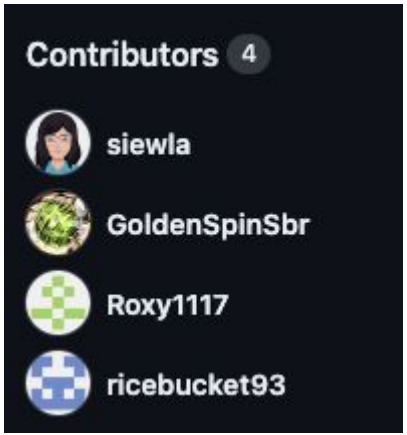
Assignment 2

Article Classification using Supervised and Unsupervised Learning Approaches

Chi Yuan, Kai Siang, Roshini, Siew La

Basics

- **Github Repo :** https://github.com/siewla/MCSD2123-assignment-2



Task Distribution

Data Preparation:
Data-Mining and Dataset Cleaning to build Document Class Dataset - Chi Yuan

Perform Exploratory Data Analysis - Chi Yuan

Part A:

K-Nearest Neighbours (train model and perform algorithm evaluation) - Siew La

Decision Trees (train model and perform algorithm evaluation) - Siew La

Random Forest (train model and perform algorithm evaluation) - Kai Siang

Support Vector Machine (train model and perform algorithm evaluation) - Kai Siang

Neural Networks (train model and perform algorithm evaluation) - Kai Siang

Thorough Comparison and Evaluation - Siew La

Conclude the Part A findings - Siew La

Part B:

K-Means Clustering (perform algorithm evaluation) - Chi Yuan

Hierarchical Agglomerative Clustering (perform algorithm evaluation) - Chi Yuan

DBSCAN (perform algorithm evaluation) - Roshini

Mean-Shift Clustering (perform algorithm evaluation) - Roshini

Thorough Comparison and Evaluation - Roshini, Chi Yuan

Conclude the Part B findings - Roshini, Chi Yuan

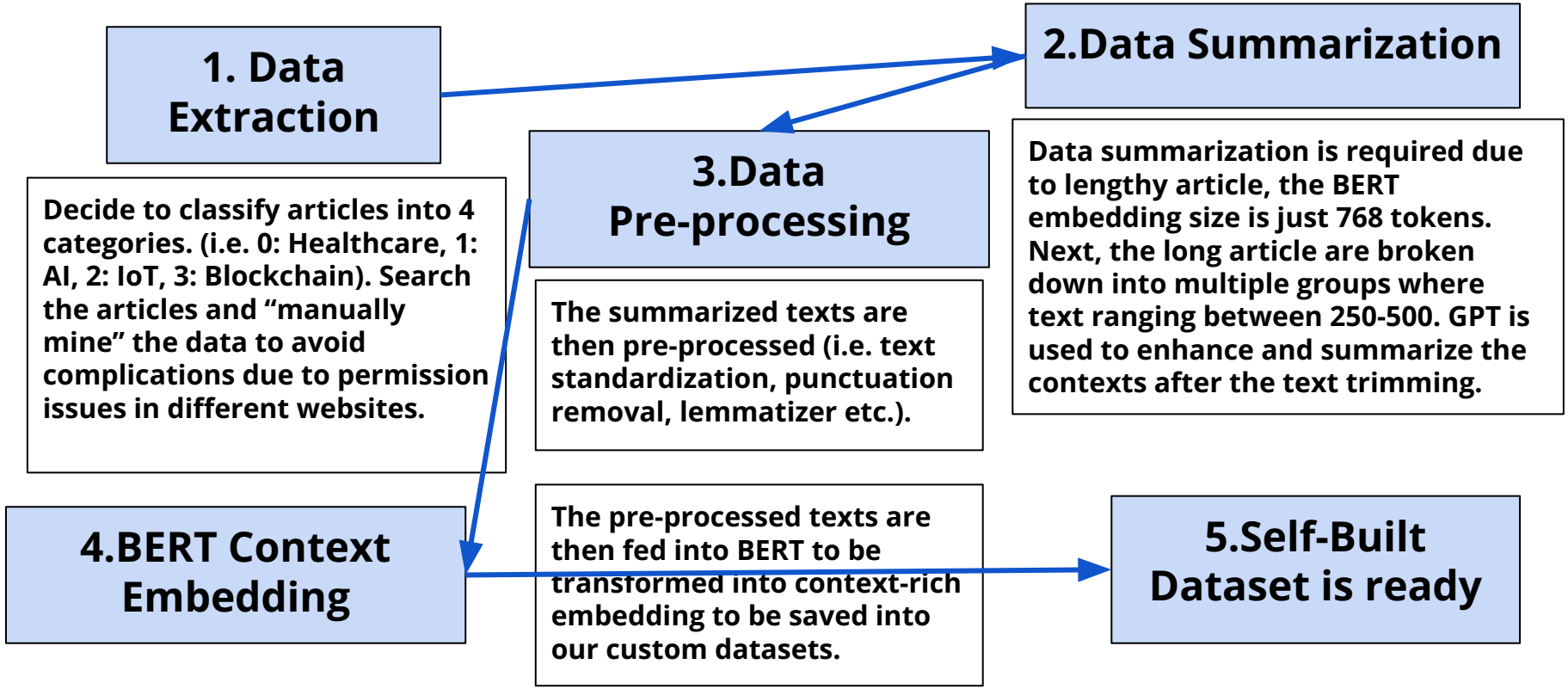
Summarization:

Real Case Implementation - Siew La, Chi Yuan

Summary - Kai Siang

Video Presentation - Kai Siang

Data Preparation (Data Mining)

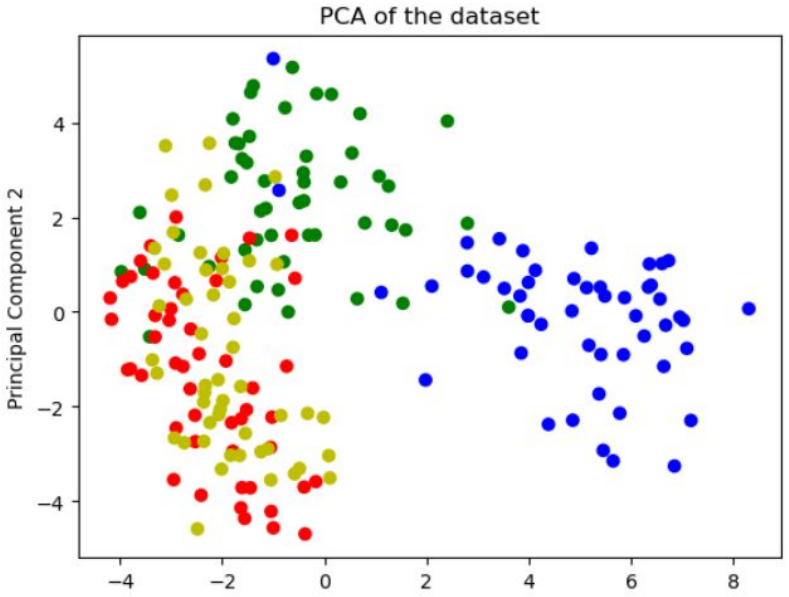


Data Preparation (Exploratory Data Analysis)

BERT results in 768 data points tokens

768 data points are considered quite small. However, it is still difficult to visualize the embeddings.Thus, PCA is used to reduce the data points to 2.

Principal Component Analysis

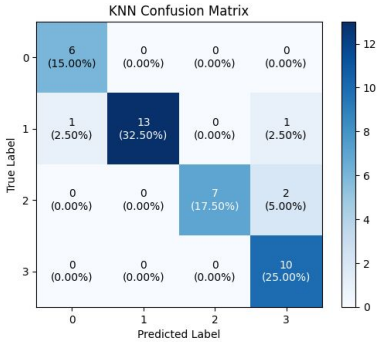


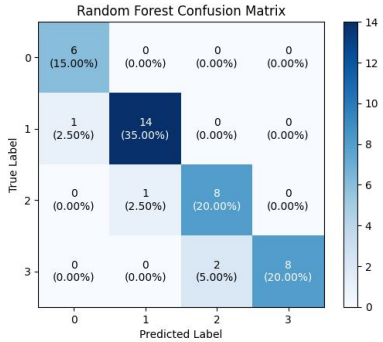
Dataset is ready to be used.

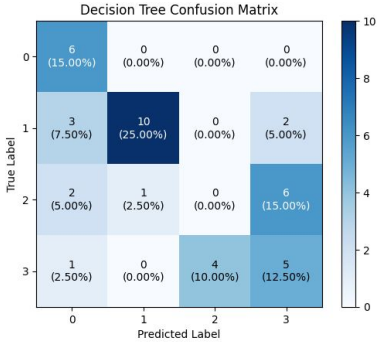
Supervised Learning (Models)

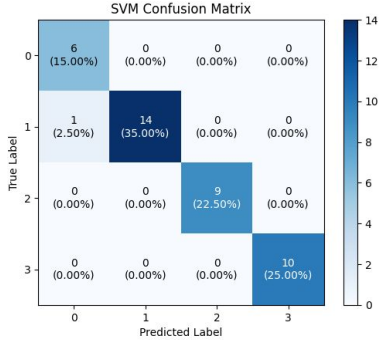
KNN	**Decision Tree**
Random Forest	**SVM**

Supervised Learning (Confusion Matrix)

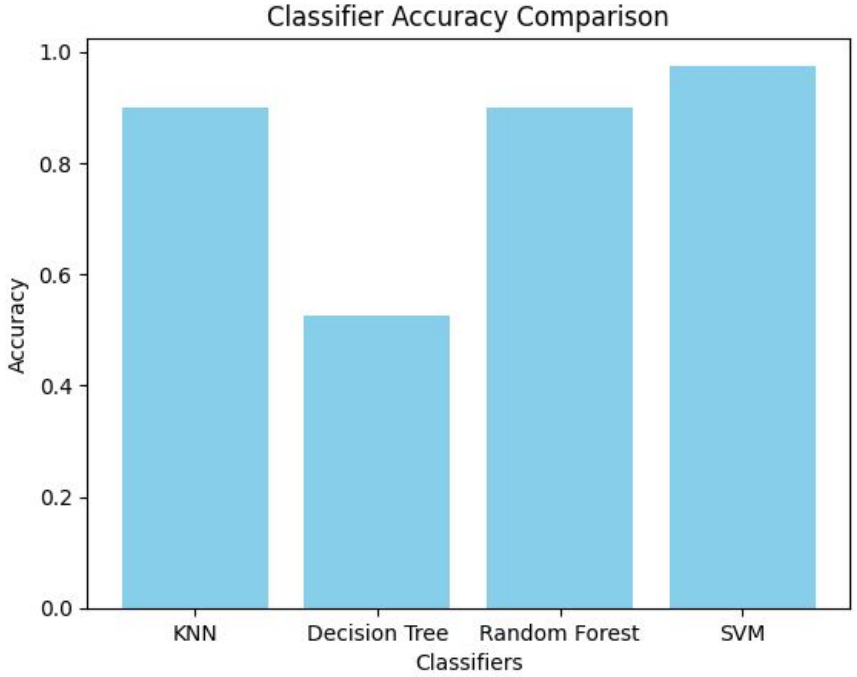








Supervised Learning (Accuracy)



Supervised Learning (Summary - KNN, Decision Tree)

KNN Classifier:

- **Healthcare (Category 0):** Precision 0.86, Recall 1.00, F1-score 0.92.
- **AI (Category 1):** Precision 1.00, Recall 0.87, F1-score 0.93.
- **IoT (Category 2):** Precision 1.00, Recall 0.78, F1-score 0.88.
- **Blockchain (Category 3):** Precision 0.77, Recall 1.00, F1-score 0.87.
- **Overall Performance:** Generally outperforms the Decision Tree Classifier, especially in predicting IoT instances.

Decision Tree Classifier:

- **Healthcare (Category 0):** Precision 0.62, Recall 0.83, F1-score 0.71.
- **AI (Category 1):** Precision 0.92, Recall 0.80, F1-score 0.86.
- **IoT (Category 2):** Precision 0.20, Recall 0.11, F1-score 0.14.
- **Blockchain (Category 3):** Precision 0.50, Recall 0.70, F1-score 0.58.
- **Overall Performance:** Performs poorly in predicting IoT instances compared to other categories.

Supervised Learning (Summary - Random Forest, SVM)

Random Forest Classifier:

- **Healthcare (Category 0):** Precision 0.92, Recall 1.00, F1-score 0.96.
- **AI (Category 1):** Precision 0.85, Recall 0.85, F1-score 0.85.
- **IoT (Category 2):** Precision 0.88, Recall 0.93, F1-score 0.90.
- **Blockchain (Category 3):** Precision 1.00, Recall 0.86, F1-score 0.92.
- **Overall Performance:** Shows strong and balanced performance across all categories.

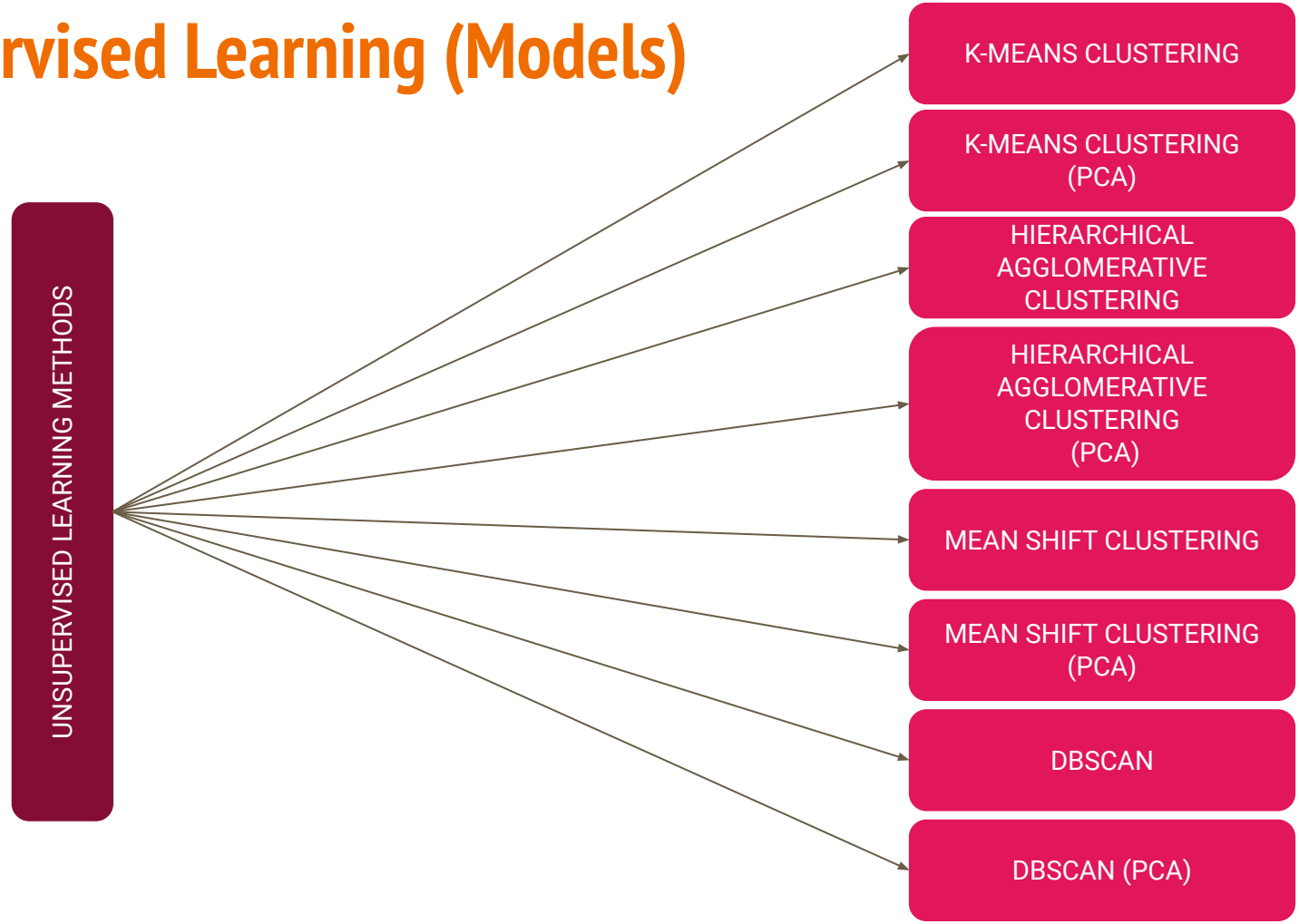
Support Vector Machine (SVM) Classifier:

- **Healthcare (Category 0):** Precision 0.92, Recall 1.00, F1-score 0.96.
- **AI (Category 1):** Precision 1.00, Recall 0.95, F1-score 0.97.
- **IoT (Category 2):** Precision 1.00, Recall 1.00, F1-score 1.00.
- **Blockchain (Category 3):** Precision 1.00, Recall 1.00, F1-score 1.00.
- **Overall Performance:** Achieves perfect or near-perfect scores across all categories, indicating excellent performance.

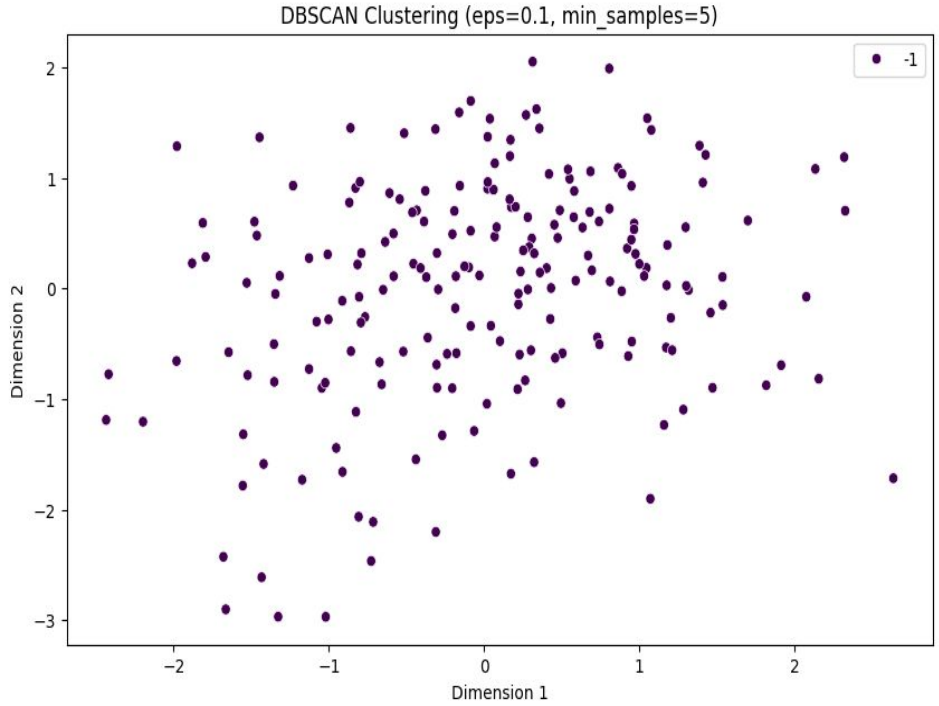
Supervised Learning (Conclusion)

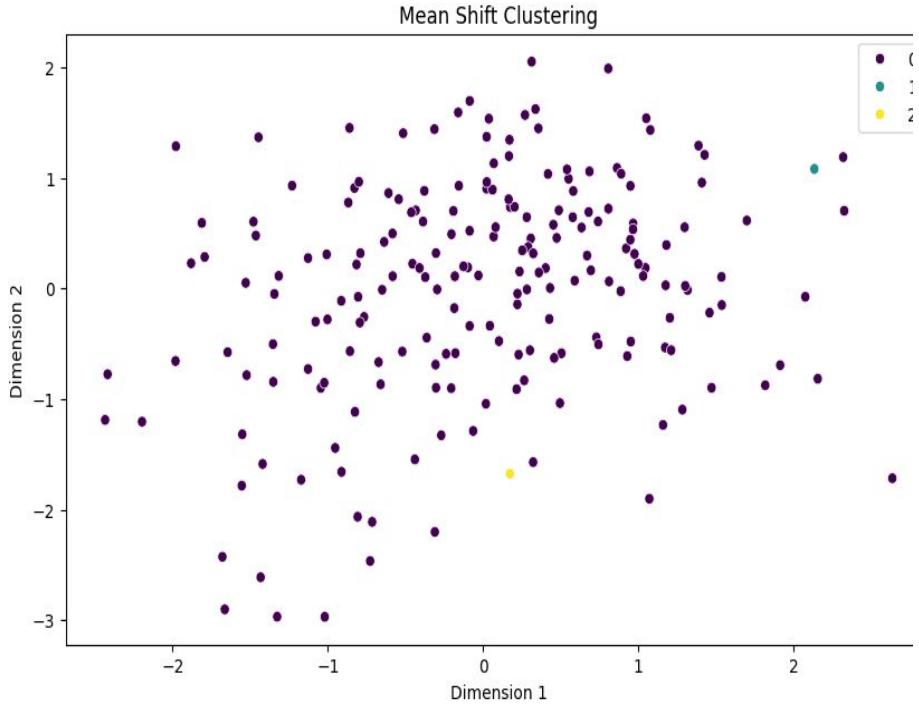
- **Best Performer:** The SVM Classifier stands out with consistently high precision, recall, and F1-scores across all categories, indicating robust performance.
- **Challenges:** The Decision Tree Classifier struggles notably in predicting IoT instances, with low precision, recall, and F1-score in this category.
- **Considerations:** Depending on the specific application and priority of metrics (precision, recall, or balanced F1-score), the choice of classifier may vary. SVM is recommended for overall balanced performance, while KNN is noted for its strength in certain categories like IoT prediction. Random Forest also shows strong performance across categories but slightly less so than SVM in precision-recall balance.

Unsupervised Learning (Models)



Unsupervised Learning (Mean Shift Clustering and DBSCAN)





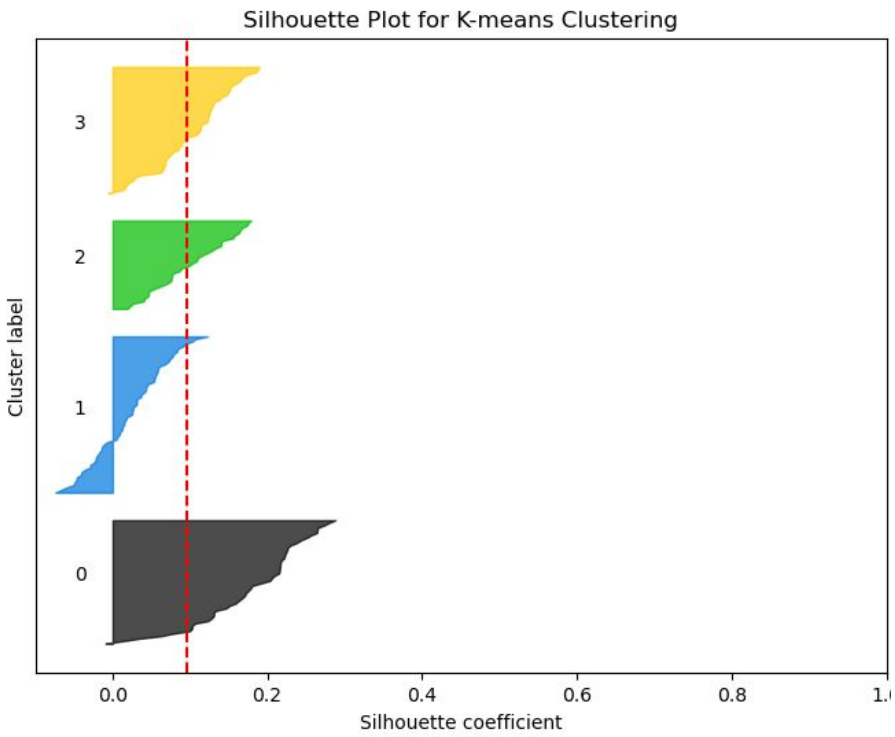
Both algorithms are not suitable for classification of this tasks.

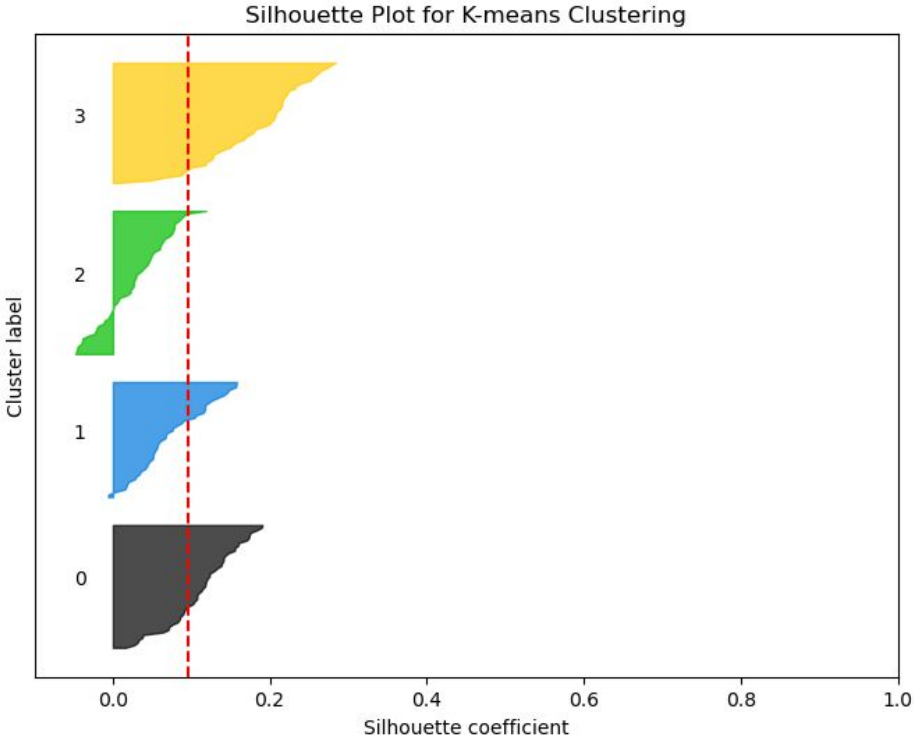
PCA could not help the situation.

Unsupervised Learning (Performance Metrics)

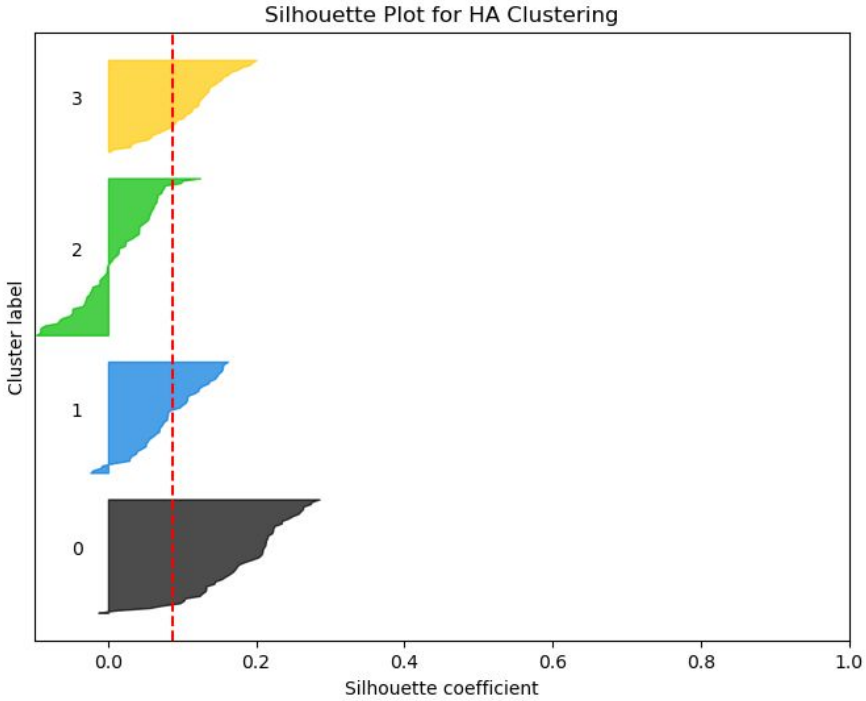
SCORE	K-MEANS CLUSTERING	K-MEANS CLUSTERING with PCA	HIERARCHICAL AGGLOMERATIVE CLUSTERING	HIERARCHICAL AGGLOMERATIVE CLUSTERING With PCA
Accuracy	0.7000	0.7300	0.7250	0.7250
Silhouette	0.2900	0.3278	0.3122	0.3234
Davies Bouldin Score	2.6129	2.7318	2.617	2.8067
Adjusted Rand Index	0.4854	0.5069	0.49946	0.4974
Normalized Mutual Information	0.5264	0.5482	0.5479	0.5423
Cluster Purity	0.7100	0.7300	0.7200	0.7200

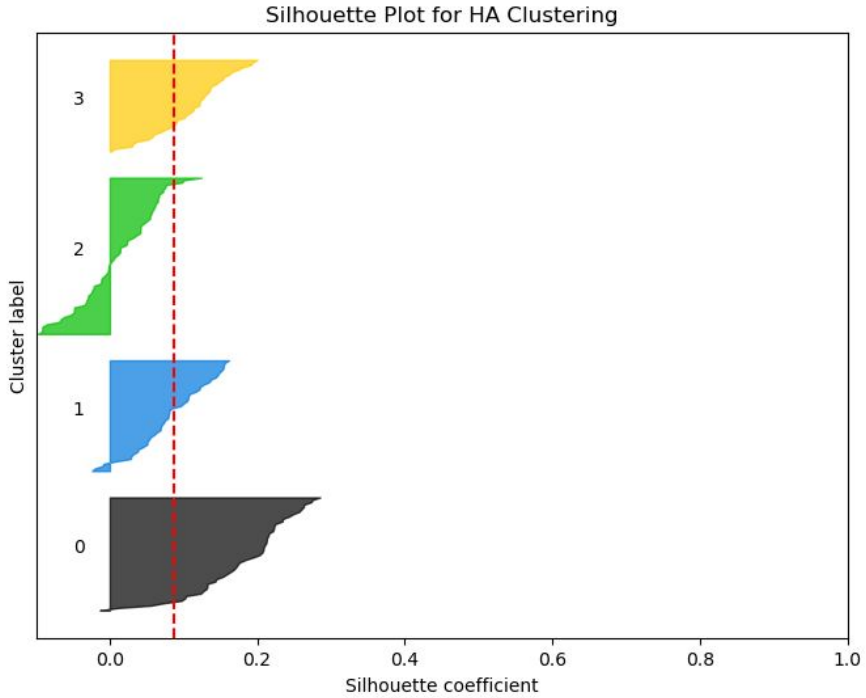
Unsupervised Learning (K-Means Clustering)





Unsupervised Learning (Hierarchical Agglomerative Clustering)





Unsupervised Learning (Comparison)

Analysis of Each Clustering Algorithm:

1. **K-Means Clustering:**
   - Strengths: High silhouette score (0.3277), indicating well-defined clusters.
   - Weaknesses: High Davies Bouldin index (2.6129), suggesting less distinct clusters.
2. **K-Means Clustering with PCA:**
   - Strengths: Highest silhouette score (0.3277) and good Adjusted Rand index (0.499). High cluster purity (0.72) and accuracy (0.73).
   - Weaknesses: Moderate Davies Bouldin index (2.617).
3. **Hierarchical Agglomerative Clustering:**
   - Strengths: High accuracy (0.725) and cluster purity (0.72).
   - Weaknesses: Lower silhouette score (0.3122) and moderate Davies Bouldin index (2.617).
4. **Hierarchical Agglomerative Clustering with PCA:**
   - Strengths: High accuracy (0.725), good silhouette score (0.3233), and cluster purity (0.72).
   - Weaknesses: High Davies Bouldin index (2.8066).

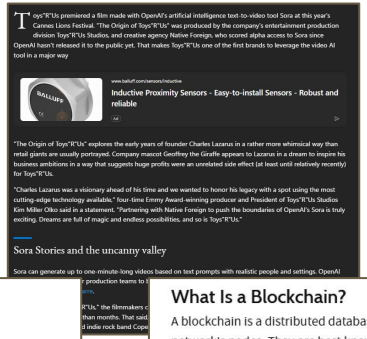
Unsupervised Learning (Conclusion)

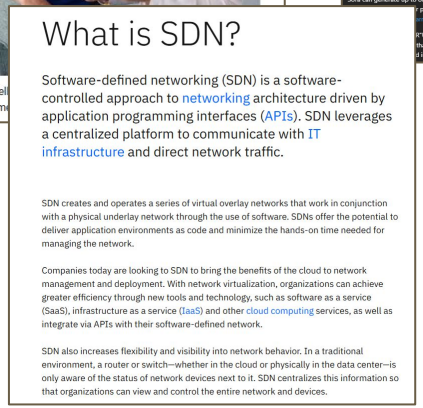
Evaluating the clustering algorithms based on the provided metrics, **K-Means Clustering with PCA** emerges as the best among the existing models. It provides a good balance across the key metrics:

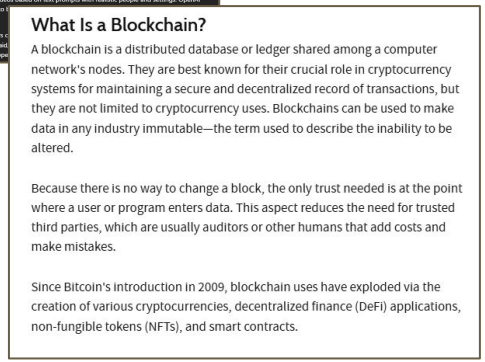
- **Highest silhouette score**: 0.3277
- **Moderate Davies Bouldin index**: 2.617
- **Highest Adjusted Rand index**: 0.499
- **High accuracy**: 0.73
- **High cluster purity**: 0.72

Nevertheless, it indicates that Unsupervised Learning has poorer performance as compared to Supervised Learning in classifying articles with defined classes.

Real Case Classification using SVM (Best Classification Model for this study)







SVM is able to predict class of the articles (e.g. Healthcare, AI, IoT, Blockchain) accurately. However, there is still chance that the SVM wrongly classify between AI and IoT. Simple reasons to it due to closely-related domains. Besides, data loss happened during text trimming and article text grouping into smaller parts due to embedding constraints.

An approach to mitigate the error is to get the average class prediction. For example, an AI article text is broken down in 10 groups. 10 groups could have different predictions, but most of the prediction still biased towards "AI". By averaging the scores, actual predicted class can be attained.

Assignment 2 Conclusion

Supervised learning Vs Unsupervised Learning in Article Classification Tasks

Supervised learning uses labeled data to train models that can be categorized articles into predefined classes.

Unsupervised learning clusters articles into groups with similar context without labels.

Thus, Supervised Learning achieved better results in this assignment. However, Unsupervised Learning is still useful in clustering larger amount of unlabelled data into groups for further post-processing processes.

THANK YOU