**BIO542: Machine Learning for Biomedical Applications**

**Assignment 2**

**Plagiarism:** All submitted codes are expected to be the result of your individual effort. You should never misrepresent someone else’s work as your own. In case any plagiarism is detected you will get one grade reduction in the final examination. Cite the resource wherever using other’s work.

**Instructions:**

1)   Allowed programming language is Python. Name of the input and output files should be given as argument and proper documentation is required inside the code.

2)   Mention your roll number and name at the start of each file. Write a clean code with proper comments at appropriate places as it will be checked.

3)  Store each problem with rollno\_questionX.py where X is the question no. and upload a zip folder with rollno\_name\_assignmentX.zip containing all the codes. Codes won’t be checked if they don't follow the guidelines.

4)  You all need to submit only zipped files. RAR files or file with any other extension will not be evaluated.

5) Create a docker container (rollno\_name\_A2) on docker hub and submit the zipped files into your own docker container. We pull your docker container to evaluate Assignment 2.

**Link for the dataset:**

You can download the data from the given link

<https://drive.google.com/drive/u/0/folders/1KrFWz_4GQDIqDKTlwA30DrFP5kJakINM>

At the provided link, there is a folder named mlba\_data\_2019\_A2, which contains two zip files, one is PRROM.zip and another is PTD.zip.

Unzip both the files, PRROM.zip and another is PTD.zip. Post unzipping, you will get two folders named PRROM and PTD. Both folders contain 74 different folders.

There is a folder for each student with their name and Roll no mentioned clearly.

Download your respective folder.

**(A) Questions on Expression Dataset (PTD)**

control\_sPTD folder contains two files (training and validation dataset (70:30)); for positive samples and negative samples file. Each file contains the expression data of the genes corresponding to each sample, where rows represents the sample and columns represents the gene expression value and the column name (Group) represent the label (control vs sPTD (preterm delivery)).

These values are different for each student.

For a given dataset you need to answer the following questions:

Question1: Write a python code for feature selection from the given dataset by excluding these (GA, GADel, Group, TTD) features. Write the selected features into output file named as (rollno\_name\_q1.csv) and submit the python code (rollno\_name\_q1.py)

Question2: Write a python code for developing a classification model, using different machine learning techniques you can use (Scikit learn library). Train the model on training dataset and check the performance on validation dataset and submit the python code (rollno\_name\_q2.py)

Question3: Compute performance of your best models on testing dataset in term of Accuracy, MCC, Specificity, Sensitivity and AUC-ROC Value and submit it into a separate file (rollno\_name\_q3.csv).

Question 4: Repeat the above process by including 3 features such as "GA (gestational age), GADel ( gestational age at delivery), TTD (interval in weeks from sample to delivery (GADel-GA)" with selected features in combination, either one of these three features and train the model again including these features and compute the performance. Write the output in .csv file (rollno\_name\_q4.csv)

**(B) Questions on Expression Dataset (PRROM)**

Control\_PRROM folder contains two files (training and validation dataset (70:30)); for positive samples and negative samples file. Each file contains the expression data of the genes corresponding to each sample, where rows represents the sample and columns represents the gene expression value and the column name (Group) represent the label (control vs PRROM (Premature rupture of membranes)). These values are different for each student.

For a given dataset you need to answer the following questions:

Question5: Write a python code for feature selection from the given dataset by excluding these (GA, GADel, Group, TTD) features. Write the selected features into output file named as (rollno\_name\_q5.csv) and submit the python code (rollno\_name\_q5.py)

Question6: Write a python code for developing a classification model, using different machine learning techniques you can use (Scikit learn library). Train the model on training dataset and check the performance on validation dataset and submit the python code (rollno\_name\_q6.py)

Question7: Compute performance of your best models on testing dataset in term of Accuracy, MCC, Specificity, Sensitivity and AUC-ROC Value and submit it into a separate file (rollno\_name\_q7.csv).

Question8: Repeat the above process by including 3 features "GA (gestational age), GADel ( gestational age at delivery), TTD (interval in weeks from sample to delivery (GADel-GA)" with selected features and train the model again including these features and compute the performance. Write the output in .csv file (rollno\_name\_q8.csv)