ML Problem

Dataset Description:

Mammography is the most effective method for breast cancer screening available today. However, the low positive predictive value of breast biopsy resulting from mammogram interpretation leads to approximately 70% unnecessary biopsies with benign outcomes. To reduce the high number of unnecessary breast biopsies, several computer-aided diagnoses (CAD) systems have been proposed in the last years. These systems help physicians in their decision to perform a breast biopsy on a suspicious lesion seen in a mammogram or to perform a short-term follow-up examination instead.

This data set can be used to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. It contains a BI-RADS assessment, the patient's age and three BI-RADS attributes together with the ground truth (the severity field).

Attribute Information:
1. BI-RADS assessment: 1 to 5 (ordinal, non-predictive!)
2. Age: patient's age in years (integer)
3. Shape: mass shape: round=1 oval=2 lobular=3 irregular=4 (nominal)
4. Margin: mass margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5 (nominal)
5. Density: mass density high=1 iso=2 low=3 fat-containing=4 (ordinal)
6. Severity: benign=0 or malignant=1 (binominal, goal field!)

Pre-requisite: Python environment setup for Jupyter notebook is mandatory.

Instructions:
a. You can use any libraries of your choice in python.
b. Provide the code in a notebook with the format Your_Name.ipynb
c. The notebook should run without errors and should also display outputs and visualizations.

Evaluation Task:
Download the dataset from attached file and perform the following tasks:

1. Build Statistical Classification model to detect severity
2. What considerations have been used for model selection?
3. What features would you want to create for your prediction model based on data provided?
4. How have you performed hyper-parameter tuning and model optimization? What are the reasons for your decision choices for these steps?
5. What is your model evaluation criteria? What are the assumptions and limitations of your approach?
6. Determine whether the data is normally distributed visually and statistically.
7. Comment on EDA of variables in data.
8. How are you detecting and treating outliers in the dataset for better convergence?
9. What techniques have been used for treating missing values to prepare features for model building?
10. What is the distribution of target with respect to categorical columns?
11. Comment on any other observations or recommendations based on your analysis.