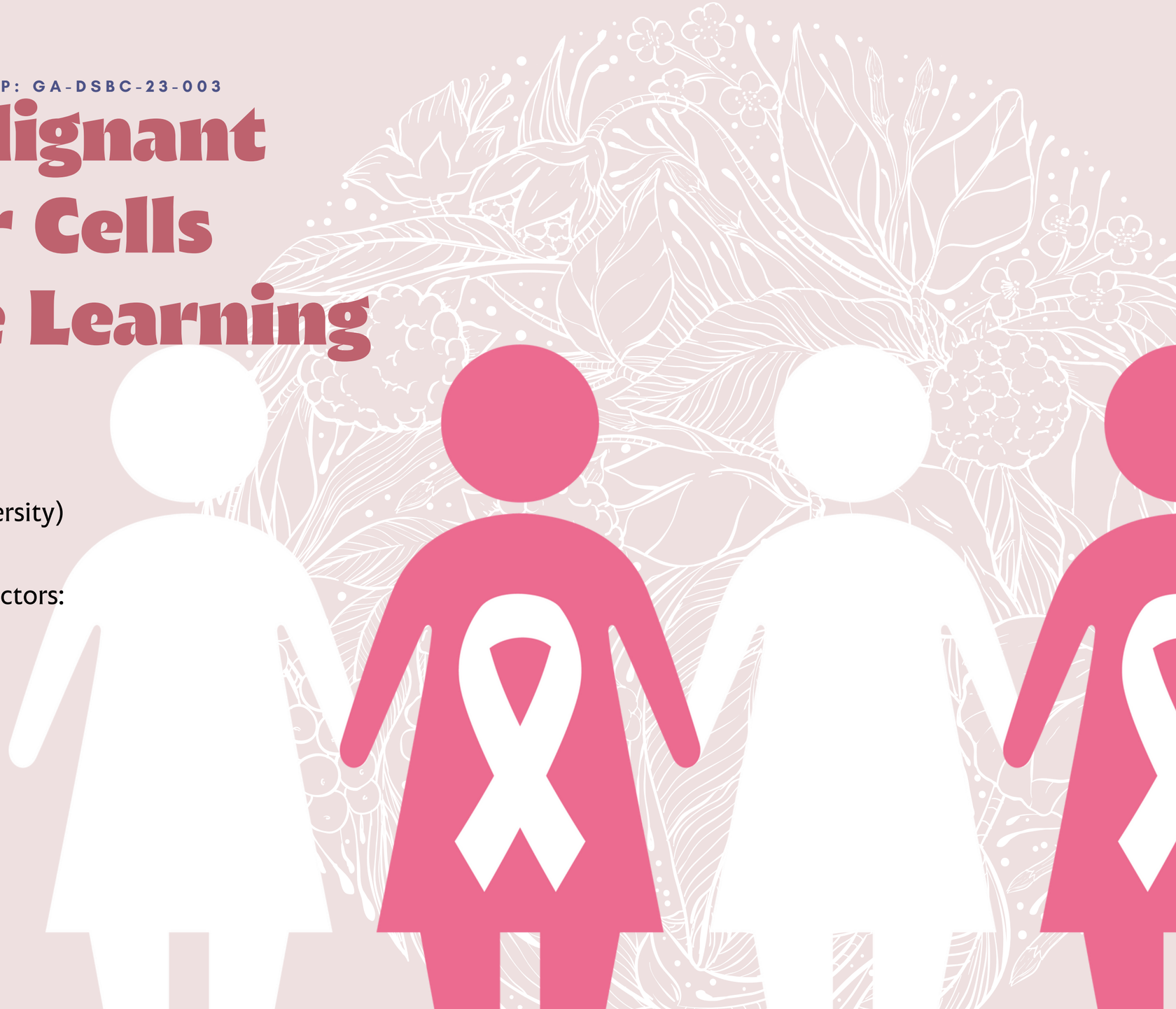
GENERAL ASSEMBLY DATA SCIENCE BOOTCAMP: GA-DSBC-23-003

Detecting Malignant Breast Cancer Cells with Machine Learning

Capstone Project
Presented by
Goh Png Ee
(BSc. Psychology, HELP University)

Under the Guidance of Instructors:
Ng Shu Min
Ts.Dr. Mogana Darshini

PRESENTED ON: 16TH OCTOBER 2023



Today's agenda.

We will be covering the steps taken for the capstone project, from problem statement, to evaluation after getting results from modelling

01. Introduction

02. Data Gathering and Preparation

03. Exploratory Data Analysis

04. Modelling

05. Evaluation

06. Recommendations

GENERAL ASSEMBLY DATA SCIENCE BOOTCAMP: GA-DSBC-23-003

01. Introduction

Angelina Jolie, an iconic role as Lara Croft in Tomb Raider

Olivia Newton John, iconic performance as Sandy in the musical film “Grease”

Linda McCartney, ex spouse of The Beatle’s Paul McCartney

What did they have in common?

Why Breast Cancer?

Breast cancer is a widespread and highly heterogeneous disease, with various sub-types that respond differently to treatments. Accurate classification of breast cancer cells can lead to tailored therapies, reducing the burden of unnecessary treatments and minimizing adverse effects on patients. Developing a model for the accurate classification of breast cancer cells is a critical step towards better understanding, managing, and ultimately conquering breast cancer.

Not only that..

-Cardiovascular Diseases: Obesity alongside lifestyle factors can be a predictor for breast cancer and cardiovasular diseases (American Cancer Study, 2022)
-Diabetes: Individuals with type 2 diabetes are at greater risk of developing cancers including breast cancer (American Association for Cancer Research, n.d.)
-Thyroid Disorders: Individuals with tyroid cancer, autoimmune thyroiditis(AITD) are also at a much greater risk of breast cancer (Chen et al., 2021)

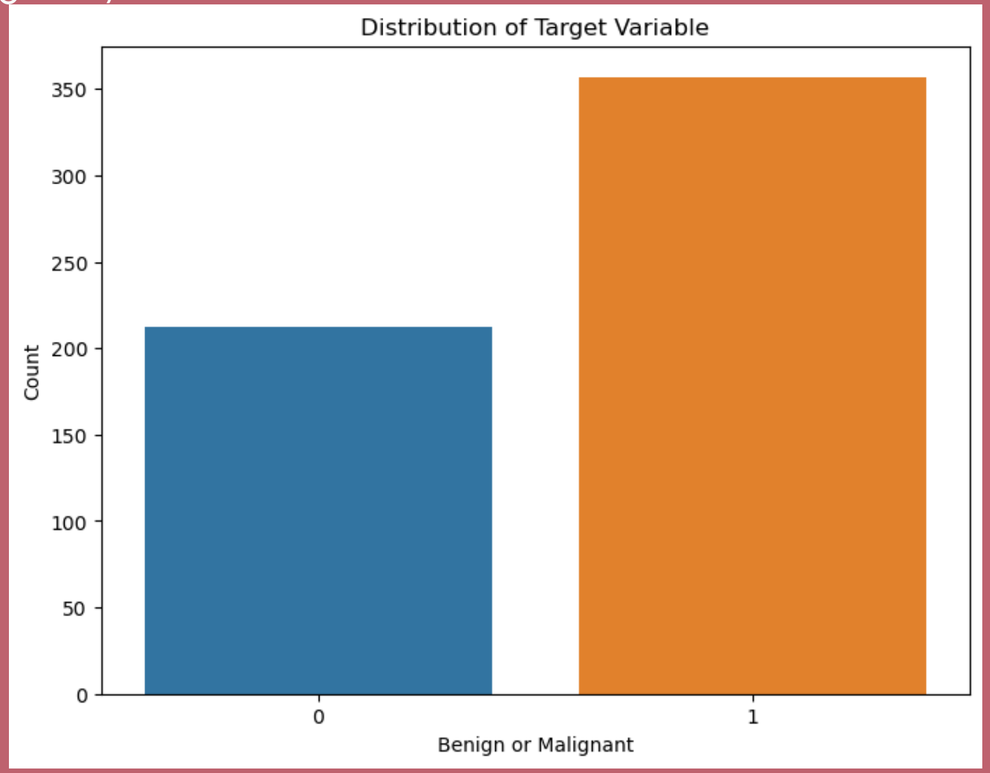
02. Data Gathering and Preparation

-The dataset obtained from Kaggle, was originally provided by UC Irvine's Breast Cancer Wisconsin (Diagnostic) machine learning repository.

-This dataset was edited and summarized into 31 columns, with features of the cells ranging from mean radius, mean texture, mean perimeter, mean area, et cetera.

-This data set consists of 569 entries of cells with different information (features) about it in the 31 columns.

-The target variable, originally labelled as target is slightly imbalanced as the distribution between the two classes (0 = benign, 1 = malignant) are as such:



02. Data Gathering and Preparation

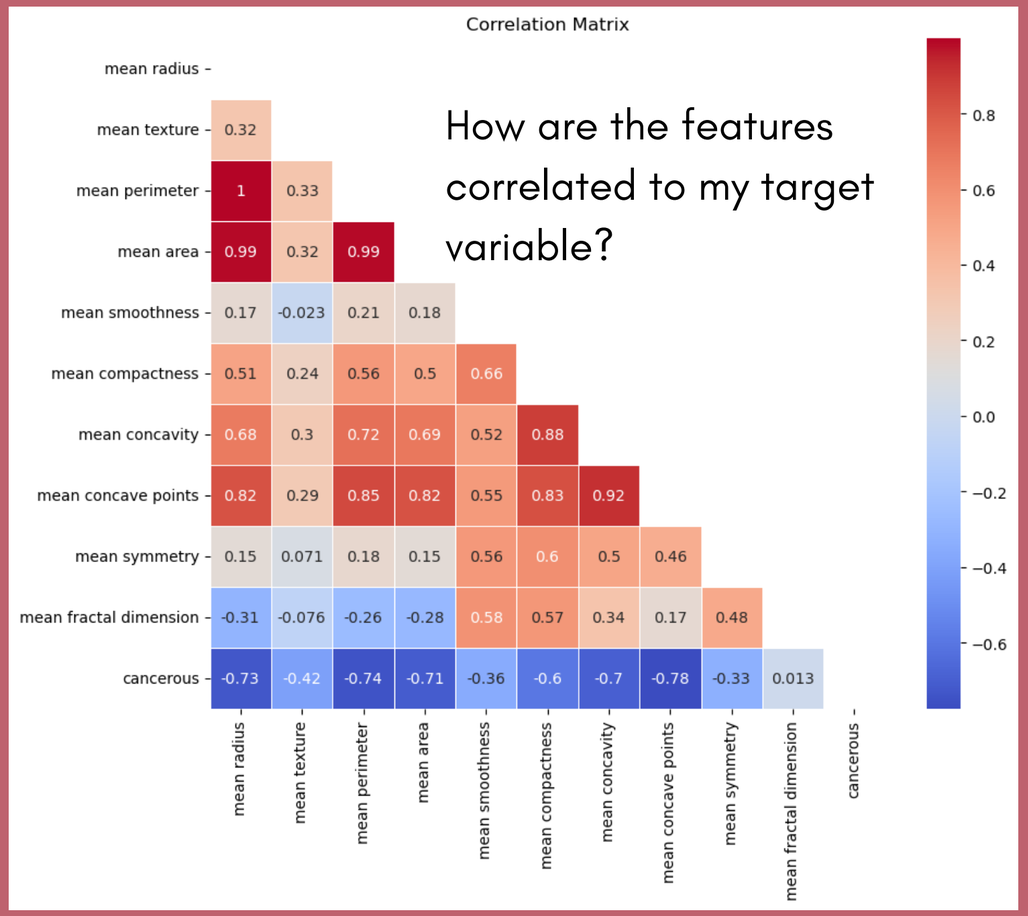
Data Cleaning:

- columns information
- missing values
- data type
- choosing meaningful features

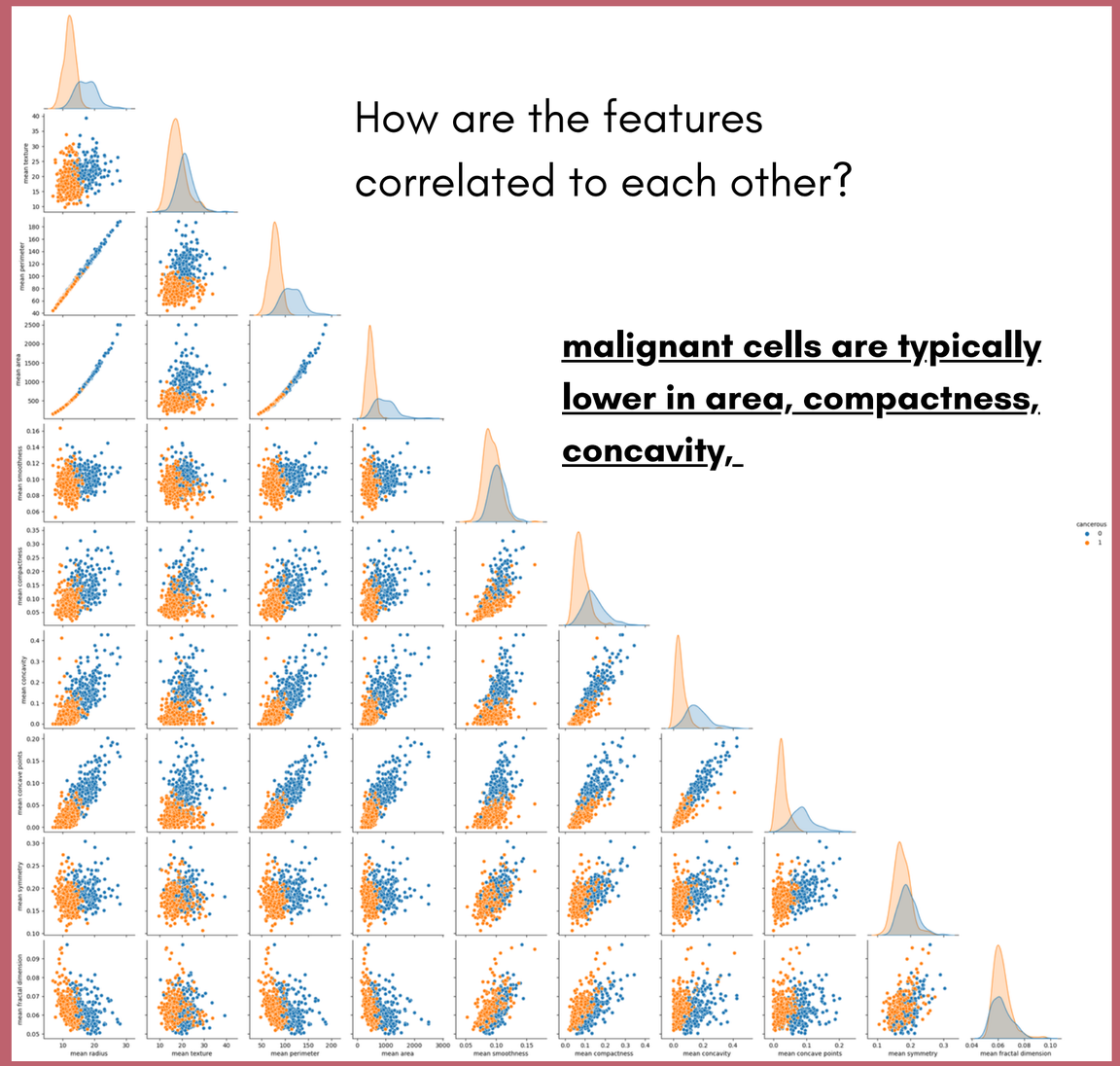
Data Transformation:

- StandardScalar()

03. Exploratory Data Analysis



The features that have the highest correlation with our target variable is mean concave point, mean perimeter, mean radius and mean area.

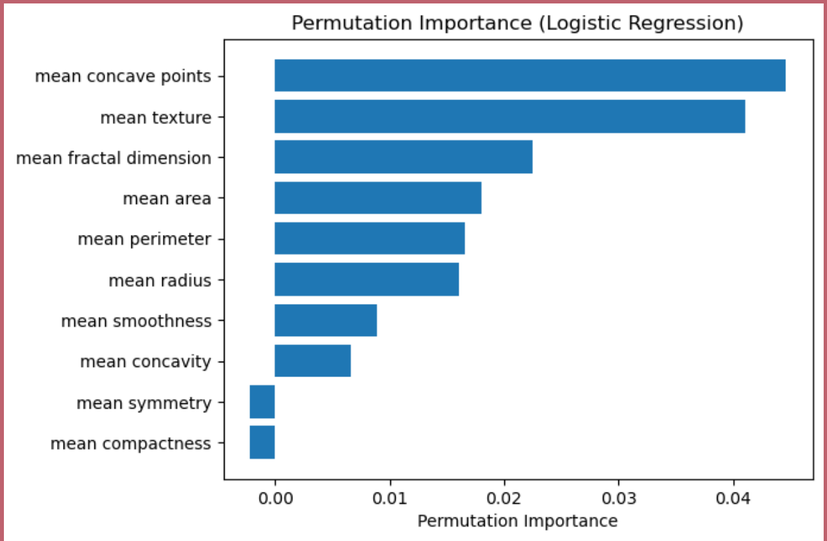


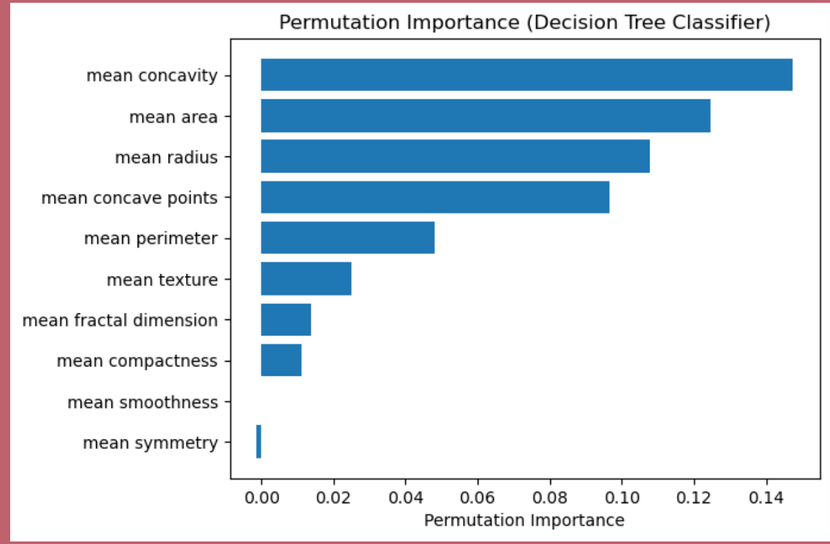
The two classes for cancerous are seperated for features: mean radius, mean perimeter, mean area, mean concavity, mean concave points, and mean compactness. Opposite is true for features: mean symmetry, mean fractal dimension, mean smoothness, and mean texture.

04. Modelling

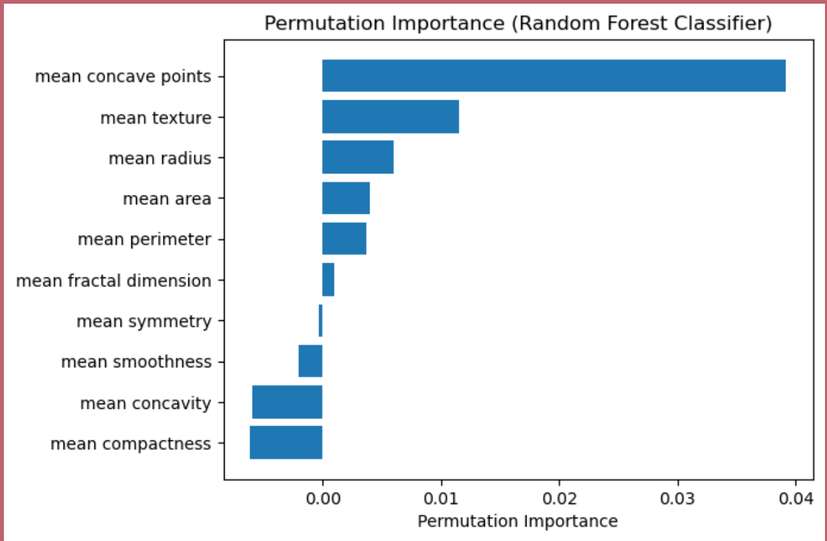
1. Firstly, split my dataset to training (60%) and testing (40%) sets, and then scale my input variables using StandardScalar().
2. Run all input features and test accuracy for all 4 models (Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and KNeighbors Classifier).
3. Then, we check for feature importance with permutation importance across all 4 models, and removed features that are under .005 importance value.
4. Next, we perform hyperparameter tuning, and find the best parameters for each model using randomsearchCV() and scoring method as 'recall' as we prioritize avoiding the labeling of malignant cells as benign.
5. Run the best parameters with the model and cross validate the accuracy of the model with 5 folds to ensure that our initial accuracy was not by luck.

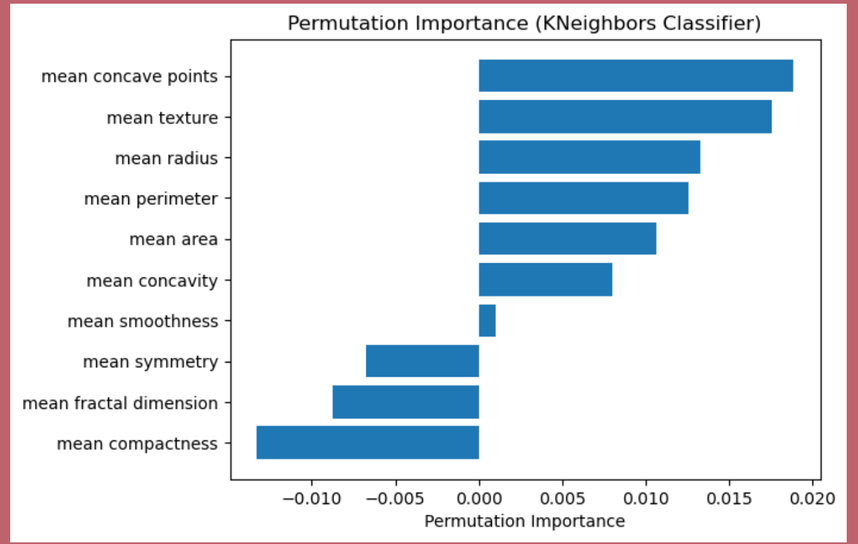
05. Evaluation (Feature Importance)



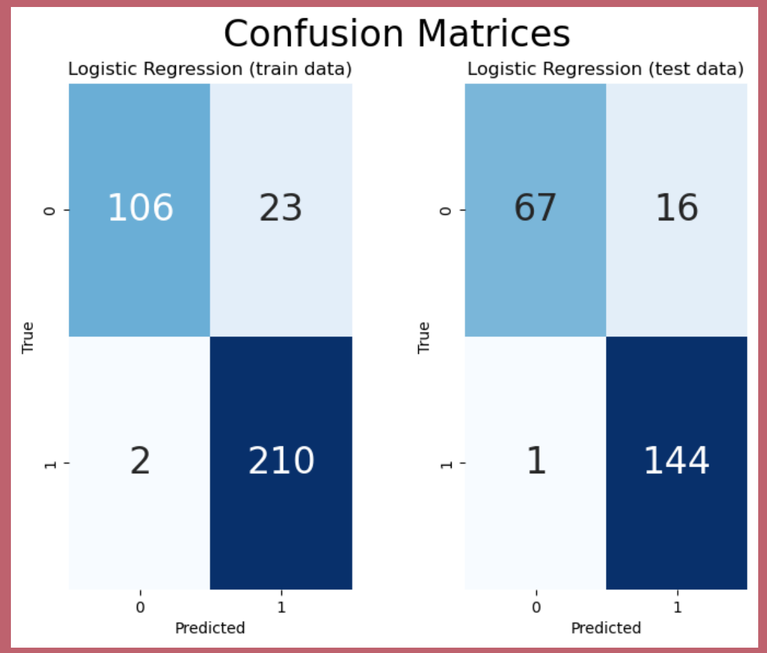


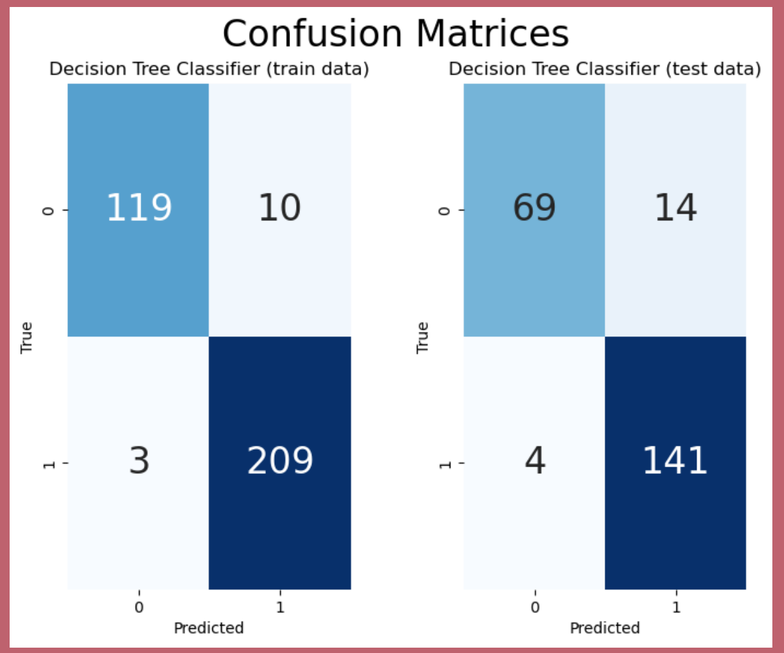
Dropping features that are <.005 importance score

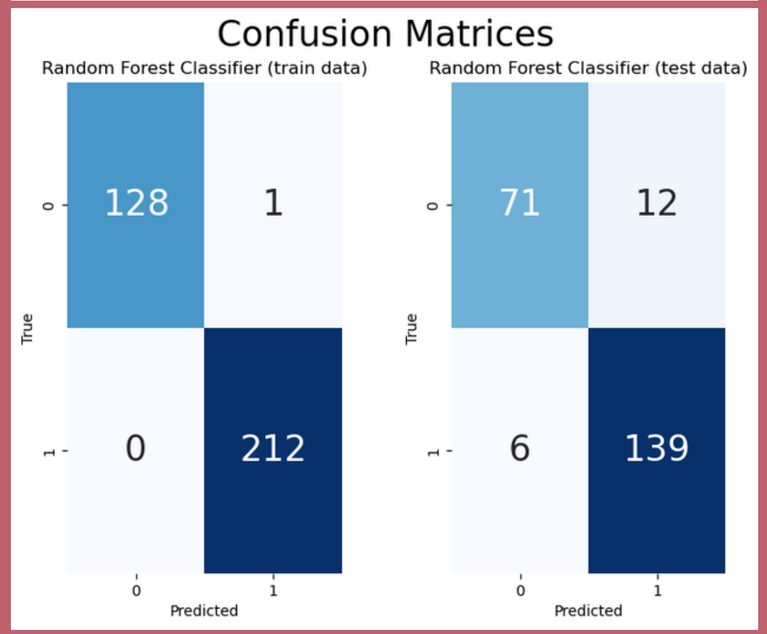


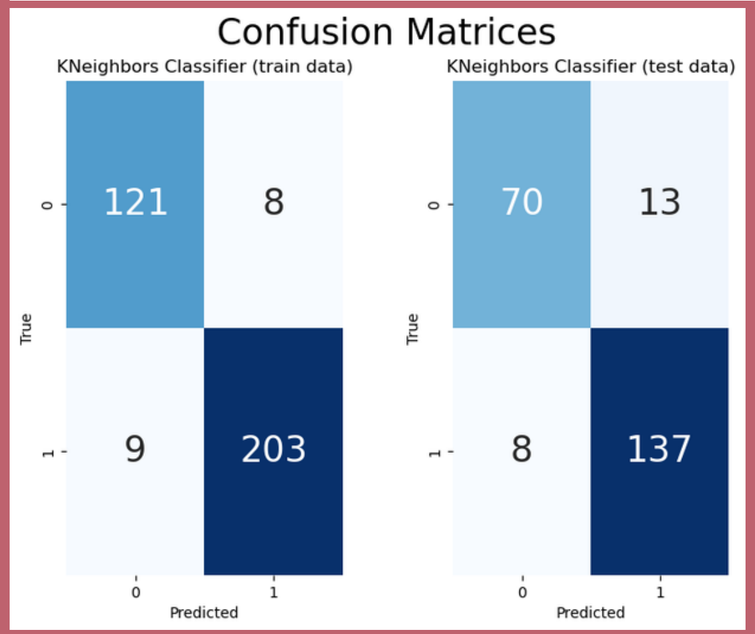


05. Evaluation (Confusion Matrix)









We can see that Logistic Regresison performs the best on recall.

05. Evaluation (Tabulated)

	Model Name	Parameters	Accuracy (Train)	Precision (Train)	Recall (Train)	F1 Score (Train)	Log Loss (Train)	Accuracy (Test)	Precision (Test)	Recall (Test)	F1 Score (Test)	Log Loss (Test)	Cross-validated Accuracy (Test)
0	Logistic Regression	Default	95.014663	94.930876	97.169811	96.037296	1.396549e-01	93.859649	92.810458	97.931034	95.302013	0.143682	96.068075
1	Logistic Regression	Best Params	92.668622	90.128755	99.056604	94.382022	2.406548e-01	92.543860	90.000000	99.310345	94.426230	0.234627	99.436620
2	Decision Tree Classifier	Default	100.000000	100.000000	100.000000	100.000000	2.220446e-16	91.666667	92.000000	95.172414	93.559322	3.003638	92.996870
3	Decision Tree Classifier	Best Params	96.187683	95.433790	98.584906	96.983759	8.086386e-02	92.105263	90.967742	97.241379	94.000000	1.198936	92.719092
4	Random Forest Classifier	Default	100.000000	100.000000	100.000000	100.000000	4.298869e-02	92.543860	92.666667	95.862069	94.237288	0.158104	96.075900
5	Random Forest Classifier	Best Params	99.706745	99.530516	100.000000	99.764706	6.158698e-02	92.105263	92.052980	95.862069	93.918919	0.158749	95.794210
6	KNeighbors Classifier	Default	94.721408	95.327103	96.226415	95.774648	1.017188e-01	91.666667	93.750000	93.103448	93.425606	0.597930	95.786385
7	KNeighbors Classifier	Best Params	95.014663	96.208531	95.754717	95.981087	1.257131e-01	90.789474	91.333333	94.482759	92.881356	0.166804	96.075900

As we can see here from our model table, logistic regression with the best parameters resulted in a cross validated accuracy of 99.44% and with a recall of 99.31%, which tells us that logistic regression works very well with very high accuracy for the given dataset to predict and classify malignant or benign breast cancer cells in the healthcare and oncology industry.

06. Recommendations

1. Since breast cancer is related to other diseases (American Association for Cancer Research, n.d.; American Cancer Study, 2022; Chen et al., 2021), future studies and researchers can make use of the model to correctly classify malignant cancer cells, together with characteristics of other illnesses for early detection and prevention of other conditions and diseases.
2. One limitation to this model is that it was developed with a relatively small dataset (569 values), future researchers and data scientists could use a larger dataset to more accurately classify malignant cancer cells.
3. A more balanced distribution of classes for the target variable in the future would help in accuracy of the model in classifying benign and malignant cancer cells.

-fin-

Thank you for listening! :)