Total No. of Pages: 03 *Roll No. .....*

FIFTH SEMESTER **B.Tech. [CSE]**

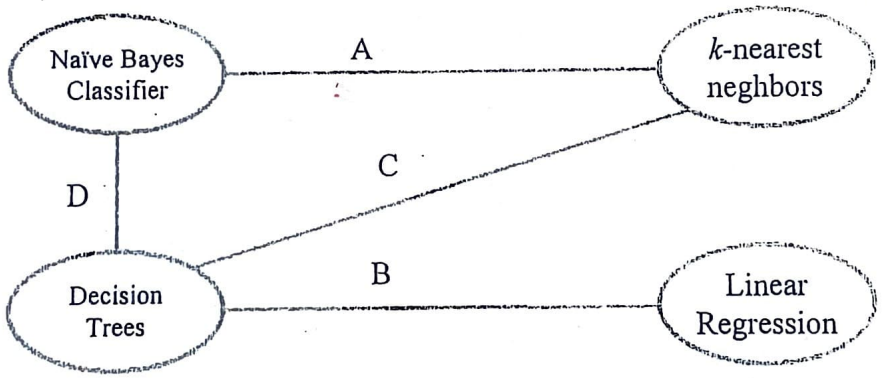
MID SEMESTER EXAMINATION *Sept-Oct, 2023*

CO327 MACHINE LEARNING

Time: 1:30 Hours *Max. Marks: 20*

Note: Answer ***ALL*** questions.
Assume suitable missing data, if any.
CO# is course outcome(s) related to the question.

1[a] Consider the figure below where four machine learning algorithms are connected by edges labeled A, B, C, and D. Label each edge with one characteristic or the difference the algorithms share. These labels should be concise and address basic concepts, such as Types of Learning Problems, Loss Functions, and Hypothesis Space (the set of all possible models or hypotheses that can be chosen by a learning algorithm to make predictions or classifications).



Note: When labeling edges between algorithms for example: SVM and Random Forest, pinpoint key differences. SVM, a linear model, excels in binary classification by optimizing a separating hyperplane, while Random Forest, an ensemble method, combines decision trees. Thus, labels like "Linear Model vs Ensemble Model" or "Hyperplane vs Tree Structure". **[2] [CO2]**

[b] Suppose, you are a ML engineer at TechPioneer, working on two distinct projects, each with unique goals and datasets. Determine, which one among the Supervised, Unsupervised or Reinforcement Learning, is most appropriate for each project based on the provided scenarios.

Project Beta: TechPioneer aims to create a recommendation system for a music app. The available dataset contains user interactions with different songs, but there are no labels indicating user preferences. The objective is to group users with similar music tastes to offer personalized song recommendations.

Project Gamma: TechPioneer is developing an intelligent agent for a game, where the agent needs to navigate through various levels, adapt, and make decisions to maximize the score, receiving points as rewards and penalties. **[1+1] [CO1]**

2[a] You are provided with a dataset (Table-I) containing profiles of customers, including their age group, income range, gender, and purchase history. Construct a decision tree to predict whether a given customer will make a purchase (Yes) or not (No), using the dataset. Clearly illustrate each step involved in the construction process. Use the GINI index to evaluate splits in the dataset.

Table I

Customer	Age Group	Income Range	Gender	Purchase Decision
1	Young	Low	Male	No
2	Young	Medium	Female	Yes
3	Middle-Aged	Low	Male	No
4	Middle-Aged	Medium	Female	Yes
5	Senior	Low	Female	No
6	Senior	Medium	Female	Yes
7	Young	Low	Male	Yes
8	Young	Medium	Female	No

Predict the Purchase Decision for a new customer with the following profile: Age Group: Middle-Aged, Income Range: Medium, Gender: Male using the above decision tree. Note: In case of tie between 'Yes' and 'No', break the tie by assigning it to 'Yes' class (Purchase Decision). **[3] [CO2]**

[b] Perform testing of the above decision tree on the training dataset (Table-II) and calculate its accuracy. Note: In case of tie between the 'Yes' and 'No' classes, break the tie by assigning it to 'Yes' class (Purchase Decision). **[2] [CO2]**

[c] Prune the decision tree, constructed in 2[a], to reduce its depth by one level. Assess whether this pruning will impact the accuracy of the decision tree. **[2] [CO2]**

3[a] Consider the dataset in Table-I again and design a Naïve Bayes Classifier and apply it to predict the Purchase Decision for a new customer with the following profile: Age Group: Middle-Aged, Income Range: Medium, Gender: Male. **[3] [CO2]**

[b] Why is the Naive Bayes classifier considered 'naive'? Also, comment on prediction (Purchase Decision) by two classifiers (Decision tree and Naïve Bayes Classifier) for same data instance: Age Group: Middle-Aged, Income Range: Medium, Gender: Male. **[2] [CO2]**

4[a] Suppose, you are a ML engineer at a cybersecurity firm. System X is designed to detect potential phishing websites. A failure to detect a phishing website could lead to significant data breaches, making it crucial to identify as many phishing websites as possible, even if some legitimate websites are mistakenly classified as phishing. You are provided with the results of system X on a sample set: Number of websites correctly identified as phishing: 120, Number of legitimate websites incorrectly identified as phishing: 30, Number of phishing websites missed (classified as non phishing websited) by the system: 10. Find the precision and recall of the system X and identify which is more crucial measure of performance precision or recall for system X.

[2] [CO2, CO3]

[b] Now, suppose there is another system Y, developed to identify the illegal sharing of copyrighted material. Incorrectly classifying legal content as illegal could lead to legal complications and loss of reputation, so it's vital to be as accurate as possible in the positive cases to avoid false accusations. You are provided with the results of system Y on a sample set: Number of instances correctly identified as illegal sharing: 50, Number of legal content instances incorrectly identified as illegal: 5, Number of illegal content instances missed by the system: 20. Find the precision and recall of the system Y and identify which is more crucial measure of performance precision or recall for system Y.

[2] [CO2, CO3]

---Best of Luck---