CT – 3 Mini Project ML – USC PA2312052010010

Subject Name: UNSUPERVISED MODEL- MACHINE LEARNING.

Subject Code: 20PAIE51J

Introduction to the Dataset

- The dataset comprises information on chemical compounds and their associated properties, spanning 780 rows.
- Each row represents a unique chemical compound, characterized by various molecular descriptors and bioconcentration factors.
- The dataset is structured with 13 columns, detailing different aspects of each chemical compound.

Key Features of the Dataset

Chemical Identifiers (ÿCAS, SMILES):Unique identifiers and SMILES notation representing the chemical structures.

Molecular Descriptors:

1. **piPC09:**Molecular multiple path count.
2. **PCD:** Difference between multiple path count and path count.
3. **X2Av:** Average valence connectivity.
4. **MLOGP:** Moriguchi octanol-water partition coefficient.
5. **ON1V:** Overall modified Zagreb index by valence vertex degrees.
6. **N-072:** Frequency of RCO-N< / >N-X=X fragments.
7. **B02[C-N]:** Presence or absence of C-N atom pairs.
8. **F04[C-O]:** Frequency of C-O atom pairs.

Target Variable:

logBCF: Bioconcentration Factor in log units, serving as the response variable in QSAR modeling.

Purpose of the Dataset

- The dataset aims to facilitate research in chemoinformatics and quantitative structure-activity relationship (QSAR) studies.
- It provides valuable insights into the relationship between chemical structure and bioconcentration factors, aiding in the prediction of environmental behaviour and biological activity of chemical compounds.

Theoretical Inference:

1. Display Top Five Rows:

- Displaying the top five rows of the dataset provides an initial glimpse into its structure and content.
- This step allows us to quickly inspect the data's format, column names, and some sample values.

2. Dropping Unnecessary Columns:

- Removing columns such as ÿCAS, SMILES, Set, and Class eliminates redundant or irrelevant information from the dataset.
- These columns may not contribute to the analysis or may contain identifying information that is not required for modeling.

3. Dropping the Target Variable:

- Excluding the target variable, in this case, 'logBCF,' is essential when performing unsupervised learning tasks.
- Unsupervised learning aims to identify patterns and structures in the data without relying on labelled outcomes.
- By dropping the target variable, we ensure that the clustering or dimensionality reduction techniques focus solely on the features without bias from the target.

4. Dataset Description:

- Describing the dataset provides an overview of its contents, including the number of observations and features.
- It summarizes the data's characteristics, such as data types, missing values, and statistical summaries of numerical columns.

5. Exploratory Data Analysis (EDA):

- EDA involves a comprehensive examination of the dataset to understand its underlying patterns, relationships, and distributions.

- This process typically includes visualizations, summary statistics, and correlation analyses to uncover insights and potential anomalies.

6. Applying Scaling Technique and Transforming the Data:

- Scaling is a crucial preprocessing step that standardizes the range of features in the dataset.
- Techniques such as StandardScaler or MinMaxScaler are commonly used to scale features to a consistent range.
- Scaling ensures that features with larger magnitudes do not dominate the analysis, particularly in distance-based algorithms like K-means clustering.

7. Applying KMeans Clustering:

- KMeans clustering partitions the dataset into K clusters based on feature similarity.
- By iterating over different values of K and computing the Within-Cluster Sum of Squares (WSS), we can determine the optimal number of clusters.
- The Elbow Method visually identifies the inflection point in the WSS curve, indicating the optimal K value where additional clusters provide diminishing returns in terms of reducing WSS.
- KMeans clustering with a specific K value assigns each data point to the nearest centroid, forming distinct clusters based on feature similarity.
- Visualizing the clustering results allows for intuitive interpretation and assessment of cluster quality.

8. Applying Agglomerative Clustering:

- Agglomerative clustering is a hierarchical clustering technique that iteratively merges similar clusters.
- It does not require specifying the number of clusters beforehand, making it suitable for exploratory analysis.
- Evaluation metrics such as silhouette score or Davies–Bouldin index can help determine the optimal number of clusters.
- Visualizing the clustering results aids in understanding the hierarchical structure and cluster assignments.

9. Applying DBSCAN Technique:

- Density-Based Spatial Clustering of Applications with Noise (DBSCAN) identifies clusters based on regions of high density separated by areas of low density.

- Parameters such as epsilon (eps) and minimum samples (min_samples) control the cluster formation process.

- Visualizing the clustering results provides insights into the density-based clustering pattern and outlier detection.

10. Applying GMM Soft Clustering:

- Gaussian Mixture Model (GMM) clustering assumes that data points are generated from a mixture of several Gaussian distributions.
- Soft clustering assigns probabilities of data points belonging to each cluster, allowing for more flexible cluster assignments.
- Evaluation metrics such as Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC) can assess model fit and determine the optimal number of components.
- Visualizing the soft clustering results provides a probabilistic view of cluster assignments, enabling nuanced interpretation.

11. Applying FCM Soft Clustering:

- Fuzzy C-Means (FCM) clustering assigns fuzzy membership values to data points, indicating the degree of association with each cluster.
- It allows for overlapping clusters and accommodates uncertainty in data point assignments.
- Evaluation metrics such as fuzzy silhouette score or Dunn index can evaluate the quality of fuzzy clustering.
- Visualizing the fuzzy clustering results provides insights into cluster overlap and membership degrees.

12. Applying PCA and SVD for Dimensionality Reduction:

- Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) are dimensionality reduction techniques used to capture the most significant variance in the data.

- Explained variance and cumulative variance plots illustrate the amount of information retained by each principal component or singular vector.
- Techniques such as scree plots or cumulative explained variance can determine the optimal number of components or singular values to retain.
- Decomposing the dataset while retaining a certain percentage of information ensures that the reduced-dimensional data adequately represents the original dataset.

13. Cluster Analysis with PCA Data:

- Cluster analysis is performed on the decomposed PCA data to identify optimal clusters using both hard and soft clustering methods.
- Hard clustering assigns each data point to a single cluster based on proximity to cluster centroids, while soft clustering assigns membership probabilities to multiple clusters.
- Evaluation metrics such as silhouette score or Davies–Bouldin index can assess the quality of cluster assignments.
- Visualizing the clustering results provides insights into cluster separability and compactness.

14. Cluster Analysis with SVD Data:

- Similar to PCA data, cluster analysis is conducted on the decomposed SVD data to identify optimal clusters using both hard and soft clustering methods.
- Evaluation metrics such as silhouette score or Davies–Bouldin index can evaluate the quality of cluster assignments.
- Visualizing the clustering results allows for the interpretation of cluster patterns and the assessment of clustering effectiveness.

15. **Inference for Cluster Analysis Results:**

- ✓ The results of cluster analysis provide valuable insights into the underlying structure of the dataset.
- ✓ Interpretation of cluster assignments and cluster characteristics enables the identification of distinct data patterns and groups.
- ✓ These insights can inform decision-making processes, such as targeted marketing strategies, customer segmentation, or anomaly detection.
- ✓ Understanding cluster characteristics aids in identifying potential areas for further investigation or intervention.
- ✓ Overall, cluster analysis facilitates data-driven decision-making by uncovering hidden patterns and structures within the dataset.