|  |  |  |
| --- | --- | --- |
| **Lab no.** | **Category** | **Practical Definition** |
| 1 | [A]  [B]  [C] | **Introduction to panda library.**   * Read\_csv, Series, DataFrame, Conditional Selection, Setting and Resetting index, MultiIndexing. * Groupby, drop(), describe(), adding new column, max(), Loc(), info(), shape(), head(), tail() * plot(), value\_counts(), join(), merge(), xs(). |
| 2 | [A]  [B]  [C] | **Perform the following exercise using the function of panda library.**   * Import the necessary libraries. * Import the dataset from this address.   <https://raw.githubusercontent.com/justmarkham/DAT8/master/data/u.user>   * Assign it to a variable called users and use the 'user\_id' as index. * Display the first 25 entries. * Display the last 10 entries. * What is the number of observations in the dataset? * What is the number of columns in the dataset? * Print the name of all the columns. * How is the dataset indexed? * What is the data type of each column? * Print only the occupation column. * How many different occupations are in this dataset? * What is the most frequent occupation? * Summarize the Data Frame. * Summarize all the columns. * Summarize only the occupation column * What is the mean age of users? * What is the age with least occurrence? * Perform the all above mentioned exercise for the following data set. * Occupation * Chipotle * World Food Facts   Link: https://github.com/guipsamora/pandas\_exercises/tree/master/  01\_Getting\_%26\_Knowing\_Your\_Data |
| 3 | [A]  [B]  [C] | **Data Exploration**   * First, you need to read the titanic dataset from local disk and display first five records. * Identify Nominal, Ordinal, Binary and Numeric attributes from data sets and display all values. * Identify symmetric and asymmetric binary attributes from data sets and display all values. * For each quantitative attribute, calculate its average, standard deviation, minimum, mode, range and maximum values. * For the qualitative attribute (class), count the frequency for each of its distinct values. * It is also possible to display the summary for all the attributes simultaneously in a table using the describe() function. If an attribute is quantitative, it will display its mean, standard deviation and various quantiles (including minimum, median, and maximum) values. If an attribute is qualitative, it will display its number of unique values and the top (most frequent) values. * For multivariate statistics, you can compute the covariance and correlation between pairs of attributes. * Display the histogram for Age attribute by discretizing it into 8 separate bins and counting the frequency for each bin. * A boxplot can also be used to show the distribution of values for each attribute. * Display scatter plot for any 5 pair of attributes, we can use a scatter plot to visualize their joint distribution. * Perform the all above mentioned exercise for the following data set. * Occupation * Chipotle * World Food Facts   Link: https://github.com/guipsamora/pandas\_exercises/tree/master/  01\_Getting\_%26\_Knowing\_Your\_Data |
| 4 | [A]  [B]  [C] | **Data Preprocessing**   * Import necessary library and read dataset. * Calculate total number of instances and attributes in dataset. * Calculate the total number of missing data in each attribute. * Replace the missing value by the mean and median of that attribute. * Replace the missing value by category wise mean value ( e.g. missing value of female candidate should be replaced by the mean value of female.) * Calculate the number of duplicate rows and remove that duplicate row. * Perform data aggregation, sampling, and discretization on the dataset * Perform the all above mentioned exercise for the following data set. * Occupation * Chipotle * World Food Facts   Link: https://github.com/guipsamora/pandas\_exercises/tree/master/  01\_Getting\_%26\_Knowing\_Your\_Data |
| 5 | [A]  [B]  [C] | Apply apriori algorithm on given dataset, calculate support and confidence than find the association.  Implement the same problem with its solution using python or java programming.   |  |  | | --- | --- | | **TID** | **Items** | | 100 | 1 3 4 | | 200 | 2 3 5 | | 300 | 1 2 3 5 | | 400 | 2 5 |   Apply apriori algorithm on the dataset given in the link below, calculate support and confidence than find the association.  Link: https://www.kaggle.com/datasets/aslanahmedov/market-basket-analysis |
| 6 | [A]  [B]  [C] | Generate the Frequent Pattern Tree for the given dataset.  Implement the same problem with its solution using python or java programming.   |  |  | | --- | --- | | **TID** | **Items** | | 100 | 1 3 4 | | 200 | 2 3 5 | | 300 | 1 2 3 5 | | 400 | 2 5 |   Apply FP Tree algorithm on the following dataset and generate FP Tree.  Link: https://www.kaggle.com/datasets/priteshraj10/covid-vaccination-all-countries-data |
| 7 | [A]  [B]  [C] | Apply Naïve Bayes Classifiers for the given dataset.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Day** | **Outlook** | **Temperature** | **Humidity** | **Wind** | **PlayTennis** | | 1 | Sunny | Hot | High | Weak | No | | 2 | Sunny | Hot | High | Strong | No | | 3 | Overcast | Hot | High | Weak | Yes | | 4 | Rain | Mild | High | Weak | Yes | | 5 | Rain | Cool | Normal | Weak | Yes | | 6 | Rain | Cool | Normal | Strong | No | | 7 | Overcast | Cool | Normal | Strong | Yes | | 8 | Sunny | Mild | High | Weak | No | | 9 | Sunny | Cool | Normal | Weak | Yes | | 10 | Rain | Mild | Normal | Weak | Yes | | 11 | Sunny | Mild | Normal | Strong | Yes | | 12 | Overcast | Mild | High | Strong | Yes | | 13 | Overcast | Hot | Normal | Weak | Yes | | 14 | Rain | Mild | High | Strong | No |   **x**’=(Outlook=*Sunny,* Temperature=*Cool,* Humidity*=High,* Wind=*Strong*)?  is it Yes or No? find it using naïve bayes classifiers.  Implement the same problem with its solution using python or java programming.  Apply Naïve Bayes Classifiers for the given dataset and do the prediction.  Link: https://www.kaggle.com/datasets/priteshraj10/covid-vaccination-all-countries-data |
| 8 | [A]  [B]  [C] | Apply K-nearest neighbour classifiers for the given dataset.   |  |  |  |  | | --- | --- | --- | --- | | **Name** | **Age** | **Healthy** | **Game** | | Ajay | 10 | Yes | Cricket | | Sanjay | 20 | Yes | Cricket | | Vijay | 30 | Yes | Cricket | | Harjay | 40 | No | Chess | | Samay | 50 | No | Chess |   Now for the new data, person with Name: Dhananjay, Age:60 and Healthy: Yes. What game will be played by the person Dhananjay? find it by applying KNN classifiers.  Implement the same problem with its solution using python or java programming.  Apply kNN algorithm on the dataset of the Occupation as given in the link below.  https://github.com/guipsamora/pandas\_exercises/tree/master/  01\_Getting\_%26\_Knowing\_Your\_Data |
| 9 | [A]  [B]  [C] | Apply the decision tree algorithm for the given dataset and generate the decision tree,   |  |  |  |  | | --- | --- | --- | --- | | **Age** | **Competition** | **Type** | **Profit** | | Old | Yes | Software | Down | | Old | No | Software | Down | | Old | No | Hardware | Down | | Mid | Yes | Software | Down | | Mid | Yes | Hardware | Down | | Mid | No | Hardware | Up | | Mid | No | Software | Up | | New | Yes | Software | Up | | New | No | Hardware | Up | | New | No | Software | Up |   Implement the same problem with its solution using python or java programming.  Apply Decision Tree algorithm on given dataset and do the classification according to disease.  Link: https://www.kaggle.com/datasets/taranvee/covid-19-dataset-till-2222022 |
| 10 | [A]  [B]  [C] | Apply the k-mean clustering on the following data and form the cluster for the value of k=2.   |  |  | | --- | --- | | **Height** | **Weight** | | 100 | 90 | | 80 | 70 | | 60 | 50 | | 40 | 30 | | 20 | 10 |   Implement the same problem with its solution using python or java programming.  Apply k-mean clustering on the given data set and generate cluster based on disease.  Link: https://www.kaggle.com/datasets/taranvee/covid-19-dataset-till-2222022 |
| 11 | [A]  [B]  [C] | Apply the agglomerative clustering on the following data and form the cluster.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | P1 | P2 | P3 | P4 | P5 | | P1 | 0 |  |  |  |  | | P2 | 9 | 0 |  |  |  | | P3 | 3 | 7 | 0 |  |  | | P4 | 6 | 5 | 9 | 0 |  | | P5 | 11 | 10 | 2 | 8 | 0 |   Implement the same problem with its solution using python or java programming.  Apply the agglomerative clustering on the following data and form the cluster.  https://www.kaggle.com/datasets/rajasankar/hierarchical-clustering-of-7-million-proteins |
| 12 | [A]  [B]  [C] | **Apply outlier detection techniques for the spam email filtering using any classifier.**  Training dataset is as per below.  spam: “send us your password”  ham: “send us your review”  ham: “password review”  spam: “review us ”  spam: “send your password”  spam: “send us your account”  spam: “send us your password”  ham: “send us your review”  ham: “password review”  spam: “review us ”  spam: “send your password”  spam: “send us your account”  Implement same problem using K-NN classifier.  **Apply outlier detection techniques for the spam email filtering for the given dataset.**  https://www.kaggle.com/datasets/nitishabharathi/email-spam-dataset |