2-2bx_2) -1=0$$

$$cx_1^2 + dx_2^2 - 2acx_1 - 2bdx_2 + (a^2c + b^2d - 1) = 0$$

weights → $(2ac, 2bd, a^2c, b^2d) 0)$

intercept → $a^2+b^2-r^2$

The elliptical boundary looks linearly seperable in this feature space