Programming for Data Analytic
SOFT8032
Second Examination

November 2022

1 Third Assessment. Second Project

This project contributes 50% in your final mark. This is an individual project and has to be all done by yourself solely. You may be called for a zoom meeting to explain different parts of your submission, if needed.

Please submit your project via canvas as *ONE PYTHON FILE ONLY.*

Any question regarding the project should be communicated with farshad.toosi@mtu.ie or Canvas message.

1.1 Dataset Overview

For this project we are going to perform a number analytic tasks on the **humanDetails.csv** file. This dataset contains details about individuals such as their age, salary range, relationship, country and etc.

1.2 Project Specification

The objective of this project is to provide an insight into the underlying pattern of the dataset.

Please perform the following tasks:

1. **Task1:** For each known county use Work-class and Age to predict Income with the following setting:
   (a) Any record associated with an unknown cell in work-class should be removed from the dataset.
   (b) Some values in Age columns are represented as decade, e.g., 20s or 30s. You are required to clean these values by removing the s and convert it into an integer value.
   (c) Income is either less or equal than 50k or greater than 50k.
   (d) Cross validation using 5 folds and 20% test size.

For each country run decision tree classifier and try different values for the depth of the tree. Pick the depth that produces highest accuracy for test set. And finally pick those countries that their process of learning still suggest overfitting and visualize them using appropriate visualization technique to display the gap between training and test. In this task overfitting occurs if the gap between training and test is more than 20%.

Interpret the results.

2. **Task2:** Use hours-per-week, Occupation, Age and relationship to predict income. Apply *Decision Tree* and *K Nearest Neighbour* classifiers to compare their accuracy (test and training).

   (a) Any unknown cell in the Occupation column should be filled with the Occupation with the highest frequency.

   (b) Some values in Age columns are represented as decade, e.g., 20s or 30s. You are required to clean these values by removing the s and convert it into an integer value.

   (c) One of the values of relationship attribute is *Other-relative* that needs to be removed from the dataset.

   (d) Some values in hours-per-week attribute mentioned only once. Those values should be removed from the dataset.

   (e) Income attribute should have two unique values.

   (f) Use cross-validation with 5 folds.

   (g) Try different values for the classifiers' parameters.

   (h) Visualize the results and interpret any pattern you find.

   Interpret the results.

3. **Task3:** Use *age*, *fnlwgt*, *education-num* and *hours-per-week* to train a model with K-Means. Apply the following settings:

   (a) Some values in Age columns are represented as decade, e.g., 20s or 30s. You are required to clean these values by removing the s and convert it into an integer value.

   (b) Cluster the data into two clusters.

   Reduce the number of features to two features. Use the first new feature as X coordinate and the second new feature as Y coordinate and visualize them twice using appropriate visualisation technique as follows:

   (a) The color of each individual in the first visualization is decided based on the value in *Income* attribute.

   (b) The color of each individual in the second visualization is decided based on the label that is generated by the clustering algorithm.

   (c) Visualize the two plots next to each other.

   Interpret the results.

Note that visualization plots need to have proper labels and annotations.

1.3 Submission and Deadline

Please use the python file template that is provided for you and complete your project in that file. Each task needs to be implemented as a separated function with interpretation as a comment below the function.

Please write your name and student ID as a comment in the designated area in the provided template python file.

The template file should be re-named at the end using your student ID followed by letter s, for example if your student ID is: 1234567 the the python file should be named: s1234567.py

The deadline for this project is 16th Dec 2022. One-week late submission with 10 marks penalty would be accepted and the deadline would be 23rd Dec 2022. Two weeks late submission with 20 marks penalty would be accepted and the deadline would be 30th Dec 2022.

Any question about this project should be communicated with Farshad Ghassemi Toosi farshad.toosi@mtu.ie or via Canvas.

Please submit your project via Canvas.

1.4 Rubric

This rubric is subject to change.

1. Correct task implementation (model training, accuracy reporting, visualization if needed and etc). (100%)
2. Relatively correct task implementation (model training, accuracy reporting, visualization if needed and etc). (70%)
3. Partly correct task implementation (model training, accuracy reporting, visualization if needed etc). (40%)
4. Wrong task implementation. (0%)