Machine Learning – Final 12/16/2024

You must submit a CLEAN runnable Jupyter notebook showing all your work and your responses to (a)-(h) below. Name the file your_lastname_final.ipynb

Full credit will be given for clean, readable, executable notebooks and clear explanations only. ***Make sure you submit everything before 12.15 PM – I will not be responsible if you get locked out from Brightspace after the time expires.***

The file *data.csv* (on Brightspace) contains data on 30 features of a system and a "target" value of interest. Missing data, if any, is marked with "NA". You want to develop a Logistic Regression model to predict the target from the features. **Use a random_state=104 for all calculations.**

(a) **List** the *data preparation* (eliminate missing data rows, encoding, scaling, etc.) steps you will need to use, in the order you will do them.

(b) Carry out the steps you list in (a) in an attached Jupyter Notebook. Only steps listed above must be in the notebook. Steps should be clear, and commented to show what you are doing.

(c) Split the prepared data into a training (80%) and a test set(20%). Train a logistic regression model with the training data, and explore the performance metrics of the model with the test data.

(e) If you want to reduce the number of features to less than 30, you can do a PCA. To ensure a 95% variance ratio, how many PCA components do you need?

(f) Carry out the feature reduction with PCA indicated by (e). For the PCA-reduced features set, carry out the same Logistic regression modeling. What are the metrics for the new model against the test data? How do they compare with the metrics of the model with all features included?

(g) **Using the PCA-reduced model**, predict the target for the new data (in terms of original target variables) given in the attached *new_data.csv* file.

(h) Assume that in (g) you are predicting tumors to be benign, "B", or malignant, "M", based on tumor dimensions. How confident are you in those predictions? What will you tell the patients about their wellness?