Assignment 2

Section - A

a) In a random forest, as the diversity increases, the correlation decreases and vice versa. This is the tradeoff between diversity and correlation in a random forest. If the correlation is too high and the diversity is too low, the variance will be high, thus causing overfitting. Similarly, if the trees are completely uncorrelated or they have very low correlation, the diversity will be high, however it will not be possible to form aggregates easily as all decision trees will act independently and hence we will not get a reliable output. That is why, It is important for the trees to be correlated to some extent in order to both minimize the variance and also reach a consensus among trees, to form reliable aggregate output, thus having both generalization and reliability.

b) The curse of dimensionality can become an issue in Naive Bayes when we have data with a very large number of features and correspondingly not enough data points. This will result in overfitting, as the model is too complex and is not generalizable with the amount of data available.

    To solve this problem, we can use dimensionality reduction. Techniques like Principal component analysis can be used to reduce the dimensionality of the data and choose those features that are most strongly correlated with the output and thus have the most effect on it.

c) If the Naive bayes classifier encounters a sample which contains attribute values that is not present in the training dataset, it can lead to the zero probability problem. In other words, the classifier will assign it 0 probability of occurring, as it has not encountered this value earlier while training.

    To resolve this problem, we can use Laplace smoothing. In laplace smoothing, using a parameter p (not equal to 0), we assign a non zero probability to all data points.

    For example: let's say we have a feature X which predicts a target variable Y, X consists of values {0, 1, 2}, Y consists of values {0, 1}. We don't have any value such that, {Y = 0 | X = 0}. When we get such a sample without laplace smoothing, P(Y = 1 | X = 0} = 0.

    However, if we use laplace smoothing with parameter k, then P(Y = 1 | X = 0) = (Count(Y = 1, X = 0) + k) / Count(Y = 1) + k * n, where n = number of classes. Thus, in this way we have resolved the zero probability problem.

d) Yes, splitting a decision tree node using Information gain is biased if some attributes have more cardinality than others.This can lead to overfitting as an attribute with high cardinality will split the data further using several branches, and this will result in lack of generalizability.

    To resolve this issue, we also need to take cardinality into account when splitting a node. One such criterion for that is gain ratio. Gain ratio mitigates the

effect of cardinality by also taking into consideration the information gained from a split along with its cardinality, thus giving a more balanced result.

For example, let's consider a case where we have two features, F1 with values {0, 1, 2} and F2 with values {3, 4, 5}, and an output variable O with values {0, 1}. In this case, the formulae for Information Gain and Gain Ratio can be expressed as:

Gain(F1) = Entropy(O) - ∑v∈F1 P(F1 = v) · Entropy(O|F1 = v) GainRatio(F1) = Gain(F1) / IntrinsicInformation(F1)

In these formulas, 'Entropy(O)' represents the entropy of the output variable 'O,' and 'P(F1 = v)' represents the probability that the attribute 'F1' takes on the value 'v.' The Gain Ratio effectively takes both the information content of 'F1' and its cardinality into account, ensuring a more balanced attribute selection process.

2)

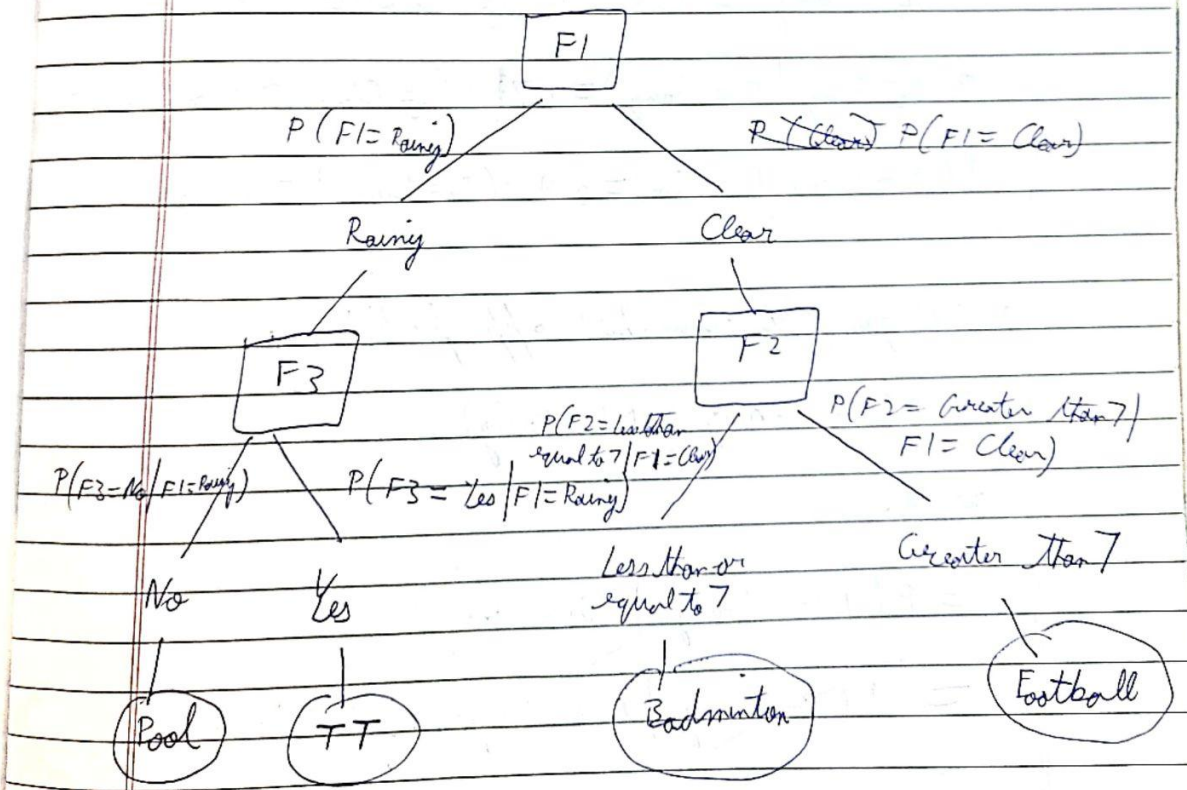
A2) (a) Initially, there are 3 features for classification

F1 ⇒ Weather — Rainy / Clear

F2 ⇒ Outdoor activity friends — Greater than 7 / Less than or equal to 7

F3 ⇒ Can borrow racket? — Yes / No

Based on these 3 features and the possibilities mentioned in the problem, here is the decision tree.



Outcome = { Pool, TT, Badminton, Football }

$P(\text{Pool}) = P(F2 = \text{No} \mid F1 = \text{Rainy}) \cdot P(F1 = \text{Rainy})$

$P(TT) = P(F2 = \text{Yes} \mid F1 = \text{Rainy}) \cdot P(F1 = \text{Rainy})$

$P(\text{Badminton}) = P(F3 = \text{less than or equal to } 7 \mid F1 = \text{Clear}) \cdot P(F1 = \text{Clear})$

$P(\text{Football}) = P(F3 = \text{Greater than } 7 \mid F1 = \text{Clear}) \cdot P(F1 = \text{Clear})$

(b) let B_t be the opp's prediction,

$P(B_t = \text{Rainy}) = 0.3, \quad P(B_t = \text{Clear}) = 0.7$

$P(B_t = \text{Rainy} \mid F1 = \text{Rainy}) = 0.8, \quad P(B_t = \text{Clear} \mid F1 = \text{Rainy}) = 0.2$

$P(B_t = \text{Clear} \mid F1 = \text{Clear}) = 0.9, \quad P(B_t = \text{Rainy} \mid F1 = \text{Clear}) = 0.1$

~~$P(\text{It's going to rain but opp predicts rainy})$~~

~~$= P(B_t = \text{Rainy} \mid F1 = \text{Rainy})$~~

~~$= 0$~~

$= P(F1 = \text{Rainy} \mid B_t = \text{Rainy})$

$= P(B_t = \text{Rainy} \mid F1 = \text{Rainy}) \cdot \frac{P(F1 = \text{Rainy})}{P(B_t = \text{Rainy})}$

$= 0.8 \times \frac{P(F1 = \text{Rainy})}{0.3}$

~~$P(F1 = \text{Rainy}) =$~~

P (It is going to rain but o/p products rainy)

$= P(Pr = \text{Rainy} \mid F1 = \text{Rainy})$

$= P(F1 = \text{Rainy} \mid Pr = \text{Rainy})$

$= P(Pr = \text{Rainy} \mid F1 = \text{Rainy}) \cdot \frac{P(F1 = \text{Rainy})}{P(Pr = \text{Rainy})}$

$= \boxed{0.8 \times \frac{P(F1 = \text{Rainy})}{0.3}}$

Assumption: If $P(F1 = \text{Rainy}) = P(Pr = \text{Rainy}) = 0.3$,

then

$P(F1 = \text{rainy} \mid Pr = \text{rainy}) = 0.3$

(Q2). ~~P(Rainy) =~~

$P(Br = Rainy) = P(Rainy) \cdot P(Br = Rainy \mid Fl = Rainy)$

$+ P(Fl = Clear) \cdot P(Br = Rainy \mid Fl = Clear)$

$0.3 = P(Fl = Rainy) \cdot (0.8) + (1 - P(Fl = Rainy)) \cdot$

$P(Fl = Rainy) = \frac{0.2}{0.7} = \frac{2}{7}$

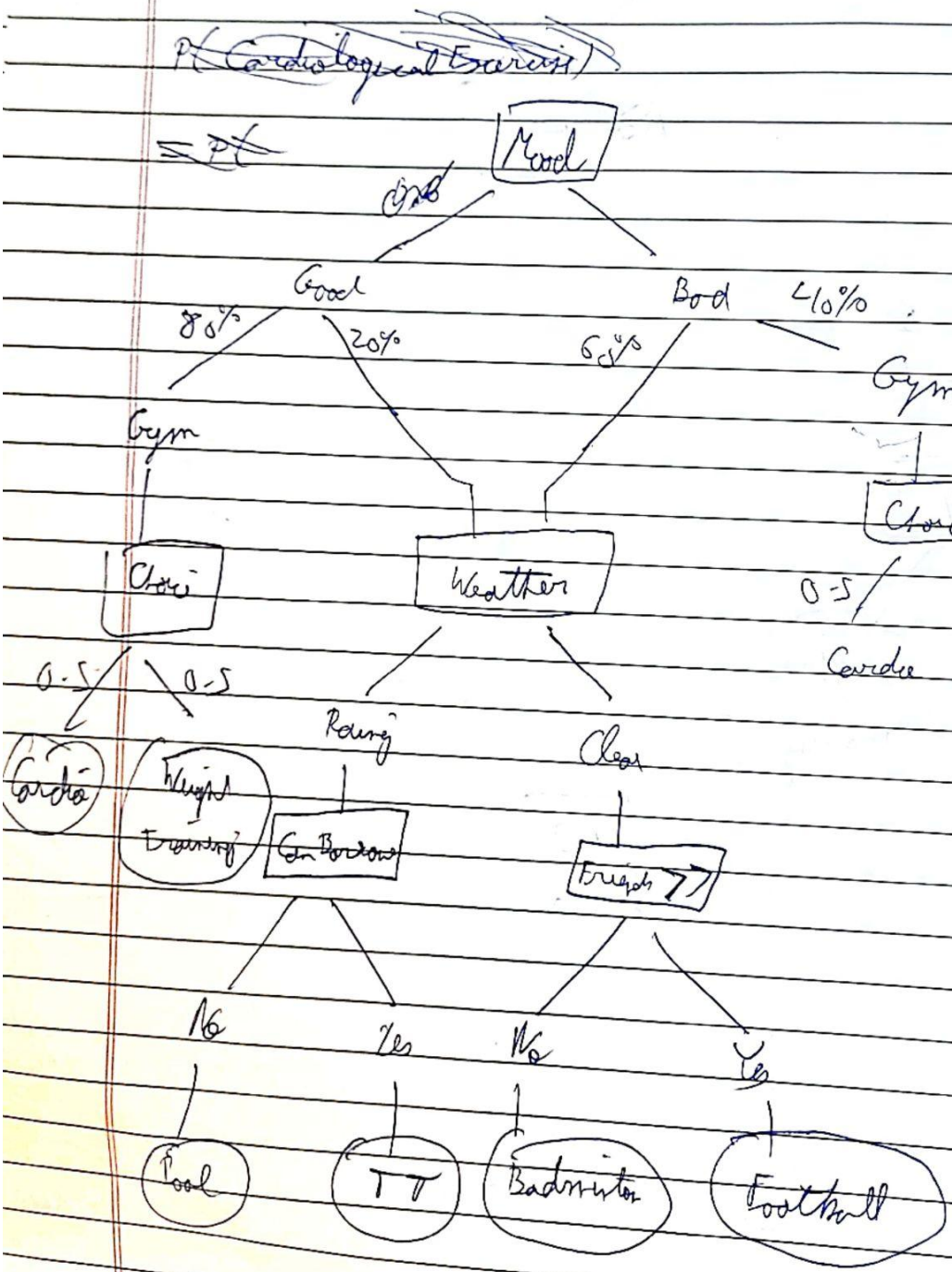
$P(Fl = Rainy \mid Br = Rainy)$

$= P(Br = Rainy \mid Fl = Rainy) \cdot \frac{P(Fl = Rainy)}{P(Br = Rainy)}$

$= 0.8 \times \frac{2}{7} \times \frac{1}{0.3}$

$= \frac{16}{21} = 0.762$

Outcome = { Cardiological Exercise, Weight Training, Pool, TT, Badminton, Football }



P (Cardiological Exercise)

= P(Cardiological Exercise | Gym) . P(Gym | Mood = Good).
P (Mood = Good) + P(Cardiological Exercise | Gym)
. P (Gym | Mood = Bad). P(Mood = Bad)

~~= (0.5 × 0.8 × 0.4)~~

P (Weight Training)

= P(Weight Training | Gym) . P(Gym | Good Mood).
P(Good Mood)

+ P(Weight Training | Gym) . P(Gym | Bad Mood).
P (Bad Mood)

~~P(Pool) = P(Can Borrow~~

P(Pool) = P(Cannot Borrow | Rainy) . P(Rainy | Good Mood) . P(Good Mood)

P(TT) = P(Can Borrow | Rainy) . (P(Rainy | Good Mood) . P(Good Mood)

$P(Tol) = P(\text{Cannot Borrow} \mid \text{Rainy}) \cdot P(\text{Rainy} \mid \text{Good Mood}) \cdot P(\text{Good Mood})$

$+ P(\text{Cannot Borrow} \mid \text{Rainy}) \cdot P(\text{Rainy} \mid \text{Bad Mood}) \cdot P(\text{Bad Mood})$

~~$P(\text{Badminton}) = P($~~

$P(TT) = P(\text{Can Borrow} \mid \text{Rainy}) \cdot P(\text{Rainy} \mid \text{Good Mood}) \cdot P(\text{Good Mood})$

$+ P(\text{Can Borrow} \mid \text{Rainy}) \cdot P(\text{Rainy} \mid \text{Bad Mood}) \cdot P(\text{Bad Mood})$

$P(\text{Badminton}) = P(\text{Friends} \leq 7 \mid \text{Clear}) \cdot P(\text{Clear} \mid \text{Good Mood}) \cdot P(\text{Good Mood})$

$+ P(\text{Friends} \leq 7 \mid \text{Clear}) \cdot P(\text{Clear} \mid \text{Bad Mood}) \cdot P(\text{Bad Mood})$

$P(\text{Football}) = P(\text{Friends} > 7 \mid \text{Clear}) \cdot P(\text{Clear} \mid \text{Good Mood}) \cdot P(\text{Good Mood})$

$+ P(\text{Friends} > 7 \mid \text{Clear}) \cdot P(\text{Clear} \mid \text{Bad Mood}) \cdot P(\text{Bad Mood})$

$P(\text{Good Mood}) = 0.6, \quad P(\text{Bad Mood}) = 0.4$

$P(F=7 \mid \text{Good Mood}) = 0.7, \quad P(F=7 \mid \text{Bad Mood}) = 0.45$

$$P(\text{Good Mood} \mid F=7) = \frac{P(F=7 \mid \text{Good Mood}) P(\text{Good Mood})}{P(F=7)}$$

$$= \frac{0.7 \times 0.6}{P(F=7)}$$

$$= \frac{0.42}{P(F=7)}$$

$$P(\text{Bad Mood} \mid F=7) = \frac{P(F=7 \mid \text{Bad Mood}) P(\text{Bad Mood})}{P(F=7)}$$

$$= \frac{0.45 \times 0.4}{P(F=7)}$$

$$= \frac{0.18}{P(F=7)}$$

$$P(\text{Good Mood} \mid F=7) + P(\text{Bad Mood} \mid F=7) = 1$$

$$\frac{0.42}{P(F=7)} + \frac{0.18}{P(F=7)} = 1$$

$$P(\text{Good Mood} \mid F=7) = \frac{0.42}{0.6} = 0.7$$

$$P(\text{Bad Mood} \mid F=7) = \frac{0.18}{0.6} = 0.3$$

$$P(\text{Gym}) \propto P(\text{Gym} \mid F=7) = P(\text{Good Mood} \mid F=7) \cdot P(\text{Gym} \mid \text{Good Mood}, F=7) + P(\text{Bad Mood} \mid F=7) \cdot P(\text{Gym} \mid \text{Bad Mood}, F=7)$$

$$= 0.7 \times 0.8 + 0.3 \times 0.4$$

$$= 0.68$$

$$P(\text{Not Gym} \mid F=7) = P(\text{Good Mood} \mid F=7) \cdot P(\text{Not Gym} \mid \text{Good Mood}, F=7) + P(\text{Bad Mood} \mid F=7) \cdot P(\text{Not Gym} \mid \text{Bad Mood}, F=7)$$

$$= 0.7 \times 0.2 + 0.3 \times 0.6$$

$$= 0.32$$

Therefore gym is the most likely option as $P(\text{Gym} \mid F=7) > P(\text{Not Gym} \mid F=7)$

Also, as Weight Training and cardiological sources are equally probable, they are both equally likely outcomes

Also, $P(\text{Cardio}) = 0.68 \times \frac{1}{2} = 0.34$

$P(\text{Weight Training}) = 0.68 \times \frac{1}{2} = 0.34$

$P(\text{Tired} \mid$

Section - C

1) In EDA, we found that, Features ca and thal have missing values. We used the median value to replace the missing values.

2) The class imbalance is very low as the percentage difference between classes 0 and 1 is only 16.5%. Therefore, we can carry on without resampling.

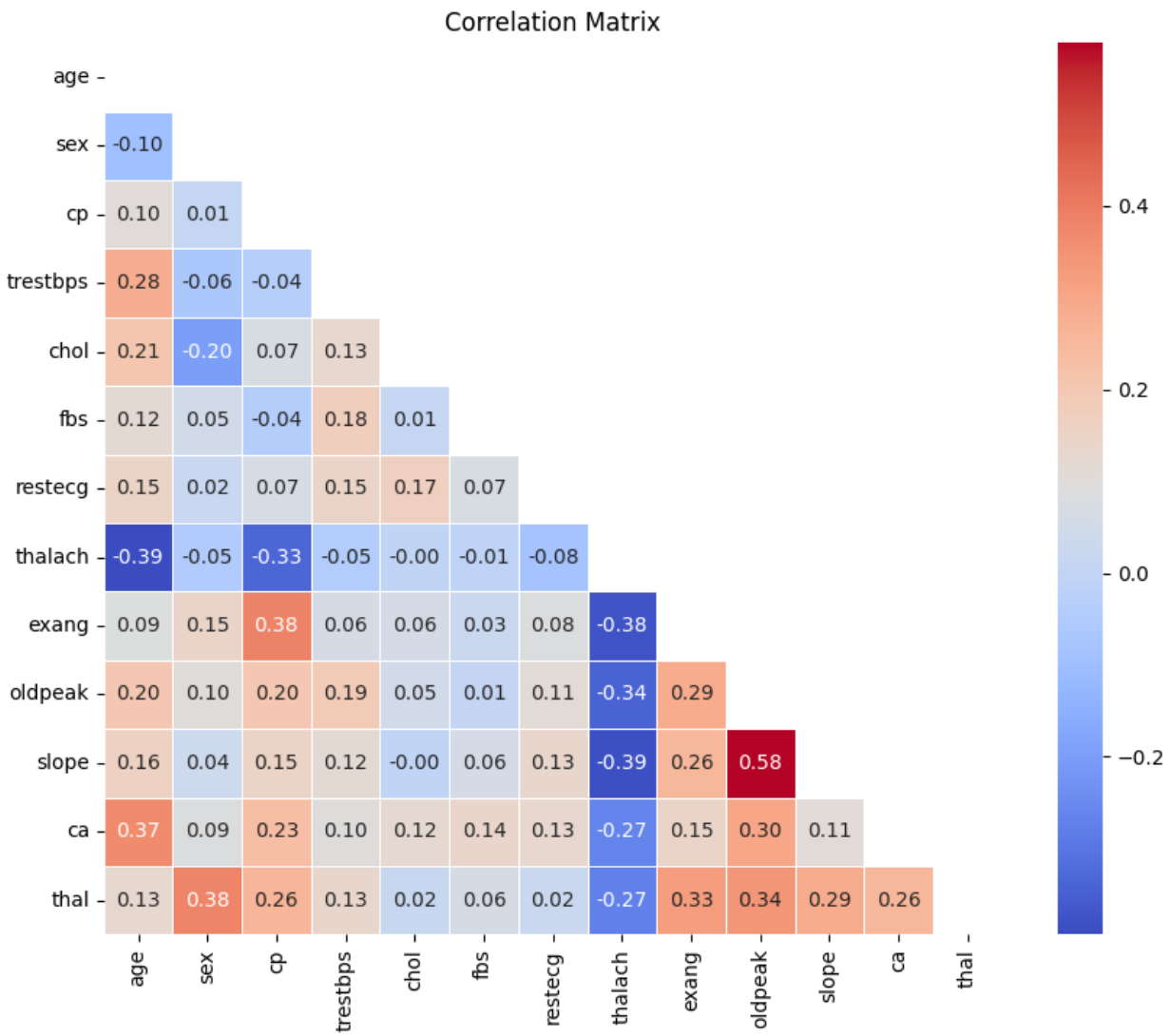
3) The data seems correlated. The feature thalach seems most strongly correlated with all the other features.

4) During Training and Testing, the following were our conclusions Entropy (77%) gives a higher accuracy than Gini(72%) , therefore moving forward we will be using it as our criterion.

5) For decision tree, after hypertuning we got accuracy of 80%.

6) Random Forrest, gave an accuracy of upwards of 82%, slightly higher than decision tree.


Class Distribution of the Target Variable
Frequency
0
20
40
60
80
100
120
140
160
0
1
Class Type



Final Classification Report on test data:

	precision	recall	f1-score	support
0	0.78	0.88	0.82	32
1	0.84	0.72	0.78	29
accuracy			0.80	61
macro avg	0.81	0.80	0.80	61
weighted avg	0.81	0.80	0.80	61