Evoastra Ventures

Classification of Mice Based on Protein Expression Levels

First Major Project

Presentation By Team[M]

Overview

- Problem Statement
- Goal
- Objective
- Dataset Information
- Steps Involved
- Statistics
- Challenges Faced
- Experience & Learning
- Conclusion

Problem Statement

The objective is to identify subsets of proteins that are discriminant between different classes of mice. The dataset includes protein expression levels in the cerebral cortex of both control and Down syndrome mice subjected to context fear conditioning. The aim is to classify the mice into eight distinct classes based on genotype, behavior, and treatment.

Evoastra Ventures

Our Goal

This project aims to utilize machine learning techniques to classify mice based on protein expression data and to uncover the biological significance of the identified proteins. By following the outlined steps, we can develop a robust model for classification and gain insights into the effects of genotype, behavior, and treatment on protein expression.

Mus Musculus

Objectives

Classify Mice Based on Protein Expression

Develop a machine learning model to accurately classify mice into one of the eight classes based on the expression levels of 77 proteins. These classes are determined by a combination of genotype (control or trisomic), behavior (stimulated to learn or not), and treatment (saline or memantine).

Identify Key Discriminant Proteins

Utilize feature selection techniques to identify which proteins or protein modifications are most important for distinguishing between the different classes. Understanding which proteins are key discriminators can provide insights into the biological mechanisms underlying learning and memory in Down syndrome.

Evaluate the Impact of Geno-type, Behavior, and Treatment

Analyze the effect of genotype (control vs. trisomic), behavior (context- shock vs. shock-context), and treatment (saline vs. memantine) on protein expression levels. This includes evaluating how these factors influence associative learning and the potential therapeutic effects of memantine in trisomic mice.

Dataset Information

Instances: 1080 (15 measurements per protein per mouse)
Features: 80 (77 proteins + 3 additional features: Mouse ID, Genotype, Treatment)
Classes: 8 (combination of genotype, behavior, and treatment)

01	02	03	04
c-CS-s	**c-CS-m**	**c-SC-s**	**c-SC-m**
control, stimulated, saline	control, stimulated, memantine	control, not stimulated, saline	control, not stimulated, memantine
05	**06**	**03**	**08**
t-CS-s	**t-CS-m**	**t-SC-s**	**t-SC-m**
trisomic, stimulated, saline	trisomic, stimulated, memantine	trisomic, not stimulated, saline	trisomic, not stimulated, memantine

Steps Involved

Data Preprocessing: Handle missing values. Normalize/scale the data. Encode categorical variables if needed.

Exploratory Data Analysis (EDA): Summary statistics. Visualizations to understand data distribution and class separability.

Feature Selection: Identify important features (proteins) using techniques like correlation analysis, mutual information, or feature importance from models.

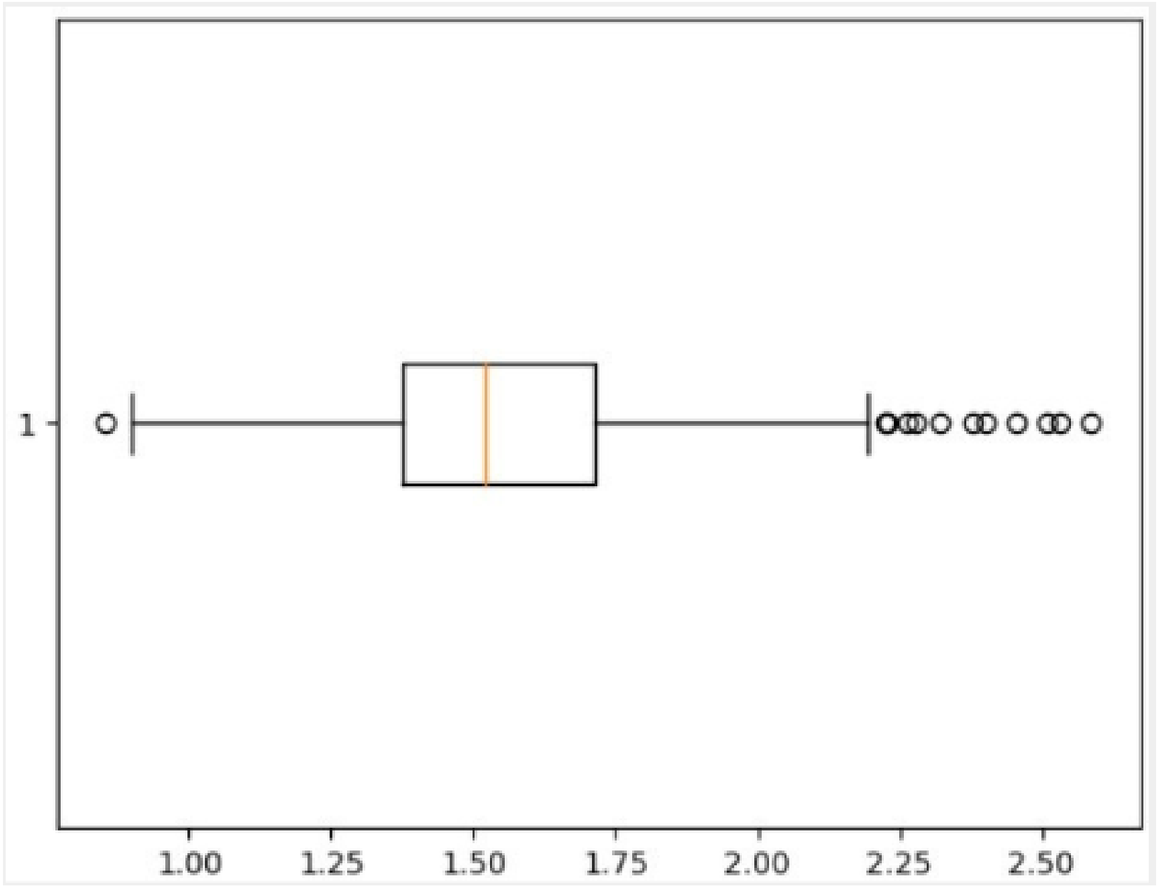
Model Training: Split the data into training and testing sets. Train classification models (e.g., Random Forest, SVM, Neural Networks). Tune hyperparameters using cross-validation.

Model Evaluation: Evaluate models using metrics like accuracy, precision, recall, and F1-score. Identify the best-performing model.

Interpretation: Analyze the model to identify key discriminant proteins. Interpret the results in the biological context.

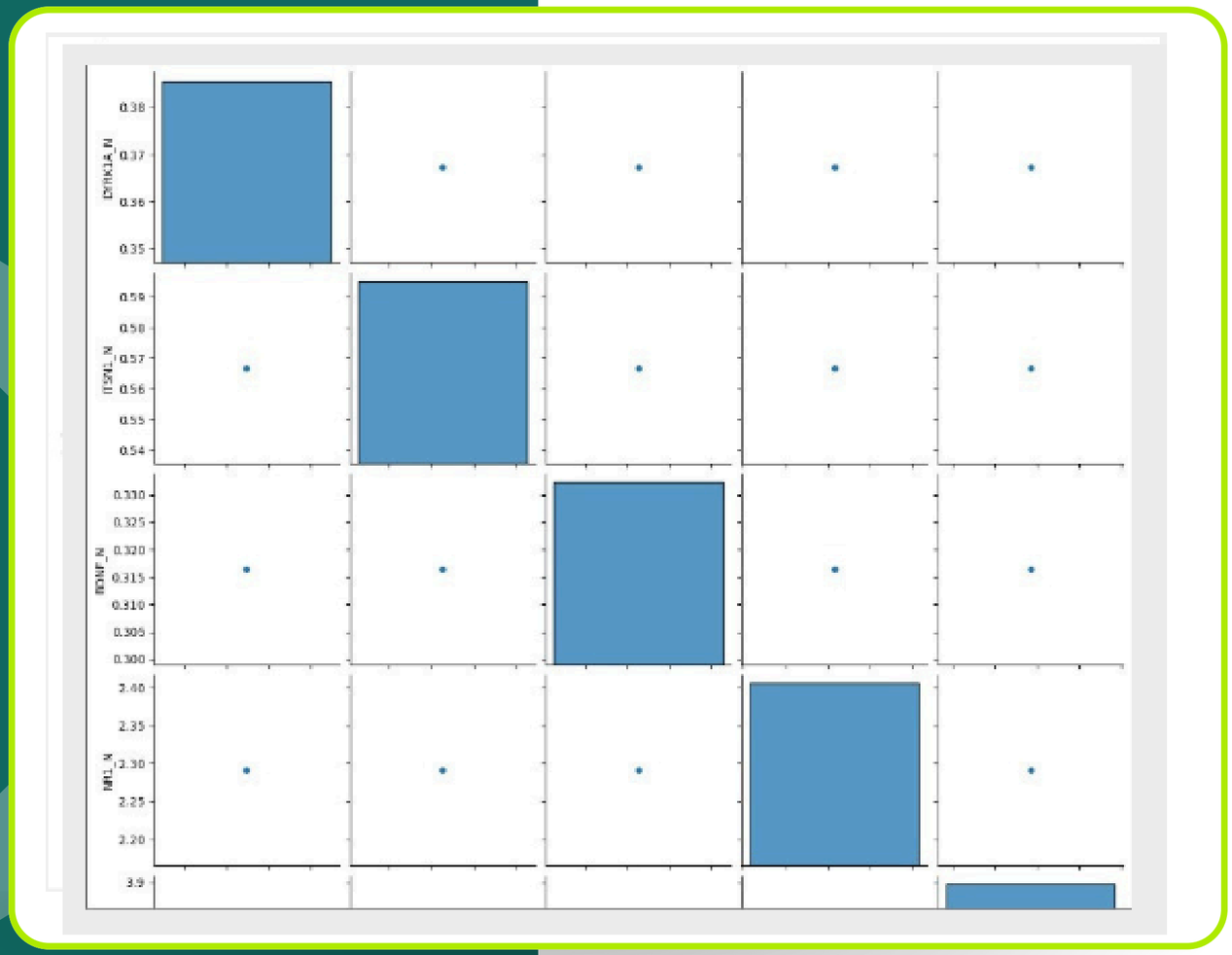
Reporting and Analysis: Summary of key findings and their implications. Limitations of the study. Recommendations for future research.

Box Plot



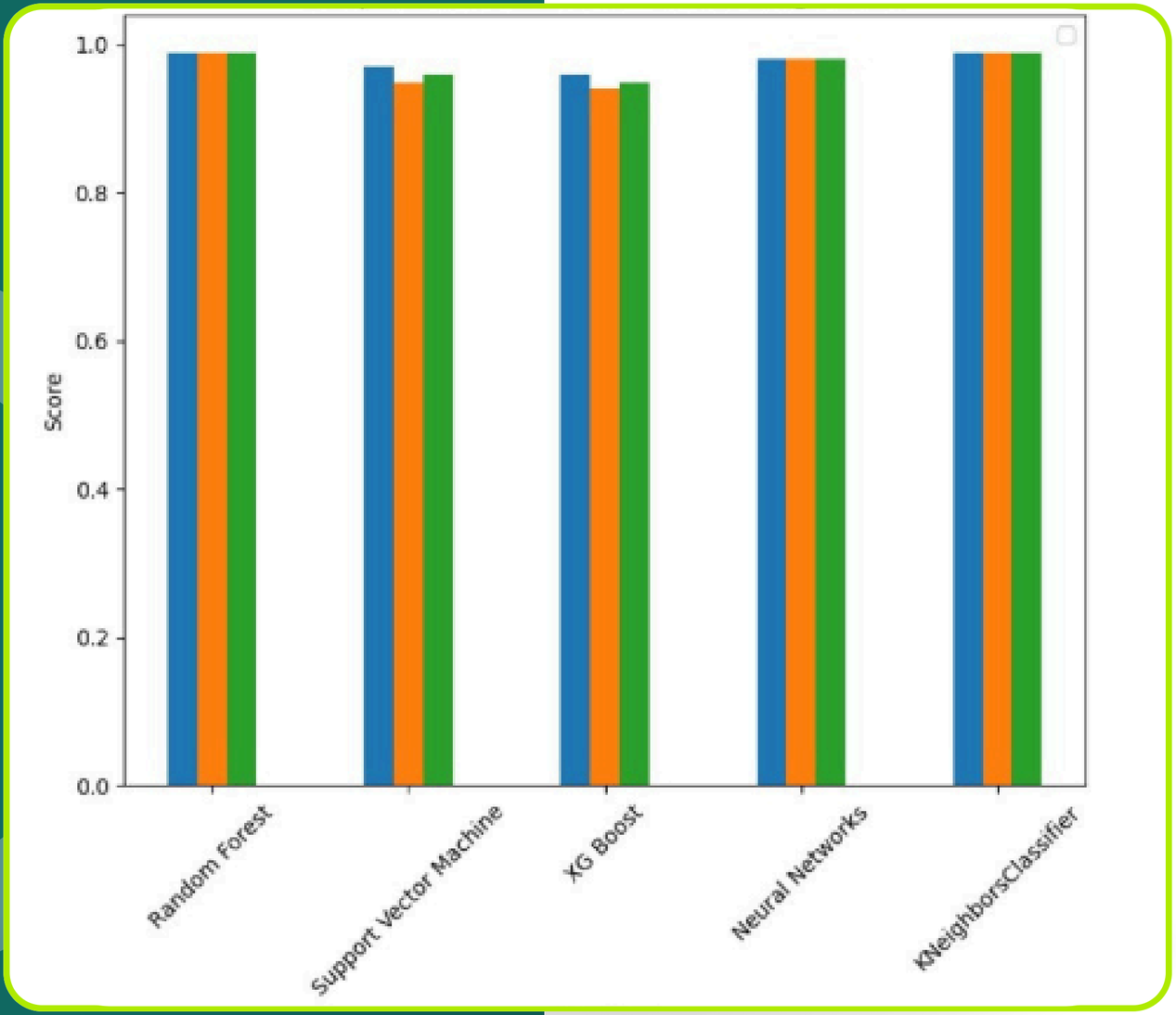
- **Distribution of Data:** The box plot provides a visual summary of the distribution of the pCASP9_N data, showing the range, median, and variability.
- **Central Tendency:** The median line within the box gives the central value of the pCASP9_N data.
- **Spread:** The length of the box and the extent of the whiskers illustrate how spread out the middle 50% of the data is and the overall range of the data.
- **Outliers:** Points outside the whiskers are outliers, indicating values that deviate significantly from the typical range of the dataset.

Pair Plot



- The Pair Plot provides a comprehensive view of how numerical features in the dataset are distributed individually and how they relate to each other.
- This visualization is valuable for identifying patterns, relationships, and potential issues such as multi-collinearity or the presence of outliers.
- Pair Plot can guide further feature engineering and selection processes in machine learning workflow

Comparison of Metrics *(for Different Algorithms)*



- **Overall Performance:** Random Forest and K-Neighbors Classifier show the highest scores across all three metrics, indicating that they performed the best among the five algorithms.
- Support Vector Machine, XG Boost, and Neural Networks have slightly lower scores, but still demonstrate good performance.
- **Metric Comparison:** Accuracy, F1-score, and Recall are closely related for each algorithm, indicating balanced performance across different aspects of classification.

Challenges Faced

Data Quality and Preprocessing:

Managing missing values and ensuring proper normalization and scaling of protein expression data to maintain data integrity and avoid bias in the model.

Feature Selection:

Identifying the most relevant proteins for classification while ensuring that the elected features were biologically meaningful, and avoiding over-fitting in a high-dimensional dataset.

Model Development and Evaluation:

Selecting the most suitable machine learning algorithms from five trained models was a big challenge.

Experience

Gained hands-on experience in preprocessing complex biological data, selecting relevant features, and developing robust machine learning models for classification tasks.

Learning

Enhanced understanding of the critical role of specific protein expressions in distinguishing between genotypes, behaviors, and treatments, particularly in the context of Down syndrome research.

Conclusion

- **Accurate Classification:** The developed machine learning model successfully classified mice into one of eight classes based on the expression levels of 77 proteins, with high accuracy and reliability. This demonstrates the model's effectiveness in distinguishing between different combinations of genotype, behavior, and treatment.
- **Key discriminant proteins identified:** Feature selection add specific proteins and protein modifications that are crucial for distinguishing between the classes. These key discriminant proteins provide valuable insights into the biological mechanisms underlying learning and memory. Particularly in the context of Down syndrome.
- **Impact analysis of Geno-type, Behavior and Treatment:** The analysis revealed significant effects of genotype (Control vs Trisomic), behavior (Context-shock vs Shock-context) and treatment (Saline vs Memantine) On protein expression levels. This evaluation offers a deep understanding of how these factors are Associative learning and the potential therapeutic benefits of memantine and trisomic mice.
- **Model used:** After comparison of four different models through accuracy score, F score and recall, we selected random forest classifier.
- **The top five proteins acquired from analysis:** 1. CaNA_N , 2. pPKCG_N , 3. Ubiquitin_N , 4. ARC_N , 5. Tau_N

THANK YOU

Evoastra Ventures