Prediction of benign or malignant cancer tumors:

The goal of this project is to classify breast cancer tumors into malignant or benign groups using the provided database and machine learning skills. In other words, we try to predict the probability of a tumor being benign based on the historical data (feature and target variables) that are already synthetized.

The data for this study is a modified version of a dataset that is collected from UCI Machine Learning Repository [1]. In current version of the data, all values are synthetized, and they are not real-valued features. The only purpose of this dataset is to test machine learning skills of the applicants.

[1] Source: Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Attribute Information:

1) ID number
2-31) Ten synthetic-valued features are computed for each cell nucleus:
    a) radius (mean of distances from center to points on the perimeter)
    b) texture (standard deviation of gray-scale values)
    c) perimeter
    d) area
    e) smoothness (local variation in radius lengths)
    f) compactness (perimeter^2 / area - 1.0)
    g) concavity (severity of concave portions of the contour)
    h) concave points (number of concave portions of the contour)
    i) symmetry
    j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features.
32) Diagnosis (M = malignant, B = benign)

- Please show all your work in Python Jupyter notebook.
- Using data visualization tools, please explain how we can understand the data structure.
- Please explain if dimensionality reduction is required/possible or not. How did you check?
- Which classification methods are you using? How do you decide among different methods?
- Please provide confusion matrix and explain how it can help us to check reliability of result.
- Please provide the learning curve and explain how it can help us in determining whether the model is being over-fit or under-fit.
- When do you consider adding “regularization parameter” to the model? and how it will help to improve the model performance?