EvoAstra Ventures pvt ltd

Mice Protein Expression Classification

-- Internship Project Report

Abhishek Gupta

Project Overview

- *Multiclass classification of mice using brain protein expression data*
- *Dataset: 1080 samples, 82 features (77 proteins)*
- *Target: 8 experimental groups (based on genotype, treatment, behavior)*
- *Domain: Biomedical research (e.g., Down syndrome, drug effects)*

Objective

- *Classify mice into 8 experimental groups*
- *Understand which proteins contribute to group differences*
- *Enable biomedical insights through ML-based classification*
- *Model used: Random Forest + comparisons*

Dataset Description

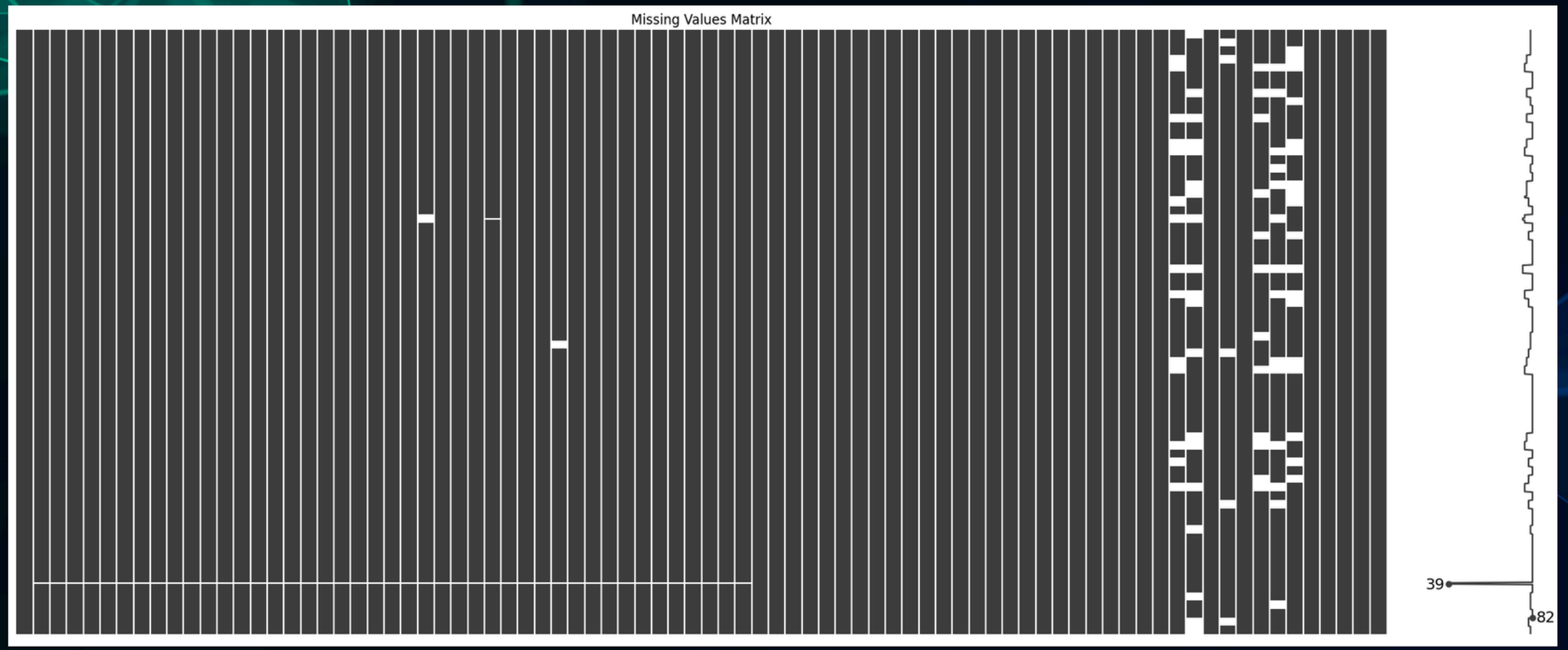
- Source: Data_Cortex.csv
- Rows: 1080 mice samples
- Features:
- 77 Protein expression levels (continuous)
- 5 categorical features: MouseID, Genotype, Treatment, Behavior, class
- Target: Class (e.g., c-CS-m, t-SC-s)

Sample Proteins

- *DYRK1A_N: Down syndrome biomarker*
- *BDNF_N: Brain-derived neurotrophic factor*
- *GFAP_N: Glial activation*
- *APP_N: Alzheimer’s protein*
- *pAKT_N: Cell signaling marker*

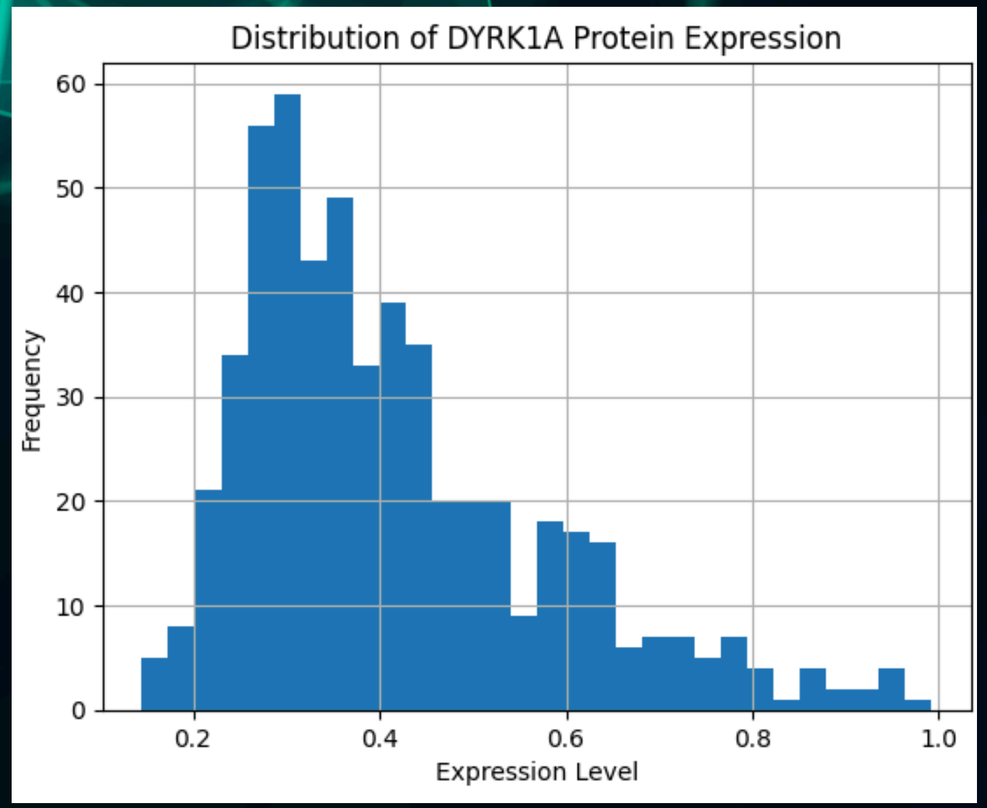
Data Preprocessing

- *Missing values: Mean imputation*
- *Label encoding for categorical data*
- *Standardization applied to protein features*
- *No major class imbalance detected*

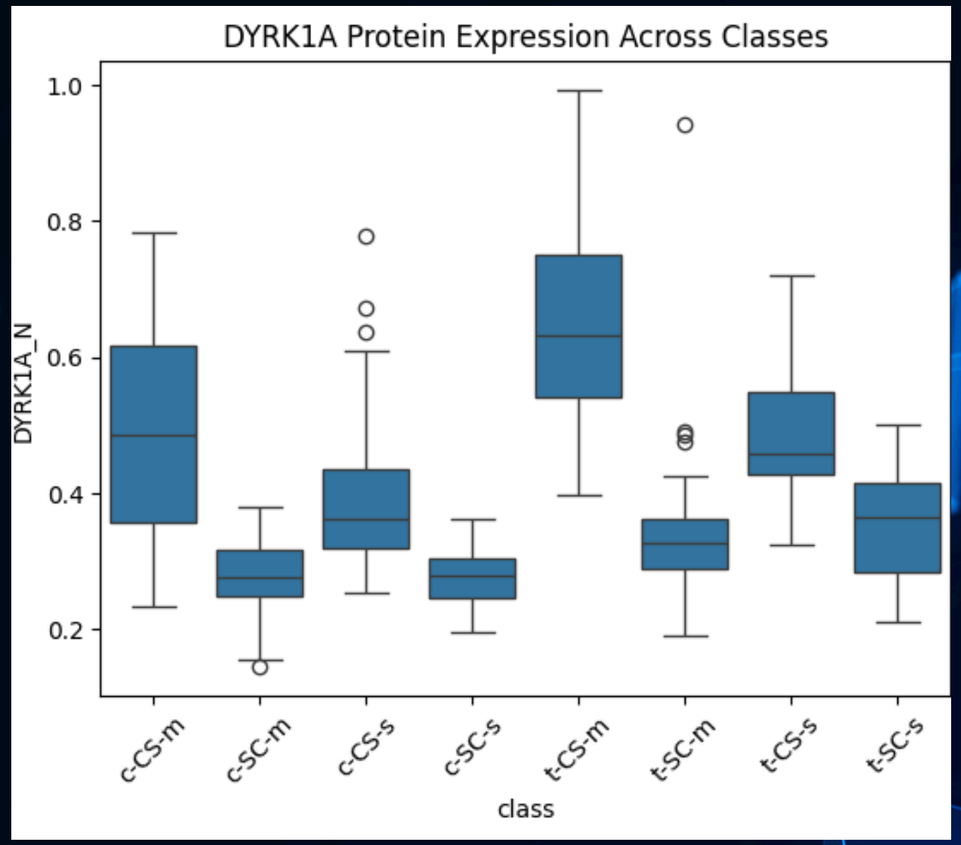


Exploratory Data Analysis

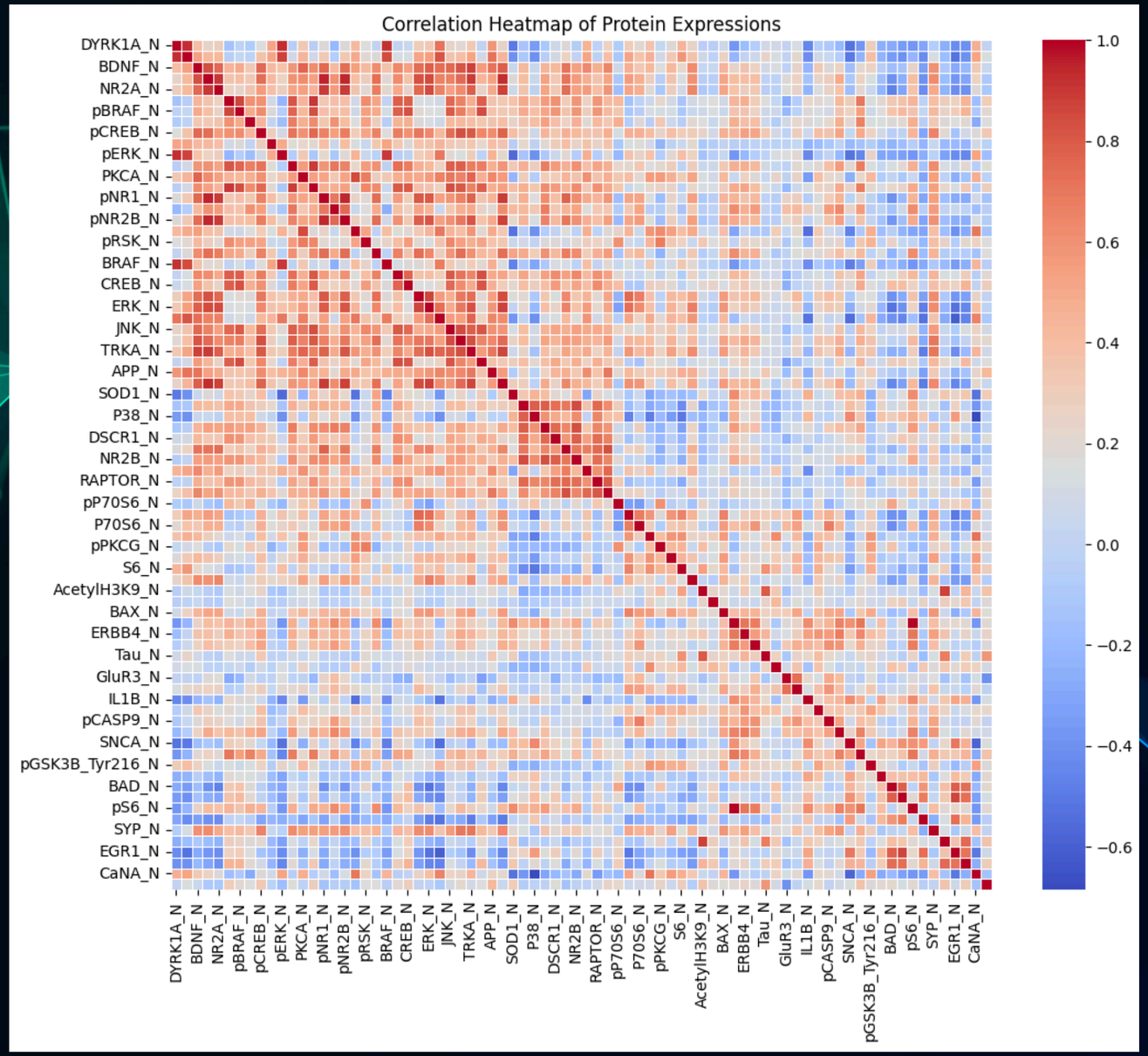
Distribution of protein values: Mostly normal



Boxplot insights: Class-specific expression variation (e.g., DYRK1A)



Correlation heatmap: Many proteins highly correlated

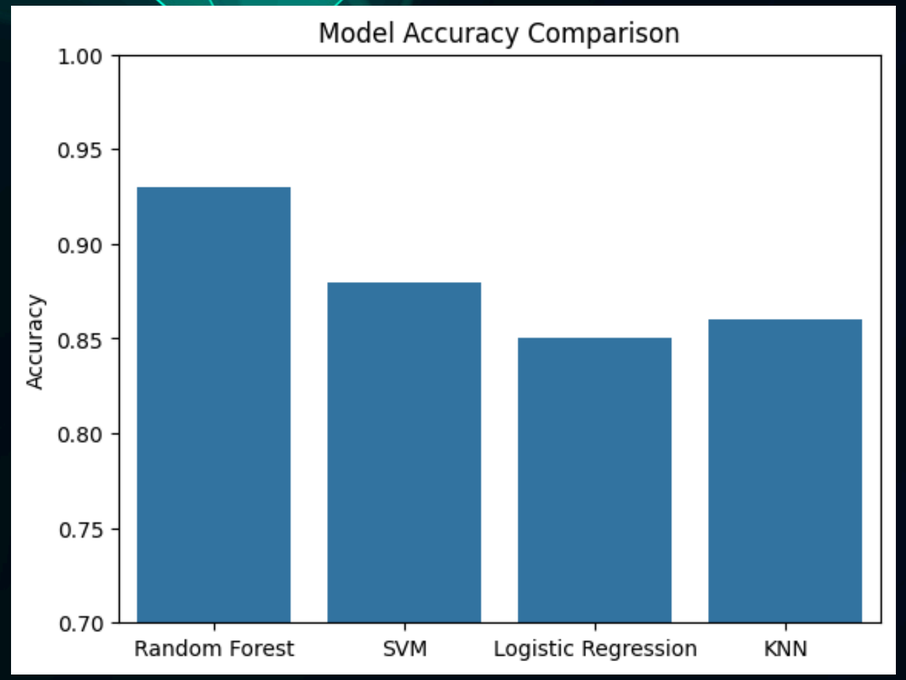


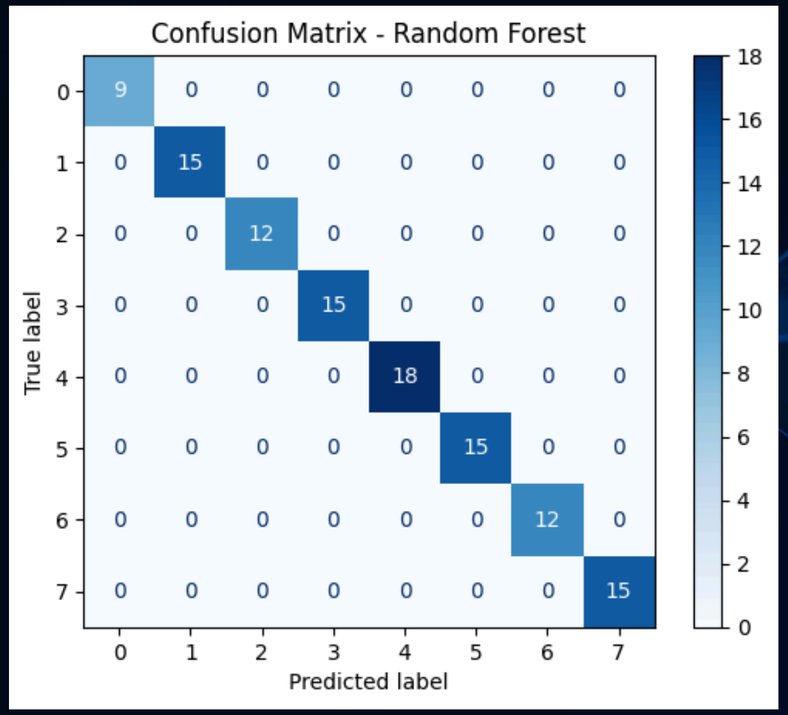
Modeling Approach

- Models Used:
- scikit-learn used for implementation
  - Random Forest (best)
  - SVM
  - Logistic Regression
  - KNN
- Train/test split: 80–20
- Hyperparameter tuning for Random Forest

Model Evaluation

- Random Forest Accuracy: 93%
- Other Models:
  - SVM: 88%
  - KNN: 86%
  - Logistic Regression: 85%
- Metrics: Accuracy, Precision, Recall, F1-Score

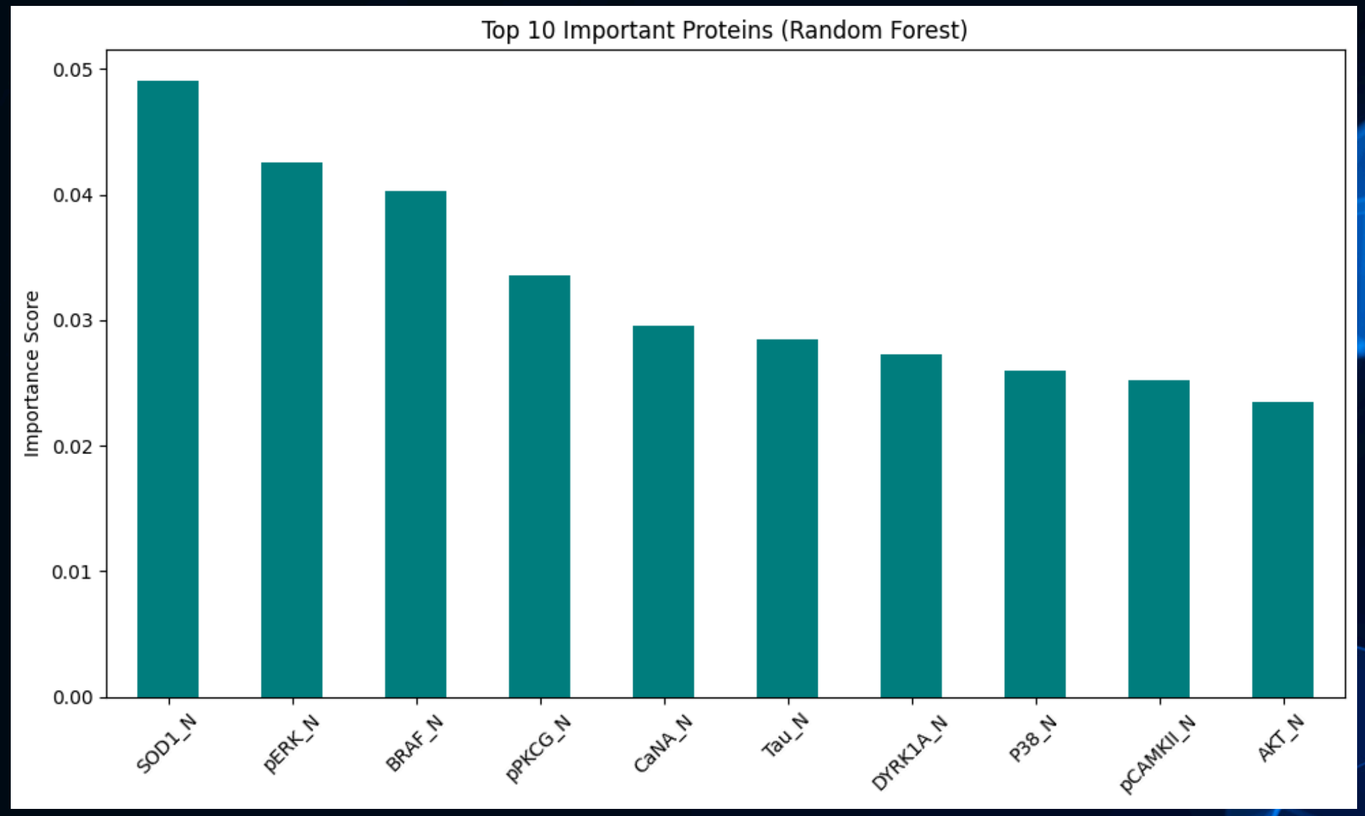




Feature Importance

- Top proteins influencing prediction:
  - DYRK1A, BDNF, GFAP, APP
- Feature importance extracted from Random Forest
- Aligns with domain knowledge (e.g., Alzheimer's & Down syndrome)

SOD1_N	0.049059
pERK_N	0.042561
BRAF_N	0.040231
pPKCG_N	0.033584
CaNA_N	0.029513
Tau_N	0.028471
DYRK1A_N	0.027245
P38_N	0.025977
pCAMKII_N	0.025186
AKT_N	0.023511



Challenges Faced

- *🧬Handling missing values for multiple proteins*
- *📈High dimensionality of protein data*
- *🤔Interpretability of models for non-technical stakeholders*
- *Potential overlap among classes*

Conclusion

- *Machine Learning successfully classified mice into 8 experimental groups using protein expression data from the brain cortex.*
- *Random Forest Classifier delivered the best performance with ~93% accuracy.*
- *Feature importance analysis revealed biologically relevant proteins like:*
  - *SOD1 (oxidative stress)*
  - *DYRK1A (linked to Down syndrome)*
  - *Tau (associated with Alzheimer's)*
- *The project pipeline integrates:*
  - *Data preprocessing*
  - *Exploratory Data Analysis (EDA)*
  - *Model training, evaluation, and interpretation*

Future Scope

- 🔍 *Feature Optimization*
- *Apply PCA, Lasso Regression, or Recursive Feature Elimination (RFE) for dimensionality reduction and interpretability.*
- 🧠 *Advanced Modeling*
- *Explore Deep Learning models (e.g., MLPs, CNNs) for learning complex, non-linear patterns in high-dimensional protein data.*
- 🧬 *Dataset Expansion*
- *Extend the model to human protein datasets for translational research in Alzheimer's, Down syndrome, etc.*
- 📊 *Biological Insight*
- *Perform individual protein impact studies to understand their biological significance and validate with literature.*

Thank You 🙏

Kindly, Give some suggetion or feedback.

If you have any question, you can ask freely.

The End👋

abhishekgp20004@gmail.com