**Github Link:** [**amberwilliams181/Machine\_Learning\_CW: Data Science and Machine Learning Coursework - Amber Williams**](https://github.com/amberwilliams181/Machine_Learning_CW)

**Introduction:**

This project investigates the impact of salicylic acid (SA) application on gene expression in *Arabidopsis thaliana* leaves, with a focus on identifying genes that show the most significant variation in expression between control and treatment groups. Additionally, machine learning models, including logistic regression and neural networks, were employed to predict whether genes will be upregulated or downregulated under treatment based on their expression profiles under control conditions. Through this analysis, the project aims to contribute to a better understanding of the molecular response of *Arabidopsis thaliana* leaves to salicylic acid application.

**Methods:**

**Data Cleaning and Pre-processing:** The dataset was pre-processed by removing double identification columns and then setting the ID column as the index and removing missing values. Column names were updated to reflect the treatment conditions. Averages of control and treatment groups were computed and added as new columns ("Control Avg." and "Treatment Avg.") for each gene, creating a dataset of gene expression averages for subsequent analysis.

**Variance Thresholding:** Variance thresholding was applied using the gene expression averages dataset to select genes that showed significant variability across control and treatment groups. Various thresholds were tested, with a value of 2 chosen to retain a manageable number of genes for further analysis if a literature review were to be later undertaken.

**Heatmap Generation:** A heatmap was created to visualise the expression of genes meeting the selected variance threshold, representing genes that varied the most between control and treatment groups.

**Hierarchical Clustering:** Hierarchical clustering was performed on the genes based on Euclidean distances by grouping genes with similar expression patterns. The clustering results were observed with a dendrogram to identify gene clusters with similar expression activity across samples.

**Principle Component Analysis (PCA):** PCA was used to reduce the dimensionality of the dataset to two components (PC1 and PC2), visualizing the clustering of control and treatment samples. The explained variance for each component was calculated to assess their role in separating the groups.

**Distribution of upregulated and downregulated genes:** The distribution of upregulated and downregulated genes was calculated to establish the balance of data, as imbalanced data can lead to model bias towards the majority class, leading to model accuracy becoming less interpretable.

**Machine Learning Models for Gene Regulation Prediction:** Logistic regression and neural network models were utilised to predict gene upregulation or downregulation based on control expression data alone. The models were evaluated using classification accuracy and visualised via confusion matrices and a learning curve plot for the neural network model.

**Results:**

**Variance Thresholding:** Genes exhibiting the most significant changes in expression were successfully selected. A variance threshold of 2 was selected to balance the number of significant genes for further analysis. A heatmap was then generated, highlighting the top 19 genes that showed the greatest variation which could provide an appropriate starting point for investigating the functional effects of salicylic acid treatment.

**Hierarchical Clustering:** Several genes clustered together very closely, such as GSTU24 and UGT75B1, and AT4G22530 and GRX480, indicating similar expression responses between control and treatment groups, while others were more distantly related thus having more divergent expression profiles. These insights into gene similarities could be used to inform further investigation into potential co-regulated genes or involvement in related biological pathways.

**Principle Component Analysis:** The proportion of variance explained by PC1 and PC2 was 40.73% and 35.73%, respectively, which accounted for 76.46% of the total variance. Control and treatment groups showed some separation along PC1, though clustering was not tight. Control 2 and 3, along with treatment 1 and 3 showed the closest clustering, whereas control 1 and treatment 2 were more dispersed. Additionally, treatment 1 and control 1 clustered closely along PC2, as did control 3 and treatment 2, suggesting partial overlap in their gene expression profiles. This overlap along PC2 could reflect environmental factors or experimental variations which influenced gene expression independently of SA treatment.

**Logistic Regression:** The distribution of data was balanced and therefore modelling was able to be applied without modifying the dataset. The model achieved an accuracy of 61%, which is only slightly better than chance. Precision and recall were relatively balanced, with a slightly better recall for class 0 (downregulation) compared to class 1 (upregulation). The F1-scores were close, with class 0 scoring 0.64 and class 1 scoring 0.58, indicating no significant bias towards either class. Given the limited results, a more complex model was tested.

**Neural Network:** The neural network model achieved an accuracy of 62%, showing little improvement over the logistic regression model. Its performance remained similar to chance, indicating limited effectiveness for predicting upregulation or downregulation under treatment.

**Conclusions:**

In conclusion, 19 genes were identified that showed significant changes in expression between the control and treatment groups, indicating their importance as potential candidates for salicylic acid response or stress pathway investigation. Further exploration of their relationships through hierarchical clustering could also prove beneficial for future study into their functions. The inability of the models to accurately predict gene regulation may stem from the inherent complexity of gene expression regulation, where multiple interacting pathways influence the outcome. This suggests that additional features, such as environmental variables or protein interaction data, might improve predictions.