Week 3 Assignment - MSc AIDA UoM

Professor Nikolaos Samaras

Academic Year: 2021-2022

1. Titanic Data Set

The Titanic data set (Titanic.xlsx) includes 2201 entries that represent the passengers and crew of the Titanic when it sank. Each record is defined by four attributes:

- Type of accommodation on the ship (first, second, third, crew)
- Age (adult, child)
- Sex (male, female)
- The fourth attribute is the outcome attribute, indicating whether the person who survived narrated their tragic story to the authorities.

Write Python code that will:

1. Read the file `Titanic.xlsx`.
2. Randomly select 80% of the records as the training set with the criterion: the training set should maintain the same ratio between the two classes as in the entire data set. The remaining 20% of the records will be used as the test set.
3. Use the `DecisionTreeClassifier` algorithm to develop a prediction model where the dependent variable is the outcome (whether the person survived and narrated their story) and the independent variables are the type of accommodation, age and sex.
4. Compute the confusion matrix to assess the performance of the model.

2. Cardiology Data Set

The Cardiology data set (Cardiology.xlsx) contains 303 entries. Out of these, 165 entries include information about patients who do not suffer from heart disease, while the remaining 138 entries include information about patients suffering from heart disease. The data set includes 13 input features (columns A to M) and one output feature (column N).

The features are described as follows:

Feature	Type	Description
Age	Numeric	Age in years
Sex	Categorical	Male, Female
Chest Pain	Categorical	Type: Angina, Abnormal Angina, Not Anginal (Non-Anginal), Asymptomatic
Blood Pressure	Numeric	Resting blood pressure upon hospital admission
Cholesterol	Numeric	Serum cholesterol
Blood Sugar	Binary	Fasting blood sugar < 120 mg/dL (True, False)
Resting ECG	Categorical	Results: Normal, Abnormal, Hypertrophy
Max Heart Rate	Numeric	Maximum heart rate achieved
Induced Angina	Binary	Yes, No
Old peak	Numeric	ST depression induced by exercise relative to rest
Slope	Categorical	Up, Flat, Down
Numbered Colored Vessels	Numeric	Number of major vessels colored by fluoroscopy (0-3)
Thal	Categorical	Normal, Fixed defect, Reversible defect
Concept Class	Categorical	Healthy, Sick

Write Python code that will:

1. Read the file Cardiology.xlsx.
2. Randomly select 75% of the records as the training data set, ensuring the same ratio between the two classes as seen in the entire data set. Use the remaining 25% as the test data set.
3. Use the DecisionTreeClassifier algorithm to develop a prediction model where the dependent variable is the 'Class' and the independent variables are the remaining ones.
4. Calculate the confusion matrix to assess the performance of the model.