**NOTE:**

1. This is a private dataset, please do not share with anyone else.
2. Please submit both your Jupyter notebook and a PDF version
3. Please name your assignment with the names of the group members
4. Please add group member names on the document as well.
5. The dataset has 1303 samples with 39 columns. The target variable is 'SARCOPENIA'. Do not use `STAR’ variable in your analysis.

**Background:**

Age-related loss of skeletal muscle mass and function, defined as sarcopenia, is significantly related to adverse health outcomes e.g. increased risk of falls and fractures, physical frailty, mobility limitation, and even premature mortality.1 Besides, aging, smoking, physical inactivity, malnutrition and chronic comorbid diseases such as hypertension (HT), diabetes mellitus (DM), coronary artery disease (CAD), hyperlipidemia and obesity also have deteriorative effects on muscle mass and function.2 Prevention of sarcopenia-related morbidity and mortality, among aging people in particular, depends on early detection and treatment, more importantly on unraveling the relationships between muscle loss, decline in muscle function and physical performance and the aforementioned potential predictor

YOU WILL PREDICT IF A USER WILL BE DEVELOPING SARCOPENIA.

The following is a guideline for you, you can approach the problem any way you like. You don’t have to follow the suggestions below.

**You will do:**

1. **Preprocessing**

Read the data and print the data information, the shape of the data and explore it. You may need to create new variables such as Education level, Smoking, Gender, Exercise, etc.

1. **MACHINE LEARNING**

Implement Logistic Regression, Random Forest, and Gradient Boosting on this dataset. You might plan to develop two separate models for men and women, as gender plays a significant role in Sarcopenia.

You will develop three models:

**Model1:** Determine the probability that the patient has Sarcopenia based on variables available when the patient is home, such as age, weight, height, exercise, Hiper tension, DM, Education, Smoking.. Values for these variables can be obtained without the patient reporting physically to a clinic.

At the end of Model1, diagnose patients with very high probability of developing Sarc as positive and very low probability of Sarc as negative. The remaining patients need further testing, so send them to the send model.

**Model 2:** Predict the Sarcopenia probability including CST and GSP. To measure CST, “Chair Stand Test”, a physician times a patient when she sits and stands up on a chair five times. To measure GDP, “Gait speed test”, a physician times a patient when she walks a certain distance. We are provided with these two features in the dataset. In the second model, in addition to the variables we used in the first model, we will include CST and GDP.

**Model 3:** Predict the Sarcopenia including ‘GRS’, Grip Strength Test. A physician measures the grp strength of a patient using a gyro hand grip. We will include GRS into our model with all the variables from Model2.

So, in sum, we start with a patient pool, we first eliminate the ones that we are almost certain whether they have Sarcopenia or not. Next, the remaining patients go through Model 2 and we again eliminate the patients that we are almost certain whether they have Sarcopenia or not. We will repeat the same procedure for Model 3. The remaining patients after Model 3 will go through ultrasound process to measure their muscle thickness, which will increase the certainty for diagnosing whether they have Sarcopenia. Our goal is, making the diagnosis as much as possible and reduce number of patients going through extra tests.

In your deliverable, report precision, recall, f1, and auc scores. Report feature importance with a bar graph.

Create a simple web application which takes in data from the user and calculates the probabilities based on three models.