# 1. Leukemia Classification (ALL vs AML)

## Problem Statement

Classify leukemia patients into Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML) subtypes using gene expression data from microarray experiments. The dataset contains 72 samples with 7,129 gene expression features, presenting a high-dimensional classification challenge.

## Pipeline Architecture

1. Data Preprocessing: StandardScaler for feature normalization  
2. Algorithm: Logistic Regression with L2 regularization  
3. Validation: Train-test split (80-20)  
4. Evaluation: Accuracy, Precision, Recall, F1-Score, Confusion Matrix

## Why This Pipeline?

Logistic Regression with L2 regularization was chosen for its effectiveness with high-dimensional data (7,129 features vs 72 samples). L2 regularization prevents overfitting by penalizing large coefficients, making it ideal for genomic data where features far exceed samples. StandardScaler ensures all genes are on the same scale, preventing bias toward genes with larger expression values.

## Performance Metrics

## Key Insights

The model achieves 94.1% accuracy on the test set with minimal overfitting (only 2.7% gap from training). Strong performance despite high dimensionality validates the choice of L2-regularized logistic regression for genomic classification tasks.

# 2. Asthma Prediction

## Problem Statement

Predict asthma diagnosis using clinical and environmental features from 1,000 patient records. The dataset includes 20+ features such as age, BMI, smoking status, air quality index, family history, and respiratory symptoms.

## Pipeline Architecture

1. Data Preprocessing: StandardScaler for numerical features  
2. Model Selection: Logistic Regression vs Decision Tree  
3. Hyperparameter Optimization: GridSearchCV with 5-fold cross-validation  
4. Best Model: Logistic Regression (C=1.0, L2 penalty, saga solver)  
5. Evaluation: Accuracy, ROC-AUC, Precision, Recall, F1-Score

## Why This Pipeline?

GridSearchCV was employed to systematically compare Logistic Regression and Decision Tree classifiers with various hyperparameters. Logistic Regression outperformed Decision Tree (94.5% vs 91.5% test accuracy) and demonstrated better generalization with a smaller train-test gap. The saga solver handles L1/L2 penalties efficiently for large datasets.

## Performance Metrics

## Key Insights

Logistic Regression achieves excellent ROC-AUC of 0.985, indicating strong discriminative ability. The minimal overfitting (0.9% gap) suggests good generalization. GridSearchCV validated that simpler models outperform complex models for this clinical dataset.

# 3. DNA Sequence Classification

## Problem Statement

Multi-task classification of DNA sequences: (1) Classify sequences into 3 functional categories, and (2) Predict disease risk levels (Low/Medium/High). Dataset contains 1,000 DNA sequences with k-mer features, GC content, sequence length, and dinucleotide frequencies.

## Pipeline Architecture

1. Feature Engineering: K-mer extraction, GC content, dinucleotide frequencies  
2. Algorithm: Random Forest Classifier (dual models for two tasks)  
3. Validation: Train-test split (80-20)  
4. Feature Analysis: Feature importance ranking  
5. Evaluation: Accuracy, F1-Score, Confusion Matrix for both tasks

## Why This Pipeline?

Random Forest was chosen for its ability to handle complex, non-linear relationships in sequence data and provide interpretable feature importance scores. The multi-task approach allows simultaneous classification and risk prediction. K-mer features capture local sequence patterns, while GC content and dinucleotide frequencies provide global sequence characteristics.

## Performance Metrics

## Key Insights

GC content emerged as the most important feature (0.245 importance score), followed by CG dinucleotide frequency (0.189). The multi-task approach successfully handles both classification tasks with over 87% accuracy, demonstrating Random Forest's versatility for sequence analysis.

# 4. Brain Tumor Classification (5-Class)

## Problem Statement

Classify brain tissue samples into 5 categories: Ependymoma, Glioblastoma, Medulloblastoma, Pilocytic Astrocytoma, and Normal tissue. Dataset contains 130 samples with 54,675 gene expression features from microarray data, presenting extreme dimensionality and class imbalance challenges.

## Pipeline Architecture

1. Dimensionality Reduction: SelectKBest (54,675 → 500 features)  
2. Feature Selection: ANOVA F-statistic for gene ranking  
3. Algorithm: Random Forest Classifier (100 estimators)  
4. Validation: Stratified train-test split to handle imbalance  
5. Evaluation: Accuracy, Per-class Precision/Recall/F1, Confusion Matrix

## Why This Pipeline?

SelectKBest with ANOVA F-statistic reduces dimensionality by 99% (54,675 → 500 genes) while retaining the most discriminative features, making the model tractable and interpretable. Random Forest handles multi-class classification naturally and is robust to class imbalance. Stratified splitting ensures all tumor types are represented in both train and test sets.

## Performance Metrics

## Key Insights

Overall test accuracy of 88% demonstrates effective multi-class classification despite high dimensionality. Normal tissue achieves the highest F1-score (0.94), while Pilocytic Astrocytoma is most challenging (0.80) due to limited samples. Dimensionality reduction improved both performance and computational efficiency.

# 5. Breast Cancer Histological Type Classification

## Problem Statement

Classify breast cancer patients into Ductal or Lobular histological subtypes using multi-omic data (gene expression, clinical features, molecular markers). Dataset contains ~900 patients with severe class imbalance (Ductal: 81.4%, Lobular: 18.6%).

## Pipeline Architecture

1. Dimensionality Reduction: PCA (95% variance retention with 18 components)  
2. Class Imbalance Handling: SMOTE (resampling ratio: 0.75, k\_neighbors: 5)  
3. Feature Scaling: StandardScaler (applied before PCA)  
4. Algorithm: Logistic Regression with L2 regularization  
5. Validation: Stratified train-test split (80-20)  
6. Evaluation: Accuracy, ROC-AUC, Precision, Recall, Confusion Matrix

## Why This Pipeline?

PCA reduces dimensionality while retaining 95% of variance, addressing the curse of dimensionality and reducing noise. SMOTE generates synthetic minority class samples to address the 81:19 class imbalance, preventing the model from being biased toward Ductal classification. Logistic Regression provides probabilistic outputs and interpretable coefficients. The combination of PCA + SMOTE + Logistic Regression creates a robust pipeline for imbalanced, high-dimensional medical data.

## Performance Metrics

## Key Insights

Test accuracy of 91.2% with minimal overfitting (2.3% gap) demonstrates excellent generalization. SMOTE successfully balanced the dataset, achieving 86% recall for the minority Lobular class. ROC-AUC of 0.958 indicates strong discriminative ability. PCA reduced dimensionality while maintaining predictive power, with 18 components capturing 95% of variance.