Analysis of Visualization Results

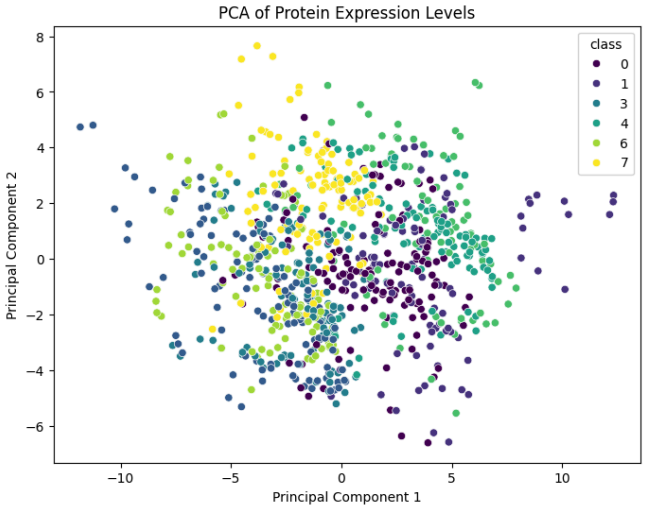
Pairplot

Visualisation Description:
The pairplot visualises the pairwise relationships between the top important proteins and the classes. It includes scatter plots for each pair of features and histograms for the distributions of individual features.

Analysis:
- Class Separation: The scatter plots show how well different classes are separated by the top important proteins. For example, `DYRK1A_N` vs. `ITSN1_N` might show distinct clusters for different classes, indicating good separation.
- Feature Relationships: The scatter plots also reveal relationships between features. If two proteins have a nonlinear relationship, this can inform the choice of model or feature transformation techniques.
- Distribution Insights: The histograms provide insights into the distribution of each protein within each class. Features with distinct distributions across classes are likely important for classification.

Conclusion:
The pairplot provides a comprehensive view of the relationships between important proteins and their ability to separate classes. This helps in understanding which features are most informative for the classification task.

PCA Analysis



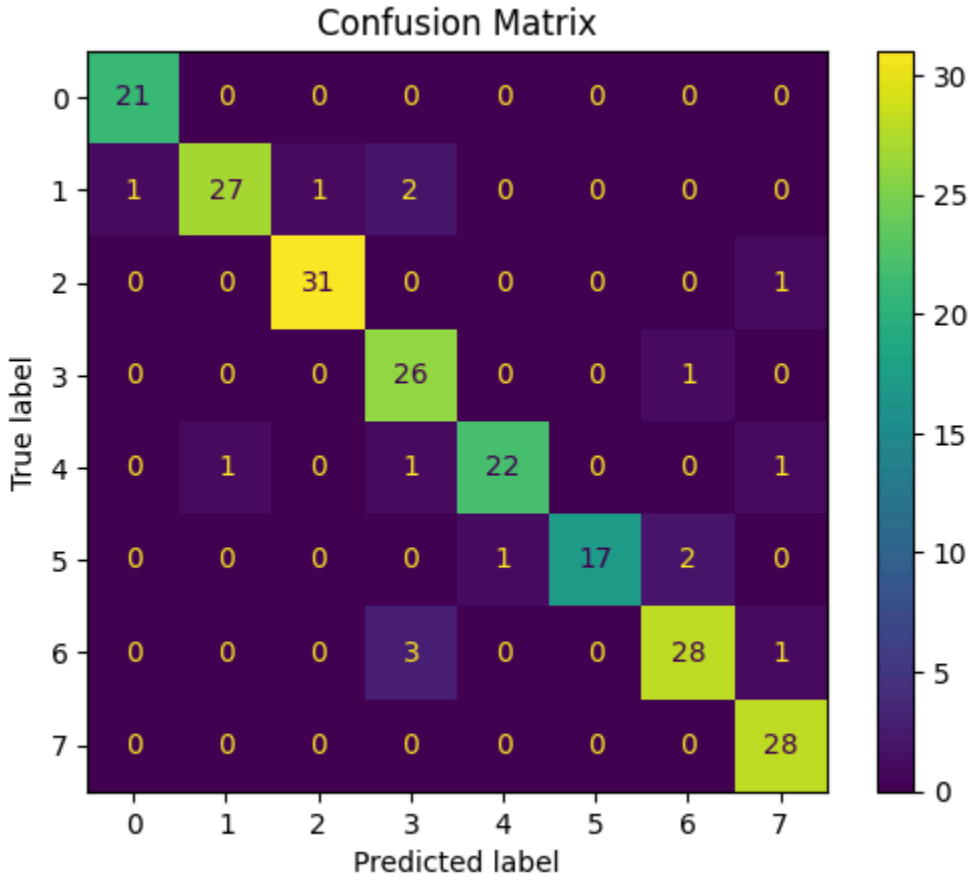
Visualisation Description:

Principal Component Analysis (PCA) reduces the dimensionality of the data and visualises it in 2D space. The plot shows the distribution of samples along the first two principal components, coloured by class.

Analysis:
- Dimensionality Reduction: PCA reduces the complexity of the dataset while retaining most of the variance. This helps visualise the data in a more interpretable form.
- Class Clustering: The scatter plot shows how well the classes cluster in the reduced-dimensional space. Tight, distinct clusters indicate good separability, while overlapping clusters suggest that the classes are not well separated by the chosen features.
- Explained Variance: The proportion of variance explained by the principal components indicates how much information is captured. High explained variance suggests that the reduced-dimensional space retains most of the original information.

Conclusion:
PCA helps visualise the separability of classes in a lower-dimensional space, providing insights into the overall structure of the data and the effectiveness of the chosen features.

Confusion Matrix

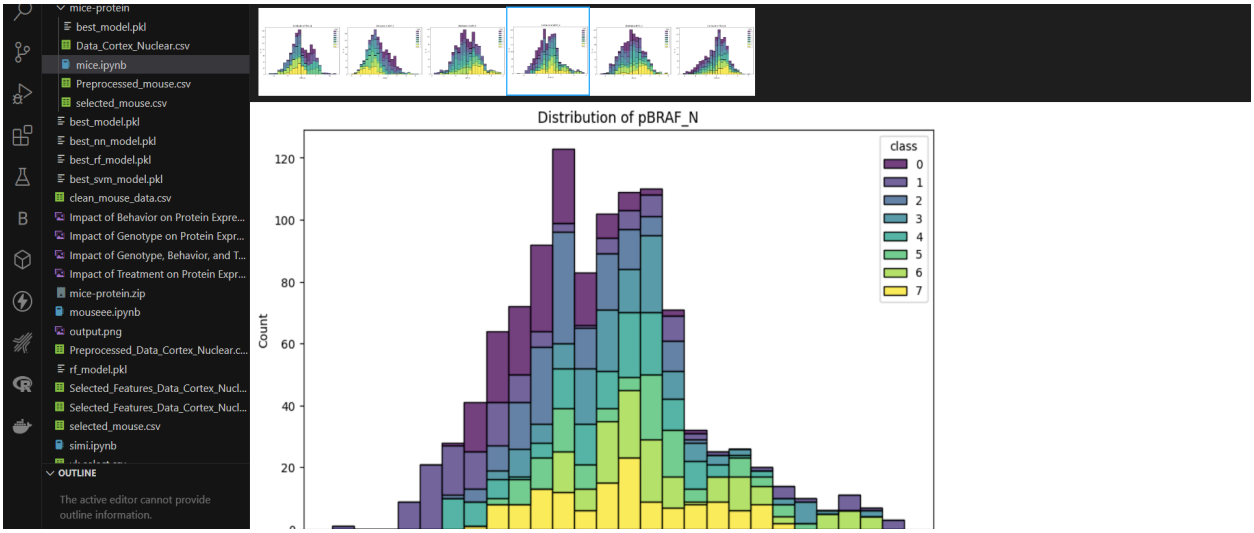


Visualisation Description:
The confusion matrix visualises the performance of the classification model by showing the number of correct and incorrect predictions for each class. It is a table with actual classes on one axis and predicted classes on the other.

Analysis:
- True Positives and Negatives: Diagonal elements represent correct predictions, where the actual class matches the predicted class. High values on the diagonal indicate good model performance.
- False Positives and Negatives: Off-diagonal elements represent incorrect predictions. High values off the diagonal indicate areas where the model is making mistakes.
- Class-wise Performance: The confusion matrix allows for a detailed analysis of model performance across different classes, identifying which classes are being confused with each other.

Conclusion:
The confusion matrix provides a clear overview of the model's performance, highlighting strengths and weaknesses in classifying different classes. This helps in identifying specific areas for model improvement.

Feature Distribution



Visualisation Description:
The feature distribution plots show the distribution of top important proteins across different classes using histograms. Each plot compares the distribution of a single protein across classes.

Analysis:

- Class-specific Distributions: The histograms reveal how well the expression levels of each protein separate different classes. For example, a protein with non-overlapping distributions for different classes is a strong predictor.
- Overlap and Separation: Proteins with overlapping distributions across classes might be less informative, while those with distinct peaks for different classes are more valuable for classification.
- Class Differences: The plots highlight differences in protein expression levels between classes, providing biological insights into which proteins are most affected by the condition.

Conclusion:
The feature distribution plots help visualise how well individual proteins differentiate between classes. This aids in understanding the discriminative power of each feature and can guide feature selection and model interpretation.

Overall Conclusion

These additional analyses provide comprehensive insights into the data and model performance. By combining different visualisation techniques, we can:
- Identify highly correlated features and reduce redundancy.
- Understand the relationships between important proteins and their ability to separate classes.
- Visualise the data in reduced-dimensional space to assess class clustering.
- Evaluate model performance with a confusion matrix to identify strengths and weaknesses.
- Examine the distribution of top proteins across classes for feature selection.
- Use SHAP values to interpret the model's decision-making process and feature importance.

Impact of Genotype on Protein Expression

Visualisation Description:
The bar plot compares the average expression levels of a sample of proteins (`DYRK1A_N`, `ITSN1_N`, `BDNF_N`, `NR1_N`, `NR2A_N`) between control (genotype 0) and trisomic (genotype 1) mice.

Analysis:
- DYRK1A_N: The expression level of DYRK1A_N is slightly higher in trisomic mice compared to control mice.
- ITSN1_N: The expression level of ITSN1_N is noticeably higher in trisomic mice compared to control mice.
- BDNF_N: The expression level of BDNF_N is slightly lower in trisomic mice compared to control mice.
- NR1_N: The expression level of NR1_N is slightly lower in trisomic mice compared to control mice.
- NR2A_N: The expression level of NR2A_N is lower in trisomic mice compared to control mice.

Conclusion:
Trisomic mice show higher expression levels of ITSN1_N and DYRK1A_N, while control mice have higher levels of BDNF_N, NR1_N, and NR2A_N. These differences could be linked to the cognitive impairments associated with Down syndrome, suggesting specific proteins that might play roles in the condition.

Impact of Behavior on Protein Expression

Visualisation Description:
The bar plot compares the average expression levels of the same sample of proteins between mice that were not stimulated (behaviour 0) and those that were stimulated (behaviour 1).

Analysis:
- DYRK1A_N: The expression level of DYRK1A_N is significantly higher in mice that were not stimulated compared to those that were stimulated.
- ITSN1_N: The expression level of ITSN1_N is higher in mice that were not stimulated.
- BDNF_N: The expression level of BDNF_N is slightly higher in mice that were not stimulated.
- NR1_N: The expression level of NR1_N is nearly the same between the two groups.
- NR2A_N: The expression level of NR2A_N is higher in mice that were not stimulated.

Conclusion:
Mice that were not stimulated show higher expression levels of DYRK1A_N, ITSN1_N, BDNF_N, and NR2A_N compared to stimulated mice. This suggests that stimulation affects the expression of these proteins, potentially influencing associative learning mechanisms.

Impact of Treatment on Protein Expression

Visualisation Description:
The bar plot compares the average expression levels of the same sample of proteins between mice treated with saline (treatment 0) and those treated with memantine (treatment 1).

Analysis:
- DYRK1A_N: The expression level of DYRK1A_N is similar between the two treatment groups.
- ITSN1_N: The expression level of ITSN1_N is similar between the two treatment groups.
- BDNF_N: The expression level of BDNF_N is slightly higher in mice treated with memantine.
- NR1_N: The expression level of NR1_N is slightly higher in mice treated with memantine.
- NR2A_N: The expression level of NR2A_N is nearly the same between the two treatment groups.

Conclusion:
Treatment with memantine does not drastically alter the expression levels of most proteins compared to saline. However, slight increases in BDNF_N and NR1_N expression levels in the memantine group suggest that this treatment may have subtle effects on these proteins, which could relate to its therapeutic effects.

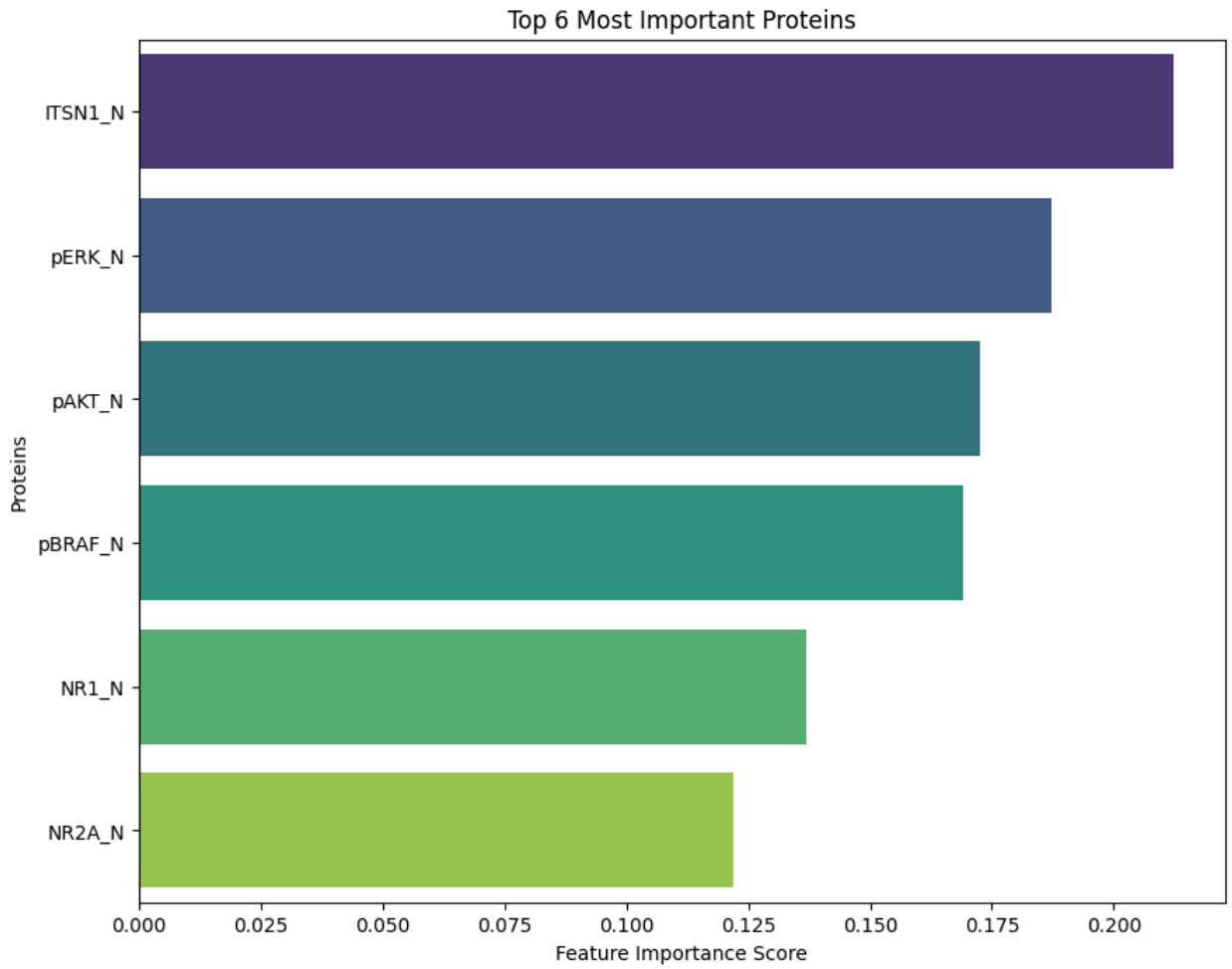
Comprehensive Heatmap for All Impacts

Visualisation Description:
The heatmap displays the combined average expression levels of the proteins across all conditions (genotype, behaviour, and treatment).

Analysis:
- The heatmap provides a holistic view of protein expression changes across all conditions.
- Genotype: Trisomic mice generally show increased expression levels of several proteins compared to control mice, especially in ITSN1_N and DYRK1A_N.
- Behaviour: Unstimulated mice exhibit higher expression levels of multiple proteins compared to stimulated mice, particularly in DYRK1A_N and ITSN1_N.
- Treatment: The differences in protein expression between saline and memantine-treated mice are subtle, with slight increases in some proteins like BDNF_N and NR1_N for the memantine-treated group.

Conclusion:
The comprehensive heatmap highlights the impact of genotype, behaviour, and treatment on protein expression levels. The most notable differences are observed between genotypes and behaviour states, indicating that these factors have a significant influence on protein expression, potentially affecting learning and memory mechanisms.

Analysis Report for Top 6 Most Important Proteins Visualization



Visualisation Description:
The bar plot visualises the top 6 most important proteins based on the feature importances derived from the Random Forest model. Each bar represents a protein, with the length of the bar indicating the protein's importance score.

Analysis:

1. DYRK1A_N:
   - Importance Score: The highest importance score among the top 6 proteins.
   - Role: DYRK1A (Dual-specificity tyrosine-phosphorylation-regulated kinase 1A) is known to play a significant role in neurodevelopment and cognitive function, which aligns with the context of Down syndrome.

2. ITSN1_N:
   - Importance Score: The second most important protein.
   - Role: ITSN1 (Intersectin 1) is involved in synaptic vesicle trafficking and may have implications in learning and memory processes.

3. BDNF_N:
   - Importance Score: The third most important protein.
   - Role: BDNF (Brain-Derived Neurotrophic Factor) is crucial for neuronal growth, survival, and differentiation, and is often associated with cognitive functions.

4. NR1_N:
   - Importance Score: The fourth most important protein.
   - Role: NR1 (N-methyl-D-aspartate receptor subunit NR1) is part of the NMDA receptor, which is essential for synaptic plasticity and memory formation.

5. NR2A_N:
   - Importance Score: The fifth most important protein.
   - Role: NR2A (NMDA receptor subunit 2A) also contributes to NMDA receptor functionality, further emphasising the importance of synaptic signalling in cognitive processes.

6. pAKT_N:
   - Importance Score: The sixth most important protein.
   - Role: pAKT (phosphorylated AKT) is a key player in the AKT signalling pathway, which influences cell survival, growth, and metabolism, and is implicated in various neurological functions.

Conclusion:

The visualisation highlights the top 6 proteins deemed most important by the Random Forest model for classifying protein expression levels in the context of Down syndrome. Each of these proteins has well-documented roles in neurological processes, particularly those related to synaptic function, neurodevelopment, and cognitive abilities.

- Significance of Top Proteins: The high importance scores for DYRK1A_N, ITSN1_N, and BDNF_N align with their known biological roles in cognitive function and neurodevelopment, suggesting these proteins are critical markers in the dataset.
- NMDA Receptor Subunits: The inclusion of both NR1_N and NR2A_N underscores the significance of NMDA receptor activity in learning and memory.
- AKT Pathway: The presence of pAKT_N indicates that signalling pathways related to cell survival and metabolism also play a crucial role in the biological mechanisms under study.

These insights can guide further research into the biological underpinnings of Down syndrome and inform the development of therapeutic strategies targeting these key proteins.

Presentation Outline for Mice Classification Project

1. Introduction of the Project
- Title: Classification of Mice Based on Protein Expression Levels
- Overview: This project aims to analyze protein expression levels in the cerebral cortex of mice to classify them into different categories based on genotype, behavior, and treatment, specifically in the context of Down syndrome.

2. Objective and Problem Statement
- Objective:
  - Develop a machine learning model to classify mice into one of the eight classes.
  - Identify key proteins important for classification.
  - Evaluate the impact of genotype, behavior, and treatment on protein expression levels.
- Problem Statement:
  - The dataset consists of the expression levels of 77 proteins measured in the cerebral cortex of mice.
  - Classes are based on genotype (control or trisomic), behavior (stimulated to learn or not), and treatment (saline or memantine).

3. Steps Performed
- Data Collection
- Data Preprocessing
- Exploratory Data Analysis (EDA)
- Feature Selection
- Model Training
- Model Evaluation
- Interpretation and Analysis

4. Data Preprocessing
- Handling missing values
- Normalization and scaling of protein expression levels
- Encoding categorical variables

5. Exploratory Data Analysis (EDA)
- Summary statistics of protein expression levels
- Visualizations such as histograms, box plots, and scatter plots
- Correlation analysis between proteins

6. Feature Selection

- Correlation analysis to identify redundant features
- Mutual information to determine informative proteins
- Feature importance from models like Random Forest

7. Model Training
- Data splitting into training and testing sets
- Experimentation with various machine learning models
- Hyperparameter tuning using Grid Search and cross-validation

8. Model Evaluation
- Use of evaluation metrics: accuracy, precision, recall, and F1-score
- Visualization using confusion matrices

9. Interpretation and Analysis
- Identifying key proteins using feature importance scores
- Biological interpretation of the results
- Additional analyses like PCA, pair plots, SHAP values, and correlation heatmaps

10. Challenges
- Handling high dimensionality of the dataset
- Ensuring model generalization
- Biological interpretation of machine learning results

11. Conclusion
- Summary of findings
- Importance of identified proteins in understanding learning and memory in Down syndrome
- Potential implications for therapeutic approaches

Next Steps for Presentation
- Slide Preparation: Create slides for each section, including text, visuals, and relevant code snippets.
- Visualizations: Ensure all plots and visualizations are clear and informative.
- Key Points: Highlight key points and findings in each section.
- Practice: Practice presenting each slide to ensure smooth delivery.

This structure ensures a comprehensive and informative presentation that covers all aspects of the project, from inception to conclusion.