MOHAMMED VI POLYTECHNIC UNIVERSITY

Breast Cancer Detection Using Machine Learning

Réalisé par :

Imane YASSIRI

Chaimaa ELALAMI

Encadré par :

Mr. Younes JABRANE

Année universitaire : 2020 - 2021

Filière : Modélisation et Sciences des Données

Remerciements

Nous tenons tout d'abord à remercier Dieu le tout puissant, qui nous a donné la force et la patience d'accomplir ce travail dans les meilleures conditions.

À tous nos enseignants pour leurs efforts qui nous ont guidées et qui ont enrichi nos travaux tout au long de cette année universitaire.

Nous tenons également à exprimer nos vifs remerciements et notre profonde gratitude à notre cher encadrant **Mr. Younes JABRANE** pour son attention, son orientation et son aide pendant la réalisation de ce travail. Il nous a accompagnées en nous guidant pas à pas avec patience et bienveillance, ce qui nous a permis d'avancer et de progresser. Il a également su éveiller en nous la curiosité et l'envie de réaliser et d'accomplir des tâches que nous croyions être inaccomplissables.

Nous tenons aussi à exprimer notre gratitude et nos chaleureux remerciements à nos parents et nos proches pour leurs soutiens. Et finalement, nous remercions tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail.

Table des figures

Table des matières

Résumé

[8] Dans le parcours de soin, la phase de diagnostic est essentielle pour l'orientation du patient et son suivi. Le Machine Learning apporte de nouvelles solutions aux professionnels de santé pour gagner du temps et optimiser le bon diagnostic. Il ouvre de nouvelles perspectives dans le repérage des maladies. Par exemple il peut aider les médecins à détecter plus facilement les anomalies sur les radios des patients.

Le cancer du sein est la seconde cause de décès des femmes à travers le monde. Toute femme a une chance sur huit de développer un cancer du sein durant sa vie. La mammographie est la méthode la plus efficace pour la détection précoce des maladies du sein.

C'est notre cas dans ce projet. Notre travail consiste à détecter, en se basant sur des images d'histologie mammaire, la présence ou non d'anomalie. En utilisant les techniques de traitement d'images, les algorithmes de classification, nous allons construire des modèles et évaluer leur performance à l'aide de plusieurs métriques.

Abstract

The diagnostic phase is essential for patient orientation and follow-up. Machine Learning brings new solutions to healthcare professionals to help them save time and optimize the correct diagnosis. It opens new perspectives in the detection of diseases. For example, it can help doctors to easily detect abnormalities on patient x-rays.

Breast cancer is the second leading cause of death for women around the world. Every woman has a one in eight chance of developing breast cancer in her lifetime. Mammography is the most effective method for the early detection of breast disease.

In this project, our mission is to detect, based on breast histology images, whether or not an abnormality is present. Using image processing techniques, and classification algorithms, we will build models and evaluate their performance using several metrics.

Liste des abréviations

RCGC :Registre des cancers du Grand Casablanca

IRM : Imagerie par résonance magnétique

ML : Machine learning

KDD: Knowledge Discovery in Database

KNN : K-nearest neighbors

SVM : Support vector machine

IA : Intelligence artificielle

ROC : Receiver operating characteristic curve

TN : True negative

FN : False negative

FP : False positive

TP : True positive

IDC : Invasive ductal carcinoma

DT : Decision tree

LR : Logistic Regression

RF : Random forest

ROC : Receiver operating characteristic curve

Introduction générale

Plus que toute autre maladie, le cancer fait peur. En effet, il est la première cause de mortalité, et il poursuit sa propagation dans le monde avec 18,1 millions de nouveaux cas et 9,6 millions de décès enregistrés en 2018. Au Maroc, plus de 48.000 nouveaux cas de cancer ont été diagnostiqués en 2019. 65% des cancers diagnostiqués touchent les femmes contre 35% pour les hommes.

[2] Le cancer du sein en tête de liste chez les femmes. En fait, il est l'un des cancers les plus fréquent et répandu. Dans le cas des retards de diagnostic, ce cancer cause un nombre important de décès. Au Maroc, le cancer du sein constitue 35.8% des cancers chez les femmes. En 2004, ce chiffre était de 5.465. Les femmes âgées entre 45 et 49 ans sont les plus touchées par ce type de cancer. En 2019, 10.414 nouveaux cas de cancer du sein ont été diagnostiqués dans le Royaume. En 2030, ce chiffre passera à 16.018, selon les prévisions du Registre des cancers du Grand Casablanca (RCGC).

[3] *La Fondation Lalla Salma* est au service de lutte contre le cancer au Maroc. Cette fondation s'est tracée pour objectif de mettre en place un dispositif national de lutte contre le cancer qui bénéficie des meilleures pratiques dans le domaine, en mettant en œuvre une stratégie adaptée aux spécificités du pays.
Créée à l'initiative de son Altesse Royale la Princesse Lalla Salma, la Fondation Lalla Salma de lutte contre le cancer œuvre depuis 2005 à améliorer la prise en charge des patients, à encourager les actions de prévention et à faire de la lutte contre le cancer une priorité de santé publique au Maroc.

Ce présent rapport est composé de trois chapitres :

- Le premier chapitre consiste à donner un aperçu général sur le cancer du sein.
- Le deuxième chapitre présente l'état d'art du Machine Learning, les différents algorithmes utilisés et les métriques d'évaluation de performance de ces algorithmes.
- Le troisième chapitre consiste à présenter le jeu de données, les étapes du prétraitement, les résultats des algorithmes de classification obtenus, et l'évaluation de leur performance.

Chapitre I - *Etat de l'art I : Cancer du sein*

Introduction :

Ce premier chapitre donne un aperçu global sur le cancer du sein dans ses différents contextes (ses facteurs de risques, ses moyens de dépistages, etc...), ainsi qu'une visualisation de la conduite du projet.

1. *Cancer du sein :*

[1] Le diagnostic précoce du cancer du sein peut améliorer considérablement les chances de survie, car il peut favoriser un traitement clinique rapide des patientes. Une classification plus précise des tumeurs bénignes peut empêcher les patientes de subir des traitements inutiles. Ainsi, le diagnostic correct du cancer du sein et la classification en groupes malins ou bénins font l'objet de nombreuses recherches.

1.1. Anatomie du sein :

[1] Le sein est composé de graisse, de tissu conjonctif, de glandes et de canaux et comprend ce qui suit :

- Les ligaments sont des bandes de tissu conjonctif soutenant les seins et traversant la peau pour joindre les muscles.
- Les lobules sont les glandes qui produisent le lait (chaque sein a de 15 à 25 lobules).
- Les canaux acheminent le lait des lobules au mamelon.
- Le mamelon, situé au centre de l'aréole, se compose de fibres musculaires.
- L'aréole, une surface ronde, entoure le mamelon.

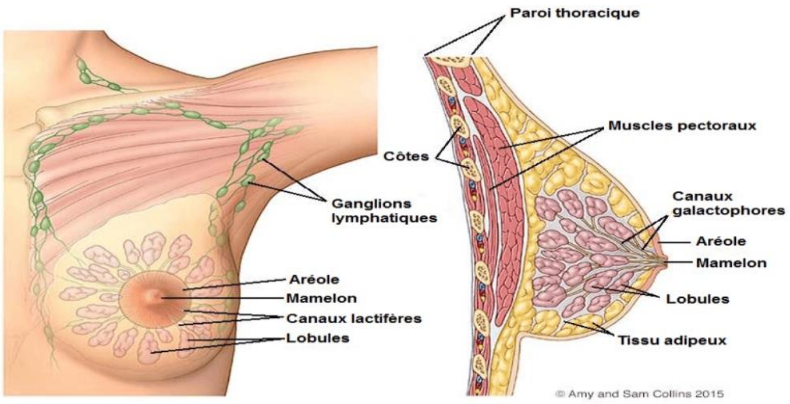


Figure 1: Anatomie du sein

1.2. Cancers du sein suivant leur localisation :

[1] Le sein est composé de nombreux lobes glandulaires (eux-mêmes constitués de plusieurs lobules), structures de production du lait. Chacun se poursuit par un canal lactifère, qui l'amène jusqu'au mamelon. Les lobes glandulaires sont entourés de tissu adipeux (graisse), ainsi que de vaisseaux sanguins et lymphatiques.
Les vaisseaux lymphatiques conduisent la lymphe au niveau des ganglions axillaires, situés sous le bras. Ces ganglions, sortes de réservoirs de cellules immunitaires, peuvent parfois être atteints par les cellules tumorales.
Il existe plusieurs cancers du sein différant selon leur localisation et leur extension. Les trois formes les plus fréquemment rencontrées sont les suivantes :

- Les carcinomes **in situ** : les cellules cancéreuses restent dans les canaux et les lobules, et n'ont pas diffusé dans les tissus environnants.

- Les carcinomes **infiltrants** : ici, les cellules cancéreuses ont envahi les tissus entourant les canaux et lobules. Si ce type de cancer n'est pas pris en charge à temps, il conduit à la formation de métastases dans les ganglions axillaires et le reste du corps.

- Les carcinomes **inflammatoires** : ils se situent en surface, au niveau de la peau.

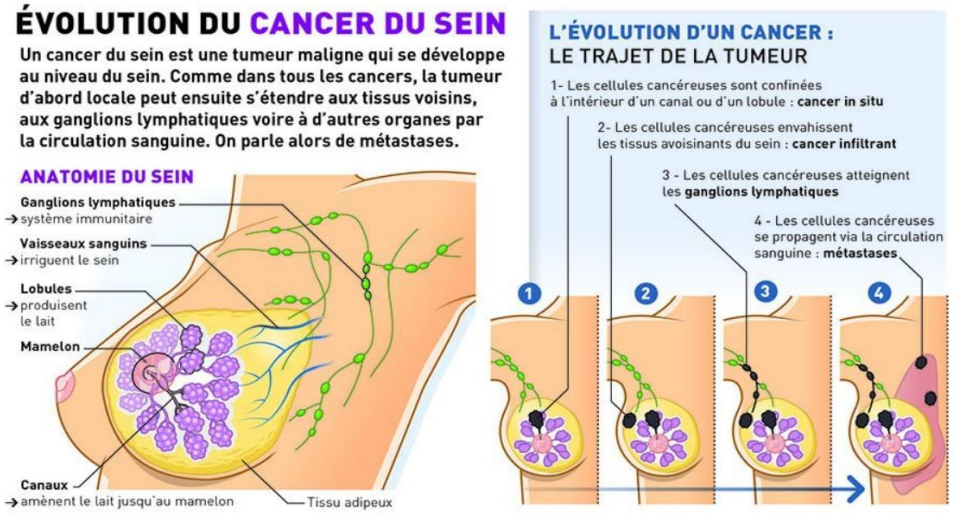


Figure 2: Evolution du cancer du sein

1.3. Facteurs de risque du cancer du sein :

[1] Le cancer du sein est une maladie multifactorielle, dont l'apparition résulte d'une combinaison entre des facteurs environnementaux et génétiques.

Les quatre principaux facteurs de risque sont :

- L'âge. Le risque est maximal entre 65 et 74 ans.
- Les prédispositions génétiques. Lorsque plusieurs personnes d'une même famille sont atteintes du même cancer, il peut s'agir d'un cancer héréditaire : dans ce cas, une mutation dite de prédisposition au cancer du sein est transmise de génération en génération.
- Les antécédents familiaux de cancers du sein (présence d'un cas isolé dans la famille proche).
- L'existence d'antécédents personnels de cancers du sein ou de maladies bénignes du tissu mammaire.

D'autres facteurs de risque plus secondaires ont également été identifiés ou suspectés :

- La puberté précoce et la ménopause tardive.
- Le traitement hormonal substitutif de la ménopause suivi pendant plus de 5 ans.
- L'absence de grossesse ou les grossesses tardives, absence d'allaitement.
- La consommation régulière d'alcool.
- Le tabagisme.

1.4. Moyens de dépistage du cancer du sein :

[1] Avant 50 ans, il est essentiel qu'une femme consulte chaque année son médecin pour qu'il puisse procéder à un examen clinique de ses seins. En cas de doute ou d'anomalie, il pourra alors programmer des examens complémentaires. Depuis 2004, les femmes âgées de 50 à 74 ans sont invitées à se faire dépister tous les deux ans (mammographie et examen clinique).
Plusieurs outils existent pour établir un diagnostic de la pathologie :

- La palpation : elle permet la mise en évidence d'une grosseur anormale.
- L'imagerie :
  - La mammographie : examen radiologique révélant des lésions de quelques millimètres indétectables par la palpation.
  - L'échographie : examen utilisant les ultrasons, prescrit lorsque la mammographie a révélé une anomalie.
  - L'imagerie par résonance magnétique (IRM) : elle est réalisée pour obtenir des renseignements complémentaires aux informations données par la mammographie et l'échographie.

- Les prélèvements : suite à une image suspecte, seul l'examen au microscope d'un prélèvement permet de confirmer le diagnostic d'un cancer du sein. Il existe plusieurs techniques de prélèvement :
    - La cytoponction : prélèvement de quelques cellules avec une aiguille très fine afin de les analyser.
    - La biopsie : prélèvement d'un fragment de tissu réalisé sous anesthésie locale pour un examen microscopique.

1.5. Traitements des cancers du sein :

[1] La prise en charge des patients dépend des caractéristiques de la tumeur et de son extension : suivant la démarche thérapeutique adoptée, cinq types de traitements peuvent-être utilisés, seuls ou en combinaison.

- La chirurgie : c'est une étape incontournable, qui permet d'enlever la masse tumorale. Durant l'opération, le chirurgien peut prélever les ganglions lymphatiques les plus proches de la tumeur, les ganglions sentinelles. Ils sont maintenant analysés directement afin de savoir si les cellules cancéreuses se sont disséminées. En fonction du résultat, le chirurgien adapte son geste.
- La radiothérapie : elle est souvent réalisée en complément de la chirurgie au niveau de la région de la tumeur, souvent en association avec la chimiothérapie. On peut aussi la prescrire avant l'opération pour réduire le volume de la tumeur et faciliter son ablation.
- La chimiothérapie : donnée dans de rares cas en pré-opératoire, elle peut faire diminuer la taille de la tumeur. On fait aussi appel à la chimiothérapie post-opératoire pour prévenir le développement de métastases ou les éliminer en cas d'extension de la maladie.
- L'hormonothérapie : elle est utilisée chez des patientes dont les cellules tumorales expriment des récepteurs hormonaux.
- Les thérapies ciblées : ces traitements s'attaquent de manière spécifique aux tumeurs.

2. *Conduite du projet :*

Une bonne gestion du projet affecte positivement sa conduite et sa réalisation. Le diagramme de Gantt représenté dans la *figure 3* présente graphiquement les tâches effectuées sur une échelle de temps.

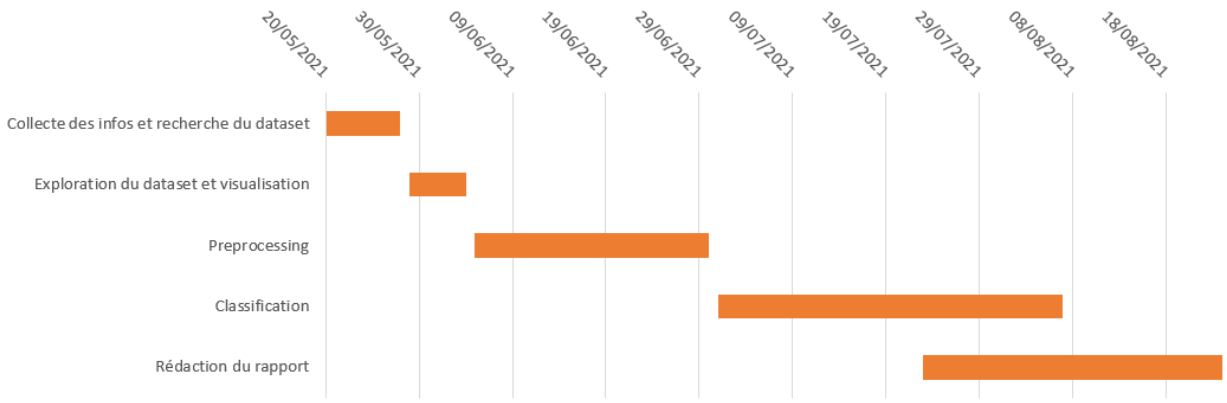


Figure 3 : Diagramme de Gantt

Conclusion :

Dans ce premier chapitre, nous avons défini le champ de notre étude suivi d'un planning de travail afin de préciser nos objectifs à atteindre.

Chapitre II - *Etat de l'art II : Machine Learning*

Introduction :

Dans ce deuxième chapitre, nous présentons l'état d'art du Machine Learning, les différents algorithmes utilisés et les métriques d'évaluation de performance de ces algorithmes.

1. Machine Learning :

Le (ML) est largement reconnu comme la méthodologie de choix dans la classification des modèles de détection du cancer de sein. Les méthodes de classification et de Data Mining sont un moyen efficace de classer les données. Surtout dans le domaine médical, où ces méthodes sont largement utilisées dans le diagnostic et l'analyse pour prendre des décisions.

[6] L'apprentissage automatisé ou apprentissage statistique (Machine Learning en anglais), champ d'étude de l'intelligence artificielle, concerne la conception, l'analyse, le développement et l'implémentation de méthodes permettant à une machine d'évoluer par un processus base sur les données plutôt que par des algorithmes déterministes classiques.

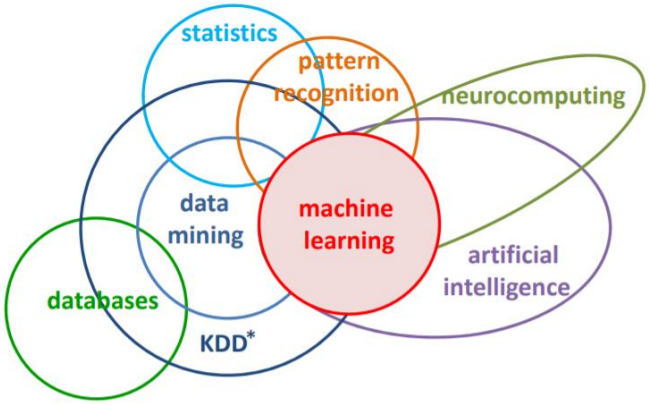


Figure 4: Intersection du ML avec d'autres disciplines

[5] Un programme informatique traditionnel effectue une tâche en suivant des instructions précises, et donc systématiquement de la même façon. Au contraire, un système Machine Learning ne suit pas d'instructions, mais apprend à partir de l'expérience. Par conséquent, ses performances s'améliorent au fil de son " entraînement " à mesure que l'algorithme est exposé à davantage de données.

2. Pourquoi le Machine Learning ?

[5] Le monde évolue aujourd'hui, tout comme les besoins et les exigences des gens. De plus, nous assistons à une quatrième révolution industrielle des données. Afin de tirer des enseignements utiles de ces données et de tirer des leçons de la façon dont les gens et le système interagissent avec les données, nous avons besoin d'algorithmes de calcul qui peuvent produire les données et nous fournir des résultats qui nous seraient utiles de diverses façons. L'apprentissage automatique a révolutionné des industries comme la médecine, la santé, la fabrication, la banque et plusieurs autres industries. Par conséquent, l'apprentissage automatique est devenu un élément essentiel de l'industrie moderne.

3. Champs d'application :

[5] Le Machine Learning concerne tous les secteurs d'activité, notamment l'industrie, le commerce, la santé et les sciences de la vie, le tourisme et l'hôtellerie, les services financiers, l'énergie, les matières premières et les services publics. Domaines d'utilisation :

- **Secteur industriel :** maintenance prédictive et surveillance des équipements.
- **Commerce :** upselling et marketing cross-canal.
- **Santé et sciences de la vie :** diagnostic et réduction des risques.
- **Tourisme et hôtellerie :** tarification dynamique.
- **Services financiers :** analyse et régulation des risques.
- **Énergie :** optimisation de la demande et de l'approvisionnement.

4. Types d'algorithmes de Machine Learning :

[7] Les algorithmes d'apprentissage automatique sont souvent classés comme supervisés ou non supervisés.

- Les **algorithmes d'apprentissage automatique supervisés** peuvent appliquer ce que l'on a appris dans le passé à de nouvelles données en utilisant des exemples étiquetés pour prédire des événements futurs. À partir de l'analyse d'un ensemble de données d'apprentissage connu, l'algorithme d'apprentissage produit une fonction inférée permettant de prédire les valeurs de sortie. Le système est capable de fournir des cibles pour toute nouvelle entrée après un apprentissage suffisant. L'algorithme d'apprentissage peut également comparer sa sortie avec la sortie correcte prévue et trouver des erreurs afin de modifier le modèle en

conséquence.

- En revanche, des **algorithmes d'apprentissage automatique non supervisés** sont utilisés lorsque les informations utilisées pour entraîner ne sont ni classées ni étiquetées. L'apprentissage non supervisé étudie comment les systèmes peuvent inférer une fonction permettant de décrire une structure cachée à partir de données non étiquetées. Le système ne trouve pas le bon résultat, mais il explore les données et peut tirer des déductions à partir de jeux de données pour décrire les structures cachées à partir de données non étiquetées.

- Les **algorithmes d'apprentissage automatique semi-supervisés** se situent quelque part entre l'apprentissage supervisé et l'apprentissage non supervisé, car ils utilisent à la fois des données étiquetées et non étiquetées pour l'apprentissage – généralement une petite quantité de données étiquetées et une grande quantité de données non étiquetées. Les systèmes qui utilisent cette méthode sont capables d'améliorer considérablement la précision de l'apprentissage.

- Les **algorithmes d'apprentissage automatique par renforcement** sont une méthode d'apprentissage qui interagit avec son environnement en produisant des actions et en découvrant des erreurs ou des avantages. La recherche par essais et erreurs et la récompense différée sont les caractéristiques les plus pertinentes de l'apprentissage par renforcement. Cette méthode permet aux machines de déterminer automatiquement le comportement idéal dans un contexte spécifique afin d'optimiser ses performances.

5. Algorithmes de classification utilisés :

Les problèmes de classification consistent à prédire les classes ou étiquettes d'un ensemble de données à partir d'une base d'apprentissage pré-étiquetée. Pour ce faire, nous avons utilisé un ensemble de classifieurs :

- **KNN :**

L'algorithme des K plus proches voisins ou K-nearest neighbors (kNN) est un algorithme de Machine Learning qui appartient à la classe des algorithmes d'apprentissage supervisé simple et facile à mettre en œuvre qui peut être utilisé pour résoudre les problèmes de classification et de régression.

- **Arbre de décision :**

Cet outil d'aide à la décision ou d'exploration de données permet de représenter un ensemble de choix sous la forme graphique d'un arbre. C'est une des méthodes

d'apprentissage supervisé les plus populaires pour les problèmes de classification de données.

- **SVM :**

Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais support vector machine, SVM) sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de classification et de régression. Les SVM sont une généralisation des classifieurs linéaires.

- **Forêts aléatoires :**

[4] Le random forest est composé de plusieurs arbres de décision, travaillant de manière indépendante sur une vision d'un problème. Chacun produit une estimation, et c'est l'assemblage des arbres de décision et de leurs analyses, qui va donner une estimation globale. En somme, il s'agit de s'inspirer de différents avis, traitant un même problème, pour mieux l'appréhender. Chaque modèle est distribué de façon aléatoire aux sous-ensembles d'arbres décisionnels.

- **Régression logistique :**

La régression logistique est un modèle mathématique qui combine un ensemble de variables prédictives (X) avec une variable aléatoire binomiale (Y). Elle est couramment utilisée dans le domaine de l'intelligence artificielle (IA) et du Machine Learning. Elle est considérée comme l'un des modèles d'analyse multivariée les plus simples à déchiffrer et analyser.

6. Métriques d'évaluation utilisées :

Une évaluation rigoureuse des performances d'un algorithme est une étape indispensable. Nous avons donc utilisé un ensemble de métriques :

- **Matrice de confusion :**

Cette matrice est la base de tout. Les calculs suivants et la courbe ROC se basent sur les *tn*, *fn*, *fp* et *tp* que l'on voit sur l'image ci-dessous.

		Predicted class: 0	Predicted class: 1
True class	0	9 true negative (tn)	1 false positive (fp)
	1	0 false negative (fn)	1 true positive (tp)

Figure 5: Matrice de confusion

- **Accuracy :**

L'accuracy permet de connaitre la proportion de bonnes prédictions par rapport à toutes les prédictions. L'opération est simplement :

$$\text{Accuracy} = \frac{tn + tp}{tn + fp + fn + tp}$$

- **Précision :**

La précision correspond au nombre de documents correctement attribués à la classe i par rapport au nombre total de documents prédits comme appartenant à la classe i (total predicted positive).

$$\text{Precision} = \frac{tp}{tp + fp}$$

- **Rappel :**

Le rappel correspond au nombre de documents correctement attribués à la classe i par rapport au nombre total de documents appartenant à la classe i (total true positive).

$$\text{Recall} = \frac{tp}{tp + fn}$$

- **F1-score :**

Le F1-Score combine subtilement la précision et le rappel. Il est intéressant et plus intéressant que l'accuracy car le nombre de vrais négatifs (tn) n'est pas pris en compte.

$$\text{F1-Score} = 2\,\frac{precision * recall}{(precision + recall)}$$

- **Borda count :**

Cette technique consiste à assigner les scores aux modèles, afin de trouver le meilleur. Elle est utilisée dans le but de classer les résultats en se basant sur toutes les métriques d'évaluation utilisés. Cette méthode est souvent adaptée aux problèmes de classification par fusion de classifieurs.

Conclusion :

Nous avons donc pu découvrir dans ce chapitre les outils utilisés pour réaliser notre projet.

Chapitre III - *Contribution : Automatisation de la classification des patients atteints ou non du cancer du sein :*

Introduction :

Ce troisième chapitre consiste à présenter le dataset, les étapes du preprocessing, les résultats des méthodes de classification obtenus, et l'évaluation de leur performance.

1. Description et exploration du dataset :

1.1. Exploration du dataset :

Notre dataset contient plusieurs dossiers : chacun est nommé selon l'*id* de la patiente. Nous avons au total 279 patientes, chacune a un ensemble de patchs qui ont été extraits. Toutes les patientes souffrent du cancer du sein. Donc pour chaque patiente, on a 2 sous-dossiers :

1 : Contient des images des zones contenant les cellules cancéreuses, repérées par les coordonnées x et y.
0 : Contient des images des zones contenant les cellules non cancéreuses (zones saines).

Au total, nous avons 277524 patchs, tels que 198738 appartiennent à la classe 0, et 78786 à la classe 1.

Avant d'entamer le processus de classification, nous avons effectué une analyse du dataset. Afin d'effectuer les taches de la visualisation, nous avons transformé notre dataset en dataframe.

Out[43]:

	patient_id	x	y	target	path
0	10253	1001	1001	0	C:/Users/hp/Documents/MSD/PFA/Dataset/IDC_regu...
1	10253	1001	1051	0	C:/Users/hp/Documents/MSD/PFA/Dataset/IDC_regu...
2	10253	1001	1101	0	C:/Users/hp/Documents/MSD/PFA/Dataset/IDC_regu...
3	10253	1001	1151	0	C:/Users/hp/Documents/MSD/PFA/Dataset/IDC_regu...
4	10253	1001	1201	0	C:/Users/hp/Documents/MSD/PFA/Dataset/IDC_regu...
...	...	...	...	...	...
277519	9383	2051	901	1	C:/Users/hp/Documents/MSD/PFA/Dataset/IDC_regu...
277520	9383	2051	951	1	C:/Users/hp/Documents/MSD/PFA/Dataset/IDC_regu...
277521	9383	2101	1001	1	C:/Users/hp/Documents/MSD/PFA/Dataset/IDC_regu...
277522	9383	2101	901	1	C:/Users/hp/Documents/MSD/PFA/Dataset/IDC_regu...
277523	9383	2101	951	1	C:/Users/hp/Documents/MSD/PFA/Dataset/IDC_regu...

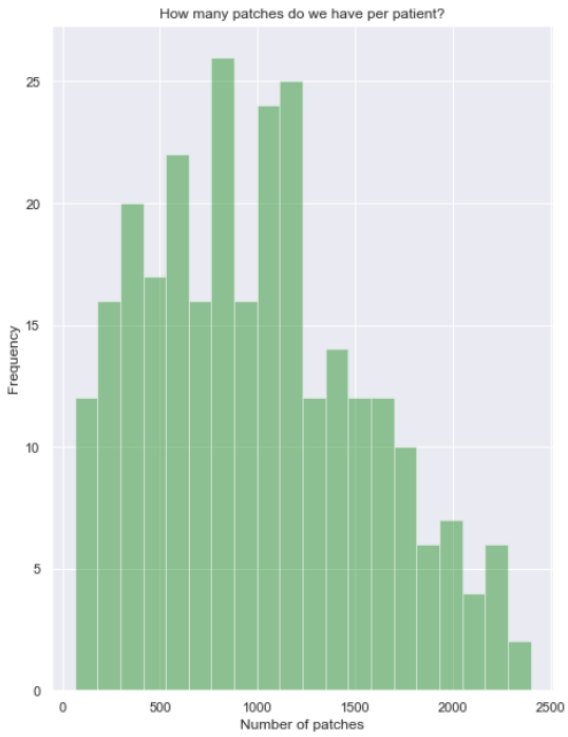
277524 rows × 5 columns

Figure 6: Dataframe

Les caractéristiques sont :

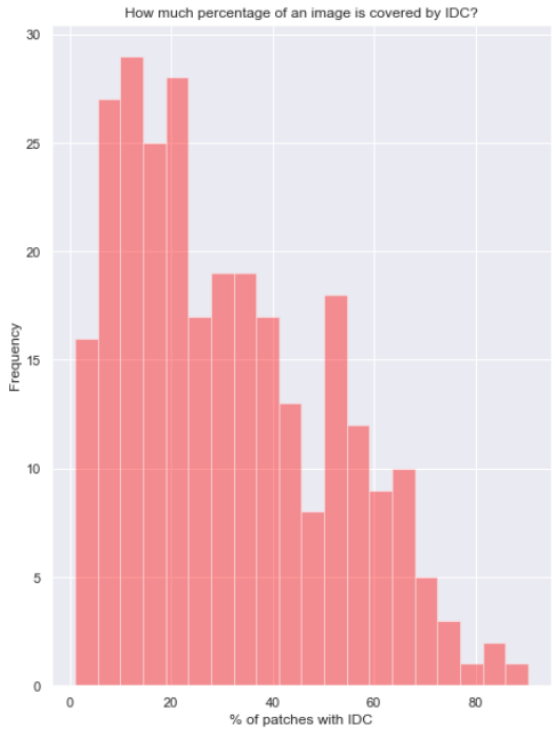
- *patient_id* : c'est l'id de la patiente.
- *x* et *y* : les coordonnées d'où le patch a été rogné.
- *Target* : 0 ou 1.
- *Path* : le chemin d'accès de chaque patch.

1.2. Statistiques :



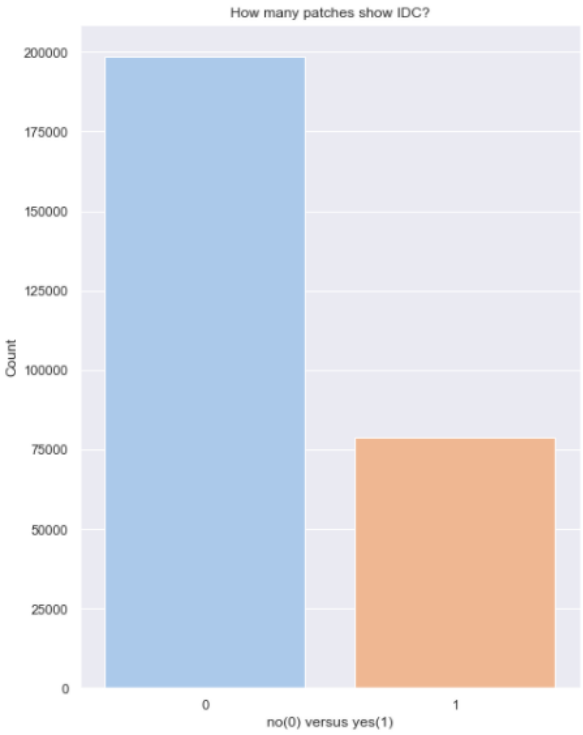
Le nombre de patchs d'images par patiente varie. Cela conduit à se demander si toutes les images montrent la même résolution des cellules tissulaires ou si cela varie entre les patientes.

Figure 7: Visualisation - Le nombre de patchs d'images



Cette figure représente le % des patchs avec le cancer (affecté).

Figure 8: Visualisation - Le % des patchs avec le cancer



Notre dataset est déséquilibré.

Figure 9: Visualisation - Données déséquilibrées

1.3. Visualisation des patchs :

Voici une visualisation des patchs cancéreux :

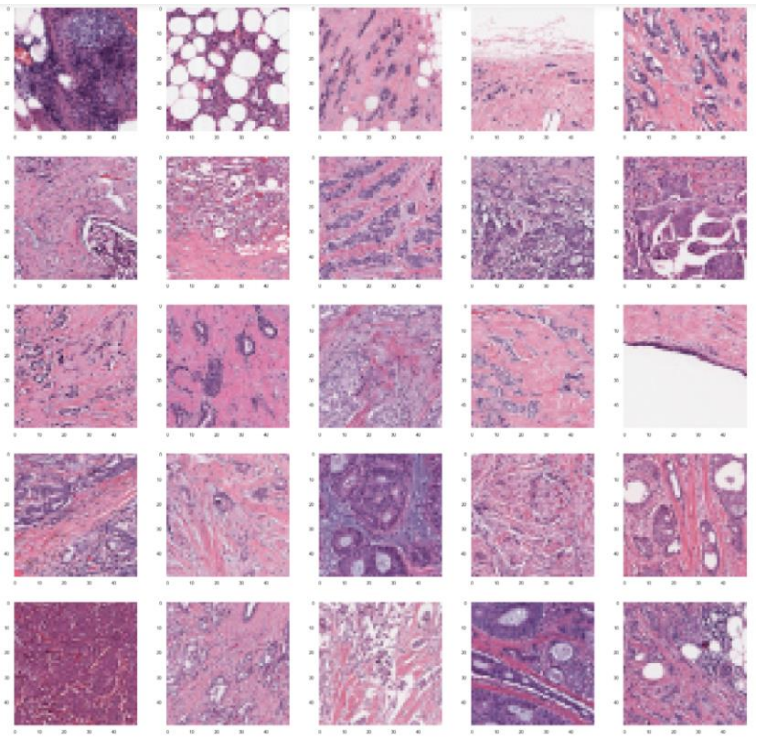


Figure 10: Visualisation - Les patchs cancéreux

Et ceux qui ne sont pas cancéreux :

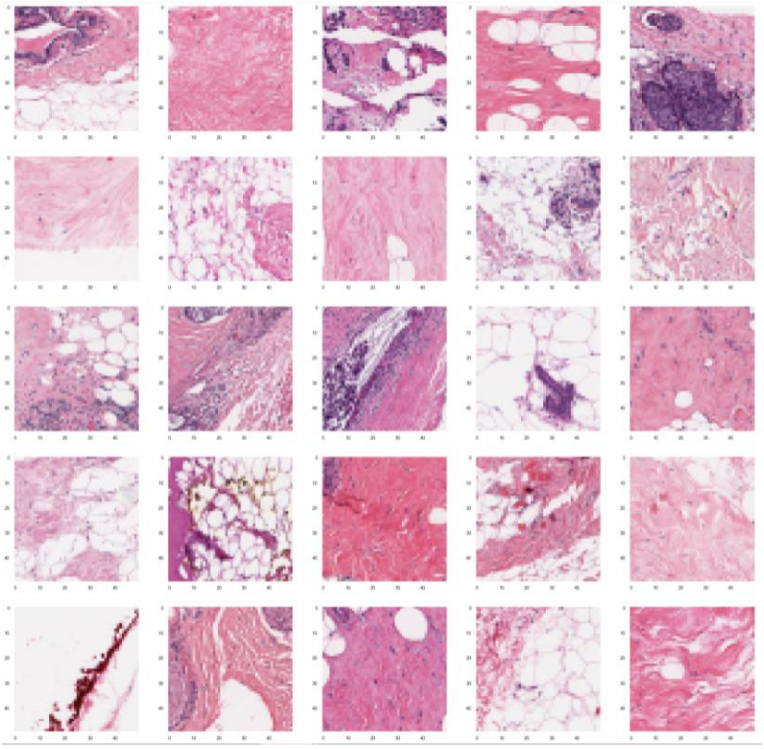


Figure 11: Visualisation - Les patchs non cancéreux

Nous remarquons que les patchs positifs sont plus violets que ceux négatifs.

Pour chaque patiente, nous avons construit une image complète de son tissu mammaire, en combinant l'ensemble de ses patchs.

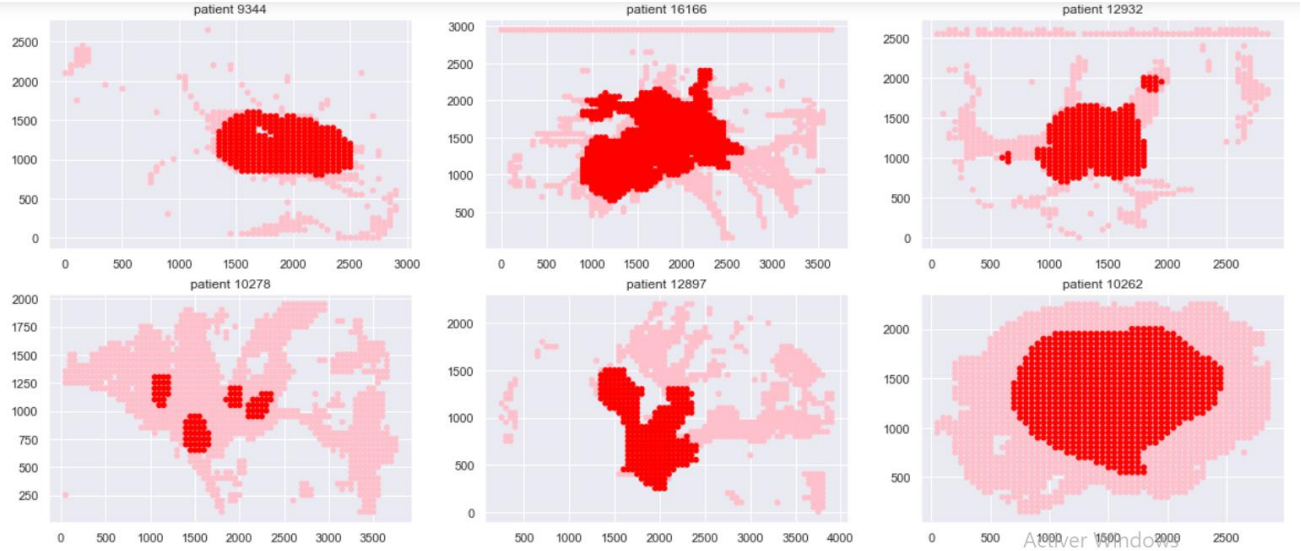


Figure 12: Visualisation - Images complètes des tissus mammaires

Nous remarquons que les tissus cancéreux ont tendance à apparaître en clusters plutôt que d'être dispersés partout.

Voici les images réelles complètes :

Figure 13: Visualisation - Images réelles complètes des tissus mammaires

En comparant les résultats des deux visualisations, il semble que les tissus plus foncés et plus violets ont plus de chances d'être cancéreux que ceux de couleur rose.

1.4. Corrélation entre les images :

Nous avons vérifié par la suite la corrélation entre ces patchs. Pour ce faire, nous avons étudié la similarité entre les images. Nous avons commencé par faire un test, en choisissant 2 images similaires.

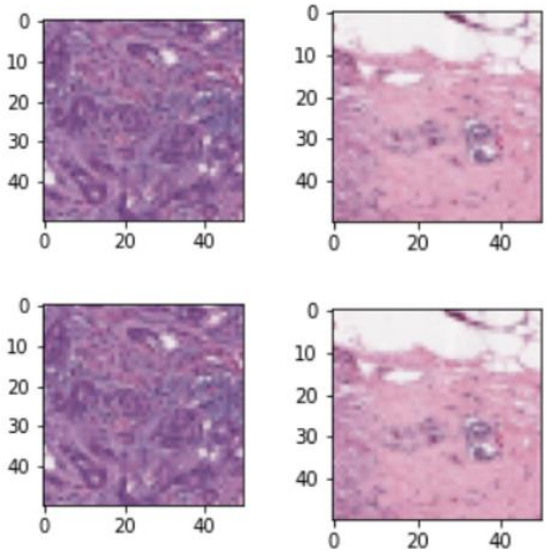


Figure 14: Similarité entre le même patch

Nous avons eu comme résultat $similarité = 1$, ceci prouve que les deux images sont bel et bien similaires. Nous avons refait la même opération mais pour des images différentes, choisies aléatoirement, appartenant aux classes 0 et 1.

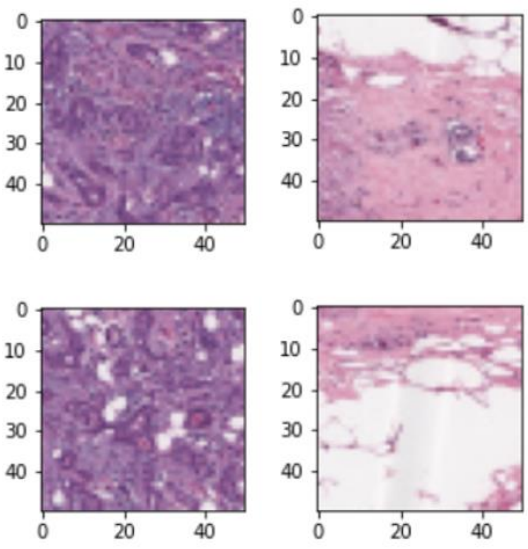


Figure 15: Similarité entre des patchs différents

Nous avons trouvé dans ce cas que la similarité tend vers 0. Les images sont donc différentes. En se basant sur ces résultats, on suppose qu'il n'a pas de corrélation entre l'ensemble des images de notre dataset.

2. Processus expérimental :

Nous allons décrire dans cette section toutes les étapes de prétraitement effectuées, les techniques de classification utilisées et l'évaluation de leur performance.

2.1. Prétraitement :

Le prétraitement des données dans le Machine Learning est une étape cruciale qui contribue à améliorer la qualité des données afin d'extraire des informations significatives à partir des données. Le prétraitement fait référence à la technique de préparation (nettoyage et organisation) des données brutes pour les rendre adaptées à la construction et à la formation des modèles de Machine Learning.

Le prétraitement sert alors à transformer les données brutes en un format compréhensible et lisible.

2.1.1. Normalisation :

[9] La normalisation est une étape essentielle du prétraitement des données dans toute application de Machine Learning et d'ajustement de modèle. Le but de la normalisation est de transformer les données pour qu'elles soient sans dimension et/ou aient des distributions similaires. Ce processus est connu sous d'autres noms tels que standardisation, mise à l'échelle des caractéristiques, etc.

L'histogramme d'une image numérique est une fonction décrivant la répartition des niveaux de gris des pixels de l'image. Cette fonction est définie par : $\forall I \in \{0,\dots,255\}, h(I) = Nombre\ de\ pixels\ ayant\ le\ niveau\ de\ gris\ I$.
L'histogramme suit la forme du graphe de densité de probabilité des niveaux de gris de l'image.

Lorsque l'histogramme est normalisé, il indique en ordonnée la probabilité pi de trouver un pixel de niveau de gris i dans l'image : $\forall i \in \{0,\dots,255\}, p_i = \frac{nombre\ de\ pixels\ d'intensité\ i}{nombre\ total\ de\ pixels}$

La modification de l'histogramme permet d'ajuster la dynamique des niveaux de gris ou des couleurs dans une image afin de la rendre plus agréable visuellement.

Plusieurs applications sont utilisées :

- **Étirement d'histogrammes :** consiste à corriger la luminosité, ou exposition, de l'image. Analysons la forme des histogrammes pour des images dont l'exposition est mauvaise.
- **Égalisation d'histogrammes :** concerne l'amélioration du contraste de l'image. Cette technique consiste à équilibrer le mieux possible la distribution des dynamiques des pixels.

Nous avons généré l'histogramme avec la fonction *hist* de *matplotlib.pyplot*. La figure ci-dessous montre que l'échelle de l'histogramme va de 0 à 256.

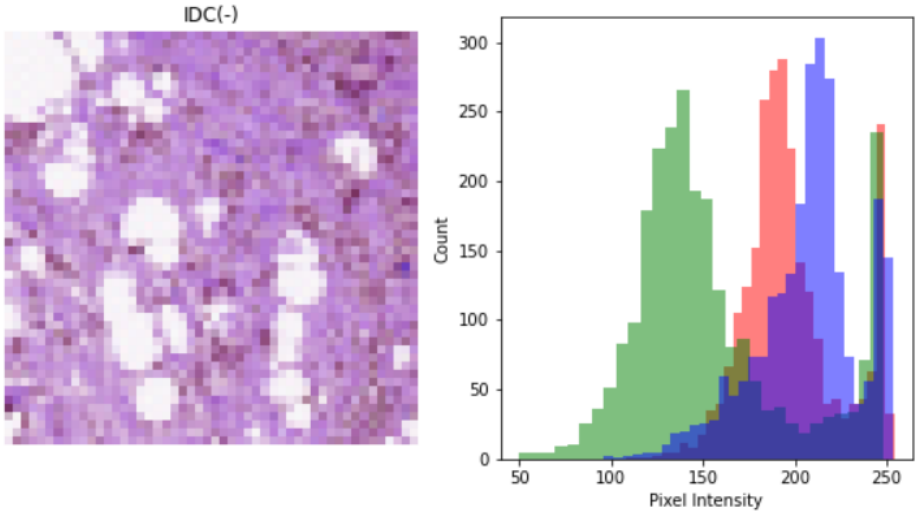


Figure 16: Histogramme non normalisé

Nous avons alors effectué une normalisation, et rendu l'échelle 0 à 1, en appliquant une égalisation d'histogramme. L'objectif de cette opération de normalisation est de faciliter par la suite l'utilisation des algorithmes de classification.

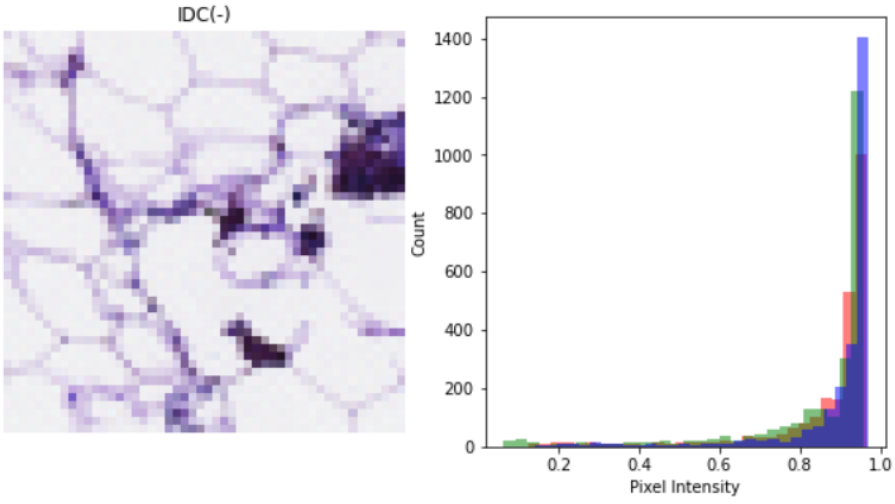


Figure 17: Histogramme normalisé

2.1.2. Data balancing :

La plupart des algorithmes d'apprentissage automatique fonctionnent mieux lorsque le nombre d'échantillons dans chaque classe est à peu près égal. En effet, la plupart des algorithmes sont conçus pour maximiser la précision et réduire les erreurs.

Cependant, si les données sont déséquilibrées, nous obtenons une précision assez élevée simplement en prédisant la classe majoritaire, mais nous ne parvenons pas à capturer la classe minoritaire.

Les techniques les plus adoptées pour traiter les données déséquilibrées sont :

- **Le suréchantillonnage (Oversampling) :** consiste à dupliquer des données aléatoirement de la classe minoritaire.
- **Le sous-échantillonnage (Undersampling) :** cette technique est la plus simple, elle consiste à supprimer des données aléatoirement de la classe majoritaire.

Notre dataset contient 65279 images appartenant à la classe 0, et 24721 à la classe 1. Nous avons donc fait le sous-échantillonnage rendu donc les données équilibrées.

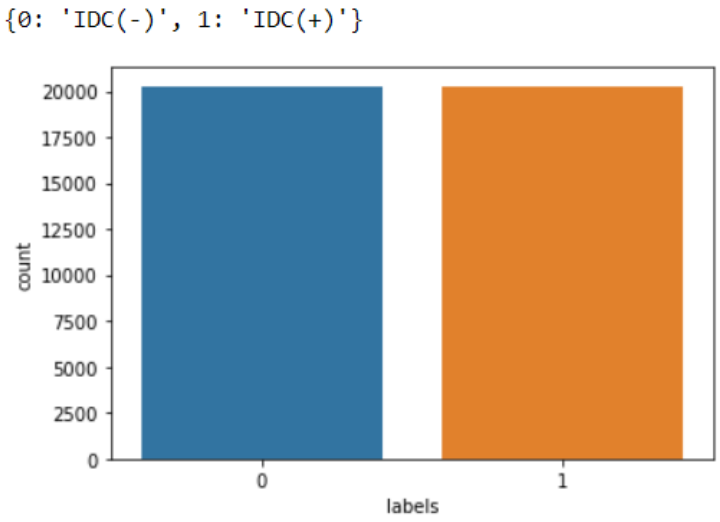


Figure 18: Dataset équilibré

2.2. Résultats des méthodes de classification:

2.2.1. Résultats :

Dans cette partie, nous présentons les résultats obtenus depuis l'expérimentation, en se basant sur les métriques d'évaluation.

Algorithmes	Precision		Recall		F1-score		Support		Matrice de confusion	Accuracy
	0	1	0	1	0	1	0	1		
DT	0.68	0.70	0.69	0.69	0.69	0.69	5039	5257	$\begin{pmatrix}3500 & 1539\\1648 & 3609\end{pmatrix}$	**0.69**
LR	0.74	0.68	0.70	0.73	0.72	0.70	5442	4854	$\begin{pmatrix}3819 & 1623\\1329 & 3525\end{pmatrix}$	**0.71**
KNN	0.66	0.84	0.81	0.72	0.73	0.77	4232	6064	$\begin{pmatrix}3421 & 811\\1727 & 4337\end{pmatrix}$	**0.75**
RF	0.81	0.79	0.79	0.81	0.80	0.80	5283	5013	$\begin{pmatrix}4181 & 1102\\967 & 4046\end{pmatrix}$	**0.80**
SVM	0.83	0.80	0.81	0.83	0.82	0.81	5245	4931	$\begin{pmatrix}4230 & 1015\\858 & 4073\end{pmatrix}$	**0.82**

Maintenant nous appliquons la méthode *Borda count* sur nos différents classifieurs. Nous obtenons le résultat suivant :

```
Le classement des modèles :  [('SVM', 4), ('RF', 8), ('KNN', 12), ('LR', 16), ('DT', 20)]
```

2.2.2. Discussion :

En se basant sur les différentes métriques utilisées, nous pouvons constater que le *SVM* donne les meilleurs résultats.

Comme l'accuracy est égale à 0.82, et en pourcentage 82% après que nous avons équilibré notre dataset, nous pouvons dire que le modèle du SVM est performant.

La précision, le recall et le F1-score sont de 0.82, elles tendent vers 1, alors le modèle peut être considéré comme étant un bon classifieur.

Selon la matrice de confusion :

- **Vrai positif (TP) :** 4073 patchs cancéreux sont prédits positivement par le modèle.
- **Vrai négatif (TN) :** 4230 patchs non cancéreux sont prédits négativement par le modèle.

- **Faux positif (FP) :** 858 patchs sont prédits comme positifs d'être cancéreux, bien qu'ils ne le soient pas.
- **Faux négatif (FN) :** 1015 patchs cancéreux sont prédits comme négatifs.

Le test statistique non paramétré *Borda count* affirme que le *SVM* est bel est bien le modèle le plus performant.

Conclusion :

Après avoir effectué les tâches : le prétraitement, les algorithmes de classification et les métriques d'évaluation, nous avons pu conclure l'algorithme qui donne les meilleurs résultats.

Conclusion générale

Le Machine Learning a présenté un réel potentiel d'impact dans les soins de santé, et en particulier dans le domaine du diagnostic médical. En fait, il rend les soins de santé plus intelligents. Ce puissant sous-ensemble d'intelligence artificielle peut être aussi appliqué à de nombreux d'autres cas d'utilisation, tels que la reconnaissance vocale utilisée par les assistants vocaux et la création d'expériences d'achat en ligne personnalisées grâce à sa capacité à apprendre des associations.

Le cancer du sein est considéré comme l'une des principales causes de décès chez les femmes. La détection précoce du cancer du sein joue un rôle essentiel pour sauver la vie des femmes. La détection du cancer du sein peut être effectuée à l'aide d'algorithmes modernes de Machine Learning. Dans notre projet, nous nous sommes concentrées sur l'application d'un ensemble de méthodes de classification. En appliquant des métriques différentes, nous avons pu évaluer nos algorithmes et déterminer celui qui donne les meilleurs résultats.

Comme perspective de ce travail, nous comptons déployer notre modèle en utilisant le framework *Flask*.

Webographie

[1] https://www.frm.org consulté le 26/05/2021

[2] https://www.medias24.com consulté le 26/07/2021

[3] https://contrelecancer.ma consulté le 28/07/2021

[4] https://www.journaldunet.fr consulté le 26/07/2021

[5] https://www.lebigdata.fr consulté le 01/08/2021

[6] http://www.campus-industrie.eu consulté le 01/08/2021

[7] https://analyticsinsights.io consulté le 04/08/2021

[8] https://techtomed.com consulté le 04/08/2021

[9] https://openclassrooms.com consulté le 04/08/2021