National Forensics Sciences University, Goa Campus
TA-1 Examination

Program Name – MTECH AIDS
Subject Name- Fundamentals of DS and ML
Time- 45 mins
Sem – I
Subject Code - CTMTAIDS SI P4
Date- 10-09-2024
Instructions - 1) Answer all questions. 2) Assume suitable data.
Max. Marks- 25

Q.1 Multiple Choice Questions (1 mark each) 10 marks

1. **In the Naive Bayes algorithm, what assumption is made about the features used for classification?**
   A. All features are correlated and dependent.
   B. Features are conditionally independent given the class label.
   C. Features must follow a Gaussian distribution.
   D. The class label is independent of the features.

2. **What is the main drawback of using the Naive Bayes classifier for a dataset with continuous variables?**
   A. It requires all features to be categorical.
   B. It cannot model data with missing values.
   C. It assumes independence between continuous features, making Gaussian approximations necessary.
   D. It always suffers from overfitting in high-dimensional spaces.

3. **Which of the following best describes the process of data wrangling in the data science lifecycle?**
   A. Transforming data into a more understandable format
   B. Eliminating irrelevant features from the dataset
   C. Creating new features based on existing ones
   D. Visualizing data using exploratory plots

4. **Which type of data would be most appropriately represented using a one-hot encoding technique?**
   A. Continuous numerical data
   B. Ordinal categorical data
   C. Nominal categorical data
   D. Ratio-scaled numerical data

5. **Which of the following best describes the p-value in hypothesis testing?**
   A. The probability that the null hypothesis is true
   B. The probability of obtaining a test result at least as extreme as the observed one, assuming the null hypothesis is true
   C. The probability of rejecting the alternative hypothesis
   D. The probability of making a Type II error

6. **Given two independent events A and B, which of the following is the correct way to calculate the probability of both A and B occurring?**
A. $P(A) + P(B)$
B. $P(A) \times P(B)$
C. $P(A) / P(B)$
D. $P(A \cup B)$

7. **Which of the following types of data would be categorized as discrete?**
A. The temperature of a city over a week
B. The number of cars in a parking lot
C. The weight of a person
D. The height of students in a class

8. **Which term refers to the difference between the expected value of an estimator and the true value of the parameter being estimated?**
A. Variance
B. Bias
C. Standard Error
D. Mean Squared Error

9. **If you have a dataset where 60% of the rows contain missing values in one column, which of the following is the best approach?**
A. Drop all rows with missing values
B. Impute missing values using the median
C. Impute missing values using a random value
D. Drop the column entirely

10. **In the context of Bayes' theorem, which of the following best describes posterior probability?**
A. The probability of the hypothesis given the prior knowledge
B. The probability of the data given the hypothesis
C. The updated probability of the hypothesis after observing new data
D. The probability of observing the data, marginalizing over all possible hypotheses

Q.2 **Answer any 3 questions (3x5 marks each)** **15 Marks**

A. Write a note on different types of Data in Data Science

B. A standard deck of 52 cards contains 4 suits: hearts, diamonds, clubs, and spades. Each suit has 13 cards. What is the probability of drawing a spade or a queen from a shuffled deck?

C. What data transformation technique scales the data to fall within a specific range, usually [0, 1]. Write the formula. Also, explain one more data transformation technique with formula.

D. Explain different Data Processing techniques.

END