***MILESTONE – 1***

**TOPIC NAME:** **Clustering Mixed Data for Heart Failure Patient Profiles in Clinical Records Using Machine Learning**

**GROUP FORMATION*-***

***1-*** Vansh Tyagi (E23CSEU0532)

***2-*** Vansh Khanna (E23CSEU0518)

3- Ronit Agarwal (E23CSEU0540)

**PROBLEM STATEMENT IDENTIFICATION-**

Heart disease is a leading cause of mortality worldwide, with heart failure representing a significant clinical challenge. Analyzing clinical records can offer valuable insights for early diagnosis, prognosis, and personalized treatment. However, these records often contain mixed data types, including numerical, categorical, and ordinal features, which complicates traditional clustering techniques. This project aims to develop an efficient clustering model for mixed data to group patients with similar clinical characteristics. By applying machine learning and artificial intelligence techniques, the project seeks to identify meaningful patient subgroups from a heart failure dataset. The resulting clusters can assist healthcare professionals in understanding patterns within patient profiles, potentially informing better treatment strategies and resource allocation.

**TITLE AND DESCRIPTION-**

**Clustering Mixed Data for Heart Failure Patient Profiles in Clinical Records Using Machine Learning:**

This project clusters heart failure patient data, which includes both numerical and categorical features, using machine learning techniques for mixed data. By grouping patients with similar clinical characteristics, the project aims to reveal patterns that can help healthcare providers understand risk factors and guide treatment decisions, supporting more effective care for heart failure patients.

***MILESTONE – 2***

**Literature Survey-**

We have selected a total of 9 research papers on clustering to conduct our research survey for our project. An overview of the 9 papers are as follows:

* **Research Paper on Cluster Techniques of Data Variations**

**Domain:** Data clustering for analysis and summarization.

**Dataset:** Not specified; focuses on general clustering techniques.

**Methodology:** Discusses cluster definitions and popular methods like well-separated, center-based, contiguous, density-based, and similarity-based clustering. Hierarchical and partitional approaches are explained with emphasis on K-means and hierarchical clustering.

**Performance Metrics:** Focus on the concept of cluster quality, addressing challenges like the "curse of dimensionality" and proximity metrics.

**Key Aspects:** Provides foundational insights into clustering applications, including biological and psychological data analysis.

* **An Analytical Study on Behavior of Clusters Using K-Means, EM, and K Means Algorithm\***

**Domain:** Information retrieval and clustering performance analysis.

**Dataset:** Heart Spect dataset with 267 instances related to cardiac images.

**Methodology:** Compares k-means, k\*-means, and Expectation Maximization (EM) clustering algorithms based on purity, entropy, CPU time, and mean value analysis.

**Performance Metrics:** Clustering quality measured by purity and entropy, cluster-wise analysis, and computational efficiency.

**Key Aspects:** EM clustering outperforms k-means and k\*-means, showing enhanced cluster quality and lower computational time.

* **Review Paper on Clustering Techniques**

**Domain:** Data mining and unsupervised learning for clustering applications.

**Dataset:** General discussion without specific datasets; focuses on algorithm categorization.

**Methodology:** Reviews partition-based, hierarchical, density-based, and grid-based clustering algorithms, discussing applications and challenges like high dimensionality.

**Performance Metrics:** Evaluates algorithms based on scalability, noise handling, and computational efficiency.

**Key Aspects:** Highlights strengths of density-based methods for arbitrary-shaped clusters and grid-based algorithms for faster processing.

* **Comparative Study of Various Clustering Algorithms in Data Mining**

**Domain:** Data mining with a focus on clustering algorithm evaluation.

**Dataset:** General overview, without specific datasets; comparison of clustering methods.

**Methodology:** Discusses partition-based (K-means, K-medoids), hierarchical (CURE, BIRCH, CHAMELEON), and density-based (DBSCAN, OPTICS, DENCLUE) clustering techniques. Highlights the strengths and limitations of each method, focusing on factors like data size, noise handling, and high-dimensional clustering.

**Performance Metrics:** Examines algorithm efficiency, scalability, and sensitivity to noise, with a detailed analysis of convex and arbitrary-shaped clusters.

**Key Aspects:** Provides a comprehensive comparison, showing that no single algorithm suits all clustering scenarios, thus helping researchers choose algorithms based on specific data characteristics and clustering needs.

* **Clustering Techniques for Research Papers: A Detailed Review**

**Domain:** Text mining and clustering applied to research paper organization.

**Dataset:** Focused on clustering methods for research papers; discusses semantic and statistical text clustering.

**Methodology:** Explores three approaches: statistical (using centroid and relationship-based features), graph-based (using citation graphs), and semantic-based (leveraging WordNet and lexical chains). Evaluates methods for representing document similarity.

**Performance Metrics:** Evaluated with metrics like F-measure, entropy, and purity, showing improvements in clustering performance with methods that combine text and citation data.

**Key Aspects:** Emphasizes that incorporating citation relationships and semantic concepts can significantly enhance clustering quality for text data, especially in research contexts with interconnected documents.

* **A Comprehensive Survey of Clustering Algorithms**

**Domain:** Clustering methods for data science applications.

**Dataset:** General survey without specific datasets; covers multiple domains like biology and computer science.

**Methodology:** Examines traditional (e.g., K-means, hierarchical) and modern (e.g., kernel-based, ensemble) clustering methods. Discusses similarity measures, distance functions, and evaluation criteria.

**Performance Metrics:** Compares time complexity, clustering accuracy, and scalability for each method.

**Key Aspects:** Provides a detailed comparative analysis of clustering algorithms and their applicability across diverse data types and scales.

* **Exploring Clustering Techniques in Machine Learning**

**Domain:** Machine learning clustering applications for various sectors, such as image segmentation and anomaly detection.

**Dataset:** No specific datasets; general exploration of clustering approaches.

**Methodology:** Discusses partitioning-based, density-based, hierarchical, and model-based clustering techniques. Focuses on evaluation metrics like cohesion and separation to assess quality.

**Performance Metrics:** Examines algorithm performance based on scalability, interpretability, and cluster cohesion.

**Key Aspects:** Highlights applications in real-world scenarios, as well as the challenges associated with initialization sensitivity and computational demands.

* **Research on K-Value Selection Method of K-Means Clustering Algorithm**

**Domain:** Optimization of the K-means algorithm for various clustering applications.

**Dataset:** Iris dataset for experimentation.

**Methodology:** Evaluates K-value selection methods including Elbow, Gap Statistic, Silhouette Coefficient, and Canopy for improving K-means clustering results.

**Performance Metrics:** Assesses methods by accuracy of cluster formation, convergence rate, and cluster cohesion.

**Key Aspects:** Offers insights on optimizing K-means clustering by choosing the best K-value, which is critical for accurate clustering outcomes.

* **An Analysis on Clustering Algorithms in Data Mining**

**Domain:** Data mining and machine learning clustering applications.

**Dataset:** No specific datasets; focuses on clustering methodologies.

**Methodology:** Reviews clustering types, including hierarchical and partitional algorithms. Analyzes key techniques like linkage methods and nearest neighbor clustering.

Performance Metrics: Discusses time complexity, cluster purity, and accuracy.

**Key Aspects:** Emphasizes clustering’s role in exploratory data analysis, with insights on algorithm efficiency and effectiveness for large datasets.

**Data Pre-processing -**

**Data Loading and Column Selection:**

The dataset was loaded into a DataFrame named df. Initially, all columns were included, but a subset of columns relevant for clustering was selected, specifically: age, sex, serum\_sodium, platelets, and smoking.

**Defining and Scaling Features:**

The selected columns were divided into numerical (age, serum\_sodium, platelets) and categorical (sex, smoking) features.

**Scaling Numerical Features:**

The StandardScaler from sklearn.preprocessing was applied to standardize the numerical features, ensuring they have a mean of 0 and a standard deviation of 1. The scaled data was stored in a new DataFrame called df\_scale.

**Data Visualization / Exploratory Data Analysis (EDA)-**

EDA allows us to examine the underlying patterns, distributions, and relationships within the dataset. Though specific visualizations were not included in the notebook, here are recommended approaches and how they would aid the clustering analysis of heart failure records:

**Distribution Analysis:**

Visualizing the distributions of scaled numerical features, such as age, serum\_sodium, and platelets, would help identify skewness or outliers. For instance, a histogram or box plot for age could reveal if the patient ages cluster around certain values, indicating typical heart failure age groups in this dataset. Similarly, observing the distribution of serum\_sodium can show if extreme values align with specific health outcomes, as serum sodium levels play a role in heart function.

**Correlation Matrix and Heatmap:**

Calculating the correlation matrix for numerical features, followed by visualizing it with a heatmap, would highlight any strong positive or negative correlations between variables. For example, a heatmap can reveal if high serum sodium levels correlate with higher or lower age values or if platelets and serum\_sodium interact. These insights guide feature selection and might identify redundant features to remove before clustering.

**Principal Component Analysis (PCA):**

Applying PCA helps reduce the dimensionality of the data while preserving as much variance as possible. By projecting the data into a 2D or 3D space, PCA provides a way to visualize potential clusters. For instance, plotting the first two or three principal components might show if patients with higher serum\_sodium or older age group together, suggesting a natural clustering pattern. PCA not only facilitates visualization but also reduces computational costs if applied before clustering.

**Box Plots and Violin Plots:**

These visualizations compare the distribution of each feature across categorical groups, such as smoking and sex. A box plot could illustrate whether smokers tend to have lower or higher platelet levels on average compared to non-smokers. Such insights may reveal variables that contribute to the formation of distinct patient clusters, indicating that these factors are critical to include in the clustering model.

**⁠Model Creation and Testing-**

This section focuses on implementing a clustering model to group patients based on their clinical characteristics, optimizing the model to find the best clustering structure, and assessing cluster quality.

**Finding the Optimal Number of Clusters:**

In the code, the clustering model’s effectiveness can be enhanced by determining the most appropriate number of clusters. Common methods to identify this number include:

**Elbow Method:** The Elbow Method involves plotting the Within-Cluster Sum of Squares (WCSS) against the number of clusters (k). As k increases, the WCSS typically decreases since more clusters fit the data more precisely. However, after a certain point, adding more clusters results in diminishing returns, causing an "elbow" shape on the plot. The location of this elbow indicates the optimal k.

**Silhouette Analysis:** This method calculates the average silhouette score, which measures how similar a data point is to its own cluster compared to others. Silhouette scores range from -1 to 1, with higher values indicating well-defined clusters. Plotting silhouette scores for different values of k can reveal the cluster structure that best represents the data.

**Implementing the Clustering Model:**

Based on the nature of the dataset, K-means clustering would be an effective choice. K-means is well-suited for scaled numerical data and quickly identifies patterns in medium-sized datasets. In this notebook, K-means could be implemented by iterating over different values of k, assessing the cluster quality using WCSS and silhouette scores, and choosing the most effective k based on these metrics.

**Evaluating and Interpreting Clusters:**

Once the clusters are formed, examining the centroids of each cluster can reveal insights about the health profiles of grouped patients. For instance, a cluster with higher average age and lower serum\_sodium may indicate older patients with potential heart issues related to sodium imbalance. By analyzing the means of each feature within clusters, the clusters can be interpreted in terms of risk factors, health characteristics, or possible treatment requirements. This final analysis can be presented with bar charts or radar plots that compare feature averages across clusters, providing a clear view of each cluster's health profile.