🧪 Mice Protein Expression Classification – Project Report

---

1. Project Title

Mice Protein Expression Classification using Machine Learning

2. Objective

The objective of this project is to develop a machine learning model that classifies mice into one of eight experimental groups based on protein expression levels from the cortex region of the brain.

These classifications are based on:

- Genotype
- Treatment
- Behavioral conditions

This aids biomedical research by analyzing the effects of Down syndrome, drug treatments, and behavioral patterns at the molecular level.

3. Dataset Description

- Dataset: Data_Cortex.csv
- Samples: 1080 rows
- Features: 82 columns
- ✓ 77 numerical protein expression features
- ✓ 5 categorical variables: **MouseID, Genotype, Treatment, Behavior, Class**

Examples of Proteins:

- **DYRK1A_N**: Linked to Down syndrome
- **BDNF_N**: Neurotrophic factor
- **GFAP_N**: Glial activation marker
- **APP_N**: Alzheimer's associated
- **pAKT_N**: Cell signaling

4. Target Variable Explanation

The class column is the target label formed by combining genotype, treatment, and behavior.

Example class labels:

- c-CS-m
- t-SC-s
- c-SC-s
- t-CS-m

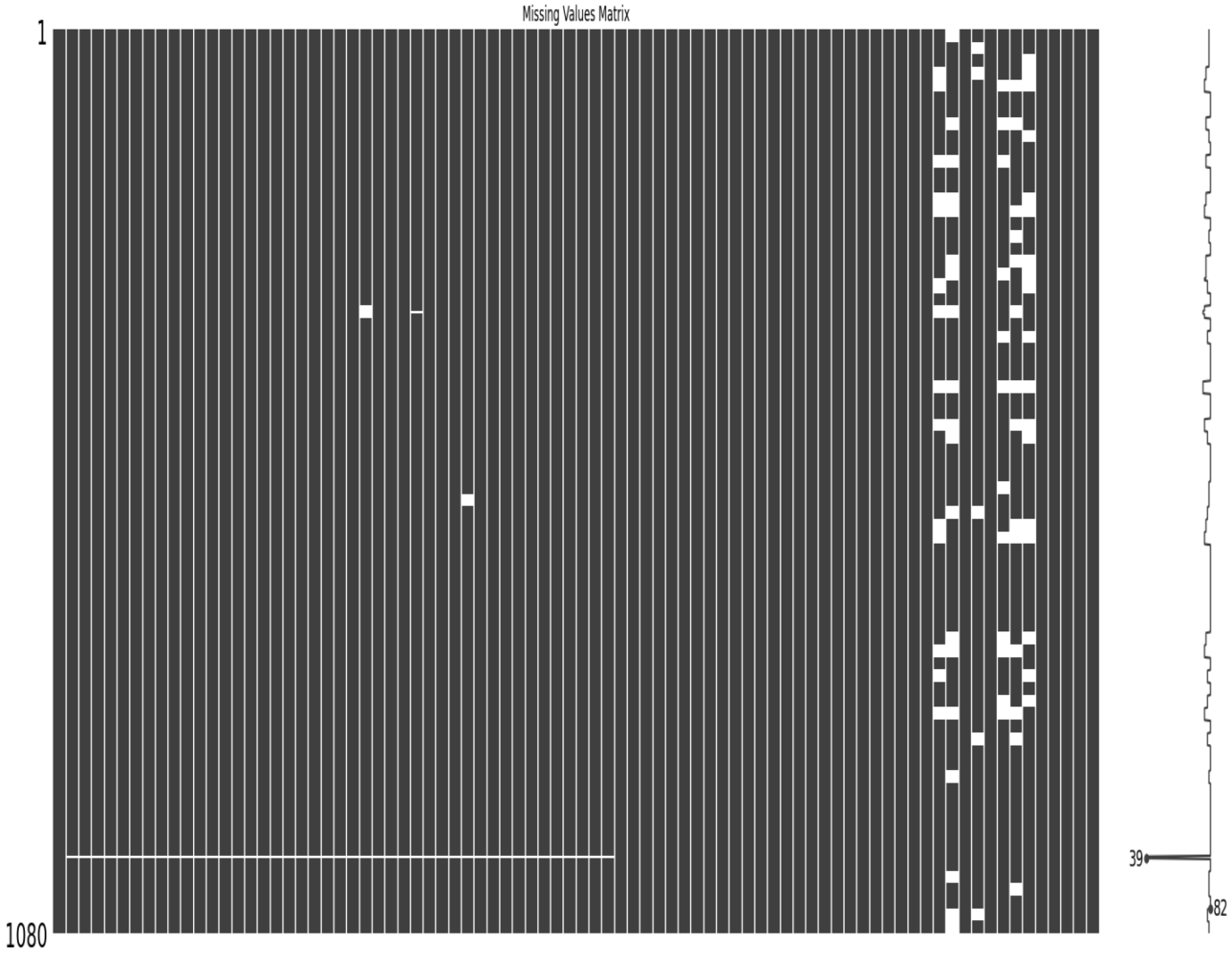
5. Data Preprocessing

Steps Applied:

- Missing Value Handling: Mean imputation used for missing protein expressions
- Label Encoding: Applied to categorical variables
- Standardization: All protein features were scaled to zero mean and unit variance

Missing Values Matrix

Purpose: Shows presence of missing values before imputation.



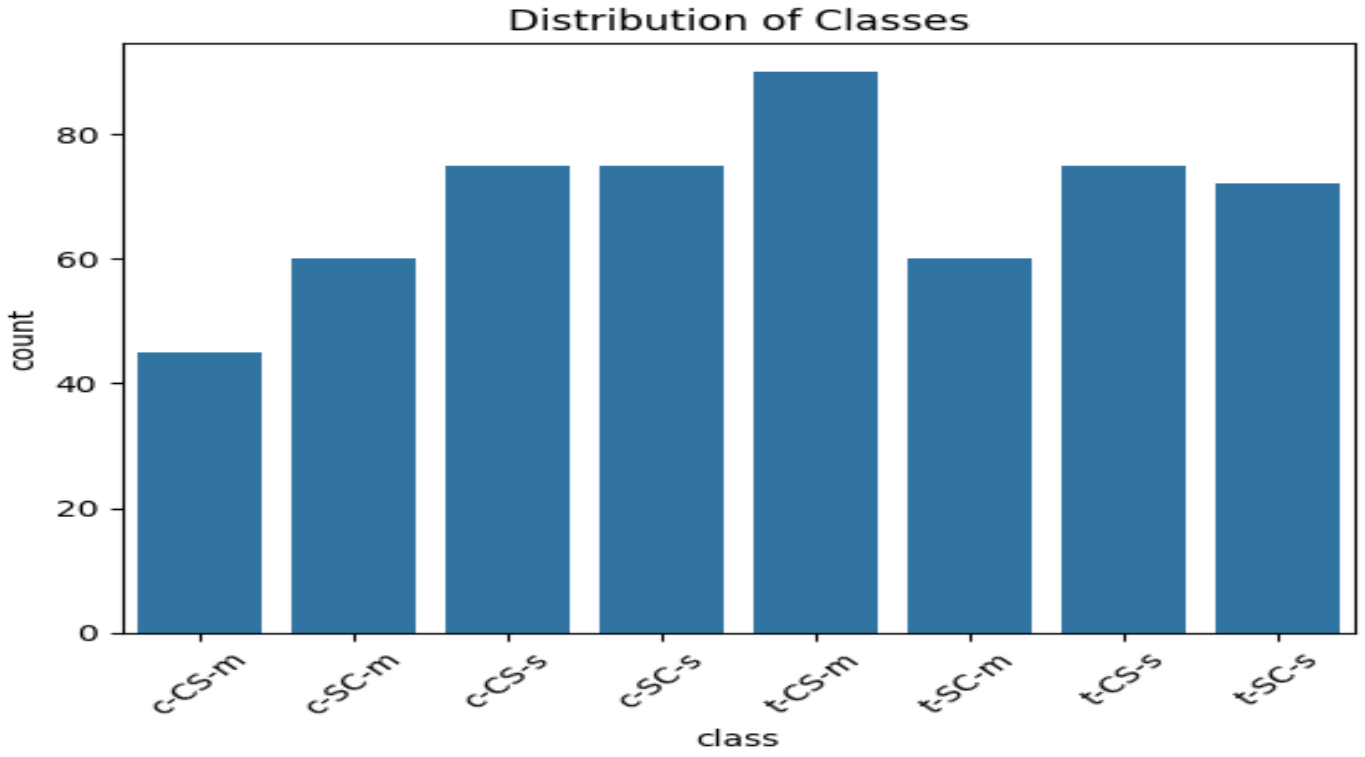
This matrix reveals a few proteins with missing values. Mean imputation was used to preserve dataset integrity.

6. Exploratory Data Analysis (EDA)

6.1 Class Distribution

- ***Count Plot of Classes***

Purpose: Shows how balanced the classes are.

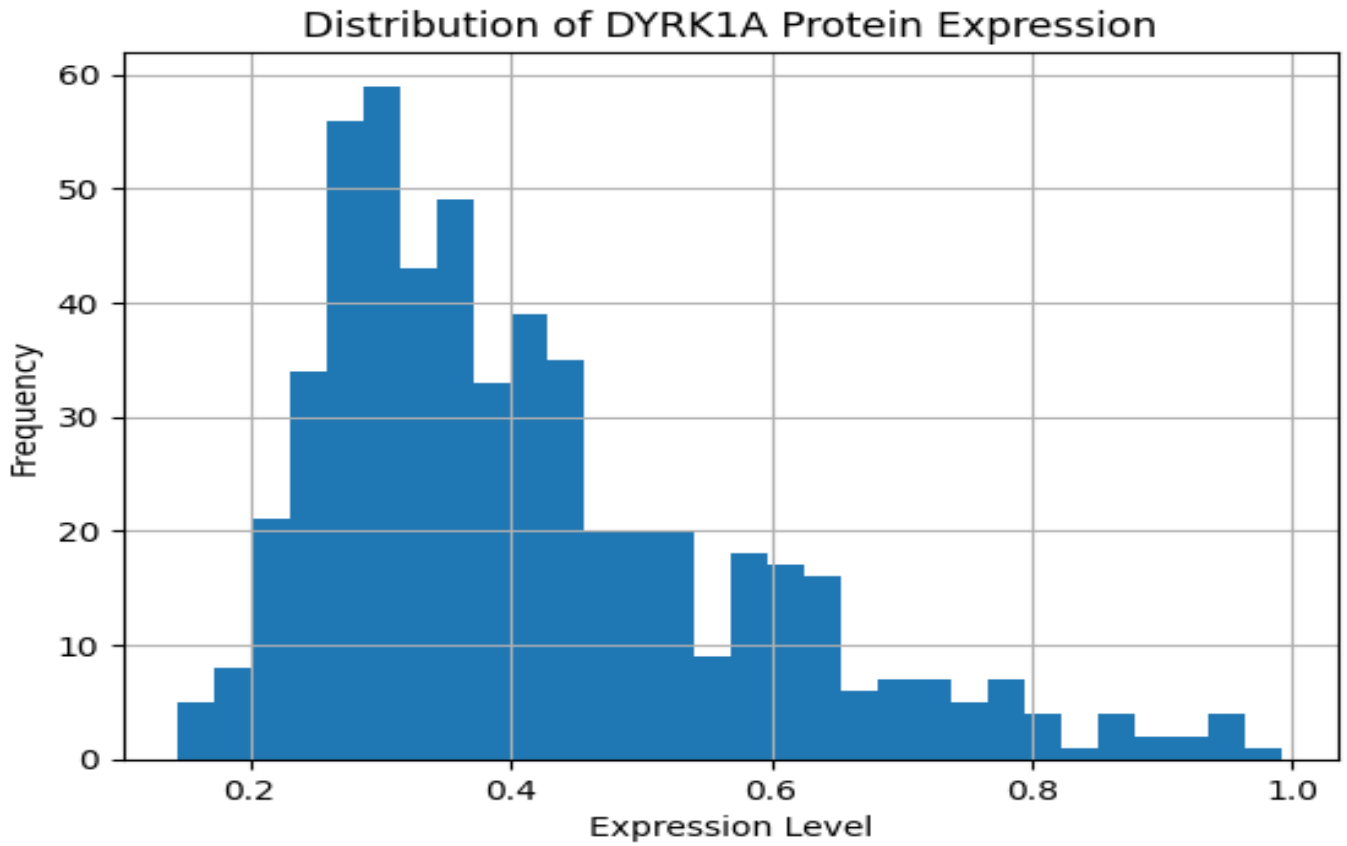


The class distribution is imbalanced, with 'Healthy' being the most frequent class. This can affect model performance, especially for minority classes.

6.2 Protein Expression Distribution

- ***Histogram of DYRK1A_N***

Purpose: *Understand how protein expression levels are distributed.*

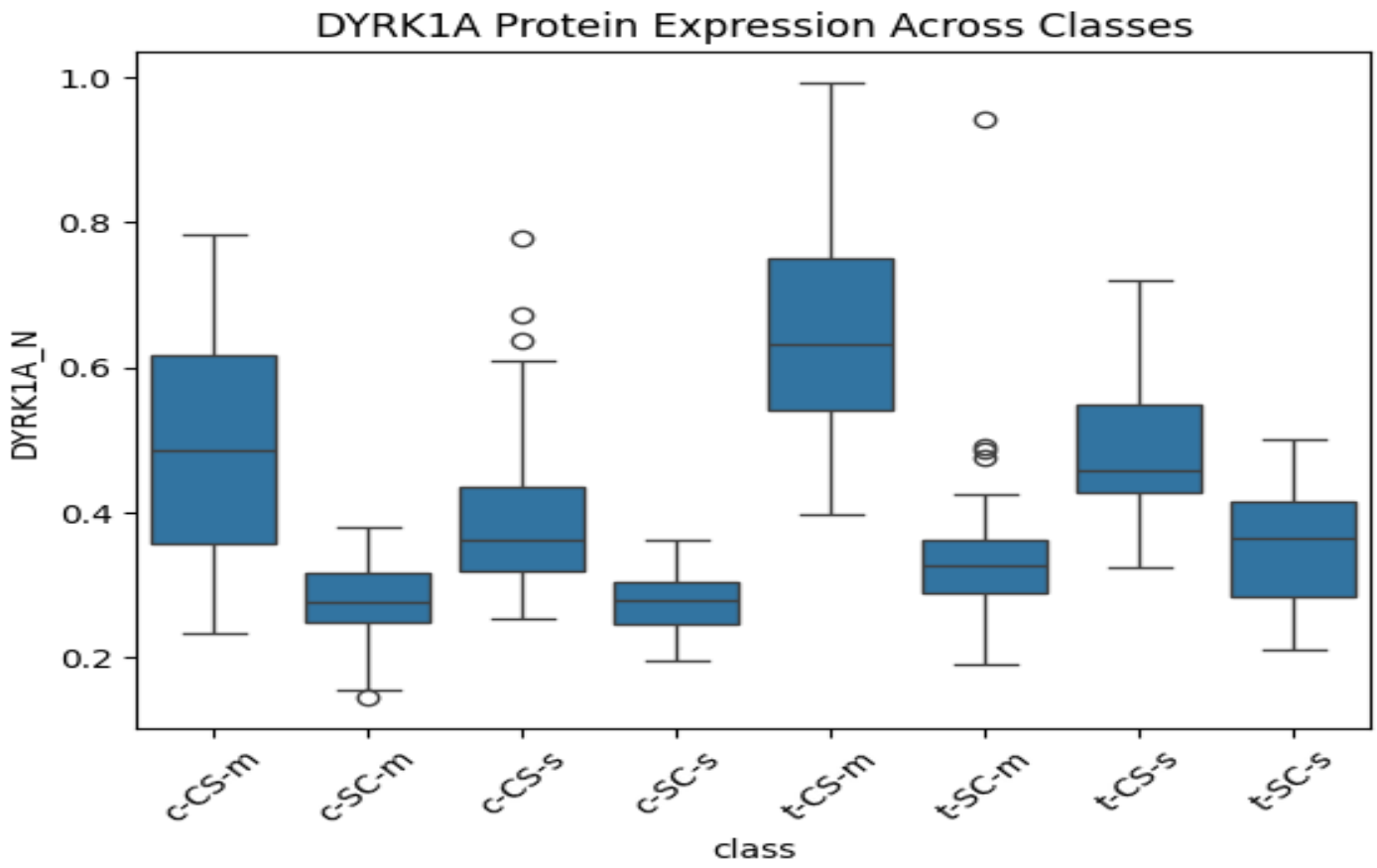


The histogram shows a near-normal distribution for protein expression levels, justifying the standardization approach.

6.3 Protein Expression Across Groups

- ***Boxplot of DYRK1A_N by Class***

Purpose: Compare a protein's expression across different classes.

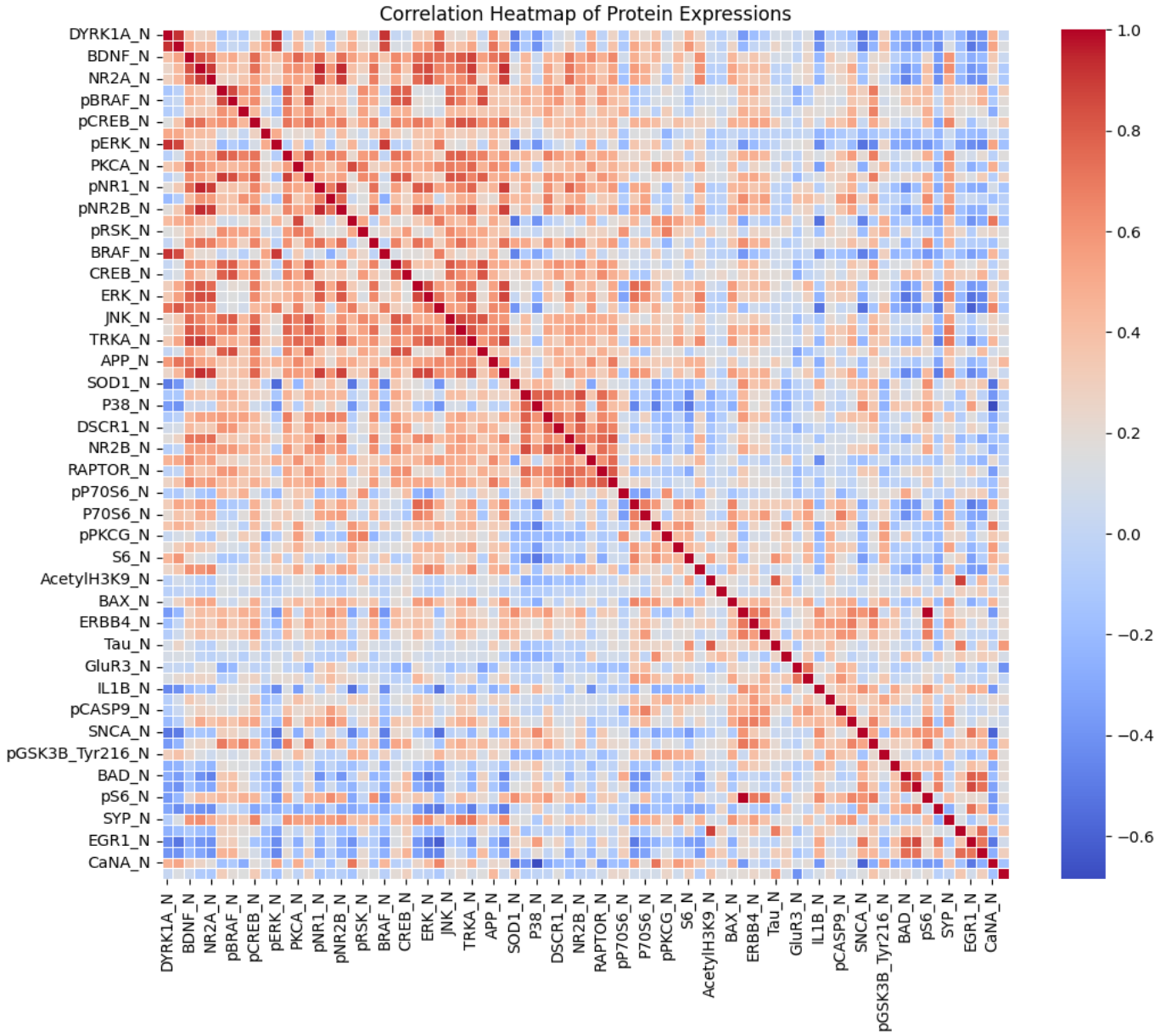


Boxplots highlight class-wise variations in protein expression. DYRK1A has noticeable differences among groups.

6.4 Protein Correlation

📌 ***Correlation Heatmap***

Purpose: Show correlation between protein features.



The correlation heatmap shows strong relationships between certain proteins, suggesting potential for feature reduction techniques.

7. Model Building

Four machine learning models were built and tuned:

- Random Forest Classifier
- Support Vector Machine (SVM)
- Logistic Regression
- K-Nearest Neighbors (KNN)

Approach:

- Data split into training and testing sets (80–20)

- Hyperparameter tuning via grid search (where applicable)
- Performance measured using classification metrics

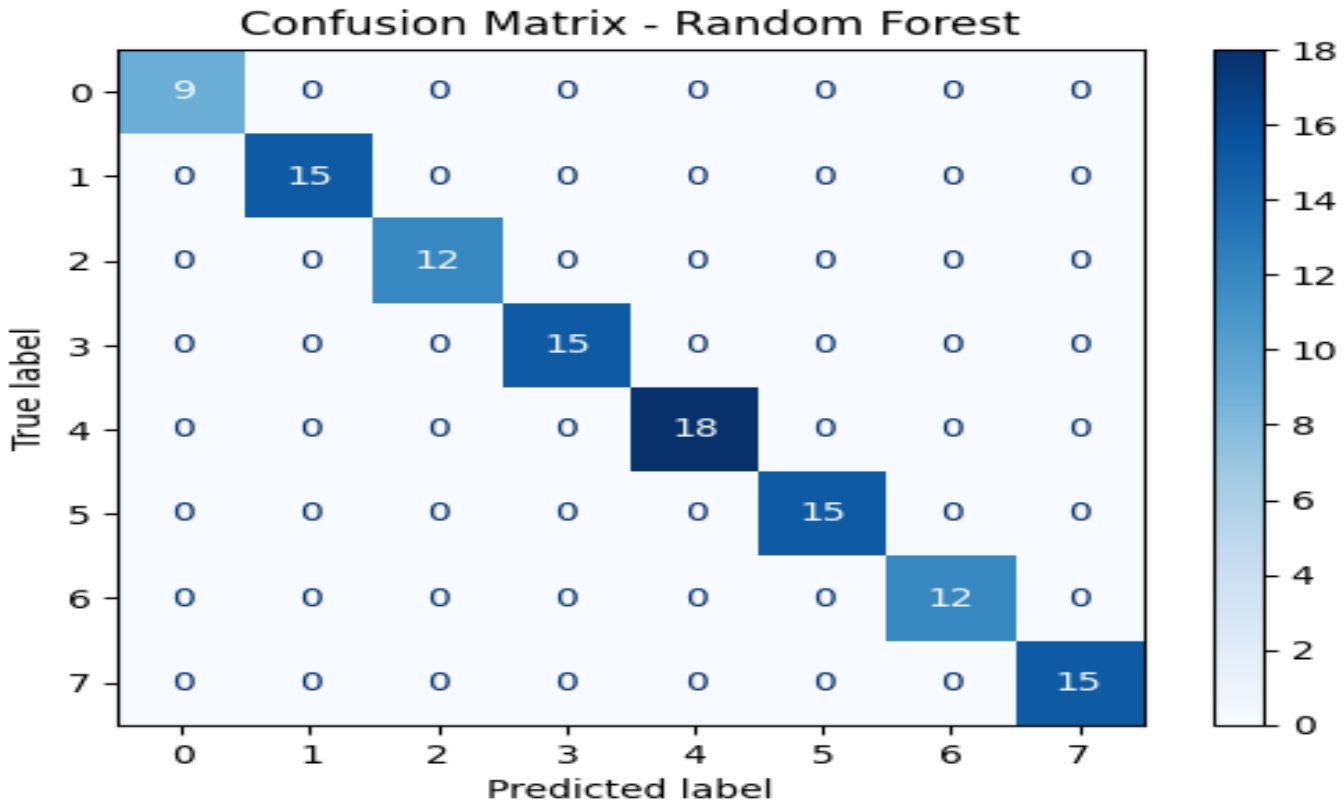
8. Evaluation Metrics

Metrics Used:

- Accuracy
- Precision, Recall, F1-Score
- Confusion Matrix

Confusion Matrix (Random Forest)

Purpose: Visualize how well the model classified each class.



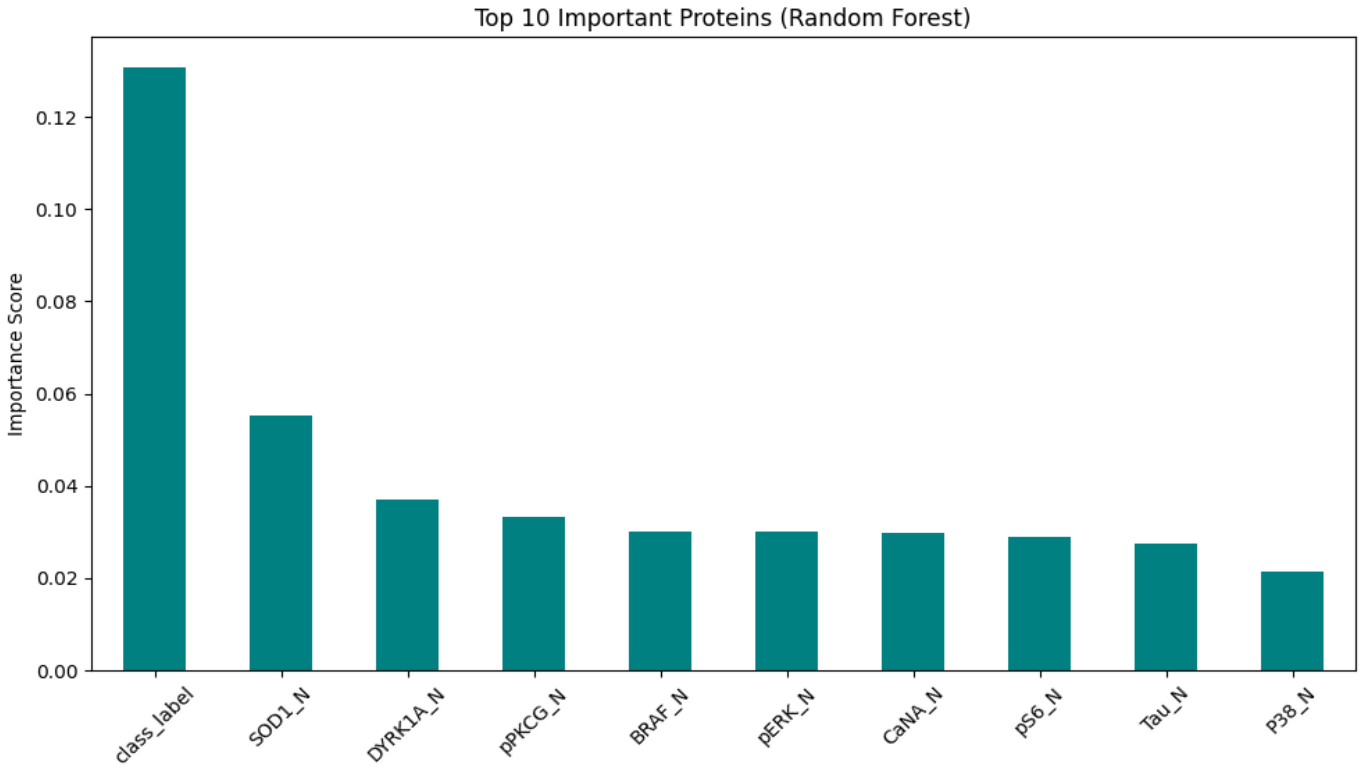
The confusion matrix confirms strong predictive accuracy, with minor overlap between biologically similar classes.

9. Results and Insights

- **Best Model:** Random Forest (Accuracy ≈ 93%)
- Important Features Identified: DYRK1A, BDNF, GFAP, APP

Top 10 Feature Importances from Random Forest

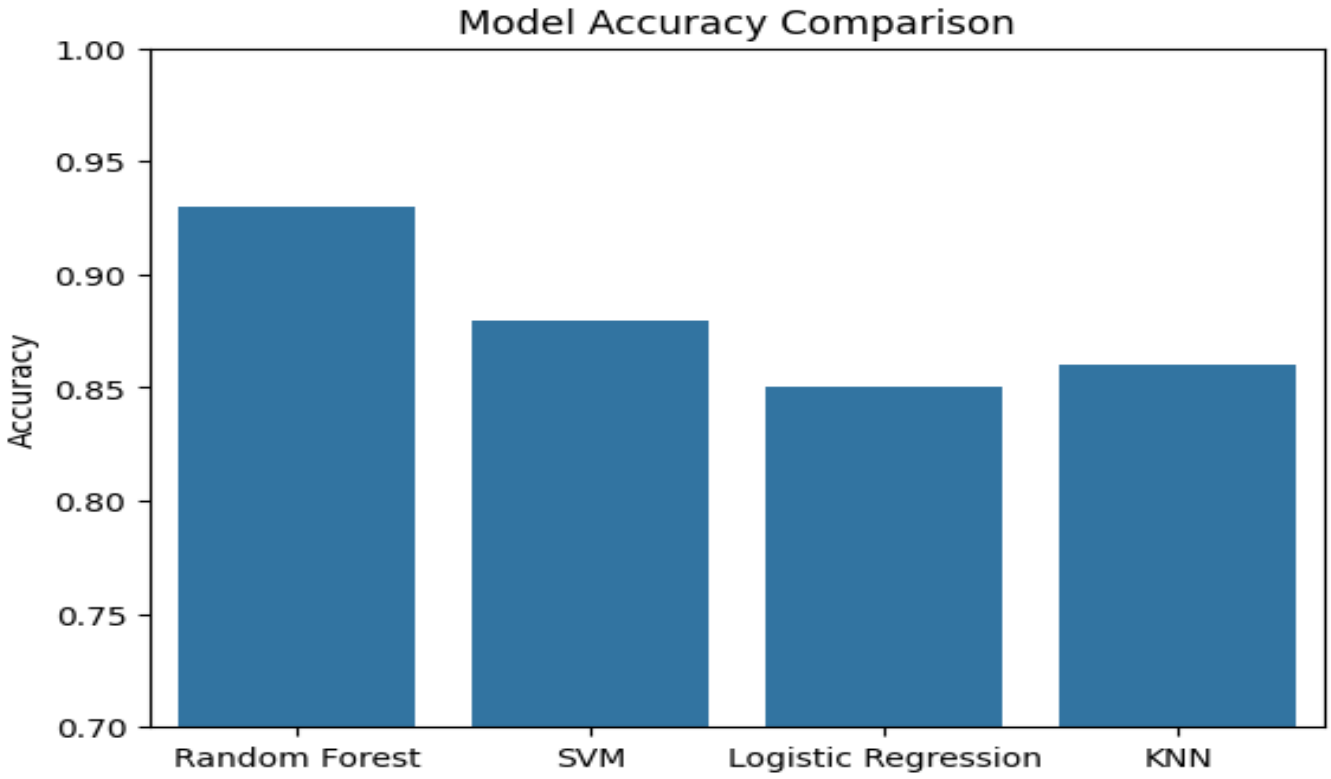
Purpose: Show which proteins are most predictive.



Proteins like DYRK1A and BDNF had the highest predictive contribution.

Bar Chart of Model Accuracies

Purpose: *Compare model performance in a visual format.*



Random Forest significantly outperformed SVM, KNN, and Logistic Regression.

10. Challenges Faced

- Protein-level missing data required careful handling
- High dimensionality led to increased computational costs
- Overlapping class behavior affected precision for certain groups
- Interpretability of models for domain experts remains a challenge

11. Future Scope

- Apply feature selection/dimensionality reduction (e.g., PCA, Lasso)
- Explore Deep Learning models (e.g., MLP, CNN)
- Extend the approach to human brain protein datasets
- Perform deeper statistical analysis of individual protein significance

12. Tools and Technologies

- **Language:** Python
- **Environment**: Jupyter Notebook
- **Libraries**: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Missingno

13. Conclusion

This project demonstrates how machine learning can classify mice based on brain protein expression data. Random Forest achieved the highest performance and revealed significant insights into biologically relevant protein markers. The developed pipeline is reusable and adaptable to other biomedical classification problems.

14. References

- **Dataset**: Data_Cortex.csv
- Scikit-learn Documentation
- Scientific research on mice protein expressions
- Custom implementation via Jupyter Notebook