الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي والبحث العلمي

جامعة سعيدةد. مولاي الطاهر

كلية التكنولوجيا

قسم: الإعلام الآلي

Mémoire de Master

Spécialité : MICR

Thème

Apprentissage supervisé pour classification d'images médicales: méthodes SVM et KNN

Présenté par :

BOUARFA OUSSAMA

BELLIL HADJIRA

Dirigé par :

Me Derkaoui ORKIA

Promotion 2021 - 2022

Remerciements

A Dieu
En premier lieu, je tiens à remercier man encadreur Dr.Derkaoui Orkia qui m'a aidé et conseillé durant ce travail.
Je remercie également tous les enseignants du département de l'informatique de l'université.
Enfin, je remercie tous ceux qui en soutenu,encouragé et donné l'envie de mener à terme ce travail.

Résumé

Dans ce travail,nous avons abordé les problémes de la classification des données.
L'une dans les questions principales que nous avons considérer concerne la classification d'image qui la tache d'attibuer à une image d'entrée X un label Y à partir d'un ensemble fixé de catégories,et intérésser spécialement les images médicale Brain Tumor MRI,nous allons choisir une base de recherche (Dataset) composée de petites images Gray scale.Cette base se composé de 1222 petites images Gray scale.Nous utilisons l'apprentissage supervisé avec les deux algorithmes les plus populaires en classification:SVM(Support Vector Machine) et KNN(K-Nearest Neighbors).Des tests de comparaison entre les deux méthodes nous permettrons d'évaluer le principe de chacune de ces méthodes.

L'objectif de la classification d'images est d'élaborer un systeme capable d'affecter une classe automatiquement à une image.Ainsi, ce systeme permet d'éffectuer une tache d'expertise qui peut s'avére couteuse à acquérir pour un étre humain en raison notament de contraintes physiques comme la concentration, la fatigue ou le temps nécessité par un volume important de données images .

Mots Clés : *Apprentissage supervisé,classification d'images ,SVM , KNN*

Abstract

In this work, we have addressed the problems of data classification.
One of the main issues we have considered concerns image classification, which tasks it with assigning an input image X a label Y from a fixed set of categories, and is of particular interest to medical images Brain Tumor MRI, we will choose a research base (Dataset) composed of small Gray scale images. This base consists of 1222 small Gray scale images. We use supervised learning with the two most popular classification algorithms: SVM (Support Vector Machine) and KNN(K-Nearest Neighbors). Comparison tests between the two methods will allow us to evaluate the principle of each of these methods.

The objective of image classification is to develop a system capable of automatically assigning a class to an image. human due in particular to physical constraints such as concentration, fatigue or the time required by a large volume of image data.

Key Words : *Supervised Learning , Classification Image , SVM , KNN .*

Table des matières

1 Introduction générale **7**
1.0.1 Problématique . . . 7

2 Etat de l'art **8**
2.1 Qu'est le machine learning ? . . . 8
2.1.1 Les différents types de Machine Learning . . . 8
2.2 Apprentissage Supervisé et Non supervisé . . . 8
2.2.1 Régression et Classification . . . 8
2.2.2 Apprentissage par Renforcement . . . 8
2.2.3 Apprentissage Semi-Supervisé . . . 9
2.3 Les différents types d'algorithm . . . 9
2.3.1 La Régression Linéaire . . . 9
2.3.2 Les k plus proches voisins . . . 10
2.3.3 Le Classifieur Naïf de Bayes . . . 11
2.3.4 K-means . . . 12
2.3.5 Les arbres de décision . . . 12
2.3.6 Les forêts aléatoires . . . 13
2.3.7 Les machines à vecteur de support . . . 13
2.3.8 Les Réseaux de Neurones . . . 14
2.3.9 Performance et sur-apprentissage . . . 15

3 Conclusion **17**

4 L'Apprentissage Supervisé **18**
4.1 Définition Mathématique . . . 18
4.1.1 Avantages et Inconvénients . . . 19
4.1.2 Applications Apprentissage Supervisé . . . 19
4.1.3 Méthodes D'apprentissage Supervisé . . . 20
4.1.4 Le principe de fonctionnement de l'algorithme KNN . . . 20
4.1.5 Les Machines à Vecteur de Support . . . 21
4.2 La Classification . . . 22
4.2.1 Les Types de Classification . . . 23
4.2.2 La Classification Binaire . . . 23
4.2.3 Classification à Classes Multiples . . . 23
4.2.4 Mesure d'évaluation pour les modéles de classification . . . 24
4.2.5 Mesures de Classification . . . 25
4.3 Classification des Images . . . 26
4.3.1 Les Motivations de Classification d'images . . . 26
4.3.2 Extraction de caractéristiques . . . 27
4.3.3 La Méthode de résolution . . . 27
4.3.4 Le traitement d'images . . . 27
4.3.5 C'est quoi une image . . . 28
4.3.6 Définition de l'image numérique . . . 28
4.3.7 Types d'images . . . 28
4.3.8 Filtrage des images . . . 30
4.3.9 Rehaussement des images . . . 30
4.3.10 L'ihstogramme d'une image . . . 31
4.3.11 Transformation des images . . . 31
4.3.12 Méthodes de classification supervisé . . . 31

5 Conclusion **32**

6 Introduction **33**
6.1 Les types de Tumeurs cérébrale . . . 33
6.1.1 Tumeur Gliome . . . 33
6.1.2 Tumeur Méningiome . . . 33
6.1.3 Tumeur Hypophysaire . . . 34
6.1.4 Qu'est-ce qu'une IRM ? . . . 34
6.1.5 Comment un système IRM produit-il une image diagnostique ? . . . 34
6.2 Prétraitement du dataset . . . 35
6.3 Méthodologie de mise en œuvre . . . 35
6.3.1 Présentation des outils . . . 35
6.3.2 Les Paquets Python . . . 35
6.3.3 Le hardware . . . 36
6.3.4 acquisition d'image . . . 36
6.3.5 Kaggle Dataset : . . . 36
6.4 Les étapes pour créer un système d'apprentissage automatique . . . 37
6.4.1 Méthode Proposée . . . 37
6.5 Mise en œuvre et résultats . . . 38
6.5.1 Charger les dépendances . . . 38
6.5.2 Split data . . . 38
6.5.3 Feature Scaling . . . 40
6.5.4 Visualisation de données . . . 40
6.5.5 Model Training . . . 40
6.5.6 Evaluation . . . 40
6.5.7 Testing dataset . . . 41
6.5.8 Evaluation les metrics en Table . . . 43
6.5.9 Confusion Matrix . . . 43

7 Conclusion **44**
7.1 Conclusion et Perspectives . . . 44

Table des figures

1 Le processus typique du ML . . . 8
2 Apprentissagr par Renforcement . . . 9
3 (a) représente le prix des maisons en fonction de leur surface qui sont en vent àBerkeley. (b) montre un exemple d'ensemble d'apprentissage de n points sur le plan x,y . . . 10
4 KNN classe majoritaire . . . 11
5 Naive Bayes. . . . 11
6 K-means Clustering . . . 12
7 Exemple d'arbre de décision . . . 13
8 Illustration d'un hyperplan de séparation optimale . . . 14
9 Support Vector Machine . . . 14
10 Les Réseaux de Neurones . . . 15
11 Sur-Apprentissage . . . 15
12 La Validation Croisée . . . 16
13 Schéma fonctinnel illustrant la formde L'apprentissage Supervisé . . . 18
14 L'algorithme du K-NN et son princip de fonctionnement par l'éxemple . . . 20
15 Distance entre deux points :distance Euclidienne versus distance de Manhattan . . . 21
16 Les SVM permettent de Projeter les dpnnées dans une espace de plus grande dimension via une fonction noyau pour les séparer linéairement . . . 22
17 Processus d'Apprentissage avec modele SVM . . . 22
18 Classification Binaire . . . 23
19 Classification à Classes Multiples . . . 24

20 Matrice de Confusion pour la classification binaire . . . 25
21 Exemple d'image matricielle . . . 29
22 Diagramme résume l'ensemble des traitements qui peuventetre appliqués à l'image . . . 30
23 Exemple sur image rehaussée . . . 30
24 L'histogramme d'une image . . . 31
25 Les étapes de classification supervisées . . . 32
26 Représentation d'images d'imagerie par résonance magnétique (IRM) normalisées montrant différents types de tumeurs dans différents plans . . . 34
27 Kaggle Dataset exemple . . . 36
28 Résultat de splite data train . . . 39
29 Résultat de splite data test . . . 39
30 resultat de visualisation . . . 40
31 Resultat de classification SVM . . . 42
32 nouveau cas avec SVM . . . 42
33 nouveau cas avec KNN . . . 43
34 Histogramme de Comparaison entre SVM et KNN . . . 43

Liste des tableaux

1 Dataset pour la classification . . . 38
2 Régularisation de K pour meilleurs résultats . . . 40
3 Régularisation de C pour meilleurs résultats . . . 41
4 Résultats de evaluation les metrics . . . 43
5 Confusion Matrix KNN . . . 44
6 Confusion matrix SVM . . . 44

Acronyms

LR : Linéaire Regression
IA : Intelligence Artificielle
ML : Machine Learning
K-NN : K Nearest Neighbors
RF : Random Forest
SVM : Support Vector Machine
TP : True Positive
TN : True Negative
FP : False Positive
FN : False Negative
CV : Computer Vision
PCA : Principal Composant Analyse
IRM : Image Résonance Magnétique
UI : User Interface

Introduction générale

Intelligence Artificielle (IA)est l'un des domaines les plus récents de la science et de l'ingénierie.

Les travaux ont sérieusement débute après la seconde guerre mondial,et le nom lui mémé a été inventé en 1956. Régulièrement cité comme "domaine ou j'aimerai bien y être" par les scientifique dans d'autres disciplines. Un étudiant en physique peut raisonnablement se dire que toutes les bonnes idées ont été trouvées par Galilée,Newton,Einstein et le reste. De l'autre coté,tout reste ouvert en IA.

Historiquement,quatre approche de l'IA ont été suivies,chacune par des gens différents avec des méthodes différentes.Une approche axée sur l'homme doit être en partie une science empirique,impliquant observation et hypothèse sur le comportement humaine.une approche rationnelle implique une combinaison de mathématique et d'ingénierie[38].

Les différents groupes se sont décriés et se sont aides mutuellement,voici les quatre approches :

1)**Acting humanly** : le test de Turing "l"art de créer des machines qui exécutent des fonctions requérant une intelligence lorsqu'elles sont exécutées par des êtres humains".

2)**Thinking humanly** :la modélisation cognitive"L'excident nouveau défi de construire des ordinateurs qui pensent,des machines avec des consciences, au sens figurré".

3)**Thinking rationally** :les lois de la pensée "L'étude des calculs qui rendent possibles la perception,le rationnement et les actes"[].

4)**Acting rationnaly** :les agents rationnels "L'IA es l'étude de la conception des agents intelligents"[74]

L'IA englobe plusieurs sous domaines allant du plus générale (apprentissage et perception)au plus spécifique,comme jouer aux échecs,démontrer des théorèmes mathématique,conduire une voiture ou diagnostique des maladies .L'IA se révèle être utile dans toutes les taches intellectuelles.C'est vraiment un domaine universel et pluri-disciplinaire.
Le but de la recherche en IA est de créer une technologie qui permette aux ordinateurs et aux machines de fonctionner d'une manière intelligente.

Le problème générale de la création d'une intelligence a été divisé en plusieurs sous problèmes.Celles-ci consistent en des capacités que les chercheurs espèrent qu'un système intelligent pourra exécuter[74].

Il existe plusieurs stratégies utilisées en IA ; entre autres l'Apprentissage automatique.
On peut citer trois types d'algorithme d'apprentissage automatique :

- Apprentissage Supervisé.
- Apprentissage non Supervisé.
- Apprentissage par Renforcement.

1.0.1 Problématique

La classification est un probléme central de l'apprentissage automatique (ML) et de l'IA.Une régle de classification est une procédure permettant d'affecter à un objet l'étiquette du groupe auquel il appartient,autrement dit de le reconnaitre.

Nous allons nous intéresser à la problématique de **la classification d'image** qui est la tache d'attribuer à une image d'entrée X un label Y à partir d'un ensemble fixe de catégories.C'est l'un des problemes fondamentaux de la vision par ordinateur.,nous allons choisir une base de recherche (Dataset) composée de petites images Gray scale.Cette base se composé de 1222 petites images Gray scale.Nous utilisons l'apprentissage supervisé avec les deux algorithmes les plus populaires en classification :SVM(Support Vector Machine) et KNN(K Neighbors).Des tests de comparaison entre les deux méthodes nous permettrons d'évaluer le principe de chacune de ces méthodes.
Ce mémoire est structure en 3 Chapitres :
Dans le premier chapitre nous presentons les différents types de machine learning et les algorithmes .
Dans le deuxieme chapitres : nous presentons les traitement d'images pour faire la classification .
Dans le troisiéme chapitres :nous présentons le domaine de la classification d'image et le résolution de ces problémes.

Chapitre 1

Etat de l'art

2.1 Qu'est le machine learning ?

Le ML est une discipline de l'IA qui offre aux ordinateurs la possibilité d'apprendre à partir d'un ensemble d'observations que l'on appelle ensemble d'apprentissage.
Le Machine Learning(ML)a connu un essor de son utilisation et de son application à des problèmes d'automatisation dans divers domaines[67].

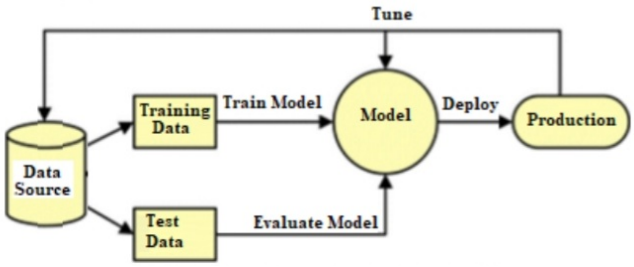

FIGURE 1 – Le processus typique du ML

2.1.1 Les différents types de Machine Learning

2.2 Apprentissage Supervisé et Non supervisé

-L'apprentissage **supervisé** est la tache d'apprentissage automatique la plus simple et la plus connue.

-L'apprentissage **Non supervisé** ou Clustering ne demande aucun étiquetage préalable des données.Le but est que le modèle réussisse à regrouper les observations disponibles en catégories par lui même[69].

2.2.1 Régression et Classification

— Un modèle de **Classification** est un modèle de ML dont les sorties y appartiennent à un ensemble fini de valeurs (exemple :bon,myen,mauvais)[50]

— Un modèle de **Régression** est un modèle de ML dont les sorties y sont des nombres (exemple :la température de demain)[49]

2.2.2 Apprentissage par Renforcement

Apprentissage par Renforcement

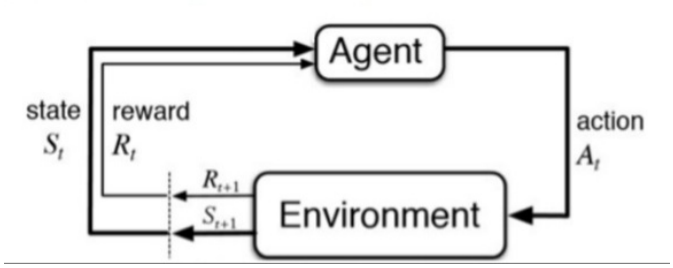

FIGURE 2 – Apprentissagr par Renforcement

2.2.3 Apprentissage Semi-Supervisé

L'apprentissage Semi-Supervisé est une combinaison de méthodes d'apprentissage automatique supervisé et non supervisé,il peut etre fructueux dans les domaines de l'apprentissage automatique et de l'exploration de données ou les données Non étiquetées son déja présentes et ou l'obtention des données étiquetées est un processus fastidieux avec des méthodes d'apprentissage automatique plus courantes.vous formez un algorithme d'apprentissage automatique sur un ensemble de données étiqueté dans lequel ,chaque enregistrement comprende les infirmations sur les résultats[70].

2.3 Les différents types d'algorithm

2.3.1 La Régression Linéaire

-Une régression linéaire est un modèle de ML supervisé,avec x en entrée et y en sortie elle est de la forme[21]

$$y = ax + b \tag{1}$$

1)-**Regression Linéaire Simple**
est l'une des techniques les plus utilisées dans l'apprentissage automatique et cela revient principalement à sa simplicité et la facilité d'interprétation de ses résultats.Comme on a pris l'habitude d'applique le modéle d'apprentissage sur un exemple pour le voir en action , on va procéder ainsi pour l'algorithme de régression linéaire simple.Tout d'abord,on importe le clsse **LinearRegression** et définit les données x et y sur lesquelles le modele va performer[21].

2)-**Régression Linéaire Multiple**

est une méthode de régression mathématique étendant la régression linéaire simple pour décrire les variations d'une variable endogéne associée aux variations de plusieurs variables exogénes[61].

Avantages :
-Le modele est facile à interpréter.

Inconvénients :
-Sensible aux bruits.
-Négligence des interactions entre les variables prédictives.

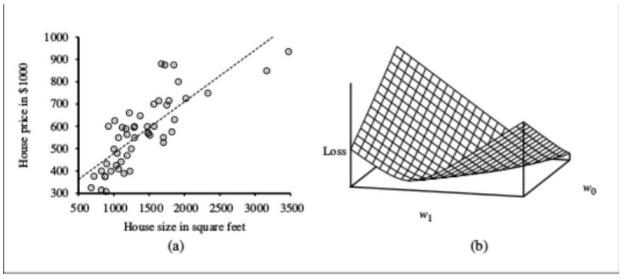

FIGURE 3 – (a) représente le prix des maisons en fonction de leur surface qui sont en vent àBerkeley.
(b) montre un exemple d'ensemble d'apprentissage de n points sur le plan x,y

2.3.2 Les k plus proches voisins

L'algorithme des **K-Nearest Neighbors (KNN)(K plus proches voisins)**est un algorithme de classification supervisé.Chaque observation de l'ensemble d'apprentissage est représentée par un point dans un espace à n dimensions ou n est le nombre de variables prédictives.Pour prédire la classe d'une observation ,on cherche les k points les plus proches de cet exemple [18].

Avantages :
-Apprentissage rapide.
-Méthode facile à comprendre.
-Adapté aux domaines ou chaque classe est représentée par plusieurs prototypes et ou les frontiéres sont irréguliéres .

Inconvénients :
-Prédiction lente car il faut revoir tous les exemples à chaque fois.
-Méthode gourmande en place mémoire .
-sensible aux attibuts non pertinents et corrélés.
-Particuliérement vulnérable au fléau de la dimensionnalité.

Pour k=3 la classe majoritaire du point central est la classe B ,mais si change la valeur du voisinage k=6 la

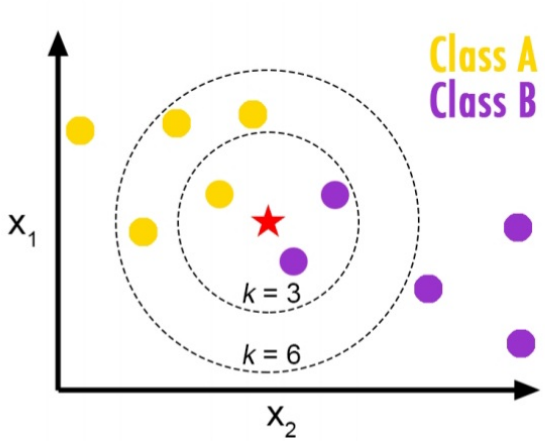

FIGURE 4 – KNN classe majoritaire

class majoritaire devient la classe A.

2.3.3 Le Classifieur Naïf de Bayes

Le **classifieur naïf de bayes**est un algorithme supervisé probabiliste que la présence d'une caractéristique particulière dans une classe n'est pas liée à la présence de toute autre caractéristique.
il est principalement utilisé à des fins de regroupement et de classification dépend de la probabilité conditionnelle de se produire[34].

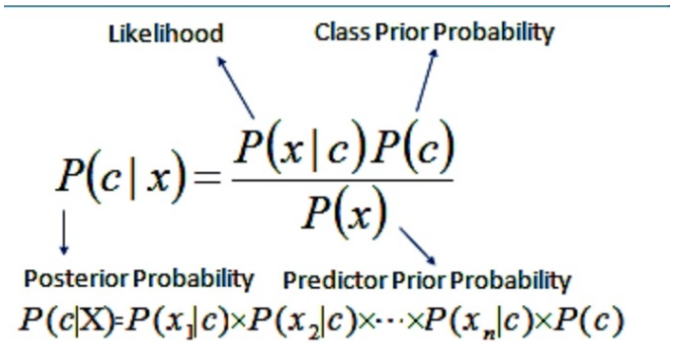

FIGURE 5 – Naive Bayes.

Avantages :
-L'algorithme offre de performance.

Inconvénients :
-La prédiction devient erronée si l'hypothése indépendance conditionnelle est invalide .

2.3.4 K-means

L'algorithme des **K-moyennes (k-means)** est un algorithme non Supervisé le plus simples qui résolvent le problème de clustering bien connu.La procédure suit un moyen simple et facile de classer un ensemble de données donné(Dataset)à travers un certain nombre de clusters.Les étapes de algorithme sont[78] :

— Choisir k points qui représentent la position moyenne des cluster.

— répéter jusqu'à stabilisation des points centraux :
-affecter chacun des M points au plus proche des k points centraux.
-metter à jour les points centraux en calculant les centres de gravité des k cluster.

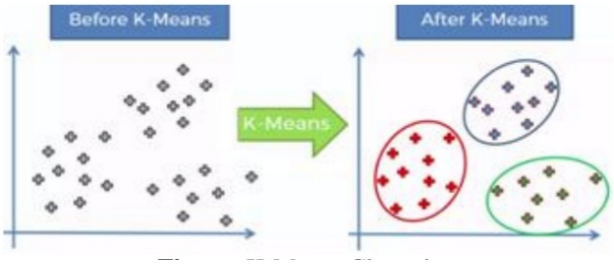

FIGURE 6 – K-means Clustering

Avantages :
-Implementable pour des grandes volumes de données.

Inconvénients :
-Le choix du parametre K n'est pas découvert mais choisi par l'utilisateur.
-La solution dépend des K centre de gravité choisi lors de l'initialisation.

2.3.5 Les arbres de décision

Un **arbre de décision (decision tree)** est un algorithme d'apprentissage supervisé qui va permettre la prise de décision en prenant en entrée une population ,un échantillon pour ensuite procéder à une catégorisation basée sur des facteurs discriminants.Cet outil va donc répartir les individus en groupes homogènes et va émettre des prédictions à partir de données connu[60].Un arbre de décision se présente comme sur la figure

Avantages :
-Peu de préparation des données.
-Performance sur de grandes jeux de données.

Inconvénients :
-L'existence d'un risque de sur-apprentissage si l'arbre devient trés complex.

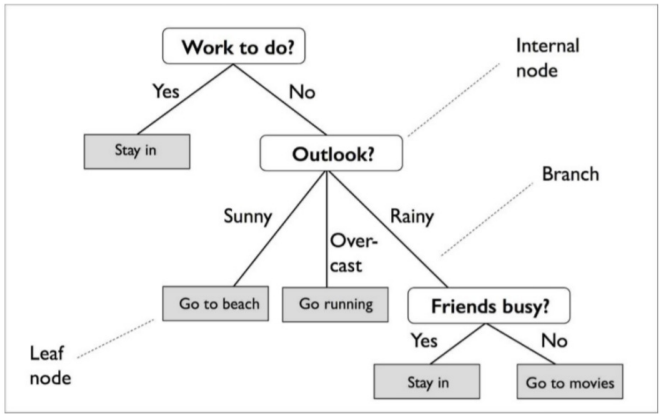

FIGURE 7 – Exemple d'arbre de décision

2.3.6 Les forêts aléatoires

Les algorithmes de **forêts aléatoires(Random Forest ou RF)** sont connus pour être des outils très efficaces de classification dans de nombreux domaines,notamment en finance .Il s'agit d'une méthode de classification d'ensemble,qui établit un ensemble de classificateurs,contrairement aux arbres de décision CART et C5.0 qui ne construisent qu'un classificateur[16].

Avantages :
-C'est un des meilleur algorithme pour ce qui est de la précision.
-Incorporation de la validation croisée.

Inconvénients :
-Une implémentation difficile.

2.3.7 Les machines à vecteur de support

Les machines à vecteurs de support**(Support Vector Machine ou SVM)** sont des algorithmes d'apprentissage Supervisé ,utiles tant pour les problèmes de classification que régression , et dont l'objectif est de séparer les données en classe à l'aide d'un séparateur que l'on appellera "frontière" et qui va maximiser la distance entre ces classes,appelée "Marge"[29] .
cas[51].

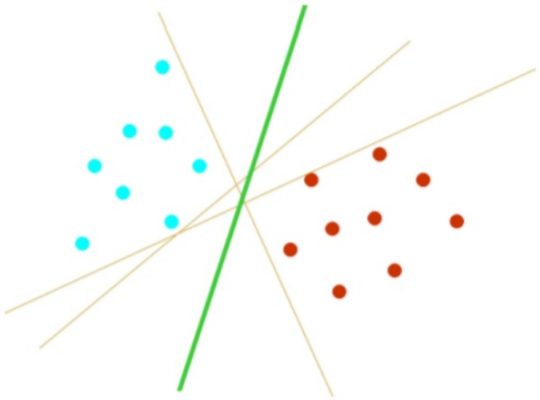

FIGURE 8 – Illustration d'un hyperplan de séparation optimale

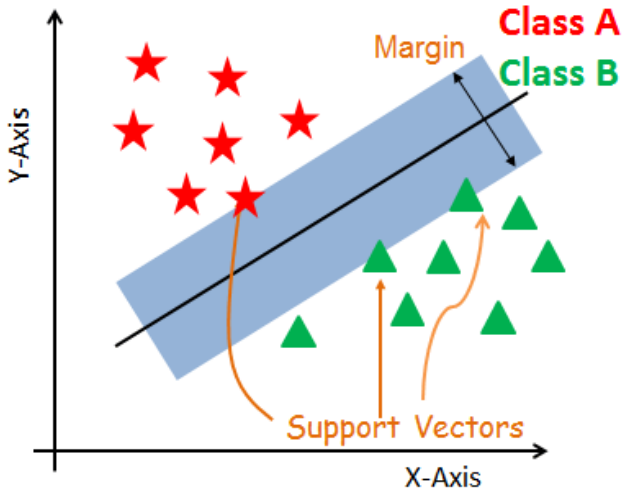

FIGURE 9 – Support Vector Machine

Avantages :
-Sa grande précision de prédiction .
-Fonction bien sur de plus petits data sets .
-Ils peuvent etre plus efficace car ils utilisent un sous-ensemble de points d'etrainement.

Inconvénients :
-Ne convient pas à des jeux de données plus voluminex,car le temps d'entrainement avec les SVM peut etre long.
-Moins efficace sur les jeux de données contenant du bruits et beaucoup d'outiliers[19].

2.3.8 Les Réseaux de Neurones

Les **Réseaux de Neurones**
sont une série d'algorithmes qui s'efforcent de reconnaitre les relations sous jacentes dans un ensemble de données grace àun processus qui imite le fonctionnement du cerveau humain. En ce sens,les réseaux de neu-

rones font référence à des systemes de neurones,qu'ils soient de nature organique ou artificielle[25].

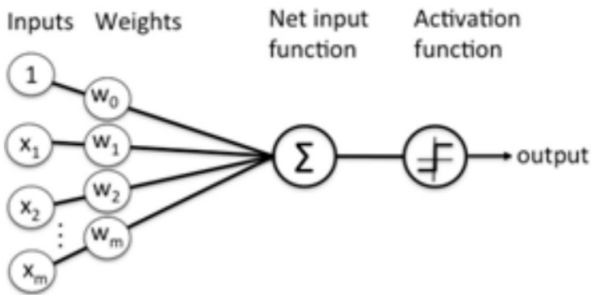

FIGURE 10 – Les Réseaux de Neurones

2.3.9 Performance et sur-apprentissage

Le **sur-apprentissage ou (Overfiting)**désigne le processus dans lequel un modèle s'adapte tellement aux données historiques qu'il en devient inefficace pour des prédictions futures.L'algorithme trouvera donc des relations dans les données d'entrainement qui au final ne s'appliquent pas dans le cas des données étudiés[71].

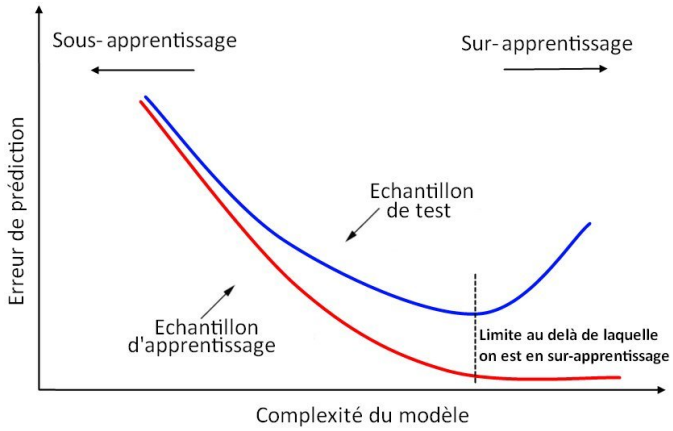

FIGURE 11 – Sur-Apprentissage

Le **sous-apprentissage(Underfitting)** est l'inverse et désigne le cas ou l'algorithme n'apprend pas assez de relations que pour faira des prédictions précise,et se caractérise par une faible variance,mais un biais élevé .

Le meilleur modéle est celui du juste milieu,il ne doit souffrir ni d'Underfitting ni d'Overfitting.

Pour résoudre ce ,on divise les données en deux Groupes distincts. Le premier sera **L'ensemble d'apprentissage**.Le deuxième sera **l'ensemble de test**.

Pour avoir une bonne séparation des données en données d'apprentissage et données de test ,on utilise **La validation Croisée**.

L'idée c'est séparer aléatoirement les données dont on dispose en k parties,une fera office d'ensemble de

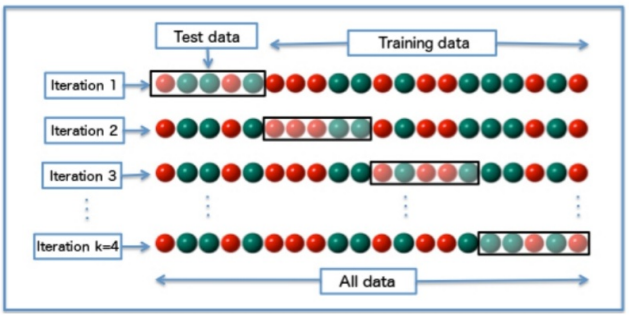

FIGURE 12 – La Validation Croisée

test et les autres constitueront l'ensemble d'apprentissage.
-Après que chaque échantillon ait été utilisé une fois comme ensemble de test[9] .

3 Conclusion

L'apprentissage automatique **(Machine Learning)** peut étre Supervisé ou Non Supervisé ,si nous avons moins de données et des données Clairement étiqutées pour la formation **(Training)**,optez pour l'apprentissage Supervisé.

L'apprentissage Non supervisé donnerait généralement de meilleures performances et résultats pour les grands ensemble de données.

Chapitre 2

L'Apprentissage Supervisé

L'apprentissage Supervisé est un paradigme d'apprentissage Automatique permettant d'acquérir les informations de relation d'entrée-sortie d'un systeme sur la base d'un ensemble de donné d'échantillons d'apprentissage d'entrée-sortie appariés[54].
comme la sortie est considée comme l'étiquette des données d'entrée ou de la supervision ,un échantillon d'apprentissage d'entrée-sortie est également appelé données d'apprentissage étiquetées ou données supervisées[46].
le but de l'apprentissage supervisé est de construire un systeme artificiel qui peut apprendre le mappage entre l'éntrée et la sortie, et peut prédire la sortie du systeme compte tenu de nouvelles entrées . si la sortie prende un ensemble fini d e v aleurs d iscrétes q ui i ndiquent l es é tiquettes d e c lasse d e l 'entrée,le mappage appris conduit à la classification des données d'entrée . si la sortie prend des valeurs continues, cela conduit à une régression de l'entrée. Les informations sur la relation entrée-sortie sont fréquemment représentées avec des paramètres de modèle d'apprentissage. Lorsque ces paramètres ne sont pas directement disponibles à partir d'échantillons d'apprentissage, un système d'apprentissage doit passer par un processus d'estimation pour obtenir ces paramètres[4].

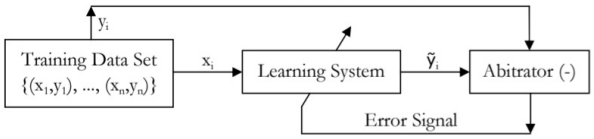

FIGURE 13 – Schéma fonctinnel illustrant la formde L'apprentissage Supervisé

Figure 13 montre un schéma fonctionnel qui illustre la forme d'apprentissage supervisé dans ce schéma x_i, y_i) est un exemple de formation supervisée où'x' représente l'entrée du système,'y' représente la sortie du système, et 'i' est l'indice de l'échantillon d'apprentissage , pendant un processus d'apprentissage supervisé, une entrée d'apprentissage x_i est transmise au système d'apprentissage, et le système d'apprentissage génère une sortie y_i[12].

4.1 Définition Mathématique

Une base de données d'apprentissage (ou ensemble d'apprentissage) est un ensemble de couples entrée-sortie $(x_n, y_n)_{1 \leq n \leq N} avec x_n \in X x_n \in X et y_n \in Y y_n \in Y$[47].
La méthode d'apprentissage supervisé utilise cette base d'apprentissage pour déterminer une estimation de f notée g et appelée indistinctement fonction de prédiction, hypothèse ou modèle qui à une nouvelle entrée x associe une sortie $g(x)$.
On distingue trois types de problèmes solubles avec une méthode d'apprentissage automatique supervisée :
-$Y \subset R$: lorsque la sortie que l'on cherche à estimer est une valeur dans un ensemble continu de réels, on parle d'un problème de **régression** . La fonction de prédiction est alors appelée **un régresseur**.
-$Y = \{1, \ldots, I\}$: lorsque l'ensemble des valeurs de sortie est fini, on parle d'un problème de **classification** La fonction de prédiction est alors appelée **un classifieur**.
-Lorsque$Y$ est un ensemble de données structurées, on parle d'un problème de prédiction structurée, qui revient à attribuer une sortie complexe à chaque entrée[17].

4.1.1 Avantages et Inconvénients

le principal avantage de l'apprentissage supervisé est que toutes les classes ou sorties analogiques manipulées par l'algorithme de ce paradigme sont significatives pour les humains.et il peut être facilement utilisé pour la classification des modèles discriminatoires et pour la régression des données. Mais il présente également des inconvénients serval.le premier est causé par la difficulté à collecter les supervisions ou les labels[31]. lorsqu'il y a un énorme volume de données d'entrée, par exemple, ce n'est pas une tâche triviale d'étiqueter un énorme ensemble d'images pour la classification des images.deuxièmement, car tout dans le monde réel n'a pas une étiquette distinctive, il y a des incertitudes et des ambiguïtés dans la supervision ou les étiquettes.par exemple, la marge pour séparer les deux concepts,ces difficultés peuvent limiter l'application du paradigme d'apprentissage supervisé dans certains scénarios[62].

4.1.2 Applications Apprentissage Supervisé

L'apprentissage supervisé permet à une machine d'apprendre le comportement humain ou le comportement d'un objet dans certaines tâches.les connaissances apprises peuvent ensuite être utilisées par la machine pour effectuer des actions similaires sur ces tâches.l'apprentissage supervisé ont été utilisés avec succès dans des domaines tels que la recherche d'informations[12].

1-Vision Par Ordinateur

La vision par ordinateur est un domaine scientifique et branche de l'intelligence artificielle qui traite de la façon dont les ordinateurs peuvent acquérir une compréhension de haut niveau à partir d'images ou de vidéos numériques[40].

2-Reconnaissance De Formes

La reconnaissance de formes (ou parfois reconnaissance de motifs) est un ensemble de techniques et méthodes visant à identifier des motifs informatiques à partir de données brutes afin de prendre une décision dépendant de la catégorie attribuée à ce motif[24].

3-Reconnaissance de l'écriture Manuscrite

La reconnaissance de l'écriture manuscrite (en anglais, handwritten text recognition ou HTR) est un traitement informatique qui a pour but de traduire un texte écrit en un texte codé numériquement[18].

4-Reconnaissance automatique de la parole

La reconnaissance automatique de la parole (souvent improprement appelée reconnaissance vocale) est une technique informatique qui permet d'analyser la voix humaine captée au moyen d'un microphone pour la transcrire sous la forme d'un texte exploitable par une machine[2].

5-Traitement Automatique des Langues

est un domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle, qui vise à créer des outils de traitement de la langue naturelle pour diverses applications. Il ne doit pas être confondu avec la linguistique informatique, qui vise à comprendre les langues au moyen d'outils informatiques[55].

6-Bio-informatique

La bio-informatique, ou bioinformatique, est un champ de recherche multi-disciplinaire de la biotechnologie où travaillent de concert biologistes, médecins, informaticiens, mathématiciens, physiciens et bio-informaticiens, dans le but de résoudre un problème scientifique posé par la biologie[58].

4.1.3 Méthodes D'apprentissage Supervisé

1-La régression Linéaire.

2- Les K plus proches voisins.

3- Le Classifieur Naif de Bayes.

4- K-means .

5- Les Arbres de Décision.

6- Les Foréts Alétoires .

7- Les Machines àVecteur de support .

-Nous avons déja discuté dans le section précédente ,mais nous avons définir deux algorithmes importants en classification avec détails à savoir KNN et SVM :

4.1.4 Le principe de fonctionnement de l'algorithme KNN

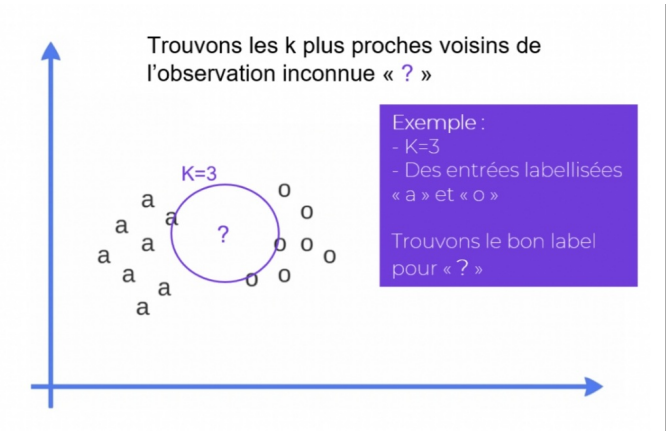

FIGURE 14 – L'algorithme du K-NN et son princip de fonctionnement par l'éxemple

En résumé les étapes da l'algorithme sont les suivantes :

- On choisit une fonction de définition pour la distance entre observation.
- On fixe une valeure pour K , nombre de plus proches voisins.

Pour une nouvelle observation inconnue en entrée dont on veut prédire sa variable de sortie, il faut faire :
-**Etape 1** Calculer toutes les distances entre cette observation en entrée et les autres observation du jeu de données.
-**Etape 2** Conserver les K du jeu de données qui sont les plus proches de l'observation à prédire.
-**Etape 3** Prendre les valeurs des observations rtenues :
-Si on effectue une régression , l'algorithme calcule la moyenne (ou la médiane) des valeurs des observations retenues.
-Si on effectue une Classification , l'algorithme assigne le label de la classe Majoritaire à la données qui était inconnue.
-**Etape 4** Retourner la valeur qui a été prédite poar l'algorithme pour l'observation en entée qui était inconnue[7].

2)-**Métriques d'évaluation de la similarité entre les observations**
Comme on vient de le dire,pour mesurer la proximité entre les observations,on doit imposer une fonction de similarité à l'algorithme.
Cette fonction qui calcule la **Distance** entre deux observations estime l'affinité entre l'observation[37].

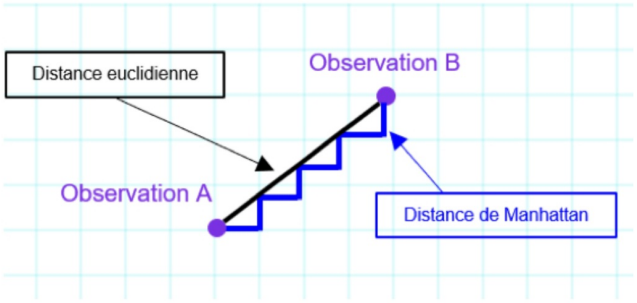

FIGURE 15 – Distance entre deux points :distance Euclidienne versus distance de Manhattan

Parmi les fonctions de similarité les plus connues,il y a la **Distance Euclidienne**.C'est cette fonction que nous avons utilisées dans notre exemple plus haut.

La **Distance de Manhattan** peut etre intéressante pour des données qui ne sont pas du méme type (c'est-àdire des données qui n'ont pas été misee sur la meme échelle)[75]

4.1.5 Les Machines à Vecteur de Support

-Ce modèle a été rapidement adopté en raison de sa capacité à travailler avec des données de grandes dimensions, ses garanties théoriques et les bons résultats réalisés en pratique. Requérant un faible nombre de paramètres, les SVM sont appréciées pour leur simplicité d'usage[29].
Le Principe des SVM consiste à ramener un problème de classification ou de discrimination à **un hyperplan (feature space)** dans lequel les données sont séparées en plusieurs classes dont la frontière est la plus éloignée possible des points de données (ou "marge maximale"). D'où l'autre nom attribué aux SVM :

Ces méthodes reposent sur deux idées clés : la notion de marge maximale et la notion de fonction noyau.
les support vector machines font appel à des noyaux, c'est-à-dire des fonctions mathématiques permettant de projeter et séparer les données dans l'espace vectoriel, les "vecteurs de support" étant les données les plus proches de la frontière. C'est la frontière la plus éloignée de tous les points d'entraînement qui est optimale, et qui présente donc la meilleure capacité de généralisation[51].

La fonction mathématique utilisée pour la transformation est appelée **noyau**.Ces fonctions permettent de

séparer les données en les projetant dans un feature space(un espace vectoriel de plus grande dimension comme Figure ci dessous [52].

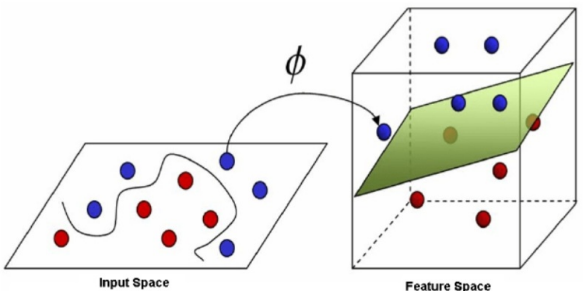

FIGURE 16 – Les SVM permettent de Projeter les dpnnées dans une espace de plus grande dimension via une fonction noyau pour les séparer linéairement

La technique de maximisation de marge permet quant à elle de garantir une meilleure robustesse face bruit et donc un modele plus généralisable.

-Les types de noyaux suivants :

-Linéaire.
-Polynomial.
-Fonction radial de base (RBF).
-Sigmoide[76].

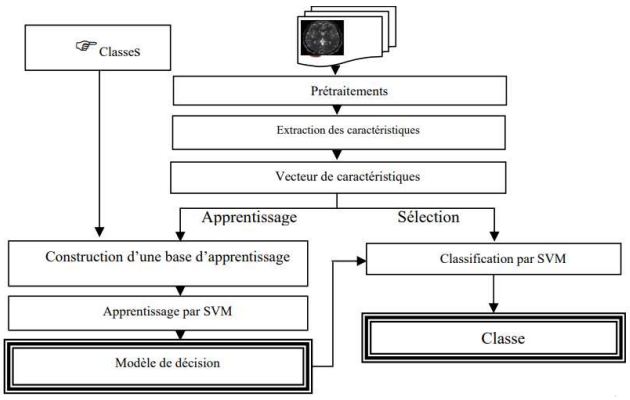

FIGURE 17 – Processus d'Apprentissage avec modele SVM

4.2 La Classification

Le classement automatique ou classification supervisée est la catégorisation algorithmique d'objets. Elle consiste à attribuer une classe ou catégorie à chaque objet (ou individu) à classer, en se fondant sur des données statistiques. Elle fait couramment appel à l'apprentissage automatique et est largement utilisée en reconnaissance de formes[45].

4.2.1 Les Types de Classification

4.2.2 La Classification Binaire

La classification Binaire est la tache de classer les éléments d'un ensemble en deux groupes sur la base d'un régle de classification[66].

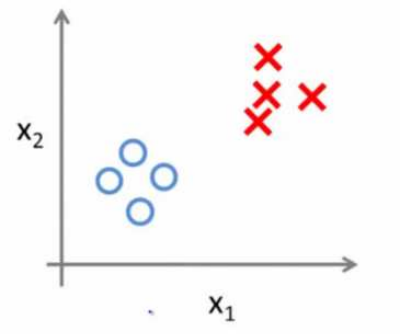

FIGURE 18 – Classification Binaire

-Problémes typiques de classification Binaire :

-1) Des **Tests Médicaux** pour déterminer si un patient souffre ou non de certaines maladies.
-2) **Le Controle de la Qualité** dans l'industrie pour déterminer si une spécification a été respectée.
-3) dans **la Rcherche d'informations** décider si une page doit étre dans l'ensemble de résultats d'une recherche ou non[36].

4.2.3 Classification à Classes Multiples

Classification à Classes Multiples est un processus de réparition d'un lot de propositions entre plus de deux ensembles[28].

Technique de transformation des problémes

-Cette technique traite des stratégies pour réduire le probleme de la classification Multiclasses à de multiples problemes de la classification binaire.

1)-**One-vs Rest** One vs All

La stratégie consiste à former à un classificateur unique perclass avec les échantillons de ce classé comme échantillons positifs et tous les autres échantillons negativs .

2)-**One -vs One**

Un classificateur binaire de trains pour le probléme de multiclasse de maniére à K ,chacun reçoit les chantillons d'une paire de classes de l'ensemble d'entrainement original,et doit apprendre à distinguer ces deux classes[63].

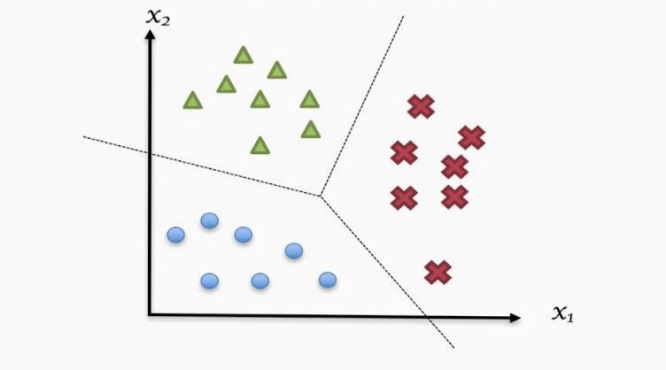

FIGURE 19 – Classification à Classes Multiples

4.2.4 Mesure d'évaluation pour les modéles de classification

-L'une des maniéres les plus rependues pour Mesurer la performance d'un modele de classification est **la Matrice de Confusion**.Cette derniére correspond à un résume tabulaire du nombre de prédictions correctes et non corrects.faites par le modele .Dans cette matrice chaque ligne correspond a une classe réelle et chaque colonne correspond a une classe estimée.
Elle inclut les valeures suivantes :

1)-**Vrais Positifs ou (True Positive TP)** soit lorsque la classe réelle et la classe estimée sont toutes les deux positives.

2)-**Vrais Négatifs ou (True Negative TN)** soit lorsque la classe réelle et la classe estimée sont toutes les deux negatives.

3)-**Faux Positifs ou (False Positive FP)** soit lorsque la classe réelle est négative mais que la classe estimée est positive.On appelle ceci une Erreur de Tupe 1.

4)-**Faux Négatifs ou (False Negative)** soit lorsque la classe réelle est positive mais que la classe estimée negative.On appelle ceci une Erreur de Type 2.

Dans le cas d'une classification binaire,la Matrice Confusion sera une matrice de deux par deux,avec quatre valeurs,comme dans la photo suivante[39] :

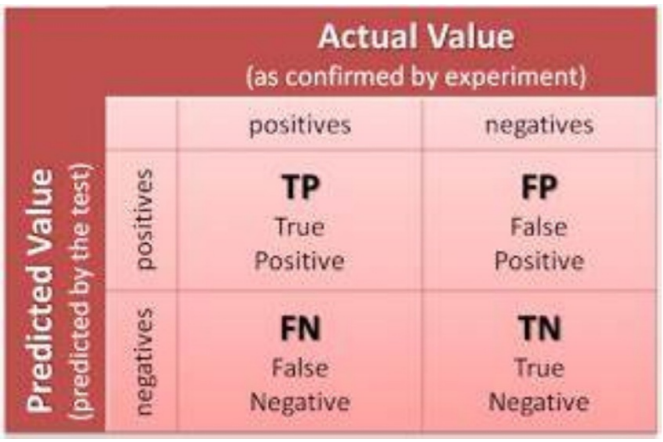

FIGURE 20 – Matrice de Confusion pour la classification binaire

4.2.5 Mesures de Classification

Parmi les Mesures de classification,on trouve la accuracy,la précision,le rappel,la spécificité et le score F1[48].
a. Accuracy :

La Accuracy correspond au nombre de prédictions correctes faites par le modéle.Elle représente le ratio entre le nombre de prédictions correctes et le nombre total de prédiction.En utilisant la formule suivante :
$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$

b. Précision :

La précision correspond au nombre d'éléments corrects rendus par le modéle.Cela correspond au ratio entre le nombre de classifications positives correctes et le nombre total de précision positives.Elle peut etre calculée avec la formule suivante :
$Precision = \frac{TP}{TP+FP}$

c. Rappel ou (Recall) :

La rappel détermine la proportion des valeurs positives qui ont été prédites avec précision.Cette mesure correspond donc au ratio entre le nombre total de classifications de classe positives.On peut utiliser la formule suivante :
$Sensitivity = \frac{TP}{TP+FN} = 1 - Type2error$

d. Spécificité :

L spécificité correspond au nombre de classes négatives prédites par le modéle.Cette mesure est déterminée par le ratio entre le nombre de prédictions négatives correctes et le nombre total de prédictions négatives.Elle peut se calculer de la maniére suivante :
$Specificity = \frac{TN}{TN+FP} = 1 - Type1error$

e. Score F1 :

Le score F1 correspond à une combinaison des mesures de rappel et de précision.Cette mesure est utilisée lorsqu'une distinction claire ne peut pas etre faite entre ces deux mesures ou lorsque les False Negative et False Positive sont les plus important.
$F1score = 2 * \frac{Sensitivity*Precision}{Sensitivity+Precision}$

4.3 Classification des Images

La Classification d'Images est un probleme fondamental en **Vision Par Ordinateur** ,qui a de nombreuses applications concrétes,et la reconnaissance de formes consistant à attribuer automatiquement une classe à une image à l'aide d'un systeme de la classification.Dans l'aproche Supervisé ,chaque image est associée à une étiquette qui décrit sa classe d'appartenance.

Un systeme de classification automatique d'images est composé des étapes suivantes :
-L'étape de **Prétraitement** permettant de "nettoyer"les images[57].
-La phase **d'extraction de caractéristiques** permettant de décrire l'information pertinente contenue dans l'image à l'aide d'opérateurs ou de descripteures discriminants.
-La phase **d'apprentissage** permettant de construire une frontiére de décision pour identifier la classe d'une image présentée à l'entrée du systeme.Ces trois phases sont essentielles dans la construction du systeme de classification[75].

Les techeniques d'extraction de caractéristiques peuvent également se scinder en caractéristiques bas-niveau utilisant par exemple l'iformation au niveau du pixel dans l'image et en caractéristique de plus haut niveau avec des descripteurs utilisant notamment une représentation texturée de l'image.

La phase d'apprentissage consiste en la construction d'une régle de décision soit à partir d'un modele.Les méthodes standards de la littérature utilisent un seul classifieur pour la construction de la régle de décision,et l'éstimation de la fonction de décision se fait à partir d'une seul hypothese.Des approches plus flexiblrs que la formulation d'une hypothese considérrent plutot une combinaison de classifieurs permettant d'améliorer les performances de ces memes classifieurs pris individuellement[1].

La phase déxtraction de caractéristiques peut etre précédée d'une phase dite de prétraitement.Cette phase a pour but de nettoyer l'image,c'est-à-dire d'isoler le contenu informatif ou d'intéret dans l'image,et permet ainsi d'occulter ou d'atténuer toute information sisceptible de nuire à la description du contenu pertinent lors de la phase d'extraction de caractéristiques.On retrouve ainsi des techeniques d'atténuation de bruits,de renforcement de contours,des techeniques d'amélioration de l'image comme le réhaussement de contrast,la réduction de la dimension de l'image par la binarisation,la réduction de l'image à ses primitives visuelles comme la squelettisation ou encore l'extraction de contours à l'aide de techeniques de filtrage[**?**].

Enfin,pour connaitre les performances d'un systeme de classification d'images,il est nécessaire de définir une procédure d'évaluation consistant notamment dans le choix d'une mesure donnant les performances du systeme et des données d'évaluation.La problématique de l'évaluation d'un systeme de classification est d'autant plus dificile qu'il n'existe pas de consenus sur la méthodes,les mesures et les procédés d'évaluation[30].

4.3.1 Les Motivations de Classification d'images

L'objectif de **la classification d'images** est d'élaborer un systeme capable d'affecter une classe automatiquement à une image.Ainsi,ce systeme permet d'effectuer une tache d'expertise qui peut s'avérer couteuse à acquérir pour un etre humain en raison notamment de contraintes physiques comme la concentration,la fatigue ou le temps nécessité par un volume important de données images[33].

4.3.2 Extraction de caractéristiques

La phase d'extraction de caractéristiques constitue généralement l'une des phases le plus importantes dans l'élaboration du system.Il s'agit en effet de déterminer un espace numérique de description dans lequel les données images seront projetées et permerttent une séparation optimale des classes en présence.Nous retrouvons des descripteurs de **bas niveau** s'intéressant à l'information contenue dans l'image au niveau du pixel et des descripteurs de plus **haut niveau** nécessitant une réprésentation intermédiaire de l'image plus adaptée[42].

-Nous présentons ci-aprés ces deux approches :
1)Extracteurs de bas niveau :

les Extracteurs de bas niveau permettent de traduire l'information présente au niveau du pixel,sans tenir compts des formes ou des patterns dans l'image.Parmi les caractéristiques extraites,nous retrouvons l'intensité du pixel brut,l'histogramme des intensités de pixels,les statistique sur cet histogramme(moyenne,entropie,variance, coefficient d'aplatissement,asymétrie),la densité de pixels.On rencontre également des extracteurs de niveau intermédiaire traduisant des informations comme des liaisons entre les pixels,des distances,leur localisation,le contraste dans l'image.

2)Extracteurs de plus haut-niveau :

Les méthodes d'extraction de plus haut niveau tiennent compte des formes et des structures dans l'image,des relations spatiales entre les pixels ou ces structures.Les propriétés les plus recherchées dans ces extracteurs sont outre leur pouvoir descriptif,l'invariance et la bustesse à différentes transformations pouvant affecter l'image.Ainsi,la description obtenue demeure relativement inchangée face à ces transformations pouvevant plus ou moins affecter des contenus identiques ou similaires dans les images.On retrouve couramment les invariances au changement d'échelle,de prespective,aux transformations affines comme la translation,la rotation et la robustesse au changement de luminosité ou de contraste[30].

4.3.3 La Méthode de résolution

Le probléme de classification d'images est posé formellement de la maniére suivante :
-Il ya K classes d'images possibles.L'ensemble $\{0, 1, ..., K-1\}$ définit les labels des différentes classes (exemple : 0 = "oiseau" et 1 = "chien".
-Nous avons une collection de N images en entrée : $\{X_i\}_{i\in\{1,...,N\}}$

-Les classes des N images sont connues à l'avance : chaque image X_i est étiquetée par $y_i \in \{0, 1, ..., K-1\}$.

-Le but est de classifier correctement une nouvelle image,dont on ne connait pas la classe :on veut trouver la bonne étiquette y' de X'[68].

4.3.4 Le traitement d'images

Est une branche du traitement de signal dédiée aux images et vidéo.

Le traitement d'images est l'ensemble des opérations effectuées sur l'image,afin d'en améliorer la lisibilité et d'en faciliter l'interprétation.C'est,par exemple,la cas des opérations de rehaussement de contraste,élimination du bruit et correction d'un flou.

C'est aussi l'ensemble d'opérations effecttuées pour extraire des informations de l'image comme la segmentation et l'extraction de contours.

Avant le traitement d'images,on peut aussi effectuer des opérations de prétraitement qui sont toutes les techeniques visant à améliorer la qualité d'une image.De ce fait,la donnée de départ est l'image initiale et le resultat est également une image[35].

4.3.5 C'est quoi une image

Une **image** est une représentation planaire d'un objet quelconque.
Mathématiquement,c'est une fonction bidimensionnelle de la forme $f(x, y)$, ou $f(x0, y0)$ .Donc,c'est un processus continu 2D résultat d'une mesure physique.L'amplitude de f est appelée intensité ou niveau de gris de l'image au point de coordonnées (x, y)[22] .

4.3.6 Définition de l'image numérique

Le terme d'image numérique désigne dans son sens le plus générale,toute image qui a été acquise,traitée et sauvgardée sous une foeme codée representable par des nombres.

La numérisation est le processus qui permet de passer de l'état d'image phisyque qui est caractérisée par l'aspect continu du signal qu'elle représente à l'état d'image numérique qui est caractérisée par l'aspect discret(l'intensité lumineuse ne peut prendre que des valeurs quantifiées en un nombre fini de points distincts).C'est cette forme numérique qui permet une exploitation ultérieure par des outils logiciels sur ordinateur[3].

4.3.7 Types d'images

1)-**Images matricielles :**

Dans la description que nous avons faite jusqu'a présent des images en utilisé une matrice.On dit alors que l'image est matricielle .Ce type d'image est adapté à l'affichage sur écran mais peu adapté pour l'impression car bien souvent la résolution est faible[11].

2)-**Images vectorielles :**

Le principe des images vectorielles est de représenter les données de l'image à l'aide de formules mathématique.Permet d'agrandir l'image indéfiniment sans pert de qualité et d'obtenir un faible encombrement[27].

3)-**Représentation spatiale :**

Elle est fait directement à partir des échantillons d'une image dans le domaine spatial.Une image 2D $f(x, y)$ scalaire réelle peut etre vue comme une surface en 3D.Ce qu'on voit lorsqu'on regarde l'image est une correspondance entre niveau de gris et grandeur physique.

4)-**Représentation fréquentielle :**

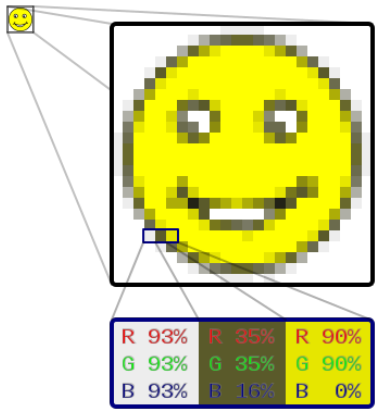

FIGURE 21 – Exemple d'image matricielle

C'est une représentation obtenue à partir d'une transformation de l'image dans le domaine fréquentiel[73].

-explication c'est digramme :

Acquisition : c'est premier étape dans le traitement de l'image.Elle est essentielle on ne peut décrire,extraire ou améliorer quelques choses qui n'existent pas.

Amélioration : parmi les traitements les plus simples et les plus utilisé mmettre en relief les détails ou faire ressortir certaines caractéristiques.

Restauration d'images : amélioration de l'images ayant subies des dégradations bougé de caméra.

Traitement des images couleurs : domaine qui prend de l'importance en raison du développement d'internet.

Ondelettes et multirésolutions : fondements pour la représentation à différents degrés de résolution représentation,extraction d'attributs, compression...

Compression : réduction de la quantité d'informations véhiculées par une images stockage,transmission de données...

Traitements morpholigiques : ensemble d'outils pour extraire des composantes de l'image représentation et description des formes.

Segmentation : procédure de partitionnement de l'image en ses composantes ou objets reconnaissance des formes.

Représentation et description : intervient généralement aprés une segmentation,conversion des résultats obtenus sous une forme convenable pour la suite du traitement.La description peut etre vue comme une sélection de caractéristiques, classification.

Reconnaissance des formes : assignation d'un label à un objet en se basant sur ses descripteurs.

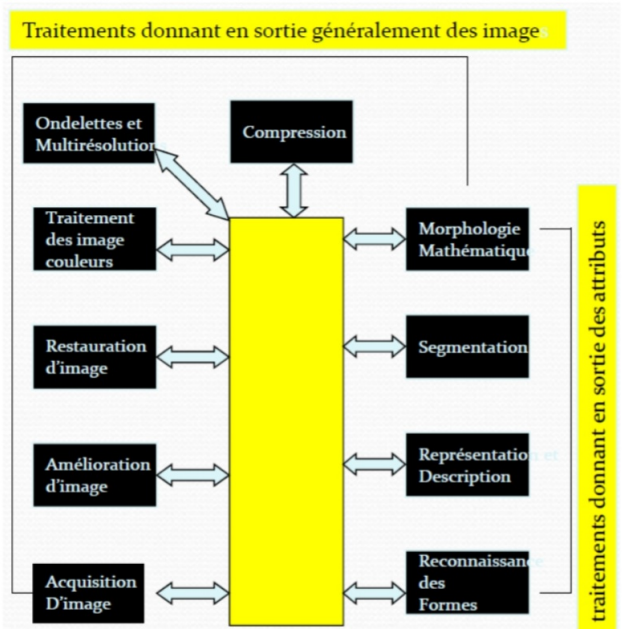

FIGURE 22 – Diagramme résume l'ensemble des traitements qui peuventetre appliqués à l'image

4.3.8 Filtrage des images

Le principe du **filtrage** est de modifier la valeur des pixels d'une image,généralement dans le but d'améliorer son aspect.En pratique ,il s'agit de créer une nouvelle image en se servant des valeurs des pixels de l'image d'origine[6].

4.3.9 Rehaussement des images

Le rehaussement des images est essentiellement un processus qui permet de faciliter l'interprétation visuelle d'une image.Le rehaussement des contrastes se fait en changeant les valeurs initiales de façon à utiliser toutes les valeurs possibles,ce qui permet d'augmenter le contraste entre les cibles et leur environnement.Pour bien comprendre comment fonctionne ce type de rehaussement ,il faut premierement comprendre le concept de **l'histogramme** d'une image[53].

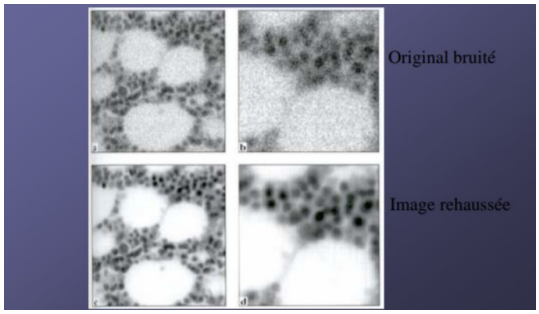

FIGURE 23 – Exemple sur image rehaussée

4.3.10 L'ihstogramme d'une image

Un histogramme d'une image est une représentation graphique des valeurs numériques d'intensité qui composent une image.Ces valeurs apparaissent le long de l'axe des x du graphique .La fréquence d'occurrence de chacune de ces valeurs est présenté le long de l'axe des y .Il est possible de produire différents types de rehaussement[8] .

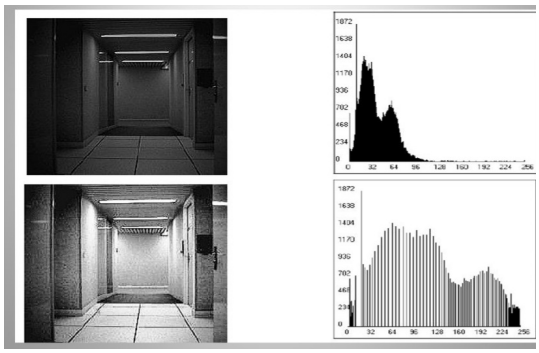

FIGURE 24 – L'histogramme d'une image

4.3.11 Transformation des images

La transformation d'images est un procédé qui implique la manipulation de plusieurs bandes de données,que ce soit pour transformer une image provenant d'un capteur multispectral ou pour transformer plusieurs images de la mème région prises à des moments différents .
La transformation d'images génère une "nouvelle" image en combinant les différentes dources d'information de manière à rehausser certaines caractéristiques ou certaines propriétés des données que sont,moins évidentes dans l'image originale[5].

4.3.12 Méthodes de classification supervisé

L'utilisation d'une méthode de classification supervisé,l'analyste identifie des échantillons assez homogénes de l'image qui sont représentatifs de différents types de surfaces.Ces échantillons forment un ensemble de données-tests.La sélection de ces données-tests est basée sur les connaissances de l'analyse,sa familiarité avec les régions géographiques et les types de surfaces présents dans l'image.L'analyste supervise donc la classification d'un ensemble specifique de classes[20].
Les informations numériques pour chacune des bandes et pour chaque pixel de ces ensemble sont utilisées pour l'ordinateur puisse définir les classes et ensuit reconnaitre des rigions aux propriétés similaires à chaque classe.
L'ordinateur utilise un programme spécial ou algorithme afin de déterminer la "signature" numérique de chacune des classes.Plusieurs algorithmes différents sont possibles[13].

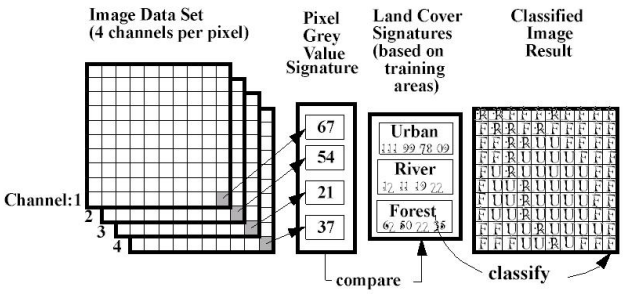

FIGURE 25 – Les étapes de classification supervisées

5 Conclusion

Dans cette chapitre ,nous avons vu qu'est ce que l'apprentissage supervisé ,et méthodes de classification SVM et KNN.
-L'utilisation des images numériques et le traitement de l'image peut viser à réduire le bruit.
Nous avons vu quelque techenique de traitement et l'analyse d'image pour faire la classification supervisé .

Chapitre 3

Contribution

Introduction

La tumeur cérébrale ou Brain Tumor est l'une des maladies les plus rigoureuses de la science médicale. Une analyse efficace et efficiente est toujours une préoccupation essentielle pour le radiologue dans la phase prématurée de la croissance tumorale. Le classement histologique, basé sur un test de biopsie stéréotaxique, est l'étalon-or et la convention pour détecter le grade d'une tumeur cérébrale. La procédure de biopsie nécessite que le neurochirurgien perce un petit trou dans le crâne à partir duquel le tissu est prélevé. Il existe de nombreux facteurs de risque impliquant le test de biopsie, notamment des saignements de la tumeur et du cerveau provoquant une infection, des convulsions, une migraine sévère, un accident vasculaire cérébral, le coma et même la mort. Mais le principal problème avec la biopsie stéréotaxique est qu'elle n'est pas précise à 100, ce qui peut entraîner une grave erreur de diagnostic suivie d'une mauvaise gestion clinique de la maladie[32].

La biopsie tumorale étant un défi pour les patients atteints de tumeurs cérébrales, des techniques d'imagerie non invasives telles que l'imagerie par résonance magnétique (IRM) ont été largement utilisées pour diagnostiquer les tumeurs cérébrales. Par conséquent, le développement de systèmes de détection et de prédiction du grade des tumeurs à partir des données IRM est devenu nécessaire. Mais à première vue de la modalité d'imagerie comme dans l'imagerie par résonance magnétique (IRM), la bonne visualisation des cellules tumorales et sa différenciation avec ses tissus mous voisins est une tâche quelque peu difficile qui peut être due à la présence d'un faible éclairage dans les modalités d'imagerie ou sa grande présence de données ou plusieurs complexité et variance de la forme non structurée de type tumeur, de la taille viable et des emeur[43].
La détection automatisée des défauts en imagerie médicale à l'aide de l'apprentissage automatique est devenue le domaine émergent dans plusieurs applications de diagnostic médical. Son application dans la détection des tumeurs cérébrales en IRM est très cruciale car elle fournit des informations sur les tissus anormaux nécessaires à la planification du traitement. Des études dans la littérature récente ont également rapporté que la détection et le diagnostic informatisés automatiques de la maladie, basés sur l'analyse d'images médicales, pourraient être une bonne alternative car cela permettrait au radiologue de gagner du temps et d'obtenir une précision testée. De plus, si les algorithmes informatiques peuvent fournir des mesures robustes et quantitatives de la représentation des tumeurs, ces mesures automatisées aideront grandement à la gestion clinique des tumeurs cérébrales en libérant les médecins du fardeau de la représentation manuelle des tumeurs[64].

6.1 Les types de Tumeurs cérébrale

6.1.1 Tumeur Gliome

Un gliome est un type de tumeur qui prend naissance dans les cellules gliales du cerveau ou de la colonne vertébrale. Les gliomes représentent environ 30 / de toutes les tumeurs cérébrales et du système nerveux central, et 80 / de toutes les tumeurs cérébrales malignes.

6.1.2 Tumeur Méningiome

également connue sous le nom de **tumeur méningée**, est généralement une tumeur à croissance lente qui se forme à partir des méninges, les couches membraneuses entourant le cerveau et la moelle épinière.Les symptômes dépendent de l'emplacement et se produisent à la suite de la pression de la tumeur sur les tissus voisins.

6.1.3 Tumeur Hypophysaire

les tumeurs hypophysaires sont des excroissances anormales qui se développent dans votre glande pituitaire. Certaines tumeurs hypophysaires entraînent une trop grande quantité d'hormones qui régulent des fonctions importantes de votre corps. Certaines tumeurs hypophysaires peuvent amener votre glande pituitaire à produire des niveaux inférieurs d'hormones[10].

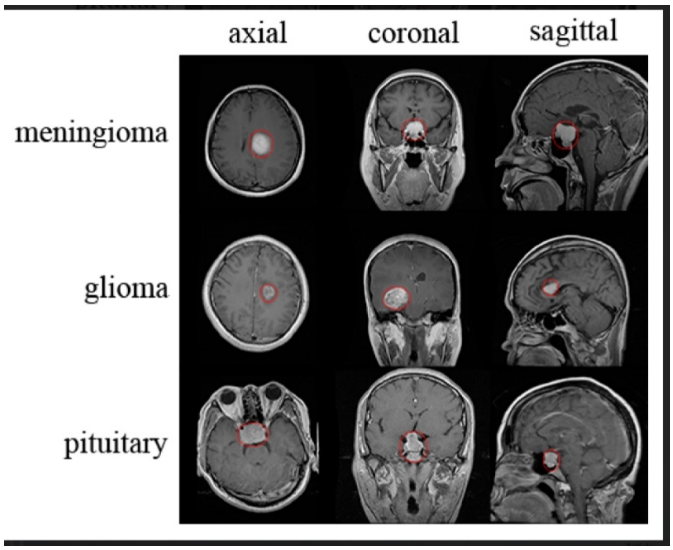

FIGURE 26 – Représentation d'images d'imagerie par résonance magnétique (IRM) normalisées montrant différents types de tumeurs dans différents plans

6.1.4 Qu'est-ce qu'une IRM ?

Une IRM, ou **image par résonance magnétique** , est créée par un puissant champ magnétique, des impulsions de radiofréquence et un ordinateur. Une IRM peut capturer une image diagnostique claire des organes, des tissus mous, des os et de la structure interne d'un patient. Grâce à un champ magnétique robuste et à des ondes radio, les examens IRM évitent complètement l'utilisation de rayonnements pour produire une image diagnostique[59].

6.1.5 Comment un système IRM produit-il une image diagnostique ?

Pour capturer une image, **le système IRM** utilise et envoie des ondes magnétiques et radiofréquences dans le corps du patient. L'énergie émise par les atomes dans le champ magnétique envoie un signal à un ordinateur. Ensuite, l'ordinateur utilise des formules mathématiques pour convertir le signal en une image. Les patients devront rester allongés sur une table qui glisse dans une machine pendant environ 20 minutes pour les études de la colonne vertébrale et environ 30 minutes pour les extrémités, le cerveau et tous les autres types d'études. Habituellement, l'espace où la table se glisse est étroit et compact. Heureusement, la nouvelle technologie a permis aux appareils d'IRM d'être ouverts, facilitant une expérience patient plus confortable[44].

6.2 Prétraitement du dataset

Le système proposé capture l'image d'une image (scanner médical) pour l'utiliser dans apprentissage.
Chaque image est traitée et ses caractéristiques essentielles sont extraites et stockées dans une table de vecteurs avec le libellé de la classe. Une fois l'acquisition des images de toutes les classes terminée, On utilise la méthode SVM pour trouver un modèle de décision qui permette de bien distinguer les types les uns des autres et enregistre ce modèle pour l'utiliser lors de la sélection. En mode sélection, les caractéristiques de l'image en question sont extraites puis exposées au modèle utilisé pour déterminer son type. Le type détecté est utilisé pour classer la nouvelle image dans la bonne classe.

6.3 Méthodologie de mise en œuvre

6.3.1 Présentation des outils

Python 3
1)- Python 3 : nous allons utilisé python qui est un langage de programmation mathématique statistique comme R.
-le code python est plus compact et lisible.

-la structure de données python est supérieure.
-il est open source et fournit également plus de packages graphiques et d'ensembles de données[72].

Jupyter Notebook
2)-Jupyter Notebook est un outil open source permettant d'écrire du code informatique et de le partager pour collaborer[77].

6.3.2 Les Paquets Python

1-scikit Learn
Scikit-learn est une bibliothèque libre Python destinée à l'apprentissage automatique.

Elle propose dans son framework de nombreuses bibliothèques d'algorithmes à implémenter, clé en main. Ces bibliothèques sont à disposition notamment des data scientists.

Elle comprend notamment des fonctions pour estimer des forêts aléatoires, des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support. Elle est conçue pour s'harmoniser avec d'autres bibliothèques libres Python, notamment **NumPy** et **SciPy**[56].

2-Numpy
NumPy est une bibliothèque pour langage de programmation Python, destinée à manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableau[15].

3-Matplotlib
Matplotlib enfin est un paquet pour créer des graphiques. Avec Matplotlib, il est possible de tracer des courbes de fonctions en 2D et 3D, de tracer des "heat maps", voire même de faire des animations[41].

4-Pandas
pandas est un outil d'analyse et de manipulation de données open source rapide, puissant, flexible et facile à utiliser, construit sur le langage de programmation Python[65].

5-Open CV
OpenCV (pour Open Computer Vision) est une bibliothèque graphique libre, initialement développée par Intel, spécialisée dans le traitement d'images et videos en temps réel[23].

6.3.3 Le hardware

— **Processor :** Intel® Core™ i5-10th Gen CPU @ 5.
— GHzInstalled memory (RAM) :4.00GB .
— **System Type :** 64-bit Operating System.

6.3.4 acquisition d'image

les images IRM sont acquises puis ces images sont transmises en entrée à l'étape de prétraitement.

6.3.5 Kaggle Dataset :

Remote source :(https ://github.com/sartajbhuvaji/brain-tumor-classification-dataset)

Le dossier contient des données IRM. Les images sont déjà divisées en dossiers de **Train** et de **test**.
Chaque dossier a plus de deux sous-dossiers. Ces dossiers ont des IRM des classes de tumeurs respectives.

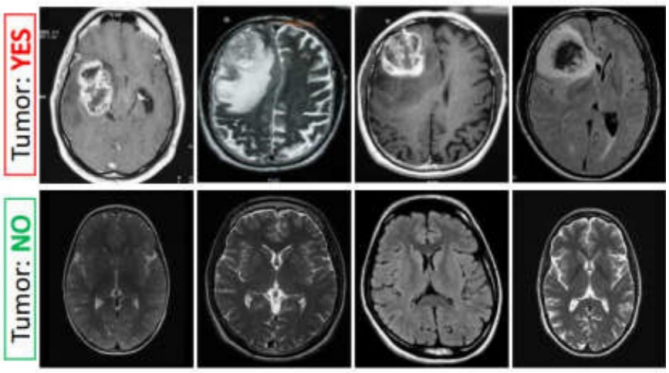

FIGURE 27 – Kaggle Dataset exemple

1) Testing :

a)Non-Tumor.

b)Tumor.

2) Training :

a)Non-Tumor.

b)Tumor.

6.4 Les étapes pour créer un système d'apprentissage automatique

1)collecte de données

2)pré-traitement des données

3)Feature extraction et selection

4)Splitting data entre training et testing

5)Algorithme de selection Classification(SVM-KNN)

6)Training

7)Testing

6.4.1 Méthode Proposée

le travail proposé vise à améliorer les performances du classificateur traditionnel. Ces classificateurs nécessitent les ensembles des données pour la formation et ont une faible complexité en temps de calcul, ce qui convient donc au diagnostic et à la classification des tumeurs cérébrales assistées par ordinateur.Nous proposons une méthode d'ensemble utilisant.
-Plusieurs classificateurs basés sur l'apprentissage ont déjà été utilisés pour la classification, notamment **la machine à vecteurs de support (SVM)** et **K Nearest Neighbors (KNN)**.

Les machines à vecteurs de support une meilleurs précision de recherche pour des problèmes classification des images, peuvent bien généraliser et rapidement adopté en raison de sa capacité à travailler avec des données de grandes dimensions ,les svm sont appréciées pour leur simplicité d'usage .

k-nn basé sur une fonction de distance et une fonction de vote, la métrique utilisée est **la distance euclidienne** . Il vise à calculer la superficie de la région tumorale et à classer les tumeurs cérébrales comme **Positives Tumeurs** et Tumeurs.

La réduction des fonctionnalités est effectuée à l'aide de **PCA** et la classification à l'aide de **SVM**.

L'extraction des caractéristiques des images IRM sera effectuée par échelle de gris**(Gry Scale)**.

Ce système de détection et de segmentation tumorale comporte plusieurs étapes :

(a)Images d'IRM cérébrales d'entrée.
(b)Le prétraitement des images est utilisé pour améliorer la qualité des images.

(c)Les caractéristiques seront extraites des images segmentées.
(d)Les caractéristiques réduites sont soumises à un classificateur de machine à vecteurs de support pour identifier la tumeur.

Analyse en Composantes Principales (PCA)
Analyse en composantes principales L'ACP est une transformation orthogonale de variables corrélées en variables linéairement non corrélées appelées composantes principales.
La première composante principale a les variances les plus élevées ; les composantes successives contiennent la variance la plus importante possible de sorte qu'elle reste orthogonale aux composantes précédentes. Il est utilisé pour réduire les dimensions ou les fonctionnalités avec lesquelles nous devons former notre classificateur, ce qui aide finalement à réduire la complexité temporelle et spatiale pour le calcul de données données[26].

6.5 Mise en œuvre et résultats

6.5.1 Charger les dépendances

1) importer :

a)- **python** .
b)- **sklearn** (pip install sklearn).
c)- **openCV** (pip install opencv-python).
d)- **numpy** (pip install numpy).
e)- **matplotlib** (pip install matplotlib).

DataSet	Trainig dataset	Testing datase
NO-Tumor	395	105
Tumor	827	115
Total	1222	220

TABLE 1 – Dataset pour la classification

Data Analysis array([0, 1] 0 :"NO-Tumor".
1 :"Tumor".
)
X.shape==>(1222, 200, 200)
X-updated.shape==>(1222, 40000)

6.5.2 Split data

dans cette étape, nous avons diviser les données en deux parties (Training et Testing)
-**Training** = 0.8
-**Testing** = 0.2

‘ **xtrain.shape** ==>(977, 40000)
xtest.shape ==>(245, 40000)

-Train set comme résultats ci-dessus :

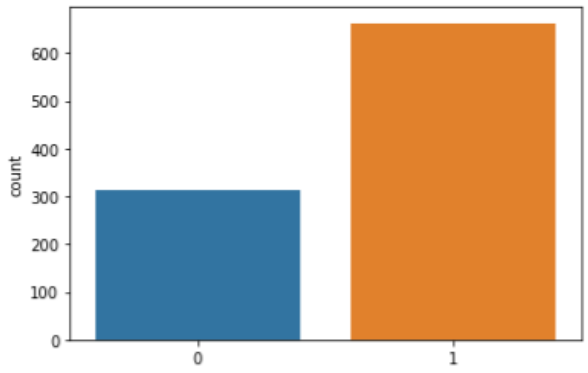

FIGURE 28 – Résultat de splite data train

-Test set comme figure ci-dessus :

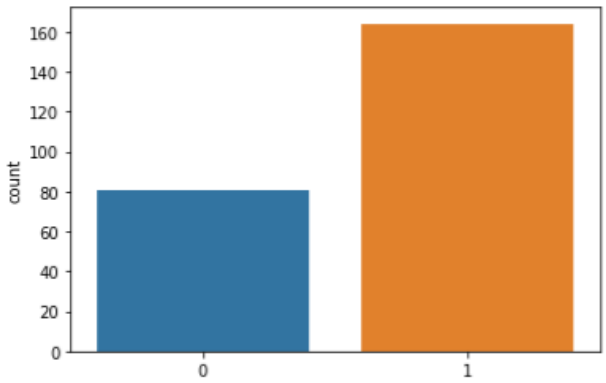

FIGURE 29 – Résultat de splite data test

6.5.3 Feature Scaling

Nous avons utiliser la technique de Scaling minmax pour amener toutes les valeurs des caractéristiques à moins ou égales à 1.

6.5.4 Visualisation de données

est un ensemble de méthode permettant de résumer de manière graphique de données statistiques.

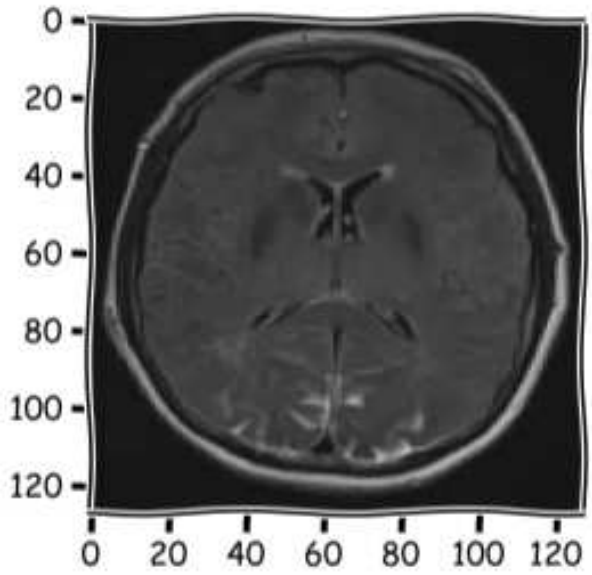

FIGURE 30 – resultat de visualisation

6.5.5 Model Training

comme nous avons fait avec la partie prétraitement . Je vais former le modèle en utilisant les algorithmes SVM et KNN , puis nous comparerons les performances de ces deux modèles différents.

6.5.6 Evaluation

nous comparerons les scores des deux modéles ci-dessus.

1) **KNN** :

K	Training	Testing	Accuracy
k=1	1.0	0.95	0.75
k=3	1.0	0.96	0.83
k=5	0.94	0.94	0.94
k=10	0.93	0.94	0.80

TABLE 2 – Régularisation de K pour meilleurs résultats

D'après les résultats présentés dans le tableau , nous pouvons conclure que le choix de de valeur K à utiliser pour effectuer une bonne valeur de Accuracy ,on utiliser de voisins un nombre K petit plus on sera sujette au **Sous apprentissage (Underfitting)** comme le cas K=1 et K= 3.Par ailleurs,plus on utiliser de voisins un nombre K grand ,on risque d'avoir du **Overfitting** comme le cas K=10,par conséquent un bonne résultat dans le cas K=5,car se mieux accuracy.

2) **SVM** :

C	Training	Testing	Accuracy
C=1	0.99	0.96	0.96
C=3	1.0	0.96	0.95
C=6	1.0	0.94	0.93
C=20	1.0	0.78	0.78

TABLE 3 – Régularisation de C pour meilleurs résultats

Les valeurs Gamma et C sont des hyperparamètres clés qui peuvent être utilisés pour former le modèle SVM le plus optimal à l'aide du noyau linéaire. **C**pour contrôler l'erreur.Nous pouvons conclure que une valeur plus élevée de C se traduira par un modèle qui a une très grande accuracy mais qui peut ne pas être généralisable comme le cas C=6 ,par contre ,la valeur inférieure de C est petite, plus la tolérance de mauvaise classification est grande et ce qui est formé est un classificateur à marge souple qui généralise mieux que le classificateur à marge maximale comme le cas C=1.

Interprétation des résultats concernant SVM avec noyau Gaussian :

Le noyau Gassien tel qu'il est définit utilise deux paramètres C et gamma. Ces paramètres sont choisis d'une façon empirique après plusieurs essais sur les échantillons du dataset. Le résultat de ces essais est que les paramètres C et gamma qui donnent les meilleurs résultats sont C=10 et gamma=0.01.

6.5.7 Testing dataset

il s'agit donc de créer un modèle prédictif à l'aide de sklearn sur un ensemble de données sur les tumeurs cérébrales.

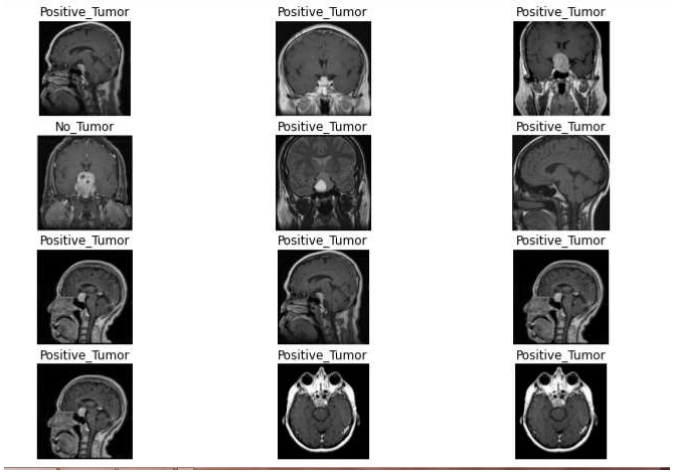

FIGURE 31 – Resultat de classification SVM

Nous avons tester une nouvelle échantillon (nouvelle image) :
Résultat avec Modéle SVM

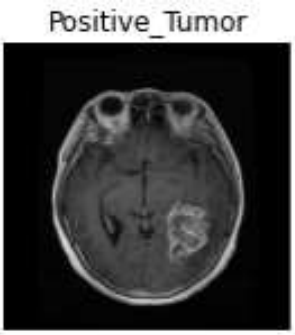

FIGURE 32 – nouveau cas avec SVM

Résultat avec KNN

à partir des résultats présentes dans les figures (31) et (32).On remarque que SVM predict correct ,par contre knn predict incorrect.Et nous en concluons le modéle SVM c'est la donner un bon résultat réalisé.

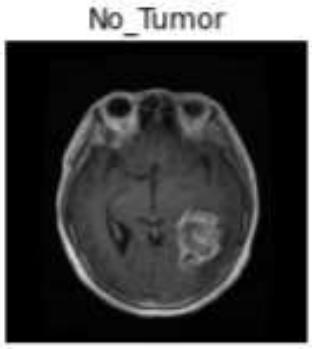

FIGURE 33 – nouveau cas avec KNN

6.5.8 Evaluation les metrics en Table

les résultats obtenus en termes de métriques d'évaluation considérées sont présentés dans le tableau 4

Classifier	Accuracy	Precision	Recall	F1-score
SVM	0.96	0.96	0.98	0.90
KNN	0.91	0.93	0.79	0.85

TABLE 4 – Résultats de evaluation les metrics

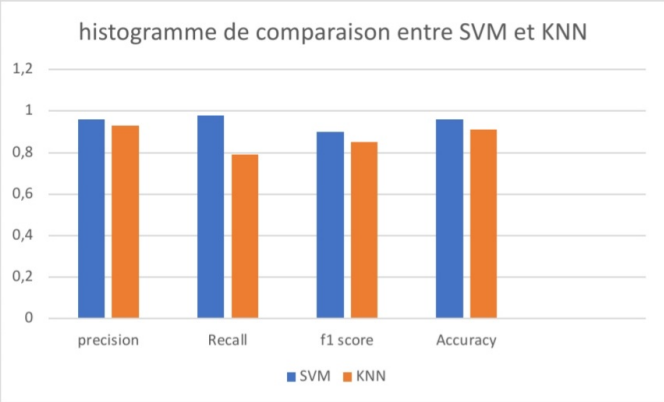

FIGURE 34 – Histogramme de Comparaison entre SVM et KNN

à partir des résultats présentés dans le tableau , nous pouvons conclure que la méthode proposée SVM avec une Accuracy de 0.96 surpasse le classificateur comparé, car ils offrent les meilleures performances en termes de divers paramètres d'évaluation par rapport au classificateur KNN.

6.5.9 Confusion Matrix

La méthode proposée SVM donneé d'excellentes performances sur divers paramètres d'évaluation comme décriten Table 5 et Table 6 , en comparaison avec la méthode existante KNN. Ainsi, la SVM est efficace pour

Prediction	Tumor	68	13
	Non-Tumor	3	161

Table 5 – Confusion Matrix KNN

Prediction	Tumor	75	6
	No-Tumor	3	161

Table 6 – Confusion matrix SVM

travailler pour la classification de No-Tumor et Positive-Tumor.

7 Conclusion

-Dans cette partie,nous avons spécifier l e d omaine d e l a c lassification d' image ," Brain Tu mor MR I" avec l'apprentissage automatique pour effectuer une classification sur cette d'images.et nous avons commencé par les algorithmes de ML traditionnelles SVM et KNN pour faire la classification supervsé et vu quelques étapes et expliqué le principe de chaque algorithme .

7.1 Conclusion et Perspectives

Dans ce travail, nous avons mis en oeuvre des techniques de l'apprentissage supervisé pour la classifucation d'image médicales.Nous avons utilisé les deux méthodes KNN(K Nearest neighbors) et SVM(Support Vector Machine).
Les résultats expérimentaux obtenus par la classification par SVM sont prometteurs et peuvent étre améliorés en optimisant les hyper paramétres C et Gamma.
La comparaison des résultats robustesse du modéle SVM.
C**omme perspectives** on peut tester application sur de nouveaux échantillons et aussi développer la méthode SVM pour mieux l'adapter à une meilleure classification.

Références

[1] Bahman Abbassi and Li Zhen Cheng. Sfe2d : Un outil d'extraction de caractéristiques spectrales de géo-images.

[2] Martine Adda-Decker. De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. *JEP*, pages 389–400, 2006.

[3] David Ameisen. Qu 'est-ce qu 'une image numérique. In *Conference Paper*, volume 57, pages 169–172, 2013.

[4] Massih-Reza Amini. Principes de base en apprentissage supervisé, 2020.

[5] John Ashburner and Karl J Friston. Spatial transformation of images. *Human brain function*, pages 43–58, 1997.

[6] Fatiha Barigou, Baghdad Atmani, Youcef Bouziane, and Naouel Barigou. Accélération de la méthode des k plus proches voisins pour la catégorisation de textes. In *EGC*, volume 2013, pages 241–246, 2013.

[7] FAIROUZ BELILITA. *Contribution à l'implémentation d'algorithmes de traitement numérique du signal sur des architectures parallèles très performantes. Application à la TFD et aux filtres numériques RII.* PhD thesis, Université de M'sila, 2017.

[8] Maïtine Bergounioux. Quelques méthodes de filtrage en traitement d'image. 2011.

[9] P Besse and L Ferré. Sur l'usage de la validation croisée en analyse en composantes principales. *Revue de statistique appliquée*, 41(1) :71–76, 1993.

[10] Peter McL Black. Brain tumors. *New England Journal of Medicine*, 324(22) :1555–1564, 1991.

[11] Philippe Bolon, Jean-Marc Chassery, Jean-Pierre Cocquerez, Didier Demigny, Christine Graffigne, Annick Montanvert, Sylvie Philipp, Rachid Zéboudj, Josiane Zerubia, and Henri Maître. *Analyse d'images : filtrage et segmentation*. Masson, 1995.

[12] Amel Borgi. *Apprentissage supervisé par génération de règles : le système SUCRAGE*. PhD thesis, Paris 6, 1999.

[13] Charles Bouveyron and Stephane Girard. Classification supervisée et non supervisée des données de grande dimension. *La revue MODULAD*, 40 :81–102, 2009.

[14] Bruno Bouzy. Apprentissage par renforcement (3). *Cours de d'apprentissage automatique*, 2005.

[15] Eli Bressert. Scipy and numpy : an overview for developers. 2012.

[16] Yves Brostaux. *Etude du classement par forêts aléatoires d'échantillons perturbés à forte structure d'interaction*. PhD thesis, FUSAGx-Faculté Universitaire des Sciences agronomiques de Gembloux, 2005.

[17] Olivier Caelen. *Sélection Séquentielle en Environnement Aléatoire Appliquéea l'Apprentissage Supervisé*. PhD thesis, PhD thesis, ULB, 2009.

[18] Faïcel Chamroukhi. Classification supervisée : Les k-plus proches voisins. *mémoire de fin d'étude, Université du Sud Toulon–Var*, 2013.

[19] Clement Chatelain. Les support vector machine (svm). Technical report, Technical report, 2003.

[20] Marie Chavent, Christiane Guinot, Yves Lechevallier, and Michel Tenenhaus. Méthodes divisives de classification et segmentation non supervisée : Recherche d'une typologie de la peau humaine saine. *Revue de statistique appliquée*, 47(4) :87–99, 1999.

[21] Pierre Cornillon and Eric Matzner-Lober. *Régression : théorie et applications*. Springer, 2007.

[22] Gilles Deleuze. *L'image-temps*. Les Ed. de minuit, 1985.

[23] Aniket Dhanawade, Abhishek Drode, Gifty Johnson, Aadesh Rao, and Savitha Upadhya. Open cv based information extraction from cheques. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 93–97. IEEE, 2020.

[24] Edwin Diday. Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques. *Revue de statistique appliquée*, 19(2) :19–33, 1971.

[25] Gérard Dreyfus, JM Martinez, M Samuelides, MB Gordon, F Badran, S Thiria, and L Hérault. *Réseaux de neurones*, volume 39. Eyrolles Paris, 2002.

[26] Camille Duby and Stéphane Robin. Analyse en composantes principales. *Institut National Agronomique, Paris-Grignon*, 80, 2006.

[27] Ludovic Dugas. Recherches expérimentales sur les différents types d'images. *Revue Philosophique de la France et de l'Étranger*, 39 :285–292, 1895.

[28] Alberto Fernández, Victoria LóPez, Mikel Galar, MaríA José Del Jesus, and Francisco Herrera. Analysing the classification of imbalanced data-sets with multiple classes : Binarization techniques and ad-hoc approaches. *Knowledge-based systems*, 42 :97–110, 2013.

[29] Dominik Francoeur. Machines à vecteurs de support : une introduction. *CaMUS (Cahiers Mathématiques de l'Université de Sherbrooke)*, 1 :7–25, 2010.

[30] Syntyche Gbèhounou, François Lecellier, and Christine Fernandez-Maloigne. Extraction de l'impact emotionnel des images. *Traitement de Signal*, pages 409–432, 2012.

[31] André Gide. Chapitre 3 l'apprentissage automatique en télédétection. *rat*, page 63, 2017.

[32] Nelly Gordillo, Eduard Montseny, and Pilar Sobrevilla. State of the art survey on mri brain tumor segmentation. *Magnetic resonance imaging*, 31(8) :1426–1438, 2013.

[33] Philippe-Henri Gosselin. *Méthodes d'apprentissage pour la recherche de catégories dans des bases d'images*. PhD thesis, Université de Cergy Pontoise, 2005.

[34] Romain Guigourès and Marc Boullé. Optimisation directe des poids de modèles dans un prédicteur bayésien naïf moyenné. In *EGC*, pages 77–82, 2011.

[35] Ali Haddad. *Méthodes variationnelles en traitement d'image*. PhD thesis, École normale supérieure de Cachan-ENS Cachan, 2005.

[36] Jean Hagendorf. Pavages, carrelages, forçage, hypomorphie et classification de gallaï dans les relations binaires. 2016.

[37] Amel Haliche. *Classification supervisée à base de KNN avec pondération d'attributs par l'algorithme génétique*. PhD thesis, 2015.

[38] J-P Haton, Nadjet Bouzid, François Charpillet, Marie-Christine Haton, Brigitte Lâasri, Hassan Lâasri, Pierre Marquis, Thierry Mondot, and Amedeo Napoli. *Le raisonnement en intelligence artificielle*. InterEditions, 1991.

[39] Laurent Henriet. *Systèmes d'évaluation et de classification multicritères pour l'aide à la décision : Construction de modèles et procédures d'affectation*. PhD thesis, Université Paris Dauphine-Paris IX, 2000.

[40] Radu Horaud and Olivier Monga. *Vision par ordinateur : outils fondamentaux*. Editions Hermès, 1995.

[41] John D Hunter. Matplotlib : A 2d graphics environment. *Computing in science & engineering*, 9(03) :90–95, 2007.

[42] Guillaume Joutel, Véronique Eglin, Stéphane Bres, and Hubert Emptoz. Extraction de caractéristiques dans les images par transformée multi-échelle. In *21° Colloque GRETSI, Troyes, FRA, 11-14 septembre 2007*. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, 2007.

[43] Michael R Kaus, Simon K Warfield, Arya Nabavi, Peter M Black, Ferenc A Jolesz, and Ron Kikinis. Automated segmentation of mr images of brain tumors. *Radiology*, 218(2) :586–591, 2001.

[44] Bryan Kolb, Ian Q Whishaw, and Gordon Campbell Teskey. *Cerveau et comportement*. De Boeck Supérieur, 2019.

[45] Th AM Kruip and SJ Dieleman. Macroscopic classification of bovine follicles and its validation by micromorphological and steroid biochemical procedures. *Reproduction Nutrition Développement*, 22(3) :465–473, 1982.

[46] Stéphane Lallich, Philippe Lenca, and Benoît Vaillant. Construction d'une entropie décentrée pour l'apprentissage supervisé. In *EGC 2007 : 7èmes journées francophones" Extraction et gestion des connaissances", Atelier Qualité des Données et des Connaissances, 23 janvier, Namur, Belgique*, pages 45–54, 2007.

[47] Thomas Laloë. *Sur quelques problèmes d'apprentissage supervisé et non supervisé*. PhD thesis, Université Montpellier II-Sciences et Techniques du Languedoc, 2009.

[48] Philippe Leray, Hugo Zaragoza, and Florence d'Alché Buc. Pertinence des mesures de confiance en classification. In *Conférence francophone RFIA*, 2000.

[49] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3) :18–22, 2002.

[50] Wei-Yin Loh. Classification and regression trees. *Wiley interdisciplinary reviews : data mining and knowledge discovery*, 1(1) :14–23, 2011.

[51] Jérôme Louradour. *Noyaux de sequences pour la verification du locuteur par machines a vecteurs de support*. PhD thesis, Toulouse 3, 2007.

[52] Jonathan Milgram, Robert Sabourin, and Mohamed Cheriet. Système de classification à deux niveaux de décision combinant approche par modélisation et machines à vecteurs de support. In *Conférence Internationale Francophone sur l'Ecrit et le Document (CIFED 04)*, 2004.

[53] Noura Mounib. *Une approche co-évolutionnaire proie-prédateur pour le réhaussement d'images*. PhD thesis, Batna, Université El Hadj Lakhdar. Faculté des sciences de l'ingénieur, 2007.

[54] Sébastien Mustière. *Apprentissage supervisé pour la généralisation cartographique*. PhD thesis, Paris 6, 2001.

[55] Fiammetta Namer. *Morphologie, lexique et traitement automatique des langues*. Hermès-Lavoisier, 2009.

[56] Fabio Nelli. Machine learning with scikit-learn. In *Python Data Analytics*, pages 313–347. Springer, 2018.

[57] Roberta B Oliveira, Joao P Papa, Aledir S Pereira, and Joao Manuel RS Tavares. Computational methods for pigmented skin lesion classification in images : review and future trends. *Neural Computing and Applications*, 29(3) :613–636, 2018.

[58] Pavel Pevzner and Nicolas PUECH. *Bio-informatique moléculaire : Une approche algorithmique*. Springer, 2006.

[59] IRM Qu'est-ce qu'une. Comprendre l'examen d'imagerie par résonance magnétique (irm).

[60] Ricco Rakotomalala. Arbres de décision. *Revue Modulad*, 33 :163–187, 2005.

[61] Ricco Rakotomalala. Pratique de la regression lineaire multiple. *Diagnostic et selection de variables*, 2011.

[62] L Rouvière. Apprentissage supervisé-machine learning. 2022.

[63] Matthias Seeger and Michael Jordan. Sparse gaussian process classification with multiple classes. Technical report, Department of Statistics, University of Berkeley, CA, 2004.

[64] Edward Shaw, Charles Scott, Luis Souhami, Robert Dinapoli, Robert Kline, Jay Loeffler, and Nancy Farnan. Single dose radiosurgical treatment of recurrent previously irradiated primary brain tumors and brain metastases : final report of rtog protocol 90-05. *International Journal of Radiation Oncology* Biology* Physics*, 47(2) :291–298, 2000.

[65] Harvey S Singer and Christopher Loiselle. Pandas : a commentary. *Journal of psychosomatic research*, 55(1) :31–39, 2003.

[66] Aized Amin Soofi and Arshad Awan. Classification techniques in machine learning : applications and issues. *Journal of Basic and Applied Sciences*, 13 :459–465, 2017.

[67] Joel TANKEU, Philippe ADIABA, Steve ELANGA, and Nadia TOUATI. Comment utiliser le machine learning pour gagner des marchés publics ? *Management & Datascience*, 4(6), 2020.

[68] Claire Thomas. *Fusion d'images de résolutions spatiales différentes*. PhD thesis, École Nationale Supérieure des Mines de Paris, 2006.

[69] Fabien Torre. Globo : un algorithme stochastique pour l'apprentissage supervisé et non-supervisé. In *Actes de la Première Conférence d'Apprentissage*, pages 161–168. Citeseer, 1999.

[70] Apprentissage Transductif and Arnaud Revel. Apprentissage semi-supervisé.

[71] Ujjwal Ujjwal. *Gestion du compromis vitesse-précision dans les systèmes de détection de piétons basés sur apprentissage profond*. PhD thesis, Université Côte d'Azur (ComUE), 2019.

[72] Guido Van Rossum and Fred L Drake Jr. *Python tutorial*, volume 620. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

[73] Asparukh Velkov. *Les filigranes dans les documents ottomans : divers types d'images*. Stoyan Shivarov, 2005.

[74] Cédric Villani, Yann Bonnet, Charly Berthet, François Levin, Marc Schoenauer, Anne Charlotte Cornut, and Bertrand Rondepierre. *Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne*. Conseil national du numérique, 2018.

[75] Rémi Viola, Rémi Emonet, Amaury Habard, Guillaume Metzler, Sébastien Riou, and Marc Sebban. Une version corrigée de l'algorithme des plus proches voisins pour l'optimisation de la f-mesure dans un contexte déséquilibré. In *Conférence sur l'Apprentissage automatique (CAp 2019)*, 2019.

[76] Haifeng Wang and Dejin Hu. Comparison of svm and ls-svm for regression. In *2005 International Conference on Neural Networks and Brain*, volume 1, pages 279–283. IEEE, 2005.

[77] Jiawei Wang, KUO Tzu-Yang, Li Li, and Andreas Zeller. Assessing and restoring reproducibility of jupyter notebooks. In *2020 35th IEEE/ACM international conference on automated software engineering (ASE)*, pages 138–149. IEEE, 2020.

[78] Jyoti Yadav and Monika Sharma. A review of k-mean algorithm. *Int. J. Eng. Trends Technol*, 4(7) :2972–2976, 2013.