Predicting Breast Cancer Diagnosis Using Machine Learning

Valeria Solozobova

Project report(2023-12-01)

Github URL: https://github.com/ValeriaMalin/Breast_cancer_wisconsin

Objective and dataset:

•**Objective**:
•The aim of this project is to build a robust machine learning model to classify breast cancer diagnosis (Malignant or Benign) based on the Breast Cancer Wisconsin Dataset.

•**Dataset Overview**:
•Source: UCI Machine Learning Repository and https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data .
•Dataset contains 569 samples with 30 features representing various diagnostic measurements 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.
Ten real-valued features are computed for each cell nucleus:

a)radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features.

Methodology:

Exploratory Data Analysis (EDA):

-Overview of Data:

-Number of samples, features, and target class distribution (e.g., benign vs malignant).

-Statistical summary (mean, standard deviation, etc., skewness, kurtosis).

-Key Findings:

-Correlation heatmap to identify multicollinearity.

-Distribution plots for key features.

-Observations regarding class imbalance or any outliers.

-Density plots

Feature Selection and Preprocessing:

-Feature Selection Methods:

-Manual selection according the correlation coefficients, features importances and PCA

-SelectKBest

-Data Preprocessing:

-Handling missing values and duplokates(if any).

-Normalization or standardization of features, if needed.

-Splitting data into training and test sets

Methodology:

Modelling

-Random Forest

-SVR

-**XGBoost**

For all the models the various features sets were analysed, the hyper-parameters were fine-tuned with GridSearchCV, the crucial metric for evaluation was considered as well

Conclusion and Outlook:

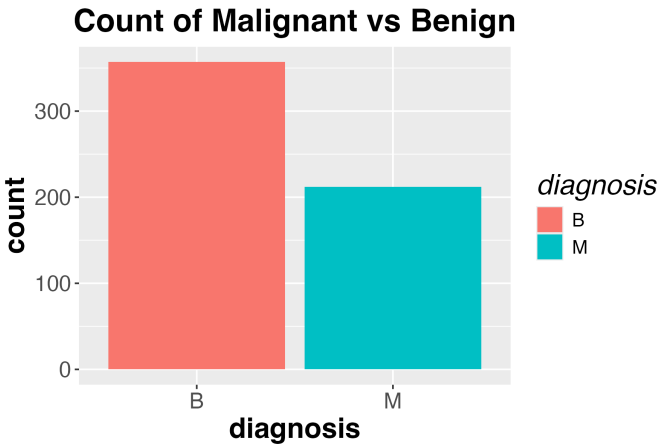
-Comparison between models:

-Metric comparison

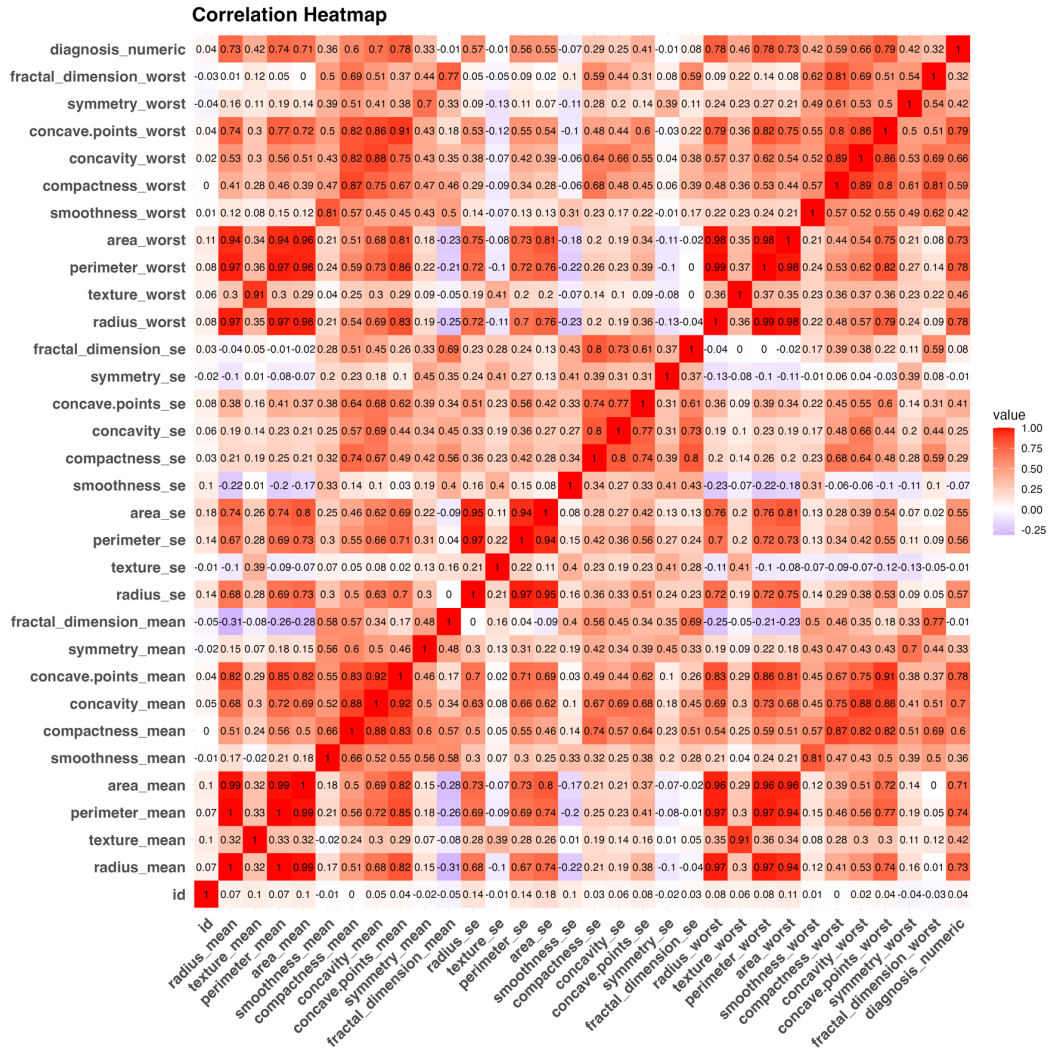
-Error analysis

-Future enhancements

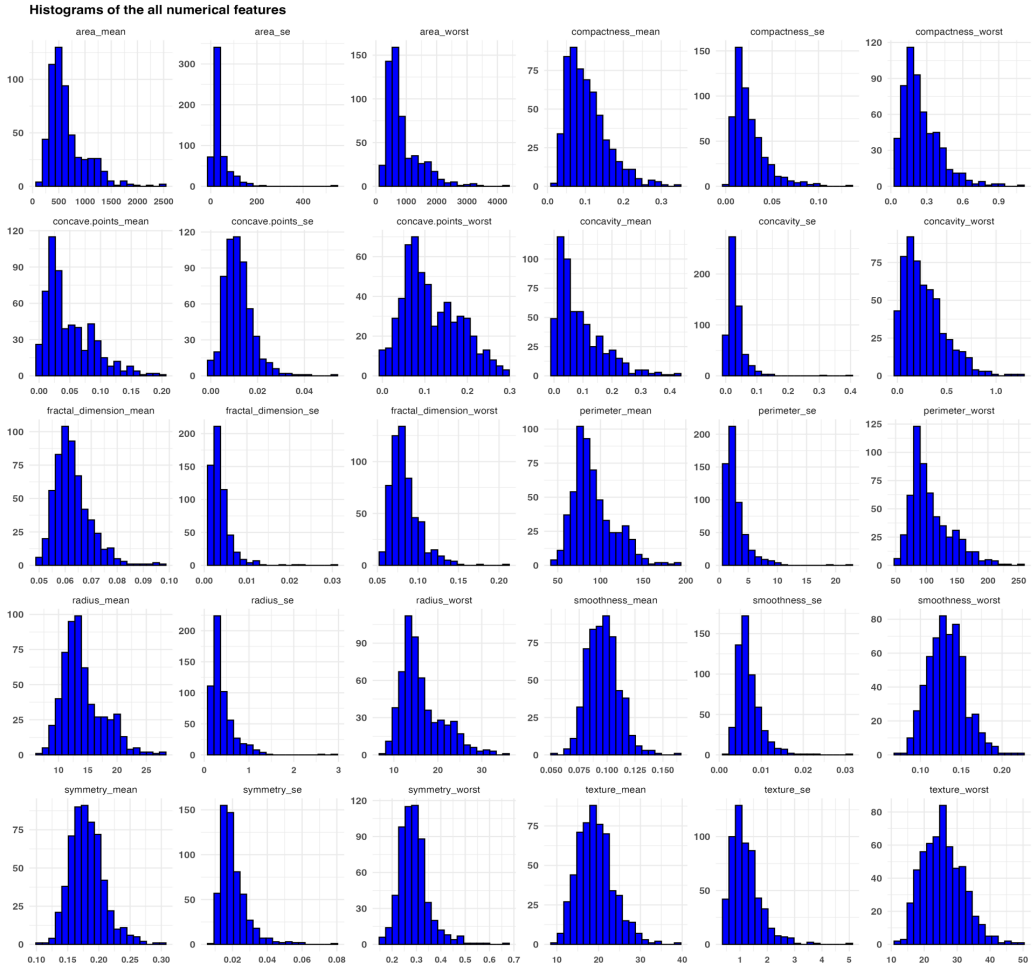
Exploratory Data Analysis:



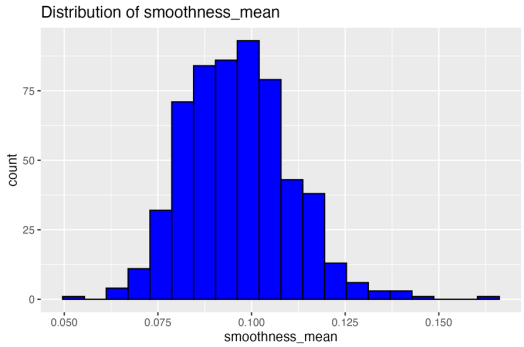
Class distribution: 357 benign, 212 malignant
30 features
ID
Class: M or B

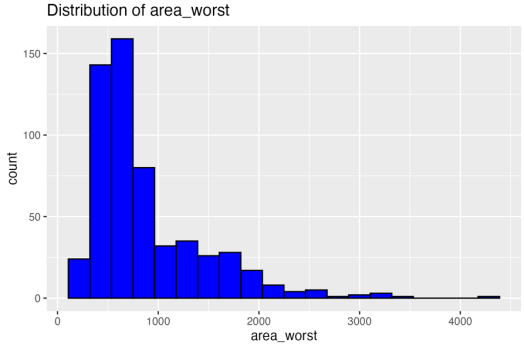


Exploratory Data Analysis:



Some conclusion and examples histoplots!!1!



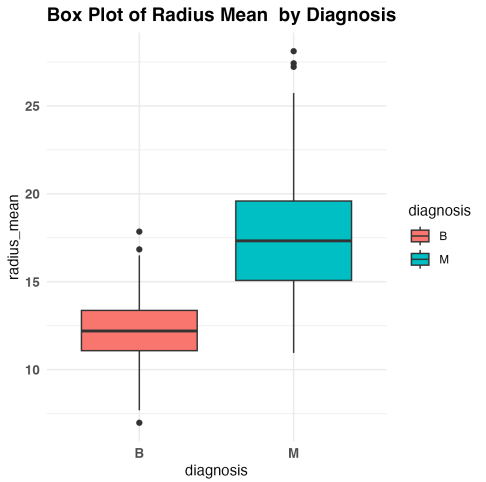


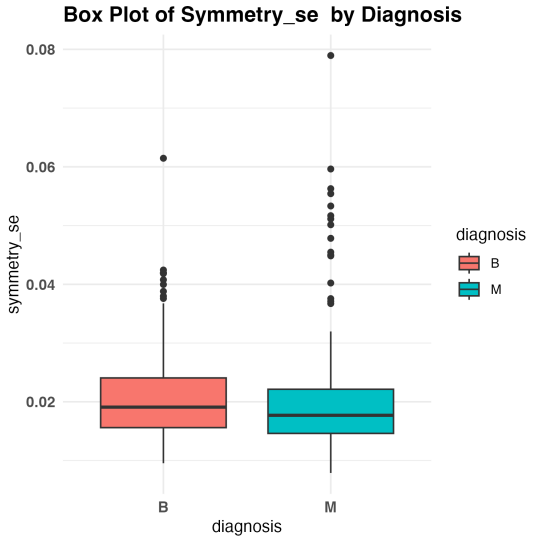
Exploratory Data Analysis: skewness and kurtosis

	Feature	Skewness	Kurtosis
id	id	6.45667315953913	44.812812751863
radius_mean	radius_mean	0.939893445557632	3.82758367391404
texture_mean	texture_mean	0.648733567286701	3.74114542022429
perimeter_mean	perimeter_mean	0.988036954167723	3.9531650486735
area_mean	area_mean	1.64139050920442	6.60976126482313
smoothness_mean	smoothness_mean	0.455119920357176	3.83794535026572
compactness_mean	compactness_mean	1.18698332397452	4.62513951631127
concavity_mean	concavity_mean	1.39748323528266	4.97059165157889
concave points_mean	concave points_mean	1.16809034810126	4.04668022395321
symmetry_mean	symmetry_mean	0.723694717715498	4.26611697471047
fractal_dimension_mean	fractal_dimension_mean	1.30104739278909	5.96901689828588
radius_se	radius_se	3.0804639853352	20.5211621896265
texture_se	texture_se	1.64210026494796	8.29175288832731
perimeter_se	perimeter_se	3.43453047461471	24.2037748104257
area_se	area_se	5.43281586295193	51.7671956105044
smoothness_se	smoothness_se	2.30834422104598	13.3675371954662
compactness_se	compactness_se	1.89720239140839	8.05096602262541
concavity_se	concavity_se	5.09698094901704	51.4225620924194
concave points_se	concave points_se	1.4408668862917	8.070839733647
symmetry_se	symmetry_se	2.18934183892583	10.8163879926805
fractal_dimension_se	fractal_dimension_se	3.91361665467527	29.0399497684146
radius_worst	radius_worst	1.10020503727222	3.92528760461351
texture_worst	texture_worst	0.497006669761678	3.21180937803143
perimeter_worst	perimeter_worst	1.12518762087713	4.05024268384629
area_worst	area_worst	1.85446799160962	7.34733080463865
smoothness_worst	smoothness_worst	0.41433004572343	3.50275974713195
compactness_worst	compactness_worst	1.46966746109576	6.00212020851537
concavity_worst	concavity_worst	1.1472023399942	4.5905680713088
concave points_worst	concave points_worst	0.491315939794396	2.45863292998209
symmetry_worst	symmetry_worst	1.43014486775407	7.39507329375575
fractal_dimension_worst	fractal_dimension_worst	1.65819315504776	8.18811128241064

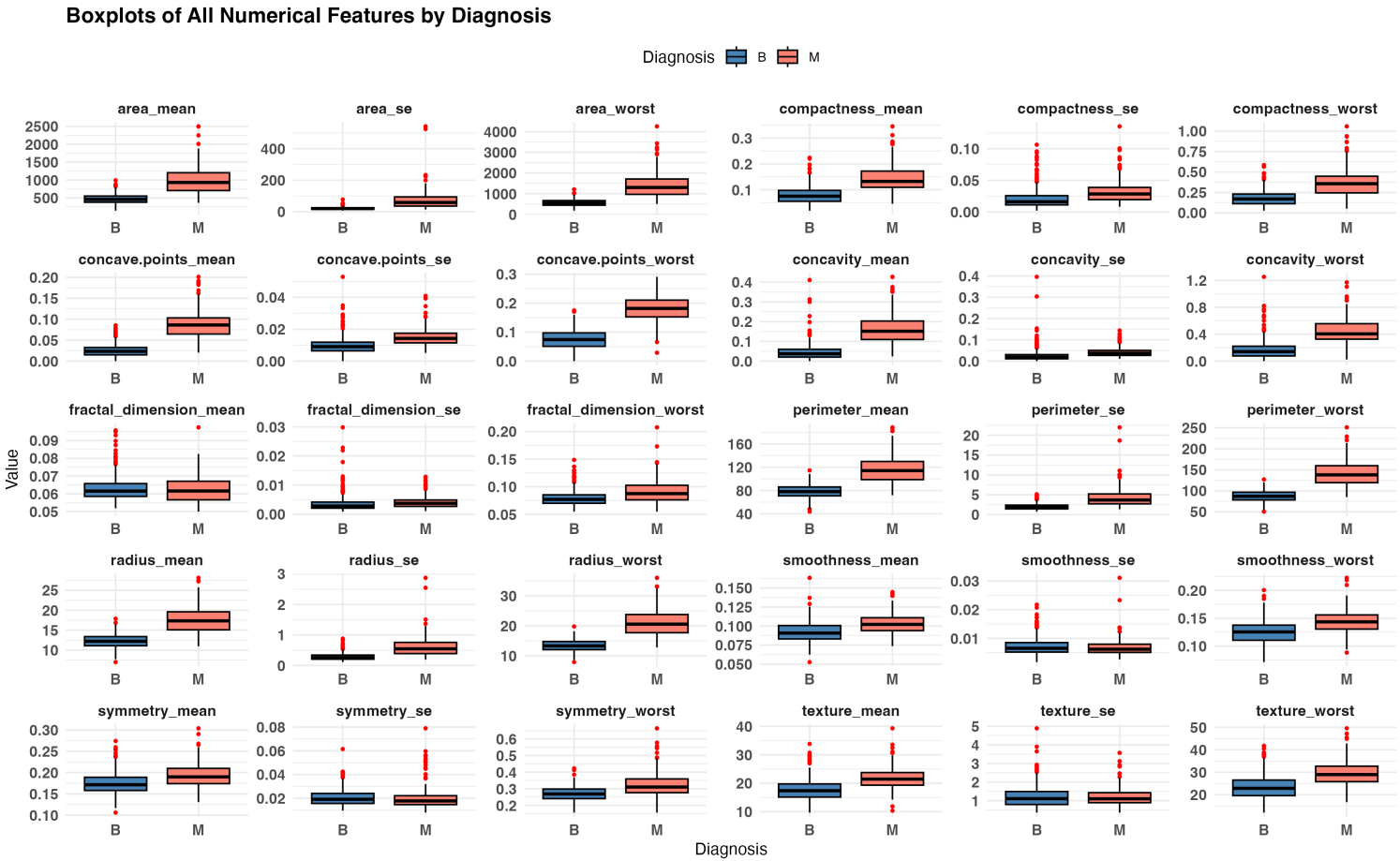
Exploratory Data Analysis:

Variable	Min	Q1	Median	Mean	Q3	Max	NA_Count
id	8670	869218	906024	30371831.4323374	8813129	911320502	0
radius_mean	6.981	11.7	13.37	14.1272917398946	15.78	28.11	0
texture_mean	9.71	16.17	18.84	19.2896485061511	21.8	39.28	0
perimeter_mean	43.79	75.17	86.24	91.9690333919156	104.1	188.5	0
area_mean	143.5	420.3	551.1	654.889103690685	782.7	2501	0
smoothness_mean	0.05263	0.08637	0.09587	0.0963602811950791	0.1053	0.1634	0
compactness_mean	0.01938	0.06492	0.09263	0.104340984182777	0.1304	0.3454	0
concavity_mean	0	0.02956	0.06154	0.0887993158172232	0.1307	0.4268	0
concave points_mean	0	0.02031	0.0335	0.0489191458699473	74	0.2012	0
symmetry_mean	106	0.1619	0.1792	0.181161862917399	0.1957	304	0
fractal_dimension_mean	0.04996	0.0577	0.06154	0.0627976098418278	0.06612	0.09744	0
radius_se	0.1115	0.2324	0.3242	0.405172056239016	0.4789	2.873	0
texture_se	0.3602	0.8339	1.108	1.21685342706503	1.474	4.885	0
perimeter_se	757	1.606	2.287	2.86605922671353	3.357	21.98	0
area_se	6.802	17.85	24.53	40.337079086116	45.19	542.2	0
smoothness_se	0.001713	0.005169	0.00638	0.00704097891036907	0.008146	0.03113	0
compactness_se	0.002252	0.01308	0.02045	0.0254781388400703	0.03245	0.1354	0
concavity_se	0	0.01509	0.02589	0.031893716344464	0.04205	396	0
concave points_se	0	0.007638	0.01093	0.0117961370826011	0.01471	0.05279	0
symmetry_se	0.007882	0.01516	0.01873	0.0205422987697715	0.02348	0.07895	0
fractal_dimension_se	0.0008948	0.002248	0.003187	0.00379490386643234	0.004558	0.02984	0
radius_worst	7.93	13.01	14.97	16.2691898066784	18.79	36.04	0
texture_worst	12.02	21.08	25.41	25.677223198594	29.72	49.54	0
perimeter_worst	50.41	84.11	97.66	107.261212653779	125.4	251.2	0
area_worst	185.2	515.3	686.5	880.583128295255	1084	4254	0
smoothness_worst	0.07117	0.1166	0.1313	0.132368594024605	146	0.2226	0
compactness_worst	0.02729	0.1472	0.2119	0.254265043936731	0.3391	1.058	0
concavity_worst	0	0.1145	0.2267	0.272188483304042	0.3829	1.252	0
concave points_worst	0	0.06493	0.09993	0.114606223198594	0.1614	291	0
symmetry_worst	0.1565	0.2504	0.2822	0.290075571177504	0.3179	0.6638	0
fractal_dimension_worst	0.05504	0.07146	0.08004	0.0839458172231986	0.09208	0.2075	0



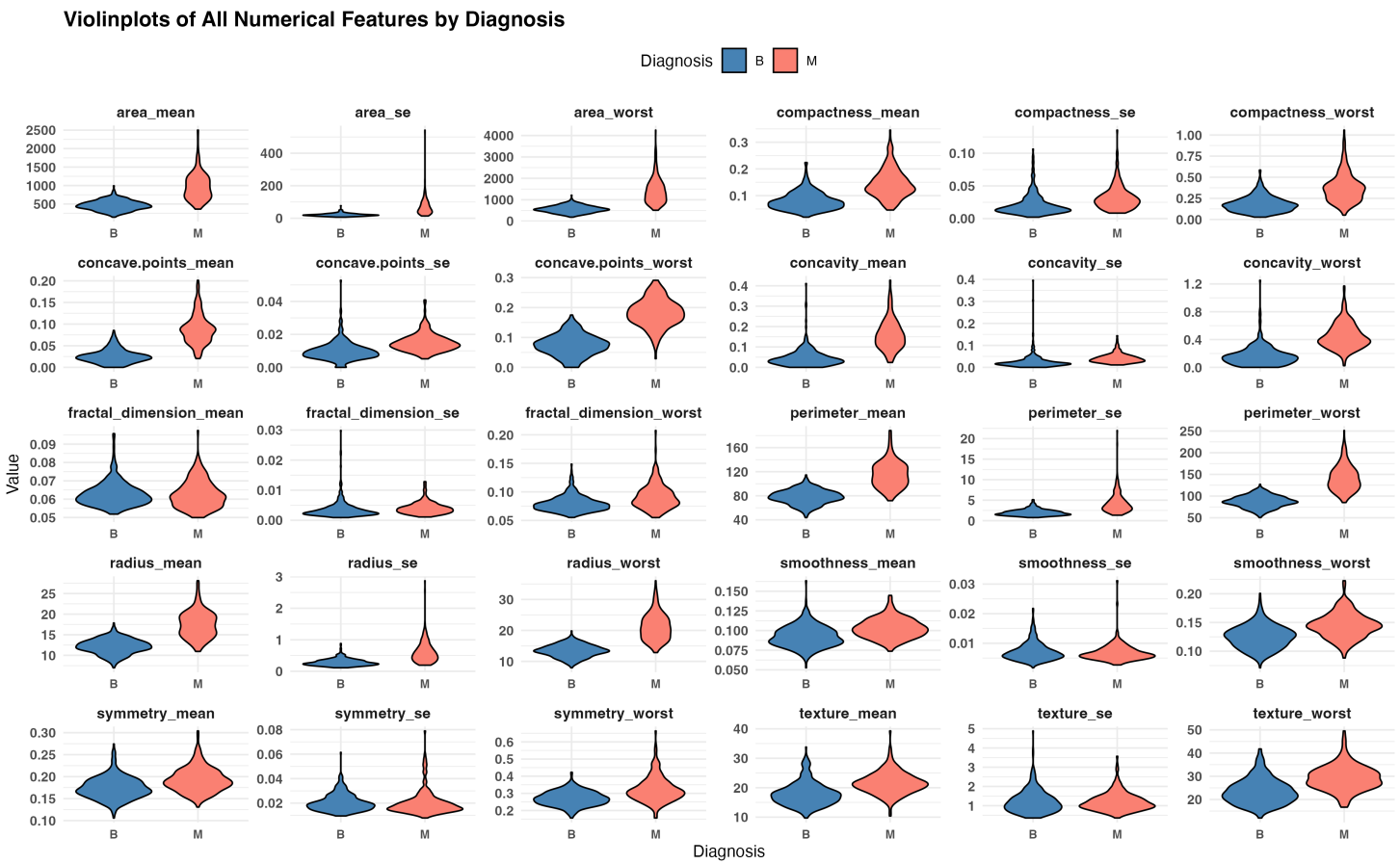


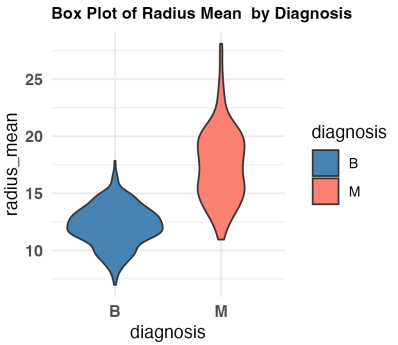
Exploratory Data Analysis:

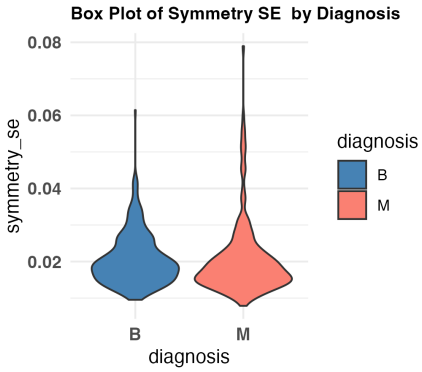


Exploratory Data Analysis:

Some conclusion and separated histoplots!!1!







Exploratory Data Analysis:

t_test_results_breast_cancer

Variable	t_statistic	Degrees_of_Freedom	p_value	Mean_Group_B	Mean_Group_M	CI_Lower	CI_Upper	Interpretation
radius_mean	-22.21	289.71	< 1e-04	12.15	17.46	-5.79	-4.85	Significant difference
texture_mean	-11.02	463.07	< 1e-04	17.91	21.6	-4.35	-3.03	Significant difference
perimeter_mean	-22.94	285.41	< 1e-04	78.08	115.37	-40.49	-34.09	Significant difference
area_mean	-19.64	244.79	< 1e-04	462.79	978.38	-567.29	-463.88	Significant difference
smoothness_mean	-9.3	466.21	< 1e-04	0.09	0.1	-0.01	-0.01	Significant difference
compactness_mean	-15.82	310.39	< 1e-04	0.08	0.15	-0.07	-0.06	Significant difference
concavity_mean	-20.33	296.43	< 1e-04	0.05	0.16	-0.13	-0.1	Significant difference
concave.points_mean	-24.84	265.55	< 1e-04	0.03	0.09	-0.07	-0.06	Significant difference
symmetry_mean	-8.11	406.09	< 1e-04	0.17	0.19	-0.02	-0.01	Significant difference
fractal_dimension_mean	0.3	403.64	0.7667	0.06	0.06	0	0	No significant difference
radius_se	-13.3	237.95	< 1e-04	0.28	0.61	-0.37	-0.28	Significant difference
texture_se	0.21	511.68	0.8354	1.22	1.21	-0.08	0.1	No significant difference
perimeter_se	-12.83	233.8	< 1e-04	2	4.32	-2.68	-1.97	Significant difference
area_se	-12.16	216.22	< 1e-04	21.14	72.67	-59.89	-43.18	Significant difference
smoothness_se	1.62	463.68	0.1053	0.01	0.01	0	0	No significant difference
compactness_se	-7.08	403.03	< 1e-04	0.02	0.03	-0.01	-0.01	Significant difference
concavity_se	-6.92	561.41	< 1e-04	0.03	0.04	-0.02	-0.01	Significant difference
concave.points_se	-10.74	455.45	< 1e-04	0.01	0.02	-0.01	0	Significant difference
symmetry_se	0.14	333.28	0.8871	0.02	0.02	0	0	No significant difference
fractal_dimension_se	-2.04	553.22	0.0422	0	0	0	0	Significant difference
radius_worst	-24.83	265.48	< 1e-04	13.38	21.13	-8.37	-7.14	Significant difference
texture_worst	-12.26	447.18	< 1e-04	23.52	29.32	-6.73	-4.87	Significant difference
perimeter_worst	-25.33	264.69	< 1e-04	87.01	141.37	-58.59	-50.14	Significant difference
area_worst	-20.57	229.91	< 1e-04	558.9	1422.29	-946.08	-780.69	Significant difference
smoothness_worst	-10.82	412.57	< 1e-04	0.12	0.14	-0.02	-0.02	Significant difference
compactness_worst	-15.16	285.62	< 1e-04	0.18	0.37	-0.22	-0.17	Significant difference
concavity_worst	-19.6	360.54	< 1e-04	0.17	0.45	-0.31	-0.26	Significant difference
concave.points_worst	-29.12	360.42	< 1e-04	0.07	0.18	-0.12	-0.1	Significant difference
symmetry_worst	-9.53	290.63	< 1e-04	0.27	0.32	-0.06	-0.04	Significant difference
fractal_dimension_worst	-7.32	315.23	< 1e-04	0.08	0.09	-0.02	-0.01	Significant difference

Non significant differences:

Radius_se

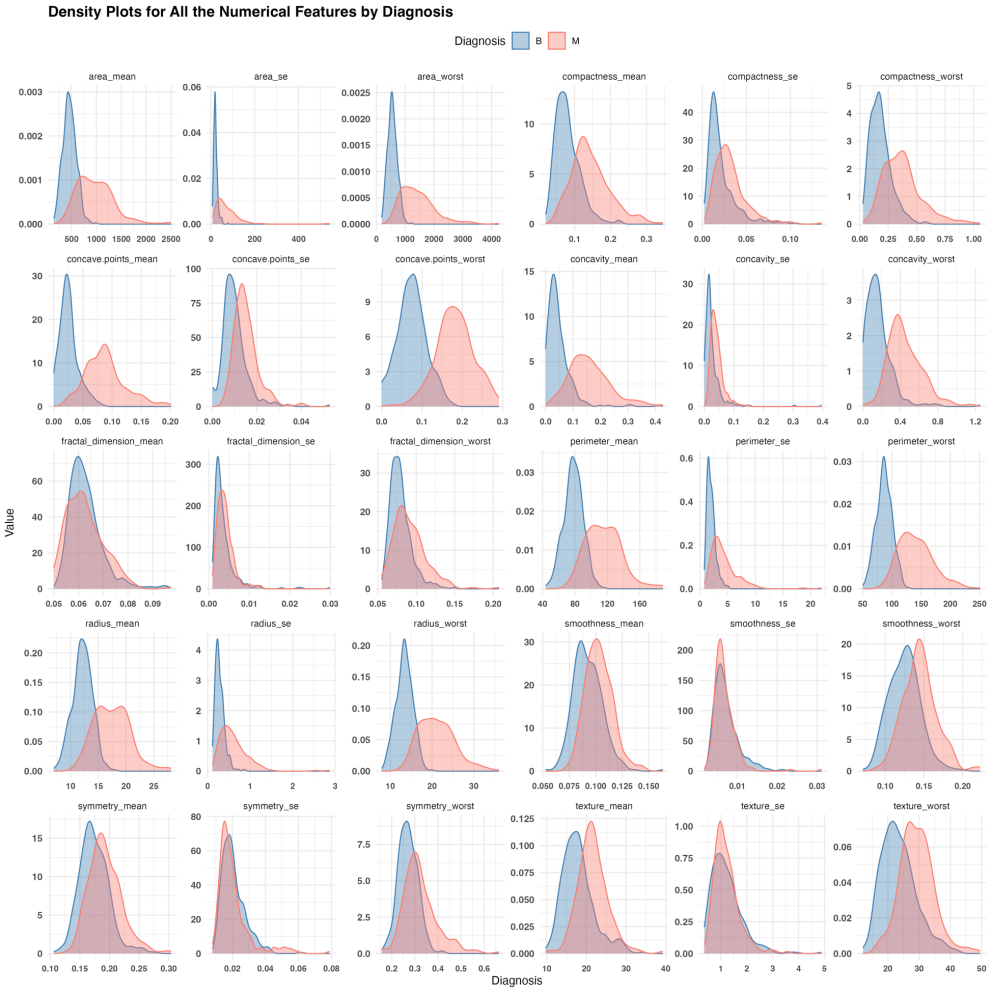
Texture_se

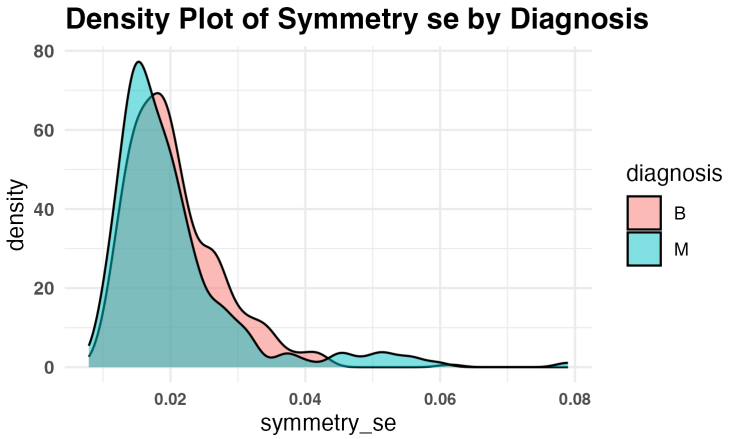
Smoothness_se

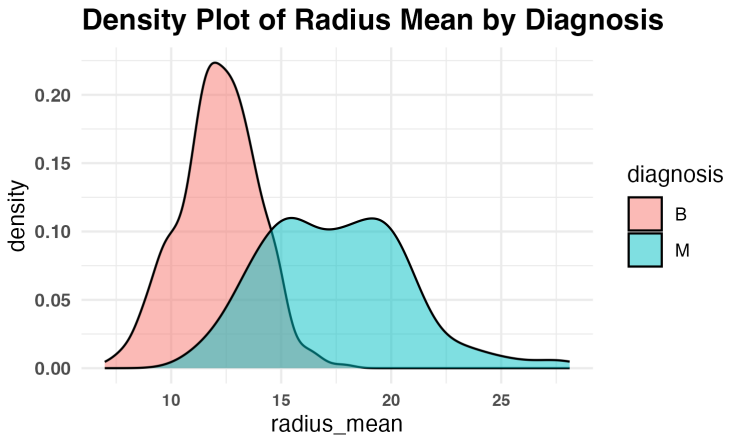
Symmetry_se

Fractal_dimension_se has as well quite high p-value (0.0422), but the difference is still significant

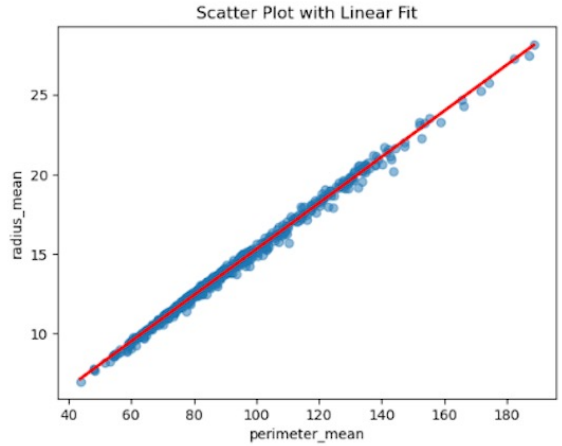
Exploratory Data Analysis:

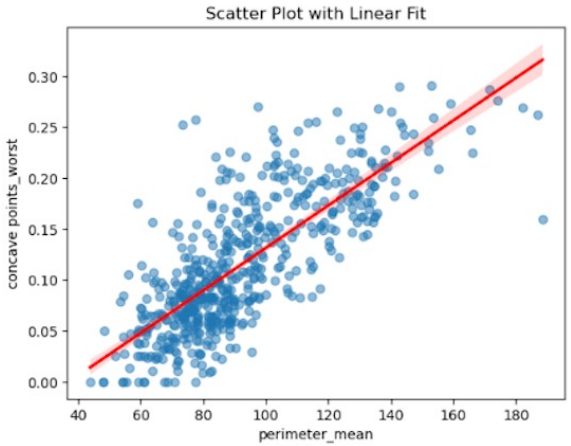


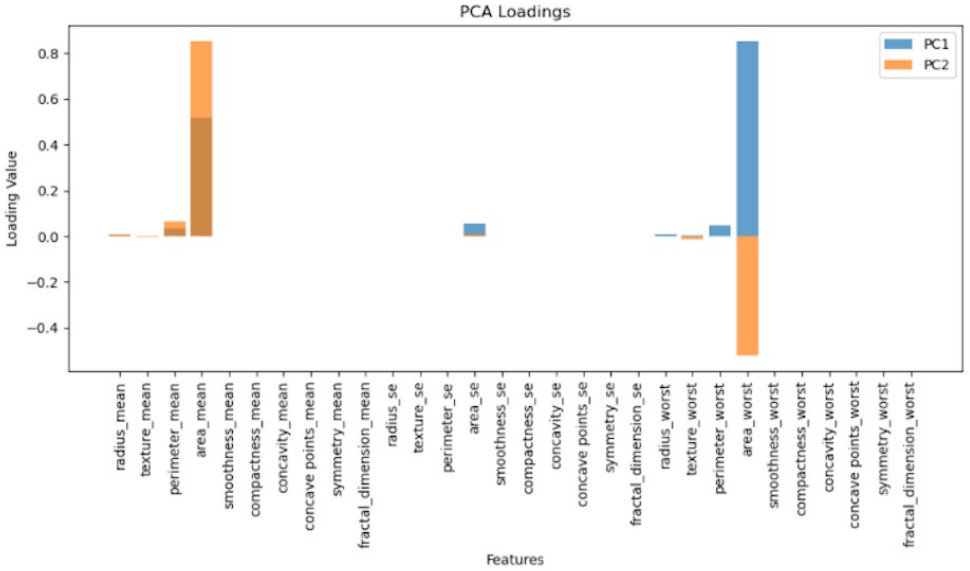




Exploratory Data Analysis: Some thoughts about colinearity of features, consideration about PCA loadings



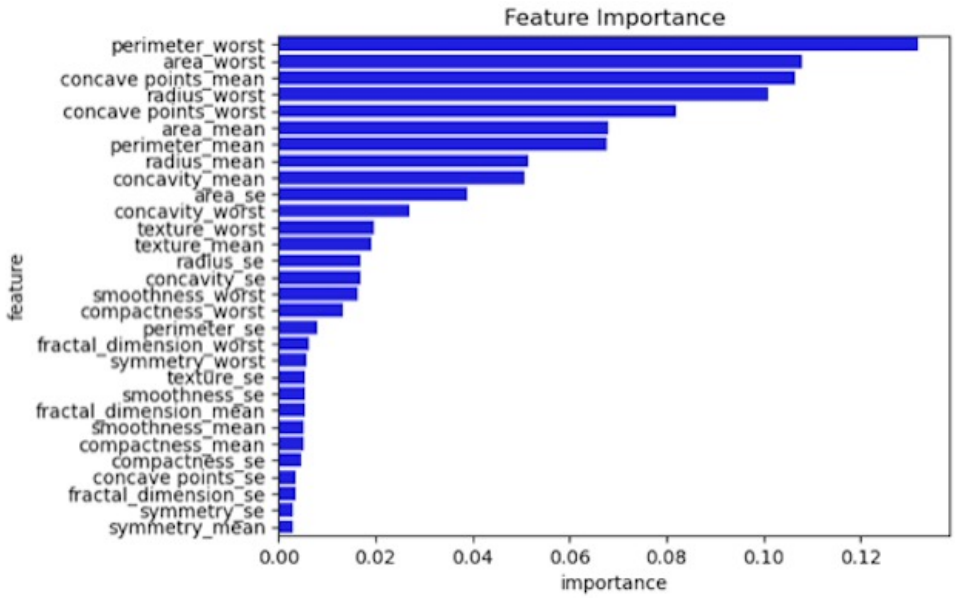




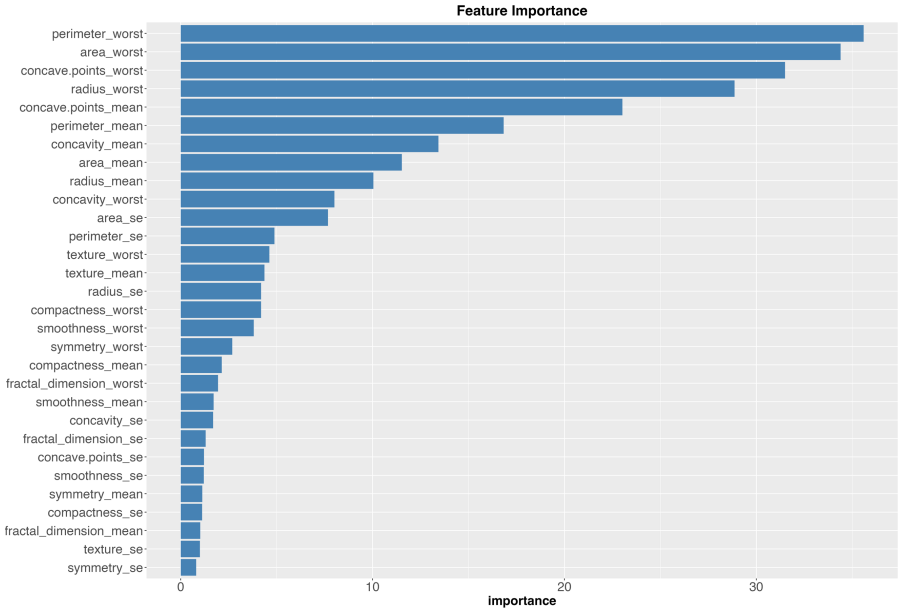
Machine Learning: Random Forest Model

Feature Importances in Random Forest

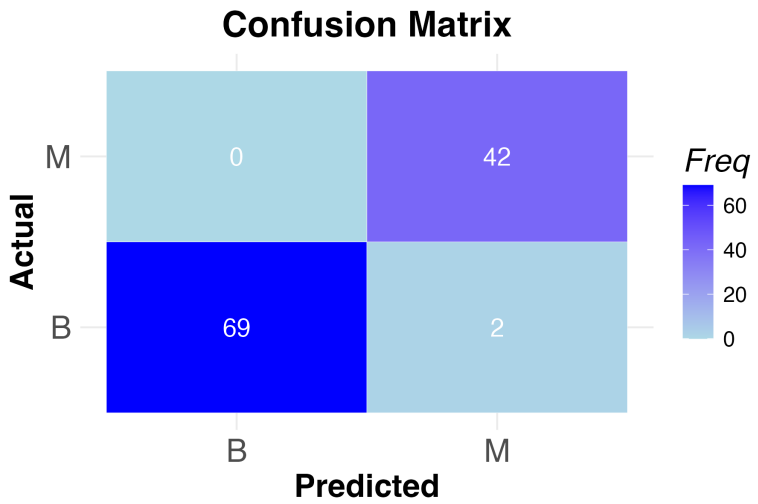
Python:



R:



Machine Learning: Random Forest Model ®



rf_model_metrics_breast_can

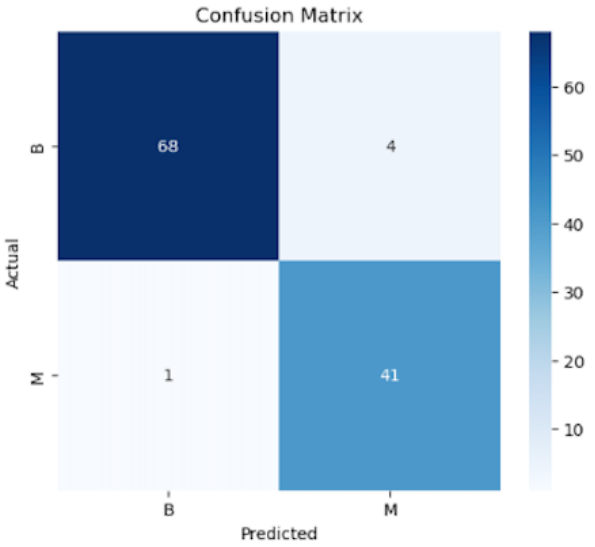
Metric	Value
Accuracy	0.982300884955752
Sensitivity	0.971830985915493
Specificity	1

1.Overall Accuracy: 98.23%
The model correctly predicted almost all samples.
2.Sensitivity (Recall) for Class B: 97.18%
The model identified 97.18% of true "B" cases correctly.
3.Specificity for Class B: 100%
All true "non-B" cases were classified correctly.
4.Precision (Positive Predictive Value) for Class B: 100%
Every case predicted as "B" was actually "B."
5.Negative Predictive Value (NPV): 95.45%
Most cases predicted as "non-B" were correct.
6.Balanced Accuracy: 98.59%
Average performance across sensitivity and specificity.
7.Kappa Score: 0.9625
Near-perfect agreement between predictions and actuals.

•**False Positives (FP)**: Predicted M but actually B → **2**.Misdiagnosing benign as malignant could lead to unnecessary treatment.
•**False Negatives (FN)**: Predicted B but actually M → **0**.No missed malignant cases, excellent outcome for this critical metric.

Interpretation: This is an **excellent model**, with high accuracy, sensitivity, and specificity. The low **false-negative rate**(only 2 missed "non-B" cases) and zero **false positives** for "B" are impressive. Further improvements might not be necessary unless I want to play with the numbers of features....(! Still to check in R)

Machine Learning: Random Forest Model (Python)




```
Classification Report:
              precision    recall  f1-score   support

           B       0.99      0.94      0.96        72
           M       0.91      0.98      0.94        42

    accuracy                           0.96       114
   macro avg       0.95      0.96      0.95       114
weighted avg       0.96      0.96      0.96       114

Confusion Matrix:
[[68  4]
 [ 1 41]]
```

For Random Model in Python I had to use GridSearchCV in order to improve it. Reduction of features (manual and SelectKbest with various k did not bring improvement.

Overall Accuracy: 96%

1. Slightly lower than the first model's 98.23%.

1.Class B:

1. **Precision**: 99% (very high, similar to the first model).
2. **Recall (Sensitivity)**: 94% (lower than the first model's 97.18%).
3. **F1-Score**: 96% (slightly lower than the first).

2.Class M:

1. **Precision**: 91% (lower than the first model's perfect precision for Class M).
2. **Recall (Sensitivity)**: 98% (higher than the first model's Class M recall).
3. **F1-Score**: 94% (comparable).

3.Confusion Matrix:

1. 1 false negative and 4 false positives for Class B.
2. Only 1 false negative for Class M

•**False Positives (FP)**: Predicted M but actually B → **4**.Misdiagnosing benign as malignant could lead to over-treatment.

•**False Negatives (FN)**: Predicted B but actually M → **1**.One missed malignant case—less ideal for a medical context compared to the first model

Which is Better?

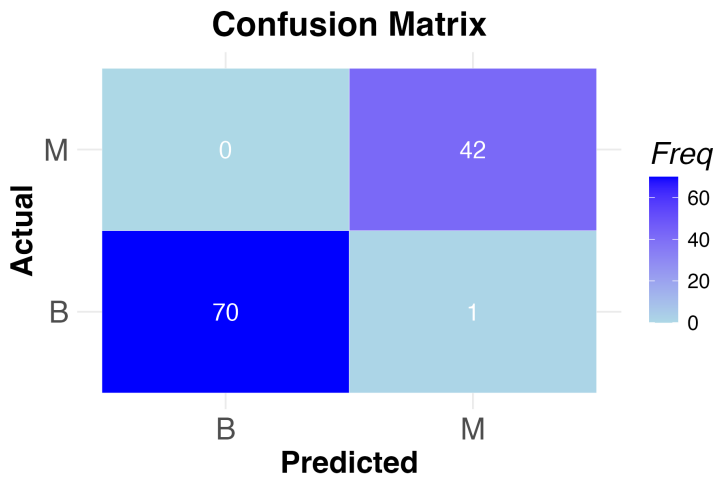
In this medical context, the **1st Model (R)** is more suitable because:

•It ensures no malignant cases are missed (**FN = 0**).

•Slightly higher false positives are a lesser concern compared to the critical risk of missing malignant cases.

Machine Learning: Support Vector machine (R)

Here the features (they are all numerica) were scaled with Standart Scaler.



svc_model_metrics_breast_c

Metric	Value
Accuracy	0.991150442477876
Sensitivity	0.985915492957746
Specificity	1

- **False Positives (FP)**: Predicted M but actually B → **1**.
- **False Negatives (FN)**: Predicted B but actually M → **0**.

- No missed malignant cases, which is excellent for a medical task.

Comparison with first R model:

Metric	Random Forest	SVC Model
Accuracy	0.9823	**0.9912**
False Negatives (FN)	0	**0**
False Positives (FP)	2	**1**
Sensitivity (Recall)	0.9718	**0.9859**
Specificity	1.0000	1.0000
Kappa	0.9625	**0.9811**
Balanced Accuracy	0.9859	**0.9930**

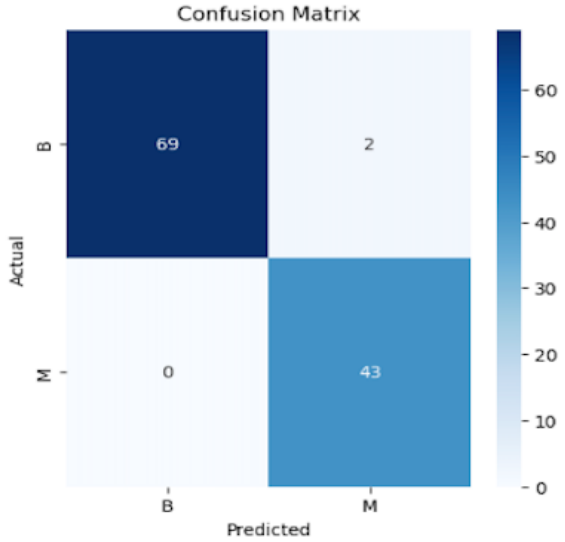
Conclusion

The **SVC model in R** is slightly better than the Random Forest model for this task:

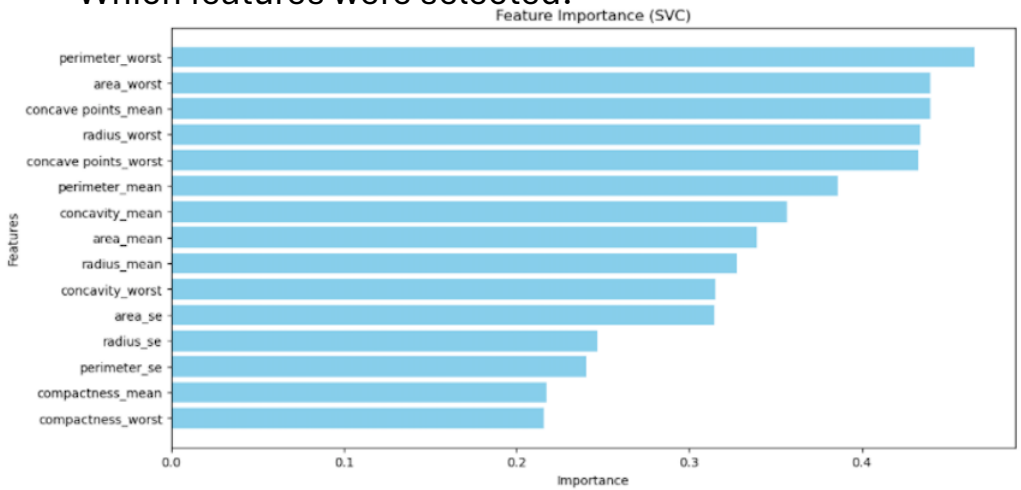
- It achieves **higher accuracy** and **lower false positives**, while maintaining zero false negatives

Machine Learning: Support Vector machine (Python)

Here the features were scaled with Standart Scaler. I need improvement of the intial model with GridSeacrCV and manual adjustmen, SelectKBest (the best k=15).
Still it slightly worse as SVC model made in R.



Which features were selected:



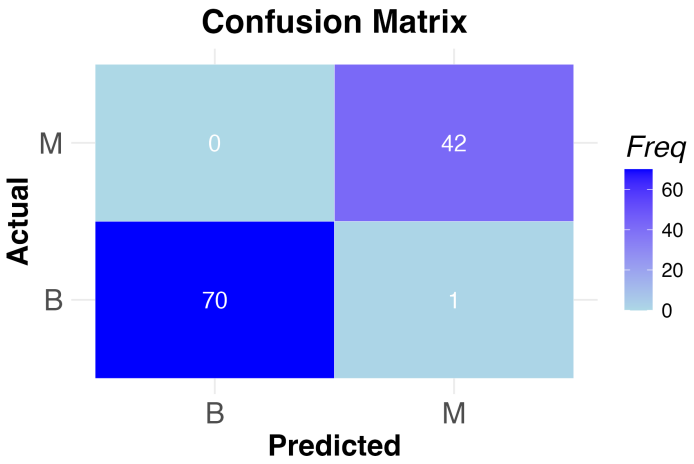
However it is the best model made in Python.

```
Classification Report:
              precision    recall  f1-score   support

           B       1.00      0.97      0.99        71
           M       0.96      1.00      0.98        43

    accuracy                           0.98       114
   macro avg       0.98      0.99      0.98       114
weighted avg       0.98      0.98      0.98       114

Confusion Matrix:
[[69  2]
 [ 0 43]]
```

- **False Positives (FP)**: Predicted M but actually B → **2**.
- **False Negatives (FN)**: Predicted B but actually M → **0**.

- No missed malignant cases, which is excellent for a medical task.

Machine Learning: XGBoost (R)



xgb_model_metrics_breast_c

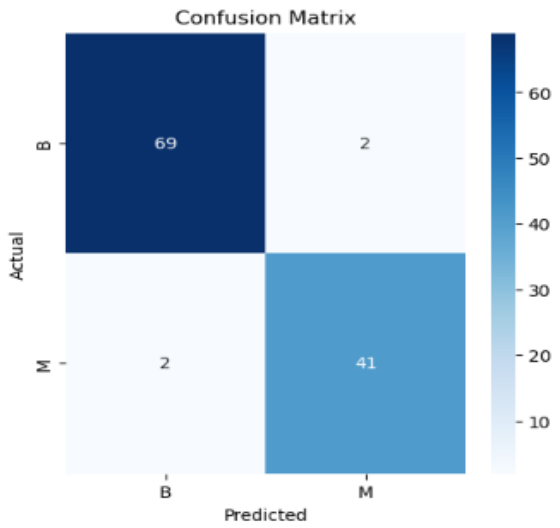
Metric	Value
Accuracy	0.991150442477876
Sensitivity	0.985915492957746
Specificity	1

The model has absolutely the same metric, which I got with SVC in R,
It is an excellent model as well.
I still can play with Features.

•**False Positives (FP)**: Predicted M but actually B → **1**.
•**False Negatives (FN)**: Predicted B but actually M → **0**.

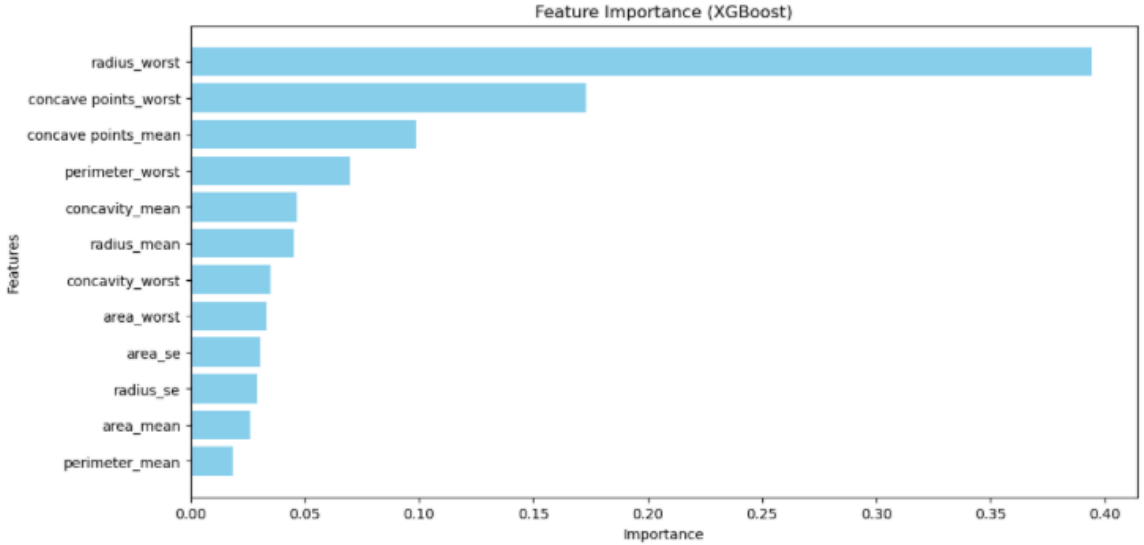
•No missed malignant cases.

Machine Learning: XGBoost (Python)




```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.97      0.97        71
           1       0.95      0.95      0.95        43

    accuracy                           0.96       114
   macro avg       0.96      0.96      0.96       114
weighted avg       0.96      0.96      0.96       114


Confusion Matrix:
[[69  2]
 [ 2 41]]
```

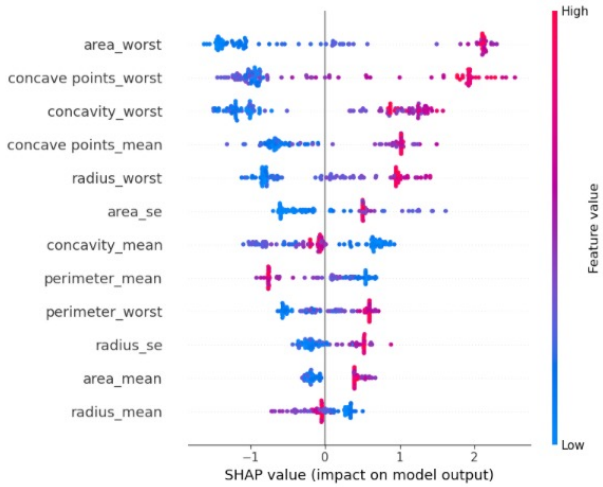
Selected features:



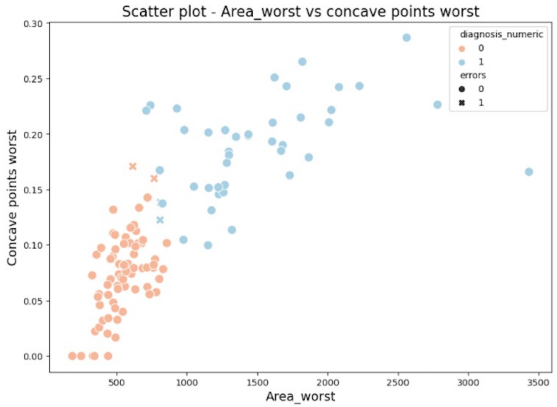
This model is not that good as the previous, and It need for further improvement. Additionally, I will make here error analysis (I did for almost all the model, but I show it here, since it produces the highest number of errors (2 false positive and 2 false negative)

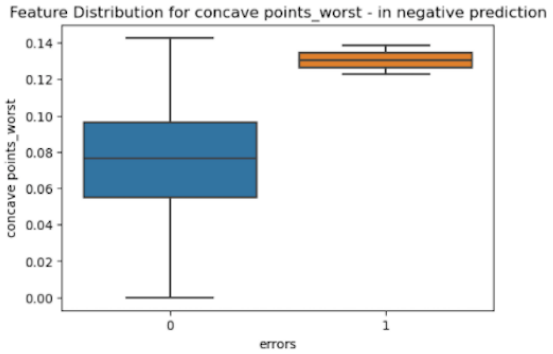
Machine Learning: XGBoost (Python)

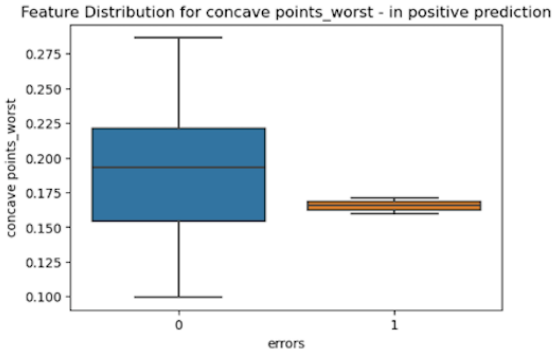
SHAP Values:



Errors analysis:



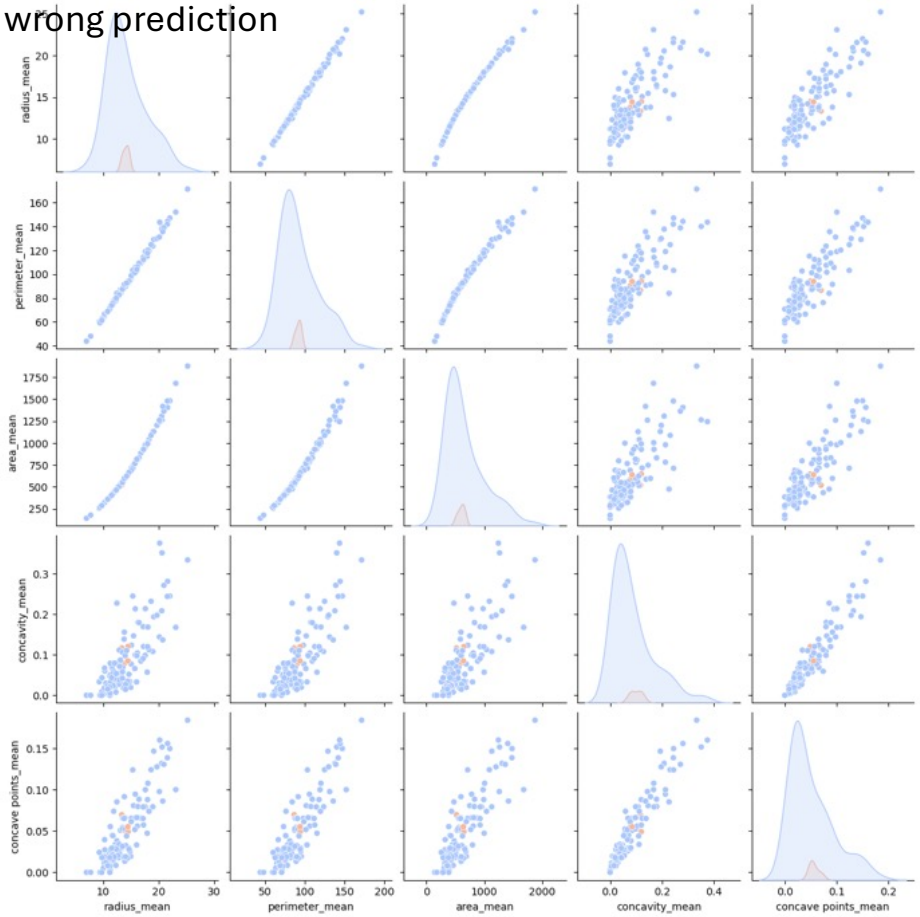




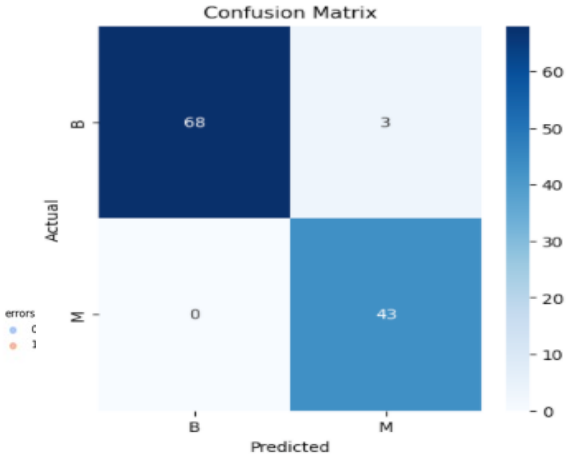
Features are ordered by their importance. The most impactful features are at the top. For example, area_worst has the highest impact on model predictions, followed by concave points_worst, concavity_worst, and so on.
SHAP Values (Left to Right): The x-axis represents the SHAP value, which indicates the direction and magnitude of a feature's contribution to the prediction: Negative SHAP values (< 0): The feature pushes the prediction towards class B (or the negative class in your dataset). Positive SHAP values (> 0): The feature pushes the prediction towards class M (or the positive class in your dataset).

Machine Learning: XGBoost (Python)

Error analysis: densitiy plot for some of the features for correct and wrong prediction



Improved model : I adjusted the decision threshold for classification




```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.96      0.98        71
           1       0.93      1.00      0.97        43

    accuracy                           0.97       114
   macro avg       0.97      0.98      0.97       114
weighted avg       0.98      0.97      0.97       114

Confusion Matrix:
[[68  3]
 [ 0 43]]

ROC AUC Score: 0.9789
```

What Does Changing the Threshold Do?

•By adjusting the threshold (e.g., to 0.2), you're saying:

- Predict **class 1** if the probability is **≥ 0.2**, rather than the default 0.5.
- This **lowers the bar** for classifying an instance as class 1.

Why Adjust the Threshold?

•**Medical Applications**: False negatives (e.g., missing a malignant tumor) are often more dangerous than false positives. Lowering the threshold helps catch more cases of class 1 (malignant).