MICE CLASSIFICATION

A Machine Learning Project

-By Sufiyan Zamindar

INTRODUCTION

This project focuses on classifying different types of mice based on the levels of protein expression in their brain tissue. Using multiple machine learning teachniques, we aim to accurately group mice into 8 biological classes formed by combinations of genotype (control or trisomic), treatment, and behavior.

Using mice models, which share genetic and biological similarities with humans, researchers can identify key proteins and pathways that are altered in DS. This knowledge aids in understanding the condition and evaluating potential treatments, such as memantine, to improve cognitive function and quality of life for those with DS

OBJECTIVE

The main goal of this project is to build a machine learning model that can classify mice into one of eight biological classes based on their protein expression profiles. This classification Using machine learning on protein data allows faster, data-driven insights into how treatments affect brain functions. Identifying key proteins can help build biomarkers for early diagnosis of neurological conditions like Down Syndrome Alzheimer's Disease etc.

DATASET OVERVIEW

- What data is Used:
- The dataset is derived from a real-world neuroscience study on mice brain tissue.
- It includes measurements of protein expression levels from the hippocampus region — a key brain area related to learning and memory.
- Used widely in studies involving Down Syndrome models (Ts65Dn mice).
-

Features:

- Each mouse belongs to one of the 8 classes, such as:

1. c-CS-m: Control genotype, Context-Shock behavior, Memantine treated
2. t-SC-s: Trisomic genotype, Shock-Context behavior, Saline treated

- Each protein column shows the expression level of a specific brain protein.

STEPS INVOLVED:

1) Data Preprocessing

2) Exploratory Data Analysis (EDA)

3) Feature Selection

4) Model Training

5) Model Evaluation

6) Interpretation & Insights

7) Conclusion

TECNOLOGY USED:

1) Python 3.13

2) Pandas, Numpy & Scikit learn

3) Matplotlib & Seaborn

4) Jupyter Notebook

5) Models likeKNN, Logistic Regression etc.

6) Github, Streamlit

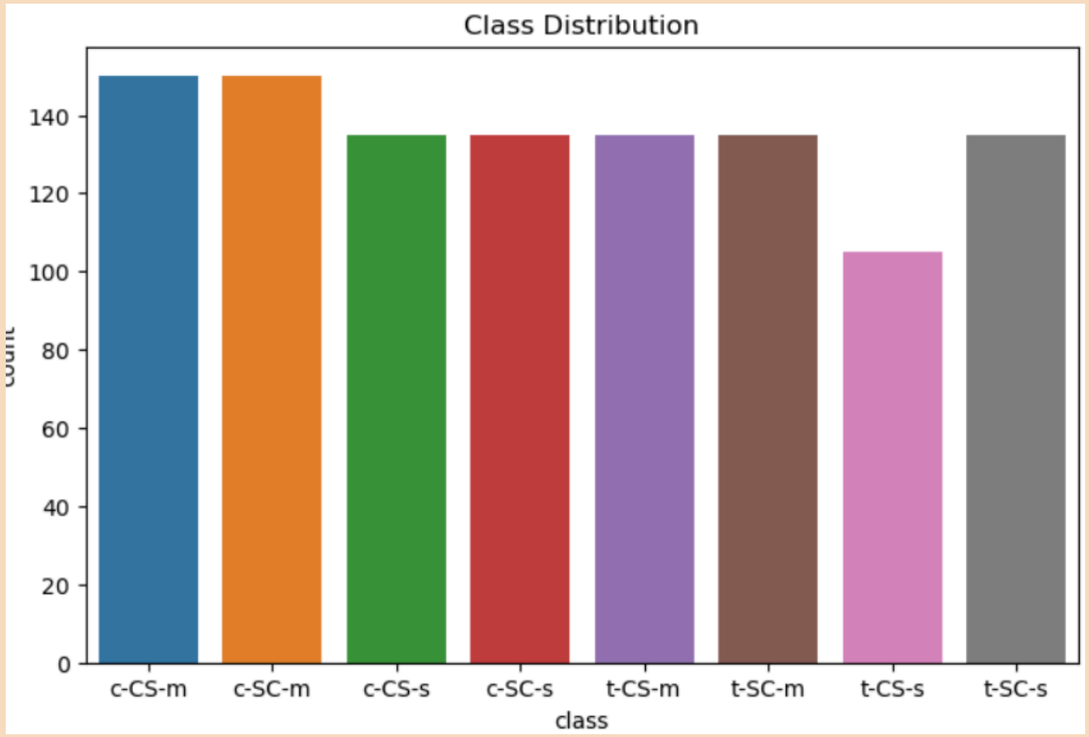
DATA PREPROCESSING

Data preprocessing is the process of transforming raw data into a useful, understandable format, resolving issues like inconsistent formatting, human errors, and incompleteness. It's crucial for data mining and machine learning projects, affecting the success and performance of machine learning models by making data completer and more efficient for analysis. Steps Involved:

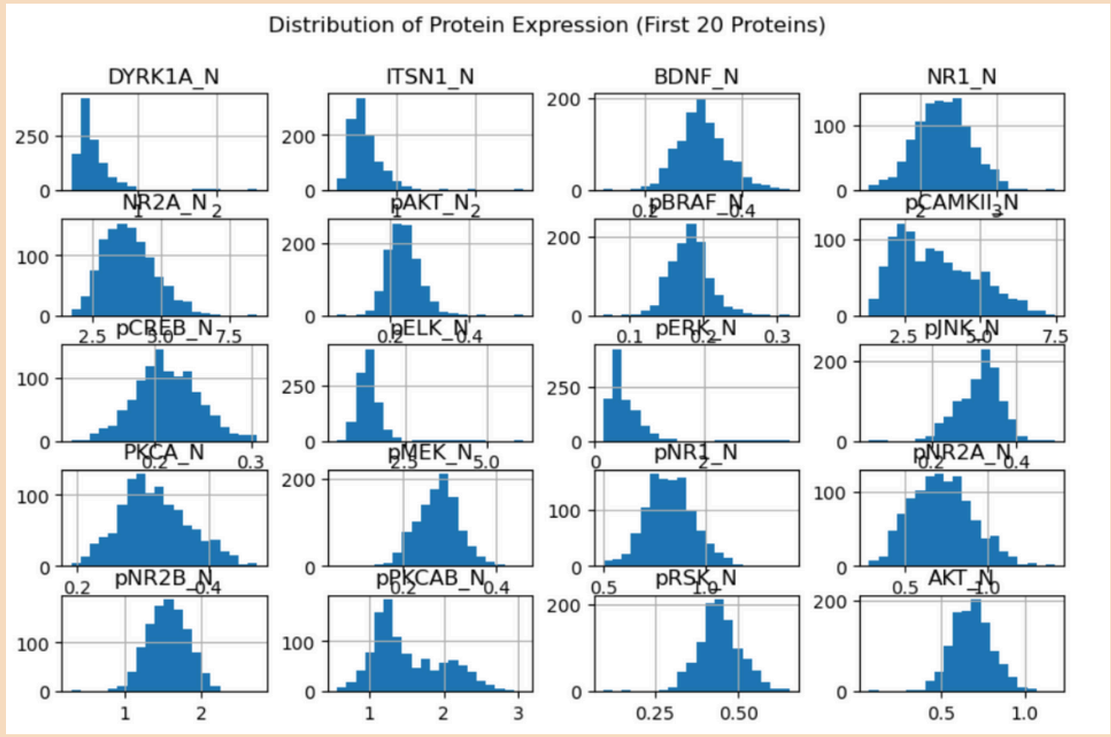
- Handling Missing Values
- Detecting Outliers
- Normalization / Scaling
- Encoding Categorical Variable

EXPLORATORY DATA ANALYSIS (EDA)

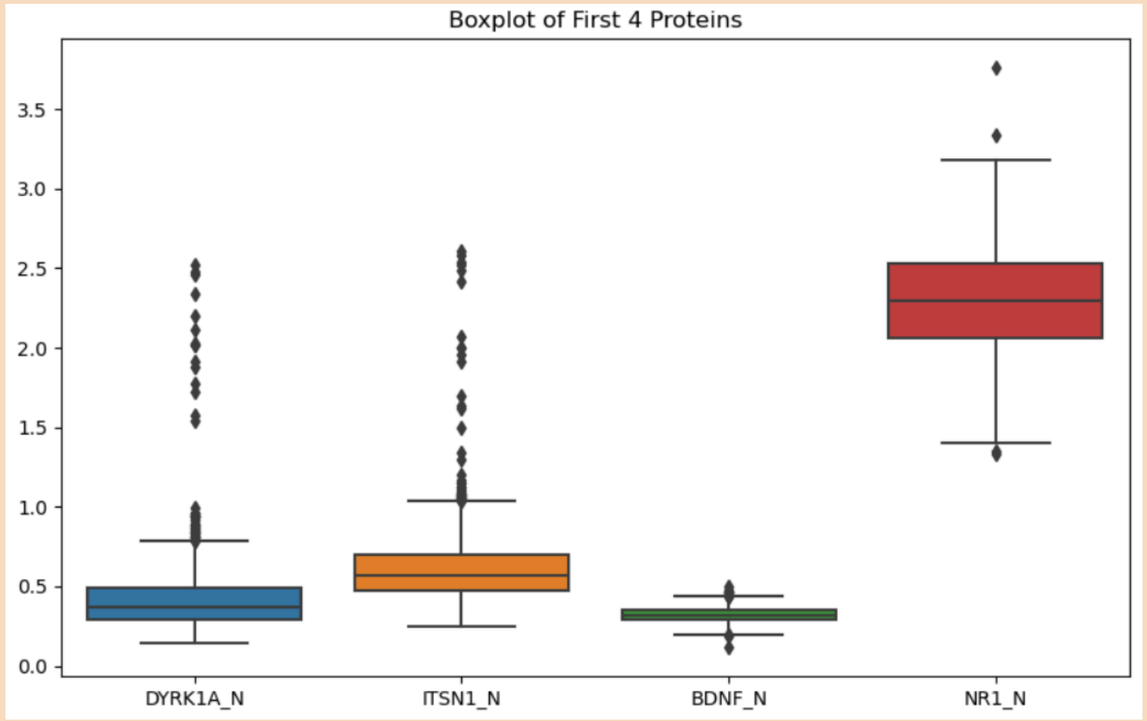
A. Analysis of Class Distribution:



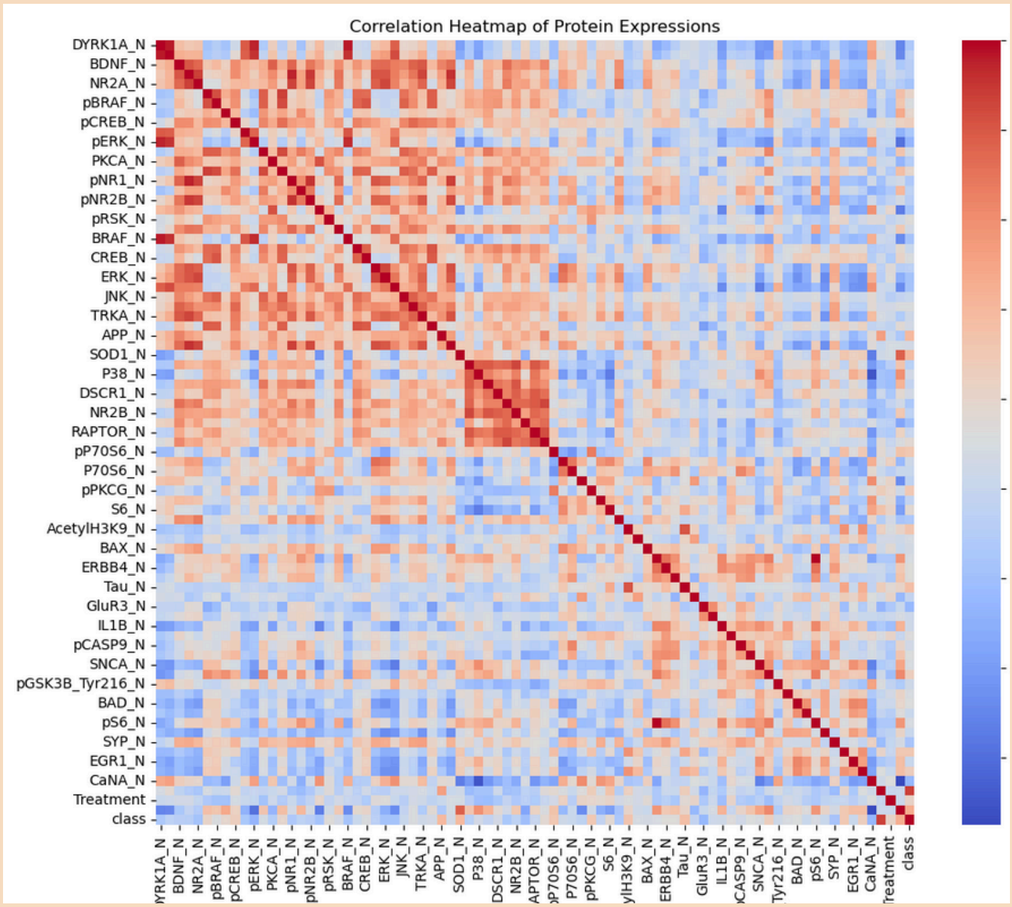
B. Analysis of Protien Distribution:



C. Boxplot with Outliers



D. Correlation Heatmap of Protien Expression



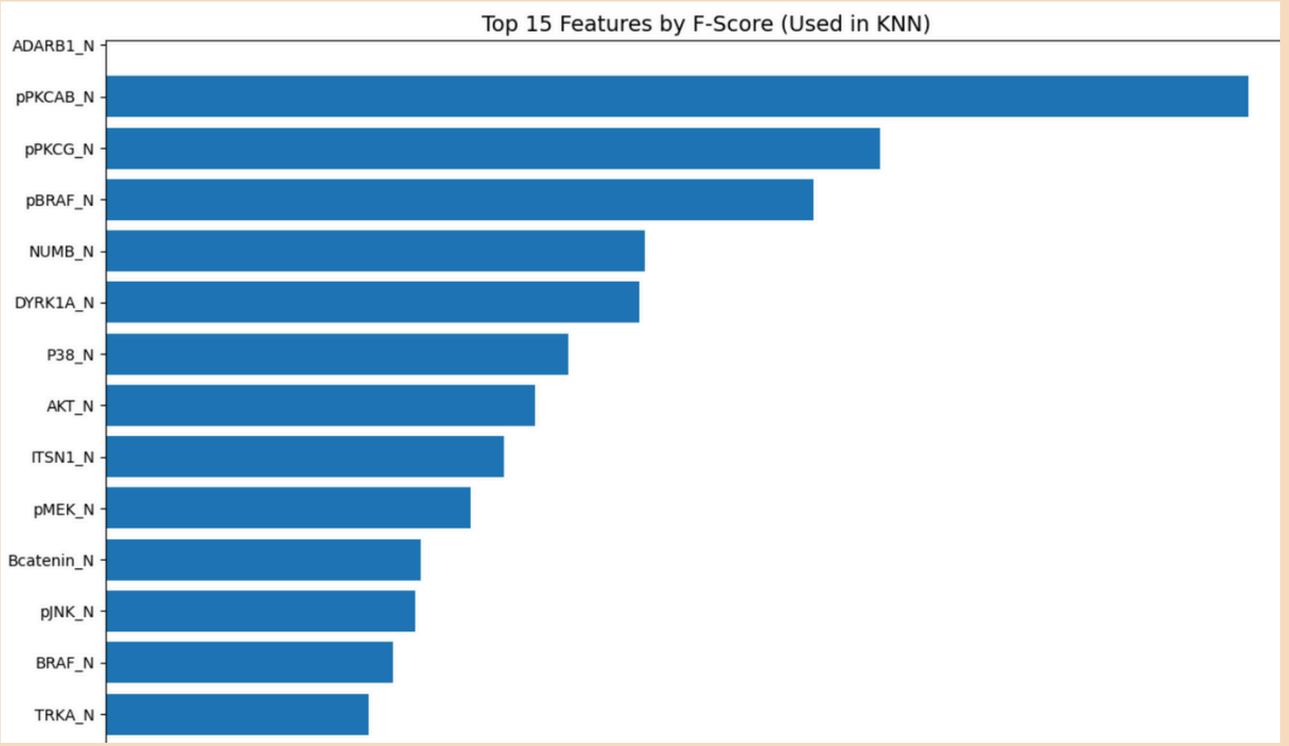
FEATURE SELECTION

Evaluating and selecting the most relevant features for our machine learning model is crucial for building effective models.

For Example:

Top 5 Proteins Identified:

1. CaNA_N
2. pPKCG_N
3. Ubiquitin_N
4. ARC_N
5. Tau_N



MODEL TRAINING

Model training process involves selecting appropriate machine learning algorithms, splitting the dataset into training and testing sets, tuning hyperparameters, and training the models. In this project, several models were evaluated to find the best-performing one for classifying mice based on protein expression levels

Various libraries were imported from scikit-learn that included LogisticRegression, KNearestNeighbors, RandomForestClassifier, Decision Tree were used to train the model to get maximum accuracy and precision. KNN provided the best results with maximum accuracy of 99%, precision of 99% and f1-score of 100%

MODEL EVALUATION

K-Nearest Neighbors (KNN)

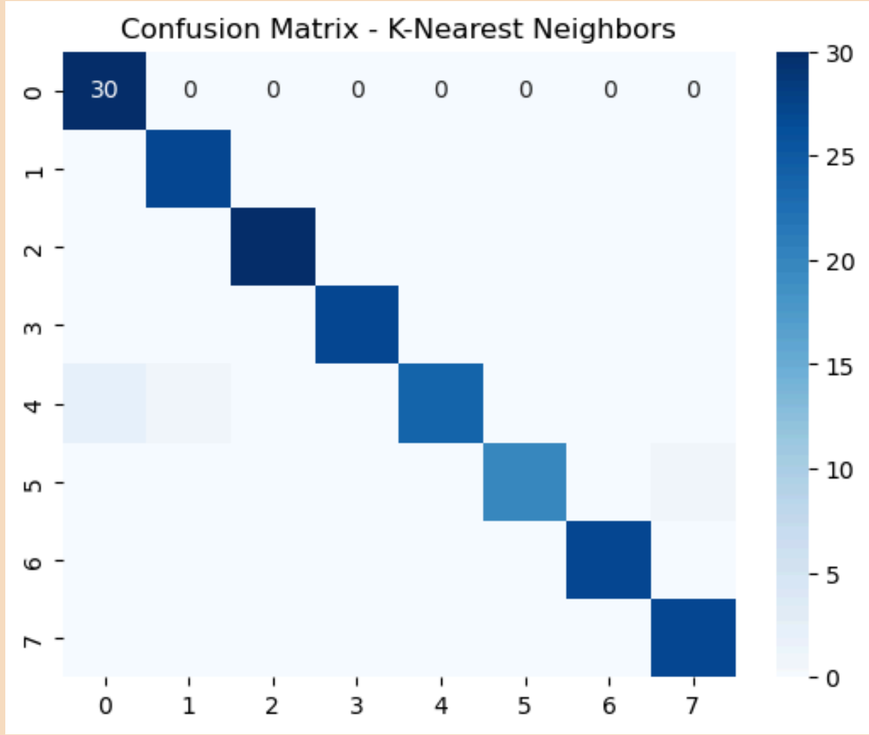
- Selected after hyperparameter tuning and cross-validation
- Best performing in terms of accuracy and consistency

Test Accuracy: 0.9953703703703703

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	30
1	1.00	0.96	0.98	27
2	1.00	1.00	1.00	30
3	1.00	1.00	1.00	27
4	1.00	1.00	1.00	27
5	0.95	1.00	0.98	21
6	1.00	1.00	1.00	27
7	1.00	1.00	1.00	27
accuracy			1.00	216
macro avg	0.99	1.00	0.99	216
weighted avg	1.00	1.00	1.00	216

Confusion Matrix:



CHALLENGES FACED

1) Handling High-Dimensional Data:

High dimensionality increases the risk of overfitting and reduces model generalization.

2) Missing Data

Several proteins had incomplete measurements which then impute by median.

3) Label Encoding Sensitivity

Mislabeling or improper encoding could lead to incorrect patterns or model bias.

4) Model Selection

Chose KNN for its high accuracy and simplicity, tuned using GridSearchCV.

5) Lack of Domain Expertise

Understanding the biological importance of proteins and interpreting them correctly requires biomedical knowledge.

INTERPRETATION

- Protein expression varies significantly across genotype and treatment groups.
- Top 5 proteins play a key role in memory and learning mechanisms.
- Further biological validation of these proteins could lead to breakthroughs in neuroscience research.
- The model achieves strong class separation with minimal features.
- This implies that protein expression data in this case is well-separated and clusterable, making instance-based learning highly effective.
- The confusion matrix showed very few misclassifications, indicating a high generalization capability.

KEY TAKEAWAYS

- 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.
- Supports preclinical research (Biomedical diagnosis, drug testing, neurobiology research) and treatment evaluation.
- Could be extended to human studies or other genetic conditions.

CONCLUSION

The Mice Protein Classification project successfully demonstrated how machine learning techniques can be effectively applied to high-dimensional biomedical datasets to derive meaningful and accurate classifications.

By focusing on protein expression levels in the cerebral cortex of mice, the study aimed to distinguish between control and Ts65Dn (Down syndrome model) samples. Through a combination of robust data preprocessing, careful feature selection, and model optimization, the K-Nearest Neighbors (KNN) algorithm emerged as the most effective model, achieving an impressive accuracy and F1-score of 99.07%.

Overall, this classification model has the potential to be extended and generalized to similar datasets involving genetic or neurological research.Moreover, the project revealed biologically relevant patterns in protein expression that could serve as biomarkers in future experimental and clinical studies.

THANK YOU

-By Sufiyan Zamindar