WEEK ONE. REPORT ON ADULT .CSV FILE.

By

SAMUEL ATEKHA OSAGHAE

STUDENT ID 2120094

UNIVERSITY OF WOLVERHAMPTON

SCHOOL OF MATHEMATICS AND COMPUTER SCIENCE.

MARCH 2023

WEEK ONE REPORT ON ADULT .CSV FILE

In the Adult. Csv file dataset we have originally (32561 rows and 15 columns which means the datasets contain data collected from 32,561 adults on their age, occupation, native country their race, marital status. I got this details through this code "data.shape" .

I decided to select 30,000 samples from the 32561 whole datasets set and analyse the data the sample with a random state of 94 through this code /"data = data.sample(n=30000, random_state = 48)"/

The summary of the data sets hence is out of the 30, 000 sample, there are 20097 Male while the Female are 9903. The average age of the sample is 38.588. meanwhile the oldest person is 90years.and while the youngest is 17 years. They are of different working class from private to self employed to state government employed and the average hour worked by week is 40.433 while the lowest hour hours worked by one adult is 1 and maximum hours worked per week is 99hours per week.

There are six types of data types in Data mining group into two categories. They are Nominal attributes, Ordinal attributes, Binary attributes grouped as **Qualitative attributes type** while the .

Numeric attributes, Discrete attributes, Continuous attributes are grouped into **Quantitative attributes types.**

From the adult datasets we have Numeric attributes types which are numbers which can be counted and the columns with these are attributes are "age " columns which is in numbers and has continuous attributes. The other Numeric attributes columns in the datasets are the "capital gain ", " capital loss" "capital-gain", "hours-per-week" ," education-num" columns.

Nominal Attribute are symbols and name of things. The nominal attributes in the adult datasets are in the "sex", "education", "marital-status", "race", "relationship", "occupation", "native-country" columns

The Ordinal attributes provides sufficient information to order the objects example of this attributes in thee datasets is the "education" column.

A binary attribute is a nominal attribute with only two elements or states such as 0 or 1, such column with such data in the datasets is the “class-label” column where it is each adult either earn above 50k or less than 50k and below . when the “sex ”columns is transforms it can also possess this attribute where 1 is for male and where 0 is for female.

Also from the Adult.csv datasets the country with the highest migrants are from the United states of America with the sum of 26857 migrants we can get this figure with this code “data['native-country'].value_counts()”.

The occupations that represents more males than females Craft-repair, Exec-managerial Prof-specialty, Sales, Other-service, Transport-moving, Machine-op-inspct, Adm-clerical Handlers-cleaners,?, Farming-fishing, Tech-support, Protective-serv . I got this through this code “data['occupation'].groupby([data['sex']]).value_counts()”

The difference between **data.head()** and **data.tail()** is that data.head() shows the top of the data that is from the very value of the datasets at the beginning of the whole datasets , while data.tail() shows the bottom of the datasets include the very last data value; that’s the last few numbers of dataset to the very last one

REFERENCE

Ronny, K. and Barry, B. (1996) Adult Data set. UCI . Available at https://archive.ics.uci.edu/ml/datasets/Adult (Accessed: 18 January 2023).