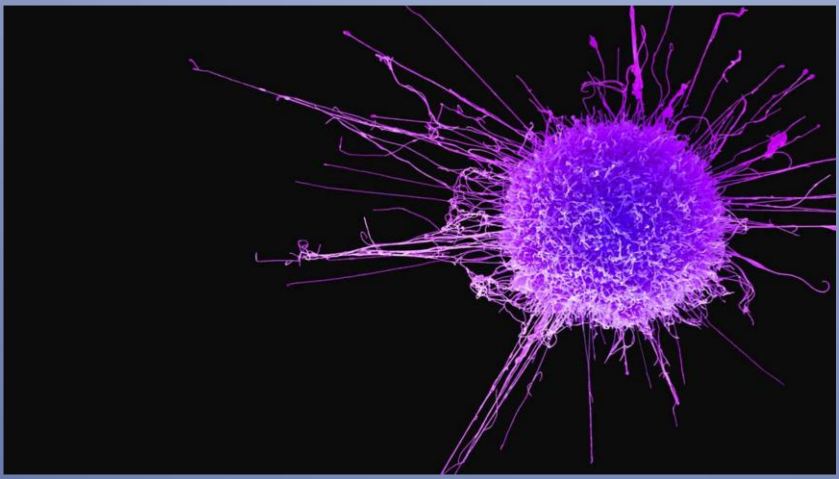


Breast Cancer

It is classification problem where the cancer is malign or baning. It is pure of work of mine.

Content

Introduction: Breast cancer arises in the lining cells (epithelium) of the ducts (85%) or lobules (15%) in the glandular tissue of the breast. Over time, these in situ (stage 0) cancers may progress and invade the surrounding breast tissue (invasive breast cancer) then spread to the nearby lymph nodes (regional metastasis) or to other organs in the body (distant metastasis). If a woman dies from breast cancer, it is because of widespread metastasis.

Breast cancer treatment can be highly effective, especially when the disease is identified early. Treatment of breast cancer often consists of a combination of surgical removal, radiation therapy and medication (hormonal therapy, chemotherapy and/or targeted biological therapy) to treat the microscopic cancer that has spread from the breast tumour through the blood.

In 2020, there were 2.3 million women diagnosed with breast cancer and 685 000 deaths globally. As of the end of 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world's most prevalent cancer. There are more lost disability-adjusted life years (DALYs) by women to breast cancer globally than any other type of cancer.

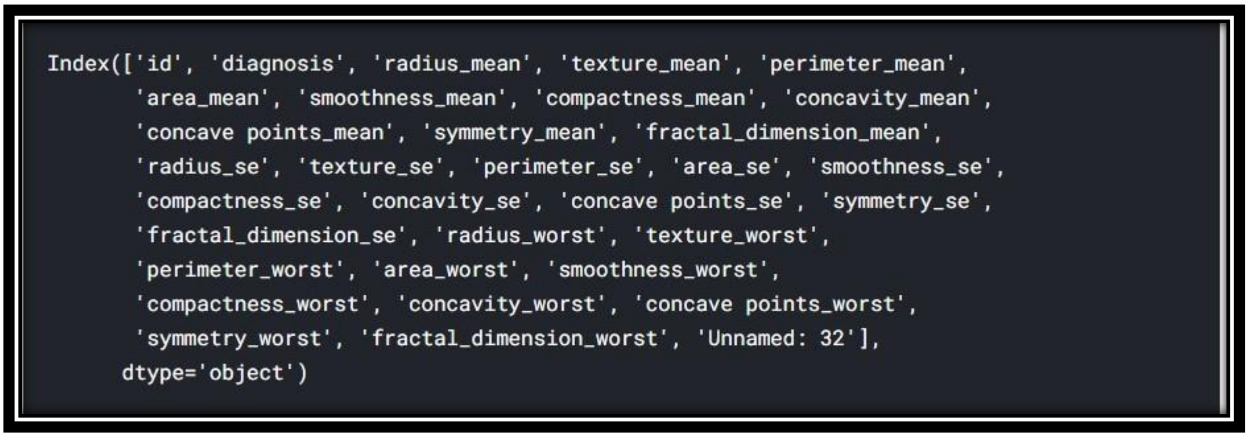
Who is at risk? Breast cancer is not a transmissible or infectious disease. There are no known viral or bacterial infections linked to the development of breast cancer. Certain factors increase the risk of breast cancer including increasing age, obesity, harmful use of alcohol, family history of breast cancer, history of radiation exposure, reproductive history (such as age that menstrual periods began and age at first pregnancy), tobacco use and postmenopausal hormone therapy.

Behavioural choices and related interventions that reduce the risk of breast cancer include:

- prolonged breastfeeding;
- regular physical activity;
- weight control;
- avoidance of harmful use of alcohol;
- avoidance of exposure to tobacco smoke;
- avoidance of prolonged use of hormones; and
- avoidance of excessive radiation exposure.

Treatment Breast cancer treatment can be highly effective, achieving survival probabilities of 90% or higher, particularly when the disease is identified early. Treatment generally consists of surgery and radiation therapy for control of the disease in the breast, lymph nodes and surrounding areas (locoregional control) and systemic therapy (anti-cancer medicines given by mouth or intravenously) to treat and/or reduce the risk of the cancer spreading (metastasis).

Dataset:- Dataset has been collected from open source and is available at through the UW CS ftp server : *ftp ftp.cs.wisc.educd math-prog/cpo-dataset/machine-learn/WDBC/.* Besides that it is also available at Kaggle. Since this dataset is used for the classification work. The features that are available in this dataset are listed below:



```
Index(['id', 'diagnosis', 'radius_mean', 'texture_mean', 'perimeter_mean',
       'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean',
       'concave points_mean', 'symmetry_mean', 'fractal_dimension_mean',
       'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se',
       'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se',
       'fractal_dimension_se', 'radius_worst', 'texture_worst',
       'perimeter_worst', 'area_worst', 'smoothness_worst',
       'compactness_worst', 'concavity_worst', 'concave points_worst',
       'symmetry_worst', 'fractal_dimension_worst', 'Unnamed: 32'],
      dtype='object')
```

Fig-1

Details of Scales:- Since we can see that “id” is one the features of the dataset which is serial number of each patients who underwent into the breast cancer diagnosis. Since “id” is serial number hence it is interval scale.

Features	***Scales***
id	Interval
diagnosis	Nominal
rest 30 features	Ratio

The above is the details of the scales of the each feature which are available in the dataset.

Explanatory and Response variable: As we can see that there

Features	variables Type
diagnosis	Response
rest 30 features	Expalanatory

Note: There are total of 32 variables. Since “diagnosis” is response variable while rest of excluding “id” are the explanatory variables.

Response Variables:- It can bee see that

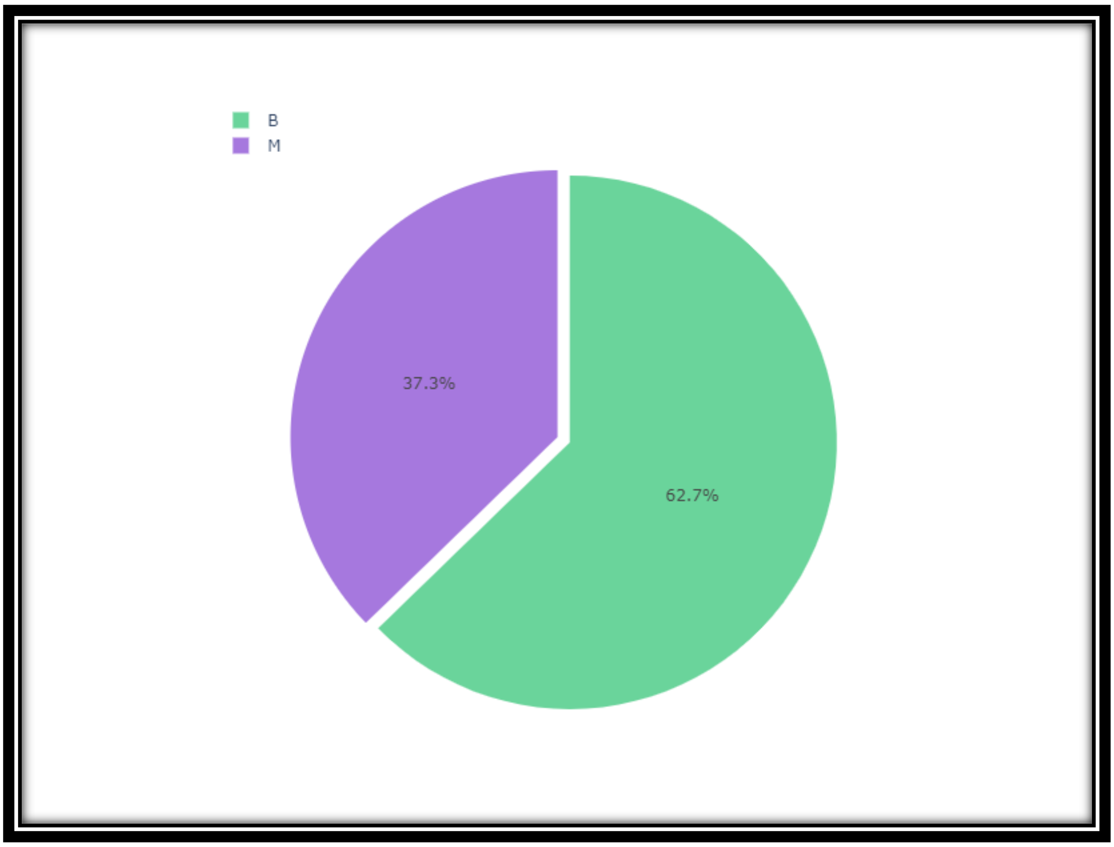


Fig-2

From the figure it can be observed that there “B” is 62% while “M” is almost 38%. We can see a kind of data imbalanced. Though in our analysis I have considered two cases:

Case-I: Where we have ignored the data imbalanced

Case-II: Where we have considered the data imbalanced and sampling is performed.

Missing Values:- After carefully analysing the dataset we have seen that it is perfect dataset which contains no null values. Hence our work became easy as no imputation has been performed here.

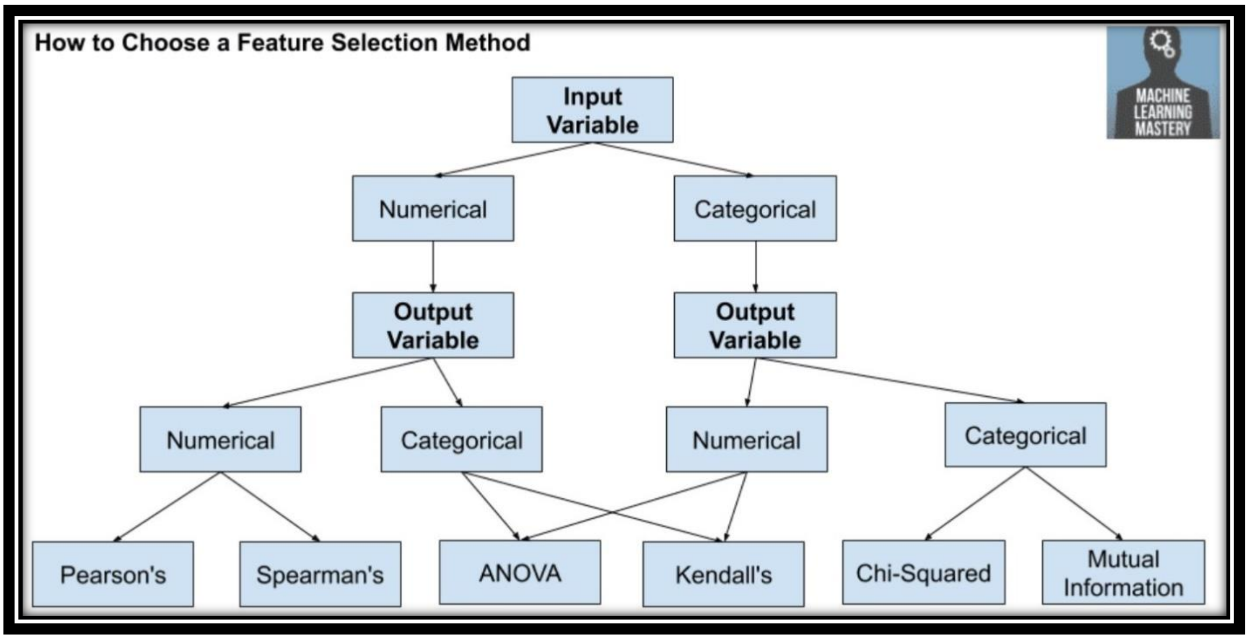
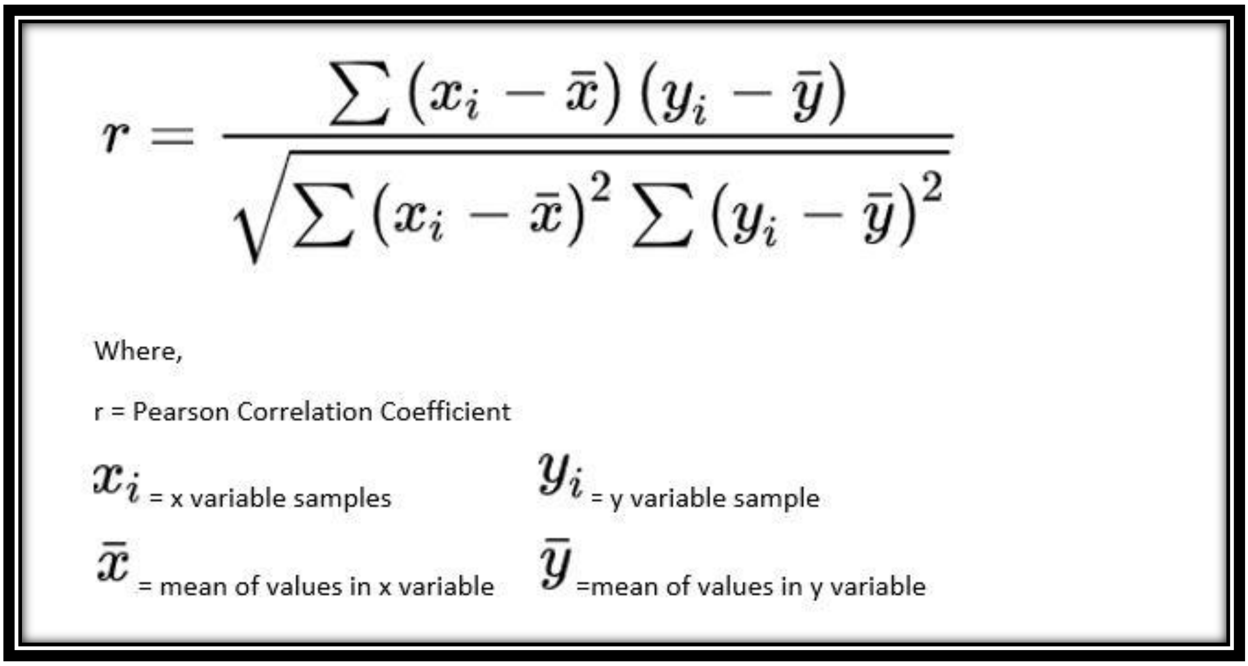


Fig-3

In our case we see that our input or explanatory variables are numerical and output or response variable is categorical one. Since in my case I have taken categorical as numerical and find out the

<u>Pearson's Correlation Coefficient:-</u> To find out the relationship between the response variable and each explanatory variable , I have calculated Pearson's coefficient. The formula for the correlation coefficient is given by



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\bar{y}$ =mean of values in y variable

The above is the formula for the correlation coefficient and it is denoted by “r”.

The value of “r” lies between -1 and 1.

- Correlation refers to the association between the observed values of two variables
- Correlation quantifies this association, often as a measure between the values -1 to 1 for perfectly negatively correlated and perfectly positively correlated. The calculated correlation is referred to as the “correlation coefficient.
- The correlation between two variables that each have a Gaussian distribution can be calculated using standard methods such as the Pearson’s correlation
- This procedure cannot be used for data that does not have a Gaussian distribution. Instead, rank correlation methods must be used.

Case-1: when *r*=-1, it shows that both the variables are negatively correlated. With the increase of the one variable other will decrease.

Case-2: when r=0, it shows that both the features are independent.

Case-3: When r=1, it shows that both the features are highly related.

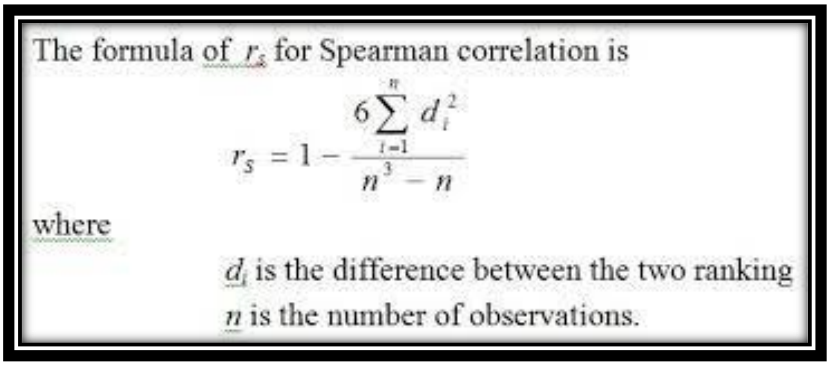
Now we have the required coefficient calculation of the all features related with response variables.

Features	Pearson Coefficient	p-value
id	0.039769	3.44E-01
diagnosis	1	0.00E+00
radius_mean	0.730029	8.47E-96
texture_mean	0.415185	4.06E-25
perimeter_mean	0.742636	8.44E-101
area_mean	0.708984	4.73E-88
smoothness_mean	0.35856	1.05E-18
compactness_mean	0.596534	3.94E-56
concavity_mean	0.69636	9.97E-84
concave points_mean	0.776614	7.10E-116
symmetry_mean	0.330499	5.73E-16
fractal_dimension_mean	-0.012838	7.60E-01
radius_se	0.567134	9.74E-50
texture_se	-0.008303	8.43E-01
perimeter_se	0.556141	1.65E-47

area_se	0.548236	5.90E-46
smoothness_se	-0.067016	1.10E-01
compactness_se	0.292999	9.98E-13
concavity_se	0.25373	8.26E-10
concave points_se	0.408042	3.07E-24
symmetry_se	-0.006522	8.77E-01
fractal_dimension_se	0.077972	6.31E-02
radius_worst	0.776454	8.48E-116
texture_worst	0.456903	1.08E-30
perimeter_worst	0.782914	5.77E-119
area_worst	0.733825	2.83E-97
smoothness_worst	0.421465	6.58E-26
compactness_worst	0.590998	7.07E-55
concavity_worst	0.65961	2.46E-72
concave points_worst	0.793566	1.97E-124
symmetry_worst	0.416294	2.95E-25
fractal_dimension_worst	0.323872	2.32E-15

Observation:- We can see that some of the features which are correlated or independent on the basis of Pearson’s correlation values.

Spearman’s Correlation Coefficient- it is same as with Pearson’s correlation coefficient.



The formula of r_s for Spearman correlation is

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

where

d_i is the difference between the two ranking
n is the number of observations.

We get almost the similar values of correlation as that we get in case of Pearson’s correlation coefficient.

Difference between Spearman and Pearson’s coefficient

- Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not)

Multi-colinearity: As we can see that there many features in the dataset i.e. there are 32 features including the response variable. There is high chance that there might be multicolinearlity among the features. To detect the multicolinearlity we have multiple options.

- Variance inflation factor (VIF) measures how much the behaviour (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables.
- 1 = not correlate.
- Between 1 and 5 = moderately correlate.
- Greater than 5 = highly correlate.
- VIF> 10 is used and remove the feature which has that high VIF.

feature	VIF
0 id	1.163246
1 radius_mean	63309.441555
2 texture_mean	251.432664
3 perimeter_mean	58123.587378
4 area_mean	1287.411538
5 smoothness_mean	393.514898
6 compactness_mean	201.166758
7 concavity_mean	157.899635
8 concave points_mean	154.297834
9 symmetry_mean	184.429153
10 fractal_dimension_mean	629.688286
11 radius_se	237.257123
12 texture_se	24.758627
13 perimeter_se	211.410744
14 area_se	73.436569
15 smoothness_se	26.330278
16 compactness_se	44.948568
17 concavity_se	33.254271
18 concave points_se	54.030035
19 symmetry_se	37.207715

20 fractal_dimension_se 27.549724
21 radius_worst 9677.820028
22 texture_worst 343.005975
23 perimeter_worst 4487.783848
24 area_worst 1139.047176
25 smoothness_worst 375.678745
26 compactness_worst 132.994862
27 concavity_worst 86.311570
28 concave points_worst 148.786616
29 symmetry_worst 219.018071
30 fractal_dimension_worst 423.466387

Observation: we can make some of the observations. Those are

- VIF of most of the features are more than 10.
- S it is not possible to remove the all the features.
- There is alternative way to remove the features. The higher VIF values will be removed and the remaining features will be used in modelled.

Alternative Method to remove features:- There are two methods that I have performed in order to remove some of the features which might not be

- In our case VIF is not working well. So I have chosen an alternative option. We have to perform some statistical analysis. We have to calculate the p-value for each feature. We remove those features which have higher p-value.
- Next is normality test is performed for each of the variables. Those features which do not follow the Gaussian distribution will be not used for the modelling. To perform the normality test, we can use QQ plot, distribution plot and histogram. Beside that some KS test and chi-square goodness of fit are also performed.

Q-Q plot

- Another popular plot for checking the distribution of a data sample is the quantile-quantile plot, Q-Q plot, or QQ plot for short.
- it is a technique to compare whether two sets of sample points are from or they follow same distributions.
- one distribution is know , we have to check for the other distribution

- If the unknown sample of dataset follow given distribution, we will have a scatter plot, where data points will be in a straight line y = x.
- he idea is to plot the quintile values of two distributions/samples and see
- if they make a straight line or not. If the quintiles of two sample sets are similar or in a better case, identical then sample set is from the same distribution. # The process of QQ plot
- Arrange the dataset in increasing order.
- Calculate the percentiles for each increasing dataset
- Plot those percentiles with the help of Scatter plot
- Done

Shapiro-Wilk Test

- Evaluates a data sample and quantifies how likely it is that the data was drawn from a Gaussian distribution
- The Shapiro-Wilk test is believed to be a reliable test of normality, although there is some suggestion that the test may be suitable for smaller samples of data, e.g. thousands of observations or fewer.
- The function returns both the W-statistic calculated by the test and the p-value.

Modelling

Dataset Spliting:- Dataset is containing the whole features. Before feeding into the model we will split the dataset in train and test. Train dataset is fed into the machine learning model and accuracy score is calculated. Test dataset is unseen data which is used for test performance.

Backward stepwise

- Backward stepwise selection (or backward elimination) is a variable selection method which:
- Begins with a model that contains all variables under consideration (called the Full Model)
- Then starts removing the least significant variables one after the other
- Until a pre-specified stopping rule is reached or until no variable is left in the model
- From Full model we eliminate the least important features.

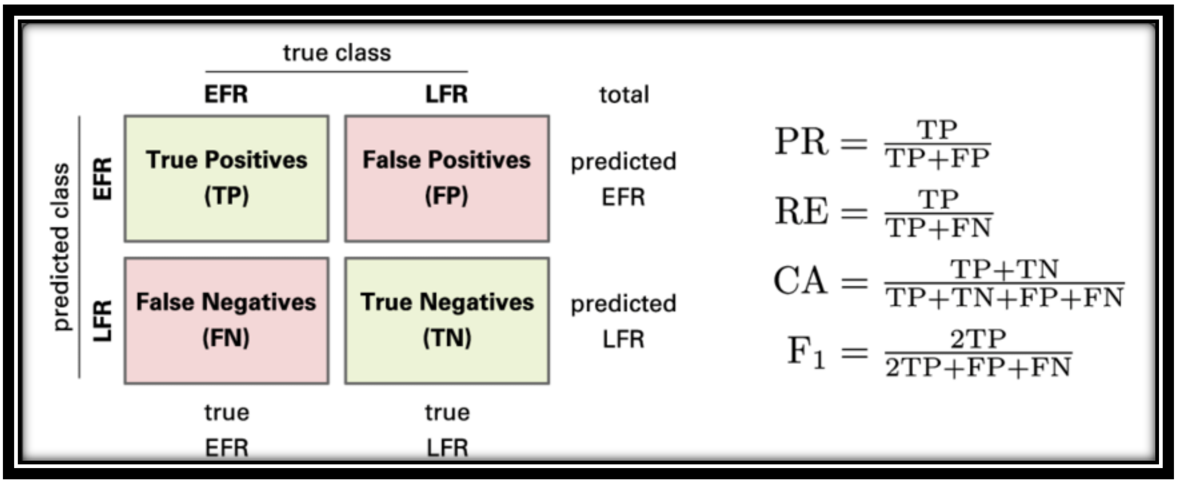
Determine the least significant variable to remove at each step

The least significant variable is a variable that:

- Has the highest p-value in the model, or
- Its elimination from the model causes the lowest drop in R2, or
- Its elimination from the model causes the lowest increase in RSS (Residuals Sum of Squares) compared to other predictors. # Choose a stopping rule The stopping rule is satisfied when all remaining variables in the model have a p-value smaller than some pre-specified threshold. When we reach this state, backward elimination will terminate and return the current step's model. # Where backward stepwise is better
- Starting with the full model has the advantage of considering the effects of all variables simultaneously.
- This is especially important in case of collinearity (when variables in a model are correlated which each other) because backward stepwise may be forced to keep them all in the model unlike forward selection where none of them might be entered
- Unless the number of candidate variables > sample size (or number of events), use a backward stepwise approach.

Metrics for Classification:- The metrics are used for obtaining the accuracy of the model is

- **Accuracy Score**- In classification accuracy is for a given points how many points are correctly classified. It is good metrics for the balanced class but it will not give good values for the case of imbalanced dataset.
- **Confusion Metrics**- It is tool to find out the accuracy of the model. It is matrix of the true class and predicted class.



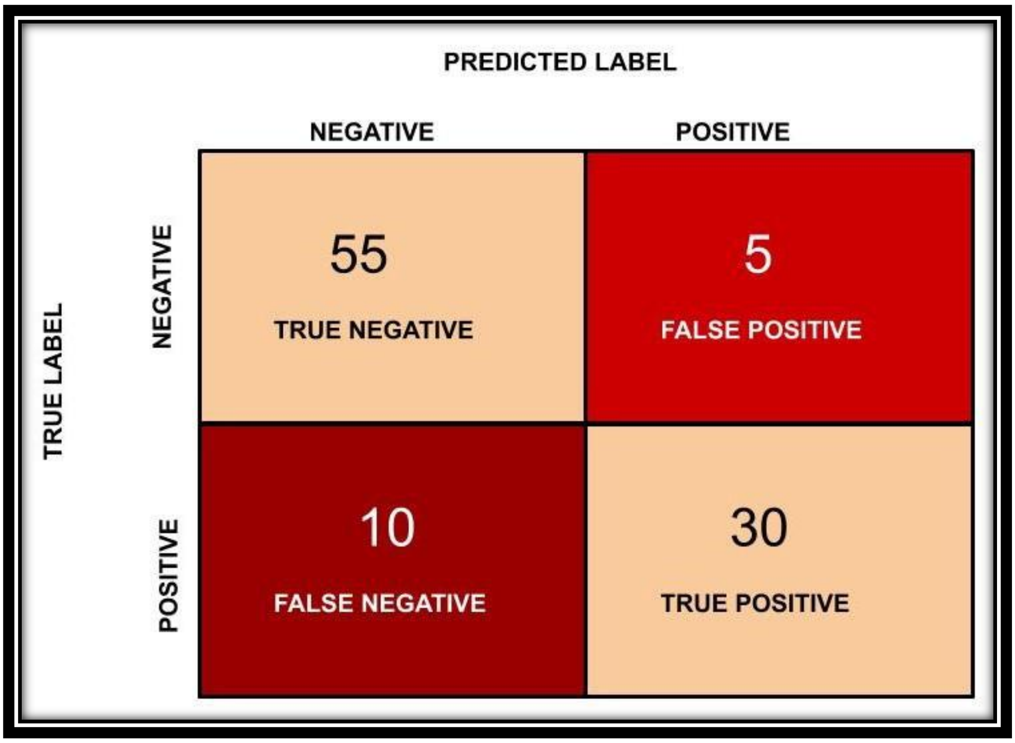
$$PR = \frac{TP}{TP+FP}$$

$$RE = \frac{TP}{TP+FN}$$

$$CA = \frac{TP+TN}{TP+TN+FP+FN}$$

$$F_1 = \frac{2TP}{2TP+FP+FN}$$

- **Precision**:- It is the ratio of True Positive/(True Positive + False Positive). It is one of the good metrics for imbalanced dataset. As we want the Precision to be 1 i.e. when FP=0.
- **Recall**:- Ratio of True Positive/(True Positive + False negative) . It is one of the good metrics for imbalanced dataset. As we want the Recall to be 1 i.e. when FN=0. Since in our case we want recall to be high. We want to reduce the Type-I error
- **F-1 score**- It is the harmonic mean of Precision and Recall.
- **ROC-AUC score:** It is stand for Receiver Operating Characteristics curve. Drawing the curve between the True Positive Rate and False Positive rate. For each threshold value of the probabilities, we get different value of FPR and TPR; according to this we plot the curve.
- **Precision Recall curve:** It is plot between the precision and recall. It is good measure for the imbalanced dataset. A precision-recall curve is a plot of the precision (y-axis) and the recall (x-axis) for different thresholds, much like the ROC curve. Dumb model gives an area of equal to 0.5.



PREDICTED LABEL

TRUE LABEL		NEGATIVE	POSITIVE
	NEGATIVE	55 TRUE NEGATIVE	5 FALSE POSITIVE
	POSITIVE	10 FALSE NEGATIVE	30 TRUE POSITIVE

Logistic Regression: It is basically a classification algorithm used to classify the binary classification mainly. Since the algorithm is developed in such a way that the outcome will predict the class on the basis of the probability. Since the class are classified on the basis of the threshold probability.

The formula is like log of odd's ratio is linearly related. This is function is called as Sigmoid Function. We know the formula.

Loss Function: The logistic loss function is calculated as

$$w *= argmin \sum \ln(1 + e^{-y*w^{tx}})$$

Where $y * w^{tx}$ is the sign distance from the plane which is optimum.

Regularization: It is a penalty which is used to reduce the loss function and helps to reduce the over fitting. As we know that sometime the machine learning model which perform very well on training dataset but on test dataset those models become over fitting. Over fitting is due to the consideration of the noise dataset. This is why we use regularization technique. As it helps to reduce the over fitting of the model.

$$w *= argmin \sum \ln\left(1 + e^{-y*w^{tx}}\right) + Regularization$$

Two type of regularization are there

Ridge Regression: Ridge regression is one of the types of linear regression in which we introduce a small amount of bias, known as Ridge regression penalty so that we can get better long-term predictions.

- In Statistics, it is known as the L-2 norm.
- In this technique, the cost function is altered by adding the penalty term (shrinkage term),
- which multiplies the lambda with the squared weight of each individual feature
- Therefore, the optimization function(cost function) becomes:
- $w *= argmin \sum \ln\left(1 + e^{-y*w^{tx}}\right) + \backslash labda \sum \beta^2$

Usage of Ridge Regression:

- When we have the independent variables which are having high collinearity (problem of multicollinearity) between them, at that time general linear or polynomial regression will fail so to solve such problems, Ridge regression can be used.

- If we have more parameters than the samples, then Ridge regression helps to solve the problems.

Limitation of Ridge Regression

- Not helps in Feature Selection: It decreases the complexity of a model but does not reduce the number of independent variables since it never leads to a coefficient being zero rather only minimizes it
- **Model Interpretability:** Its disadvantage is model interpretability since it will shrink the coefficients for least important predictors, very close to zero but it will never make them exactly zero. In other words, the final model will include all the independent variables, also known as predictors.

Lasso Regression:

- Lasso regression is another variant of the regularization technique used to reduce the complexity of the model. It stands for **Least Absolute and Selection Operator**.
- It is similar to the Ridge Regression except that the penalty term includes the absolute weights instead of a square of weights.
    - $w *= argmin \sum \ln\left(1 + e^{-y*w^{tx}}\right) + \backslash labda\ \sum \beta$
- In statistics, it is known as the **L-1 norm**.
- In this technique, the L1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero which means there is a complete removal of some of the features for model evaluation when the tuning parameter λ is sufficiently large.
- Therefore, the lasso method also performs **Feature selection** and is said to yield **sparse models**.

Limitation of Lasso Regression:

- **Problems with some types of Dataset:** If the number of predictors is greater than the number of data points, Lasso will pick at most n predictors as non-zero, even if all predictors are relevant.
- **Multicollinearity Problem:** If there are two or more highly collinear variables then LASSO regression selects one of them randomly which is not good for the interpretation of our model?

Key Differences between Ridge and Lasso Regression

- In Ridge Regression only overfit is overcome but in Lasso regression both Overfit and feature selection problem are resolved.
- Lasso Regression tends to make coefficients to absolute zero whereas Ridge regression never sets the value of coefficient to absolute zero.

Important points about λ:

- λ is the tuning parameter used in regularization that decides how much we want to penalize the flexibility of our model i.e, **controls the impact on bias and variance**.

  - λ value incrasses , variance decreases and hence overfiting is avaoided . But after further increase in value causes, bian to increase and undrefitting happens.
  - When **λ = 0**, the penalty term has no effect, the equation becomes the cost function of the linear regression model. Hence, for the minimum value of λ i.e, λ=0, the model will resemble the linear regression model. So, the estimates produced by ridge regression will be equal to least squares
  - However, as **λ**→∞ (tends to infinity), the impact of the shrinkage penalty increases, and the ridge regression coefficient estimates will approach zero.