Gene Expression Cancer RNA-Seq Multi-Model Analysis

MAI272 - Advanced Machine Learning Project

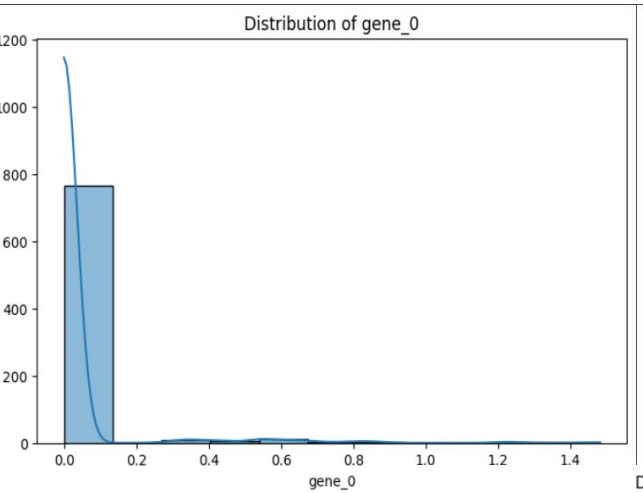
Manav Gupta: 2348529
Priyansha Upadhyay: 2348541
Satyam Kumar: 2348555

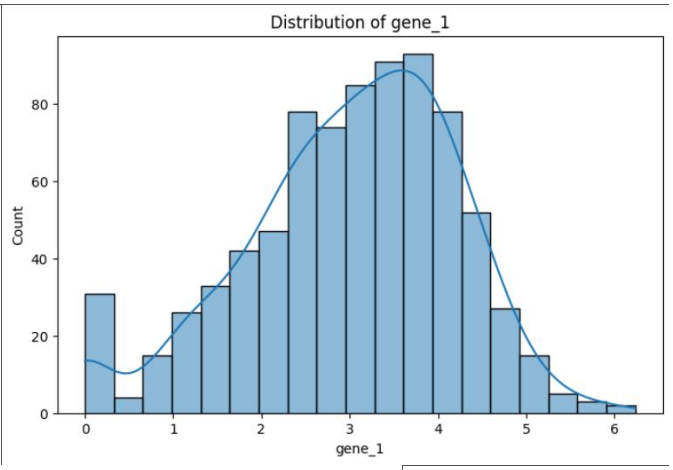
Introduction

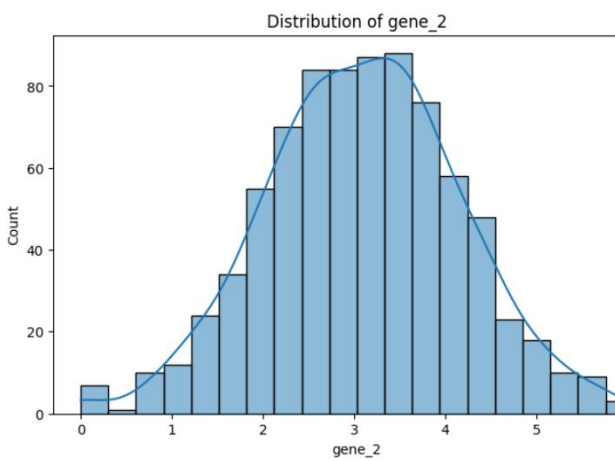
- Gene expression analysis is a cornerstone of modern molecular biology, offering unparalleled insights into cellular processes, disease mechanisms, and potential therapeutic targets.
- This project focuses on a gene expression dataset featuring 801 samples, each characterized by a staggering 20,532 features.
- To address the challenges posed by high dimensionality and class imbalance inherent in gene expression datasets, the project employs advanced dimensionality reduction techniques such as t-SNE and PCA.
- Through this comprehensive approach, the project aims to provide valuable insights into the relative strengths and limitations of various machine learning algorithms and research-driven methodologies in the context of gene expression analysis.

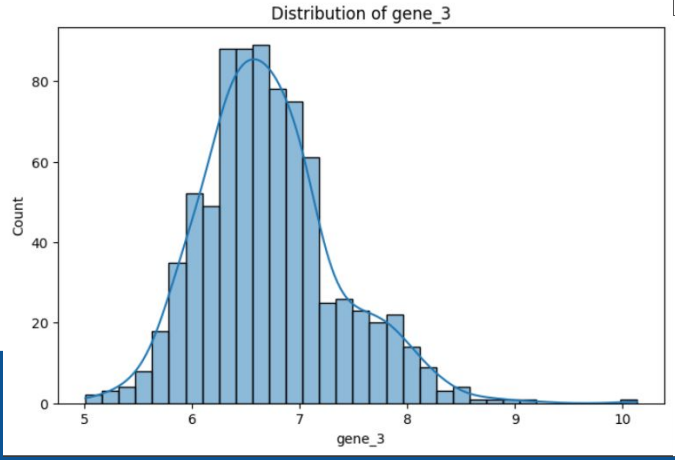
2. Data Processing and Exploration

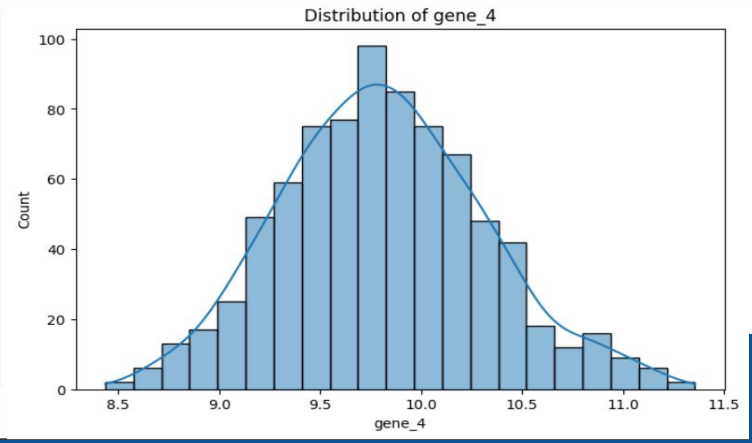
- The initial assessment of the dataset reveals a robust quality, with no missing values and no duplicate columns present. This absence of data gaps and redundancy streamlines the data cleaning process and ensures a solid foundation for subsequent analyses.





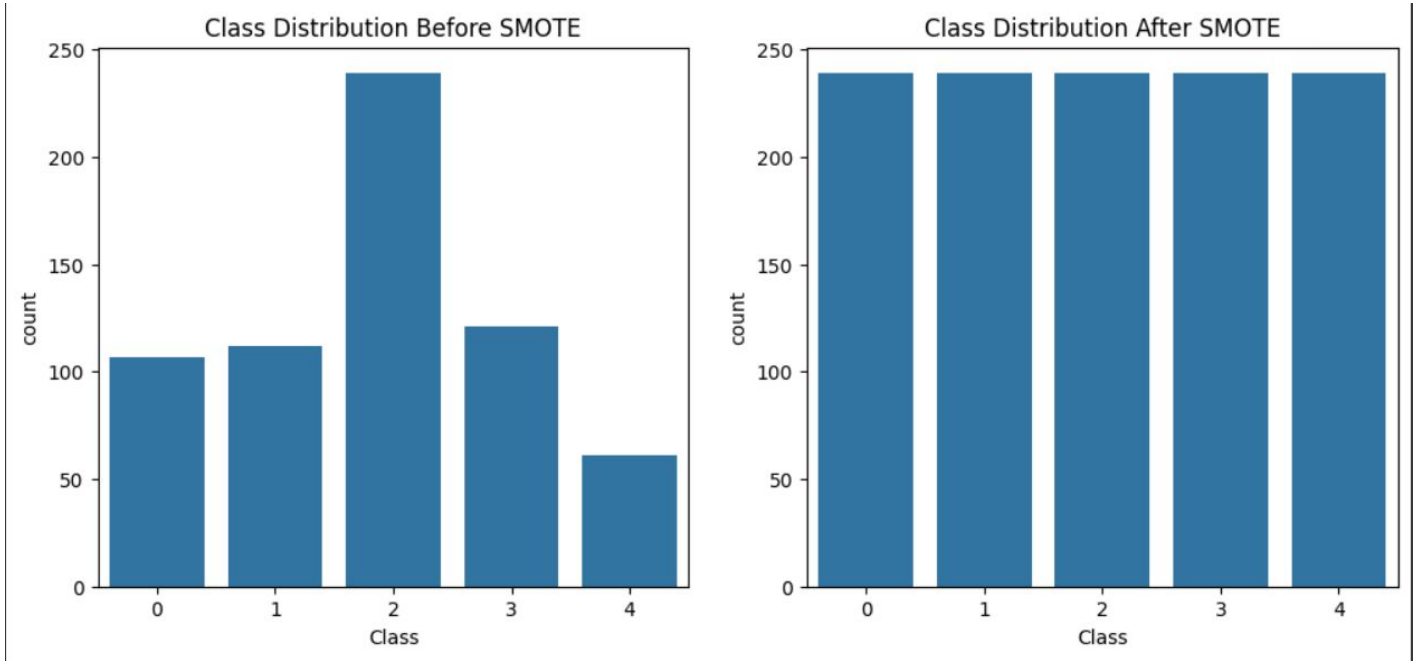






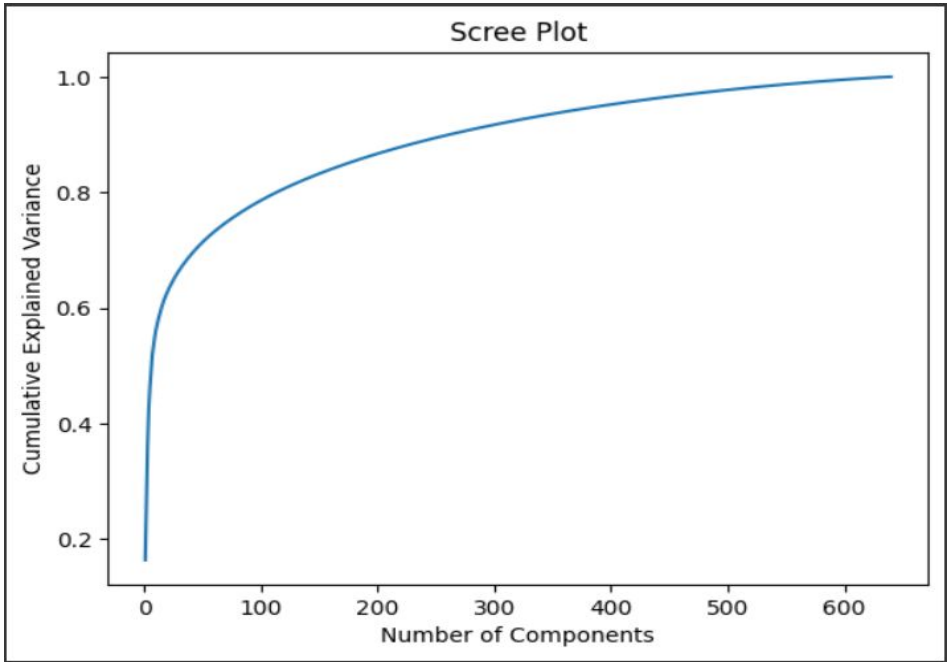
Continued..

- The Synthetic Minority Over-sampling Technique (SMOTE) was applied, effectively correcting the imbalanced dataset. This transformation led to an updated dataset with 239 features.



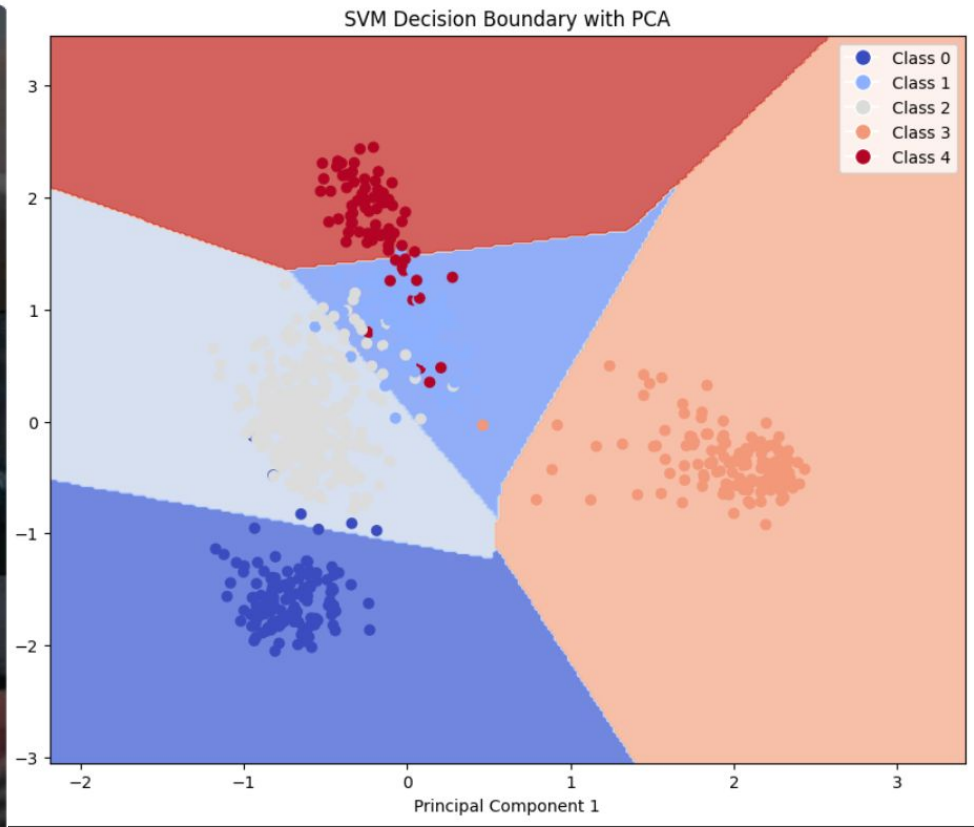
Continued..

- To further enhance the modeling process, Principal Component Analysis (PCA) was employed for dimensionality reduction. The scree plot, based on cumulative explained variance, was instrumental in determining the appropriate number of components.



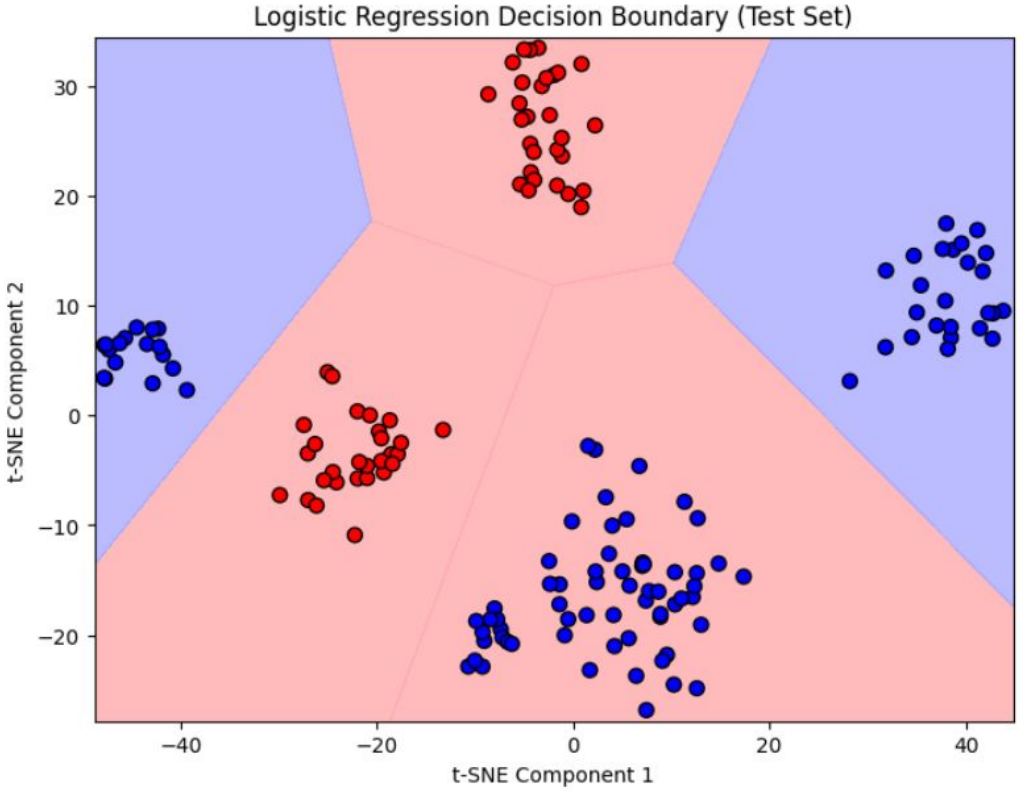
3. Algorithms Implemented: SVM

- Support Vector Machines (SVM), a powerful classification algorithm, were employed for the gene expression dataset. The dataset, enriched through SMOTE and dimensionality reduction using PCA, served as the input for SVM.



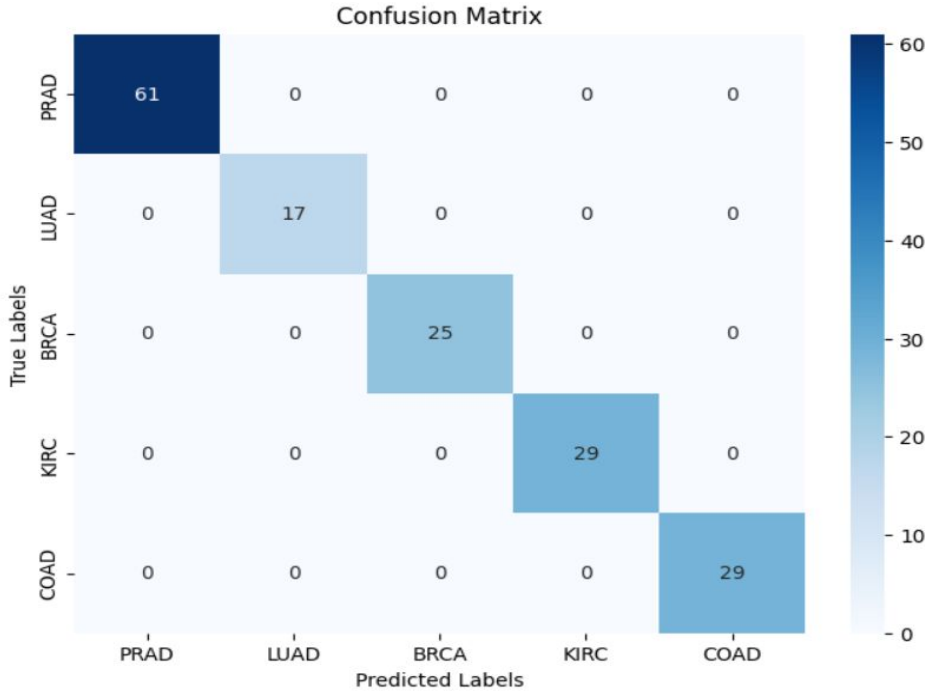
Algorithms Implemented: Logistic Regression

- Logistic Regression, a versatile classification algorithm, was applied to the gene expression dataset after preprocessing with PCA and t-SNE for dimensionality reduction. The logistic regression model aimed to classify instances within the dataset, leveraging the insights gained from these reduction techniques.



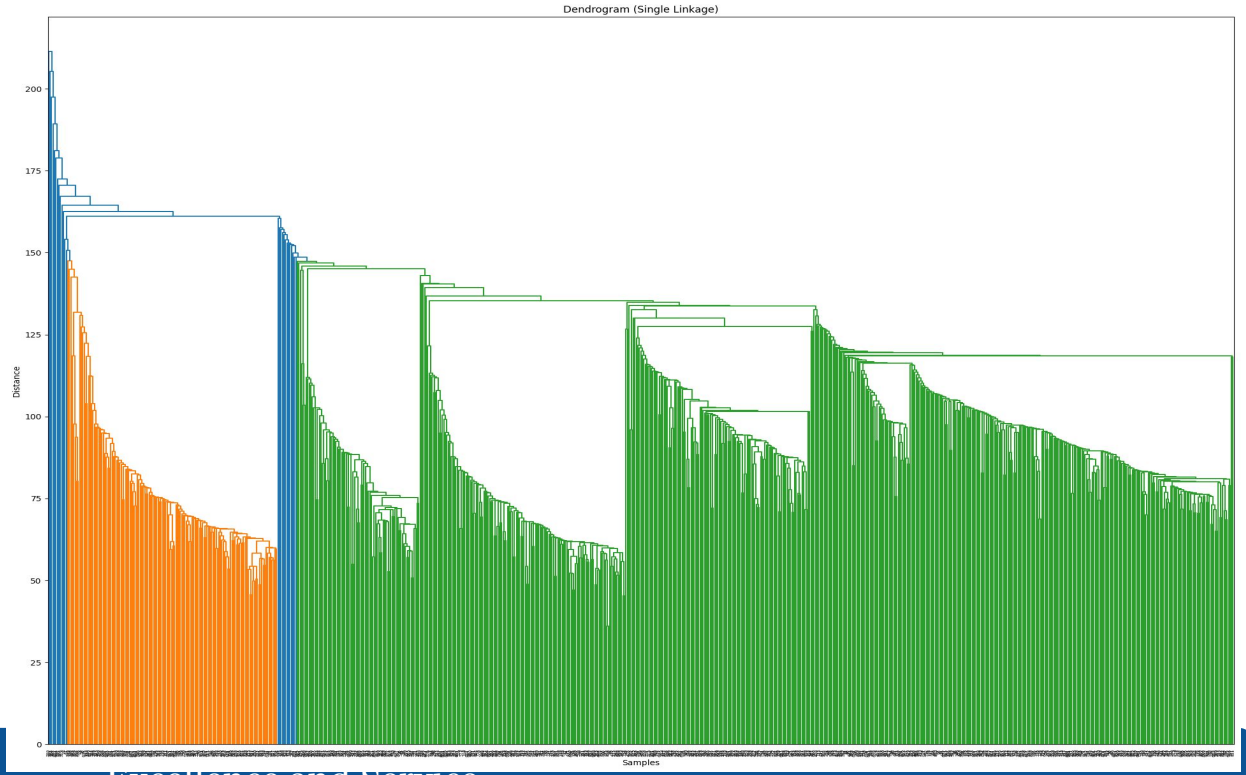
Algorithms Implemented: K-Nearest Neighbours

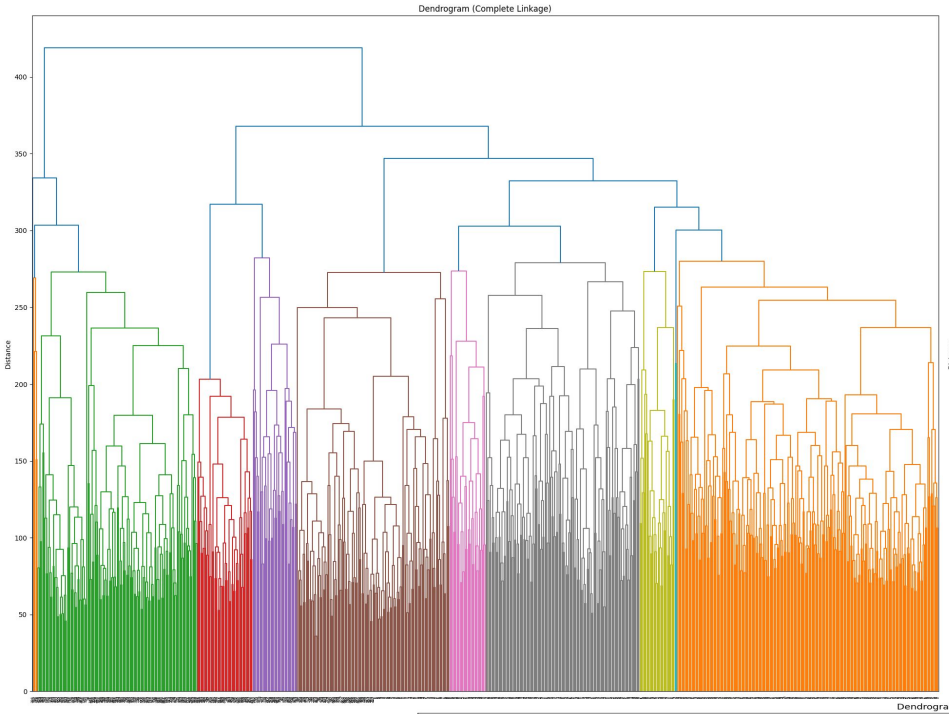
- k-Nearest Neighbors (KNN), a versatile and intuitive classification algorithm, was applied to the gene expression dataset to discern patterns and relationships within the data.

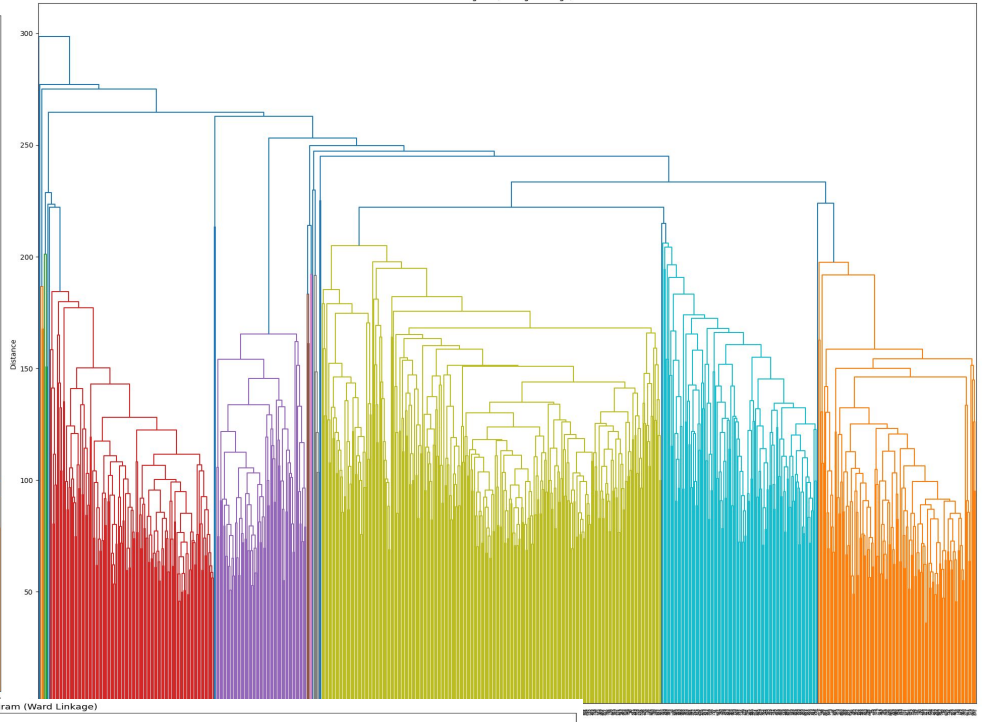


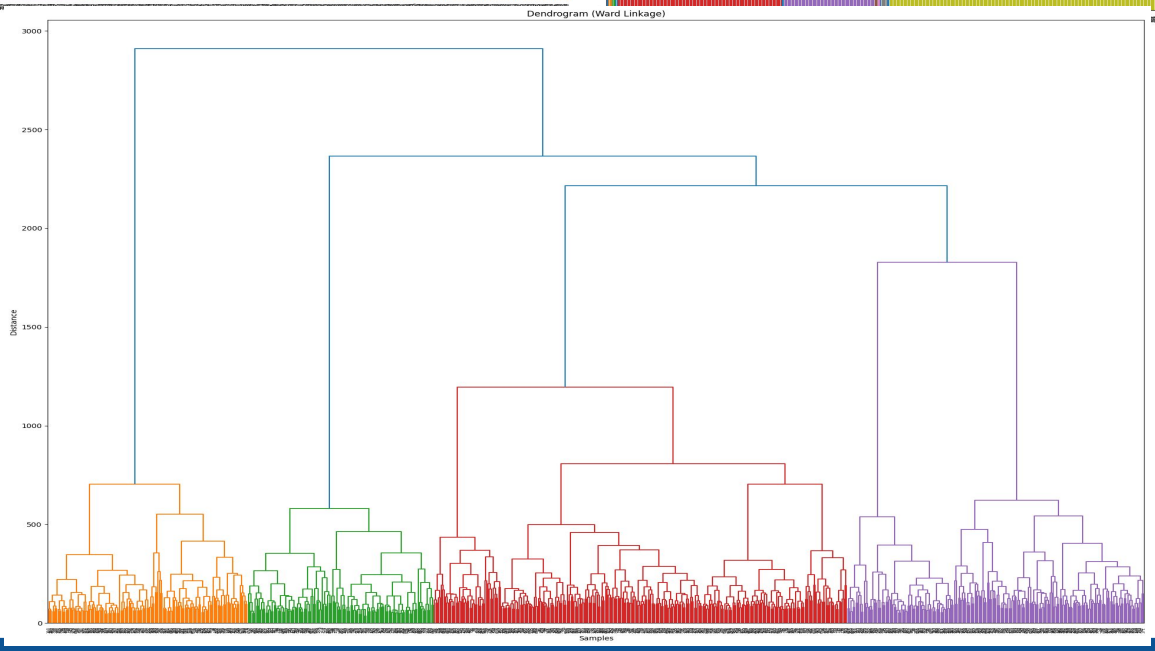
Algorithms Implemented: Hierarchical Clustering

- Hierarchical Clustering, a powerful unsupervised clustering algorithm, was employed to unravel inherent patterns within the gene expression dataset. The algorithm was executed using different linkage methods, including:
  - Single Linkage
  - Ward Linkage
  - Complete Linkage
  - Average Linkage



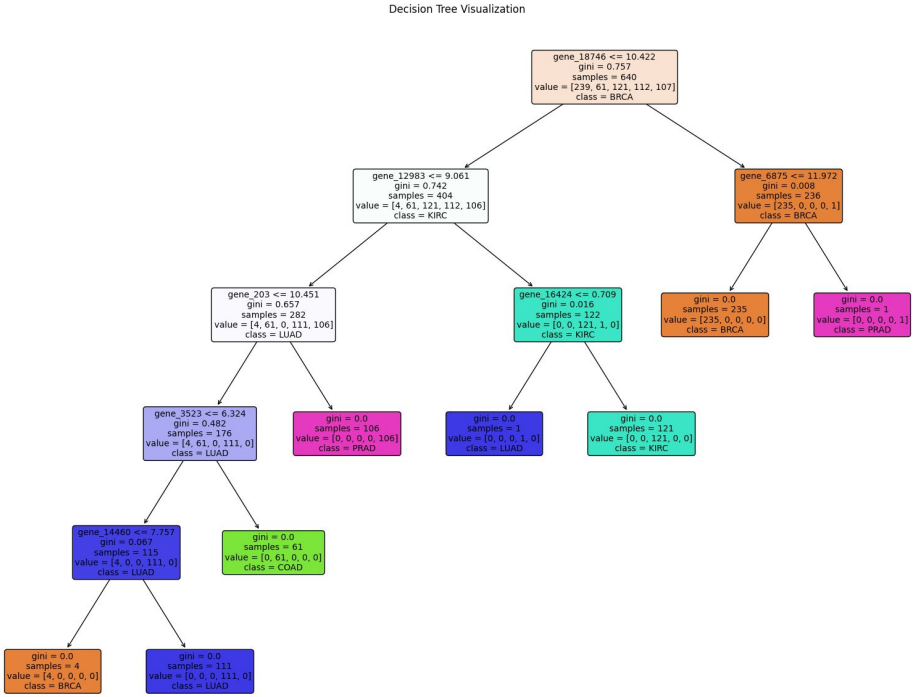

Dendrogram (Complete Linkage)
Distance
0
50
100
150
200
250
300
350
400


Distance
50
100
150
200
250
300


Dendrogram (Ward Linkage)
Distance
0
500
1000
1500
2000
2500
3000
Samples

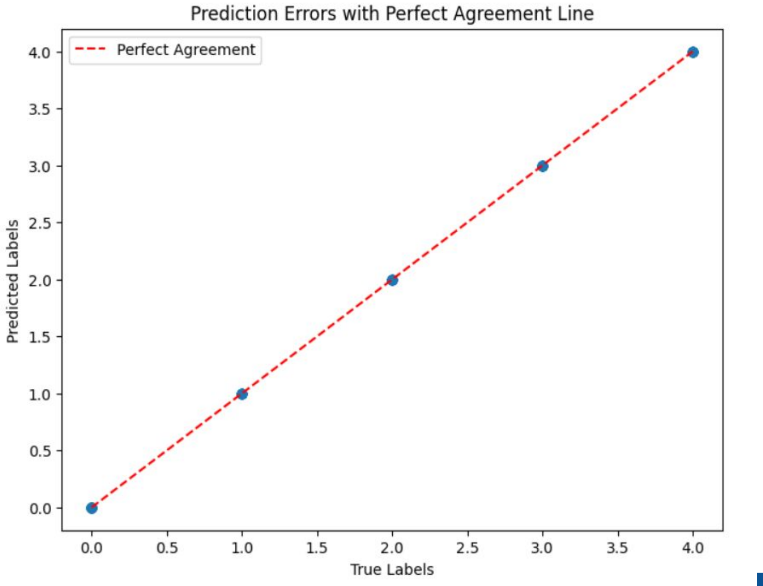
Algorithms Implemented: Decision Tree

- Decision Tree, a fundamental and interpretable machine learning algorithm, was implemented to decipher patterns within the gene expression dataset.
- The Decision Tree model was trained on the preprocessed dataset. Decision trees are known for their simplicity and interpretability, making them valuable for understanding the underlying structure of the gene expression data.



Algorithms Implemented: Light GBM Classifier (Gradient Boost)

- Light GBM, a powerful gradient boosting framework, was employed to capture complex relationships and enhance predictive performance on the gene expression dataset.
- The Light GBM Classifier was trained on the preprocessed dataset. Leveraging gradient boosting, Light GBM builds an ensemble of decision trees, iteratively refining the model's predictions.



4. Comparative analysis of different models

Model	Accuracy	Precision	Recall	F1-Score
SVM (PCA after SMOTE)	1	1	1	1
Tuned SVM (PCA after SMOTE)	1	1	1	1
Logistic Regression	1	1	1	1
Logistic Regression (t-SNE on PCA)	1	1	1	1
k Nearest Neighbors	1	1	1	1
Decision Tree	0.98	0.98	0.98	0.98
LightGBM	1	1	1	1

Table 01: Overview of Model Performance

Linkage Method	Silhouette Score
Single Linkage	0.178
Complete Linkage	0.248
Average Linkage	0.235
Ward Linkage	0.322

Table 02: k Nearest Neighbors Silhouette Results

SVM & Logistic Regression Classification Report:

```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        29
           1       1.00      1.00      1.00        29
           2       1.00      1.00      1.00        61
           3       1.00      1.00      1.00        25
           4       1.00      1.00      1.00        17

    accuracy                           1.00       161
   macro avg       1.00      1.00      1.00       161
weighted avg       1.00      1.00      1.00       161

Confusion Matrix:
[[29  0  0  0  0]
 [ 0 29  0  0  0]
 [ 0  0 61  0  0]
 [ 0  0  0 25  0]
 [ 0  0  0  0 17]]
```

Dig11:SVM Report

```
Training Set Performance:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       107
           1       1.00      0.98      0.99       112
           2       0.99      1.00      1.00       239
           3       1.00      1.00      1.00       121
           4       1.00      1.00      1.00        61

    accuracy                           1.00       640
   macro avg       1.00      1.00      1.00       640
weighted avg       1.00      1.00      1.00       640

Confusion Matrix:
[[107   0   0   0   0]
 [  0 110   2   0   0]
 [  0   0 239   0   0]
 [  0   0   0 121   0]
 [  0   0   0   0  61]]
```

```
Test Set Performance:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        29
           1       1.00      1.00      1.00        29
           2       1.00      1.00      1.00        61
           3       1.00      1.00      1.00        25
           4       1.00      1.00      1.00        17

    accuracy                           1.00       161
   macro avg       1.00      1.00      1.00       161
weighted avg       1.00      1.00      1.00       161

Confusion Matrix:
[[29  0  0  0  0]
 [ 0 29  0  0  0]
 [ 0  0 61  0  0]
 [ 0  0  0 25  0]
 [ 0  0  0  0 17]]
```

Dig13: Logistic Regression Report

K-nearest & Hierarchical Clustering Evaluation Report:

```
Accuracy: 1.00

Precision: 1.00
Recall: 1.00
F1-Score: 1.00
```

```
True Positives: 17
True Negatives: 61
False Positives: 0
False Negatives: 0
```

```
Confusion Matrix:
[[61  0  0  0  0]
 [ 0 17  0  0  0]
 [ 0  0 25  0  0]
 [ 0  0  0 29  0]
 [ 0  0  0  0 29]]
```

Dig15: K nearest neighbor Report

Silhouette Score:

- Single Linkage: 0.178
- Complete Linkage: 0.248
- Average Linkage: 0.235
- Ward Linkage: 0.322

Decision Tree & LightGBM Evaluation Report:

Accuracy: 0.9751552795031055

Classification Report:

	precision	recall	f1-score	support
BRCA	0.95	0.98	0.97	61
COAD	1.00	1.00	1.00	17
KIRC	1.00	1.00	1.00	25
LUAD	0.97	0.97	0.97	29
PRAD	1.00	0.93	0.96	29
accuracy			0.98	161
macro avg	0.98	0.98	0.98	161
weighted avg	0.98	0.98	0.98	161

Confusion Matrix:

```
[[60  0  0  1  0]
 [ 0 17  0  0  0]
 [ 0  0 25  0  0]
 [ 1  0  0 28  0]
 [ 2  0  0  0 27]]
```

Dig17:Decision Tree Report

Accuracy: 1.0

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	61
1	1.00	1.00	1.00	17
2	1.00	1.00	1.00	25
3	1.00	1.00	1.00	29
4	1.00	1.00	1.00	29
accuracy			1.00	161
macro avg	1.00	1.00	1.00	161
weighted avg	1.00	1.00	1.00	161

Confusion Matrix:

```
[[61  0  0  0  0]
 [ 0 17  0  0  0]
 [ 0  0 25  0  0]
 [ 0  0  0 29  0]
 [ 0  0  0  0 29]]
```

Dig18:LightGBM Report

5. Insightful interpretation of results

- Supervised Models:

SVM, Logistic Regression, k Nearest Neighbors, Decision Tree, and LightGBM demonstrated strong performance.
Models are effective in capturing patterns, resulting in high accuracy and minimal misclassifications.

- Unsupervised Model (Hierarchical Clustering):

Silhouette scores indicated that Ward Linkage provided the most well-defined clusters.
Hierarchical clustering is exploratory and may not directly align with supervised metrics.

Continued..

- PCA and SMOTE Impact:

SVM with PCA after SMOTE showcased enhanced performance on the test set, indicating successful handling of imbalanced data.
Other models did not explicitly utilize PCA or SMOTE but still performed exceptionally well.

- Consideration for Decision-Making:

Choice of model depends on specific requirements like interpretability, computational efficiency, or need for a black-box model.
Supervised models generally outperformed unsupervised clustering for this dataset.
Rigorous evaluation on an external dataset is recommended for a more comprehensive understanding of generalization performance.

Comparison on basis of computational cost

Algorithm	Training Cost	Factors Affecting Cost
SVM	High	High-dimensional data, kernel function choice
Logistic Regression	Moderate	Data size, regularization terms
K-Nearest Neighbors (KNN)	High	Data size, choice of k
Hierarchical Clustering	Moderate	Data size, clustering algorithm used
Decision Tree	Moderate	Data size, tree depth
LightGBM	Moderate to High	Data size, number of trees, features, learning rate

Conclusion

The project provided valuable insights into the application of various machine learning algorithms for gene expression analysis. While all models achieved high accuracy (1), individual strengths and limitations emerged:

- Support Vector Machines (SVM) and Logistic Regression delivered exceptional performance, demonstrating their suitability for classifying gene expression data preprocessed with both PCA and SMOTE.
- k-Nearest Neighbors (KNN) also achieved high accuracy, highlighting its versatility for pattern recognition within gene expression data.
- Hierarchical Clustering with Ward Linkage achieved the best Silhouette score, indicating its effectiveness in identifying meaningful clusters within the dataset.

- Decision Tree achieved slightly lower accuracy but offered advantages in interpretability and capturing complex relationships, respectively.
- LightGBM showcased effective using of Gradient Boosting.

Overall, the project underscores the versatility and power of machine learning in extracting valuable insights from gene expression data. The project also highlights the importance of data preprocessing techniques like SMOTE and dimensionality reduction for optimizing model performance.

References

1. Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review
2. A Classification Framework Applied to Cancer Gene Expression Profiles - PMC
3. Cancer classification using gene expression data - ScienceDirect
4. 1.4. Support Vector Machines — scikit-learn 1.4.0 documentation
5. sklearn.linear_model.LogisticRegression — scikit-learn 1.4.0 documentation
6. 1.6. Nearest Neighbors — scikit-learn 1.4.0 documentation
7. 2.3. Clustering — scikit-learn 1.4.0 documentation
8. 1.10. Decision Trees — scikit-learn 1.3.2 documentation
9. LightGBM's documentation!
10. http://scikit-learn.org/stable/modules/model_evaluation.html

Thank You!!