|  |  |  |  |
| --- | --- | --- | --- |
| A picture containing drawing, stop, room  Description automatically generated | Business Intelligence  Practical #8 | | |
|  |  |  |  |
| **Name** | Farman Kazi | **Roll Number** | 21302B0049 |
| **Class** | TY B. Sc. IT | **Division** | A |
| **Subject/Course:** | Business intelligence | | |
| **Topic** | Clustering | | |
|  |  |  |  |
| **Q1. Write Overview of Clustering.** | | | |
| * Clustering is basically a type of **unsupervised** learning method. * An unsupervised learning method is a method in which we draw **references** from datasets consisting of input data without labelled responses. * Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples. * Clustering is the task of dividing the population or data points into several groups such that data points in the same groups are more like other data points in the same group and dissimilar to the data points in other groups. * It is basically a collection of objects based on similarity and dissimilarity between them. * Clustering is very much important as it determines the intrinsic grouping among the unlabelled data present. * There are no criteria for good clustering. It depends on the user, and what criteria they may use which satisfy their need. * For instance, we could be interested in finding representatives for homogeneous groups (data reduction), finding “natural clusters” and describing their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection). * This algorithm must make some assumptions that constitute the similarity of points and each assumption make different and equally valid clusters. | | | |
| **Q2.** **Perform Clustering using K-means method.** | | | |
| 1. Download **clustering.xlsx** 2. Open **Google Colab** 3. Upload **clustering.xlsx** file (cluster as k). Cluster can be different.   Algorithm of Cluster:  Euclidean distance formula  **sqrt((x2-x1)^2 +(y2-y1)^2)**  **Code**  import pandas as pd  import numpy as np  import random as rd  import matplotlib.pyplot as plt  data = pd.read\_csv('/content/clustering.csv')  data.head()  **Output**    **Code**  X = data[["LoanAmount", "ApplicantIncome"]]    plt.scatter(X["ApplicantIncome"], X["LoanAmount"], c = 'black')  plt.xlabel('AnnualIncome')  plt.ylabel('Loan Amount (In Thousands)')  plt.show()  **Output**    **Code**  K = 3  Centroids = (X.sample(n = K))  plt.scatter(X["ApplicantIncome"], X["LoanAmount"], c = 'black')  plt.scatter(Centroids["ApplicantIncome"], Centroids["LoanAmount"], c = 'red')  plt.xlabel('AnnualIncome')  plt.ylabel('Loan Amount (In Thousands)')  plt.show()  **Output**    **Code**  diff = 1  j = 0  while diff != 0:  XD = X  i = 1  for index1, row\_c in Centroids.iterrows():  ED = []  for index2, row\_d in XD.iterrows():  d1 = (row\_c["ApplicantIncome"] - row\_d["ApplicantIncome"]) \*\* 2  d2 = (row\_c["LoanAmount"] - row\_d["LoanAmount"]) \*\* 2  d = np.sqrt(d1 + d2)  ED.append(d)  X[i] = ED  i = i + 1  C = []  for index, row in X.iterrows():  min\_dist = row[1]  pos = 1  for i in range(K):  if row[i + 1] < min\_dist:  min\_dist = row[i + 1]  pos = i + 1  C.append(pos)  X["Cluster"] = C  Centroids\_new = X.groupby(["Cluster"]).mean()[["LoanAmount", "ApplicantIncome"]]  if j == 0:  diff = 1  j = j + 1  else:  diff = (Centroids\_new['LoanAmount'] - Centroids['LoanAmount']).sum() + (Centroids\_new['ApplicantIncome'] - Centroids['ApplicantIncome']).sum()  print(diff.sum())  Centroids = X.groupby(["Cluster"]).mean()[["LoanAmount", "ApplicantIncome"]]  **Output**    **Code**  color = ['blue', 'green', 'cyan']  for k in range(K):  data = X[X["Cluster"] == k + 1]  plt.scatter(data["ApplicantIncome"], data["LoanAmount"], c = color[k])  plt.scatter(Centroids["ApplicantIncome"], Centroids["LoanAmount"], c = 'red')  plt.xlabel('Income')  plt.ylabel('Loan Amount (In Thousands)')  plt.show()  **Output** | | | |