BREAST CANCER

SC1015 Mini-Project

FCS6 Group 5

NG HOE PING (U2321991F)

ONG YAO SHENG (U2322398H)

EDMUND YEO ZI LONG (U2322794K)

Why Breast Cancer?

- 12.5% of all cancers globally
- 670k deaths in 2022

Who suffers from breast cancer?

IN 2023

297,790 cases in <u>women</u>

2,800 cases in <u>men</u>

Women are more susceptible to breast cancer

Our Team's Agenda

01. Identifying Cancer Tumors

- Determine whether it is "M" or "B"
  - (Malignant or Benign)

02. What affects survivability?

- Inherent parameters that can't be changed
  - E.g. Blood type
- Modifiable parameters that can be changed
  - E.g. Smoking

We want to help a patient increase their survival rate after determining it is malignant (cancerous) tumor

Sources from Kaggle

First

"Breast-cancer.csv"

↓

"original"

To identify Malignant/Begnin tumors

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840
1	842517	M	20.57	17.77	132.90	1326.0	0.08474
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960
3	84348301	M	11.42	20.38	77.58	386.1	0.14250
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030

Sources from Kaggle

Birth_control(Contraception)	\nmenstrual_age	\nmenopausal_age	\nBenign_malignant_cancer	condition
1	1	0.0	1	death
0	2	0.0	0	death
0	1	0.0	1	death
0	2	0.0	0	death
0	0	0.0	0	death

Birth_control(Contraception)	\nmenstrual_age	\nmenopausal_age	\nBenign_malignant_cancer	condition
0	2	0	1	recovered
1	2	0	0	recovered
0	1	0	0	recovered
1	2	2	1	recovered
1	1	0	0	recovered

Birth_control(Contraception)	\nmenstrual_age	\nmenopausal_age	\nBenign_malignant_cancer	condition
0	1	0.0	0	under treatment
1	1	0.0	0	under treatment
1	2	0.0	0	under treatment
1	1	2.0	1	under treatment
1	1	0.0	1	under treatment

Second

"death.csv"
"recovered.csv"
"under-treatment.csv"

↓

"death"
"recovered"
"undertreatment"

To identify parameters that affects survivability

Mean

Average

Radius
Texture
Perimeter
Area
Smoothness
Compactness
Concavity
Concave Points
Symmetry
Fractal Dimensions

Se

Standard Error

Radius
Texture
Perimeter
Area
Smoothness
Compactness
Concavity
Concave Points
Symmetry
Fractal Dimensions

Worst

Outliers

Radius
Texture
Perimeter
Area
Smoothness
Compactness
Concavity
Concave Points
Symmetry
Fractal Dimensions

"Diagnosis"
(Classifies as "M" or "B")

"Patient ID"

Mean

Average

Radius
Texture
Perimeter
Area
Smoothness
Compactness
Concavity
Concave Points
Symmetry
Fractal Dimensions

Worst

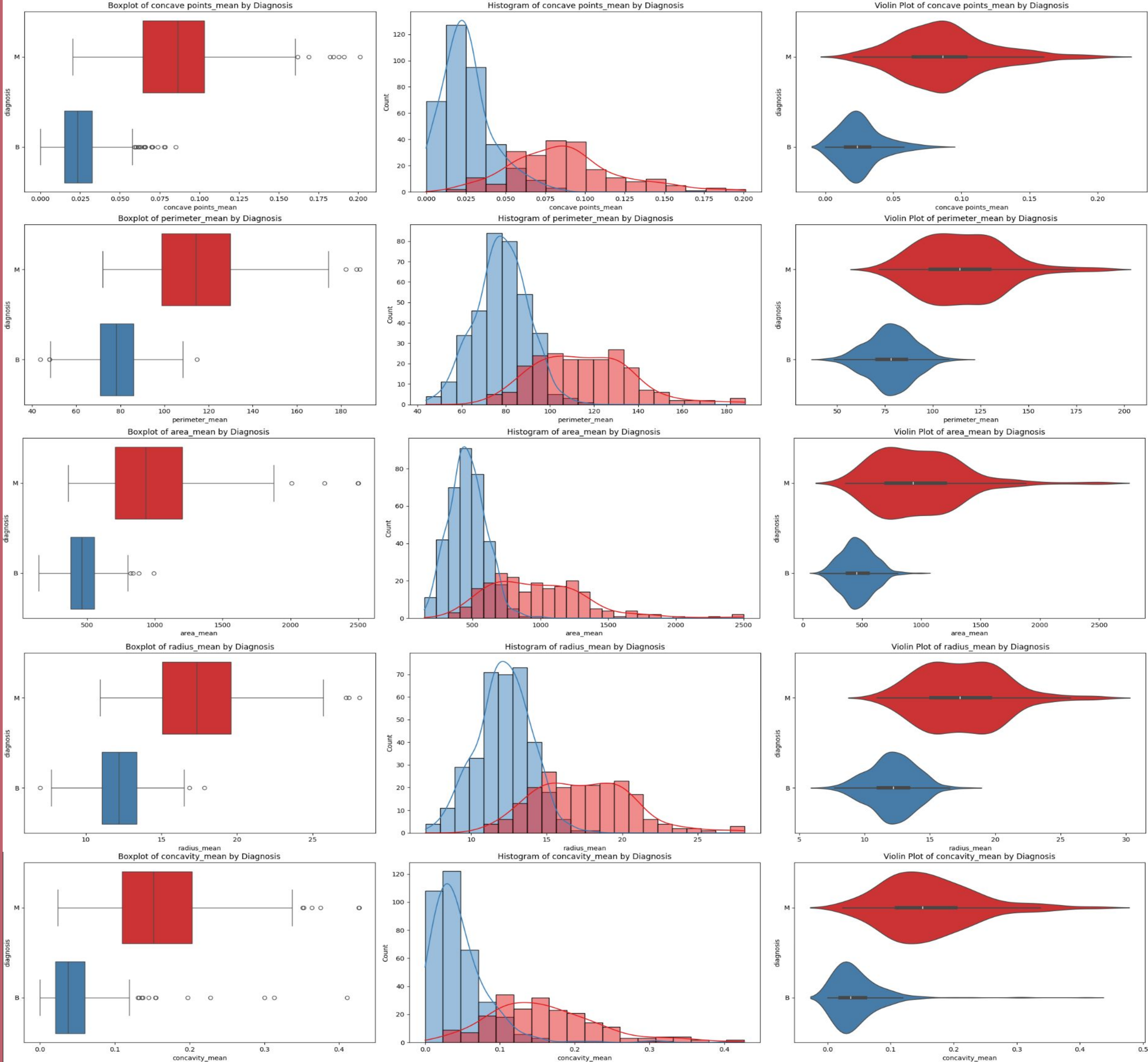
Outliers

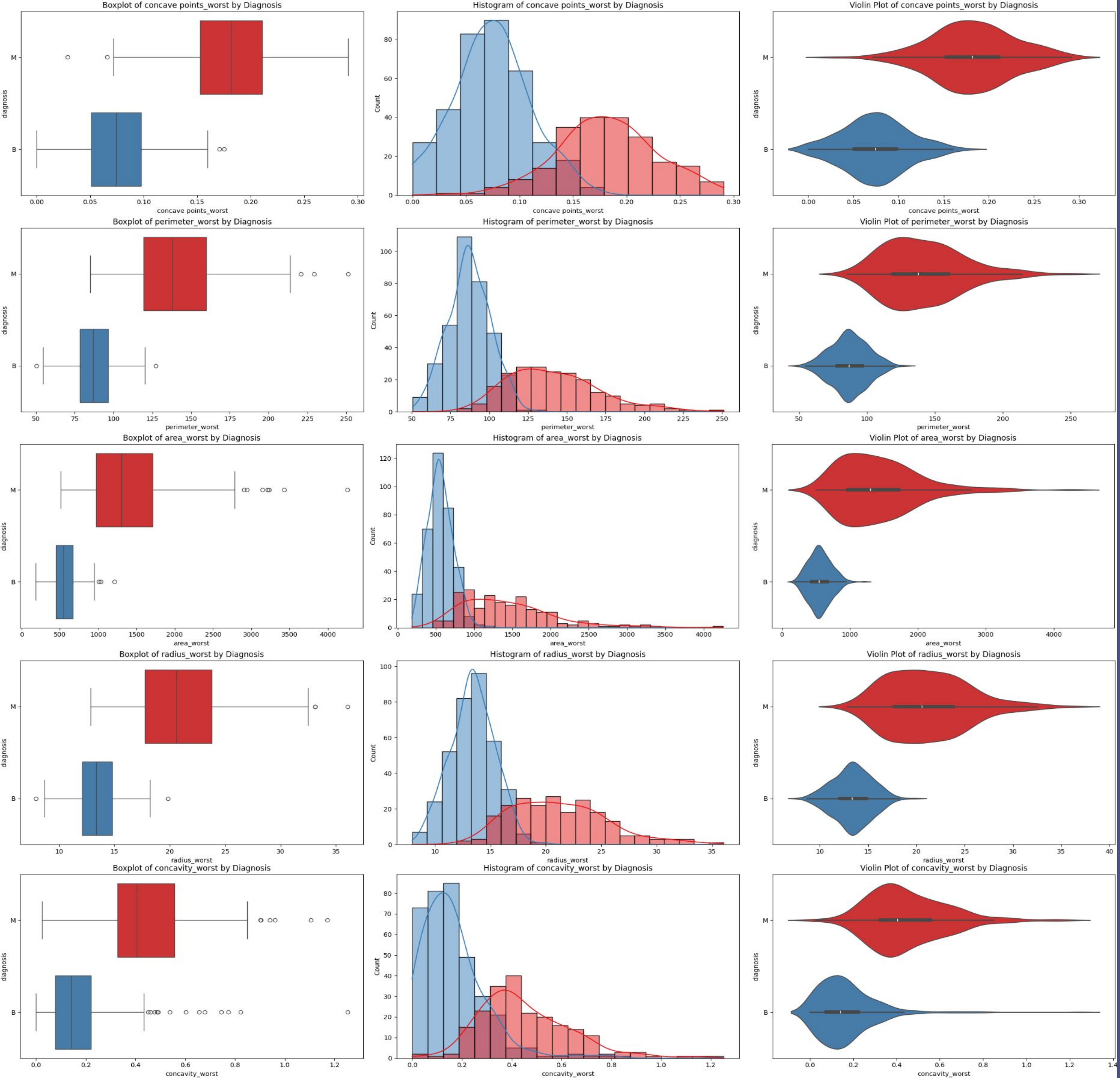
Radius
Texture
Perimeter
Area
Smoothness
Compactness
Concavity
Concave Points
Symmetry
Fractal Dimensions

"Diagnosis"
(Classifies as "M" or "B")

Mean

```
Top meanbreast predictors:
1: 0.84615 concave points_mean
2: 0.80420 concavity_mean
3: 0.78322 area_mean
4: 0.77622 perimeter_mean
5: 0.76923 radius_mean
```





Worst

```
Top worstbreast predictors:
1: 0.88112 area_worst
2: 0.86713 perimeter_worst
3: 0.86014 concave points_worst
4: 0.85315 radius_worst
5: 0.78322 concavity_worst
```

Second Source

- Death
- Recovered
- Under Treatment

01. Females only Dataset

- Removed all male data

02. Removed irrelevant parameters

- Patient ID
- Education

03. Combined "Death" and "Recovered"

- into "survival" to use machine learning

04. One-hot encoding

- Death => "1"
- Recovered => "0"

Before

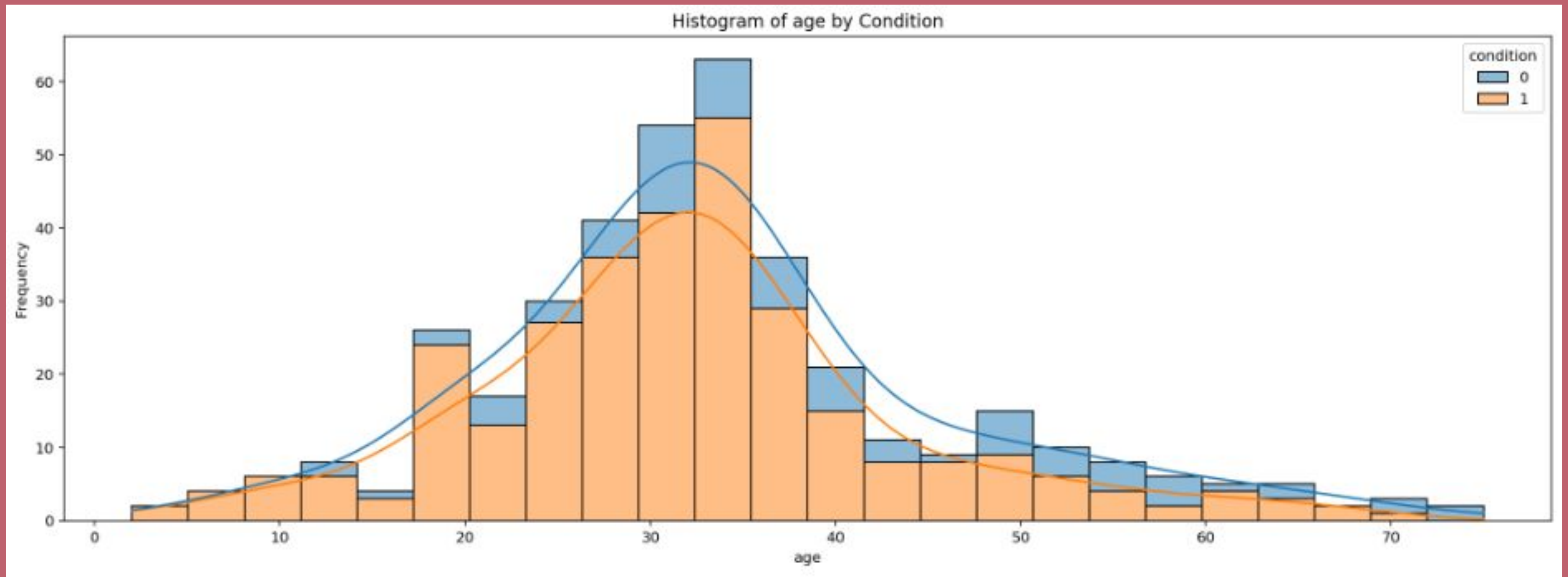
Birth_control(Contraception)	\nmenstrual_age	\nmenopausal_age	\nBenign_malignant_cancer	condition
1	1	0.0	1	death
0	2	0.0	0	death
0	1	0.0	1	death
0	2	0.0	0	death
0	0	0.0	0	death

Birth_control(Contraception)	\nmenstrual_age	\nmenopausal_age	\nBenign_malignant_cancer	condition
0	2	0	1	recovered
1	2	0	0	recovered
0	1	0	0	recovered
1	2	2	1	recovered
1	1	0	0	recovered

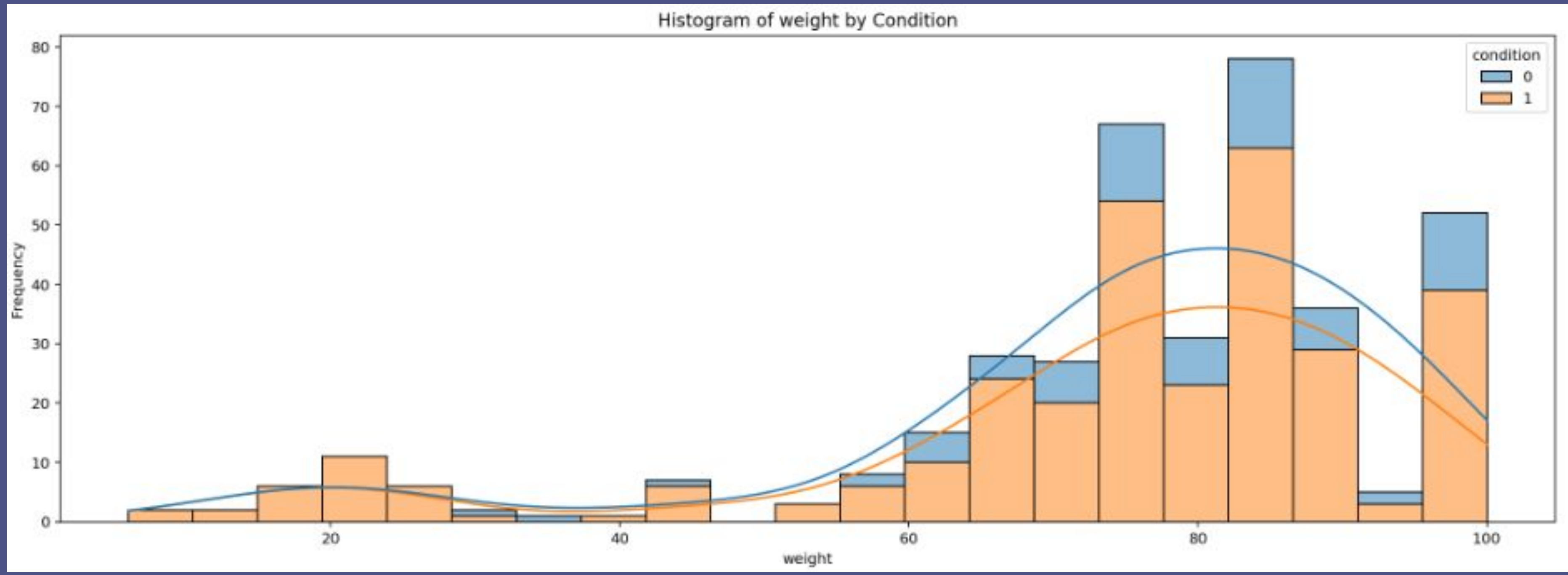
After

Birth_control(Contraception)	menstrual_age	menopausal_age	condition
1	1	0.0	1
0	1	0.0	1
1	1	0.0	1
0	2	0.0	1
0	2	0.0	1
...	...	...	...
0	2	2.0	0
1	1	0.0	0
1	2	0.0	0
1	2	1.0	0
1	2	0.0	0

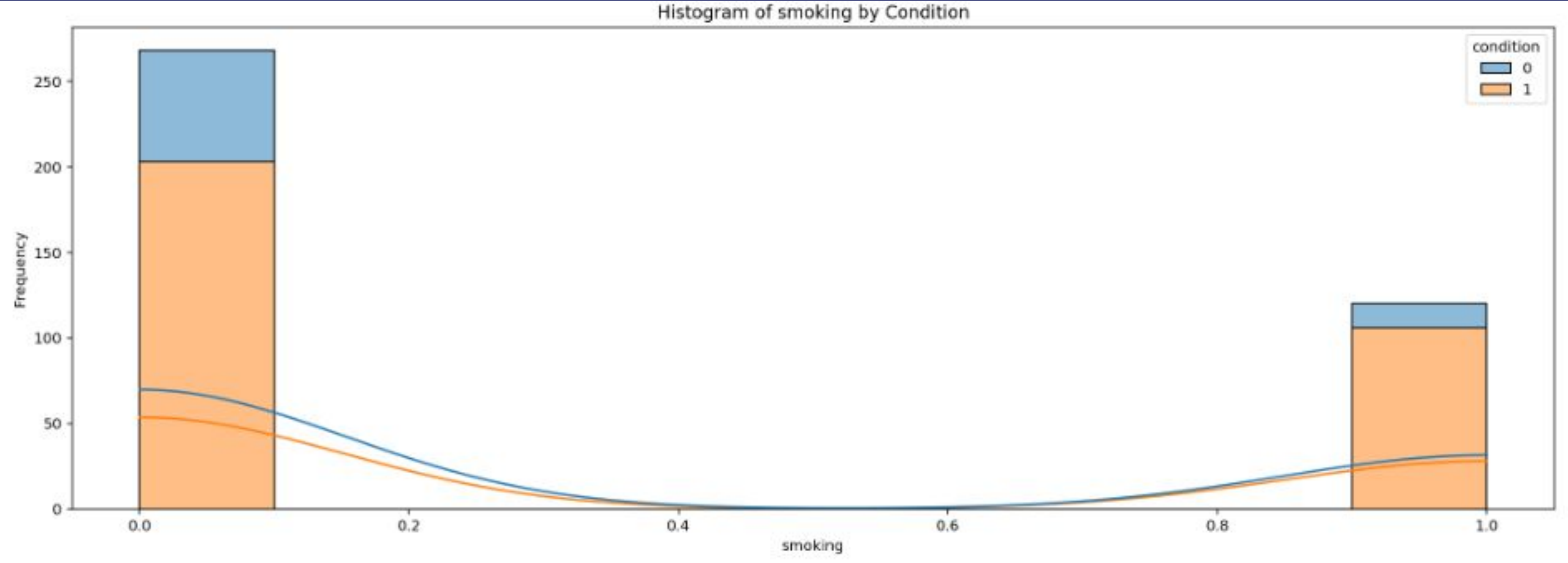
Age



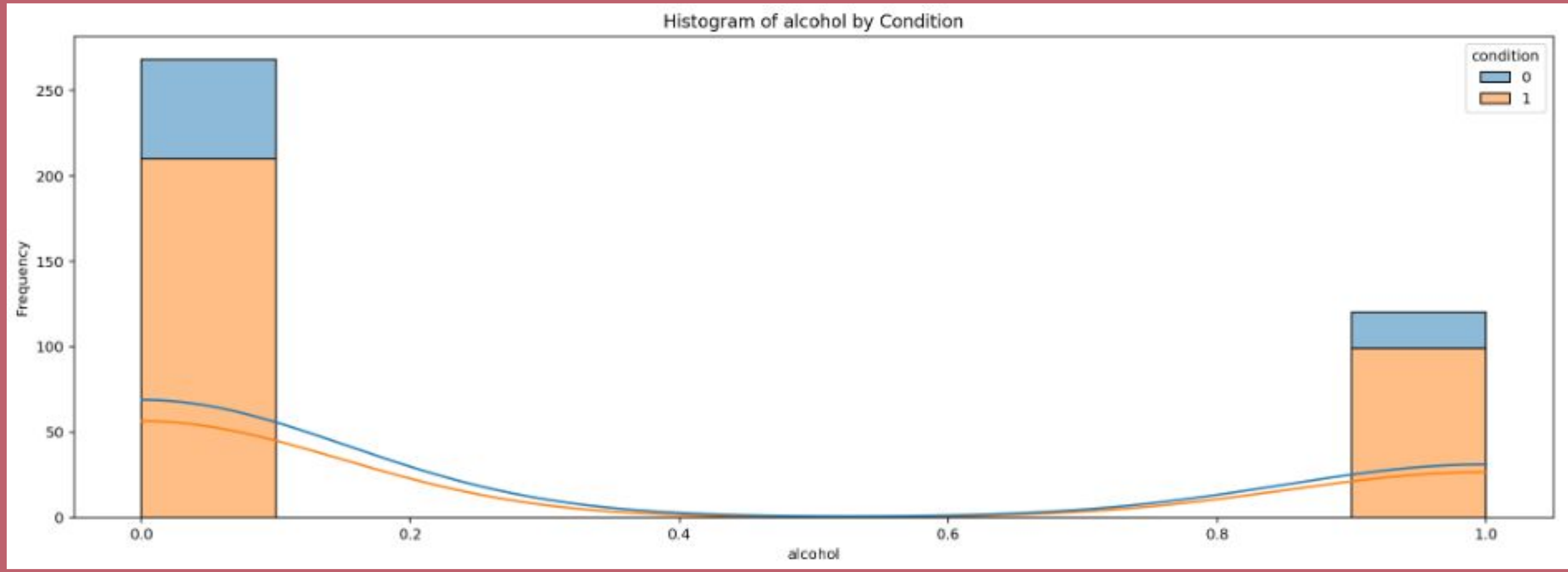
Weight



Smoking



Drinking



Applying Machine Learning on “original” Dataset

Aim for categorical outcomes from our numerical feature dataset

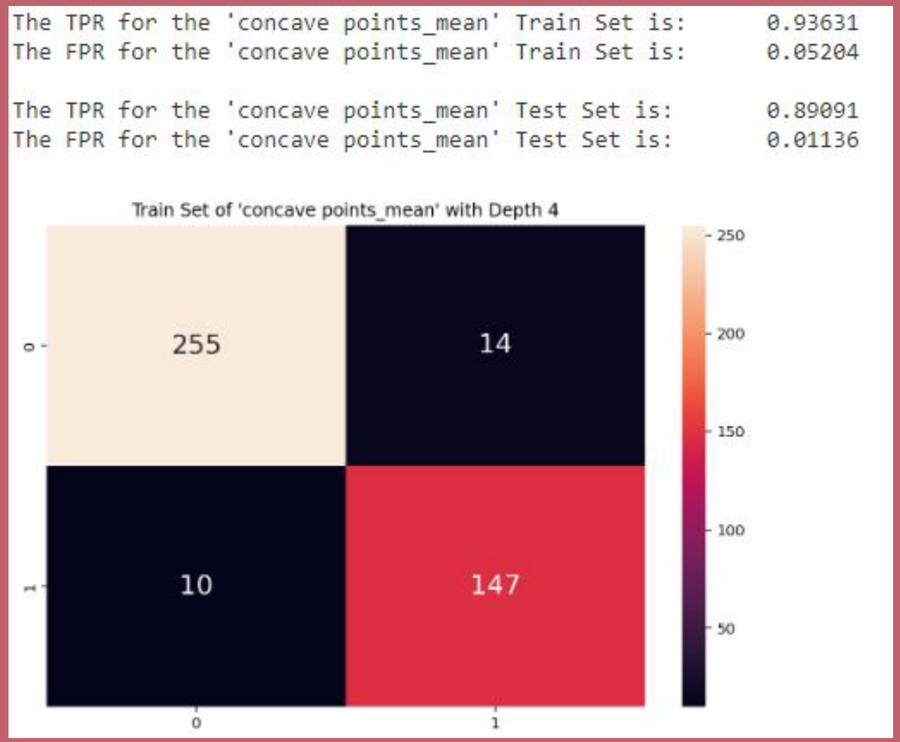
Decision Tree (Depth 4)

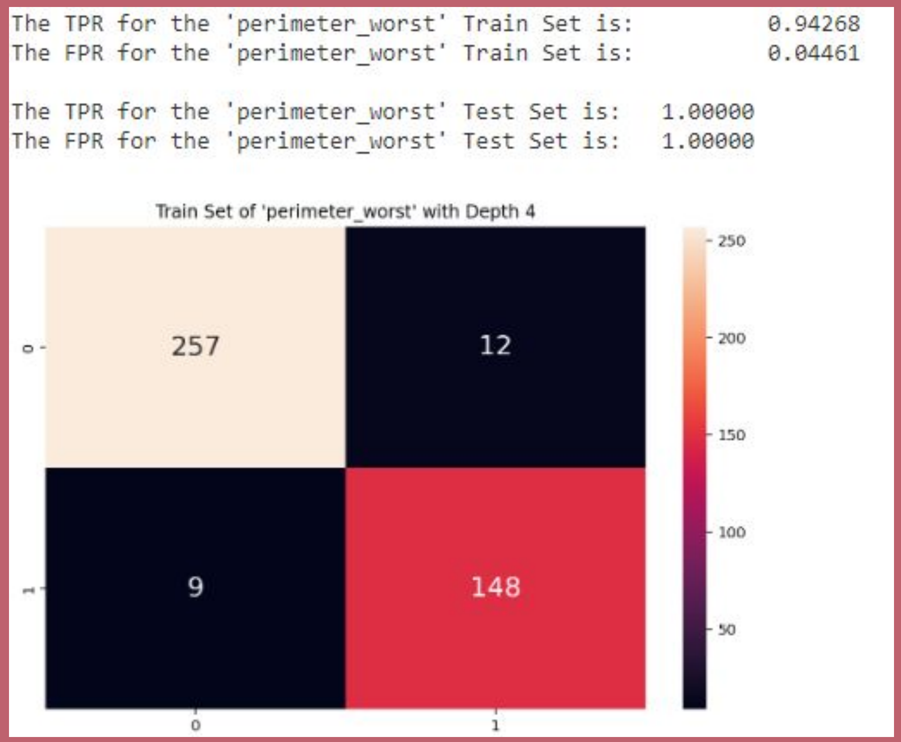
Highest Accuracy for “mean”

“Concave points”: Accuracy of 0.94366

Highest Accuracy for “worst”

“Perimeter”: Accuracy of 0.9507





Applying Machine Learning on “original” Dataset

Aim for categorical outcomes from our numerical feature dataset

Decision Tree (Depth 4)

Highest Accuracy for “mean”

“Concave points”: Accuracy of 0.94366

Highest Accuracy for “worst”

“Perimeter”: Accuracy of 0.9507

RandomForest Classifier

Highest Accuracy for “mean”

“Concave points”: Accuracy of 0.84615

Highest Accuracy for “worst”

“Area”: Accuracy of 0.88112

Although the accuracy decreased,

1. We have overcome the problem of being overfitting
2. Uphold the accuracy and prediction of correct classification

Applying Machine Learning on “survival” Dataset

Aim for categorical outcomes from our numerical feature dataset

Classifiers we used

01. Logistic Regression
02. K-Nearest Neighbours (KNN)
03. Support Vector Machine (SVM)
04. Normal Decision Tree
05. RandomForest Classifier
06. Gaussian Naive Bayes (NB)

Applying Machine Learning on “survival” Dataset

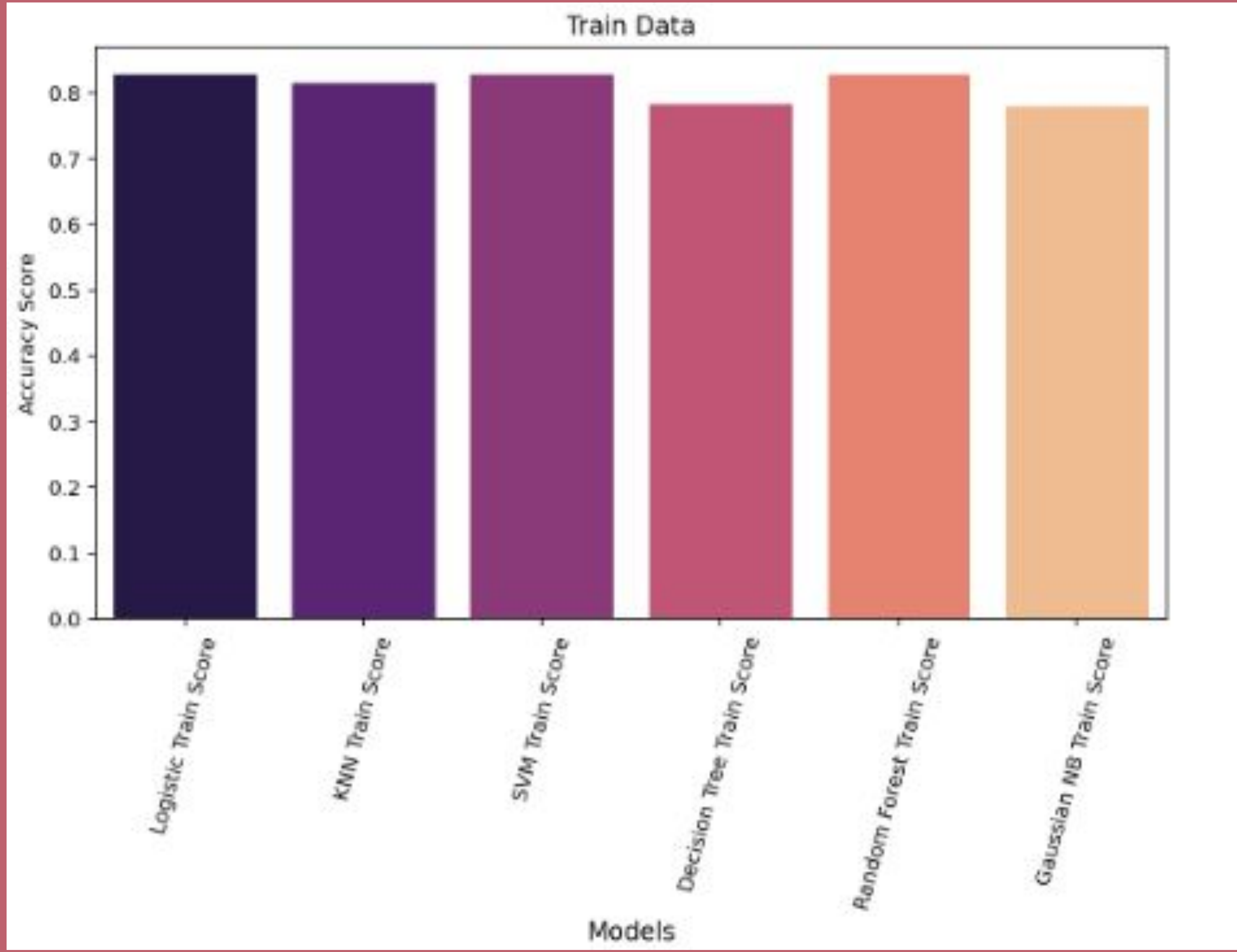
Aim for categorical outcomes from our numerical feature dataset

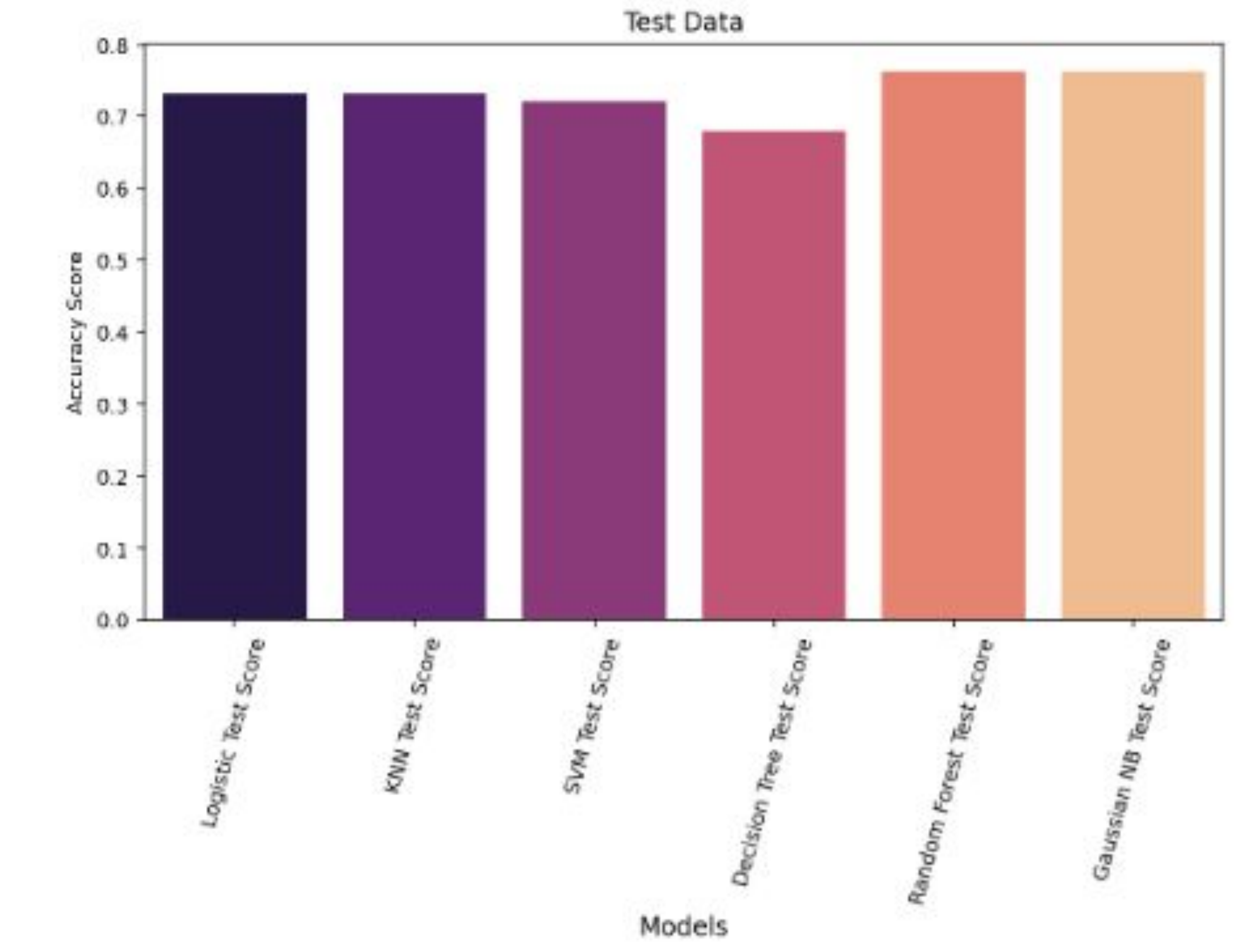
Train Scores:

	Accuracy Score
Logistic Train Score	0.797195
KNN Train Score	0.783460
SVM Train Score	0.790357
Decision Tree Train Score	0.752309
Random Forest Train Score	0.814319
Gaussian NB Train Score	0.779895

Test Scores:

	Accuracy Score
Logistic Test Score	0.762887
KNN Test Score	0.835052
SVM Test Score	0.804124
Decision Tree Test Score	0.670103
Random Forest Test Score	0.835052
Gaussian NB Test Score	0.742268





Applying Machine Learning on "survival" Dataset

Aim for categorical outcomes from our numerical feature dataset

Before SMOTE

```
Train Set : (291, 19)
Test Set  : (97, 19)
```

prediction	0	1
actual		
0	7	13
1	3	74

After SMOTE

```
Train Set : (463, 19)
Test Set  : (155, 19)
```

prediction	0	1
actual		
0	67	10
1	7	71

Applying Machine Learning on “survival” Dataset

Aim for categorical outcomes from our numerical feature dataset

Before SMOTE

Train Scores:

	Accuracy Score
Logistic Train Score	0.797195
KNN Train Score	0.783460
SVM Train Score	0.790357
Decision Tree Train Score	0.752309
Random Forest Train Score	0.814319
Gaussian NB Train Score	0.779895

Test Scores:

	Accuracy Score
Logistic Test Score	0.762887
KNN Test Score	0.835052
SVM Test Score	0.804124
Decision Tree Test Score	0.670103
Random Forest Test Score	0.835052
Gaussian NB Test Score	0.742268

After SMOTE

Train Scores:

	Accuracy Score
Logistic Train Score	0.784081
KNN Train Score	0.833731
SVM Train Score	0.840252
Decision Tree Train Score	0.745115
Random Forest Train Score	0.853132
Gaussian NB Train Score	0.779780

Test Scores:

	Accuracy Score
Logistic Test Score	0.780645
KNN Test Score	0.832258
SVM Test Score	0.819355
Decision Tree Test Score	0.774194
Random Forest Test Score	0.890323
Gaussian NB Test Score	0.812903

RandomForest Classifier is still remains the best model

Applying Machine Learning on "survival" Dataset

Aim for categorical outcomes from our numerical feature dataset

Multilayer Perceptron (MLP)

Keras Neural Network Model

```
Test Loss: 0.3581313490867615
Test Accuracy: 0.8580645322799683
```

RandomForest Classifier

```
Precision: 0.8765432098765432
Recall: 0.9102564102564102
F1: 0.8930817610062893
```

Insufficient Data for Deep Learning Model. RandomForest is still more accurate

Applying Machine Learning on "survival" Dataset

Aim for categorical outcomes from our numerical feature dataset

RandomForest Classifier

	Importance
age	0.125928
weight	0.111579
thickness_tumor	0.091978
radiation_history	0.078452
breast_pain	0.072805
smoking	0.060928
giving_birth	0.060096
blood	0.059527
alcohol	0.054868
menopausal_age	0.050498
taking_blood_pressure_medicine	0.038425
taking_heartMedicine	0.032832
taking_gallbladder_disease_medicine	0.032459
hereditary_history	0.027685
age_FirstGivingBirth	0.026463
menstrual_age	0.024086
Birth_control(Contraception)	0.023524
abortion	0.016160
pregnency_experience	0.011708

Top 3 Variables for survivability

Top 3 Modifiable Risks to increase survivability

What we learned

Using different machine learning such as:

- Random Forest
- K-Nearest Neighbors Classifier (KNN)
- Support Vector Classifier (SVC)
- Gaussian Naive Bayes
- Synthetic Minority Overlapping Technique (SMOTE)
- Multilayer Perceptron (MLP)
    - Keras Neural Network Model

Outcome of Project

By using machine learning, patient can:

- Predict if the breast tumor is cancerous (self-diagnosis)
- What characteristic or habits to reduce/stop, to increase survivability of breast cancer

In conclusion, these are the data-driven insights:

- Patients can self-examine by looking at the 5 features (concave points, area, perimeter, radius, and concavity)
- Be aware and be diagnosed earlier
- Avoid smoking and drinking
- Eat healthy and live an active lifestyle

Protect the breast.
Check the chest.
Get the test.
Early detection is best.

Control your fate – don't be the 1 in 8