Subject: 23CSE301

Lab Session: 04

Notes:

1. Please read the assignment notes carefully and comply to the guidelines provided.
2. Code should be checked into the GitHub. These details shall be provided in the Lab.
3. If you have not completed the prerequisite assignments, please complete them before the next lab session.

Coding Instructions:

1. The code should be modularized; The asked functionality should be available as a function. Please create multiple functions if needed. However, all functions should be present within a single code block, if you are using Jupyter or Colab notebooks.
2. There should be no print statement within the function. All print statements should be in the main program.
3. Please use proper naming of variables.
4. For lists, strings and matrices, you may use your input values as appropriate.
5. Please make inline documentation / comments as needed within the code blocks.

Main Section (Mandatory):

Please use the data associated with your own project. This assignment deals with classification models.

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

Please use help manuals of sklearn package to gain understanding of the model behaviors as well as ways to use various package functionalities.

Project teams who have not completed last week's assignments, please complete them.

A1. Please evaluate confusion matrix for your classification problem. From confusion matrix, the other performance metrics such as precision, recall and F1-Score measures for both training and test data. Based on your observations, infer the models learning outcome (underfit / regularfit / overfit).

A2. Calculate MSE, RMSE, MAPE and R2 scores for the price prediction exercise done in Lab 02. Analyse the results.

A3. Generate 20 data points (training set data) consisting of 2 features (X & Y) whose values vary randomly between 1 & 10. Based on the values, assign these 20 points to 2 different classes (class0 - Blue & class1 – Red). Make a scatter plot of the training data and color the points as per their class color. Observe the plot.

A4. Generate test set data with values of X & Y varying between 0 and 10 with increments of 0.1. This creates a test set of about 10,000 points. Classify these points with above training data using kNN classifier (k = 3). Make a scatter plot of the test data output with test points colored as per their predicted class colors (all points predicted class0 are labeled blue color). Observe the color spread and class boundary lines in the feature space.

A5. Repeat A4 exercise for various values of k and observe the change in the class boundary lines.

A6. Repeat the exercises A3 to A5 for your project data considering any two features and classes.

A7. Use ***RandomSearchCV()*** or ***GridSearchCV***() operations to find the ideal 'k' value for your kNN classifier. This is called hyper-parameter tuning.

Report Assignment:

1. Update your understanding of your project in the introduction section of the report.
2. Study the downloaded papers & update the literature survey section of your report.
3. Expand the methodology and results sections with outcomes of this experiments & results obtained. Please discuss your observations, inferences in results & discussion section. Please conclude the report appropriately with these experiments. Consider following points for observation analysis & inferences.
    - Do you think the classes you have in your dataset are well separated? Justify your answer.
    - Explain the behavior of the kNN classifier with increase in value of k. Explain the scenarios of over-fitting and under-fitting in kNN classifier.
    - Do you think the kNN classifier is a good classifier based on the results obtained on various metrics?
    - Do you think the model has regular fit situation? Use train and test set performances to arrive at this inference.
    - When do you think a situation of overfit happens for kNN classifier?