IDENTIFICATION OF MICROBE-RESISTING PEPTIDES

USING MACHINE LEARNING

By Neil Bhosale

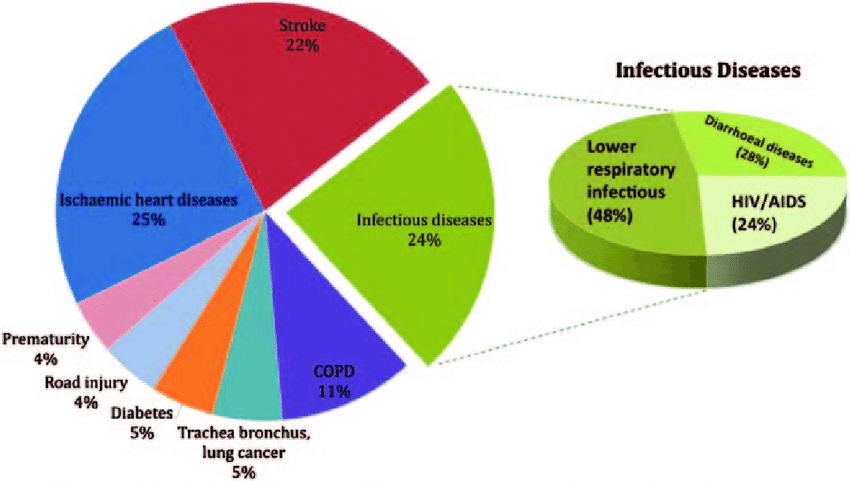
Lovely Professional University

ABSTRACT

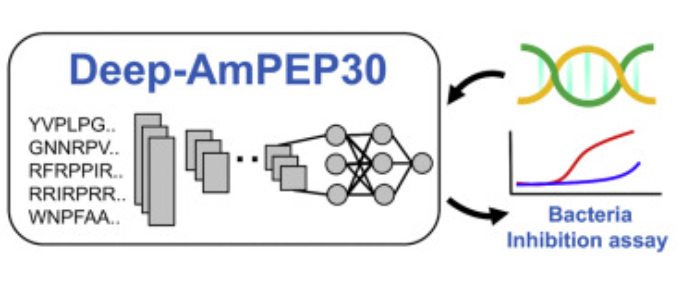
This paper presents research in the topic of finding antimicrobial peptides from a given set of peptide chains and which combination/composition

MOTIVATION

In **medical science**, there is a very urgent need to find new and better medicinal chemicals that will provide more effective cures. Every year, many new and mutated forms of pathogens (microbes) are found in sick and deceased individuals. Thus, antimicrobial resistance is an urgent and global health problem as existing drugs are becoming ineffective against the treatment of microbial infections.



GRAPHICAL ABSTRACT



Antimicrobial peptides (AMPs) are a valuable source of antimicrobial agents and a potential solution to the multi-drug resistance problem (in which, after continuous administration of a drug to a long-term patient, the patient becomes resistant to it and a new drug is required to treat the same illness and with the same effectiveness). In particular, short-length AMPs have been shown to have enhanced antimicrobial activities, higher stability, and lower toxicity to human cells. In this project I will be analyzing 2 types of short length AMPs, single amino acid chain (by composition) and dipeptide chain.

Objectives of the project:

• Create a model-feedable dataset by making functions to form csv data out of ascii/txt data.

• To build a machine learning model that, when given a new peptide chain, can predict whether it will be effective on killing harmful microbes or not (binary classification).

• To evaluate the model using various performance metrics.

• To try 3 different ML algorithms with various hyperparameters and choose the one with highest accuracy.

• To find short accuracy results on 30 ML models using LazyClassifier tool.

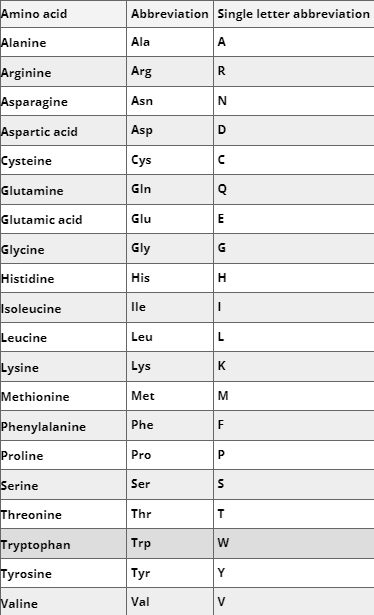


Figure: List of amino acids found in Nature (20)

THE MACHINE LEARNING PARADIGMS USED IN THE PROJECT

3.1 Proposed Algorithms:

The central problem of this project is to classify peptides into microbicidal and non-microbicidal ones. This is clearly a binary classification problem and can be solved using the following supervised learning algorithms (in that order, retrospectively): Support Vector Machine, K-Nearest-Neighbors, Decision tree, Random Forest Algorithm. Then, various model-evaluation metrics can be used to find the accuracy and scope for improvement in the model.

Support Vector Machine

Support Vector Machine or SVM is a supervised and linear Machine Learning algorithm most commonly used for solving classification problems and is also referred to as Support Vector Classification. The objective of SVM is to draw a line that best separates the two classes of data points. (This is very useful as the given problem is a binary classification one.)



If the training data is linearly separable, we can select two parallel hyperplanes that separate the two classes of data, so that the distance between them is as large as possible.



Anything on or above this boundary will have class label 1



Anything on or below this boundary will have class label 0 or -1

Nu SVC

Nu-SVC: Nu-Support Vector Classification

Nu-SVC is similar to Support Vector Classification/Machine but it uses a new parameter nu which controls the number of support vectors and training errors.

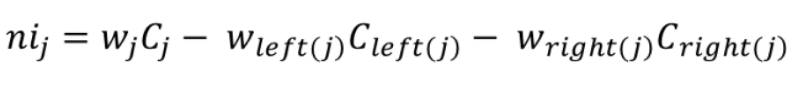
The parameter nu is an upper bound on the fraction of training errors and a lower bound of the fraction of support vectors.

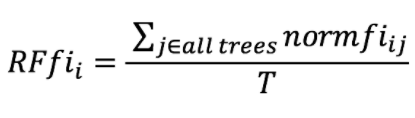
The value of nu should be in the interval (0,1] (0 non-inclusive and 1 inclusive)

Random Forest Algorithm

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

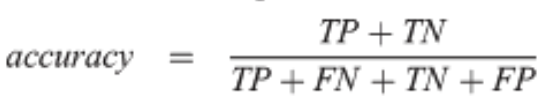
This makes it an ensemble learning technique for classification.



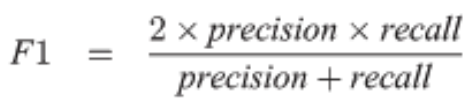


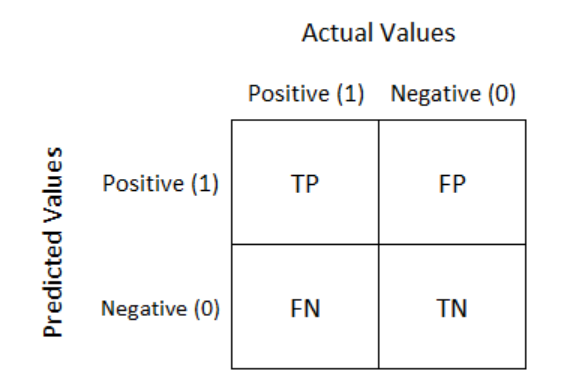
**Accuracy Score:**

This is a metric for direct ratio of right classifications to misclassifications:



**FI SCORE:**





**TOOLS AND SOFTWARE USED**

**4.1 Conda:**

Version: Miniconda3-py37\_4.8

The purpose of this technology is to be able to use the CD-HIT feature

**4.2 Python sys module:**

This helps to manipulate the runtime environment

**4.3 Pfeature Library**

Pfeature is an experimental library built in the systems biology research group of IIITD.

This is a powerful library that provides various functions to manipulate raw peptide chain data in such a way that all the possible combinations of peptide bonded chains can be formed.

These peptide chains can be for the entire raw chain or only using parts of it (The start of the chain or the end of it).

70000 such different permutations and peptide bonded combinations can be formed out of these pfeature functions from only a bare raw text file with amino acid chain Initials.

**4.4 CD-HIT**

Cd HIT is a conda functionality that

**4.5 Pandas**

All data manipulation rests on Pandas.\

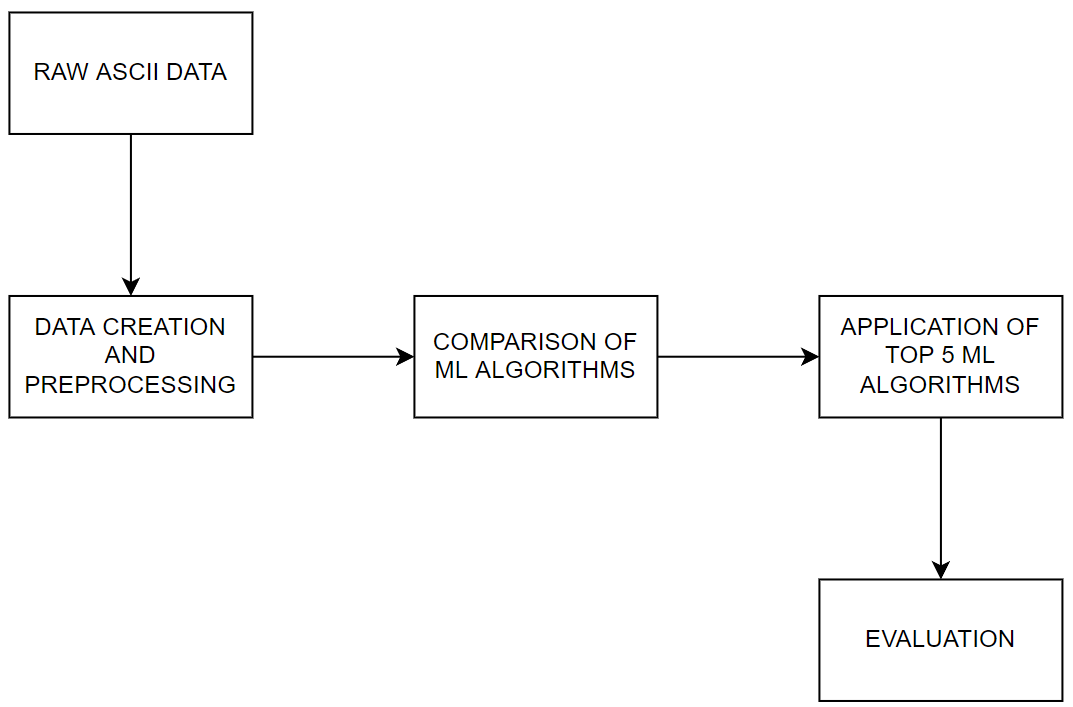
**4.6 Seaborn, Matplotlib**

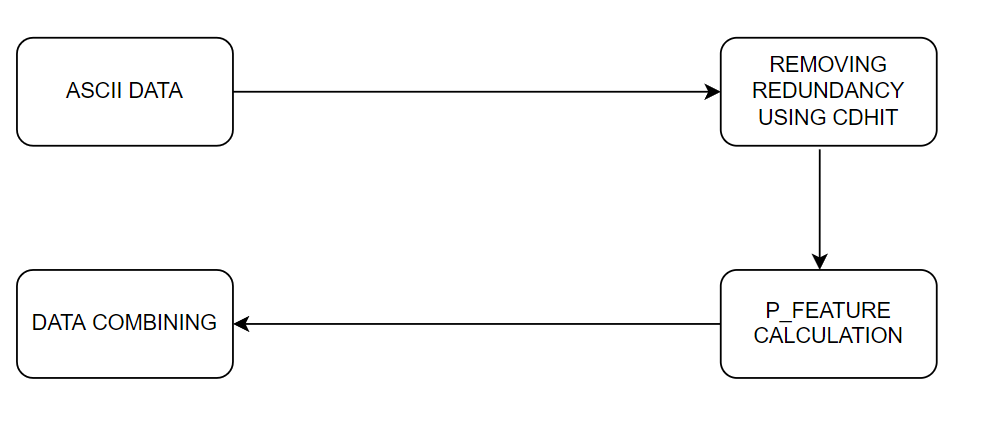
All data visualization has been done using seaborn and matplotlib for both feature selection phase and result visualization phase.

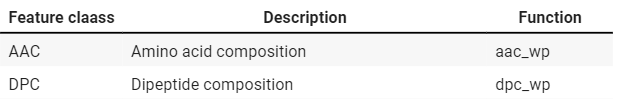
**4.8 LazyClassifier**

This is a powerful tool that aides in model selection. It has been used by me in this project for finding the top 5 performing algorithms both on the training and testing datasets. Based on that, I decided which classifiers to implement.

FLOW/STEPS OF THE PROJECT AND CIRCUIT DESCRIPTION







**DATA PREPROCESSING**

Dividing the data into Features (X) and Target (y)

Apart from the CDHIT dropping of peptide chains, the formed aac dataset was checked for variance and correlation:

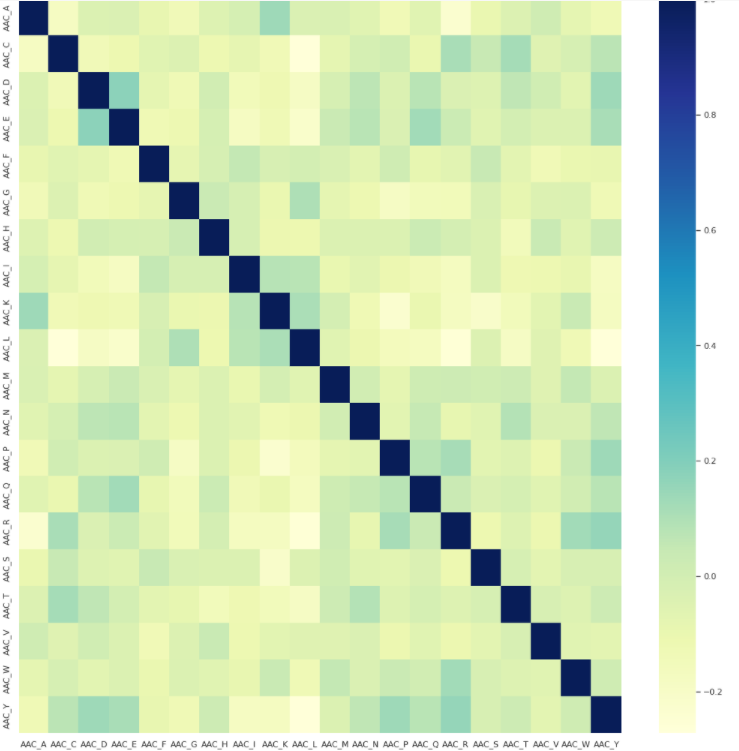
****

Figure: Correlation of all the amino-acids (features)

**COMPARISON OF ML ALGORITHMS:**

In this project I have used **lazypredict** and **LazyClassifier** to compare 27 ML algorithms:



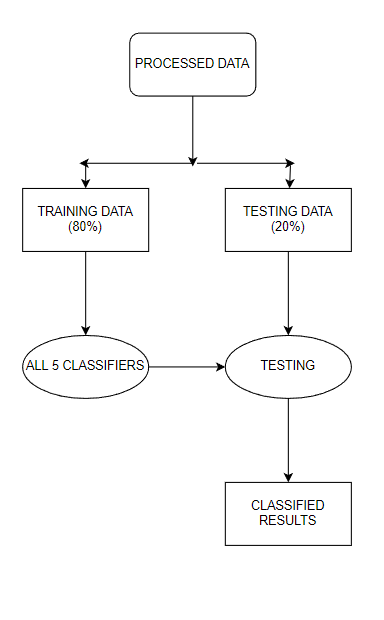
Figure: The performance of 27 different ML algorithms on the training dataset



Figure: Show the average performance of all 27 algorithms on the testing set.

BASED ON ABOVE RESULTS AND SIMPLICITY OF IMPLEMENTATION, THE FOLLOWING 5 ALGORITHMS ARE CHOSEN:

1. Support Vector Machine, 2. Nu-SVM, 3. K Nearest Neighbors, 4. Decision Tree, 5. Random Forest



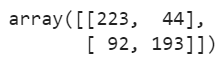
RESULTS AND HYPER PAREMETER TUNING

**6.1 Result of SVM:**

Accuracy on test data: 75.362 percent

F1 SCORE: 76.632 percent

Confusion Matrix

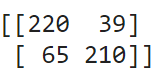


**6.2 Result of Nu-SVC:**

Accuracy on training data: 85 percent

Accuracy on testing data: 79.345 percent

Confusion Matrix



**6.3 Result of K-Nearest-Neighbor**

Accuracy Score: 74.567 percent

F1\_score: 72.958 percent

Confusion Matrix:



**6.4 Result of DECISION TREE LEARNING**

Accuracy Score: 69.993 percent

F1\_Score: 70.506 percent

Confusion Matrix:



**6.5 Result of RANDOM FOREST**

Preliminary result:

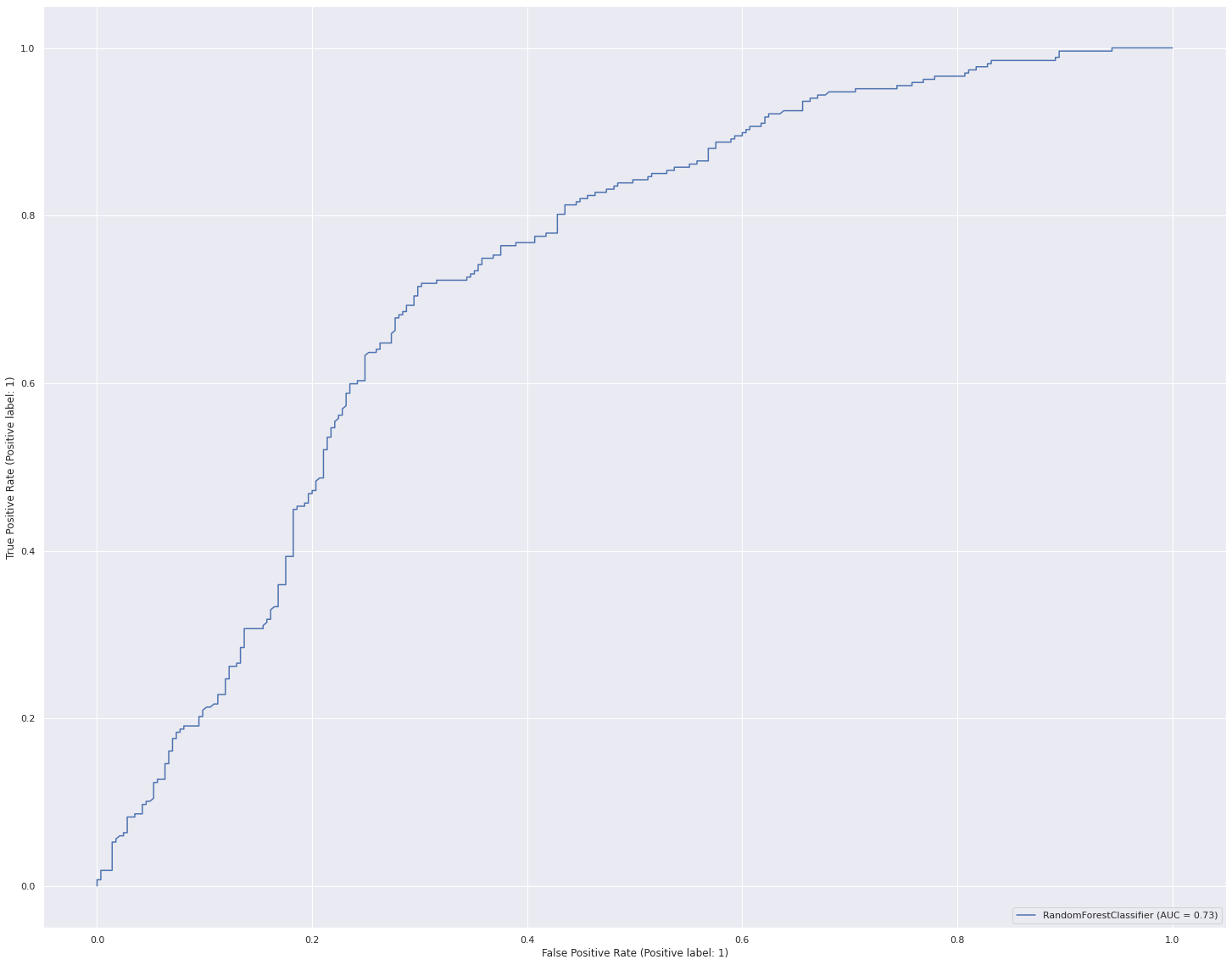
Accuracy Score: 73.913 percent

F1\_Score: 76.366 percent

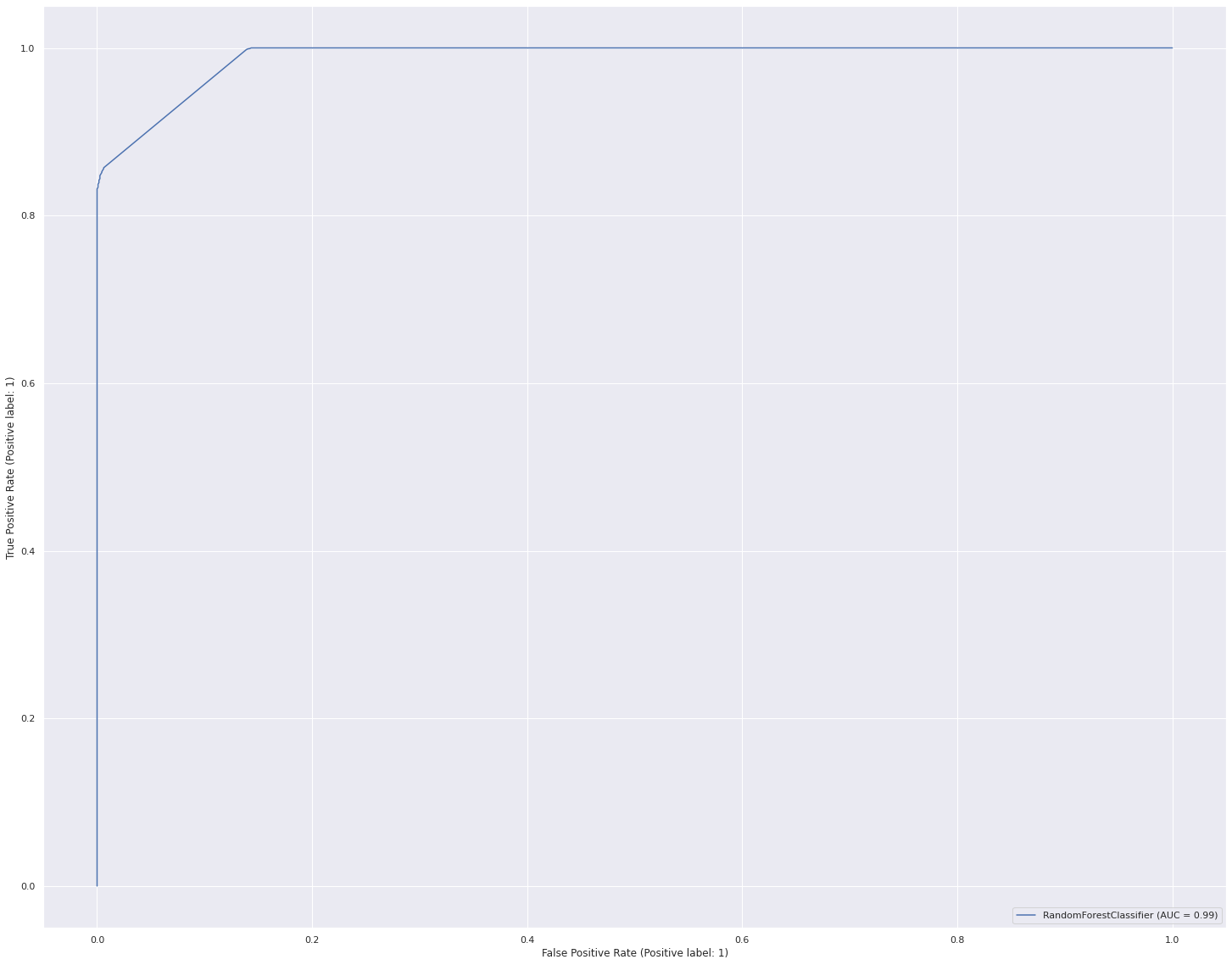
**Results after performing Hyper parameter Tuning Using Grid-Search CV:**

Accuracy Score: 81.733 percent

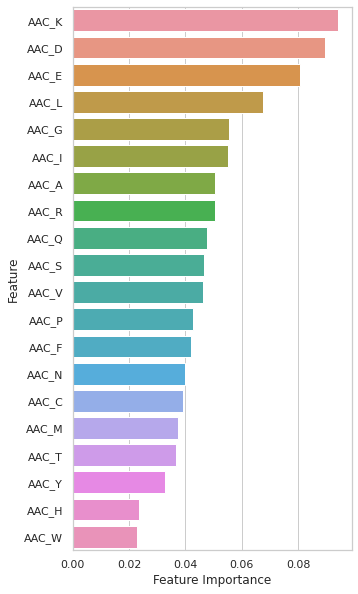
F1\_Score: 82 percent



ROC CURVE FOR Testing DATA



ROC CURVE FOR TRAINING DATA



This Bar Plot is instrumental in finding which Features have the highest GINI INDEX (plotted in descending order).

CONCLUSION AND FUTURE SCOPE

Conclusions made from the project

This project was instrumental in understanding how both the combinations and individual compositions of Amino acids in peptide chains affects the power of peptide chains in killing microbes or being completely inactive against them.

This makes it very clear that K, D and E Amino Acids are the most important Amino Acids when it comes to antimicrobial activities.

Y, H and W amino acids are least useful and their presence in the amino acid chain does not matter much.

7. Future Scope

Similar to the process performed for Amino acid composition, the same peptide chain ascii file text data can give rise to relational datasets of the following kind:

Given above are the multitudinous functionalities available in the pfeature library [Raghava Et Al]

These are all the peptide-bonded combinations that can be formed from the given chains. The functions stated next to each combination can be used to quantitatively state the composition of each combination formed in a given raw peptide chain and thus form a value for that column in the peptide row.

For example: In the 2nd combination stated in the above table: DPC (dipeptide composition):

• There are a maximum of 20 amino acids found in nature.

• Dipeptide bonds hold 2 of these together at a node (whether unique or repeated)

• So, the number of features in this new dataset will be equal to the total number of dipeptide combinations possible. i.e:

• 20\*20 = 400 Feature columns (for each peptide)

The same can be done for all other combinations as well.

The pfeature library is capable of producing 70,000 such compositional combinations, which makes it a very powerful tool in the realm of polypeptide-related systems biology applications and judging the various behaviors of peptides in the world.

This is not only limited to microbicidal properties by many other realms such as:

• Skincare products

• Hair growth products

• Antidepressants

• Anti-Indigestion (Non-microbial)

Any of the above research areas are open for the peptide behavioural researchers in the future.