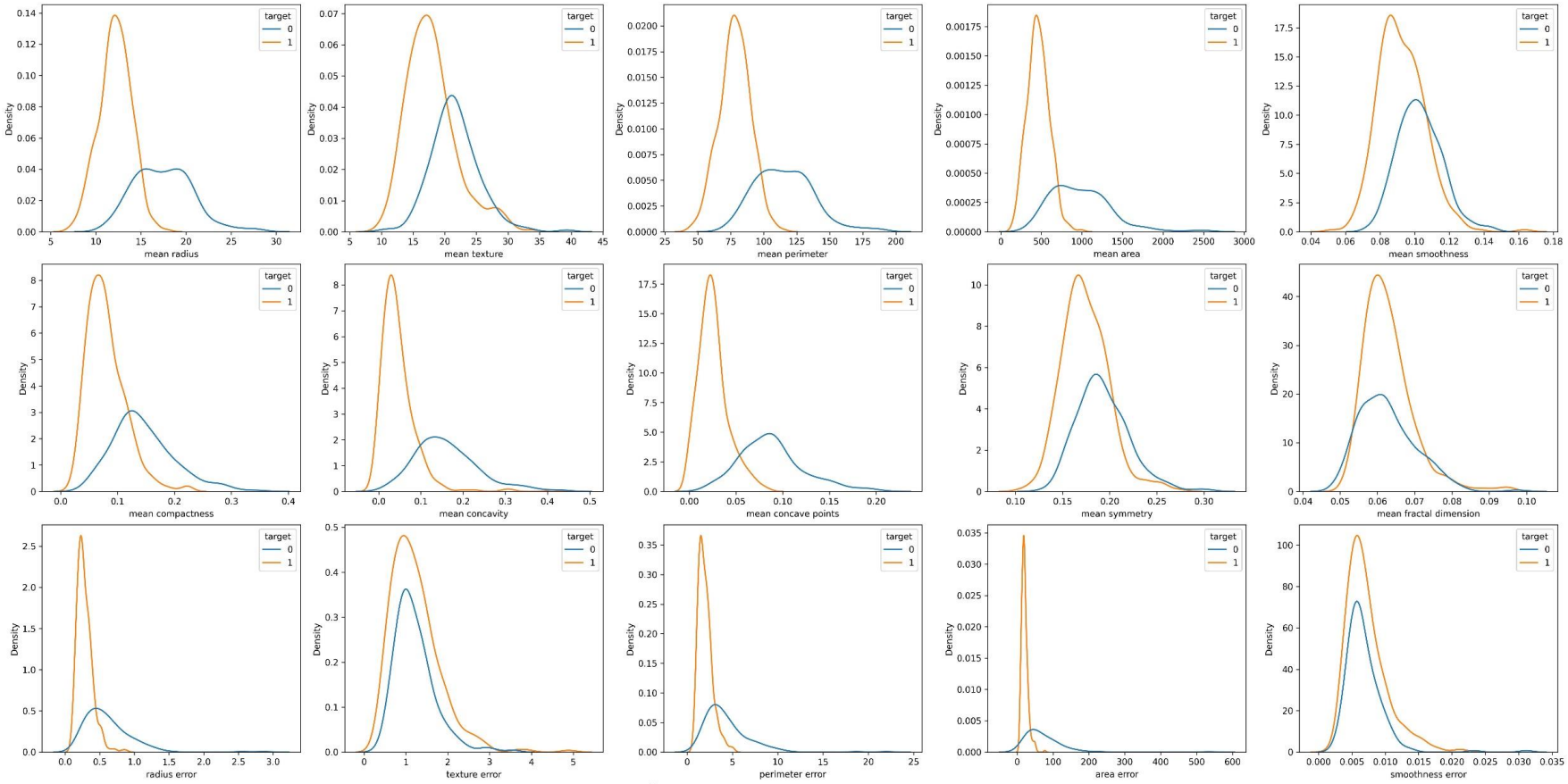
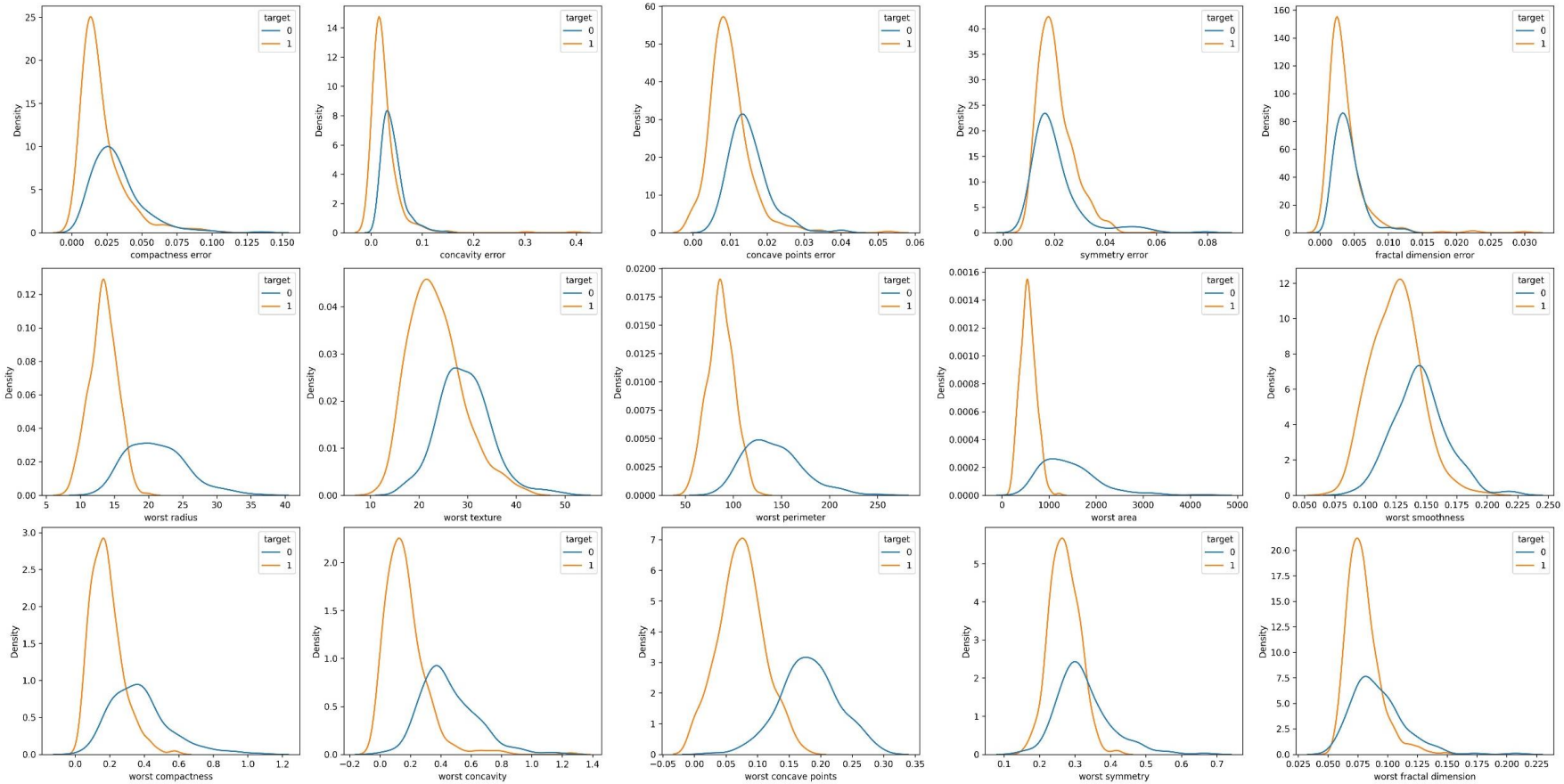
Visual representation of the features regarding the outcome

KDE plots 1



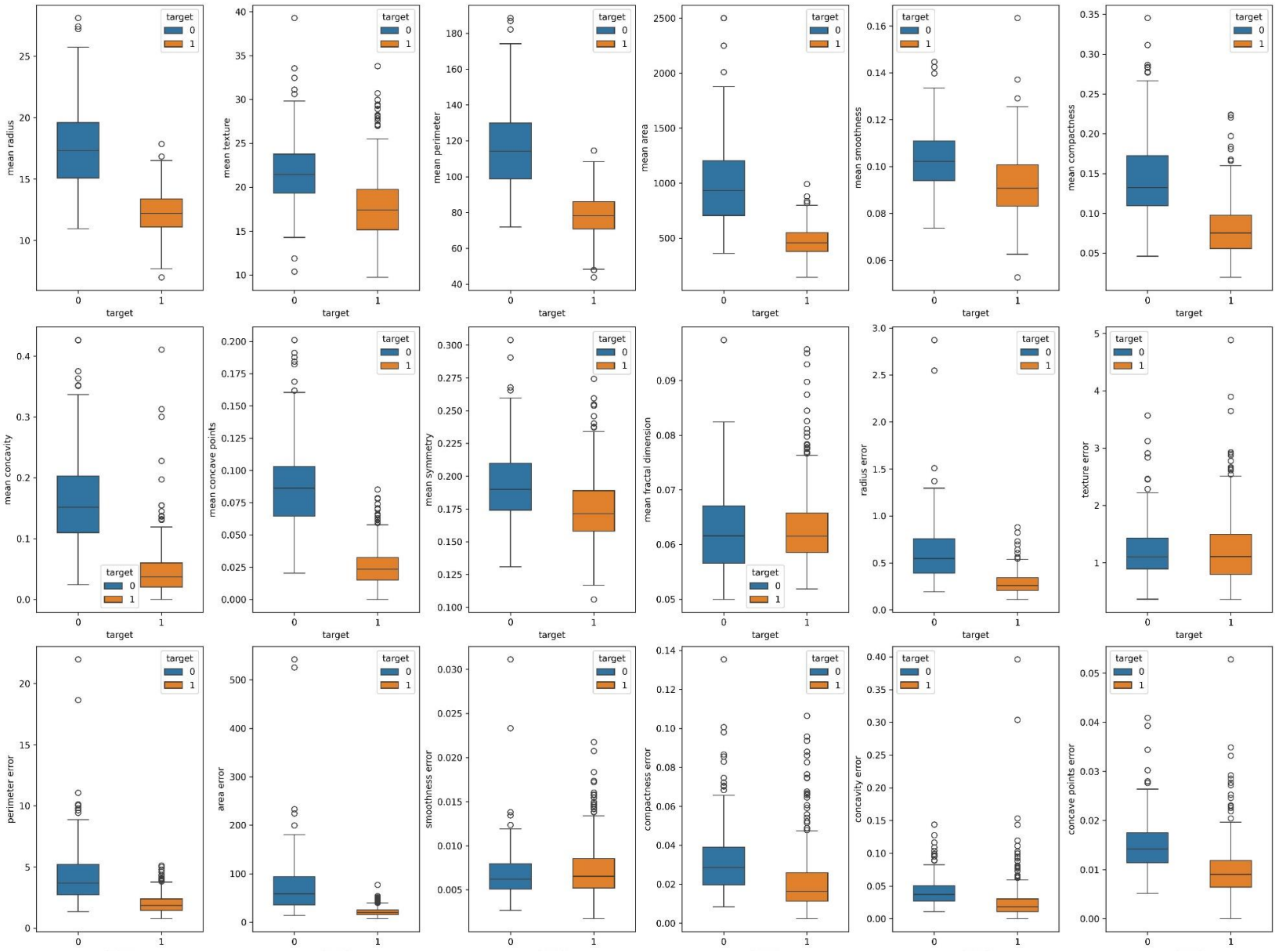
Visual representation of the features regarding the outcome

KDE plots 2



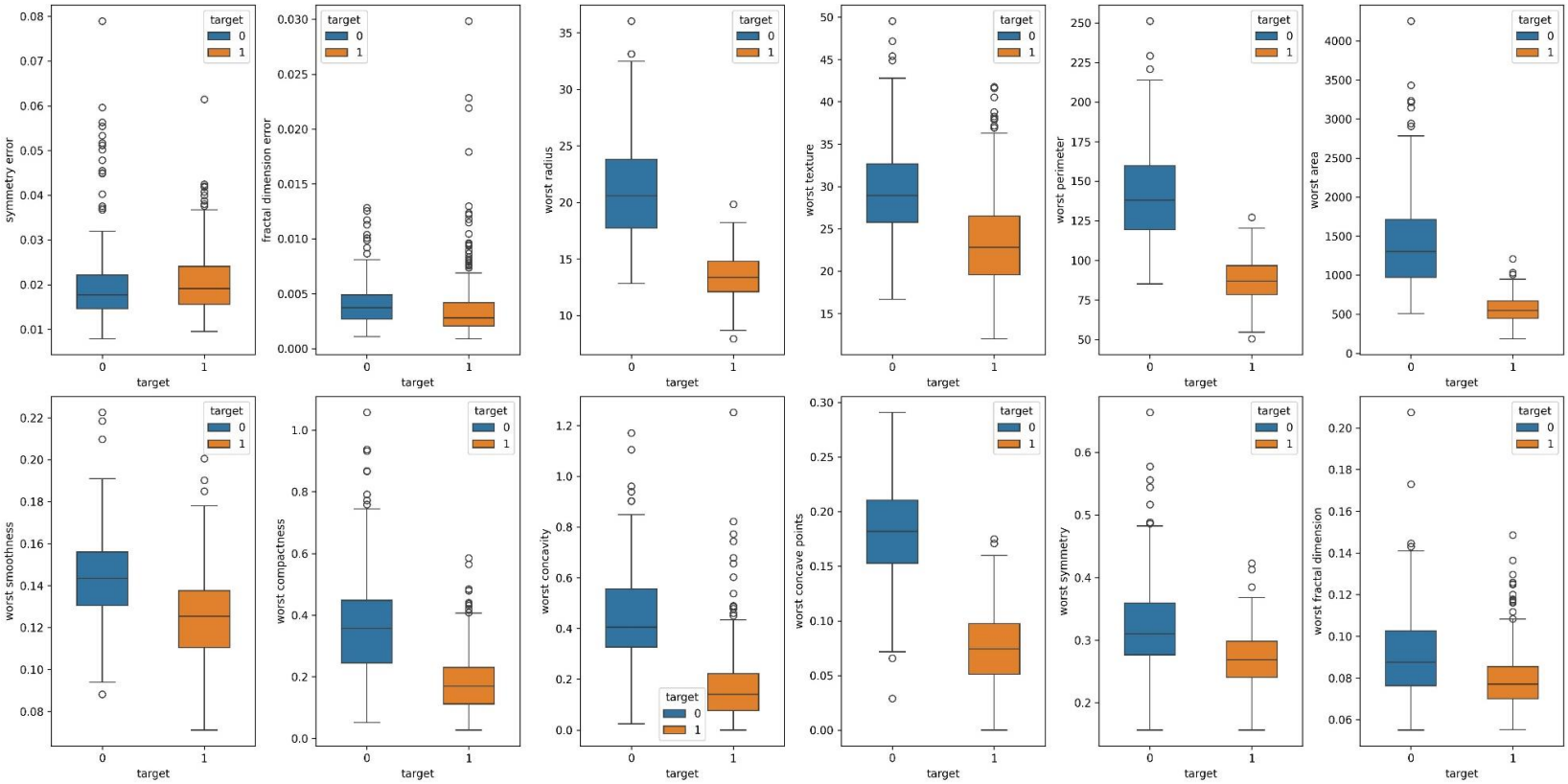
Visual representation of the features regarding the outcome

Box plots 1



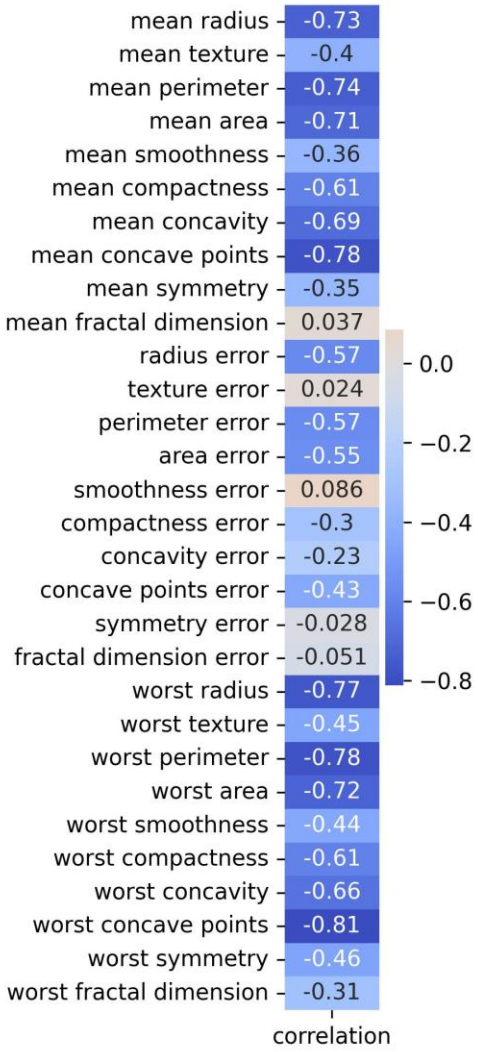
Visual representation of the features regarding the outcome

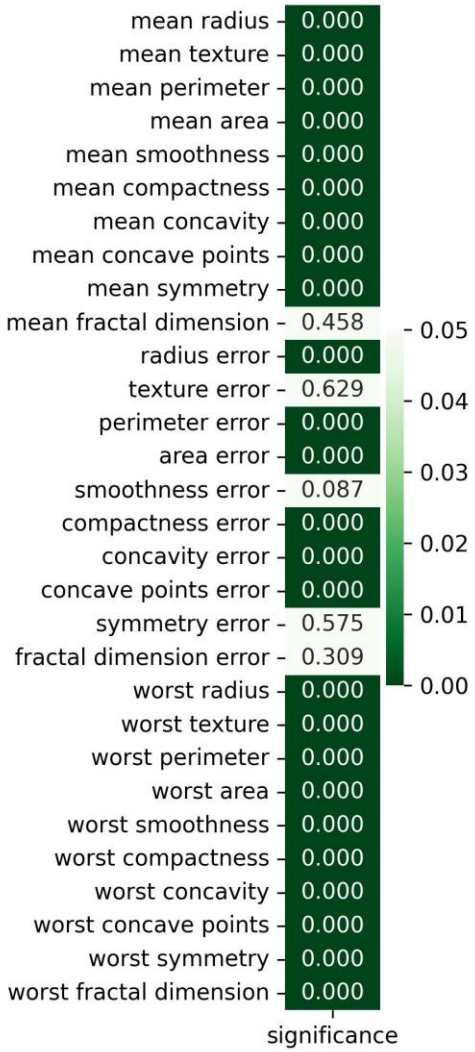
Box plots 2



Feature selection - Univariate methods

Correlation analysis: Point-Biserial correlation between each feature and the target



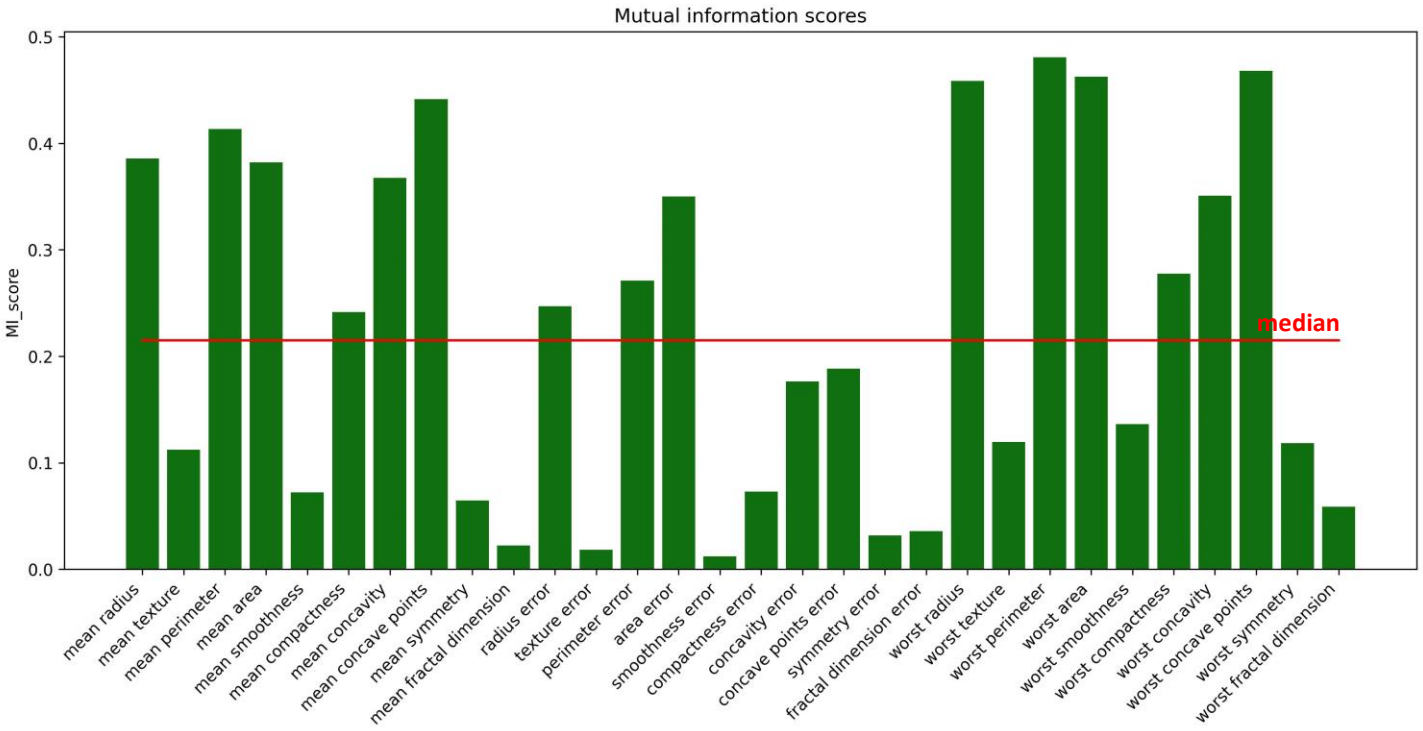


Selected features
(Corr. coefficint > 0.65)

feature	correlation	significance
worst concave points	-0.812227	9.958450e-95
mean concave points	-0.783217	9.542099e-84
worst perimeter	-0.777190	1.121979e-81
worst radius	-0.770776	1.523739e-79
mean perimeter	-0.744040	2.379762e-71
mean radius	-0.731126	9.496150e-68
worst area	-0.719659	1.009960e-64
mean area	-0.708127	7.924300e-62
mean concavity	-0.689095	2.394478e-57
worst concavity	-0.658362	8.423243e-51

Feature selection - Univariate methods

Mutual information score:

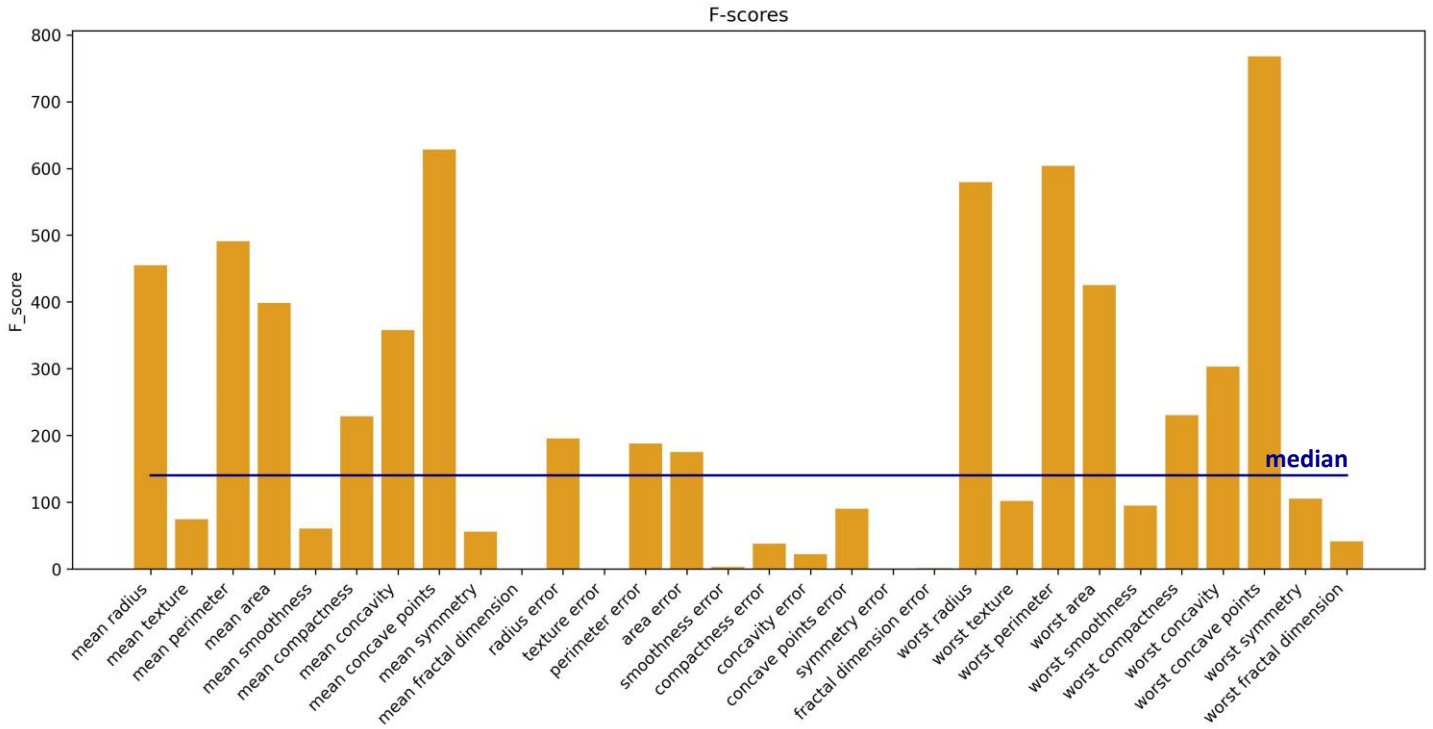


Selected features
(MI score > median)

	feature	MI_score
22	worst perimeter	0.480664
27	worst concave points	0.468007
23	worst area	0.462424
20	worst radius	0.458436
7	mean concave points	0.441503
2	mean perimeter	0.413537
0	mean radius	0.385695
3	mean area	0.382068
6	mean concavity	0.367583
26	worst concavity	0.350864
13	area error	0.350032
25	worst compactness	0.277677
12	perimeter error	0.270892
10	radius error	0.246931
5	mean compactness	0.241353

Feature selection - Univariate methods

F-score:

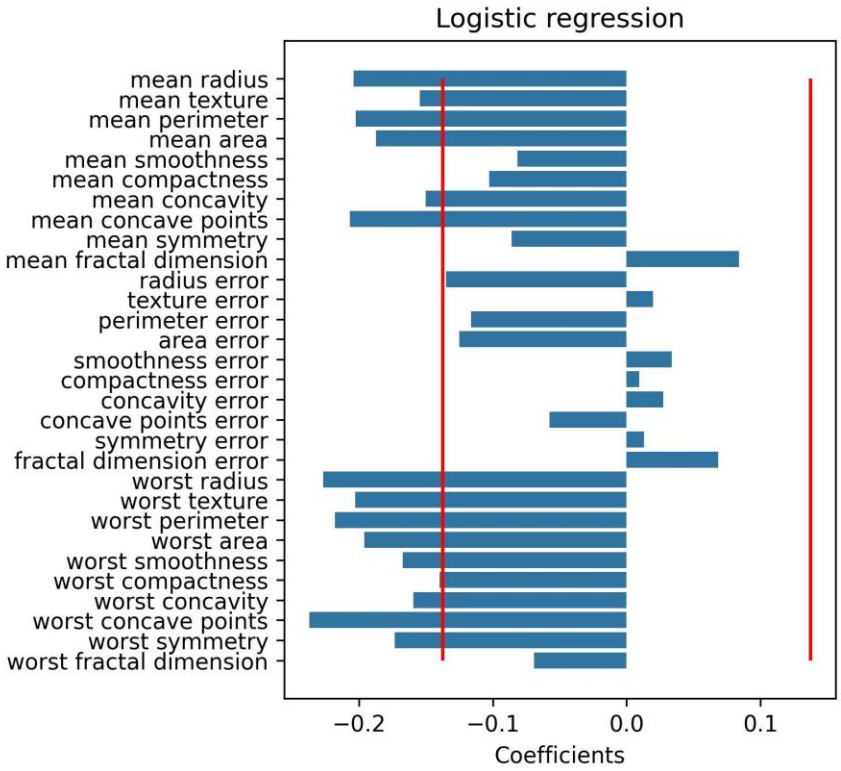


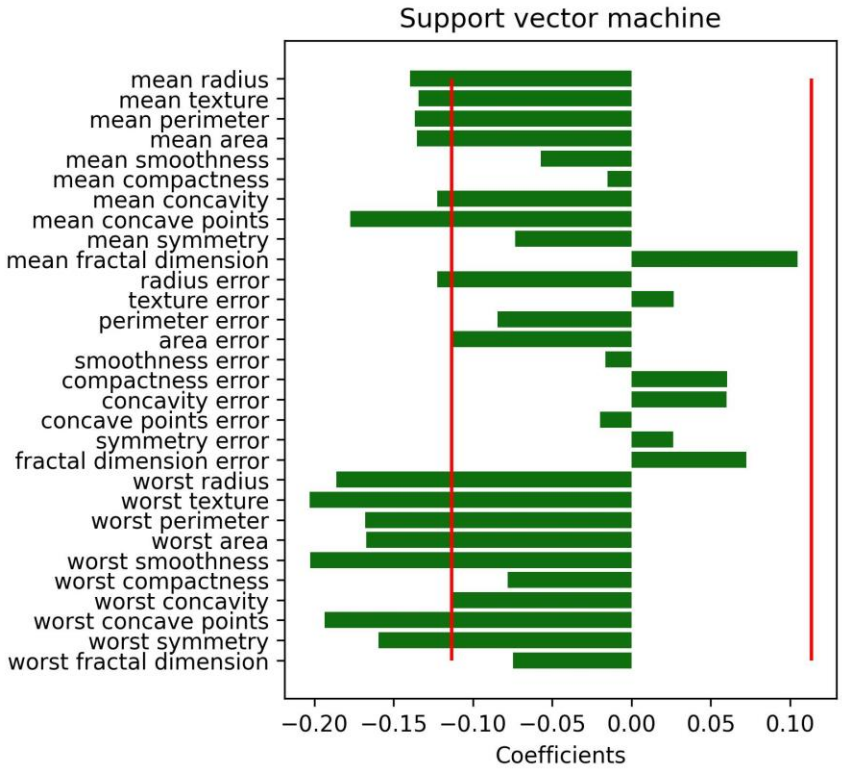
Selected features
(F-score > median)

	feature	F_score
27	worst concave points	767.720439
7	mean concave points	628.391414
22	worst perimeter	604.060947
20	worst radius	579.600781
2	mean perimeter	491.088539
0	mean radius	454.781585
23	worst area	425.420961
3	mean area	398.294193
6	mean concavity	358.072346
26	worst concavity	302.956310
25	worst compactness	230.317988
5	mean compactness	228.753334
10	radius error	195.341413
12	perimeter error	188.249626
13	area error	175.264736

Feature selection - Multivariate methods

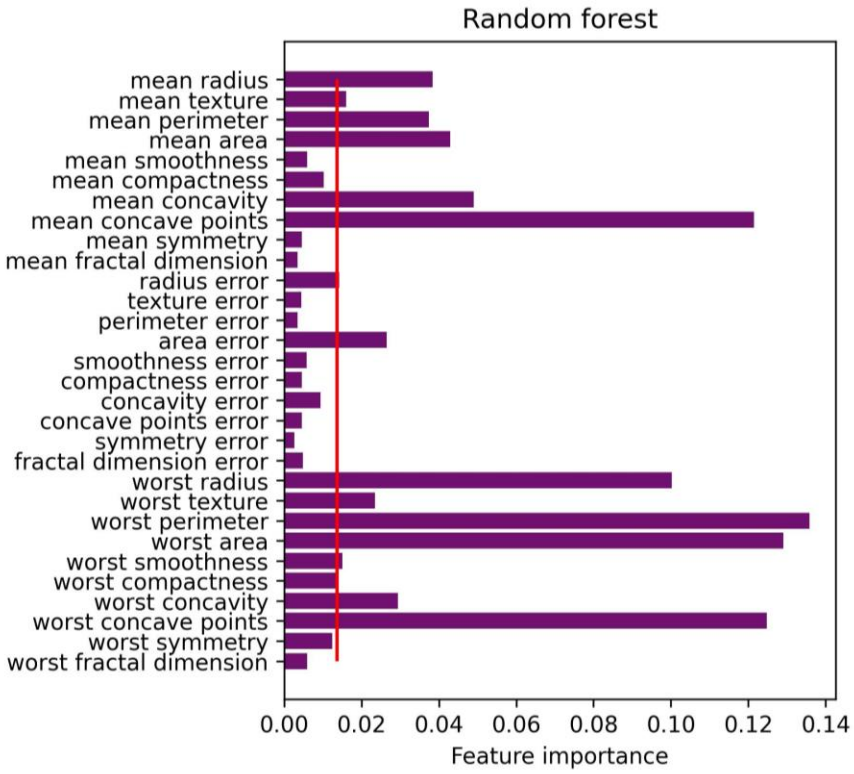
The selection is based on feature importance with **Select From Model** (SFM):
Tested ML models: logistic regression, SVM, random forest and gradient boosting
Selected features: coefficient > median (red lines)

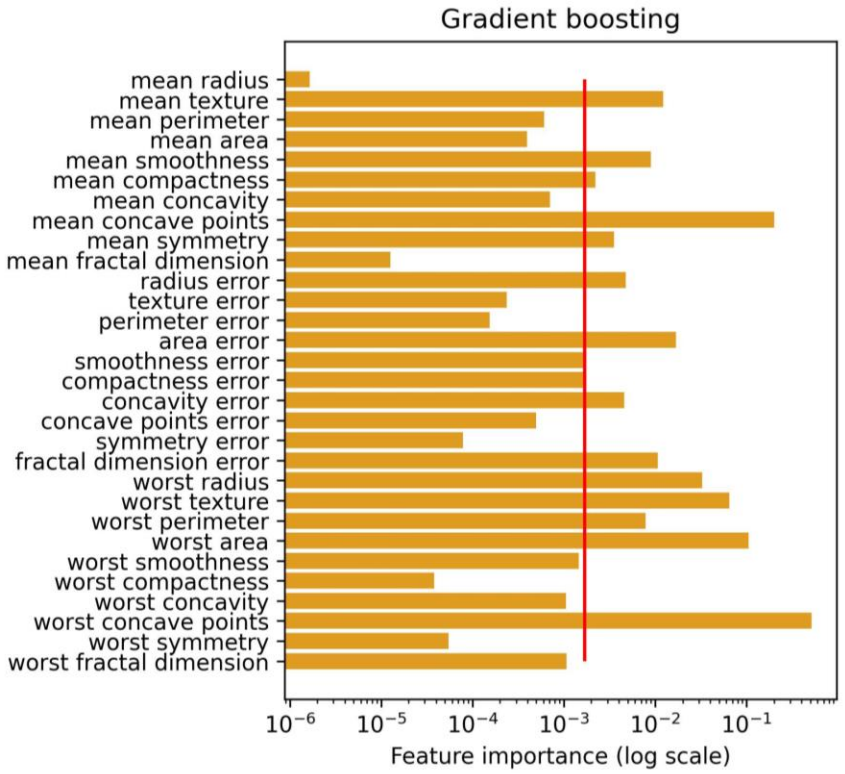




Feature selection - Multivariate methods

The selection is based on feature importance with **Select From Model** (SFM):
Tested ML models: logistic regression, SVM, random forest and gradient boosting
Selected features: feature importance > median (red lines)





Feature selection - Multivariate methods

Ranking of features with **Recursive Feature Elimination** (RFE):
Tested ML models: logistic regression, SVM, random forest and gradient boosting

Selected features
(Top 15)

rank	order_LR	order_SVM	order_RF	order_GB
1	worst concave points	worst concave points	worst perimeter	worst concave points
2	worst radius	worst radius	mean concave points	worst area
3	mean concave points	mean concave points	worst concave points	mean concave points
4	worst perimeter	worst perimeter	worst area	worst texture
5	mean perimeter	worst texture	worst radius	worst radius
6	worst texture	worst area	mean area	area error
7	worst concavity	worst smoothness	worst texture	fractal dimension error
8	worst area	mean perimeter	mean perimeter	mean texture
9	mean radius	worst symmetry	mean concavity	worst perimeter
10	worst symmetry	mean radius	mean radius	concavity error
11	mean area	radius error	worst concavity	mean compactness
12	worst smoothness	mean area	mean texture	worst smoothness
13	mean concavity	mean concavity	area error	mean perimeter
14	radius error	area error	worst smoothness	smoothness error
15	worst compactness	mean texture	worst symmetry	worst concavity

Comparison of uni- and multivariate feature selection by ranking

(Top 15 features)

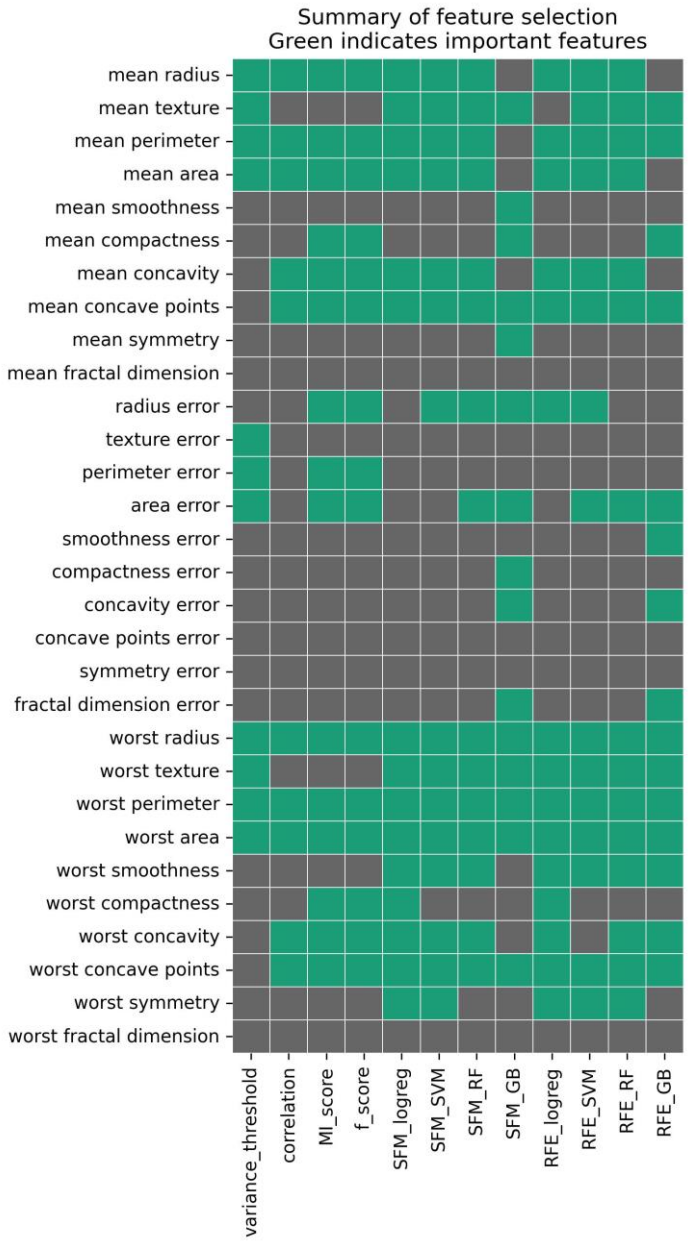
Univariate

rank	order_correlation	order_MI	order_Fscore
1	worst concave points	worst perimeter	worst concave points
2	mean concave points	worst concave points	mean concave points
3	worst perimeter	worst area	worst perimeter
4	worst radius	worst radius	worst radius
5	mean perimeter	mean concave points	mean perimeter
6	mean radius	mean perimeter	mean radius
7	worst area	mean radius	worst area
8	mean area	mean area	mean area
9	mean concavity	mean concavity	mean concavity
10	worst concavity	worst concavity	worst concavity
11	worst compactness	area error	worst compactness
12	mean compactness	worst compactness	mean compactness
13	radius error	perimeter error	radius error
14	perimeter error	radius error	perimeter error
15	area error	mean compactness	area error

Multivariate (RFE)

rank	order_LR	order_SVM	order_RF	order_GB
1	worst concave points	worst concave points	worst perimeter	worst concave points
2	worst radius	worst radius	mean concave points	worst area
3	mean concave points	mean concave points	worst concave points	mean concave points
4	worst perimeter	worst perimeter	worst area	worst texture
5	mean perimeter	worst texture	worst radius	worst radius
6	worst texture	worst area	mean area	area error
7	worst concavity	worst smoothness	worst texture	fractal dimension error
8	worst area	mean perimeter	mean perimeter	mean texture
9	mean radius	worst symmetry	mean concavity	worst perimeter
10	worst symmetry	mean radius	mean radius	concavity error
11	mean area	radius error	worst concavity	mean compactness
12	worst smoothness	mean area	mean texture	worst smoothness
13	mean concavity	mean concavity	area error	mean perimeter
14	radius error	area error	worst smoothness	smoothness error
15	worst compactness	mean texture	worst symmetry	worst concavity

Summary



Number of times a feature was selected

	n_selected
worst radius	12
worst area	12
worst perimeter	12
mean perimeter	11
worst concave points	11
mean concave points	11
mean radius	10
mean area	10
worst concavity	9
mean concavity	9
worst texture	9
area error	8
mean texture	8
worst smoothness	7
radius error	7
worst symmetry	5
worst compactness	4
mean compactness	4
perimeter error	3
fractal dimension error	2
concavity error	2
compactness error	1
texture error	1
mean symmetry	1
mean smoothness	1
smoothness error	1

Final selection of the features

Top 10:

- *worst concave points*
- *worst radius*
- *mean concave points*
- *worst perimeter*
- *mean perimeter*
- *worst texture*
- *worst area*
- *mean radius*
- *worst symmetry*
- *mean area / worst concavity*

Only from the *worst* category:

- *worst concave points*
- *worst radius*
- *worst perimeter*

+

- 4th: *worst texture*
- 5th: *worst area*
- 6th: *worst concavity*

Reasons behind the final selection of features

- *Error* metrics does not seems to be useful, both *mean* and *worst* metrics are better, therefore it seems logical to exclude all of them.

- Since logistic regression performed the best during the initial testing, I tried to optimize for that and mainly focused on the results of that model from SFM and RFE

- 10 best predictors: the selection was based on both univariate and multivariate (for logistic regression) methods, with more emphasis on SFM and RFE results. 9 features were consistently selected, however the 10th was questionable and could be either *mean area* or *worst concavity*.

- To use only one type of metrics for the prediction *worst* would be the best option. It could reduce computational cost even before the ML pipeline and could save a lot of storage space as well.

    - With this option the 3 best candidates are *worst concave points*, *worst radius* and *worst perimeter*. These 3 features are also in the top 5 with every selection method.

    - Besides these, the following features could be added to this list based on embedded and univariate methods as well (in the indicated order):
        - 4th *worst texture*
        - 5th *worst area*
        - 6th *worst concavity*