**Clustering and Classification Analysis Report**

Shashwat Baral

University of the Cumberlands

Advanced Big Data and Data Mining (MSCS-634-B01)

Date of Submission- 8/20/25

**1. Introduction**

This project aims to implement data mining techniques—specifically classification, clustering, and pattern mining—on a real-world dataset to derive significant insights.

We utilize publicly available online datasets (for instance, the Iris Dataset from the UCI Repository) to illustrate practical applications.

**Key objectives:**

* Utilize classification to forecast categorical labels.
* Employ clustering to uncover natural groupings.
* Implement pattern mining to identify frequent associations.
* Examine ethical considerations and real-world significance.

**2. Data Preparation**

**Dataset Utilized:**

* Name: Iris Flower Dataset
* Source: UCI Machine Learning Repository
* Total Records: 150
* Total Features: 4 (Sepal Length, Sepal Width, Petal Length, Petal Width)
* Target Variable: Species (Setosa, Versicolor, Virginica)

**Data Cleaning:**

* Inspected for missing values (none detected).
* Transformed categorical labels into numerical format (for modeling purposes).
* Standardized features for clustering.

**3. Methodology**

**3.1 Classification**

* Algorithm: Decision Tree Classifier
* Justification: High interpretability, appropriate for small datasets.
* Evaluation Metric: Accuracy, Confusion Matrix

**3.2 Clustering**

* Algorithm: K-Means Clustering
* Number of Clusters: 3 (determined using the Elbow Method)
* Evaluation: Silhouette Score

**3.3 Pattern Mining**

* Algorithm: Apriori (Frequent Itemset Mining)
* Thresholds: Min Support = 0.2, Min Confidence = 0.6
* Objective: Discover feature associations and patterns.

**4. Results & Evaluation**

**4.1 Classification Results**

* The Decision Tree model achieved an accuracy rate of 96%.
* Most misclassifications were observed between the Versicolor and Virginica species.

**4.2 Clustering Results**

* The K-Means algorithm successfully clustered the dataset into three natural groups.
* The clusters exhibited a close alignment with species categories, achieving approximately 90% accuracy.

**4.3 Pattern Mining Results**

* Example Rule: If Petal Length is greater than 4.5, then Species is Virginica (Confidence = 0.92).
* These rules demonstrated strong correlations between petal dimensions and species classification.

**5. Insights**

* Classification Insight: Simple decision boundaries effectively separate Iris Setosa, while there is an overlap between Versicolor and Virginica.
* Clustering Insight: The unsupervised learning approach naturally grouped the species, thereby validating biological distinctions.
* Pattern Insight: Petal dimensions serve as the most discriminative factors for species identification.

**6. Ethical Considerations**

* It is essential to avoid bias in both model selection and interpretation.
* The dataset should be utilized solely for educational and research purposes.
* Caution is advised when applying results to fields such as healthcare or hiring, where bias may lead to significant consequences.

**7. Recommendations**

* For practical applications, employing ensemble models such as Random Forest or Gradient Boosting may enhance accuracy.
* Utilizing larger and more diverse datasets will improve generalizability.
* Implementing a hybrid approach that combines clustering and classification can facilitate exploratory analysis.

**8. Challenges Encountered**

* Determining the optimal number of clusters necessitated experimentation.
* Addressing the overlapping class boundaries between Versicolor and Virginica posed challenges.
* Striking a balance between interpretability and accuracy in algorithm selection was also a concern.

**9. References**

* *UCI Machine Learning Repository: Iris Dataset*
* *Han, Kamber, & Pei (2011). Data Mining: Concepts and Techniques.*
* *Scikit-learn Documentation: https://scikit-learn.org*