**CHRIST (Deemed to be University)**

**Department of Computer Science**

**MAI171-Machine Learning**

**ESE Component Test Date:12-9-2024 Duration:1.30 hrs Maximum Marks:30 Marks**

**Evaluation Rubrics:**

**Execution:**20 Marks

**Visualization:**5 Marks

**Summary Report:**5 Marks

**Total:**30 Marks

**Instructions:**

Before taking the test, you must execute the code provided in the Google Classroom file named "Test6." Failure to do so will result in a zero mark for the test.

**Question**

You are working with the Titanic dataset to classify passengers into two categories: survived or not survived. You will build both logistic regression and decision tree models to perform this classification. After building and evaluating both models, you will compare their performance and perform a statistical analysis to gain further insights.

**1. Data Exploration**

Load and explore the Titanic dataset. Provide summary statistics, visualize data distributions, and investigate relationships between key features (e.g., Age, Sex, Fare, Pclass) and survival status.

**2.Data Splitting**

Split the Titanic dataset into training and testing sets (e.g., 70% for training, 30% for testing). Describe why this step is essential in model development.

**3. Model Building and Comparison**

**- Logistic Regression Model:** Construct a logistic regression model to predict survival based on the selected features.

**- Decision Tree Model:** Construct a decision tree model to predict survival based on the same features.

**- Model Training:**Train both the logistic regression and decision tree models on the training data. Visualize the decision boundaries for both models if applicable and create a comparison graph of their performance metrics.

**- Model Comparison:** Plot and compare the results of both models, including accuracy, precision, recall, and F1-score.

**4. Model Evaluation and Interpretation**

**- Confusion Matrix:** Compute and present the confusion matrix for both models applied to the testing data.

**- Accuracy and Precision:** Calculate and compare the accuracy and precision of both models.

**5. ROC Curve and AUC**

**- ROC Curve:** Construct an ROC (Receiver Operating Characteristic) curve for both models and calculate the AUC (Area Under the ROC Curve) scores. Interpret the results and discuss the effectiveness of each model in predicting survival.

**6. Statistical Analysis**

**- Chi-Square Test:**Conduct a Chi-square test of independence to examine the relationship between categorical variables (e.g., `Sex` and `Survived`). Provide the test statistic and p-value, and interpret the results to determine if there is a significant association between these variables.

**7. Model Selection**

**-Recommendation:** Based on your analysis and comparison of the models' performance, recommend which model (logistic regression or decision tree) provides better accuracy and overall performance for this dataset. Justify your choice with the results obtained.