Programming for Data Analytic
SOFT8032

Dec 2020

1 Third Assessment. Second Project

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

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

1.1 Dataset Overview

The required dataset (*People's dataset*) for this project contains individuals' information such as yearly income, education, age, gender and etc. For this project we will perform a number of analytic tasks on the *People's dataset* which is shared on Canvas with you. This dataset contains the details of a large number people and each person has the following features in the dataset.

1. age
2. workType
3. fnlwgt
4. education
5. marital-status
6. job
7. relationship
8. gender
9. capital-gain
10. capital-loss
11. hours-per-week
12. native-country

13. Income

Note, the dataset might contain noises such as extra space before or after values or columns' header.

1.2 Project Specification

The objective of this project is to provide analytical details on the relations between features in the dataset using Machine Learning techniques with the aid of visualization approaches.

Please perform the following tasks:

1. Apply a decision tree classifier on the dataset containing people younger than 50. This dataset has four attributes: (*education*, *workType*, *job* and *Income*) where *Income* is the class attribute.

   First use 33% of the data individuals as test set and the remaining as training set. Report the accuracy for both training and test sets. Now increase the test percentage to 95% and report the accuracy for training and test sets. Use comment to discuss your observations.

   Before training the models, you need to perform the following pre-processing steps:

   (a) Convert the categorical values in *Income* attribute to numerical values. Note, there must be only two different/unique values at the end in this column/attribute; one for the income greater than 50k and the other one for the income less than or equal to 50k.

   (b) Convert the categorical values in *education* attribute to numerical values.

   (c) Convert the categorical values in *workType* attribute to numerical values.

   (d) Convert the categorical values in *job* attribute to numerical values.

   (e) All the empty cells from *Income* attributes are filled in with the value with the highest frequency; see Figure 1 for more clarification.

   (f) All the empty cells from *workType* attributes are filled in with the value with the highest frequency.

   (g) All the empty cells from *education* attributes are filled in with the value with the highest frequency.

2. Create a new dataset with two attributes (*workType* and *Income* where *Income* is the class attribute). Calculate the Entropy for attribute *workType* for two different values as follows:

   - Entropy(workType, private)
   - Entropy(workType, State-gov)

   The following pre-processing steps are required to be done before calculating the entropies:

   (a) Convert the categorical values in *Income* attribute to numerical values. Note there must be only two different/unique values at the end in this column/attribute; one for the income greater than 50k and the other one for the income less than or equal to 50k.

   (b) All the empty cells from *Income* attributes are filled in with the value with the highest frequency; see Figure 1 for more clarification.

(c) All the empty cells from *workType* attributes are filled in with the value with the highest frequency.

Use comment and explain which one has a higher entropy and what does it indicate.

3. Create a new dataset for females older than 30 using *education*, *workType*, *age* and *job* where *workType* is the class attribute.

   Apply three different classification models using cross validation where cv=5 on this dataset as follows:

   - Decision Tree classifier
   - Naive bayes (GaussianNB)
   - Random Forest classifier

   For each classifier, calculate the average accuracy of test and training sets over all iterations and visualize the results using bar chart (e.g., Figure 2). Note appropriate visualization features are needed to be embedded in the figure.

   The following set of pre-processing steps are required to be done before training the model.

   (a) Convert the categorical values in *education* attribute to numerical values.

   (b) Convert the categorical values in *workType* attribute to numerical values.

   (c) Convert the categorical values in *job* attribute to numerical values.

   (d) Make all none-numerical cells in *age* column/attribute empty.

   (e) All the empty cells from *age* attribute are filled in with the average of existing numerical values in the column/attribute.

   (f) Finally all the rows that have at least one empty cell are removed.

4. Use an unsupervised learning model on two different data-sets as follows:

   (a) *Income*, *education* and *job*

   (b) *education* and *job*

   Try different numbers of clusters for each dataset (1, 2, 3, 4, 5, 6, 7) and use an appropriate visualization technique to discuss what is potentially the best number of clusters for each dataset. Use comments to discuses the differences and your reasoning. Use appropriate visualization features. Apply the following pre-processing steps before running the model:

   (a) All the rows that have at least one empty cell are removed.

   (b) Convert the categorical values in *education* attribute to numerical values.

   (c) Convert the categorical values in *job* attribute to numerical values.

   (d) Convert the categorical values in *Income* attribute to numerical values. Note there must be only two different/unique values in this column/attribute at the end; one for the income greater than 50k and the other one for the income less than or equal to 50k.

   Use appropriate feature scaling for attributes.

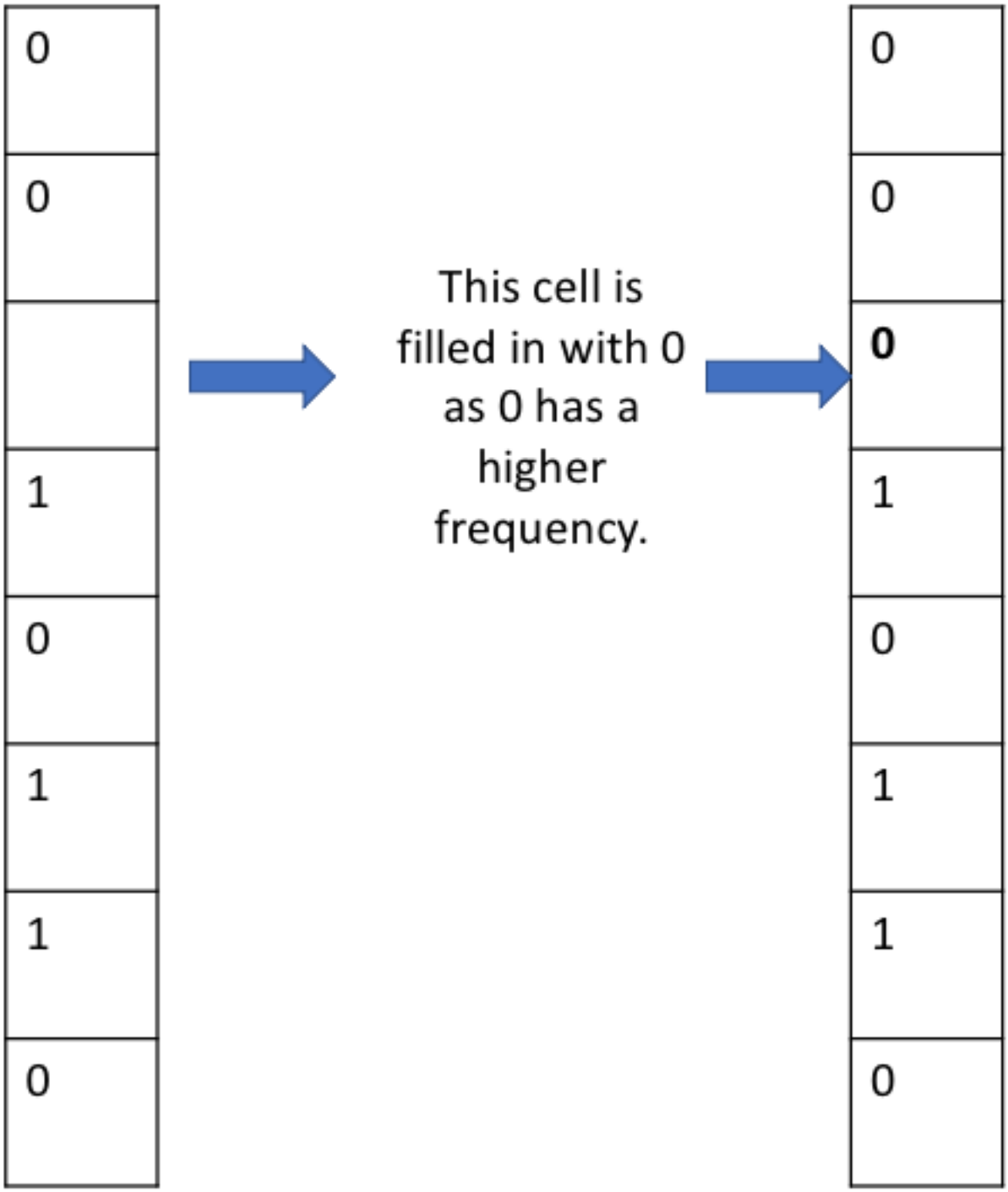

Figure 1: Question 1 hint.

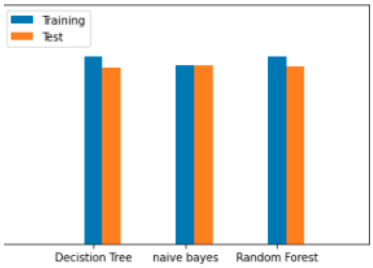

Figure 2: Models' accuracy.

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 a few lines interpretation as comment below the function.

Please write your name, student ID and your course name as comment in the designated area in the provided 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 Do NOT zip the python file for submission. The deadline for this project is 23rd of December at 23:59. Late submission is accepted with 10 marks penalty and the deadline for late submission is 30th December at 23:59.

Any question about this project should be communicated with Farshad Ghassemi Toosi farshad.toosi@cit.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, interpretation, visualization if needed etc). (100%)

2. Relatively correct task implementation (model training, accuracy reporting, interpretation, visualization if needed etc). (70%)

3. Partly correct task implementation (model training, accuracy reporting, interpretation, visualization if needed etc). (40%)

4. Fully wrong task implementation. (0%)