**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 which show the most significant expression variation 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.

**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. The column names were also updated to reflect the treatment conditions in layman terms. Averages of control and treatment groups were calculated 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 performed 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 achieve a manageable number of genes for further analysis, for example if a literature review were to be later undertaken.

**Heatmap Generation:** A heatmap was created to visualise the expression of genes which met 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, grouping genes that showed similar expression patterns. The clustering results were observed with a dendrogram in order 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), visualising the clustering of control and treatment samples. The explained variance for each component was calculated to assess their roles in splitting the groups.

**Distribution of upregulated and downregulated genes:** The distribution of upregulated and downregulated genes was calculated to establish balance of data, where imbalanced data may lead to model bias towards the class which has the majority, leading to model accuracy having lower validity.

**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 were also visualised via confusion matrices as well as 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 therefore had more divergent expression profiles. These insights into gene similarities could be used to advise further investigation into potential co-regulated genes or their 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, although clustering was not very 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 relatively balanced and therefore modelling was able to be applied without dataset modification. The model achieved an accuracy of 61%, which is only slightly better than chance. Precision and recall were also 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. As the results were limited, 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, suggesting 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 be a result of the fundamental complexity of gene expression regulation. This suggests that additional features, possibly including environmental variables or protein interaction data, might improve predictions.