A Comparative Analysis of BreastCancer Detection and Diagnosis

Using Data Visualization and Machine Learning Applications

ANNASSIRI Fatima Zahra
The National School of Applied Sciences of Tangier

Abstract

In the developing world, cancer death is one of the major problems for humankind. In the literature, there are many studies about predicting the type of breast tumors. In this paper, data about breast cancer tumors from Dr. William H. Walberg of the University of Wisconsin Hospital were used for making predictions on breast tumor types, based on the paper of dr.Muhammet Fatih Ak, has the same title.

Data visualization and machine learning techniques including logistic regression, k-nearest neighbors, support vector machine, naïve Bayes, decision tree, random forest, and rotation forest were applied to this dataset. Python seaborn and pendas were chosen to be applied to these machine learning techniques and visualization.

In this paper, we present a bibliographic and literature review about breast cancer and different machine learning and data mining techniques for the detection of breast cancer were proposed. the Results obtained with the Random Forest Classifier model with all features included showed the highest classification accuracy (96.5 per cent).

Table of Contents

Introduction

Breast cancer is the most common cancer in women worldwide, with nearly 1.7 million new cases diagnosed in 2012, representing about 25 percent of all cancers in women. Incidence rates vary widely across the world, from 27 per 100,000 in Middle Africa and Eastern Asia to 92 per 100,000 in Northern America. It is the fifth most common cause of death from cancer in women, with an estimated 522,000 deaths (6.4 % of the total).
It represents 36.9% of all female cancers in Morocco **(GLOBOCAN., 2018)**.

Introduction

Breast Cancer Overview

The causes of breast cancer are multiple, both genetic and environmental (Cordina-Duverger et al., 2016).

In addition to diet, nutrition and physical activity, other established causes of breast cancer include life events, radiation and medicament.

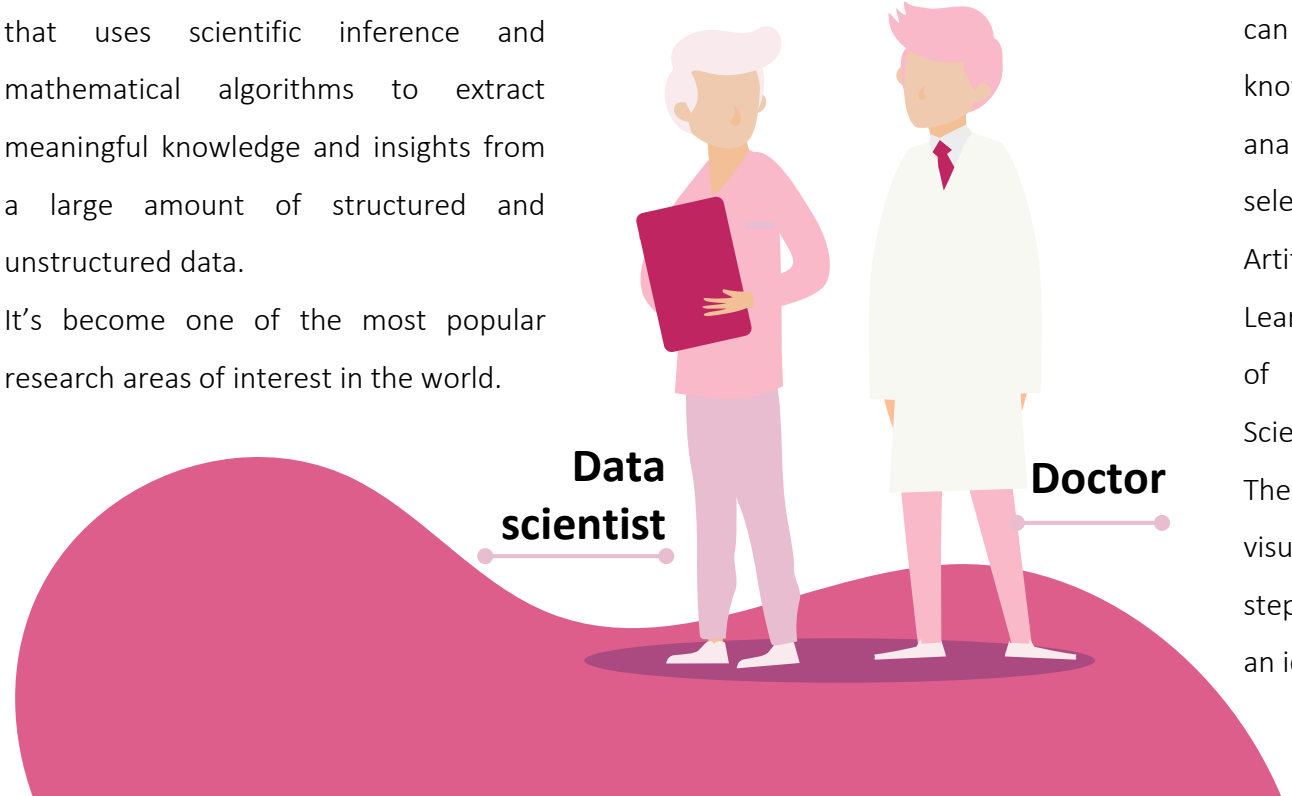
The diagnosis of breast cancer combines several examination steps. Exams additional clinical, para-clinical and laboratory tests are needed to define and specify the therapeutic strategy best suited to each patient.

Introduction

DS, DM ,ML and DV Thechniques Overview

Data Science is a multidisciplinary field that uses scientific inference and mathematical algorithms to extract meaningful knowledge and insights from a large amount of structured and unstructured data.

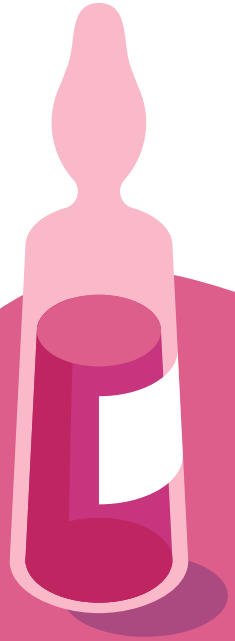
It's become one of the most popular research areas of interest in the world.



They exist many public datasets useful in the domain of health-care.

The researchers believe that they can transform this data into useful knowledge for prediction Using analysis techniques for feature selection, such as data mining, Artificial Intelligence, and Machine Learning, these techniques are one of the core elements of Data Science.

Therefore, data understanding and visualization are very important steps for preprocessing data to have an idea about the data



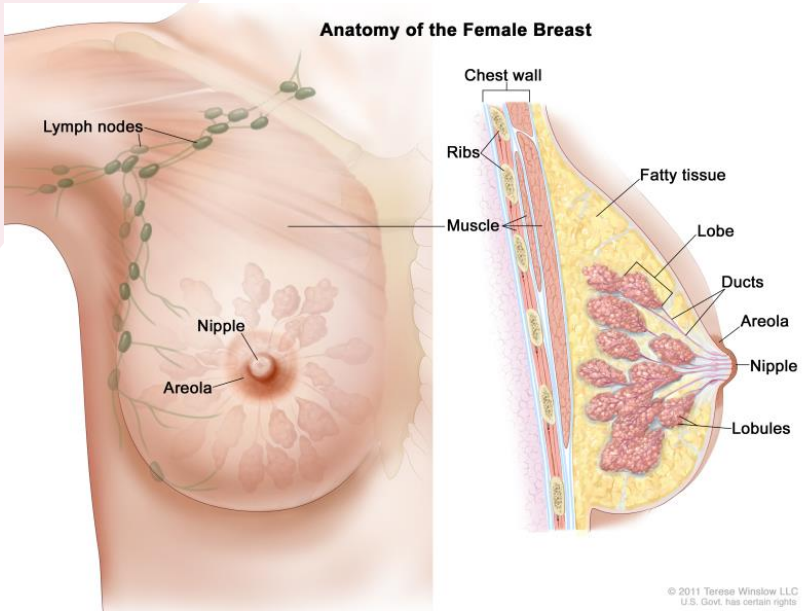
02

Breast Cancer Bibliographic Review

Breast Cancer Bibliographic Review

Anatomy and Physiology of Breaste

The breasts are under the hormonal control of the ovaries, which in turn are under the control of the pituitary gland. From puberty to menopause, the breast undergoes constant changes. These more or less visible changes are controlled by a set of hormones, the most significant of which are: estrogen and progesterone, which are secreted by the ovaries, and prolactin secreted by the pituitary gland. Estrogens in particular allow breast development during puberty and play an important role throughout pregnancy, and progesterone plays a role in particular in the differentiation of breast cells and in the menstrual cycle (Harlay et al., 1999).



Breast Cancer Bibliographic Review

Incidence and Mortality

Worldwide, an estimated 19.3 million new cancer cases and almost 10.0 million cancer deaths occurred in 2020. **Female breast cancer** has surpassed lung cancer as the most commonly diagnosed cancer, with **an estimated 2.3 million new cases (11.7%),** followed by lung (11.4%), ..

Lung cancer remained the leading **cause of cancer death**, with an estimated 1.8 million deaths (18%), followed by colorectal (9.4%), .. , **and female breast (6.9%) cancers.**

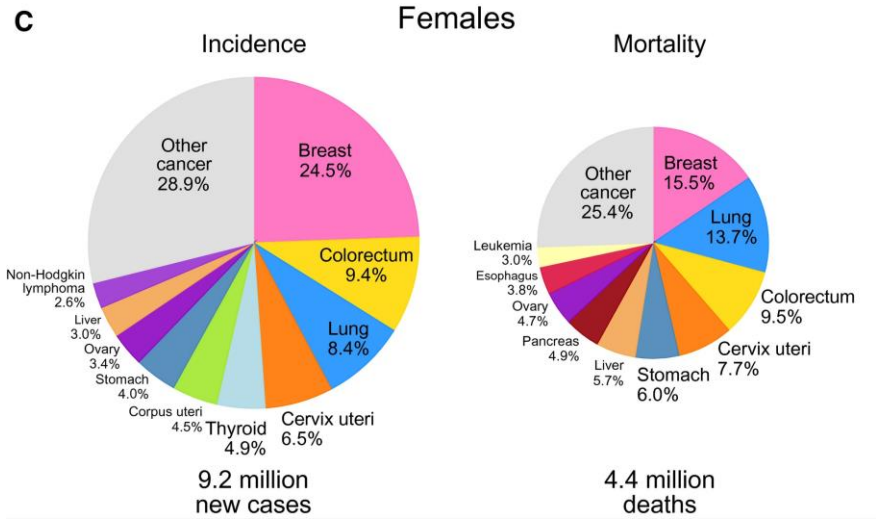


Figure : Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries

Breast Cancer Bibliographic Review

Incidence and Mortality

2.3M

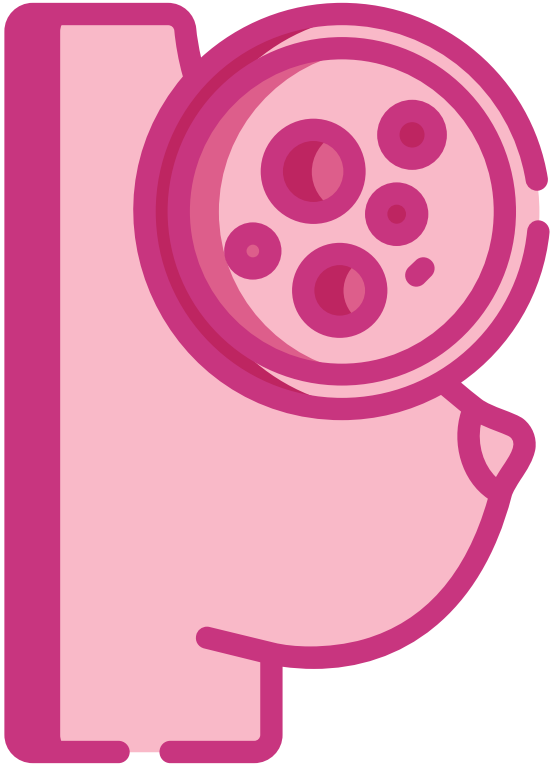
In 2020, 2,3M new case of Breast cancer
About 11;7% of worldwide new cancer cases

69K

About 6.9% deaths from the breast cncer in 2020.

Breast Cancer Bibliographic Review

Incidence



Men	10%
Women	90%

Breast Cancer Bibliographic Review

Risk factors

Advanced age

the period between 20 and 50 years, when the risk increases very rapidly with age.(Lansac et al., 2018).

Family history

The presence of a family history is a very important risk factor.

Genetic factor

the existence of a family risk: High-risk genes, Women carrying mutations in genes BRCA1 and BRCA2, tumor suppressor genes, have a high risk of developing CS. (Cordina-Duverger et al., 2016).

Ionizing radiation

Intensive population monitoring in Hiroshima and Nagasak has shown that the breast is one of the organs most sensitive to the effects radiation.

Benign breast history

are histologically divided into two groups: proliferative lesions and lesions not proliferative with or without atypia (Tabib Ibrahim et al., 2015).

hormonal factors

Reproductive factors

Alcohol

cigarette

obesity

Breast Cancer Bibliographic Review

Breast cancer symptoms

01

Dimpling

Irritation or dimpling of breast skin.

02

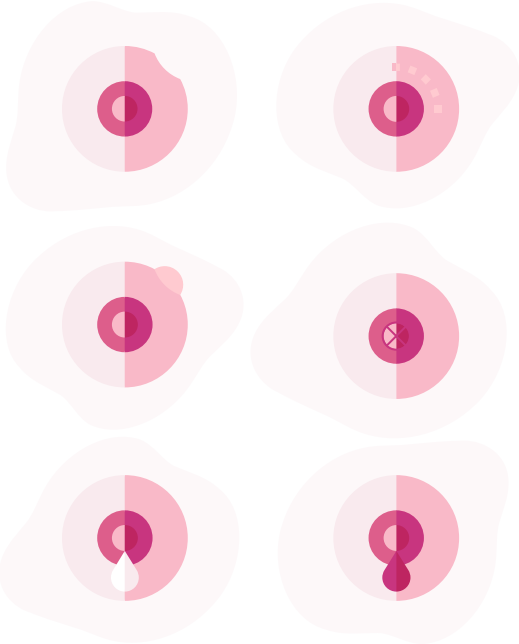
Lump

New lump in the breast or underarm (armpit).

03

Lymph discharge

A lump or swelling in the underarm lymph nodes.



04

Texture change

Any change in the size or the shape of the breast.

05

Nipple inversion

Redness or flaky skin in the nipple area or the breast. Pulling in of the nipple or pain in the nipple area.

06

Bloody discharge

Nipple discharge other than breast milk, including blood.

Breast Cancer Bibliographic Review

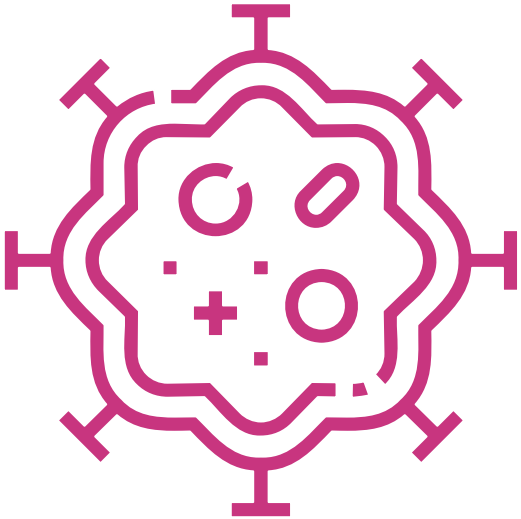
Medical diagnostic

Mammography

It is a first-line radiological examination in front of a breast nodule.

Ultrasound

Breast ultrasound is a breast imaging test that uses ultrasound to produce images of the inside of the breast



Genetic testing

The BRCA gene test is a blood test that's done to determine if you have mutations in your DNA that increase the risk of breast cancer.

Biopsy

The biopsy is the only examination that provides a definite diagnosis of the tumor.

Breast Cancer Bibliographic Review

Traitement

Surgery

There are two types of treatment:
Conservative treatment : consists of removing only the tumor (lumpectomy)
mastectomy: it is the excision of the entire mammary gland.

Targeted therapy

Anti-HER2 therapies: Trastuzumab (Herceptin), Pertuzumab (Perjeta) and Trastuzumab - emtansine (Kadcyla), Trastuzumab and pertuzumab improve the effectiveness of chemotherapy.

Radiotherapy

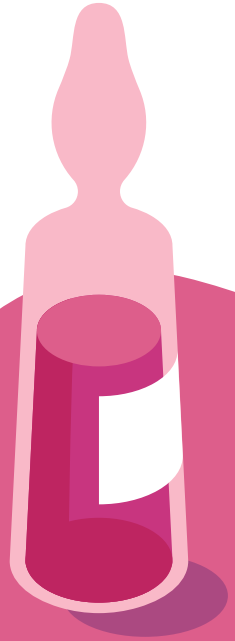
External irradiation of the breast (or chest wall in case of mastectomy) and lymph node areas is usually performed postoperatively.

Chemotherapy

Performed after surgery, it decreases the risk of relapse and increases the chances of recovery.

Hormonal therapy

It aims to prevent the action on the tumor of estrogen, produced continuously by the body. It is indicated when hormone receptors have been detected in the tumor.



03

Breast Cancer Literature Review

Breast Cancer Bibliographic Review

Introduction

The health-care environments become increasingly complex, because of the big increase in the amount of data in multiple forms and type: samples hospital reports .., this data became numerical almost time.

From Wikipedia; Health care is the maintenance or improvement of health via the prevention, diagnosis, treatment, recovery, or cure of disease, illness, injury, and other physical and mental impairments in people.

Health care is delivered by health professionals and allied health fields. Medicine, dentistry, pharmacy, midwifery, nursing, optometry, audiology, psychology, occupational therapy, physical therapy, athletic training and other health professions are all part of health care.

It includes work done in providing primary care, secondary care, and tertiary care, as well as in public health.

Breast Cancer Bibliographic Review

Recent Related Studies

Objective	Study	Technology Used
Clarification of data science and applications	Dhar	Literature reviews
	Aruna et al.	Naïve Bayes, support vector machine, decision tree
	Chaurasia et al	Naïve Bayes, SVM, neural networks, decision tree
	Asri et al	SVM, decision tree (c4.5), naïve Bayes, k-nearest neighbors
	Delen et al	Naïve Bayes, neural network, c4.5 decision tree
	Qu et al	Naïve Bayes, decision tree, random tree
	Sriniva et al	One dependency augmented naïve Bayes, naïve Bayes

Table : Recent Related Studies.(Muhammet Fatih Ak et all, 2020)

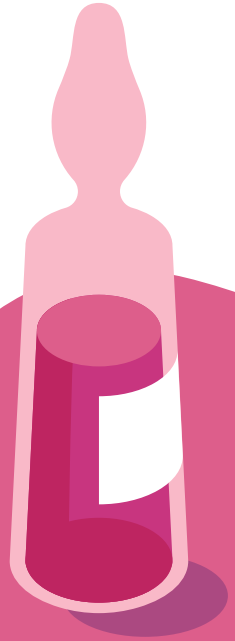
Objective	Study	Technology Used
Objective Study Approach and Methods Used Prediction of breast cancer by using data mining methods	Bernal et al.	Logistic regression, neural networks, decision tree, nearest neighbors
	Wang et al.	Support vector machine (SVM), artificial neural network (ANN), naïve Bayes classifier, adaboost tree
	Williams et al.	Naïve Bayes and the J48 decision trees
	Nithya et al.	Naïve Bayes, support vector machine-sequential minimal optimization, decision tree, multilayer perceptron
	Oyewola et al	Logistic regression, linear discriminant analysis, quadratic discriminant analysis, random forest and support vector machine
	Agarap	Agarap Gru- SVMS, linear regression, multilayer perceptron, nearest neighbor, softmax regression and support vector machine
	Westerdijk	Logistic regression, random forest, support vector machine, neural network, and ensemble models
	Vard et al	Particle swarm optimization (PSO), support vector machine (SVMs), decision tree and multilayer perceptron neural network
	Kourou et al	Artificial neural networks (ANNs), Bayesian networks (BNs), support vector machines (SVMs) and decision trees (DTs)
	Pratiwi	Extreme learning machine methods
	Shukla et al	Self-organizing map (SOM) and density-based spatial clustering of applications with noise (DBscan), multilayer perceptron (MLP), SEER program models

Breast Cancer Bibliographic Review

Resume

The main contributions of this paper are provided in the following:

- Establish an adequate model by revealing the predictive factors of early-stage breast cancer patients from a broader perspective and compare the robustness of the model by accuracy measures.
- A more comprehensive comparison and analysis using data visualization and machine learning applications for breast cancer detection and visibility to validate the model.
- Observe which features are most effective in predicting breast cancer and to understand general trends.
- A better prediction of breast cancer by using data mining methods.



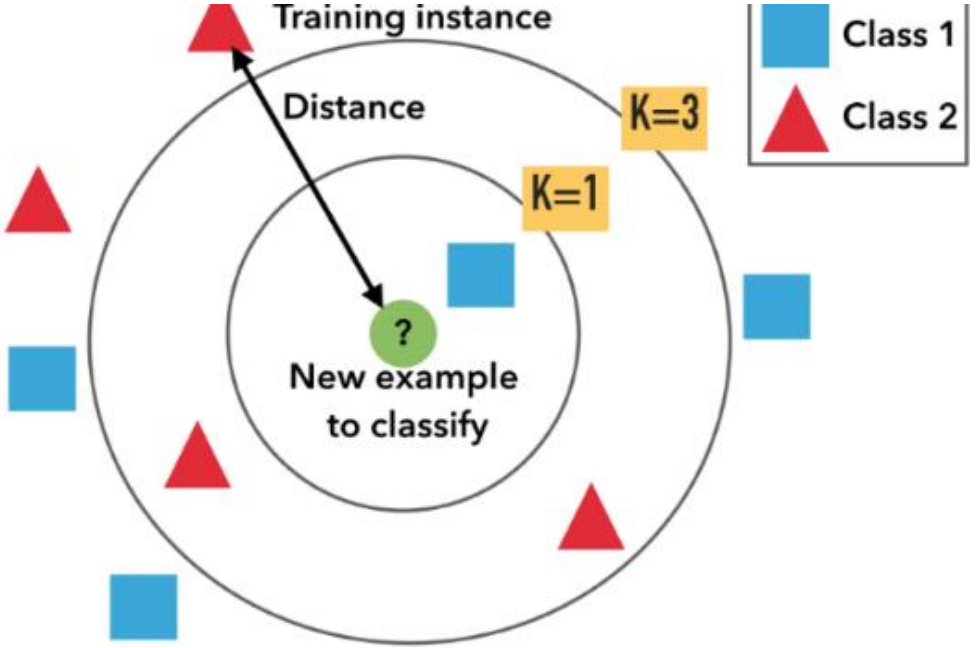
04

Methode and Application

Method and Application

k-nearest neighbors classifier (KNN)

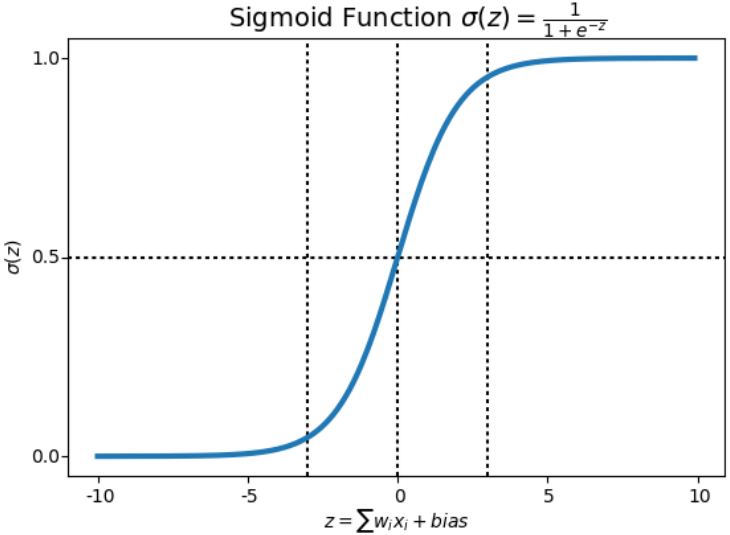
k-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically.



Method and Application

Logistic Regression

Logistic Regression is a Machine Learning algorithm which is used for classification problems, it is a predictive analysis algorithm and based on the concept of probability. Logistic regression transforms its output using the logistic sigmoid function to return a probability value.



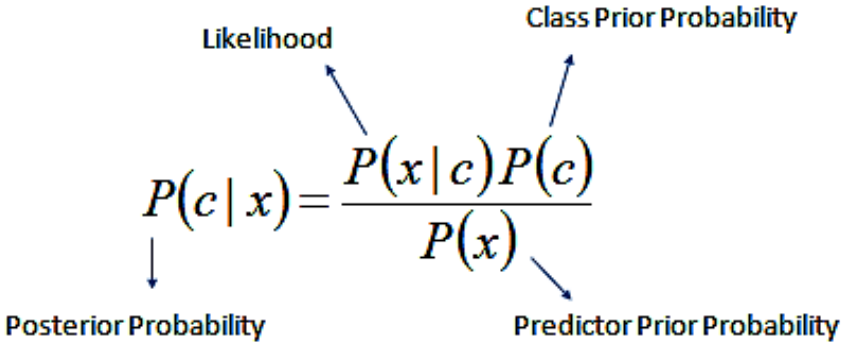
$$0 \leq h_\theta(x) \leq 1$$

We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the ‘Sigmoid function’ or also known as the ‘logistic function’ instead of a linear function.

Method and Application

Naïve Bayes

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.



$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

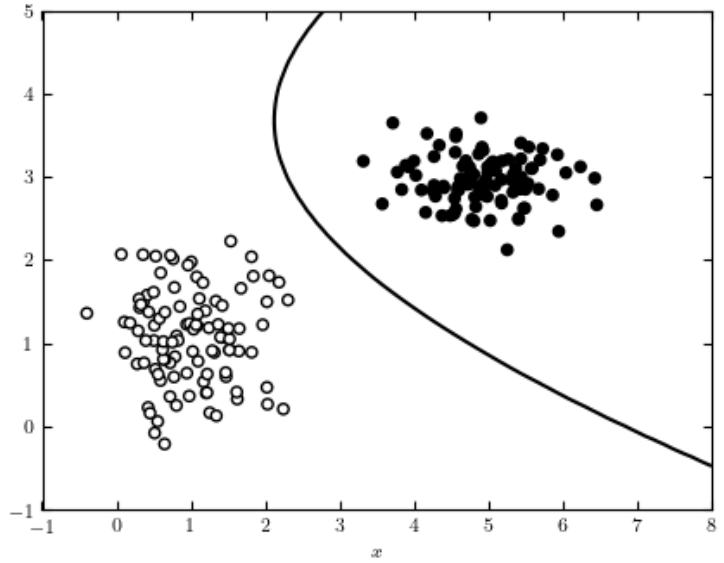
There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

Method and Application

Gaussian Naïve Bayes

Gaussian Naïve Bayes is one kind of naïve Bayes application. It assumes that features follow a normal distribution. The possibility of features is considered to be Gaussian and has a conditional probability. Gaussian naïve Bayes theorem is given below:

$$P(xi|y) = \frac{1}{\sqrt{2\pi\sigma 2y}} \exp\left(\frac{-(xi-\mu y)^2}{2\sigma^{2y}}\right)$$



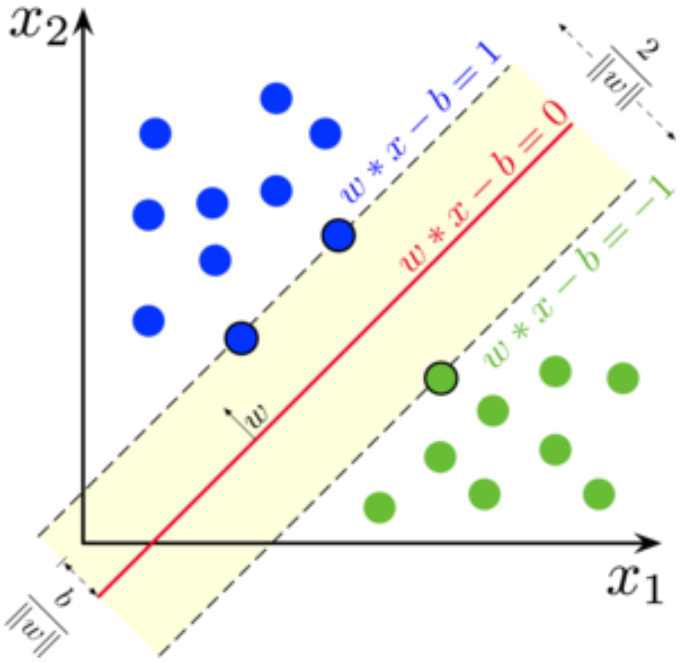
Method and Application

Support Vector Machine SVM

SVM in machine learning is supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.

SVMs are one of the most robust prediction methods, being based on statistical learning frameworks or VC theory.

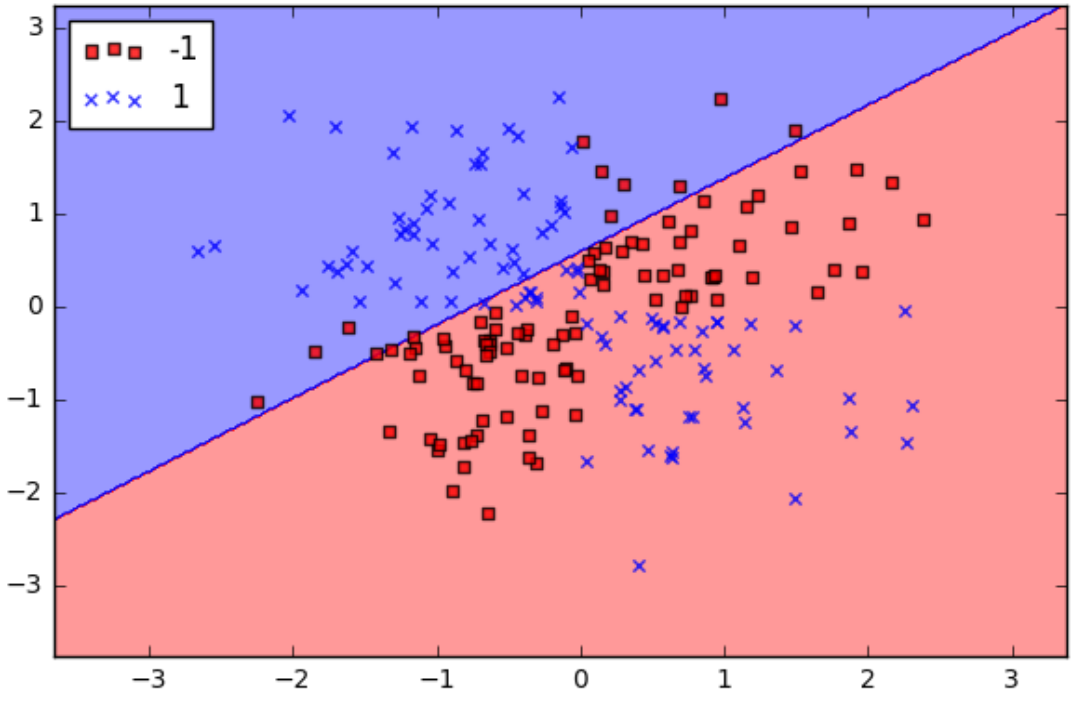
Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.



Method and Application

Linear Support Vector Classification

Similar to SVC with parameter kernel='linear', but implemented in terms of liblinear rather than libsvm, so it has more flexibility in the choice of penalties and loss functions and should scale better to large numbers of samples. This class supports both dense and sparse input and the multiclass support is handled according to a one-vs-the-rest scheme.



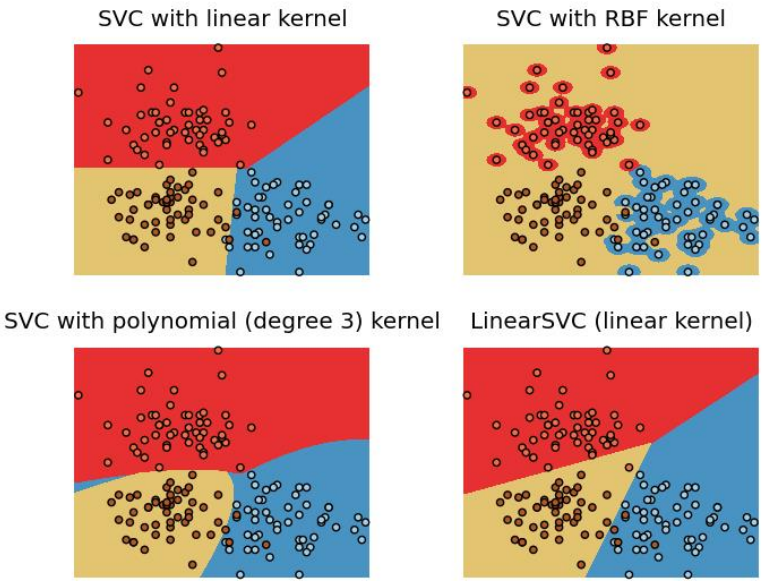
Method and Application

Support Vector Classification Using a RBF Kernel

Radial Basis Function is a commonly used kernel in SVC:

$$K(x,x')=\exp(-||x-x'||22\sigma2)$$

where ||x−x′||2 is the squared Euclidean distance between two data points x and x′. If this doesn't make sense, Sebastian's book has a full description. However, for this tutorial, it is only important to know that an SVC classifier using an RBF kernel has two parameters: gamma and C.



Method and Application

Decision Tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

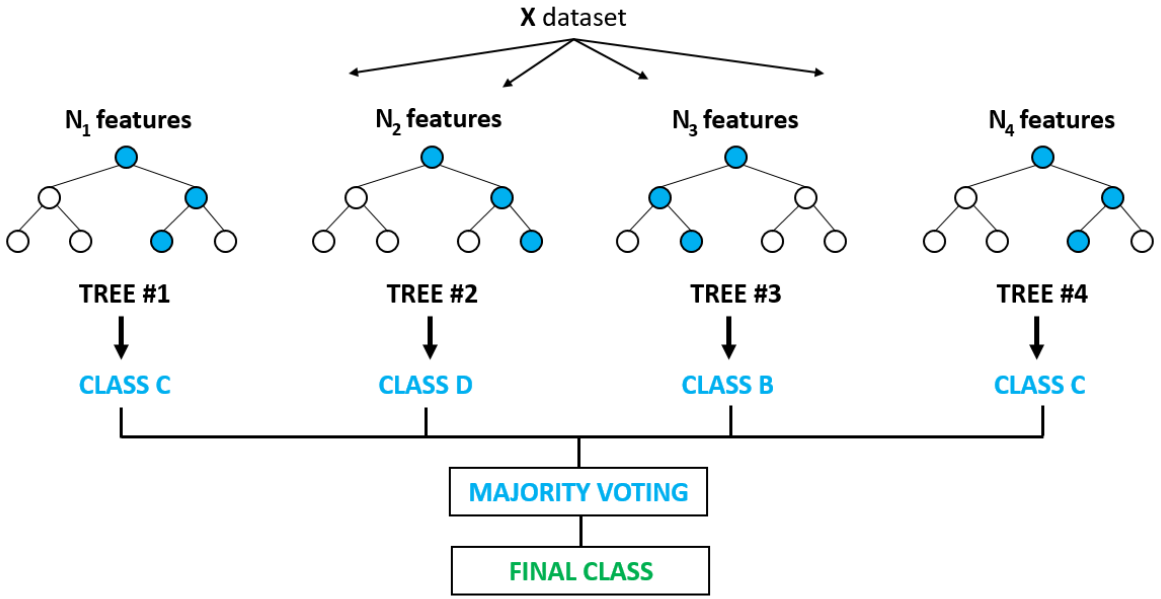
Randomness or uncertainty of feature x is defined as entropy and can be calculated as follows:

$$H(x) = Ex[I(x)] = -\sum p(x)logp(x)$$

Method and Application

Random Forest Classification algorithm

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.



Method and Application

Dataset

This data represents Wisconsin Diagnostic Breast Cancer (WDBC) dataset, published by Center for Machine Learning and Intelligent Systems.

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Data Set Characteristics:	Multivariate	Number of Instances:	569	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	32	Date Donated	1995-11-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	1443274

Method and Application

Attribute Information:

1) ID number
2) Diagnosis (M = malignant, B = benign)
3-32)

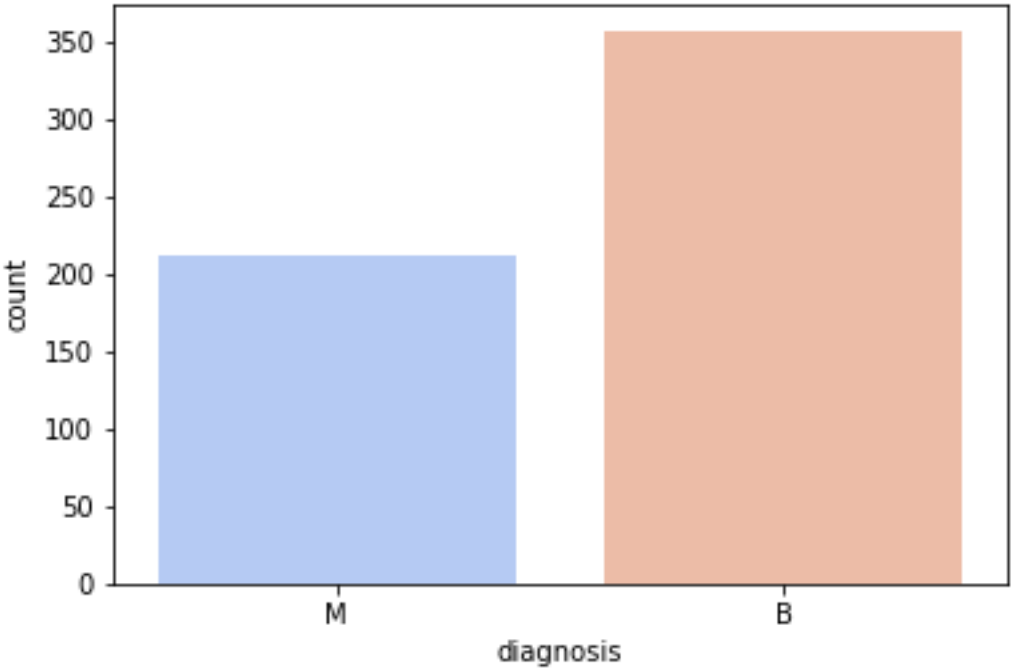
Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

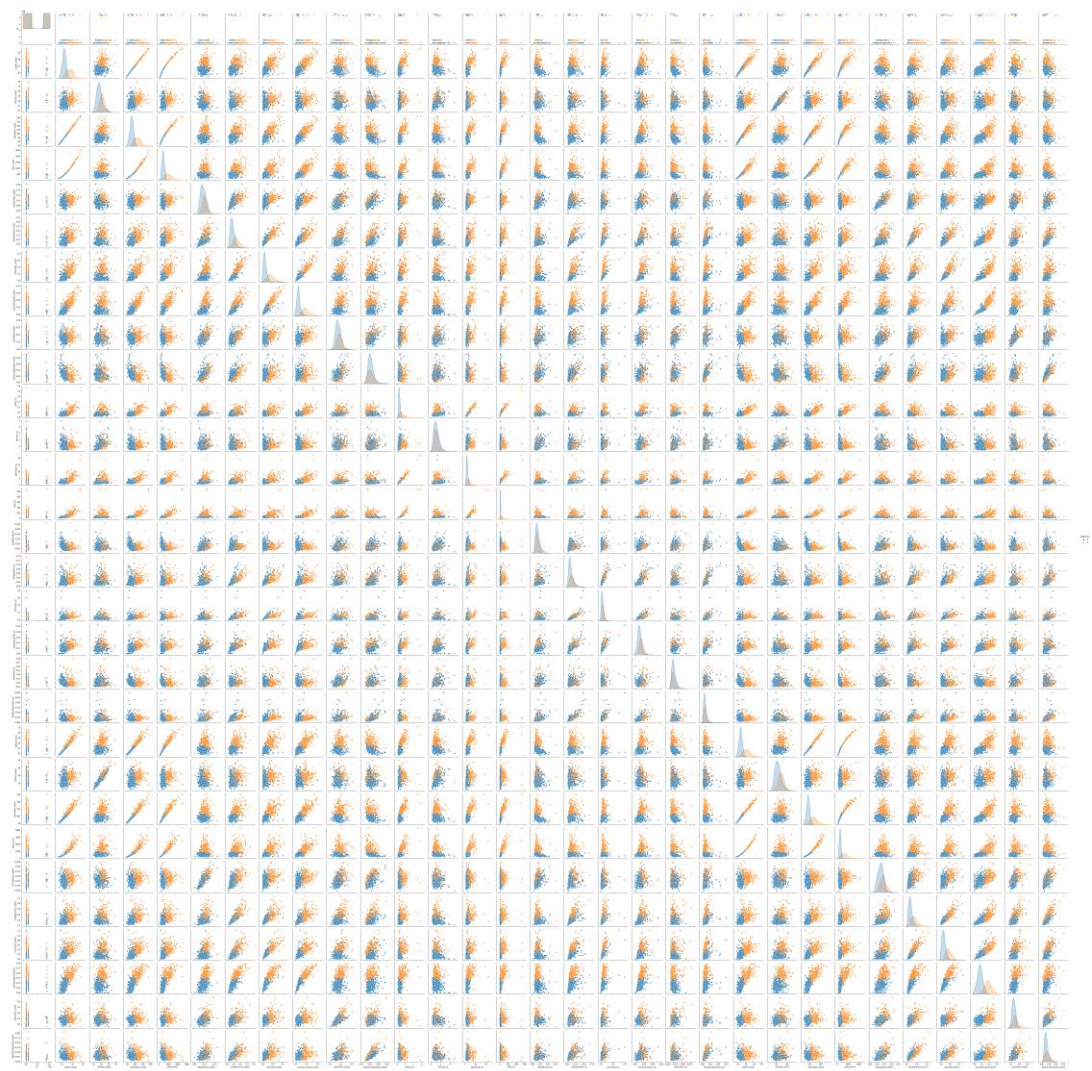
1	ID	9	Symmetry Mean	17	Smoothness Se	25	Perimeter Worst
2	diagnosis	10	concavity mean	18	compactness se	26	area worst
3	radius mean	11	concave points mean	19	concavity se	27	smoothness worst
4	texture mean	12	fractal dimension mean	20	concave points se	28	compactness worst
5	perimeter mean	13	radius se	21	symmetry se	29	concavity worst
6	area mean	14	texture se	22	fractal dimension se	30	concave points worst
7	smoothness mean	15	perimeter se	23	radius worst	31	symmetry worst
8	compactness mean	16	area se	24	texture worst	32	fractal dimension worst

Method and Application: Data visualization

A count of the number of 'M' Malignant and 'B'Benign cells



The count plot show that there is many cells Benign then Malignant.



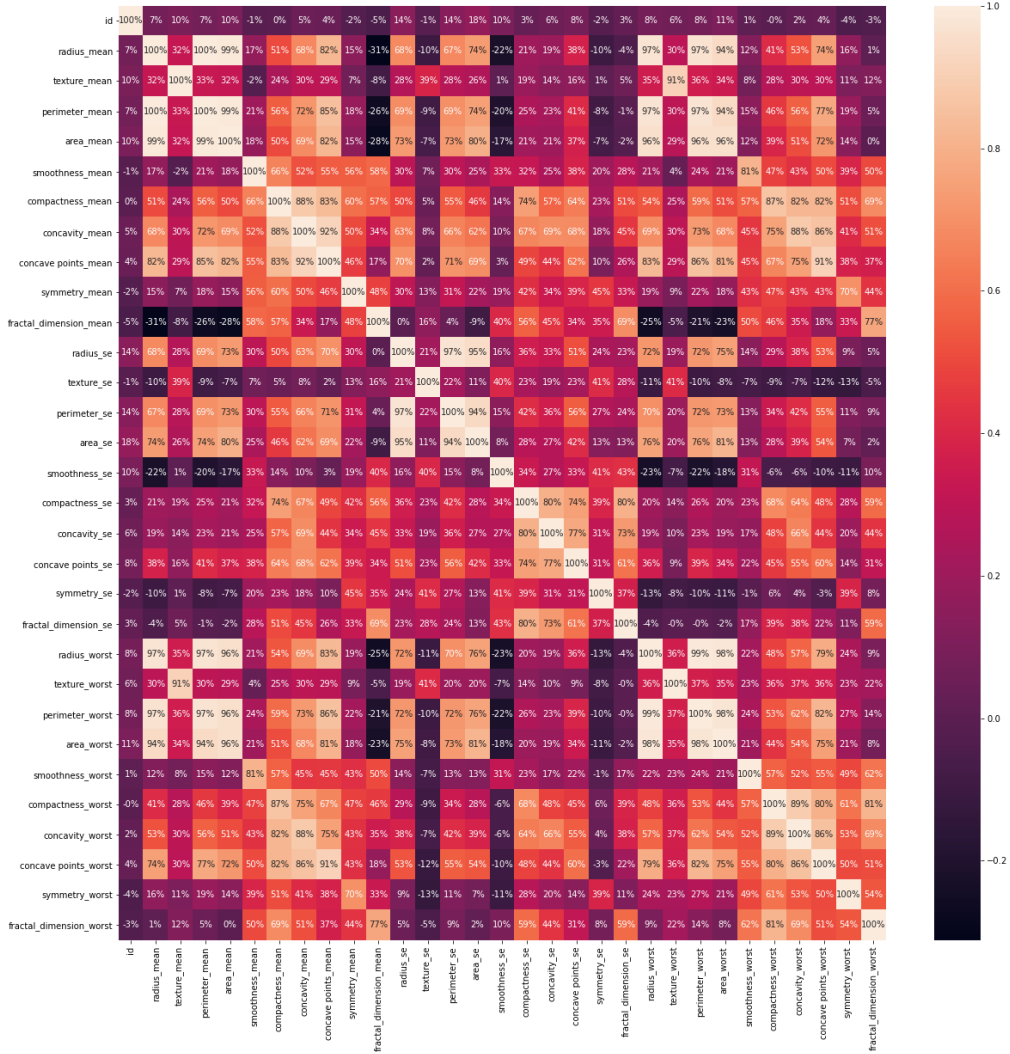
Method and Application: Data visualization

pairplot

pairPlot pairwise relationships in a dataset. Between all features .

All parameters can be useful to classify cancer.

Identify whether the data are balanced or unbalanced.



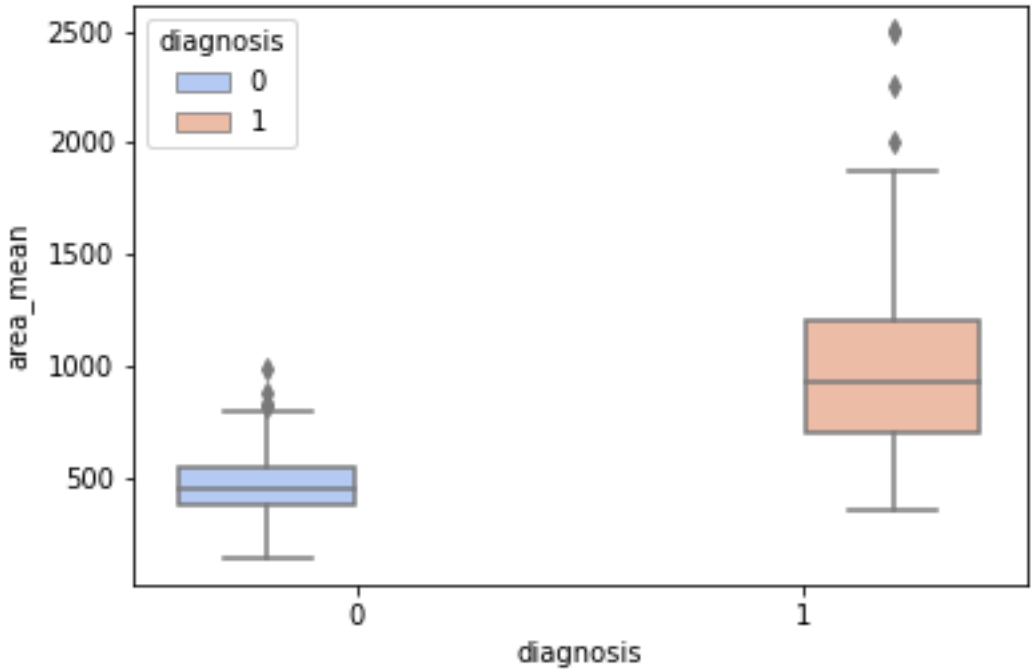
Method and Application: Data visualization

Visualize the correlation by creating a heat map

The heat map was constructed to indicate a correlation between all features. The graph is represented in figure below.

Method and Application: Data visualization

Normalise numerical value of Features and plot The box plot

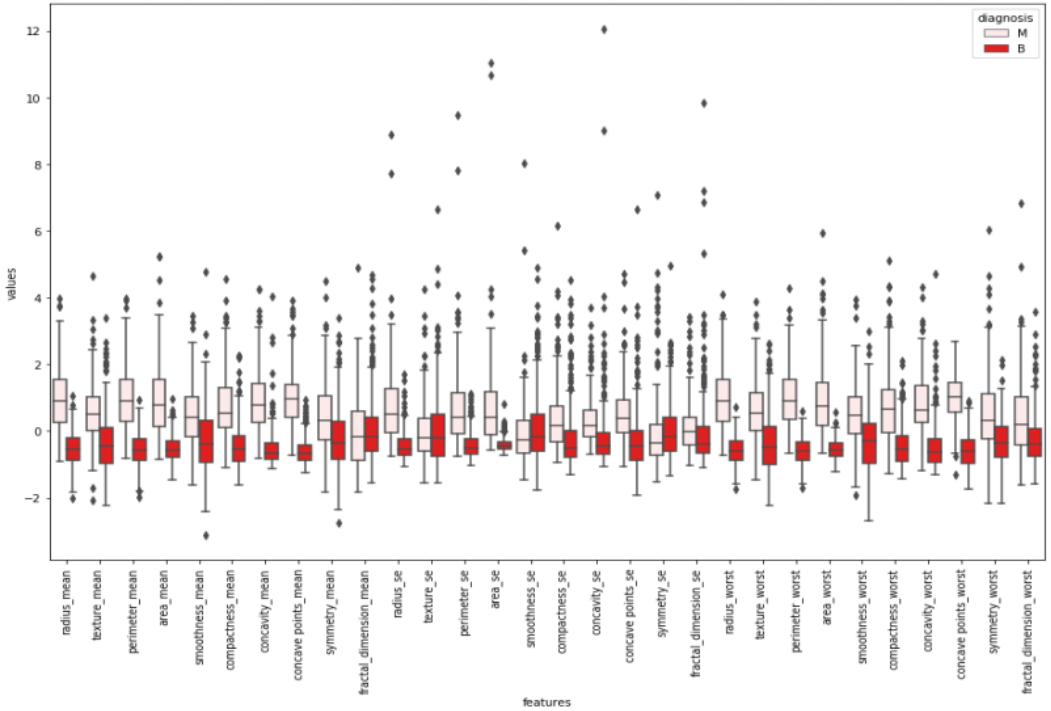


After normalizing the numerical values of features, a box plot was created for cleaning the data.

The figure below indicate a box plot of area_mean per diagnosis.

Method and Application: Data visualization

Normalise numerical value of Features and plot The box plot

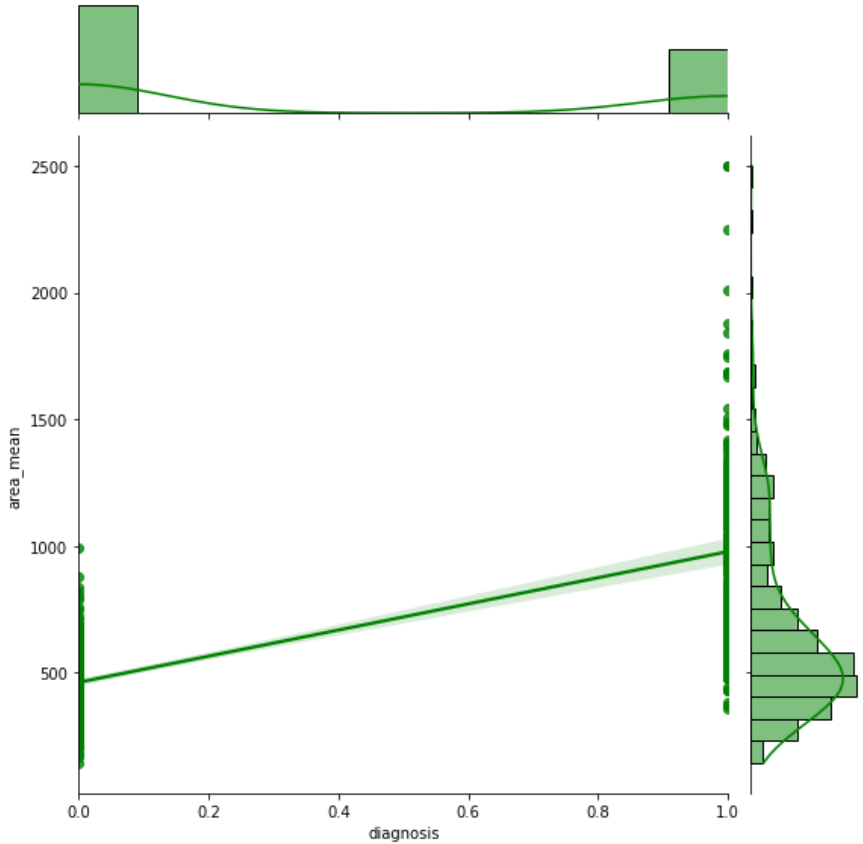


The figure below indicate a box plot of features , according to it , ther were so many outliers.

This means sufficiency was lacking to use these features while classifying by looking at the boxplot.

Method and Application: Data visualization

Draw a plot of two variables with bivariate and univariate graphs.



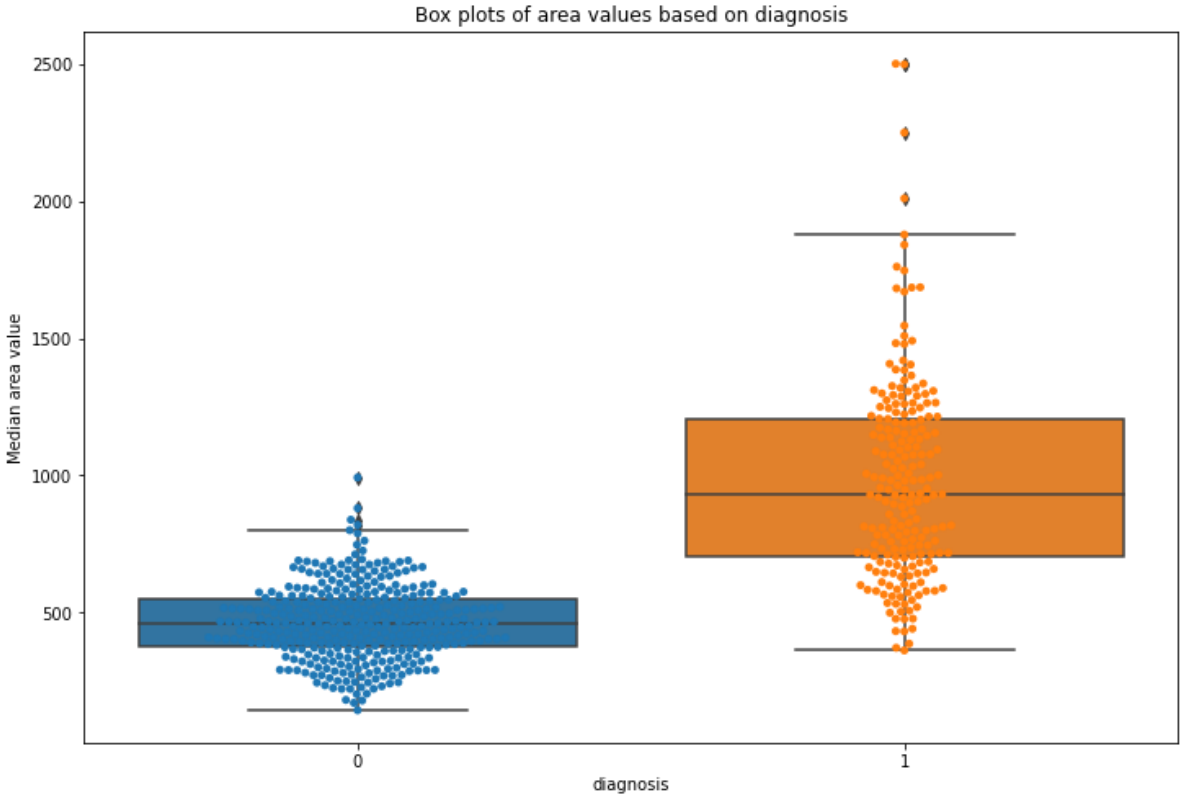
This function provides a convenient interface to the JointGrid class, with several canned plot kinds.

This is intended to be a fairly lightweight wrapper; if you need more flexibility, you should use JointGrid directly.

The figure below show the JoinPlot of area_means per diagnosis.

Method and Application: Data visualization

Box plots of area values based on diagnosis



The figure below show the join of boxplot and swarm plot of area values based on diagnosis

Method and Application: Machine learning

Requirements to Create Good Machine Learning Systems

There are many different types of machine learning algorithms, with hundreds published each day, and they're typically grouped by either learning style (i.e. supervised learning, unsupervised learning, semi-supervised learning) or by similarity in form or function (i.e. classification, regression, decision tree, clustering, deep learning, etc.)

Method and Application: Machine learning

Cleaning , split the data and feature scaling

exploring and cleaning the data. set up the data for the model by first splitting the data set into a feature data set also known as the independent data set (X), and a target data set also known as the dependent data set (Y) , then Split the data into 75% training and 25% testing data sets. Then Scale the data to bring all features to the same level of magnitude, which means the feature / independent data will be within a specific range for example 0–100 or 0–1

```
X = df.iloc[:, 2:31].values
Y = df.iloc[:, 1].values
```

```
[18] #Split the data  into 75% training and 25% testing data sets.
     from sklearn.model_selection import train_test_split
     X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.25, random_state = 0)
```

```
#Scale the data to bring all features to the same level of magnitude,
#which means the feature / independent data will be within a specific range for example 0-100 or 0-1.
#Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

Method and Application: Machine learning

Create function with different models

Create a function to hold many different models (e.g. **Logistic Regression, Decision Tree Classifier, Random Forest Classifier**) to make the classification. These are the models that will detect if a patient has cancer or not.

Within this function I will also print the accuracy of each model on the training data.

```
#create function with different models

def models(X_train,Y_train):

  #Using Logistic Regression
  from sklearn.linear_model import LogisticRegression
  log = LogisticRegression(random_state = 0)
  log.fit(X_train, Y_train)

  #Using KNeighborsClassifier
  from sklearn.neighbors import KNeighborsClassifier
  knn = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p = 2)
  knn.fit(X_train, Y_train)

  #Using SVC linear
  from sklearn.svm import SVC
  svc_lin = SVC(kernel = 'linear', random_state = 0)
  svc_lin.fit(X_train, Y_train)

  #Using SVC rbf
  from sklearn.svm import SVC
  svc_rbf = SVC(kernel = 'rbf', random_state = 0)
  svc_rbf.fit(X_train, Y_train)

  #Using GaussianNB
  from sklearn.naive_bayes import GaussianNB
  gauss = GaussianNB()
  gauss.fit(X_train, Y_train)

  #Using DecisionTreeClassifier
  from sklearn.tree import DecisionTreeClassifier
  tree = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
  tree.fit(X_train, Y_train)

  #Using RandomForestClassifier method of ensemble class to use Random Forest Classification algorithm
  from sklearn.ensemble import RandomForestClassifier
  forest = RandomForestClassifier(n_estimators = 10, criterion = 'entropy', random_state = 0)
  forest.fit(X_train, Y_train)

  #print model accuracy on the training data.
  print('[0]Logistic Regression Training Accuracy:', log.score(X_train, Y_train))
  print('[1]K Nearest Neighbor Training Accuracy:', knn.score(X_train, Y_train))
  print('[2]Support Vector Machine (Linear Classifier) Training Accuracy:', svc_lin.score(X_train, Y_train))
  print('[3]Support Vector Machine (RBF Classifier) Training Accuracy:', svc_rbf.score(X_train, Y_train))
  print('[4]Gaussian Naive Bayes Training Accuracy:', gauss.score(X_train, Y_train))
  print('[5]Decision Tree Classifier Training Accuracy:', tree.score(X_train, Y_train))
  print('[6]Random Forest Classifier Training Accuracy:', forest.score(X_train, Y_train))

  return log, knn, svc_lin, svc_rbf, gauss, tree, forest
```

Method and Application: Machine learning

Create the models

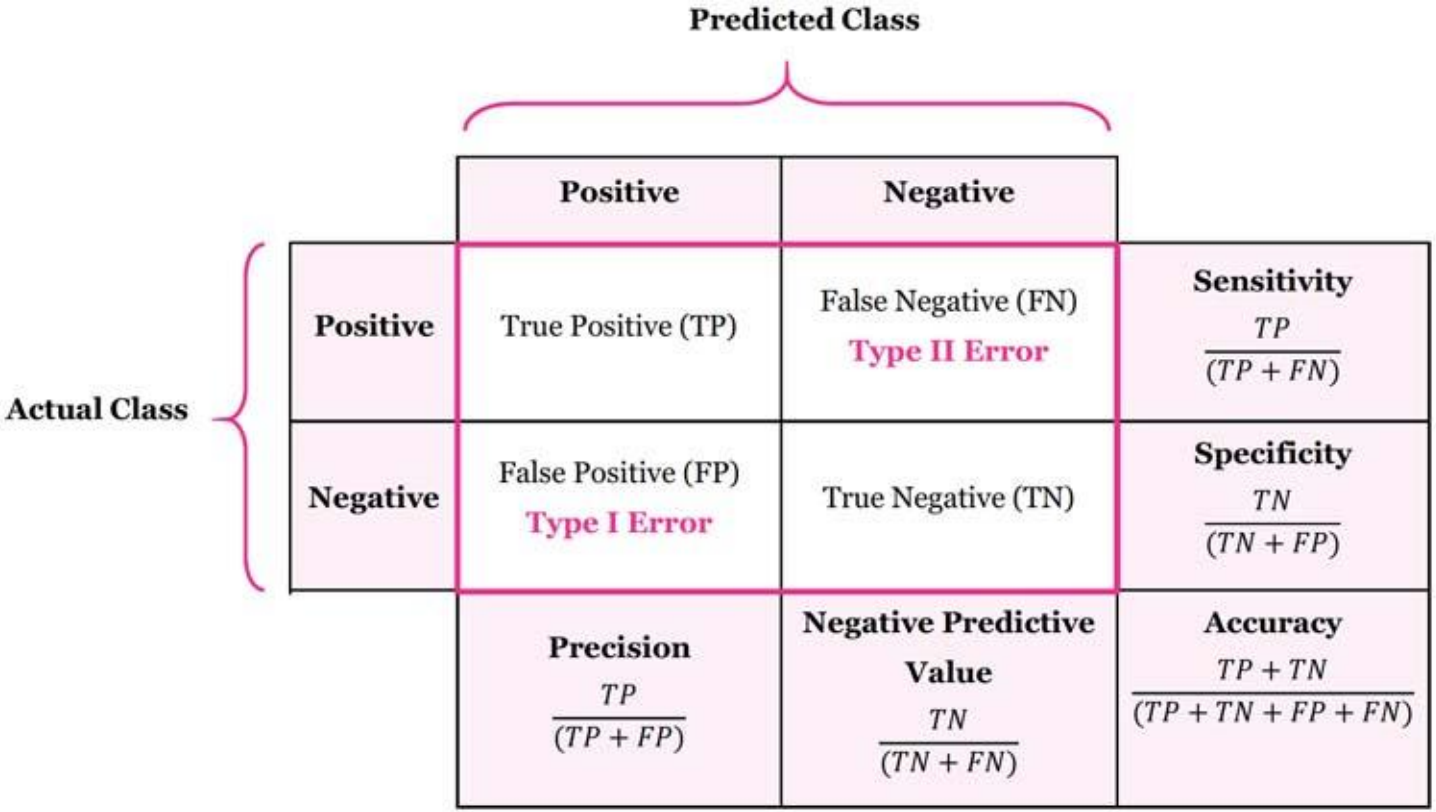
Create the model that contains all of the models, and look at the accuracy score on the training data for each model to classify if a patient has cancer or not.

```
[21] #create my model

     model = models(X_train,Y_train)
```

```
[0]Logistic Regression Training Accuracy: 0.9906103286384976
[1]K Nearest Neighbor Training Accuracy: 0.9765258215962441
[2]Support Vector Machine (Linear Classifier) Training Accuracy: 0.9882629107981221
[3]Support Vector Machine (RBF Classifier) Training Accuracy: 0.9835680751173709
[4]Gaussian Naive Bayes Training Accuracy: 0.9507042253521126
[5]Decision Tree Classifier Training Accuracy: 1.0
[6]Random Forest Classifier Training Accuracy: 0.9953051643192489
```

Method and Application: Machine learning

Confusion Matrix



Outils

PYTHON

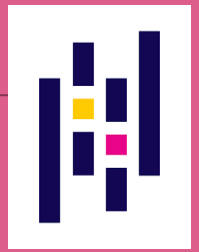
Python est un langage de programmation interprété, multi-paradigme et multiplateforme

Seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive

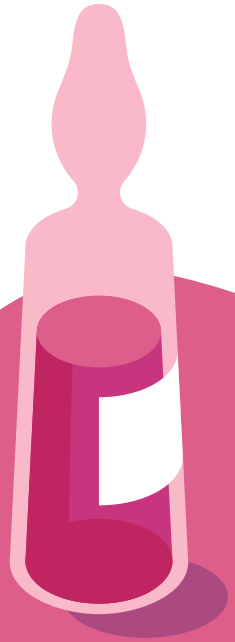
Google COLAB

Google Colab ou Collaboratory est un service cloud, offert par Google, basé sur Jupyter Notebook



pandas

pandas. pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.



05

Result

Result : Machine learning

Show the confusion matrix and the accuracy of the models in the test data

Model 0

	precision	recall	f1-score	support
0	0.96	0.96	0.96	90
1	0.92	0.92	0.92	53
accuracy			0.94	143
macro avg	0.94	0.94	0.94	143
weighted avg	0.94	0.94	0.94	143

0.9440559440559441

Model 1

	precision	recall	f1-score	support
0	0.95	0.99	0.97	90
1	0.98	0.91	0.94	53
accuracy			0.96	143
macro avg	0.96	0.95	0.95	143
weighted avg	0.96	0.96	0.96	143

0.958041958041958

Model 2

	precision	recall	f1-score	support
0	0.98	0.97	0.97	90
1	0.94	0.96	0.95	53
accuracy			0.97	143
macro avg	0.96	0.96	0.96	143
weighted avg	0.97	0.97	0.97	143

0.965034965034965

Model 3

	precision	recall	f1-score	support
0	0.97	0.98	0.97	90
1	0.96	0.94	0.95	53
accuracy			0.97	143
macro avg	0.96	0.96	0.96	143
weighted avg	0.96	0.97	0.96	143

0.965034965034965

Model 4

	precision	recall	f1-score	support
0	0.93	0.94	0.94	90
1	0.90	0.89	0.90	53
accuracy			0.92	143
macro avg	0.92	0.92	0.92	143
weighted avg	0.92	0.92	0.92	143

0.9230769230769231

Model 5

	precision	recall	f1-score	support
0	0.99	0.93	0.96	90
1	0.90	0.98	0.94	53
accuracy			0.95	143
macro avg	0.94	0.96	0.95	143
weighted avg	0.95	0.95	0.95	143

0.951048951048951

Model 6

	precision	recall	f1-score	support
0	0.98	0.97	0.97	90
1	0.94	0.96	0.95	53
accuracy			0.97	143
macro avg	0.96	0.96	0.96	143
weighted avg	0.97	0.97	0.97	143

0.965034965034965

Result : Machine learning

Comparison between confusion matrix and the accuracy of the models in the test data

Model	Testing accuracy
Logistic Regression	0.9440559440559441!
KNeighborsClassifier	0.958041958041958!
SVC linear	0.965034965034965!
SVC rbf	0.965034965034965!
GaussianNB	0.9230769230769231!
DecisionTreeClassifier	0.951048951048951!
RandomForestClassifier	0.965034965034965!

From the accuracy and metrics above, the model that performed the best on the test data was the Random Forest Classifier with an accuracy score of about 96.5%.

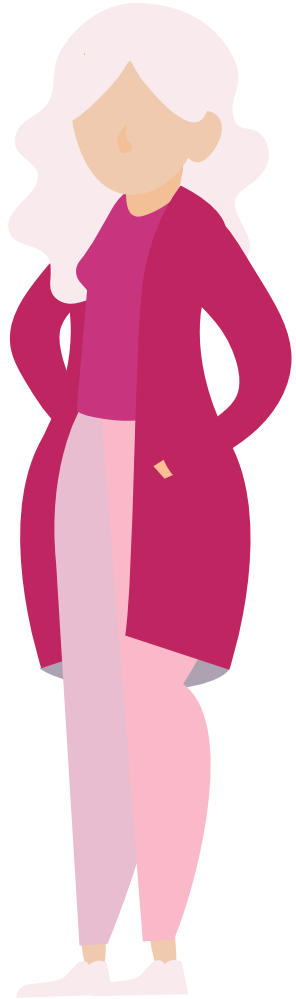
Result : Test

Make the prediction/classification on the test data and show both the Random Forest Classifier model classification/prediction

```
[24] #Print Prediction of Random Forest Classifier model
     pred = model[6].predict(X_test)
     print(pred)

     #Print a space
     print()

     #Print the actual values
     print(Y_test)
```

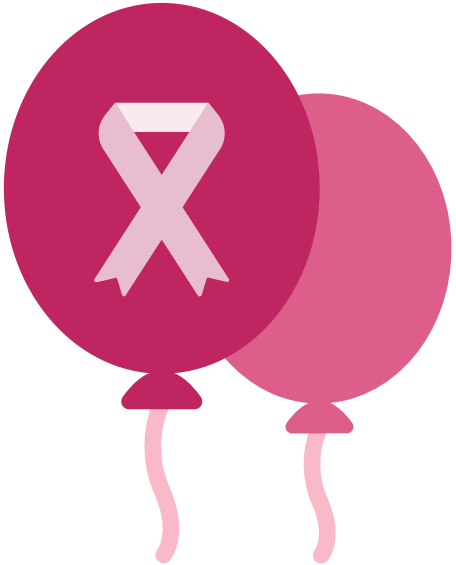
```
[1 0 0 0 0 0 0 0 0 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 0 0 1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 0 0 1 0 0
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 1 0 1 0 1 1 0
 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1]

[1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 1 0 1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 0 0 1 0 1
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 1 0 1 0 1 1 0
 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1]
```

the Model work with an accuracy score of about 96.5% per the test data .



Result : discussion

The model needs more work to improve the accuracy of the models into 100%



```
[24] #Print Prediction of Random Forest Classifier model
     pred = model[6].predict(X_test)
     print(pred)

     #Print a space
     print()

     #Print the actual values
     print(Y_test)
```

```
[1 0 0 0 0 0 0 0 0 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 0 0 1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 1 0 0 1 0 0
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 1 1 0
 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 1]

[1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 1 0 1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 1 0 0 1 0 1
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 1 1 0
 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 1]
```

Conclusions

Healthcare data science a huge domain knowledge will help a professional define what data is essential for the implementation of a certain project and interpret the received results of analytical and modeling work.

Resources

Blogs and websites:

- https://www.cdc.gov/cancer/breast/basic_info/symptoms.htm
- https://www.zeolearn.com/magazine/what-is-data-science
- https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
- https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)
- https://www.wcrf.org/dietandcancer/breast-cancer
- https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
- https://chrisalbon.com/machine_learning/support_vector_machines/svc_parameters_using_rbf_kernel/
- https://www.saedsayad.com/decision_tree.htm
- https://cio-wiki.org/wiki/Machine_Learning
-

Articles :

- TABIB IBRAHIM EL KHALIL, B. S. Etude descriptive et rétrospective des cas de cancer du sein. (2015).
- Coates AS, W. E. Tailoring therapies improving the management of early breast cancer: expert consensus on the primary therapy of early breast cancer . (2015).
- Cohen-Haguenauer et al, O. Hereditary predisposition to breast cancer (1): genetics. (2019).
- Dent R, T. M. Triple negative breast cancer: clinical features and patterns of recurrence . (2007).
- E. Cordina-Duverger, P. G. Épidémiologie des cancers du sein. (2016).
- Greenup R, B. A. Prevalence of BRCA mutations among women with triple negative breast cancer in a genetic counseling cohort. (2013).
- Grogan.G, M. G. Les néoplasies mammaires non invasives et invasives VIIe journées FrancoAfricaines de pathologie. (2003)