

**BREAST CANCER CLASSIFICATION**

**Andrea Tonello**

**Leonardo Micheluz**

**Lorenzo Cusin**

**Matteo Caiola**

**Introduction**

Breast cancer is one of the most common types of cancer among women worldwide, making early and accurate diagnosis crucial for effective treatment and improved patient outcomes. With the advancement of machine learning techniques, predictive models can significantly aid in the early detection and classification of breast cancer, distinguishing between benign and malignant tumours based on patient data.

The primary objective of this project is to develop and evaluate various machine learning models to classify breast cancer tumours as either benign or malignant. By leveraging a dataset containing features extracted from breast cancer biopsy samples, we will build, train, and compare the performance of different models, including Logistic Regression, Random Forest, Support Vector Machine, and others.

**Dataset**  
The dataset we will use is the well-known Breast Cancer Wisconsin (Diagnostic) Dataset, with 569 records. This dataset includes features computed from digitized images of fine needle aspirate of breast masses. The features describe characteristics of the cell nuclei present in the image, such as:

1. Radius (mean of distances from centre to points on the perimeter)
2. Texture (standard deviation of gray-scale values)
3. Perimeter
4. Area
5. Smoothness (local variation in radius lengths)
6. Compactness (perimeter² / area - 1.0)
7. Concavity (severity of concave portions of the contour)
8. Concave points (number of concave portions of the contour)

The target variable is a binary classification indicating whether the tumour is benign (label = 0) or malignant (label = 1).

**Methodology**

* Data Exploration:
  + Load and explore the dataset to understand its structure and characteristics.
  + Perform an exploration of the features by plotting boxplots and scatterplots
* Model Training and Evaluation:
  + - Split the dataset into training and testing sets.
    - Train each model using the training set.
    - Evaluate the performance of each model on the testing set using appropriate metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.
* Model Comparison and Selection:
  + - Compare the performance metrics of all models.
    - Select the best-performing model based on a balanced consideration of various metrics.

**Data Exploration**

Upon inspecting the dataset, we found that there were no missing values. This ensured that we could proceed with the analysis and modelling without the need for imputation or deletion of records due to missing data.

We then computed the correlation matrix to identify relationships between the features.

Immagine che contiene testo, schermata, modello

Descrizione generata automaticamente

Matrix of correlation

This matrix revealed that there were a few variables with significant correlations, such as radius\_mean and area\_worst, perimeter\_worst and concave points\_mean. It is important to state that not every correlation found in this way is necessarily important: for instance, the correlation found between the perimeter\_worst and perimeter\_mean is trivial and therefore not important.

To visualize the distribution of each feature with respect to the diagnosis (benign vs. malignant), we decided to plot boxplots. These helped us to highlight differences in the feature distributions between benign and malignant cases, allowing us to see which features might be more informative for classification.

Immagine che contiene testo, Parallelo, schermata, linea

Descrizione generata automaticamente

Boxplots of our continuous variables

We finally created scatter plots for pairs of features that were most likely correlated (concavity\_mean vs perimeter\_worst, concave points\_worst vs area\_worst and concave points\_mean vs radius\_mean).

Immagine che contiene schermata, testo, diagramma

Descrizione generata automaticamenteImmagine che contiene schermata, testo, Diagramma

Descrizione generata automaticamenteImmagine che contiene schermata, testo

Descrizione generata automaticamente

Scatterplots of different variables

These scatter plots helped us to visually assess the relationships between features and understand how they might jointly contribute to distinguishing between benign and malignant tumours.

After having a better understanding of our dataset we passed to the next phase of the project, the model evaluation.

**Model training and evaluation**

Among all models available, we chose to start with the Logistic Regression. We decided to use as a dependent variable (y) the diagnosis feature, while for the independent variable (X) we chose to use the radius\_mean and concave points\_mean features. Before starting with the training of the model, we need to know whether the sample size of our dataset is adequate. After evaluating the prevalence and applying the rule of 10 events per variable, we found out that our sample size is more than adequate.

We then proceeded to split the dataset into the train set and test set (75% for the train, 25% for the test) and initialized our X and y.

Subsequently, we created and fitted our logistic regression model, and evaluated the confusion matrix (90 TP, 5 FN, 3 FP, 45 TN)

We also evaluated specific metrics such as accuracy, precision, recall and f1-score by using the classification report package, obtaining pretty good values.

An important value that was needed to be evaluated was the Odds Ratio, which showed that for each Logistic Regression coefficient the OR value was higher than 1, implying that a raise on independent variables values increases the probability of event of interest.

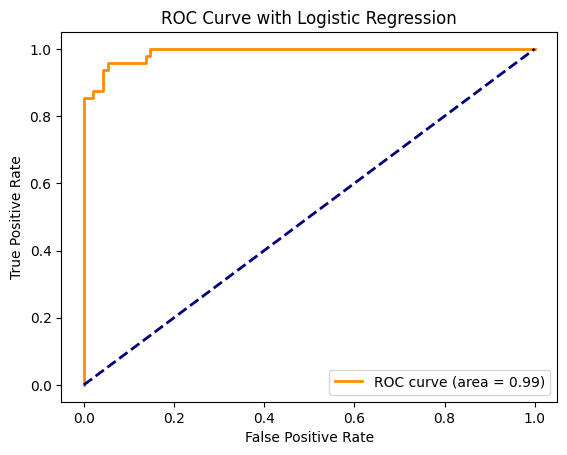
After evaluating the discrimination of the model, we needed to evaluate its calibration. Therefore, we decided to perform the Platt-Scaling method in order to calibrate the model.

Immagine che contiene testo, linea, diagramma, Diagramma

Descrizione generata automaticamente

Model predicted probabilities before and after calibration.

With this graph we show how big are the difference between before and after the calibration regarding the predicted probabilities. We finally evaluate the model by using the AUC-ROC graph to have a better understanding of our model performance.



ROC curve for the Logistic Regression model

As we can see, the Logistic Regression model demonstrated exceptional performance with an AUC-ROC score of 0.99. This score indicates an excellent ability of the model to distinguish between benign and malignant tumours.

**Other models**

In addition to the Logistic Regression model, we implemented and trained several other machine learning models to classify breast cancer tumours. These models include:

* Decision Tree: a non-linear model that splits the data into subsets based on the value of input features, creating a tree-like structure of decisions. This model performed pretty good, obtaining an accuracy of 94%, and high values of precision, recall and f1-score as well.

Immagine che contiene diagramma

Descrizione generata automaticamente

Decision Tree

* Random Forest: an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes of the individual trees. This model performed like the decision tree, obtaining a higher value of accuracy (97%), and higher values of sensitivity and specificity as well.

Immagine che contiene testo, diagramma, schermata, linea

Descrizione generata automaticamente

Random Forest

* K-Nearest Neighbours: a simple, instance-based learning algorithm that classifies a sample based on the majority class of its k nearest neighbours.

Immagine che contiene testo, schermata, diagramma, Diagramma

Descrizione generata automaticamente

K-NN

Thanks to this plot we managed to determine the number of neighbours that gave us the highest value of accuracy, which is 21. By running the K-NN algorithm with 21 neighbours we obtained the maximum value of accuracy, which was 95%, a good value. By evaluating other metrics, we found out that also precision and recall were pretty good.

* Support Vector Machine: a powerful classification algorithm that finds the hyperplane that best separates the classes in the feature space. This model had an adequate accuracy (94%) and high precision and recall as well.

Immagine che contiene schermata, testo, diagramma, Diagramma

Descrizione generata automaticamente

SVM

* Kernel Support Vector Machine: an extension of SVM that uses kernel functions to handle non-linear decision boundaries by mapping the input features into higher-dimensional space. This model was similar to the SVM, since all metrics were pretty much identical.

**Model Comparison and Selection**

Finally, we decided to evaluate the AUC score of each model and plot every ROC curve in order to determine which model performed the best.

Immagine che contiene testo, linea, schermata, Diagramma

Descrizione generata automaticamente

ROC curve for all models

As we can see, we found that the Random Forest and Logistic Regression models achieved the highest AUC-ROC scores. This indicates that these models have the best capability to distinguish between benign and malignant tumours among the tested algorithms. Moving forward, these models can be further refined and possibly integrated into clinical workflows to assist healthcare professionals in making informed decisions.