**Data used:**

1. Update\_train.csv
2. Updated\_Test.csv
3. Exports\_ldl\_rows\_135\_full.csv
4. Exports\_hdl\_rows\_162\_full.csv
5. Export\_hgb\_rows\_965\_full.csv

**Packages used:**

1. Pandas
2. Numpy
3. Scikit-learn
4. Os
5. Re
6. Lightgbm
7. Catboost
8. xgboost

**Output data and where they are stored:**

1. We run our code in Colab notebook and the file is read from our google drive
2. The output thus obtained is directly saved in our google colab itself

**Explanation of features used:**

Since we have little knowledge of each absorbance, our main feature is the target value itself. From the extra dataset given , we used ldl\_135, hdl\_162 and hgb\_965 to predict cholesterol\_ldl\_value, hemoglobin(hgb)\_value and hdl\_cholesterol\_value.

These features are used for the prediction of target variable

We have not used any external data apart from what zindi has provided to us.

**Steps followed: (seed is 88)**

1. Read the dataset as mentioned in Data used
2. Using extra dataset (3,4,5 in Data Used), LGBM Regressor model is built to predict the features cholesterol\_ldl\_value, hemoglobin(hgb)\_value and hdl\_cholesterol\_value
3. Std column is removed as it is the closeness of different scans made using spectroscopy
4. Min Max scaler is applied to each absorbance to normalize the value
5. Chi-squared value is computed between target variables to know the dependency between each target variable
6. Each class is mapped to 0,1 and 2 for low, ok and high respectively
7. 3 models LGBM, catboost and XGB is built and run for 5000 iterations to predict each class and finally the mode is taken as a final prediction (voting classifier)
8. Submission file is created based on the specifications given