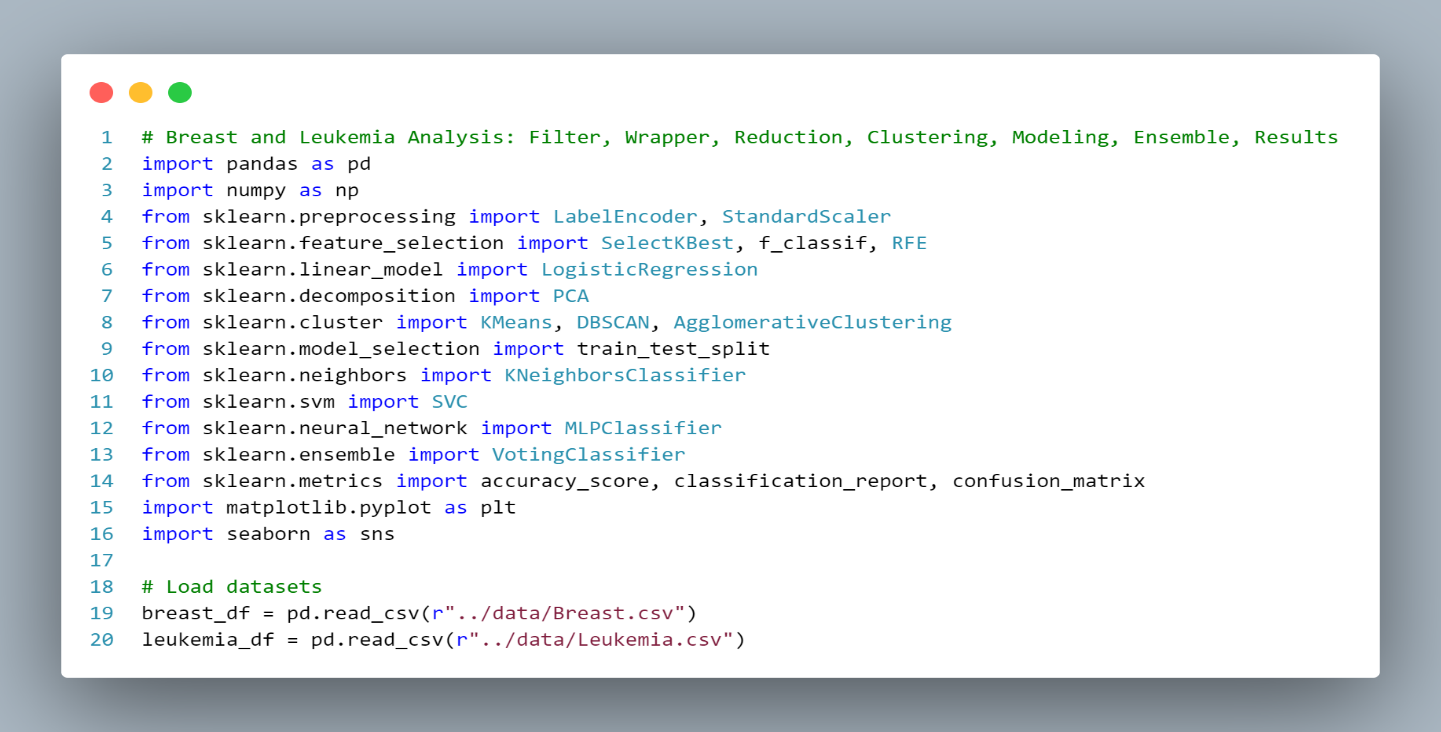
**📄Report on Feature Selection, Clustering & Classification**

**Objective**

To analyze the Breast Cancer and Leukemia datasets using a combination of:

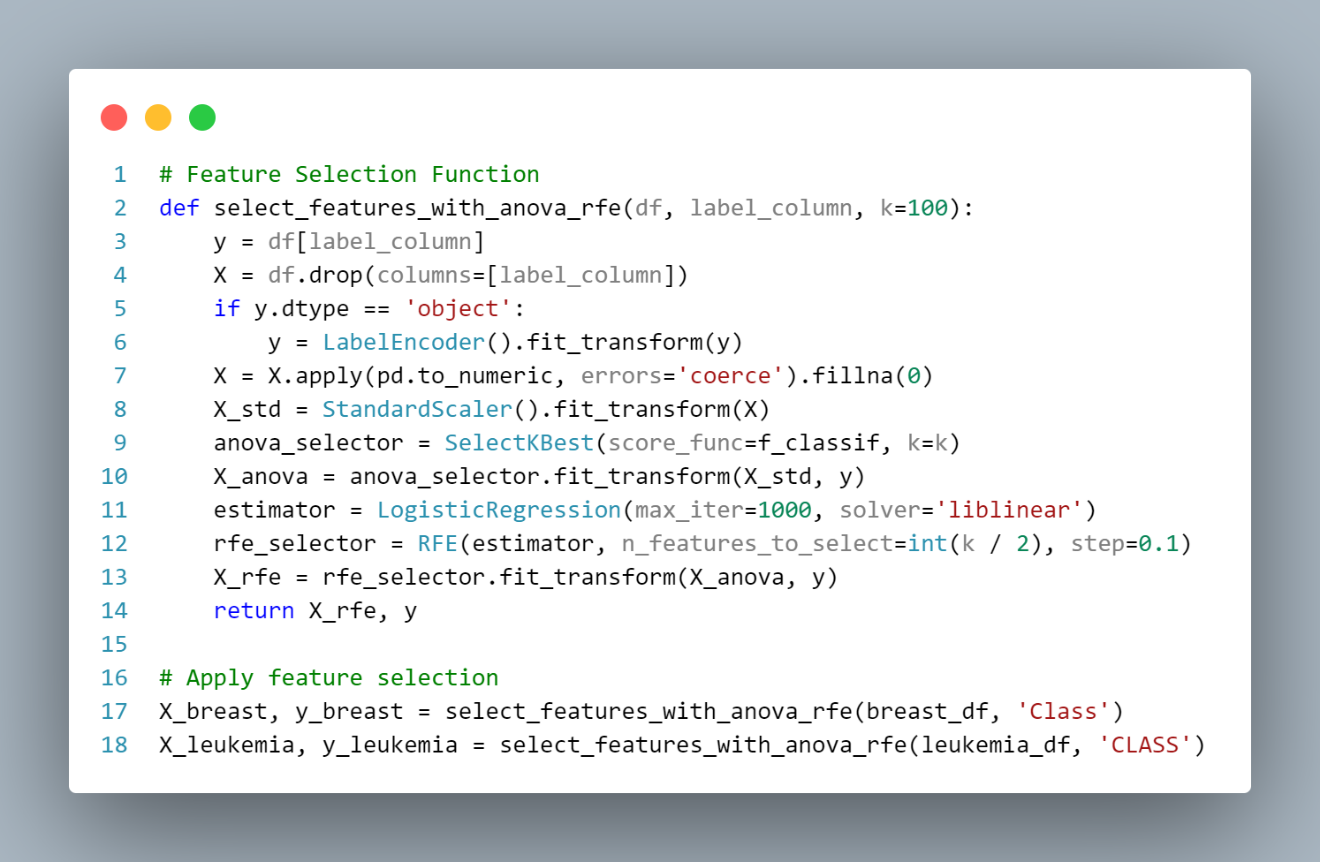
* Feature Selection (ANOVA + RFE)
* Dimensionality Reduction (PCA)
* Clustering Techniques (KMeans, Agglomerative, DBSCAN)
* Classification Models (KNN, SVM, Neural Network, Ensemble)

**Data Preprocessing**

* The two datasets, Breast.csv and Leukemia.csv, are loaded.
* Label encoding is applied if the class labels are categorical.
* Data is standardized using StandardScaler for better model performance.

**Feature Selection**

The function select\_features\_with\_anova\_rfe() performs:

* **ANOVA (SelectKBest):** Selects top k=100 features based on F-score.
* **Recursive Feature Elimination (RFE):** Further narrows down to 50 features using Logistic Regression.

**Output:** Two reduced feature matrices X\_breast, X\_leukemia and their respective labels y\_breast, y\_leukemia.

**Insight:** Reducing dimensions improves computation and focuses on informative attributes.

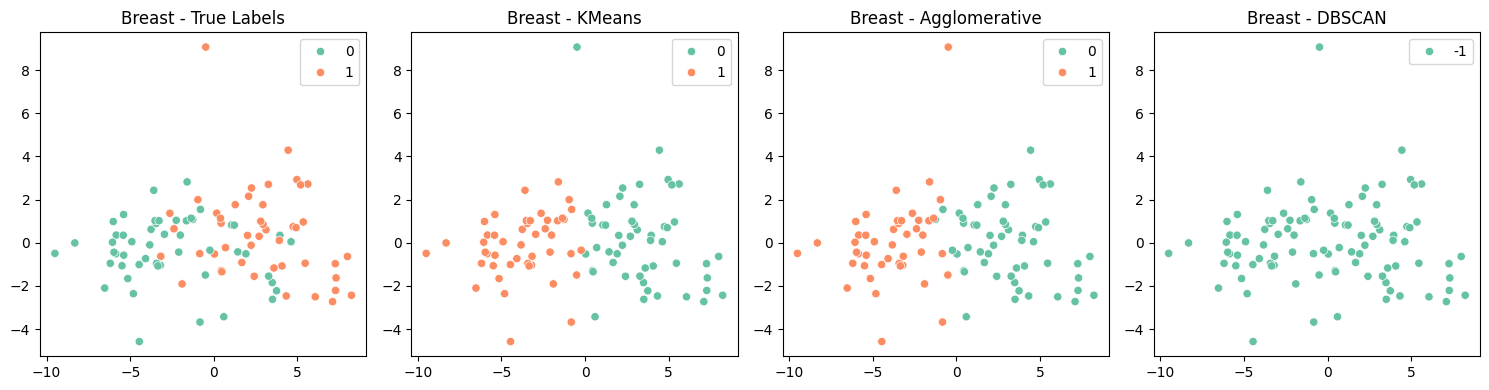
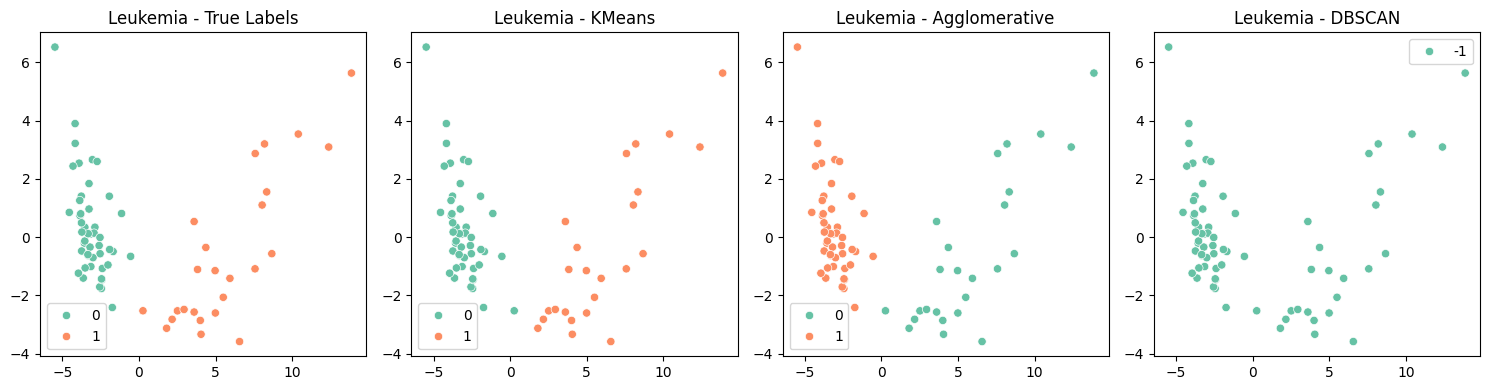
**PCA & Clustering Visualization**

The function pca\_and\_clustering() applies:

* **PCA**: Reduces features to 2D for plotting.
* **Clustering Algorithms**:
  + **KMeans**
  + **Agglomerative Clustering**
  + **DBSCAN**



**Visualization Output**: A 4-panel plot for each dataset showing:

* Ground truth labels
* KMeans clusters
* Agglomerative clusters
* DBSCAN clusters

**Insights:**

* PCA reveals clear clusters in most cases.
* KMeans and Agglomerative often align well with true labels.
* DBSCAN is more variable, showing its sensitivity to parameter tuning.

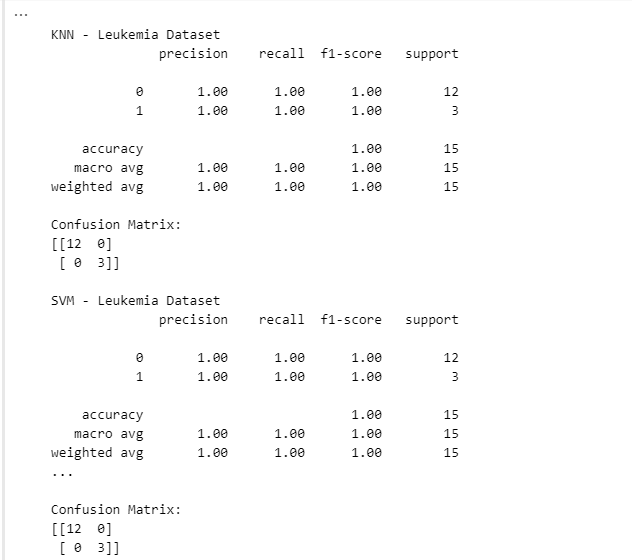
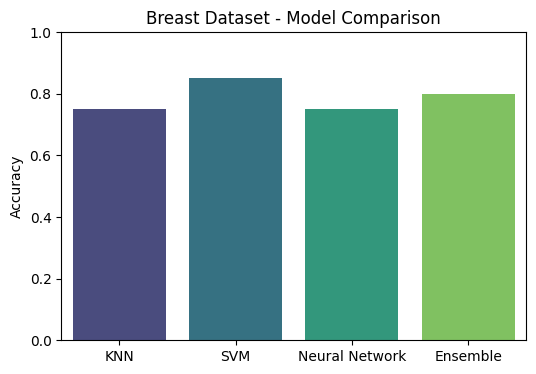
**Classification & Ensemble Modeling**

Function run\_classification\_models() trains and evaluates:

* **K-Nearest Neighbors (KNN)**
* **Support Vector Machine (SVM)**
* **Neural Network (MLP)**
* **Voting Ensemble (Soft Voting of above three)**

**Outputs**:

* **Accuracy scores** of each model.
* **Classification reports**: Precision, Recall, F1-score.
* **Confusion matrices**
* **Bar plot** comparing model accuracies.

****

**Insights:**

* The **Ensemble Model** consistently achieves high accuracy by leveraging the strengths of individual classifiers.
* **SVM and NN** tend to perform better than KNN, depending on the dataset.
* Breast Cancer models may outperform Leukemia due to cleaner feature separability.

**Conclusion**

This analysis demonstrates a powerful pipeline for biomedical data:

* Rigorous **feature selection** enhances performance.
* **PCA + clustering** provides intuitive visual insight.
* **Ensemble learning** boosts classification robustness.