MSBA Python Bootcamp
July 2022
Practical Exercise 8 – Classification and Clustering
(Additional Practice Questions)

PE08-01 – Classification – Census Income (Adult) Dataset

This question is based on the Census Income (Adult) dataset taken from the UCI Machine Learning Repository – http://archive.ics.uci.edu/ml/datasets/Adult.

The original source of the dataset is attributed to:
Kohavi, R. and Becker, B., Data Mining and Visualization, Silicon Graphics.

The dataset was extracted from the 1994 United States Census Bureau database (originally found at http://www.census.gov/ftp/pub/DES/www/welcome.html). The filtering criteria were "((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))". The prediction task is to determine **whether a person makes over $50000 (i.e., 50K) a year**.

The dataset downloaded from UCI Machine Learning Repository consists of more than 30000 observations. There is a target variable together with 14 other variables. The actual dataset given to you has the variable fnlwgt (final weight) removed. Description of the target variable and the remaining 13 predictive variables are listed below:

1. Income – <=50K (coded as lte50k) or >50K (coded as gt50k)
2. age – continuous.
3. workclass – Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
4. education – Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
5. education-num: Number of years of education – continuous.
6. marital-status – Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7. occupation – Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
8. relationship: Relationship of individuals to the head of household – Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9. race – White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10. sex – Female, Male.
11. capital-gain – continuous.
12. capital-loss – continuous.
13. hours-per-week: Hours worked per week – continuous.
14. native-country – United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

a) Perform a decision tree analysis and report your results and findings.

b) Discuss any data preparation and exploratory data analysis tasks that you have performed.

PE8-02 – Clustering – Wholesale Customers Dataset

This exercise is based on the Wholesale Customers dataset taken from the UCI Machine Learning Repository – http://archive.ics.uci.edu/ml/datasets/Wholesale+customers.

The original source of the dataset is attributed to:
Cardoso, Margarida G. M. S., ISCTE-IUL, Lisbon, Portugal

The dataset consists of the annual spending (in monetary units) of customers across six product categories – fresh products, milk products, grocery products, frozen products, detergents and paper products, and delicatessen products. Customers are classified by regions and channels.

The dataset consists of 440 observations with no missing data. There are 8 variables altogether:

1. Channel – Customer's channel; 1: Horeca (Hotel/Restaurant/Cafe), 2: Retail channel
2. Region – Customer's region; 1: Lisbon, 2: Oporto, 3: Other
3. Fresh – Annual spending (m.u.) on fresh products
4. Milk – Annual spending (m.u.) on milk products
5. Grocery – Annual spending (m.u.) on grocery products
6. Frozen – Annual spending (m.u.) on frozen products
7. DetergentsPaper – Annual spending (m.u.) on detergents and paper products
8. Delicatessen – Annual spending (m.u.) on and delicatessen products

a) Perform clustering analysis using the 6 annual spending variables to place customers into homogeneous groups. Report your results and findings.

b) Once similar customers have been grouped together, it is possible to identify market niche as well as to perform other data mining tasks within each group of similar customers. Suggests some possible use cases and briefly explain the data mining tasks that are involved.