

Task 3: Classification & Association Rule Mining Report

Student Name: Peter Kidiga

ID: 671341

Part A: Classification Analysis (Iris Dataset)

1. Model Performance

Two classification models were trained on the Iris dataset: a **Decision Tree** and a **K-Nearest Neighbors (KNN, k=5)** classifier.

- **Decision Tree Results:** The decision tree achieved an accuracy of approximately [Insert Your Accuracy, e.g., 1.00 or 0.97]. The classification report indicates high precision and recall across all three species (*setosa*, *versicolor*, *virginica*). The visualization (`decision_tree_viz.png`) reveals that the model effectively used petal length and width as the primary splitting criteria to distinguish between classes.
- **KNN Comparison:** The KNN model ($k=5$) achieved an accuracy of [Insert Your KNN Accuracy, e.g., 1.00].

2. Comparison & Conclusion

Comparing the two models, [Insert which was better, usually KNN or both are equal] performed slightly better/equally well.

- **Why?** The Iris dataset is small with well-separated clusters. KNN is often superior for such datasets because it creates smooth, non-linear decision boundaries based on local proximity. Decision Trees, while interpretable, can sometimes overfit or create rigid rectangular boundaries that may misclassify points on the edge of a cluster.
- **Selection:** For this specific dataset, **KNN** is preferred for its robustness, though the Decision Tree offers better interpretability for understanding *why* a flower is classified as a certain species.

Part B: Association Rule Mining (Market Basket Analysis)

1. Overview

Using the Apriori algorithm on a synthetic dataset of 50 transactions, we identified frequent itemsets with a minimum support of **0.2** and derived association rules with a minimum confidence of **0.5**.

2. Rule Analysis

One of the strongest rules identified was:

Rule: {Diapers} -> {Beer}

Metrics: Support: ~[Insert Support], Confidence: ~[Insert Confidence], Lift: > 1.0

Interpretation:

- **Lift > 1.0:** This indicates a strong positive correlation; customers who buy diapers are significantly more likely to buy beer than a random customer.
- **Confidence:** A confidence of [Insert Confidence, e.g., 0.8] means that in 80% of cases where diapers were purchased, beer was also purchased.

3. Business Implications

This rule suggests a specific shopping pattern (e.g., parents picking up supplies). A retailer could leverage this insight by:

- **Placement:** Placing these items closer together in the store to increase convenience and impulse purchases.
- **Promotions:** Creating "Weekend Essentials" bundles that include both items.
- **Recommendation Systems:** If a user adds diapers to their online cart, the system should automatically recommend beer or related snacks.

Submission Checklist

1. **Code Files:** `preprocessing_iris.py` (or the self-contained block), `clustering_iris.py`, `mining_iris_basket.py`, `olap_queries.py`.
2. **Screenshots/Images:**
 - o `star_schema_diagram.png`
 - o `iris_pairplot.png`
 - o `iris_heatmap.png`
 - o `elbow_curve.png`
 - o `cluster_visualization.png`
 - o `decision_tree_viz.png`
 - o `olap_visualization_uk_sales.png`
3. **Database:** `retail_dw.db` (if small enough, or just the SQL script `warehouse_schema.sql`).
4. **Reports:** The text provided above.