THE CLASSIFICATION OF VIRAL HOSTS BASED ON METAGENOMIC FEATURES.

As a part of the subject

22BIO211 Intelligence of Biological Systems 2

Submitted by Group B-9

SURIYA K P	**(CB.EN.U4AIE22164)**
IPPATAPU VENKATA SRICHANDRA	**(CB.EN.U4AIE22165)**
DURAI SINGH K	**(CB.EN.U4AIE22167)**

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
CSE(AI)

AMRITA
श्रद्धावान् लभते ज्ञानम् VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY UNDER SECTION 3 OF THE UGC ACT, 1956

Centre for Computational Engineering and Networking

AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE – 641 112 (INDIA)

JUNE – 2024

AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112

BONAFIDE CERTIFICATE

This is to certify that the thesis entitled **“The classification of viral hosts based on metagenomic features**” submitted by SURIYA K P (CB.EN.U4AIE22164), IPPATAPU VENKATA SRICHANDRA (CB.EN.U4AIE22165), DURAI SINGH K (CB.EN.U4AIE22167) for the award of the Degree of Bachelor of Technology in the “CSE(AI) ” is a Bonafede record of the work carried out by her under our guidance and supervision at Amrita School of Artificial Intelligence, Coimbatore.

Project Guide:

Dr. Harishchander A.

Submitted for the university examination held on 13/06/2024

AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112

DECLARATION

We, SURIYA K P (CB.EN.U4AIE22164), IPPATAPU VENKATA SRICHANDRA (CB.EN.U4AIE22165), DURAI SINGH K (CB.EN.U4AIE22167) hereby declare that this is entitled "**The classification of viral hosts based on metagenomic features**", is the record of the original work done by us under the guidance of **Dr. Harishchander A**, Assistant Professor, Centre for Computational Engineering and Networking, Amrita School of Artificial Intelligence, Coimbatore. To, the best of our knowledge this work has not formed the basis for the award of any degree/diploma/ associate ship/fellowship/or a similar award to any candidate in any University.

Name	**Roll Number**	**Signature**
SURIYA K P	(CB.EN.U4AIE22164)	
IPPATAPU VENKATA SRICHANDRA	(CB.EN.U4AIE22165)	
DURAI SINGH K	(CB.EN.U4AIE22167)	

Place: Coimbatore

Date: 13-06-2024

Acknowledgement

We would like to express our special thanks of gratitude to our faculty **Dr. Harishchander A,**who gave us the golden opportunity to do this wonderful project, which also helped us in doing a lot of Research and we came to know about so many new things. We are thankful for the opportunity given.We would also like to thank our group members, as without their cooperation, we would not have been able to complete the project within the prescribed time.

Content

Figures

1 Abstract

This project is the classification of viral hosts based on metagenomic features. It is a critical task in virology, offering valuable insights into virus-host interactions and enhancing our understanding of viral ecology. This project aims to develop a robust machine learning model to classify viral hosts using a dataset of viral genomes characterized by features such as genome size, GC content, and the number of coding sequences (CDS). In this project, three machine learning algorithms, Decision Tree Classifier, Support Vector Machine (SVM), and Random Forest Classifier—were selected for model training and evaluation. The models were trained on a subset of the data, with performance evaluated using metrics such as accuracy, confusion matrix, and classification report. The outcomes of this research contribute to the field of viral genomics by providing a framework for the automated classification of viral hosts, aiding in the rapid identification and characterization of viral species.

2 Introduction

Understanding virus-host interactions is a fundamental aspect of virology that provides critical insights into viral ecology, pathogenesis, and evolution. The ability to accurately predict viral hosts based on metagenomic features has significant implications for public health, agriculture, and environmental management. With the advent of high-throughput sequencing technologies, large-scale metagenomic datasets have become available, enabling the development of machine learning models to classify viral hosts with high accuracy.

This project aims to classify viral hosts using a comprehensive dataset of viral genomes. By leveraging features such as genome size, GC content, and the number of coding sequences (CDS), we aim to develop robust machine learning models capable of accurately predicting the host organisms of various viruses. The project employs three different machine learning algorithms: Decision Tree Classifier, Support Vector Machine (SVM) with linear and radial basis function (rbf) kernels, and Random Forest Classifier. These models are trained and evaluated on a subset of the dataset, with their performance assessed using metrics such as accuracy, confusion matrix, and classification report.

Dataset Description

The dataset used in this project is a rich resource for studying viral genomics and host interactions, consisting of 7,362 entries that detail various aspects of viral genomes and taxonomy. It includes information on 4,893 unique virus names classified into 17 distinct taxonomic groups, with each entry marked as "Complete" and containing a unique assembly accession number. The dataset encompasses a variety of genomic features such as genome size and GC content, which vary among the viruses. Additionally, it provides information about the host organisms, which include bacteria, fungi, plants, and vertebrates, illustrating the diverse range of virus-host interactions. The number of coding sequences (CDS) per virus is listed, offering insights into the genetic complexity of the viruses. Moreover, FTP links to GenBank and RefSeq entries are included, granting access to detailed genomic data.

3 Methodology

3.1 Loading Dataset

The first step in our analysis involved loading the dataset containing information about various viruses and their corresponding hosts. This dataset was stored in a CSV file named viruses.csv. To read and manipulate this data efficiently, we used the pandas library in Python. The data was loaded into a pandas DataFrame, which provides a powerful and flexible structure for data analysis.

3.2 Pre-Processing

a. Assigning Column Names:

To ensure clarity and usability, we assigned appropriate column names to each column in the dataset. The essential columns retained for this study were host, size_mb, gc_percent, and cds. These columns represent the following biological sequence-related terms:

- **host**: The organism that the virus infects.
- **size_mb**: The size of the viral genome in megabases.
- **gc_percent**: The GC content percentage, indicating the proportion of guanine (G) and cytosine (C) bases in the genome.
- **cds**: The number of coding sequences in the viral genome.

Unnecessary columns were dropped from the dataset to focus our analysis on the relevant features mentioned above.

b. Label Encoding:

The categorical target column host was labeled numerically for computational efficiency. The mapping used was as follows:

Here is the mapping of host categories to numerical labels:

- Bacteria: 0
- Fungi: 1
- Plants: 2
- Vertebrates: 3
- Invertebrates: 4
- Protozoa: 5
- Vertebrates, invertebrates, human: 6

- Invertebrates, plants: 7
- Algae: 8
- Vertebrates, invertebrates: 9
- Vertebrates, human: 10
- Archaea: 11
- Human: 10

c. Handling Missing Values:

Missing values in the target column were handled by removing rows with NaN values. For the continuous feature variables, we used the SimpleImputer from sklearn to impute missing values with the mean of the respective columns.

d. Scaling Continuous Features:

To ensure that the continuous features are on a similar scale, we applied Min-Max Scaling. The formula for Min-Max Scaling is:

The formula for Min-Max Scaling is as follows:

The scaled value (X') is calculated by subtracting the minimum value of the feature (X_min) from the original value (X), and then dividing by the difference between the maximum value (X_max) and the minimum value (X_min).

In this context:

- X' is the scaled value.
- X is the original value.
- X_min is the minimum value of the feature.
- X_max is the maximum value of the feature.

e. Splitting Dataset:

We computed the counts of samples under each host category to understand the distribution of the data.The dataset was split into training and testing sets with an 80%-20% split using sklearn’s train_test_split function.

3.3 Data Visualization:

a. Distribution of Viral Hosts:

We plotted a bar graph to visualize the number of data points for each unique target class (host).

b. Histogram of Genome Size:

A histogram with 100 bins was plotted to show the distribution of the genome size in megabytes.

c. Distribution of GC%:

A histogram with 50 bins was plotted to show the distribution of the GC content percentage in the viral genomes.

d. Distribution of CDS:

A histogram with 50 bins was plotted to show the distribution of the number of coding sequences (CDS) in the viral genomes.

3.4 Model Creation:

We trained five different models for the classification task. Here is a brief explanation of each model and the theory behind them:

Model 1: Support Vector Classifier (SVC) with Linear Kernel

Support Vector Classifier (SVC) is a supervised machine learning algorithm used for classification tasks. The linear kernel is used when the data is linearly separable. The decision function for a linear SVM is defined as:

Model 2: Support Vector Classifier with RBF Kernel

The Radial Basis Function (RBF) kernel is used when the data is not linearly separable. The RBF kernel function is defined as:

Model 3: Decision Tree Classifier

Decision Tree Classifier is a non-parametric supervised learning method used for classification. It creates a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Model 4: Random Forest Classifier

Random Forest Classifier is an ensemble method that operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) of the individual trees. The formula for the prediction in Random Forest is:

Model 5: Sequential MLP (Multi-Layer Perceptron)

The MLP is a class of feedforward artificial neural network (ANN). The model used in this study had four hidden layers with ReLU activation function. The backpropagation algorithm was used to train the network for 500 epochs with a batch size of 118.

3.5 Proposed Workflow

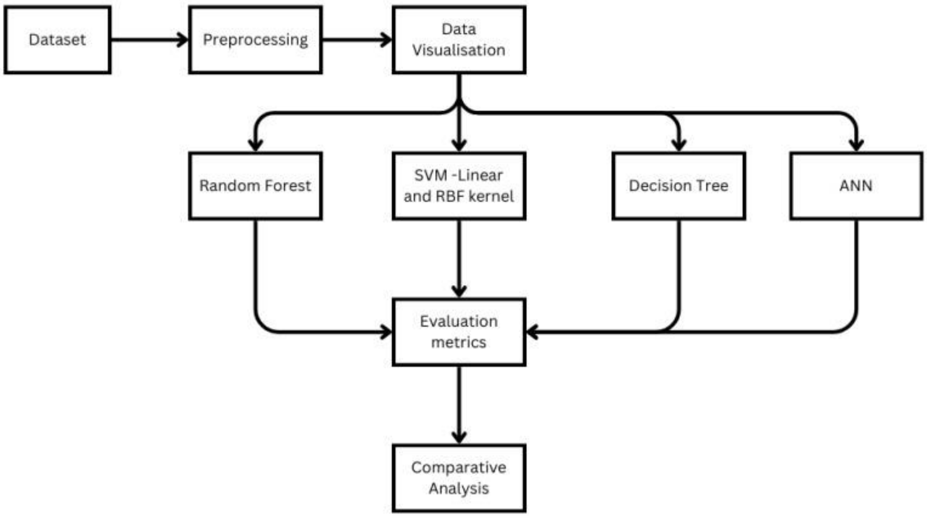


Figure 1 Workflow Methodology

In the proposed methodology the following steps involved:

In the data preprocessing step the standardizing column names and performing basic exploratory data analysis (EDA) has been done to understand the distribution and uniqueness of viruses and hosts. Categorical variables such as "organism_groups" and "host" are one-hot encoded to prepare them for machine learning algorithms, while numerical features like genome size, GC content, and CDS count are used directly.

The dataset is then split into training and testing sets using a 80-20 split ratio. Three machine learning models are trained and evaluated: a Support Vector Machine (SVM) with a radial basis function (RBF) kernel and Linear Kernel, Random Forest classifier with 50 estimators and a maximum depth of 25, a Decision Tree classifier, and Sequential MLP was being trained. Each model's performance is being evaluated using accuracy scores, confusion matrices, and classification reports and the model performance is visualized through heatmaps of confusion matrices, and the accuracies of the three models are compared to identify the best-performing one

.

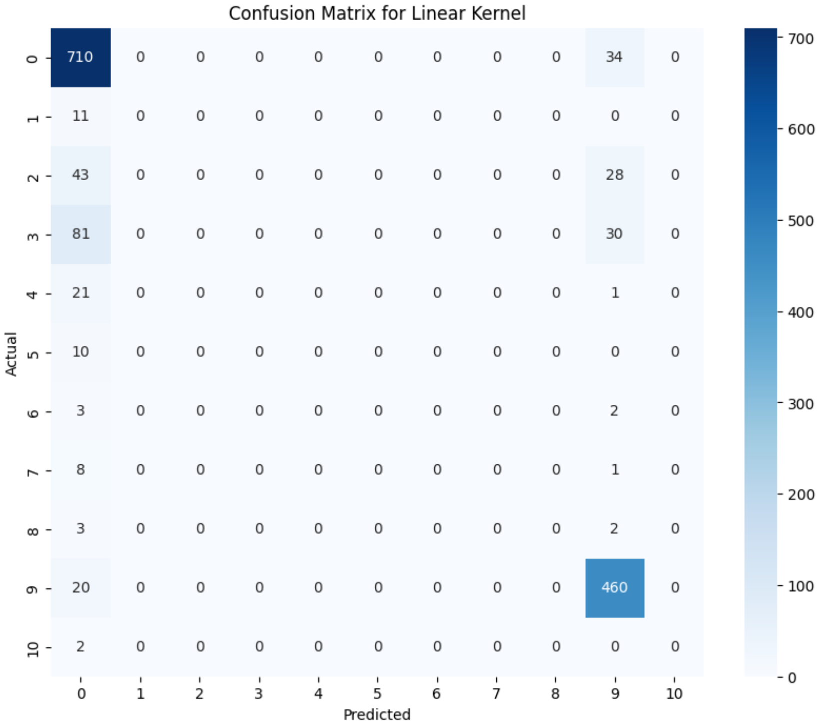
3.6 Evaluation Metrics Used

To evaluate the performance of the models, the following metrics were used:

- **Test Accuracy**: The overall accuracy of the model on the test set.
- **Multiclass Confusion Matrix**: A matrix that shows the true versus predicted classifications.
- **Precision**: The ratio of correctly predicted positive observations to the total predicted positives.
- **Recall**: The ratio of correctly predicted positive observations to the all observations in actual class.
- **F1 Score**: The weighted average of Precision and Recall.
- **Macro Averaged Metrics**: Average of the metric calculated for each class.
- **Weighted Average Metrics**: Average of the metric calculated for each class, weighted by the number of true instances for each class.

Each model's performance was assessed using these metrics to determine the most effective classifier for predicting the viral hosts based on metagenomic features.

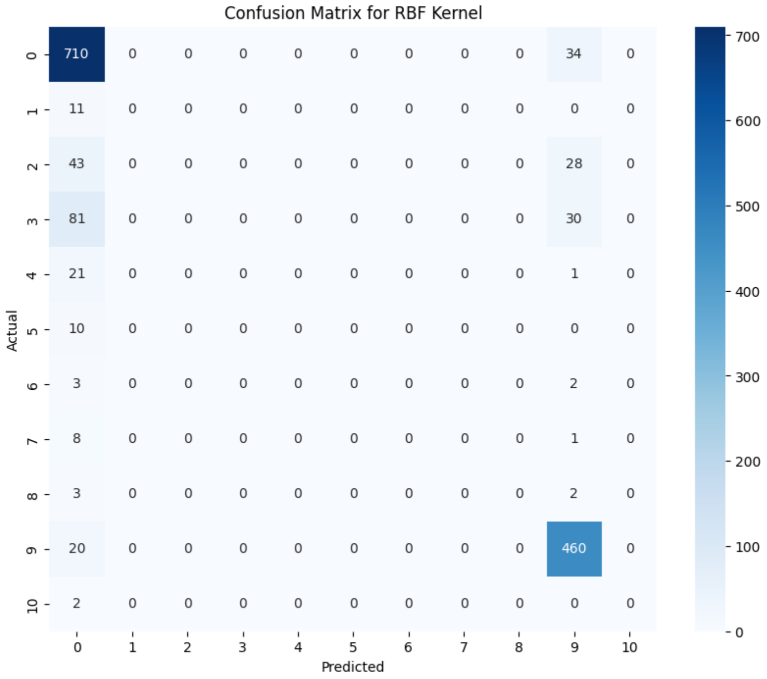
4 Results



Confusion Matrix for Linear Kernel

Actual \ Predicted	0	1	2	3	4	5	6	7	8	9	10
0	710	0	0	0	0	0	0	0	0	34	0
1	11	0	0	0	0	0	0	0	0	0	0
2	43	0	0	0	0	0	0	0	0	28	0
3	81	0	0	0	0	0	0	0	0	30	0
4	21	0	0	0	0	0	0	0	0	1	0
5	10	0	0	0	0	0	0	0	0	0	0
6	3	0	0	0	0	0	0	0	0	2	0
7	8	0	0	0	0	0	0	0	0	1	0
8	3	0	0	0	0	0	0	0	0	2	0
9	20	0	0	0	0	0	0	0	0	460	0
10	2	0	0	0	0	0	0	0	0	0	0

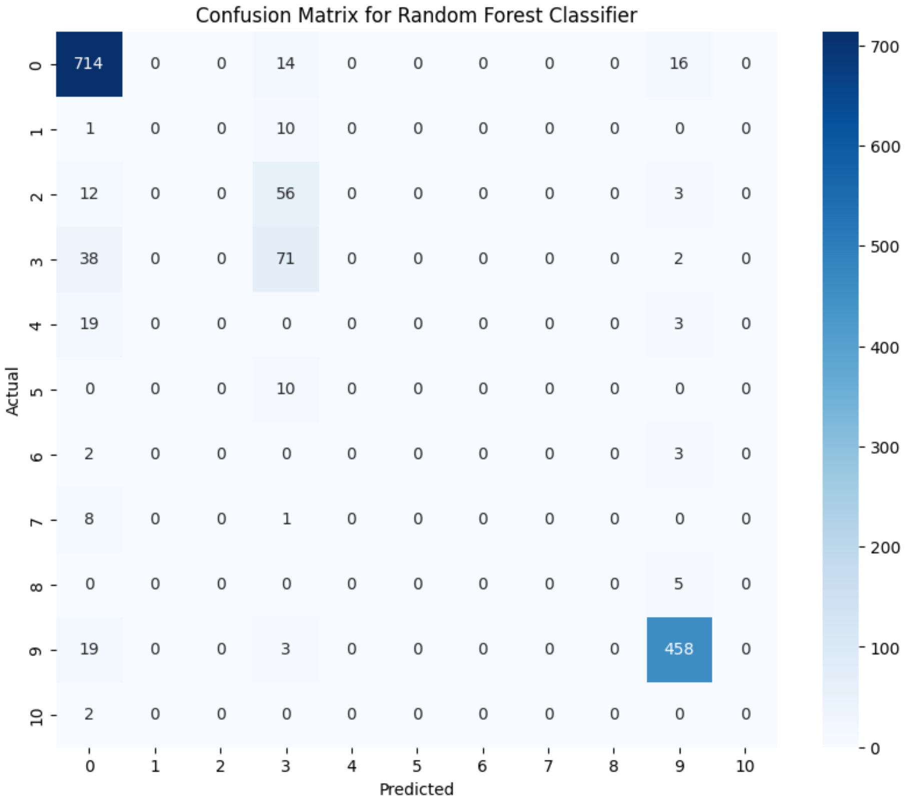
Figure 2 Confusion matrix of the svm linear kernel



Confusion Matrix for RBF Kernel

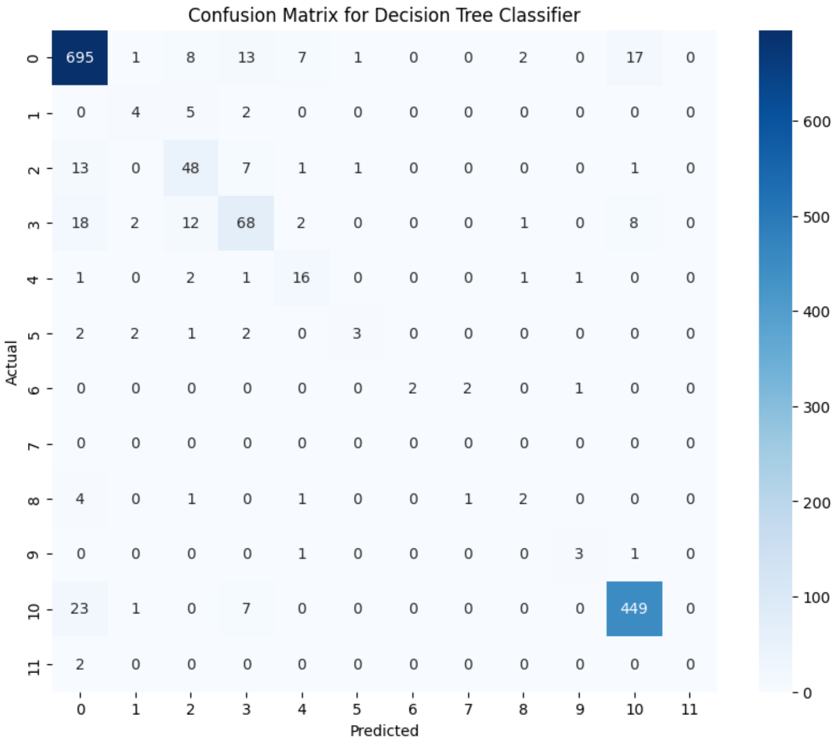
Actual \ Predicted	0	1	2	3	4	5	6	7	8	9	10
0	710	0	0	0	0	0	0	0	0	34	0
1	11	0	0	0	0	0	0	0	0	0	0
2	43	0	0	0	0	0	0	0	0	28	0
3	81	0	0	0	0	0	0	0	0	30	0
4	21	0	0	0	0	0	0	0	0	1	0
5	10	0	0	0	0	0	0	0	0	0	0
6	3	0	0	0	0	0	0	0	0	2	0
7	8	0	0	0	0	0	0	0	0	1	0
8	3	0	0	0	0	0	0	0	0	2	0
9	20	0	0	0	0	0	0	0	0	460	0
10	2	0	0	0	0	0	0	0	0	0	0

Figure 3 Confusion matrix of the svm rbf kernel



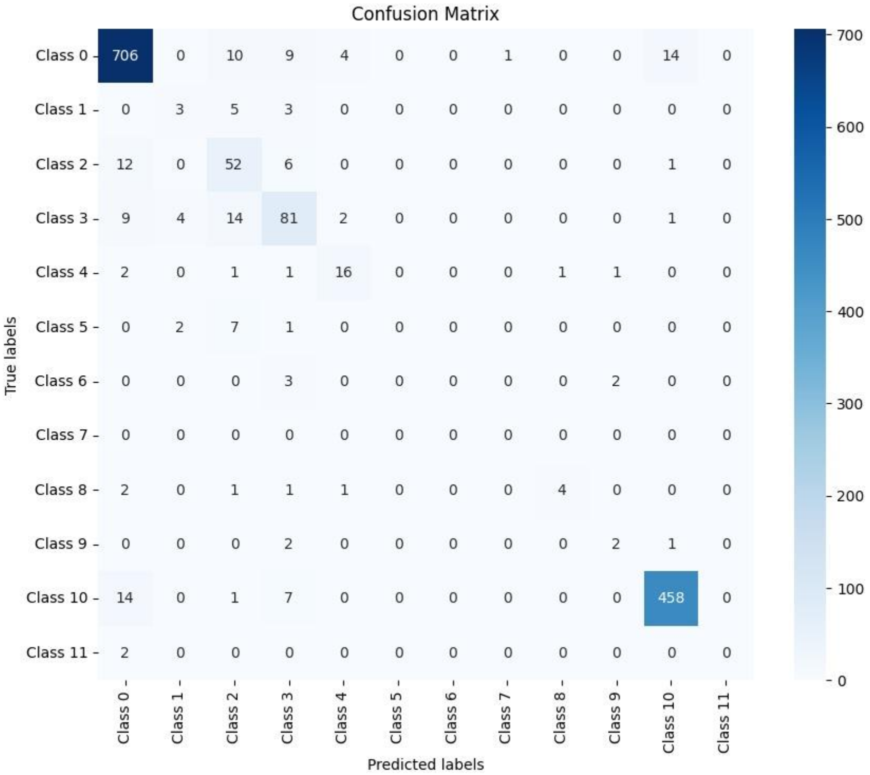
Actual \ Predicted	0	1	2	3	4	5	6	7	8	9	10
0	714	0	0	14	0	0	0	0	0	16	0
1	1	0	0	10	0	0	0	0	0	0	0
2	12	0	0	56	0	0	0	0	0	3	0
3	38	0	0	71	0	0	0	0	0	2	0
4	19	0	0	0	0	0	0	0	0	3	0
5	0	0	0	10	0	0	0	0	0	0	0
6	2	0	0	0	0	0	0	0	0	3	0
7	8	0	0	1	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	5	0
9	19	0	0	3	0	0	0	0	0	458	0
10	2	0	0	0	0	0	0	0	0	0	0

Figure 4 confusion matrix of the random forest classifier



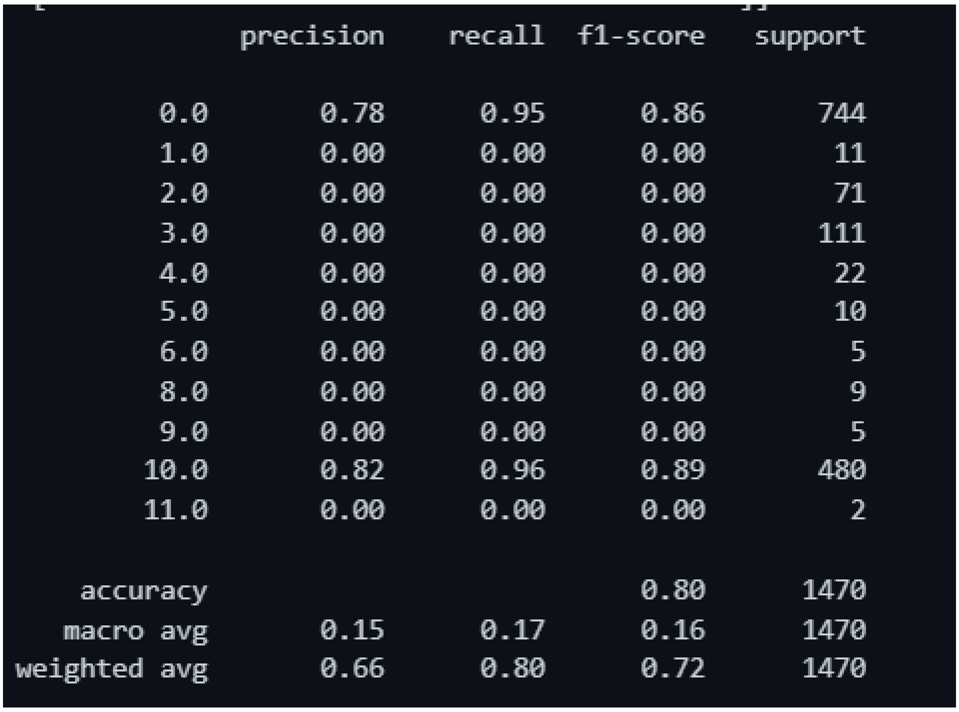
Actual \ Predicted	0	1	2	3	4	5	6	7	8	9	10	11
0	695	1	8	13	7	1	0	0	2	0	17	0
1	0	4	5	2	0	0	0	0	0	0	0	0
2	13	0	48	7	1	1	0	0	0	0	1	0
3	18	2	12	68	2	0	0	0	1	0	8	0
4	1	0	2	1	16	0	0	0	1	1	0	0
5	2	2	1	2	0	3	0	0	0	0	0	0
6	0	0	0	0	0	0	2	2	0	1	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0
8	4	0	1	0	1	0	0	1	2	0	0	0
9	0	0	0	0	1	0	0	0	0	3	1	0
10	23	1	0	7	0	0	0	0	0	0	449	0
11	2	0	0	0	0	0	0	0	0	0	0	0

Figure 5 confusion matrix of the Decision Trees



	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9	Class 10	Class 11
Class 0	706	0	10	9	4	0	0	1	0	0	14	0
Class 1	0	3	5	3	0	0	0	0	0	0	0	0
Class 2	12	0	52	6	0	0	0	0	0	0	1	0
Class 3	9	4	14	81	2	0	0	0	0	0	1	0
Class 4	2	0	1	1	16	0	0	0	1	1	0	0
Class 5	0	2	7	1	0	0	0	0	0	0	0	0
Class 6	0	0	0	3	0	0	0	0	0	2	0	0
Class 7	0	0	0	0	0	0	0	0	0	0	0	0
Class 8	2	0	1	1	1	0	0	0	4	0	0	0
Class 9	0	0	0	2	0	0	0	0	0	2	1	0
Class 10	14	0	1	7	0	0	0	0	0	0	458	0
Class 11	2	0	0	0	0	0	0	0	0	0	0	0

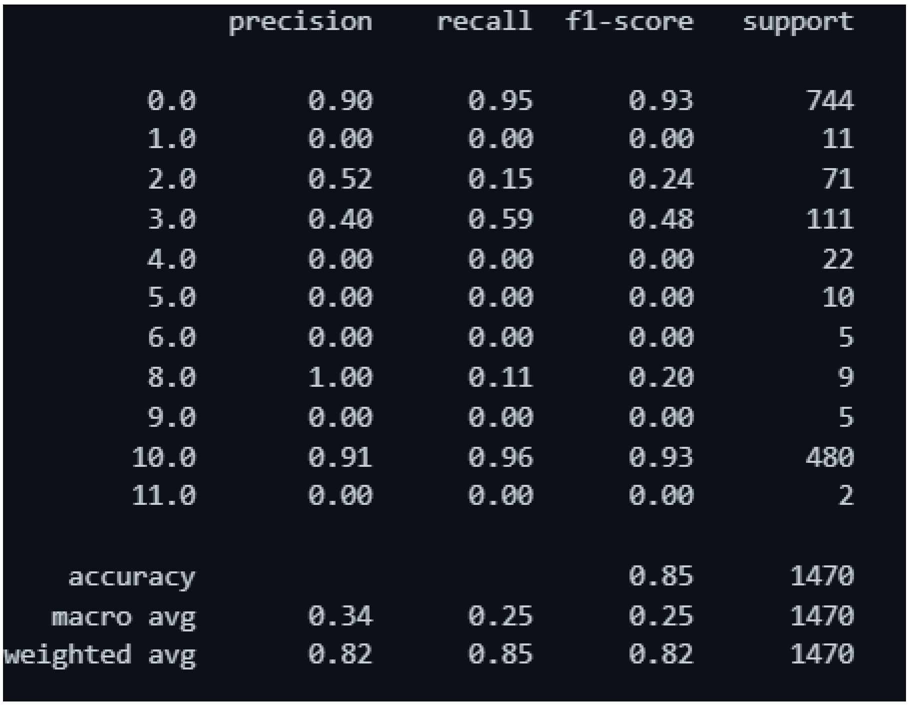
Figure 6 Figure 5 confusion matrix of the ANN


```
              precision    recall  f1-score   support

         0.0       0.78      0.95      0.86       744
         1.0       0.00      0.00      0.00        11
         2.0       0.00      0.00      0.00        71
         3.0       0.00      0.00      0.00       111
         4.0       0.00      0.00      0.00        22
         5.0       0.00      0.00      0.00        10
         6.0       0.00      0.00      0.00         5
         8.0       0.00      0.00      0.00         9
         9.0       0.00      0.00      0.00         5
        10.0       0.82      0.96      0.89       480
        11.0       0.00      0.00      0.00         2

    accuracy                           0.80      1470
   macro avg       0.15      0.17      0.16      1470
weighted avg       0.66      0.80      0.72      1470
```

Figure 7 Evaluation metrics of the svm linear kernel

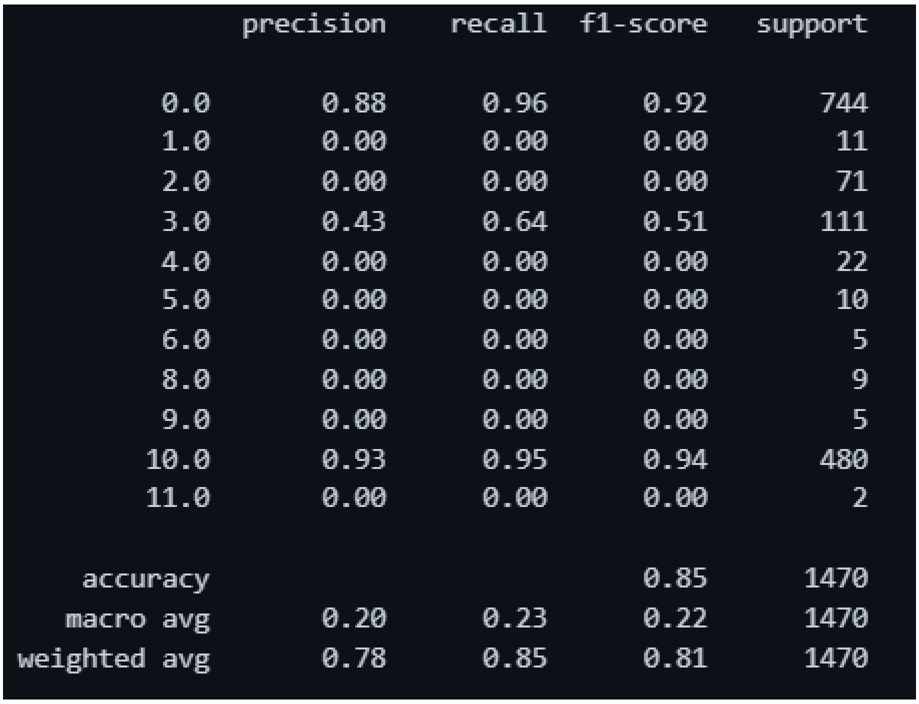


```
              precision    recall  f1-score   support

         0.0       0.90      0.95      0.93       744
         1.0       0.00      0.00      0.00        11
         2.0       0.52      0.15      0.24        71
         3.0       0.40      0.59      0.48       111
         4.0       0.00      0.00      0.00        22
         5.0       0.00      0.00      0.00        10
         6.0       0.00      0.00      0.00         5
         8.0       1.00      0.11      0.20         9
         9.0       0.00      0.00      0.00         5
        10.0       0.91      0.96      0.93       480
        11.0       0.00      0.00      0.00         2

    accuracy                           0.85      1470
   macro avg       0.34      0.25      0.25      1470
weighted avg       0.82      0.85      0.82      1470
```

Figure 8 Evaluation metrics of the svm rbf kernel

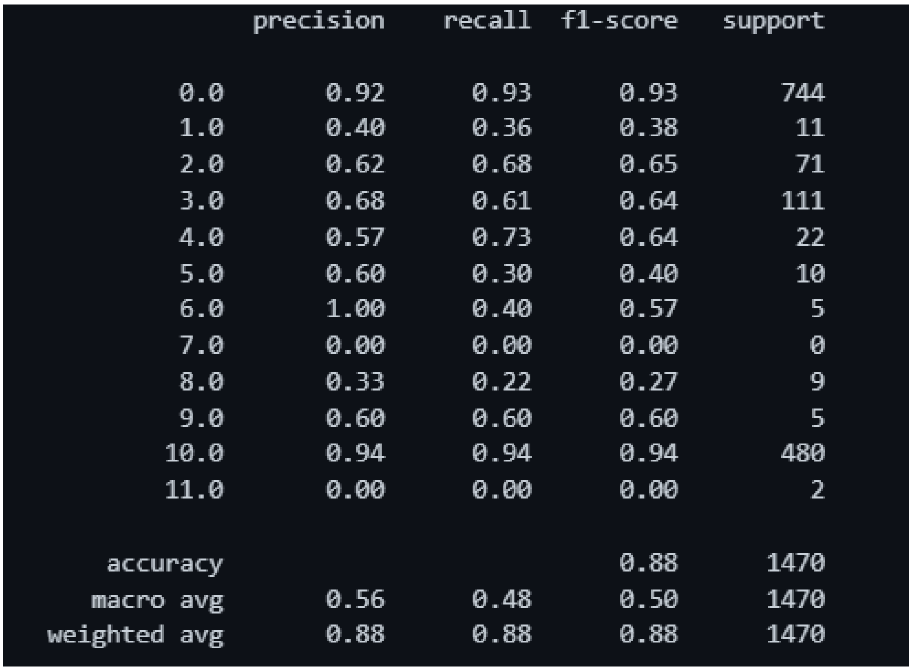


```
              precision    recall  f1-score   support

         0.0       0.88      0.96      0.92       744
         1.0       0.00      0.00      0.00        11
         2.0       0.00      0.00      0.00        71
         3.0       0.43      0.64      0.51       111
         4.0       0.00      0.00      0.00        22
         5.0       0.00      0.00      0.00        10
         6.0       0.00      0.00      0.00         5
         8.0       0.00      0.00      0.00         9
         9.0       0.00      0.00      0.00         5
        10.0       0.93      0.95      0.94       480
        11.0       0.00      0.00      0.00         2

    accuracy                           0.85      1470
   macro avg       0.20      0.23      0.22      1470
weighted avg       0.78      0.85      0.81      1470
```

Figure 9 Evaluation metrics of the random forest classifier

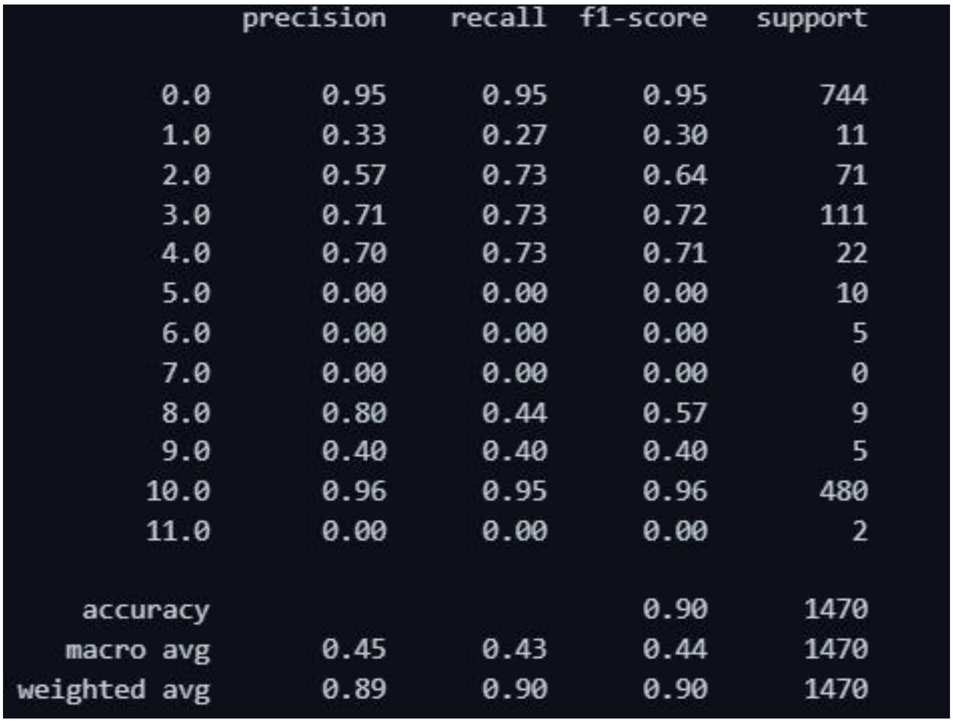


```
              precision    recall  f1-score   support

         0.0       0.92      0.93      0.93       744
         1.0       0.40      0.36      0.38        11
         2.0       0.62      0.68      0.65        71
         3.0       0.68      0.61      0.64       111
         4.0       0.57      0.73      0.64        22
         5.0       0.60      0.30      0.40        10
         6.0       1.00      0.40      0.57         5
         7.0       0.00      0.00      0.00         0
         8.0       0.33      0.22      0.27         9
         9.0       0.60      0.60      0.60         5
        10.0       0.94      0.94      0.94       480
        11.0       0.00      0.00      0.00         2

    accuracy                           0.88      1470
   macro avg       0.56      0.48      0.50      1470
weighted avg       0.88      0.88      0.88      1470
```

Figure 10 Evaluation metrics of the Decision Trees



```
              precision    recall  f1-score   support

         0.0       0.95      0.95      0.95       744
         1.0       0.33      0.27      0.30        11
         2.0       0.57      0.73      0.64        71
         3.0       0.71      0.73      0.72       111
         4.0       0.70      0.73      0.71        22
         5.0       0.00      0.00      0.00        10
         6.0       0.00      0.00      0.00         5
         7.0       0.00      0.00      0.00         0
         8.0       0.80      0.44      0.57         9
         9.0       0.40      0.40      0.40         5
        10.0       0.96      0.95      0.96       480
        11.0       0.00      0.00      0.00         2

    accuracy                           0.90      1470
   macro avg       0.45      0.43      0.44      1470
weighted avg       0.89      0.90      0.90      1470
```

Figure 11 Evaluation metrics of the Sequential MLP

5 Conclusion

From the results obtained from GridSearchCV, the Random Forest classifier emerged as the best-performing model, achieving an accuracy of 91.02% with the optimal parameters (max_depth=20, max_features='sqrt', min_samples_leaf=2, min_samples_split=4, n_estimators=70). This high performance suggests that the Random Forest model is well-suited for capturing the complexities in the data, making it the most robust and reliable option among the models tested. In contrast, the Decision Tree classifier, with an accuracy of 88.98%, also performed well but did not reach the same level of accuracy. Despite this, it offers a simpler and more interpretable model, which might be beneficial in scenarios where understanding the decision-making process is important.

The Sequential MLP achieved an accuracy of 90%. The Support Vector Classifier (SVC) with an RBF kernel achieved an accuracy of 85.51%, indicating it can effectively handle non-linear relationships in the data, although it did not perform as well as the Random Forest or Decision Tree. The SVC with a linear kernel had the lowest accuracy at 84.29%, suggesting it is less suitable for this particular dataset. Overall, the Random Forest classifier is recommended for its superior accuracy, while the Decision Tree offers a balance between performance and interpretability. The SVC models may be considered in specific cases where their unique strengths align with the data characteristics.

6 References

1. Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15:R46.
2. Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biol. 20:257. doi: 10.1186/s13059-019-1891-0.
3. Woods: a fast and accurate functional annotator and classifier of genomic and metagenomic sequences. Genomics 106, 1–6. doi: 10.1016/j.ygeno.2015.04.001
4. Steinwart, I., and Christmann, A. (2008). Support Vector Machines. Berlin: Springer Science & Business Media.
5. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733– D745. doi: 10.1093/nar/gkv1189
6. Menzel, P., Ng, K. L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat. Commun. 7:11257. doi: 10. 1038/ncomms11257
7. McIntyre, A. B. R., Ounit, R., Afshinnekoo, E., Prill, R. J., Hénaff, E., Alexander, N., et al. (2017). Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol. 18, 1–19.
8. Mathieu Alban , Leclercq Mickael , Sanabria Melissa , Perin Olivier , Droit Arnaud.TITLE=Machine Learning and Deep Learning Applications in Metagenomic Taxonomy and Functional Annotation JOURNAL Frontiers in Microbiology, VOLUME=13, YEAR=2022, URL=https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2022.811495, DOI=10.3389/fmicb.2022.811495, ISSN=1664-302X