**Héloïse ROMEO, Charles DE PUYBAUDET, Alexis DENNEULIN, Natalia GERARD**

**DIA 3**

**MACHINE LEARNING : STAGE 1**

*dataset breast cancer :* [*https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset*](https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset)

Business Scope

* **Context**

Breast cancer is a major public health issue, with over 61,000 new cases diagnosed annually in France. Current screening methods, such as mammography, have high false-positive rates, leading to increased costs and patient anxiety. The use of artificial intelligence (AI) to improve diagnostic accuracy could revolutionize this approach by reducing false positives and speeding up the management of malignant tumors.

* **Project Objectives**

Our goal is to propose a machine learning model to predict whether a tumor is benign or malignant based on medical data. We aim to develop a solution that minimizes false positives, thus avoiding wasted time and unnecessary exams in a real medical context.

The diagnosis would be faster, which could be life-saving for urgent interventions.

* **Proposed Solution**

We are developing a machine learning model using algorithms such as Support Vector Machines (SVM) and Decision Trees, based on a dataset containing tumor characteristics like size, texture, and concavity. The model aims to classify tumors more reliably and quickly than traditional methods.

* **Benefits**

*This project provides three main benefits:*

* More accurate and faster diagnoses, reducing human error.
* Fewer unnecessary exams and treatments due to more efficient diagnoses.
* Reduced anxiety and delays in patient care.
* **Target Audience**

*The target audience is the medical community as a whole :*

* Doctors and radiologists for faster and more reliable diagnoses.
* Healthcare facilities seeking to improve the quality and efficiency of screening.
* Health insurers looking to reduce the costs of unnecessary treatments.

Problem formalisation and methods

**-Roadmap**

**1. Data Exploration and Quality Analysis**

**Objective: Understand and analyse data before any transformation.**

**Loading data**: Load and preview data using .head() and .info().

**Descriptive analysis**: Descriptive statistics of data with .describe() to understand distributions.

**Visual analysis**:Histograms by class (malignant/benign) to observe differences between features.

**Correlation identification**: Correlation matrix to identify relationships between variables and reduce multicollinearity.

**2. Data Preprocessing and Management of Class Imbalance**

**Objective: Prepare the data for the models and manage any imbalances.**

**Data cleaning**: Removal of unnecessary columns or columns with a low correlation with the target variable

**Encoding of the target variable**: Transformation of the diagnosis variable into numerical values (0 for benign, 1 for malignant).

**Class imbalance management**: Check class proportions

**3. Model Selection and Performance Optimization**

**Objective: Select suitable models, optimise them and compare their performance.**

**Models tested**: **Random Forest** and **Logistic Regression** integrated in a pipeline (normalisation with StandardScaler).

**Cross-validation**: Use of cross-validation (e.g., cross\_val\_score) to assess model robustness and monitor overlearning.

**Optimisation of hyperparameters**: Implementation of **GridSearchCV** to optimise Random Forest hyperparameters.

**Model comparison**: Analysis of the performance of each model using metrics (precision, recall, F1-score) and ROC curves.

**4. Testing and Evaluation of Results**

**Objective: Test the models on a test set and analyse their performance.**

**Evaluation on the test set**: Calculation of precision, confusion matrix and classification ratio for Random Forest and Logistic Regression.

**ROC curves and AUC score**: Plot ROC curves to visualize model performance on the test set. Comparison of AUC scores (Random Forest vs Logistic Regression).

**Interpretation of results**

**-Algorithm Description**

The main objective of this project is to develop a machine learning model capable of classifying tumours as benign or malignant on the basis of tumour cell characteristics. Several classification algorithms were tested and compared in order to respond effectively to this problem.

**Logistic Regression**

Logistic regression is a supervised classification algorithm that models the relationship between features and a binary target variable (diagnosis: benign or malignant). It calculates a probability using a sigmoid function and applies a threshold (usually 0.5) to assign a class to an observation.

It has certain advantages such as its interpretability:The β coefficients give a direct indication of the impact of characteristics on the probability of classification. And its simplicity: Ideal for linear relationships between characteristics and the target.

However, it islimited to linear relationships and may therefore perform less well on datasets with complex or non-linear relationships.

**Random Forest**

Random Forest is an ensemble algorithm based on a collection of decision trees. It combines the predictions of several trees to improve accuracy and reduce the risk of overlearning.

Several decision trees are constructed from subsamples of the dataset. Each tree is trained independently on part of the data (bootstrap sampling).

During prediction, each tree votes for a class, and the majority class is retained as the final result (classification by majority vote).

The trees use a random selection of features at each division to reinforce the diversity between them.

It has several key hyperparameters to define:

**n\_estimators**: Number of trees in the forest.

**max\_depth**: Maximum depth of trees.

**min\_samples\_split**: Minimum number of samples required to split a node.

**min\_samples\_leaf**: Minimum number of samples in a leaf.

This is why it is important to choose these parameters carefully.

To improve the performance of Random Forest, a **GridSearchCV** search was carried out to find the best combinations of hyperparameters (such as the number of trees and the maximum depth). This ensures that the model is well fitted without the risk of overlearning.

This algorithm is robust to noisy data and reduces the risk of overlearning thanks to the diversity of the trees.

However, it is less interpretable than logistic regression and requires more computational resources.

**Cross-validation and optimisation of hyperparameters**

In addition, to guarantee the robustness of the models and avoid overtraining, **cross-validation** was used, dividing the data into several subsets (folds). This makes it possible to check the model's performance on different partitions of the data and avoid biased learning.

**Why did we choose these algorithms for our dataset?**

For the problem of binary classification of breast cancer (benign or malignant), we chose **logistic regression** and **Random Forest** because they are suitable for tabular datasets with well-defined numerical characteristics, as in our case. Logistic regression is ideal for establishing a baseline because of its simplicity, interpretability and ability to model linear relationships. Random Forest, on the other hand, is a robust model capable of capturing non-linear and complex relationships between features, while still being able to handle moderate-sized datasets. These two models offer a good balance between efficiency, complexity and adaptability to the data available.

**-Limitations**

**Limitations of the Breast Cancer dataset**

Although the dataset used for tumour classification is well suited to a preliminary machine learning analysis, it has several limitations that can affect the performance and generalisation of the models. Here are the main limitations identified:

**-Dataset size**

The dataset contains a total of **569 observations**, which is relatively small for training machine learning models. A small dataset size has the following disadvantages:

**Risk of overlearning**: Complex models, such as Random Forest, may memorise the training data instead of generalising correctly.

**High variance**: With a small sample size, model performance can vary widely depending on the split between training and test sets.

**Difficulty capturing rare patterns**: Subtle or rare relationships in the data may not be well learned by the models.

**-Class imbalance**

The dataset shows an **imbalance between classes**:

62.7% of tumours are benign.

37.3% of tumours are malignant.

This imbalance can lead models to favour the majority class (benign) to the detriment of detecting malignant tumours, which is particularly critical in a medical context.

**-No additional data**

The dataset contains only numerical characteristics extracted from the medical images, with no additional metadata (for example: age, family history, patient lifestyle). This information could enrich analyses and improve model performance.

The data comes from a single source or specific population, which may limit its representativeness for other geographical or demographic contexts.

Finally,the dataset does not contain any temporal or longitudinal tracking of patients. This makes it impossible to assess the evolution of tumours or to explore temporal dynamics, which could be crucial for a more accurate diagnosis or prognosis.

Methodology

**-Data description and exploration**

We started by loading and exploring the structure of the data. After an initial examination, we noted that there were no missing values in the dataset, and all columns were of a numeric type, except for the target variable, diagnosis, which indicated whether a tumor was benign or malignant. This simplified our preprocessing as we didn’t need to handle any missing values.

To better understand how each feature was distributed based on the target variable diagnosis, we visualized each feature using histograms and box plots. This helped us identify features with significant distribution differences between benign and malignant classes, which are likely useful for classification. Features that showed no meaningful variation between the classes were removed, as they wouldn’t contribute to distinguishing between benign and malignant cases.We confirmed that there were no missing values in the dataset. This is a critical step in data preprocessing, as missing values can introduce bias or errors in model training. Since all values were present, we were able to proceed without imputing or handling any missing data.We next checked the balance of the target classes by calculating the percentage of each class (benign and malignant).

An imbalanced dataset can affect model performance, as models tend to favor the majority class. In this case, the classes were reasonably balanced, so we proceeded without needing to implement any resampling techniques, such as oversampling the minority class or undersampling the majority class.We analyzed each feature to identify potential outliers using box plots.

Outliers can impact the performance of some models, so it was important to address them. In this case, we decided to keep the outliers, as they might represent significant observations that could help differentiate between the benign and malignant diagnoses.

To further refine the feature selection, we calculated the correlation matrix to identify any pairs of highly correlated features. High correlations can introduce redundancy in the data and lead to overfitting in certain models. By analyzing the correlation matrix, we were able to identify and exclude some highly correlated features, reducing dimensionality while preserving relevant information.

Finally, we applied Principal Component Analysis (PCA) for dimensionality reduction. PCA helped us capture the majority of variance in the dataset with a reduced number of principal components, which contributed to model efficiency and performance. This step was especially valuable in minimizing the risk of overfitting and reducing computation time, while retaining the essential information required for accurate classification.

**-Methodology for data Splitting for train/test**

**Function used**: Scikit-learn ‘s train\_test\_split function was used to divide the dataset randomly.

**Proportion of data**:

**67%** of the data was used to train the model (train).

**33%** of the data was reserved for final evaluation on unseen data (test).

**Specific parameters**:

**test\_size=0.33**: Defines that 33% of the data is used for the test set.

**random\_state=42**: Used to make the split reproducible by setting a random seed. This ensures that the same split is performed each time.

**Evaluation of generalisation**:

The test set simulates unseen real data, on which the model must be able to predict correctly.

Correct separation helps to detect possible overfitting or underfitting problems.

**Classic ratio**:

The 67/33 ratio is a common choice that offers a good compromise between having enough data to train the model and reserving enough data for meaningful evaluation.

**-Algorithm Implementation and Hyperparameters**

In this project, two main algorithms were implemented to solve the binary classification problem (benign or malignant): **Logistic Regression** and **Random Forest**. These models were integrated into pipelines to include data pre-processing and normalization steps, and their performance was evaluated using standard metrics. The following is a description of the implementations and hyperparameters used.

**Logistic Regression**

**Implementation**

Logistic regression was used as the basic model to establish an initial performance comparison. A pipeline was created by combining a data normalisation step with StandardScaler and the LogisticRegression model to ensure that all features were on the same scale before training.

**Hyperparameters used**

* **penalty**: L2 (default) - adds a regularization to prevent overfitting.
* **solver**: lbfgs (default) - a solver optimized for medium-sized datasets.
* **C**: 1.0 - controls the strength of the regularization (default).

These hyperparameters were not explicitly adjusted in this project, but their default settings gave good baseline performance.

**Random Forest**

**Implementation**

Random Forest was used as the main model because of its robustness and ability to capture complex, non-linear relationships in the data. A similar pipeline was used, combining StandardScaler and RandomForestClassifier. Performance was evaluated before and after hyperparameter optimisation.

**Hyperparameters before optimisation (default)**

* **n\_estimators**: 100 (number of trees in the forest).
* **max\_depth**: None (no limit on tree depth).
* **min\_samples\_split**: 2 (minimum number of samples to split a node).
* **min\_samples\_leaf**: 1 (minimum number of samples in a leaf).

**Optimising hyperparameters**

To improve performance, a search for hyperparameters was carried out using GridSearchCV. The following parameters were tested:

* **n\_estimators**: [50, 100, 200].
* **max\_depth**: [None, 10, 20, 30].
* **min\_samples\_split**: [2, 5, 10].
* **min\_samples\_leaf**: [1, 2, 4].

The **best set of hyperparameters** found was :

* **n\_estimators**: 50.
* **max\_depth**: None.
* **min\_samples\_split**: 2.
* **min\_samples\_leaf**: 4.

These parameters improved robustness and reduced the risk of overfitting, while maintaining high performance on the test set.

**Model comparison**

**Performance evaluation**

The two models were evaluated using the following metrics:

* **Overall accuracy (Accuracy)**.
* **Classification ratio**: Includes precision, recall and F1-score for each class.
* **ROC and AUC curves**: Used to visualise the models' ability to distinguish between benign and malignant classes.

**Results**

* **Random Forest** slightly outperformed Logistic Regression in terms of classification metrics and AUC score.
* However, **Logistic Regression** remains a simpler and more interpretable model, although it performs less well for complex relationships in the data.

**Conclusion**

The combination of pipelines to normalise the data, the choice of suitable algorithms and the optimisation of hyperparameters has resulted in high-performance, robust models. The search for hyperparameters for Random Forest demonstrated a significant gain in efficiency, while logistic regression provided a rapid and interpretable baseline approach. These approaches offer a good balance between simplicity, performance and adaptability to the dataset used.

Results

**-Metrics**

Especially In the context of cancer detection, it is essential to evaluate the performance of a model using multiple metrics to get a complete picture of its effectiveness. We calculated metrics such as accuracy, confusion matrix, classification report, and possibly ROC-AUC to assess model performance, avoid false positives, correctly identify all actual malignant cases and the f1 score to balances precision and recall, offering a single metric that accounts for both avoiding false positives and capturing true positives, which is vital in medical applications where both are critical.

Globally these metrics show that the model performs very well, achieving high accuracy, precision, and recall for both classes.

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

*Results for Logistic Regression.*

*Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement*

*Results for Random Forest.*

*Une image contenant texte, ligne, Tracé, diagramme

Description générée automatiquement*

*ROC curves for Random Forest and Logistic Regression.*

**-Overfitting**

To prevent overfitting, steps like **PCA** (Principal Component Analysis) and **cross-validation** were used. PCA helps reduce the feature space, lowering the chances of the model capturing noise instead of meaningful patterns. Cross-validation allows us to validate the model on multiple data splits, helping confirm that the model generalizes well rather than memorizing patterns specific to a single training set. In fact, the scores across the five folds are very close to each other, ranging from approximately 0.956 to 0.974. This consistency indicates that the model’s performance is stable and isn’t significantly influenced by the specific subset of data used in each fold.

**-Evaluation**

The model’s performance was evaluated by comparing metrics across training and test datasets, along with cross-validation results. This approach provides a well-rounded view of model accuracy, precision, and recall and helps us identify if the model maintains performance on unseen data (test set) as it did on training data, indicating reliable generalization.

Discussion and conclusion

**Model Performance**:  We compared the Random Forest and logistic regression models, assessing which model better distinguished between malignant and benign cases. The discussion considers metrics like sensitivity (ability to correctly identify malignancies) and specificity (limiting false positives), as these are crucial in medical contexts.

Overall our results are pretty much similar for our 2 models with an accuracy between 0.93 and 0.95 and all precision, recall and f1 score between 0.94 and 0.97.

The model performs well overall, with high accuracy, balanced precision, and recall, particularly critical for cancer detection. However, to further reduce false negatives (missed malignant cases), adjustments like **threshold tuning** or **cost-sensitive learning** could improve recall. Enhancing **feature engineering**, exploring **alternative models** like boosting algorithms, and validating the model on **diverse datasets** could also strengthen its reliability and generalizability in clinical applications, ensuring it’s both accurate and robust across different patient populations.

**Conclusion**

In this initial stage, our model has demonstrated solid performance in classifying malignant and benign cases with high accuracy and balanced precision and recall metrics, which are essential for cancer detection applications. However, opportunities for further refinement and improvement remain, particularly in enhancing recall for malignant cases to minimize false negatives. Additionally, alternative modeling approaches, threshold optimization, and validation across diverse datasets could further boost reliability and robustness.

As we move into Stage 2, our evaluation will focus on building upon these initial results through targeted improvements. This will include exploring more sophisticated algorithms, tuning hyperparameters, and refining our methodology to enhance model performance. We'll critically examine assumptions made in Stage 1, aiming to address any limitations in data or model choice.