Breast Cancer Prediction

Code available here: https://github.com/mtvz42/Breast_Cancer_Prediction/tree/main

Project goal: Predict whether a breast tumor is benign or malignant by building multiple ML classification algorithms to help physicians classify new breast tumor measurements.

Dataset: Breast Cancer Wisconsin Dataset
https://www.kaggle.com/datasets/rahmasleam/breast-cancer?select=breast-cancer.csv

Author: Matteo Vezzelli, PhD
https://www.linkedin.com/in/matteovezzelli/

Background

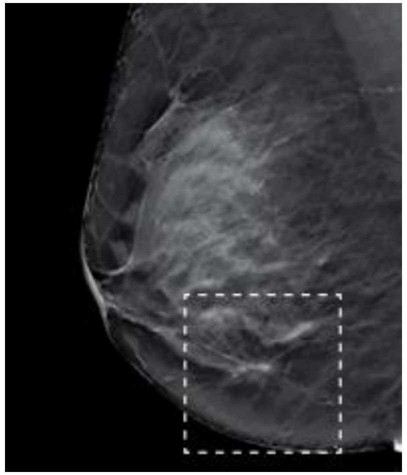
Breast cancer: most prevalent cancer among women globally.

Disease begins when cells in the breast grow uncontrollably, forming tumors.
Tumorst can be detected via X-ray or felt as lumps.

Good **article** to start with: https://mayoclinic.elsevierpure.com/en/publications/breast-cancer-6

From the **X-ray image**, the tumours can be measured in different characteristics:

- **radius** of the tumor.
- **perimeter** of the tumor.
- **area** of the tumor.
- **texture** of the tumor's surface (variation in pixel - intensity within the tumor region).
- **smoothness** of the tumor’s surface (how much it deviates from being spherical).
- **compactness** of the tumor (how compact the shape is, related to the ratio of perimeter to are).
- etc. (the full list of features is available in the code notebook).
- **diagnosis: benign or malignant: this is what we want to predict based on all measurements.**

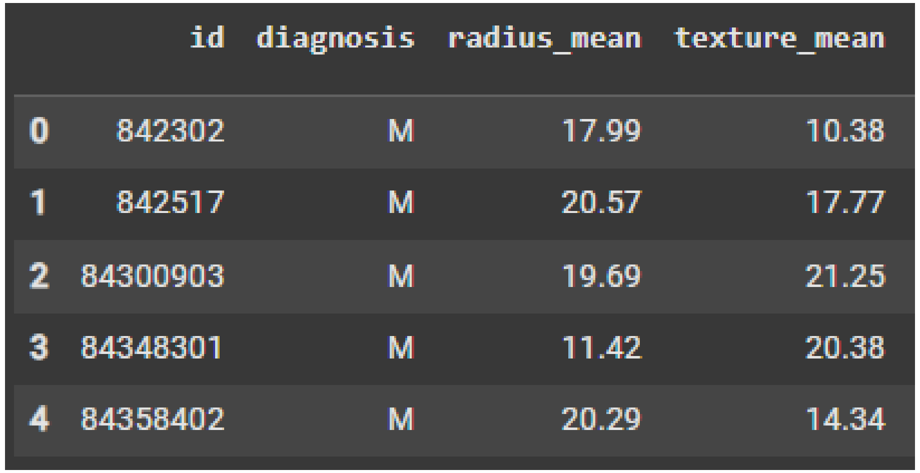


X-Ray image
Harbeck et al, 2019

EDA and Feature Engineering

Dataset: Breast Cancer Wisconsin Dataset

569 measurements (rows)



	id	diagnosis	radius_mean	texture_mean
0	842302	M	17.99	10.38
1	842517	M	20.57	17.77
2	84300903	M	19.69	21.25
3	84348301	M	11.42	20.38
4	84358402	M	20.29	14.34

32 features (columns)

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	…		
12	…		

EDA and Feature Engineering

Encoding

diagnosis 569 non-null   object

LabelEncoder
B (Bening): 0
M (Malignant): 1

Correlation with diagnosis

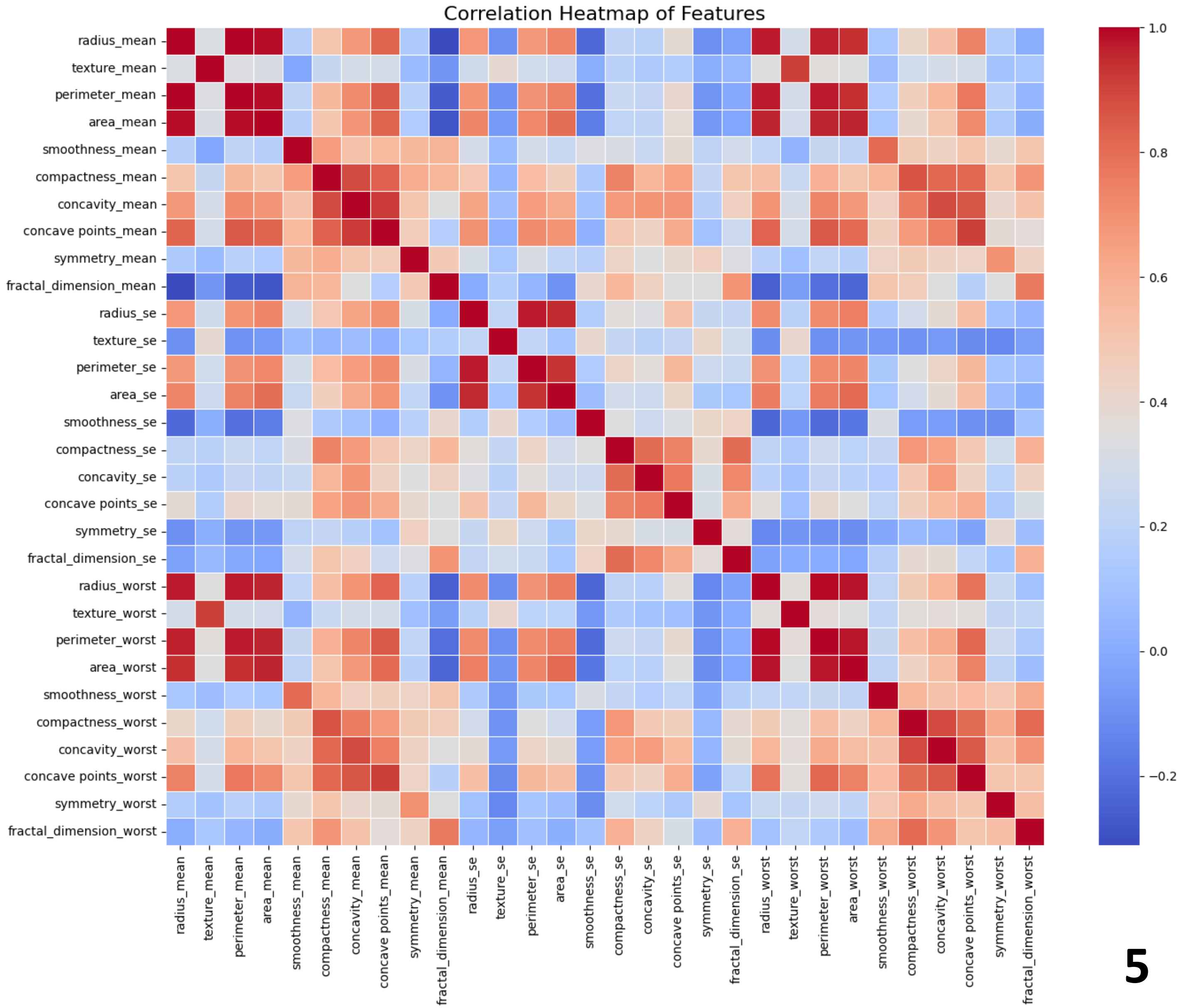
diagnosis	1.00
concave points_worst	0.79
perimeter_worst	0.78
concave points_mean	0.78
radius_worst	0.78
perimeter_mean	0.74
area_worst	0.73
radius_mean	0.73
area_mean	0.71
concavity_mean	0.70
…	

Exploratory Data Analysis

Correlation between features (without id, diagnosis)

Strong correlations among features like radius_mean, perimeter_mean, area_mean, indicating larger tumors tend to have proportional measurements.

Next:
data distributions…



EDA and Feature Engineering

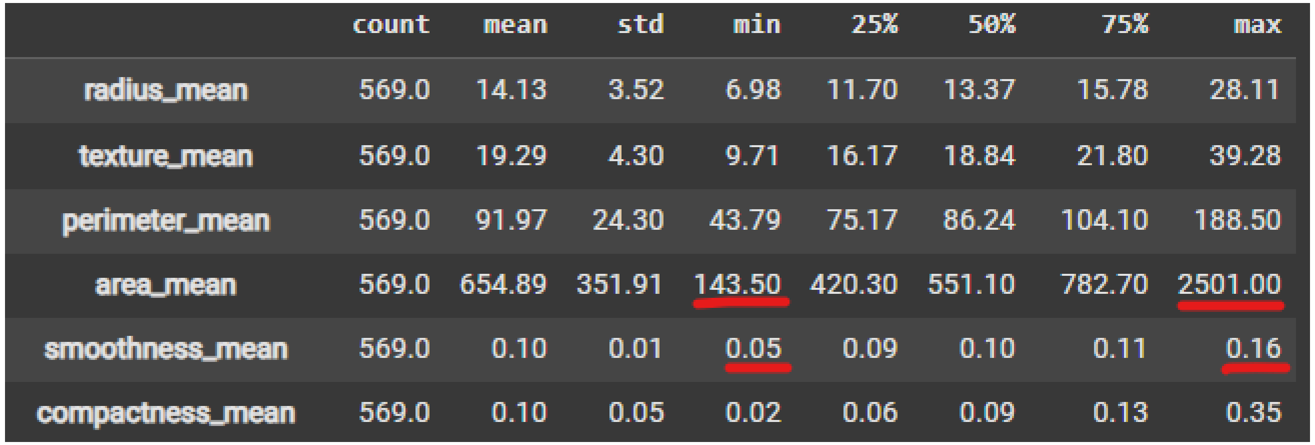
Dataset split
X: all features without id, diagnosis
y: diagnosis

y distribution
0: 62.7%
1: 37.2%

Unbalanced distribution
Split the train and test sets using one of the following:

- StratifiedShuffleSplit
- train_test_split with the stratify parameter

X distribution



	count	mean	std	min	25%	50%	75%	max
radius_mean	569.0	14.13	3.52	6.98	11.70	13.37	15.78	28.11
texture_mean	569.0	19.29	4.30	9.71	16.17	18.84	21.80	39.28
perimeter_mean	569.0	91.97	24.30	43.79	75.17	86.24	104.10	188.50
area_mean	569.0	654.89	351.91	143.50	420.30	551.10	782.70	2501.00
smoothness_mean	569.0	0.10	0.01	0.05	0.09	0.10	0.11	0.16
compactness_mean	569.0	0.10	0.05	0.02	0.06	0.09	0.13	0.35

Different value scales. Each model has a better way of scaling:

- StandardScaler for Logistic Regression, SVM
- MinMaxScaler for KNN
- No scaling for Decision Trees, Random Forest, Extra Trees, XGBoost

Models

Each model was built using:

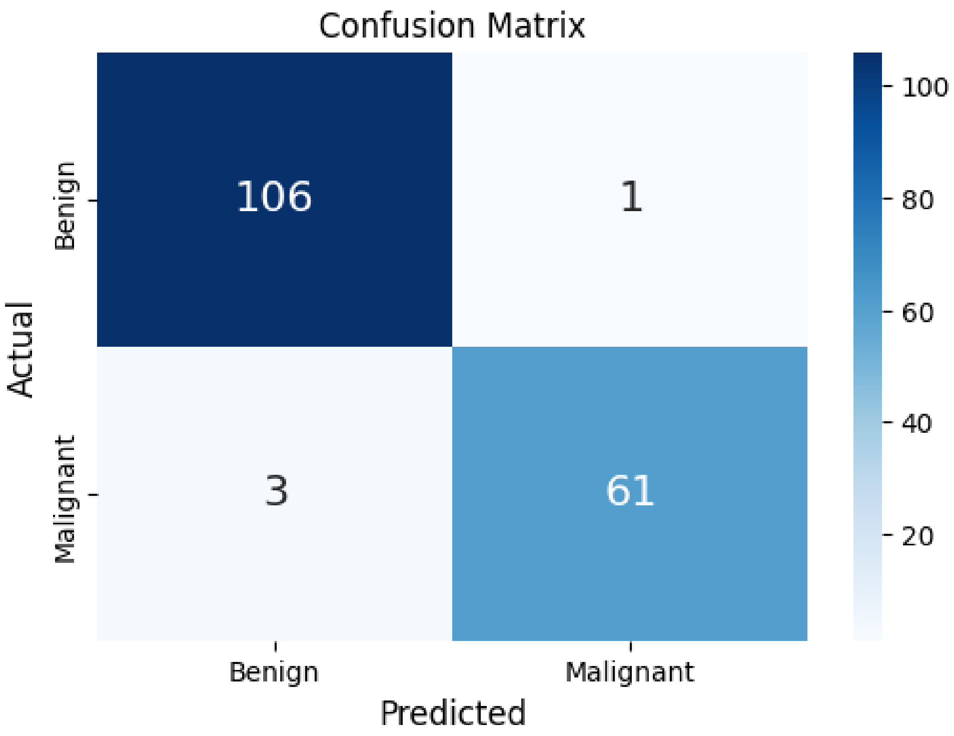
- **GridSearchCV** for hyperparameter tuning and cross-validation
- Function to calculate the **execution time** of the training
- **Scoring** F1

Models:

- Logistic Regression
- KNN
- SVM
- Decision Trees
- Random Forest
- Extra Trees
- XGBoost (not run due to a bug in the libraries)

Models

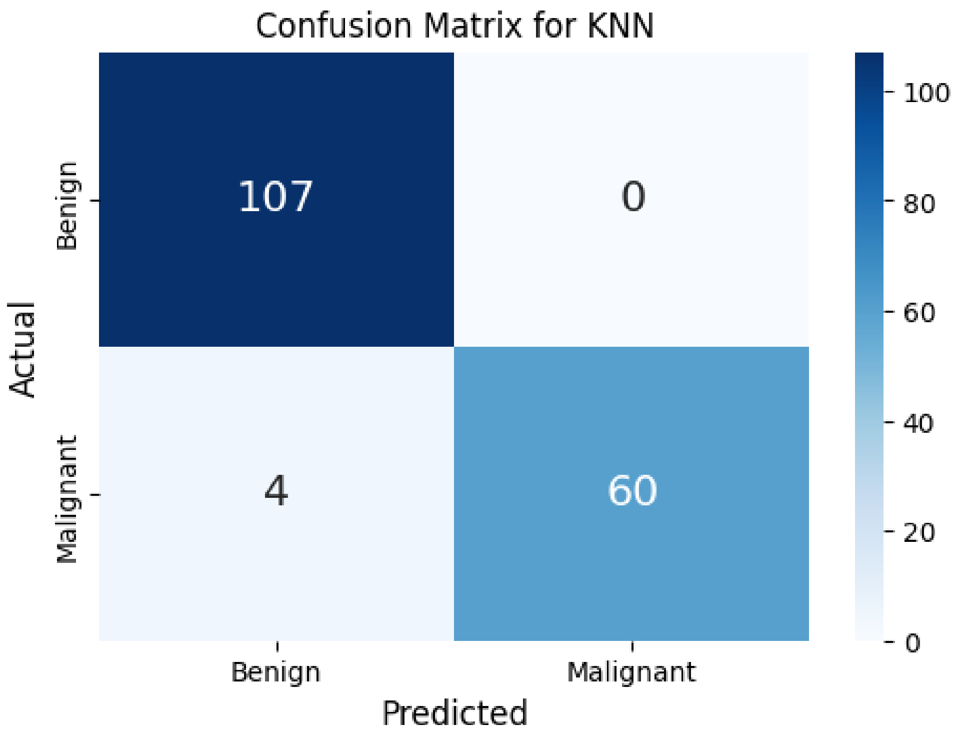
Logistic Regression			
Metric	Class 0	Class 1	Mean
Precision	0.97	0.98	0.98
Recall	0.99	0.95	0.97
F1-score	0.98	0.97	0.97
Support	107	64	171
Best parameters	C=1, penalty=l2, solver=liblinear		
Test set accuracy	0.968		
CV score	0.964		



Class 0 recall: 0.99 (1 case)
Class 1 recall: 0.95 (3 cases), most important metric: malignant tumor classified as bening

Models

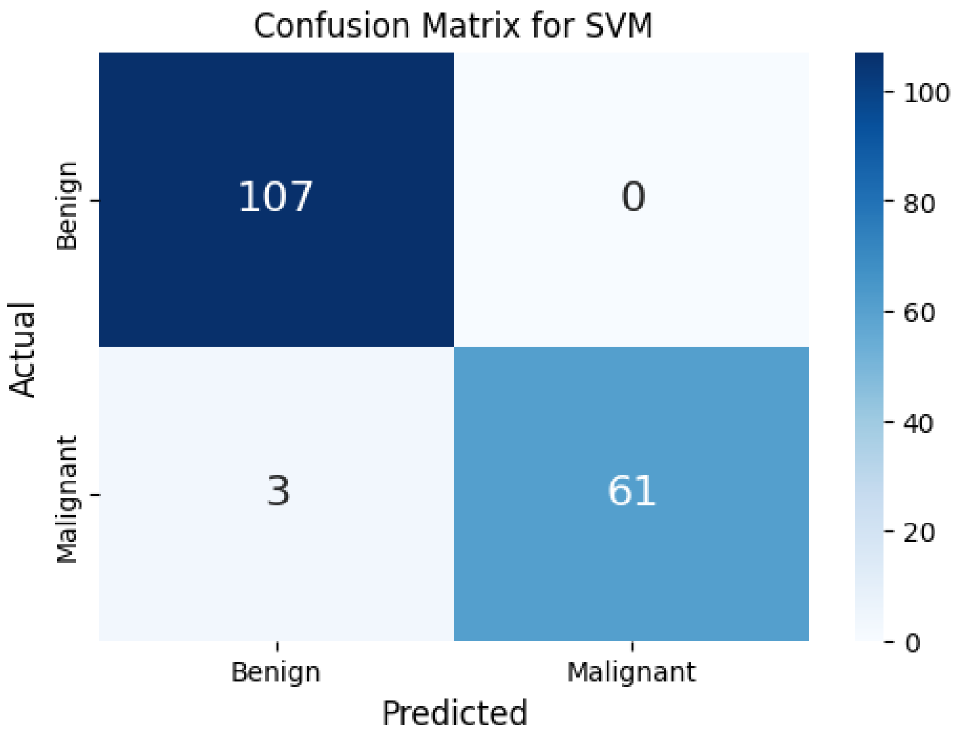
KNN			
Metric	Class 0	Class 1	Mean
Precision	0.96	1	0.98
Recall	1	0.94	0.97
F1-score	0.98	0.97	0.97
Support	107	64	171
Best parameters	metric=manhattan, n_neighbors=3, weights= uniform		
Test set accuracy	0.968		
CV score	0.968		



Class 0 recall: 1 (0 cases), better that Logistic Regression
Class 1 recall: 0.94 (4 cases), worse than Logistic Regression

Models

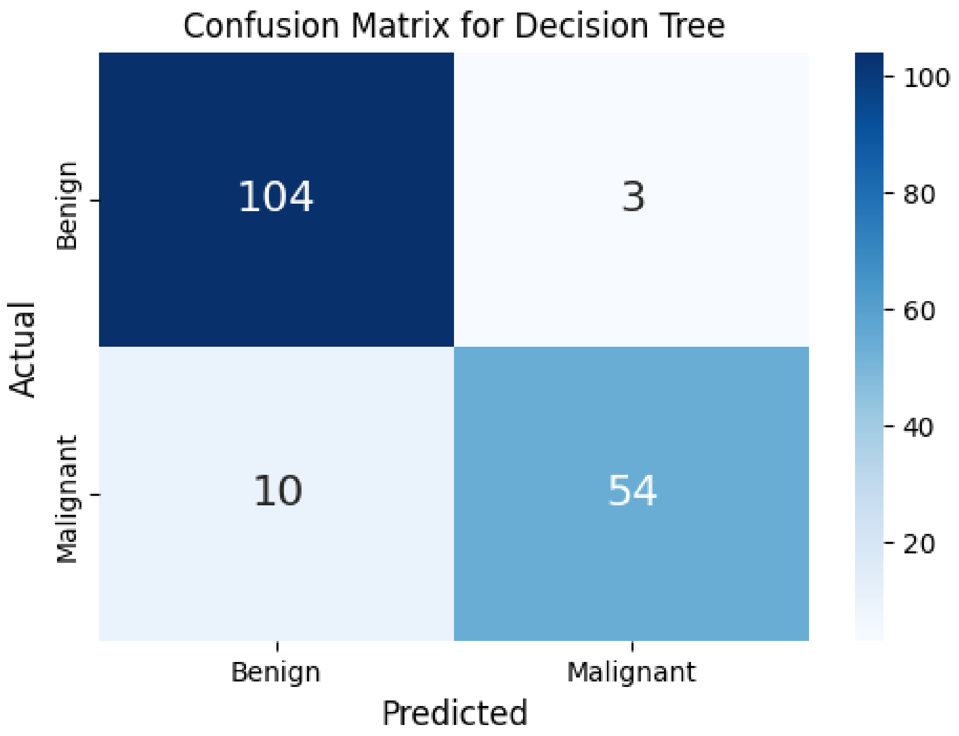
SVM			
Metric	Class 0	Class 1	Mean
Precision	0.97	1	0.99
Recall	1	0.95	0.98
F1-score	0.99	0.98	0.98
Support	107	64	171
Best parameters	C=0.1, gamma=1, kernel=linear		
Test set accuracy	0.976		
CV score	0.958		



Class 0 recall: 1 (0 cases), same as KNN
Class 1 recall: 0.95 (3 cases), better than Logistic Regression

Models

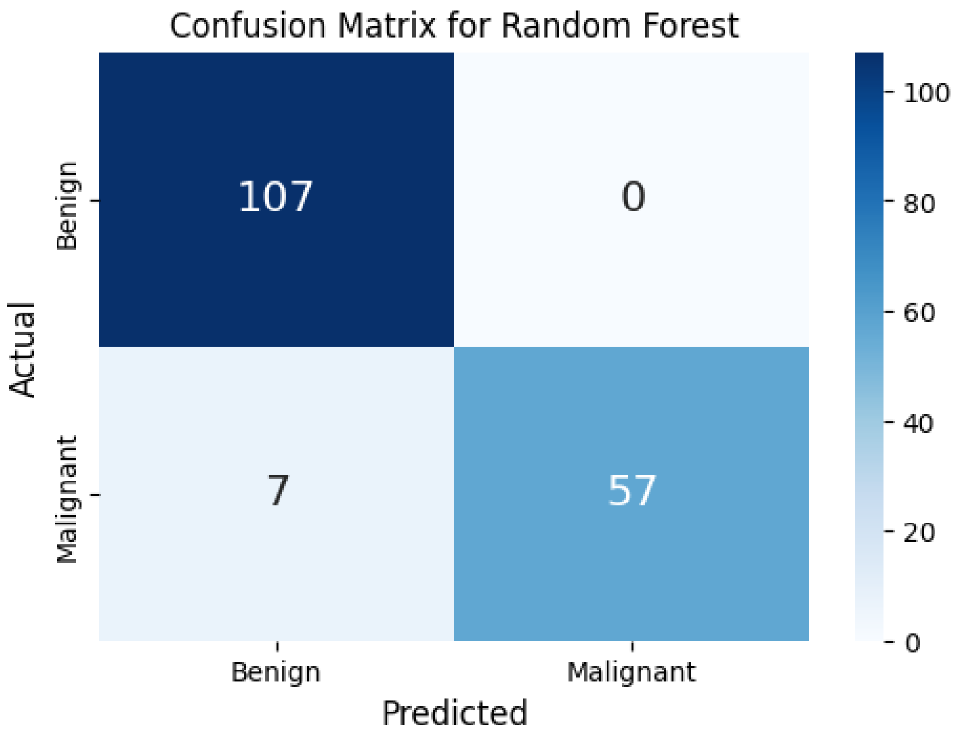
Decision Tree			
Metric	Class 0	Class 1	Mean
Precision	0.91	0.95	0.93
Recall	0.97	0.84	0.91
F1-score	0.94	0.89	0.92
Support	107	64	171
Best parameters	class_weight=None, criterion=gini, max_depth=6, max_features=log2, min_samples_leaf=1		
Test set accuracy	0.893		
CV score	0.918		



Class 0 recall: 0.97 (3 cases), worse that others
Class 1 recall: 0.84 (10 cases), worse than others

Models

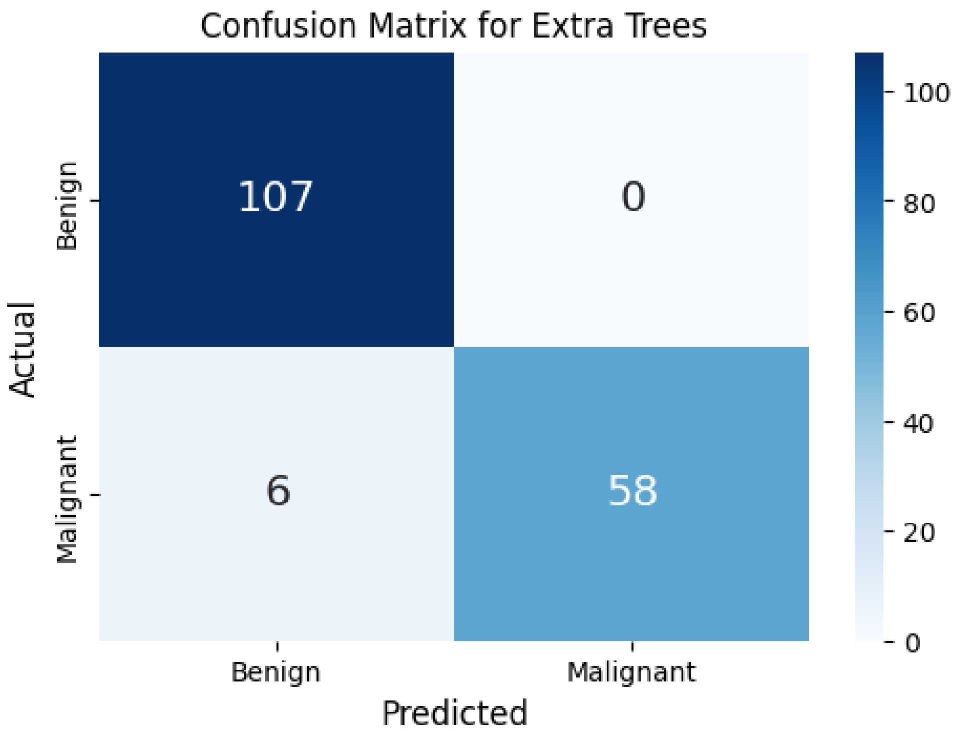
Random Forest			
Metric	Class 0	Class 1	Mean
Precision	0.94	1	0.97
Recall	1	0.89	0.95
F1-score	0.97	0.94	0.96
Support	107	64	171
Best parameters	bootstrap=True, class_weight=None, criterion= entropy, max_depth=7, max_features=sqrt, min_samples_leaf=1, n_estimators=50		
Test set accuracy	0.942		
CV score	0.955		



Class 0 recall: 1 (0 cases), same as SVM
Class 1 recall: 0.89 (7 cases), worse than SVM

Models

Extra Trees			
Metric	Class 0	Class 1	Mean
Precision	0.95	1	0.97
Recall	1	0.91	0.95
F1-score	0.97	0.95	0.96
Support	107	64	171
Best parameters	criterion=gini, max_depth=15, max_features=log2, min_samples_leaf=1, min_samples_split=2, n_estimators=50		
Test set accuracy	0.951		
CV score	0.961		



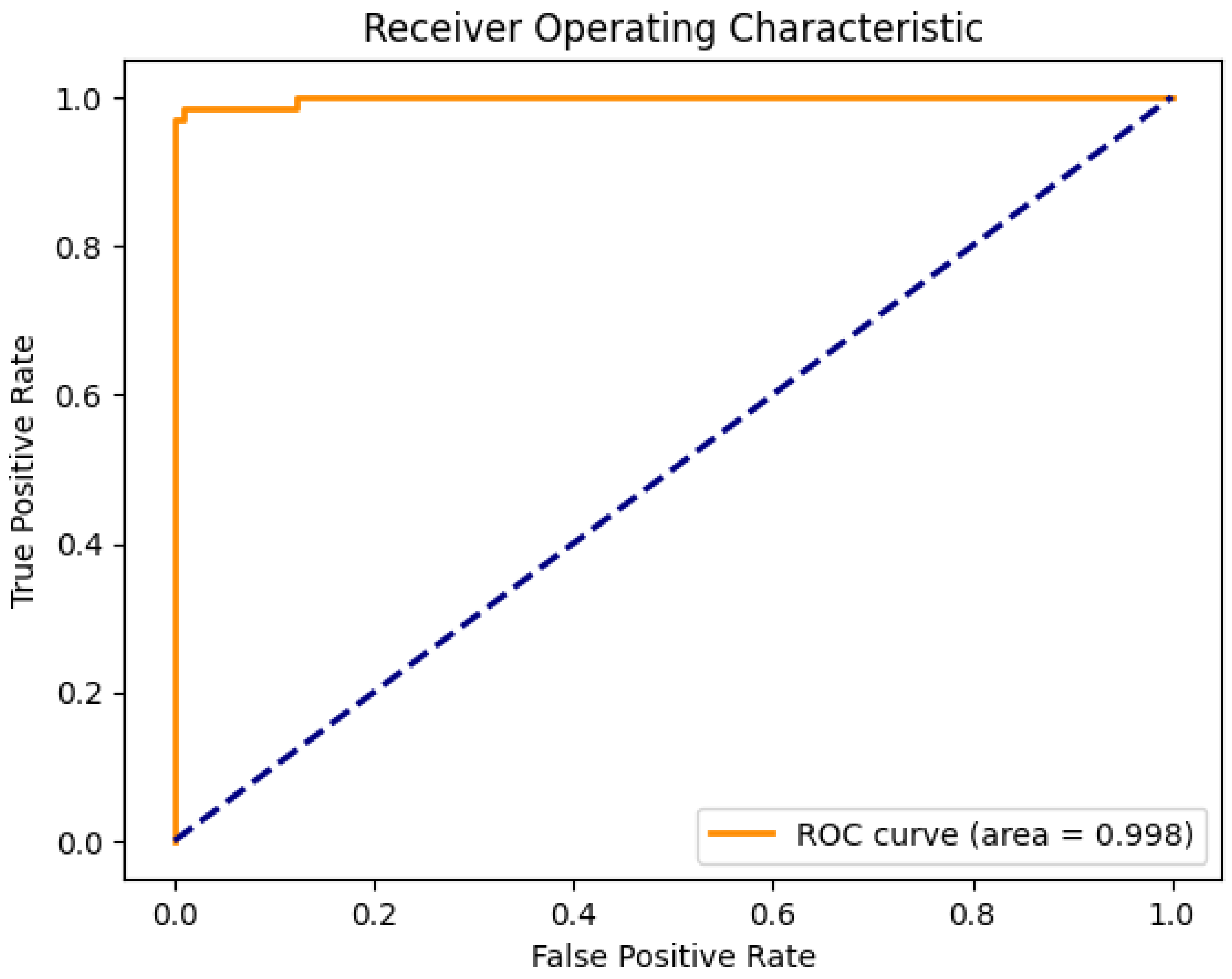
Class 0 recall: 1 (0 cases), same as SVM
Class 1 recall: 0.91 (6 cases), worse than SVM

Summary

Model	Test Accuracy	F1-score (Class 0)	F1-score (Class 1)	Execution Time (s)
Logistic Regression	0.9683	0.9815	0.9683	25.5064
KNN	0.9677	0.9817	0.9677	11.5324
SVM	0.9760	0.9862	0.9760	14.9485
Decision Tree	0.8926	0.9412	0.8926	41.5725
Random Forest	0.9421	0.9683	0.9421	818.2637
Extra Trees	0.9508	0.9727	0.9508	55.4807

- The best model in terms of accuracy and F1 score is **SVM**.
- Note also that the higher execution time is for **Random Forest**, which took more than 800 seconds (13 minutes). It's very important to balance the accuracy of the model with the computational cost. Sometimes it is fair to reduce accuracy to speed up training.

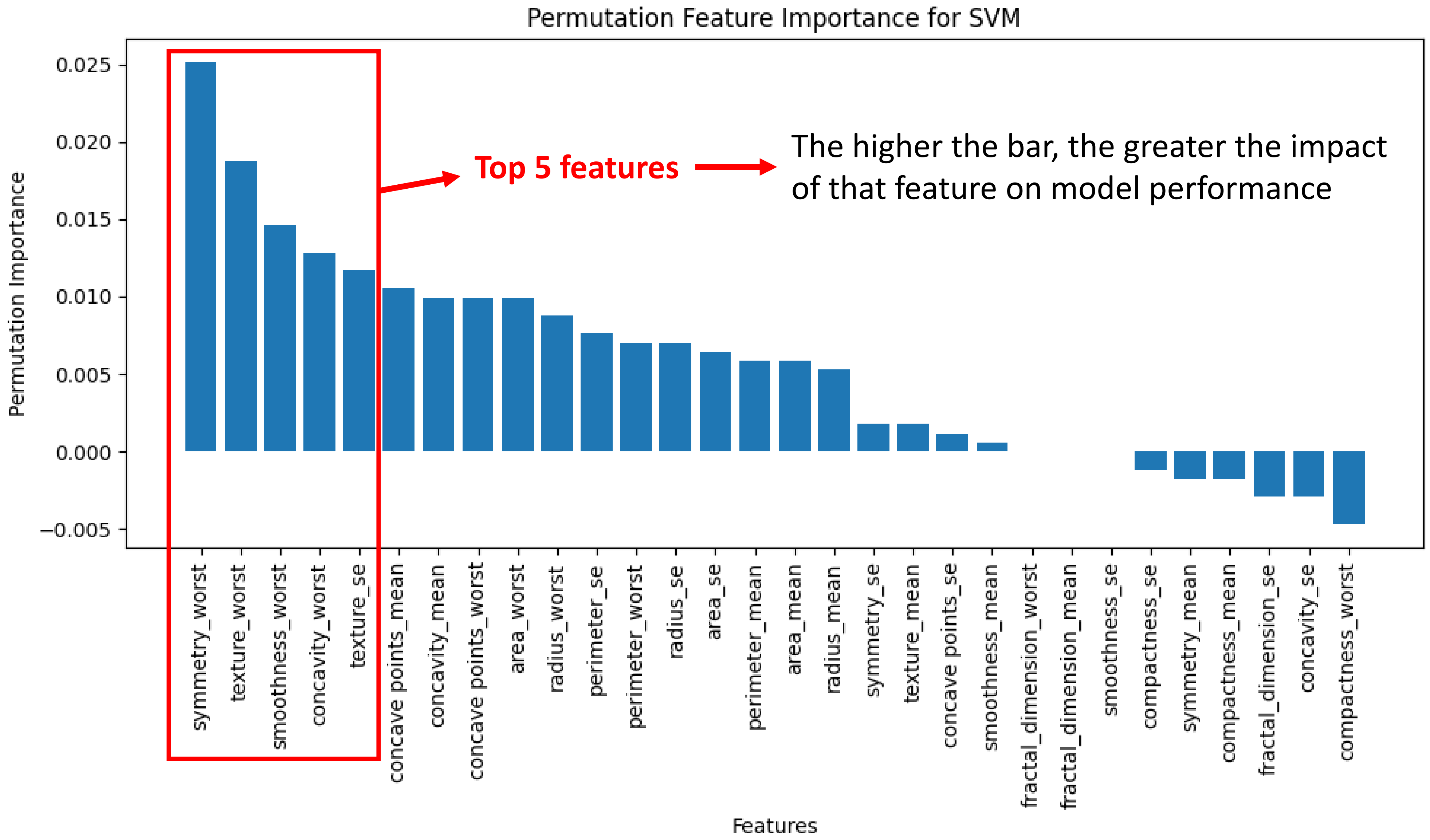
Summary



ROC curve and AUC score for SVM

The SVM model performs very well in a binary classification, with the ROC curve closer to the upper left part (meaning a high true positive rate and a very low false positive rate).
The area under the curve is 0.998, very close to 1, which means an excellent model that correctly classifies the classes.

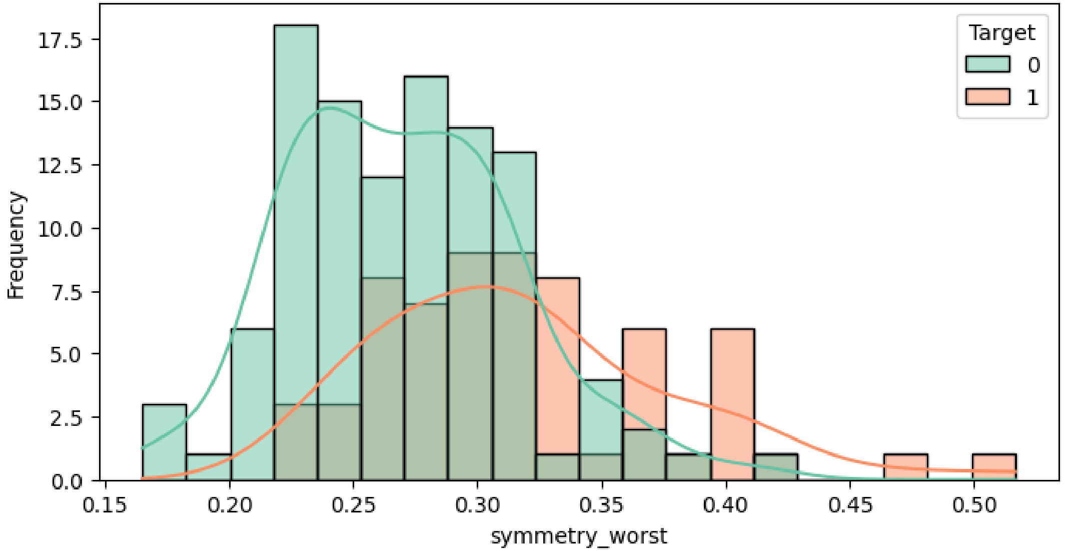
Analysis of Results



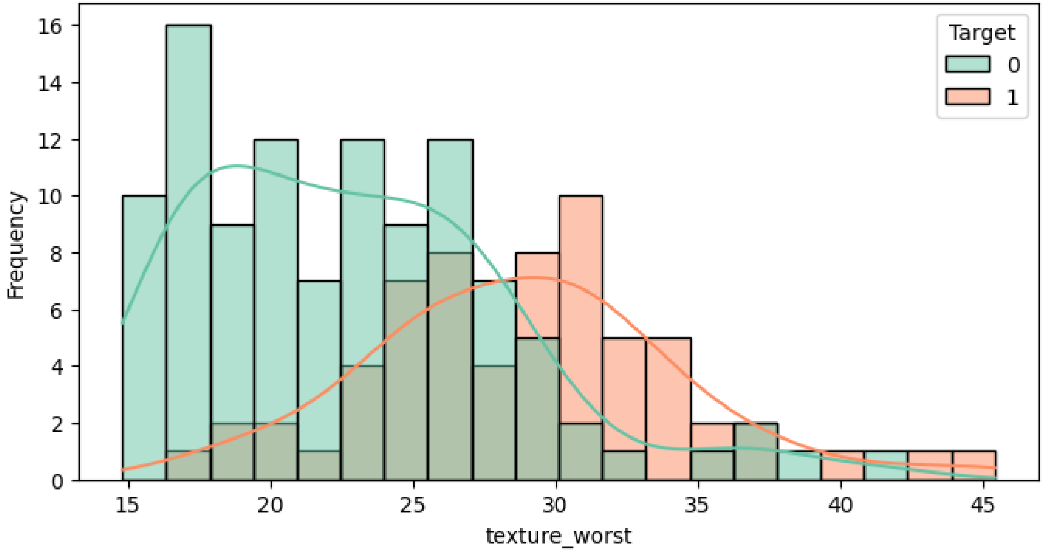
Top 5 features →

The higher the bar, the greater the impact of that feature on model performance

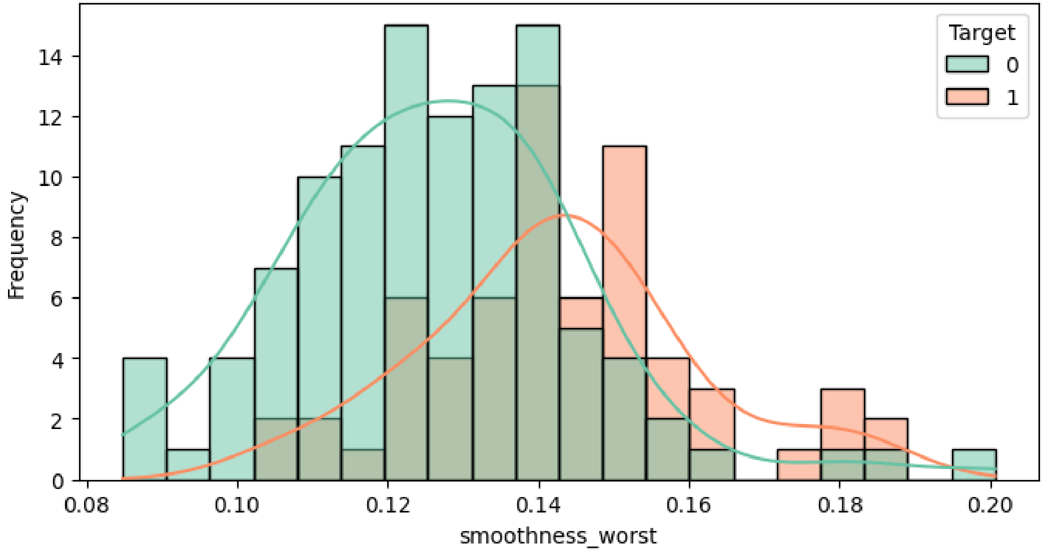
Analysis of Results



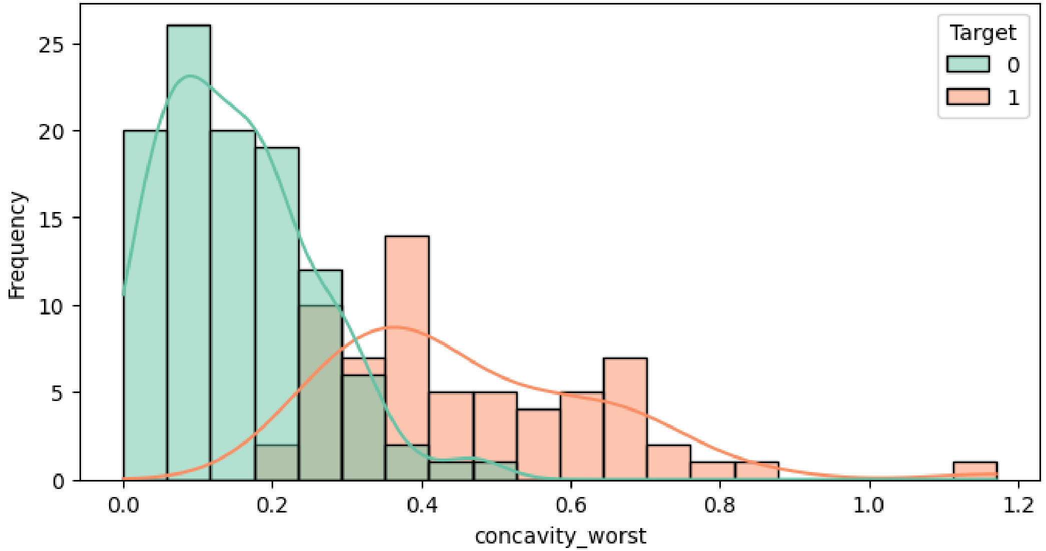
Distribution of symmetry_worst



Distribution of texture_worst

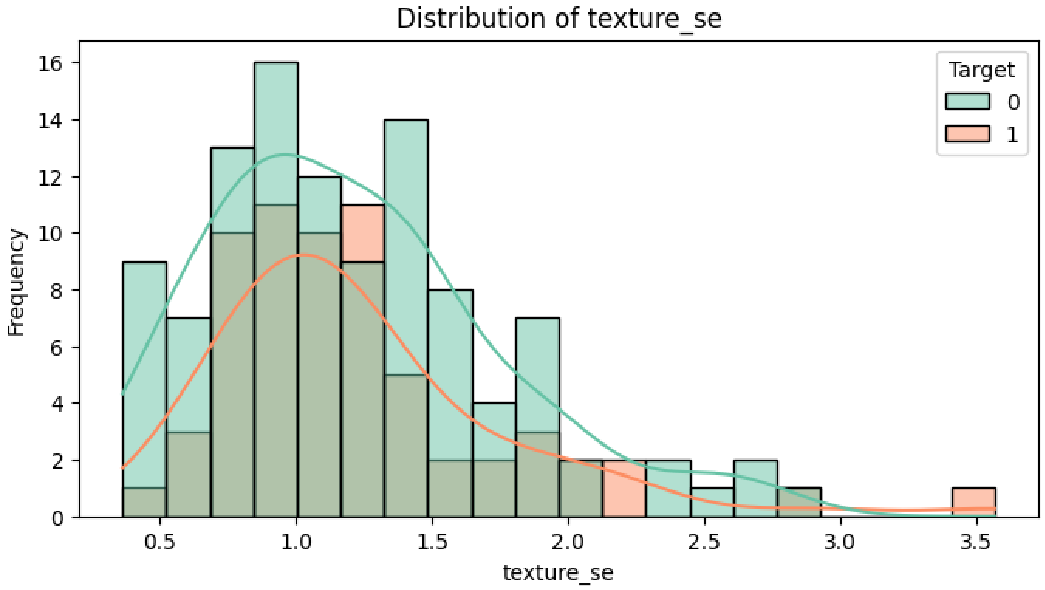


Distribution of smoothness_worst



Distribution of concavity_worst

Analysis of Results



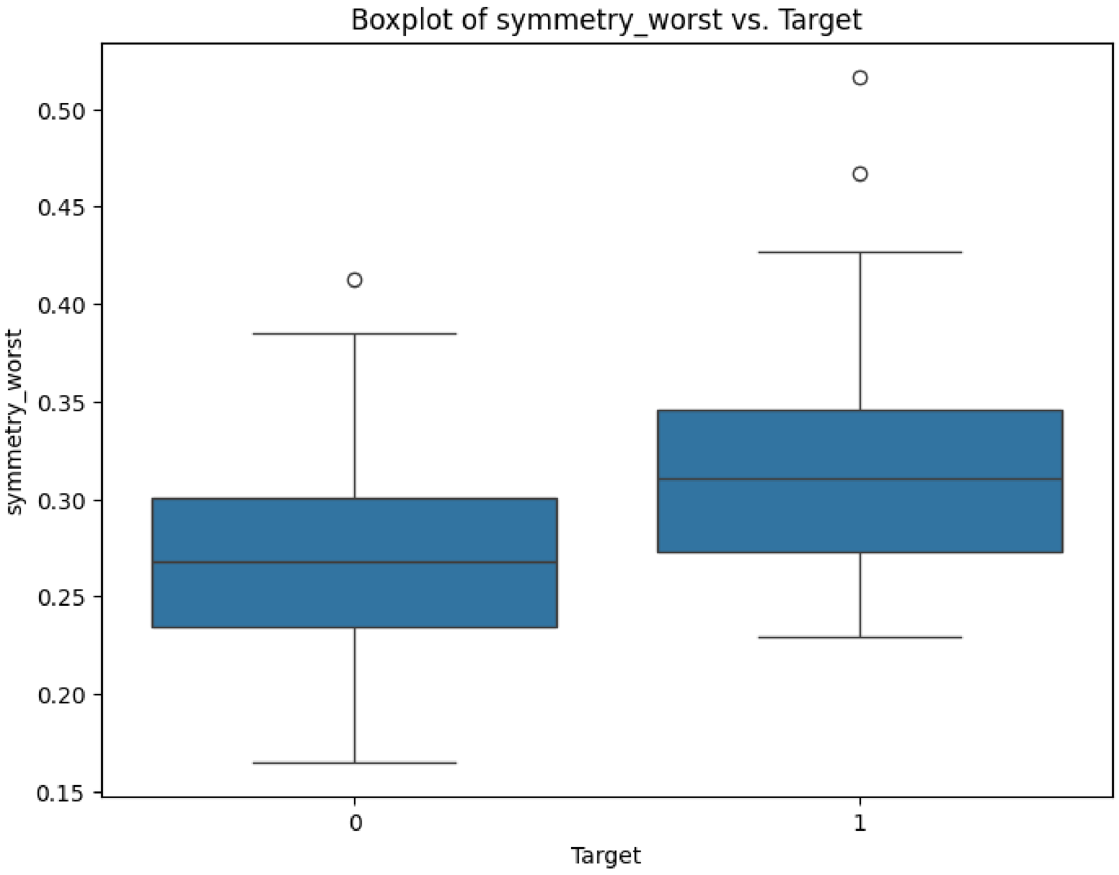
	Ranking	T-Statistic	p-value
symmetry_worse	1	-5.5738	1.8696e-07
texture_worse	2	-7.0766	7.5496e-11
smoothness_worst	3	-5.4987	1.7750e-07
concavity_worst	4	-12.5671	3.1665e-21
texture_se	5	-0.3590	7.2022e-01

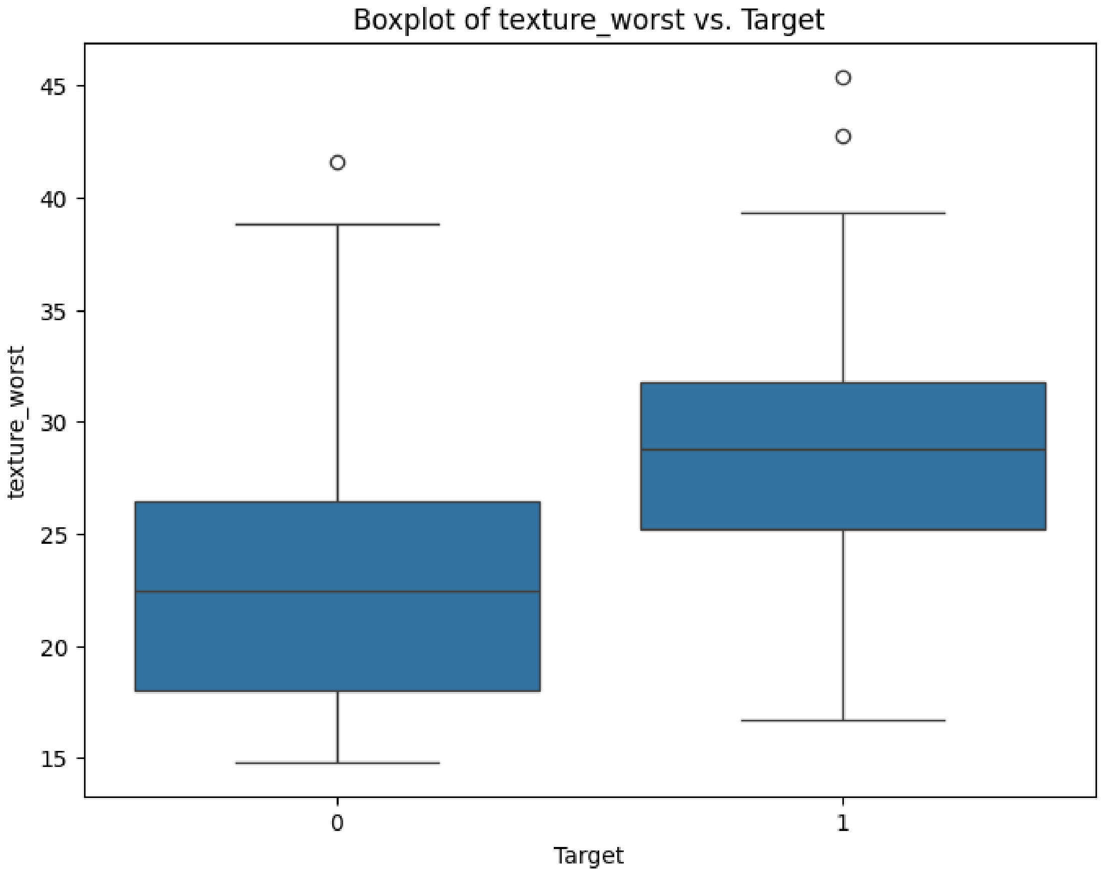
T-test: a large absolute value indicates a strong difference in means between benign and malignant classes.
p-value: if < 0.05 suggests this difference is statistically significant.

symmetry_worst: highly relevant feature for classification.
texture_worst: highly relevant feature for classification.
smoothness_worst: highly relevant feature for classification.
concavity_worst: one of the most important features for classification.
texture_se: is not a significant feature for classification.

Analysis of Results

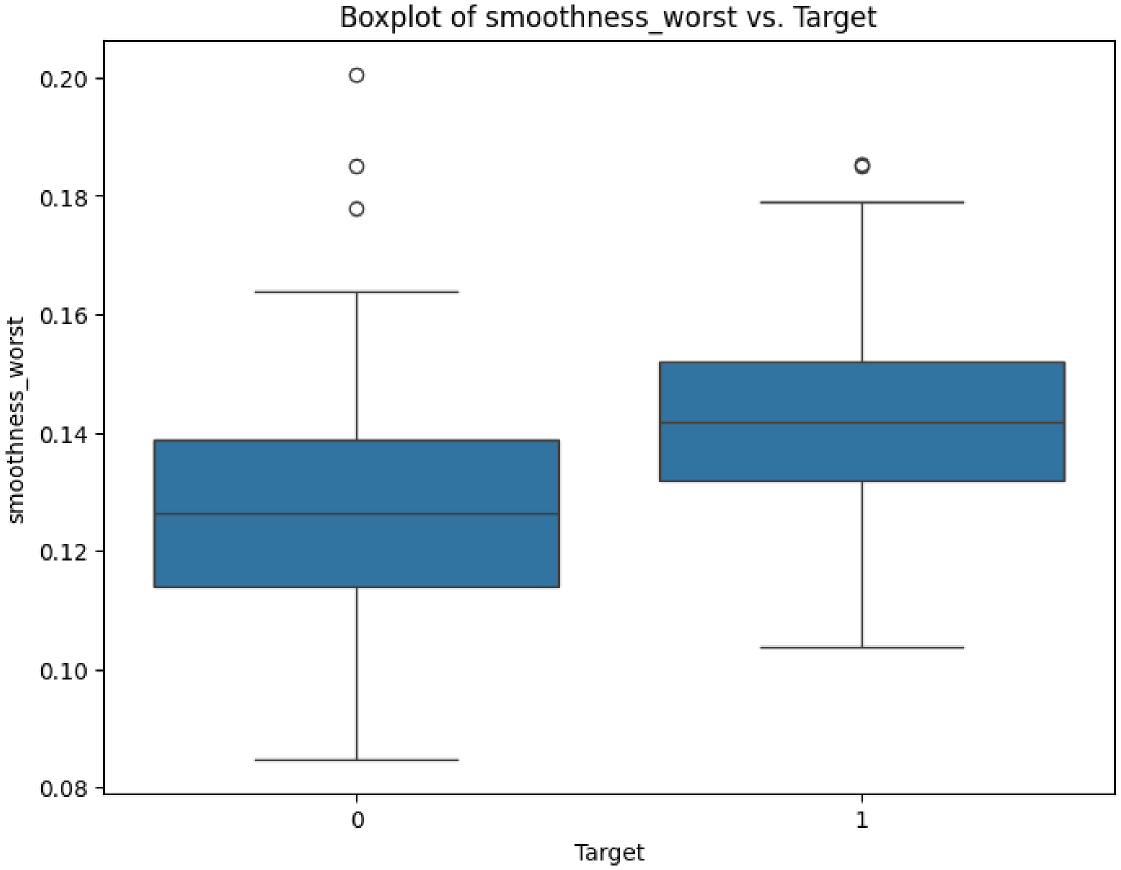
Relationship between top features and target variable

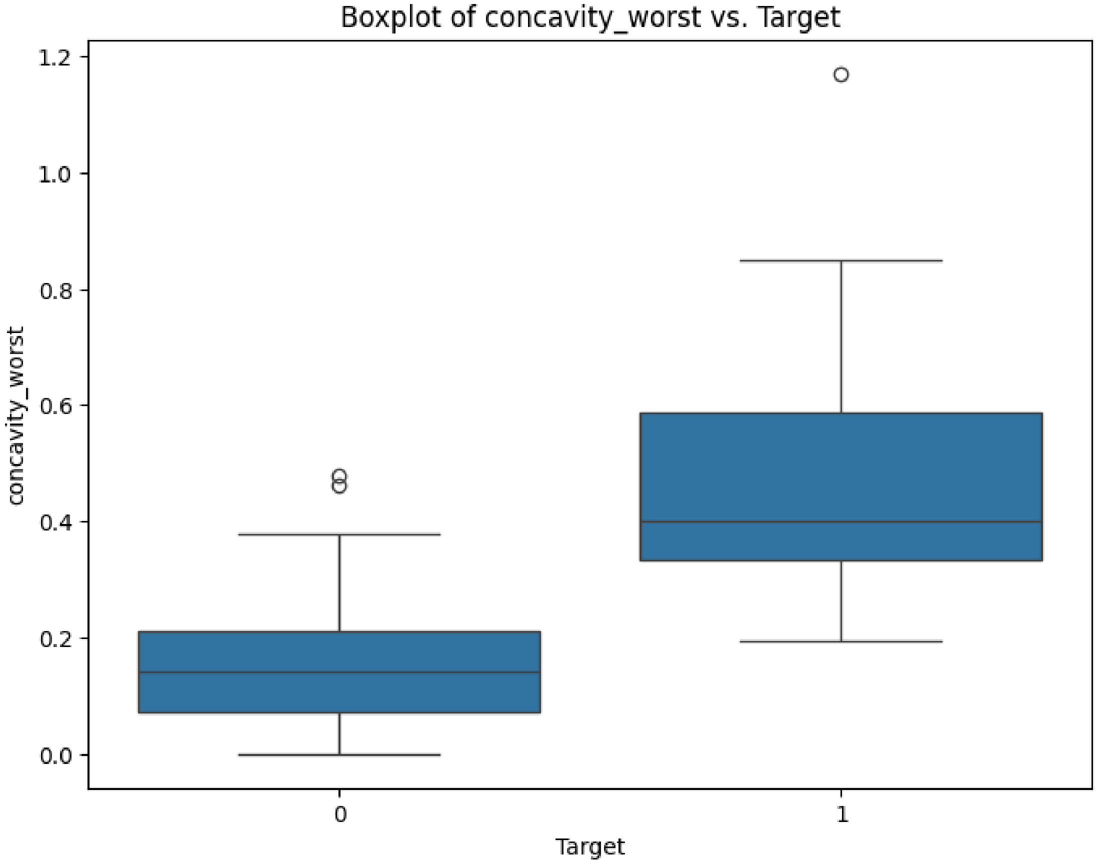




Analysis of Results

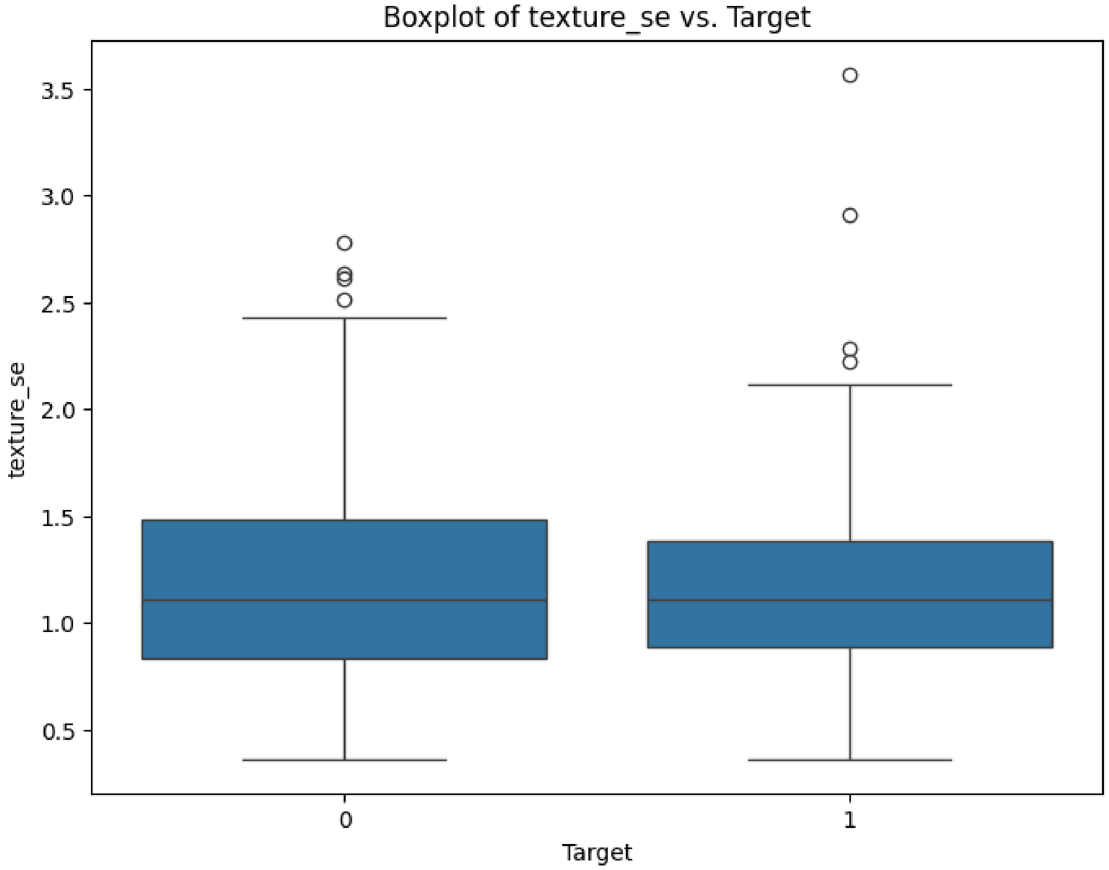
Relationship between top features and target variable





Analysis of Results

Relationship between top features and target variable

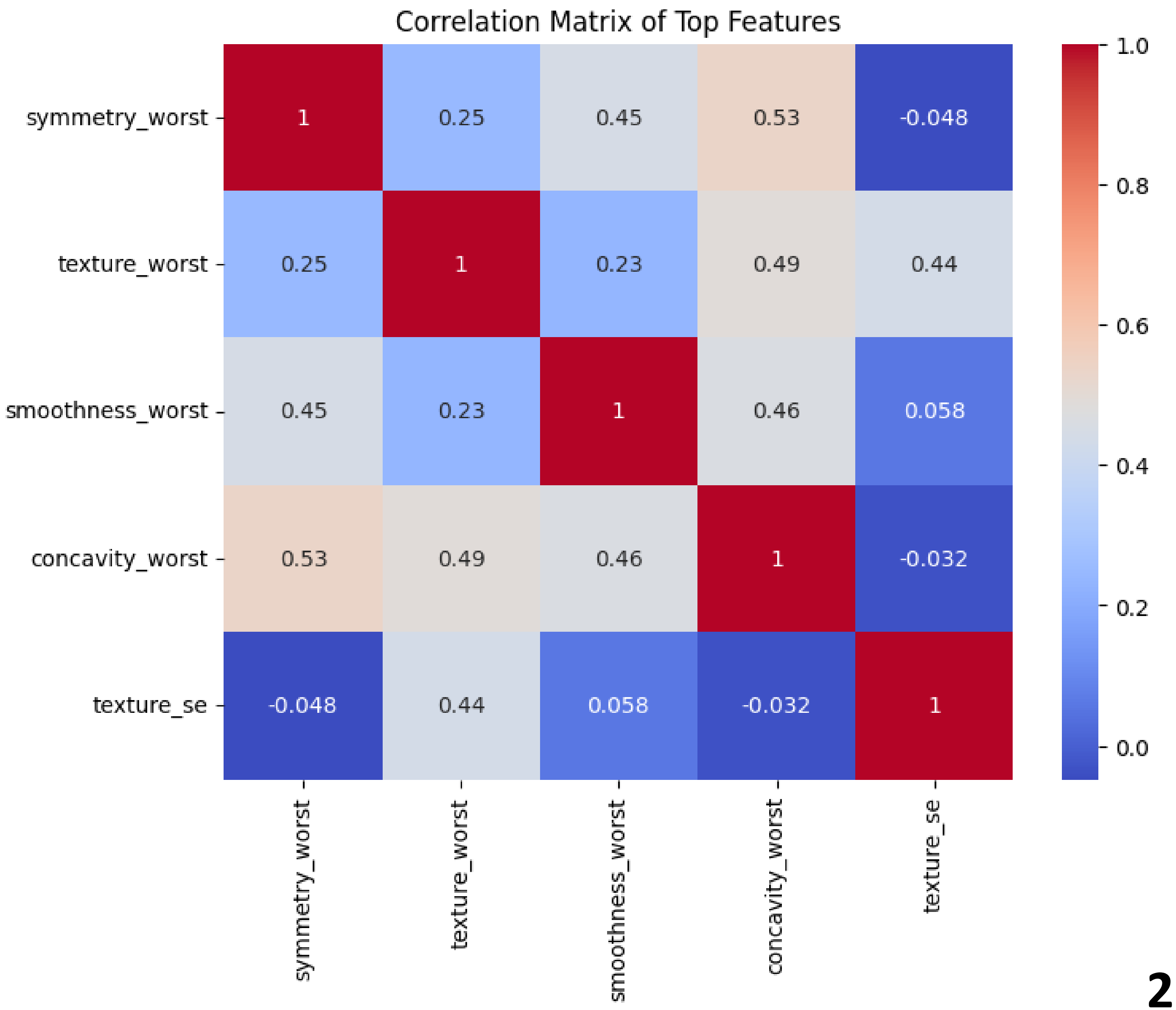


It can be seen that for each relevant feature, the malignant tumour (1) has the high mean value for that feature. This is not the case for texture_se, which is not a relevant feature.

Analysis of Results

Correlation analysis among top features

From the correlation matrix of the top features, the most correlated (<0.5) are **symmetry_worst** and **concavity_worst**, which are relevant features for classification. However, the correlation is not very strong.

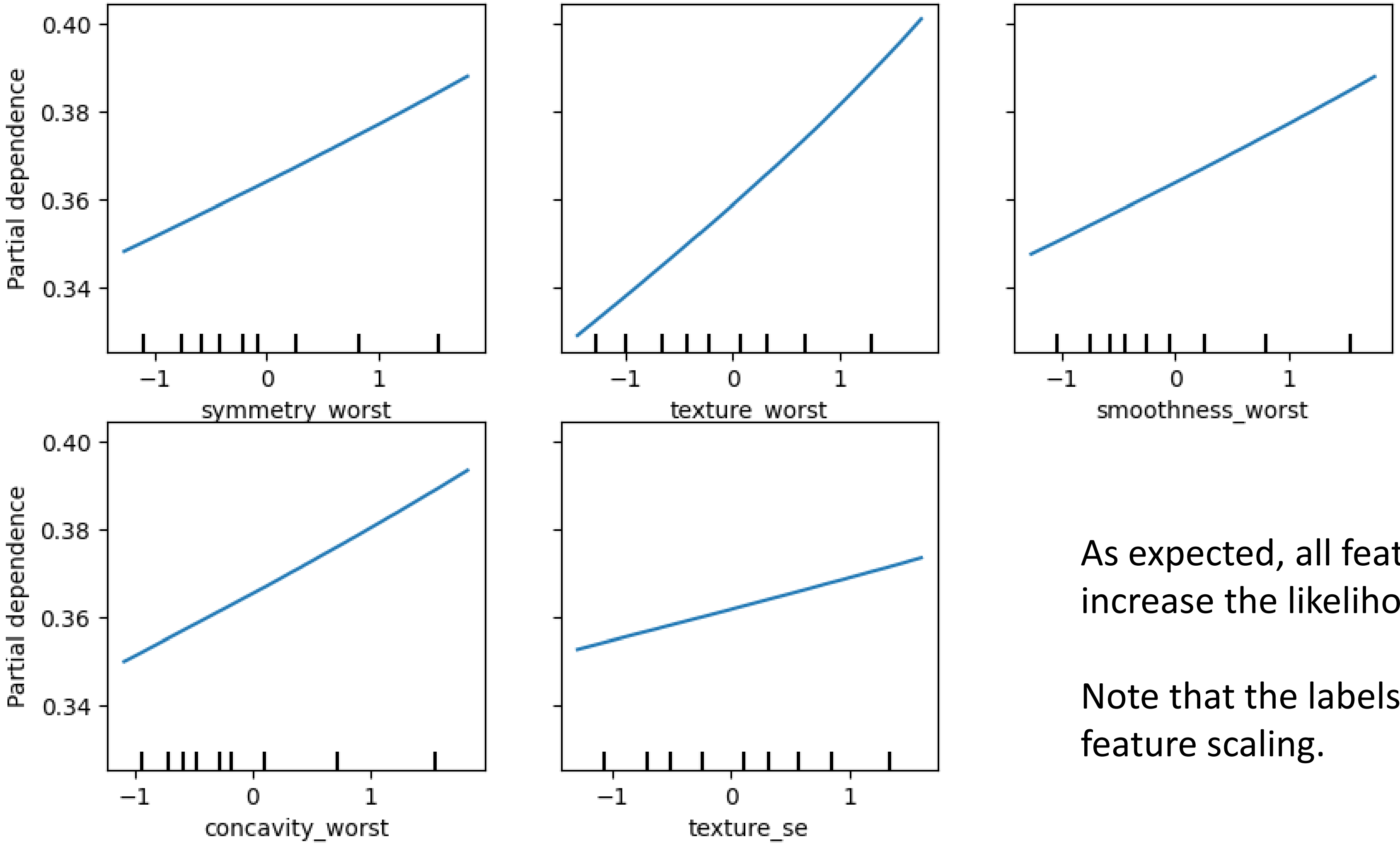


Correlation Matrix of Top Features

	symmetry_worst	texture_worst	smoothness_worst	concavity_worst	texture_se
symmetry_worst	1	0.25	0.45	0.53	-0.048
texture_worst	0.25	1	0.23	0.49	0.44
smoothness_worst	0.45	0.23	1	0.46	0.058
concavity_worst	0.53	0.49	0.46	1	-0.032
texture_se	-0.048	0.44	0.058	-0.032	1

Analysis of Results

Partial Dependency Plot (PDP)



As expected, all features (except texture_se) increase the likelihood of a malignant tumour.

Note that the labels range is -1 to 1 because of feature scaling.

Conclusion

Breast Tumour Classification Analysis

Objective: Predict whether a tumour is benign or malignant based on feature measurements.

Approach:

- Data exploration and analysis.
- Evaluation of various ML models.
- Best model: SVM with 97.6% test accuracy.
    - Misclassifications: 0 benign as malignant, 3 malignant as benign.
    - F1 Scores: Class 0 (benign) - 98.6%, Class 1 (malignant) - 97.6%.

Key Features Influencing Malignant Tumour Classification:

- **symmetry_worst**: Tumour shape irregularity.
- **texture_worst**: Pixel intensity variation.
- **smoothness_worst**: Surface deviation from sphericity.
- **concavity_worst**: Areas of inward curvature.

Conclusion

Further improvements

- Wait for the **XGBoost** bug to be fixed and add this model to the analysis.
- Change the **scaling method** to check if the model performance improves.
- Improve the SVM model by removing some **irrelevant features**.
- Combine different models in an **ensemble learning** algorithm to achieve better recall.
- Implement some other feature explanation such as **LIME** or **SHAP** methods.
- Build a **deep learning** model.