## Breast Cancer Prediction

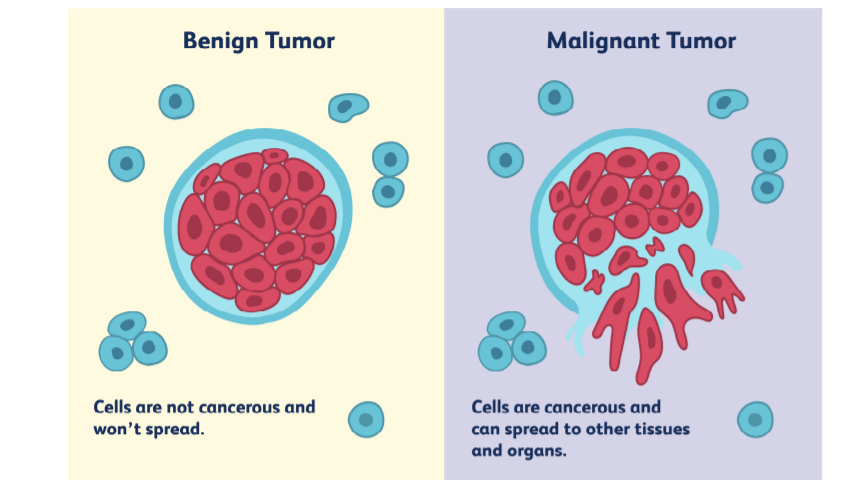
## I. Introduction

A number of recent cases of cancer and cancer-related fatalities happen because of breast cancer, which is the most prevalent disease among women globally. Breast cancer, therefore, represents a serious public health issue in today's society. Early detection of breast cancer may vastly enhance prognosis and mortality rate since prompt therapy can be provided to victims earlier. Appropriate tumour categorization may help avoid needless therapies. Breast cancer and classifying people as malignant or benign requires a lot of studies, so getting this right is critical. Machine learning is generally regarded as the approach of preference in breast cancer pattern categorization and prediction modelling because of its key features detection benefits. Uncontrolled tissue growth that outpaces and is disorderly with normal tissue growth results in a neoplasm. This aberrant mass of tissue continues to grow uncontrollably even after the stimulus that caused it to change has been removed. Cancer can begin practically anywhere in the 37.200 billion cells that make up the human body. Some cancers can break out from these tumours and move to different parts of the body as they grow. Tumours can spread through the blood or lymphatic system and develop in new locations far from the original. benign tumours are not cancerous.

It is impossible for tumours to metastasize to other parts of the body or to infiltrate surrounding tissues. Breast cancer is a disease in which a tumour develops in the breast. occurs in the breast. Because it affects both women and men equally, breast cancer is one of the most frequent cancers.

It was first discovered in 1975, and each year, over 411,000 people die as a result of it. Statistically, it is expected that the global incidence of In 2030, there will be an additional 23,6 million cancer cases, or a 68 percent rise. a greater number of cases than in 2012

Machine Learning classification techniques are a quick and efficient approach to organize and categorize data. In particular, since such techniques are extensively utilized in diagnosis and monitoring to reach conclusions, the medical profession sees them as very valuable. In this project, we aim to use traditional Machine Learning methods to classify patients affected with malignant breast cancer with the help of features that were given in the dataset.



## Dataset

The dataset is available on Kaggle and UCI Machine Learning Repository and is free to use for the purpose of research and education. The dataset consists of ten real-valued features that are computer for the nucleus.

- Radius: a feature that describes the mean of distances from centre to points on the perimeter

- Texture: a feature that describes the standard deviation of grey-scale values.

-Perimeter : This is a integer variable which describes about the perimeter of the cells

- Area : This feature describes about the area of the circle.

- Smoothness: This feature describes the local variation in radius lengths.

- Compactness: perimeter^2 / area - 1

- Concavity: This feature describes the severity of concave portions of the contour.

- Concave points: This feature describes the number of concave portions of the contour

- Symmetry: This feature describes about the symmetry

- Fractal dimension: This features is the coastline approximation – 1

## 1.2 Project Aim :

An important goal is to develop machine learning models that can distinguish between benign and malignant cancer cells in women. For a more thorough analysis, data will be altered and shrunk in order to identify patterns. As already said, the most suitable model will be chosen after AUC-ROC Scores of all elements to consider. These metrics will be defined at a later time. We are capable of doing this. To classify cancer cell nuclei images, researchers used a machine learning method for extracting key elements from the images. It would have that effect. This information can be useful for differentiating between benign (B) and malignant (M) samples. Using machine learning models, we hope to construct a classifier that can help identify superior precision with a minimal incidence of false negatives (high sensitivity).

## I.3. Objectives

* Use the data visualization techniques to visualize the data and draw interpretations from the data.
* Use the different Machine Learning models to train on the data and compare the results.
* Classify the breast cancer with high accuracy as any corruption in the result may lead to threat to patients.

## I.4. Research Questions

## A research topic is a topic that a study tries to find answers for. This refers to something in the study that's addressed by examining the data and telling the storey it reveals. The research objective is typically formulated in a way that highlights a variety of aspects, such as the study's research population and factors, and also the study's main purpose. Research is commonly centred around scientific research. It's not surprising that researchers frequently revisit and revamp their research questions: Research questions tend to be evolving rather than static. Researchers must reassess and adjust questions as they conduct literatures and build a framework for the study. The research questions that we are trying to address in this project are:

* How does the use of different parameter values will affect the performance of the models in detecting the breast cancer among patients?
* What features has high contribution in predicting the breast cancer among patients?
* How does different classification model behave in terms of performance on this data and what will be the most efficient model among the machine learning models used in the project?

## 1.5. Tools

There are several tools employed as phase of this project to achieve its main objective. a few essential tools utilized during this Endeavour include:

NumPy: NumPy is a Library in python designed to help with array management. It has features for linear algebra, Fourier transforms, and matrices, as well.

Pandas: Pandas is a Python-based library that works with data analysis and manipulation. The focus is on the operations and data structures required to manipulate tables and time - series data.

Scikit-learn: Scikit-learn is a library for Python that helps with machine learning. It is open source and available to anyone. It includes several algorithms, such as support vector machines, among others.

Matplotlib: Matplotlib is a tremendous 2D plotting library in Python, perfect for visualizing array data. Matplotlib is a library built on the concept of NumPy arrays, and it is made to work with the other components of the Scipy stack. Matplotlib has numerous plots, including lines, bars, scatter plots, histograms, and more.

Seaborn: Seaborn is an example of a Python library that works with the matplotlib data visualization framework and integrates with pandas data structures. Seaborn is Seaborn's central visualization system, which is crucial in helping the exploration of data. See how the distribution is univariate and bivariate.

## 2. Ethical, Legal, Professional and Social issues

It is highly important to consider the ethical, legal, and social issues in research. This is to identify the threats and avoid these threats at the initial stages of the project such that they will not cause any serious harm to the research.

## 2.1. Ethical Issues

According to the Ethical OS Toolkit, the possible issues that are involved in this project are as follows:

Risk Zone 7 Implicit Trust & User Understanding –

The data used in the research/project contains the health-related information of the individuals that can be misused by the companies for their benefit. The data needs to be protected and should not be used without the acknowledgment of the individuals for any purposes since this can violate ethical laws.

Risk Zone 4 Machine Ethics & Algorithmic Biases –

Any bias observed in the model that might be because of various reasons such as overfitting, or underfitting, etc., should be identified in the initial stages before sending the algorithm into production since the bias behaviour of the algorithms can misinterpret the classifications and misclassify the patients and cause serious harm to their lives.

Risk Zone 8 Hateful & Criminal Actors –

Since the data belongs to the individual health, this data can be misused by the hackers for their own benefit by selling the data on the dark web. This can be dangerous for the participants. Hence this data needs to be protected with all the latest firewalls and with the help of cybersecurity teams.

## 2.2. Social Issues

Social Issues deals with the problems that might occur in the society and cause harm to the society as a result of the research. Such issues can be considered as social issues.

The important factor to consider is the impact of a decision made by less accurate algorithms on the mental health of individuals. Any wrong decisions taken by the algorithms can affect the patient’s mental health and can cause serious harm to the individual. It is highly important to make sure the performance of the algorithms is high in the medical scenarios.

## 2.3. Legal Issues

In the similar case of an algorithm with low accuracy making mistakes and wrongly predicting the cases might end up in legal challenges. The individuals might files lawsuits against the organizations practicing these machine learning algorithms in the process of treatment.

Hence it is highly important to be careful and build robust algorithms and not to completely rely on the algorithms even at the times of production.

## 3. Methodology

## 3.1. Installing set-up

I have used Python 3.7, which I downloaded from the official Python website, for my project. I installed Python on my machine by setting it up on my hard drive and adding Python to my path. Additionally, I've installed Anaconda on my computer to get everything ready for running Python via Jupyter. After that, I was able to execute this project by using command prompt to install a few libraries.

Example : Pandas, Numpy, Scikit-learn etc.

## 2.2.Data Exploration

The data for this project is available at Kaggle [- https://www.kaggle.com/uciml/breast-cancer-wisconsin-data](-%20https:/www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes). The dataset includes 569 records with 33 variables combination data kinds make up the information.

The dataset is further explored to identify Exploratory data Analysis.

## 2.3 Exploratory Data Visualization

Due to the greater number of useful features in the data, the data gives us a high score of exploratory data analysis on the data. Data Analysis will help us understand the patterns in each feature with respect to the dependent variable and helps us understand how each feature is contributing to the rainfall prediction.

It is planned to perform a univariate analysis on each of the important features in the data. Using Data Visualizations, we can interpret the data much more efficiently.

In Exploratory data Analysis we will perform the visualization part and we will gain insights from the data. Exploratory Data Visualization (EDA) is very important part of the data science pipeline or any data science project. Exploratory Data Analysis is a vital method that involves conducting initial investigations on data in order to identify trends, identify discrepancies, evaluate assumptions, and verify conclusions using summary statistics and data visualizations. Exploratory Data Analysis (EDA) is a computational data analysis technique focused on John Tukey's pioneering work. EDA offers a basis for a wide variety of data analytic activities and addresses the diverse types of data and architecture encountered by applied researchers. EDA's fundamental conceptual and computational tools include the use of graphics and interactive data visualization, a focus on model creation, diagnosis, and evaluation, addressing fundamental measurement issues associated with various distributions.

Although these methods serve as a foundation for all research, EDA places a high value on data - based learning from data to enhance standard hypothesis testing procedures that may neglect critical unanticipated aspects of data and their effect on modelling and estimation. The EDA, it is claimed, is critical both in the early stages of science, where hypotheses and model development must be well-informed.

## 2.3.1 Univariate Analysis

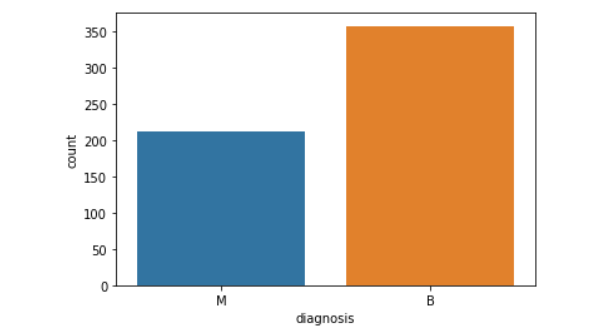
The technique of univariate analysis is for comparing and analyzing the relationship between a single feature and response variable. The prefix "uni" highlights the analysis only covering a single variable and its impact on a parameter.

For Example, the study can focus on a variable such as "gender," "height," or "weight."

However, only one variable is examined each time.

## 2.3.1.1 diagnosis Variable

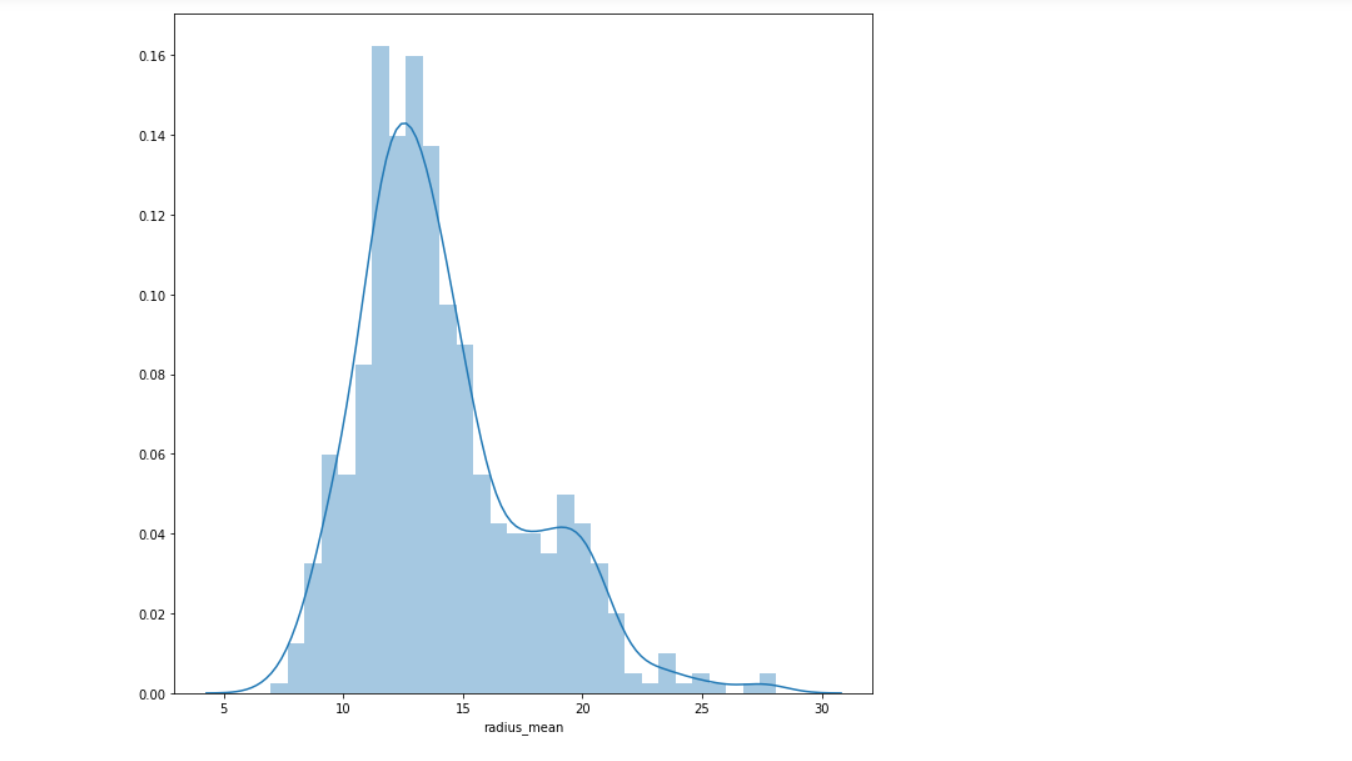
Diagnosis Variable is a Target Variable. Diagnosis is a dependent variable in our dataset which is categorical feature. There are two categories M (Malignant) and B (Benign). For counting the number of categories of Malignant and Benign i have plot the Countplot.



Diagnosis Variable consist of 569 records in which there are around 350 Benign and more than 200 Malignant.

## 2.3.1.2 Radius\_ Mean Variable

Radius mean is a integer variable which describes about the mean of distances from center to points on the perimeter.



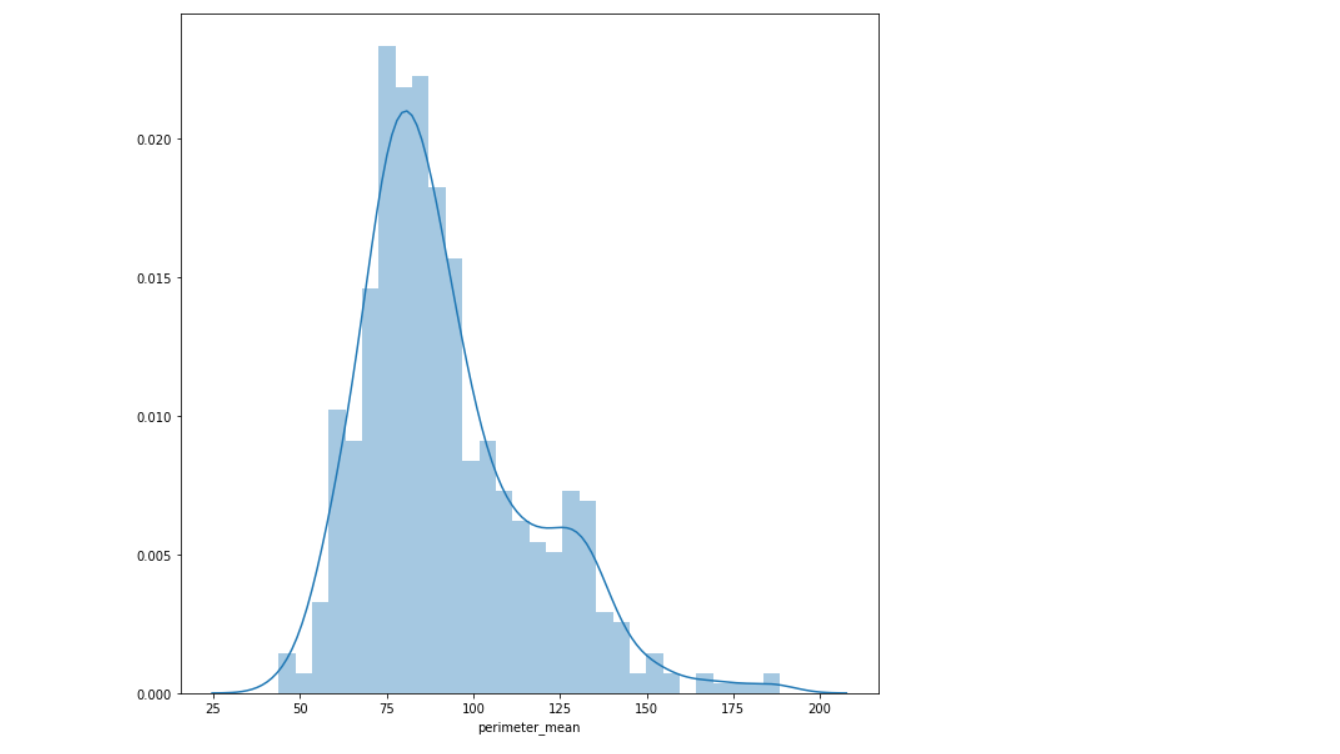
We can see that the graph is normally distributed there is no outliers in this feature. The stats of this is variable is :

|  |  |
| --- | --- |
| Count | 569.000 |
| Mean | 14.127 |
| Std | 3.524 |
| min | 6.98200 |
| 25% | 11.7000 |
| 50% | 13.3700 |
| 75% | 15.7800 |
| max | 28.1100 |

The minimum value of this column is 6.98 and the maximum value is 28.1100.

## 2.3.1.3 Perimeter mean

Perimeter mean is a integer variable which describes about the perimeter.



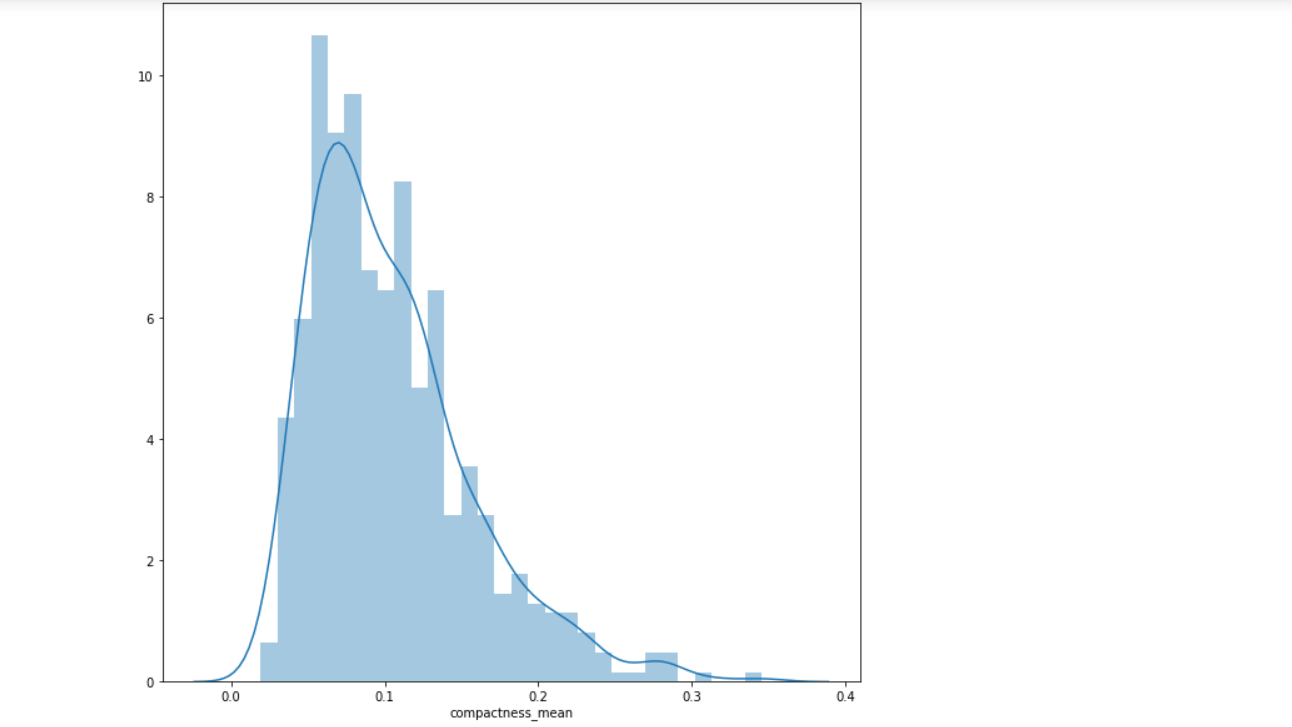
We can see that the graph is normally distributed there is no outliers in this feature. The stats of this is variable is :

|  |  |
| --- | --- |
| Count | 569.000 |
| Mean | 91.96 |
| Std | 24.29 |
| min | 43.79 |
| 25% | 75.17 |
| 50% | 86.24 |
| 75% | 104.1000 |
| max | 188.5000 |

The minimum value of this column is 43.79 and the maximum value is 188.500.

## 2.3.1.3 Compactness mean

Compactness mean is a integer variable which describes about the dimensionless number is at a minimum with a circular disk and increases with the irregularity of the boundary, but this measure also increases for elongated cell nuclei, which is not indicative of malignancy.



|  |  |
| --- | --- |
| Count | 569.000 |
| Mean | 0.104 |
| Std | 0.052 |
| min | 0.019 |
| 25% | 0.064 |
| 50% | 0.092 |
| 75% | 0.130 |
| max | 0.345 |

## 3.4 Data Cleaning

Data cleaning is a process that takes in flawed, inconsistent, irrelevant, overlapping, or flawed data, and transforms it into clean data, which is ready for a team to run numbers on. Data that is irrelevant or potentially problematic for research purposes tends to slow down the work and produce bad results. There are several different approaches to cleaning the data depending on the answers desired and the storage method being used. Data cleaning isn't just about removing information; it's about being able to remove information without destroying a dataset's value and finally, the ultimate objective of data cleaning is to provide a standardised and uniform data set to make it easy for data analytics and data dashboards to get and use the appropriate data each time a query is made.

Data cleansing, unlike data removal, is used to improve data by eliminating errors, standardising data sets, and cleaning up extraneous data points.

**Steps involved in Data Cleaning:**

3.4.1.Handling missing data

This is the main step of data cleaning . there are no missing values in our dataset. So there is no need to apply this step.

## 3.4.2. Encoding of Categorical Features

A category variable does have two or more categories. Categorical variables can belong to either nominal type or ordinal type. Categories of a nominal variable do not have any built-in ranking. Gender is often a categorical data that is simply composed of two categories with no natural order. The ordinal variable exhibits a well-defined order.

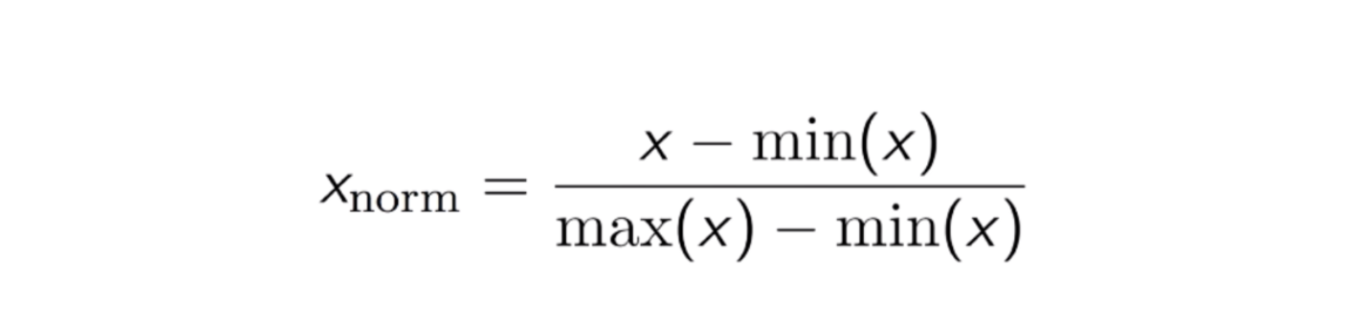
Many machine learning algorithms can't understand the categorical data on their own. On the other hand, Decision trees can learn straight from the data itself. They require all parameters to be numeric because of this the categorical data must be transformed into numbers.

**Label Encoding**: In Label Encoding simply we will give the numbers to each category.

In this project I have done Label encoding for converting the categorical features into numerical features. In this project I have used label encoding for Diagnosis variable.

## 3.4.3 Feature Scaling

Feature Scaling is also very important step of Data Cleaning if our feature lies on different –different scale. It is better to apply feature Scaling. A common practice when preparing data for machine learning is normalization. To normalize, it helps to keep everything on the same scale, to keep different value ranges constant. To make the best use of machine learning, datasets do not have to be normalized.



In this project I have applied feature Scaling on all the features except dependent variable (diagnosis). It will convert all the features in between 0 to 1 range.

## Machine Learning

Machine learning can be addressed as a subset of artificial intelligence that allows machines to learn and develop automatically without any supervision of the humans. Machine learning is concerned with the development of computer programmes capable of having access to data and use it to train themselves. The learning phase starts with insights or facts, such as instructions to search for correlations in data and make more informed potential judgments based on the examples provided. The aim of the machine learning is to allow computers to train on the data automatically and change their activities as per the requirements without any human involvement or assistance.

Machine learning allows the processing of enormous amounts of data. Although it usually produces more reliable and timely outcomes for identifying lucrative opportunities or risky threats, correctly training it may take additional time and effort. Combining machine learning and artificial intelligence with variety of technologies have the potential to render it much more efficient at analysing vast amounts of data. Within the realm of machine learning, there are several algorithms that are published every day and they are organised based on whether they focus on supervised or unsupervised learning or feature similarities. Machine learning algorithms can be classified as two types. Namely, supervised machine learning algorithms and unsupervised machine learning algorithms.

## 3.5.1 Feature Importance in Machine Learning Algorithms:

When we talk about feature importance, we're talking about methods that rate the usefulness of input information in terms of predicting a target variable. Although notable examples include rank correlation scores, coefficients derived in linear regression, decision trees, and permutations importance scores, feature significance scores can take many different forms and be derived from many different places. Ratings on feature importance are critical in predictive modelling projects, since these scores provide information about both the problem and its solution. They also serve as the basis for dimension reduction and feature selection, both of which increase the predictive model's performance.

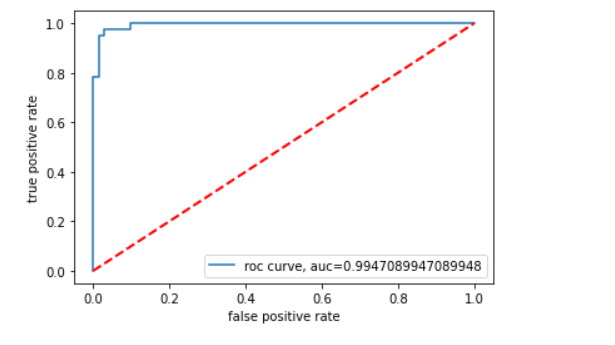
## 3.5.1.1 Logistic Regression

Logistic Regression is a predictive model technique where, using pattern recognition, a previously recorded dataset is used to deduce the future likelihood of a certain event. One way to use the Logistic Regression algorithm is to find the independent variable(s) and dependent variable that you want to forecast, and then use the regression to make predictions. In the presence of several explanatory variables, logistic regression is employed to obtain the odds ratio. Binomial regression has similarities to other linear regression models, except the dependent variables is binary. The effect of each variable is calculated to show the effect on the odd values of the observed event.

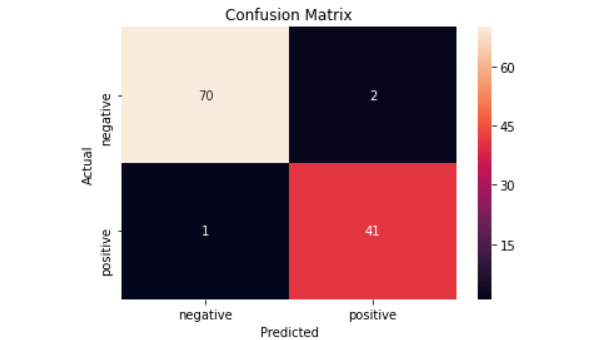
## Hyper Parameter Tuning in Logistic Regression

I have applied hyper parameter tuning in Logistic Regression Algorithm. Basically hyper parameter tuning is used to find the hyper parameters which gives higher accuracy and also reduce the chances of overfitting. For C=1000 the Training AUC Score is 99.41% which is highest among all and for C=1 the AUC Score on testing data is 97.42%.

The ROC-AUC Curve between True positive and false positive rate is:



I have also plot the confusion matrix for the dataset.

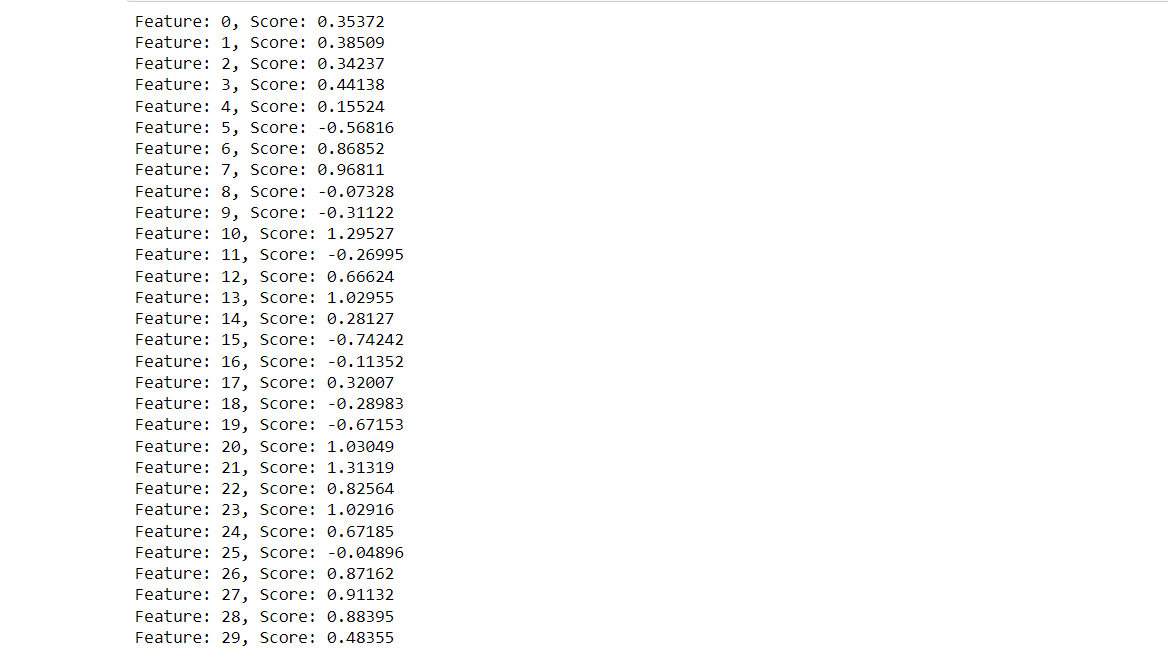


As we can see that The false positive values in Confusion matrix are 2 and there is only 1 false negative.

## 3.5.1.2.Feature Importance in Logistic Regression

After training and testing the model I have applied feature importance for logistic regression.

The feature importance for all the features are :



We can derive these conclusions from the feature importance of logistic regression

* For Some Algorithms Standarization affect the features importance of the dataset
* As in above graph we can see that there are few features whose values were either too negative or positive. Rest of the features had almost 0
* After standarization we can see that a lot of the features whose coefficeint values are much better.
* Here negative values indicate that it tries to push the model towards the negative side
* Same case with the positive value which tends to push the model in positive side.

## 3.5.2. Xg- Boost

XGBoost is a variation of the ensemble learning technique. The findings of a single machine learning model may not be enough to rely on sometimes. Ensemble learning is a solution to combining the predictive potential of several learners, and it is more orderly than many other alternatives. The end result is a model that puts together all of the results from multiple models.

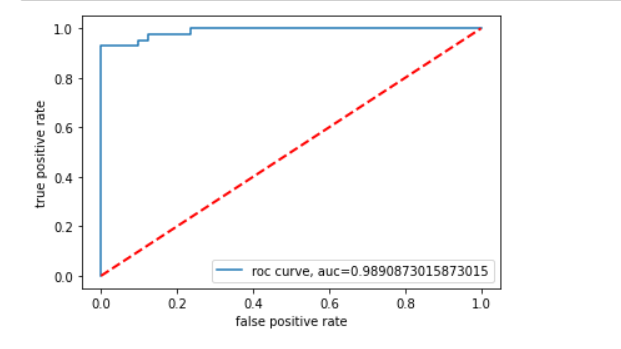
Base learners are generally split between multiple algorithms, though occasionally the same method will be replicated for simplicity. Two kinds of ensemble learners with a lot of traction are boosting and bagging. Decision trees are the most common statistical methodology to which these two methods have been applied.

## Hyperparameter Tuning in Xg-Boost Classifier

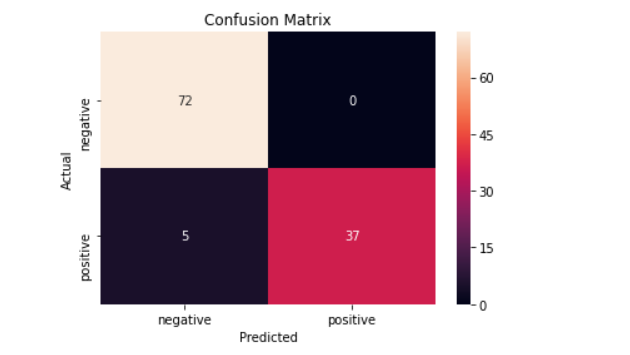
I have Applied Hyperparameter tuning on Xg-boost classifier for finding the best parameters. After Applying Hyperparameter tuning we can see that the best parameters are:

Base learner=500, Optimal depth=1

For these parameters the AUC on training data is 100% and for testing data the AUC Score is 94.04%. hence we can say that there is no overfitting in the model.



The AUC score between true positive rate and false positive rate is 0.9890. The Confusion Matrix for Xg boost algorithm is :

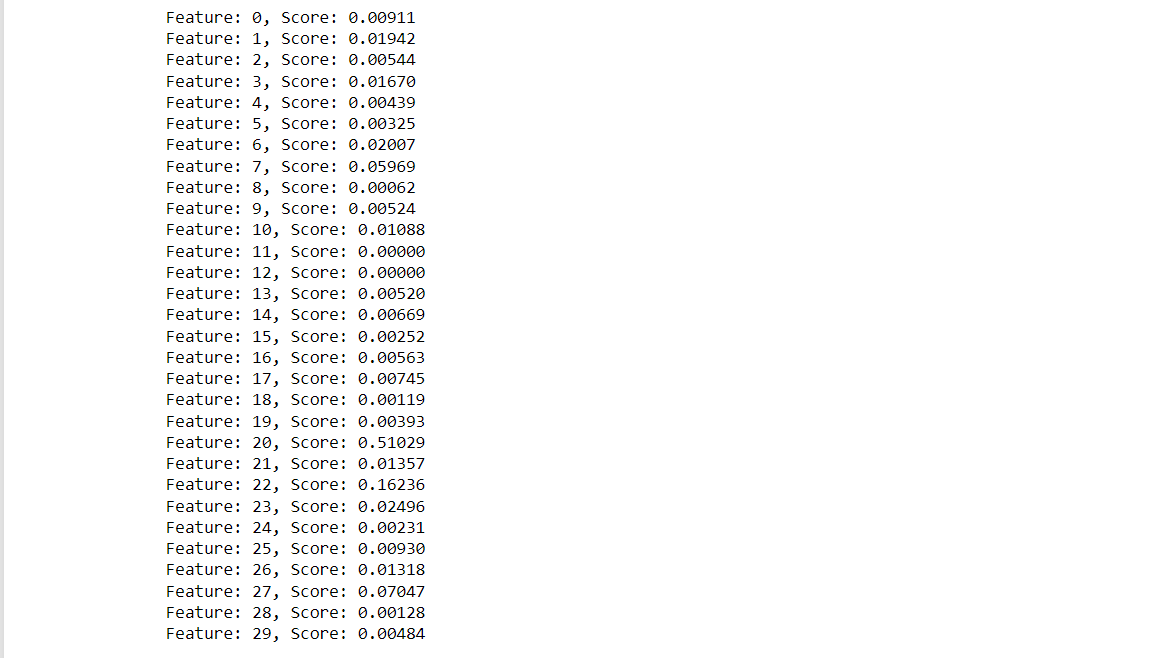


As we can see that The false positive values in Confusion matrix are 0 and there is only 5 false negative.

## 3.5.2.1 Feature Importance for Xg-Boost

After training and testing the data I have applied feature importance for Xg-Boost. These are the conclusions

The feature importance scores are:



* The most important features is Radius Worst
* As we can see that Concave points worst is 2nd important features.

## 3.5.3. Decision Tree

Decisions trees are the most popular and efficient classifiers and predictors available. Decisions are created by dividing a group of information into "decision branches," which separate results into the appropriate categories based on attribute tests.

These are the following benefits of Decision Tree:

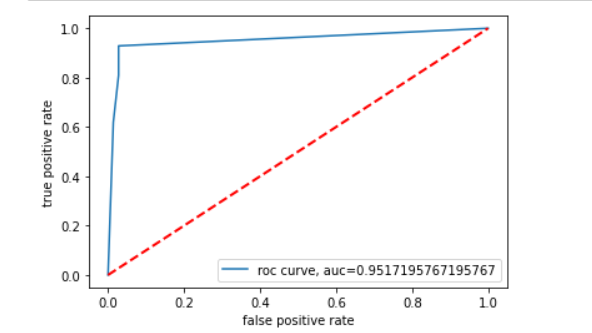
* The final outcome of a decision tree is usually well defined.
* A decision tree does not require significant processing for classification.
* For both continuous and categorical, decision trees can be used.
* Decision trees show the which variables to use when predicting or classifying.

## Hyperparameter Tuning in Decision Tree

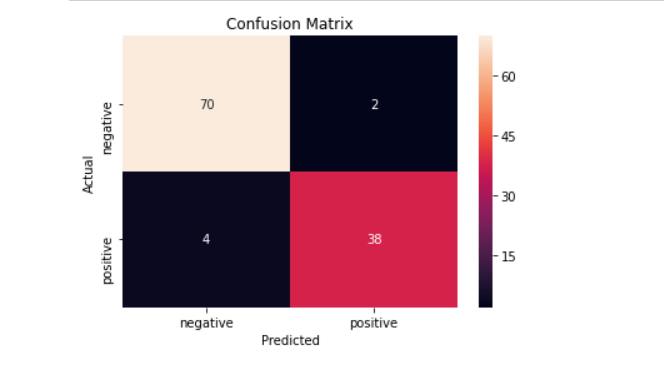
I have applied Hyperparameter tuning on Decision Tree. After Appling Hyperparameter tuning we can see that

{'criterion': 'gini', 'max\_depth': 5, 'min\_samples\_leaf': 5, 'min\_samples\_split': 4}

These are the best parameters. For Training data the algorithm gives 98.94% AUC value and with testing data the algorithm gives 91.46.



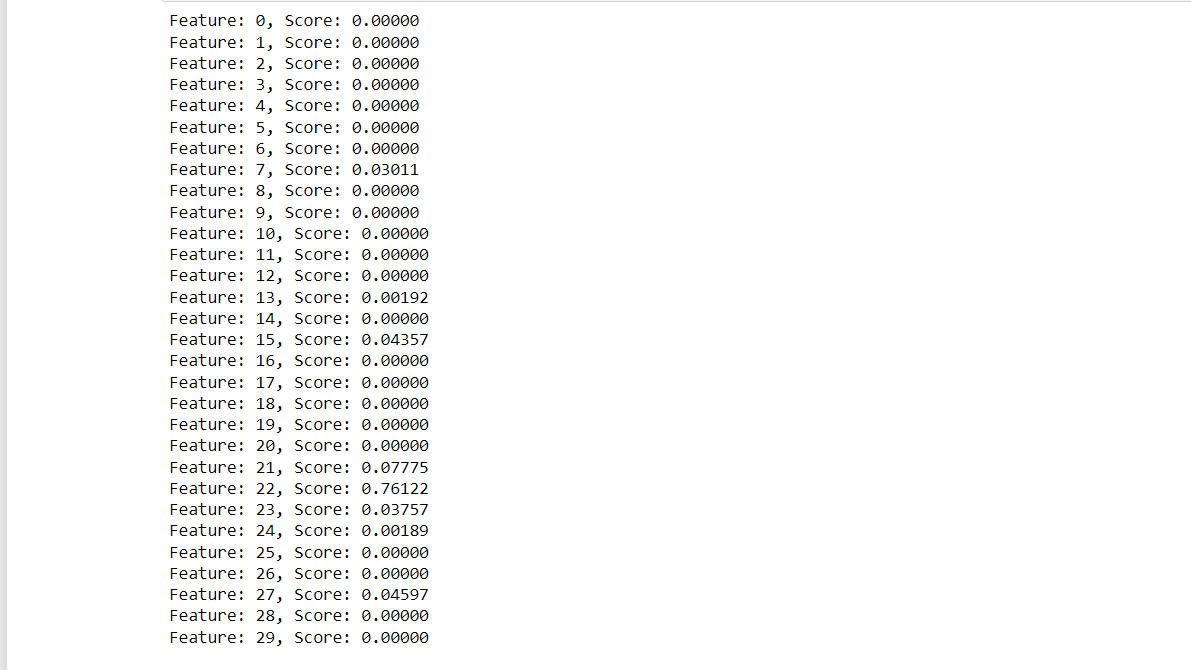
The AUC score between true positive rate and false positive rate is 0.9512 and the confusion matrix for Decision Tree is :



## 3.5.3.1. Feature Importance in Decision Tree :

After training and testing the data I have applied feature importance for Xg-Boost. These are the conclusions

The feature importance scores are:



* The most important features are "radius\_worst"
* The list of the important features are --Radius\_worst, texture\_worst, concave point worst, texture mean, concavity features, etc
* Those features whose value is almost equal to 0 , are least important features.
* Least features can be removed form the dataset and the dataset will be used for prediction.

## 3.5.4. K-Nearest Neighbors

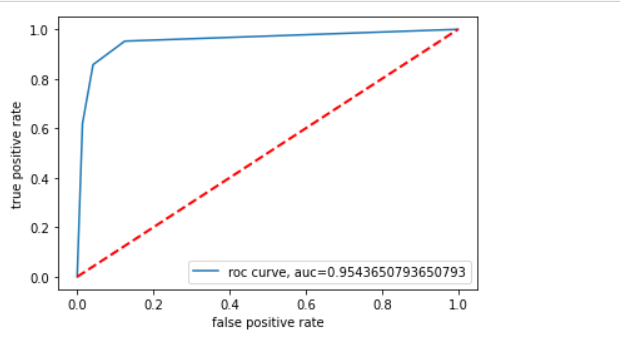
The Supervised Learning algorithm K-Nearest Neighbor is one of the easiest algorithms, with no prior assumptions on the properties of data. The K-NN method groups the new data/case based on resemblance to other, existing categories, assigning the new case to the category most similar to the current accessible cases. KNN method uses data from all the previous data points to compare similarities and classify new data points. The data can be categorized quickly with the help of K-NN, since the new information is easy to understand.

K-NN is best for Classification problems but may also be useful for Regression. Non-parametric algorithms like K-NN assume nothing about their data. It is sometimes referred to as a lazy learner algorithm since it takes a long time to learn from the training set, instead of doing so instantly. This method conducts an action on the dataset at the time of classification, rather than learning from of the training set immediately. When getting fresh data, KNN algorithms only classify it into a category with a similarity to the new data.

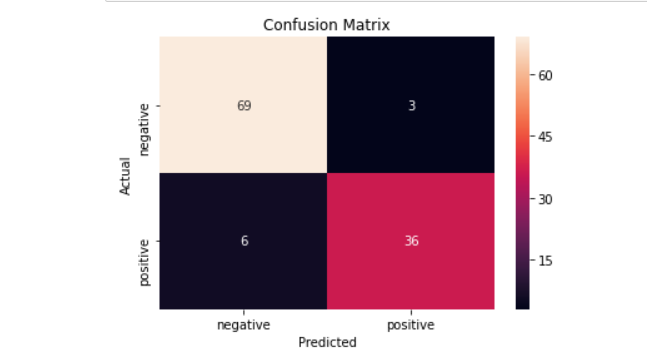
## Hyperparameter Tuning on KNN Algorithm

I have applied hyper parameter tuning in K Nearest Neighbors Algorithm. Basically hyper parameter tuning is used to find the hyper parameters which gives higher accuracy and also reduce the chances of overfitting.

For n\_neighbors=3 the algorithm gives higher AUC value with training and testing data. The AUC with the training data is 94.12% and the accuracy with the testing data is 90.77%

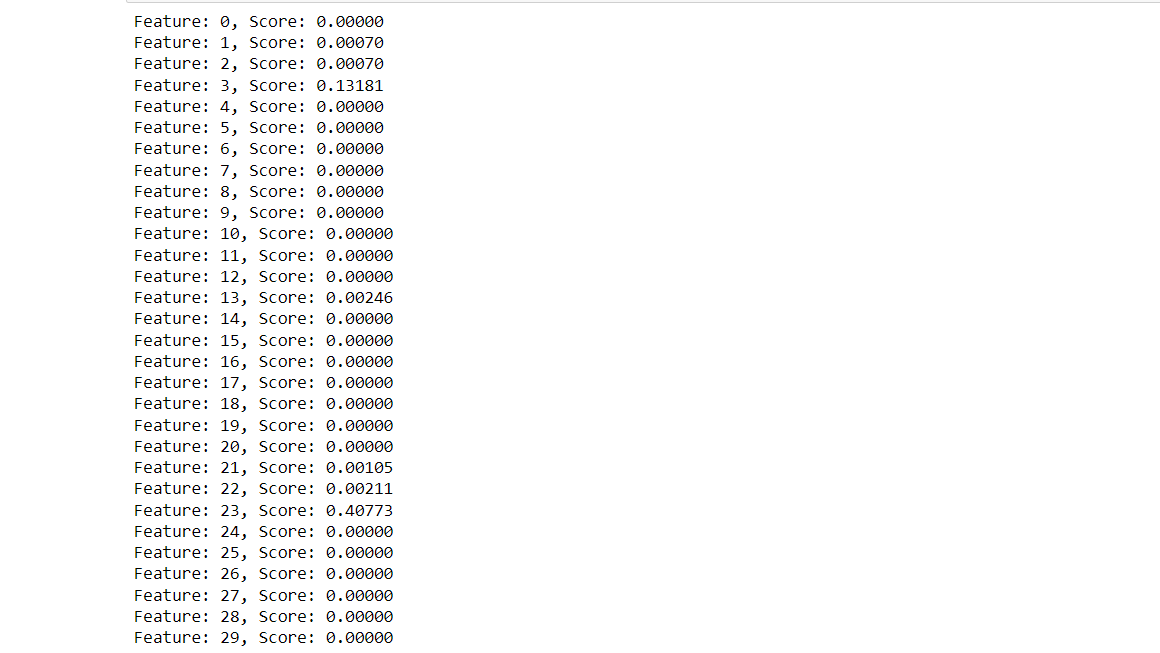


. The AUC score between true positive rate and false positive rate is 0.9543 and confusion matrix for the knn algorithm is :



## 3.5.4.1 Feature Importance for KNN:

I have applied feature importance for k nearest neighbors . Model score drops when an attribute value is randomly shuffled, which is defined as the relevance of the permutation feature. In other words, the method weakens the link between the attribute and its aim, which is why the model's performance plummets. These are the feature importance values for each feature :



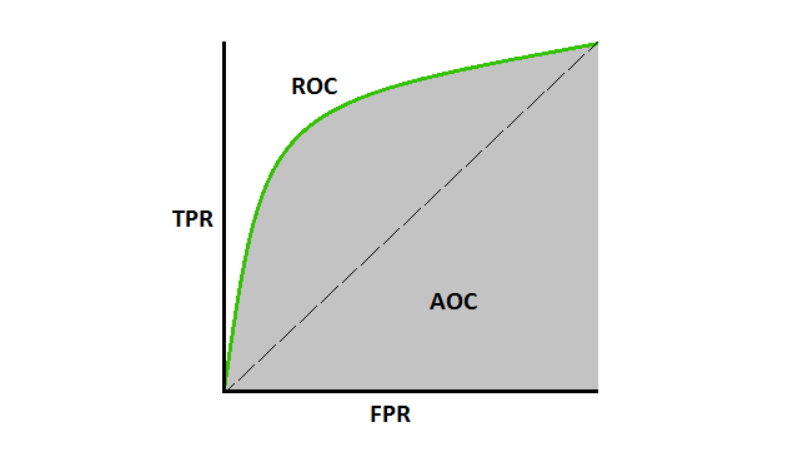
## 3.6 Performance Evaluation Metrics

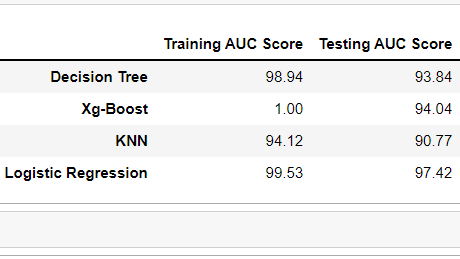
## 1. ROC-AUC Curve

While it is important to validate an evaluation of a model's skill, using metrics is necessary to evaluate the performance of a model. Metrics are employed to determine the most suitable problem-solving method. These metrics are what matter most.

At varied threshold values, AUC - ROC curve measures classification issue performance. AUC indicates the degree of separability, while ROC is a probabilistic curve. A model's class separation capability is indicated by this value. It's easier to forecast classes of 0 and 1 when the AUC is higher. AUC measures how well a model can tell which patients have the condition and which do not. AUC approaching 1 indicates that a model has a high degree of separability. AUC approaching 0 indicates a weak model with the lowest level of separability. As a matter of fact, it implies that the outcome is being re-expressed. It predicts 0s to be 1s and 1s to be 0s, respectively. The model has zero ability to separate classes when AUC equals 0.5.

The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis.

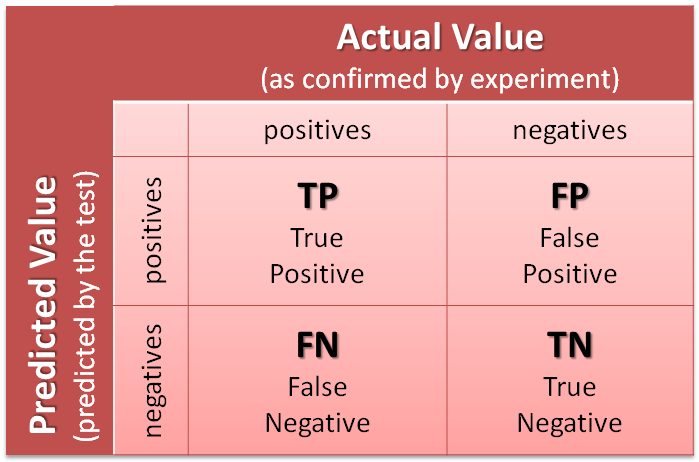
The AUC-ROC Scores for each of the models is :



2. Confusion Matrix:

A classification algorithm's performance can be summarised using a confusion matrix. A dataset with unequal numbers of samples in each class or more than two classes can be misled solely by classification accuracy.

It is possible to acquire a better understanding of your classification model's accuracy by computing a confusion matrix.



1. True Positive Rate : The true positive rate, also known as sensitivity or recall, is a metric used in machine learning to assess the proportion of real positives that are accurately detected.

2. False Positive Rate : There are two ways to compute the false positive rate: as FP/FP+TN and as FP/FP+TN (FP+TN being the total number of negatives). It's the likelihood that a false alarm will be sounded, resulting in a positive result when the underlying value is actually negative.

3. False Negative Rate : A false negative result that is actually positive. It's an example of a false negative when the results of a cancer test come back negative even though the person has cancer.

4. True Negative Rate : They don't have the sickness, which is exactly what we thought would be the case. Although we projected that they would be positive, they don't have the condition.... When a test results in a false negative (FN), it means that we projected the patient did not have the condition. A "Type II mistake" is another name for this phenomenon.)

## 

## Conclusion

In this paper, I have worked with different models of Machine Learning to predict the cancer disease from the given data to identify the better performing models. It was hoped to get multiple perspectives by doing various models, and also compare the performances. The goal of this study was to predict the cancer by utilising a dataset containing 33 predictors and over 569 observations. The dataset was revealed through the data visualisations and data exploration, and then deep exploration of the features was performed. I have also performed feature importance for each of the algorithm. So that we can remove the irrelevant features. It was investigated how features related to each other.

Initially, the exploratory data analysis was performed to understand the insights in the data. In Exploratory data Analysis I have performed various data visualization to understand the data. Further, some feature engineering was performed. In feature engineering firstly i have checked the null values . there is no null values in the dataset . Then i have applied Label encoding for converting the categorical variables into numerical ones. Then i have applied feature scaling for normalizing purpose. Then I have divided the data into training and testing data. I have performed feature importance for each of the algorithm and standardization only affects the Logistic Regression and KNN Algorithms. Then we will give this training data to the Various Machine learning Algorithms( Decision Tree, Xg-Boost, Logistic Regression, KNN)

To evaluate these models, I have used RUC-AUC Curve as the metric. By analyzing the performances of the models, it has been observed that Xg-Boost has recorded the Highest ROC-AUC values.

To answer the research questions posed, after performing the analysis on the data with different models with all the evidence we can conclude that :

* Feature importance plays a very important role in this project. Because there are 33 number of features and some features are unnecessary. That’s why i have applied feature importance with each of the algorithms.
* Radius worst and area\_mean are the most important features of the dataset.
* Xg-boost gives higher AUC-RUC values with training and testing data. Other Algorithms also performs well with the dataset. The number of False positives are False Negatives are 0 and 5 respectively.