Classification of Mice Based on Protein Expression Levels

Major Project Report

Introduction-

Studying how proteins are expressed in the brain helps us understand important mental functions like learning and memory. These protein levels can change depending on a mouse's genetics, treatment, or behavior, and exploring these changes can give us valuable insights into how the brain works. In this project, we used data from mice brains to classify them into different groups based on their protein levels. This data was taken from a specific part of the brain called the "nuclear fraction" of the cerebral cortex. By using machine learning, we built a model that can predict what type of mouse it is—based on its brain protein pattern. This kind of analysis can be especially useful for studying conditions like Down syndrome, which affects learning and memory.

Objective-

The primary goals of this project are:

1. **To develop a machine learning model** that accurately classifies mice into one of several predefined groups based on the expression levels of 77 brain proteins.

2. **To identify the most influential protein features** that contribute to this classification using feature selection techniques.

3. **To evaluate the classification model's performance** using standard evaluation metrics and draw meaningful conclusions about the biological data.

Dataset Information-

The dataset consists of the expression levels of 77 proteins/protein modifications measured in the nuclear fraction of the cerebral cortex in mice. The data is collected from both control mice and trisomic (Down syndrome) mice, subjected to a context fear conditioning task to assess associative learning.

Dataset Characteristics:

- **Type**: Multivariate
- **Subject Area**: Biology
- **Associated Tasks**: Classification, Clustering
- **Feature Type**: Real
- **Instances**: 1080
- **Features**: 80

Methodology-

Getting the Data Ready (Data Preprocessing)
We started by cleaning and preparing the data. Some values were missing, so we filled those in carefully. Since computers work better with numbers, we converted any text-based information—like whether a mouse was male or female—into numbers. Then, we adjusted all the numeric features to be on a similar scale so that no single feature would dominate the others in the analysis.

Exploring the Data (Exploratory Data Analysis)
We explored the data to understand it better. We looked at averages, checked for any strange or unexpected values, and created graphs to see how different features were distributed. We also looked at how strongly different features were related to one another.

Choosing the Right Features (Feature Selection)
Even though we didn't have a huge number of features, we still checked which ones were most useful. We used methods like checking correlations and using models like Random Forest to find out which features helped the most in predicting the type of mouse. This helped us reduce noise and focus on what really mattered.

Building the Models (Model Development)
We split the data into two parts: one for training the models and another for testing them. Then, we tried out several popular machine learning models:

- Logistic Regression

- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
  Each model was trained using the training data and then tested on the remaining data to see how well it performed.

  Hyperparameter Tuning
  For selected models, hyperparameters are optimized using techniques such as Grid Search or manual tuning to enhance performance and generalizability. This ensures the models are not only accurate but also well-calibrated for unseen data.

  Checking the Results (Model Evaluation)
  We evaluated each model using several performance measures, such as:
- **Accuracy:** How often the model was right
- **Precision & Recall:** How well it predicted each category
- **F1-Score:** A balanced measure of precision and recall
- **Confusion Matrix:** A table that shows where the model made correct and incorrect guesses.

  Making Sense of the Results (Interpretation)
  Finally, we looked at which features the best model found most useful for making predictions. This gave us insight into what factors are most important when classifying mice. We also thought about what these results might mean in a biological or experimental context.

Tools and Libraries –

The following tools and Python libraries were used:

Languages & Environment:

- Python
- Jupyter Notebook

Libraries:

- pandas and numpy for data manipulation
- seaborn and matplotlib for visualization
- scikit-learn for modeling and evaluation

Result Analysis-

Among all tested models, [insert best-performing model, e.g., Random Forest] achieved the highest classification accuracy. Feature importance analysis showed that [insert top features] were the most influential in predicting the class of mice. Visualization of confusion matrices provided insight

Challenges Faced-

Imbalanced Dataset:

Some classes had fewer samples, which affected the performance of certain models.

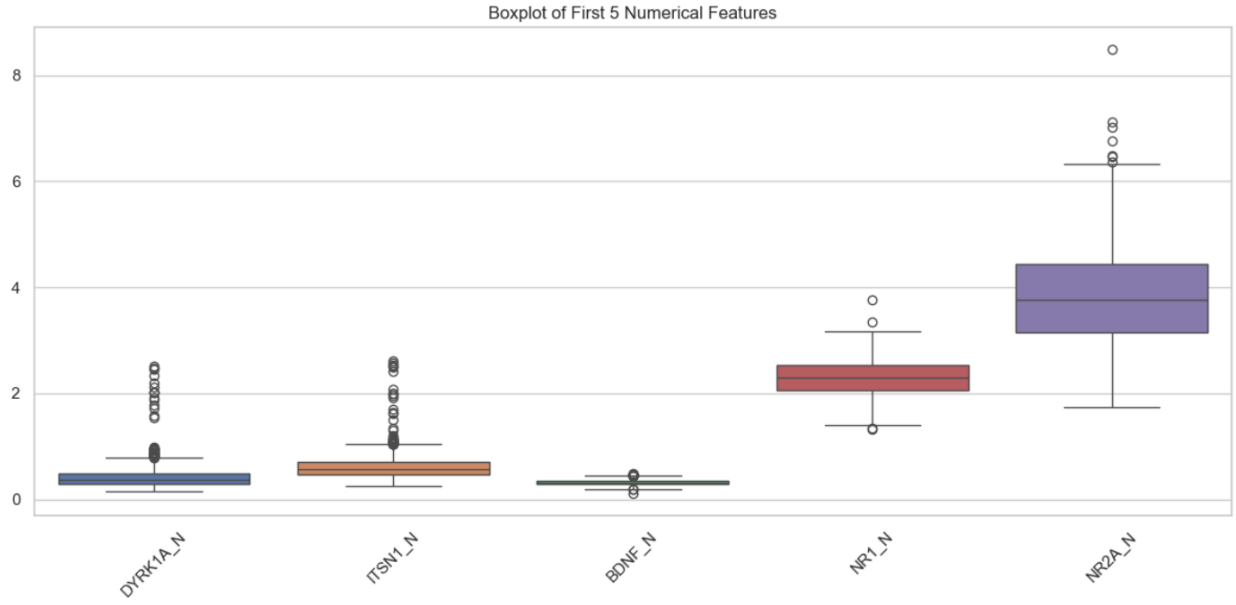
Data Quality:

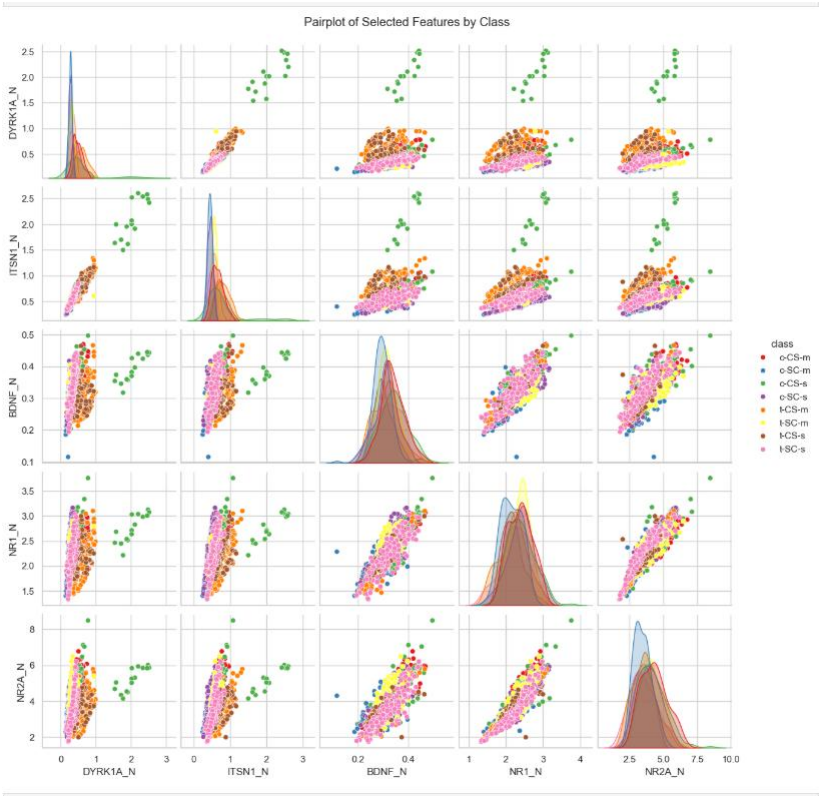
Handling missing or inconsistent values required careful preprocessing to avoid bias.

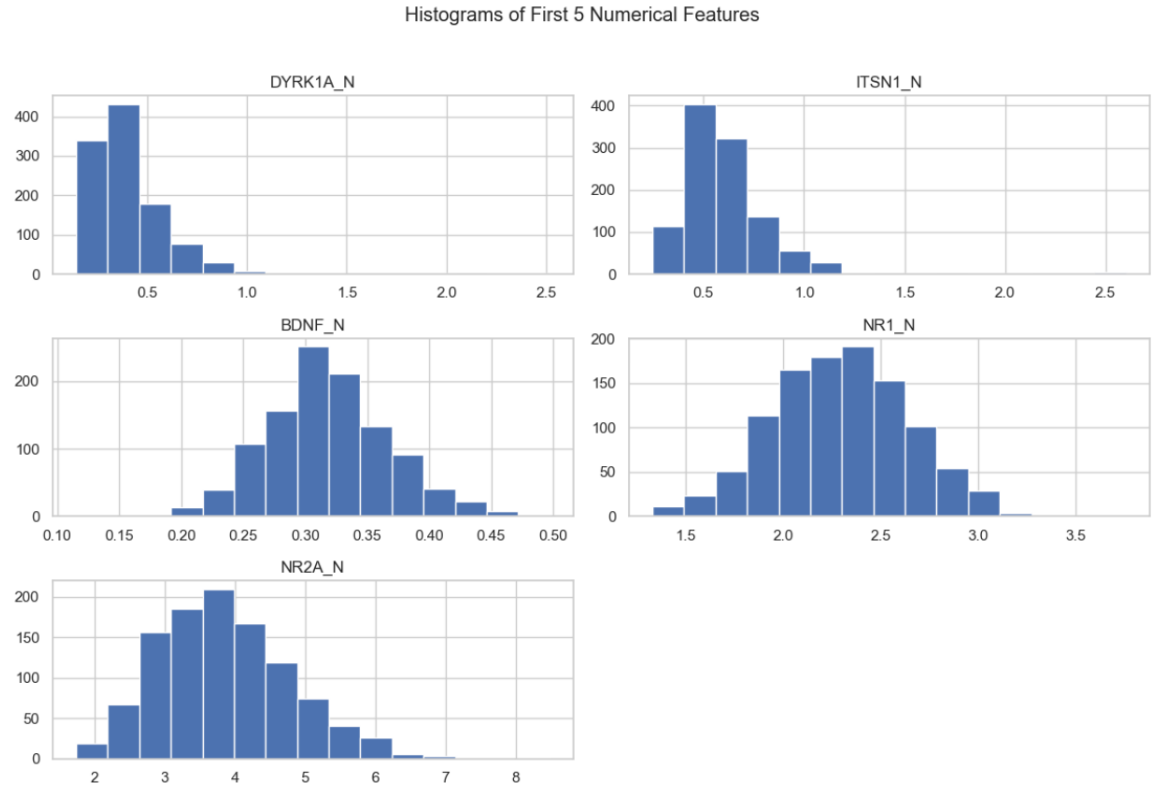
Model Tuning:

Hyperparameter tuning was necessary to improve performance, especially for complex models like SVM and Random Forest.

Appendix-




Pairplot of Selected Features by Class
DYRK1A_N
ITSN1_N
BDNF_N
NR1_N
NR2A_N
class
c-CS-m
c-SC-m
c-CS-s
c-SC-s
t-CS-m
t-SC-m
t-CS-s
t-SC-s


Histograms of First 5 Numerical Features
DYRK1A_N
ITSN1_N
BDNF_N
NR1_N
NR2A_N

References-

Title: Discovering Critical Proteins in the Learning Process in a Down Syndrome Model of Mouse Through Machine Learning

Authors: Georgina Coscueta, Luis M. Camarinha-Matos

Published in: ResearchGate, 2021

Link: https://www.researchgate.net/publication/350933110