|  |  |
| --- | --- |
| Semester | T.E. Semester VI – Computer Engineering |
| Subject | Data Warehousing and Mining |
| Subject Professor In-charge | Prof. Kavita Shirsat |
| Assisting Teachers | Prof. Kavita Shirsat |
| Laboratory | Lab 312 A |

|  |  |  |
| --- | --- | --- |
| Student Name | Deep Salunkhe 21102A0014  Omkar Patil 21102A0003  Pranav Redij 21102A0005  Sahil Pokharkar 21102A0009 | |
| Grade and Subject  Teacher’s Signature |  |  |

|  |  |  |
| --- | --- | --- |
| Assignment Number | 05 | |
| Assignment Title | Clustering | |
| Resources / Apparatus Required | Hardware:  Computer system | Software:  Python |
| Description | **K-Means Clustering:**  K-means clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into K distinct, non-overlapping subgroups or clusters. The primary goal of K-means is to group similar data points together while keeping them as distinct as possible from data points in other clusters. Here are some key concepts and steps:   1. **Initialization:** K-means starts with the initialization of K cluster centroids. These centroids are typically chosen randomly from the data points in the dataset. The value of K represents the number of clusters you want to form. 2. **Assignment Step:** In this step, each data point is assigned to the nearest centroid based on some distance metric, usually Euclidean distance. This assignment forms initial clusters. 3. **Update Step:** After assigning data points to clusters, the centroids of the clusters are recalculated as the mean of all data points within each cluster. 4. **Iterations:** Steps 2 and 3 are repeated iteratively until convergence, which occurs when the centroids no longer change significantly or a maximum number of iterations is reached. 5. **Result:** The final result is K clusters with their respective centroids.   **Visualization of Clusters:**  Visualizing the results of a K-means clustering analysis is crucial for gaining insights and presenting findings effectively. Scatter plots are a common technique for this purpose:   * **Scatter Plot:** A scatter plot is a two-dimensional data visualization that represents individual data points as dots on a Cartesian plane. Each dot's position is determined by the values of two features or dimensions.   **How to Present the Clusters using Scatter Plot:**   1. **Color-Coding:** Assign a unique color or marker style to each cluster. This makes it easy to distinguish data points belonging to different clusters. 2. **Data Points:** Plot the data points on the scatter plot, with their positions determined by the two dimensions of interest. 3. **Centroids:** Overlay the centroids of each cluster on the scatter plot. You can represent centroids as larger points or a different marker style. 4. **Title and Labels:** Include a title for the plot, axis labels, and a legend that explains the color-code or marker style for each cluster.   **Interpreting the Clusters:**  After visualizing the clusters, you can draw insights from the results. For example, you might observe:   * **Cluster Separation:** Analyze how well-separated the clusters are. Are data points within a cluster close to each other and distant from data points in other clusters? * **Cluster Size:** Examine the size of each cluster. Do some clusters have significantly more data points than others? * **Centroid Locations:** Investigate the location of cluster centroids. What does the location of a centroid represent in the context of your data? For example, in customer segmentation, it might indicate a typical customer profile. * **Patterns:** Look for patterns or trends within clusters. Are there common characteristics or behaviors shared by data points within a cluster? * **Outliers:** Identify any outliers or data points that don't seem to belong to any cluster. They might warrant further investigation. | |
| Program |  | |
| Output |  | |
| Conclusion: | By combining the theory of K-means clustering with effective visualization techniques, you can present your analysis results clearly and derive meaningful insights from your data, whether it's related to crime-prone areas or customer segmentation. | |