Lab 4 – Non-Parametric Regression

You are provided data related to disease progression in patients (for data privacy seasons, the disease and the feature names have been anonymized). The data is in two files – one contains records for patients who were diagnosed with the disease, the other is for patients who were not diagnosed.

Using this data, perform the following tasks:

1. Load the data and concatenate the two data frames (1 point). Plot the percentage of missing values for each feature (1 point) and a histogram of the number of missing features per record (1 point).
2. Create a second data frame with the missing values filled with the mean value for that feature (1 point). Split both the filled and unfilled data into training (80%) and testing (20%) sets using the same split for both (1 point).
3. Using the data with filled missing values, train each of the following models to predict if a patient should be diagnosed with the disease (1 point each). For each model, record the time taken to fit the data.
    a. A logistic regression (using sklearn or statsmodels)
    b. A support vector classifier (using sklearn)
    c. A random forest classifier (using sklearn)
    d. A boosting tree classifier (using lightgbm)
    e. A boosting tree classifier – *trained on the data with missing values* (using lightgbm)
4. Create a feedforward neural network with input dimensions matching the data and an output dimension of one. You can use any hidden size, activation function, and network depth you desire, but the output layer should have a size of 1 and a sigmoid activation function (3 points). Train this model using any optimizer and number of epochs you choose (3 points). Record the time taken to fit this model.
5. Using each of the six models above, calculate the following metrics using the appropriate test set:
    a. F1 score (1 point)
    b. Log loss (1 point)
    c. The area under the ROC curve (1 point)
6. Repeat step 5, limiting your analysis to patients with zero to eight missing values (*i.e.* 0 missing, 1 missing, etc.). Plot for each model the log loss and F1 score for each. How do you interpret these values? (3 points)
7. Based on the results above, which model do you think is the most appropriate? Justify your answer (2 points).