A diagnostics team is developing a model to classify a tumor as either “benign” or “malignant”. Attached is the data set.



The data dictionary is provided below:

1) ID number  
2) Diagnosis (M = malignant, B = benign)

Ten real-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. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All feature values are recoded with four significant digits.

(a) Conduct an Exploratory Data Analysis and explain your findings.

**(b)** Develop at least three classification models using different machine learning techniques.

(c) Explain in detail which model is better and why?

(d) Develop principal components for all independent variables, develop the ML models on principal components. Compare these models with previous models. What are the advantages (if any) of using PCA in classification model building?