Exercise 5 : Classification Tree

Setup of Working Environment

1. Create a folder on your Desktop and name it Cx1015_[LabGroup], where [LabGroup] is the name of your Group
2. Download the .ipynb files and data files posted corresponding to this exercise and store in the aforesaid folder
3. Open Jupyter Notebook (already installed on the Lab computer) and navigate to the aforesaid folder on Desktop
4. Open and explore the .ipynb files (notebooks) that you downloaded, and go through "Preparation", as follows
5. Create a new Jupyter Notebook, name it Exercise5_solution.ipynb, and save it in the same folder on the Desktop

Preparation for the Exercises

M4 ClassificationTree.ipynb	Check how to perform basic Classification on the Pokemon data (pokemonData.csv)

Objective

Note that our Housing Data has a Binary (two-level) Categorical Variable named "CentralAir", with values "Y" and "N". In the last Example Class, we have identified and analyzed some of the most relevant numeric variables in this dataset. In this Example Class, we will try to predict if a House has Central Air Conditioning or not using those Numeric Variables.

Problems

Problem 1 : Predicting CentralAir using SalePrice

Download the Kaggle dataset "train.csv" from NTU Learn, posted corresponding to this Example Class.
Import the complete dataset "train.csv" in Jupyter, as `houseData = pd.read_csv('train.csv')`

a) Plot the binary distribution of `houseData['CentralAir']` using `catplot` to check the ratio of Y against N. Plot `houseData['CentralAir']` against `houseData['SalePrice']` using `boxplot`, and note the strong relationship. You may also want to check the mutual relationship by plotting the two variables using a `swarmplot`.

b) Import Classification Tree model from Scikit-Learn : `from sklearn.tree import DecisionTreeClassifier`

c) Partition the complete dataset `houseData` into `houseData_train` (1100 rows) and `houseData_test` (360 rows).

d) Training : Fit a Decision Tree model for classification of CentralAir using SalePrice using the following variables.

```
y_train = pd.DataFrame(houseData_train['CentralAir'])
X_train = pd.DataFrame(houseData_train['SalePrice'])
```

e) Visualize the Decision Tree model using graphviz (needs the packages to be installed; check if they are installed).

f) Predict CentralAir for the train dataset using the Decision Tree model, and plot the Two-Way Confusion Matrix. Predict CentralAir for the test dataset using the Decision Tree model, and plot the Two-Way Confusion Matrix.

g) Discuss with your Friends all the accuracy parameters of the decision tree model, including its Classification Accuracy, True Positive Rate, True Negative Rate, False Positive Rate and False Negative Rate.

Problem 2 : Predicting CentralAir using Other Variables

Perform all the above steps on `'CentralAir'` against each of the variables `'GrLivArea'`, `'LotArea'`, `'TotalBsmtSF'` one-by-one to obtain individual Decision Trees. Discuss with your Friends about the models, compare the Classification Accuracy, check the True Positives and False Positives, and determine which model is the best to predict `'CentralAir'`.