<https://github.com/DeltaOptimist/EDA_Breast_Cancer_Descriptive/blob/main/EDA_Data_Visualization_Breast_Cancer_Use_case.ipynb>

<https://github.com/DeltaOptimist/Handling_Missing_Value_EDA>

<https://github.com/DeltaOptimist/Label_Encoding_Label_Encoding>

<https://github.com/DeltaOptimist/Linear_Regression_Potato_Price>

<https://github.com/DeltaOptimist/Spam_Mail_Prediction_Using_Logistic_Regression>

<https://github.com/DeltaOptimist/Feature-Transformation>

<https://github.com/DeltaOptimist/Decision_Tree_Pima_India_Dataset>

<https://github.com/DeltaOptimist/Kmeans-Mall-Customer-Clustering>

<https://github.com/DeltaOptimist/Audi_price_prediction_Random_Forest_Regressor>

**Label encoding 🡪** Assigns unique integer based on alphabetical order

**One-hot encoding 🡪** Creates dummy variables ( var are highly correlated to eachother) it leads to multicollinearity. It occurs where there is a dependency between independent features. To overcome this one of the dummy var has to be dropped.

If categorical features are ordinal use LABEL, not ordinal use ONEHOT

If num of features are more use LABEL, if less use ONEHOT

Axis=0 🡪 rows , Axis=1 🡪 columns

Inplace =true 🡪 data modified in place (dataframe updation ) , false – return copy of the object which needs to be saved.

Isnull.sum 🡪 return if any null, add returns sum of each row

Mean>median/mode 🡪 +

Mean<median/mode🡪 -

Fit\_transform 🡪 both simultaneously will increase model’s efficiency

Groupby🡪 to group all rows by specified col

Displot 🡪 plots a univariate distribution

Central Tendencies:

1. Mean
2. Median
3. Mode

 TfidfVectorizer 🡪 It helps us in dealing with most frequent words.

sklearn.metrics import accuracy\_score 🡪 (TP+TN)/ (TP+FN+TN+FP)

notnull() function detects existing/ non-missing values in the dataframe.

random\_state is basically used for reproducing your problem the same every time it is run

min\_df is 1 , which means "ignore terms that appear in less than 1 document"

Types Of Transformation

1. Normalization And Standardization
2. Scaling to Minimum And Maximum values
3. Scaling To Median And Quantiles
4. Guassian Transformation
5. Logarithmic Transformation
6. Reciprocal Trnasformation
7. Square Root Transformation
8. Exponential Trnasformation
9. Box Cox Transformation

**Feature transformations** on machine learning training are used to massage the data into a format that is better conducive to rapid learning.

**Gradient Descent** is an iterative algorithm that is used to minimize a function by finding the optimal parameters.

A **local minimum** of a function is a point where the function value is smaller than at nearby points, but possibly greater than at a distant point.

A **global minimum** is a point where the function value is smaller than at all other feasible points.

Eucledian Distance 🡪 Algorithms like KNN,K Means,Hierarichal Clustering

Every Point has some vectors and Direction

* Artificial Neural Networks (ANN)
* Convolution Neural Networks (CNN) -  learns the filters automatically without mentioning it explicitly. These filters help in extracting the right and relevant features from the input data
* Recurrent Neural Networks (RNN) - RNN captures the sequential information present in the input data i.e. dependency between the words in the text while making predictions

Standardization comes into picture when features of input data set have large differences between their ranges, or simply when they are measured in different measurement units (e.g., Pounds, Meters, Miles … etc). We try to bring all the variables or features to a similar scale. standarisation means centering the variable at zero

**MinMaxScaler** 🡪 Transform features by scaling each feature to a given range

**The towers or bars of a histogram are called bins**.

robust scaler 🡪 **Scale features using statistics that are robust to outliers**. This Scaler removes the median and scales the data according to the Interquartile Range (IQR). The IQR is the range between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile).

IQR = 75th quantile - 25th quantile.

* Guassian Transformation
* logarithmic transformation - oftenly used for reducing right skewness.
* reciprocal transformation - defined as 1/x,  It isn't defined for the value 0, but it can be applied to negative numbers.
* square root transformation - method is typically used when your data is moderately skewed.
* exponential transformation (more general, you can use any exponent)
* boxcox transformation: technique to transform non-normal data into normal shape.

T(Y)=(Y exp(λ)−1)/λ

where Y is the response variable and λ is the transformation parameter. λ varies from -5 to 5. In the transformation, all values of λ are considered and the optimal value for a given variable is selected.

**scipy.stats** is the SciPy sub-package. It is mainly used for probabilistic distributions and statistical operations.

For plotting tree, you also need to install graphviz and pydotplus.

pip install **graphviz** - function converts decision tree classifier into dot file

pip install **pydotplus** - convert this dot file to png

**StringIO** module is an in-memory file-like object.

we always assume the max number of cluster would be 10

you can judge the number of clusters by doing averaging

**elbow method** is a heuristic used in determining the number of clusters in a data set.

Within **Cluster Sum of Squares** (WCSS), K-means prefer range (1,11)

optimal number of clusters is 5

cluster\_centers\_ is called **the code book** and each value returned by predict is the index of the closest code in the code book.

remainder='passthrough' 🡪 remainder is initialized to passthrough means the column which we have not used in any transformation process should be pass as it is.

**StandardScaler** is used to resize the distribution of values ​​so that the mean of the observed values ​​is 0 and the standard deviation is 1

**r2\_score 🡪** regression score function. Best possible score is 1.0

In this type of learning we train our machine with the labelled dataset and then our machine predicts and provide us the labels for the new set of data.

The most important Supervised Learning Algorithms:  
· Linear Regression  
· Logistic Regression  
· Support Vector Machines (SVM)  
· Decision Trees and Random Forests  
· k-Nearest Neighbours  
· Neural networks

In unsupervised learning algorithm, we provide the unlabelled dataset to

our machine and it tries to analyze and find patterns within the dataset

The most important Unsupervised learning algorithms are:  
Clustering  
· k-Means  
· Hierarchical Cluster Analysis (HCA)

Semi Supervised Learning

Algorithms that can deal with partially labelled training data (Supervised) & a lot of unlabelled data (Unsupervised).

Classification. In classification, the algorithm assigns labels to data based on the predefined features. This is an example of supervised learning. More about classification algorithms you can read in our blog.

Clustering- An algorithm splits data into a number of clusters based on the

similarity of features. This is an example of unsupervised learning.

Regression - Regression algorithms try to find a relationship between variables

and predict unknown dependent variables based on known data. It is based on supervised learning.

Training Data—a subset to train a model.  
Testing Data—a subset to test the trained model.

Step 1: Import Scikit-learn library

Step 2: Import and load the Dataset  
(Identify Features, Target and Data)

Step 3: Organizing data into sets:  
(Train, Test Split)

Step 4:Building the model  
(Using dedicated Algorithm)

Step 5: Evaluating the model and its accuracy  
(Confusion Matrix, Accuracy etc)

Correct positive predictions (true positives),

Correct negative predictions (true negatives),

Incorrect positive predictions (false positives),  
Type 1 error  
  
Incorrect negative predictions (false negatives).  
Type 11 error

**Precision** is a measure that tells us what proportion of patients that we diagnosed as having cancer, actually had cancer.  
The predicted positives (People predicted as cancerous are TP and FP) and the people actually having a cancer are TP.

*[how many did we miss]*

**Recall** is a measure that tells us what proportion of patients that actually had cancer was diagnosed by the algorithm as having cancer.  
The actual positives (People having cancer are TP and FN) and the people diagnosed by the model having a cancer are TP.

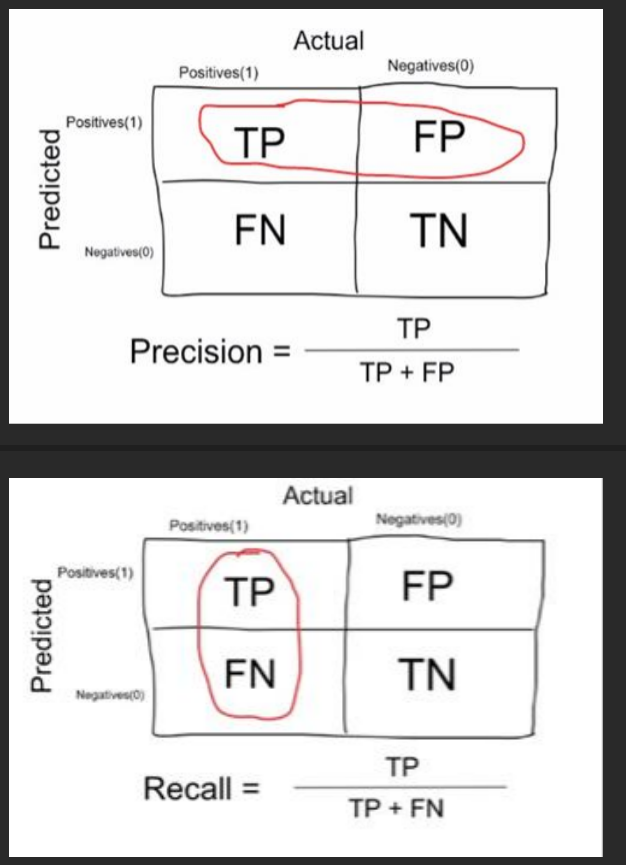
*[how many did we caught*

**Specificity –** opp of recall

Did not had cancer , model predicted non-cancerous

TN/TN+FP

**F1 SCORE -**harmonic mean of Precision & recall🡪 2\*P\*R/P+R, f scprr is 5% then accuracy is 95%.. not sure with P or R,, go for F1score



Overfitting – High variance and low bias

Techniques to reduce overfitting :  
  
1. Increase training data.  
2. Reduce model complexity.  
3. Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).  
4. Ridge Regularization and Lasso Regularization  
5. Use dropout for neural networks to tackle overfitting.

**Underfitting** – High bias and low variance

Techniques to reduce underfitting :  
  
1. Increase model complexity  
2. Increase number of features, performing feature engineering  
3. Remove noise from the data.  
4. Increase the number of epochs or increase the duration of training to get better results.

Decision tree 🡪 MAE mean absolute error

Linear R🡪 RMSE value

Classification 🡪 Confsn metrics {Precision,recall,support,specificity, sensitivity}

Unseen data 🡪 Y test [Evaluation] , X test 🡪 Prediction

Label🡪 output var / dependent var

Feature🡪 input var/independent var

(Fixed) Discrete data 🡪 classification , Continuous data 🡪 Regression

Classification ka unsupervised is clustering

Feature eng 🡪 transformation + selection

Euclidean🡪 similar type of variables

Manhattan 🡪 Different type of variables

Types of models:

* 1. GEOMETRIC🡪 Euclidean & Manhattan
  2. PROBALISTIC🡪something will happen given that something else has happend
  3. LOGICAL 🡪 If-then implication

Types of feature:

1.QUANTITATIVE🡪 mean tendency

2.ORDINAL🡪 median tendency

3.CATEGORICAL🡪 mode tendency

4.BOOLEAN