Ans 1) a) Correlation means similarity in the structure and prediction of the trees, in forest, which implies if trees are correlated, they have similar predictions whereas Diversity means differences among the trees, if trees are diverse, they have different prediction.
More correlation implies less diversity and more diversity implies less correlation. Increase in one leads to ~~of~~ decrease of other and hence are inversely related.
More correlation leads to overfitting of data as prediction ~~shifts~~ have better influence of coorelated trees. If diversity is large, the ensemble process suffers from underfitting as it is not able to learn data patterns. Therefore, it is important for the trees to be correlated upto a certain extent while maintaining diversity

b) Curse of Dimensionality is the situation in which data has too many features. More features leads to higher risk of overfitting. ~~Larger~~
Also, more feature leads to larger amount of data need to generalize accurately ~~increasing~~ increases exponentially.
The "Curse of dimensionality" can become an issue in Naive bayes if amount of data is limited and data points are sparse. Since, naive bayes algorithm calculates probability for all possible combinations of features, the computational complexity can become expensive & hence suffers from "curse of dimensiolity"

The following strategies can be employed to mitigate this problem in practice :-

1) Selecting most relevant features.

2) Reducing dimensionality while retaining information with techniques such as t-SNE

3) In case of continuous data, we can change numerical range into categories and hence reducing number of unique value of each feature

4) Using technique such as K-fold cross validation, bagging, boostin can help in counter problem by fixing problem of large data requirement

~~Ans 8)~~ c) When a naive bayes classifier encounters a value attribute not present in the training ~~dataset~~ as It may face challenge as it assumes conditional independence. The absence may lead to probability estimate of zero for that particular value and can lead to affect on classifier results.

For example, we are provided dataset of diseases. In training, naive model doesn't ever encountered virus_type. If prediction is to be made based on virus_type ~~for new~~ (a new attribute for classifier), the model may assign a probability of zero.

Approaches to mitigate problem:

1) Using different variant of Naive Bayes, which treat unseen value similary such as Gaussian N.B.

2) Grouping unseen data into a specific category

3) Assigning non zero probability to unseen data.

to mitigate

rmation with

change

reducing

validation,

r problem.

equirement

a value

as

onditional

probability

alue and

ts.

diseases.

ncountered

ade based

ttribute for

probability

yes,

such as

category

data.

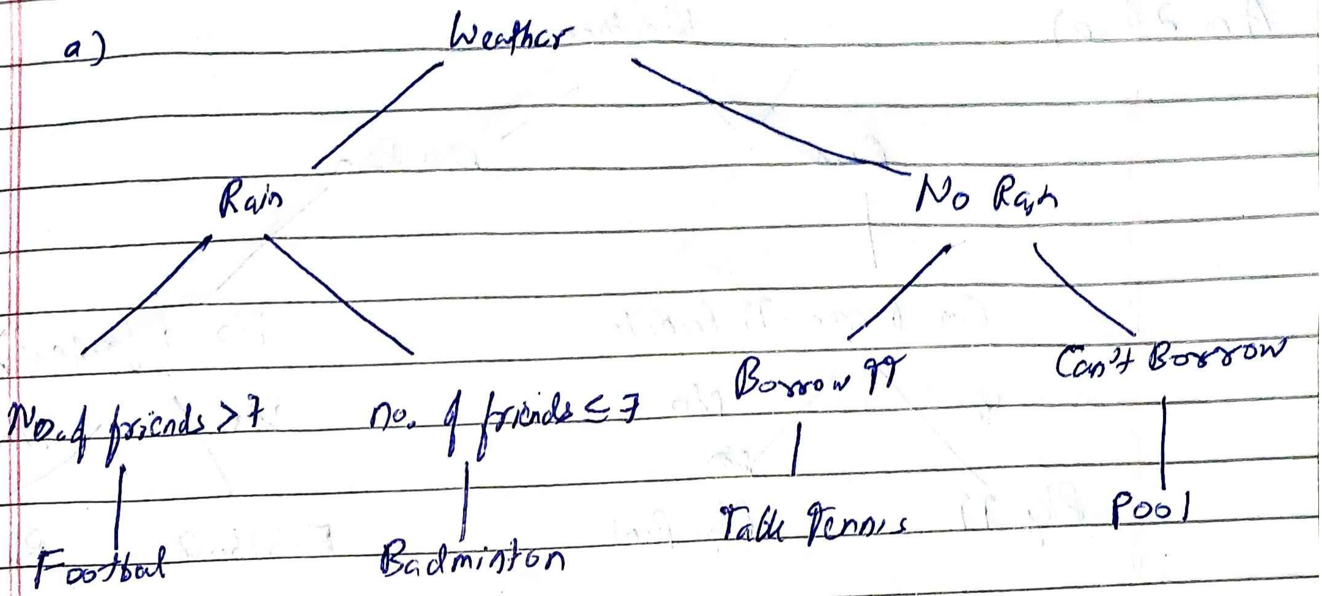
→ If missing values ~~are~~ have minimal impact on classification task and classification decision doesn't rely on that particular feature, the impact of missing value will be not Significant

d) Yes, decision tree node split using information gain can be biased towards attributes with more cardinality as information gain measures the reduction in entropy, and attributes with more cardinality have more information splits and hence higher information gain

To avoide the bias, we can use criteries as Gain ration, gini index.

Example: Suppose, we want to build a decision tree to predict whether a patient is sick or healthy based on their age, gender and symptoms. If we use Information gain and symptoms has highest cardinality, it is more likely to select symptoms as split attribute but the symptom attribute may not be the most informative attribute for predicting whether a patient is sick or healthy.

Ans2) a)



All possible outcomes with respective conditions are :

1) Play football : No Rain & No. of Friends > 7

P ~~(Play Football)) = P(No Rain) × P(Number of Friends >~~

P(Play Football) = P(No Rain) × P(Number of Friends > 7 | No Rain)

2) Play Badminton : No Rain & No. of friends ≤ 7

P(Play Badminton) = P(No Rain) × P(No. of Friends ≤ 7 | No Rain)

3) Play T.T. : Rain & Borrow TT

P(Play T.T) = P(~~No~~ Rain) × P(Borrow TT | Rain)

4) Play Pool : Rain & Can't Borrow TT

P(Play Pool) = P(Rain) × P(Can't Borrow TT | Rain)

b) ~~P('Rainy') = 0.3~~

~~P('Clear') = 0.7~~

~~P('Accurate'|'Rainy') = 0.8~~

~~P('Accurate'|'Clear')~~

Ans2) b) P(App Pred Rainy | Rainy) = 0.8

P(App Pred Clear | Clear) = 0.9

P(App Pred Rainy | Clear) = 1 − P(App Pred Clear | Clear)

= 1 − 0.9

= 0.1

P(Rainy) = 0.3

P(Clear) = 0.7

By Bayes Theorem,

$$P(\text{Rainy} \mid \text{App pred Rainy}) = \frac{P(\text{Rainy}) \times P(\text{App pred Rainy} \mid \text{Rainy})}{\text{Total Probability})}$$

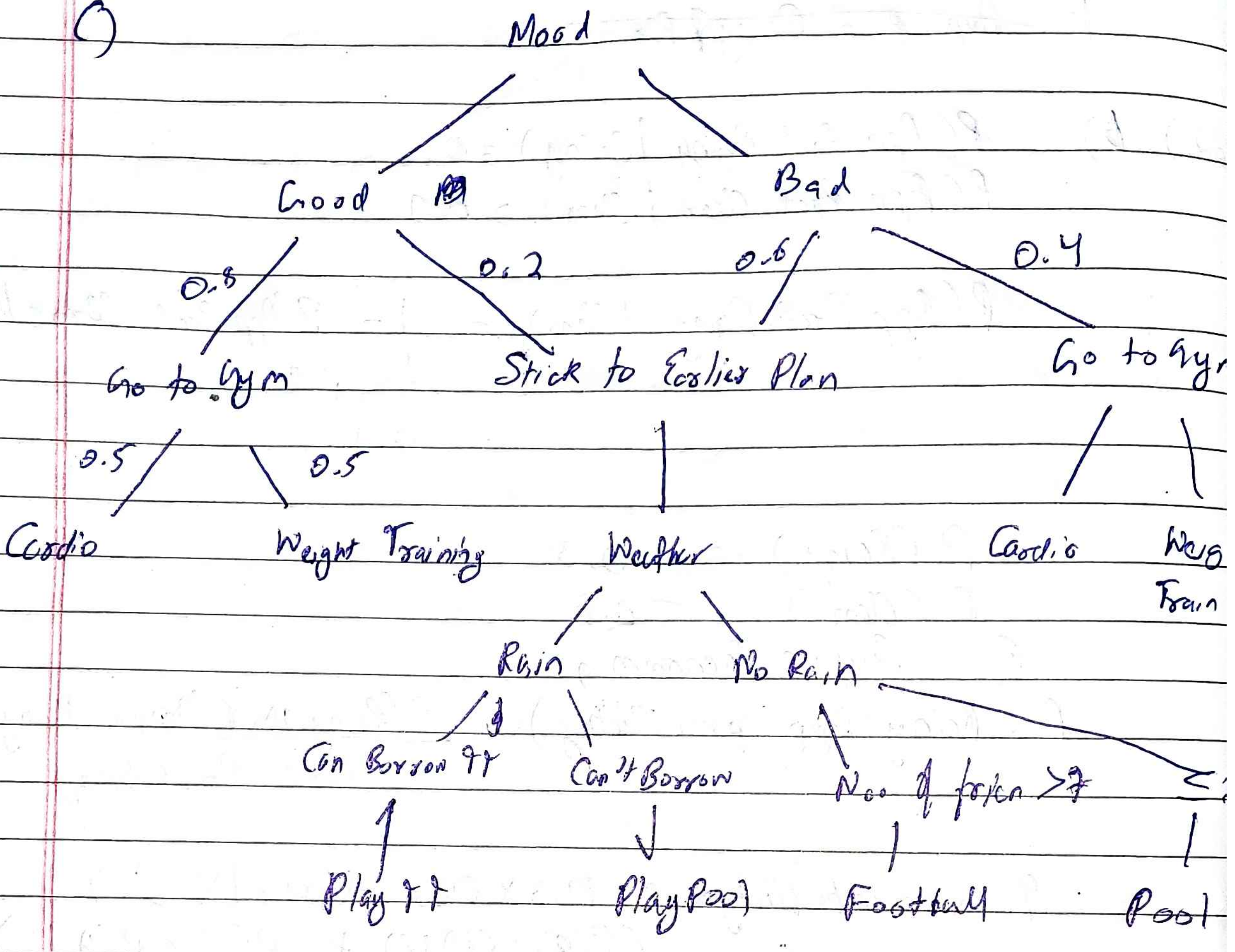
Total probability = 0.3 × 0.8 + 0.1 × 0.7

(P(R) × P(A|R) + P(C) × P(A|C)

= 0.31

$$\therefore P(\text{Rainy} \mid \text{App Pred Rainy} = \frac{0.3 \times 0.8}{0.31} = \frac{0.24}{0.31} \approx 0.77$$

c)



All possible Outcomes: are :

Cardio Exercise, Weight Training, Play Football, Play Badminton, Play T.T. , Play Pool

~~1) P(Cardio Exercise) = P(Good Mood) × P(Go to Gym | Good Mood) +~~

Let Good Mood = GM , Bad Mood = BM

Go to Gym = GTM , No Rain = NR

Weight training = WT Rain = R

Play Football = PF Num of friend > 7 = NY

Play Badminton = PB " ≤ 7 = NN

Play T.T = PT

Play Pool = PP

Cardion exercise = CE

Stick to Earlier Plan = S

Conditional Probabilities :

$P(CE) = P(GM) \times P(GTM \mid GM) \times P(CE \mid GTM) +$
$P(BM) \times P(GTM \mid BM) \times P(CE \mid GTM)$

$P(WT) = P(GM) \times P(GTM \mid GM) \times P(WT \mid GTM) +$
$P(BM) \times P(GTM \mid BM) \times P(WT \mid GTM)$

~~P(PF) = P(GM) × P(S | GM) × Pf~~

All probabilies from 1 will be multiplied ~~not~~

with $\left(P(GM) \times P(S \mid GM) + P(BM) \times P(S \mid BM)\right)$

d) $P(GTM \mid GM) = 0.8$
$P(\text{No } GTM \mid GM) = 0.2$
$P(GTM \mid BM) = 0.4$
$P(\text{No } GTM \mid BM) = 0.6$
$P(CE \mid GTM) = 0.5$
$P(CE \mid GTM) = 0.5$

let $P(GM) = p$ and $P(BM) = 1-p$

$P(GTM) = P(GTM \mid GM) \times P(GM) + P(GTM \mid BM) \times P(BM)$
$= 0.8p + 0.4(1-p)$

Similarly,
$P(\text{No } GTM) = 0.2p + 0.6(1-p)$

$P(WT) = 0.4p + 0.2(1-p)$

$P(CE) = 0.4p + 0.2(1-p)$

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

$\Rightarrow P(GM) = 0.7$

$P(BM \mid F=7) = \frac{0.45 \times 0.4}{0.6} = 0.3$

$P(\text{No } GM) = 0.14 + 0.18$
$= 0.32$

$P(NT) = 0.28 + 0.06$
$= 0.34$

$P(CE) = P(CE \mid Gym) \times P(Gym)$
$= 0.28 + 0.06$
$= 0.34$

Most Likely Outcome: Go To Gym with equal probability of weight training and cardio since go to gym> stick to earlier plan(no gym)