Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
First Semester 2024-2025

Comprehensive Exam
(EC-3 Regular)

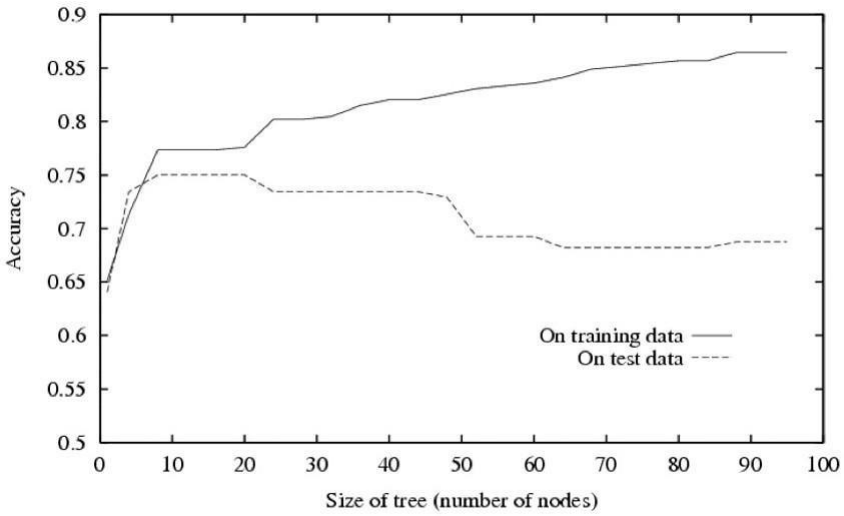
Course No.	: DSECLZG565/ AIMLCLZ565
Course Title	: MACHINE LEARNING
Nature of Exam	: Open Book
Weightage	: 40%
Duration	:
Date of Exam	:

No. of Pages = 3

Note:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Question 1: [6 marks]

a) You are training a **decision tree classifier** to learn a function f: X→Y, where both the **training** and **test examples** are independently sampled from an unknown data distribution P(X), and each instance is labeled with output Y. As the number of **nodes** in the tree (i.e., tree size) increases, the **training** and **test accuracy** curves evolve accordingly, as shown in the figure.



Now, based on the general behavior of training and test accuracy during tree growth, answer the following question.

Note: Marks will be awarded only if correct and relevant explanation is provided.

1) State whether the given statements are True/False. Justify your answer.
   *1.1) "The training error observed on the curve is an **unbiased estimate** of the true error."* [1 marks]

1.2) "*Increasing the depth or number of nodes in a decision tree increase the likelihood of overfitting."* [1.5 marks]

2) Based on the plotted training and test accuracy curves, which decision tree size (i.e., number of nodes) would you select for making predictions on new data, and why? Provide a brief justification for your answer. [2]
3) List and briefly describe at least **three pre-pruning strategies** that can be applied during decision tree construction to limit model complexity and reduce the risk of overfitting. [1.5]

1.1) False.

[0.25 marks for correct answer, 0.75 marks for the explanation]

1.2) True.

[0.25 marks for correct answer, 1.25 marks for the explanation]

2) 10 nodes **[1 mark]**. This has the highest test accuracy of any of the trees, and hence the highest expected true accuracy.[**1 mark for explanation]**

3) 1. Maximum Depth of the Tree
2. Minimum Number of Instances for a Node Split
3. Maximum Number of Leaf Node
4. Threshold: minimum threshold for information gain or Gini impurity reduction, that must be achieved for a node to split. If the best possible split at a node results in a gain or impurity reduction below the specified threshold, the split is **not made**, and the node becomes a **leaf**.

[Full Marks should be awarded if any 3 criteria with explanation are mentioned, 0.5 marks for each criteria only if explanation is provided.]

Question 2: [5 marks]

A telecom company is developing a logistic regression model to predict customer churn. Among the input features used are "Monthly Call Minutes"**,** which represents the total duration of calls made by a customer in a month, and "Total Number of Calls"**,** which counts how many calls the customer made. These two features are known to be highly positively correlated, as customers who make more calls generally accumulate more call time. How could this high correlation affect the model's performance and interpretability? Suggest at least **two ways** to address or reduce this issue in the model. [5]

It leads to multicollinearity, since these features are highly positively correlated. Multicollinearity does not affect the model's ability to make predictions, but it can have serious implications for interpretability and stability. The logistic regression algorithm struggles to determine the individual effect of each correlated feature on the target variable (customer churn), leading to unstable or misleading coefficient estimates. This makes it difficult to understand which feature is truly influencing the prediction. Moreover, the redundancy introduced by such features can cause

the model to become overly sensitive to small variations in the data, which may degrade performance on unseen data.

[1 marks only, if student has mentioned "Multi-collinearity issue", 1.5 marks for explaining its impact on the performance and interpretability]

To address this issue,

1) remove one of the correlated features.
2) Derived feature, combination of both the features.
3) Regularization techniques, like L2 (Ridge) regularization, which penalize large coefficients and help mitigate the effects of multicollinearity.
4) In more complex cases, dimensionality reduction techniques

[2.5 marks if any 2 approaches are mentioned and explanation is also provided]

Question 3: [5 marks]

You are working as a data scientist at a real estate analytics company. Your current task is to build a predictive model that estimates whether a newly listed property is likely to be classified as "High Interest" or "Low Interest", based on patterns observed in historical property listings. A listing marked as "High Interest" indicates strong buyer engagement (e.g., frequent inquiries, visits, or early offers), while "Low Interest" suggests limited attention from potential buyers.

To tackle this problem, your team has chosen to implement the k-Nearest Neighbors (k-NN) algorithm, using Gower's distance to handle the dataset's mixed feature types (numerical, categorical, and binary), assuming equal weights for all features. Use the below formula for Gowers computations.

$$d(i,j) = \frac{\sum_{c=1}^{n} \omega_c \delta_{ij}^{(c)} d_{ij}^{(c)}}{\sum_{c=1}^{n} \omega_c \delta_{ij}^{(c)}}$$

$d(i,j)$ = dissimilarity between row i and row j

c = the cth column

n = number of columns in the dataset

ω_c = weight of cth column = $\frac{1}{\text{nrows in dataset}}$

$$\delta_{ij}^{c} = \begin{cases} 0 & \text{if column c is missing in row i or j} \\ 0 & \text{if column c is asymmetric binary and both values in row i and j are 0} \\ 1 & \text{otherwise} \end{cases}$$

$$d_{ij}^{c}(\text{categorical}) = \begin{cases} 0 & \text{if i and j are equal in column c} \\ 1 & \text{otherwise} \end{cases}$$

$$d_{ij}^{c}(\text{continuous/ordinal}) = \frac{|\text{row i in column c - row j in column c}|}{\text{max(column c) - min(column c)}}$$

Consider the dataset provided below:

- **Location Type** — *(Categorical: "Urban", "Suburban", "Rural")*
- **Price per square foot (₹) (in 1,000)** — *(Numerical: Range is 4-7)*
- **Has Garden**? — *(Binary: Yes / No)*
- **Buyer_Interest?** — *(Target: High Interest / Low Interest)*

Listing	Price/ft^2 (in 1,00)	Location Type	Has Garden?	Buyer_Interest?
A	4	Urban	Yes	High
B	4.5	Suburban	No	High
C	5	Urban	No	Low
D	5.5	Suburban	Yes	Low
E	6	Urban	Yes	High
F	7	Rural	Yes	High

Based on the problem described above, predict the target variable (Buyer_Interest?) for the following query instance using a weighted 3-Nearest Neighbors (3-NN) model. Apply inverse distance weighting, where: weights, $w_i = 1/d_i$, and d_i, is the Gower distance between the query instance and the i^{th} neighbor.

Query Instance: <Price/ft^2 (in 1,000) = 5.3, Location Type = "Suburban", Has Garden = No>

Note: Full marks will be awarded only if all intermediate steps and calculations are clearly shown.

3 nearest neighbors as per Gowers computations: B, C, D

Weight = 1/Gower_distance

Weighted Voting:
For High Interest: 11.25
For Low Interest: 2.81+2.72 ~=5.5

The predicted class for the query instance **<Price/ft² = 5300, Location Type = "Suburban", Has Garden = No>** using **weighted 3-NN** with **Gower distance** is, High

[3 marks for Gower computation and finding 3 nearest neighbors. Full marks will be awarded if all 3 nearest neighbors are identified correctly, and all steps are shown properly. Otherwise, 1 marks for mentioning 3 nearest neighbours with no computations]
[1 mark for weighted voting]
[1 mark for predicting class based on weighted voting]

Question 4: [5 marks]

Naive Bayes classifier is used to categorize text into two classes: Technology and Non-Technology. The following table shows the training data:

Text	class
"The new smartphone is amazing"	Technology
"The event was well organized"	Non-Technology
"Groundbreaking AI algorithm"	Technology

Predict the class of the sentence "A new AI gadget" using the Naive Bayes algorithm (Multinomial model) with Laplace smoothing. [5]
Solution:

Vocabulary = the, new, smartphone, is, amazing, event, was, well, organized, Groundbreaking, AI, algorithm

Total Vocabulary Size (V) = 12 [0.5M]

Prior [0.5M]
P(Technology)=2/3
P(Non-Technology)=1/3

[0.5M]
$N_c = N_{tech}$ = no of words in technology class =8
$N_c = N_{nontech}$ = no of words in Non-Technology class =5

Conditional Probabilities for Each Class (with Laplace Smoothing)
[2M for table]

Word	Count (Technology)	P(word|Technology)	Count (Non-Technology)	P(word|Non-Technology)
a	0	0+1 / 8+12 = 1/20 = 0.05	0	0+1 / 5+12 = 1/17 = 0.059
new	1	1+1 / 8+12 = 2/20 = 0.10	0	0+1 / 5+12 = 1/17 = 0.059
AI	1	1+1 / 8+12 = 2/20 = 0.10	0	0+1 / 5+12 = 1/17 = 0.059
gadget	0	0+1 / 8+12 = 1/20 = 0.05	0	0+1 / 5+12 = 1/17 = 0.059

Posterior Probabilities for the Test Sentence "A new AI gadget" [1M]

Technology Class:

$P(\text{Technology}|\text{"A new AI gadget"}) \propto P(\text{A}|\text{Tech}) \times P(\text{new}|\text{Tech}) \times P(\text{AI}|\text{Tech}) \times P(\text{gadget}|\text{Tech}) \times P(\text{Tech})$
$=0.10\times0.10\times0.05\times0.05\times0.667= 0.000016675$

Non-Technology Class:

$P(\text{NonTechnology}|\text{"A new AI gadget"}) \propto P(\text{A}|\text{NonTech}) \times P(\text{new}|\text{NonTech}) \times P(\text{AI}|\text{NonTech}) \times P(\text{gadget}|\text{NonTech}) \times P(\text{NonTech})$
$=0.059\times0.059\times0.059\times0.059\times0.333= 0.000004035$

Final Answer: The sentence **"A new AI gadget"** is classified as **Technology**. [0.5 M]

Question 5: [1.5 * 4 = 6 marks]

Suppose you have a dataset that has 10 features and 10,000 training instances. You have applied logistic regression with gradient descent on this dataset and trained a model. Unfortunately, this ML model exhibits poor performance on both the training data and test data. To address this issue, your team members have proposed several solutions, as mentioned below. Suggest which of the following looks promising in the given scenario and provide reasons for your choice?

1. Use SVM with linear kernel without adding any new feature
2. Increase the regularization parameter λ, in logistic regression
3. Use SVM with RBF kernel
4. Transform the dataset using polynomial transformation and then use logistic regression

True/False	Answer	Explanation
False	Use an SVM with a linear kernel, without introducing new features.	An SVM with only the linear kernel is comparable to logistic regression, so it will likely underfit the data as well.
True	Use an SVM with a Gaussian Kernel	By using a Gaussian kernel, your model will have greater complexity and you can avoid underfitting the data.
False	Increase the regularization parameter λ	You are already underfitting the data and increasing the regularization parameter only makes underfitting stronger.
True	Create / add new polynomial features	When you add more features, you increase the variance of your model, reducing your chances of underfitting.

[0.25 for True/False and 1.25 marks for justification in each case]

Question 6: [7 marks]

a) Gradient Boosting models use a combination of multiple weak learners, where each new learner attempts to correct the mistakes made by the previous ones. One of the most important hyper-parameters in this process is the *learning rate*. Answer the following questions: **[1+2+2 = 5marks]**

Note: Marks will be awarded only if correct and relevant explanation is provided.

1) Explain what the learning rate represents in the context of Gradient Boosting.
2) Describe what might happen if the learning rate is set too small or too large.
3) Why is it important to find the right balance between the learning rate and the number of estimators (i.e., number of boosting rounds)? Support your explanation with the trade-offs involved.

1) [1 mark if correct explanation is provided]

2) What might happen if the learning rate is set too small or too large:

The learning rate controls how quickly a boosting model learns. A very small learning rate leads to slow learning and potential underfitting (especially if #estimators are also low), while a very large one can cause overfitting or instability. Choosing the right learning rate is key to achieving a good balance between model performance and training efficiency.

[1 mark for each if correct explanation is provided. Marks will not be awarded by simply mentioning overfitting and underfitting without explanation]

3) There's a trade-off between learning rate and the number of estimators in boosting models. A low learning rate needs more estimators for good performance, while a high learning rate trains faster but risks overfitting. Tuning both together using cross-validation helps find the right balance.

[Full marks will be awarded only if the trade-offs are correctly identified and clearly explained]

b) Consider an ensemble of 3 independent 2-class classifiers, each of which has an error rate of 0.3. The ensemble predicts class of a test case based on majority decision among the classifiers. Calculate the error rate of this ensemble classifier. Show all calculations clearly. (2 marks)

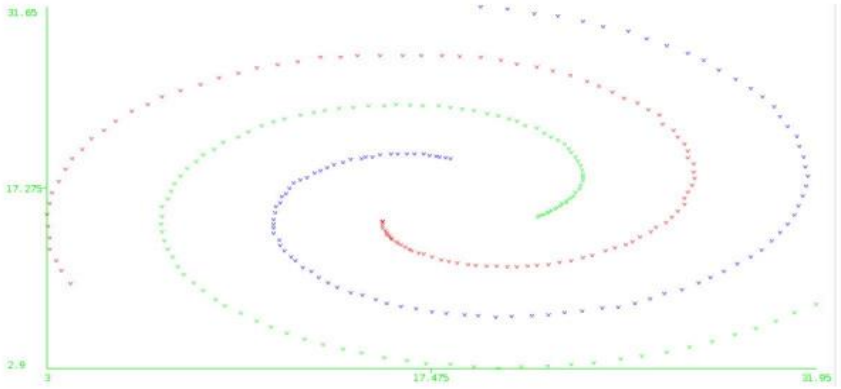
*Answer : New error rate = (0.3)**3 + 3*((0.3)**2)*0.7 = 0.216*

[full marks for the correct answer and if all steps are shown]

Question 7: .[3 + 3 = 6 Marks]

a) Consider the following data distribution. It shows that the data points belong to 3 different clusters, each cluster is denoted by different color. Answer the following questions and comment on clustering results in each case. [3]

1) What will happen if you apply k-means on the given data to cluster the dataset, assuming k is 3
2) What will happen if you apply GMM using EM on the given data to cluster the dataset, assuming 3 Gaussian components.



K-means clustering assumes spherical, linearly separable clusters and relies on Euclidean distance, so it performs poorly on datasets with complex shapes like spirals. It groups nearby points even if they belong to different spiral, leading to incorrect clustering. Gaussian Mixture Models (GMM) with Expectation-Maximization are more flexible than K-means, as they consider covariance and allow elliptical clusters, but they also struggle with non-linear, non-Gaussian structures like spirals. While GMM may slightly outperform K-means, both methods fail to capture the true spiral patterns in the data.

[1.5 marks in each case should be awarded if correct explanation is provided]

b) Randomly initializing the means and covariance matrices in a Gaussian Mixture Model (GMM) can lead to suboptimal results. Explain the potential issues that may arise from such initialization. How can these issues be addressed during the initialization phase? [3]

Random initialization in Gaussian Mixture Models (GMM) can lead to poor clustering results, as the EM algorithm may converge to local optima or slow down significantly. To address this, using K-Means for initializing means provides a more informed starting point, while multiple random restarts help increase the likelihood of finding a better solution by selecting the model with the highest log-likelihood.

[1.5 marks will be awarded for highlighting any one of the potential issues along with an explanation, and 1.5 marks for mentioning any one of the solutions.]