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

* The data set does not have any “Null values”, “Character values” or large outliers so no need to clean up the data. From observation the mean values of the columns like perimeter, concavity, area, compactness, radius are used to the classification of the diagnosis. From the parameter, larger value defined the correlation with malignant tumors.
* The mean value from the fractual dimension, smoothness or texture, it does not show particular preference of one diagnosis over the other.

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

Developed the below classification models for the data set

* Logistic Regression
* KNeighborsClassifier
* Random forest

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

Based on the feature importance I have selected the columns Concave points, Concavity, Perimeter, Area, Radius as a predictor. It gives the accuracy of 98.995% while using the model **“RandomForestClassifier”** with the columns Concave points, Concavity, Perimeter, Area, Radius as a predictor. **“Random Forest Classifier”** gives accuracy of 96.737% with the column radius as predictor.

(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 (ifany) of using PCA in classification model building?

It is very useful to reduce the number of columns in the dataset with the importance and helps to improve the accuracy level of the analysis.