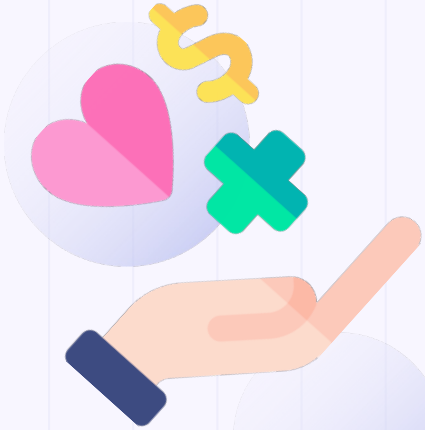
Breast Cancer Prediction

SC1015 DSAI Project

Wayne Tan Jing Heng
Chan Ming Han

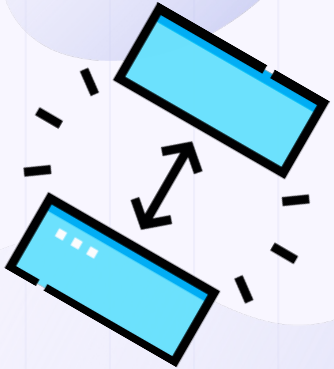
Table of contents



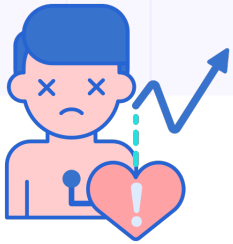
01

Problem Formulation

The need and the gap



Background

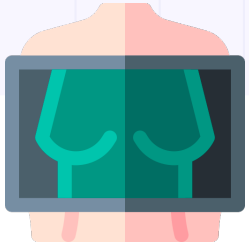


685,000

Deaths worldwide (2020)[1]

1st

Most common cancer overall[2]



1 in 8

People in the US develop breast cancer in their lifetime[3]

Background

Breast cancer is the most prevalent cancer among women in Singapore.

1 in 13 women will get breast cancer in their lifetime.

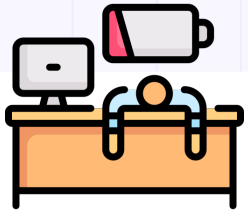
More needs to be done to tackle this problem with the resources and techniques we have.

(2020)[1]

overall[2]

develop breast cancer
in their lifetime[3]

Gap in current diagnostic methods

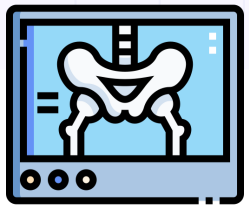


75%

surveyed doctors showed signs of disengagement and exhaustion[4]

40 million

radiologist errors per annum[5]



12.5%

screening mammograms fail to detect breast cancer[6]

Practical Motivation

We note that the diagnosis of breast cancer can be a data-driven problem and can be predicted with a variety of patient-specific data.

Ultimately, a medical professional should confirm the results, but machine learning can augment the initial prediction phase.

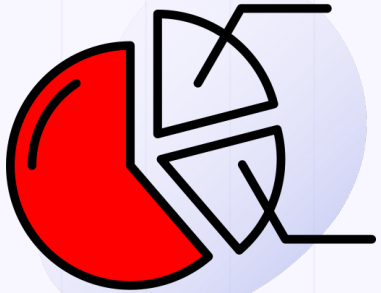
Problem Statement

How can we use machine learning models and a patient dataset to classify and predict breast cancer, to provide a reference for the early diagnosis of breast cancer?

02

Dataset

Exploratory Data Analysis



Breast Cancer Wisconsin (Diagnostic) Data Set

Our dataset was obtained from Kaggle

279239 downloads and 2190 total unique contributors

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.

30 predictor columns

These columns were split into 3 categories: mean, standard, and worst.

Mean

We chose the 10 columns for the mean category

569 rows

Before data cleaning

Data Cleaning & Preparation

Goal

To clean the dataset, remove outliers, and select the top 5 features/columns as the best 5 predictors for the response of Benign/Malignant breast cancer

Data Cleaning & Preparation

Steps

1

Data Visualisation

Provided us with a clearer understanding of the dataset

2

Outlier Removal

Outliers were removed by trimming data points outside the whiskers

3

Feature Selection

Obtain top 5 features through analyzing skewness and ANOVA hypothesis test

Data Cleaning & Preparation

Redundant Columns

A column was dropped since it had inconsistent data formatting and did not value-add

Outlier Removal

Observed outliers from boxplots and by printing the count of values beyond the whiskers

These data points were trimmed from the dataset

Skew Analysis

Skewness can lead to inaccuracies in statistical models, especially regression-based ones

Our analysis showed that 5 columns had a positive skew but none had a negative skew

Feature Selection

To determine whether or not a numerical variable is correlated to a categorical variable

We selected the top 5 features to use for correlation to the diagnosis of Benign or Malignant

Data Cleaning & Preparation

Redundant Columns

An ‘unnamed’ column was dropped since it had inconsistent data formatting and did not value-add

Outlier Removal

From boxplots and by printing the count of values beyond the whiskers, we noted outliers that may lead to inaccurate results

These data points were trimmed from the dataset

Skew Analysis

Skewness can lead to inaccuracies in statistical models, especially regression-based ones

Our analysis showed that 5 columns had a positive skew but none had a negative skew

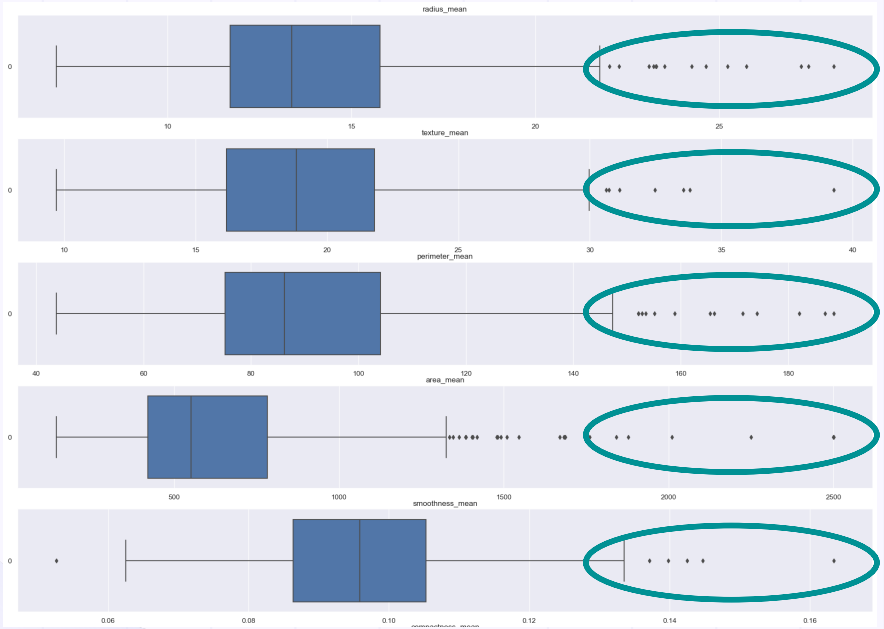
ANOVA Hypothesis Test

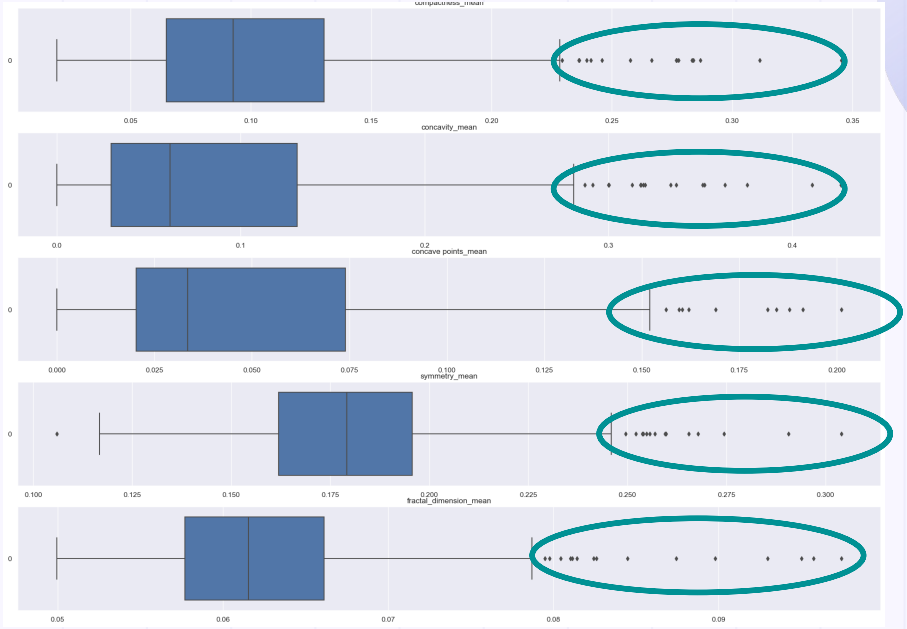
To determine whether or not a numerical variable is correlated to a categorical variable

We selected the top 5 features to use for correlation to the diagnosis of Benign of Malignant

Outlier Removal

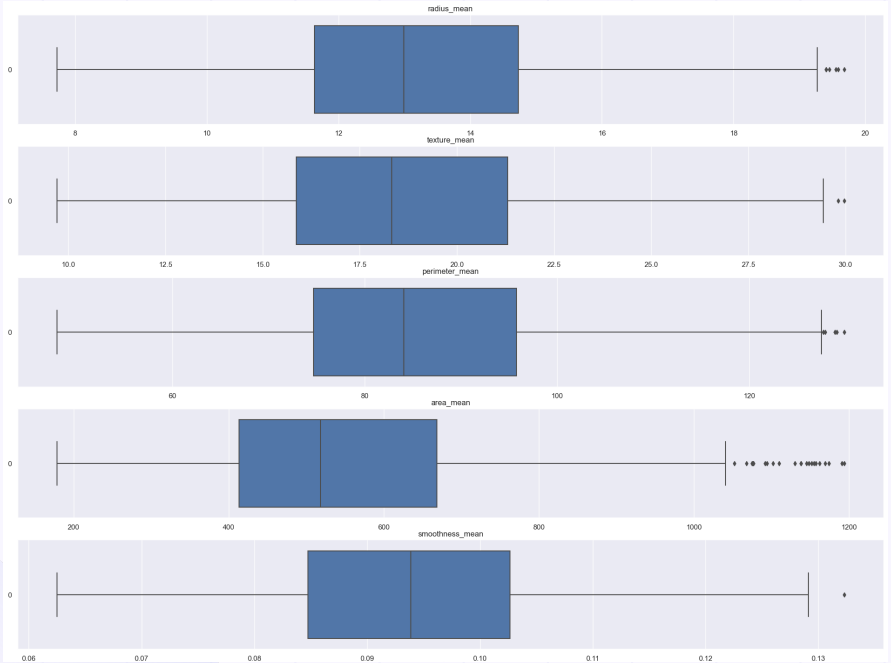
Boxplots for Outlier Analysis

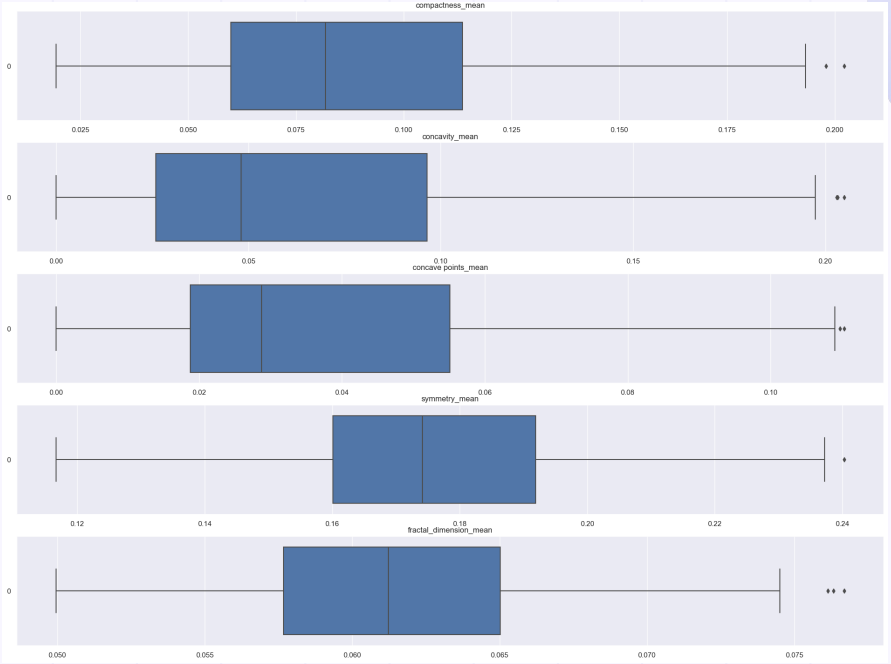




Outlier Removal

Boxplots: Outliers Removed





Data Cleaning & Preparation

Redundant Columns

An ‘unnamed’ column was dropped since it had inconsistent data formatting and did not value-add

Outlier Removal

From boxplots and by printing the count of values beyond the whiskers, we noted outliers that may lead to inaccurate results

These data points were trimmed from the dataset

Skew Analysis

Skewness can lead to inaccuracies in statistical models, especially regression-based ones

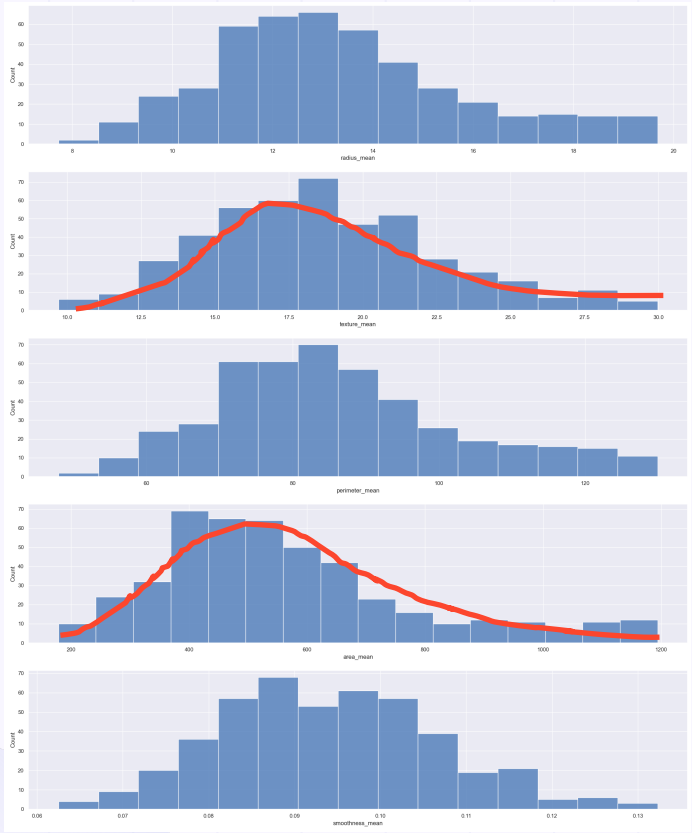
Our analysis showed that 5 columns had a positive skew but none had a negative skew

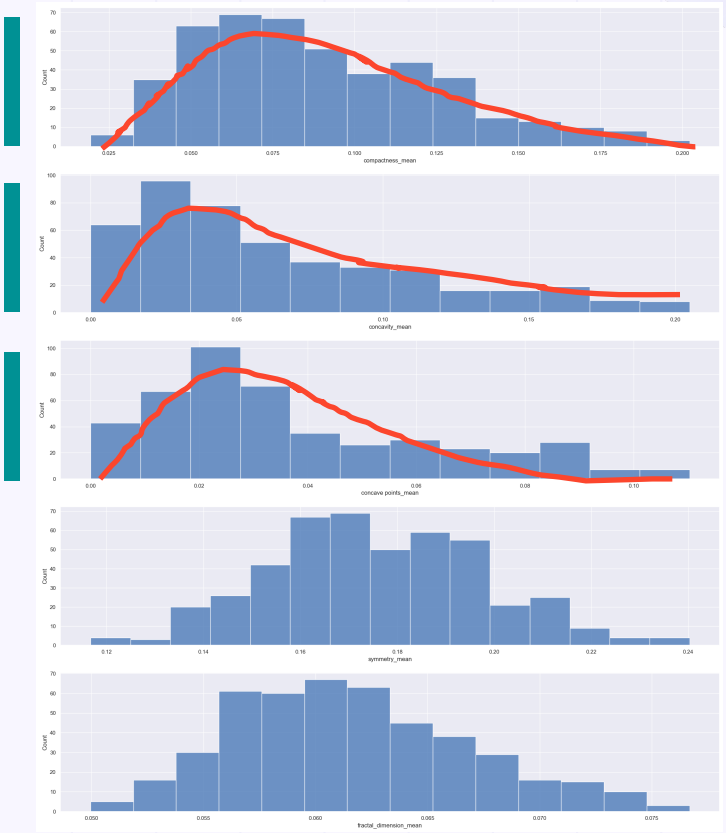
ANOVA Hypothesis Test

To determine whether or not a numerical variable is correlated to a categorical variable

We selected the top 5 features to use for correlation to the diagnosis of Benign of Malignant

Skew Analysis





Skew Analysis

However, this only provides an early sensing as to which variables might be more suitable.

We cannot choose our variables based on this information yet.

To arrive at a more convincing answer, we have to dive deeper by examining the swarm plots for each variable.

Data Cleaning & Preparation

Redundant Columns

An ‘unnamed’ column was dropped since it had inconsistent data formatting and did not value-add

Outlier Removal

From boxplots and by printing the count of values beyond the whiskers, we noted outliers that may lead to inaccurate results

These data points were trimmed from the dataset

Skew Analysis

Skewness can lead to inaccuracies in statistical models, especially regression-based ones

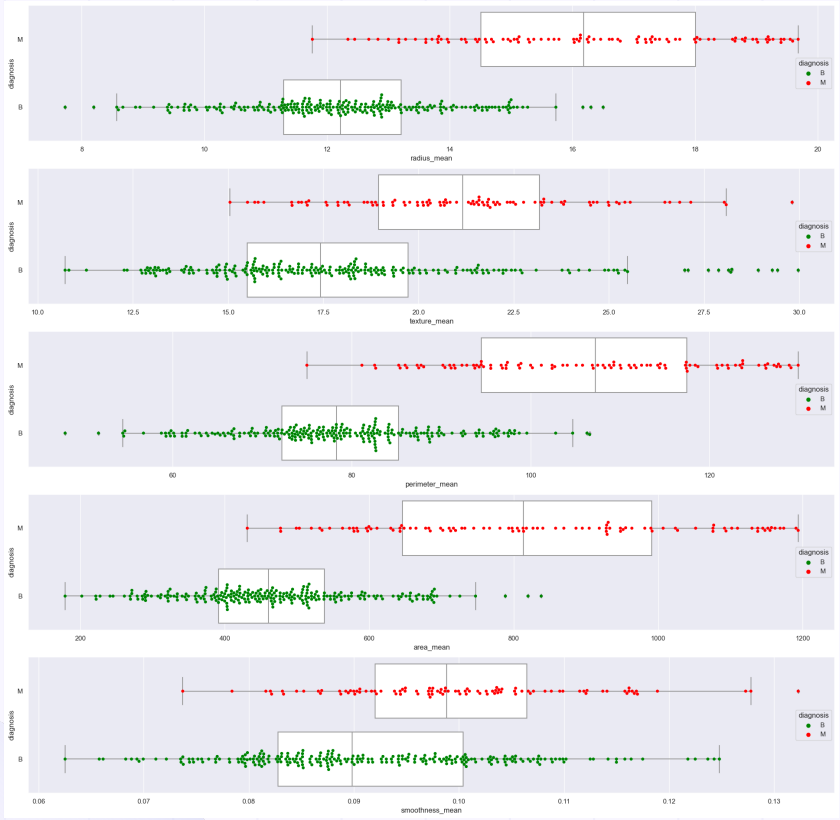
Our analysis showed that 5 columns had a positive skew but none had a negative skew

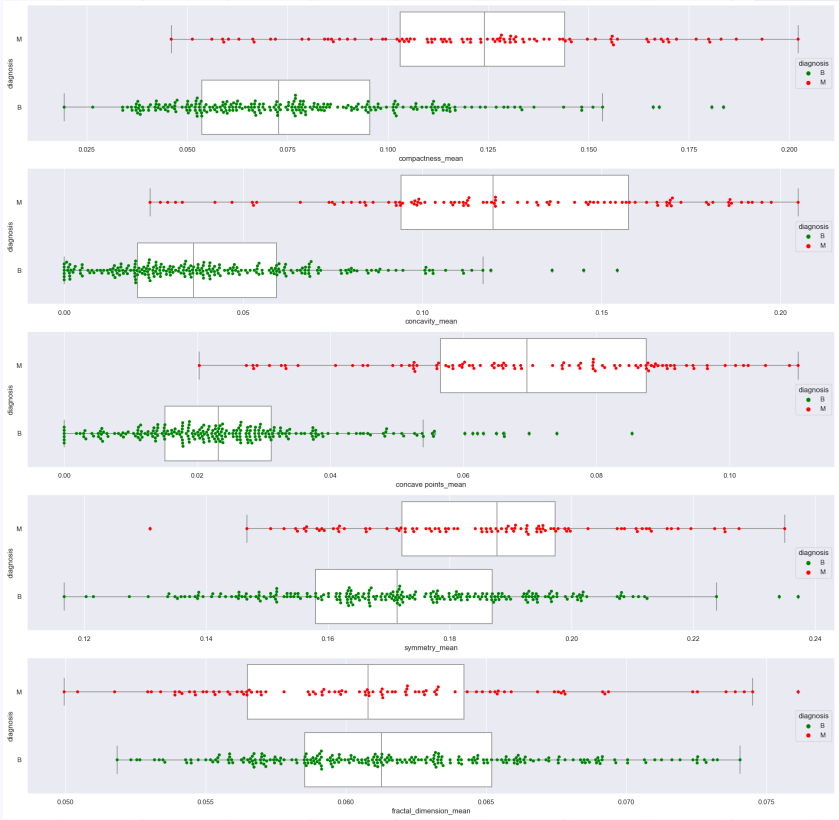
Feature Selection

Determines whether a numerical variable is correlated to a categorical variable

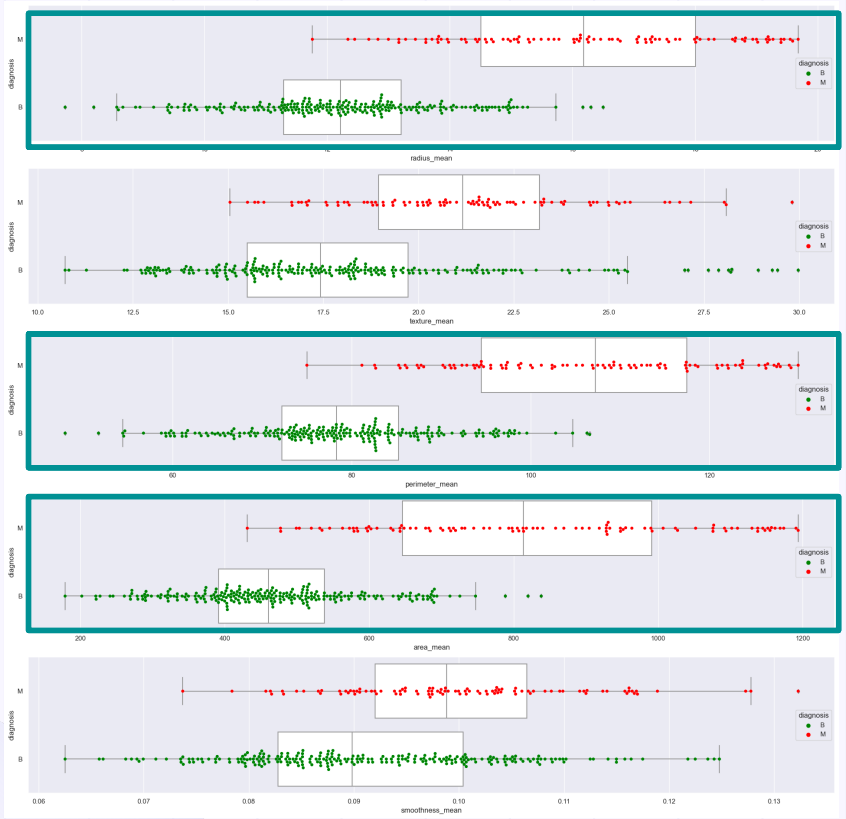
We selected the top 5 features to use for correlation to the diagnosis of Benign of Malignant

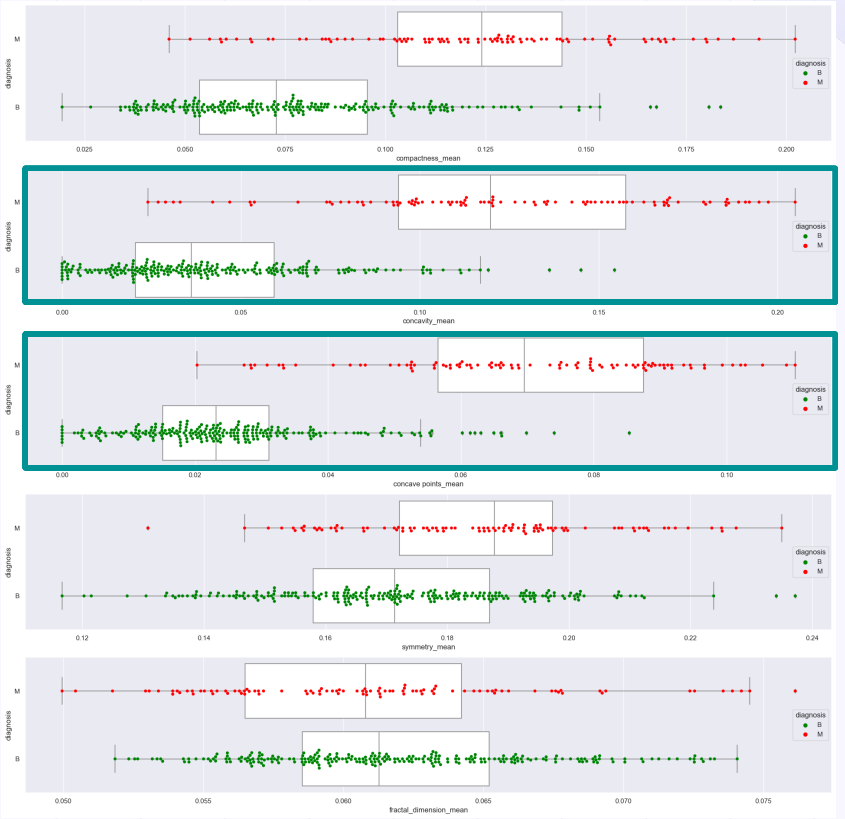
Feature Selection: Swarm Plots





Feature Selection: Swarm Plots





Feature Selection: Initial Top 5

1. Radius Mean
2. Concavity Mean
3. Concave Points Mean
4. Perimeter Mean
5. Area Mean

Feature Selection: ANOVA Hypothesis Test

Most suitable in our case as our data set comprised of a categorical target variable (MALIGNANT or BENIGN), and other numerical predictors, such as perimeter, area, etc

The ANOVA Hypothesis test was conducted with the help of the f_oneway function in the scipy.stats module

The P-value is defined as the probability of obtaining a result equal to or more extreme than what was actually observed (null hypothesis)

Hence, we decided to rank the features by their P-value, and then selecting the top 5 features based on the smallest P-values

Feature Selection: ANOVA Hypothesis Test

	Feature	P-Value
0	concave points_mean	4.734656e-68
1	area_mean	3.246072e-58
2	perimeter_mean	9.893705e-58
3	concavity_mean	4.462167e-57
4	radius_mean	1.245378e-54
5	compactness_mean	5.309499e-28
6	texture_mean	1.516865e-12
7	smoothness_mean	5.330512e-09
8	symmetry_mean	3.646212e-08
9	fractal_dimension_mean	1.608488e-01

Feature Selection: ANOVA Hypothesis Test

	Feature	P-Value
0	concave points_mean	4.734656e-68
1	area_mean	3.246072e-58
2	perimeter_mean	9.893705e-58
3	concavity_mean	4.462167e-57
4	radius_mean	1.245378e-54
5	compactness_mean	5.309499e-28
6	texture_mean	1.516865e-12
7	smoothness_mean	5.330512e-09
8	symmetry_mean	3.646212e-08
9	fractal_dimension_mean	1.608488e-01

Feature Selection: Final Top 5

0	concave points_mean	4.734656e-68
1	area_mean	3.246072e-58
2	perimeter_mean	9.893705e-58
3	concavity_mean	4.462167e-57
4	radius_mean	1.245378e-54

Concave Points Mean

Area Mean

Perimeter Mean

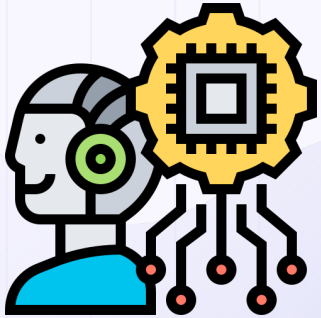
Concavity Mean

Radius Mean

03

Machine Learning

Models and Rationale



Machine Learning Models Chosen

Decision Tree Classifier

Logistic Regression

Random Forest

Support Vector Machine with Grid Search

Rationale For Model Selection

Decision Tree Classifier

Logistic Regression

Random Forest

Support Vector Machine with Grid Search

Dataset Types

Our dataset comprised predominantly continuous variables, and our goal was to predict a categorical variable.

Interpretability

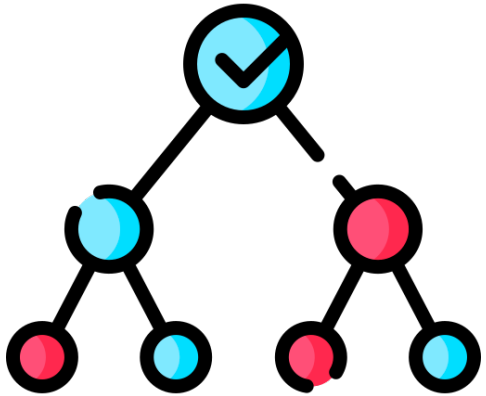
Because the model is meant to be used in a medical context, it is important for the model to be interpretable.

Performance

The model needed to perform well on metrics which were used for validation in the healthcare industry.

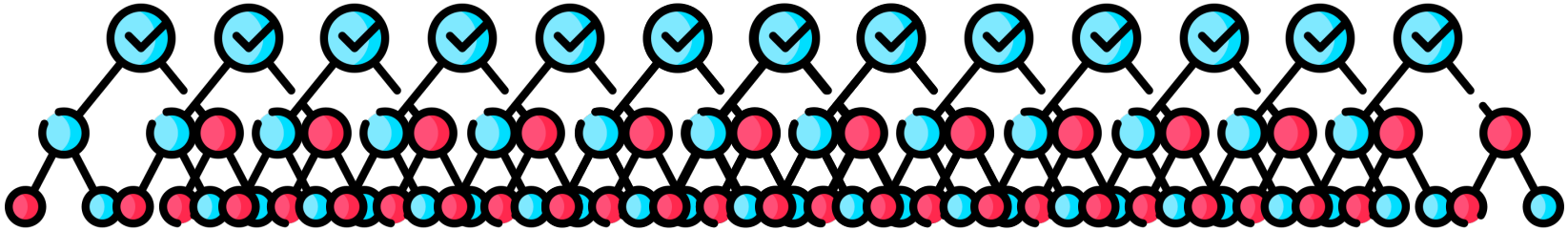
Decision Tree

- A graphical representation of decision-making processes or decision rules that can be used to predict outcomes based on input features of variables.
- **Automatic Feature Selection:** Decision Trees can automatically select important variables from the input data.
- **Interpretability:** Decision Trees are highly interpretable, giving them an advantage in the healthcare setting.
- **Ease of use:** Decision Trees are easy to implement, and do not require complex calculations or intensive computational resources.



Random Forest

- An ensemble learning technique which *combines multiple decision trees* to create a more accurate and robust model.
- **Improved Accuracy** compared to a single decision tree.
- **Resilience to Overfitting[7]:** Random Feature selection of 400-1200 decision trees make Random Forest less prone to overfitting.

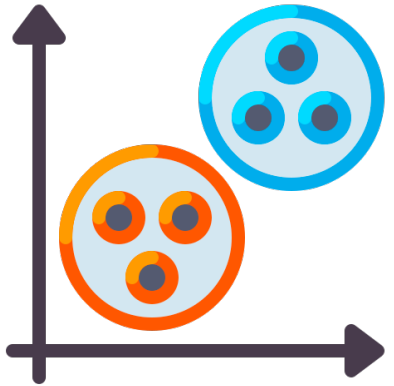


Logistic Regression

- Logistic regression is a statistical method used for binary classification, where the goal is to predict the probability of an input instance belonging to one of two classes.
- **Simplicity:** Logistic Regression models are easily implementable and highly interpretable.
- **Versatility:** Able to be used for a wide range of binary classification tasks

Support Vector Machine[8]

- Support Vector Machine (SVM) is a supervised machine learning technique which are good for classification problems with small dataset sizes.
- **Non-Linearity:** SVM is good with dealing with non-linear data as a result of their inbuilt kernels (polynomial, rbf, sigmoid etc.)
- **Hyperparameters:** SVM has a few hyperparameters, making it easier to tune.
- **Grid Search:** Computationally inexpensive in SVM implementation due to the few hyperparameters.

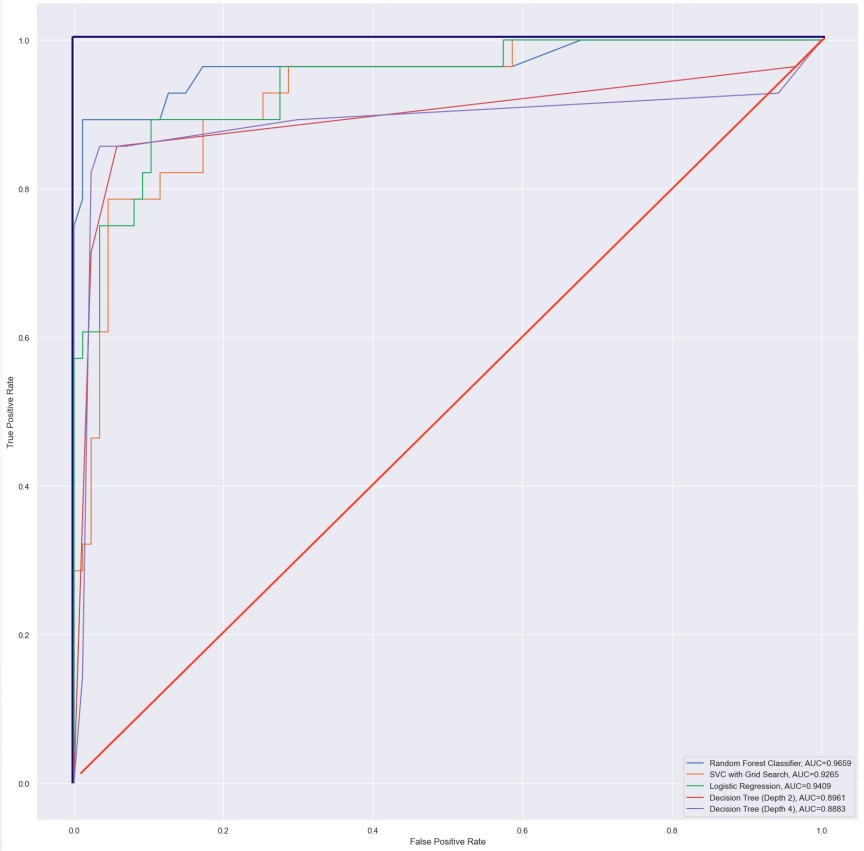


04

Model Comparison

Findings and Insights

Model Comparison (ROC)



An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.

- Widely used by clinicians in clinical trials[9]

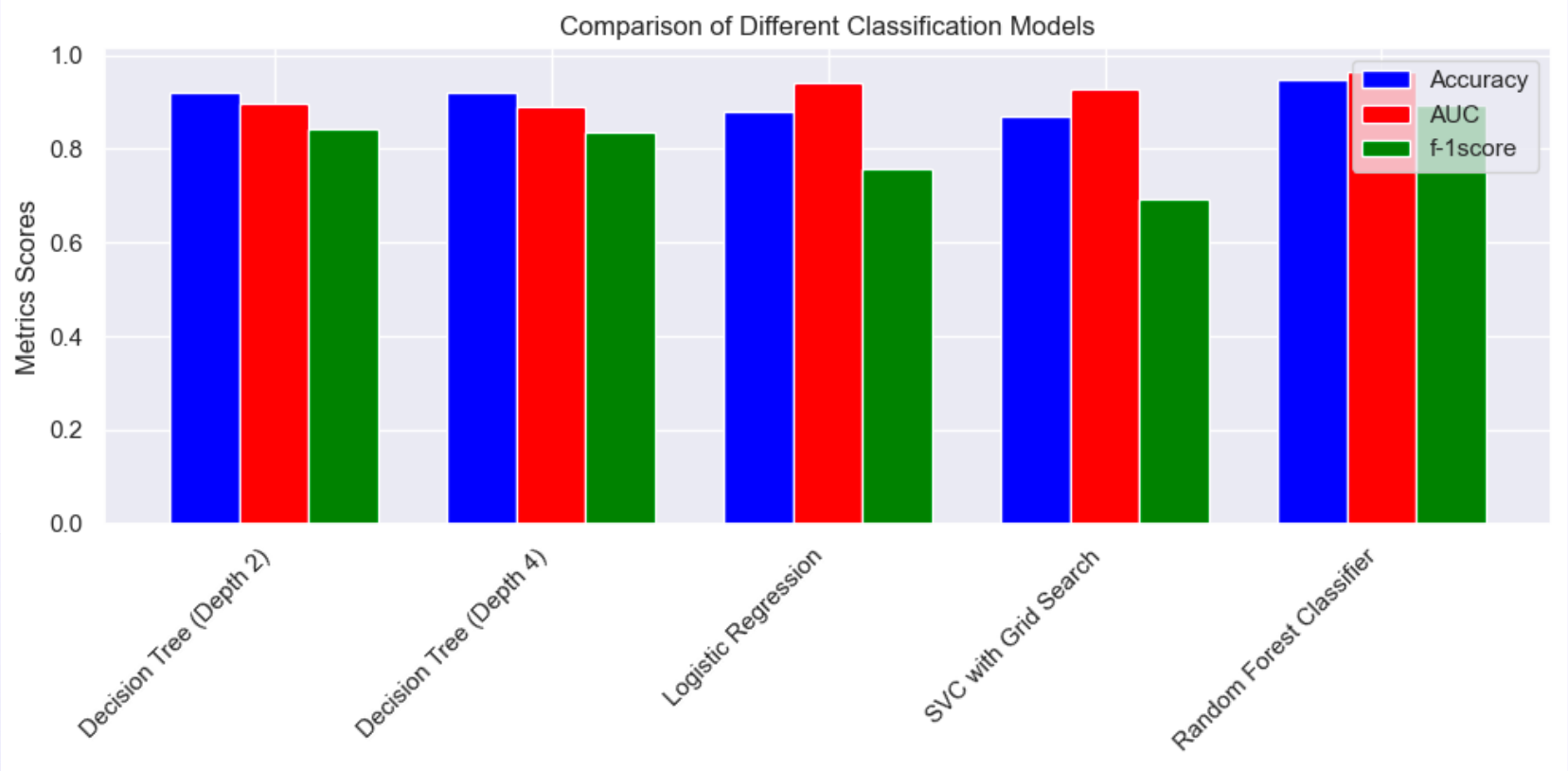
Metrics

- **TPR = FPR :** Random Classifier
- **FPR = 0, TPR = 1** : Perfect Classifier

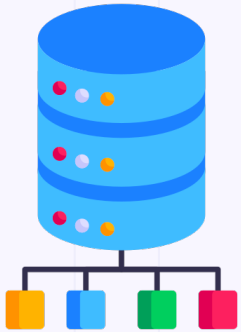
Model Comparison

	Decision Tree (depth=2)	Decision Tree (depth=4)	Logistic Regression	SVM with grid search	Random Forest
Accuracy	0.9217	0.9217	0.8783	0.8696	0.9478
AUC	0.8961	0.8883	0.9409	0.9265	0.9659
f1-Score	0.8421	0.8364	0.7586	0.6939	0.8928

Model Comparison

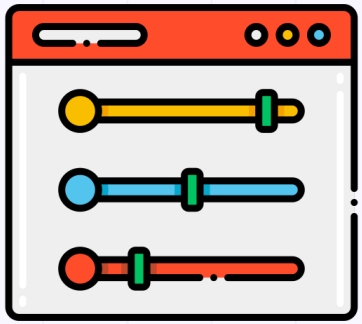


Recommendations



Dataset Selection

The models can be trained on larger and more extensive datasets, to improve the robustness of the models.



Hyperparameter Tuning

Hyperparameters of the various models can also be automatically tuned during training using informed search strategies to reduce training time.

References

[1] *Breast Cancer - Statistics*. (2023, February 23). Cancer.Net. https://www.cancer.net/cancer-types/breast-cancer/statistics#:~:text=In%202020%2C%20an%20estimated%20684%2C996,United%20States%20after%20lung%20cancer.

[2] *Common Cancer Types*. (2023, March 7). National Cancer Institute. https://www.cancer.gov/types/common-cancers#:~:text=The%20most%20common%20type%20of,are%20combined%20for%20the%20list.

[3] *Breast Cancer Statistics | How Common Is Breast Cancer?* (n.d.). https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html#:~:text=Lifetime%20chance%20of%20getting%20breast,she%20will%20develop%20breast%20cancer.

[4] Auto, H. (2021, August 6). More doctors in Singapore face burnout, anxiety amid the pandemic. *The Straits Times*. https://www.straitstimes.com/life/more-doctors-in-singapore-face-burnout-anxiety-amid-the-pandemic#:~:text=It%20surveyed%203%2C075%20healthcare%20workers,signs%20of%20disengagement%20and%20exhaustion.

[5] Brady, A. P. (2017). Error and discrepancy in radiology: inevitable or avoidable? *Insights Into Imaging*, *8*(1), 171–182. https://doi.org/10.1007/s13244-016-0534-1

[6] *Limitations of Mammograms | How Accurate Are Mammograms?* (n.d.). https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms/limitations-of-mammograms.html#:~:text=Overall%2C%20screening%20mammograms%20miss%20about,when%20in%20fact%20they%20do.

[7] Dubey, A. (2023, April 4). Feature Selection Using Random forest - Towards Data Science. *Medium*. https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f#:~:text=How%20does%20Random%20forest%20select,random%20extraction%20of%20the%20features.

[8] Yadav, A. (2018, October 22). SUPPORT VECTOR MACHINES(SVM) - Towards Data Science. *Medium*. https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589

[9] Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. *Korean Journal of Anesthesiology*, *75*(1), 25–36. https://doi.org/10.4097/kja.21209