

**Projet :**

**Breast Cancer Detection**

***REALISE par :***

* *El-asri Nossaiba*
* *charaFi asmaa*

***encadre par :***

* *mme. Radgui*

**Remerciement :**

Au terme de ce travail, nous désirons exprimer nos remerciements à tous ceux qui nous ont encouragés dans notre projet.

Tout d’abord nous tenons à exprimer notre gratitude à Madame **A. Radgui** qui nous a aidés et qui nous a consacré du temps. Et aussi d’avoir donné l’opportunité de réalisé ce projet et acquérir toute ces compétences et développé notre information, pour faciliter notre assimilation du cours Machine Learning

**1. Description de l’objectif de l’apprentissage et de la problématique traitée :**

Le cancer du sein (C.-B.) est l’un des cancers les plus courants chez les femmes dans le monde, représentant la majorité des nouveaux cas de cancer et des décès liés au cancer selon les statistiques mondiales, ce qui en fait un problème de santé publique majeur dans la société d’aujourd’hui.

Le diagnostic précoce de la CB peut améliorer considérablement le pronostic et les chances de survie, car il peut favoriser un traitement clinique rapide des patients. Une classification plus précise des tumeurs bénignes peut empêcher les patients de subir des traitements inutiles. Ainsi, le diagnostic correct de la CB et la classification des patients en groupes malins ou bénins font l'objet de nombreuses recherches. En raison de ses avantages uniques dans la détection des caractéristiques critiques à partir d'ensembles de données BC complexes, l'apprentissage automatique (ML) est largement reconnu comme la méthodologie de choix dans la classification des modèles BC et la modélisation des prévisions.

Les méthodes de classification et d'exploration de données 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.

La classification fait partie des méthodes les plus courantes de l’apprentissage supervisé. Il utilise des données historiques étiquetées pour développer un modèle qui sera ensuite utilisé pour les prévisions futures. Dans le domaine médical, les cliniques et les hôpitaux maintiennent de grandes bases de données qui contiennent des enregistrements de patients avec leurs symptômes et leur diagnostic. Par conséquent, les chercheurs utilisent ces connaissances pour développer des modèles de classification qui peuvent faire des inférences basées sur des cas historiques. L'inférence médicale est donc devenue une tâche beaucoup plus simple avec un support basé sur une machine utilisant la seule quantité de données médicales disponibles aujourd'hui. Il est utile de noter que toutes les techniques utilisées dans cet article relèvent des modèles de sous-classification.

**2. Description du Dataset :**

On a utilisé le référentiel UCI Machine Learning pour l'ensemble de données sur le cancer du sein.

Cet ensemble de données a été collecté périodiquement sur trois ans par le Dr. William H. Wolberg de l'Université du Wisconsin Hospitals et se compose de 669 cas, où les cas sont classés comme malins ou bénins. 458 des cas sont bénins et 241 sont malins. Les 10 attributs sont :

• Clump Thickness

• Cell Size Uniformity

• Cell Shape Uniformity

• Marginal Adhesion

• Single Epithelial Cell Size

• Bare Nuclei

• Bland Chromatin

• Normal Nuclei

• Mitoses

• Class

Tous ces attributs, à l'exception de la classe, sont de type numérique et leurs valeurs sont comprises entre 1 et 10. La classe a une valeur de 2 pour bénin et 4 pour malin

Cette analyse vise à observer quelles caractéristiques sont les plus utiles pour prédire le cancer malin ou bénin et à voir les tendances générales qui peuvent nous aider dans la sélection du modèle et la sélection des hyper paramètres. Le but est de classer si le cancer du sein est bénin ou malin. Pour ce faire, on a utilisé des méthodes de classification d'apprentissage automatique pour adapter une fonction qui peut prédire la classe discrète de la nouvelle entrée.

**3. Présentation de l’approche d’apprentissage adoptée et justification du choix :**

SVM est l'une des techniques de classification ML supervisées qui est largement appliquée dans le domaine du diagnostic et du pronostic du cancer. SVM fonctionne en sélectionnant des échantillons critiques de toutes les classes connues sous le nom de vecteurs de support et en séparant les classes en générant une fonction linéaire qui les divise aussi largement que possible à l'aide de ces vecteurs de support. Par conséquent, on peut dire qu'un mappage entre un vecteur d'entrée et un espace de haute dimensionnalité est réalisé à l'aide de SVM qui vise à trouver l'hyperplan le plus approprié qui divise l'ensemble de données en classes. Ce classificateur linéaire vise à maximiser la distance entre l'hyperplan de décision et le point de données le plus proche, que l'on appelle distance marginale, en trouvant l'hyperplan le mieux adapté.

Différente des autres techniques, la SVM dépend des vecteurs de support, qui sont les ensembles de données les plus proches de la frontière de décision, dans leurs algorithmes. En effet, la suppression d'autres points de données qui sont plus éloignés de l'hyperplan de décision ne modifiera pas autant la limite que si les vecteurs de support étaient supprimés.

**4. Evaluation des performances de l’apprentissage :**

**Accuracy :**

La précision du classificateur est une mesure de la capacité du classificateur à prédire correctement les observations dans leur catégorie correcte.Il s'agit du nombre de prédictions correctes divisé par le nombre total d'instances dans l'ensemble de données. Il est à noter que la précision dépend fortement du seuil choisi par le classifieur et peut donc changer pour différents ensembles de tests.Par conséquent, ce n'est pas la méthode optimale pour comparer différents classificateurs mais peut donner un aperçu de la classe

**Recall :**

Recall également connu sous le nom de sensibilité, est le taux d'observations positives qui sont correctement prédites comme positives. Cette mesure est souhaitable, notamment dans le domaine médical, car le nombre d'observations est correctement diagnostiqué .Dans cette étude, il est plus important d'identifier correctement une tumeur a maligne que de mal identifier une tumeur bénigne.

**Precision :**

La précision, également connue sous le nom de confiance, est le taux de vrais positifs et de vrais négatifs qui ont été identifiés comme de vrais positifs. Cela montre à quel point le classificateur gère les observations positives mais ne dit pas grand-chose sur les observations négatives.

**5. Analyse des résultats :**

On a comparé les résultats obtenues par la méthode SVM et celles de KNN et logistic regression

* On a utilisé la visualisation de pandas qui est construite au-dessus de matplotlib, pour trouver la distribution des données des fonctionnalités.
* On a utilisé la bibliothèque sklearn pour importer toutes les méthodes d'algorithmes de classification.
* Après avoir appliqué les différents modèles de classification, nous avons obtenu les précisions ci-dessous avec différents modèles :

**6. Conclusion :**

Les ML techniques ont été largement utilisées dans le domaine médical et ont servi d'outil de diagnostic utile qui aide les médecins à analyser les données disponibles ainsi qu'à concevoir des systèmes experts médicaux. Cet article présentait trois des techniques de ML les plus populaires couramment utilisées pour la détection et le diagnostic du cancer du sein, à savoir Support Vector Machine (SVM), Random Forest (RF) et Bayesian Networks (BN). Les principales caractéristiques et la méthadologie de chacune des trois techniques de ML ont été décrites. La comparaison des performances des techniques étudiées a été réalisée à l'aide de l'ensemble de données original du WisconsinBreast Cancer.Les résultats de la simulation obtenus ont prouvé que les performances de classification varient en fonction de la méthode choisie