Protein Expression Classification in Mice Using Machine Learning

# 1. Project Overview

This project aims to apply classical machine learning models to classify mouse samples based on protein expression levels in the cerebral cortex. The ultimate goal is to identify the model that most accurately distinguishes between experimental groups (genotypes, treatments, behaviors) and the derived biological classes.

# 2. Dataset Description

Source: Data\_Cortex\_Nuclear.csv

Samples: ~1080 mouse protein expression profiles

Features:

- 77 protein expression levels (e.g., pAKT, BRAF, etc.)

- 4 categorical attributes: Genotype, Treatment, Behavior, class

Target variable: class (binary classification: Control vs Ts65Dn)

Missing Data: Present in protein features; handled with mean imputation.

# 3. Data Preprocessing

To ensure data quality and model performance, several preprocessing steps were conducted:

- Dropped MouseID column.

- Imputed missing values using mean.

- Encoded categorical features with LabelEncoder.

- Scaled features using StandardScaler.

# 4. Feature Selection

- Applied SelectKBest with ANOVA F-test (f\_classif) to select top 50 features.

- Reduced dimensionality and improved model generalization.

# 5. Model Building & Tuning

Trained the following models:

- K-Nearest Neighbors (KNN)

- Random Forest

- Logistic Regression

- Decision Tree

**Hyperparameter tuning** was applied using GridSearchCV to optimize each model's performance.

# 6. Model Evaluation

Among all tested models, **KNN outperformed others** with the following metrics:

* **Accuracy**: 99.07%
* **F1-Score**: 99.07%

Evaluation was conducted using:

* Accuracy & F1-score comparison charts
* **Confusion Matrices** to analyze prediction performance
* Heatmaps and correlation matrices for better feature understanding

# 7. Visual Analysis

- Heatmaps of missing values and correlation matrix.

- Bar charts comparing accuracy and F1 scores.

- Confusion matrices for model evaluation.

**8. Key Outcomes & Benefits:**

-Enables more efficient analysis of Down Syndrome models through protein profiling.

-Developed a **highly accurate classifier (KNN)** suitable for biological data.

-Identifies key proteins contributing to class separability.

-Supports further **biological investigation** into genotype-treatment interactions.

# 8. Challenges Faced

- Missing Data: Imputation was essential.

- High Dimensionality: Needed feature selection.

- Close Model Scores: Required detailed metric comparison.

- Label Mapping: Careful label encoding and decoding was needed.

# 9. Conclusion

The Mice Protein Classification project demonstrates how machine learning can provide powerful tools for bioinformatics and biomedical applications. Through careful preprocessing, feature selection, and model tuning, a robust KNN classifier was developed. This model not only achieved exceptional performance metrics but also opened avenues for deeper biological understanding and practical application in healthcare R&D.