CA-2

INT-234

Predictive Analytics

Roll No:\_\_\_\_\_\_\_\_\_\_\_\_ Section:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Note:**

**1. All questions are compulsory.**

**2. Copy cases will be awarded zero without any explanation.**

**3. Each question is of 15 marks.**

Q1. Apply a decision tree classifier to the 'mushrooms.csv' dataset to predict whether a mushroom is edible or poisonous based on its characteristics. Split the dataset into 80% for training and 20% for testing. Train the model on the training set, then evaluate its performance on the test set by generating a confusion matrix and calculating the model's accuracy.Analyze the output in 2 points.

Q2. Analyze the 'heart.data.csv' dataset to determine whether there is a statistically significant linear relationship between the independent variables—smoking and biking—and the dependent variable, heart disease. Perform a multiple linear regression analysis, using smoking and biking as predictor variables and heart disease as the outcome variable." Analyze the output in 2 points.

CA2

INT-234

Predictive Analytics

Roll No:\_\_\_\_\_\_\_\_\_\_\_\_ Section:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Note:**

**1. All questions are compulsory.**

**2. Copy cases will be awarded zero without any explanation.**

**3. Each question is of 15 marks.**

Q1. Use a naive bayes classifier to analyze the "titanic\_data.csv" dataset and predict the likelihood of passenger survival after the Titanic struck the iceberg. First, split the dataset into two parts: 80% for training and 20% for testing. Once the model is trained on the training data, evaluate its performance on the test set by creating a confusion matrix and calculating the accuracy of the decision tree model.

Additionally, provide a detailed explanation of the decision tree structure, emphasizing the key criteria used for splitting (such as age, gender, and passenger class). Highlight the relative importance of each feature in predicting survival. Discuss insights gained from the model, including how it forecasts survival, the decision boundaries it forms, and the thresholds used for making predictions.Analyze the output in 2 points.

Q2. Utilize R's built-in 'Cars' dataset to analyze the relationship between an automobile's speed and stopping distance. Use 80% of the dataset for training to model the effect of speed on stopping distance through simple linear regression. Assess the strength of the relationship between these variables by examining the regression coefficients and calculating the correlation between speed and stopping distance. Analyze the output in 2 points.

CA2

INT-234

Predictive Analytics

Roll No:\_\_\_\_\_\_\_\_\_\_\_\_ Section:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Note:**

**1. All questions are compulsory.**

**2. Copy cases will be awarded zero without any explanation.**

**3. Each question is of 15 marks.**

Q1. In the given dataset (Candidate.csv), Data is given for exam centers where a government exam has been conducted. Find the correlation using regression to represent the most significant and least significant features. Plot the so formed output. Analyze the output in 2 points.

Q2. Using a cancer dataset that contains various features such as tumor size, texture, perimeter, and other clinical attributes, how can you implement and evaluate a Naive Bayes classifier in R to predict whether a tumor is benign or malignant? Start by loading and preprocessing the dataset, handling any missing values and splitting it into training and test sets. Train the Naive Bayes model using the training data, ensuring that both categorical and continuous attributes are appropriately handled. Once the model is trained, use it to predict the classifications of the test set. Finally, evaluate the model’s performance using a confusion matrix, and calculate metrics such as accuracy, precision, and recall. What do these evaluation metrics reveal about the effectiveness of the model in predicting the malignancy of tumors?"