



SCI1015 PROJECT: Breast Cancer Prediction

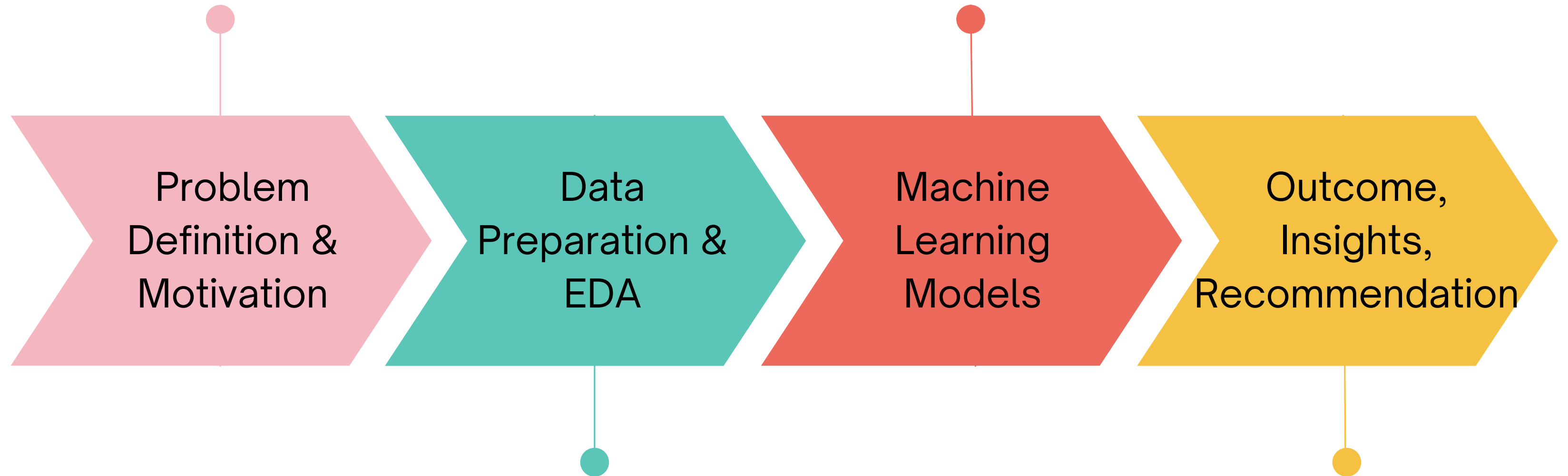


FDAE - Team 5

Fong Zheng Feng Victor
Thuvaarakesh Kiruparan
Tiang Soon Yong



CONTENTS



PROBLEM DEFINITION & MOTIVATION



Statistics of cancers in Singapore



Ten most common cancers affecting men & women (2017 - 2021)

Men	No. of cases		Women	No. of cases
Prostate	6,912		Breast	12,735
Colon & rectum	6,697		Colorectal & rectum	5,542
Lung	5,567		Lung	3,388
Lymphoid neoplasms	2,986		Corpus uteri (uterus)	3,133
Liver	2,984		Lymphoid neoplasms	2,221
Non-melanoma skin	2,136		Ovary & fallopian tube	1,855
Kidney	1,734		Non-melanoma skin	1,713
Stomach	1,684		Thyroid	1,666
Myeloid neoplasms	1,430		Pancreas	1,187
Pancreas	1,417		Stomach	1,111
			Cervix uteri	1,106

12,735 cases, highest cases among the rest.
Forms up 18.4% of all the cases

Statistics of cancers in Singapore



Overview. Breast cancer is the most commonly occurring cancer among women in Singapore accounting for 29.7% of all cancers diagnosed in females*. *1 in 13 women will get breast cancer in their lifetime. Breast cancer usually originates from the cells lining the milk ducts and glands.



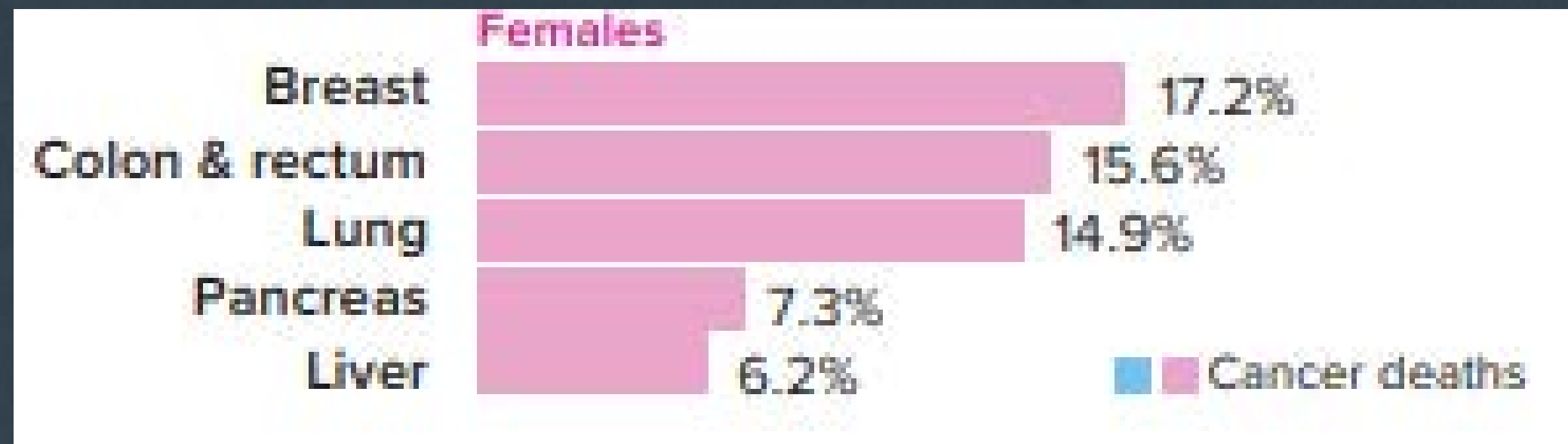
Singapore Cancer Society

<https://www.singaporecancersociety.org.sg/breast-can...>

Breast Cancer - Singapore Cancer Society



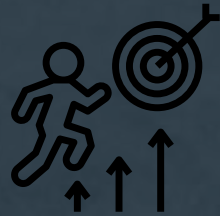
About featured snippets • Feedback



Breast Cancer is the leading cause of cancer deaths in Singapore

Problem Definition & Motivation

Motivation:



Breast cancer is one of the leading causes of cancer-related deaths among women. Early detection plays a crucial role in improving treatment outcomes and reducing mortality rates. By accurately predicting the malignancy of breast tumors, healthcare providers can intervene promptly and initiate appropriate treatment strategies.

Problem Definition:



Develop a predictive model to classify breast tumors as malignant or benign based on relevant clinical and diagnostic features. The model should accurately distinguish between malignant and benign tumors to assist healthcare professionals in early diagnosis and treatment decision-making.

DATA PREPARATION & EXPLORATORY DATA ANALYSIS



Initial Insights



of Data Columns: 32

of Data Samples: 569 non-null

“id” is irrelevant

of Features: 30 Numerical

Response: 1 Categorical, “Diagnosis”

Classification Problem

Mixed Structured Data

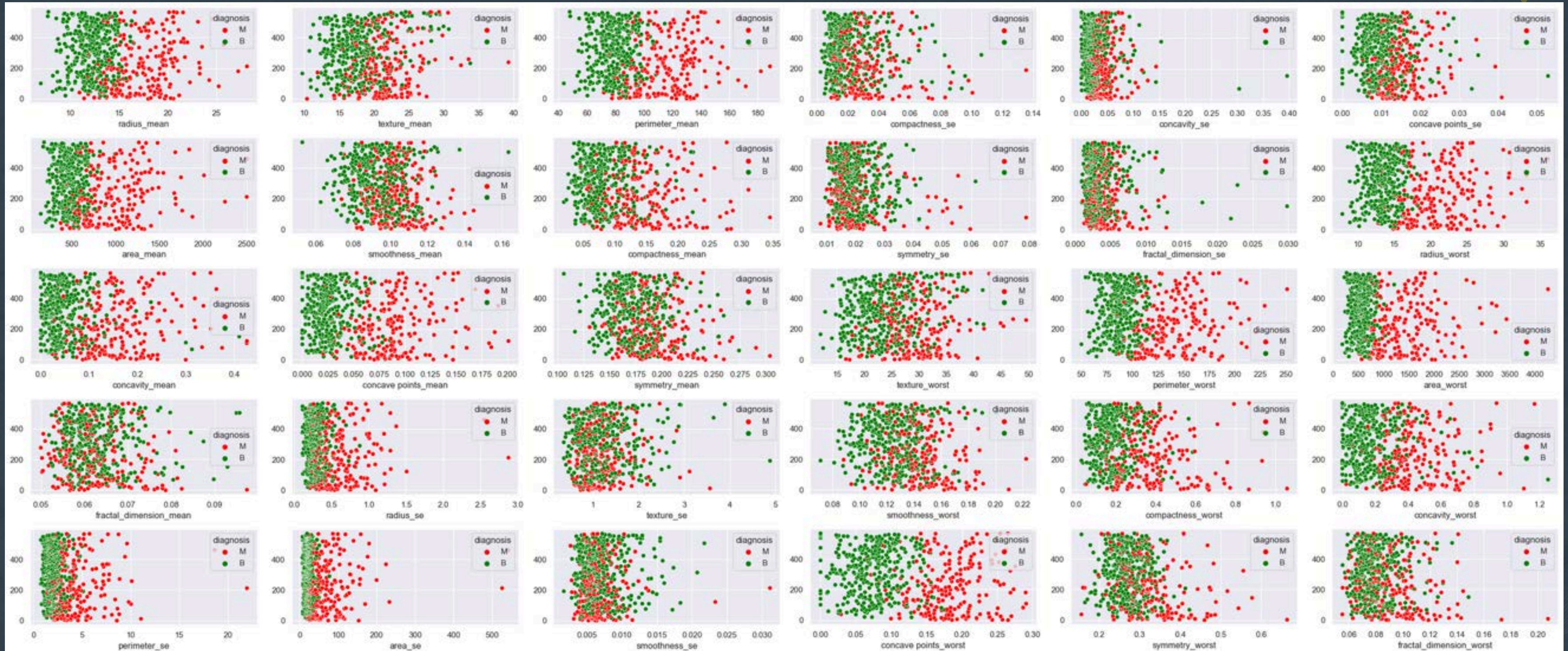


	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	rac
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	

5 rows × 32 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 32 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   id                                     569 non-null    int64
1   diagnosis                             569 non-null    object
2   radius_mean                           569 non-null    float64
3   texture_mean                           569 non-null    float64
4   perimeter_mean                         569 non-null    float64
5   area_mean                             569 non-null    float64
6   smoothness_mean                       569 non-null    float64
7   compactness_mean                      569 non-null    float64
8   concavity_mean                        569 non-null    float64
9   concave points_mean                   569 non-null    float64
10  symmetry_mean                         569 non-null    float64
11  fractal_dimension_mean                 569 non-null    float64
12  radius_se                             569 non-null    float64
13  texture_se                             569 non-null    float64
14  perimeter_se                           569 non-null    float64
15  area_se                               569 non-null    float64
16  smoothness_se                         569 non-null    float64
17  compactness_se                        569 non-null    float64
18  concavity_se                          569 non-null    float64
19  concave points_se                     569 non-null    float64
20  symmetry_se                           569 non-null    float64
21  fractal_dimension_se                   569 non-null    float64
22  radius_worst                          569 non-null    float64
23  texture_worst                         569 non-null    float64
24  perimeter_worst                       569 non-null    float64
25  area_worst                            569 non-null    float64
26  smoothness_worst                      569 non-null    float64
27  compactness_worst                     569 non-null    float64
28  concavity_worst                       569 non-null    float64
29  concave points_worst                   569 non-null    float64
30  symmetry_worst                        569 non-null    float64
31  fractal_dimension_worst                 569 non-null    float64
dtypes: float64(30), int64(1), object(1)
```

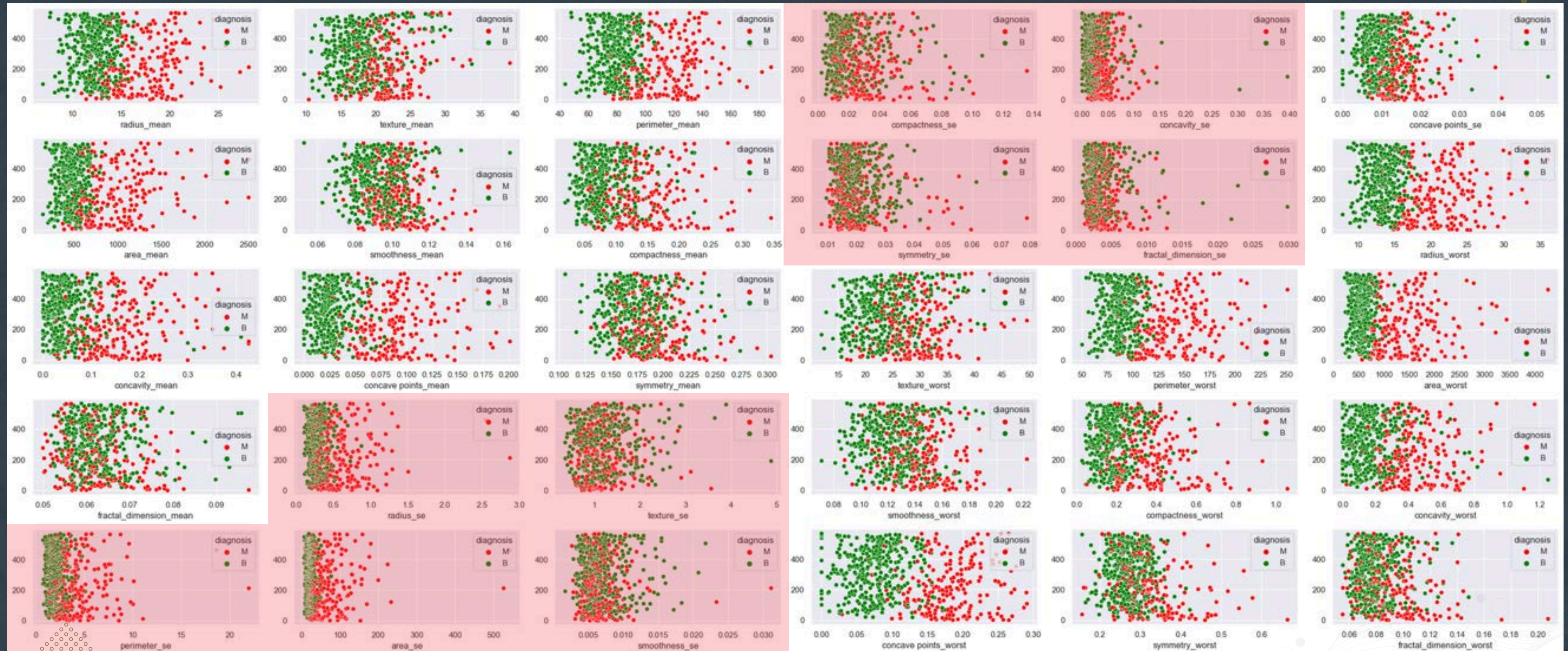

Initial Insights



Scatterplots of Features against Response

Legend: ● Malignant
● Benign

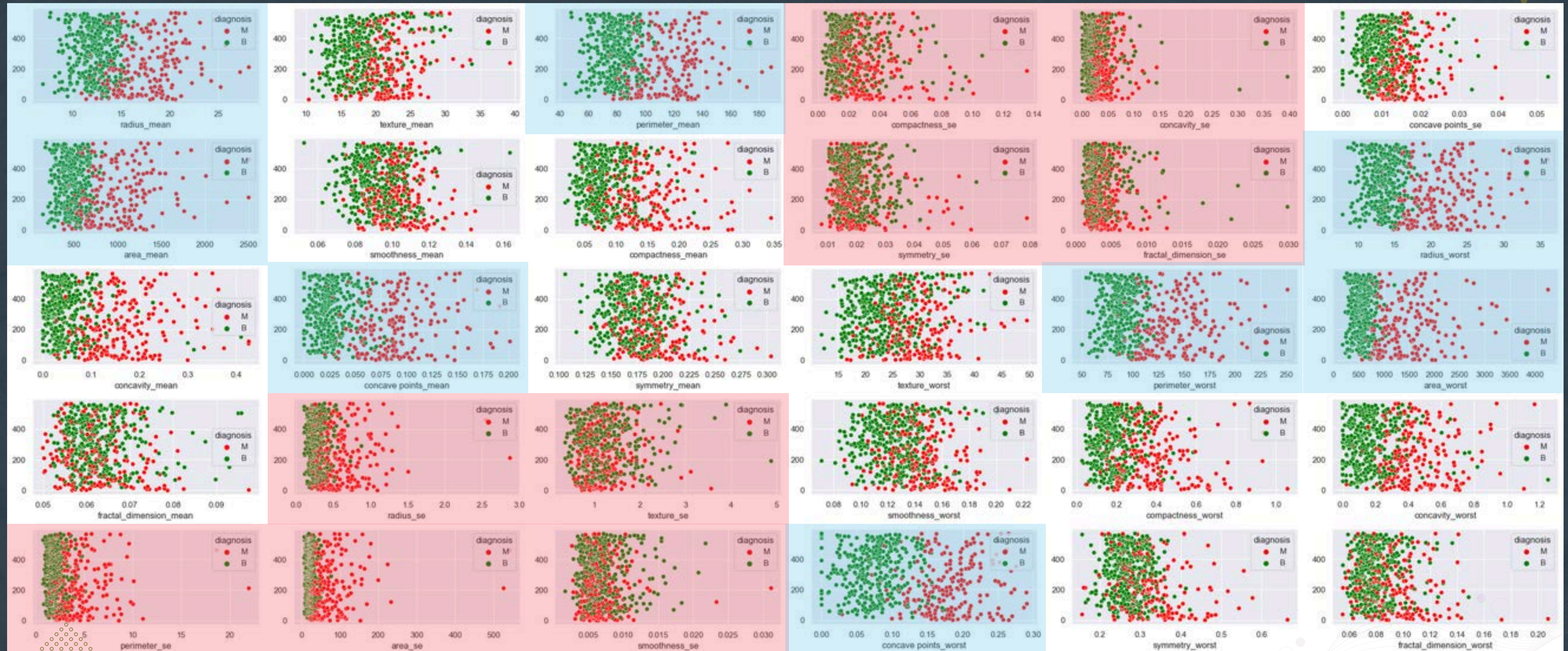
Initial Insights



Scatterplots of Features against Response

Legend: ● Malignant
● Benign

Initial Insights



Scatterplots of Features against Response

Legend: ● Malignant
● Benign

Initial Insights



Outliers Cleaning



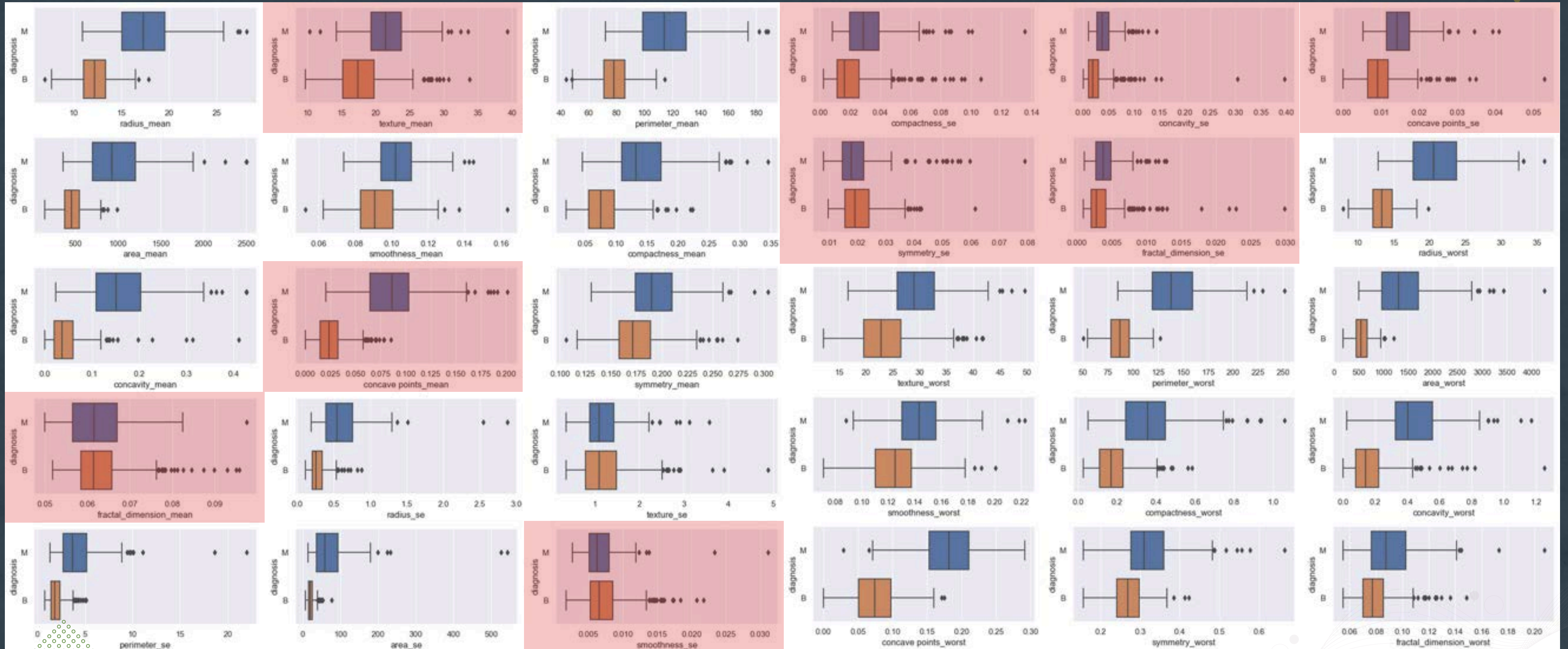
Data Balancing



Feature Selection



Outlier Cleaning



Boxplots of Features against Response

Outlier Cleaning

Method:

IQR Threshold: $1.5 \times \text{IQR}$

Z-Score Threshold: $2 \times \text{S.D}$

Findings:

Outliers removal causes

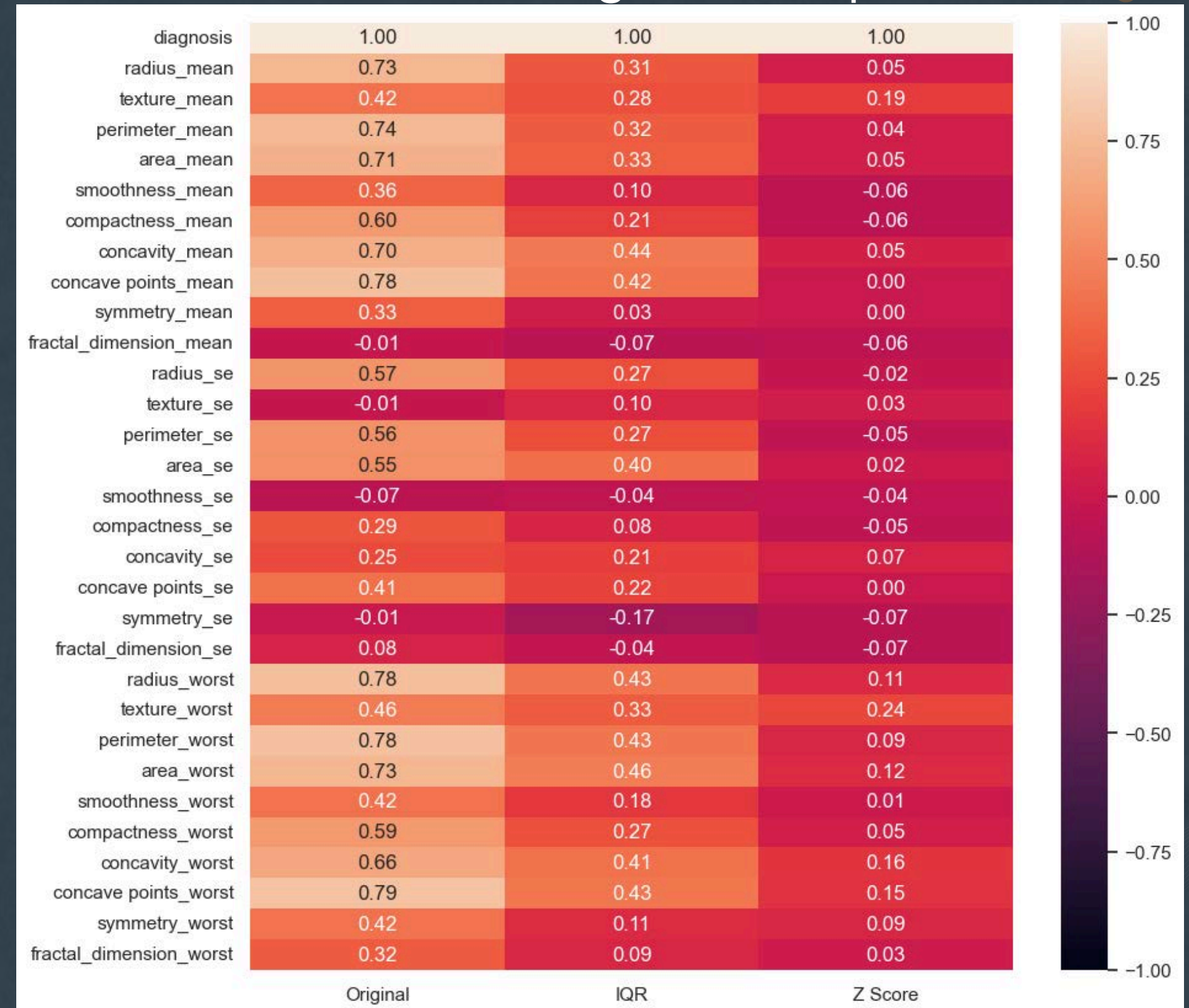
- Correlation **decreased**
- Lost of **meaningful** data
- Reduced **variability**

Decision:

Proceed with Original dataset



Correlation against Response



Data Balancing

Question:

Upsample before/after train_test_split()?

Method:

DecisionTreeClassifier()

Findings:

of data samples before upsample: 569

of data samples after upsample: 714

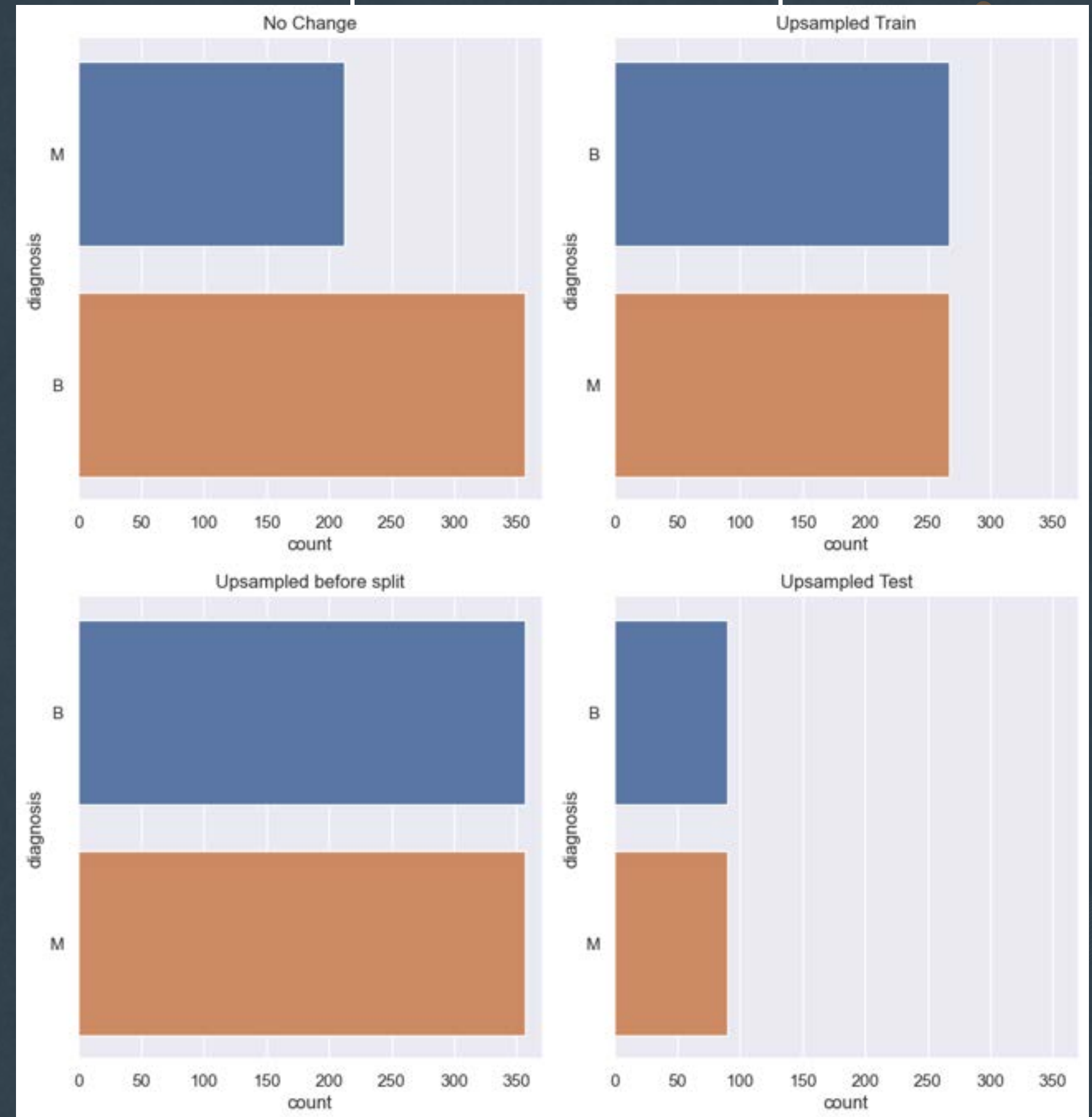
Classification Accuracy

	Train	Test
Original	1.0	0.909091
Upsample Before	1.0	0.966480
Upsample After	1.0	0.922222

Decision:

Proceed with the dataset upsampled before train_test_split()

Boxplots of Data Sample



Feature Selection

Question:

How many features should we use?

Method:

Recursive feature elimination with cross-validation (**RFECV**)

Findings:

Accuracy starts relatively **high** (0.85) even with 1 feature

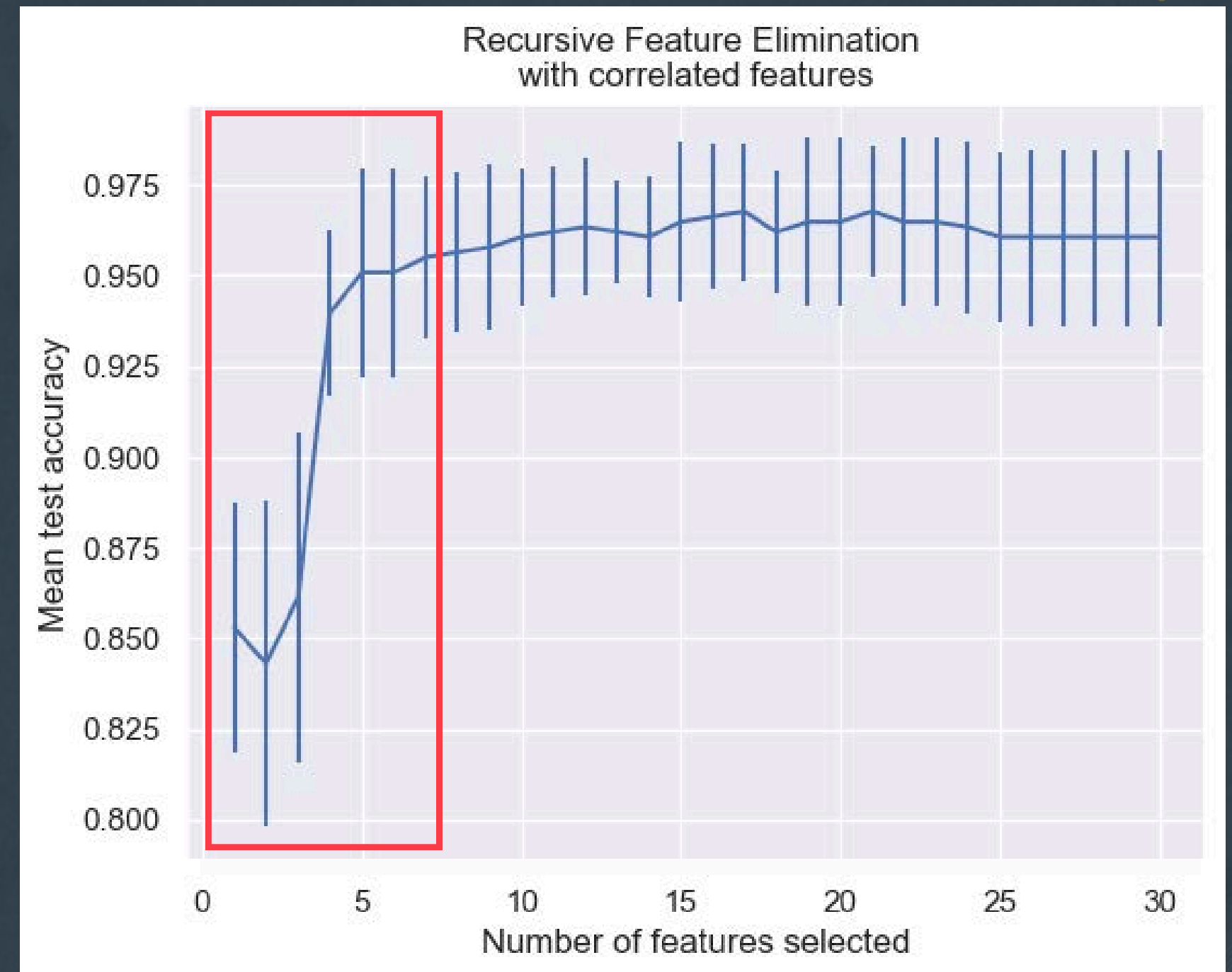
Accuracy **stagnates** after 5 features

More Features = More **Vulnerable** to errors

Decision:

Test our models with [1, 3, 5, 7] Features

Line Plot with Errorbar



Feature Selection

Question:

Which features should we use?

Method:

SelectKBest
r_regression

Findings:

Sorted Feature List according to its score

Decision:

Feature List 0 = Top 7

Feature List 1 = Top 5

Feature List 2 = Top 3

Feature List 3 = Top 1

Top 7

	Features	F_Scores	Abs_Corr
27	concave points_worst	1268.834068	0.800347
22	perimeter_worst	1064.277962	0.774055
20	radius_worst	1043.774301	0.771026
7	concave points_mean	1026.771557	0.768450
2	perimeter_mean	844.058725	0.736501
0	radius_mean	788.349625	0.724875
23	area_worst	761.055501	0.718784
6	concavity_mean	668.240159	0.695807
3	area_mean	635.460792	0.686730
26	concavity_worst	591.151026	0.673522
5	compactness_mean	412.370230	0.605604
25	compactness_worst	374.511520	0.587105
10	radius_se	281.321498	0.532178
12	perimeter_se	254.017559	0.512790
13	area_se	232.364209	0.496038
21	texture_worst	211.132272	0.478239
1	texture_mean	174.173304	0.443334
24	smoothness_worst	166.290037	0.435125
17	concave points_se	157.303993	0.425387
28	symmetry_worst	155.327814	0.423188
4	smoothness_mean	130.955007	0.394148
8	symmetry_mean	111.988029	0.368659
29	fractal_dimension_worst	92.296412	0.338754
15	compactness_se	77.120465	0.312617
16	concavity_se	67.103586	0.293478
19	fractal_dimension_se	9.574370	0.115190
14	smoothness_se	4.053819	0.075242
18	symmetry_se	0.268262	0.019407
11	texture_se	0.125171	0.013258
9	fractal_dimension_mean	0.016431	0.004804

MACHINE LEARNING MODELS

- Logistic Regression
- Binary Decision Tree
- Random Forest
- K Nearest Neighbour



ACCURACY & FALSE NEGATIVE RATE (FNR)

LOGISTIC REGRESSION

- Supervised Learning Model
- Test for binary outcomes
- Logistic Regression without parameter tuning
- Logistic Regression with parameter tuning



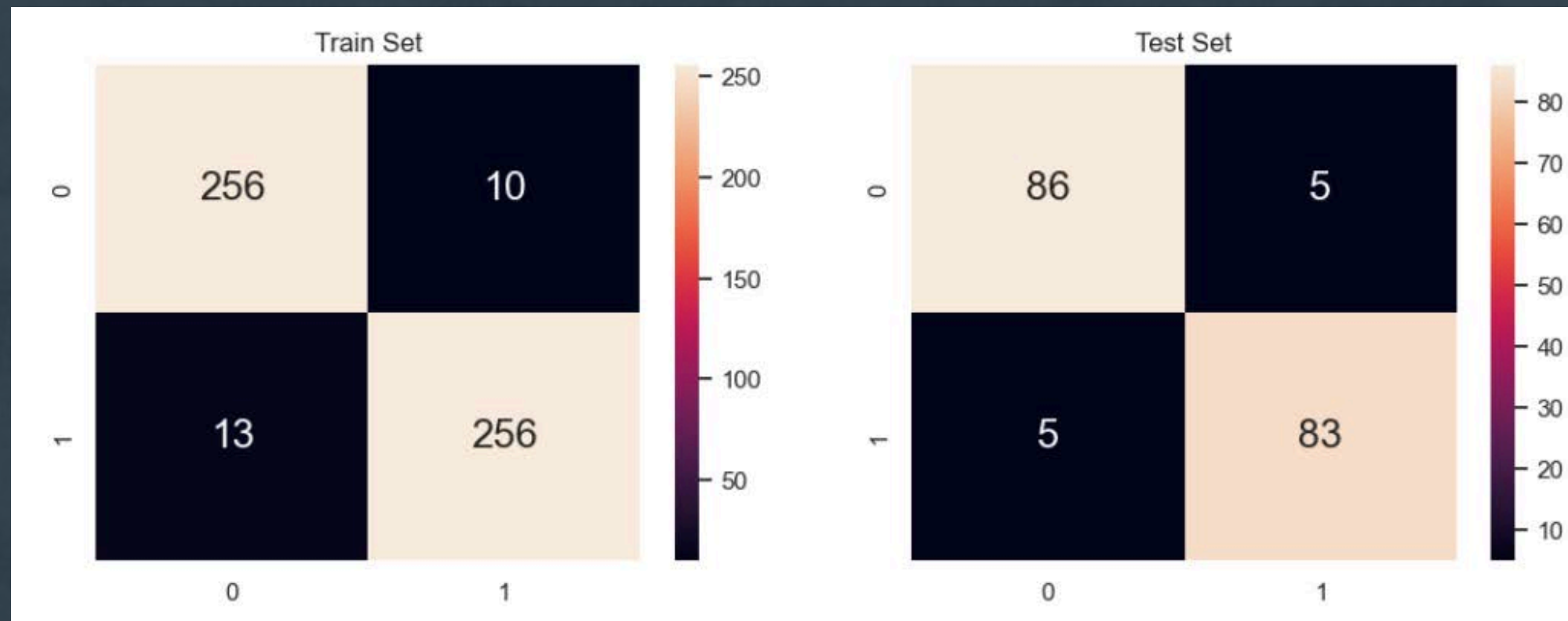
Cross Validation Grid Search



LOGISTIC REGRESSION

Model without Parameter Tuning

➡ Feature List 0 was the best - 7 Best features



Train Set
Accuracy : 0.957
FNR : 0.0483

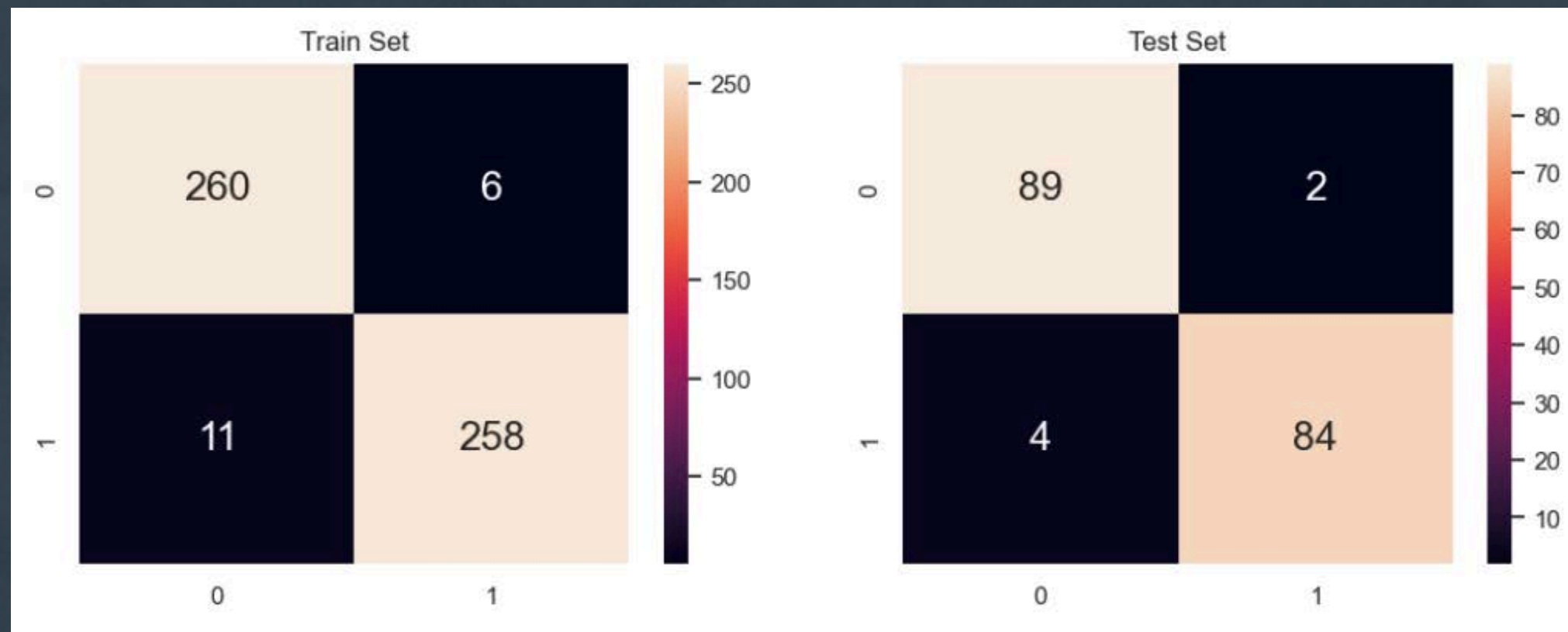
Test Set
Accuracy : 0.944
FNR : 0.0568



LOGISTIC REGRESSION

Model with Parameter Tuning using GridSearchCV

➡ Feature List 0 was the best - 7 Best features



Train Set
Accuracy : 0.968
FNR : 0.0409

Test Set
Accuracy : 0.966
FNR : 0.0455

LOGISTIC REGRESSION

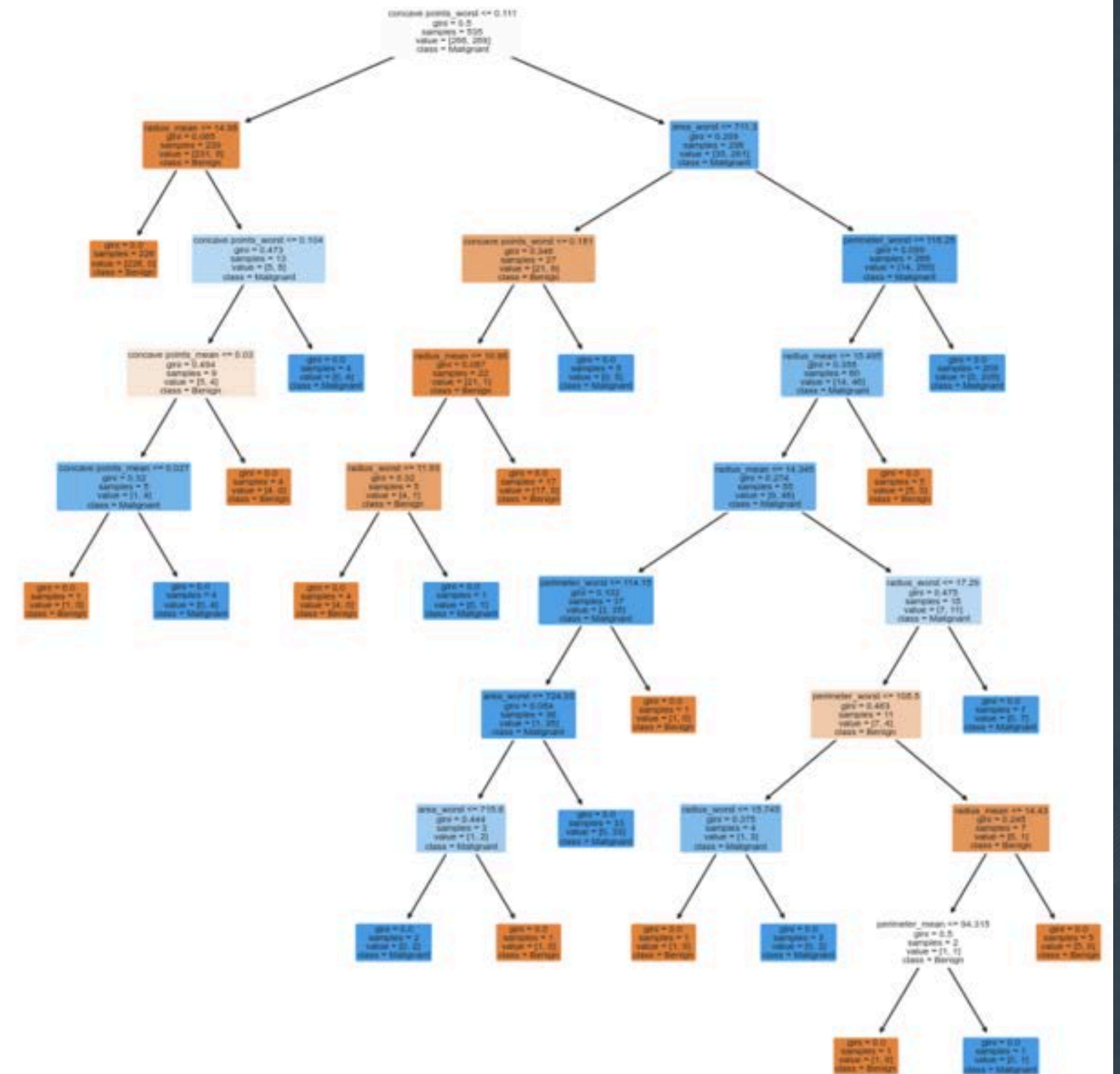
➡ Tuned Logistic Regression (feature List 0) - 7 Best Features

- Accuracy increased from 0.944 to 0.966
- FNR decreased from 0.0568 to 0.0455

	accuracy	f1_score	precision	recall	balanced_accuracy	FNR
Logistic Regression (Feature List 0)	0.944134	0.943182	0.943182	0.943182	0.944118	0.056818
Tuned Logistic Regression (Feature List 0)	0.966480	0.965517	0.976744	0.954545	0.966284	0.045455

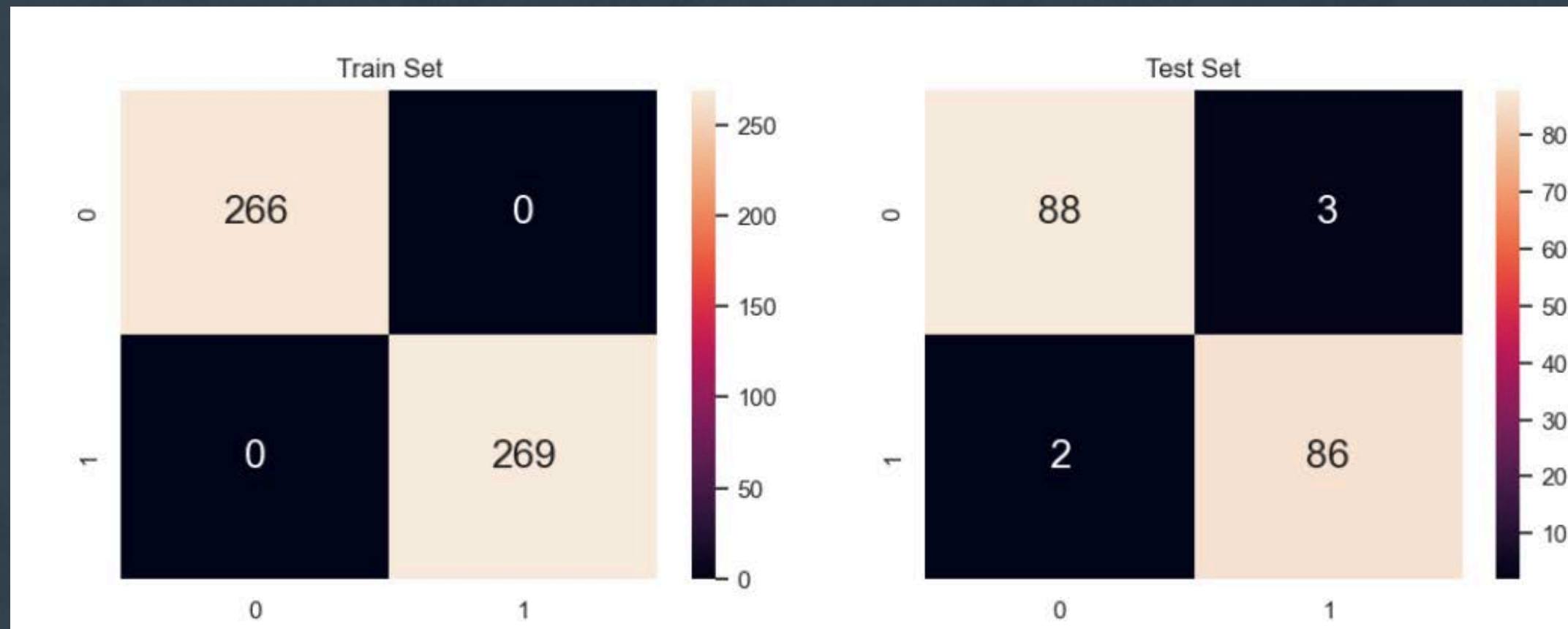
BINARY DECISION TREE

- Evaluate the accuracy of 4 different Trees
- Changed the optimal depth to “None” - expands until all leaf nodes are pure (no errors)



BINARY DECISION TREE

➡ Feature List 0 was the best - 7 Best features



Train Set
Accuracy : 1.0
FNR : 0

Test Set
Accuracy : 0.972
FNR : 0.0227



RANDOM FOREST CLASSIFIER

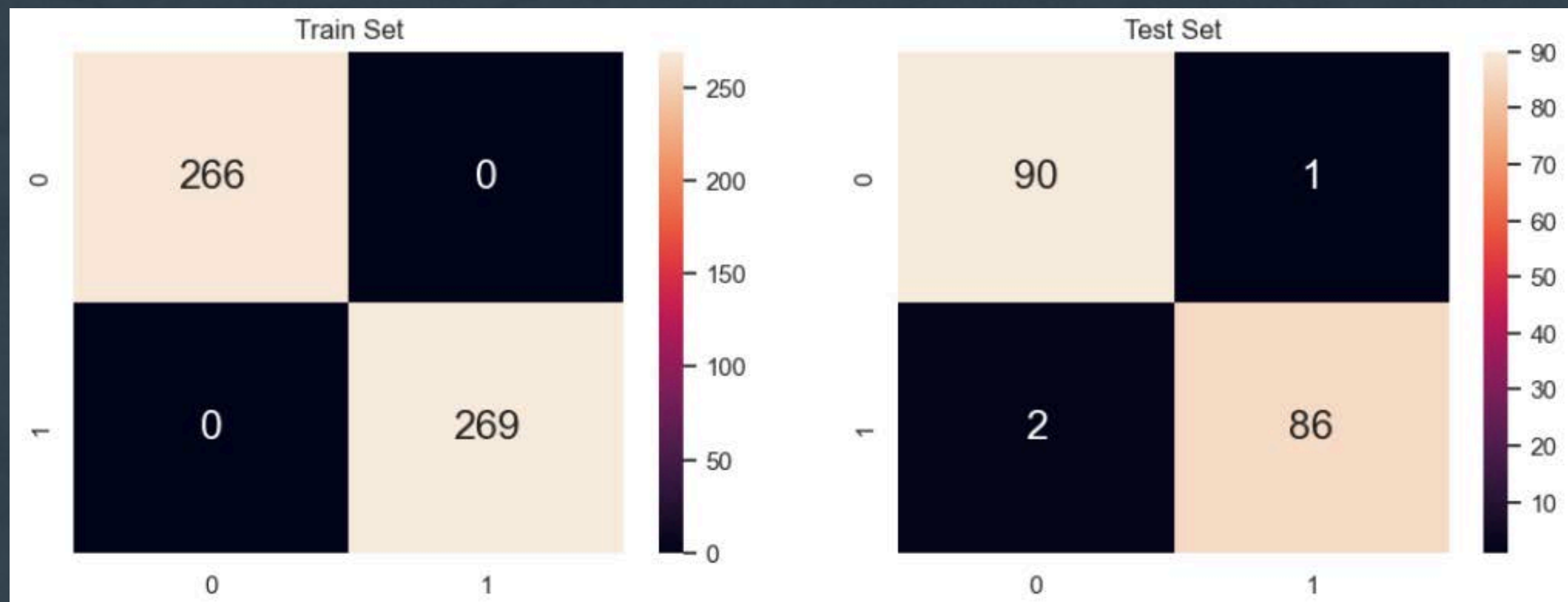
- Trains multiple decision trees on random subsets of train data
- Compares outcomes of all trees and derives a final prediction
- `n_estimator = 200`



RANDOM FOREST CLASSIFIER

- ➡ Feature List 0 - 7 Best Features
- ➡ Feature List 1 - 5 Best Features

+

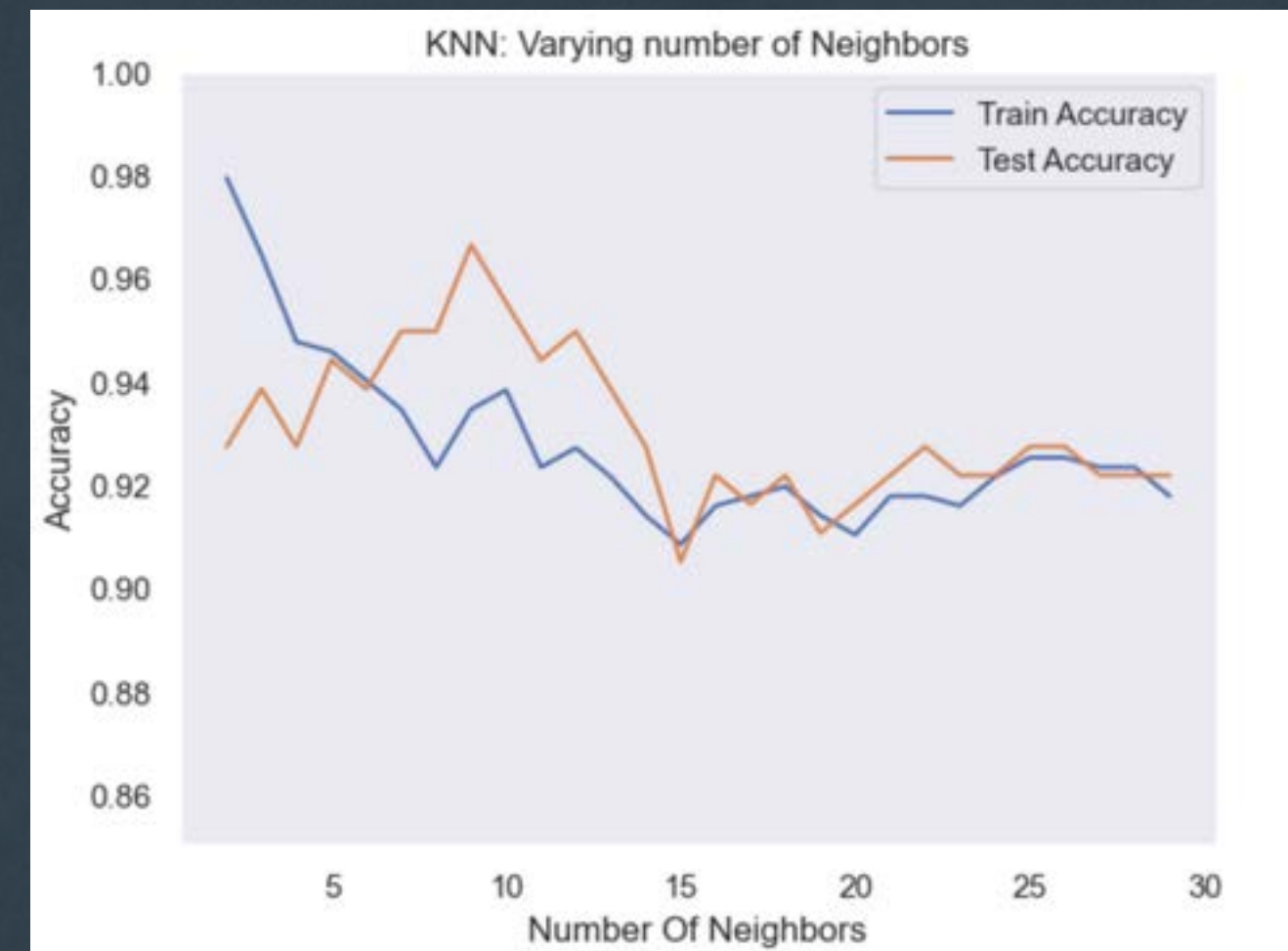


Train Set
Accuracy : 1.0
FNR : 0

Test Set
Accuracy : 0.983
FNR : 0.0227

K-NEAREST NEIGHBOUR CLASSIFICATION

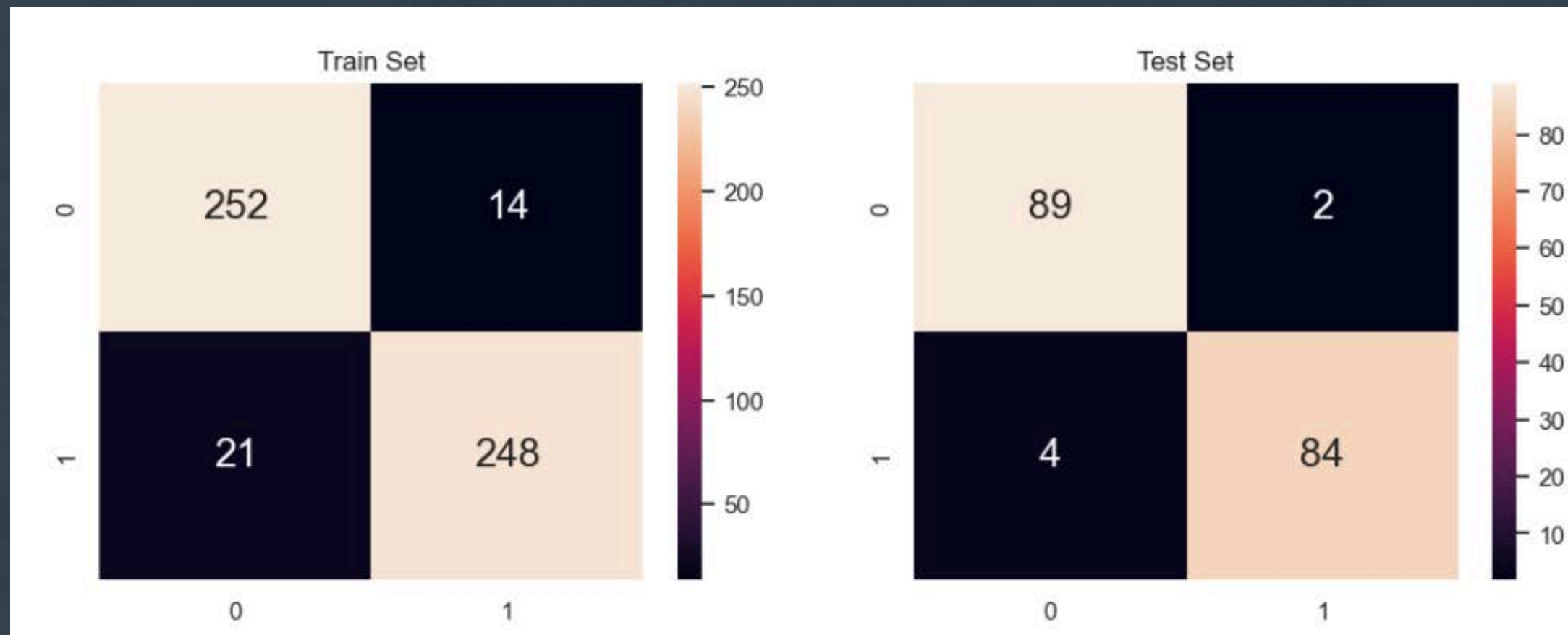
- Uses proximity to make classifications and predictions
- 1 parameter - “n_neighbours”
- Tuned the parameter by iteratively plotting classification accuracy while incrementing number of neighbour.
- Considered the Test Set
- Number of Neighbours = 9



NEAREST NEIGHBOUR CLASSIFICATION



Feature List 0 was the best - 7 Best features



Train Set
Accuracy : 0.935
FNR : 0.0781

Test Set
Accuracy : 0.966
FNR : 0.0455



FINAL DECISION

	accuracy	f1_score	precision	recall	balanced_accuracy	FNR
Random Forest (Feature List 0)	0.983240	0.982857	0.988506	0.977273	0.983142	0.022727
Random Forest (Feature List 1)	0.983240	0.982857	0.988506	0.977273	0.983142	0.022727
Decision Tree (Feature List 0)	0.972067	0.971751	0.966292	0.977273	0.972153	0.022727
Random Forest (Feature List 2)	0.972067	0.971751	0.966292	0.977273	0.972153	0.022727
Tuned Logistic Regression (Feature List 0)	0.966480	0.965517	0.976744	0.954545	0.966284	0.045455
KNearNeighbours (Feature List 0)	0.966480	0.965517	0.976744	0.954545	0.966284	0.045455
Decision Tree (Feature List 1)	0.960894	0.960000	0.965517	0.954545	0.960789	0.045455
Decision Tree (Feature List 2)	0.955307	0.955056	0.944444	0.965909	0.955482	0.034091
KNearNeighbours (Feature List 1)	0.949721	0.948571	0.954023	0.943182	0.949613	0.056818
Logistic Regression (Feature List 0)	0.944134	0.943182	0.943182	0.943182	0.944118	0.056818
KNearNeighbours (Feature List 2)	0.938547	0.937853	0.932584	0.943182	0.938624	0.056818
Random Forest (Feature List 3)	0.916201	0.917127	0.892473	0.943182	0.916646	0.056818
KNearNeighbours (Feature List 3)	0.916201	0.914286	0.919540	0.909091	0.916084	0.090909
Decision Tree (Feature List 3)	0.910615	0.911111	0.891304	0.931818	0.910964	0.068182



Random Forest
(Feature List 0)

Accuracy : 0.983
FNR : 0.0227

OUTCOME , INSIGHTS AND RECOMMENDATIONS



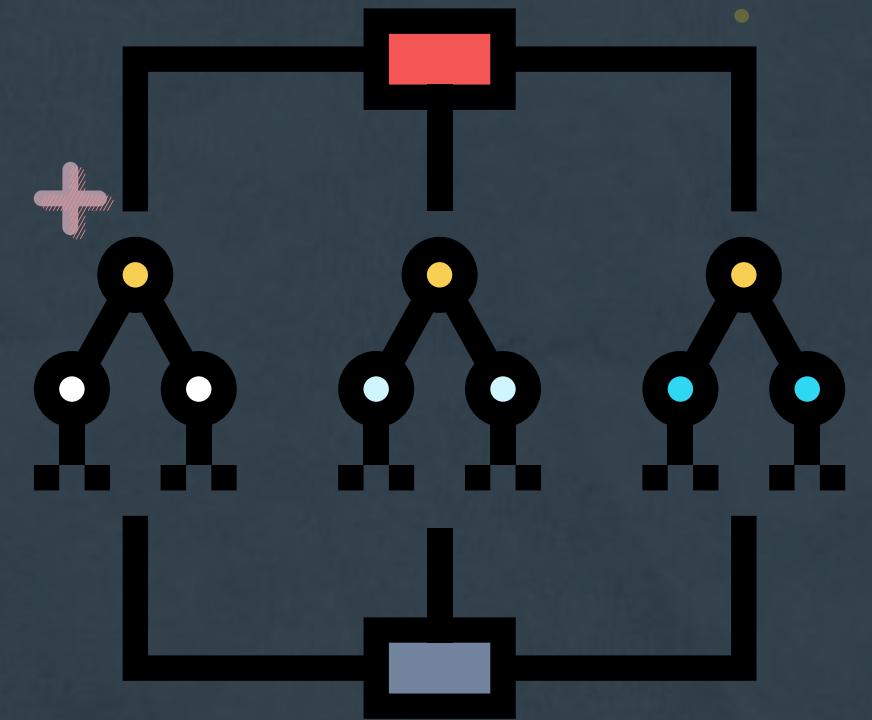


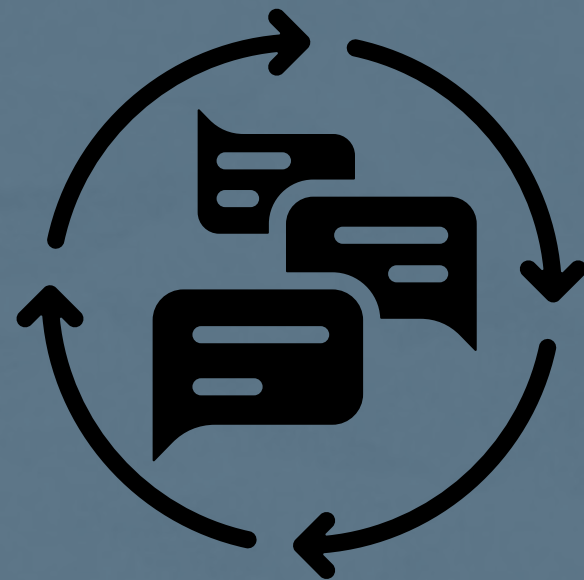
Outcome and Insights



- concave points_worst
- perimeter_worst
- radius_worst
- concave points_mean
- perimeter_mean
- radius_mean
- area_worst

Random Forest





Recommendations

- Validate the performance of the developed models through collaboration with healthcare professionals. Ensure that the model generalize well to diverse patient populations and is robust.
- Regular updates to incorporate new data, emerging research findings. Continuous improvement ensures that the models remain relevant and effective over time.
- Many models are available for a classification problem, can possibly look into other models that can also achieve a high if not better accuracy score



WE WANT TO SAY

THANK YOU

FOR YOUR ATTENTION

