**Section – A**

**1)d**

**2)d**

**3)c**

**4)c**

**5)c**

**6)d**

**7)a**

**8)b**

**9)d**

**10)b**

**11)a**

**12)a**

**13)d**

**Section – B**

What is the modelling technique you chose and why?

I choose to use Logistic Regression Technique.

# We know we have 2 types of Algorithms:

A) Supervised-Y or Target variable is there.

Here it is of two types. Regression and classification

B) Unsupervised- Y or Target variable is not there.

## For supervised, we can have either classification or Regression problems.

For classification problems, Target Variable will be discrete or categorical variable.  
For Regression problems, Target variable will be continuous or scale or Numerical Variable.

Since here y or target variable is there,so we need to use Supervised Learning Technique.

Since, Target Variable is discrete or categorical variable, so I need to use Modelling Algorithm meant to be used for classification problems.

Again,for classification problem we can either use Decision Tree or Logistic Regression.However, Decision trees are prone to be over fitting.

Since,here data sample is small we can use Logistic Regressio to build the model.l

2)What pre- processing you considered on this data and why?

**Initially I tried to find whether there is any missing values.**

**From df.info(),I got the following Results:**

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 699 entries, 0 to 698

Data columns (total 10 columns):

clump\_thickness 699 non-null int64

cell\_shape\_uniformity 699 non-null int64

cell\_size\_uniformity 699 non-null int64

marginal\_adhesion 699 non-null int64

single\_epithelial\_cell\_size 699 non-null int64

bare\_nuclei 699 non-null object

bland\_chromatin 699 non-null int64

normal\_nucleoli 699 non-null int64

mitoses 699 non-null int64

class 699 non-null int64

dtypes: int64(9), object(1)

memory usage: 54.7+ KB

from Here,I came to know that none of the variables had missing values.

If I had missing values then I had to compute Mode and replacing missing values with that for Categorical variables.

I might had used mean or median to replace missing values for scale variables.

Since there were no missing values,I was not needed to compute missing values.

Also I came to know that Bare\_nuclei is a categorical variable which is also an independent variable

Rest are all scale variables.

**However,Bare\_Nuclei was not needed to be used in Dummy coding as only if categorical variable values are not in terms of 0 and 1 we need to do dummy coding.**

**For just the sake of showing how to do the dummy coding,I used them in my coding.**

**Using Pandas Profiling I also observed the EDA report to make some general EDA.**

4)What metrics you have computed to evaluate model performance and why? How is the model performance using these metrics? What are the most important features for prediction of Breast Cancer?

**The metrics that I have used to Evaluate Model performance are the following:**

**1)** ACCUIRACY SCORE

0.9714285714285714

Here,it shows what is the overall Model Accuracy which turns out to be 97%.

Since the model Accuracy is too high,We need to get little careful as it means that even by Random Guessing we could have arrived with some similar results.

But to check whether The Model Built is good or not we need to check for the next Metric which would be precision and Recall.

precision/recall Metrics

precision recall f1-score support

0 0.97 0.99 0.98 148

1 0.97 0.94 0.95 62

accuracy 0.97 210

macro avg 0.97 0.96 0.97 210

weighted avg 0.97 0.97 0.97 210

Precision for 0 and 1 both are equal high and are respectively 0.97,0.94 and 0.97 and 0.94 respectively.

Since precision and recall for both 0 and 1 are equally good.hence this assures that Model built is good.

Next,I need to consider AUC which is :

0.9609851787271143

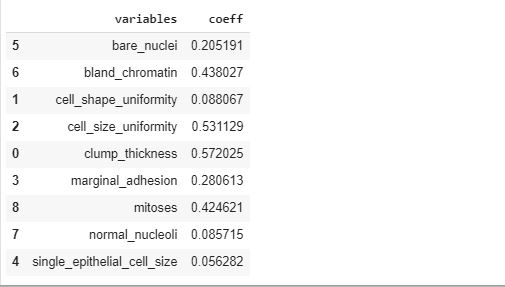
Hence AUC looks pretty good.

Finally I can plot ROC chart and Confusion Matrix to explain the Model metrics in a more better way.

If it would have been Linear Regression Model I might had used R-square for both Training and Test data to Judge Model performance and then I might had used MSE AND MAE to determine Model is stable or not with respect to the errors.

Since ,it is Logistic Regression I can’t use R-square,MAE and MSE to determine Model Performance.

I found the following Regression coefficients as given by the Model:



From here I can find out that that cell\_shape \_uniformity is most prominent with regression coefficient of 0.088067 followed by normal\_nucleoli with coefficient as 0.85715.

The other significant variables are

cell\_size\_uniformity:0.5311290

clump\_thickness:0.5720253

single\_epithelial\_cell\_size:0.056282