SEM-VII-ML-MSE-1 QP-2024-2025

Vidyalankar Institute of Technology
Semester 7 – CMPN - Mid Semester Assessment – 1

Date: 05/08/2024	Machine Learning	30 Marks /1 hour

1	**Solve any two (5 marks each)**		**CO**
	A	Justify why Linear Regression is not used for performing classification.	CO1
	B	Elaborate on: (i) Karl Pearsons Coefficient of Co relation. (ii) R square method.	CO2
	C	Discuss the importance of the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) in evaluating binary classifiers. How do these metrics help in comparing models?	CO2
2	**Solve any two (5 marks each)**		
	A	Outline the steps involved in developing a Machine Learning application for predicting housing prices.	CO1
	B	Justify the significance of the F1 score in imbalanced classification problems. How does it address the limitations of using precision or recall alone? Provide a scenario where the F1 score might still be misleading.	CO2
	C	Justify the bias-variance trade-off in the context of model complexity. How does this trade-off influence the choice of model in a real-world ML application? Provide examples to support your explanation.	CO2
3	**Solve anyone (10 marks each)**		
	A	For the dataset given below, construct a decision tree using Gini Index, and determine which attribute is a root attribute.	CO2

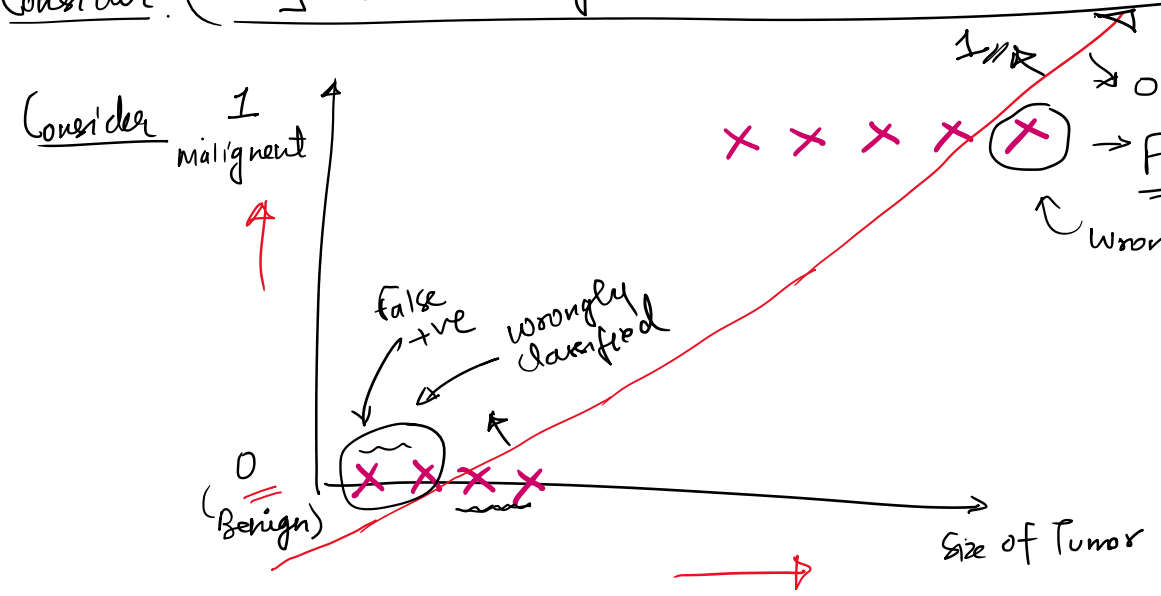
Sr. No.	Weather	Parent	Money	Decision
1	Sunny	Yes	Rich	Cinema
2	Sunny	No	Rich	Tennis
3	Windy	Yes	Rich	Cinema
4	Rainy	Yes	Poor	Cinema
5	Rainy	No	Rich	Stay In
6	Rainy	Yes	Poor	Cinema
7	Windy	No	Poor	Cinema
8	Windy	No	Rich	Shopping
9	Windy	Yes	Rich	Cinema
10	Sunny	No	Rich	Tennis

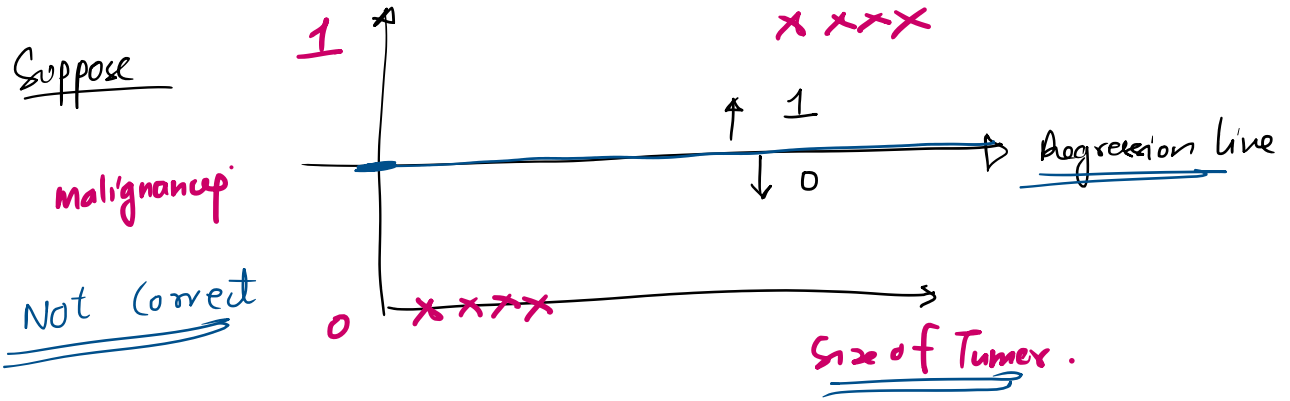
	B	Justify the significance of Cross Validation. List the type and discuss k-fold cross validation in detail? CO3

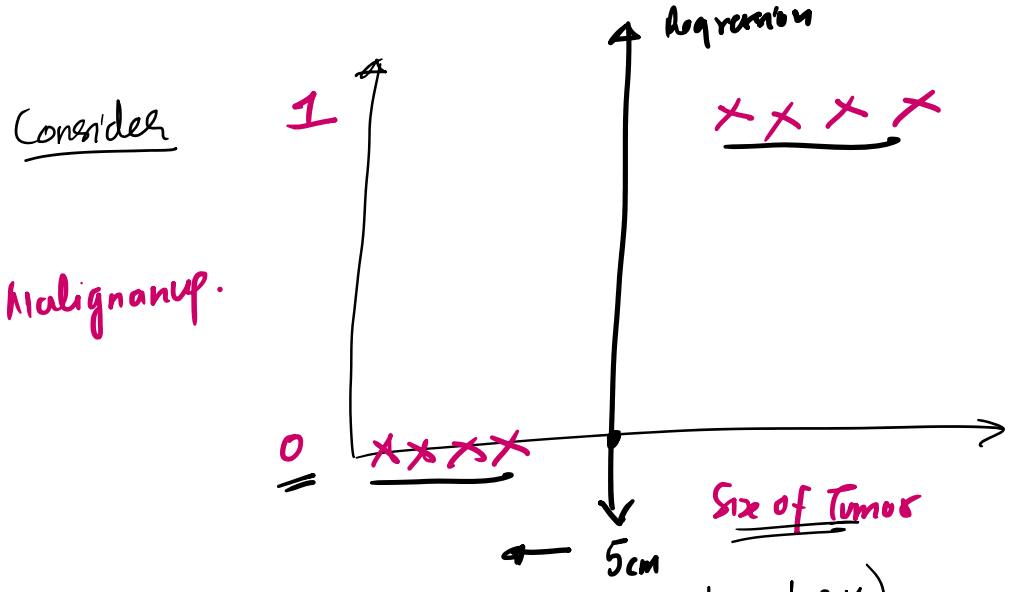
CO1	Gain knowledge about basic concepts of Machine Learning and understand the difference between supervised and unsupervised techniques
CO2	To select, apply and evaluate an appropriate machine learning model for the given
CO3	Ability to understand regression techniques.

Q1. A Justify why Linear Progression is not used for performing classification

Consider (Why linear Regression is not used for Classification)







Here we can observe for above Regression line

If tumor size $\leq$ 5cm $\rightarrow$ Yes (1) Malignant

If tumor size ≤ 5cm $\longrightarrow$ Yes (1) Malignant

tumor size > 5m $\longrightarrow$ No (0) Benign

We can say let 'P' denote Probability that $y = 1$ when $X = x$.

$$y = P_r(y = 1 \mid X = x) = [\beta_0 + \beta_1 x]$$

In linear Reg

P = probability lies bet^n 0 to 1

But linear function are unbounded.

and Expected o/p here is 0 or 1

So we cannot use Regression to build Classifier

∴ linear Regression is not suitable for Classification.

Q1. B

Elaborate on:
(i) Karl Pearsons Coefficient of Co relation.
(ii) R square method.

* To Evaluate Performance of M.L models
* To find how good the model fits on given data

(1) Karl Pearsons Coefficient of Correlation (γ)

→ To calculate Relationship bet two variable

$$\gamma = \frac{N\Sigma xy - \Sigma x \Sigma y}{\left(N*\Sigma x^2 - (\Sigma x)^2\right)\left(N*\Sigma y^2 - (\Sigma y)^2\right)}$$

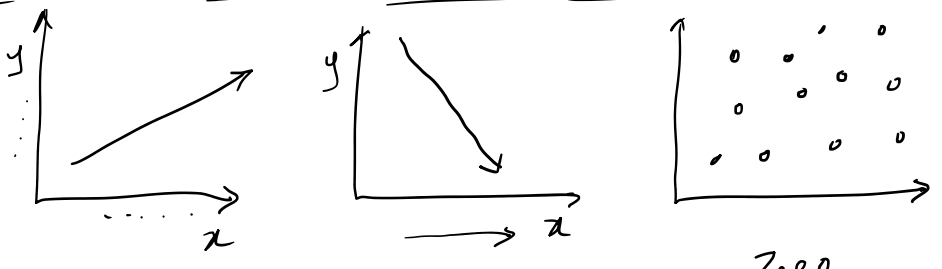
γ = quantifies the strength of relationship bet^n two variables.

The value of γ be between $+1$ and -1

If $\gamma = 1$ ⇒ Total +ve correlation ⇒ If $x\uparrow$ then $y\uparrow$

If $\gamma = -1$ ⇒ Total -ve correlation ⇒ If $x\downarrow$ then $y\uparrow$

> It gives strength (degree) & direction of correlation. or $x\uparrow$ then $y\downarrow$

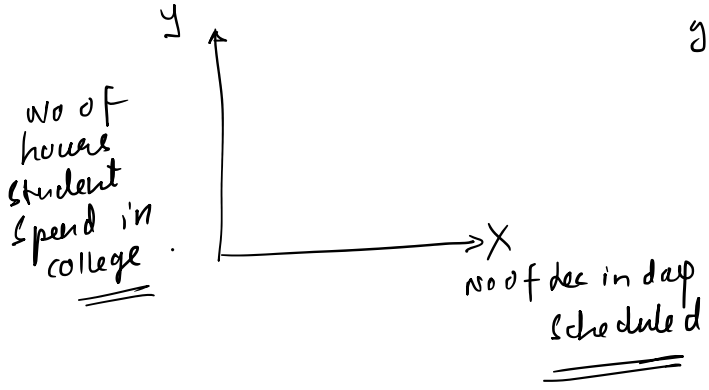


+ve Correlation — −ve Correlation — Zero Correlation.

② R^2 method (R Square)

↳ It gives information about good of fit feature of the model.

↳



If $r^2 = 0.85$

Variation in no of hours that students spend in college is 85% dependent on no of lec scheduled.

↳ Indicate percentage of variance in dependent and Independent variable pair

> value varies from 0 to 1

If $R^2 = 1$ => no diff bet^n actual & predicted value.

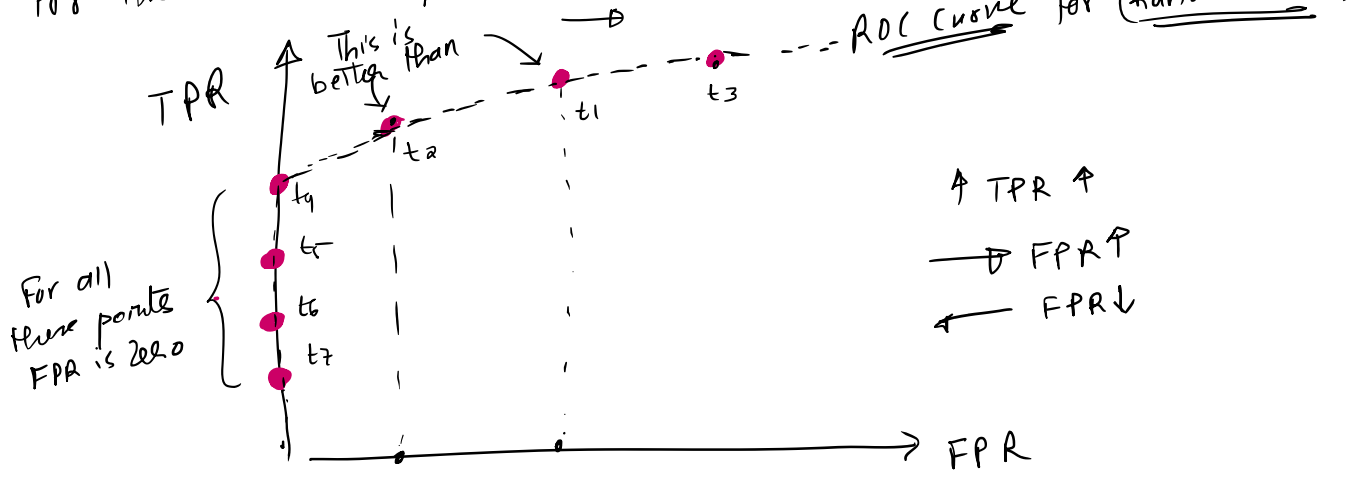
$r^2 = 0$ means the model does not learn any relationship bet^n variables.

Q1.C

Discuss the importance of the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) in evaluating binary classifiers. How do these metrics help in comparing models?

ROC (Reciever Operator characteristic)

* Consider for logistic Regression, where we identify a threshold point and prepare confusion matrix and calculate Sensitivity & Specificity.
* If threshold changes then the confusion matrix and accordingly the Sensitivity and Specifity changes.
  we can have many such thresholds bet^n $0 \to 1$
* we want to analyze the performance at diff threshold. and want to identify the best of it.
* For this we will plot TPR and FPR for diff threshold.

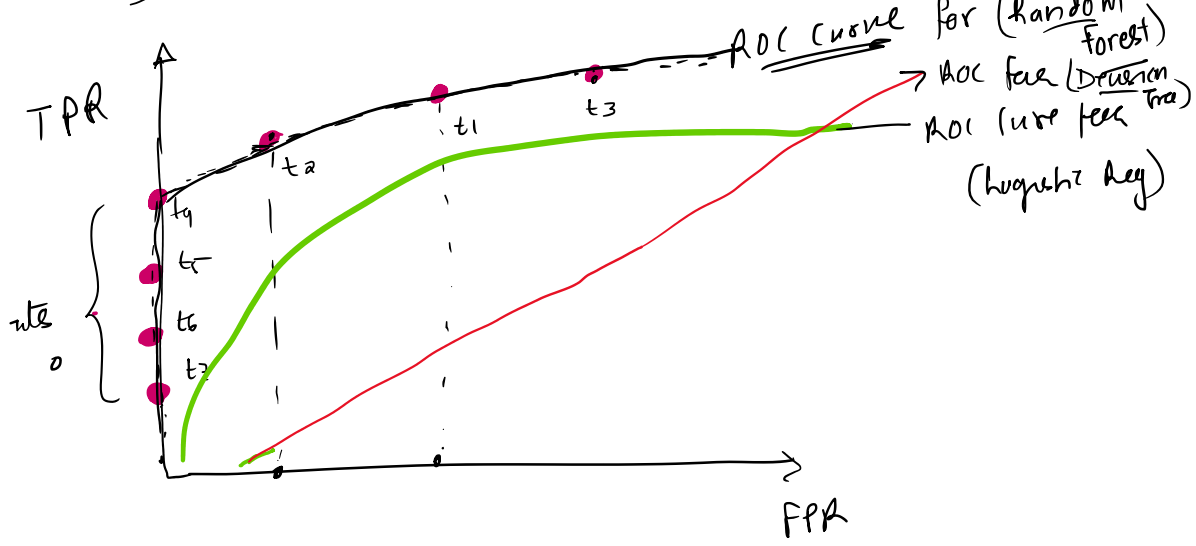


→ From the above ROC Curve for Random Forest model, it will help us to find the best threshold

AUC [Area Under Curve] → It is a method to compare ROC for more than one method and will help to judge which one is better.

* → Here will find Area Under ROC for each method.

→ ROC that has maximum Area, the corresponding method will be best.



From above ROC Curves, we find that there is maximum area under ROC Curve for Random forest, hence the method Random forest will be the best method.

Q2.A

Outline the steps involved in developing a Machine Learning application for predicting housing prices. Include data collection, preprocessing, model selection, training, evaluation, and deployment.

Solution:-

Steps

① Data collection → identify source to collect data
ie from real estate website, govt databases or real estate agencies.

* Scrape or download data containing relevant features and historical prices

② Data Preprocessing:

* Data Cleaning → Handling missing values
removing duplicates
Correcting inconsistencies

* Feature Engineering: → Create new features
* Normalize or standardize data
* Encode categorical variables

* EDA → Visualize data distributions
Correlations
Identify patterns and anomalies.

③ Model Selection * Algorithm choice
* Justification

(4) Data Splitting
* Train Test Split (approx 80% train, 20% test)
* Validation Set (opp 70% train
10% validation
20% Test)

(5) Model Training
* model Training Algo
* Hyperparameter tuning

(6) Model Evaluation
* Performance metrics
* Validation.

(7) Model Improvement
* feature Selection
* Algo optimization
* Regularization

(8) Deployment
* model Serialization
* API Development
* Integration.

Q2.B

Justify the significance of the F1 score in imbalanced classification problems. How does it address the limitations of using precision or recall alone? Provide a scenario where the F1 score might still be misleading.

F1-Score

* In reality we need a metric that takes into account both precision and recall.
* F1 score is a metric that takes into account both precision & recall.
* F1 score is harmonic mean of Precision & Recall.

$$\text{F1-Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Harmonic mean of two variables a & b

$$H = \frac{2}{\frac{1}{a} + \frac{1}{b}} = \frac{2ab}{a+b}.$$

for n variables

$$H = \frac{n}{\frac{1}{n_1} + \frac{1}{n_2} + \cdots \frac{1}{n_n}} = \frac{n}{n \sum_{i=1} \frac{1}{n_i}}$$

If F1-Score = 1 $\Rightarrow$ when Precision = 1, Recall = 1

* When Precision and Recall both are high then F1-Score is High

When to use F1-Score →

→ Accuracy is not a good metric to use when we have class imbalance.

Ex → Let say 99% of people visiting site are onlookers and

Ex → Let say 99% of people visiting site are onlookers and not purchasing anything.

→ Suppose we have a model that predicts that 100% people visiting site are onlookers.

→ The model is 1% wrong, Generally 1% Error is Acceptable

→ But such model in this case is useless

→ In such case instead of accuracy, we will prefer F1-Score.

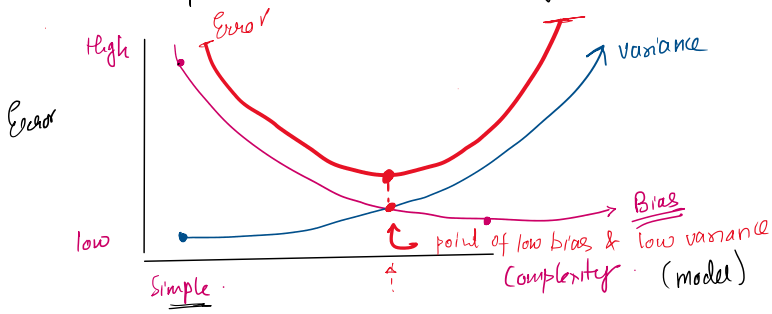
Q2.C

Justify the bias-variance trade-off in the context of model complexity. How does this trade-off influence the choice of model in a real-world ML application?

Bias - Variance Tradeoff

(1) If model is too simple and very few parameter → Underfitting → High Bias → Low Variance

(2) If model is complex and has large no of parameters → Overfitting → Low Bias → High Variance.



Bias Variance Tradeoff says

we need a model that gives

(1) low Bias

(2) low variance.

ie we need to find point of low Bias and low variance.

The ods to achieve this :->

* Regularization (Penalize 'Q' parameter)
* Boosting
* Bagging

*.

Impact on choice of models.

* Simple model (High Bias, Low Variance)

  Ex → Logistic Regression

  Linear Regression

* Complex model (Low Bias, High Variance)

  Ex → Decision Tree

Polynomial Regression

* Ensemble method : Random forest
  Gradient Boost machines } Can achieve low bias & variance as compared to individual models.

* Regularization Techniques
  * Helps preventing overfitting
  * Need careful tuning of regularizⁿ parameters

  Ex: Ridge Regression
  Lasso Regression.

1 meth

Q3.A

For a dataset given below, construct a decision tree using Gini Index, and determine which attribute is a root attribute.

Sr. No.	Weather	Parent	Money	Decision
1	Sunny	Yes	Rich	Cinema
2	Sunny	No	Rich	Tennis
3	Windy	Yes	Rich	Cinema
4	Rainy	Yes	Poor	Cinema
5	Rainy	No	Rich	Stay In
6	Rainy	Yes	Poor	Cinema
7	Windy	No	Poor	Cinema
8	Windy	No	Rich	Shopping
9	Windy	Yes	Rich	Cinema
10	Sunny	No	Rich	Tennis

Solution →

Independent features: Weather, Parent, Money

Decision/Outcome: Cinema, Tennis, Stay In, Shopping.

Shopping.

Step 1>

We will calculate Gini Index for Overall collection of Outcomes of Training Examples.

There are 4 possible outcomes for decision
- Cinema — 6 instances
- Tennis — 2 instances
- Stay In — 1 instance
- Shopping — 1 instance.

$$Gini_{(decision)} = 1 - \left(\left(\frac{6}{10}\right)^2 + \left(\frac{2}{10}\right)^2 + \left(\frac{1}{10}\right)^2 + \left(\frac{1}{10}\right)^2\right)$$
$$= 1 - \left(\frac{42}{100}\right) = 0.58$$

Step 2: To find Gini Index for money

Gini (money) → a possible value → Rich, poor

① Gini (money = Rich) → 7 instance

possible Decision:
- Cinema → 3 time
- Tennis → 2 time
- Stay In → 1 time
- Shopping → 1 time.

$$Gini_{(money=Rich)} = 1 - \left(\left(\frac{3}{7}\right)^2 + \left(\frac{2}{7}\right)^2 + \left(\frac{1}{7}\right)^2 + \left(\frac{1}{7}\right)^2\right)$$
$$= 0.694$$

② Gini → → Cinema.

(2) Gini (money=poor) → 3 instance → Decision → Cinema.

$= 1 - \left((3/3)^2\right) = 0$ //

Weighted Average $\text{Gini}_{(\text{money})} = \left(\text{Gini}_{(\text{money=Rich})} * \text{Proportion of rich}\right) + \left(\text{Gini}_{(\text{money=poor})} * \text{Proportion of poor}\right)$

$= (0.694 * 7/10) + (0 * 3/10)$

$\text{Gini}_{(\text{money})} = 0.485$

Step 3 Gini Index on Parent

For Parent feature → Possible values → Yes, NO

Gini (parent = Yes) = 5 instances → Possible Decision → Cinema

$= 1 - \left((5/5)^2\right) = 0$ //

Gini (parent = NO) = 5 instance → possible decision → Tennis → 2 times; Stay In → 1 times; Shopping → 1 time; Cinema → 1 time.

$= 1 - \left((2/5)^2 + (1/5)^2 + (1/5)^2 + (1/5)^2\right)$

$= 1 - (7/25) = 0.72$

Weighted Average of $\text{Gini}_{(\text{parent})} = (0 * 5/10) + (0.72 * 5/10) = 0.36$

$\text{Gini}_{(\text{parent})} = 0.36$

$$Gini_{(parent)} = 0.36$$

Step 4) Gini Index for Weather.

Weather → Possible values
- Sunny = 3 instances
- Windy = 4 instances
- Rainy ⇒ 3 instances.

Gini (weather = sunny) ⇒ 3 instances → possible outcomes
- Cinema → 1 time
- Tennis ⇒ 2 time

$$= 1 - \left((1/3)^2 + (2/3)^2 \right) = 0.444$$

Gini (weather = windy) → 4 instances → possible outcome
- 3 time Cinema
- 1 time Shopping

$$= 1 - \left((3/4)^2 + (1/4)^2 \right) = 0.375$$

Gini (weather = Rainy) → 3 instance → possible outcomes
- 2 cinema
- 1 Stay In

$$= 1 - \left((2/3)^2 + (1/3)^2 \right) = 0.444$$

$$\text{Weighted Average } Gini_{(Weather)} = (0.44 * 3/10) + (0.375 * 4/10) + (0.44 * 3/10)$$

$$= 0.414$$

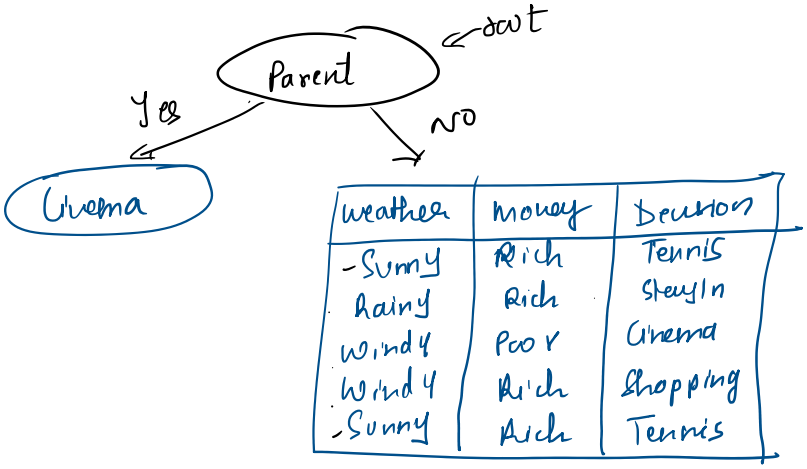
$$= 0.486$$

Gini (money) = 0.486
Gini (parent) = 0.36
Gini (weather) = 0.416

Minimum Gini ⟶ Minimum Impurity in Decision.

Here Minimum Gini Value = Gini (parent) = 0.36

So the root of Decision is Parent



weather	money	Decision
Sunny	Rich	Tennis
Rainy	Rich	Stay In
Windy	Poor	Cinema
Windy	Rich	Shopping
Sunny	Rich	Tennis

We need to find Gini (parent = No and weather)

Also Gini (parent = No and money)

Steps) Gini (parent = No and Weather) ⟶ Possible value ⟶ Sunny – 2 time, Windy – 2 time, Rainy – 1 time

5 instances

Gini (parent = No and weather = Sunny) ⟶ 2 instance ⟶ Tennis = $1 - \left(\left(\frac{2}{2}\right)^2\right) = 0$ // possible Outcome

Gini (parent = No and weather = Windy) ⟶ 2 instances ⟶ Cinema, Shopping = $1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right) = 0.5$ possible Outcome

Gini (parent = No and Weather = Rainy) → 1 instance → Stay in = $1 - ((1/1)^2) = 0$ (possible outcome)

Weighted Average Gini (parent = No and Weather) = $0.5 \times 2/5 = 0.2$

Step 6) Gini (parent = NO and Money) — 5 instance

- Rich = 4 time
- poor = 1 time

Possible values.

Gini (parent = NO and Money = Rich) → possible outcome (4)

- Tennis – 2
- Stay In – 1
- Shopping – 1

$= 1 - \left((2/4)^2 + (1/4)^2 + (1/4)^2 \right) = 0.625$

Gini (parent = NO and Money = Poor) → possible outcome

Cinema $= 1 - ((1/1)^2)$

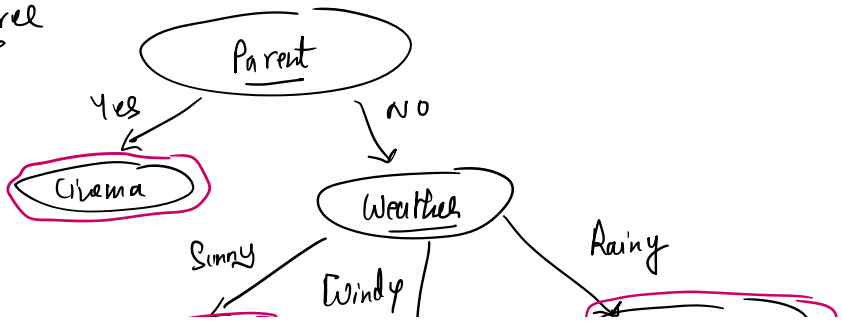
$= 0$

Weighted Average Gini (parent = No and Money) = $0.625 * 4/5 = 0.5$

Here Gini (parent = NO and Weather) has smallest value

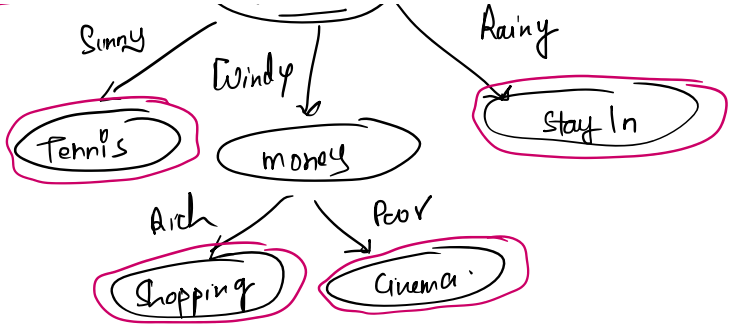
So now Next Node = Weather

* Updated Decision Tree



Ans.

Ans


Sunny
Windy
Rainy
Tennis
money
Stay In
Rich
Poor
Shopping
Cinema

Q3.B

Can you delve into the significance of Cross Validation .List the type and discuss k-fold cross validation in detail?

Cross Validation :→

→ In Supervised M.L

→ Train a model on a Dataset

→ Trained model is used to predict the target given new sample.

→ How to know if model we have trained will produce effective and accurate result on new input

Cross Validation → It is process that ensures the model will perform well on new Data.

Split the Dataset
- Training Set (60%)
- Validation Set (Hold-out set) (20%)
- Test Set (20%)

Training Set → part of data on which model is trained.
(This dataset will help to * build model)

* Validation Set → * Evaluate the Model
→ Will help to chk if model overfits or Underfits.
→ Update the parameters and again train the model.

→ Will help to [illegible]
→ Update the parameters and again train the model.
→ Repeat this until the model performs best on Validation set

Test Set ⇒ * Prediction.

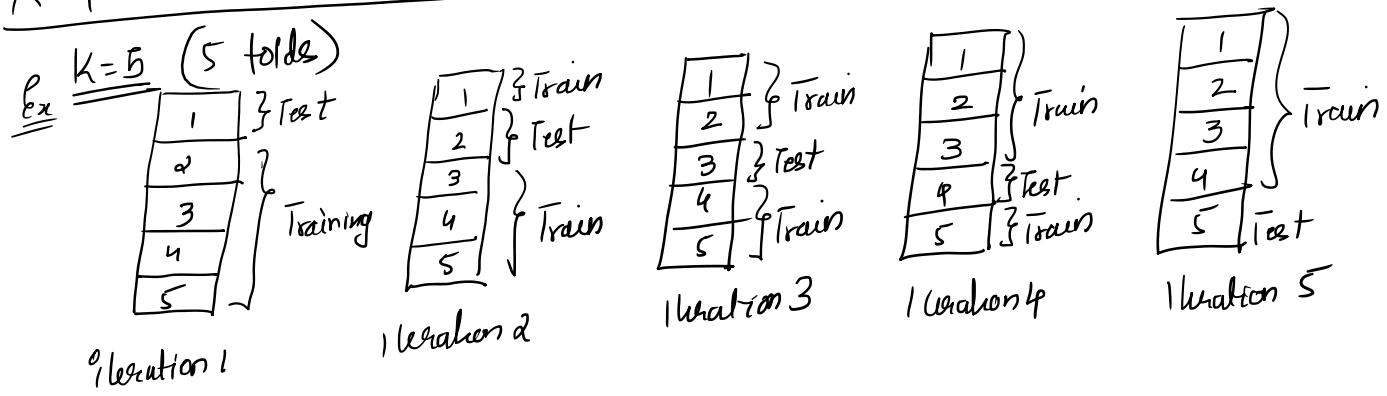
→ The fully trained model after being evaluated on validation set can be used on test set to generate Estimation.

Q) Types of Cross Validation ⇒

① The standard Validation Set Approach
② (LOOCV) Leave One Out Cross Validation.
③ K-fold Cross Validation.

K-fold Cross Validation ⇒

Ex K=5 (5 folds)



K fold ⇒ → K fold helps us to build the model in generalized form.

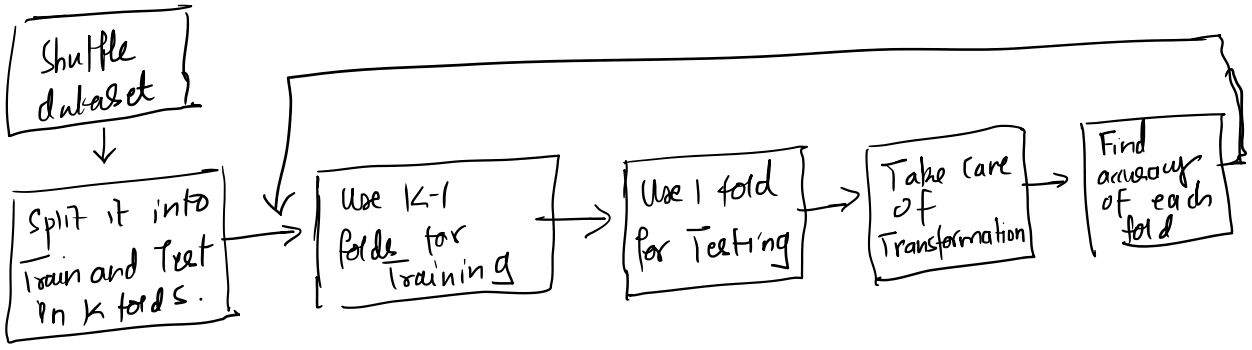
→ To achieve this K fold Cross Validation split the dataset into Training Testing & Validation.

→ To divide ... [illegible]

dataset into Training Testing & Validation.

→ Here Test and Train data will support building the model.

→ Life cycle of K-fold Cross Validation.



* The No of iterations ideally is K time
* Finding mean of accuracy score of each iteration will give the consistency of the Trained model.

Rules

① $K \geq 2$

if K == 2 → just 2 iterations.

if $K = n >> 2$ ⇒ n-1 for Training
1 for Testing

② most Commonly used value of $K = 10$

③ If K is very large then the running time of process will increase.

④ The value of K is inversely proportional to size of data i.e if dataset size is small then number of

data i'e if dataset Size is Small then number of folds can increase.