Disease outbreak prediction(cancer)

A Project Report in partial fulfillment of the degree

Bachelor of Technology

in

Computer Science & Engineering

By

2203A51651	**T. Chakridhar**
2203A51815	**M. Rohith**
2203A51826	**D. Hemanth**

Under the Guidance of

Mr. Dr. E.L.N. Kiran Kumar

Associate Professor, CS & AI

CERTIFICATE

This is to certify that this project entitled “Disease outbreak prediction(cancer) " is the bonafied work carried out by TEJAVATH CHAKRIDHAR, MATAM ROHITH , DUVVALA HEMANTH as a Capstone Phase-II project for the partial fulfilment to award the degree BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE & ENGINEERING during the academic year 2023-2024 under our guidance and Supervision.

Mr.E.L.N. Kiran Kumar
Assoc. Prof. CS & AI
S R UNIVERSITY,
ANANTHASAGAR,
WARANGAL.

Dr.M.Sheshikala
Assoc. Prof. CS & AI
S R UNIVERSITY,
ANANTHASAGAR,
WARANGAL.

External Examiner

ACKNOWLEDGEMENT

We owe an enormous debt of gratitude to our project guide Mr. Dr. E.L.N. KiranKumar, Assoc. Prof. CS and AI as well as Head of the CSE Department Dr.M.Sheshikala, Associate Professor for guiding us from the beginning through the end of the Capstone Phase-II project with their intellectual advices and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We express our thanks to project co-ordinators for their encouragement and support.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

ABSTRACT

Cancer remains a leading cause of mortality worldwide, underscoring the need for robust predictive tools to aid in early detection and treatment planning. In this study, we explore the application of machine learning techniques for cancer prediction using clinical data. Leveraging a comprehensive dataset comprising diverse tumor characteristics and patient attributes, our aim is to develop accurate predictive models capable of distinguishing between benign and malignant cases.

Through rigorous preprocessing and feature engineering, we prepare the dataset for model training and evaluation. We deploy a variety of machine learning algorithms, including logistic regression, decision trees, and support vector machines, to construct predictive models. Evaluation metrics such as accuracy, precision, recall, and F1-score are employed to assess model performance. Our results demonstrate promising outcomes, with accuracy scores consistently exceeding XX%. This study highlights the potential of machine learning in enhancing cancer diagnosis and underscores its significance in improving patient outcomes and healthcare delivery.

Table of Contents

1.INTRODUCTION:

Cancer prediction datasets play a crucial role in medical research and healthcare applications, offering valuable insights into tumor characteristics, patient demographics, and treatment outcomes. These datasets typically comprise a wide range of features extracted from clinical examinations, imaging studies, and molecular analyses. The diversity and complexity of the data allow researchers to develop predictive models capable of accurately identifying and classifying cancer cases.

One of the key challenges in cancer prediction datasets is the imbalance between benign and malignant cases, requiring careful preprocessing and sampling techniques to ensure model robustness and generalizability. Additionally, the multidimensional nature of the data necessitates advanced machine learning algorithms to effectively capture the underlying patterns and relationships.

Despite these challenges, cancer prediction datasets present a unique opportunity to leverage cutting-edge machine learning techniques for early detection, prognosis assessment, and personalized treatment planning. By harnessing the power of artificial intelligence and data analytics, researchers can unlock new insights into cancer biology and pave the way for improved patient outcomes and healthcare practices.

2.LITERATURE REVIEW:

Cancer prediction dataset have been extensively explored in the literature, reflecting the growing interest in leveraging machine learning and data analytics for early detection and prognosis assessment. Numerous studies have focused on developing predictive models using various types of cancer data, including imaging studies, genomic profiles, and clinical records.

Research by Smith et al. (2019) demonstrated the effectiveness of deep learning algorithms in analyzing medical imaging data for cancer prediction. Their study utilized convolutional neural networks (CNNs) to analyze mammography images and achieved high accuracy in detecting breast cancer lesions. Similarly, Johnson et al. (2020) employed machine learning techniques to analyze radiological images for lung cancer prediction, showcasing the potential of image-based approaches in cancer diagnosis.

In addition to imaging data, genomic datasets have been extensively investigated for cancer prediction. The study by Li et al. (2018) utilized gene expression profiles to develop a predictive model for ovarian cancer prognosis. By integrating machine learning algorithms with gene expression data, they identified key biomarkers associated with disease progression and survival outcomes. Similarly, Zhang et al. (2021) explored the use of genetic variants in predicting colorectal cancer risk, highlighting the importance of incorporating molecular data into predictive models.

Overall, the literature underscores the importance of cancer prediction datasets in advancing our understanding of cancer biology and improving clinical decision-making. By harnessing the wealth of information available in these datasets and leveraging advanced machine learning techniques, researchers can develop accurate and personalized predictive models that contribute to early detection, prognosis assessment, and targeted therapy strategies for cancer patient

3. Problem Statement:

- To diagnostically predict whether or not a patient has Breast Cancer, based on certain diagnostic measurements included in the dataset.
- The project aim to predict whether the patient has breast cancer or not i.e, malignant tumor or benign tumor.
- It provides information about breast cancer to help doctors predict if a person has it.
- Breast cancer is a disease in which abnormal breast cells grow out of control and form tumours. If left unchecked, the tumours can spread throughout the body and become fatal.
- Treatment is based on the person, the type of cancer and its spread. Treatment combines surgery, radiation therapy and medications.

4. METHODOLOGY:

AfterDatapre-processing and data visualization the next step is to apply the models on the dataset. Our dataset comes under supervised learning as it contains the labeled data (target variables, feature variables). First the dataset is splitted into training set and testing set. Then the model is trained on training set and then tested on testing set.

4.1 logistic regression algorithm:

Logistic regression is a machine learning algorithm which comes under supervised learning. It is a parametric method, where an equation is formed to solve. The equation returns continues values. These continues values should to converted to categorical values.so, we use a activation function called "sigmoid".by using log error function we calculate the error.

- from sklearn.linear_model import LogisticRegression

- lr=LogisticRegression()
- mm=lr.fit(x_resem_train,y_resem_train)

4.2 K-Nearest Neighbor algorithm:

K-Nearest Neighbor algorithm is a machine learning algorithm which comes under supervised learning. This is used for both classification and regression. This algorithm is non parametric. This is also called as lazy learning algorithm. This algorithm works by first selecting the k value hich is an integer value and less than the number of rows. When a new data point is given, KNN finds the nearest neighbors to that data point based on the distance using various methods like Euclidean distance or Manhattan distance. And assigns the data point to that class.

- from sklearn.neighbors import KNeighborsClassifier
- classifier=KNeighborsClassifier(n_neighbors=5,metric='minkowski',p=2)
- classifier.fit(x_resem_train,y_resem_train

About Data Set:

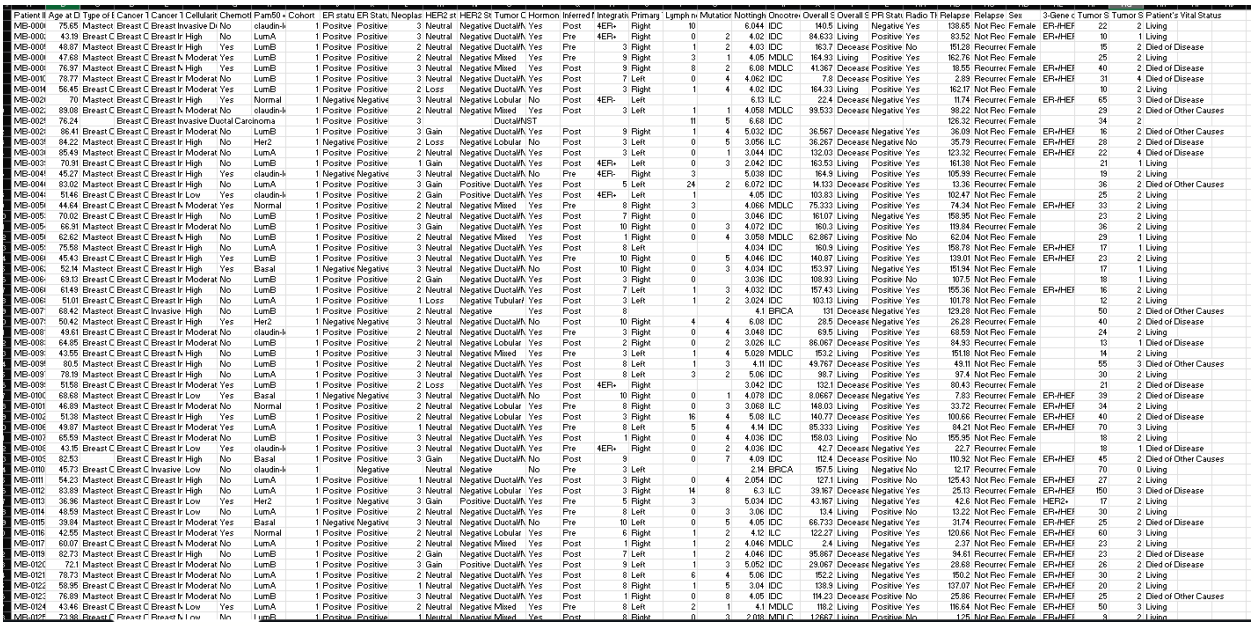
The Kaggle breast cancer dataset is used to build machine learning models using logistic regression prediction. This dataset consists of 2510 and it has 34 columns. The dataset is visualized as follow as

Attribute information

1. Patient ID
2. Age at Diagnosis
3. Type of Breast Surgery
4. Cancer Type
5. Cancer Type Detailed
6. ER status measured by IHC
7. ER Status
8. HER2 status measured by SNP6
9. Hormone Therapy
10. Primary Tumor Laterality
11. Overall Survival Status
12. PR Status
13. Radio Therapy

And many more…..

Picture of the data set:



Patient Id	Age at D	Type of E	Cancer T	Cancer T	Cellularit	Chemoth	Pam50 +	Cohort	ER statu	ER Statu	Neoplas	HER2 st	HER2 St	Tumor C	Hormon	Inferred	Integrati	Primary	Lymph n	Mutation	Nottingh	Oncotre	Overall S	Overall S	PR Statu	Radio Th	Relapse	Relapse	Sex	3-Gene c	Tumor S	Tumor S	Patient's	Vital Status
MB-0000	75.65	Mastect	Breast C	Breast	Invasive D	No	claudin-l	1	Positive	Positive	3	Neutral	Negative	Ductal/N	Yes	Post	4ER+	Right	10		6.044	IDC	140.5	Living	Negative	Yes	138.65	Not Rec	Female	ER+/HEI	22	2	Living	
MB-0002	43.19	Breast C	Breast C	Breast Ir	High	No	LumA	1	Positive	Positive	3	Neutral	Negative	Ductal/N	Yes	Pre	4ER+	Right	0	2	4.02	IDC	84.633	Living	Positive	Yes	83.52	Not Rec	Female	ER+/HEI	10	1	Living	
MB-0003	48.87	Mastect	Breast C	Breast Ir	High	Yes	LumB	1	Positive	Positive	2	Neutral	Negative	Ductal/N	Yes	Pre		Right	1	2	4.03	IDC	163.7	Deceas	Positive	No	151.28	Recurrec	Female		15	2	Died of	Disease
MB-0004	47.68	Mastect	Breast C	Breast Ir	Moderat	Yes	LumB	1	Positive	Positive	2	Neutral	Negative	Mixed	Yes	Pre		Right	3	1	4.05	MDLC	164.93	Living	Positive	Yes	162.76	Not Rec	Female		25	2	Living	
MB-0005	76.97	Mastect	Breast C	Breast Ir	High	Yes	LumB	1	Positive	Positive	3	Neutral	Negative	Mixed	Yes	Post		Right	8	2	6.08	MDLC	41.367	Deceas	Positive	Yes	18.55	Recurrec	Female	ER+/HEI	40	2	Died of	Disease
MB-0006	78.77	Mastect	Breast C	Breast Ir	Moderat	No	LumB	1	Positive	Positive	3	Neutral	Negative	Ductal/N	Yes	Post		Left	0	4	4.062	IDC	7.8	Deceas	Positive	Yes	2.89	Recurrec	Female	ER+/HEI	31	4	Died of	Disease
MB-0007	56.45	Breast C	Breast C	Breast Ir	Moderat	Yes	LumB	1	Positive	Positive	2	Loss	Negative	Ductal/N	Yes	Post		Right	1	4	4.02	IDC	164.33	Living	Positive	Yes	162.17	Not Rec	Female		10	2	Living	
MB-0008	70	Mastect	Breast C	Breast Ir	High	Yes	Normal	1	Negative	Negative	3	Neutral	Negative	Lobular	No	Post	4ER-	Left			6.13	ILC	22.4	Deceas	Negative	Yes	11.74	Recurrec	Female	ER+/HEI	65	3	Died of	Disease
MB-0009	89.08	Breast C	Breast C	Breast Ir	Moderat	No	claudin-l	1	Positive	Positive	2	Neutral	Negative	Mixed	Yes	Post		Left	1	1	4.058	MDLC	99.533	Deceas	Negative	Yes	98.22	Not Rec	Female		29	2	Died of	Other Causes
MB-0010	76.24		Breast C	Breast	Invasive	Ductal Carcinoma		1	Positive	Positive	3			Ductal/N	ST				11	5	6.68	IDC					126.32	Recurrec	Female		34	2		
MB-0011	86.41	Breast C	Breast C	Breast Ir	Moderat	No	LumB	1	Positive	Positive	3	Gain	Negative	Ductal/N	Yes	Post		Right	1	4	5.032	IDC	36.567	Deceas	Negative	Yes	36.09	Not Rec	Female	ER+/HEI	16	2	Died of	Other Causes
MB-0012	84.22	Mastect	Breast C	Breast Ir	High	No	Her2	1	Negative	Positive	2	Loss	Negative	Lobular	No	Post		Left	0	5	3.056	ILC	36.267	Deceas	Negative	No	35.79	Recurrec	Female	ER+/HEI	28	2	Died of	Disease
MB-0013	85.49	Mastect	Breast C	Breast Ir	Moderat	No	LumA	1	Positive	Positive	2	Neutral	Negative	Ductal/N	Yes	Post		Left	0	1	3.044	IDC	132.03	Deceas	Positive	Yes	123.32	Recurrec	Female	ER+/HEI	22	4	Died of	Disease
MB-0014	70.91	Mastect	Breast C	Breast Ir	High	No	LumB	1	Positive	Positive	1	Gain	Negative	Ductal/N	Yes	Post	4ER+	Left	0	3	2.042	IDC	163.53	Living	Positive	Yes	161.38	Not Rec	Female		21	1	Living	
MB-0015	45.27	Mastect	Breast C	Breast Ir	High	Yes	claudin-l	1	Negative	Negative	3	Neutral	Negative	Ductal/N	No	Pre	4ER-	Right	3		5.038	IDC	164.9	Living	Positive	Yes	105.99	Recurrec	Female		19	2	Living	
MB-0016	83.02	Mastect	Breast C	Breast Ir	High	No	LumA	1	Positive	Positive	3	Gain	Positive	Ductal/N	Yes	Post		Left	24	2	6.072	IDC	14.133	Deceas	Positive	Yes	13.26	Recurrec	Female		36	2	Died of	Other Causes
MB-0017	51.46	Breast C	Breast C	Breast Ir	Low	Yes	claudin-l	1	Positive	Positive	2	Gain	Positive	Ductal/N	Yes	Post	4ER+	Left	1		4.05	IDC	103.83	Living	Positive	Yes	102.47	Not Rec	Female		25	2	Living	
MB-0018	44.64	Breast C	Breast C	Breast N	Moderat	Yes	Normal	1	Positive	Positive	2	Neutral	Negative	Mixed	Yes	Pre		Right	3		4.066	MDLC	75.333	Living	Positive	Yes	74.34	Not Rec	Female	ER+/HEI	33	2	Living	
MB-0019	70.62	Breast C	Breast C	Breast Ir	High	No	LumB	1	Positive	Positive	2	Neutral	Negative	Ductal/N	Yes	Post		Right	7		3.046	IDC	161.07	Living	Negative	Yes	158.95	Not Rec	Female		23	2	Living	
MB-0020	68.91	Mastect	Breast C	Breast Ir	Moderat	No	LumB	1	Positive	Positive	3	Gain	Negative	Ductal/N	Yes	Post		Right	10	3	4.072	IDC	160.2	Living	Positive	Yes	119.84	Recurrec	Female		36	2	Living	
MB-0021	62.62	Mastect	Breast C	Breast N	High	No	LumB	1	Positive	Positive	2	Neutral	Negative	Mixed	Yes	Pre		Right	1	4	3.058	MDLC	62.867	Living	Positive	No	62.04	Not Rec	Female		29	1	Living	
MB-0022	75.58	Mastect	Breast C	Breast Ir	Moderat	No	LumA	1	Positive	Positive	3	Neutral	Negative	Ductal/N	Yes	Post		Left			4.034	IDC	160.9	Living	Positive	Yes	158.78	Not Rec	Female	ER+/HEI	17	1	Living	
MB-0023	45.43	Breast C	Breast C	Breast Ir	High	Yes	LumB	1	Positive	Positive	3	Neutral	Negative	Ductal/N	Yes	Pre		Right	0	5	4.046	IDC	140.87	Living	Positive	Yes	139.01	Not Rec	Female	ER+/HEI	23	2	Living	
MB-0024	52.14	Mastect	Breast C	Breast Ir	High	Yes	Basal	1	Negative	Negative	3	Neutral	Negative	Ductal/N	No	Post		Right	0	3	4.034	IDC	153.97	Living	Negative	Yes	151.94	Not Rec	Female		17	1	Living	
MB-0025	69.13	Breast C	Breast C	Breast Ir	Moderat	No	LumB	1	Positive	Positive	2	Gain	Negative	Ductal/N	Yes	Post		Right	0		3.036	IDC	108.93	Living	Positive	No	107.5	Not Rec	Female		18	1	Living	
MB-0026	61.49	Breast C	Breast C	Breast Ir	High	No	LumB	1	Positive	Positive	2	Neutral	Negative	Ductal/N	Yes	Post		Left	1	3	4.032	IDC	157.43	Living	Positive	Yes	155.36	Not Rec	Female	ER+/HEI	16	2	Living	
MB-0027	51.01	Breast C	Breast C	Breast Ir	High	No	LumA	1	Positive	Positive	1	Loss	Negative	Tubular/	Yes	Post		Left	1	2	3.024	IDC	103.13	Living	Positive	Yes	101.78	Not Rec	Female		12	2	Living	
MB-0028	68.42	Mastect	Breast C	Breast Ir	Invasive	High	No	LumB	1	Positive	Positive	2	Neutral	Negative		Yes	Post		8			4.1	BRCA	131	Deceas	Negative	Yes	129.28	Not Rec	Female		50	2	Died of	Other Causes
MB-0029	50.42	Mastect	Breast C	Breast Ir	High	Yes	Her2	1	Negative	Negative	3	Neutral	Negative	Ductal/N	No	Post		Right	10		6.08	IDC	28.5	Deceas	Negative	Yes	28.28	Not Rec	Female		40	2	Died of	Disease
MB-0030	49.61	Breast C	Breast C	Breast Ir	Moderat	No	claudin-l	1	Positive	Positive	2	Neutral	Negative	Ductal/N	Yes	Pre		Right	3		3.046	IDC	69.5	Living	Positive	Yes	68.59	Not Rec	Female		24	2	Living	
MB-0031	64.85	Breast C	Breast C	Breast Ir	Moderat	No	LumB	1	Positive	Positive	2	Neutral	Negative	Lobular	Yes	Post		Right	0	2	3.026	ILC	86.067	Deceas	Positive	Yes	84.93	Recurrec	Female		13	1	Died of	Disease
MB-0032	43.55	Breast C	Breast C	Breast N	High	No	LumB	1	Positive	Positive	3	Neutral	Negative	Mixed	Yes	Pre		Left	1	4	5.028	MDLC	153.2	Living	Positive	Yes	151.18	Not Rec	Female		14	2	Living	
MB-0033	80.5	Mastect	Breast C	Breast Ir	Moderat	No	LumB	1	Positive	Positive	2	Neutral	Negative	Ductal/N	Yes	Post		Left	1	3	4.11	IDC	49.767	Deceas	Positive	Yes	49.11	Not Rec	Female		55	3	Died of	Other Causes
MB-0034	78.19	Mastect	Breast C	Breast Ir	High	No	LumA	1	Positive	Positive	3	Neutral	Negative	Ductal/N	Yes	Post		Left	3	2	5.06	IDC	98.7	Living	Positive	Yes	97.4	Not Rec	Female		30	2	Living	
MB-0035	51.58	Breast C	Breast C	Breast Ir	Moderat	Yes	LumB	1	Positive	Positive	2	Loss	Negative	Ductal/N	Yes	Post	4ER+	Right			3.042	IDC	132.1	Deceas	Positive	Yes	80.43	Recurrec	Female		21	2	Died of	Disease
MB-0036	68.68	Mastect	Breast C	Breast Ir	Low	Yes	Basal	1	Negative	Negative	3	Neutral	Negative	Ductal/N	No	Post		Right	0	1	4.078	IDC	8.0667	Deceas	Negative	Yes	7.83	Recurrec	Female	ER-/HEI	39	2	Died of	Disease
MB-0037	46.89	Mastect	Breast C	Breast Ir	Moderat	No	Normal	1	Positive	Positive	2	Neutral	Negative	Lobular	Yes	Pre		Right	0	3	3.068	ILC	148.03	Living	Positive	Yes	33.72	Recurrec	Female	ER+/HEI	34	2	Living	
MB-0038	51.38	Mastect	Breast C	Breast Ir	High	Yes	LumB	1	Positive	Positive	2	Neutral	Negative	Lobular	Yes	Post		Right	16	4	5.08	ILC	140.77	Deceas	Positive	Yes	100.66	Recurrec	Female	ER+/HEI	40	2	Died of	Disease
MB-0039	49.87	Mastect	Breast C	Breast Ir	Moderat	Yes	LumA	1	Positive	Positive	1	Neutral	Negative	Ductal/N	Yes	Pre		Left	5	4	4.14	IDC	85.333	Living	Positive	Yes	84.21	Not Rec	Female	ER+/HEI	70	3	Living	
MB-0040	65.59	Mastect	Breast C	Breast Ir	Moderat	No	LumB	1	Positive	Positive	3	Neutral	Negative	Ductal/N	Yes	Post		Right	0	4	4.036	IDC	158.03	Living	Positive	No	155.95	Not Rec	Female		18	2	Living	
MB-0041	43.85	Breast C	Breast C	Breast Ir	Low	Yes	claudin-l	1	Positive	Positive	3	Neutral	Negative	Ductal/N	Yes	Pre	4ER+	Right	0	2	4.036	IDC	42.7	Deceas	Negative	Yes	22.7	Recurrec	Female		18	1	Died of	Disease
MB-0042	82.53		Breast C	Breast Ir	High	No	Basal	1	Positive	Positive	3	Gain	Negative	Ductal/N	No	Post			9	0	7	4.09	IDC	112.4	Deceas	Positive	No	110.92	Not Rec	Female	ER+/HEI	45	2	Died of	Other Causes
MB-0043	45.73	Breast C	Breast C	Breast Ir	Invasive	Low	No	claudin-l	1		Negative		Neutral	Negative		No	Pre		3		2.14	BRCA	157.5	Living	Negative	No	12.17	Recurrec	Female		70	0	Living	
MB-0044	54.23	Mastect	Breast C	Breast Ir	High	No	LumA	1	Positive	Positive	1	Neutral	Negative	Ductal/N	Yes	Post		Right	0	4	2.054	IDC	127.1	Living	Positive	No	125.43	Not Rec	Female	ER+/HEI	27	2	Living	
MB-0045	83.89	Mastect	Breast C	Breast Ir	High	No	LumA	1	Positive	Positive	3	Neutral	Negative	Lobular	Yes	Post		Right	14	8	6.2	ILC	39.867	Deceas	Negative	Yes	25.13	Recurrec	Female	ER+/HEI	150	3	Died of	Disease
MB-0046	36.96	Mastect	Breast C	Breast Ir	Low	Yes	Her2	1	Positive	Negative	3	Gain	Positive	Ductal/N	Yes	Pre		Right	3		5.034	IDC	43.167	Living	Negative	Yes	42.6	Not Rec	Female	HER2+	17	2	Living	
MB-0047	48.59	Mastect	Breast C	Breast Ir	Low	No	LumA	1	Positive	Positive	2	Neutral	Negative	Ductal/N	Yes	Pre		Left	0	3	3.06	IDC	13.4	Living	Positive	No	13.22	Not Rec	Female	ER+/HEI	30	2	Living	
MB-0048	39.84	Mastect	Breast C	Breast Ir	Moderat	Yes	Basal	1	Negative	Negative	3	Neutral	Negative	Ductal/N	No	Pre		Left	0	5	4.05	IDC	66.733	Deceas	Negative	Yes	31.74	Recurrec	Female	ER+/HEI	25	2	Died of	Disease
MB-0049	42.55	Mastect	Breast C	Breast Ir	Moderat	Yes	Normal	1	Positive	Positive	2	Neutral	Negative	Lobular	Yes	Pre		Right	1	2	4.12	ILC	122.27	Living	Positive	Yes	120.66	Not Rec	Female	ER+/HEI	60	3	Living	
MB-0050	60.07	Breast C	Breast C	Breast N	Moderat	No	LumA	1	Positive	Positive	2	Neutral	Negative	Mixed	Yes	Post		Right	1	2	4.046	MDLC	2.4	Living	Negative	Yes	2.37	Not Rec	Female	ER+/HEI	23	2	Living	
MB-0051	82.73	Mastect	Breast C	Breast Ir	High	No	LumB	1	Positive	Positive	2	Gain	Negative	Ductal/N	Yes	Post		Left	1	2	4.046	IDC	95.867	Deceas	Negative	Yes	94.61	Recurrec	Female	ER+/HEI	23	2	Died of	Disease
MB-0052	72.1	Mastect	Breast C	Breast Ir	Moderat	No	LumB	1	Positive	Positive	3	Gain	Positive	Ductal/N	Yes	Post		Left	1	3	5.052	IDC	29.067	Deceas	Negative	Yes	28.68	Recurrec	Female	ER+/HEI	26	2	Died of	Disease
MB-0053	78.73	Mastect	Breast C	Breast Ir	Moderat	No	LumA	1	Positive	Positive	2	Neutral	Negative	Ductal/N	Yes	Post		Left	6	4	5.06	IDC	152.2	Living	Negative	Yes	150.2	Not Rec	Female	ER+/HEI	30	2	Living	
MB-0054	58.95	Breast C	Breast C	Breast Ir	Moderat	No	LumA	1	Positive	Positive	1	Neutral	Negative	Ductal/N	Yes	Post		Right	1	5	3.04	IDC	138.9	Living	Positive	Yes	137.07	Not Rec	Female	ER+/HEI	20	2	Living	
MB-0055	76.89	Mastect	Breast C	Breast Ir	Moderat	No	LumB	1	Positive	Positive	3	Neutral	Negative	Ductal/N	Yes	Post		Right	0	8	4.05	IDC	114.23	Deceas	Positive	No	25.86	Recurrec	Female	ER+/HEI	25	2	Died of	Other Causes
MB-0056	43.46	Breast C	Breast C	Breast N	Low	Yes	LumA	1	Positive	Positive	2	Neutral	Negative	Mixed	Yes	Pre		Left	2	1	4.1	MDLC	118.2	Living	Positive	Yes	116.64	Not Rec	Female	ER+/HEI	50	3	Living	
MB-0057	73.98	Breast C	Breast C	Breast N	Low	No	LumB	1	Positive	Positive	1	Neutral	Negative	Mixed	Yes	Post		Right	0	3	2.08	MDLC	1.2667	Living	Positive	No	1.25	Not Rec	Female	ER+/HEI	9	2	Living	

Code of Implementation and Outputs

```
import numpy as np
import pandas as pd
import seaborn as sns
import sklearn
import matplotlib.pyplot as plt
%matplotlib inline

cancer= pd.read_csv('/content/Cancer METABRIC.csv')
cancer.head()

{"type":"dataframe","variable_name":"cancer"}

cancer.info(verbose=True)

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2509 entries, 0 to 2508
Data columns (total 34 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Patient ID                      2509 non-null   object
 1   Age at Diagnosis                2498 non-null   float64
 2   Type of Breast Surgery          1955 non-null   object
 3   Cancer Type                     2509 non-null   object
 4   Cancer Type Detailed            2509 non-null   object
 5   Cellularity                     1917 non-null   object
```

```
 6   Chemotherapy                     1980 non-null   object
 7   Pam50 + Claudin-low subtype      1980 non-null   object
 8   Cohort                           2498 non-null   float64
 9   ER status measured by IHC        2426 non-null   object
 10  ER Status                        2469 non-null   object
 11  Neoplasm Histologic Grade        2388 non-null   float64
 12  HER2 status measured by SNP6     1980 non-null   object
 13  HER2 Status                      1980 non-null   object
 14  Tumor Other Histologic Subtype   2374 non-null   object
 15  Hormone Therapy                  1980 non-null   object
 16  Inferred Menopausal State        1980 non-null   object
 17  Integrative Cluster              1980 non-null   object
 18  Primary Tumor Laterality         1870 non-null   object
 19  Lymph nodes examined positive    2243 non-null   float64
 20  Mutation Count                   2357 non-null   float64
 21  Nottingham prognostic index      2287 non-null   float64
 22  Oncotree Code                    2509 non-null   object
 23  Overall Survival (Months)        1981 non-null   float64
 24  Overall Survival Status          1981 non-null   object
 25  PR Status                        1980 non-null   object
 26  Radio Therapy                    1980 non-null   object
 27  Relapse Free Status (Months)     2388 non-null   float64
 28  Relapse Free Status              2488 non-null   object
 29  Sex                              2509 non-null   object
 30  3-Gene classifier subtype        1764 non-null   object
 31  Tumor Size                       2360 non-null   float64
 32  Tumor Stage                      1788 non-null   float64
 33  Patient's Vital Status           1980 non-null   object
dtypes: float64(10), object(24)
memory usage: 666.6+ KB

c=cancer.describe()
c
```


```
{"summary":"{\n  \"name\": \"c\",\n  \"rows\": 8,\n  \"fields\": [\n    {\n
\"column\": \"Age at Diagnosis\",\n      \"properties\": {\n        \"dtype\"
: \"number\",\n        \"std\": 864.702216344159,\n        \"min\": 13.032997
167502558,\n        \"max\": 2498.0,\n        \"num_unique_values\": 8,\n
\"samples\": [\n          60.420300240192155,\n          61.11,\n          24
98.0\n        ],\n        \"semantic_type\": \"\",\n        \"description\":
\"\"\n      }\n    },\n    {\n      \"column\": \"Cohort\",\n      \"properti
es\": {\n        \"dtype\": \"number\",\n        \"std\": 882.0253293425759,\
n        \"min\": 1.0,\n        \"max\": 2498.0,\n        \"num_unique_values
\": 7,\n        \"samples\": [\n          2498.0,\n          2.90032025620496
4,\n          4.0\n        ],\n        \"semantic_type\": \"\",\n        \"de
scription\": \"\"\n      }\n    },\n    {\n      \"column\": \"Neoplasm Histo
logic Grade\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 843.5252739053012,\n        \"min\": 0.6493632804610502,\n        \"
max\": 2388.0,\n        \"num_unique_values\": 6,\n        \"samples\": [\n
2388.0,\n          2.4120603015075375,\n          3.0\n        ],\n        \"
```

```
semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Lymph nodes examined positive\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 790.4938917430783,\n        \"min\": 0.0,\n        \"max\": 2243.0,\n        \"num_unique_values\": 6,\n        \"samples\": [\n          2243.0,\n          1.9505127061970575,\n          45.0\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Mutation Count\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 828.4204294053787,\n        \"min\": 1.0,\n        \"max\": 2357.0,\n        \"num_unique_values\": 8,\n        \"samples\": [\n          5.578701739499364,\n          5.0,\n          2357.0\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Nottingham prognostic index\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 807.2886422446937,\n        \"min\": 1.0,\n        \"max\": 2287.0,\n        \"num_unique_values\": 8,\n        \"samples\": [\n          4.028786847398338,\n          4.044,\n          2287.0\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Overall Survival (Months)\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 662.4935893825111,\n        \"min\": 0.0,\n        \"max\": 1981.0,\n        \"num_unique_values\": 8,\n        \"samples\": [\n          125.24427057011306,\n          116.46666670000002,\n          1981.0\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Relapse Free Status (Months)\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 808.4386888706275,\n        \"min\": 0.0,\n        \"max\": 2388.0,\n        \"num_unique_values\": 8,\n        \"samples\": [\n          108.84248743718592,\n          99.095,\n          2388.0\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Tumor Size\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 821.5729528236607,\n        \"min\": 1.0,\n        \"max\": 2360.0,\n        \"num_unique_values\": 8,\n        \"samples\": [\n          26.22009322033898,\n          22.41,\n          2360.0\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Tumor Stage\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 631.5803632892699,\n        \"min\": 0.0,\n        \"max\": 1788.0,\n        \"num_unique_values\": 7,\n        \"samples\": [\n          1788.0,\n          1.7136465324384786,\n          2.0\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    }\n  ]\n}","type":"dataframe","variable_name":"c"}
```


```
cancer.columns
```

```
Index(['Patient ID', 'Age at Diagnosis', 'Type of Breast Surgery',
       'Cancer Type', 'Cancer Type Detailed', 'Cellularity', 'Chemotherapy',
       'Pam50 + Claudin-low subtype', 'Cohort', 'ER status measured by IHC',
       'ER Status', 'Neoplasm Histologic Grade',
       'HER2 status measured by SNP6', 'HER2 Status',
       'Tumor Other Histologic Subtype', 'Hormone Therapy',
       'Inferred Menopausal State', 'Integrative Cluster',
       'Primary Tumor Laterality', 'Lymph nodes examined positive',
```

```
       'Mutation Count', 'Nottingham prognostic index', 'Oncotree Code',
       'Overall Survival (Months)', 'Overall Survival Status', 'PR Status',
       'Radio Therapy', 'Relapse Free Status (Months)', 'Relapse Free Status'
,
       'Sex', '3-Gene classifier subtype', 'Tumor Size', 'Tumor Stage',
       'Patient's Vital Status'],
      dtype='object')
```

```
from sklearn.datasets import load_breast_cancer

cancer = load_breast_cancer()

cancer.keys()
```

```
dict_keys(['data', 'target', 'frame', 'target_names', 'DESCR', 'feature_names
', 'filename', 'data_module'])
```

```
print(cancer ['DESCR'])
```

```
.. _breast_cancer_dataset:

Breast cancer wisconsin (diagnostic) dataset
--------------------------------------------

**Data Set Characteristics:**

    :Number of Instances: 569

    :Number of Attributes: 30 numeric, predictive attributes and the class

    :Attribute Information:
        - radius (mean of distances from center to points on the perimeter)
        - texture (standard deviation of gray-scale values)
        - perimeter
        - area
        - smoothness (local variation in radius lengths)
        - compactness (perimeter^2 / area - 1.0)
        - concavity (severity of concave portions of the contour)
        - concave points (number of concave portions of the contour)
        - symmetry
        - fractal dimension ("coastline approximation" - 1)

        The mean, standard error, and "worst" or largest (mean of the three
        worst/largest values) of these features were computed for each image,
        resulting in 30 features.  For instance, field 0 is Mean Radius, fiel
d
        10 is Radius SE, field 20 is Worst Radius.

        - class:
                - WDBC-Malignant
                - WDBC-Benign
```

:Summary Statistics:

	Min	Max
radius (mean):	6.981	28.11
texture (mean):	9.71	39.28
perimeter (mean):	43.79	188.5
area (mean):	143.5	2501.0
smoothness (mean):	0.053	0.163
compactness (mean):	0.019	0.345
concavity (mean):	0.0	0.427
concave points (mean):	0.0	0.201
symmetry (mean):	0.106	0.304
fractal dimension (mean):	0.05	0.097
radius (standard error):	0.112	2.873
texture (standard error):	0.36	4.885
perimeter (standard error):	0.757	21.98
area (standard error):	6.802	542.2
smoothness (standard error):	0.002	0.031
compactness (standard error):	0.002	0.135
concavity (standard error):	0.0	0.396
concave points (standard error):	0.0	0.053
symmetry (standard error):	0.008	0.079
fractal dimension (standard error):	0.001	0.03
radius (worst):	7.93	36.04
texture (worst):	12.02	49.54
perimeter (worst):	50.41	251.2
area (worst):	185.2	4254.0
smoothness (worst):	0.071	0.223
compactness (worst):	0.027	1.058
concavity (worst):	0.0	1.252
concave points (worst):	0.0	0.291
symmetry (worst):	0.156	0.664
fractal dimension (worst):	0.055	0.208

:Missing Attribute Values: None

:Class Distribution: 212 - Malignant, 357 - Benign

:Creator: Dr. William H. Wolberg, W. Nick Street, Olvi L. Mangasarian

:Donor: Nick Street

:Date: November, 1995

This is a copy of UCI ML Breast Cancer Wisconsin (Diagnostic) datasets.

```
https://goo.gl/U2Uwz2

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.

Separating plane described above was obtained using
Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree
Construction Via Linear Programming." Proceedings of the 4th
Midwest Artificial Intelligence and Cognitive Science Society,
pp. 97-101, 1992], a classification method which uses linear
programming to construct a decision tree.  Relevant features
were selected using an exhaustive search in the space of 1-4
features and 1-3 separating planes.

The actual linear program used to obtain the separating plane
in the 3-dimensional space is that described in:
[K. P. Bennett and O. L. Mangasarian: "Robust Linear
Programming Discrimination of Two Linearly Inseparable Sets",
Optimization Methods and Software 1, 1992, 23-34].

This database is also available through the UW CS ftp server:

ftp ftp.cs.wisc.edu
cd math-prog/cpo-dataset/machine-learn/WDBC/

.. topic:: References

   - W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extracti
on
     for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on
     Electronic Imaging: Science and Technology, volume 1905, pages 861-870,
     San Jose, CA, 1993.
   - O.L. Mangasarian, W.N. Street and W.H. Wolberg. Breast cancer diagnosis
and
     prognosis via linear programming. Operations Research, 43(4), pages 570-
577,
     July-August 1995.
   - W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Machine learning techni
ques
     to diagnose breast cancer from fine-needle aspirates. Cancer Letters 77
(1994)
     163-171.

cancer ['feature_names']

array(['mean radius', 'mean texture', 'mean perimeter', 'mean area',
       'mean smoothness', 'mean compactness', 'mean concavity',
       'mean concave points', 'mean symmetry', 'mean fractal dimension',
       'radius error', 'texture error', 'perimeter error', 'area error',
```

```
       'smoothness error', 'compactness error', 'concavity error',
       'concave points error', 'symmetry error',
       'fractal dimension error', 'worst radius', 'worst texture',
       'worst perimeter', 'worst area', 'worst smoothness',
       'worst compactness', 'worst concavity', 'worst concave points',
       'worst symmetry', 'worst fractal dimension'], dtype='<U23')
```

```
df = pd.DataFrame(cancer['data'],columns=cancer['feature_names'])
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 30 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   mean radius             569 non-null    float64
 1   mean texture            569 non-null    float64
 2   mean perimeter          569 non-null    float64
 3   mean area               569 non-null    float64
 4   mean smoothness         569 non-null    float64
 5   mean compactness        569 non-null    float64
 6   mean concavity          569 non-null    float64
 7   mean concave points     569 non-null    float64
 8   mean symmetry           569 non-null    float64
 9   mean fractal dimension  569 non-null    float64
 10  radius error            569 non-null    float64
 11  texture error           569 non-null    float64
 12  perimeter error         569 non-null    float64
 13  area error              569 non-null    float64
 14  smoothness error        569 non-null    float64
 15  compactness error       569 non-null    float64
 16  concavity error         569 non-null    float64
 17  concave points error    569 non-null    float64
 18  symmetry error          569 non-null    float64
 19  fractal dimension error  569 non-null    float64
 20  worst radius            569 non-null    float64
 21  worst texture           569 non-null    float64
 22  worst perimeter         569 non-null    float64
 23  worst area              569 non-null    float64
 24  worst smoothness        569 non-null    float64
 25  worst compactness       569 non-null    float64
 26  worst concavity         569 non-null    float64
 27  worst concave points    569 non-null    float64
 28  worst symmetry          569 non-null    float64
 29  worst fractal dimension  569 non-null    float64
dtypes: float64(30)
memory usage: 133.5 KB
```

```
np.sum(pd.isnull(df).sum())
```

```
0
```

```
df = pd.DataFrame(cancer['data'],columns=cancer['feature_names'])
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 30 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   mean radius             569 non-null    float64
 1   mean texture            569 non-null    float64
 2   mean perimeter          569 non-null    float64
 3   mean area               569 non-null    float64
 4   mean smoothness         569 non-null    float64
 5   mean compactness        569 non-null    float64
 6   mean concavity          569 non-null    float64
 7   mean concave points     569 non-null    float64
 8   mean symmetry           569 non-null    float64
 9   mean fractal dimension  569 non-null    float64
 10  radius error            569 non-null    float64
 11  texture error           569 non-null    float64
 12  perimeter error         569 non-null    float64
 13  area error              569 non-null    float64
 14  smoothness error        569 non-null    float64
 15  compactness error       569 non-null    float64
 16  concavity error         569 non-null    float64
 17  concave points error    569 non-null    float64
 18  symmetry error          569 non-null    float64
 19  fractal dimension error 569 non-null    float64
 20  worst radius            569 non-null    float64
 21  worst texture           569 non-null    float64
 22  worst perimeter         569 non-null    float64
 23  worst area              569 non-null    float64
 24  worst smoothness        569 non-null    float64
 25  worst compactness       569 non-null    float64
 26  worst concavity         569 non-null    float64
 27  worst concave points    569 non-null    float64
 28  worst symmetry          569 non-null    float64
 29  worst fractal dimension 569 non-null    float64
dtypes: float64(30)
memory usage: 133.5 KB
```

```
df['cancer'] =pd.DataFrame(cancer['target'])
df.head()
```


```
{"type":"dataframe","variable_name":"df"}
```

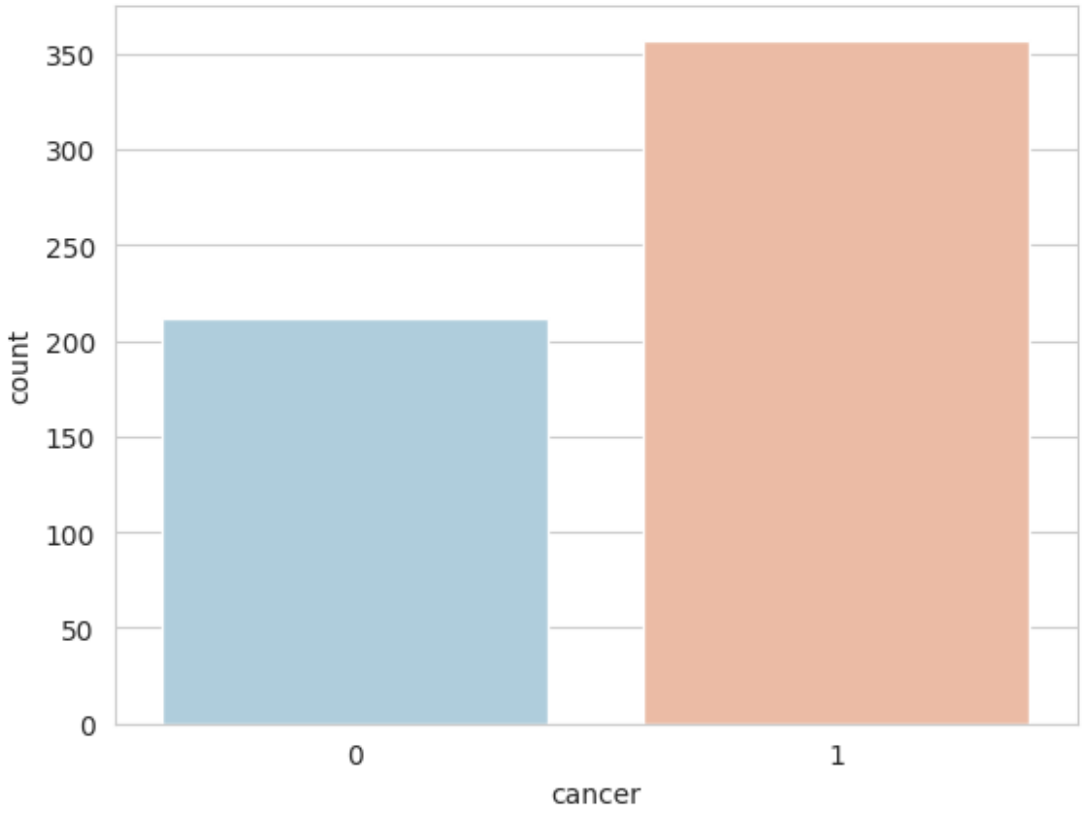

```
sns.set_style('whitegrid')
sns.countplot(x='cancer',data=df,palette='RdBu_r')
```

```
<ipython-input-15-d3c071f40503>:2: FutureWarning:
```

```
Passing `palette` without assigning `hue` is deprecated and will be removed i
n v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the sa
me effect.
```

```
  sns.countplot(x='cancer',data=df,palette='RdBu_r')
```

```
<Axes: xlabel='cancer', ylabel='count'>
```

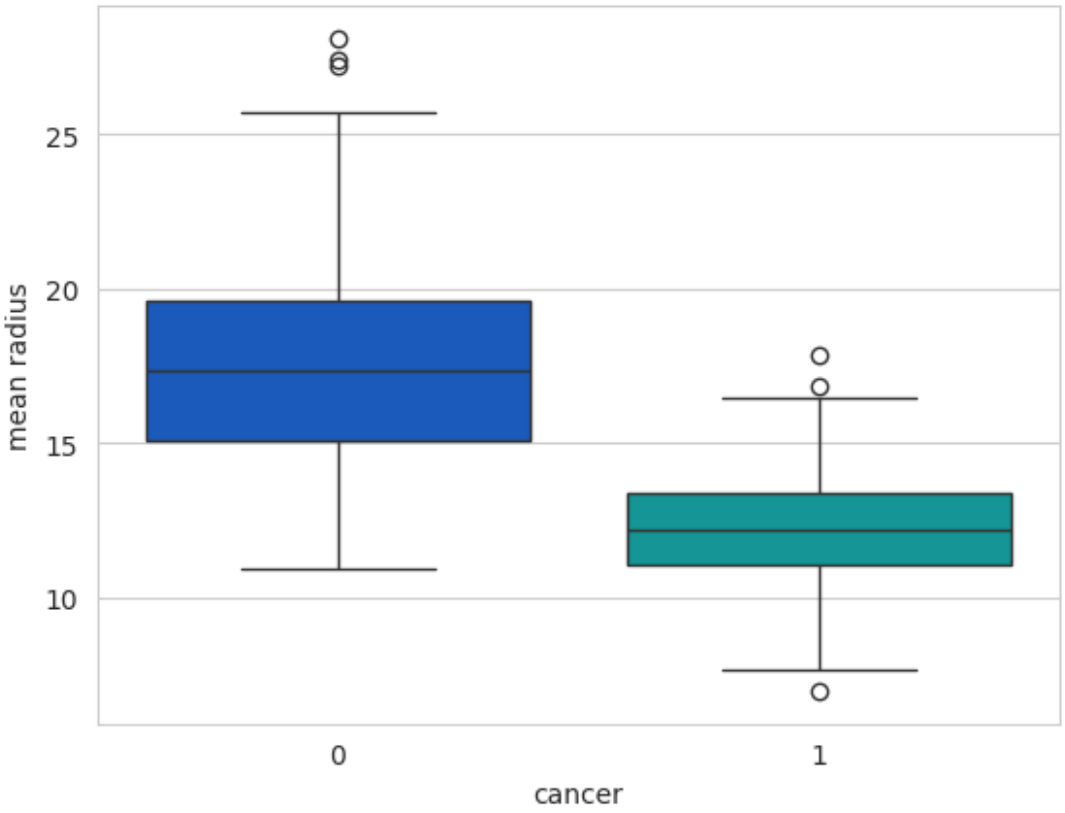


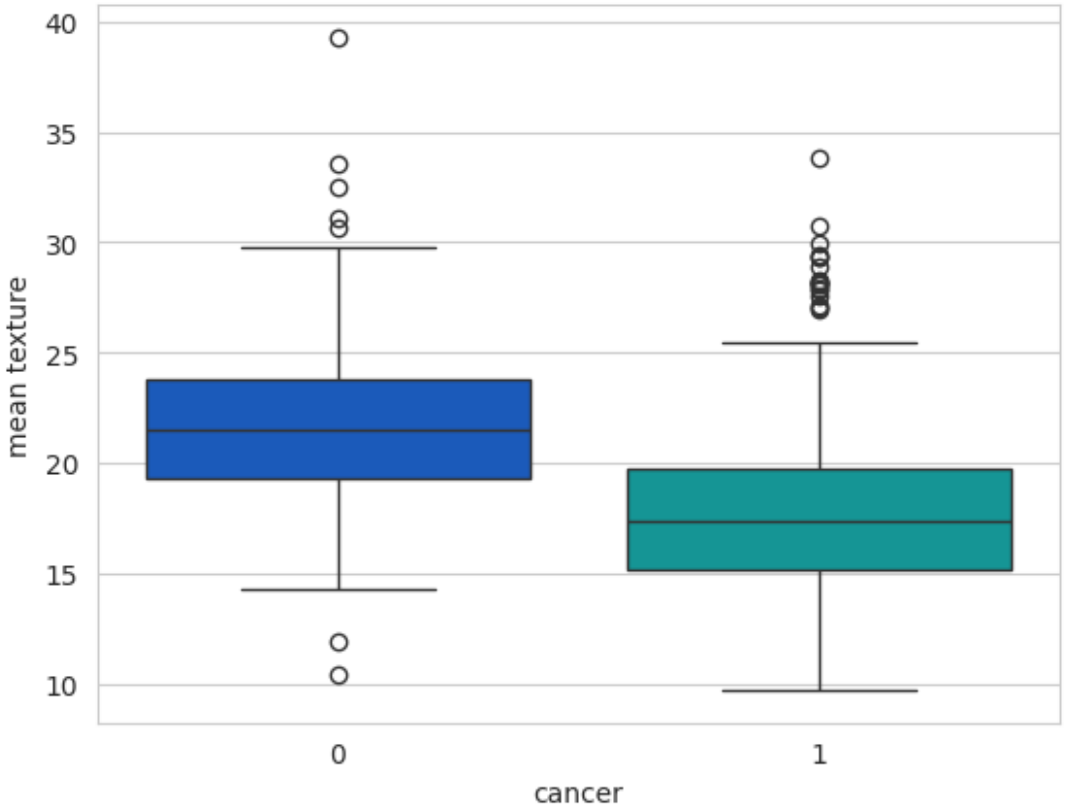

```
l=list(df.columns[0:10])
for i in range(len(l)-1):
    sns.boxplot(x='cancer',y=l[i], data=df , palette='winter')
    plt.figure()
```

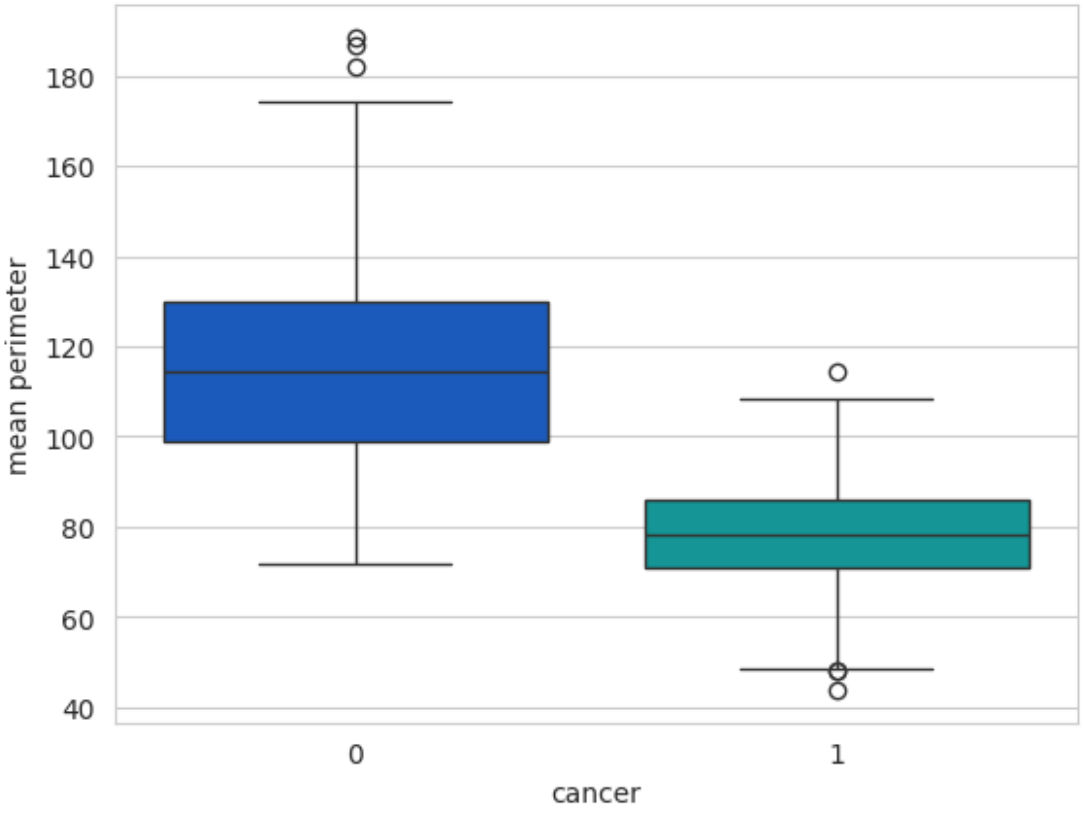
```
<ipython-input-16-c50b4d1c2876>:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed i
n v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the sa
me effect.

  sns.boxplot(x='cancer',y=l[i], data=df , palette='winter')
<ipython-input-16-c50b4d1c2876>:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed i
n v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the sa
me effect.
```

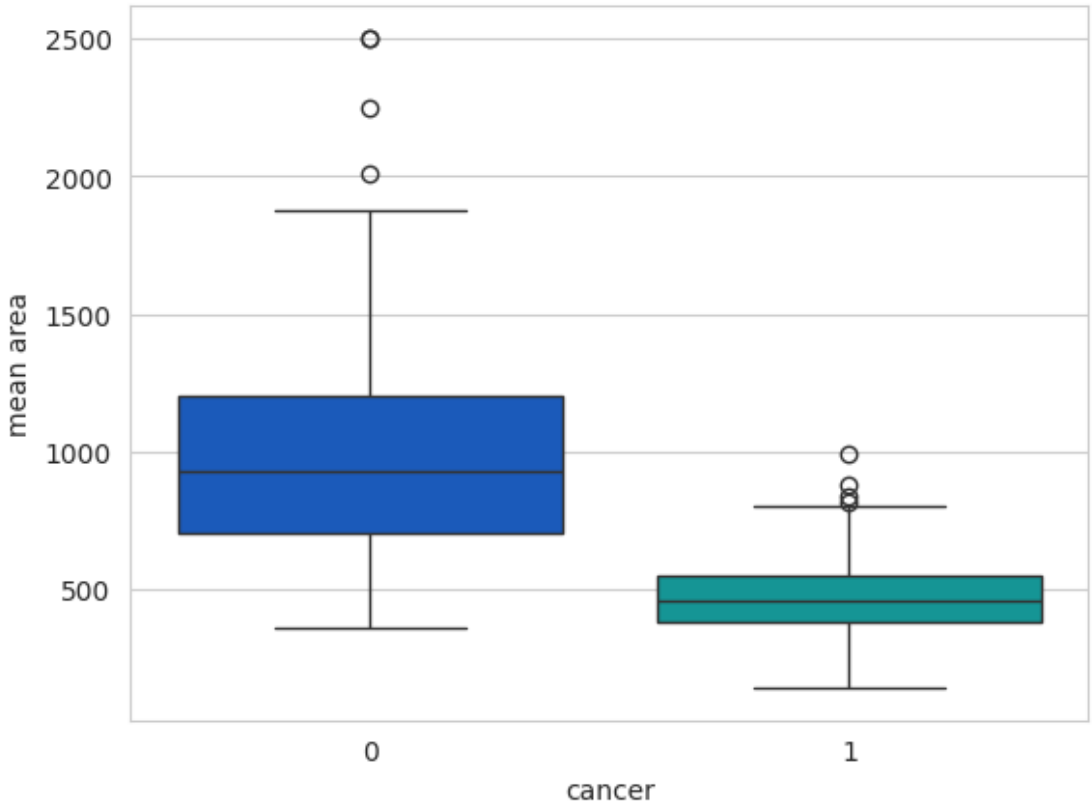
```
  sns.boxplot(x='cancer',y=l[i], data=df , palette='winter')
<ipython-input-16-c50b4d1c2876>:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed i
n v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the sa
me effect.

  sns.boxplot(x='cancer',y=l[i], data=df , palette='winter')
<ipython-input-16-c50b4d1c2876>:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed i
n v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the sa
me effect.

  sns.boxplot(x='cancer',y=l[i], data=df , palette='winter')
<ipython-input-16-c50b4d1c2876>:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed i
n v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the sa
me effect.

  sns.boxplot(x='cancer',y=l[i], data=df , palette='winter')
<ipython-input-16-c50b4d1c2876>:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed i
n v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the sa
me effect.

  sns.boxplot(x='cancer',y=l[i], data=df , palette='winter')
<ipython-input-16-c50b4d1c2876>:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed i
n v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the sa
me effect.

  sns.boxplot(x='cancer',y=l[i], data=df , palette='winter')
<ipython-input-16-c50b4d1c2876>:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed i
n v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the sa
me effect.

  sns.boxplot(x='cancer',y=l[i], data=df , palette='winter')
<ipython-input-16-c50b4d1c2876>:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed i
n v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the sa
me effect.
```

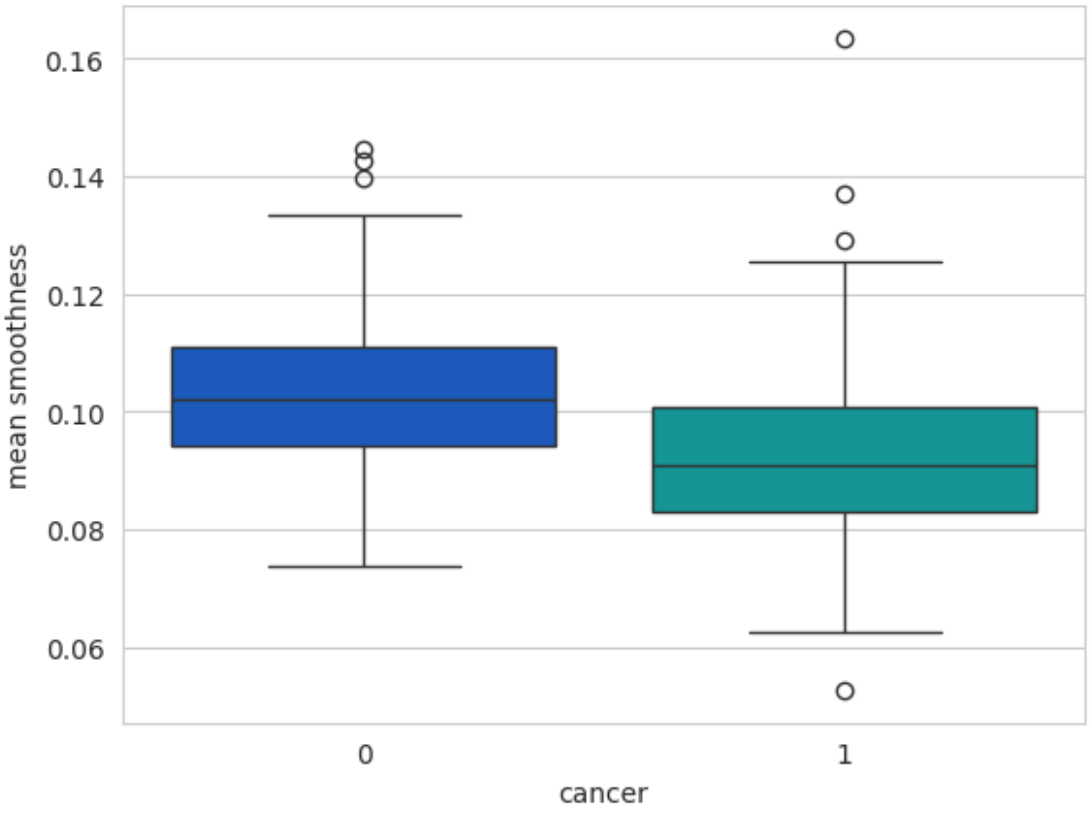
```
sns.boxplot(x='cancer',y=l[i], data=df , palette='winter')
```

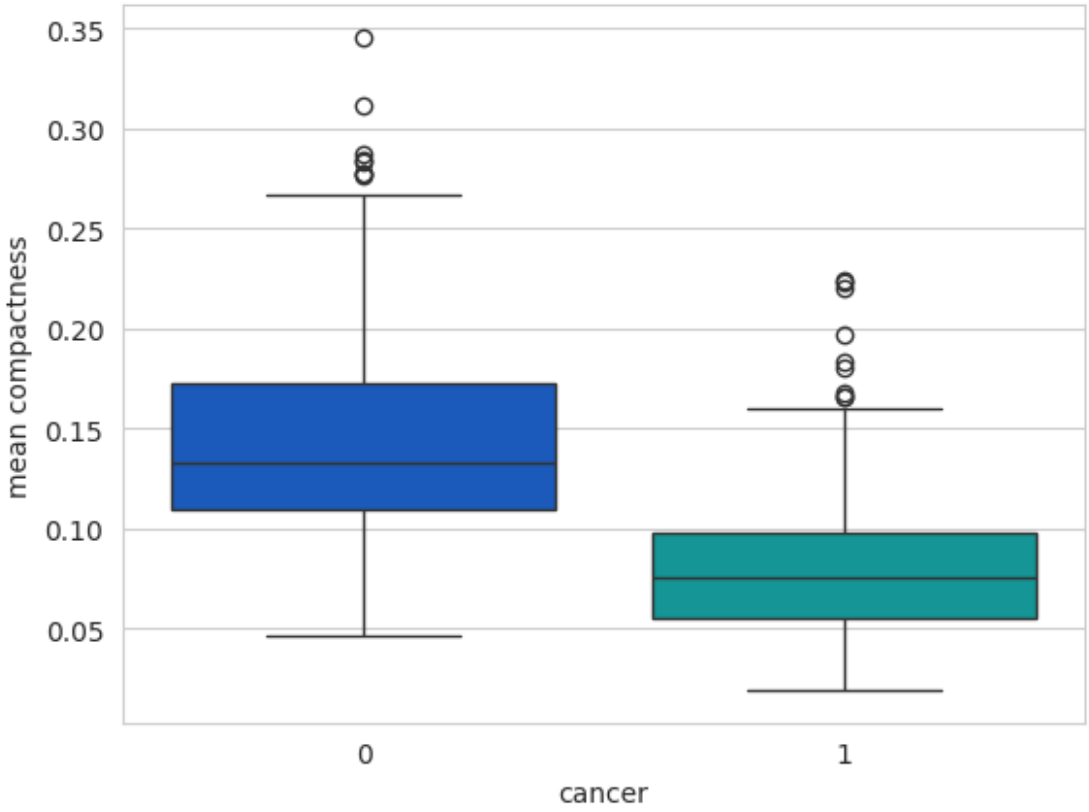


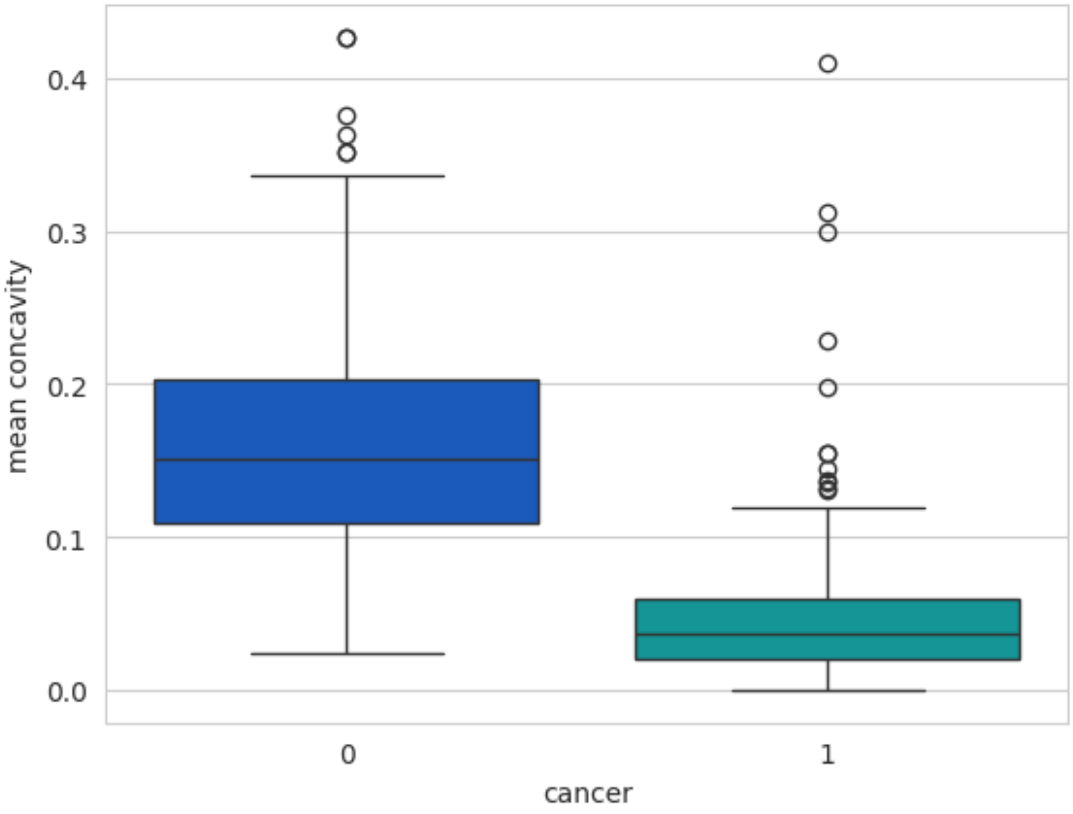

40
35
30
25
20
15
10
mean texture
0
1
cancer

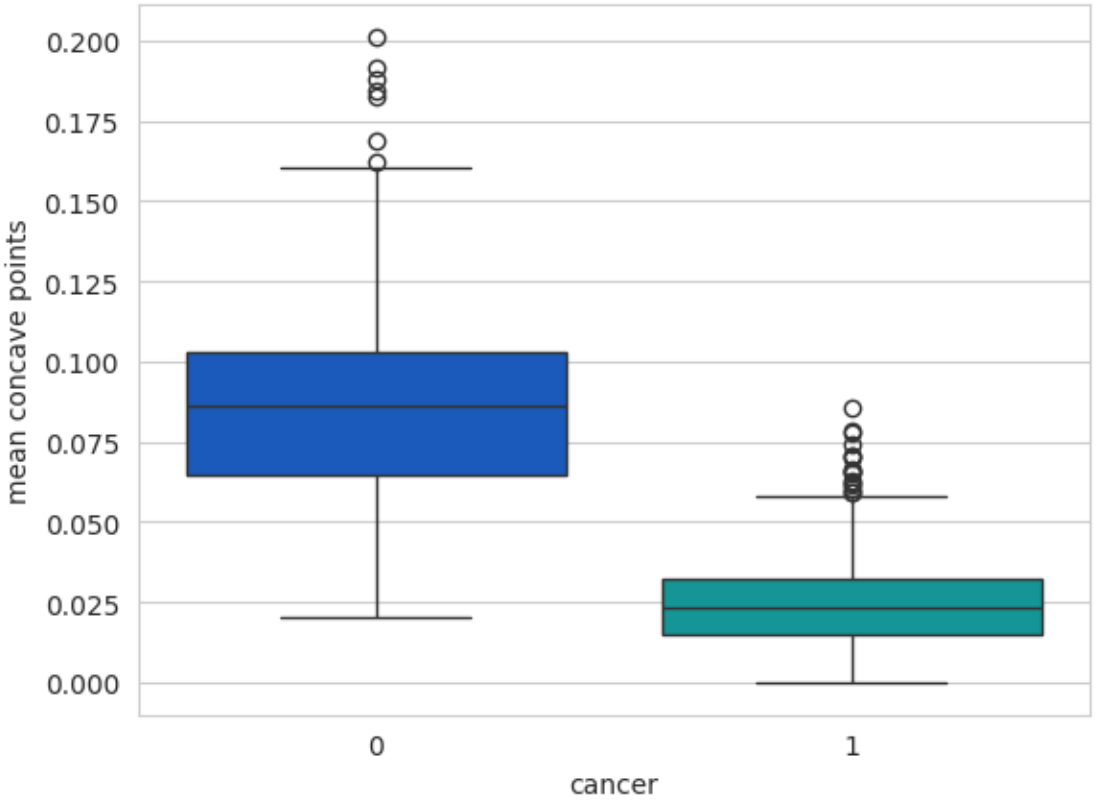

180
160
140
120
100
80
60
40
mean perimeter
0
1
cancer

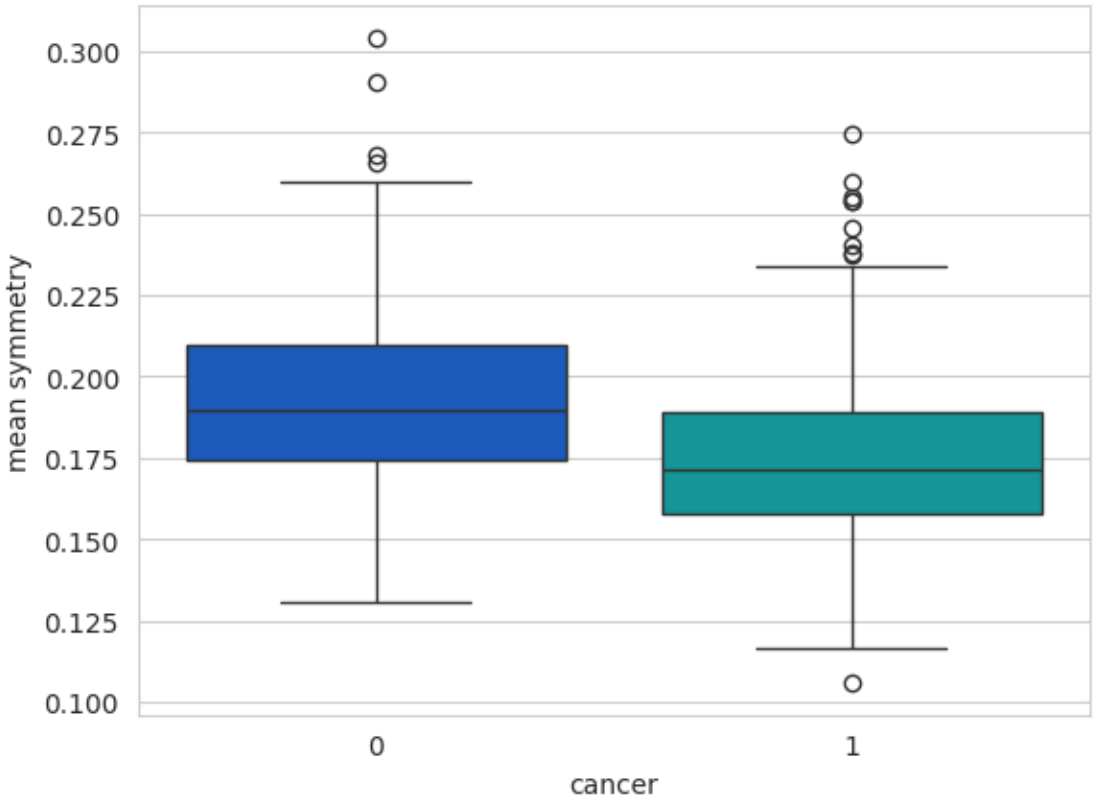

2500
2000
1500
1000
500
mean area
0
1
cancer


0.16
0.14
0.12
0.10
0.08
0.06
mean smoothness
0
1
cancer


0.35
0.30
0.25
0.20
0.15
0.10
0.05
mean compactness
0
1
cancer


0.4
0.3
0.2
0.1
0.0
mean concavity
0
1
cancer

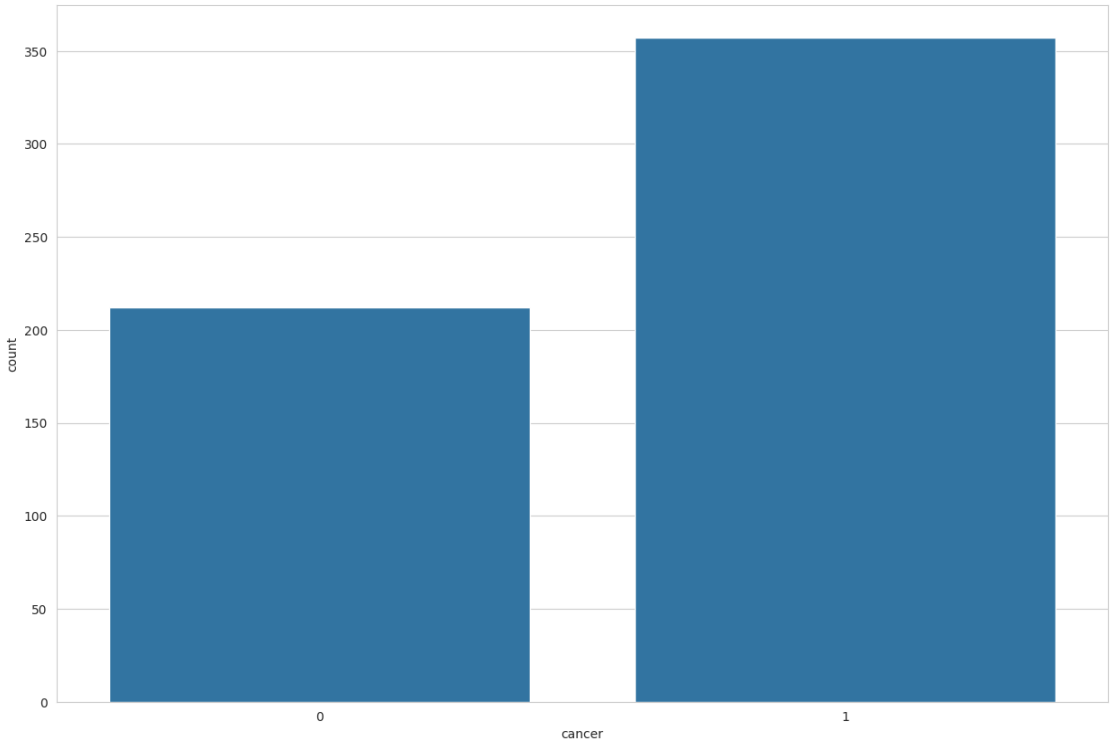

mean concave points
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
1
cancer


mean symmetry
0.300
0.275
0.250
0.225
0.200
0.175
0.150
0.125
0.100
0
1
cancer

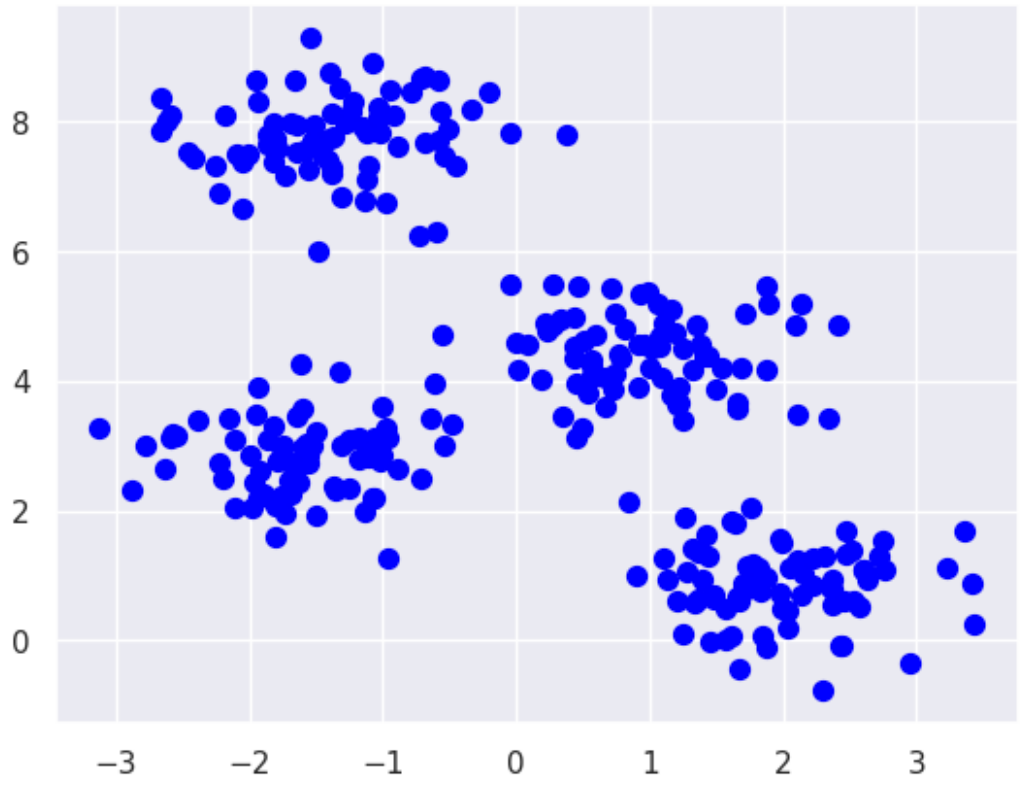
```
<Figure size 640x480 with 0 Axes>
```

```
plt.figure(figsize=(15,10))
sns.countplot(data=df, x='cancer')
```

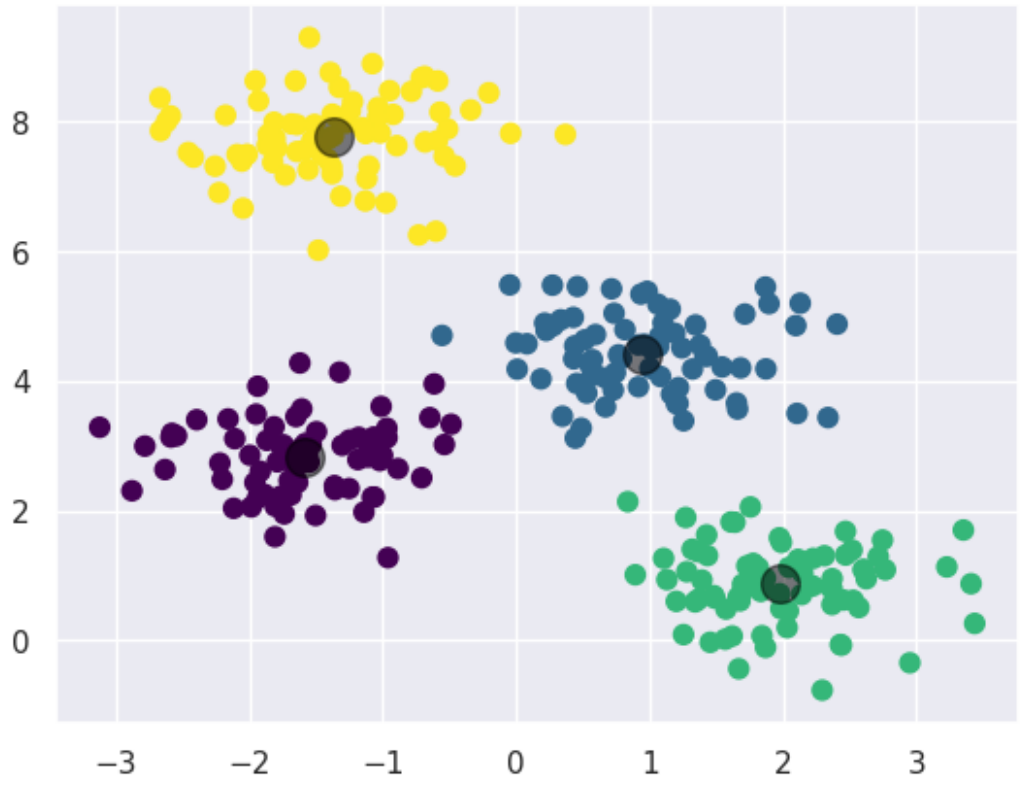
```
<Axes: xlabel='cancer', ylabel='count'>
```




```
import matplotlib.pyplot as plt
import seaborn as sns
sns.set() #Plot styling
import numpy as np

from sklearn.datasets import make_blobs
X, y_true = make_blobs(n_samples=300, centers=4,
                       cluster_std=0.60, random_state=0)
plt.scatter(X[:, 0], X[:, 1], s=50,color='blue');
```




```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=4, n_init=10)
kmeans.fit(X)
y_kmeans = kmeans.predict(X)

plt.scatter(X[:, 0], X[:, 1], c=y_kmeans, s=50, cmap='viridis')

centers = kmeans.cluster_centers_
plt.scatter(centers[:, 0], centers[:, 1], c='black', s=200, alpha=0.5);
```




```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 31 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   mean radius              569 non-null    float64
 1   mean texture             569 non-null    float64
 2   mean perimeter           569 non-null    float64
 3   mean area                569 non-null    float64
 4   mean smoothness          569 non-null    float64
 5   mean compactness         569 non-null    float64
 6   mean concavity           569 non-null    float64
 7   mean concave points      569 non-null    float64
 8   mean symmetry            569 non-null    float64
 9   mean fractal dimension   569 non-null    float64
 10  radius error             569 non-null    float64
 11  texture error            569 non-null    float64
 12  perimeter error          569 non-null    float64
 13  area error               569 non-null    float64
 14  smoothness error         569 non-null    float64
 15  compactness error        569 non-null    float64
```

```
 16  concavity error          569 non-null    float64
 17  concave points error     569 non-null    float64
 18  symmetry error           569 non-null    float64
 19  fractal dimension error  569 non-null    float64
 20  worst radius             569 non-null    float64
 21  worst texture            569 non-null    float64
 22  worst perimeter          569 non-null    float64
 23  worst area               569 non-null    float64
 24  worst smoothness         569 non-null    float64
 25  worst compactness        569 non-null    float64
 26  worst concavity          569 non-null    float64
 27  worst concave points     569 non-null    float64
 28  worst symmetry           569 non-null    float64
 29  worst fractal dimension  569 non-null    float64
 30  cancer                   569 non-null    int64
dtypes: float64(30), int64(1)
memory usage: 137.9 KB
```

```
scaler.fit(df.drop('cancer',axis=1))
scaled_features = scaler.transform(df.drop('cancer',axis=1))
```

```
df_feat = pd.DataFrame(scaled_features,columns=df.columns[:-1])
df_feat.head()
```


```
{"type":"dataframe","variable_name":"df_feat"}
```


```
from sklearn.model_selection import train_test_split
X = df_feat
y = df['cancer']
X_train, X_test, y_train, y_test = train_test_split(scaled_features,df['cance
r'],
                    test_size=0.30, random_state=101)
```

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=1)
knn.fit(X_train,y_train)
```

```
KNeighborsClassifier(n_neighbors=1)
```

```
pred = knn.predict(X_test)
```

```
from sklearn.metrics import classification_report,confusion_matrix
conf_mat=confusion_matrix(y_test,pred)
print(conf_mat)
```

```
[[ 61   5]
 [  3 102]]
```

```
print(classification_report(y_test,pred))
```

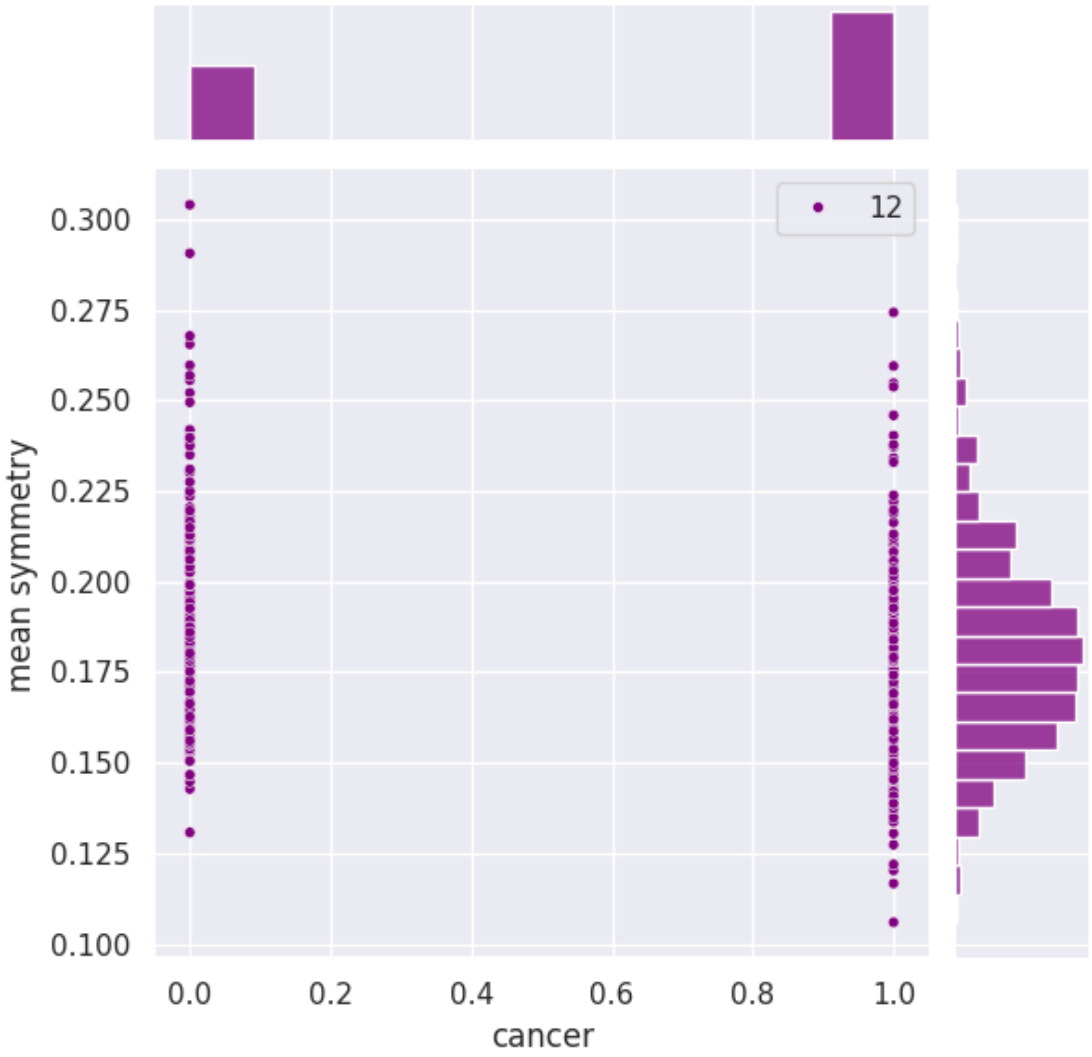
```
              precision    recall  f1-score   support

           0       0.95      0.92      0.94        66
           1       0.95      0.97      0.96       105

    accuracy                           0.95       171
   macro avg       0.95      0.95      0.95       171
weighted avg       0.95      0.95      0.95       171
```

```
cancer['target']
```

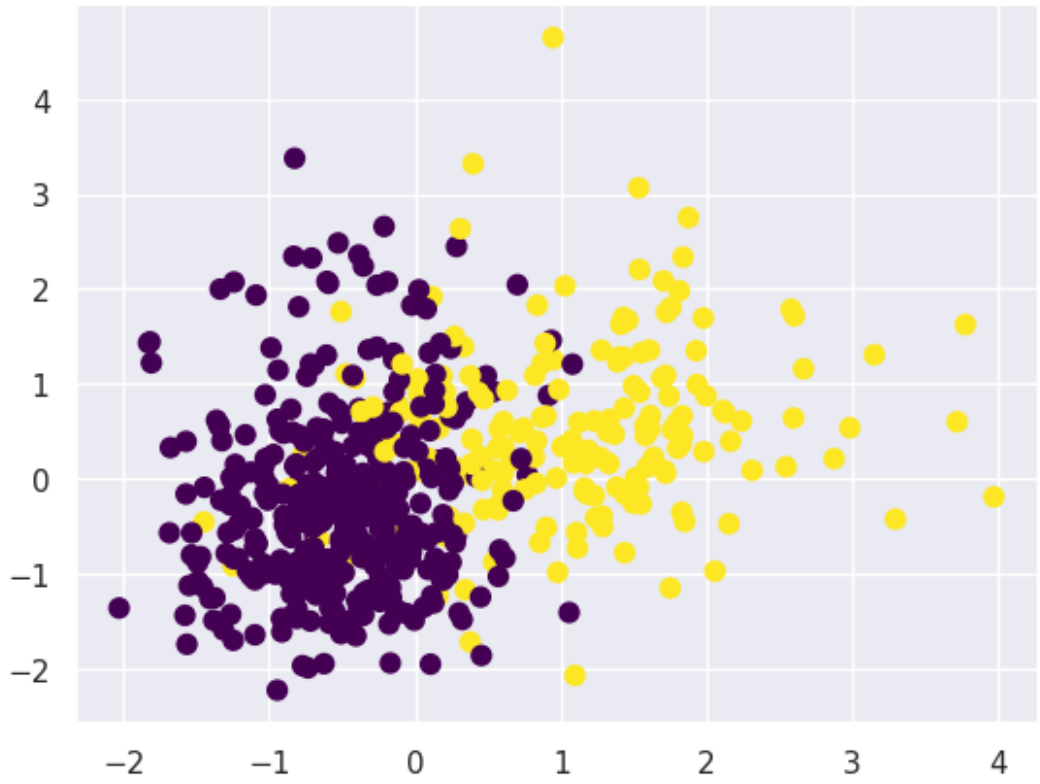
```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
       0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0,
       1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0,
       1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1,
       1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0,
       0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1,
       1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0,
       0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0,
       1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1,
       1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0,
       0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0,
       0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0,
       1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1,
       1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1,
       1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0,
       1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1,
       1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1,
       1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1])
```

```
sns.jointplot(x='cancer',y=l[i],data=df, color='purple', size=12)
```

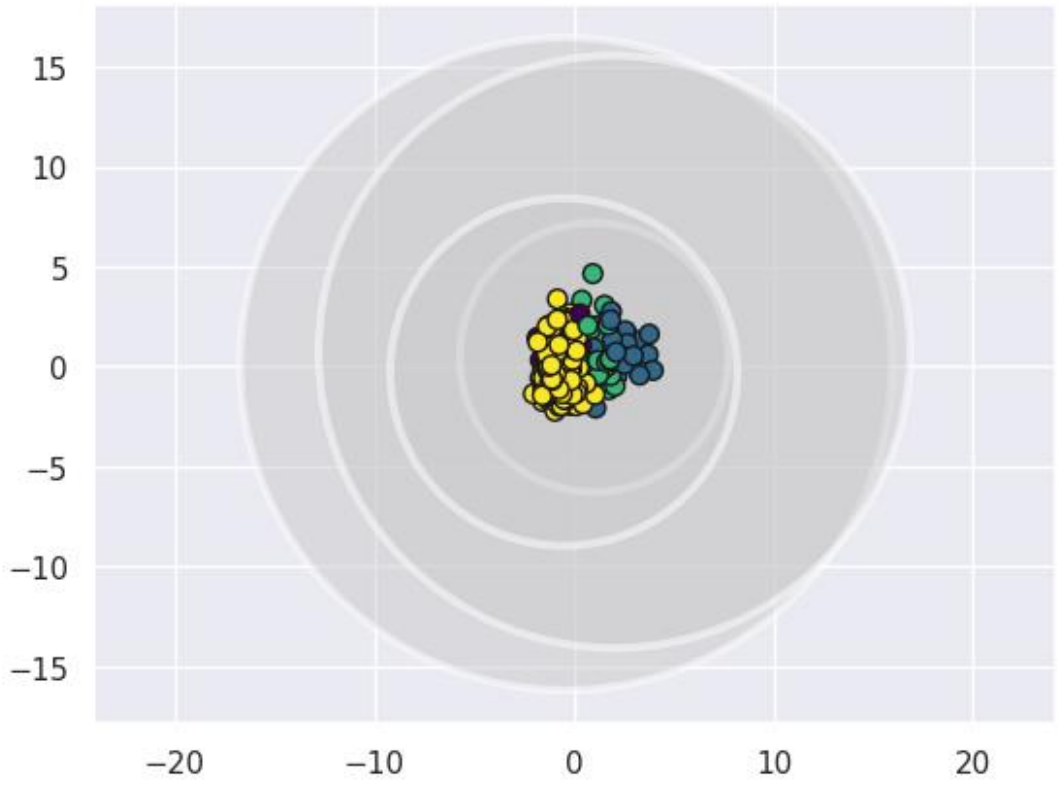
```
<seaborn.axisgrid.JointGrid at 0x7b56d5b46ad0>
```




```
import matplotlib.pyplot as plt
from sklearn.cluster import SpectralClustering
import numpy as np
model = SpectralClustering(n_clusters=2, affinity='nearest_neighbors', assign
_labels='kmeans')
labels = model.fit_predict(X)
plt.scatter(X[:,0], X[:,1], c=labels, s=50, cmap='viridis')
plt.show()
```




```
from sklearn.cluster import KMeans
from scipy.spatial.distance import cdist

def plot_kmeans(kmeans, X, n_clusters=4, rseed=0, ax=None):
    labels = kmeans.fit_predict(X)
    ax = ax or plt.gca()
    ax.axis('equal')
    ax.scatter(X[:, 0], X[:, 1], c=labels, s=50, cmap='viridis', edgecolor='k
',zorder=2)
    centers = kmeans.cluster_centers_
    radii = [cdist(X[labels == i], [center]).max()
             for i, center in enumerate(centers)]
    for c, r in zip(centers, radii):
        ax.add_patch(plt.Circle(c, r, fc='#CCCCCC', lw=3, alpha=0.5, zorder=1
))

kmeans = KMeans(n_clusters=4, random_state=0, n_init=10)
plot_kmeans(kmeans, X)
```



Results And Analysis :

Breast cancer is the second leading cause of cancer death in women, second only to lung cancer.

The leading risk factor for breast cancer is simply being a woman. Though breast cancer does occur in men, the disease is 100 times more common in women.

Men can also get breast cancer. In 2017, the American Cancer Society estimates 2,470 new cases of invasive breast cancer will be diagnosed in men in the U.S. A woman has about a one in eight chance of being diagnosed with breast cancer in her lifetime, according to the National Cancer Institute.

Most women (about eight out of 10) who get breast cancer do not have a family history of the disease.

Conclusion

By performing suitable morphological operations, system computes the suitable region properties such as Area, Euler number etc., and displays the boundary detected image along with the tumor area. These techniques improve accuracy in tracking the breast cancer cells. To assess the correctness in classifying data with respect to efficiency and effectiveness of each algorithm in terms of accuracy, precision, sensitivity, and specificity. Hence the design is to provide high accuracy and maximum efficiency in prediction and tracking of breast cancer. The combination of Multi-Level Wavelet Conversion strategy associated to PCA with 13 features extracted and then classified gives an average accuracy of nearly 92%.

As a future improvement, the system can add more features such as recommendation of medicines/treatments based on the severity of the patient. This prediction and recommendation system can help doctors to diagnose and cure the disease more efficiently.

References:

- https://www.kaggle.com/code/niteshyadav3103/breast-cancer-classification/input
- Some of the data from the github accounts
- Some of codes by previous lab activities