Final Project Report for COMP 4910

BIO Project

Nisha Puthiyedth

Simranjit Kaur (T00605906) - Team Lead

Samuel Bedu-Annan (T00581897)

Divyatej Khurana (T00637205)

Table of Contents

Summary

Feature selection method plays a prominent role in the elimination of redundant and irrelevant data to improve the performance and reduce the cost of data analysis. Analysis of Variance (ANOVA), chi-square, Permutational Multivariate Analysis of Variance (PERMANOVA) and Support Vector Machines (SVM). In our project, feature selection methods are applied to the microbiome dataset by using the sklearn python package to analyze and compare the list of features resulting from each method in the Python programming language. 50, 100 and 200 best features were analyzed and compared for each feature selection method. Out of 200 features, 12 features were common in all of them. Cluster Maps, Scatterplots and Venn Diagrams are included to show a visual comparison of all feature selection methods. This report includes the problem description, requirements, project details, feature selection methods, analysis, design, implementation, testing, project presentations, acknowledgement, and references of our project.

Problem

Analysis and comparison of the performance of four feature selection methods which include Support Vector Machines (SVM), Chi-square, Analysis of Variance (ANOVA), Permutational multivariate analysis of variance (PERMANOVA).

Requirements

Functional Requirements

- Literature review of the feature selection methods.
- Implementation of feature selection methods in Python by using python packages.
- Searching and selecting biological dataset.
- Perform data cleaning.
- Applying the feature selection methods on the selected dataset.
- Comparison of list of features resulted from each selection method.
- Perform analysis of resulting features.
- Report writing and updating after each step.

Non-Functional Requirements

- Project shall be hosted on GitHub repository.
- Project code shall be easily compatible with PyCharm or Jupyter Notebook IDE.
- Compare execution time of all feature selection methods (removed at the end of the project. See page 28 for detailed reason).
- Properly commented code shall be written for easy maintenance.
- Implemented feature selection method shall be compatible with small or medium sized datasets.

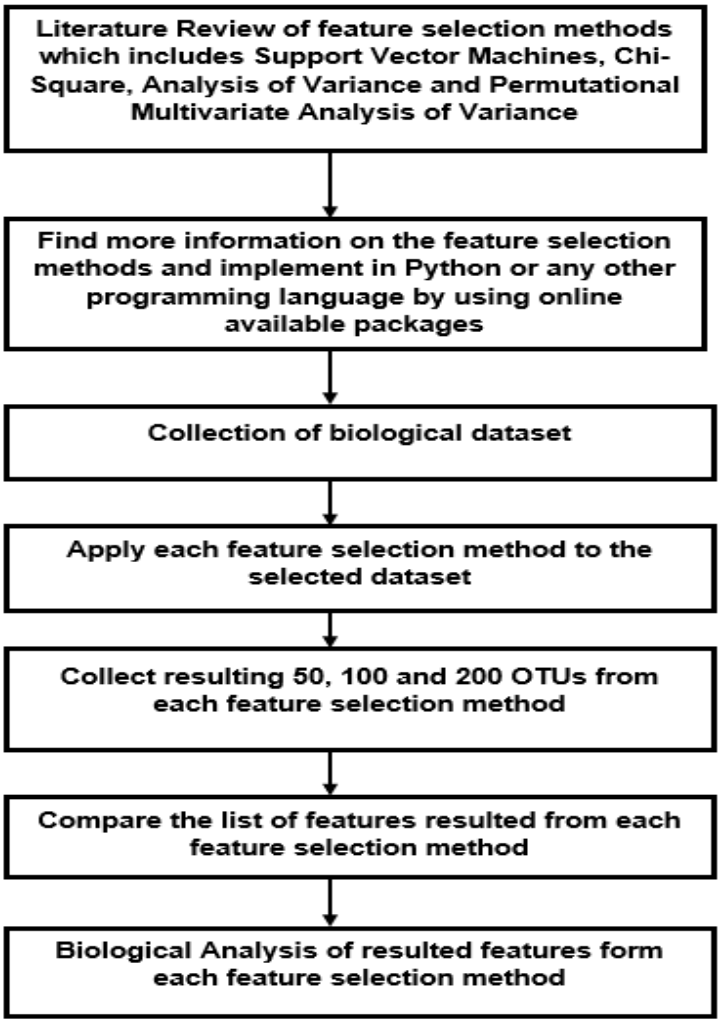
Project Details

In this project, analysis, and comparison of the performance of four feature selection methods (Support Vector Machines, chi-square, Analysis of Variance and Permutational Multivariate Analysis of Variance) was done by using Agile software development lifecycle.

First, research was conducted on the feature selection methods and their importance in machine learning. The project was mainly divided into 6 phases. In phase 1 of the project, Literature Review was performed on the above-mentioned feature selection methods by using online journals, peer-reviewed articles, and other information available online. More information on the feature selection methods was collected and implemented in python in the 2nd phase of the project. Online available python packages were used for the implementation of the feature selection methods.

Phase 3 of the project was to find the biological dataset for the project. The dataset was provided by our client to work on. Next, phase 4 of the project was done which included the application of each feature selection method to the selected dataset. List of features resulting from each feature selection method were compared in the phase 5 of the project by creating Cluster Maps, Scatterplots and Venn diagrams for best 50, 100 and 200 features. Phase 6 included the biological analysis in which the resulting features were annotated, and biological information was gathered on the resulting features.

The flowchart of different phases of our project is:



Feature Selection Methods

ANOVA

ANOVA stands for “Analysis of Variance” [6]. It is a feature selection method used to compare the means of more than two groups by performing t-test. In ANOVA, group mean differences are inferred by analyzing variances [6]. It uses a variance-based F-test to check the mean equality. It also tests the null hypothesis. i.e., all group means are equal. It can be performed in two ways: One-way ANOVA (one factor) and Two-way ANOVA (factor is an independent variable).

Chi-Square

chi-square test is a statistical test which measures the association between two categorical variables [9]. We can use this to statically determine whether the observed variable is dependent or independent in comparison to the expected variables. The value of x2 can be calculated by using a simple formula.

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

After calculating the x2 value we compare it to the critical value corresponding to the degrees of freedom. Degree of freedom can be calculated by (No of rows -1) *(No of Columns -1). If the value of x2 is smaller than the critical value, then the null hypothesis is true and concludes that there is NO significant association between the variables. Whereas, If the calculated value is higher than the critical value in the table, we reject the null hypothesis and conclude that there is a significant association between the variables.

PERMANOVA

PERMANOVA, which is an acronym for “Permutational Multivariate Analysis of Variance” [13], is defined as a geometric partitioning of multivariate variation in the space of a chosen dissimilarity measure based on how ANOVA is designed. It produces a set of p-values using distribution-free permutation techniques [13].

In other words, PERMANOVA is a feature selection method which compares the groups of objects based on the centroid and the dispersion of the groups and tests the null hypothesis that these are equal for all groups. It chooses the similarity based on distance measure. The rejection hypothesis of PERMANOVA is that either the centroid and/or the spread of the objects is different between the groups.

SVM

Support Vector Machine refers to a statistical and machine learning technique used on a variety of applications such as prediction [4], pattern recognition [1] and biological data processing [5]. From the diagram below, SVM works by identifying a hyperplane that separates different classes (green and blue). It constructs the hyperplane by maximizing the margin of the decision boundary based on the distance of the support vectors. For a 2D problem, it uses a 1D line. For a 3D problem, it uses a 2D plane. SVM is regarded as one of the most accurate machine learning algorithms among many others due to its high generalization ability. Studies found in the early 21st & 19th centuries on the experimental success and general features of SVM have highlighted the important role it plays in different academic fields.

Analysis

Problem and requirements

We defined a problem along with the functional and non-functional requirements of the project and validated them with the client. They are written in the Problem and Requirements sections above.

Literature Review

In our literature review, which is phase 1 of our project, we got a few articles on different feature selection methods from our client and we also found some articles on our own to get a basic understanding of how the feature selection methods worked and the purpose of each feature selection method. We wrote a summary of what we learned from those articles in our own words which will be helpful in writing the final report for our client. This report can be found here: Final Project Report.

To find general information about feature selection methods and how it has been used by others, we explored the following articles:

1. Stability of feature selection algorithm: A Review
2. Performance Comparison of Feature Selection Methods

The useful information found from these resources is written at the beginning of the Literature Review section of our Final Project Report.

Following are the links of the articles that we read for each feature selection method:

SVM

1. Support Vector-Networks
2. SVM based method for predicting HLA-DRB1∗0401 binding peptides in an antigen sequence
3. Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification

Chi-square

1. BIO-STATISTICS: A BRIEF OVERVIEW
2. Chi-Square Test is Statistically Significant: Now what?
3. Conceptual model on application of chi-square test in education and social sciences

ANOVA

1. An introduction to analysis of variance (ANOVA) with special reference to data from clinical experiments in optometry
2. Analysis of variance-why it is more important than ever
3. Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor

PERMANOVA

1. Permutational analysis of variance - Wikipedia
2. Association of Oral Microbiome with Risk for Incident Head and Neck Squamous Cell Cancer

The summary of important information obtained from these articles is written in the Literature Review section of our report which can be found in our Final Project Report.

Python Packages

We found various python packages which includes Pandas, SciPy, Pingouin, StatsModels, NumPy, Sklearn, Skbio via online tutorials, which are useful for the implementation of the feature selection methods. In phase 2 of our project, we found more information on the feature selection methods via Kaggle, GitHub and Stackoverflow, implemented them using the python packages and applied it on a sample dataset. It yielded great results for most of the feature selection methods i.e., SVM, Chi-square, ANOVA. The code that we tried with the sample dataset can be found on our GitHub repository.

Likewise, we used the existing code from the sample dataset on the approved dataset for the project. We used Pandas, NumPy, SciPy, Pingouin, StatsModels python libraries for the implementation of SVM, Chi-Square and ANOVA feature selection methods. The results of this implementation did not show the results that we expected. The code can be found here on our GitHub repository.

Then, we decided to use the Sklearn python library which shows great results. The code of successful implementation of SVM, Chi-square and ANOVA feature selection methods can be found on our GitHub repository.

We tried some online tutorials for the implementation of the PERMANOVA feature selection method with the help of the Skbio package in Python. We were only able to create a distance matrix of our selected dataset. The code for creating the distance matrix from our dataset can be found on our GitHub repository. We were not able to implement the PERMANOVA feature selection method properly. So, our client provided us with the results of the PERMANOVA feature selection method which were implemented using the Skbio python library.

Data collection

Phase 3 of our project is the collection of biological dataset. Our client provided us with a real dataset. We applied the feature selection methods on the selected dataset.

Dataset Details

Microbiome Dataset for Brassica and Wheat		
Samples	**Features**	**Sample Classes Names**
6	5477 (OTUs)	Wheat
5		Brassica

For our project, we worked on the microbiome dataset for Brassica and Wheat which includes 11 samples and 5477 features. The dataset was divided into two main sample classes: Wheat and Brassica. There are five Brassica samples namely Brassica_1, Brassica_2, Brassica_3, Brassica_4, Brassica_5 and six Wheat samples that are Wheat_1, Wheat_2, Wheat_3, Wheat_4, Wheat_5, Wheat_6. OTUs are essentially different soil samples which contain various microorganisms for helping in the growth of the Wheat and Brassica.

Data Preprocessing

Our client provided us with the preprocessed dataset to work on as it is real and confidential. We applied the feature selection methods on the provided dataset in python. This is phase 4 of our project. This phase consists of implementing feature selection methods on the dataset and getting the result of the implementation which has been completed for SVM, Chi-square and ANOVA. Our client provided us with the results of the PERMANOVA feature selection method.

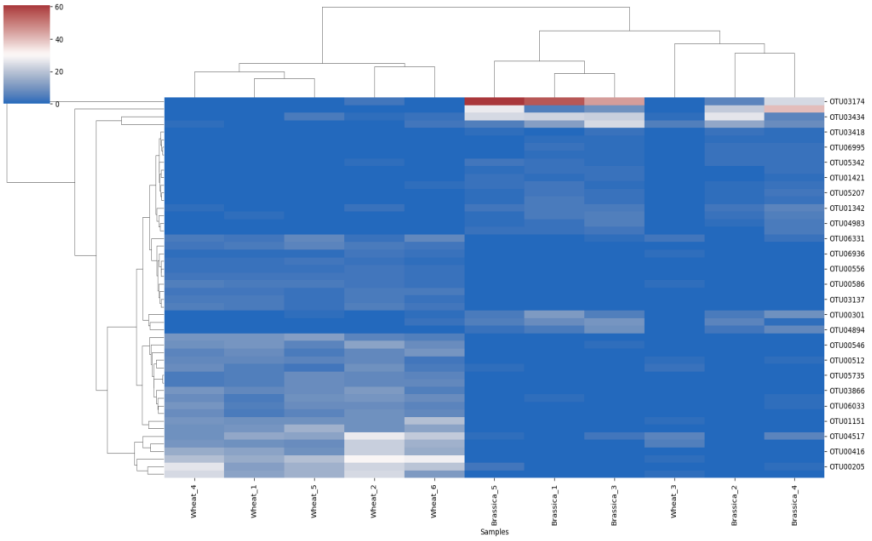
Comparison of resulting features

We did a comparison of the resulting best 50, 100 and 200 features of SVM, Chi-square, ANOVA and PERMANOVA feature selection methods by creating VENN diagrams, Scatter Plots and Heat maps. After getting feedback from our client, we needed to re-do the scatter plots for the Chi-square method and the heatmaps we created were wrong. So, we decided to create cluster maps. The heatmaps for each feature selection method that are rejected by our client can be found in the Results folder under Not Working folder of our GitHub repository.

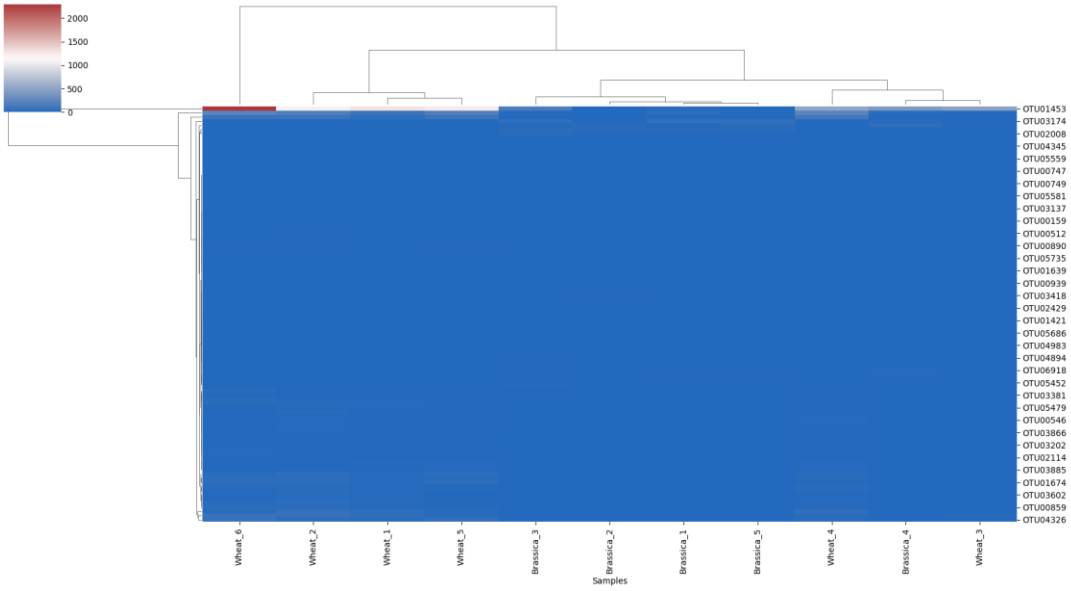
The results of successful comparison by creating Cluster maps, Scatterplots and Venn diagrams can be found in the Comparison Of Results folder on our GitHub repository. The results of each feature selection methods are explained below:

Cluster Maps

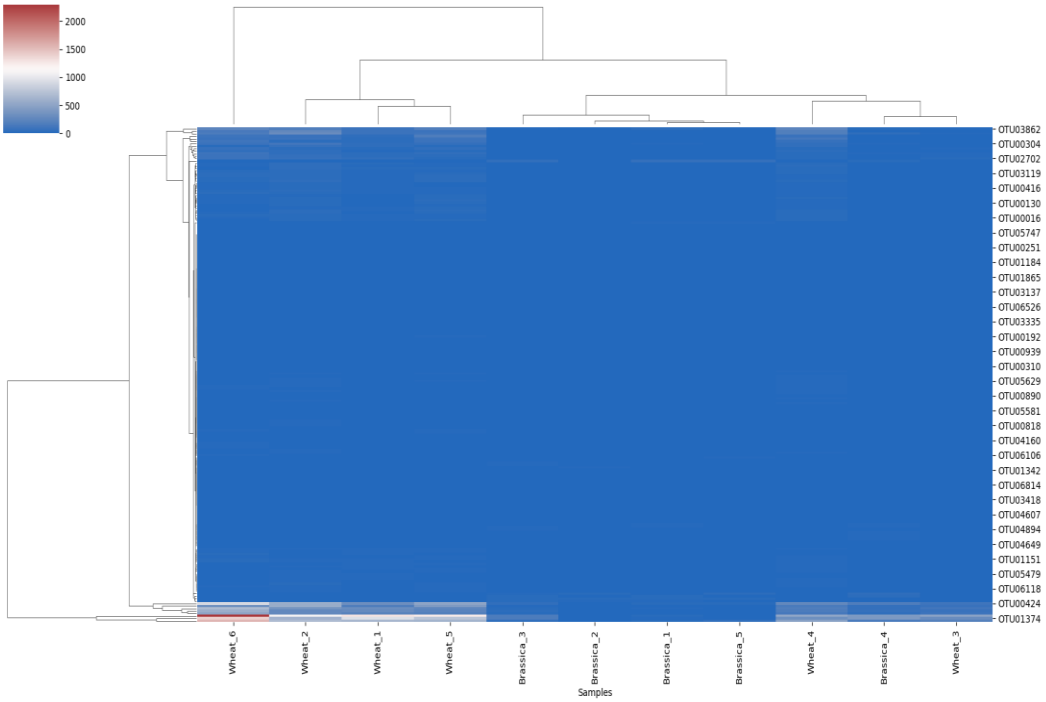
ANOVA - 50 OTUs



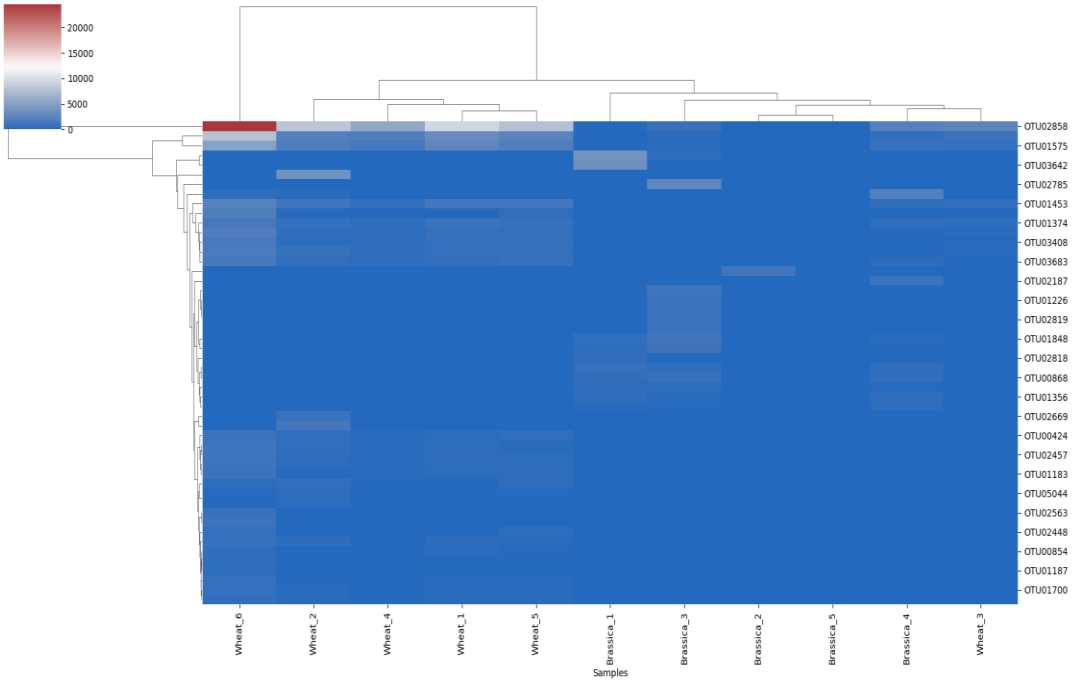
ANOVA - 100 OTUs



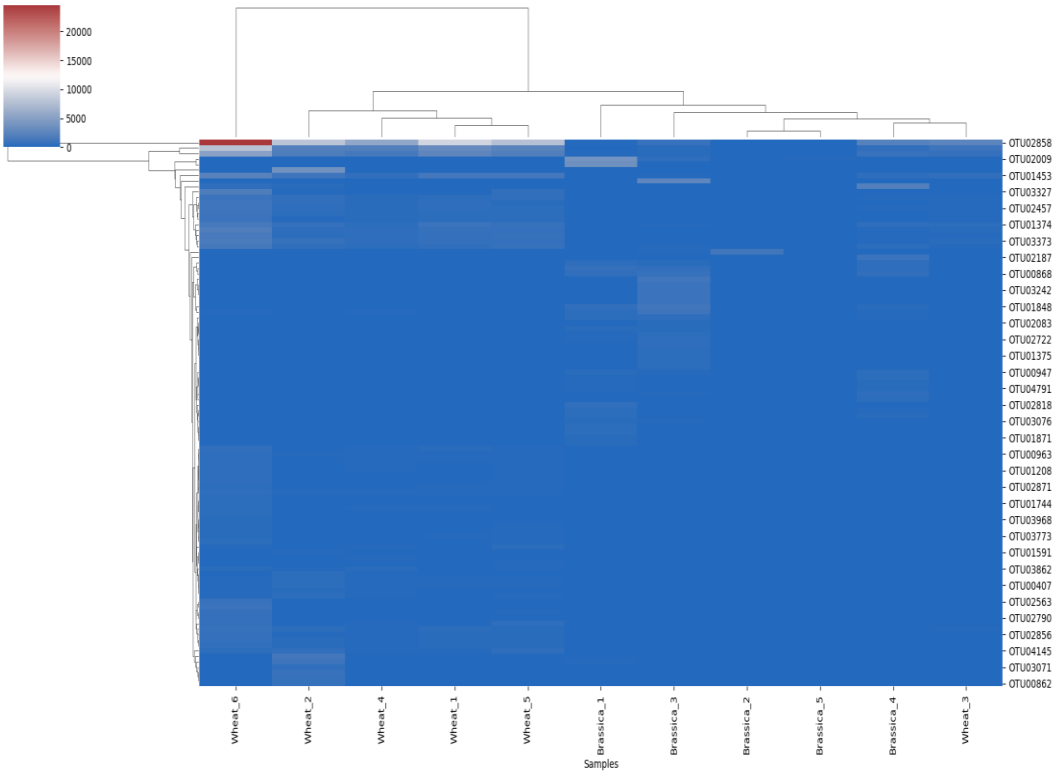
ANOVA - 200 OTUs



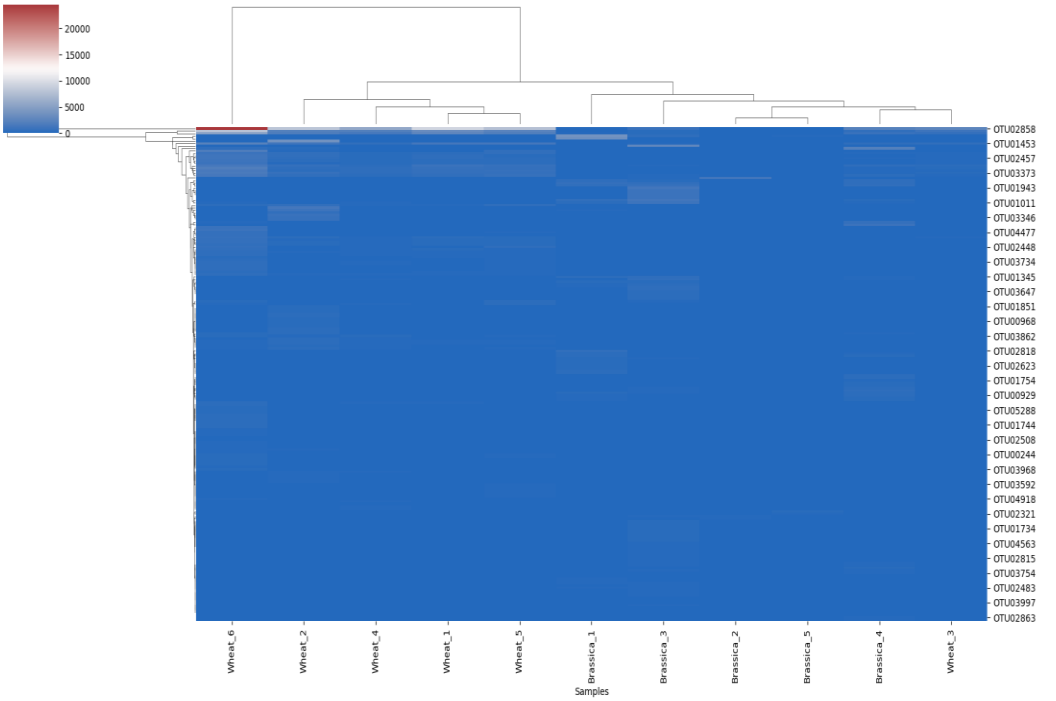
Chi-Square - 50 OTUs



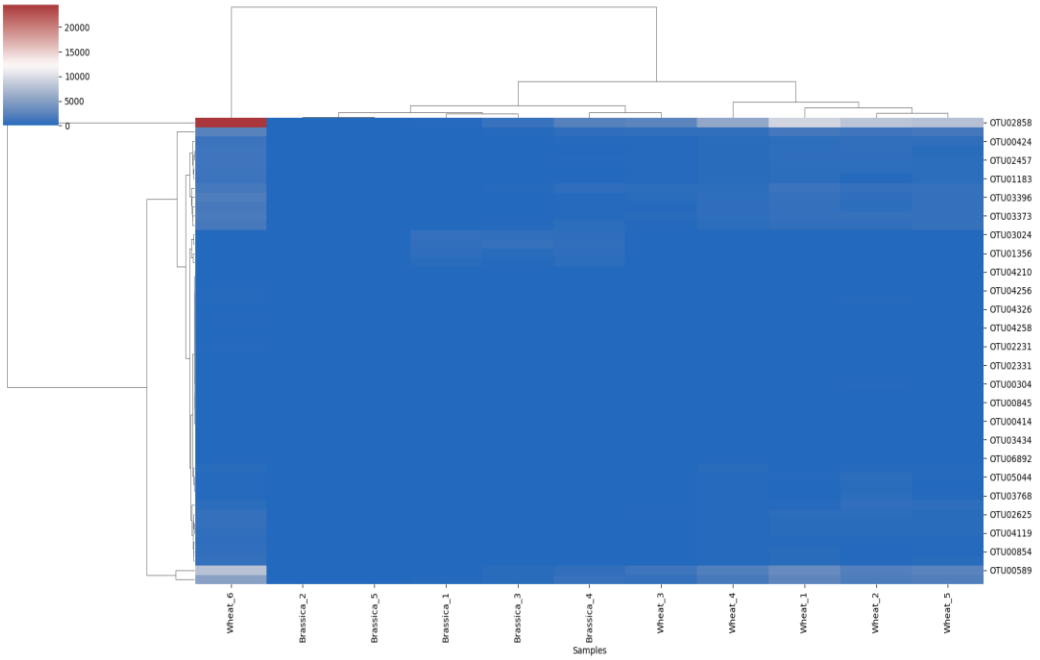
Chi-Square - 100 OTUs



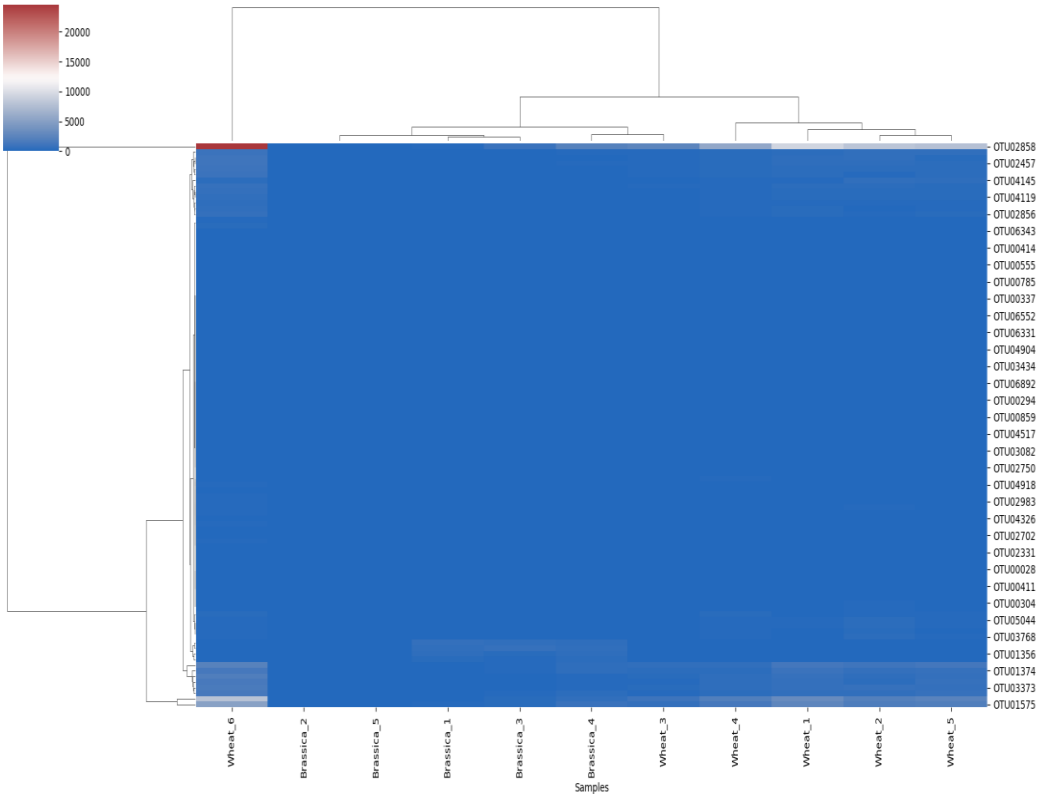
Chi-Square - 200 OTUs



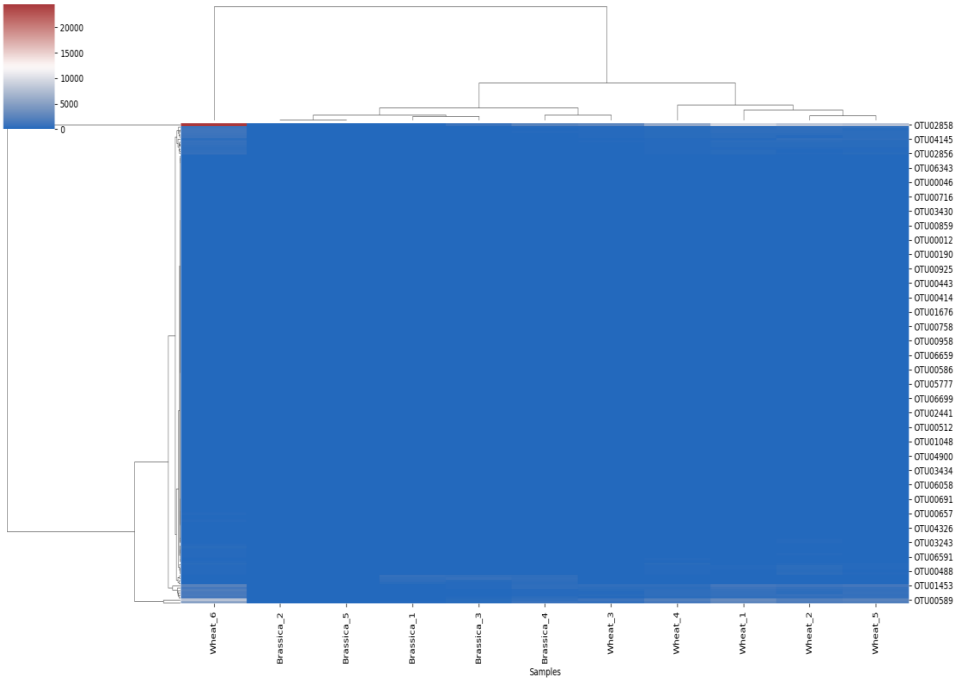
PERMANOVA - 50 OTUs



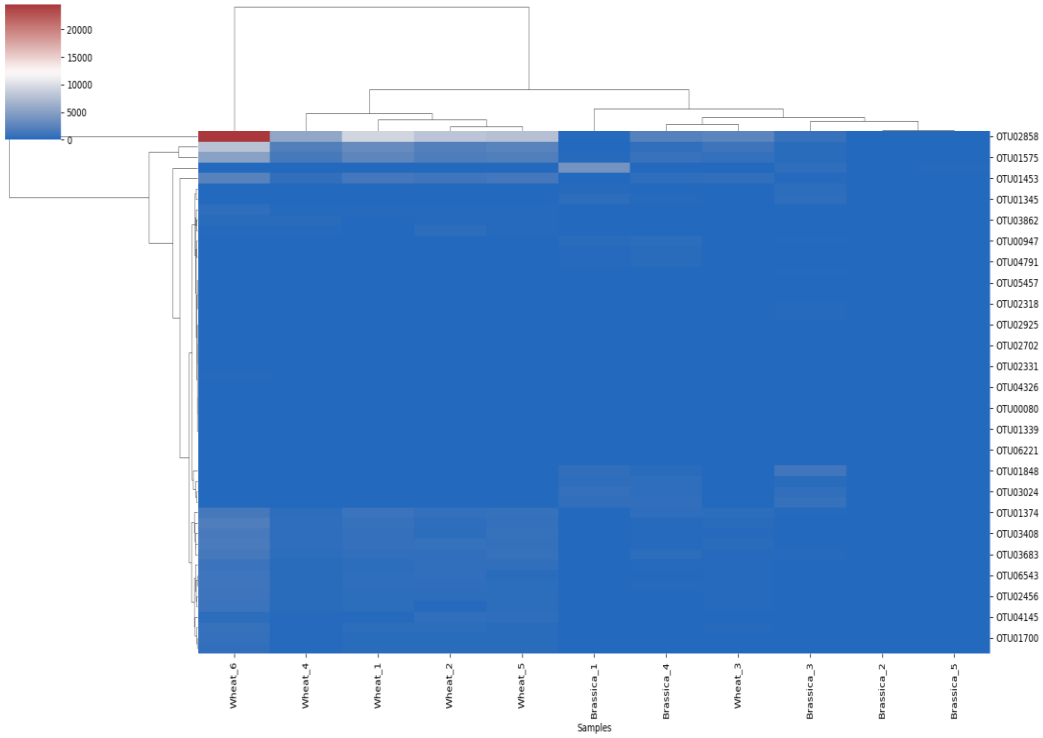
PERMANOVA - 100 OTUs



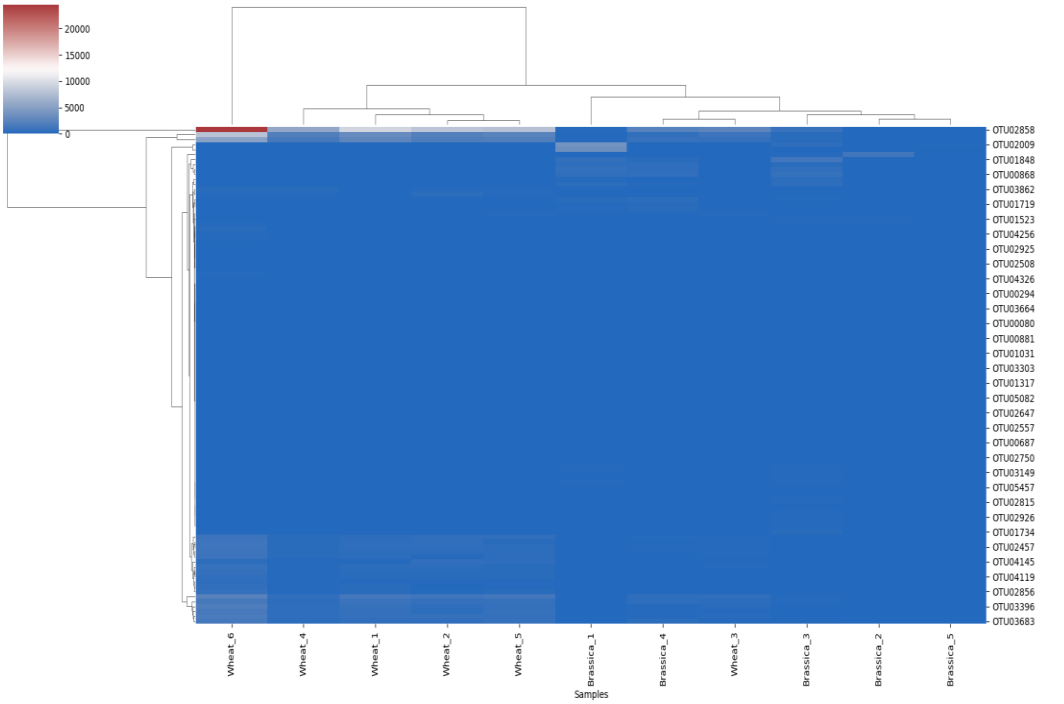
PERMANOVA - 200 OTUs



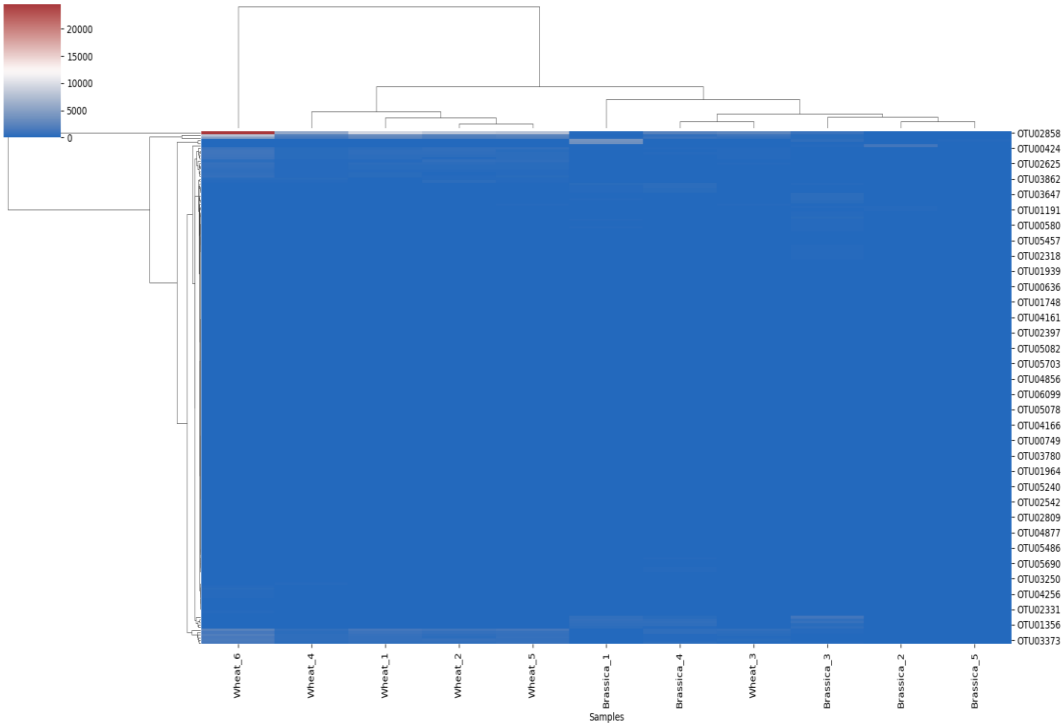
SVM - 50 OTUs



SVM - 100 OTUs



SVM - 200 OTUs

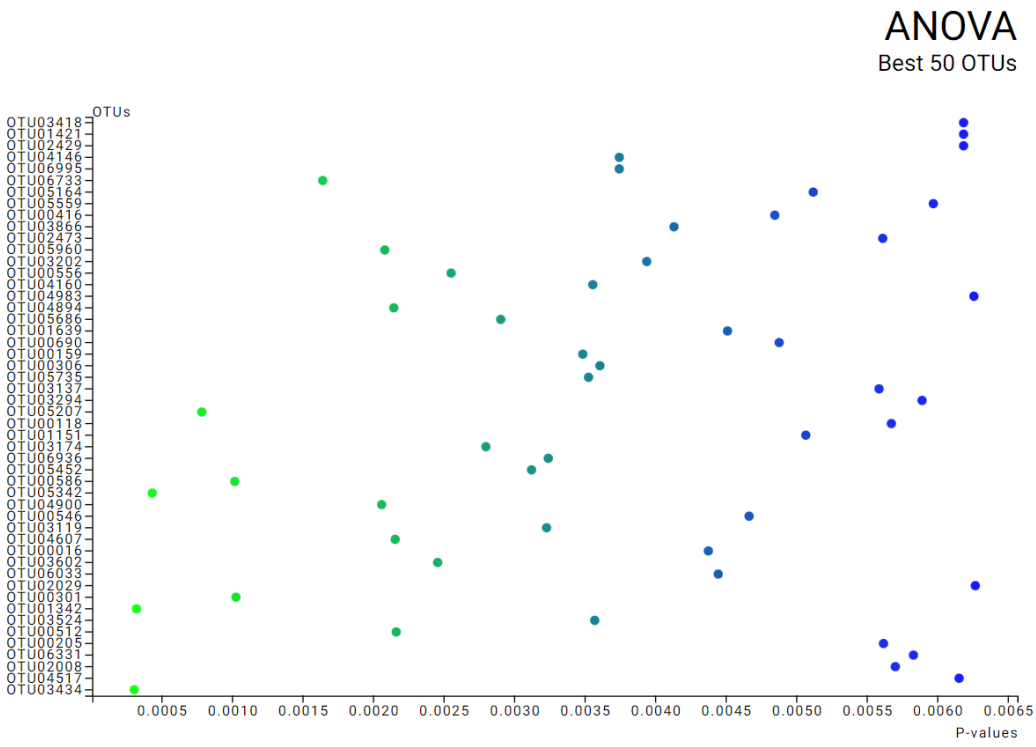


We were able to get the cluster maps ready. For 50 OTU results of feature selection methods, ANOVA shows that Wheat_1, Wheat_2, Wheat_4, Wheat_5 and Wheat_6 samples behave in a similar way and Brassica_1, Brassica_2, Brassica_3, Brassica_4 and Wheat_3 samples behave in a similar way. On the other hand, Wheat_6 samples show significantly different behavior for chi-square, PERMANOVA and SVM feature selection methods in comparison to all other 10 samples.

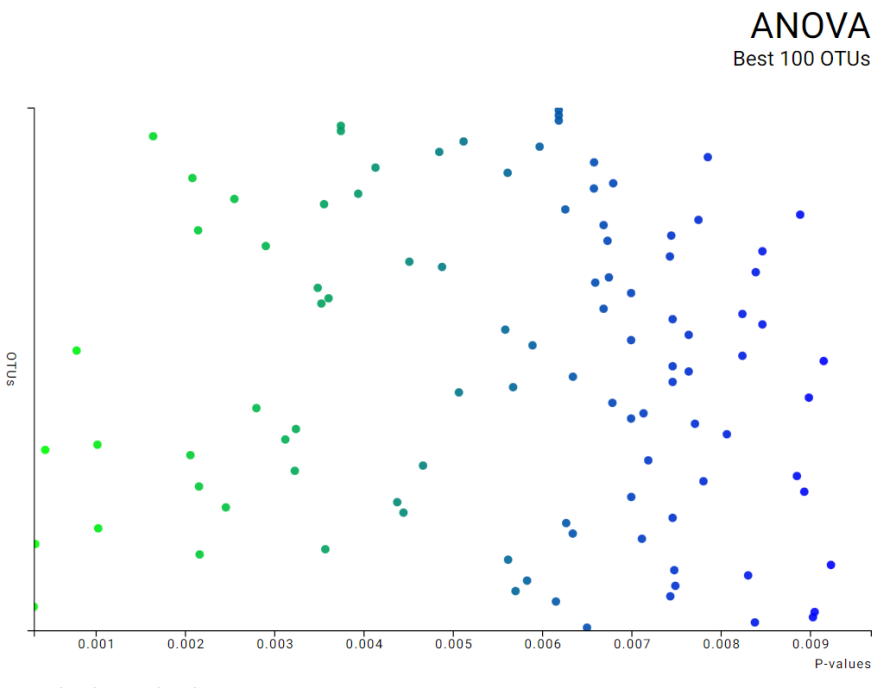
Regarding 100 OTU results, Wheat_6 samples behave significantly differently in comparison with the other samples for all feature selection methods. Wheat_6 samples show significantly different behavior as compared to the other samples for all feature selection methods for 200 OTU results.

Scatter Plots

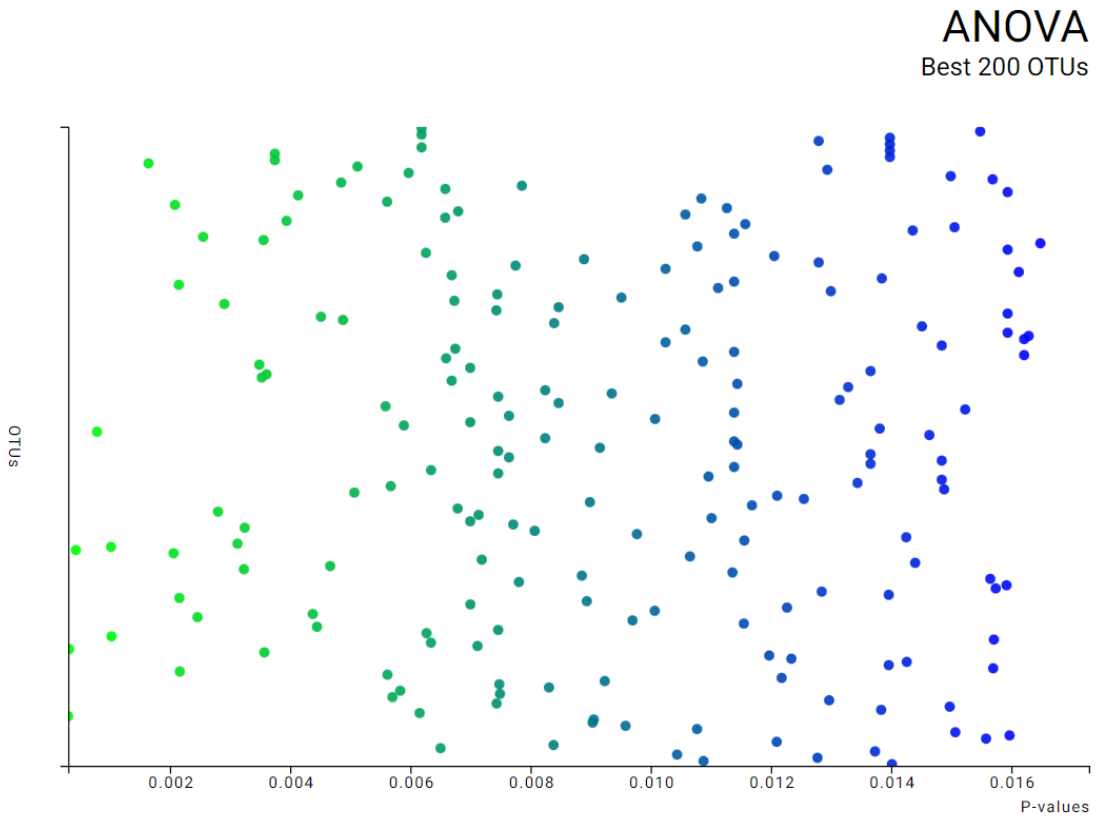
ANOVA - 50 OTUs



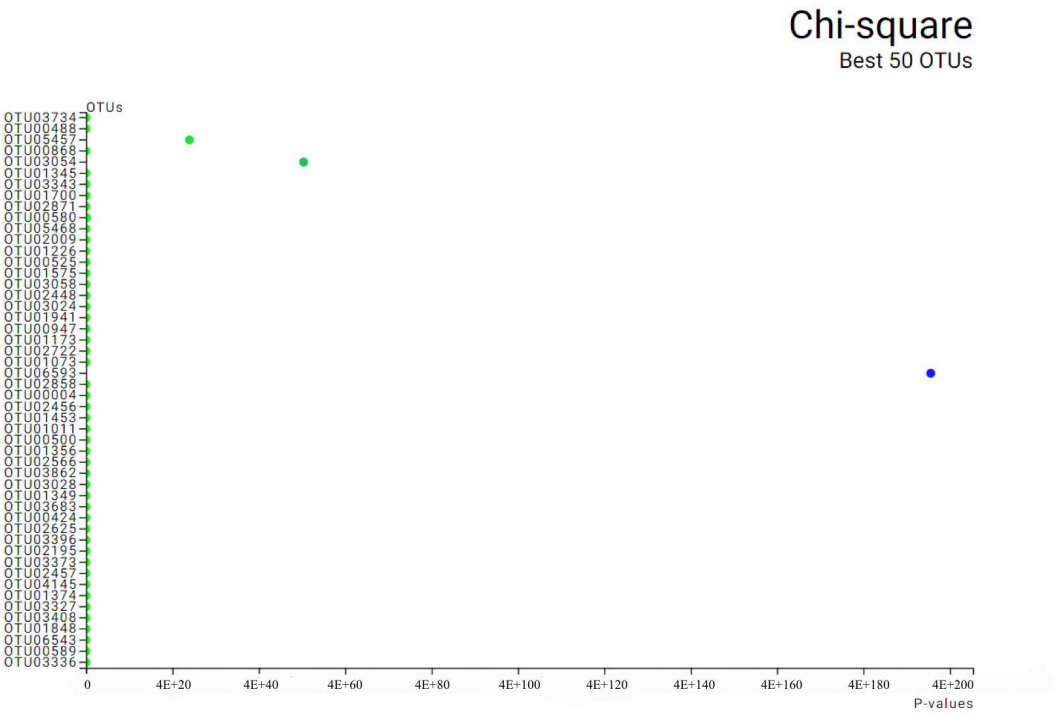
ANOVA - 100 OTUs



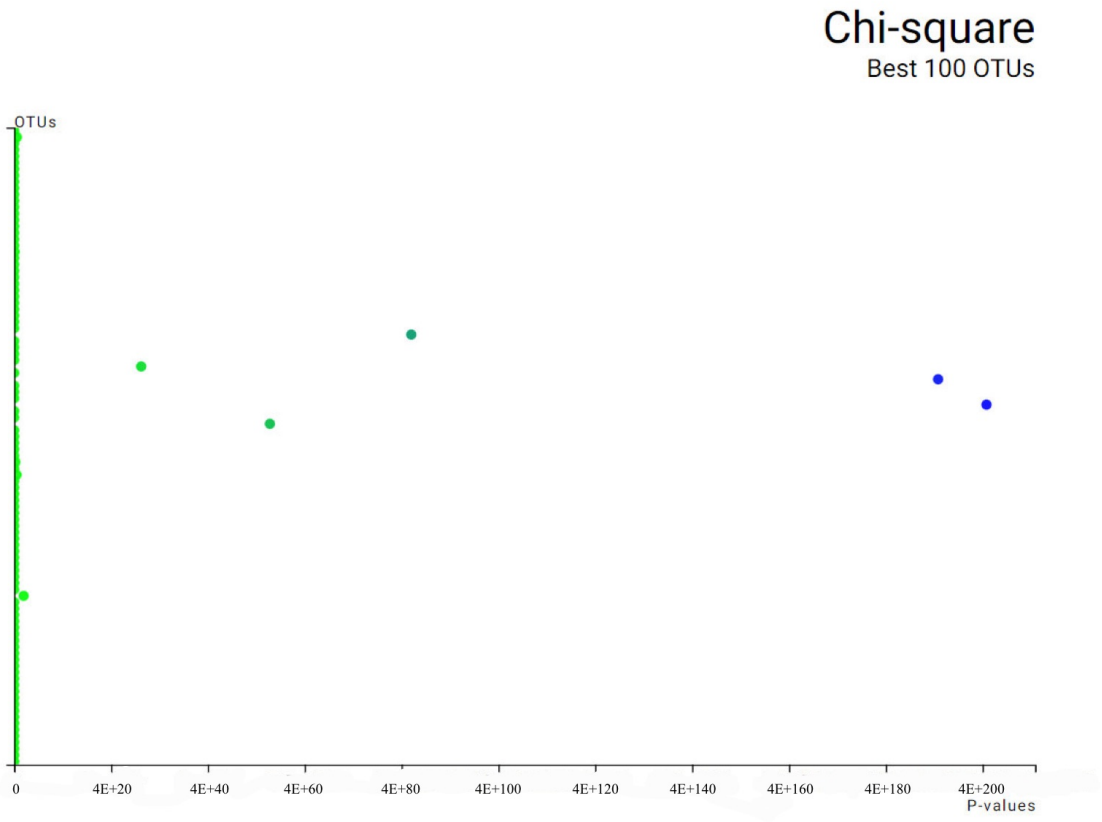
ANOVA - 200 OTUs



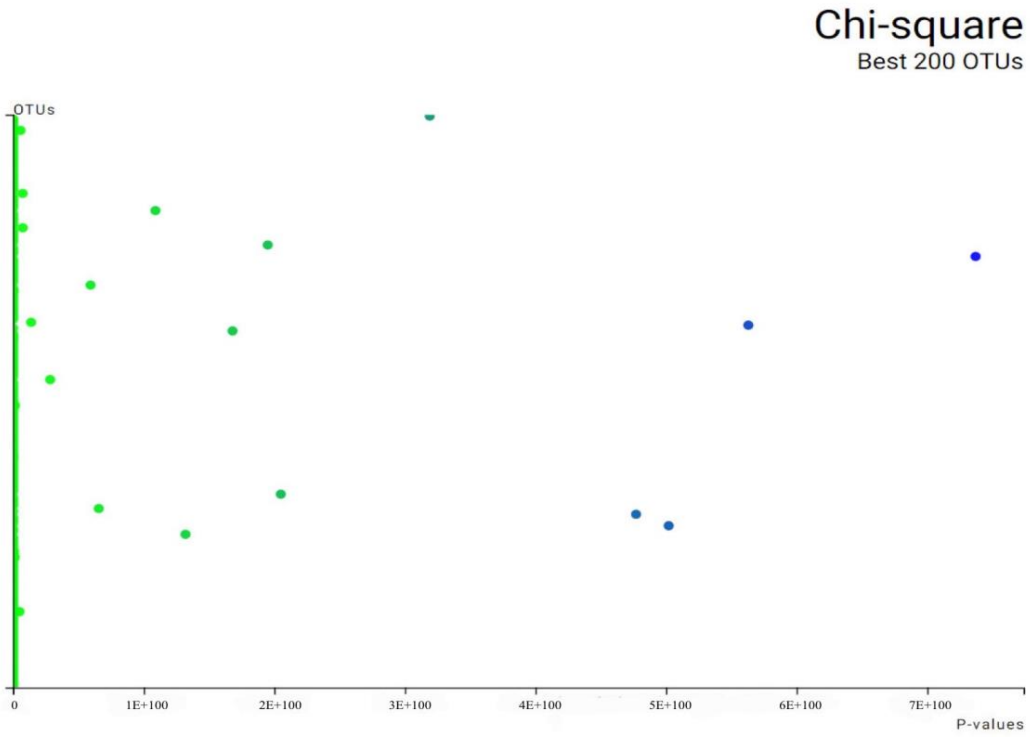
Chi-Square - 50 OTUs



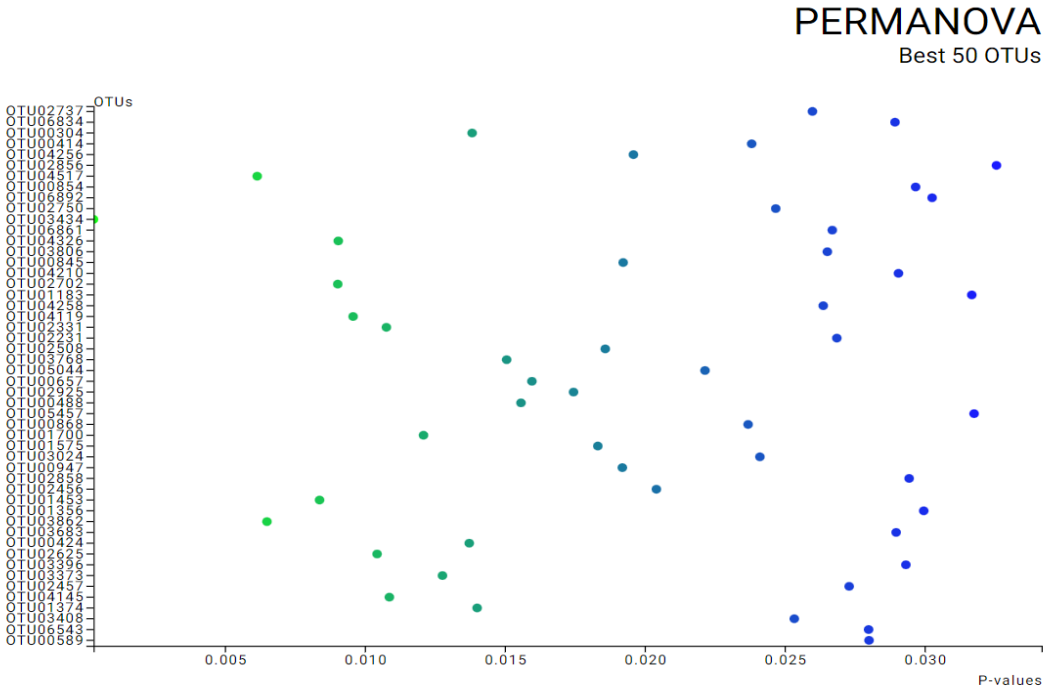
Chi-Square - 100 OTUs



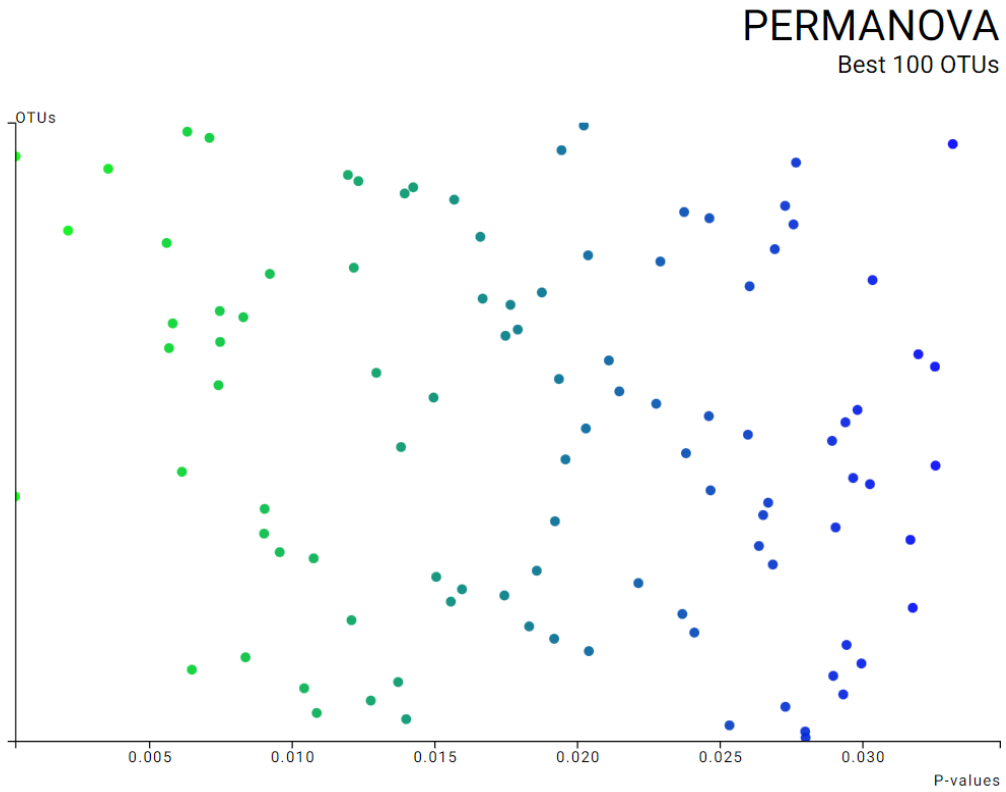
Chi-Square - 200 OTUs



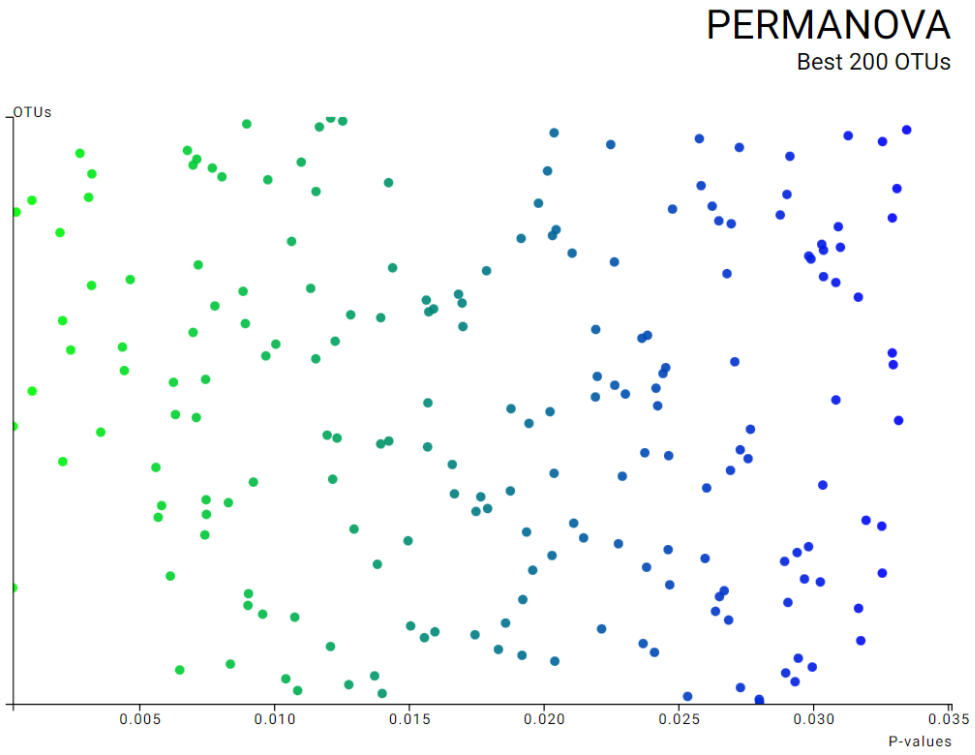
PERMANOVA - 50 OTUs



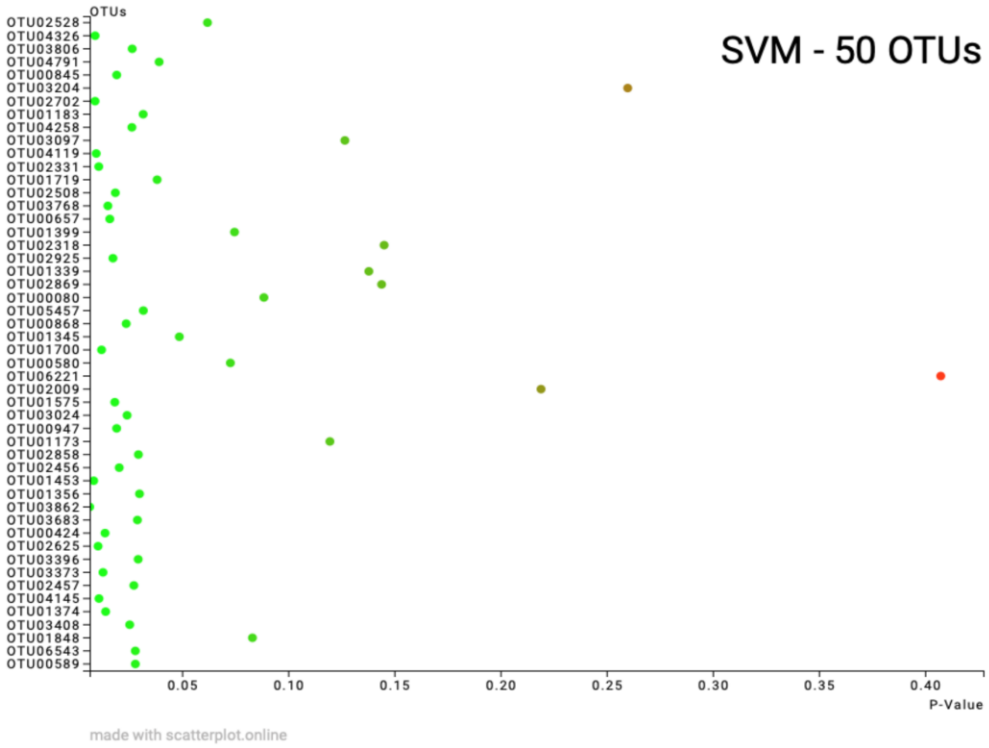
PERMANOVA - 100 OTUs



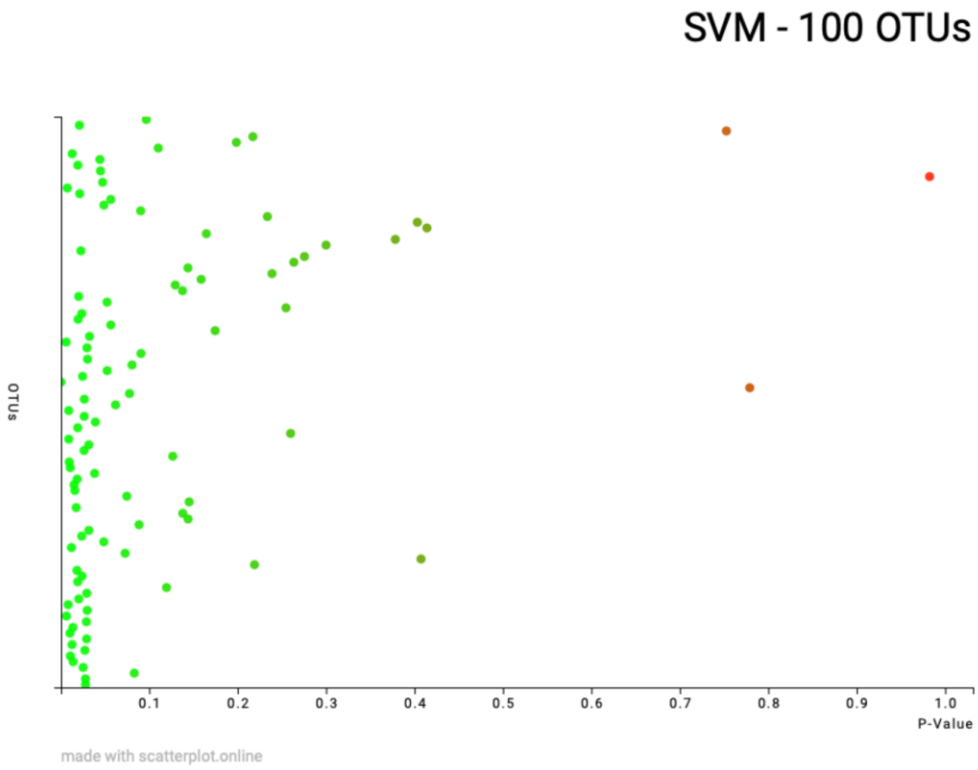
PERMANOVA - 200 OTUs



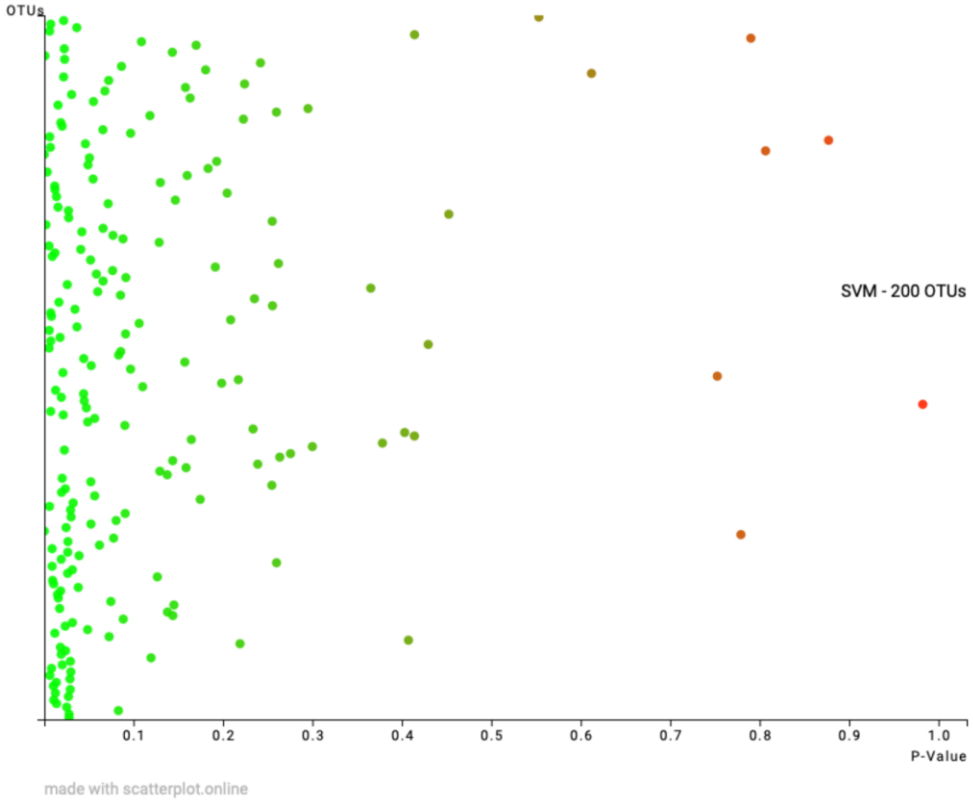
SVM - 50 OTUs



SVM - 100 OTUs



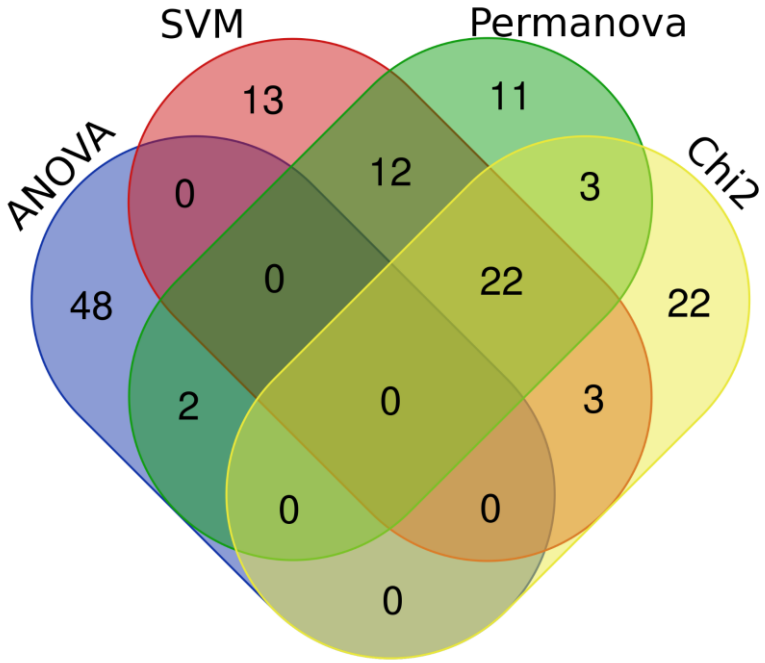
SVM - 200 OTUs



The diagrams for scatter plots for each feature selection method shows a measure of the probability for how the best 50, 100 and 200 OTUs could have occurred by a random chance. Every data point represents an OTU. For each OTU which has a p-value greater than 0.05 this means it is not statistically significant or more likely it was chosen by a random chance. If less than 0.05, then they are statistically significant or were purposely chosen.

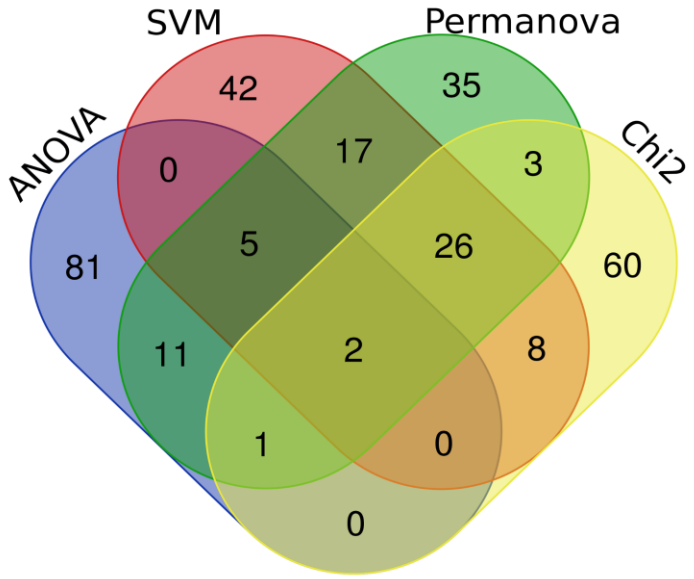
Venn Diagrams

50 OTUs



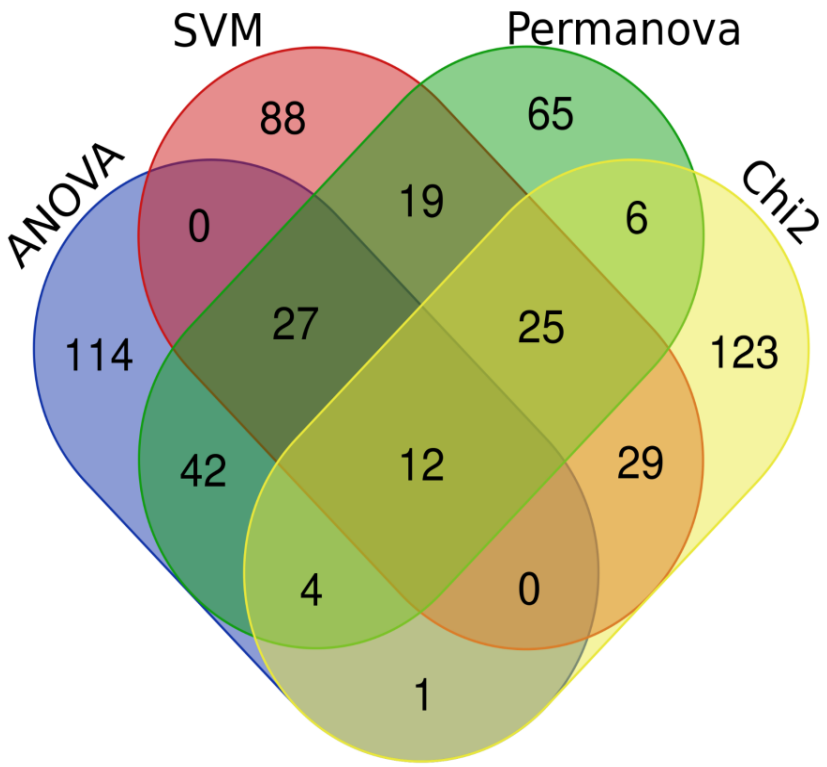
Names	total	elements
Chi2 Permanova SVM	22	OTU01575 OTU02457 OTU01700 OTU01374 OTU04119 OTU04145 OTU03408 OTU02625 OTU02858 OTU03024 OTU06543 OTU00424 OTU03373 OTU01183 OTU00589 OTU00868 OTU02456 OTU00947 OTU01453 OTU01356 OTU03683 OTU03396
ANOVA Permanova	2	OTU03434 OTU04517
Permanova SVM	12	OTU02508 OTU00657 OTU03862 OTU02925 OTU05457 OTU00845 OTU03768 OTU03806 OTU04326 OTU04258 OTU02702 OTU02331
Chi2 SVM	3	OTU02009 OTU01848 OTU01345
Chi2 Permanova	3	OTU00854 OTU05044 OTU02856
ANOVA	48	OTU04607 OTU02473 OTU00416 OTU04894 OTU00301 OTU06033 OTU01342 OTU05207 OTU05452 OTU03866 OTU03137 OTU06733 OTU00306 OTU00546 OTU06331 OTU04900 OTU05735 OTU05559 OTU02429 OTU04160 OTU03174 OTU00016 OTU05960 OTU00512 OTU03294 OTU03418 OTU06936 OTU04983 OTU00556 OTU04146 OTU00586 OTU02008 OTU05164 OTU00159 OTU03602 OTU05342 OTU01639 OTU00205 OTU00118 OTU01151 OTU02029 OTU01421 OTU03119 OTU00690 OTU05686 OTU03524 OTU03202 OTU06995
SVM	13	OTU01339 OTU04791 OTU01173 OTU06221 OTU00580 OTU02318 OTU02869 OTU03204 OTU00080 OTU03097 OTU01719 OTU01399 OTU02528
Permanova	11	OTU00414 OTU00488 OTU06892 OTU00304 OTU04210 OTU06834 OTU06861 OTU02750 OTU02737 OTU02231 OTU04256
Chi2	22	ï»¿OTU03336 OTU03642 OTU01187 OTU00407 OTU02818 OTU00994 OTU02669 OTU02070 OTU01011 OTU03242 OTU02785 OTU02667 OTU00963 OTU00767 OTU01943 OTU01226 OTU02187 OTU02448 OTU02563 OTU02819 OTU03327 OTU04973

100 OTUs



Names	total	elements
ANOVA Chi2 Permanova SVM	2	OTU03862 OTU01453
ANOVA Permanova SVM	5	OTU04517 OTU05486 OTU04326 OTU03434 OTU02702
ANOVA Chi2 Permanova	1	OTU01867
Chi2 Permanova SVM	26	OTU01575 OTU01374 OTU04119 OTU03768 OTU03968 OTU02858 OTU03024 OTU06543 OTU00424 OTU03373 OTU00589 OTU02456 OTU00947 OTU03396 OTU02457 OTU01700 OTU04145 OTU03408 OTU00854 OTU03806 OTU02625 OTU02856 OTU01183 OTU00868 OTU01356 OTU03683
ANOVA Permanova	11	OTU01342 OTU06331 OTU05844 OTU02008 OTU00205 OTU03524 OTU00859 OTU03382 OTU00512 OTU00758 OTU04904
Permanova SVM	17	OTU00414 OTU06892 OTU05457 OTU00685 OTU02235 OTU04166 OTU04258 OTU02750 OTU00294 OTU02508 OTU00657 OTU02925 OTU01317 OTU00845 OTU06861 OTU02331 OTU04256
Chi2 SVM	8	OTU03642 OTU02009 OTU00994 OTU01345 OTU04791 OTU01173 OTU01848 OTU01719
Chi2 Permanova	3	OTU00488 OTU05044 OTU00407
ANOVA	81	OTU02473 OTU00890 OTU00749 OTU00301 OTU06033 OTU03866 OTU03137 OTU05479 OTU06918 OTU05305 OTU05735 OTU05559 OTU03174 OTU05747 OTU05960 OTU00016 OTU03294 OTU04983 OTU04903 OTU06118 OTU05581 OTU00556 OTU02010 OTU00586 OTU04980 OTU05164 OTU00925 OTU01579 OTU00159 OTU01865 OTU03602 OTU01151 OTU03135 OTU01421 OTU03119 OTU00747 OTU06995 OTU04607 OTU01770 OTU00240 OTU00416 OTU04894 OTU00124 OTU00939 OTU05207 OTU05452 OTU06733 OTU00306 OTU06961 OTU00546 OTU06774 OTU04900 OTU04287 OTU01522 OTU04345 OTU02114 OTU02429 OTU04160 OTU02127 OTU06929 OTU03418 OTU06936 OTU03885 OTU04146 OTU05619 OTU00889 OTU05806 OTU00406 OTU01674 OTU05342 OTU01639 OTU00118 OTU02029 OTU04636 OTU03381 OTU05686 OTU00690 OTU00268 OTU01239 OTU03202 OTU00251
SVM	42	OTU01734 OTU01227 OTU00687 OTU06516 OTU01502 OTU01110 OTU06225 OTU00580 OTU04856 OTU03341 OTU02318 OTU02869 OTU03204 OTU02483 OTU00080 OTU03097 OTU02926 OTU02809 OTU02988 OTU02495 OTU01399 OTU02196 OTU03303 OTU03250 OTU03664 OTU01339 OTU02557 OTU06221 OTU00881 OTU03149 OTU05082 OTU06590 OTU00636 OTU06166 OTU03402 OTU02815 OTU03118 OTU02536 OTU01031 OTU02647 OTU01523 OTU02528
Permanova	35	OTU03082 OTU00785 OTU00555 OTU01461 OTU02983 OTU00028 OTU00012 OTU00420 OTU00304 OTU03243 OTU03556 OTU06591 OTU04210 OTU06558 OTU00366 OTU00958 OTU03571 OTU07027 OTU02231 OTU00408 OTU06552 OTU03428 OTU01964 OTU03366 OTU06343 OTU00337 OTU00411 OTU04918 OTU05332 OTU06834 OTU02542 OTU03335 OTU02737 OTU02679 OTU00423
Chi2	60	OTU03336 OTU04450 OTU01187 OTU02818 OTU00244 OTU01030 OTU03501 OTU02826 OTU02871 OTU04477 OTU02513 OTU01375 OTU03346 OTU00661 OTU01011 OTU03076 OTU01744 OTU02790 OTU06438 OTU02785 OTU02566 OTU03315 OTU00767 OTU01943 OTU01226 OTU02819 OTU03071 OTU02083 OTU00862 OTU03520 OTU04973 OTU03773 OTU03070 OTU02669 OTU02070 OTU00525 OTU01208 OTU02237 OTU01871 OTU00929 OTU03242 OTU03023 OTU00754 OTU02667 OTU00963 OTU02722 OTU01941 OTU05468 OTU02187 OTU05754 OTU03028 OTU02448 OTU02563 OTU01591 OTU03734 OTU00524 OTU03327 OTU02756 OTU00482 OTU03647

200 OTUs



Names	total	elements
ANOVA Chi2 Permanova SVM	12	OTU04119 OTU03862 OTU00424 OTU01867 OTU02625 OTU02702 OTU01453 OTU01374 OTU03768 OTU03373 OTU01700 OTU04145
ANOVA Permanova SVM	27	OTU04517 OTU00749 OTU03556 OTU02008 OTU00205 OTU03434 OTU01964 OTU02029 OTU00301 OTU01342 OTU06331 OTU05486 OTU04326 OTU05844 OTU03524 OTU00859 OTU03382 OTU00657 OTU01317 OTU00512 OTU00337 OTU03335 OTU00758 OTU04904 OTU00814 OTU02331 OTU00423
ANOVA Chi2 Permanova	4	OTU00488 OTU00304 OTU04918 OTU03174
Chi2 Permanova SVM	25	OTU05457 OTU03024 OTU03396 OTU02508 OTU03408 OTU01183 OTU02679 OTU04256 OTU01575 OTU02983 OTU02235 OTU03968 OTU02858 OTU06543 OTU05622 OTU00589 OTU02456 OTU00947 OTU02457 OTU00854 OTU03806 OTU02856 OTU00868 OTU01356 OTU03683
ANOVA Permanova	42	OTU06033 OTU05305 OTU01152 OTU04903 OTU00586 OTU00130 OTU04665 OTU00925 OTU03119 OTU00747 OTU02314 OTU01770 OTU00939 OTU05452 OTU00908 OTU03310 OTU01913 OTU03892 OTU06936 OTU06526 OTU05293 OTU01674 OTU05830 OTU06439 OTU05629 OTU05479 OTU06918 OTU00016 OTU04649 OTU03602 OTU00408 OTU01860 OTU04607 OTU00546 OTU04900 OTU05407 OTU01522 OTU03430 OTU05342 OTU05106 OTU03381 OTU00716
ANOVA Chi2	1	OTU03023
Permanova SVM	19	OTU01676 OTU06892 OTU00685 OTU06558 OTU04258 OTU02750 OTU07027 OTU00294 OTU02925 OTU02397 OTU02542 OTU06594 OTU06861 OTU00414 OTU04471 OTU04166 OTU00958 OTU04819 OTU00845
Chi2 SVM	29	OTU00994 OTU04621 OTU01375 OTU00580 OTU05690 OTU02083 OTU01754 OTU01345 OTU01191 OTU02863 OTU06270 OTU03149 OTU03402 OTU02815 OTU00061 OTU01719 OTU01523 OTU03647 OTU03642 OTU01734 OTU02009 OTU01502 OTU02318 OTU02483 OTU04791 OTU01173 OTU01848 OTU06590 OTU00929
Chi2 Permanova	6	OTU06591 OTU02231 OTU02737 OTU05044 OTU03243 OTU00407
ANOVA	114	OTU02473 OTU00890 OTU00732 OTU04851 OTU06032 OTU05559 OTU05606 OTU05747 OTU06530 OTU05960 OTU06680 OTU04983 OTU05581 OTU04980 OTU05164 OTU01579 OTU00159 OTU01865 OTU01151 OTU01421 OTU03135 OTU03227 OTU06995 OTU02766 OTU00416 OTU03537 OTU06733 OTU06961 OTU05429 OTU04287 OTU04345 OTU02055 OTU01184 OTU06814 OTU06929 OTU03885 OTU04146 OTU05619 OTU02561 OTU00889 OTU05806 OTU01653 OTU00406 OTU00129 OTU05836 OTU04874 OTU01639 OTU00118 OTU00310 OTU05686 OTU00573 OTU00268 OTU03202 OTU00251 OTU05437 OTU02555 OTU02054 OTU05835 OTU03866 OTU03137 OTU05735 OTU05621 OTU02021 OTU03294 OTU06118 OTU06286 OTU00556 OTU02010 OTU06016 OTU03834 OTU06665 OTU00157 OTU05604 OTU00951 OTU00106 OTU00240 OTU04894 OTU06106 OTU06817 OTU01154 OTU00124 OTU05207 OTU03618 OTU01728 OTU00306 OTU06774 OTU02114 OTU00070 OTU02429 OTU04160 OTU01884 OTU00818 OTU00781 OTU02127 OTU05874 OTU03418 OTU01615 OTU06217 OTU03543 OTU06020 OTU06215 OTU05729 OTU00358 OTU03822 OTU06341 OTU04636 OTU05350 OTU02463 OTU04502 OTU00192 OTU00690 OTU01049 OTU01239 OTU04168
SVM	88	OTU00687 OTU02630 OTU04161 OTU01110 OTU03415 OTU06893 OTU00457 OTU04049 OTU03573 OTU03341 OTU02869 OTU03780 OTU01357 OTU01190 OTU03097 OTU00869 OTU02809 OTU02988 OTU01178 OTU03405 OTU02355 OTU01399 OTU02196 OTU03664 OTU01540 OTU06221 OTU00881 OTU02518 OTU05082 OTU06099 OTU03644 OTU06166 OTU00896 OTU01444 OTU02334 OTU02536 OTU03118 OTU01031 OTU05198 OTU01559 OTU02647 OTU02199 OTU05078 OTU01227 OTU03583 OTU06516 OTU05204 OTU01977 OTU05703 OTU06225 OTU02867 OTU01435 OTU04856 OTU03317 OTU01114 OTU01748 OTU06283 OTU03204 OTU00080 OTU01809 OTU05145 OTU02926 OTU02495 OTU01939 OTU03303 OTU03250 OTU02624 OTU06825 OTU01339 OTU02775 OTU02557 OTU00942 OTU00738 OTU00461 OTU01373 OTU02944 OTU05297 OTU00223 OTU03377 OTU00731 OTU03658 OTU00636 OTU00697 OTU04877 OTU06164 OTU02511 OTU02528 OTU05240
Permanova	65	OTU01485 OTU01461 OTU06058 OTU00028 OTU01380 OTU00012 OTU03800 OTU06695 OTU00047 OTU04523 OTU00271 OTU03698 OTU06826 OTU04210 OTU03571 OTU04081 OTU04373 OTU03226 OTU00046 OTU04185 OTU04369 OTU01452 OTU03366 OTU06325 OTU00165 OTU00145 OTU05777 OTU00411 OTU05332 OTU06834 OTU06659 OTU06920 OTU01219 OTU03638 OTU00691 OTU00190 OTU02036 OTU03082 OTU00785 OTU00555 OTU06699 OTU03301 OTU00976 OTU00420 OTU03350 OTU00443 OTU00366 OTU02509 OTU03576 OTU00521 OTU06552 OTU04465 OTU03428 OTU01536 OTU03034 OTU01048 OTU04013 OTU02263 OTU06872 OTU06343 OTU05456 OTU02441 OTU01148 OTU00281 OTU06735
Chi2	123	OTU03336 OTU03058 OTU03754 OTU04450 OTU03144 OTU00924 OTU02818 OTU01030 OTU04563 OTU02623 OTU02871 OTU00447 OTU02439 OTU03346 OTU03592 OTU03641 OTU01011 OTU02558 OTU06566 OTU02790 OTU03031 OTU02785 OTU06741 OTU02566 OTU00440 OTU03315 OTU01943 OTU01226 OTU02819 OTU01779 OTU03997 OTU00862 OTU02287 OTU01825 OTU03520 OTU04973 OTU03773 OTU04322 OTU01163 OTU03070 OTU02336 OTU02851 OTU02872 OTU01851 OTU01102 OTU03222 OTU02070 OTU01349 OTU00500 OTU01871 OTU04490 OTU00754 OTU02667 OTU00963 OTU02722 OTU03045 OTU03054 OTU05468 OTU06593 OTU05503 OTU05754 OTU02448 OTU02962 OTU03734 OTU02756 OTU03343 OTU01187 OTU00244 OTU00135 OTU00937 OTU03501 OTU02826 OTU03710 OTU04477 OTU02513 OTU03053 OTU00661 OTU01360 OTU03076 OTU03443 OTU01744 OTU01203 OTU01694 OTU06438 OTU00968 OTU00767 OTU02321 OTU02726 OTU06358 OTU03071 OTU05288 OTU00638 OTU00004 OTU05860 OTU06398 OTU03692 OTU03411 OTU00884 OTU03142 OTU02669 OTU01222 OTU00525 OTU01092 OTU01208 OTU02237 OTU02486 OTU02333 OTU03242 OTU02195 OTU02688 OTU01073 OTU04238 OTU02643 OTU01941 OTU02187 OTU03028 OTU02563 OTU04964 OTU01591 OTU00524 OTU03327 OTU00482 OTU03191

These Venn diagrams show the common OTUs between the four feature selection methods ANOVA, PERMANOVA, Chi-square and SVM. They tell us how similar the methods are operating, and which ones are completely different from each other. If we take the 200 OTUs Venn diagram for instance, it tells us that PERMANOVA and ANOVA act very similarly to each other as they have 85 OTUs which are common in their set of best 200 OTUs. On the other hand, ANOVA and CHI2 methods are completely different to each other as they only have 17 OTUs which are common between them.

Execution time of the implementation of each feature selection method was also compared. It was not worth a comparison according to our client because the feature selection method implementations were executed on different hardware specifications. Therefore, this non-functional requirement was removed at the end of the project.

Biological Analysis

At the beginning of our project, we decided if time permits, we will perform biological analysis by gathering more biological information. We were expecting to get the metadata file for the samples included in our dataset, but we did not receive that file from our client. So, our client provided us with four OTUs that were thoroughly tested in the lab and were concluded as the best features in our dataset. Below is the table which includes the OTUs tested in the lab along with the ones that are found in the resulting OTUs of each feature selection method.

Biologically Tested OTUs	OTU00845, OTU03024, OTU05622, OTU01719
Matched OTUs	
ANOVA	**0/4**
Chi-Square	**3/4** (OTU03024, OTU05622, OTU01719)
PERMANOVA	**3/4** (OTU00845, OTU03024, OTU05622)
SVM	**4/4** (OTU01719, OTU00845, OTU03024, OTU05622)

Conclusion

Findings

ANOVA, chi-square, PERMANOVA and SVM feature selection methods were applied to a microbiome dataset to compare the performance and results of the methods against each other. After implementing these methods by using the microbiome dataset in Python programming language, it yielded 12 common OTUs between all 4 feature selection methods among 200 OTUs results. Out of 5477 OTUs, there were only 12 common OTUs between the methods so we can say that these feature selection methods have different effects on this particular dataset.

According to the OTUs that were tested in the lab, Chi-Square, PERMANOVA and SVM feature selection methods overperformed ANOVA. Therefore, selecting an accurate feature selection method is very important and requires lots of considerations.

Future Improvements

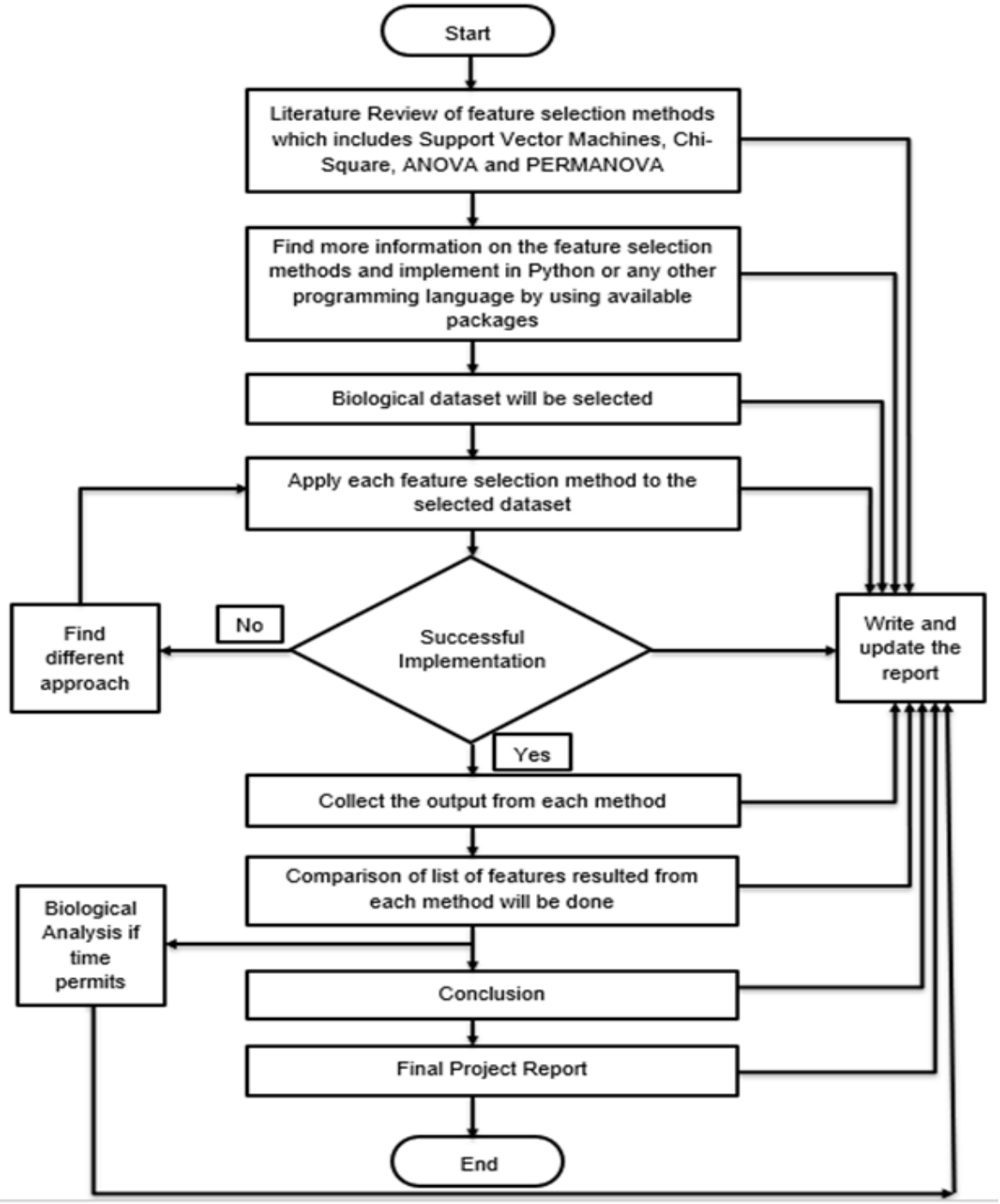
As the PERMANOVA feature selection method is mainly used for microbiome datasets to choose the best features, similar feature selection methods like MANOVA, ANCOVA and MANCOVA could be tried to see if these feature selection methods behave in a similar way. On the contrary, the same dataset could be used to implement the embedded feature selection methods like Lasso, Ridge and Decision Tree to compare and analyze the performance.

Project Report

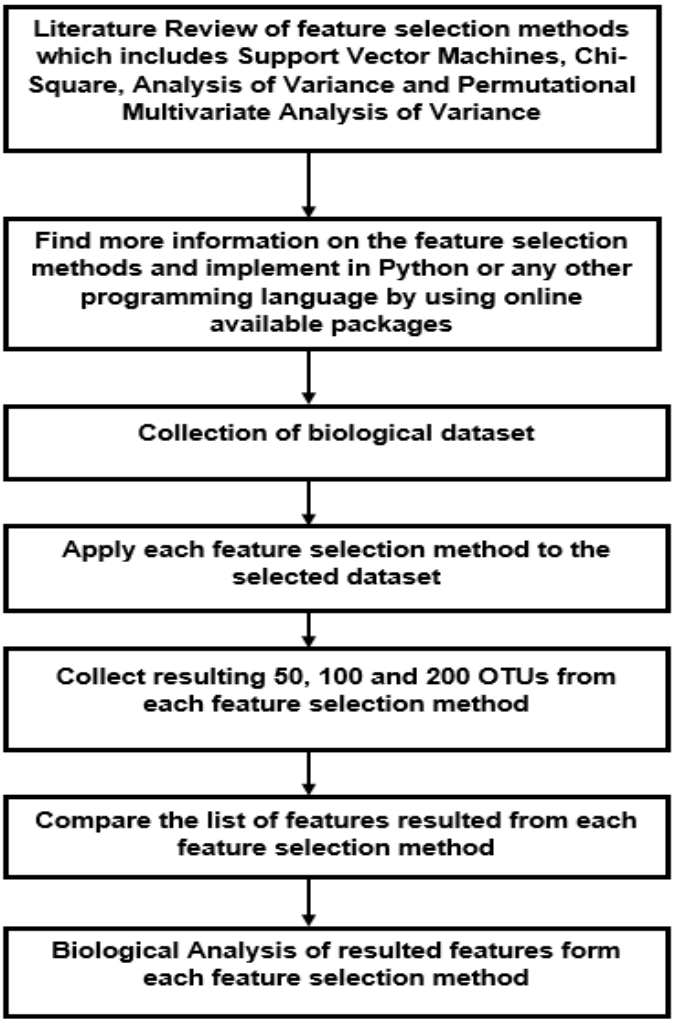
We wrote the final project report of our project which includes Abstract, Introduction, Literature Review, Methods, Experiments and Results, Evaluation, Conclusion, Future Plan, Acknowledgement and References. Our final report can be found on our Google Doc.

Design

The flowchart of the design of our project is:



The flowchart of different phases of our project is:



Implementation

We implemented three feature selection methods i.e., SVM, Chi-square, and ANOVA using the Python programming language. SVM and Chi-square are being implemented in PyCharm IDE whilst Jupyter Notebook is used for ANOVA. We have also leveraged packages such as Pandas, matplotlib, scikit-learn, scipy.stats, Statsmodels and Pingouin to fully implement these feature selection methods. Each method has its own python file with the dataset in a csv format. The dataset consists of 5,477 rows with 12 columns which includes 11 samples. It was preprocessed by our client to make sure it met the requirements of the project and to maintain confidentiality of the dataset. The project repository currently sits on GitHub so our client can have easy access to the files. All of them worked effectively with our online sample dataset.

We were able to implement the SVM, Chi-square and ANOVA feature selection methods using Sklearn, Pandas and NumPy python packages. The code of these methods can be found in the Feature Selection Method working folder on our GitHub repository. Below are the flowcharts of each feature selection method implemented:

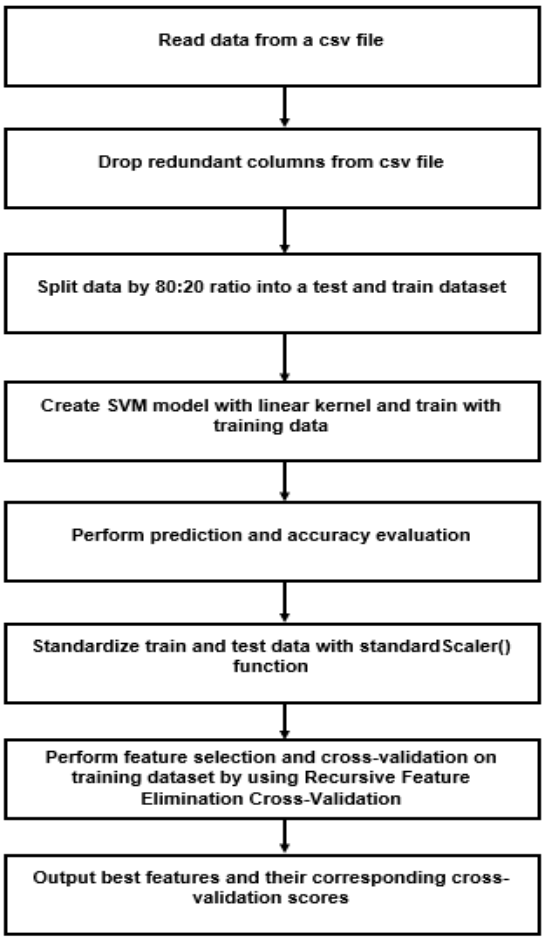
SVM

Support Vector Machine refers to a statistical and machine learning technique used on a variety of applications such as prediction (Guenther & Schonlau, 2016), pattern recognition (Cortes & Vapnik, 1995) and biological data processing (Evgeniou & Pontil, 2001).

The packages used for SVM implementation includes sklearn.train_test_split, sklearn.RFECV, sklearn.StandardScaler, sklearn.SVC, sklearn.accuracy_score. The corresponding parameters include X_train, y_train, cv, estimator, step, min_features_to_select, scoring ('accuracy'), kernel ('linear'), X_test, y_test and y_pred.

The biological dataset we received was first preprocessed by the client to make sure there were no null values and that they belonged to a specific data type. For the SVM algorithm to efficiently use the dataset, I first transposed the data from the original shape into a new shape in a new CSV file. Also, I manually created two general classes (B for Brassica and W for Wheat) for 11 samples to help effectively train the data with a sklearn library called train.test.split. Also, I tried to separate the two classes' data into two pandas data frames to find the best features in each class, which eventually produced errors. Subsequently, the raw data was used without separation, SVM classification with a linear kernel was applied to the raw data which yielded an accuracy of 0.66. Later, I experimented with the algorithm by using it together with a model named RFECV (Recursive Feature Elimination Cross Validation) to obtain the best features from the dataset. It yielded an accuracy of 100 percent with 3 optimal numbers of features (OTU05392, OTU03294, OTU06203). After discussion with the client, she requested that we generate the top 50 best features. I experimented with a similar sklearn model, RFE (Recursive Feature Elimination), which allowed for manual tuning of the number of features. The only setback was that it failed to output the probabilities/percentages of why the selected features were the top 50 features. After two-three hours of experimenting with different approaches from stackoverflow and scikit learn website, I went back to the RFECV algorithm and was able to come up with a solution that ranked the best 50 features as well as provide their corresponding percentages of why they were the top 50. It yielded an accuracy result of 100 percent.

Overall workflow of SVM feature selection method



Chi-Square

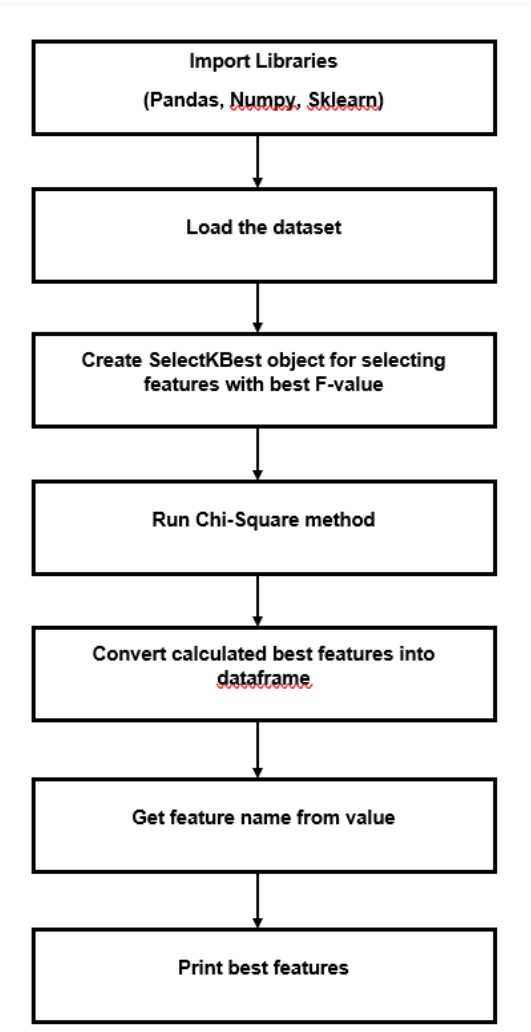
There were various articles on implementing the chi-square algorithm in python but due to lack of proper documentation, it was very difficult to jump into implementation on the original dataset. At first, we started implementing chi-square in python using a dummy dataset. We used the SciPy library, defined a dataset using an array and then implemented chi-square by passing that array to the function. Successful implementation using this library was a little step further towards implementing chi-square on the real dataset. After that, the dataset was preprocessed, and it was converted into a csv file from a xlsv (excel) file. Then the csv file containing the dataset was imported to the source code file using the Pandas library.

The data frame is then passed to the chi-square function, but instead of getting the best p values for 50 datasets, the method was showing p values for all 5477 entries. We tried extracting the p values from the result and then getting 50 best values, but it didn't work well. All the code is on the GitHub repository.

After researching for a while, we came across a new library called sklearn which had some packages solely built for feature selection. After installing the sklearn package, the data frame was passed to the chi-square function. F-value is extracted for the best 50 OUT's and then it is stored into a variable. The best f value is compared with the original dataset to get the OTU name for the 50 best f values and stored in another variable called feature_name.

This method was repeated to extract the best 100,200 values by just changing the parameter for SelectKBest function (K). After getting the best 50,100,200 values, three different csv files are made to further use those values and create cluster maps, scatter plots, graphs, and Venn diagrams. The packages used for the implementation of chi-square are sklearn.feature_selection , numpy, pandas.

Overall workflow of Chi-Square feature selection method



ANOVA

ANOVA stands for Analysis of variance. ANOVA is a statistical feature selection method which compares the means of two or more groups to determine if there are any statistically significant differences between them. The assumptions of the test are:

- H0 (Null Hypothesis) i.e., the means of all groups are equal.
- H1 (Alternate Hypothesis) i.e., at least one mean of the groups is different

ANOVA uses an f-test to see if there is any significant difference between the groups. The result of the ANOVA f-ratio will be close to 0 if there is no significant difference between the groups such that all variances are equal. ANOVA can be performed in two ways:

- One-way ANOVA
- Two-way ANOVA

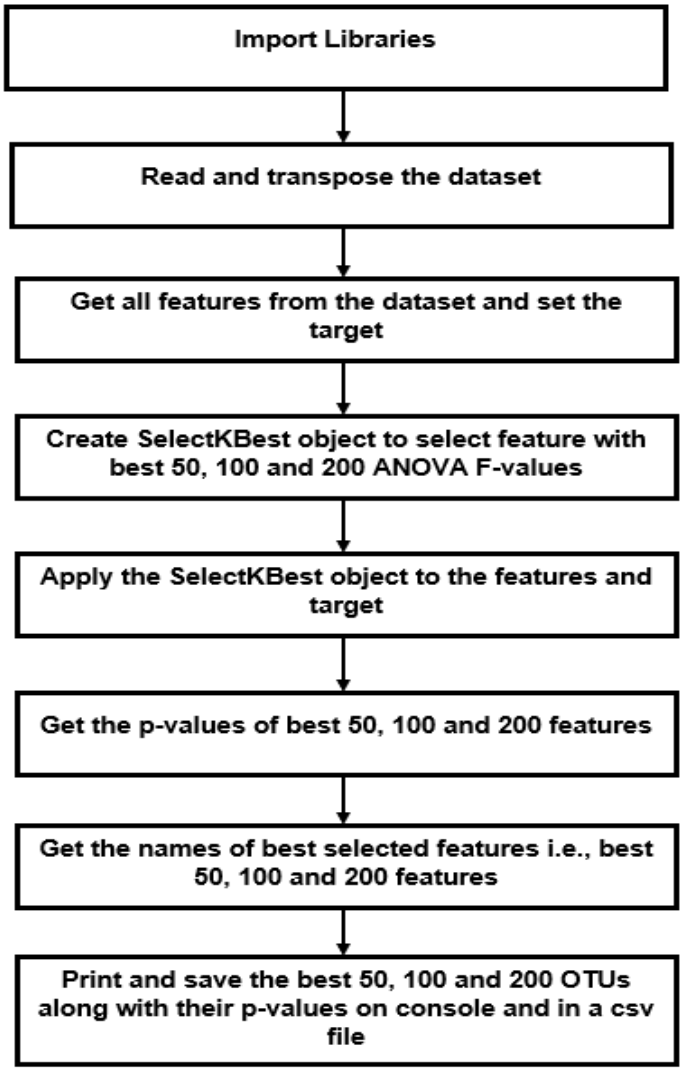
One-way ANOVA can tell us that there were two or more groups that were statistically different from each other but cannot tell which ones were. We can perform post hoc tests to find out which specific groups were different from each other. We used One-way ANOVA in our experiment to find out the best features in our dataset.

Many experiments were done in the Python programming language to get the best 50, 100 and 200 OTUs from our dataset by using different python packages.

1. Python Package StatsModels was used to compare the samples in our dataset. It compared two samples at a time but would have worked better if more than two samples could have been compared at one time.
2. Pingouin Python package was also used to set up the model for ANOVA. 5477 features were used during the implementation. It was not giving accurate p-values (probability values) as more than 5000 OTUs were used. It was tried again with only 5000 OTUs but it did not compute the p-values as our dataset included many 0's and 1's which were making them infinite.
3. SciPy python package was also tried to implement ANOVA, but the model was not set up properly. So, it ended up with neither any warning nor a meaningful output.
4. Best 50, 100 and 200 OTUs were successfully chosen by implementing ANOVA feature selection method with the help of Sklearn.feature_selection module. SelectKBest object was imported to remove all but the k highest scoring features. F_classif function was used to compute the ANOVA f-value for the dataset.

The packages and parameters used for ANOVA Implementation includes Pandas, NumPy, Sklearn.feature_selection module, SelectKBest Object, f_classif parameter and k parameter, which is the number of top features to select which is 50, 100 and 200 for ANOVA.

Overall workflow of ANOVA feature selection method



PERMANOVA

PERMANOVA stands for Permutational Multivariate Analysis of Variance [16]. PERMANOVA feature selection method is very similar to ANOVA except that it operates in a distance matrix which allows for multivariate analysis. PERMANOVA tests if two or more groups of objects are significantly different based on a categorical factor. This method computes a pseudo-F statistic [16].

We tried some online tutorials for the implementation of the PERMANOVA feature selection method with the help of the Skbio package in Python. We were having few issues with the implementation as there is not much information available online and we are relying on the documentation of the feature selection method and the libraries for implementation. We were only able to create a distance matrix of our selected dataset. The code can be found in our

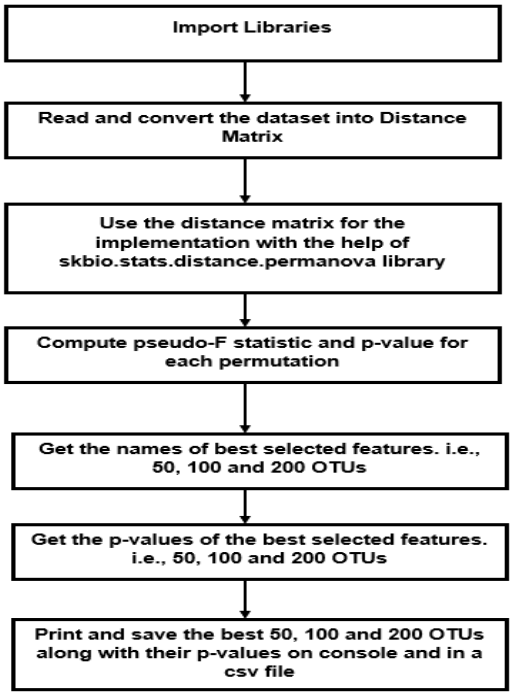
GitHub repository. The libraries that are available are not getting updated and are not compatible with python. Therefore, our client provided us with the results of the PERMANOVA feature selection method. Below are the details of the implementation of the PERMANOVA feature selection method:

For the implementation of the PERMANOVA feature selection method by using the selected dataset, Scipy python library was used. Firstly, the whole dataset was converted into a distance matrix by using scipy.spatial library and distance_matrix function. Implementation after creating the distance matrix was not successful due to mismatch in the versions of libraries and the Python version. Due to the unsuccessful implementation, our client provided us with 50, 100 and 200 OTUs results for PERMANOVA.

The packages and parameters that were used by our client for the implementation of PERMANOVA feature selection method are skbio.stats.distance, PERMANOVA, numpy, pandas, distance_matrix, grouping, column and permutations.

During PERMANOVA implementation, statistical significance was accessed via a permutation test. The assignment of objects to groups (grouping parameter used) is randomly permuted several times which was controlled via permutations parameter. A pseudo-F statistic is computed for each permutation and the p-value is the proportion of permuted pseud-F statistics that are equal to or greater than the original which means unpermuted pseudo-F statistic. PERMANOVA returns the statistical test which includes the test statistic and the p-value (probability value).

Overall workflow of PERMANOVA feature selection method



Testing

- For SVM feature selection method

  The package (sklearn's SelectKBest) we initially used automatically produced only 3 best features instead of 10 features the client asked for. We modified our code to use sklearn's RFECV (Recursive Feature Elimination Cross Validation) package which allowed us to manually select how many features the algorithm yielded. We subsequently went on to produce 50, 100 & 200 best features from our dataset.

- For Chi-square feature selection methods

  SciPy python library was used for the implementation of Chi-square feature selection method. For the testing of our project, we mainly discussed the results of each feature selection method and the comparison results with our client to see if our results meet the client's requirements. Throughout the project, we did the following:

  An expected value was required for the implementation. Therefore, after discussion with Nisha, we decided to change the python package and used Sklearn.feature_selection for the implementation of Chi-square feature selection method.

- For ANOVA feature selection methods

  The Pingouin python package was used in the beginning for the implementation of ANOVA feature selection method. First, it was giving a warning that the probability values that it gives are not accurate because the dataset file has more than 5000 entries.

  After discussion with Nisha, we decided to split the dataset into 2 parts and then tried it. But then the implemented model for ANOVA did not compute the p-values when post hoc tests were performed because our dataset had so many 0's and 1's. After discussion with Nisha, we decided to remove the unnecessary data from the original dataset to try this again.

  It worked perfectly. But our client was looking for the sample OTUs instead of the names of the plants. So, the whole model implemented was wrong.

  At last, ANOVA was implemented with the Sklearn.feature_selection python package. While getting the probability values of the best selected OTUs, we used scores_ function which was giving us very high values. After discussion with Nisha and reading the documentation, another function pvalues_ was used to successfully get the probability values for best features of ANOVA.

- For comparison of our project

  We created scatter plots and double checked with our Nisha for feedback. According to our client, the scatter plots for Chi-square were not visually appealing due to the extremely low p-values of the OTUs and the scale of the scatter plot. We modified the OTUs p-values by multiplying each of them by 10^200 to standardize the whole data and make the data points on the graph more spread out. Our client accepted this change.

  Also, we created heatmaps for all the OTUs in the respective methods, but we got feedback from our client that it did not meet the requirements of the project. Cluster maps were then suggested by our client. We created the cluster maps to see the behavior of the OTUs in each of the 11 samples and to see which samples had OTUs behaving in a similar way. We tested this with the client to see if the diagram met her requirement and she insisted the color of the maps could be lighter instead of dark. We then modified our seaborn cluster map code to include the parameter (cmap="vlag") which modified the color of the cluster maps to show low data point clusters as blue, middle data point as white and high data points as red. Our client was pleased with this change.

- We tested the resulting 50, 100 and 200 OTUs of each feature selection method with the OTUs that are Biologically tested in the lab. The already tested OTUs were provided by our client. Below is the table of the resulting OTUs of each feature selection method and the biologically tested OTUs.

Biologically Tested OTUs	OTU00845, OTU03024, OTU05622, OTU01719
Matched OTUs	
ANOVA	**0/4**
Chi-Square	**3/4** (OTU03024, OTU05622, OTU01719)
PERMANOVA	**3/4** (OTU00845, OTU03024, OTU05622)
SVM	**4/4** (OTU01719, OTU00845, OTU03024, OTU05622)

Success, Challenges and Lessons Learned from the project

Throughout the project, we went through some challenges and learned a lot about research methods, report writings and overall project development and management before the successful implementation of our project. The table below shows the list of some of the challenges, lessons learned and the success of our project.

Challenges	Lessons Learned	Success
• We faced few challenges in the implementation of ANOVA. For example, we tried StatsModels, Pingouin, and SciPy python libraries. They did not work in a way we wanted. So, we tried the Sklearn python package. • We faced many challenges in the implementation of the PERMANOVA feature selection method. The Skbio python library is not updating. Therefore, we were not able to implement it in a new version of python. • The probability values of chi-square implementation were too small to be represented on the graph. So, we manually treated the data in a way that it can be easily visualized. • We struggled to generate the scores of each OTU we got from running the SVM method. Hence, we decided to leverage the code from ANOVA method to get the p-values for each OTU from SVM.	• We learned about overall project management. • We learned how to resolve issues from our client. • We learned about the overall implementation of Agile software development lifecycle. • We learned how to work in a team, how to clear the doubts in a team and about the communication at different phases of the project. • We learned how to break down a complex problem into smaller solvable parts which are easier to work on and then merge them into a bigger problem.	• We were able to create cluster maps to see how the OTUs behave in each of the 11 soil samples • We successfully implemented the feature selection methods to yield 200 OTUs for each • We compared the OTUs from each feature selection method using a Venn Diagram too which resulted in 12 common OTUs amongst all four methods. • We also compared the resulting OTUs with the biologically proven OTUs that were tested in the lab. ANOVA feature selection method underperforms as compared to all other feature selection methods which performed significantly better.

Appendix A - First Presentation Slides

AGENDA

- Team
- Company
- Project
- Problem
- Key Requirements

TEAM

Nisha Puthiyedth	Client
Simranjit Kaur	T00605906 - Team Lead
Samuel Bedu-Annan	T00581897
Divyatej Khurana	T00637205

COMPANY

RESEARCH PROJECT

NISHA PUTHIYEDTH

PROJECT

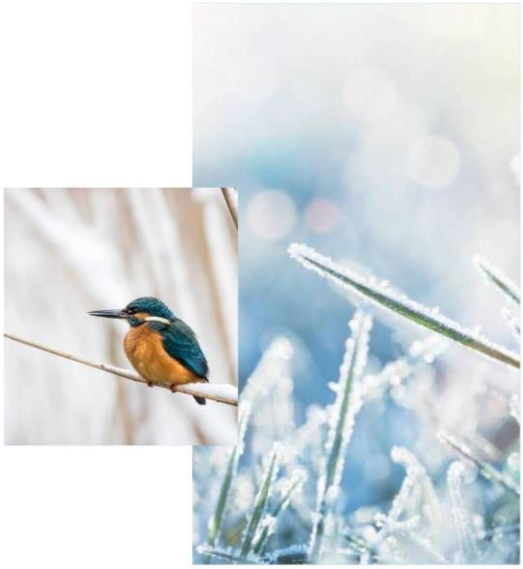
Research based Machine Learning Project

- Implementation in Python
- Report Writing

PROBLEM

Analysis and comparison of the performance of four feature selection methods:

- Support Vector Machines
- Chi-square
- Analysis and Variance (ANOVA)
- Permutational Multivariate Analysis of Variance (PERMANOVA)



KEY REQUIREMENTS

Functional

- Literature Review of methods
- Implementation of methods using Python Packages
- Searching and selecting Biological Dataset
- Data Cleaning
- Applying methods on the dataset
- Comparison of list of features resulted from each selection method
- Performing Analysis of resulting features
- Report writing an updating after each step

Non-Functional

- Project on GitHub
- Compatible with PyCharm IDE
- Compare execution time of methods
- Commented code for maintenance
- Implemented methods compatibility with small or medium sized datasets

Appendix B - Midterm Review Slides

TEAM

Nisha Puthiyedth	Client
Simranjit Kaur	T00605906 - Team Lead
Samuel Bedu-Annan	T00581897
Divyatej Khurana	T00637205

COMPANY/CLIENT

RESEARCH PROJECT
Nisha Puthiyedth

- Assistant Teaching Professor -Thompson Rivers University, Canada
- Lecturer - University of Saskatchewan, Canada
- Postdoctoral Fellow - University of Saskatchewan, Canada
- PhD Computer Science - University of Newcastle, Australia
- MSc Bioinformatics - University of East London, UK
- B.Tech Bioinformatics - Bharath University, India

PROJECT

Research based Machine Learning Project

- Implementation in Python
- Report Writing

PROBLEM

Analysis and comparison of the performance of four feature selection methods:

- Support Vector Machines
- Chi-square
- Analysis and Variance (ANOVA)
- Permutational Multivariate Analysis of Variance (PERMANOVA)

KEY REQUIREMENTS

Functional

- Literature Review of methods
- Implementation of methods using Python Packages
- Searching and selecting Biological Dataset
- Preproccessing
- Applying methods on the dataset
- Comparison of list of features resulted from each selection method
- Performing Analysis of resulting features
- Report writing an updating after each step

Non-Functional

- Project on GitHub
- Compatible with PyCharm IDE
- Compare execution time of methods
- Commented code for maintenance
- Implemented methods compatible with small or medium sized datasets

IMPLEMENTATION

- SVM, Chi-square and ANOVA implementation has been completed successfully.
- Code is committed on GitHub
- Report has been written for the implemented methods.

IMPLEMENTATION
SVM (Support Vector Machines)

- Defined as a supervised machine learning algorithm used for classification and regression
- Works by identifying a hyperplane that separates different classes
- Implemented using libraries from scikit-learn (sklearn)
- Works well in high dimensional spaces (many columns/features)
- Works poorly when have large dataset (many rows/ data points)
- Application of SVM:

Facial Expression Classification

Text Classification

Gene Classification

IMPLEMENTATION
(Chi-Square)

- Chi square test is a statical test which measures the association between two categorial variables .
- Use Value of X2 to obtain result.

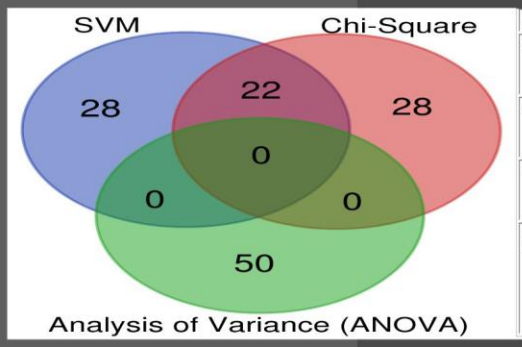
$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

- Implemented using SKLearn Library.
- Works best when using a large dataset and independent values.

IMPLEMENTATION
(ANOVA – Analysis of Variance)

- Compare the means of more than two groups
- Two types: one-way and two-way ANOVA
- Test the null hypothesis i.e., all group means are equal
- Used Pandas, NumPy and Sklearn for implementation.

Output (Work in Progress)



Names	total	elements
Chi-Square SVM	22	OTU01575 OTU02457 OTU01700 OTU01374 OTU04145 OTU03408 OTU01848 OTU02625 OTU02858 OTU03024 OTU06543 OTU00424 OTU03373 OTU00589 OTU00868 OTU02456 OTU00947 OTU01453 OTU01356 OTU03683 OTU01345 OTU03396
SVM	28	OTU01339 OTU03058 OTU00488 OTU03678 OTU06221 OTU03061 OTU05340 OTU03862 OTU01947 OTU05457 OTU05648 OTU02871 OTU06254 OTU01349 OTU00580 OTU00228 OTU02195 OTU02869 OTU00582 OTU01073 OTU02566 OTU00080 OTU02454 OTU02748 OTU03028 OTU03680 OTU03734 OTU04872
Chi-Square	28	OTU03336 OTU03642 OTU01187 OTU00407 OTU02818 OTU04119 OTU02009 OTU00994 OTU00854 OTU02669 OTU02070 OTU05044 OTU01011 OTU03242 OTU02785 OTU02667 OTU00963 OTU00767 OTU01943 OTU01226 OTU02187 OTU02856 OTU02448 OTU02563 OTU02819 OTU01183 OTU03327 OTU04973
Analysis of Variance (ANOVA)	50	OTU04607 OTU03434 OTU02473 OTU00416 OTU04894 OTU04517 OTU00301 OTU06033 OTU01342 OTU05207 OTU05452 OTU03866 OTU03137 OTU06733 OTU00306 OTU00546 OTU06331 OTU04900 OTU05735 OTU05559 OTU02429 OTU04160 OTU03174 OTU05960 OTU00016 OTU00512 OTU03294 OTU03418 OTU06936 OTU04983 OTU00556 OTU04146 OTU00586 OTU02008 OTU05164 OTU00159 OTU03602 OTU05342 OTU01639 OTU00205 OTU00118 OTU01151 OTU02029 OTU01421 OTU03119 OTU00690 OTU05686 OTU03524 OTU03202 OTU06995

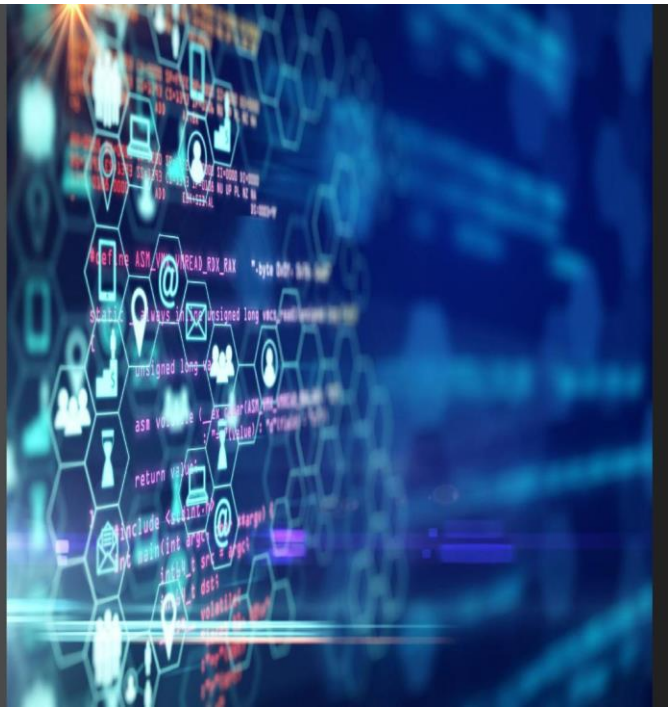
DOCUMENTATION

- Code

https://github.com/samuelbeduannan/CSResearchProject/

- Report

https://www.overleaf.com/project/61746f61d96be52920d31011



NEXT PLAN

- Implementation of PERMANOVA
- Comparison of resulted features
- Biological Analysis (If time permits)
- Conclusion
- Future Improvement
- Project Report

References

[1] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[2] E. Byvatov, U. Fechner, J. Sadowski, and G. Schneider, "Comparison of support Vector Machine and artificial neural network systems for drug/nondrug classification," Journal of Chemical Information and Computer Sciences, vol. 43, no. 6, pp. 1882–1889, Jun. 2003.

[3] M. Bhasin and G. P. Raghava, "SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence," Bioinformatics, vol. 20, no. 3, pp. 421–423, Jan. 2004.

[4] N. Guenther and M. Schonlau, "Support Vector Machines," The Stata Journal: Promoting communications on statistics and Stata, vol. 16, no. 4, pp. 917–937, 2016.

[5] T. Evgeniou and M. Pontil, "Support Vector Machines: Theory and applications," Machine Learning and Its Applications, pp. 249–257, 2001.

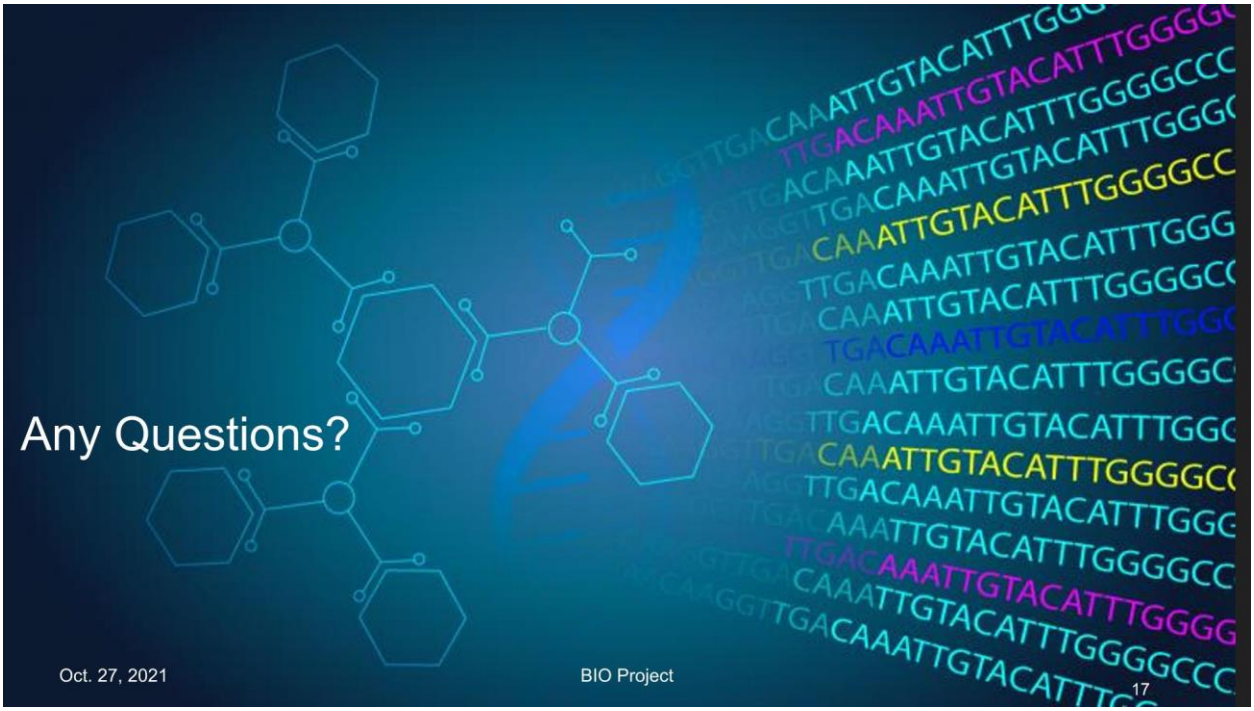
[6] A. Gelman, "Analysis of Variance - Why it is More Important Than Ever," The Analysis of Statistics, vol. 33, no. 1, pp. 1-53, Feb. 2005. [Online]. Available: https://projecteuclid.org/journals/annals-of-statistics/volume-33/issue-1/Analysis-of-variancewhy-it-is-more-important-than-ever/10.1214/009053604000000newline01048.full. [Accessed Sept. 20, 2021].

[7] M. Kumar, N. K. Rath, A. Swain and S. K. Rath, "Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor," Procedia Computer Science, vol. 54, pp. 301-310, August 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1newline1877050915013599. [Accessed Sept. 25, 2021].

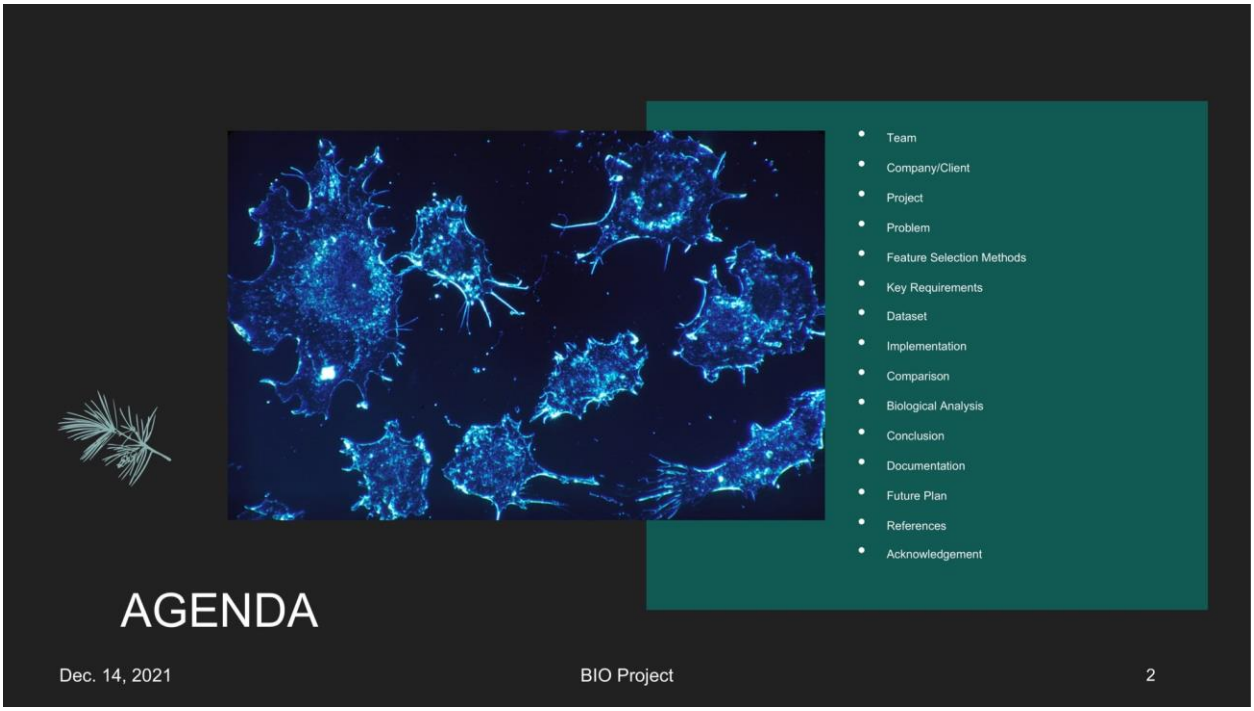
[8] R. A. Armstrong, S.V. Slade, and F. Eperjesi, "An Introduction to Analysis of Variance (ANOVA) with Special Reference to Data from Clinical Experiments in Optometry," Ophthalmic Physiol Opt., vol. 20, no. 3, pp. 235-41, May 2000. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/10897345/. [Accessed Sept. 25, 2021].

Acknowledgment

We would like to express our special thanks of gratitude to our client Nisha Puthiyedth as well as our instructor Kevin O'Neil who gave us the golden opportunity to do this wonderful project on the topic Performance Comparison Of Feature Selection Methods, which has helped us in doing lots of research and made us know about so many new things. We are utterly grateful to them.


Any Questions?

Appendix C - Final Presentation



TEAM

Nisha Puthiyedth	Client
Simranjit Kaur	T00605906 - Team Lead
Samuel Bedu-Annan	T00581897
Divyatej Khurana	T00637205

COMPANY/CLIENT

RESEARCH PROJECT
Nisha Puthiyedth

- Assistant Teaching Professor -Thompson Rivers University, Canada
- Lecturer - University of Saskatchewan, Canada
- Postdoctoral Fellow - University of Saskatchewan, Canada
- PhD Computer Science - University of Newcastle, Australia
- MSc Bioinformatics - University of East London, UK
- B.Tech Bioinformatics - Bharath University, India

PROJECT

Research based Machine Learning Project

- Implementation in Python
- Report Writing

PROBLEM

Analysis and comparison of the performance of four feature selection methods:

- Support Vector Machines
- Chi-square
- Analysis and Variance (ANOVA)
- Permutational Multivariate Analysis of Variance (PERMANOVA)

Feature Selection Methods

- Identify the set of significant features that are capable of representing the whole dataset
- Reduces the number of input variables when developing a predictive model
- Helps to reduce the computational cost of data analysis
- Improve the performance of the model in some cases by eliminating redundant and irrelevant data

KEY REQUIREMENTS

Functional

- Literature Review of methods
- Implementation of methods using Python Packages
- Searching and selecting Biological Dataset
- Preprocessing
- Applying methods on the dataset
- Comparison of list of features resulted from each selection method
- Performing Analysis of resulting features
- Report writing an updating after each step

Non-Functional

- Project on GitHub
- Compatible with PyCharm IDE
- Compare execution time of methods (removed at the end of the project)
- Commented code for maintenance
- Implemented methods compatible with small or medium sized datasets

DATASET

Microbiome Dataset for Brassica and Wheat		
Samples	**Features**	**Sample Classes Names**
6	5477 (OTUs)	Wheat
5		Brassica

IMPLEMENTATION

- SVM, Chi-square and ANOVA implementation has been completed successfully.
- PERMANOVA results were provided by our client
- Code is committed on GitHub
- Final report has been written.

IMPLEMENTATION
SVM (Support Vector Machines)

- Defined as a supervised machine learning algorithm used for classification and regression
- Works by identifying a hyperplane that separates different classes
- Implemented using libraries from scikit-learn (sklearn)
- Works well in high dimensional spaces (many columns/features)
- Works poorly when have large dataset (many rows/ data points)
- Application of SVM:

Facial Expression Classification

Text Classification

Gene Classification

IMPLEMENTATION
(Chi-Square)

- Chi square test is a statical test which measures the association between two categorial variables .
- Use Value of X2 to obtain result.

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

- Implemented using SKLearn Library.
- Works best when using a large dataset and independent values.

IMPLEMENTATION
(ANOVA – Analysis of Variance)

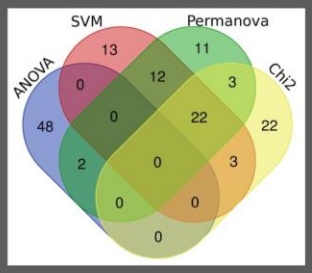
- Compare the means of more than two groups
- Two types: one-way and two-way ANOVA
- Test the null hypothesis i.e., all group means are equal
- Used Pandas, NumPy and Sklearn for implementation.

IMPLEMENTATION
(PERMANOVA – Permutational Multivariate Analysis of Variance)

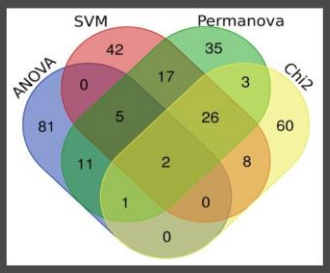
- Compares the groups of objects based on the centroid and the dispersion of the groups.
- tests the null hypothesis that these are equal for all groups
- It chooses the similarity based on distance measure
- Used Pandas, NumPy, Skbio for implementation.

Comparison (VENN Diagrams)

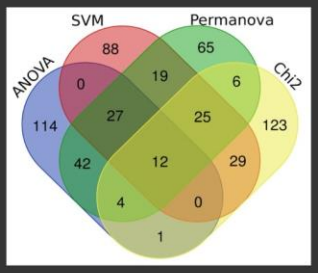
50 OTUs



100 OTUs

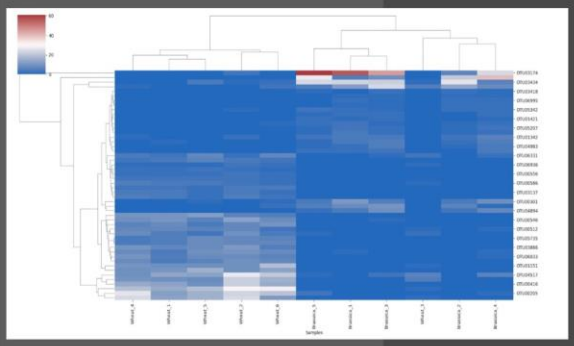


200 OTUS

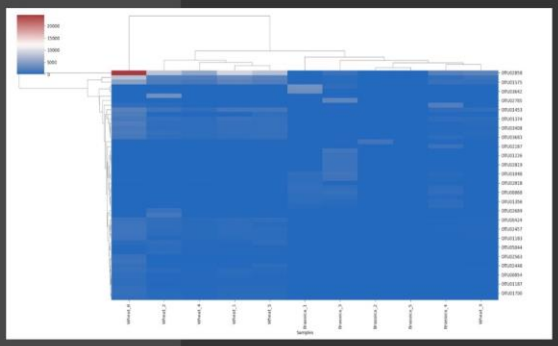


Comparison (Cluster Maps)

ANOVA

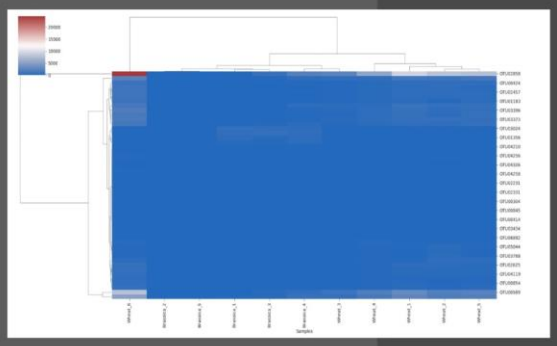


Chi-Square

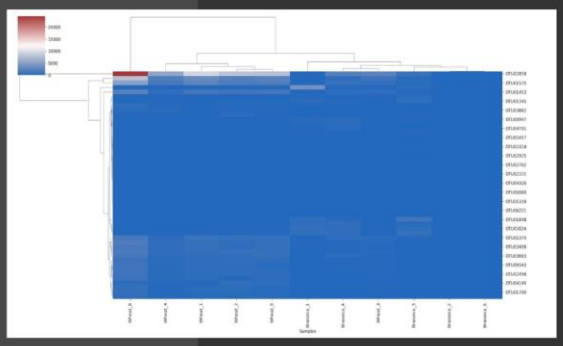


Comparison (Cluster Maps)

PERMANOVA

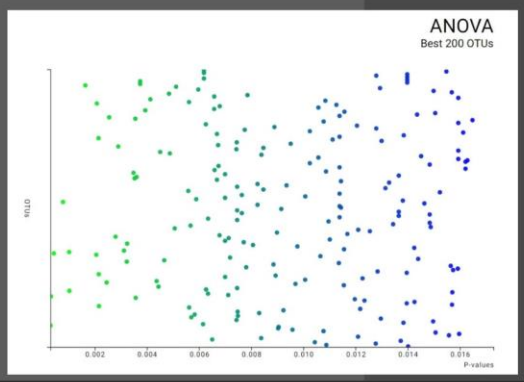


SVM

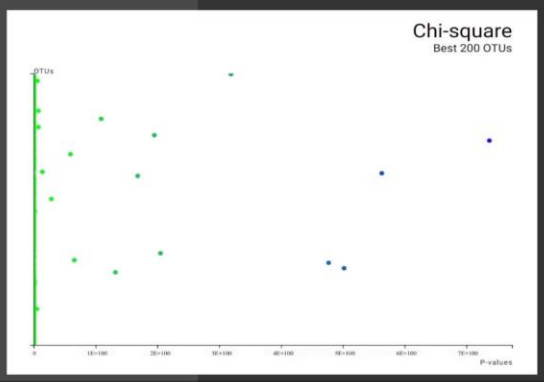


Comparison (Scatterplots)

ANOVA

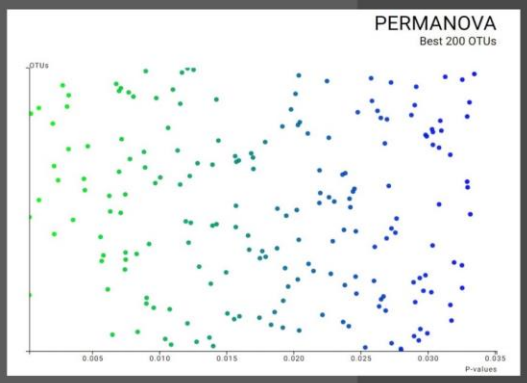


Chi-Square

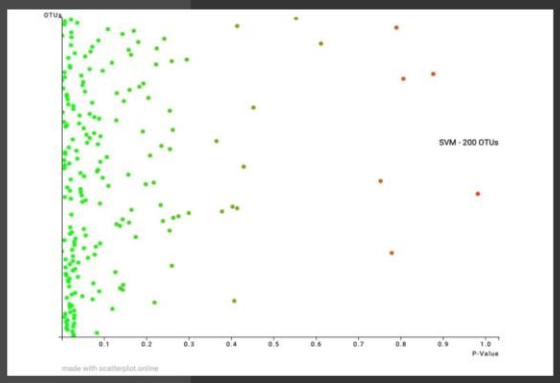


Comparison (Scatterplots)

PERMANOVA



SVM

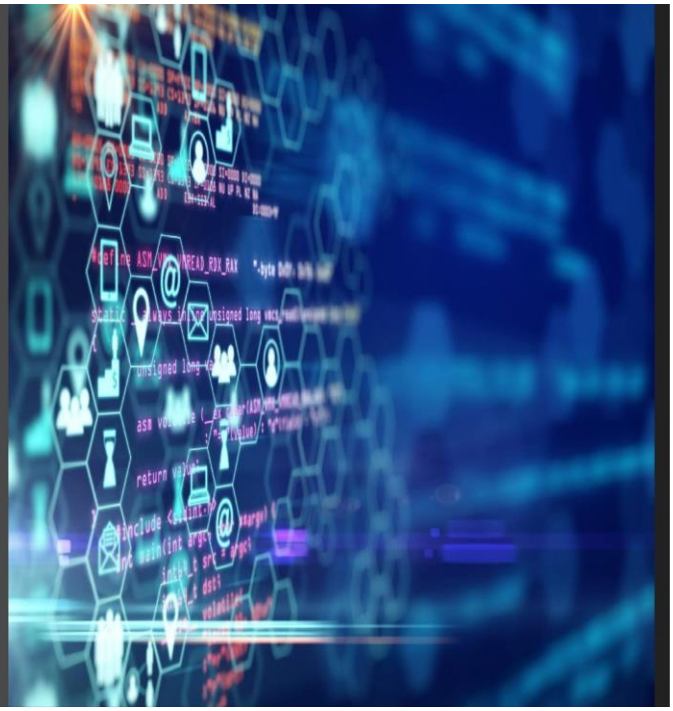


Biological Analysis

Biologically Tested OTUs	OTU00845, OTU03024, OTU05622, OTU01719
Matched OTUs	
ANOVA	**0/4**
Chi-Square	**3/4** (OTU03024, OTU05622, OTU01719)
PERMANOVA	**3/4** (OTU00845, OTU03024, OTU05622)
SVM	**4/4** (OTU01719, OTU00845, OTU03024, OTU05622)

CONCLUSION

- SVM performs well as compared to other feature selection methods
- Chi-Square and PERMANOVA behave similar a bit
- ANOVA performs uniquely as compared to the other for this dataset
- 12 common OTUs among all feature selection methods
- All methods show different effects for this dataset
- These feature selection methods might behave differently for different dataset



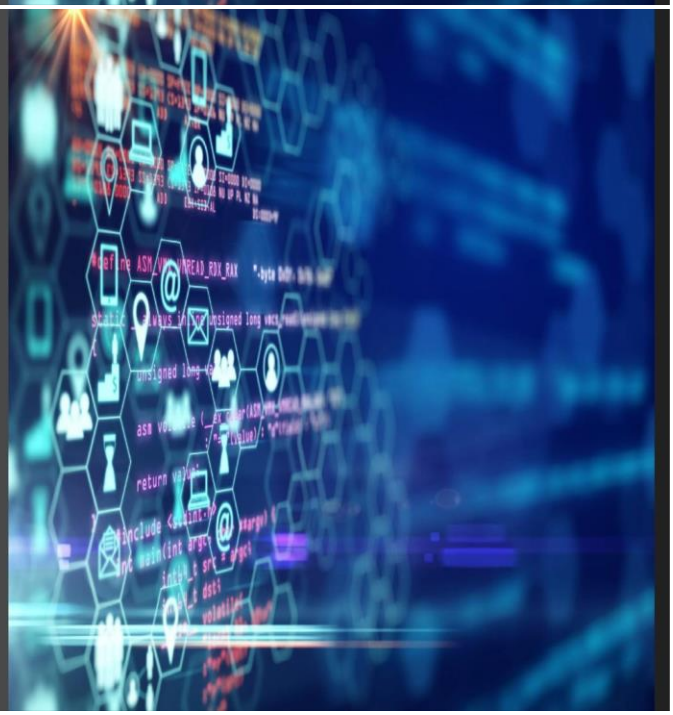
DOCUMENTATION

- Code

https://github.com/samuelbeduannan/CSResearchProject/

- Report

https://docs.google.com/document/d/1_y1vE6FpXooA67qFtpRypa0PyAIrs_Ju1CrveM0BfSc/edit



FUTURE PLAN

- As the PERMANOVA feature selection method is mainly used for microbiome datasets to choose the best features, similar feature selection methods like MANOVA, ANCOVA and MANCOVA could be tried to see if these feature selection methods behave in a similar way.
- On the contrary, the same dataset could be used to implement the embedded feature selection methods like Lasso, Ridge and Decision Tree to compare and analyze the performance.

References

[1] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995. [Online]. Available: http://homepages.rpi.edu/~bennek/class/mmld/papers/svn.pdf.

[2] E. Byvatov, U. Fechner, J. Sadowski, and G. Schneider, "Comparison of support Vector Machine and artificial neural network systems for drug/nondrug classification", Journal of Chemical Information and Computer Sciences, vol. 43, no. 6, pp. 1882-1889, Jun. 2003. [Online]. Available: https://pubs.acs.org/doi/10.1021/ci0341161.

[3] M. Bhasin and G. P. Raghava, "SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence," Bioinformatics, vol. 20, no. 3, pp. 421-423, Jan. 2004. [Online]. Available: https://academic.oup.com/bioinformatics/article/20/3/421/186291.

[4] N. Guenther and M. Schonlau, "Support Vector Machines", The Stata Journal: Promoting communications on statistics and Stata, vol. 16, no. 4, pp. 917-937, 2016. [Online]. Available: https://journals.sagepub.com/doi/10.1177/1536867X1601600407.

[5] T. Evgeniou and M. Pontil, "Support Vector Machines: Theory and applications," Machine Learning and Its Applications, pp. 249-257, 2001. [Online]. Available: https://www.researchgate.net/publ.ic ation/221621494_Support_Vector_Machines_Theory_and_Applications.

[6] A. Gelman, "Analysis of Variance - Why it is More Important Than Ever," The Analysis of Statistics, vol. 33, no. 1, pp. 1-53, Feb. 2005. [Online]. Available: https://projecteuclid.org/journals/annals-of-statistics/volume-33/issue-1/Analysis-of-variancewhy-it-is-more-important-than-ever/10.1214/0090 53604000001048.full. [Accessed Sept. 20, 2021].

[7] M. Kumar, N. K. Rath, A. Swain and S. K. Rath, "Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor," Procedia Computer Science, vol. 54, pp. 301-310, August 2015. [Online]. Available: https://www.sciencedirect.com/science/article/p ii/S1877050915013599. [Accessed Sept. 25, 2021].

[8] R. A. Armstrong, S.V. Slade, and F. Eperjesi, "An Introduction to Analysis of Variance (ANOVA) with Special Reference to Data from Clinical Experi-ments in Optometry," Ophthalmic Physiol Opt., vol. 20, no. 3, pp. 235-24, May 2000. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/1089734 5/. [Accessed Sept. 25, 2021].

[9] A. Sharma, "BIO-STATISTICS: A BRIEF OVERVIEW," *Kamls.in*, 2021. [Online]. Available: http://kamls.in/wp-content/uploads/2016/11/Vol.20_No.2_2011-1-24-28.pdf. [Accessed Nov 28, 2021].

[10] S. Onchiri, "Conceptual model on application of chi-square test in education and social sciences," Academicjournals.org, 2021. [Online]. Available: https://academicjournals.org/journal/E RR/article-full-text-pdf/[illegible]2FC76609. [Accessed Nov 28, 2021].

References

[11] D. Sharpe, "Your Chi-Square Test is Statistically Significant: Now What?," 2021 [Online]. Available: https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1288&context=pare. [Accessed Nov 28, 2021].

[12] B. J. Kelly, R. Gross, K. Bittinger, S. Sherrill-Mix, J.D. Lewis, R.G. Collman, F.D. Bushman, and H. Li, "Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA," *Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA*, vol. 31, no. 15, pp. 2461-2468, Mar. 2015. [Online]. Available: https://academic.oup.com/bioinformatics/article/31/15/2461/188732. [Accessed Nov 28, 2021]

[13] M. J Anderson, "Permutational multivariate analysis of variance (PERMANOVA)," *Wiley StatsRef: Statistics Reference Online*, pp. 1–15, Nov. 2017.[Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/9781118445112.stat07841. [Accessed Nov 28, 2021]

[14] Wikipedia, "Permutational analysis of variance," *Wikipedia*, 24-Nov-2021. [Online]. Available: https://en.wikipedia.org/wiki/Permutational_analysis_of_variance. [Accessed: 04-Dec-2021].

[15] R. B. Hayes, J. Ahn, X. Fan, B. A. Peters, Y. Ma, L. Yang, I. Agalliu, R. D. Burk, I. Ganly, M. P. Purdue, N. D. Freedman, S. M. Gapstur, and Z. Pei, "Association of oral microbiome with risk for incident head and neck squamous cell cancer," *JAMA Oncology*, vol. 4, no. 3, pp. 358–365, Jan. 2018 [Online]. Available: https://jamanetwork.com/journals/jamaoncology/fullarticle/2668530. [Accessed Sep 27, 2021]

[16] Scikit-bio development team, scikit-bio docs 0.2.3, "skbio.stats.distance.PERMANOVA," 2014. Available: http://scikit-bio.org/docs/0.2.3/generated/generated/skbio.stats.distance.PERMANOVA.html. [Accessed Dec 04, 2021]

[17] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," June 2019 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157819304379. [Accessed Dec 04, 2021]

[18] "Calculate and draw Custom Venn Diagrams," Bioinformatics and Evolutionary Genomics [Online]. Available: https://bioinformatics.psb.ugent.be/webtools/Venn/. [Accessed Nov 10, 2021]

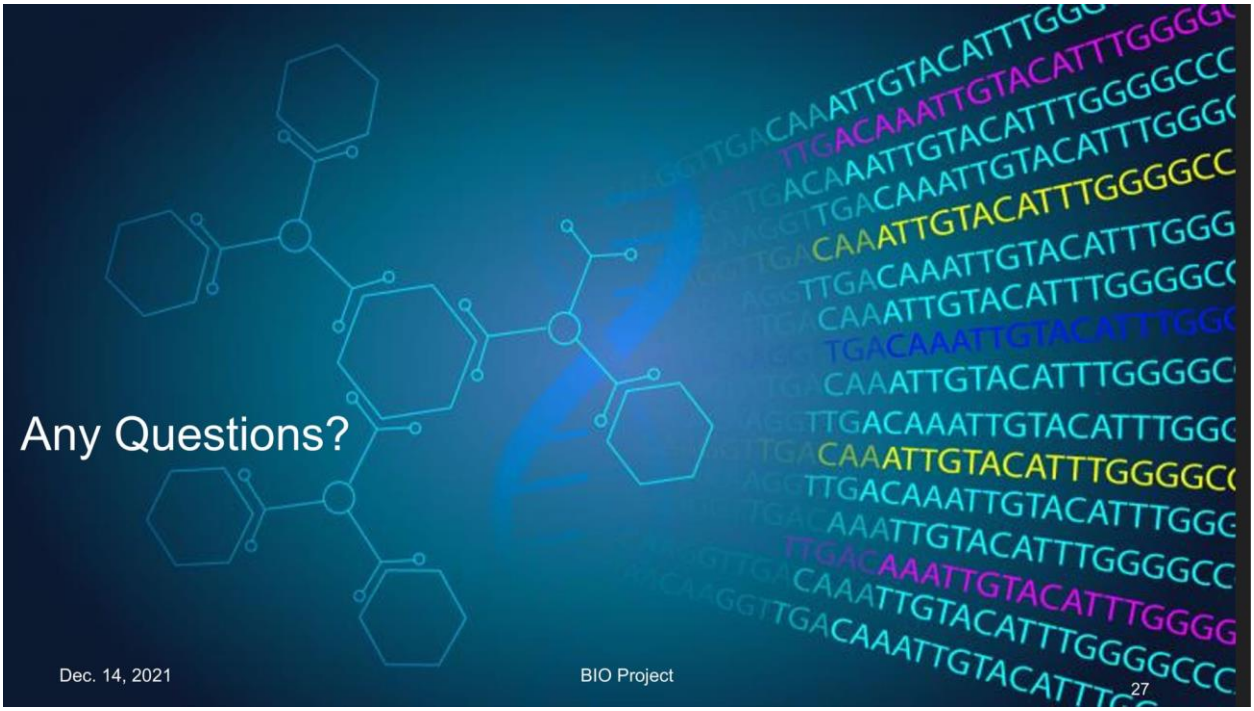
[19] T. Z. Phyu and N. N. Oo, "Performance comparison of feature selection methods," *MATEC Web of Conferences*, vol. 42, p. 06002, 2016 [Online]. Available: https://www.researchgate.net/publication/295263934_Performance_Comparison_of_Feature_Selection_Methods. [Accessed Dec 04, 2021]

[20] "Scatter Plot Maker," scatterplot.online [Online]. Available: https://scatterplot.online/. [Accessed Oct 20, 2021].

[21] M. Waskom, "seaborn.clustermap," Seaborn 0.11.2 [Online]. Available: https://seaborn.pydata.org/generated/seaborn.clustermap.html. [Accessed Nov 20, 2021]

Acknowledgment

We would like to express our special thanks of gratitude to our client Nisha Puthiyedth as well as our instructor Kevin O'Neil who gave us the golden opportunity to do this wonderful project on the topic Performance Comparison Of Feature Selection Methods, which has helped us in doing lots of research and made us know about so many new things. We are utterly grateful to them.



Appendix D - Final Project Report for Client

Comparison and Analysis of Feature Selection Methods: ANOVA (Analysis of Variance), Chi-Square, PERMANOVA (Permutational Multivariate Analysis of Variance) and SVM (Support Vector Machines)

By Simranjit Kaur, Samuel Bedu-Annan, Divyatej Khurana
Students at Thompson Rivers University, Kamloops, BC, Canada

Keywords: ANOVA, PERMANOVA, SVM, Chi-Square, Cluster Maps, Scatterplots, Venn Diagrams, OTUs

Abstract

Feature selection method plays a prominent role in the elimination of redundant and irrelevant data to improve the performance and reduce the cost of data analysis. Analysis of Variance (ANOVA), chi-square, Permutational Multivariate Analysis of Variance (PERMANOVA) and Support Vector Machines (SVM). Feature selection methods are applied to the microbiome dataset by using the sklearn python package to analyze and compare the list of features resulting from each method in the Python programming language. 50, 100 and 200 best features were analyzed and compared for each feature selection method. Out of 200 features, 12 features were common in all of them. Cluster Maps, Scatterplots and Venn Diagrams are included to show a visual comparison of all feature selection methods.

Introduction

Feature selection plays an important role in data analysis and bioinformatics applications. It reduces the number of input variables when developing a predictive model and helps to reduce the computational cost of data analysis and improve the performance of the model in some cases by eliminating redundant and irrelevant data. The main objective of the feature selection is to identify the set of significant features that are capable of representing the whole dataset. There are mainly three types of feature selection methods: Wrapper methods, Filter methods and Embedded methods. Our paper illustrates the analysis and comparison of a list of features resulting from four feature selection methods i.e., Analysis of Variance (ANOVA), Permutational Multivariate Analysis of Variance (PERMANOVA) , chi-square and Support Vector Machines(SVM) . These feature selection methods lie under both wrapper and filter methods. A microbiome dataset was used for the analysis and comparison of the resulting features of each feature selection method which included 11 soil samples of Wheat and Brassica plants and each sample had 5477 OTUs.

Literature Review

In general, filter-based techniques have less computational complexity as compared to embedded and wrapper-based techniques. Although they both employ their own learning process for feature selection, features picked by wrapper-based and embedded techniques don't always work well with other classifiers. Wrapper-based approaches have a substantial risk of overfitting due to their complexity. Whereas filter-based approaches give more stable sets of selected features due to their resistance to overfitting.[17]

According to Phyu and Oo [19], feature selection method is the technique of selecting a subset of features from the original set of features by identifying and removing features with irrelevant and redundant information. The authors compared the performance of four feature selection methods on standard datasets drawn from the UC Iravine and Waikato Environment for Knowledge Analysis (WEKA) collection [19]. A description of the datasets used per the article is as below:

Table 1. Characteristics of Datasets [19]

No	Datasets	Features	Instances	Classes
1	Vehicle	18	945	4
2	Page-blocks	11	5473	5
3	Sonar	60	208	2
4	Liver-disorder	7	345	2
5	Cylinder-band	40	512	2
6	Waveform	41	1000	3

The authors selected three feature selection methods in WEKA namely Information Gain, Symmetrical Uncertainty, Relief-F and an additional proposed method made by the authors [19]. The four algorithms were applied by the authors to the dataset to see which algorithms had the most reduction of features and better accuracy [19]. After applying the 4 algorithms to the dataset, the proposed algorithm had the most reduction of features [19]. The authors stated that the proposed algorithm reduced features from 60 to 11 in the Sonar datasets and 41 to 8 in the Waveform dataset, which were the most significant number of reductions [19]. Below is a brief summary of the number of features remaining per each method after applying the methods to the datasets:

Table 2. Number of features selected per each method [19]

Datasets	Information Gain (IG)	Symmetrical Uncertainty	Relief-F	Proposed Algorithm
Vehicle	17	19	19	14
Page-blocks	10	11	8	8
Sonar	8	22	45	11
Liver-disorder	2	2	1	2

Cylinder-band	4	21	16	19
Waveform	18	20	18	8

The proposed algorithm made by the authors performed significantly better than the existing feature selection methods from WEKA. Therefore, feature selection is prominent while removing the redundant features and in selecting the dominant features from the dataset.

ANOVA

In exploratory and confirmatory data analysis, ANOVA is the most prominent feature selection method [6]. Gelman [6] proposed an analysis which is hierarchical in nature and provides the correct ANOVA comparisons as it is not always easy to set up ANOVA models appropriately. This approach helped them to better understand the inferences of complicated models. It turned out to be very helpful while using the multilevel modeling when we have complicated data structure to set up the model [6]. Gelman [6] explains that it is quite useful for the researchers in psychology to use ANOVA in this way.

Kumar et. al [7] explained that dimensionality problems cause instability in the dataset and hinder the important information of the dataset. Therefore, selecting accurate and relevant genes plays an essential role in microarray data analysis. They proposed the ANOVA based on MapReduce and then using a Map Reduce based K-nearest Neighbor classifier instead of using feature selection or extraction along with classification which resulted in selection of more accurate features [7].
Armstrong et. al [8] explicated why and how the ANOVA was developed along with the simple implementation using a small dataset. They focused on the advanced clinical research done by eye practitioners to perform ANOVA on their data [8]. In this study, ANOVA was used to compare the treatment means and to evaluate the post hoc tests [8]. The authors concluded that ANOVA has great utility and flexibility in analyzing the data to find the number of patients required in a given experimental situation [8].

Chi-Square

Sharma [9] concluded that while practically implementing the chi-square test, it requires large sample sizes, and it may not be effective to use chi-square when sample size is less than 20 because it is very sensitive to small frequencies and chi-square can lead to erroneous conclusions. The chi-square test is used with categorical data, and actual tally numbers must be used [9]. Percentages or means of the actual dataset should not be used as it might lead to incorrect results [9].

Before performing the chi-square test, we should meet the following conditions so that we get a logically correct answer for our hypothesis testing.

1. Observations recorded and used are collected on a random basis. If observations are not randomly collected, then one of the major assumptions of inferential statistics is violated and inferences are correspondingly tenuous [10].
2. All the members (or items) in the sample must be independent i.e., they should not have any predefined correlation between them. This is to ensure that the occurrence of one individual observation (event) has no effect upon the occurrence of any other observation (event) in the sample under consideration [10].
3. No group should contain very few samples (less than 10) [10].
4. The constraints should not contain squares or higher power of frequencies [10].

Lewis and Burke wrote “In any investigation where the $\chi 2$ test is to be applied, the categories must be established in a logically defensible and reliable manner before the data are collected, if possible” [11].

PERMANOVA

According to wikipedia, PERMANOVA is used for comparing groups of data/objects and to test the null hypothesis to check that the centroids and dispersion defined for all groups are equivalent [14]. PERMANOVA and ANOVA share various similarities but also major differences as well. ANOVA bases the importance of the result on the assumption of normality while PERMANOVA determines the importance by comparing the F test result to the actual gain from random permutations between groups of objects [14].

Ahn et. al [15] performed bacterial 16S rRNA gene sequencing to determine oral microbiome composition and specific bacterial abundances by collecting mouthwash samples from various volunteers [15]. PERMANOVA was used to perform statistical analysis to determine whether overall bacterial composition differed by case or control status, different ages, sex, race, smoking status, quantity of daily cigarettes, drinking alcohol, quantity of daily ethanol intake and oral HPV-16 status [15]. The authors compared bacterial taxa between patient cases and controls by using the differential gene expression analysis based on the negative binomial distribution [15]. After comparison of the bacterial taxa, they identified 10 phyla, 22 classes, 36 orders, 62 families, 127 genera, and 439 species [15]. In analysis of oral microbiome taxa, greater abundance of phylum Actinobacteria was associated with increased risk for HNSCC [15]. When examining lower-level taxa in phylum Actinobacteria, greater abundance of order Corynebacteriales, family corynebacteriaceae, and genus Corynebacterium were associated with reduced risk for HNSCC [15]. Also, class Betaproteobacteria order Neisseriales, family Neisseriaceae, and genus Kingella in phylum Proteobacteria were significantly associated with lower risk of HNSCC [15].

SVM

Byvatov et. al [2] have shown that the prediction accuracy in SVM is slightly better than in Artificial Neural Network (ANN) regardless of the type of descriptors used for molecule encoding, the size of the training datasets and the algorithm used for training the neural network [2]. The SVM and ANN were used on drug/non-drug classification problems as an example of binary decision problems [2]. After classifying molecular compounds into either a drug or a non-drug category, SVM achieved a prediction accuracy of 82% compared to the 80% prediction accuracy achieved by ANN [2]. The main aim of the study by the authors was to compare the ability of the SVM and ANN classifiers in separating drugs from non-drugs. The reason SVM was compared to ANN is because both models are built using different sizes of training data which is used to measure the influence of the training size on the quality of the classification model. The above results show evidence that SVM outperforms ANN when a large number of features are used during the training stage [2].

Bhasin et. al [3] developed a SVM based method to identify HLA-DRBI *0401 binding peptides in an antigenic sequence [3]. The authors claimed the Radial basis function (RBF) kernel to be the best in classifying the data of DRBI *0401 binders and non-binders [3]. The regulatory parameters c & g of the RBF kernel were set to 5 and 0.1 respectively to achieve the most optimal solution [3]. In this study, the authors also stated that the SVM method achieved an accuracy score of 86 percent while ANN achieved 78 percent when used in the same study [3]. From the above, the authors concluded that SVM is superior to ANN in classifying data of MHC binders and non-binders [3]. Finally, the authors made a claim that cross-validation method is not true validation hence, performance of methods should be evaluated using data which has never been used for training/testing [3].

Project Details

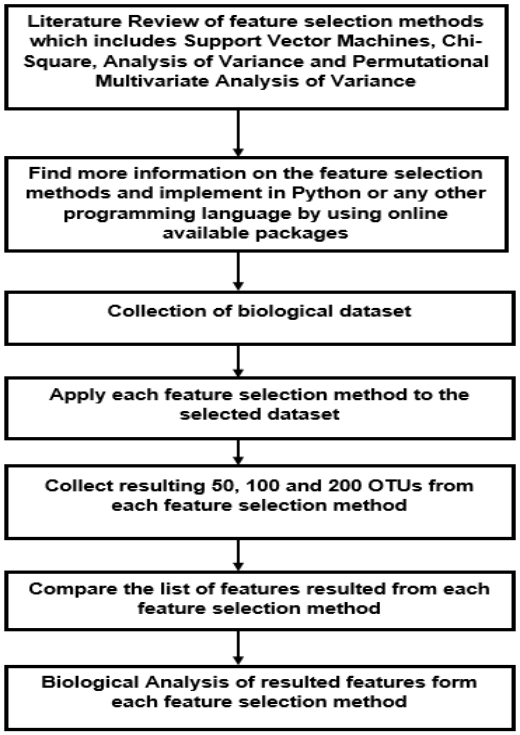
In this project, analysis, and comparison of the performance of four feature selection methods (Support Vector Machines, chi-square, Analysis of Variance and Permutational Multivariate Analysis of Variance) was done.

First, research was conducted on the feature selection methods and their importance in machine learning. The project was mainly divided into 6 phases. In phase 1 of the project, Literature Review was performed on the above-mentioned feature selection methods by using online journals, peer-reviewed articles, and other information available online. More information on the feature selection methods was collected and implemented in python in the 2nd phase of the project. Online available python packages were used for the implementation of the feature selection methods.

Phase 3 of the project was to find the biological dataset for the project. The dataset was provided by our client to work on. Next, phase 4 of the project was done which included the application of each feature selection method to the selected dataset. List of features resulting

from each feature selection method were compared in the phase 5 of the project by creating Cluster Maps, Scatterplots and Venn diagrams for best 50, 100 and 200 features. Phase 6 included the biological analysis in which the resulting features were annotated, and biological information was gathered on the resulting features.

Figure 1. Flow Chart of the project details



Methods

ANOVA

ANOVA stands for “Analysis of Variance” [6]. It is a feature selection method used to compare the means of more than two groups by performing t-test. In ANOVA, group mean differences are inferred by analyzing variances [6]. It uses a variance-based F-test to check the mean equality. It also tests the null hypothesis. i.e., all group means are equal. It can be performed in two ways: One-way ANOVA (one factor) and Two-way ANOVA (factor is an independent variable).

Chi-Square

chi-square test is a statistical test which measures the association between two categorical variables [9]. We can use this to statically determine whether the observed variable is

dependent or independent in comparison to the expected variables. The value of x2 can be calculated by using a simple formula.

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

After calculating the x2 value we compare it to the critical value corresponding to the degrees of freedom. Degree of freedom can be calculated by (No of rows -1) *(No of Columns -1). If the value of x2 is smaller than the critical value, then the null hypothesis is true and concludes that there is NO significant association between the variables. Whereas, If the calculated value is higher than the critical value in the table, we reject the null hypothesis and conclude that there is a significant association between the variables.

PERMANOVA

PERMANOVA, which is an acronym for “Permutational Multivariate Analysis of Variance” [13], is defined as a geometric partitioning of multivariate variation in the space of a chosen dissimilarity measure based on how ANOVA is designed. It produces a set of p-values using distribution-free permutation techniques [13].

In other words, PERMANOVA is a feature selection method which compares the groups of objects based on the centroid and the dispersion of the groups and tests the null hypothesis that these are equal for all groups. It chooses the similarity based on distance measure. The rejection hypothesis of PERMANOVA is that either the centroid and/or the spread of the objects is different between the groups.

SVM

Support Vector Machine refers to a statistical and machine learning technique used on a variety of applications such as prediction [4], pattern recognition [1] and biological data processing [5]. From the diagram below, SVM works by identifying a hyperplane that separates different classes (green and blue). It constructs the hyperplane by maximizing the margin of the decision boundary based on the distance of the support vectors. For a 2D problem, it uses a 1D line. For a 3D problem, it uses a 2D plane. SVM is regarded as one of the most accurate machine learning algorithms among many others due to its high generalization ability. Studies found in the early 21st & 19th centuries on the experimental success and general features of SVM have highlighted the important role it plays in different academic fields.

Experiments and Results

Table 3. Dataset Details

Microbiome Dataset for Brassica and Wheat		
Samples	**Features**	**Sample Classes Names**
6	5477 (OTUs)	Wheat
5		Brassica

For the experiments, we worked on the microbiome dataset for Brassica and Wheat which includes 11 samples and 5477 features. The dataset was divided into two main sample classes: Wheat and Brassica. There are five Brassica samples namely Brassica_1, Brassica_2, Brassica_3, Brassica_4, Brassica_5 and six Wheat samples that are Wheat_1, Wheat_2, Wheat_3, Wheat_4, Wheat_5, Wheat_6. OTUs are essentially different soil samples which contain various microorganisms for helping in the growth of the Wheat and Brassica.

ANOVA

ANOVA stands for Analysis of variance. ANOVA is a statistical feature selection method which compares the means of two or more groups to determine if there are any statistically significant differences between them. The assumptions of the test are:

- H0 (Null Hypothesis) i.e., the means of all groups are equal.
- H1 (Alternate Hypothesis) i.e., at least one mean of the groups is different

ANOVA uses an f-test to see if there is any significant difference between the groups. The result of the ANOVA f-ratio will be close to 0 if there is no significant difference between the groups such that all variances are equal. ANOVA can be performed in two ways:

- One-way ANOVA
- Two-way ANOVA

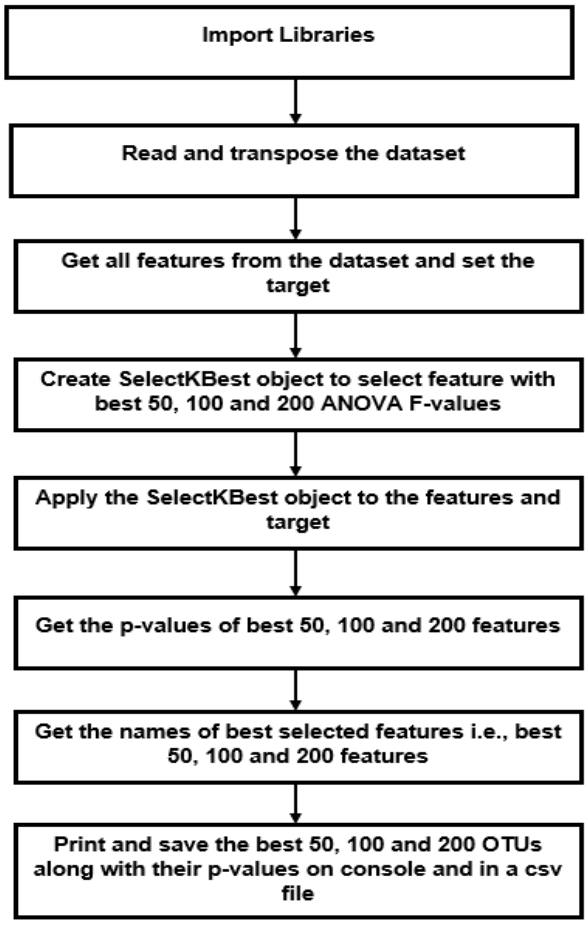
One-way ANOVA can tell us that there were two or more groups that were statistically different from each other but cannot tell which ones were. We can perform post hoc tests to find out which specific groups were different from each other. We used One-way ANOVA in our experiment to find out the best features in our dataset.

Many experiments were done in the Python programming language to get the best 50, 100 and 200 OTUs from our dataset by using different python packages.

1. Python Package StatsModels was used to compare the samples in our dataset. It compared two samples at a time but would have worked better if more than two samples could have been compared at one time.
2. Pingouin Python package was also used to set up the model for ANOVA. 5477 features were used during the implementation. It was not giving accurate p-values (probability values) as more than 5000 OTUs were used. It was tried again with only 5000 OTUs but it did not compute the p-values as our dataset included many 0's and 1's which were making them infinite.
3. SciPy python package was also tried to implement ANOVA, but the model was not set up properly. So, it ended up with neither any warning nor a meaningful output.
4. Best 50, 100 and 200 OTUs were successfully chosen by implementing ANOVA feature selection method with the help of Sklearn.feature_selection module. SelectKBest object was imported to remove all but the k highest scoring features. F_classif function was used to compute the ANOVA f-value for the dataset.

The packages and parameters used for ANOVA Implementation includes Pandas, NumPy, Sklearn.feature_selection module, SelectKBest Object, f_classif parameter and k parameter, which is the number of top features to select which is 50, 100 and 200 for ANOVA.

Figure 2. Overall workflow of ANOVA feature selection method



SelectKBest object was created to select features with best 50, 100 and 200 ANOVA F-values by setting the k value to 50, 100, and 200. The resulted 50, 100 and 200 OTUs for ANOVA feature selection method from our dataset are listed below:

Best 50 OTUs

```
['OTU03434', 'OTU04517', 'OTU02008', 'OTU06331', 'OTU00205', 'OTU00512', 'OTU03524', 'OTU01342', 'OTU00301', 'OTU02029', 'OT
U06033', 'OTU03602', 'OTU00016', 'OTU04607', 'OTU03119', 'OTU00546', 'OTU04900', 'OTU05342', 'OTU00586', 'OTU05452', 'OTU069
36', 'OTU03174', 'OTU01151', 'OTU00118', 'OTU05207', 'OTU03294', 'OTU03137', 'OTU05735', 'OTU00306', 'OTU00159', 'OTU00690',
'OTU01639', 'OTU05686', 'OTU04894', 'OTU04983', 'OTU04160', 'OTU00556', 'OTU03202', 'OTU05960', 'OTU02473', 'OTU03866', 'OTU
00416', 'OTU05559', 'OTU05164', 'OTU06733', 'OTU06995', 'OTU04146', 'OTU02429', 'OTU01421', 'OTU03418']
```

Best 100 OTUs

```
['OTU03862', 'OTU01453', 'OTU02702', 'OTU04326', 'OTU03434', 'OTU04517', 'OTU05486', 'OTU02008', 'OTU01867', 'OTU06331', 'OT
U04904', 'OTU03382', 'OTU00758', 'OTU00205', 'OTU00512', 'OTU03524', 'OTU01342', 'OTU05844', 'OTU00859', 'OTU00301', 'OTU020
29', 'OTU00749', 'OTU06033', 'OTU03602', 'OTU00016', 'OTU00925', 'OTU00747', 'OTU04607', 'OTU01770', 'OTU04903', 'OTU03119',
'OTU00546', 'OTU05479', 'OTU04900', 'OTU05342', 'OTU00586', 'OTU05452', 'OTU05305', 'OTU06936', 'OTU01674', 'OTU01522', 'OTU
00939', 'OTU03174', 'OTU06918', 'OTU03381', 'OTU01151', 'OTU00118', 'OTU01239', 'OTU02010', 'OTU06929', 'OTU04345', 'OTU0313
5', 'OTU00268', 'OTU05207', 'OTU03294', 'OTU00889', 'OTU06118', 'OTU03137', 'OTU05581', 'OTU00251', 'OTU05619', 'OTU04980',
'OTU05735', 'OTU00306', 'OTU01579', 'OTU00159', 'OTU06961', 'OTU02127', 'OTU03885', 'OTU00690', 'OTU01639', 'OTU00240', 'OTU
00124', 'OTU05686', 'OTU05806', 'OTU06774', 'OTU04894', 'OTU05747', 'OTU02114', 'OTU01865', 'OTU04983', 'OTU04160', 'OTU0055
6', 'OTU03202', 'OTU00890', 'OTU04287', 'OTU05960', 'OTU02473', 'OTU03866', 'OTU00406', 'OTU04636', 'OTU00416', 'OTU05559',
'OTU05164', 'OTU06733', 'OTU06995', 'OTU04146', 'OTU02429', 'OTU01421', 'OTU03418']
```

Best 200 OTUs

```
['OTU01374', 'OTU04145', 'OTU03373', 'OTU02625', 'OTU00424', 'OTU03862', 'OTU01453', 'OTU01700', 'OTU00488', 'OTU00657', 'OT
U03768', 'OTU02331', 'OTU04119', 'OTU02702', 'OTU04326', 'OTU03434', 'OTU04517', 'OTU00304', 'OTU00408', 'OTU05486', 'OTU013
17', 'OTU02008', 'OTU01867', 'OTU06331', 'OTU04904', 'OTU03382', 'OTU00758', 'OTU03556', 'OTU00205', 'OTU00512', 'OTU03335',
'OTU01964', 'OTU04918', 'OTU00337', 'OTU00423', 'OTU03524', 'OTU01342', 'OTU05844', 'OTU00859', 'OTU00814', 'OTU00301', 'OTU
02029', 'OTU00749', 'OTU06033', 'OTU06526', 'OTU05830', 'OTU03602', 'OTU00016', 'OTU01152', 'OTU04665', 'OTU00925', 'OTU0074
7', 'OTU04607', 'OTU01860', 'OTU00908', 'OTU06439', 'OTU05629', 'OTU01770', 'OTU03310', 'OTU04903', 'OTU05407', 'OTU03119',
'OTU00546', 'OTU03892', 'OTU05479', 'OTU00716', 'OTU04900', 'OTU05342', 'OTU00586', 'OTU05452', 'OTU05106', 'OTU04649', 'OTU
03430', 'OTU05305', 'OTU06936', 'OTU01674', 'OTU01522', 'OTU02314', 'OTU00939', 'OTU03174', 'OTU06918', 'OTU01913', 'OTU0338
1', 'OTU05293', 'OTU00130', 'OTU01151', 'OTU05836', 'OTU00118', 'OTU06530', 'OTU03834', 'OTU01884', 'OTU01239', 'OTU02010',
'OTU01653', 'OTU00310', 'OTU05621', 'OTU06929', 'OTU00781', 'OTU04345', 'OTU03135', 'OTU06032', 'OTU05835', 'OTU00268', 'OTU
02055', 'OTU05207', 'OTU04502', 'OTU03294', 'OTU00889', 'OTU00358', 'OTU06118', 'OTU06020', 'OTU06341', 'OTU03137', 'OTU0558
1', 'OTU01049', 'OTU00251', 'OTU00818', 'OTU05619', 'OTU02021', 'OTU00192', 'OTU04980', 'OTU05735', 'OTU00306', 'OTU05606',
'OTU01579', 'OTU00159', 'OTU00951', 'OTU06961', 'OTU06817', 'OTU03822', 'OTU02127', 'OTU05437', 'OTU03543', 'OTU01154', 'OTU
01615', 'OTU00157', 'OTU00129', 'OTU05874', 'OTU03885', 'OTU00690', 'OTU01639', 'OTU01728', 'OTU00240', 'OTU00124', 'OTU0568
6', 'OTU05806', 'OTU02555', 'OTU06774', 'OTU00070', 'OTU05604', 'OTU04894', 'OTU01184', 'OTU02054', 'OTU05747', 'OTU06215',
'OTU00732', 'OTU02114', 'OTU03618', 'OTU01865', 'OTU03537', 'OTU04983', 'OTU04168', 'OTU05350', 'OTU03227', 'OTU04160', 'OTU
00556', 'OTU06286', 'OTU02561', 'OTU06665', 'OTU00573', 'OTU03202', 'OTU00890', 'OTU02766', 'OTU04287', 'OTU06106', 'OTU0596
0', 'OTU02473', 'OTU06814', 'OTU03866', 'OTU06016', 'OTU00406', 'OTU04636', 'OTU00416', 'OTU03023', 'OTU00106', 'OTU05559',
'OTU05729', 'OTU05164', 'OTU06733', 'OTU06995', 'OTU06217', 'OTU04146', 'OTU04851', 'OTU02429', 'OTU04874', 'OTU06680', 'OTU
05429', 'OTU01421', 'OTU02463', 'OTU03418']
```

CHI-SQUARE

There were various articles on implementing the chi-square algorithm in python but due to lack of proper documentation, it was very difficult to jump into implementation on the original dataset.

At first, we started implementing chi-square in python using a dummy dataset. We used the SciPy library, defined a dataset using an array and then implemented chi-square by passing that array to the function. Successful implementation using this library was a little step further towards implementing chi-square on the real dataset.

After that, the dataset was preprocessed, and it was converted into a csv file from a xlsv (excel) file. Then the csv file containing the dataset was imported to the source code file using the Pandas library.

The data frame is then passed to the chi-square function, but instead of getting the best p values for 50 datasets, the method was showing p values for all 5477 entries.

We tried extracting the p values from the result and then getting 50 best values, but it didn't work well. All the code is on the GitHub repository.

After researching for a while, we came across a new library called sklearn which had some packages solely built for feature selection.

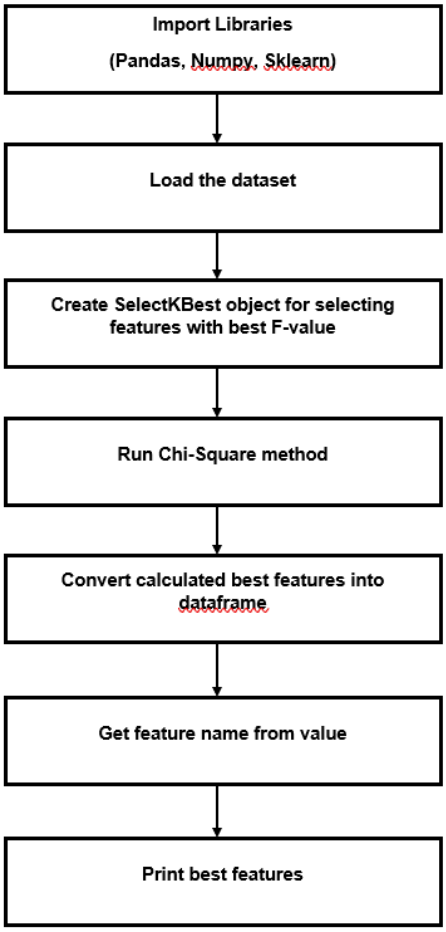
After installing the sklearn package, the data frame was passed to the chi-square function. F-value is extracted for the best 50 OUT's and then it is stored into a variable. The best f value is compared with the original dataset to get the OTU name for the 50 best f values and stored in another variable called feature_name.

This method was repeated to extract the best 100,200 values by just changing the parameter for SelectKBest function (K).

After getting the best 50,100,200 values, three different csv files are made to further use those values and create cluster maps, scatter plots, graphs and Venn diagrams.

The packages used for the implementation of chi-square are sklearn.feature_selection , NumPy, pandas.

Figure 3. Overall workflow of Chi-Square feature selection method



The resulted 50, 100 and 200 OTUs for chi-square feature selection method from our dataset are listed below:

Best 50 OTUs

['OTU03336', 'OTU00589', 'OTU06543', 'OTU01848', 'OTU03408', 'OTU03327', 'OTU01374', 'OTU04145', 'OTU02457', 'OTU03373', 'OTU03396', 'OTU02625', 'OTU00424', 'OTU03683', 'OTU01356', 'OTU01011', 'OTU01453', 'OTU02456', 'OTU02858', 'OTU00947', 'OTU03024', 'OTU02448', 'OTU01575', 'OTU01226', 'OTU02009', 'OTU01700', 'OTU01345', 'OTU00868', 'OTU00767', 'OTU05044', 'OTU04119', 'OTU01183', 'OTU00854', 'OTU02856', 'OTU00963', 'OTU00407', 'OTU00994', 'OTU03642', 'OTU01187', 'OTU02563', 'OTU04973', 'OTU02187', 'OTU02669', 'OTU01943', 'OTU02819', 'OTU02070', 'OTU02667', 'OTU03242', 'OTU02818', 'OTU02785']

Best 100 OTUs

['OTU03336', 'OTU00589', 'OTU06543', 'OTU01848', 'OTU03408', 'OTU03327', 'OTU01374', 'OTU04145', 'OTU02457', 'OTU03373', 'OTU03396', 'OTU02625', 'OTU00424', 'OTU03683', 'OTU03028', 'OTU03862', 'OTU02566', 'OTU01356', 'OTU01011', 'OTU01453', 'OTU02456', 'OTU02858', 'OTU02722', 'OTU01173', 'OTU00947', 'OTU01941', 'OTU03024', 'OTU02448', 'OTU01575', 'OTU00525', 'OTU01226', 'OTU02009', 'OTU05468', 'OTU02871', 'OTU01700', 'OTU01345', 'OTU00868', 'OTU00488', 'OTU03734', 'OTU00767', 'OTU05044', 'OTU00524', 'OTU03768', 'OTU03346', 'OTU01719', 'OTU04119', 'OTU01183', 'OTU04791', 'OTU03806', 'OTU03520', 'OTU03501', 'OTU00854', 'OTU02856', 'OTU00963', 'OTU00244', 'OTU03773', 'OTU00407', 'OTU00994', 'OTU00482', 'OTU03642', 'OTU03968', 'OTU01030', 'OTU01208', 'OTU01187', 'OTU02563', 'OTU04477', 'OTU00929', 'OTU01867', 'OTU02513', 'OTU02083', 'OTU06438', 'OTU03070', 'OTU03647', 'OTU04973', 'OTU02187', 'OTU03315', 'OTU01375', 'OTU01591', 'OTU00661', 'OTU02669', 'OTU02790', 'OTU01943', 'OTU03071', 'OTU02819', 'OTU02070', 'OTU02237', 'OTU01871', 'OTU01744', 'OTU00862', 'OTU02667', 'OTU03242', 'OTU02756', 'OTU04450', 'OTU03023', 'OTU02818', 'OTU02826', 'OTU05754', 'OTU02785', 'OTU03076', 'OTU00754']

Best 200 OTUs

['OTU03336', 'OTU00589', 'OTU06543', 'OTU01848', 'OTU03408', 'OTU03327', 'OTU01374', 'OTU04145', 'OTU02457', 'OTU03373', 'OTU02195', 'OTU03396', 'OTU02625', 'OTU00424', 'OTU03683', 'OTU01349', 'OTU03028', 'OTU03862', 'OTU02566', 'OTU01356', 'OTU00500', 'OTU01011', 'OTU01453', 'OTU02456', 'OTU00004', 'OTU02858', 'OTU06593', 'OTU01073', 'OTU02722', 'OTU01173', 'OTU00947', 'OTU01941', 'OTU03024', 'OTU02448', 'OTU03058', 'OTU01575', 'OTU00525', 'OTU01226', 'OTU02009', 'OTU05468', 'OTU00580', 'OTU02871', 'OTU01700', 'OTU03343', 'OTU01345', 'OTU03054', 'OTU00868', 'OTU05457', 'OTU00488', 'OTU03734', 'OTU00767', 'OTU02318', 'OTU06566', 'OTU04238', 'OTU05044', 'OTU00524', 'OTU05503', 'OTU03768', 'OTU03346', 'OTU03754', 'OTU02508', 'OTU01719', 'OTU02231', 'OTU04119', 'OTU00135', 'OTU06398', 'OTU01183', 'OTU02702', 'OTU02643', 'OTU04791', 'OTU03806', 'OTU01092', 'OTU03520', 'OTU01523', 'OTU03501', 'OTU01222', 'OTU02688', 'OTU01502', 'OTU00854', 'OTU02856', 'OTU02815', 'OTU03149', 'OTU04256', 'OTU00304', 'OTU00963', 'OTU00244', 'OTU02483', 'OTU03773', 'OTU02737', 'OTU02235', 'OTU01734', 'OTU03402', 'OTU00407', 'OTU00994', 'OTU00482', 'OTU03642', 'OTU06591', 'OTU03968', 'OTU03692', 'OTU04964', 'OTU01030', 'OTU01208', 'OTU02851', 'OTU06590', 'OTU01187', 'OTU00440', 'OTU02563', 'OTU05288', 'OTU02872', 'OTU04477', 'OTU00929', 'OTU01867', 'OTU03243', 'OTU02983', 'OTU03710', 'OTU02513', 'OTU02083', 'OTU06438', 'OTU01163', 'OTU01191', 'OTU03070', 'OTU02863', 'OTU01754', 'OTU02679', 'OTU04621', 'OTU06270', 'OTU00061', 'OTU02962', 'OTU03647', 'OTU03031', 'OTU04918', 'OTU04973', 'OTU02187', 'OTU05690', 'OTU03315', 'OTU01375', 'OTU05622', 'OTU01591', 'OTU04490', 'OTU00661', 'OTU01825', 'OTU03443', 'OTU02669', 'OTU02558', 'OTU02790', 'OTU01943', 'OTU03071', 'OTU02819', 'OTU00638', 'OTU02333', 'OTU03641', 'OTU00937', 'OTU03142', 'OTU03174', 'OTU03997', 'OTU03411', 'OTU02070', 'OTU02237', 'OTU03592', 'OTU06741', 'OTU05860', 'OTU01871', 'OTU01744', 'OTU00862', 'OTU01694', 'OTU00924', 'OTU04563', 'OTU02667', 'OTU03242', 'OTU02756', 'OTU03045', 'OTU04450', 'OTU00884', 'OTU03023', 'OTU02818', 'OTU02623', 'OTU02287', 'OTU02826', 'OTU02336', 'OTU03144', 'OTU03191', 'OTU04322', 'OTU00447', 'OTU02439', 'OTU05754', 'OTU03222', 'OTU02785', 'OTU02486', 'OTU03076', 'OTU00754', 'OTU01779', 'OTU02726', 'OTU06358', 'OTU01203', 'OTU02321', 'OTU00968', 'OTU01360', 'OTU03053', 'OTU01102', 'OTU01851']

PERMANOVA

PERMANOVA stands for Permutational Multivariate Analysis of Variance [16]. PERMANOVA feature selection method is very similar to ANOVA except that it operates in a distance matrix

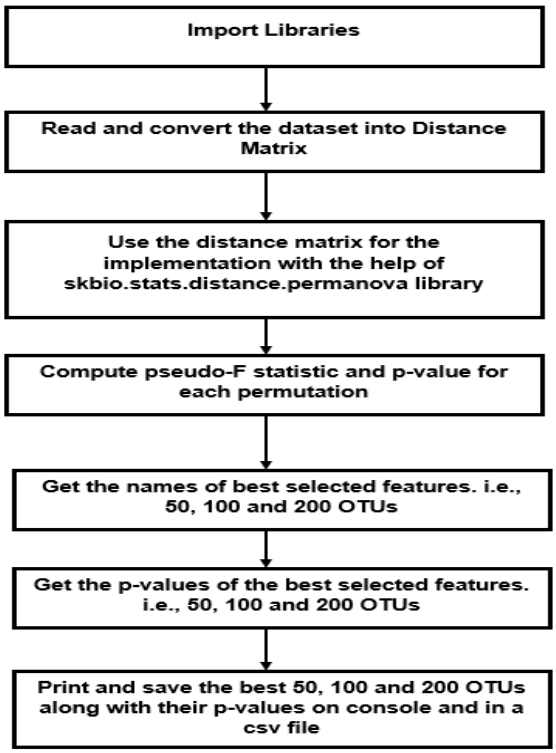
which allows for multivariate analysis. PERMANOVA tests if two or more groups of objects are significantly different based on a categorical factor. This method computes a pseudo-F statistic [16].

For the implementation of the PERMANOVA feature selection method by using the selected dataset, SciPy python library was used. Firstly, the whole dataset was converted into a distance matrix by using scipy.spatial library and distance_matrix function. Implementation after creating the distance matrix was not successful due to mismatch in the versions of libraries and the Python version. Due to the unsuccessful implementation, our client provided us with 50, 100 and 200 OTUs results for PERMANOVA.

The packages and parameters that were used by our client for the implementation of PERMANOVA feature selection method are skbio.stats.distance.PERMANOVA, NumPy, pandas, distance_matrix, grouping, column and permutations.

During PERMANOVA implementation, statistical significance was accessed via a permutation test. The assignment of objects to groups (grouping parameter used) is randomly permuted several times which was controlled via permutations parameter. A pseudo-F statistic is computed for each permutation and the p-value is the proportion of permuted pseud-F statistics that are equal to or greater than the original which means unpermuted pseudo-F statistic. PERMANOVA returns the statistical test which includes the test statistic and the p-value (probability value).

Figure 4. Overall workflow of PERMANOVA feature selection method



The resulted 50, 100 and 200 OTUs for PERMANOVA feature selection method from our dataset are listed below:

Best 50 OTUs

['OTU00589', 'OTU06543', 'OTU03408', 'OTU01374', 'OTU04145', 'OTU02457', 'OTU03373', 'OTU03396', 'OTU02625', 'OTU00424', 'OTU03683', 'OTU03862', 'OTU01356', 'OTU01453', 'OTU02456', 'OTU02858', 'OTU00947', 'OTU03024', 'OTU01575', 'OTU01700', 'OTU00868', 'OTU05457', 'OTU00488', 'OTU02925', 'OTU00657', 'OTU05044', 'OTU03768', 'OTU02508', 'OTU02231', 'OTU02331', 'OTU04119', 'OTU04258', 'OTU01183', 'OTU02702', 'OTU04210', 'OTU00845', 'OTU03806', 'OTU04326', 'OTU06861', 'OTU03434', 'OTU02750', 'OTU06892', 'OTU00854', 'OTU04517', 'OTU02856', 'OTU04256', 'OTU00414', 'OTU00304', 'OTU06834', 'OTU02737']

Best 100 OTUs

['OTU00589', 'OTU06543', 'OTU03408', 'OTU01374', 'OTU04145', 'OTU02457', 'OTU03373', 'OTU03396', 'OTU02625', 'OTU00424', 'OTU03683', 'OTU03862', 'OTU01356', 'OTU01453', 'OTU02456', 'OTU02858', 'OTU00947', 'OTU03024', 'OTU01575', 'OTU01700', 'OTU00868', 'OTU05457', 'OTU00488', 'OTU02925', 'OTU00657', 'OTU05044', 'OTU03768', 'OTU02508', 'OTU02231', 'OTU02331', 'OTU04119', 'OTU04258', 'OTU01183', 'OTU02702', 'OTU04210', 'OTU00845', 'OTU03806', 'OTU04326', 'OTU06861', 'OTU03434', 'OTU02750', 'OTU06892', 'OTU00854', 'OTU04517', 'OTU02856', 'OTU04256', 'OTU00414', 'OTU00304', 'OTU06834', 'OTU02737', 'OTU02235', 'OTU00420', 'OTU00407', 'OTU06591', 'OTU03968', 'OTU00408', 'OTU00685', 'OTU05486', 'OTU04166', 'OTU01317', 'OTU06343', 'OTU00294', 'OTU06552', 'OTU02008', 'OTU01867', 'OTU03243', 'OTU02983', 'OTU06331', 'OTU04904', 'OTU03382', 'OTU01461', 'OTU07027', 'OTU03082', 'OTU02679', 'OTU00366', 'OTU00758', 'OTU03556', 'OTU03366', 'OTU00411', 'OTU05332', 'OTU00205', 'OTU00028', 'OTU00512', 'OTU02542', 'OTU00555', 'OTU03571', 'OTU00958', 'OTU03335', 'OTU01964', 'OTU04918', 'OTU00337', 'OTU00423', 'OTU03524', 'OTU00012', 'OTU01342', 'OTU03428', 'OTU00785', 'OTU05844', 'OTU00859', 'OTU06558']

Best 200 OTUs

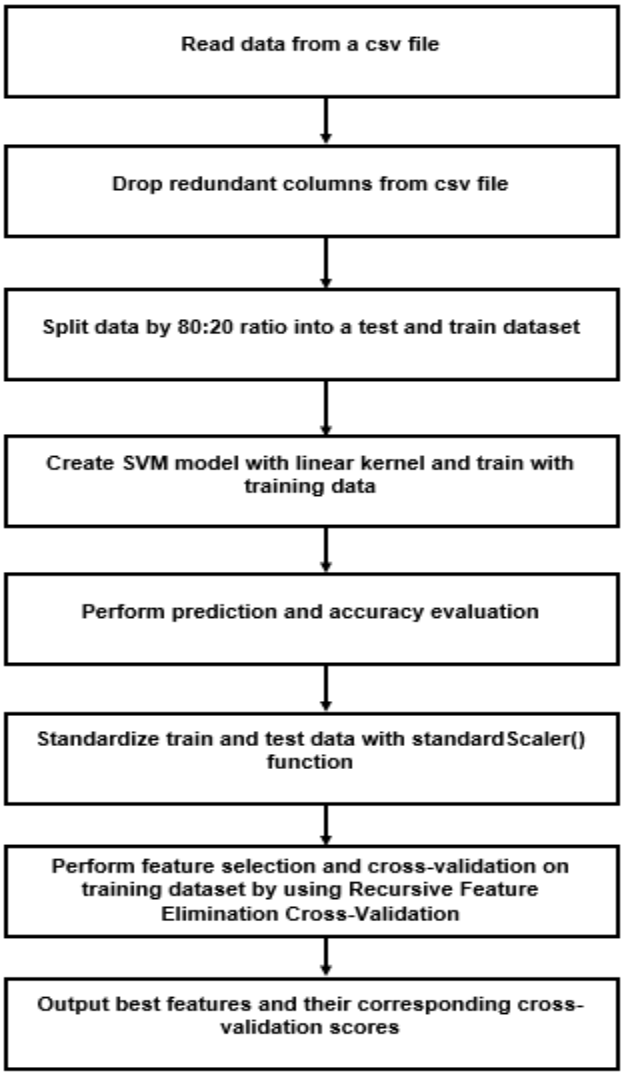
['OTU00589', 'OTU06543', 'OTU03408', 'OTU01374', 'OTU04145', 'OTU02457', 'OTU03373', 'OTU03396', 'OTU02625', 'OTU00424', 'OTU03683', 'OTU03862', 'OTU01356', 'OTU01453', 'OTU02456', 'OTU02858', 'OTU00947', 'OTU03024', 'OTU01575', 'OTU01700', 'OTU00868', 'OTU05457', 'OTU00488', 'OTU02925', 'OTU00657', 'OTU05044', 'OTU03768', 'OTU02508', 'OTU02231', 'OTU02331', 'OTU04119', 'OTU04258', 'OTU01183', 'OTU02702', 'OTU04210', 'OTU00845', 'OTU03806', 'OTU04326', 'OTU06861', 'OTU03434', 'OTU02750', 'OTU06892', 'OTU00854', 'OTU04517', 'OTU02856', 'OTU04256', 'OTU00414', 'OTU00304', 'OTU06834', 'OTU02737', 'OTU02235', 'OTU00420', 'OTU00407', 'OTU06591', 'OTU03968', 'OTU00408', 'OTU00685', 'OTU05486', 'OTU04166', 'OTU01317', 'OTU06343', 'OTU00294', 'OTU06552', 'OTU02008', 'OTU01867', 'OTU03243', 'OTU02983', 'OTU06331', 'OTU04904', 'OTU03382', 'OTU01461', 'OTU07027', 'OTU03082', 'OTU02679', 'OTU00366', 'OTU00758', 'OTU03556', 'OTU03366', 'OTU00411', 'OTU05332', 'OTU00205', 'OTU00028', 'OTU00512', 'OTU02542', 'OTU00555', 'OTU03571', 'OTU00958', 'OTU03335', 'OTU01964', 'OTU04918', 'OTU00337', 'OTU00423', 'OTU03524', 'OTU00012', 'OTU01342', 'OTU03428', 'OTU00785', 'OTU05844', 'OTU00859', 'OTU06558', 'OTU02397', 'OTU00047', 'OTU00814', 'OTU01676', 'OTU05622', 'OTU06594', 'OTU00301', 'OTU04081', 'OTU04471', 'OTU02029', 'OTU00749', 'OTU04819', 'OTU03638', 'OTU06033', 'OTU00165', 'OTU04523', 'OTU00190', 'OTU06526', 'OTU05830', 'OTU00691', 'OTU03602', 'OTU00016', 'OTU01152', 'OTU04665', 'OTU00976', 'OTU06695', 'OTU00925', 'OTU03698', 'OTU01380', 'OTU00747', 'OTU04607', 'OTU01860', 'OTU00908', 'OTU06439', 'OTU05629', 'OTU01770', 'OTU03576', 'OTU03310', 'OTU03301', 'OTU04013', 'OTU04903', 'OTU05407', 'OTU03119', 'OTU03350', 'OTU00546', 'OTU00281', 'OTU04465', 'OTU03226', 'OTU03892', 'OTU05479', 'OTU01048', 'OTU00443', 'OTU06058', 'OTU01536', 'OTU02036', 'OTU04369', 'OTU02263', 'OTU00716', 'OTU05777', 'OTU01219', 'OTU04900', 'OTU01148', 'OTU06325', 'OTU06872', 'OTU01452', 'OTU00521', 'OTU02441', 'OTU05342', 'OTU03800', 'OTU06920', 'OTU06735', 'OTU00586', 'OTU05452', 'OTU00046', 'OTU05106', 'OTU00271', 'OTU04185', 'OTU04649', 'OTU03430', 'OTU05305', 'OTU06936', 'OTU04373', 'OTU01674', 'OTU01522', 'OTU02314', 'OTU00939', 'OTU06826', 'OTU03174', 'OTU06918', 'OTU05456', 'OTU02509', 'OTU06699', 'OTU03034', 'OTU06659', 'OTU01485', 'OTU00145', 'OTU01913', 'OTU03381', 'OTU05293', 'OTU00130']

SVM

Support Vector Machine refers to a statistical and machine learning technique used on a variety of applications such as prediction (Guenther & Schonlau, 2016), pattern recognition (Cortes & Vapnik, 1995) and biological data processing (Evgeniou & Pontil, 2001).

The packages used for SVM implementation includes sklearn.train_test_split, sklearn.RFECV, sklearn.StandardScaler, sklearn.SVC, sklearn.accuracy_score. The corresponding parameters include X_train, y_train, cv, estimator, step, min_features_to_select, scoring ('accuracy'), kernel ('linear'), X_test, y_test and y_pred.

Figure 5. Overall workflow of SVM feature selection method



The biological dataset we received was first preprocessed by the client to make sure there were no null values and that they belonged to a specific data type. For the SVM algorithm to efficiently use the dataset, I first transposed the data from the original shape into a new shape in a new CSV file. Also, I manually created two general classes (B for Brassica and W for Wheat) for 11 samples to help effectively train the data with a sklearn library called train.test.split. Also, I tried to separate the two classes' data into two pandas data frames to find the best features in each class, which eventually produced errors. Subsequently, the raw data was used without separation, SVM classification with a linear kernel was applied to the raw data which yielded an accuracy of 0.66. Later, I experimented with the algorithm by using it together with a model named RFECV (Recursive Feature Elimination Cross Validation) to obtain the best features from the dataset. It yielded an accuracy of 100 percent with 3 optimal numbers of features (OTU05392, OTU03294, OTU06203). After discussion with the client, she requested that we generate the top 50 best features. I experimented with a similar sklearn model, RFE (Recursive Feature Elimination), which allowed for manual tuning of the number of features. The only setback was that it failed to output the probabilities/percentages of why the selected features were the top 50 features. After two-three hours of experimenting with different approaches from stackoverflow and scikit learn website, I went back to the RFECV algorithm and was able to come up with a solution that ranked the best 50 features as well as provide their corresponding percentages of why they were the top 50. It yielded an accuracy result of 100 percent and below is a list of the best features it produced:

Best 50 OTUs

[('OTU00589',) ('OTU06543',) ('OTU01848',) ('OTU03408',) ('OTU01374',) ('OTU04145',) ('OTU02457',) ('OTU03373',) ('OTU03396',)
('OTU02625',) ('OTU00424',) ('OTU03683',) ('OTU03862',) ('OTU01356',) ('OTU01453',) ('OTU02456',) ('OTU02858',) ('OTU01173',)
('OTU00947',) ('OTU03024',) ('OTU01575',) ('OTU02009',) ('OTU06221',) ('OTU00580',) ('OTU01700',) ('OTU01345',) ('OTU00868',)
('OTU05457',) ('OTU00080',) ('OTU02869',) ('OTU01339',) ('OTU02925',) ('OTU02318',) ('OTU01399',) ('OTU00657',) ('OTU03768',)
('OTU02508',) ('OTU01719',) ('OTU02331',) ('OTU04119',) ('OTU03097',) ('OTU04258',) ('OTU01183',) ('OTU02702',) ('OTU03204',)
('OTU00845',) ('OTU04791',) ('OTU03806',) ('OTU04326',) ('OTU02528',)]

Best 100 OTUs

[('OTU00589',) ('OTU06543',) ('OTU01848',) ('OTU03408',) ('OTU01374',) ('OTU04145',) ('OTU02457',) ('OTU03373',) ('OTU03396',)
('OTU02625',) ('OTU00424',) ('OTU03683',) ('OTU03862',) ('OTU01356',) ('OTU01453',) ('OTU02456',) ('OTU02858',) ('OTU01173',)
('OTU00947',) ('OTU03024',) ('OTU01575',) ('OTU02009',) ('OTU06221',) ('OTU00580',) ('OTU01700',) ('OTU01345',) ('OTU00868',)
('OTU05457',) ('OTU00080',) ('OTU02869',) ('OTU01339',) ('OTU02925',) ('OTU02318',) ('OTU01399',) ('OTU00657',) ('OTU03768',)
('OTU02508',) ('OTU01719',) ('OTU02331',) ('OTU04119',) ('OTU03097',) ('OTU04258',) ('OTU01183',) ('OTU02702',) ('OTU03204',)
('OTU00845',) ('OTU04791',) ('OTU03806',) ('OTU04326',) ('OTU02528',) ('OTU06861',) ('OTU01523',) ('OTU01110',) ('OTU03434',)
('OTU02750',) ('OTU01502',) ('OTU02495',) ('OTU06892',) ('OTU03118',) ('OTU00854',) ('OTU04517',) ('OTU02856',) ('OTU02815',)
('OTU03149',) ('OTU04256',) ('OTU00414',) ('OTU02988',) ('OTU02483',) ('OTU02235',) ('OTU01734',) ('OTU02557',) ('OTU03402',)
('OTU00994',) ('OTU06166',) ('OTU03341',) ('OTU03642',) ('OTU03968',) ('OTU03303',) ('OTU02809',) ('OTU02196',) ('OTU03664',)
('OTU00881',) ('OTU04856',) ('OTU00636',) ('OTU01031',) ('OTU03250',) ('OTU00685',) ('OTU05486',) ('OTU06516',) ('OTU05082',)
('OTU02647',) ('OTU04166',) ('OTU06590',) ('OTU01317',) ('OTU01227',) ('OTU00687',) ('OTU02926',) ('OTU06225',) ('OTU00294',)
('OTU02536',)]

Best 200 OTUs

[('OTU00589',) ('OTU06543',) ('OTU01848',) ('OTU03408',) ('OTU01374',) ('OTU04145',) ('OTU02457',) ('OTU03373',) ('OTU03396',)
('OTU02625',) ('OTU00424',) ('OTU03683',) ('OTU03862',) ('OTU01356',) ('OTU01453',) ('OTU02456',) ('OTU02858',) ('OTU01173',)
('OTU00947',) ('OTU03024',) ('OTU01575',) ('OTU02009',) ('OTU06221',) ('OTU00580',) ('OTU01700',) ('OTU01345',) ('OTU00868',)
('OTU05457',) ('OTU00080',) ('OTU02869',) ('OTU01339',) ('OTU02925',) ('OTU02318',) ('OTU01399',) ('OTU00657',) ('OTU03768',)
('OTU02508',) ('OTU01719',) ('OTU02331',) ('OTU04119',) ('OTU03097',) ('OTU04258',) ('OTU01183',) ('OTU02702',) ('OTU03204',)
('OTU00845',) ('OTU04791',) ('OTU03806',) ('OTU04326',) ('OTU02528',) ('OTU06861',) ('OTU01523',) ('OTU01110',) ('OTU03434',)
('OTU02750',) ('OTU01502',) ('OTU02495',) ('OTU06892',) ('OTU03118',) ('OTU00854',) ('OTU04517',) ('OTU02856',) ('OTU02815',)
('OTU03149',) ('OTU04256',) ('OTU00414',) ('OTU02988',) ('OTU02483',) ('OTU02235',) ('OTU01734',) ('OTU02557',) ('OTU03402',)
('OTU00994',) ('OTU06166',) ('OTU03341',) ('OTU03642',) ('OTU03968',) ('OTU03303',) ('OTU02809',) ('OTU02196',) ('OTU03664',)
('OTU00881',) ('OTU04856',) ('OTU00636',) ('OTU01031',) ('OTU03250',) ('OTU00685',) ('OTU05486',) ('OTU06516',) ('OTU05082',)
('OTU02647',) ('OTU04166',) ('OTU06590',) ('OTU01317',) ('OTU01227',) ('OTU00687',) ('OTU02926',) ('OTU06225',) ('OTU00294',)
('OTU02536',) ('OTU01748',) ('OTU01114',) ('OTU01435',) ('OTU00942',) ('OTU00929',) ('OTU02008',) ('OTU05078',) ('OTU01867',)
('OTU02983',) ('OTU02083',) ('OTU06331',) ('OTU03644',) ('OTU01178',) ('OTU01191',) ('OTU04904',) ('OTU03382',) ('OTU02863',)
('OTU01754',) ('OTU07027',) ('OTU01444',) ('OTU01190',) ('OTU01373',) ('OTU06893',) ('OTU02679',) ('OTU02624',) ('OTU04621',)
('OTU06099',) ('OTU00223',) ('OTU02630',) ('OTU05198',) ('OTU05240',) ('OTU00758',) ('OTU03556',) ('OTU06270',) ('OTU00205',)
('OTU02775',) ('OTU00461',) ('OTU06283',) ('OTU04877',) ('OTU00061',) ('OTU00512',) ('OTU03647',) ('OTU02542',) ('OTU02518',)
('OTU00958',) ('OTU03335',) ('OTU00896',) ('OTU03405',) ('OTU01964',) ('OTU02334',) ('OTU00337',) ('OTU00423',) ('OTU05297',)
('OTU05703',) ('OTU06164',) ('OTU03524',) ('OTU00869',) ('OTU01809',) ('OTU03415',) ('OTU05690',) ('OTU01342',) ('OTU03658',)
('OTU05844',) ('OTU06825',) ('OTU01357',) ('OTU00859',) ('OTU02355',) ('OTU03573',) ('OTU06558',) ('OTU02397',) ('OTU00738',)
('OTU04161',) ('OTU01375',) ('OTU04049',) ('OTU00814',) ('OTU01559',) ('OTU01540',) ('OTU01676',) ('OTU00697',) ('OTU00457',)
('OTU03583',) ('OTU02867',) ('OTU05622',) ('OTU02944',) ('OTU01939',) ('OTU02199',) ('OTU03317',) ('OTU06594',) ('OTU00301',)
('OTU01977',) ('OTU04471',) ('OTU00731',) ('OTU03780',) ('OTU05204',) ('OTU05145',) ('OTU02029',) ('OTU03377',) ('OTU00749',)
('OTU04819',) ('OTU02511',)]

Comparison/Evaluation

Comparison of the resulting best 50, 100 and 200 OTUs of SVM, chi-square, ANOVA and PERMANOVA feature selection methods was done by creating VENN diagrams [18], Scatter Plots [20] and Cluster Maps [21].

Cluster Maps

Cluster Maps are a visual representation of data classification that is easy to understand and manipulate. The Seaborn python library was used to create the cluster maps. Cluster map function in Seaborn was used to create hierarchically clustered heat maps with dendrograms on both rows and or columns by using hierarchical clusters to order the data by similarity.

Figure 6. Cluster Map of ANOVA - 50 OTUs

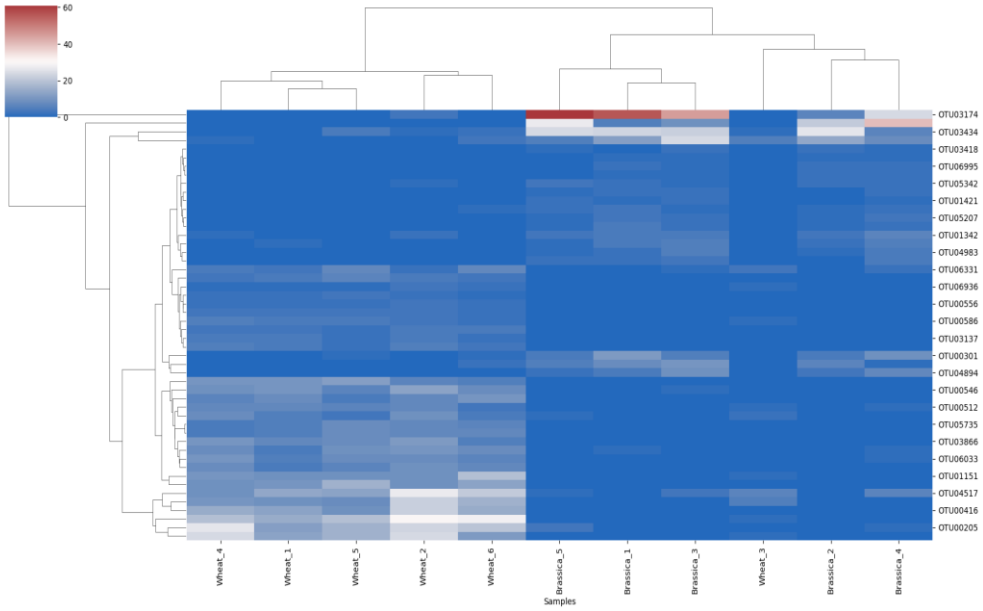


Figure 7. Cluster Map of Chi-Square - 50 OTUs

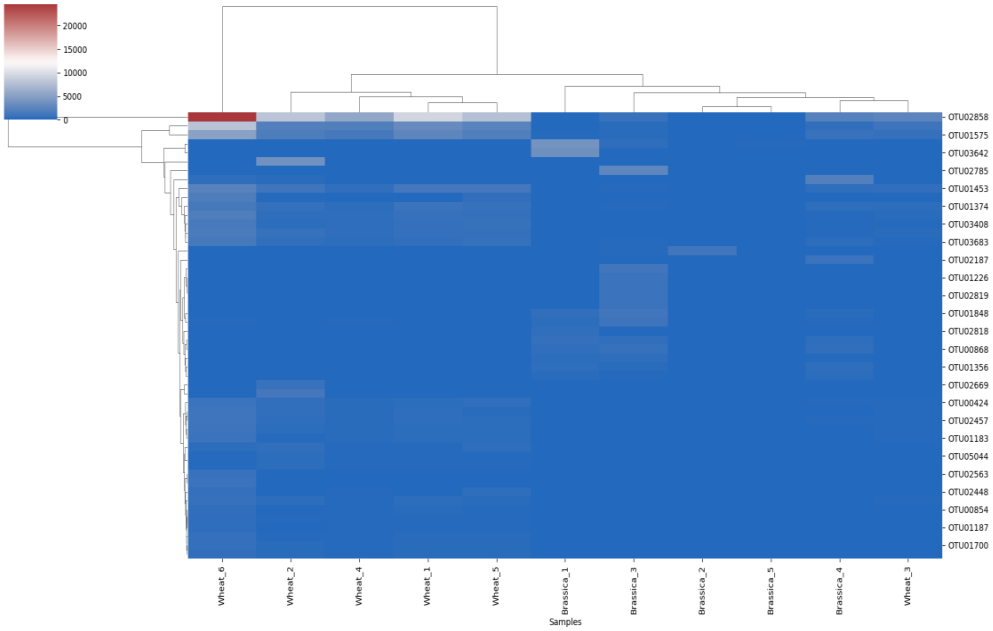


Figure 8. Cluster Map of PERMANOVA - 50 OTUs

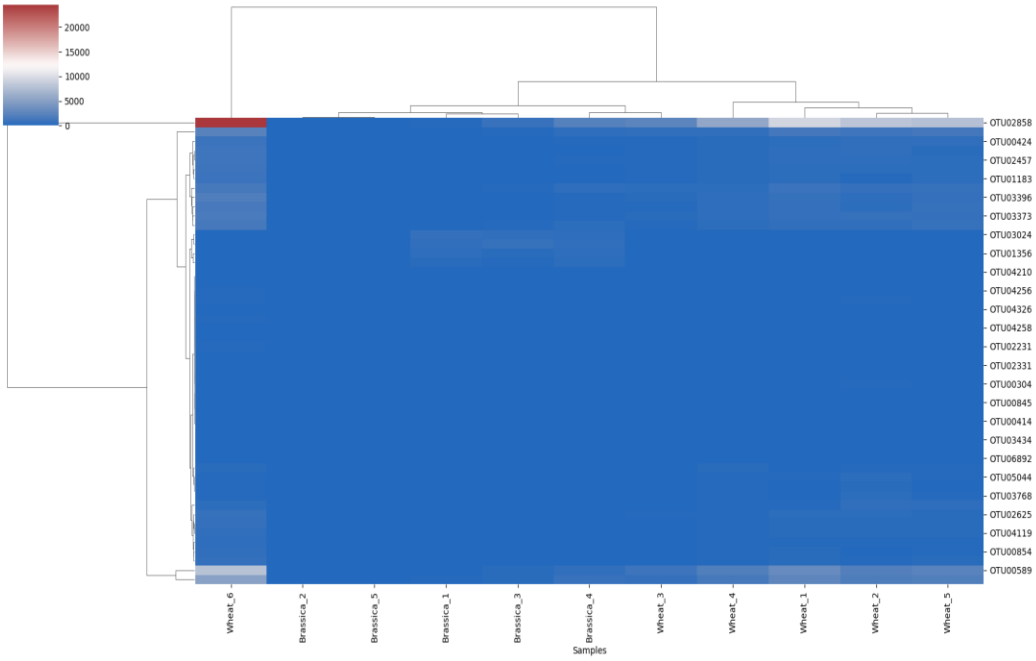
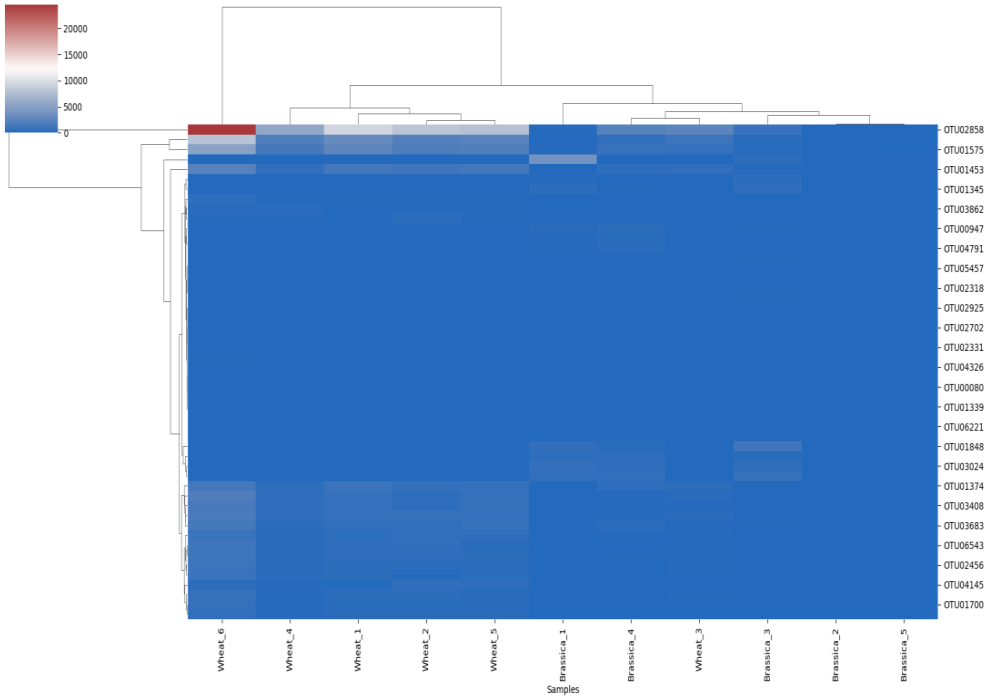


Figure 9. Cluster Map of SVM - 50 OTUs



For 50 OTU results of feature selection methods, ANOVA shows that Wheat_1, Wheat_2, Wheat_4, Wheat_5 and Wheat_6 samples behave in a similar way and Brassica_1, Brassica_2, Brassica_3, Brassica_4 and Wheat_3 samples behave in a similar way. On the other hand, Wheat_6 samples show significantly different behavior for chi-square, PERMANOVA and SVM feature selection methods in comparison to all other 10 samples.

Figure 10. Cluster Map of ANOVA - 100 OTUs

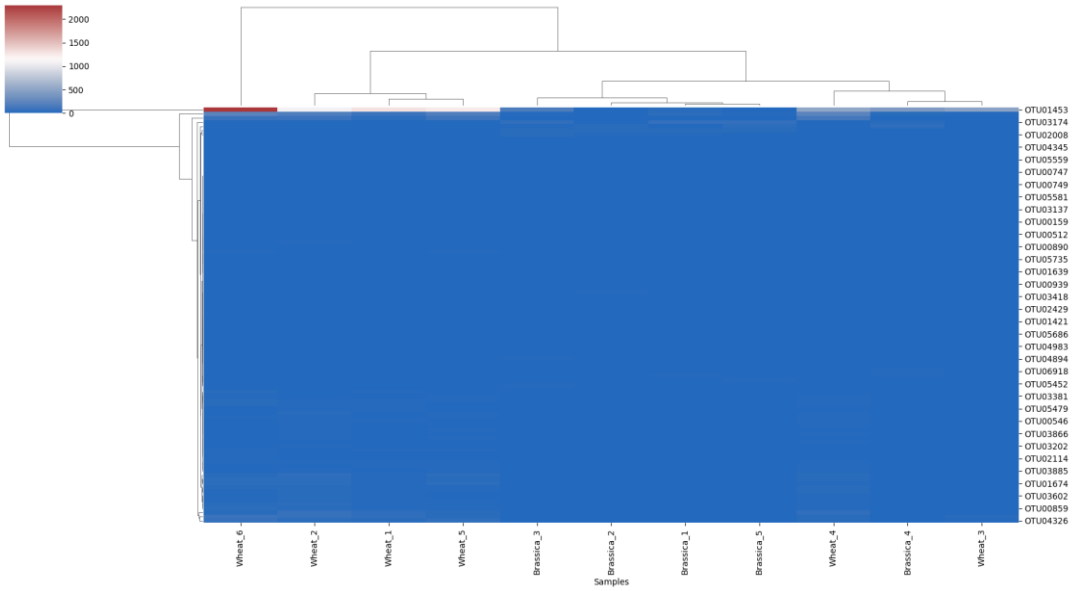


Figure 11. Cluster Map of Chi-Square - 100 OTUs

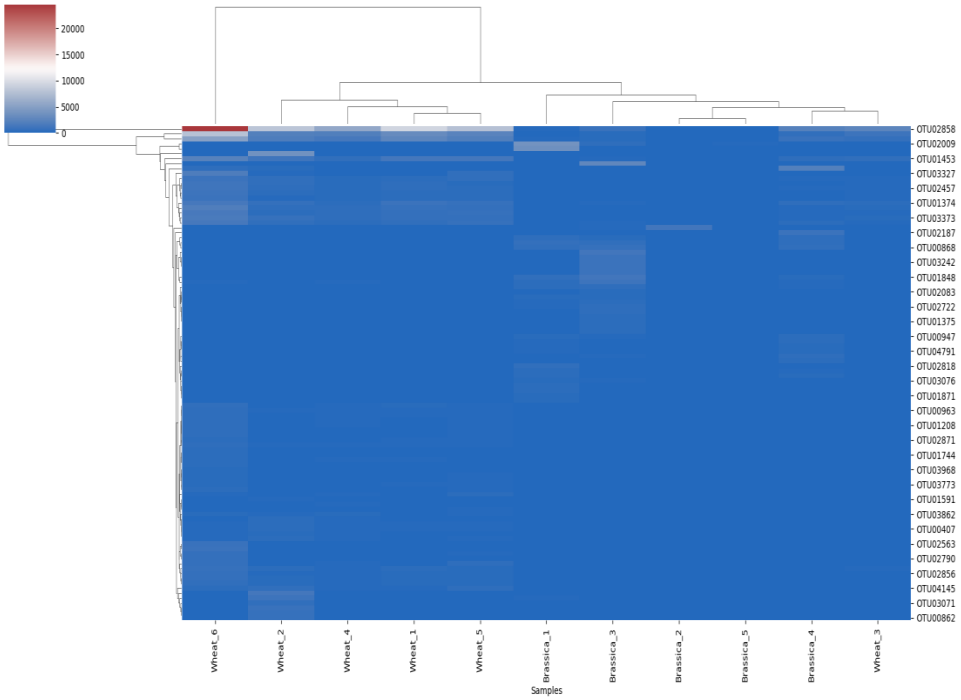


Figure 12. Cluster Map of PERMANOVA - 100 OTUs

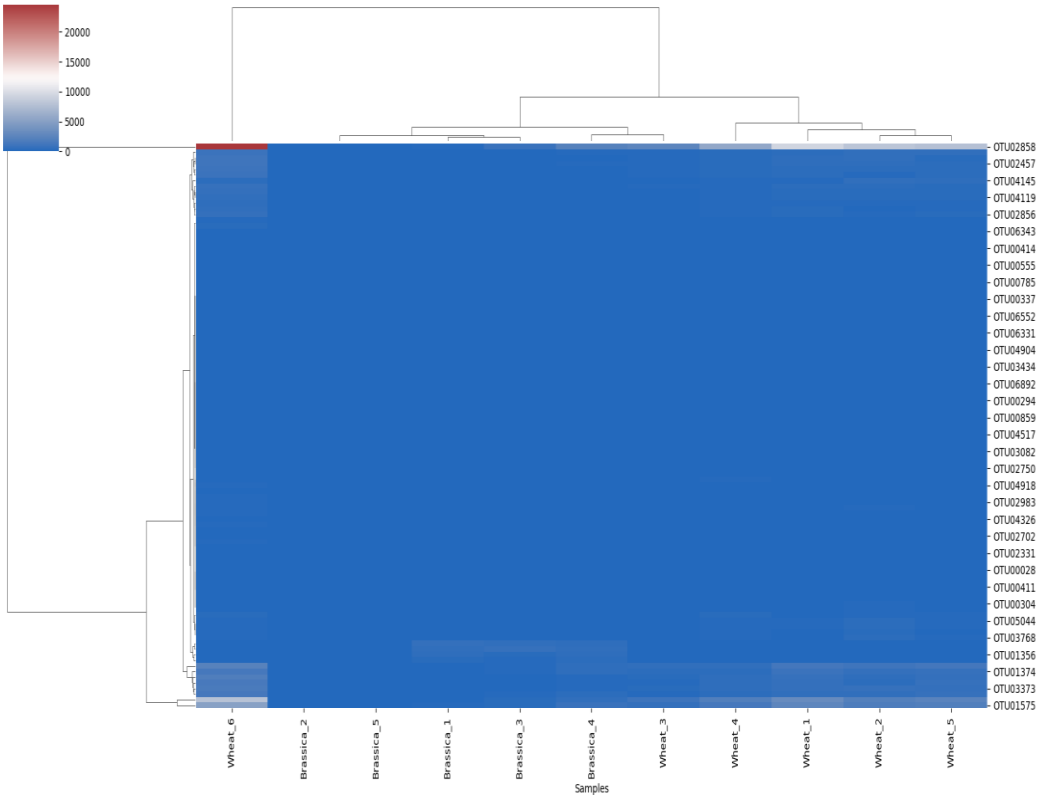
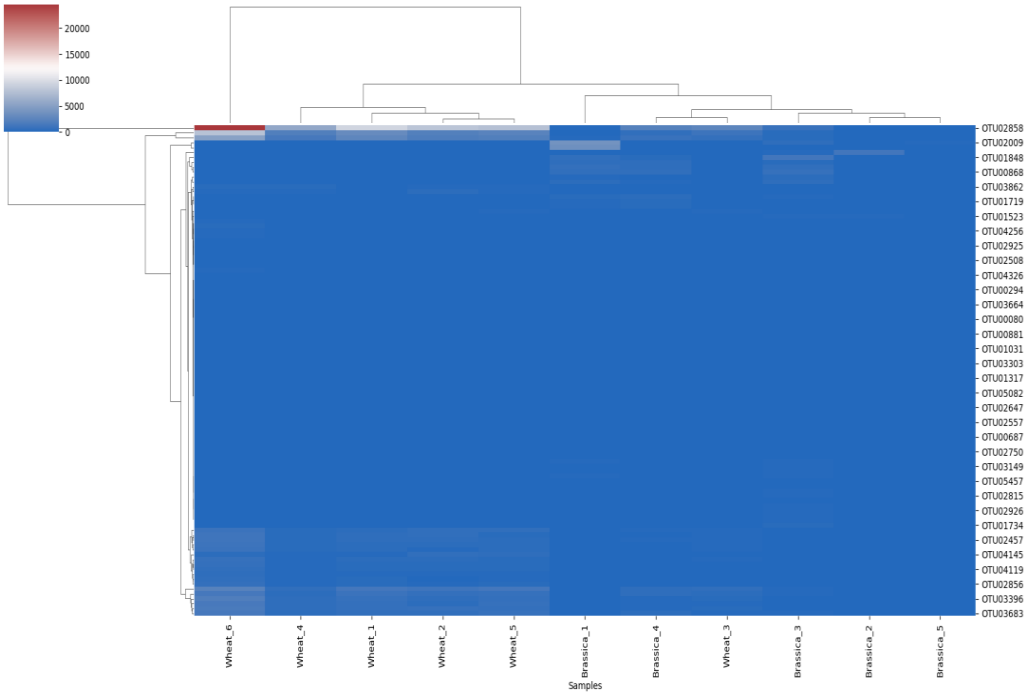


Figure 13. Cluster Map of SVM - 100 OTUs



Regarding 100 OTU results, Wheat_6 samples behave significantly differently in comparison with the other samples for all feature selection methods.

Figure 14. Cluster Map of ANOVA- 200 OTUs

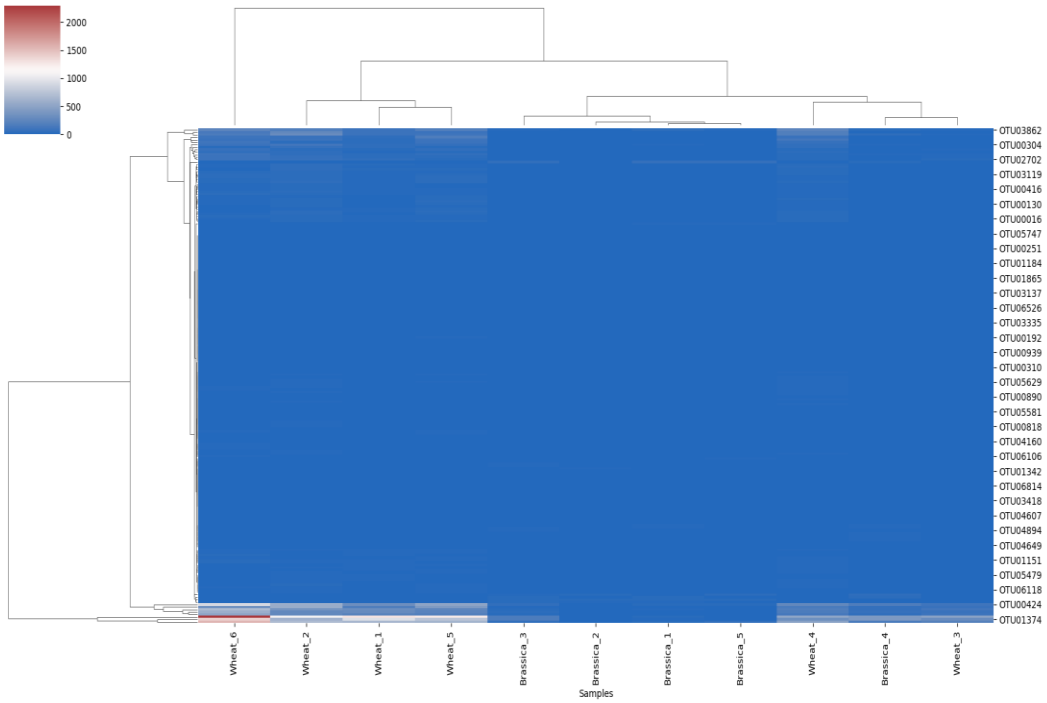


Figure 15. Cluster Map of Chi-Square - 200 OTUs

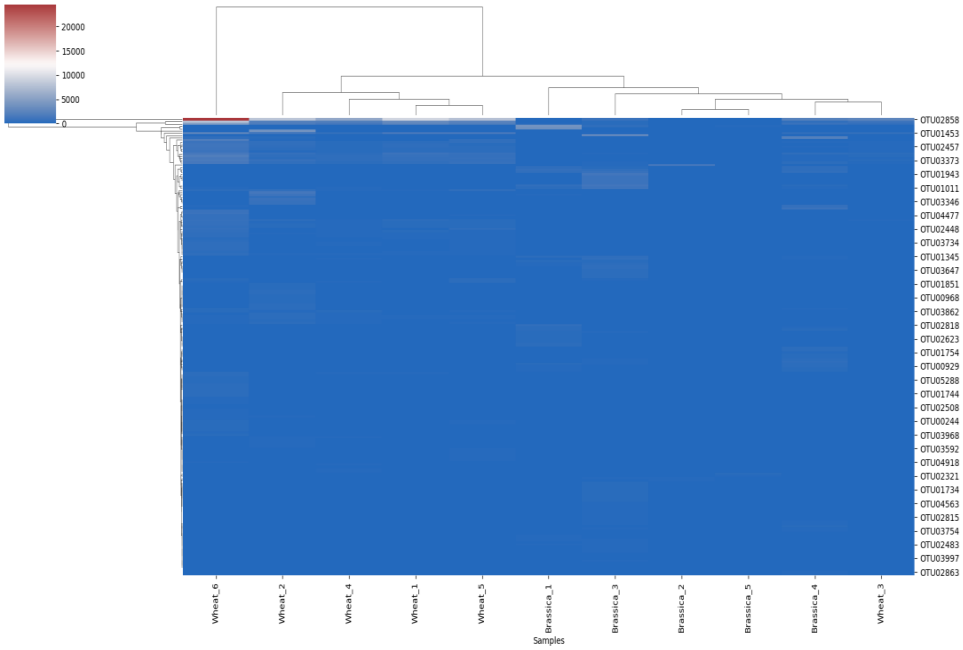


Figure 16. Cluster Map of PERMANOVA - 200 OTUs

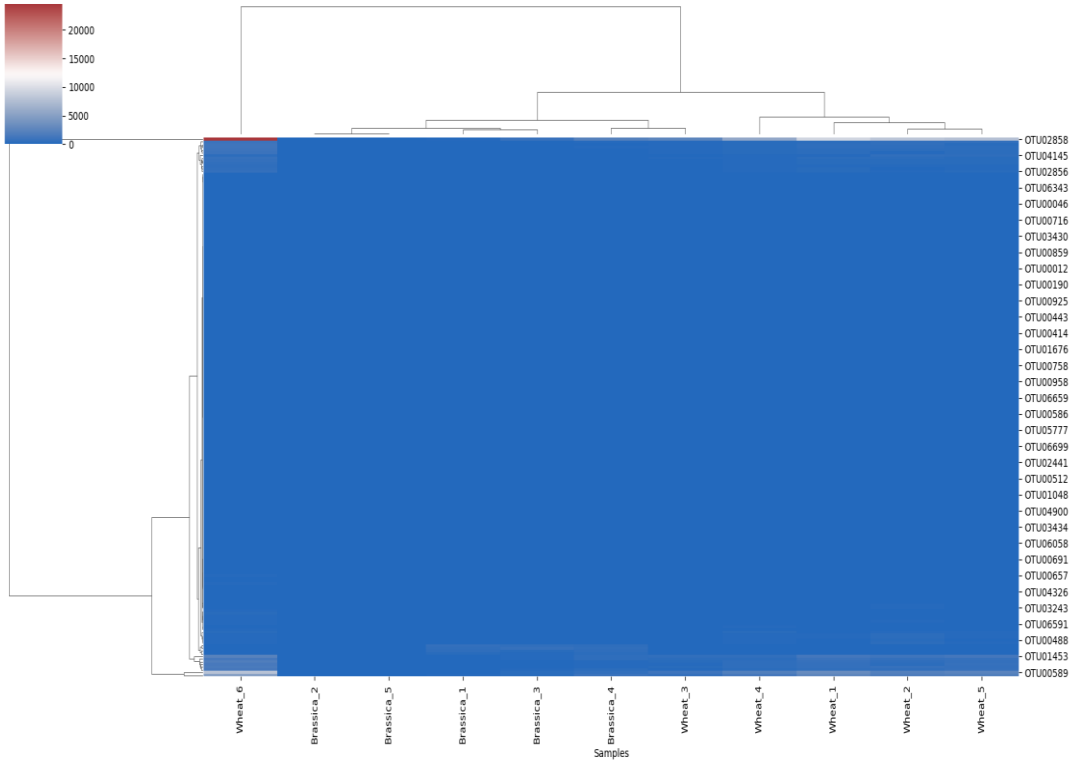
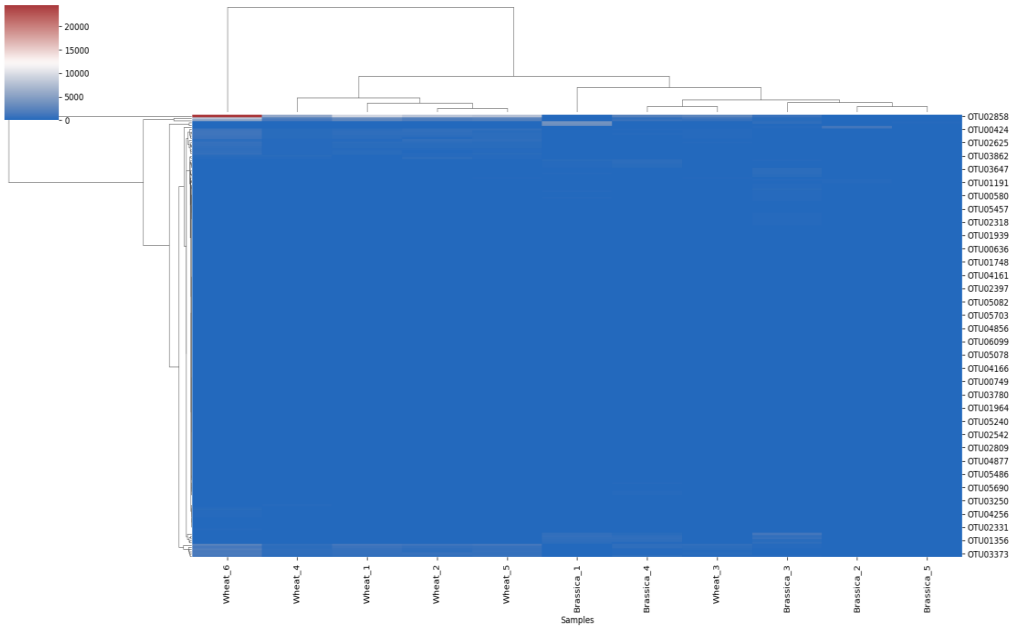


Figure 17. Cluster Map of SVM - 200 OTUs



Wheat_6 samples show significantly different behavior as compared to the other samples for all feature selection methods for 200 OTU results.

Scatter Plots

The diagrams for scatter plots for each feature selection method shows a measure of the probability for how the best 50, 100 and 200 OTUs could have occurred by a random chance. Every data point represents an OTU. For each OTU which has a p-value greater than 0.05 this means it is not statistically significant or more likely it was chosen by a random chance. If less than 0.05, then they are statistically significant or were purposely chosen.

Figure 18. Scatterplot of ANOVA - 50 OTUs

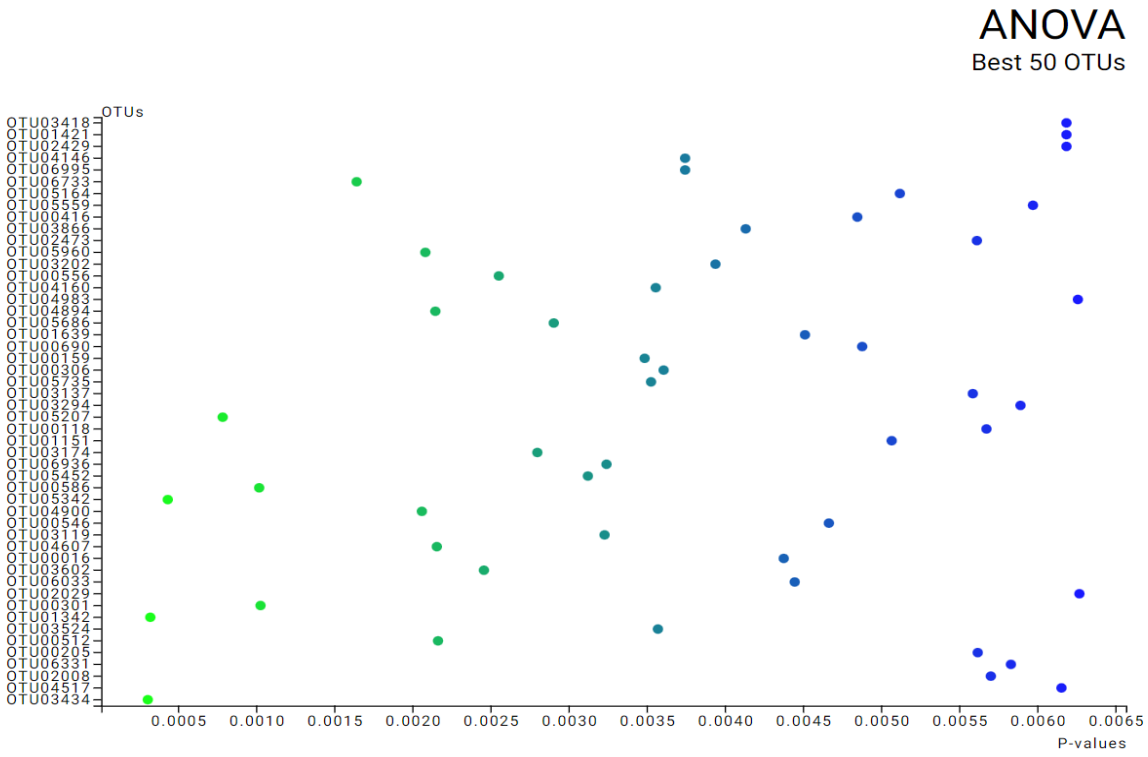


Figure 19. Scatterplot of Chi-Square - 50 OTUs

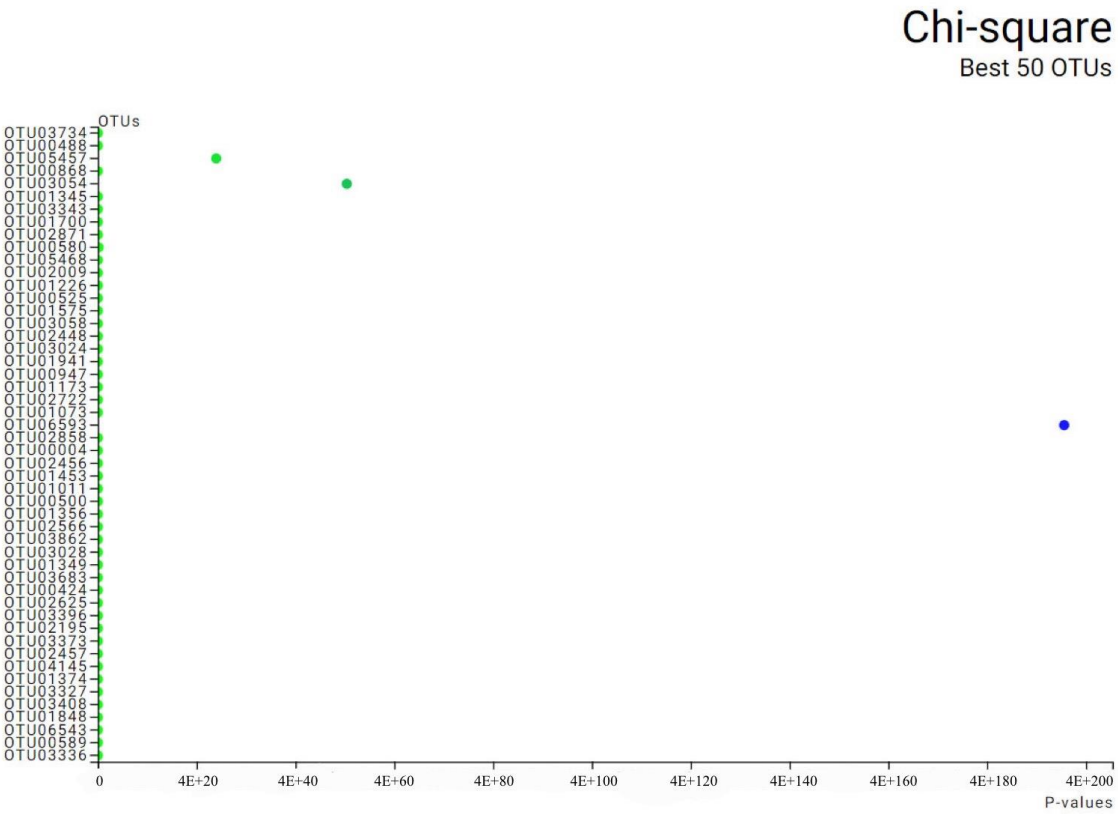


Figure 20. Scatterplot of PERMANOVA - 50 OTUs

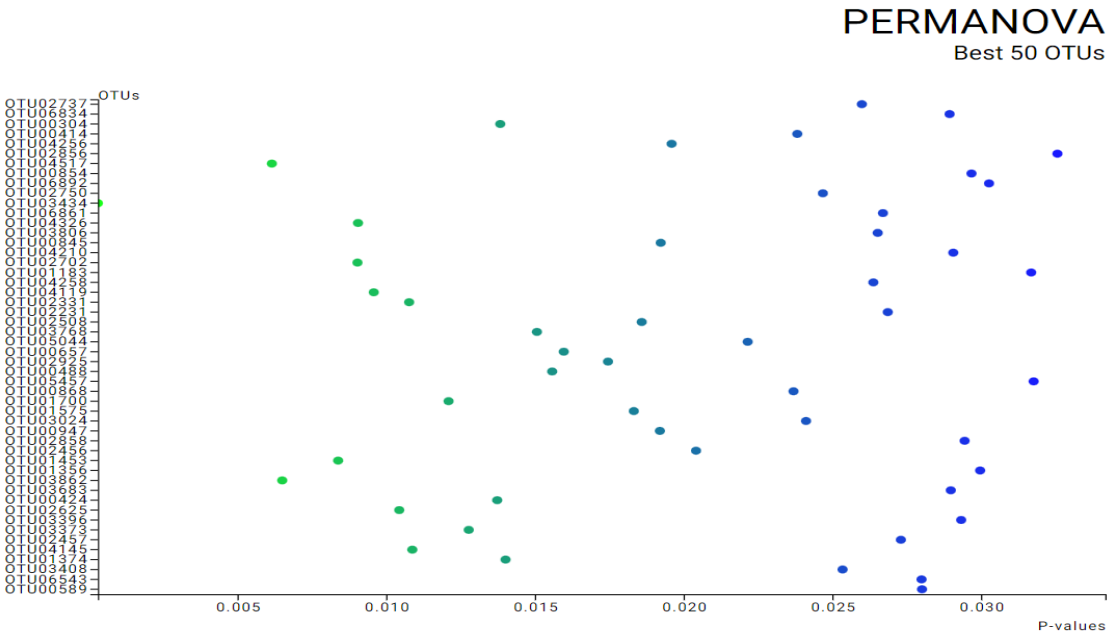
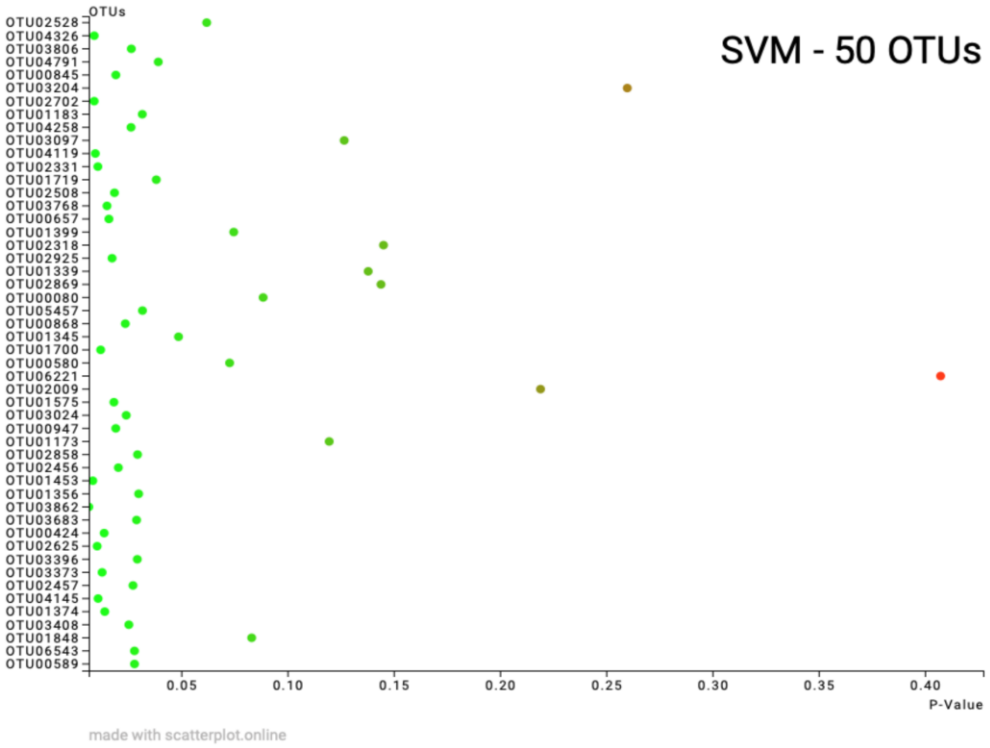


Figure 21. Scatterplot of SVM - 50 OTUs



Every point on the graph represents an OTU (Operational Taxonomic Unit). After analyzing all 4 graphs which depict the best 50 OTUs selected by each method, SVM came up with 39 OTUs with p-values less than 0.05 so the test hypothesis should be rejected and 11 OTUs greater than 0.05 so they should be accepted. On the other hand, chi-square has a lot of points very close to zero due to its way of calculating p-values which is completely different from the other three feature selection methods. So, all the OTUs should be rejected. PERMANOVA is showing that almost half of the points lie between 0.005 and 0.015 and the rest seem scattered between 0.020 to 0.065 which means that the test hypothesis should be rejected. For ANOVA 50 OTUs scatter plot, 7 p-values lie between 0.0005 and 0.0020 and rest of the p-values are scattered all over the graph and no p-value is greater than 0.0065. Hence, it is a statistically significant test result, and the test hypothesis should be rejected.

Figure 22. Scatterplot of ANOVA - 100 OTUs

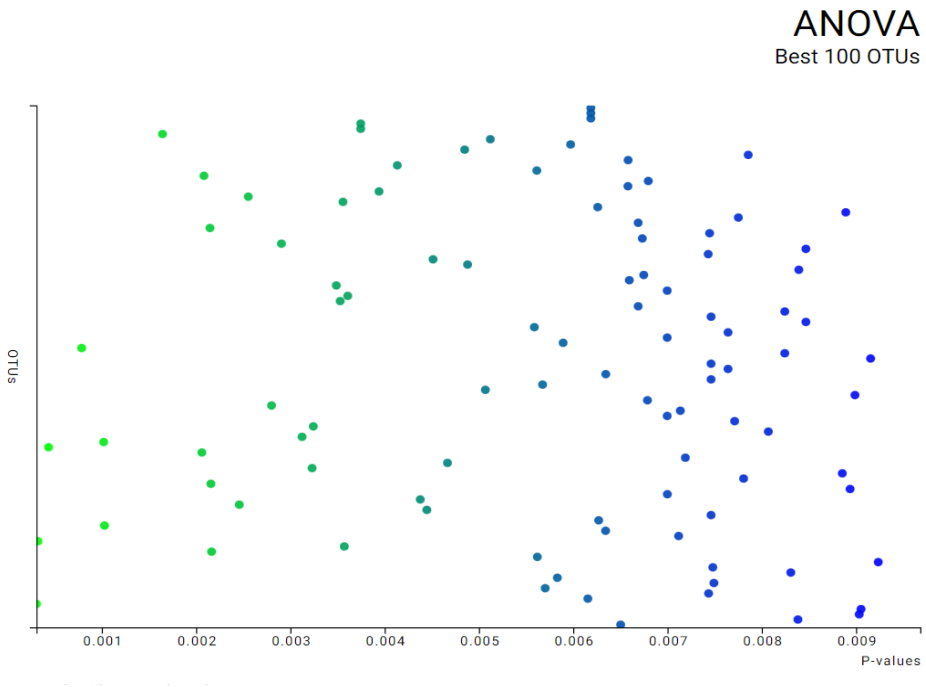


Figure 23. Scatterplot of Chi-Square - 100 OTUs

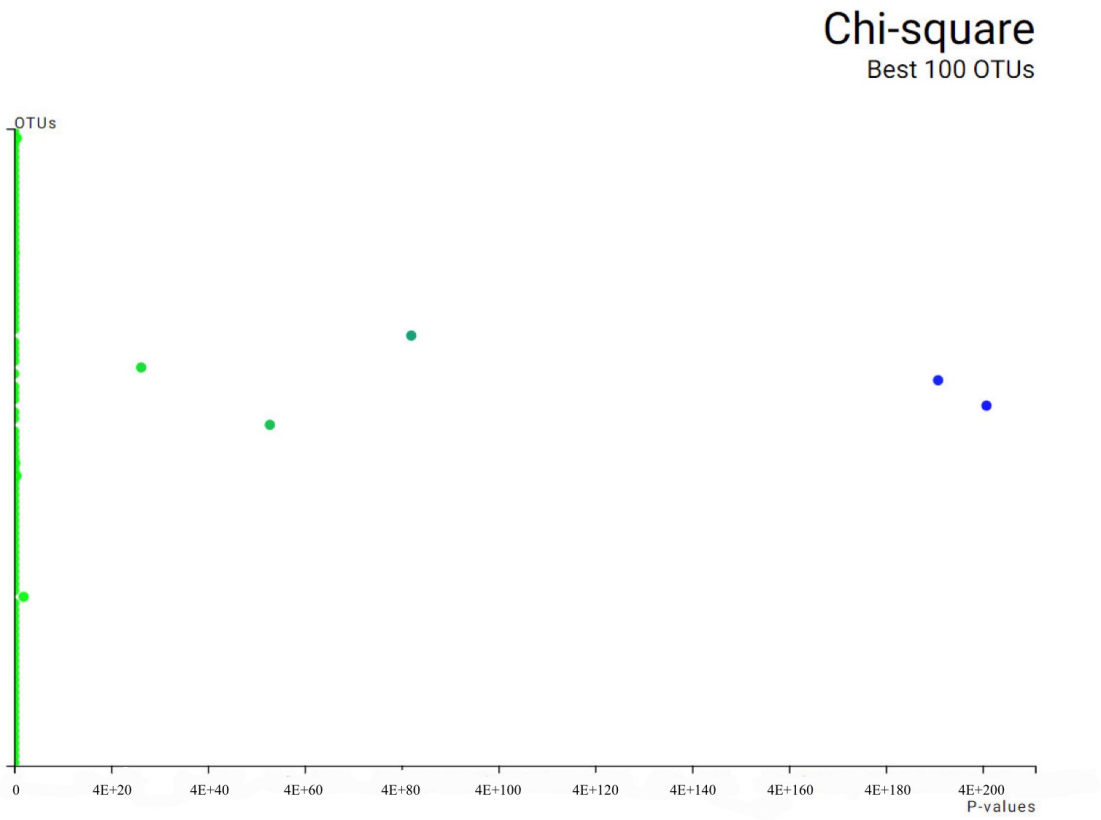


Figure 24. Scatterplot of PERMANOVA - 100 OTUs

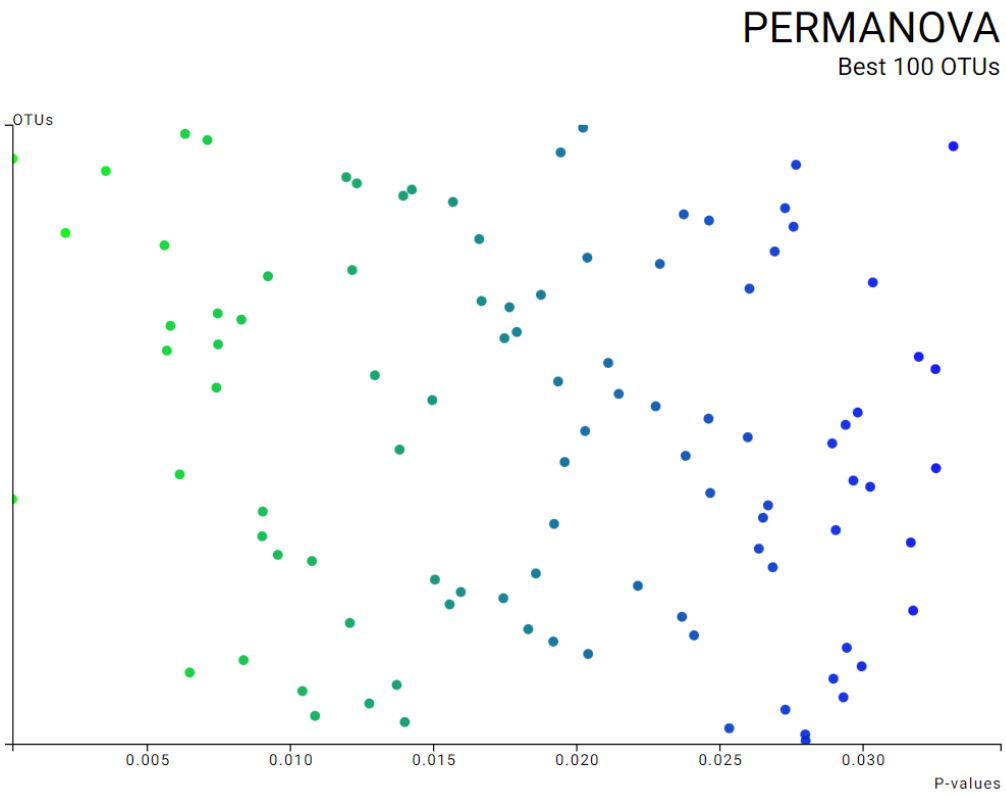
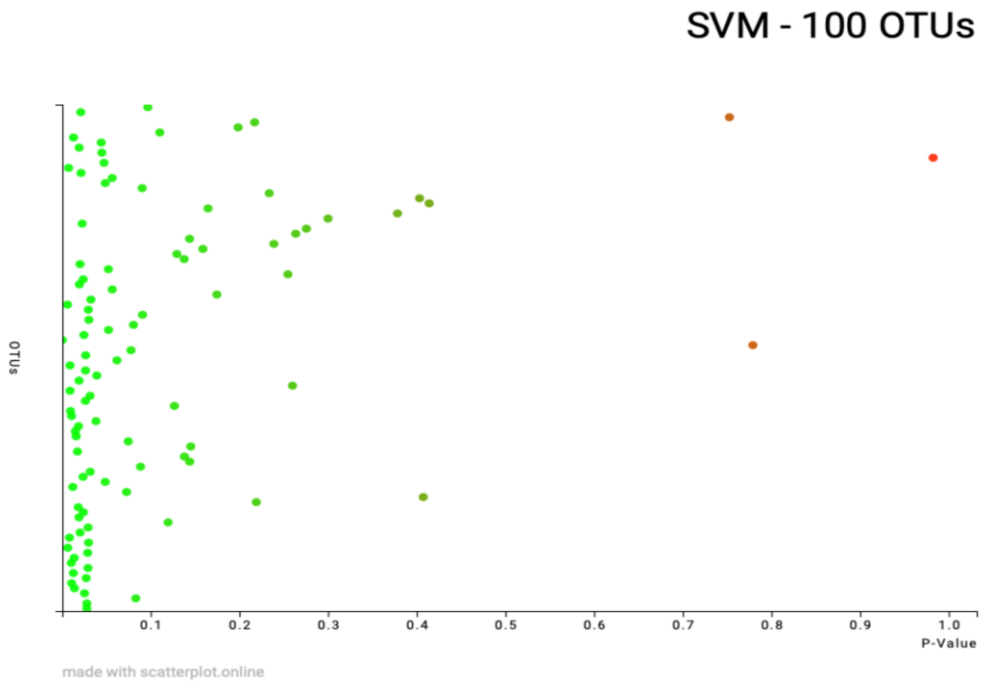


Figure 25. Scatterplot of SVM - 100 OTUs



After iterating the code to get output for the best 100 OTUs, ANOVA and PERMANOVA behave in a very similar way. The test hypothesis is false for both, and the hypothesis should be rejected as it is a statistically significant test result. The scatterplot of 100 OTUs shows that less than half of the p-values are less than 0.05 for SVM which indicates that it is a statistically significant test result, and the test hypothesis should be rejected and most of the p-values are greater than 0.05 so they should be accepted. The chi-square scatterplot for 100 OTUs manifest that all the p-values are less than 0.05 so the test hypothesis is rejected.

Figure 26. Scatterplot of ANOVA - 200 OTUs

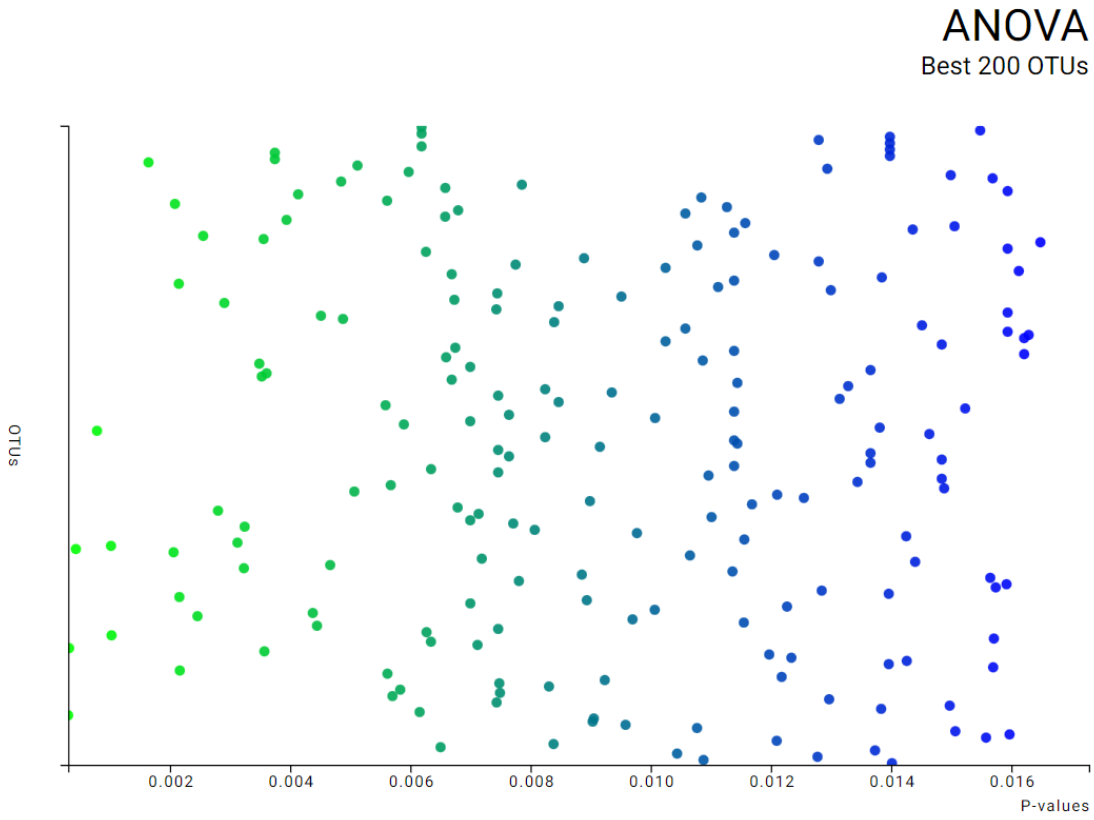


Figure 27. Scatterplot of Chi-Square - 200 OTUs

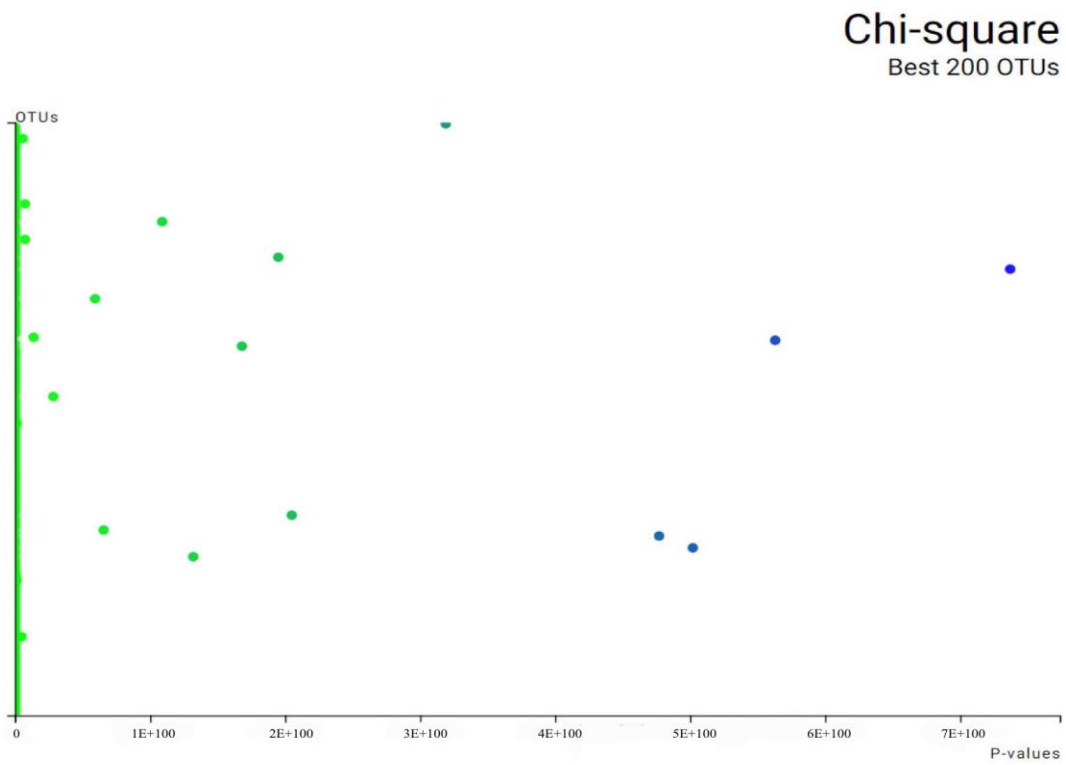


Figure 28. Scatterplot of PERMANOVA - 200 OTUs

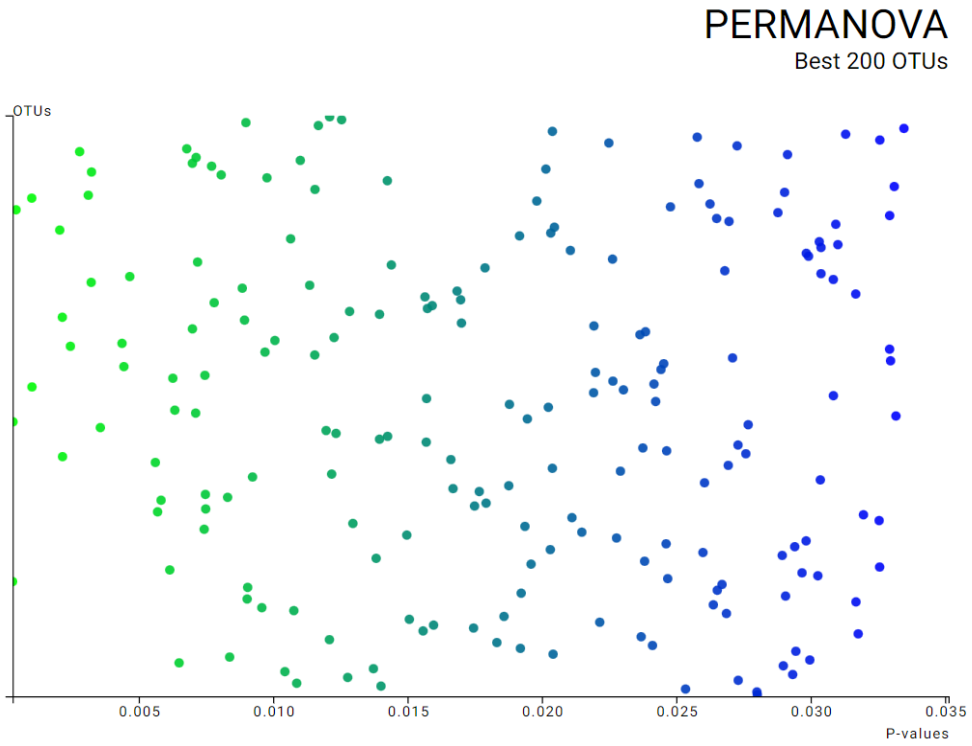
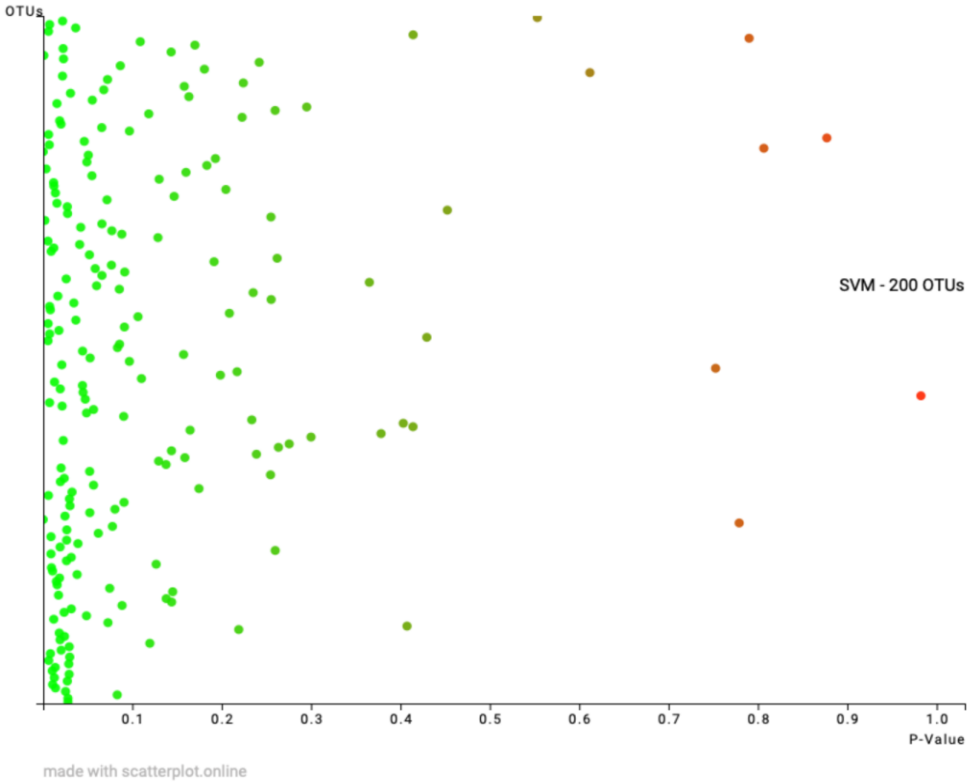


Figure 29. Scatterplot of SVM - 200 OTUs



A little scatter of points was observed in the chi-square scatter plot for 200 OTUs while most of the points are still very close to 0. Most of the probability values for SVM 200 OTUs scatterplot lie between 0 and 0.5 which means we cannot conclude that there is a statistically significant difference between them, and the test hypothesis is true or should be accepted, but the plots that lie between 0 and 0.05 shows a significant difference and should be rejected. Both ANOVA and PERMANOVA behave in a very similar way. The test hypothesis is false for both, and the hypothesis should be rejected as it is a statistically significant test result.

Venn Diagrams

The following Venn diagrams show the common OTUs between the four feature selection methods ANOVA, PERMANOVA, chi-square and SVM. They tell us how similar the methods are operating, and which ones are completely different from each other.

Figure 30. Venn Diagram of 50 OTUs

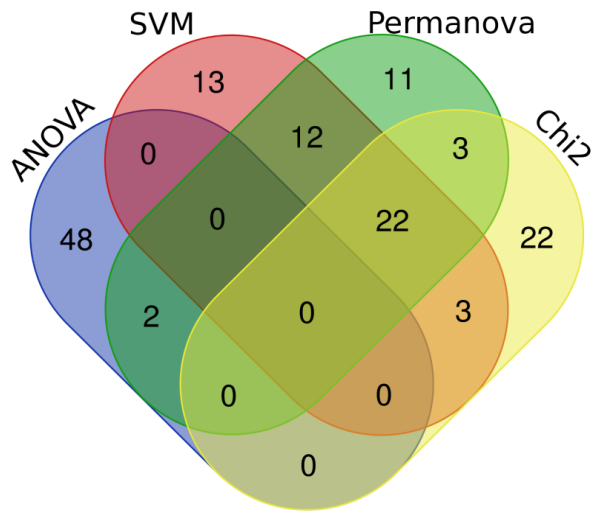


Figure 31. Chart of 50 OTUs names per above Venn Diagram

Names	total	elements
Chi2 Permanova SVM	22	OTU01575 OTU02457 OTU01700 OTU01374 OTU04119 OTU04145 OTU03408 OTU02625 OTU02858 OTU03024 OTU06543 OTU00424 OTU03373 OTU01183 OTU00589 OTU00868 OTU02456 OTU00947 OTU01453 OTU01356 OTU03683 OTU03396
ANOVA Permanova	2	OTU03434 OTU04517
Permanova SVM	12	OTU02508 OTU00657 OTU03862 OTU02925 OTU05457 OTU00845 OTU03768 OTU03806 OTU04326 OTU04258 OTU02702 OTU02331
Chi2 SVM	3	OTU02009 OTU01848 OTU01345
Chi2 Permanova	3	OTU00854 OTU05044 OTU02856
ANOVA	48	OTU04607 OTU02473 OTU00416 OTU04894 OTU00301 OTU06033 OTU01342 OTU05207 OTU05452 OTU03866 OTU03137 OTU06733 OTU00306 OTU00546 OTU06331 OTU04900 OTU05735 OTU05559 OTU02429 OTU04160 OTU03174 OTU00016 OTU05960 OTU00512 OTU03294 OTU03418 OTU06936 OTU04983 OTU00556 OTU04146 OTU00586 OTU02008 OTU05164 OTU00159 OTU03602 OTU05342 OTU01639 OTU00205 OTU00118 OTU01151 OTU02029 OTU01421 OTU03119 OTU00690 OTU05686 OTU03524 OTU03202 OTU06995
SVM	13	OTU01339 OTU04791 OTU01173 OTU06221 OTU00580 OTU02318 OTU02869 OTU03204 OTU00080 OTU03097 OTU01719 OTU01399 OTU02528
Permanova	11	OTU00414 OTU00488 OTU06892 OTU00304 OTU04210 OTU06834 OTU06861 OTU02750 OTU02737 OTU02231 OTU04256
Chi2	22	ï»¿OTU03336 OTU03642 OTU01187 OTU00407 OTU02818 OTU00994 OTU02669 OTU02070 OTU01011 OTU03242 OTU02785 OTU02667 OTU00963 OTU00767 OTU01943 OTU01226 OTU02187 OTU02448 OTU02563 OTU02819 OTU03327 OTU04973

Regarding the 50 OTUs Venn diagram, it tells us that PERMANOVA, chi-square and SVM act very similarly to each other as they have 22 OTUs which are common in their set of best 50 OTUs. On the other hand, ANOVA and PERMANOVA methods are dissimilar to each other as they only have 2 OTUs which are common between them.

Figure 32. Venn Diagram of 100 OTUs

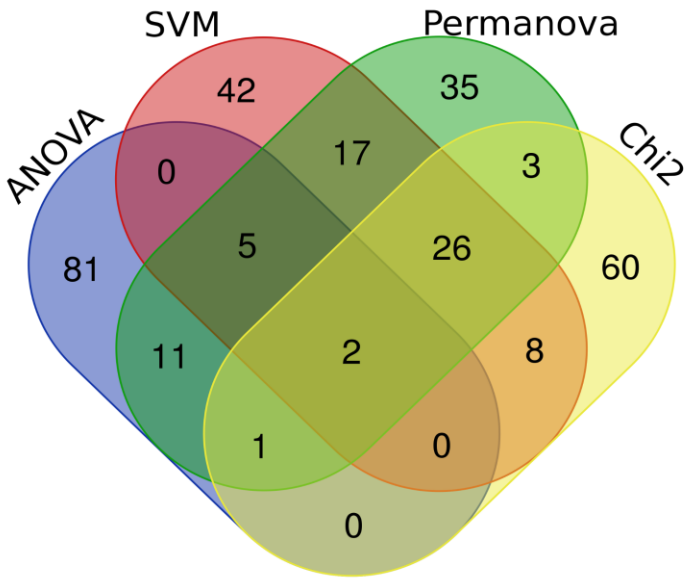


Figure 33. Chart of 100 OTUs names per above Venn Diagram

Names	total	elements
ANOVA Chi2 Permanova SVM	2	OTU03862 OTU01453
ANOVA Permanova SVM	5	OTU04517 OTU05486 OTU04326 OTU03434 OTU02702
ANOVA Chi2 Permanova	1	OTU01867
Chi2 Permanova SVM	26	OTU01575 OTU01374 OTU04119 OTU03768 OTU03968 OTU02858 OTU03024 OTU06543 OTU00424 OTU03373 OTU00589 OTU02456 OTU00947 OTU03396 OTU02457 OTU01700 OTU04145 OTU03408 OTU00854 OTU03806 OTU02625 OTU02856 OTU01183 OTU00868 OTU01356 OTU03683
ANOVA Permanova	11	OTU01342 OTU06331 OTU05844 OTU02008 OTU00205 OTU03524 OTU00859 OTU03382 OTU00512 OTU00758 OTU04904
Permanova SVM	17	OTU00414 OTU06892 OTU05457 OTU00685 OTU02235 OTU04166 OTU04258 OTU02750 OTU00294 OTU02508 OTU00657 OTU02925 OTU01317 OTU00845 OTU06861 OTU02331 OTU04256
Chi2 SVM	8	OTU03642 OTU02009 OTU00994 OTU01345 OTU04791 OTU01173 OTU01848 OTU01719
Chi2 Permanova	3	OTU00488 OTU05044 OTU00407
ANOVA	81	OTU02473 OTU00890 OTU00749 OTU00301 OTU06033 OTU03866 OTU03137 OTU05479 OTU06918 OTU05305 OTU05735 OTU05559 OTU03174 OTU05747 OTU05960 OTU00016 OTU03294 OTU04983 OTU04903 OTU06118 OTU05581 OTU00556 OTU02010 OTU00586 OTU04980 OTU05164 OTU00925 OTU01579 OTU00159 OTU01865 OTU03602 OTU01151 OTU03135 OTU01421 OTU03119 OTU00747 OTU06995 OTU04607 OTU01770 OTU00240 OTU00416 OTU04894 OTU00124 OTU00939 OTU05207 OTU05452 OTU06733 OTU00306 OTU06961 OTU00546 OTU06774 OTU04900 OTU04287 OTU01522 OTU04345 OTU02114 OTU02429 OTU04160 OTU02127 OTU06929 OTU03418 OTU06936 OTU03885 OTU04146 OTU05619 OTU00889 OTU05806 OTU00406 OTU01674 OTU05342 OTU01639 OTU00118 OTU02029 OTU04636 OTU03381 OTU05686 OTU00690 OTU00268 OTU01239 OTU03202 OTU00251
SVM	42	OTU01734 OTU01227 OTU00687 OTU06516 OTU01502 OTU01110 OTU06225 OTU00580 OTU04856 OTU03341 OTU02318 OTU02869 OTU03204 OTU02483 OTU00080 OTU03097 OTU02926 OTU02809 OTU02988 OTU02495 OTU01399 OTU02196 OTU03303 OTU03250 OTU03664 OTU01339 OTU02557 OTU06221 OTU00881 OTU03149 OTU05082 OTU06590 OTU00636 OTU06166 OTU03402 OTU02815 OTU03118 OTU02536 OTU01031 OTU02647 OTU01523 OTU02528
Permanova	35	OTU03082 OTU00785 OTU00555 OTU01461 OTU02983 OTU00028 OTU00012 OTU00420 OTU00304 OTU03243 OTU03556 OTU06591 OTU04210 OTU06558 OTU00366 OTU00958 OTU03571 OTU07027 OTU02231 OTU00408 OTU06552 OTU03428 OTU01964 OTU03366 OTU06343 OTU00337 OTU00411 OTU04918 OTU05332 OTU06834 OTU02542 OTU03335 OTU02737 OTU02679 OTU00423
Chi2	60	OTU03336 OTU04450 OTU01187 OTU02818 OTU00244 OTU01030 OTU03501 OTU02826 OTU02871 OTU04477 OTU02513 OTU01375 OTU03346 OTU00661 OTU01011 OTU03076 OTU01744 OTU02790 OTU06438 OTU02785 OTU02566 OTU03315 OTU00767 OTU01943 OTU01226 OTU02819 OTU03071 OTU02083 OTU00862 OTU03520 OTU04973 OTU03773 OTU03070 OTU02669 OTU02070 OTU00525 OTU01208 OTU02237 OTU01871 OTU00929 OTU03242 OTU03023 OTU00754 OTU02667 OTU00963 OTU02722 OTU01941 OTU05468 OTU02187 OTU05754 OTU03028 OTU02448 OTU02563 OTU01591 OTU03734 OTU00524 OTU03327 OTU02756 OTU00482 OTU03647

In relation to the 100 OTUs, chi-square, PERMANOVA and SVM still top the chart as the most similar group of methods. A new discovery since the 50 OTUs diagram is that chi-square, PERMANOVA and ANOVA now appear to be the three methods with the lowest number of common OTUs between them, which is 3 OTUs. This tells us that when the number of OTUs produced increases, the similarity between methods also increases.

Figure 34. Venn Diagram of 200 OTUs

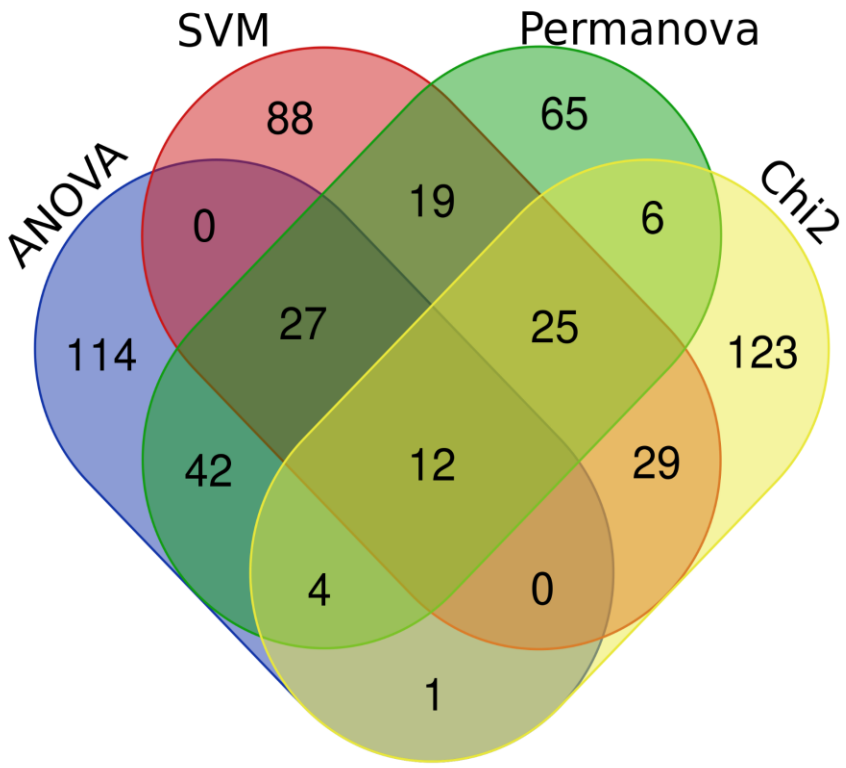


Figure 35. Chart of 200 OTUs names per above Venn Diagram

Names	total	elements
ANOVA Chi2 Permanova SVM	12	OTU04119 OTU03862 OTU00424 OTU01867 OTU02625 OTU02702 OTU01453 OTU01374 OTU03768 OTU03373 OTU01700 OTU04145
ANOVA Permanova SVM	27	OTU04517 OTU00749 OTU03556 OTU02008 OTU00205 OTU03434 OTU01964 OTU02029 OTU00301 OTU01342 OTU06331 OTU05486 OTU04326 OTU05844 OTU03524 OTU00859 OTU03382 OTU00657 OTU01317 OTU00512 OTU00337 OTU03335 OTU00758 OTU04904 OTU00814 OTU02331 OTU00423
ANOVA Chi2 Permanova	4	OTU00488 OTU00304 OTU04918 OTU03174
Chi2 Permanova SVM	25	OTU05457 OTU03024 OTU03396 OTU02508 OTU03408 OTU01183 OTU02679 OTU04256 OTU01575 OTU02983 OTU02235 OTU03968 OTU02858 OTU06543 OTU05622 OTU00589 OTU02456 OTU00947 OTU02457 OTU00854 OTU03806 OTU02856 OTU00868 OTU01356 OTU03683
ANOVA Permanova	42	OTU06033 OTU05305 OTU01152 OTU04903 OTU00586 OTU00130 OTU04665 OTU00925 OTU03119 OTU00747 OTU02314 OTU01770 OTU00939 OTU05452 OTU00908 OTU03310 OTU01913 OTU03892 OTU06936 OTU06526 OTU05293 OTU01674 OTU05830 OTU06439 OTU05629 OTU05479 OTU06918 OTU00016 OTU04649 OTU03602 OTU00408 OTU01860 OTU04607 OTU00546 OTU04900 OTU05407 OTU01522 OTU03430 OTU05342 OTU05106 OTU03381 OTU00716
ANOVA Chi2	1	OTU03023
Permanova SVM	19	OTU01676 OTU06892 OTU00685 OTU06558 OTU04258 OTU02750 OTU07027 OTU00294 OTU02925 OTU02397 OTU02542 OTU06594 OTU06861 OTU00414 OTU04471 OTU04166 OTU00958 OTU04819 OTU00845
Chi2 SVM	29	OTU00994 OTU04621 OTU01375 OTU00580 OTU05690 OTU02083 OTU01754 OTU01345 OTU01191 OTU02863 OTU06270 OTU03149 OTU03402 OTU02815 OTU00061 OTU01719 OTU01523 OTU03647 OTU03642 OTU01734 OTU02009 OTU01502 OTU02318 OTU02483 OTU04791 OTU01173 OTU01848 OTU06590 OTU00929
Chi2 Permanova	6	OTU06591 OTU02231 OTU02737 OTU05044 OTU03243 OTU00407
ANOVA	114	OTU02473 OTU00890 OTU00732 OTU04851 OTU06032 OTU05559 OTU05606 OTU05747 OTU06530 OTU05960 OTU06680 OTU04983 OTU05581 OTU04980 OTU05164 OTU01579 OTU00159 OTU01865 OTU01151 OTU01421 OTU03135 OTU03227 OTU06995 OTU02766 OTU00416 OTU03537 OTU06733 OTU06961 OTU05429 OTU04287 OTU04345 OTU02055 OTU01184 OTU06814 OTU06929 OTU03885 OTU04146 OTU05619 OTU02561 OTU00889 OTU05806 OTU01653 OTU00406 OTU00129 OTU05836 OTU04874 OTU01639 OTU00118 OTU00310 OTU05686 OTU00573 OTU00268 OTU03202 OTU00251 OTU05437 OTU02555 OTU02054 OTU05835 OTU03866 OTU03137 OTU05735 OTU05621 OTU02021 OTU03294 OTU06118 OTU06286 OTU00556 OTU02010 OTU06016 OTU03834 OTU06665 OTU00157 OTU05604 OTU00951 OTU00106 OTU00240 OTU04894 OTU06106 OTU06817 OTU01154 OTU00124 OTU05207 OTU03618 OTU01728 OTU00306 OTU06774 OTU02114 OTU00070 OTU02429 OTU04160 OTU01884 OTU00818 OTU00781 OTU02127 OTU05874 OTU03418 OTU01615 OTU06217 OTU03543 OTU06020 OTU06215 OTU05729 OTU00358 OTU03822 OTU06341 OTU04636 OTU05350 OTU02463 OTU04502 OTU00192 OTU00690 OTU01049 OTU01239 OTU04168
SVM	88	OTU00687 OTU02630 OTU04161 OTU01110 OTU03415 OTU06893 OTU00457 OTU04049 OTU03573 OTU03341 OTU02869 OTU03780 OTU01357 OTU01190 OTU03097 OTU00869 OTU02809 OTU02988 OTU01178 OTU03405 OTU02355 OTU01399 OTU02196 OTU03664 OTU01540 OTU06221 OTU00881 OTU02518 OTU05082 OTU06099 OTU03644 OTU06166 OTU00896 OTU01444 OTU02334 OTU02536 OTU03118 OTU01031 OTU05198 OTU01559 OTU02647 OTU02199 OTU05078 OTU01227 OTU03583 OTU06516 OTU05204 OTU01977 OTU05703 OTU06225 OTU02867 OTU01435 OTU04856 OTU03317 OTU01114 OTU01748 OTU06283 OTU03204 OTU00080 OTU01809 OTU05145 OTU02926 OTU02495 OTU01939 OTU03303 OTU03250 OTU02624 OTU06825 OTU01339 OTU02775 OTU02557 OTU00942 OTU00738 OTU00461 OTU01373 OTU02944 OTU05297 OTU00223 OTU03377 OTU00731 OTU03658 OTU00636 OTU00697 OTU04877 OTU06164 OTU02511 OTU02528 OTU05240
Permanova	65	OTU01485 OTU01461 OTU06058 OTU00028 OTU01380 OTU00012 OTU03800 OTU06695 OTU00047 OTU04523 OTU00271 OTU03698 OTU06826 OTU04210 OTU03571 OTU04081 OTU04373 OTU03226 OTU00046 OTU04185 OTU04369 OTU01452 OTU03366 OTU06325 OTU00165 OTU00145 OTU05777 OTU00411 OTU05332 OTU06834 OTU06659 OTU06920 OTU01219 OTU03638 OTU00691 OTU00190 OTU02036 OTU03082 OTU00785 OTU00555 OTU06699 OTU03301 OTU00976 OTU00420 OTU03350 OTU00443 OTU00366 OTU02509 OTU03576 OTU00521 OTU06552 OTU04465 OTU03428 OTU01536 OTU03034 OTU01048 OTU04013 OTU02263 OTU06872 OTU06343 OTU05456 OTU02441 OTU01148 OTU00281 OTU06735
Chi2	123	OTU03336 OTU03058 OTU03754 OTU04450 OTU03144 OTU00924 OTU02818 OTU01030 OTU04563 OTU02623 OTU02871 OTU00447 OTU02439 OTU03346 OTU03592 OTU03641 OTU01011 OTU02558 OTU06566 OTU02790 OTU03031 OTU02785 OTU06741 OTU02566 OTU00440 OTU03315 OTU01943 OTU01226 OTU02819 OTU01779 OTU03997 OTU00862 OTU02287 OTU01825 OTU03520 OTU04973 OTU03773 OTU04322 OTU01163 OTU03070 OTU02336 OTU02851 OTU02872 OTU01851 OTU01102 OTU03222 OTU02070 OTU01349 OTU00500 OTU01871 OTU04490 OTU00754 OTU02667 OTU00963 OTU02722 OTU03045 OTU03054 OTU05468 OTU06593 OTU05503 OTU05754 OTU02448 OTU02962 OTU03734 OTU02756 OTU03343 OTU01187 OTU00244 OTU00135 OTU00937 OTU03501 OTU02826 OTU03710 OTU04477 OTU02513 OTU03053 OTU00661 OTU01360 OTU03076 OTU03443 OTU01744 OTU01203 OTU01694 OTU06438 OTU00968 OTU00767 OTU02321 OTU02726 OTU06358 OTU03071 OTU05288 OTU00638 OTU00004 OTU05860 OTU06398 OTU03692 OTU03411 OTU00884 OTU03142 OTU02669 OTU01222 OTU00525 OTU01092 OTU01208 OTU02237 OTU02486 OTU02333 OTU03242 OTU02195 OTU02688 OTU01073 OTU04238 OTU02643 OTU01941 OTU02187 OTU03028 OTU02563 OTU04964 OTU01591 OTU00524 OTU03327 OTU00482 OTU03191

For 200 OTUs Venn diagram, there is a new discovery that PERMANOVA and ANOVA act very similarly to each other as they have 85 OTUs which are common in their set of best 200 OTUs. Hence, this confirms our hypothesis that the similarity between feature selection methods increases with an increasing number of features to produce. On the other hand, ANOVA and chi-square methods are completely different to each other as they only have 17 OTUs which are common between them.

There were 4 OTUs that were thoroughly tested in the lab and were concluded as the best features. Below is the table which includes the OTUs tested in the lab along with the ones that are found in the resulting OTUs of each feature selection method.

Table 4. Biological Analysis

Biologically Tested OTUs	OTU00845, OTU03024, OTU05622, OTU01719
Matched OTUs	
ANOVA	**0/4**
Chi-Square	**3/4** (OTU03024, OTU05622, OTU01719)
PERMANOVA	**3/4** (OTU00845, OTU03024, OTU05622)
SVM	**4/4** (OTU01719, OTU00845, OTU03024, OTU05622)

Conclusion

ANOVA, chi-square, PERMANOVA and SVM feature selection methods were applied to a microbiome dataset to compare the performance and results of the methods against each other. After implementing these methods by using the microbiome dataset in Python programming language, it yielded 12 common OTUs between all 4 feature selection methods among 200 OTUs results. Out of 5477 OTUs, there were only 12 common OTUs between the methods so we can say that these feature selection methods have different effects on this particular dataset. According to the OTUs that were tested in the lab, Chi-Square, PERMANOVA and SVM feature selection methods overperformed ANOVA. Therefore, selecting an accurate feature selection method is very important and requires lots of considerations.

Future Plan

As the PERMANOVA feature selection method is mainly used for microbiome datasets to choose the best features, similar feature selection methods like MANOVA, ANCOVA and MANCOVA could be tried to see if these feature selection methods behave in a similar way. On the contrary, the same dataset could be used to implement the embedded feature selection methods like Lasso, Ridge and Decision Tree to compare and analyze the performance.

Acknowledgement

We would like to express our special thanks of gratitude to our client Nisha Puthiyedth as well as our instructor Kevin O'Neil who gave us the golden opportunity to do this wonderful project on the topic Comparison and Analysis of Feature Selection Methods, which has helped us in doing lots of research and made us know about so many new things. We are utterly grateful to them.

References

[1] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995. [Online]. Available: http://homepages.rpi.edu/~bennek/class/mmld/papers/svn.pdf.

[2] E. Byvatov, U. Fechner, J. Sadowski, and G. Schneider, "Comparison of support Vector Machine and artificial neural network systems for drug/nondrug classification", Journal of Chemical Information and Computer Sciences, vol. 43, no. 6, pp. 1882-1889, Jun. 2003. [Online]. Available: https://pubs.acs.org/doi/10.1021/ci0341161.

[3] M. Bhasin and G. P. Raghava, "SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence," Bioinformatics, vol. 20, no. 3, pp. 421-423, Jan. 2004. [Online]. Available: https://academic.oup.com/bioinformatics/article/20/3/421/186291.

[4] N. Guenther and M. Schonlau, "Support Vector Machines", The Stata Journal: Promoting communications on statistics and Stata, vol. 16, no. 4, pp. 917-937, 2016. [Online]. Available: https://journals.sagepub.com/doi/10.1177/1536867X1601600407.

[5] T. Evgeniou and M. Pontil, "Support Vector Machines: Theory and applications," Machine Learning and Its Applications, pp. 249-257, 2001. [Online]. Available: https://www.researchgate.net/publication/221621494_Support_Vector_Machines_Theory_and_Applications.

[6] A. Gelman, "Analysis of Variance - Why it is More Important Than Ever," The Analysis of Statistics, vol. 33, no. 1, pp. 1-53, Feb. 2005. [Online]. Available: https://projecteuclid.org/journals/annals-of-statistics/volume-33/issue-1/Analysis-of-variancewhy-it-is-more-important-than-ever/10.1214/009053604000001048.full. [Accessed Sept. 20, 2021].

[7] M. Kumar, N. K. Rath, A. Swain and S. K. Rath, "Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor," Procedia Computer Science, vol 54, pp. 301-310, August 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050915013599. [Accessed Sept. 25, 2021].

[8] R. A. Armstrong, S.V. Slade, and F. Eperjesi, "An Introduction to Analysis of Variance (ANOVA) with Special Reference to Data from Clinical Experi-ments in Optometry," Ophthalmic Physiol Opt., vol. 20, no. 3, pp. 235-24, May 2000. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/10897345/. [Accessed Sept. 25, 2021].

[9] A. Sharma, "BIO-STATISTICS: A BRIEF OVERVIEW," *Kamls.in*, 2021. [Online]. Available: http://kamls.in/wp-content/uploads/2016/11/Vol.20_No.2_2011-1-24-28.pdf. [Accessed Nov 28, 2021].

[10] S. Onchiri, "Conceptual model on application of chi-square test in education and social sciences," *Academicjournals.org*, 2021. [Online]. Available: https://academicjournals.org/journal/ERR/article-full-text-pdf/3912FC76609. [Accessed Nov 28, 2021].

[11] D. Sharpe, "Your Chi-Square Test is Statistically Significant: Now What?," 2021 [Online]. Available: https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1269&context=pare. [Accessed Nov 28, 2021].

[12] B. J. Kelly, R. Gross, K. Bittinger, S. Sherrill-Mix, J.D. Lewis, R.G. Collman, F.D. Bushman, and H.

Li, “Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA,” *Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA*, vol. 31, no. 15, pp. 2461-2468, Mar. 2015. [Online]. Available: https://academic.oup.com/bioinformatics/article/31/15/2461/188732. [Accessed Nov 28, 2021]

[13] M. J Anderson, “Permutational multivariate analysis of variance (PERMANOVA),” *Wiley StatsRef: Statistics Reference Online*, pp. 1–15, Nov. 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/9781118445112.stat07841. [Accessed Nov 28, 2021]

[14] Wikipedia, “Permutational analysis of variance,” *Wikipedia*, 24-Nov-2021. [Online]. Available: https://en.wikipedia.org/wiki/Permutational_analysis_of_variance . [Accessed: 04-Dec-2021].

[15] R. B. Hayes, J. Ahn, X. Fan, B. A. Peters, Y. Ma, L. Yang, I. Agalliu, R. D. Burk, I. Ganly, M. P. Purdue, N. D. Freedman, S. M. Gapstur, and Z. Pei, “Association of oral microbiome with risk for incident head and neck squamous cell cancer,” *JAMA Oncology*, vol. 4, no. 3, pp. 358–365, Jan. 2018 [Online]. Available: https://jamanetwork.com/journals/jamaoncology/fullarticle/2668530. [Accessed Sep 27, 2021]

[16] Scikit-bio development team, scikit-bio docs 0.2.3, “skbio.stats.distance.PERMANOVA,” 2014. Available: http://scikit-bio.org/docs/0.2.3/generated/generated/skbio.stats.distance.PERMANOVA.html. [Accessed Dec 04, 2021]

[17] U. M. Khaire and R. Dhanalakshmi, “Stability of feature selection algorithm: A review.” June 2019 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157819304379. [Accessed Dec 04, 2021]

[18] “Calculate and draw Custom Venn Diagrams,” Bioinformatics and Evolutionary Genomics [Online]. Available: https://bioinformatics.psb.ugent.be/webtools/Venn/. [Accessed Nov 10, 2021]

[19] T. Z. Phyu and N. N. Oo, “Performance comparison of feature selection methods,” *MATEC Web of Conferences*, vol. 42, p. 06002, 2016 [Online]. Available: https://www.researchgate.net/publication/295263934_Performance_Comparison_of_Feature_Selection_Methods. [Accessed Dec 04, 2021]

[20] “Scatter Plot Maker,” scatterplot.online [Online]. Available: https://scatterplot.online/. [Accessed Oct 20, 2021]

[21] M. Waskom, “seaborn.clustermap,” Seaborn 0.11.2 [Online]. Available: https://seaborn.pydata.org/generated/seaborn.clustermap.html. [Accessed Nov 20, 2021]