A Comparative Analysis of Machine Learning Methods for Predicting the Presence of Crohn’s Disease

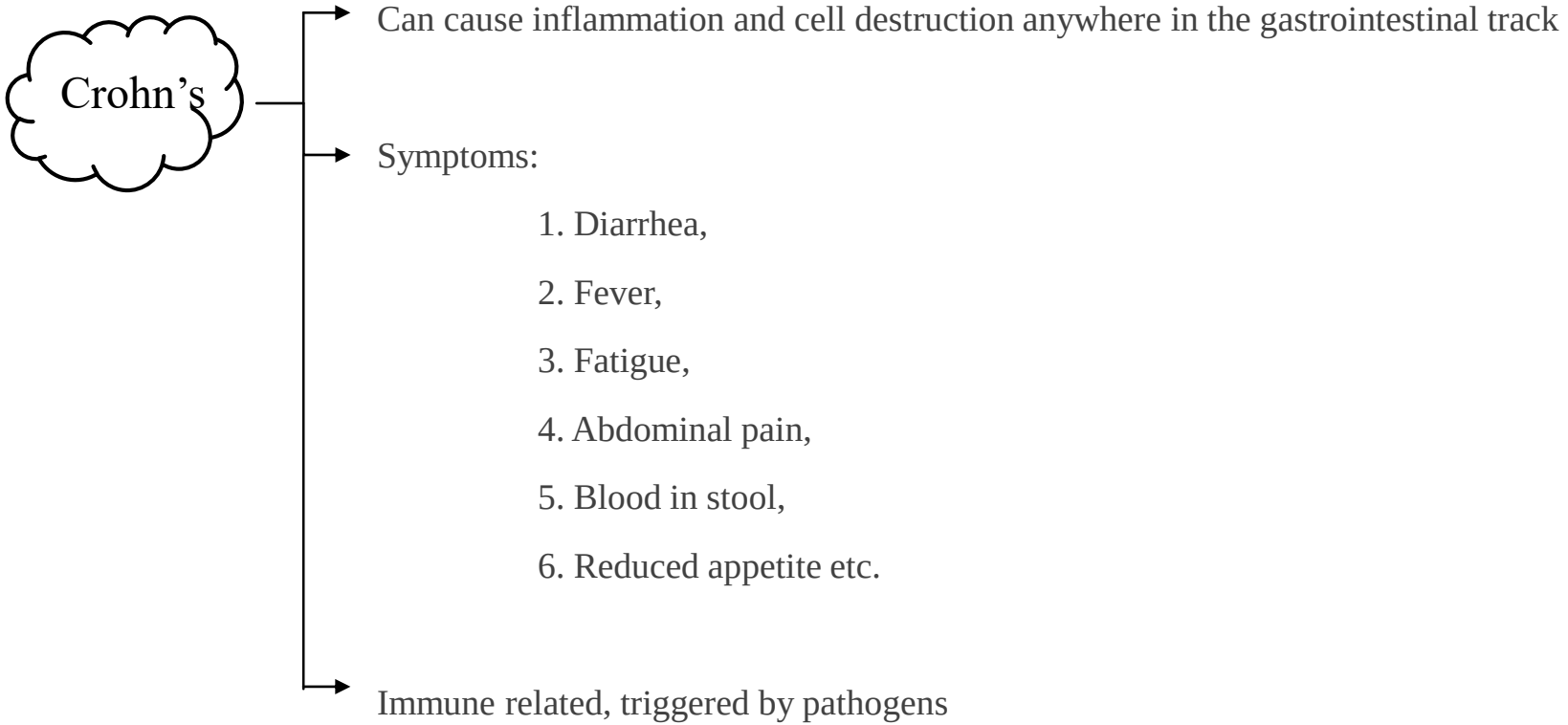
NAME: Aatreyi Pranavbhai Mehta

Carleton id: 101199821

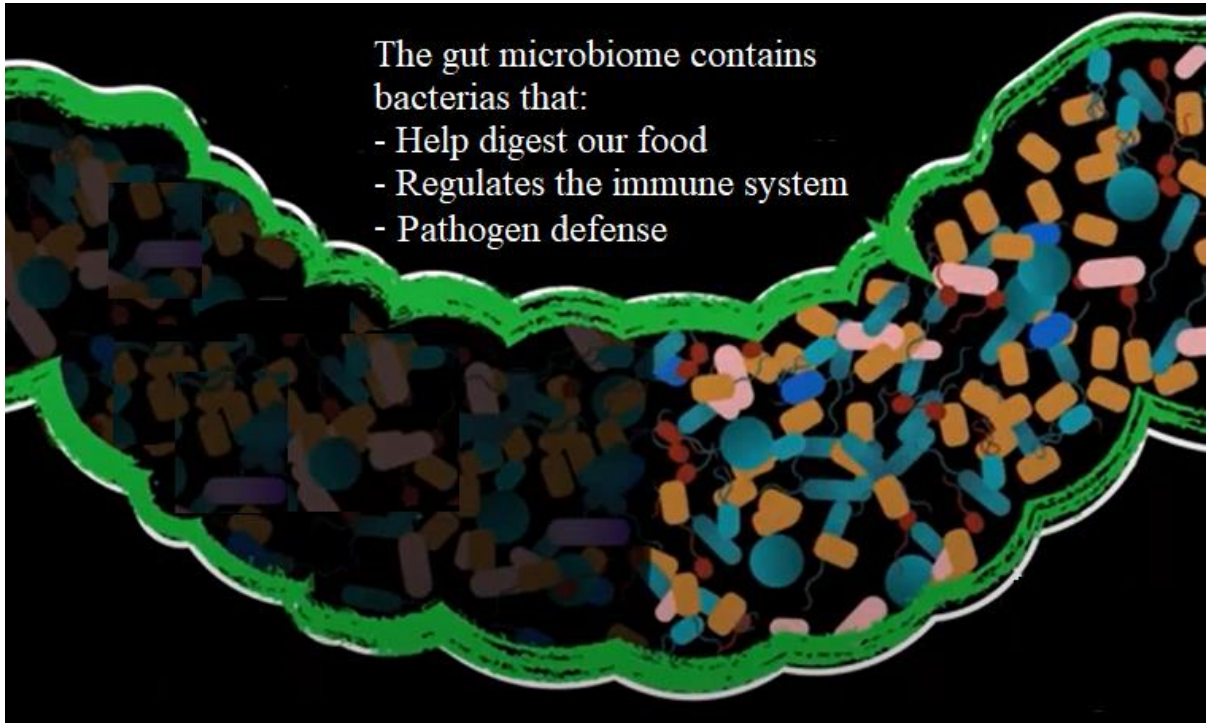
COURSE: Parallel Algorithms and Applications in Data Science [COMP 5704]

INTRODUCTION

- AIM: To provide a comparative analysis between different machine learning approaches based on their efficiency to handle a huge dataset of patients of Crohn's Disease.
- About Crohn's Disease(CD):



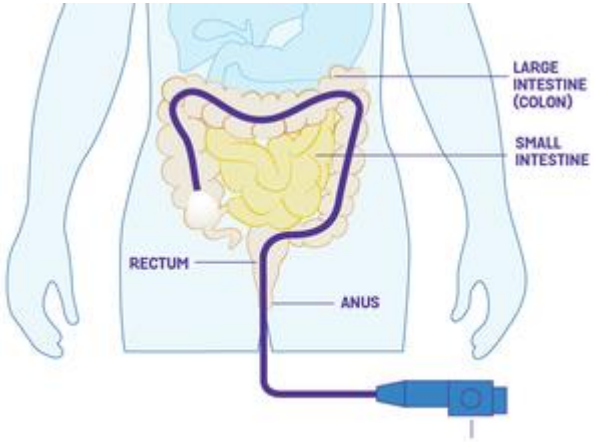
- Parallel Computing has helped to enable the study of gut microbiome.



- The human gut microbiome has up to 1,000 kinds of bacteria, majority are beneficial to your health, but some may cause illness.
- Adherent-invasive Escherichia coli (AIEC) – Type of gut bacteria that increases intestine inflammation.
- The bacteria-produced metabolite(a substance that breaks food and drugs) interacts with immune system cells in the intestinal lining, causing unregulated and out of control inflammation causing destruction of healthy tissues.

PREVIOUS METHODS USED TO TEST CD

Colonoscopy



- A tube is introduced into the rectum to examine the inside of the colon with the help of a tiny video camera at the tube's tip.
- Medication causes frequent, loose bowel movements to empty the colon.
- Use of anesthesia.

CT Scan

- Imaging test that uses computer to put a series of x-ray images together to create detailed 3D images of organs.
- Patients need to drink clear liquids and not consume food 3 hours prior to method.
- Given a liquid contrast preparation to swallow.

- According to a survey [2010], bowel preparation was stated the most unpleasant and less favorable part of these methods. The majority of people, however, recommended CT scan since it is less invasive, faster, and easier, and it does not require anesthesia.

SUGGESTED TEST METHOD

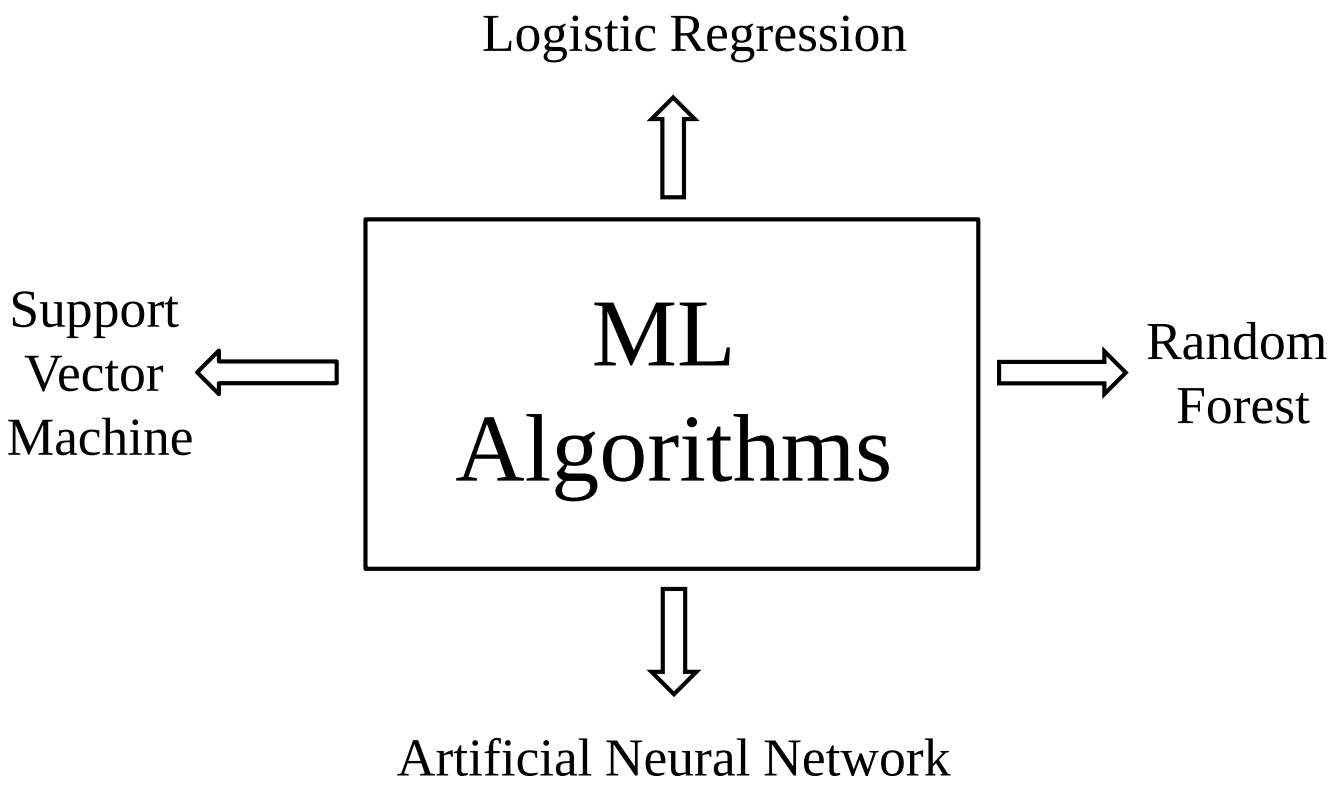
- Faecal Calprotectin - abundant protein in neutrophils and is a dimeric calcium, iron, manganese, and zinc sequestering protein.
- Released when there is inflammation in the intestine.
- Increased FC concentrations - reliable indicator of inflammatory bowel disease(IBD).
- Ideal for monitoring disease development in CD patients and lowering the frequency of needless endoscopies.

- The range of FC Level and its association with IBD:

1. < or =50.0 mcg/g - Normal
2. 50.1-250.0 mcg/g - Borderline represents a mild inflammatory process like a treatable IBD.
3. > 250.0 mcg/g - Abnormal, suggestive of an active inflammatory process within the gastrointestinal system.

- This suggests that when a patients' FC Level exceeds 250.0 mcg/g – high chance to be positive for CD.

THE ALGORITHMS OF MACHINE LEARNING

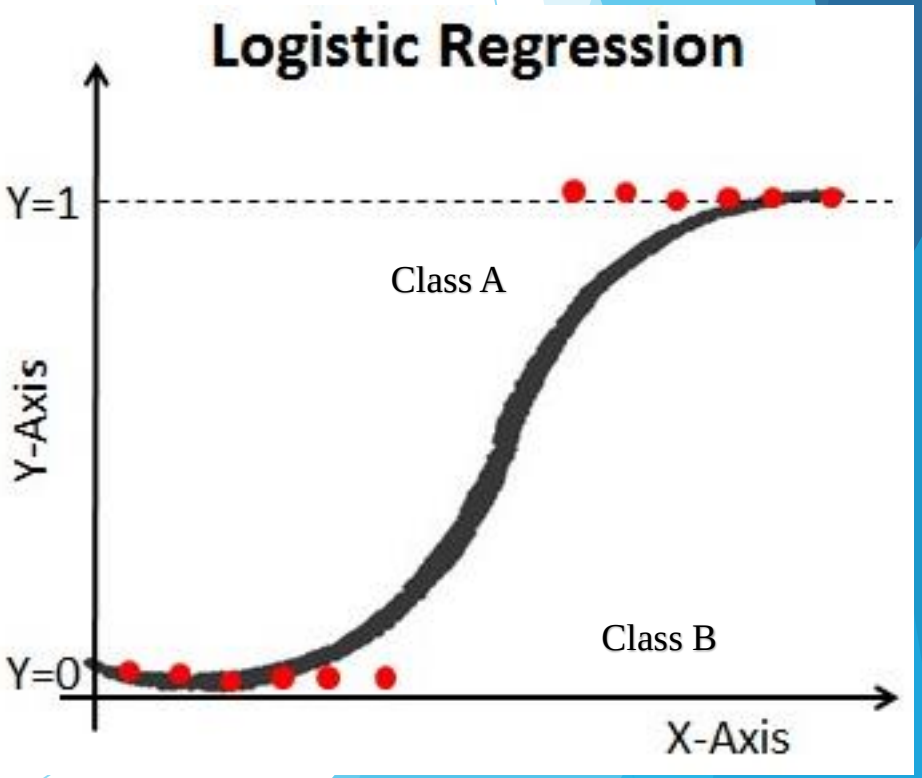


Logistic Regression

- LR can be used when the target variable is categorical. For example to predict whether the person has Crohn's disease or not.
- Binary Classification – Class A (People tested positive), Class B (People tested negative)
- Sigmoid Function: Y = 1 / (1 + e^-x)

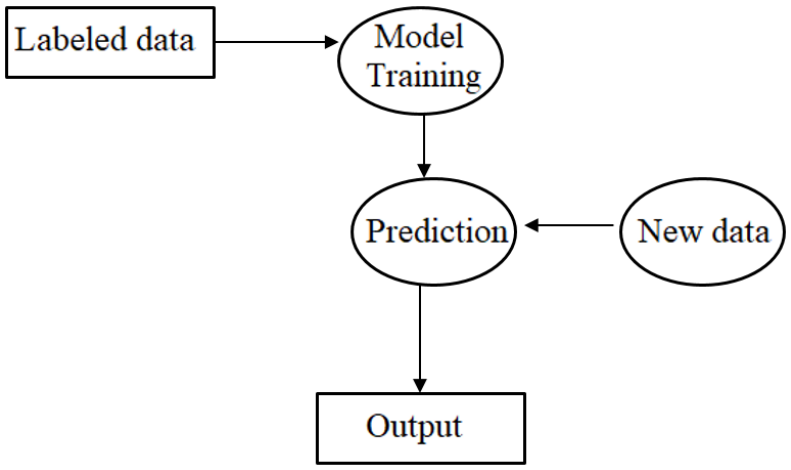
output — euler's constant (value: 2.718) — independent variable that has to be transformed

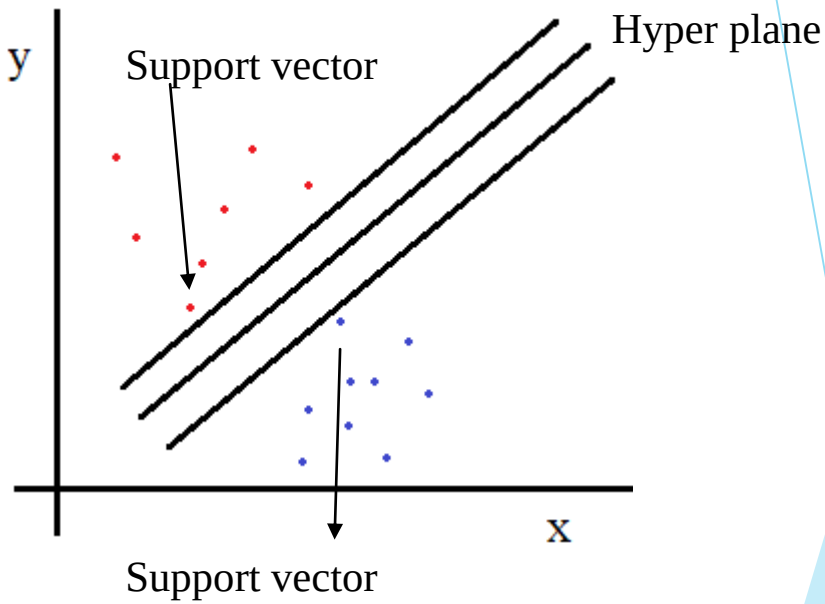
- The sigmoid function simply tries to convert the independent variable into an expression of probability ranging from 0 to 1 w.r.t. dependent variable
- Rare case: The points that fall exactly on the cutoff line will be termed as unclassifiable.
- Application: Fraud detection, Disease diagnosis, spam – no spam emails.



Support Vector Machine

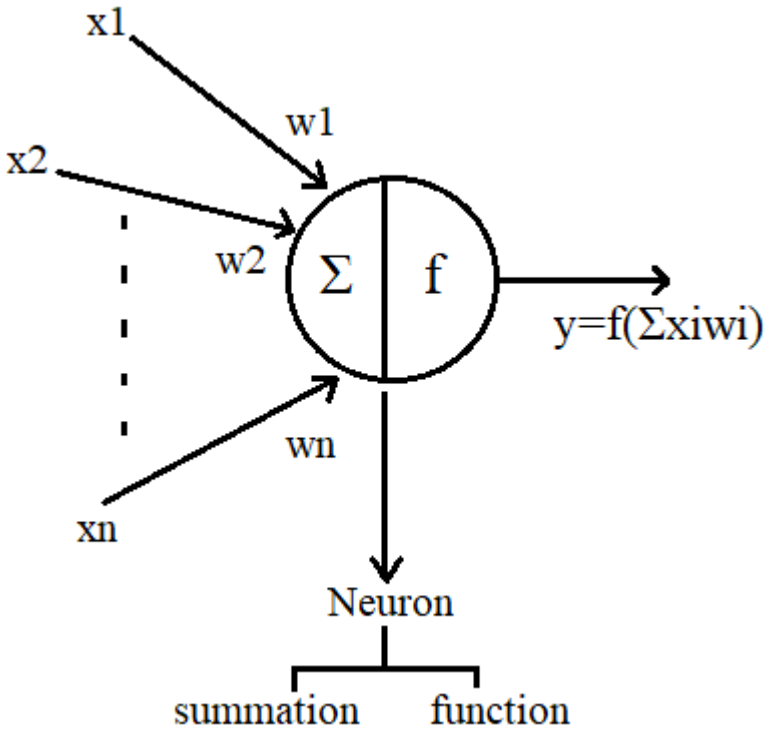
- Type of supervised learning, used for classification and regression analysis.
- To divide a data into two class, we use decision boundary(hyper plane). The hyper plane helps us decide that the new data will belong to Class A or Class B.





- A line p is constructed by the datapoint nearest to Class B(support vector), similarly another line q is constructed by the datapoint nearest to Class A(support vector).
- Distance between p and hyperplane – D-
- Distance between q and hyperplane – D+
- Margin decides which hyperplane will exist and which wouldn't, here Margin = (D-) + (D+)

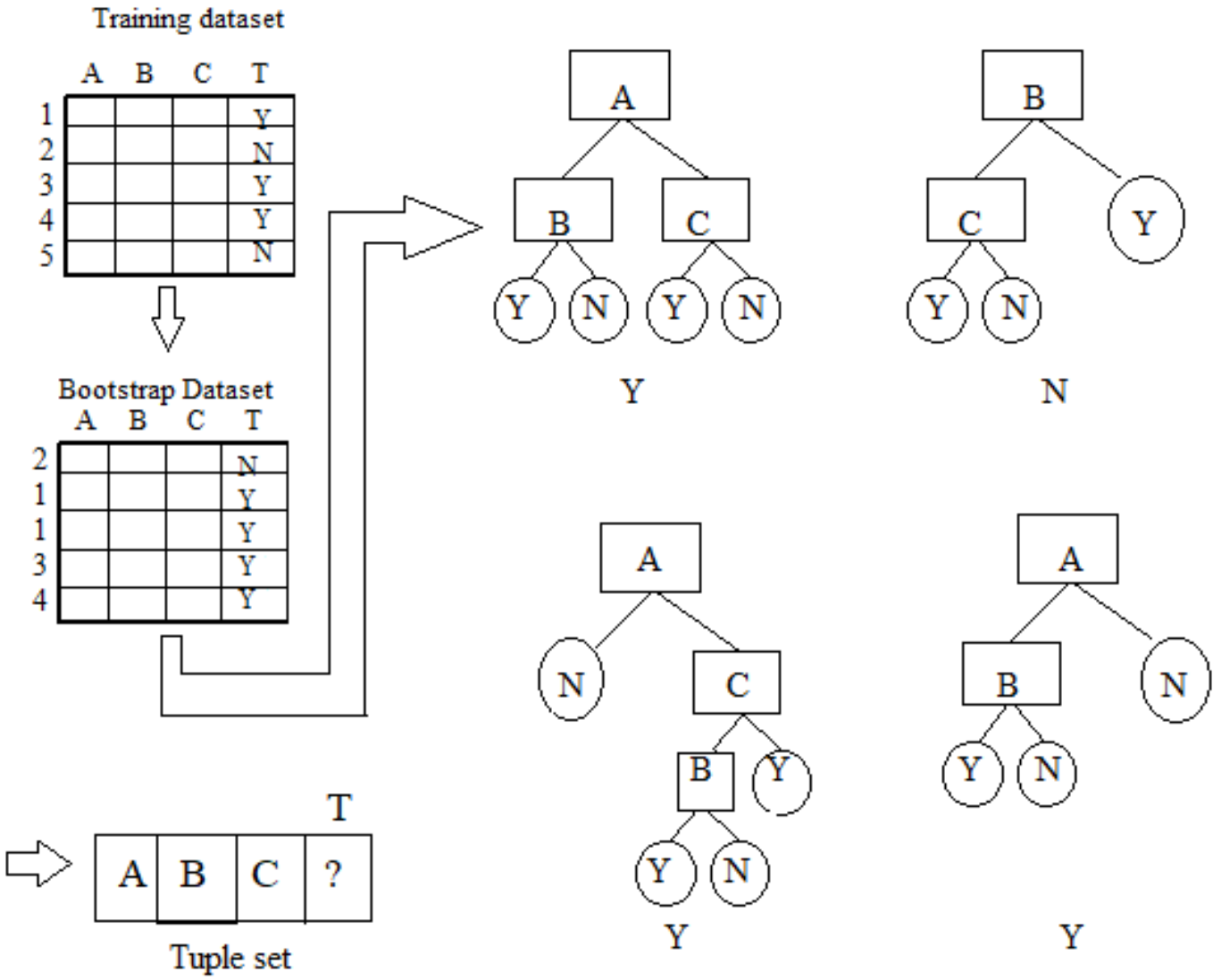
Artificial Neural Networks



- x1, x2…..xn = input signals coming from maybe various other neurons.
- Each signal has an associated weight i.e. w1, w2…..wn.
- Σ evaluates (x1w1 + x2w2….. xnwn)
- Activation function(f) generates an appropriate output for a given neuron based on the provided input

Random Forest

- It is an ensemble classifier that uses decision tree algorithm in a randomized way.



- Input: Training data
- A,B,C - Attributes; O – Target attribute
- Bootstrap Dataset(BD) is created by randomly picking and inserting any of the samples of original data, even duplication is allowed.
- Then a Decision Tree is plotted in a randomized fashion.
- For each selected node, we use subset of variables at each step.
- Example: For the root node, the possible variables are A,B,C but for deciding the root node, we consider only 2 variables out of 3 i.e. the subset of total variables. Let us consider A and B. Now assuming that A is better at splitting the samples, if so A is considered the root node.
- For the child node, we randomly consider 2 variables, B and C. Now assuming that B is better at splitting the samples, if so B is considered the left child. C will be the right child eventually. And the decision variables Y and N i.e. Yes and No will be the leaf nodes.
- Similarly, for more decision trees, the same steps are followed.
- Finally, we will get a test tuple which decides the value of target attribute with the help of final values of attribute A,B and C.
- The tuple is applied to each decision tree to know the opinion of each decision tree.
- Based on the majority opinion of each decision tree, we get the value of the test tuple.

PROGRAMMING

- Language used: Python

- Dataset: of 5000 people with 6 fields portraying their FC Level in terms of mcg/g, weight in terms of kg, height in terms of cm, BMI, gender (female = 1 and male = 0) as input and the last field represents the output that is crohn's disease (positive = 1 and negative = 0)

- Libraries used while coding:

1. numpy - a Python third-party library that makes numerical computing easier by giving a flexible N-dimensional array object to store data and powerful mathematical functions to manipulate those arrays of numbers.
2. multiprocessing – this module helps to take full advantage of multiple processors on a single machine.
3. pandas - its primary function is data analysis. It supports the import of data from a variety of file formats.
4. pyplot – for plotting 2D graphics.
5. sklearn library that has inbuilt classifiers for LR, SVM, ANN and Random Forest like LogisticRegression, SVC, MLPClassifier, RandomForestClassifier respectively.

- In the program :
- Considering each row of the dataset as an array, and taking the values of a[0], a[1]…a[4] as inputs we get the value of output as either 0 or 1 depending on the input features. For example: If the value of the fc level is 250.0 or below, there are high chances that the patient would test negative for CD, otherwise the test result would be positive.

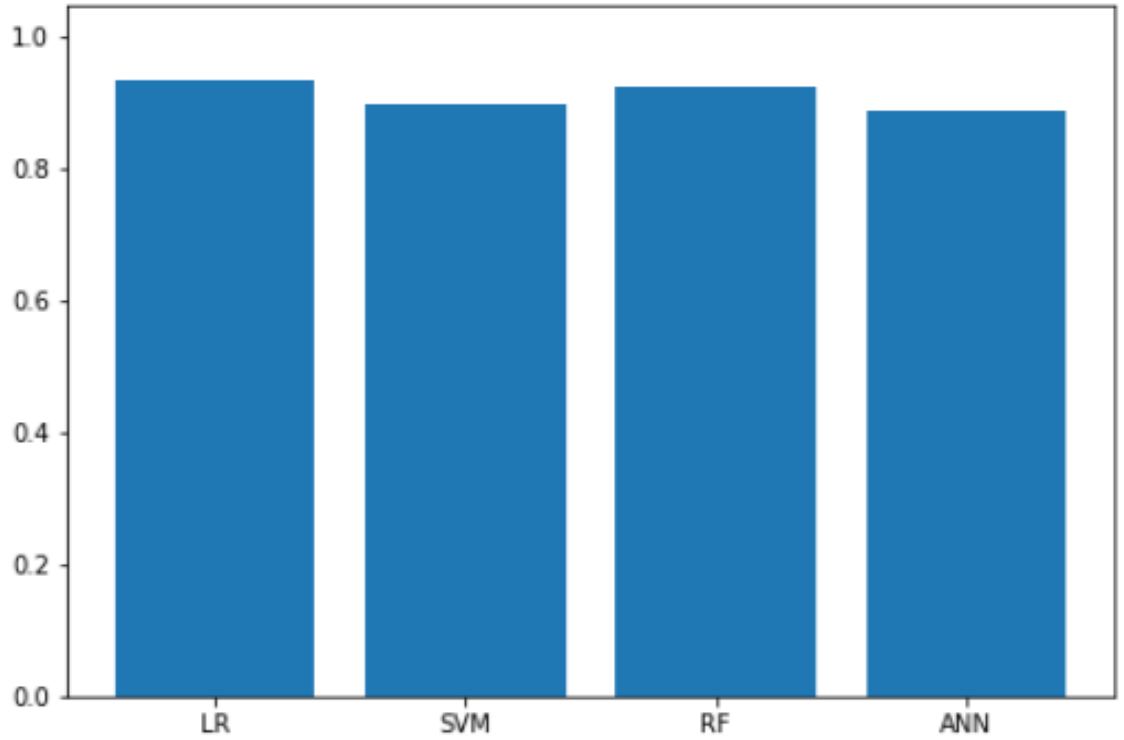
- I have used the inbuilt ML Classifies of the sklearn library to compare the accuracy of machine learning approaches and portrayed the end result in the form of a bar graph that represents the mean accuracy.

- Moreover to implement parallel execution of a function in multiprocessing, there are two main objects:

  The Pool Class and the Process Class.
- Pool runs multiple Python processes in the background and distributes your computations across multiple CPU cores, allowing them to run in parallel.
- To make any function run in parallel, Pool() provides the apply function.
- apply() accepts the parameters passed to the 'function-to-be-parallelized' as an argument in the args argument.
- Due to the ongoing parallelization, the overall execution time is improved.

RESULT

```
LR:  0.971400 (0.028600)
SVM: 0.948600 (0.051400)
RF:  0.968900 (0.031100)
ANN: 0.939900 (0.060100)
```




```
The number of patients who have tested negative for Crohn's disease:  4980
The number of patients who have tested positive for Crohn's disease:  20
<multiprocessing.pool.Pool object at 0x7fe9100bac50>
```

CONCLUSION

- The accuracy of Logistic Regression turned out to be 97.14% that is the highest as compared to all other algorithms, followed by Random Forest with the accuracy of 96.89%.
- Support Vector Machine also shows a considerably good accuracy of 94.86%.
- For the huge dataset as ours, Artificial Neural Network portrays the least accuracy as compared to the other models.
- All in all, it can be determined that Logistic Regression was the model that predicted the highest accuracy for a large dataset of Crohn’s disease patients.

REFERENCES:

1. https://www.mayoclinic.org/diseases-conditions/crohns-disease/symptoms-causes
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5390326/
3. https://bmcgastroenterol.biomedcentral.com/articles/10.1186/s12876-020-1183-x
4. https://www.kaggle.com/datasets?search=rohn%27s
5. https://www.mayocliniclabs.com/test-catalog/Clinical+and+Interpretive/63016
6. https://www.machinelearningplus.com/python/parallel-processing-python/
7. https://www.youtube.com/results?search_query=5+minute+engineering

Questions

Q1. Which bacteria of gut microbiome, if present in abundance causes inflammation leading to CD?

Q2. How is ANN different than LR, SVM and Random Forest?

Q3. What class and method is used for parallelization in python?


thank you