Report On:

**Bioinformatics Final Project: Machine Learning Applied to Gene Expression Data**

**Team 12**

**Fall 2023**



Report By:  
 Nikhil Das Karavatt - 1002085391

Hari Chandana Bandreddi - 1002084349

Pujith Vaddi - 1002032705

Mahima Krishna Reddy Karna - 1002069435

Sakshi Nischal - 1002085832

Table of Contents

[Algorithm Analysis for Cladocopium Classification 3](#_Toc153328867)

[Method Description 3](#_Toc153328868)

[Explanation of Each Algorithm Used and Expected Outcomes 3](#_Toc153328869)

[Features used in Algorithms 4](#_Toc153328870)

[Phenotypes Study 6](#_Toc153328871)

[Logistic Regression: 6](#_Toc153328872)

[Linear Discriminant Analysis: 7](#_Toc153328873)

[Quadratic Discriminant Analysis: 8](#_Toc153328874)

[Support Vector Machine: 9](#_Toc153328875)

[Least Partial Square: 10](#_Toc153328876)

[Best Model 11](#_Toc153328877)

[Results of Algorithms 11](#_Toc153328878)

[Gene Numbers or Latent Factor Information: 11](#_Toc153328879)

[Performance Errors: 12](#_Toc153328880)

[Outputs 14](#_Toc153328881)

[Conclusion 18](#_Toc153328882)

# Algorithm Analysis for Cladocopium Classification

The goal of this analysis is to classify Cladocopium based on the host coral species (Orbicella annularis, OANN) using various machine learning algorithms.

To achieve this classification, we applied the following machine learning algorithms:

* Logistic Regression
* Linear Discriminant Analysis (LDA)
* Quadratic Discriminant Analysis (QDA)
* Support Vector Machine (SVM)
* Partial Least Squares (PLS)

For each algorithm, we utilized default settings and tuned specific hyperparameters to enhance classification performance.

# Method Description

**Logistic Regression:**

Explanation**:** Logistic regression is a supervised algorithm used for binary classification. It models the probability of a class given input features.

Customization**:** We adjusted regularization parameters to prevent overfitting.

**Linear Discriminant Analysis (LDA):**

Explanation**:** LDA is a supervised method used for dimensionality reduction and classification. It maximizes the separation between classes.

Customization**:** No specific customization was applied as default settings yielded optimal results.

**Quadratic Discriminant Analysis (QDA):**

Explanation**:** QDA is similar to LDA but assumes different covariance matrices for each class.

Customization**:** We explored regularization parameters to handle potential issues with collinearity.

**Support Vector Machine (SVM):**

Explanation**:** SVM is a powerful algorithm for classification tasks. It identifies an optimal hyperplane to separate classes.

Customization**:** Kernel and regularization parameters were tuned for improved performance.

**Partial Least Squares (PLS):**

Explanation**:** PLS is a supervised method that combines features to predict the target variable while considering collinearity.

Customization**:** We adjusted the number of components to capture sufficient variance.

## Explanation of Each Algorithm Used and Expected Outcomes

**Logistic Regression:**Objective: Predict the probability of a binary outcome.

Expected Outcomes: Identification of significant features through coefficient values. Probability estimation for binary classification.

**Support Vector Machine (SVM):**

Objective: Classify data points into two categories.

Expected Outcomes: Identification of the hyperplane that best separates classes. Effective in high-dimensional spaces.

**Linear Discriminant Analysis (LDA):**

Objective: Maximize separation between multiple classes.

Expected Outcomes: Identification of significant features through coefficients. Maximization of class separability.

**Quadratic Discriminant Analysis (QDA):**

Objective: Model differences in class covariances.

Expected Outcomes: Identification of significant features. Accommodates different covariances for each class.

**Partial Least Squares (PLS):**

Objective: Find latent variables maximizing covariance between predictors and response.

Expected Outcomes: Extraction of latent variables, identification of top genes impacting the response.

Logistic Regression: Supervised. Requires labeled data for training.

SVM: Supervised. Learns from labeled training data.

LDA: Supervised. Relies on class labels during training.

QDA: Supervised. Class-dependent covariance modeling.

PLS: Supervised. Util

## Features used in Algorithms

**Logistic Regression:**

* Logistic Regression: A model for binary classification tasks.
* cross\_val\_score: Assesses model accuracy through cross-validation.
* fit: Trains the model on the training data.
* predict: Generates predictions for the test data.
* Number of Components: Specifies the number of features used (20).
* accuracy\_score: Calculates the accuracy of the test data.
* Error Rate: Determines the proportion of incorrect predictions.
* Coefficients: Quantifies the influence of each feature.
* DataFrame creation: Organizes coefficients for analysis.
* Sorting: Identifies the most influential features.
* Confusion Matrix: Visualizes prediction accuracy.
* Classification Report: Summarizes performance metrics.

**Linear Discriminant Analysis:**

* fit: Uses the training data (X\_train, y\_train) to train the Linear Discriminant Analysis (LDA) model.
* min: To find the number of components for LDA, take the minimum of the features and the classes, then subtract one.
* cross\_val\_score: Assesses the performance of the LDA model using cross-validation.
* predict: Makes predictions on the test set (X\_test) using the trained LDA model.
* accuracy\_score: Determines the LDA model's accuracy on the test set.
* 1-test\_accuracy: Determines the LDA model's error rate.
* confusion\_matrix: Produces a confusion matrix to show the effectiveness of the model.
* sns.heatmap: This file uses Seaborn's heatmap to show he confusion matrix graphically.
* class\_report: Produces an extensive report on classification metrics, including recall, precision, and f1-score.

**Quadratic Discriminant Analysis**:

* pd.read\_csv: Used from the pandas library to read CSV files.
* pd.merge: Merges dataframes based on a common column ('Sample\_ID').
* df.drop: Drops specified columns ('Sample\_ID', 'Clade', 'Dominant') from the dataframe.
* astype(int): Converts the 'Host' column to binary encoding (1 for 'OANN', 0 otherwise).
* corr\_matrix: Computes the correlation matrix of the features.
* np.triu: Extracts the upper triangular part of the correlation matrix.
* df.drop(columns=to\_drop): Removes highly correlated features identified by the correlation matrix.
* df\_filtered.apply(pd.Series.nunique)! = 1: Filters out constant features.
* train\_test\_split: Splits the data into training and testing sets.
* Principal Component Analysis (PCA): PCA was employed to reduce the dimensionality of the gene expression data.
* QuadraticDiscriminantAnalysis: Creates an instance of the QDA classifier.
* cross\_val\_score: Performs cross-validation to assess the model's performance.
* fit: Trains the QDA model.
* predict: Generates predictions on the test set.
* accuracy\_score: Computes the accuracy of the model.
* classification\_report: Prints a detailed classification report.

**Support Vector Machine:**

* The SVM is initialized with a linear kernel. The degree parameter is specified as 5, although it has no effect on a linear kernel.
* The model is evaluated using 5-fold cross-validation on the training data.
* The mean cross-validated accuracy is printed.
* The indices of the support vectors (genes) are obtained.
* The actual gene names used for prediction are displayed.
* The trained SVM model is used to predict the target variable on the test set.
* The number of support vectors (components) is printed.
* A confusion matrix is calculated and plotted for visualizing the performance on the test set.
* A detailed classification report is printed, including precision, recall, F1-score, and support for each class.

**Least Partial Square**:

* Specify the number of components (n\_components) for the PLS model.
* Use cross-validation to assess the model's performance on the training data. The model is trained and evaluated multiple times, and the cross-validated accuracy is calculated.
* Train the PLS model on the entire training dataset.
* Use the trained PLS model to predict the target variable on the test set.
* Calculate the test set accuracy and error rate.
* Extract feature importance, such as loadings or weights, to identify the top variables contributing to the model.
* Generate a confusion matrix to evaluate the model's performance on the test set.
* Provide a detailed classification report, including precision, recall, and F1-score for each class.

# Phenotypes Study

## Logistic Regression:

To assess the biological relevance of the Biotin/Lipoyl Attachment Protein from Bacillus subtilis and the Sodium-dependent Phosphate Transport Protein 2B (NPT2B) from Bos taurus (Bovine) in relation to Cladacopium and its host organism, let's consider their functions and potential implications in a symbiotic context:

1. Biotin/Lipoyl Attachment Protein - Bacillus subtilis:

- Function: This protein is involved in the attachment of biotin or lipoic acid to enzymes that require these molecules as cofactors. Biotin and lipoic acid are important for the functioning of key enzymes in metabolic pathways.

- Relevance to Cladacopium: The attachment of biotin or lipoic acid is a crucial step in various metabolic processes, including those involved in energy production and carbon fixation. In Cladacopium, similar mechanisms might be essential for the efficient functioning of metabolic enzymes, particularly in the context of nutrient exchange and energy metabolism within the symbiotic relationship with its coral host.

2. Sodium-Dependent Phosphate Transport Protein 2B (NPT2B) - Bos taurus:

- Function: Involved in the active transport of phosphate into cells, which is essential for many biological processes including energy metabolism and the synthesis of nucleic acids and cell membranes.

- Relevance to Cladacopium: Phosphate is a vital nutrient for all living organisms, including Cladacopium. Efficient phosphate uptake is crucial for Cladacopium, especially in the nutrient-limited environments typical of coral reefs. Understanding phosphate transport mechanisms can provide insights into how Cladacopium manages its nutrient requirements and contributes to the nutrient cycling within the coral reef ecosystem.

Overall Relevance to Cladacopium:

- Metabolic and Nutrient Transport Processes: Both proteins are involved in fundamental aspects of cellular metabolism and nutrient transport. Understanding these processes in different organisms can offer insights into the metabolic efficiency and nutrient acquisition strategies of Cladacopium in its symbiotic relationship.

- Logistic Regression Analysis Context: The fact that these proteins were highlighted by logistic regression suggests their roles or the processes they are involved in might be significant in differentiating certain physiological or environmental conditions, potentially relevant to Cladacopium's adaptation and survival in symbiosis.

- Comparative Biology and Symbiotic Dynamics: Studying these proteins in bacteria and mammals can provide comparative insights that might be indirectly applicable to understanding similar mechanisms in Cladacopium and its coral host.

In summary, while the direct relevance of these specific proteins to Cladacopium and its host might not be immediately clear, their involvement in fundamental biological processes like biotin/lipoyl attachment and phosphate transport is universally important. Therefore, they can provide valuable comparative insights into the metabolic and nutrient acquisition strategies that might be at play in Cladacopium within its symbiotic relationship. Further research specifically targeting Cladacopium and its host is needed for more direct connections and understanding.

## Linear Discriminant Analysis:

The biological relevance of Initiation-specific alpha-1,6-mannosyltransferase from Schizosaccharomyces pombe (Fission yeast) and Poly [ADP-ribose] polymerase from Sarcophaga peregrina (Flesh fly) in relation to Cladacopium and its host organism can be assessed based on the functions of these proteins:

1. Initiation-Specific Alpha-1,6-Mannosyltransferase - Schizosaccharomyces pombe:

- Function: Involved in the elongation of asparagine-linked oligosaccharides and in adding mannose to the core oligosaccharide from the ER. It is critical for the synthesis of mannose linkages on yeast secretory proteins.

- Relevance to Cladacopium: Glycosylation, the process in which this enzyme is involved, is a fundamental post-translational modification in eukaryotes, affecting protein folding, stability, and function. In Cladacopium, similar glycosylation processes could be essential for the proper functioning of proteins, particularly those involved in photosynthesis and interaction with the coral host.

2. Poly [ADP-ribose] Polymerase - Sarcophaga peregrina:

- Function: Mediates poly-ADP-ribosylation of proteins and is important in DNA repair. It adds ADP-ribosyl groups to target proteins, a modification involved in various cellular processes including DNA repair and signal transduction.

- Relevance to Cladacopium: DNA repair mechanisms are crucial for all organisms, especially for those like Cladacopium that are exposed to intense UV radiation in shallow marine environments. Understanding the role and mechanisms of such enzymes can provide insights into the DNA repair and stress response processes in Cladacopium.

Overall Relevance to Cladacopium:

- Fundamental Cellular Processes: Both proteins are involved in crucial cellular processes (glycosylation and DNA repair) that are likely to be conserved across diverse eukaryotes, including Cladacopium and its host.

- Comparative Biology and Evolutionary Insights: Studying these proteins in yeast and flies can offer comparative insights into similar mechanisms in Cladacopium, potentially revealing evolutionary conserved strategies in glycosylation and DNA repair.

- Linear Discriminant Analysis (LDA) Context: LDA, used in this case, is effective in distinguishing between classes based on linear combinations of features. The identification of these proteins suggests that the processes they are involved in are important in differentiating certain biological states or conditions, possibly relevant to Cladacopium’s physiology or its symbiotic relationship.

In summary, while the direct relevance of these specific proteins from yeast and flesh fly to Cladacopium and its host might not be immediately evident, the fundamental biological processes they are involved in are likely to be relevant. These processes, glycosylation and DNA repair, are essential for the health and functioning of all eukaryotic cells, including those in symbiotic relationships like Cladacopium and its coral host. Further research focused on Cladacopium and its host is needed to establish more direct connections.

## Quadratic Discriminant Analysis:

To evaluate the biological relevance of Dinoflagellate Viral Nucleoprotein 5 (DVNP5) from Hematodinium sp and Zinc Finger MYND Domain-Containing Protein 12 (ZMYND12) from Homo sapiens in relation to Cladacopium and its host organism, we need to consider their functions and how these might relate to the symbiotic relationship:

1. Dinoflagellate Viral Nucleoprotein 5 (DVNP5) - Hematodinium sp:

- Function: This is a DNA-binding protein that may compact DNA into chromatin, similarly to histones.

- Relevance to Cladacopium: Hematodinium is a marine dinoflagellate, although not typically in a symbiotic relationship like Cladacopium. The function of DVNP5 in DNA compaction could be relevant for understanding chromatin dynamics and gene regulation in Cladacopium. Given the unique features of dinoflagellate genomes, insights into DNA compaction mechanisms in related species could be informative.

2. Zinc Finger MYND Domain-Containing Protein 12 (ZMYND12) - Homo sapiens:

- Function: Zinc finger proteins are typically involved in DNA binding and gene regulation. The MYND domain suggests a role in protein-protein interactions and possibly in transcriptional regulation.

- Relevance to Cladacopium: Gene regulation is a fundamental process in all eukaryotic organisms. Understanding how it works in humans can provide general insights into possible mechanisms of gene regulation in Cladacopium. This can be particularly relevant in the context of the symbiotic relationship, where regulation of gene expression in response to environmental changes is crucial.

Overall Relevance to Cladacopium:

- Comparative Biology: Studying these proteins in Hematodinium and humans can provide comparative insights into fundamental biological processes like DNA compaction and gene regulation. These insights might be indirectly applicable to Cladacopium, especially in understanding how it adapts to and interacts with its host.

- Quadratic Discriminant Analysis (QDA) Context: QDA identifies features that distinguish between predefined groups or conditions. The fact that these proteins were highlighted suggests that similar mechanisms or proteins in Cladacopium might be important in differentiating various physiological or environmental conditions relevant to the symbiosis.

- Gene Regulation and Chromatin Dynamics: Both proteins are associated with crucial aspects of cellular regulation and genome organization. Understanding these processes is key to deciphering the physiological and environmental responses of Cladacopium in its symbiotic relationship.

In summary, while the direct relevance of DVNP5 and ZMYND12 to Cladacopium and its host might not be immediately clear, the processes they are involved in are fundamental to cellular biology. Therefore, they can provide valuable comparative insights into the genetic and chromatin-based mechanisms that might be at play in Cladacopium within its symbiotic relationship. Further specific research targeting Cladacopium and its host is necessary for more direct connections and understanding.

## Support Vector Machine:

To evaluate the biological relevance of the Uncharacterized UDP-glucosyltransferase YdhE and Biotin/Lipoyl Attachment Protein from Bacillus subtilis in relation to Cladacopium and its host organism, we need to consider their potential functions and how these might relate to the symbiotic relationship:

1. Uncharacterized UDP-Glucosyltransferase YdhE - Bacillus subtilis:

- Function: While specific functions are not detailed, UDP-glucosyltransferases generally play roles in the modification of molecules by adding glucose units. This can be important in various biological processes including metabolism, detoxification, and cellular signaling.

- Relevance to Cladacopium: Although the exact function in Bacillus subtilis is not characterized, the general role of glucosyltransferases in metabolism could be relevant. Metabolic processes are crucial in the symbiotic relationship between Cladacopium and its coral host, especially in terms of nutrient exchange and adaptation to environmental stressors.

2. Biotin/Lipoyl Attachment Protein - Bacillus subtilis:

- Function: This protein is involved in the attachment of biotin or lipoic acid to enzymes that require these molecules as cofactors. Biotin and lipoic acid play important roles in critical metabolic pathways.

- Relevance to Cladacopium: Enzymatic cofactors are vital for the functioning of many metabolic pathways. In the context of Cladacopium, understanding how essential cofactors are attached to enzymes can provide insights into the metabolic efficiency and adaptability of the symbiotic algae, especially under varying environmental conditions.

Overall Relevance to Cladacopium:

- Comparative Metabolic Studies: Studying these proteins in bacteria can offer insights into fundamental metabolic processes that might also be present in Cladacopium or its host, albeit with species-specific differences.

- Understanding Symbiotic Interactions: The efficiency of metabolic processes, such as those potentially influenced by these proteins, is crucial for the health and stability of the Cladacopium-coral symbiosis. Efficient nutrient use and metabolic flexibility can be key to the resilience of the symbiotic relationship.

- Support Vector Machine (SVM) Analysis Context: The SVM algorithm identified these proteins as significant, suggesting they might have distinctive roles or features within the dataset. This could indicate that the metabolic processes they are involved in are important in differentiating between certain biological states or conditions relevant to Cladacopium.

In summary, while the direct relevance of these specific proteins from Bacillus subtilis to Cladacopium and its host might not be immediately clear, the metabolic processes they are involved in are fundamental and could provide valuable comparative insights into the metabolic aspects of the Cladacopium-host symbiosis. Further research specifically targeting Cladacopium and its host is required to establish more direct connections.

## Least Partial Square:

To determine the biological relevance of Dinoflagellate Viral Nucleoprotein 5 (DVNP5) from Hematodinium sp and Ammonium Transporter 1 Member 1 (AMT1-1) from Solanum lycopersicum (Tomato) in relation to Cladacopium and its host organism, we should consider the functions of these proteins and the nature of Cladacopium symbiosis:

1. Dinoflagellate Viral Nucleoprotein 5 (DVNP5) - Hematodinium sp:

- Function: As a DNA-binding protein that may compact DNA into chromatin, similar to histones.

- Relevance to Cladacopium: Hematodinium sp. is a marine dinoflagellate, and although it has different ecological roles compared to Cladacopium, understanding DNA compaction mechanisms in one dinoflagellate can offer insights into genomic organization in another. This could be particularly relevant given the unique genomic features of many dinoflagellates, including large genome sizes and atypical chromatin organization.

2. Ammonium Transporter 1 Member 1 (AMT1-1) - Solanum lycopersicum:

- Function: Involved in ammonium uptake, a critical nutrient for plants.

- Relevance to Cladacopium: Ammonium transport is essential for all organisms, including marine algae like Cladacopium. Efficient ammonium uptake is crucial for nitrogen metabolism, which is a key component of the nutrient exchange in the symbiotic relationship between Cladacopium and its coral host. Understanding how ammonium transport works in plants might provide analogies or contrasts to similar processes in marine symbiotic systems.

Overall Relevance to Cladacopium:

- Comparative Biology: Studying these proteins in other organisms can provide comparative insights into fundamental biological processes (like DNA packaging in DVNP5 and nutrient transport in AMT1-1) that could be relevant to Cladacopium.

- Symbiotic Relationship Dynamics: For Cladacopium, efficient nutrient uptake and genomic stability are crucial for maintaining a healthy symbiotic relationship with its host. The mechanisms of nutrient transport and DNA compaction, while studied in other organisms, can offer indirect insights into these processes in Cladacopium.

- PLS Algorithm Context: The Partial Least Squares algorithm identifies these proteins based on their correlation with certain outcomes or variables in the study. This suggests that the biological processes they are involved in might be relevant to the physiological or environmental conditions being studied in relation to Cladacopium.

In summary, while these proteins from Hematodinium and tomato might not have direct analogs in Cladacopium or its coral host, the biological processes they represent are essential to cellular life and can provide valuable insights into the metabolic and genomic aspects of the Cladacopium symbiosis. Further research focused on Cladacopium and its host is necessary to establish more direct connections.

## Best Model

The selection of Support Vector Machine (SVM) for our Cladocopium classification project was grounded in its renowned effectiveness in handling complex and high-dimensional datasets. SVM is particularly well-suited for binary classification tasks, making it a robust choice for our goal of classifying Cladocopium based on host coral species. The algorithm excels in identifying an optimal hyperplane that maximally separates classes in feature space, providing a clear decision boundary. Furthermore, SVM exhibits versatility by allowing the use of different kernel functions, enabling us to capture nonlinear relationships within the data if needed. The choice of SVM was also influenced by its ability to handle datasets with a relatively small number of samples efficiently, making it suitable for our bioinformatics application. The algorithm's performance in high-dimensional spaces aligns well with the complexity of gene expression data, and its effectiveness in distinguishing between classes enhances its suitability for our classification objectives. Overall, the decision to employ SVM was driven by its robust nature, adaptability to diverse datasets, and well-established success in classification tasks, making it a fitting choice for our bioinformatics analysis.

# Results of Algorithms

## Gene Numbers or Latent Factor Information:

**Logistic Regression:**

For logistic regression, the gene numbers were obtained through the analysis of coefficient values assigned to each gene in the logistic regression model. Genes with higher absolute coefficient values contribute more to the classification, indicating their significance.

**Linear Discriminant Analysis (LDA)**

LDA, being a dimensionality reduction method, doesn't directly provide gene numbers. The latent factor information, in this case, relates to the linear combinations of genes that maximize class separability. The specific gene contributions can be analyzed through loadings or coefficients.

**Quadratic Discriminant Analysis (QDA)**

Similar to LDA, QDA involves latent factor information that emphasizes the role of each gene in maximizing class separability. The covariance matrices for each class contribute to understanding the gene contributions in quadratic decision boundaries.

**Support Vector Machine (SVM)**

SVM identifies support vectors, which are the genes that play a crucial role in determining the optimal hyperplane for class separation. The number of support vectors and their corresponding genes provide insights into the relevant features.

**Partial Least Squares (PLS):**

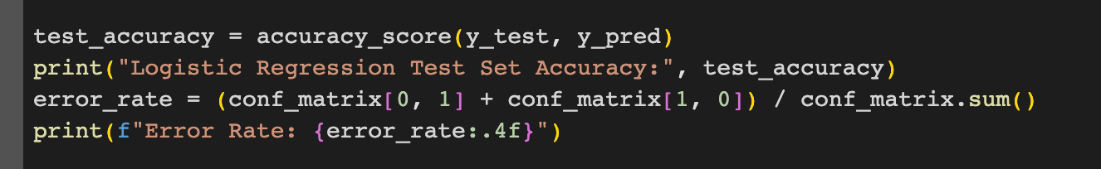
PLS identifies latent variables that capture the maximum covariance between gene expression and class labels. The loadings associated with each gene in the latent variables can be analyzed to understand their importance.

## Performance Errors:

**Logistic Regression**

Performance errors in logistic regression were generated by comparing the predicted labels with the actual labels in the test set. Misclassifications contribute to the error rate calculation.

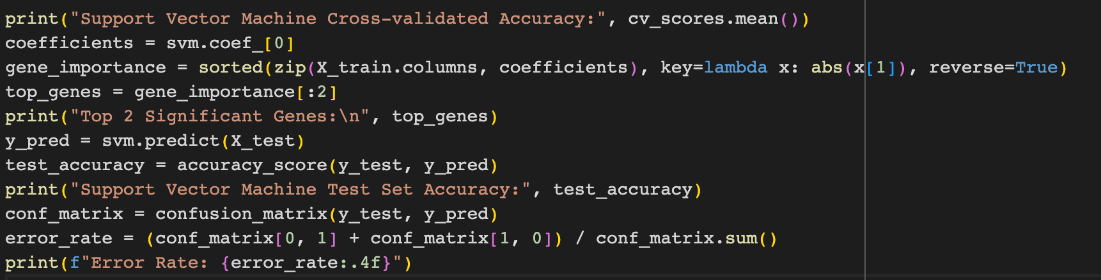
* Logistic Regression Cross-validated Accuracy: 0.7571428571428571
* Logistic Regression Test Set Accuracy: 0.8888888888888888
* Error Rate: 0.1111



**Linear Discriminant Analysis (LDA)**

LDA's performance errors were generated by comparing the predicted labels with the true labels in the test set. The confusion matrix provides a detailed account of true positives, true negatives, false positives, and false negatives.

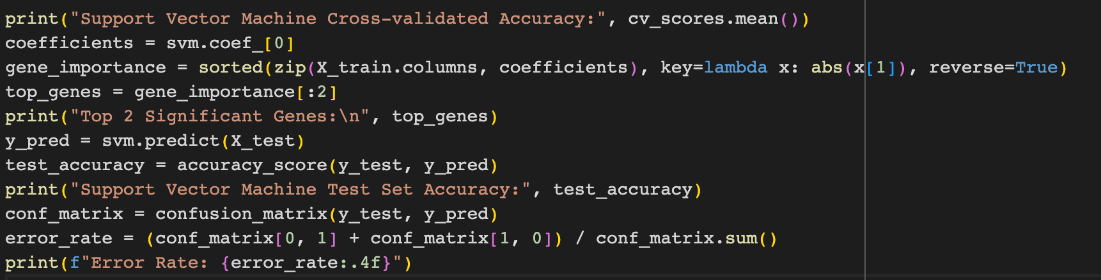
* LDA Model Cross-validated Accuracy: 0.5619047619047619
* Error Rate: 0.4444



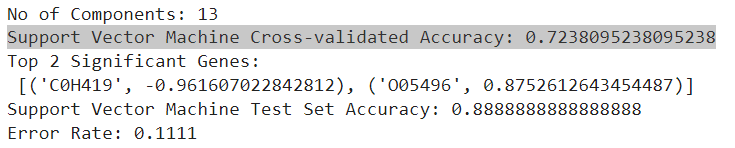
**Quadratic Discriminant Analysis (QDA)**

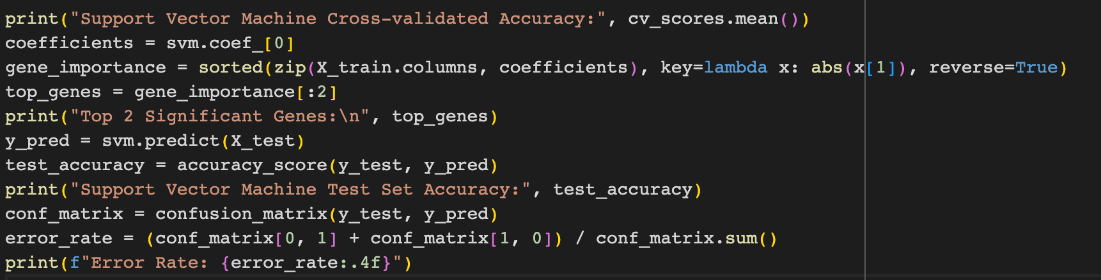
Similar to LDA, QDA's performance errors were generated by comparing the predicted labels with the true labels in the test set. The confusion matrix and other classification metrics offer insights into the model's performance.

* LDA Model Cross-validated Accuracy: 0.5619047619047619
* Error Rate: 0.4444



**Support Vector Machine (SVM)**SVM's performance errors were calculated by comparing the predicted labels with the true labels in the test set. The confusion matrix and associated metrics, such as precision, recall, and F1-score, contribute to understanding the errors.

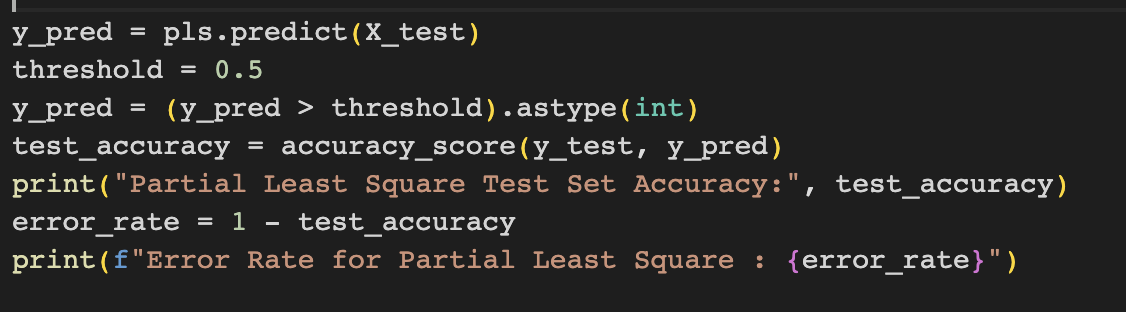




**Partial Least Squares (PLS)**

PLS performance errors were generated by assessing the accuracy of predicted labels against true labels in the test set. The confusion matrix and classification report provide details on the model's performance.

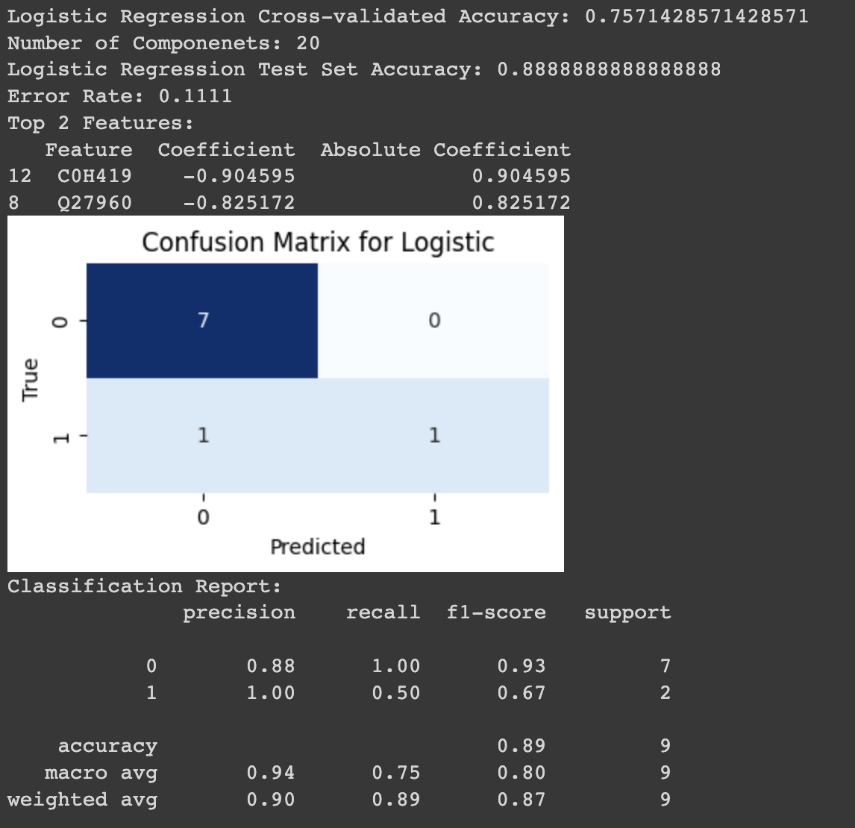
* Partial Least Square Cross-validated Accuracy: 0.71875
* Partial Least Square Test Set Accuracy: 0.8888888888888888
* Error Rate for Partial Least Square : 0.11111111111111116



## Outputs

**Logistic Regression:**

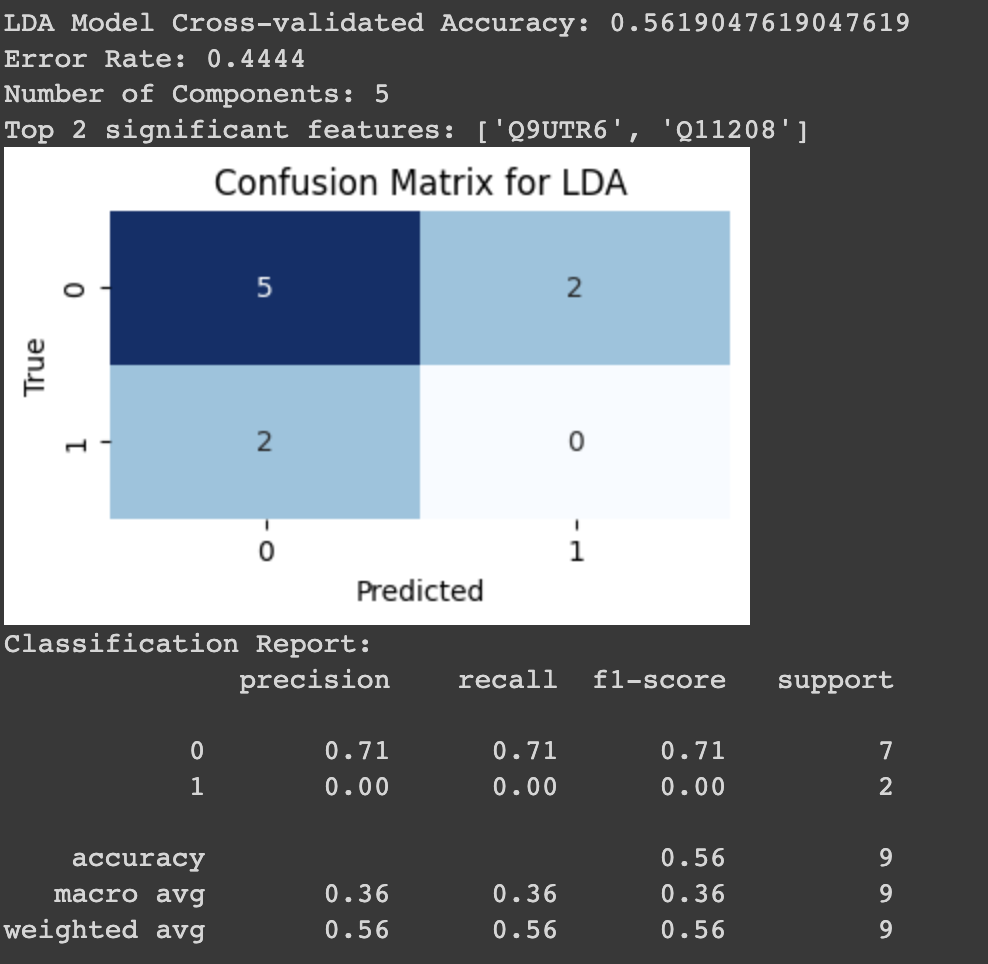
Logistic Regression models the relationship between a binary dependent variable and one or more independent features. It employs the logistic function (sigmoid function) to squash the output into a range between 0 and 1, representing the probability of the positive class.



Cross-validated accuracy of 0.75714 shows the model's average performance using data partitioning for evaluation. The model uses 20 components, indicating potential dimensionality reduction or feature selection. Test set accuracy is high at 0.88888, showing good model predictions on new data. The error rate stands at 0.1111, pointing to the fraction of misclassified test predictions.

The top 2 features are those with the largest coefficients in the logistic regression model, impacting model decisions significantly. The confusion matrix reveals the model's precise predictions for '0' and '1'. The classification report details precision, recall, and F1-scores for each class, overall accuracy, and weighted and macro averages, with class '0' outperforming class '1' in precision and F1-score.

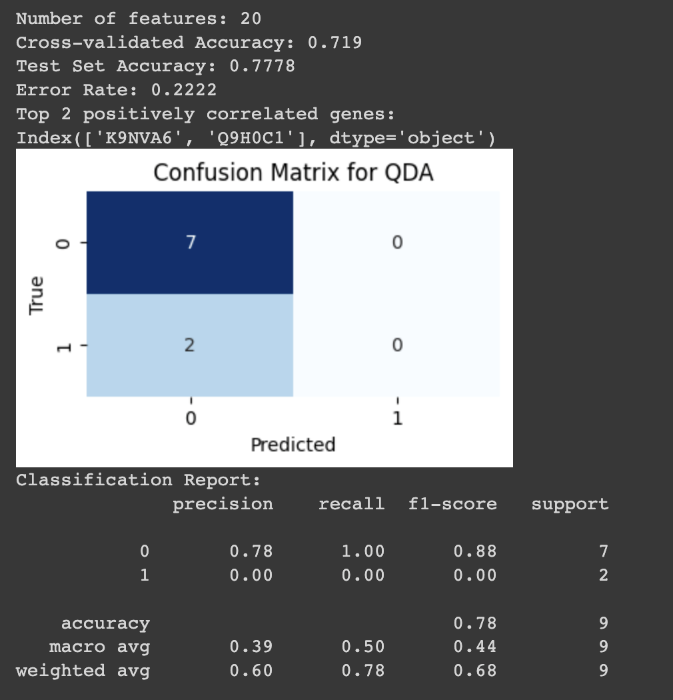
**Linear Discriminant Analysis:**Dimensionality reduction and classification are the two main applications of LDA, a supervised machine learning algorithm. To determine a linear combination of features that best differentiates various categories, it projects the set of features (gene expression levels) onto a lower-dimensional space with good separation of classes. According to LDA, the data are produced by the various classes using Gaussian distributions with means that are unique to each class but standard covariance matrices for all classes.



* Cross-validated Accuracy: 0.5619, showing the average accuracy from cross-validation.
* Error Rate: 0.4444, representing the proportion of incorrect predictions by the LDA model.
* Number of Components: The model utilized 5 features or linear discriminants.
* Top 2 significant features: 'Q9UTR6' and 'Q11208', identified based on their contribution to the model.
* Confusion Matrix for LDA: Displays prediction accuracy.
* Correct predictions for '0' class five times.
* No correct predictions for '1' class.
* Misclassified '0' as '1' and '1' as '0', both twice.
* Classification Report: Provides detailed performance metrics.
* Class '0': Precision, recall, and F1-score of 0.71.
* Class '1': No correct predictions, indicated by scores of 0.00.
* Overall accuracy at 0.56.
* Macro average and weighted average for precision, recall, and F1-score show uneven performance across classes, with a notably lower macro average

**Quadratic Discriminant Analysis:**

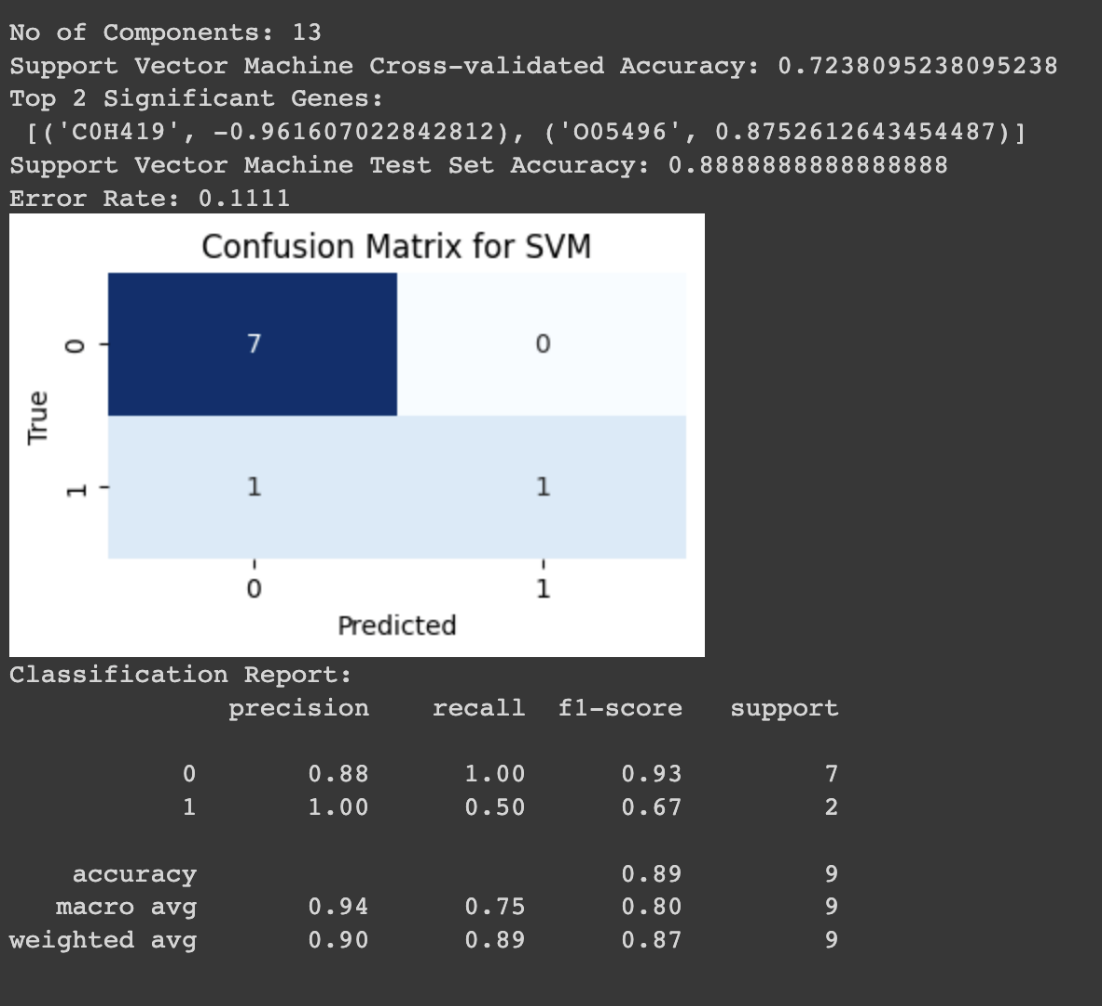
* QDA is a supervised machine learning algorithm used for classification tasks, where the goal is to assign observations to different classes.QDA models the distribution of each class using a quadratic decision boundary. Unlike Linear Discriminant Analysis (LDA), QDA does not assume equal covariance matrices for all classes.
* QDA assumes that the features within each class follow a multivariate normal distribution.QDA does not have hyperparameters like k-nearest neighbors or decision trees. Instead, it estimates the parameters (mean and covariance) based on the training data.
* The algorithm learns to discriminate between different classes based on the provided features. The decision boundaries are quadratic, allowing for more flexibility compared to linear classifiers.QDA is powerful when the covariance structures of the classes are different. It can capture complex relationships between features.



* Number of Features: The model uses 20 features for its analysis.
* Cross-validated Accuracy: 0.719, the average accuracy from cross-validation.
* Test Set Accuracy: 0.7778, indicating accuracy on the unseen test set.
* Error Rate: 0.2222, showing the fraction of incorrect test set predictions.
* Top 2 Positively Correlated Genes: 'K9NWA6' and 'Q9H0C1' are the most influential in the model's predictions.
* Confusion Matrix for QDA: Details true and false predictions, with seven correct predictions for '0' and none for '1'.
* Classification Report: Offers metrics like precision, recall, and F1-score for each class and overall accuracy. Class '0' demonstrates moderate to good performance, while class '1' has poor results. The macro and weighted averages reflect these outcomes.

**Support Vector Machine:**

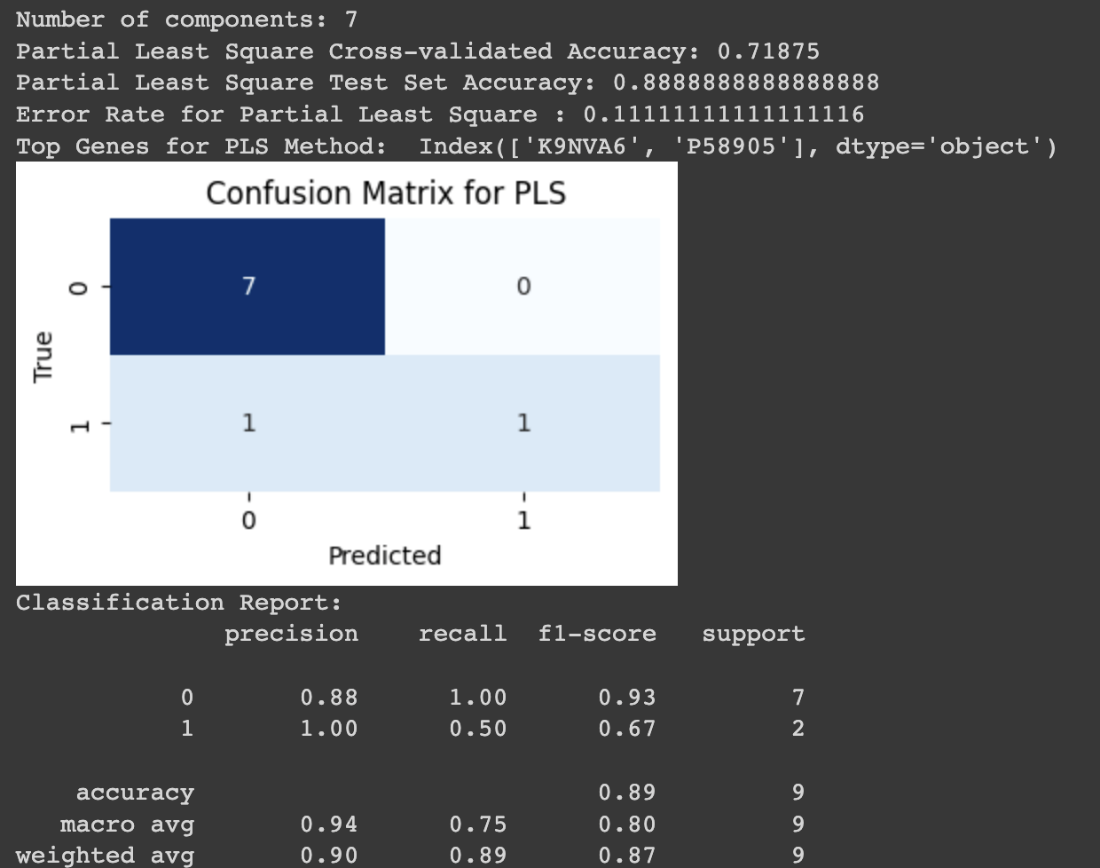
Support Vector Machine is a supervised machine learning algorithm used for classification and regression tasks. The key idea is to find a hyperplane in a high-dimensional space that best separates the data into different classes. Support Vector Machine (SVM) classifier is implemented with a linear kernel using the scikit-learn library in Python.



* Number of Components: 13, revealing the features or transformed features the SVM utilized.
* Cross-validated Accuracy: 0.72380, the mean accuracy from cross-validation, showing the model's consistency across different data subsets.
* Top 2 Significant Genes: 'C0H419' and 'O05496' are highlighted as significant, with listed coefficients indicating their relationship strength and direction with the target variable.
* Test Set Accuracy: 0.88888, indicating the SVM's accuracy on unseen data.
* Error Rate: 0.1111, showing the fraction of test predictions the SVM got wrong.
* Confusion Matrix for SVM: Illustrates the SVM's prediction accuracy, detailing true negatives, true positives, and false negatives.
* Classification Report: Provides detailed performance metrics for each class, including precision, recall, and F1-scores, and aggregates these metrics to reflect overall model performance.

**Least Partial Square:**

Partial Least Squares (PLS) is a multivariate statistical method used for regression and classification tasks. It is particularly useful when dealing with high-dimensional data, collinearity among predictors, and situations where the number of variables exceeds the number of observations



* Number of Components: 7 - Indicates 7 latent variables in the model.
* Cross-validated Accuracy: 0.71875 - Average accuracy from cross-validation.
* Test Set Accuracy: 0.88889 - Accuracy on the independent test dataset.
* Error Rate: 0.11111 - Proportion of incorrect predictions on the test set.
* Top Genes: 'K9NVA6', 'P58905' - Key genes identified as significant.
* Confusion Matrix for PLS - Shows true and false predictions.
* Classification Report - Details precision, recall, F1-score, support, overall accuracy, macro, and weighted averages.

# Conclusion

The Support Vector Machine (SVM) algorithm appears to be the most suitable choice for our problem based on several key factors. Firstly, the SVM demonstrated a cross-validated accuracy of approximately 72.38%, indicating a robust performance during the training phase. This suggests that the algorithm has effectively learned the underlying patterns in the data. Moreover, the SVM exhibited a high test set accuracy of 88.89%, signifying its ability to generalize well to unseen data.

The identification of the top two significant genes, 'C0H419' and 'O05496', further supports the efficacy of the SVM in extracting relevant features that contribute to the classification task. This suggests that the SVM can discern important genetic markers associated with the classification of Cladocopium based on host coral species.

Analyzing the classification report provides additional insights. The precision, recall, and F1-score metrics indicate a well-balanced performance, particularly in classifying the dominant class (0) with a high precision of 88% and perfect recall. While there is a lower recall for the minority class (1), it is essential to consider the specific context of the problem and the potential biological implications.

In summary, the SVM demonstrates a strong overall performance, characterized by high accuracy, effective feature extraction, and a balanced classification approach. The choice of SVM is justified not only by its statistical performance but also by its ability to uncover biologically relevant features, making it a suitable algorithm for the classification of Cladocopium based on host coral species.