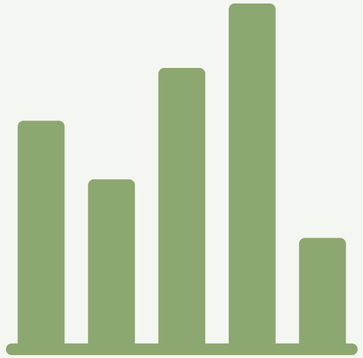
Détection du mildiou de la pomme de terre par imagerie grâce aux méthodes de Machine Learning

Yasmine BOUCHIBTI
Leslie CIETERS
Meryem GRIMAJ

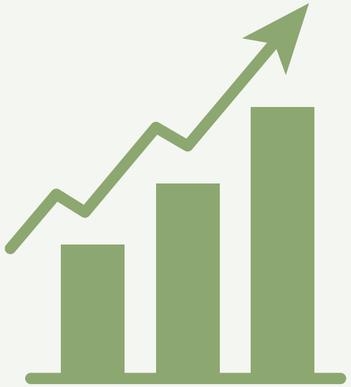
Sommaire

Introduction

- *Alternaria solani* (mildiou précoce)
- *Phytophthora infestans* (mildiou tardif)



Jusqu’à 80% de pertes

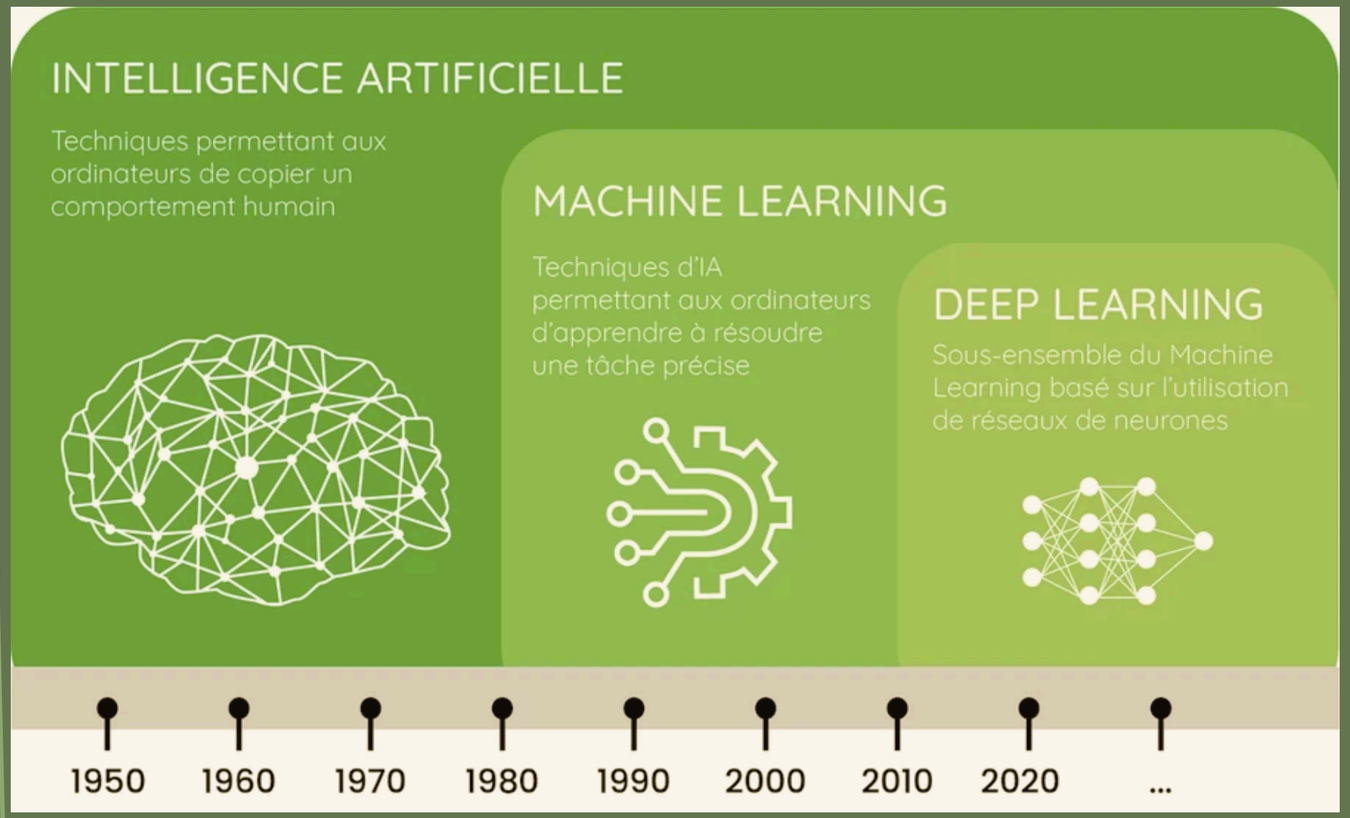


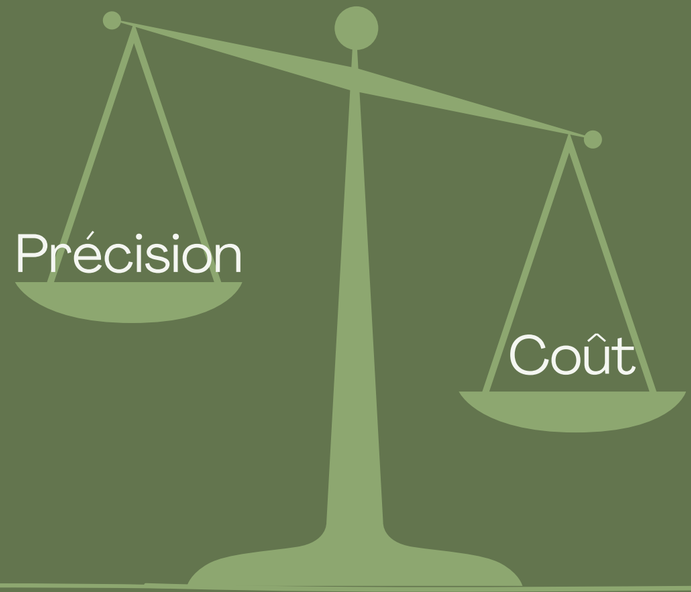
Utilisation de produits phytosanitaires

la Grande Famine 1845-1849

Sur Arte, « La Grande Famine en Irlande », dernière tragédie avant l’indépendance. (2021, novembre 30). Le Monde.

Machine Learning et Deep Learning





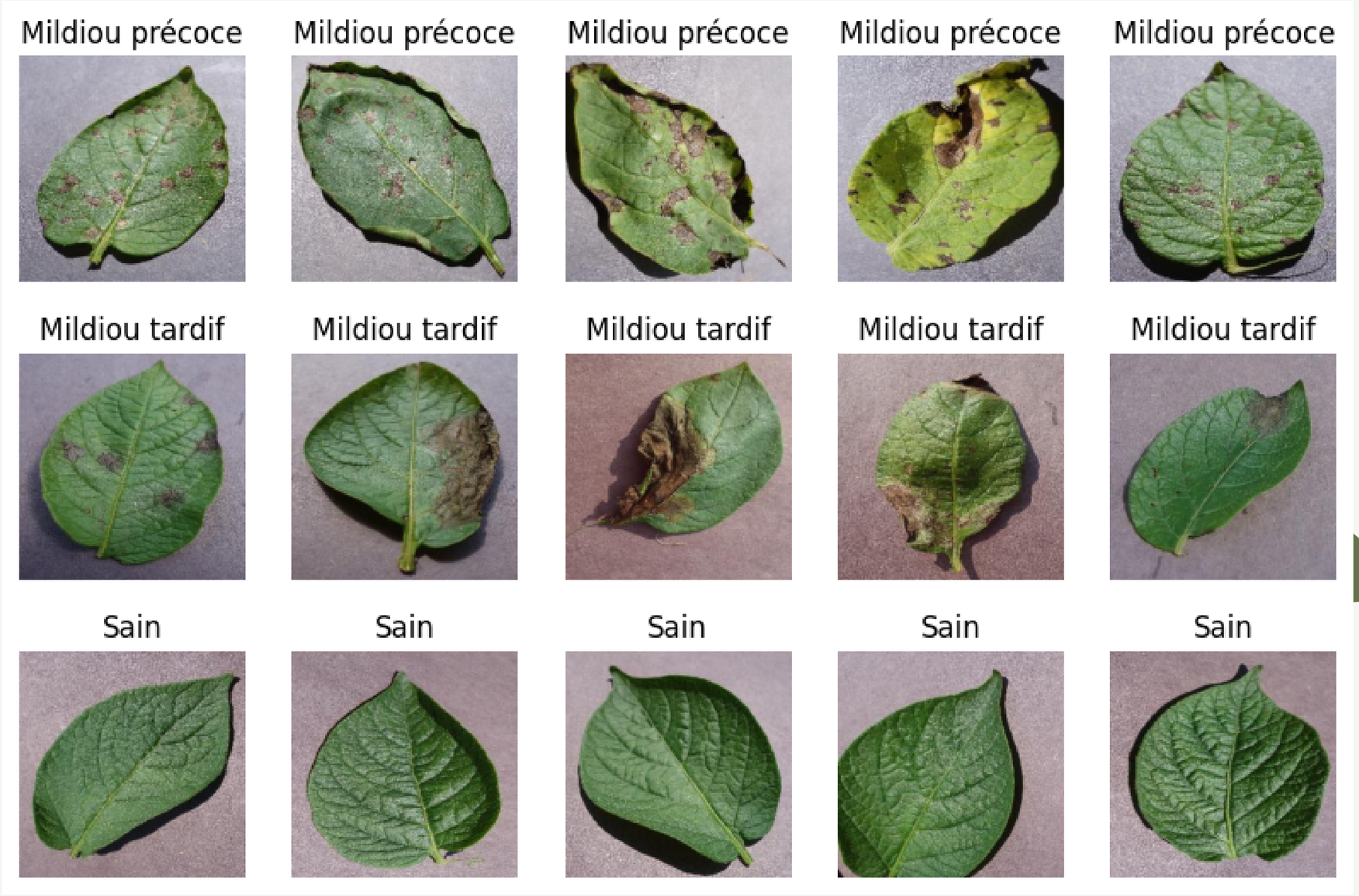
https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.dilepix.com%2Fblog%2Fdifference-intelligence-artificielle-deep-learning-agriculture&psig=AOvVaw0-A7dL2IJIZTBJJ79V6GI5&ust=1729093623304000&source=images&cd=vfe&opi=89978449&ved=0CBQQjRxqFwoTCOCKv-

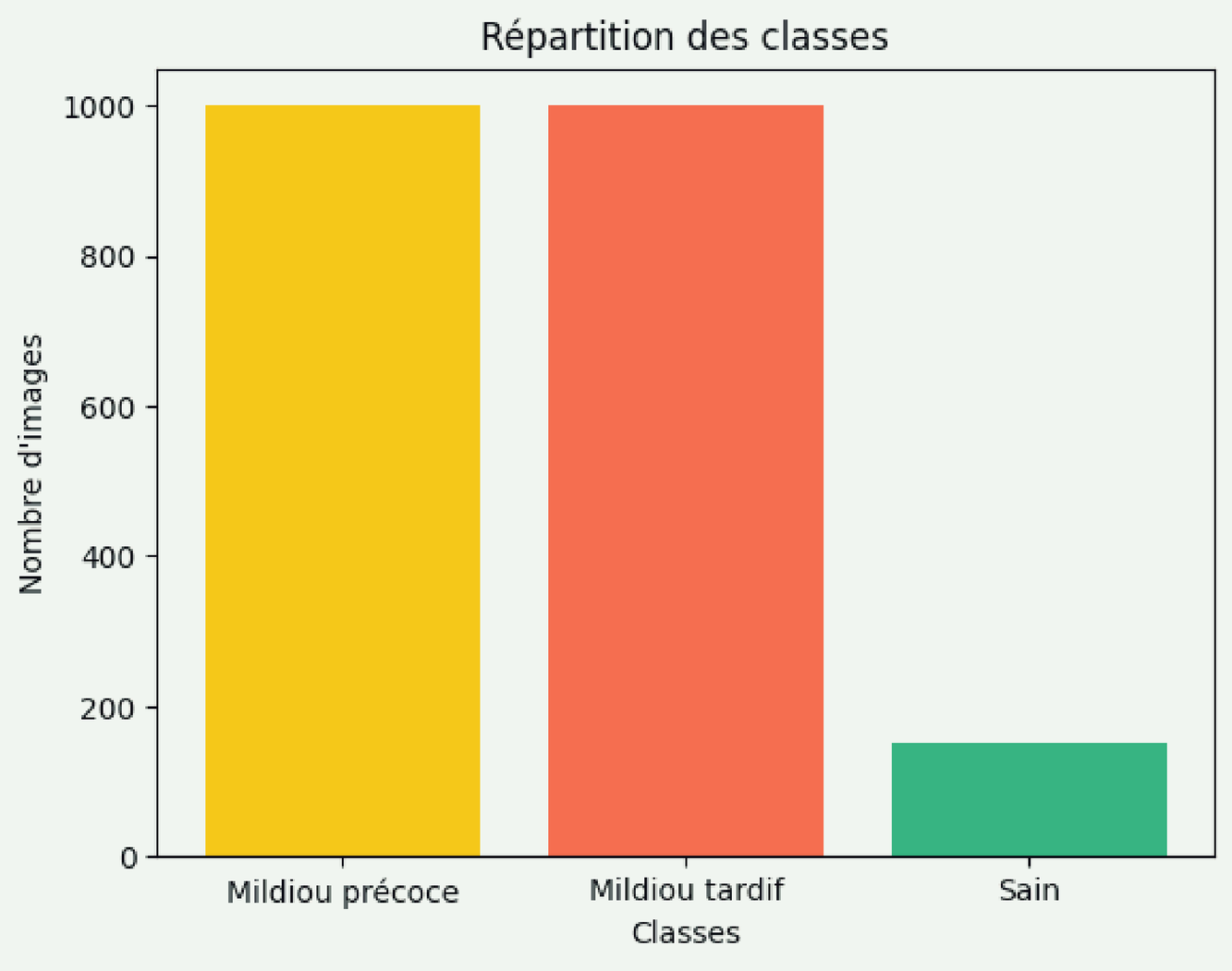
Imagerie

Le jeu de données

2152 images réparties en trois classes :

- Mildiou précoce
- Mildiou tardif
- Sain

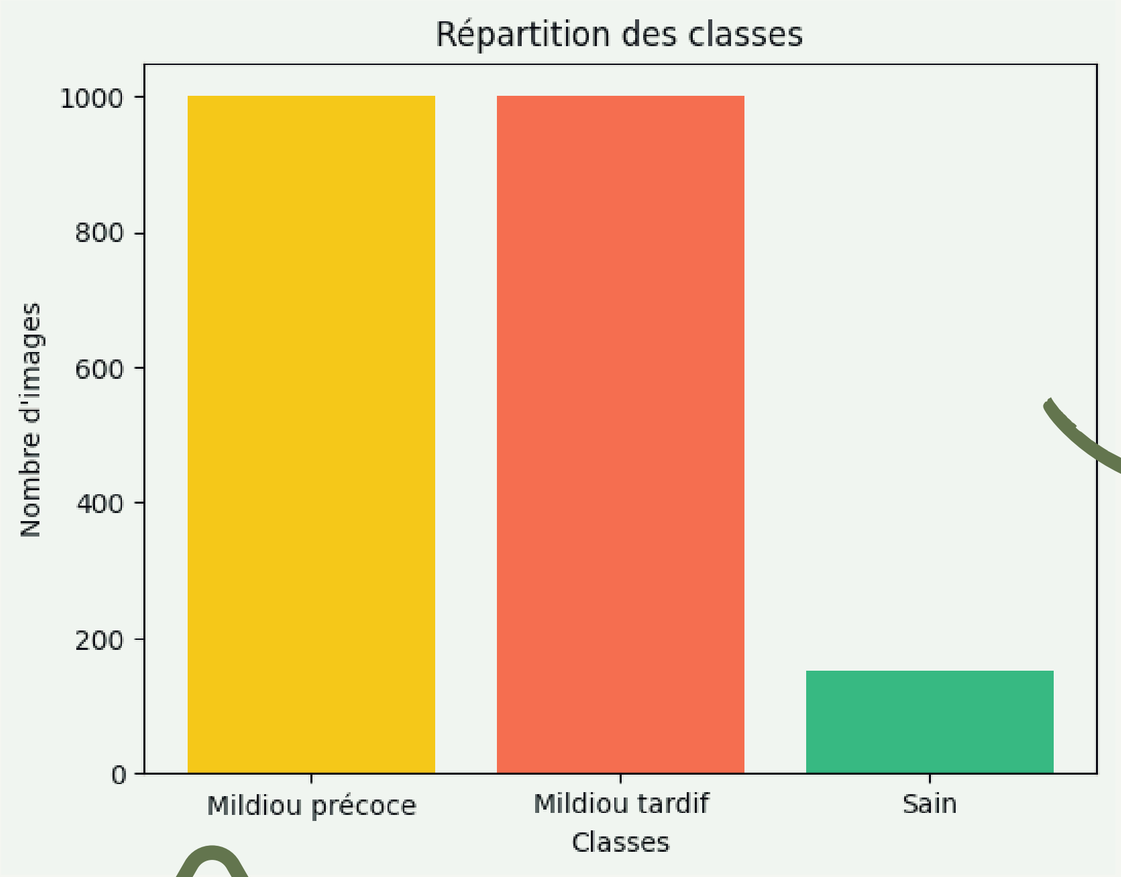

Mildiou précoce
Mildiou précoce
Mildiou précoce
Mildiou précoce
Mildiou précoce
Mildiou tardif
Mildiou tardif
Mildiou tardif
Mildiou tardif
Mildiou tardif
Sain
Sain
Sain
Sain
Sain

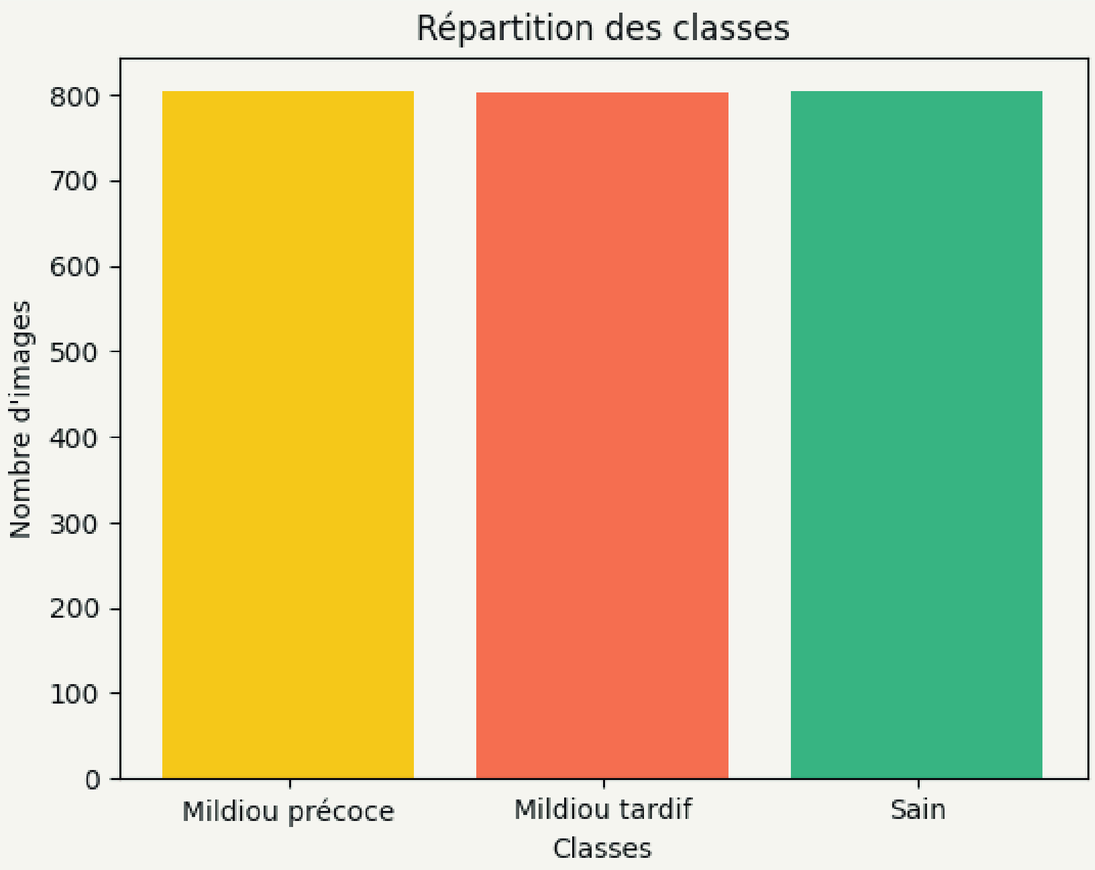


Déséquilibre important des classes

Technique SMOTE pour le suréchantillonnage de la classe minoritaire

- Séparation du jeu de données (80/20) pour l'entraînement et la validation





Uniquement sur le jeu de données d'entraînement

Modèles de Machine Learning

KNN

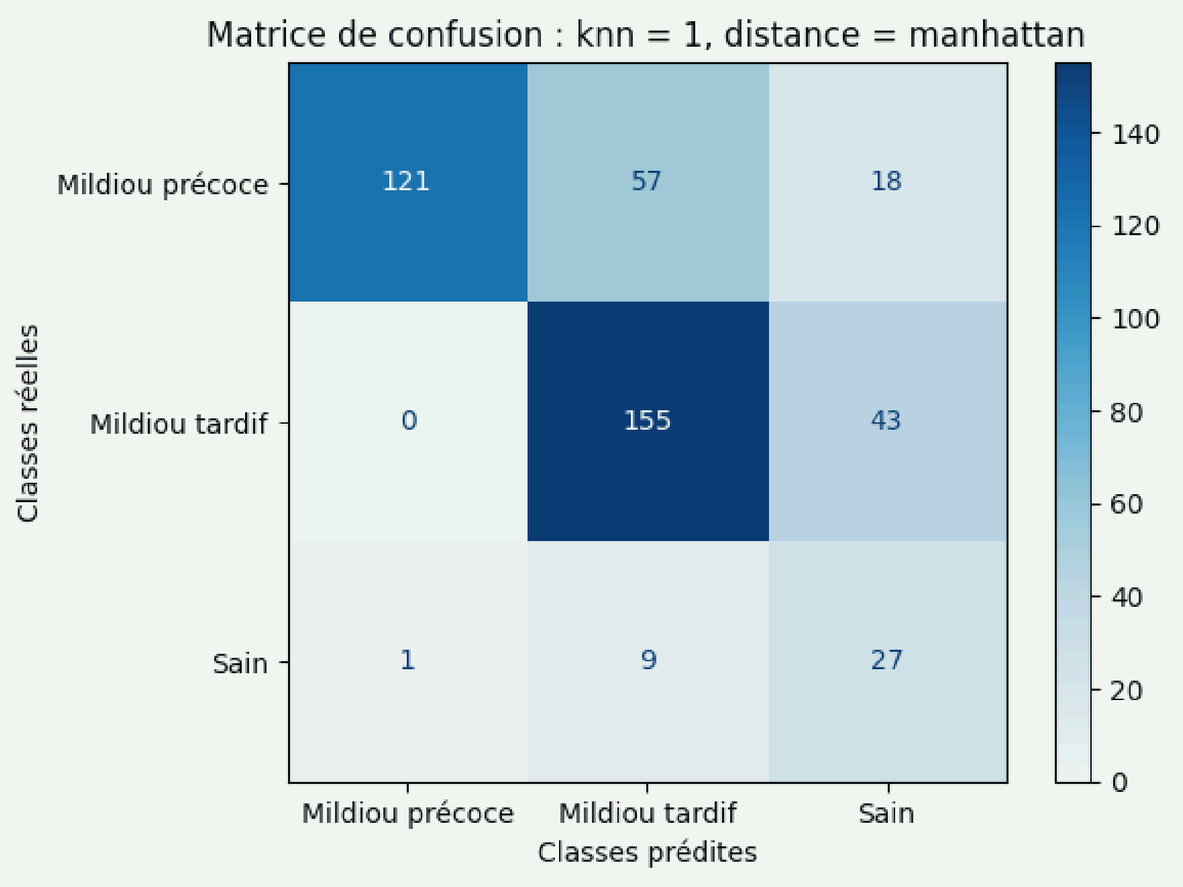
Régression logistique

SVM

Forêts aléatoires

KNN

Comparaison de plusieurs modèles afin de définir les meilleurs paramètres

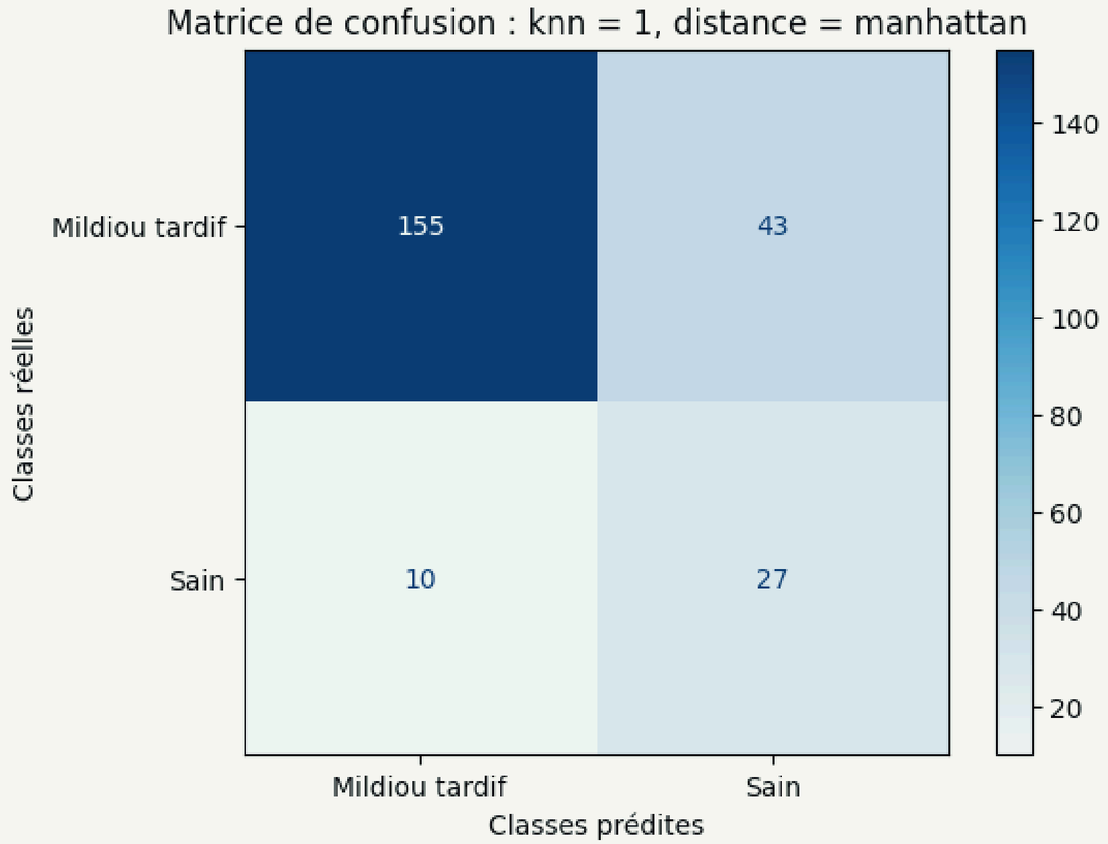


- Accuracy = 0.70
- Taux de faux positifs = 0.15
- Taux de faux négatifs = 0.27

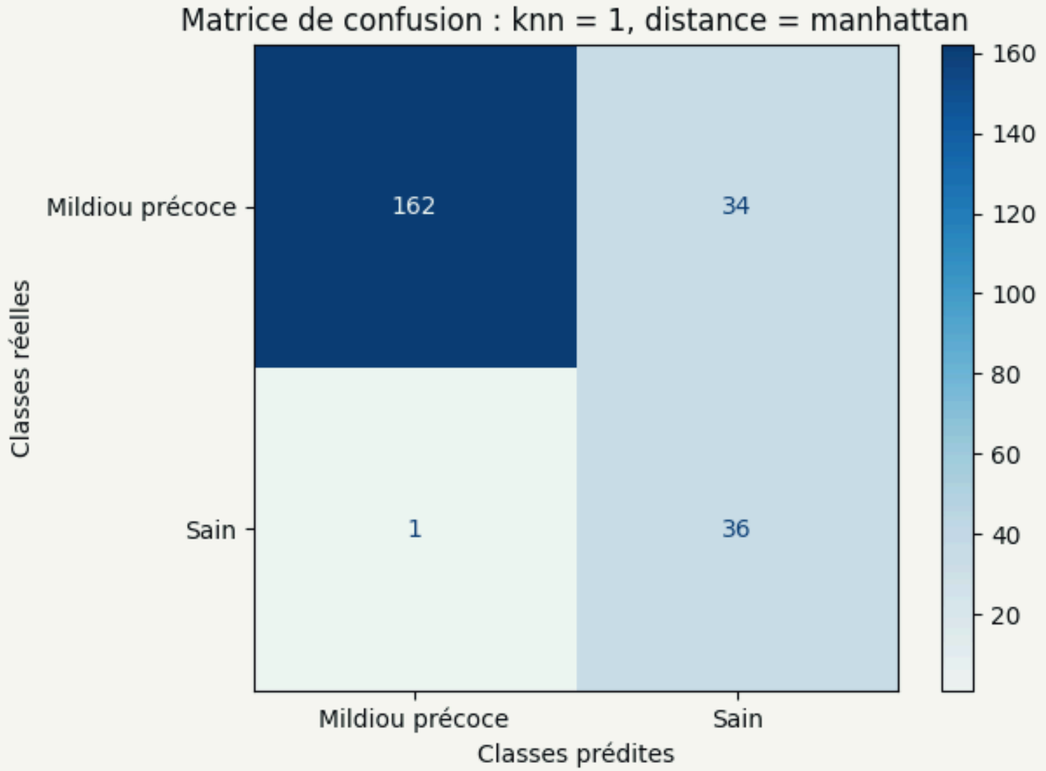
KNN

En séparant les classes

Accuracy = 0.77



Accuracy = 0.85



Régression logistique

1

Méthode one-vs-all

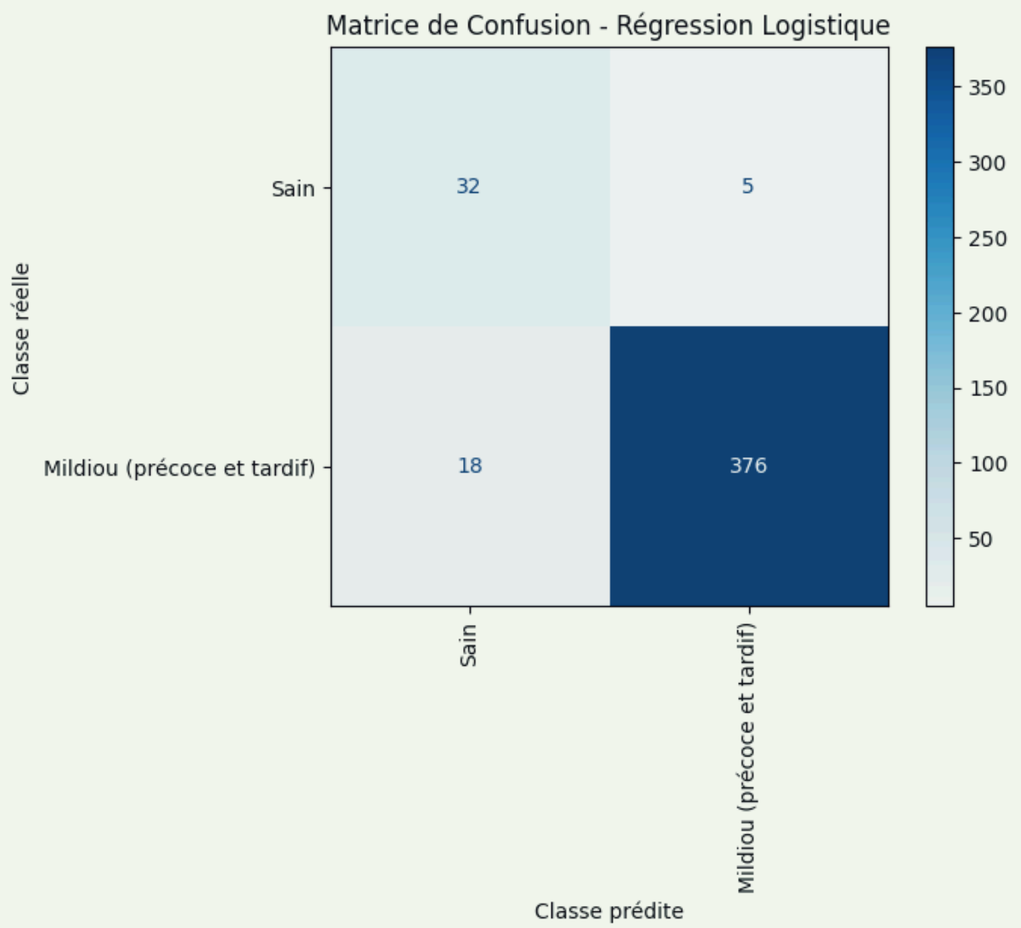
2

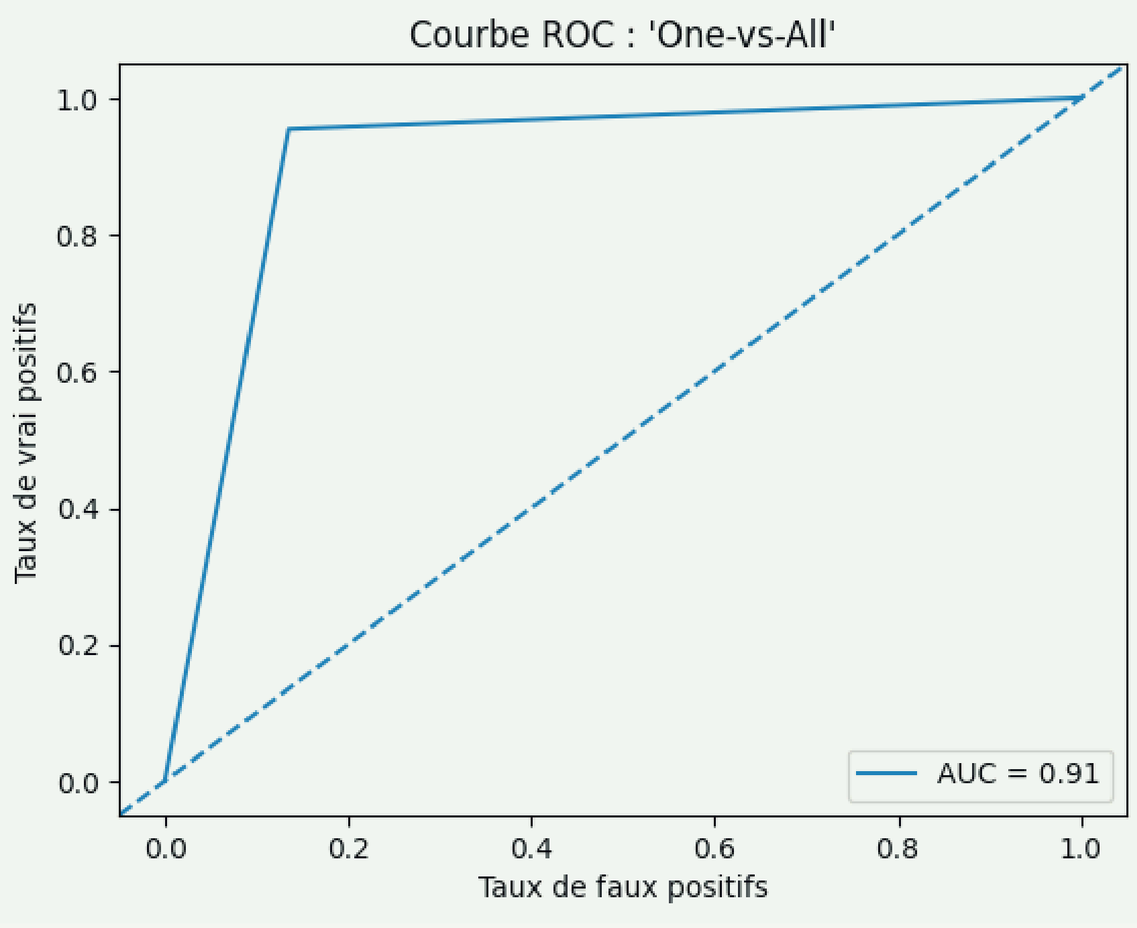
Distinction entre Mildiou tardif et précoce

Modèles de Machine Learning

Résultats

Accuracy = 0.95



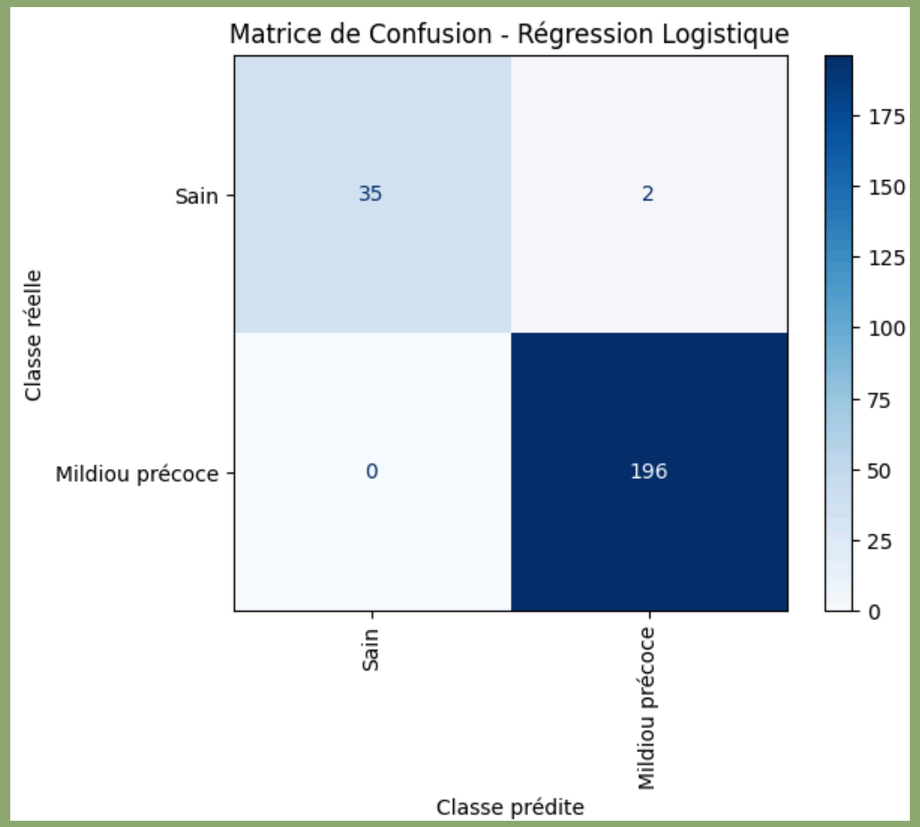


Modèles de Machine Learning

Résultats

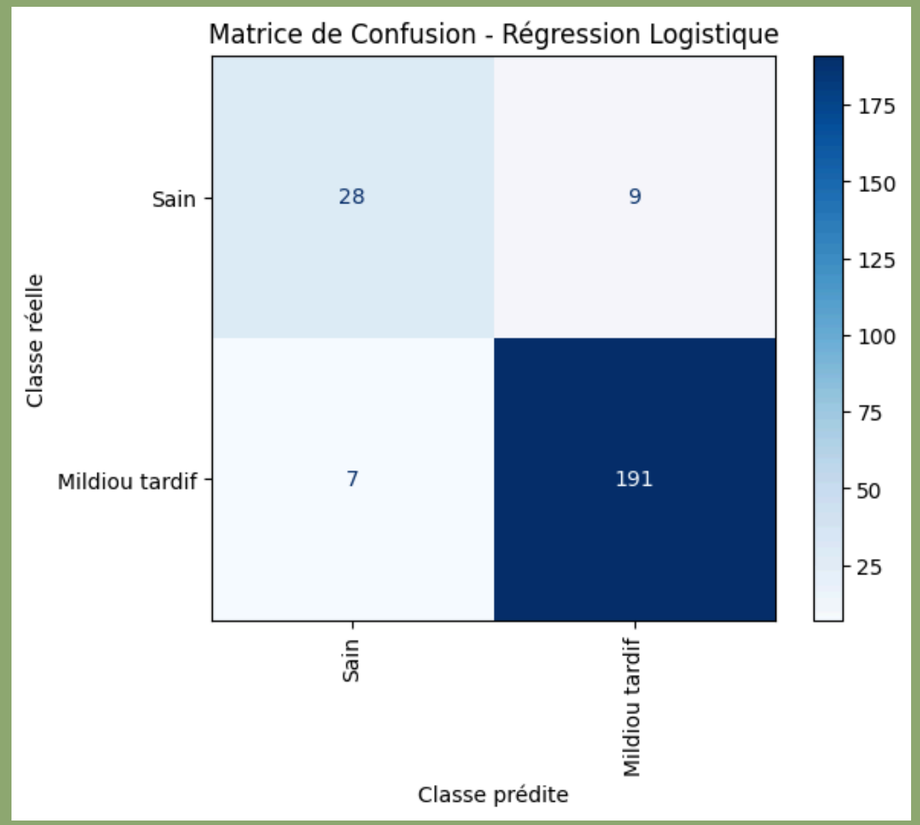
1

Accuracy = 0.99



2

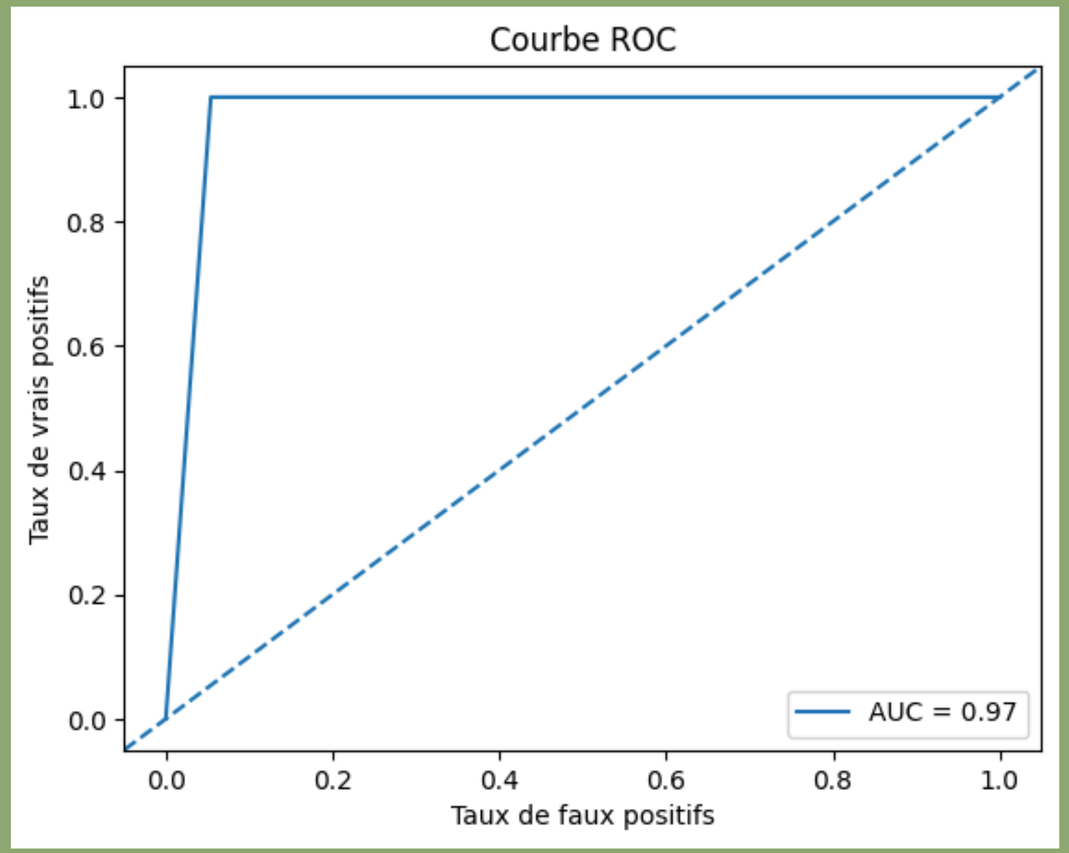
Accuracy = 0.93



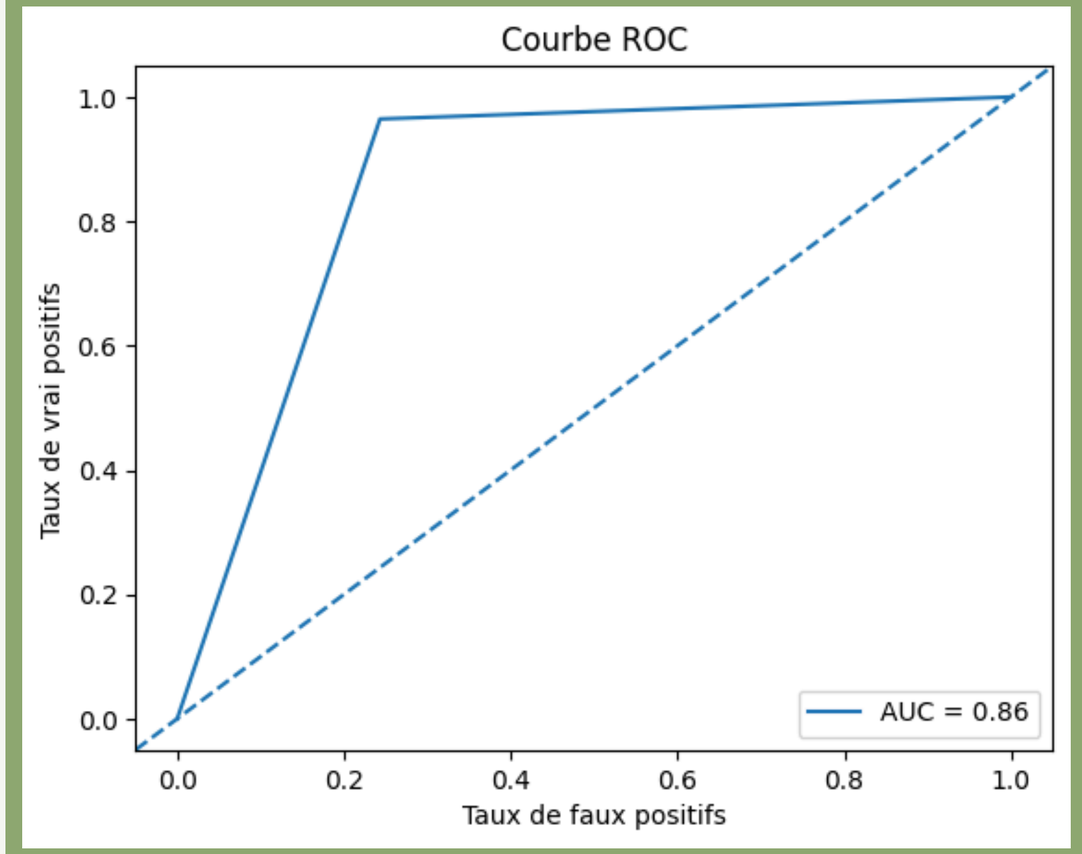
Modèles de Machine Learning

Résultats

1



2



SVM

Objectifs de la recherche

1

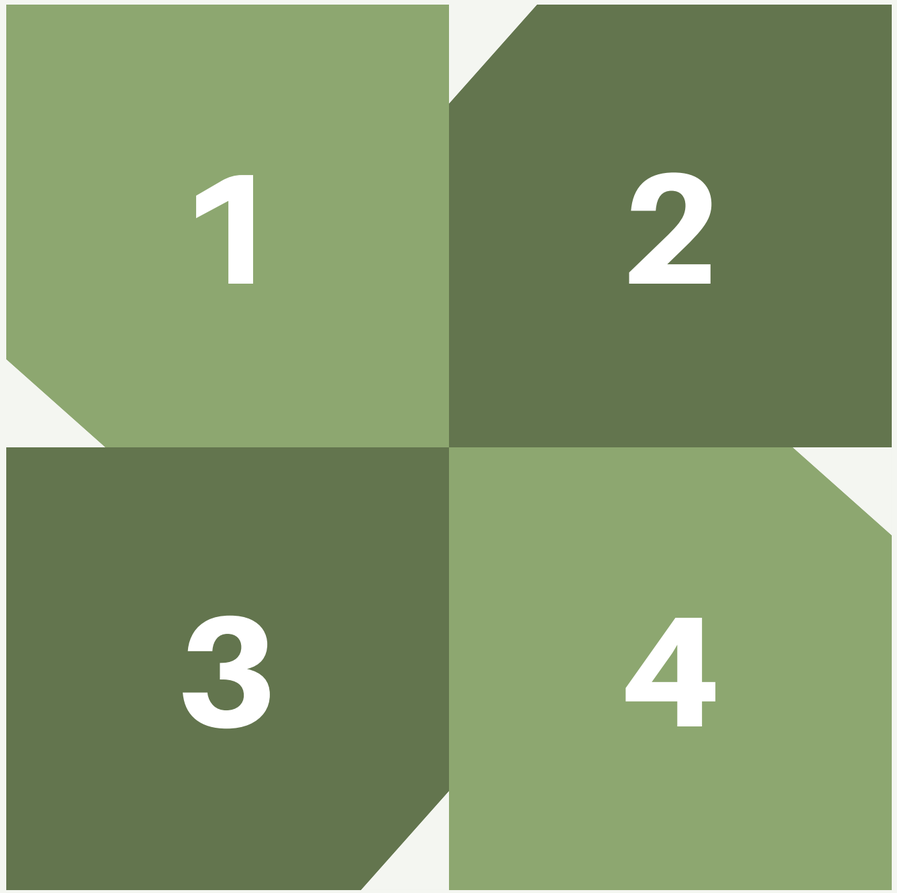
Choix de la méthode

2

Choix des hyperparamè-tres

Choix de la méthode

Ramener un problème de classification ou de discrimination à un hyperplan



Efficace dans les espaces de haute dimension

Utilise un sous-ensemble de points d’entraînement dans la fonction de décision (appelé vecteurs de support), il est donc également efficace en mémoire.

Différentes fonctions du noyau peuvent être spécifié pour la fonction de décision

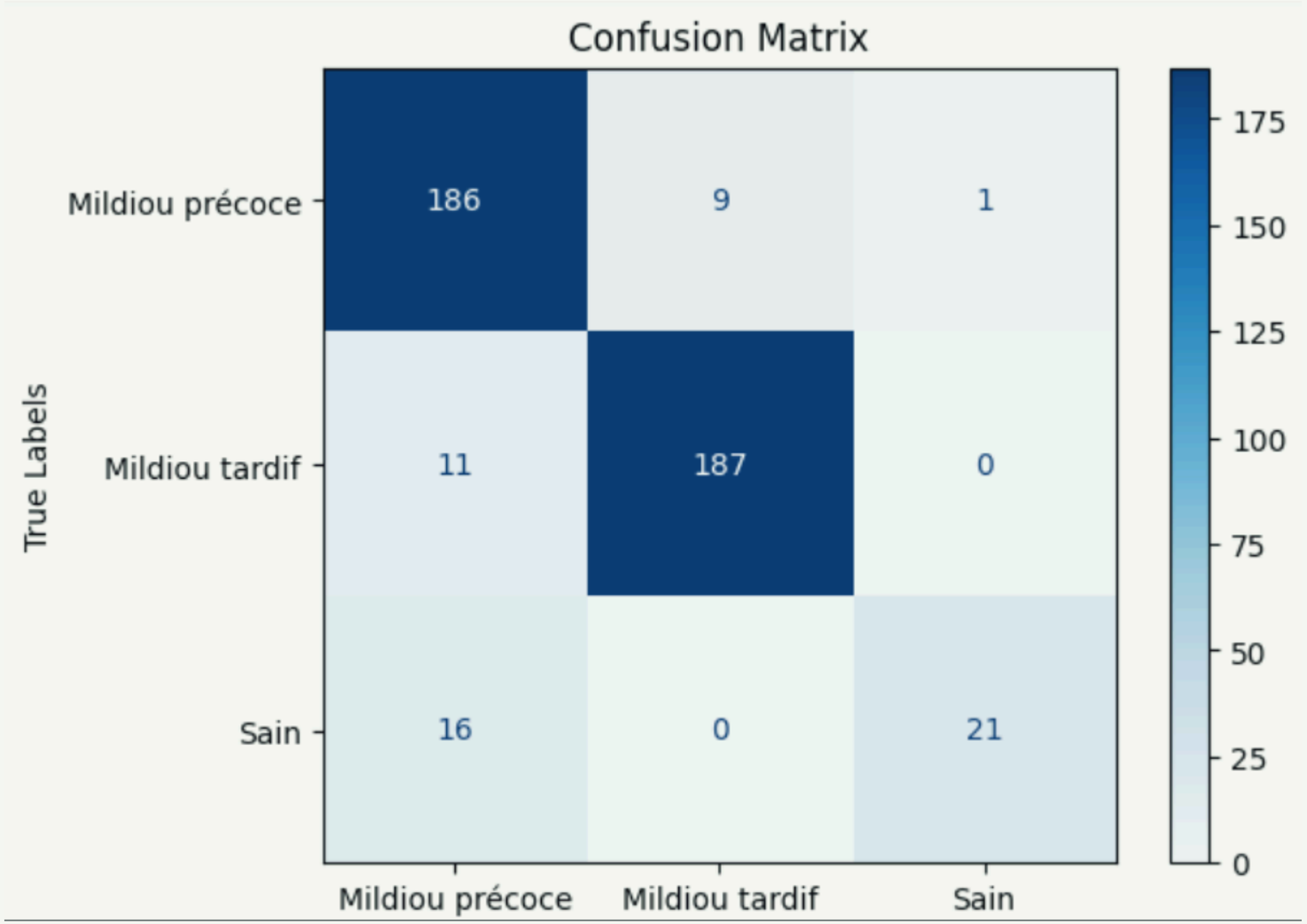
Choix des hyperparamètres

- un contre un pour SVC : des classificateurs sont construits et chacun d'entre eux forme des données à partir de deux classes.
- kernel (noyau) : P
- decision_function_shape : On choisit d'appliquer la méthode de classificateur un contre le reste au lieu d'un contre un

Résultats

Accuracy: 91.42%

	precision	recall	f1-score	support
0	0.87	0.95	0.91	196
1	0.95	0.94	0.95	198
2	0.95	0.57	0.71	37
accuracy			0.91	431
macro avg	0.93	0.82	0.86	431
weighted avg	0.92	0.91	0.91	431


Confusion Matrix
Mildiou précoce
Mildiou tardif
Sain
True Labels
186
9
1
11
187
0
16
0
21
Mildiou précoce
Mildiou tardif
Sain
175
150
125
100
75
50
25
0

CONCLUSION POUR LES SVM

Le modèle fonctionne très bien pour les classes majoritaires (0 et 1) avec des précisions, rappels et F1-scores élevés

DecisionTreeClassifier

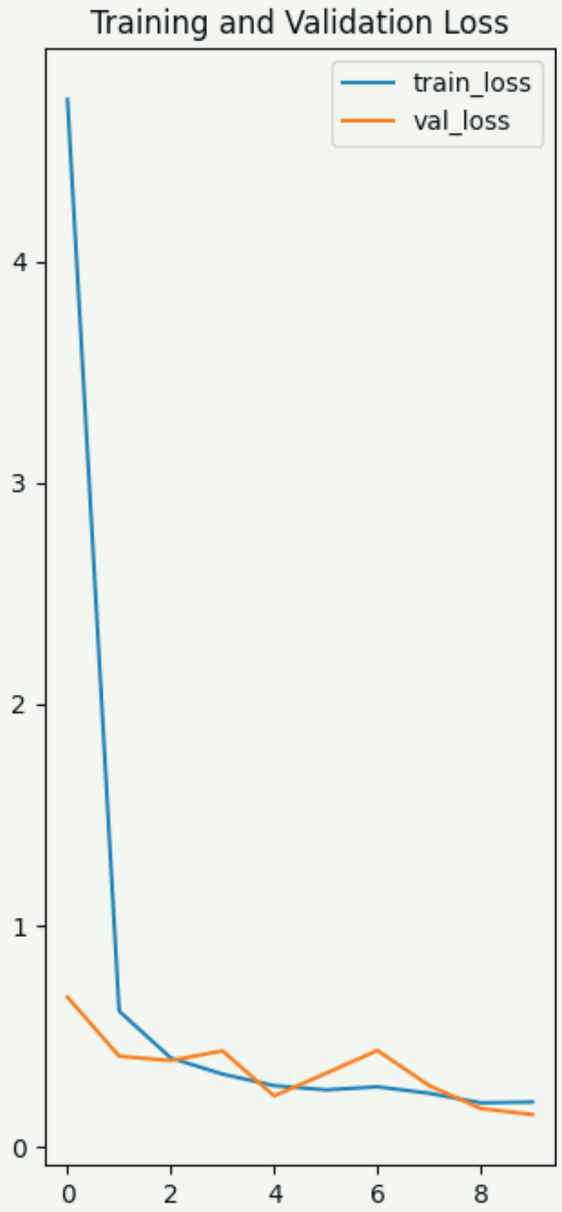
Objectifs de la recherche

1

Choix de la méthode

2

Choix des hyperparamètres


Training and Validation Loss
train_loss
val_loss
0
1
2
3
4
0
2
4
6
8

CHOIX DE LA MÉTHODE

Rapides à entraîner

robustes aux données bruitées

Permet de contrôler la complexité de l’arbre, ce qui permet de prévenir le surapprentissage

ne nécessite pas de normaliser ou de transformer les données.

Demandent moins de données et de temps de calcul

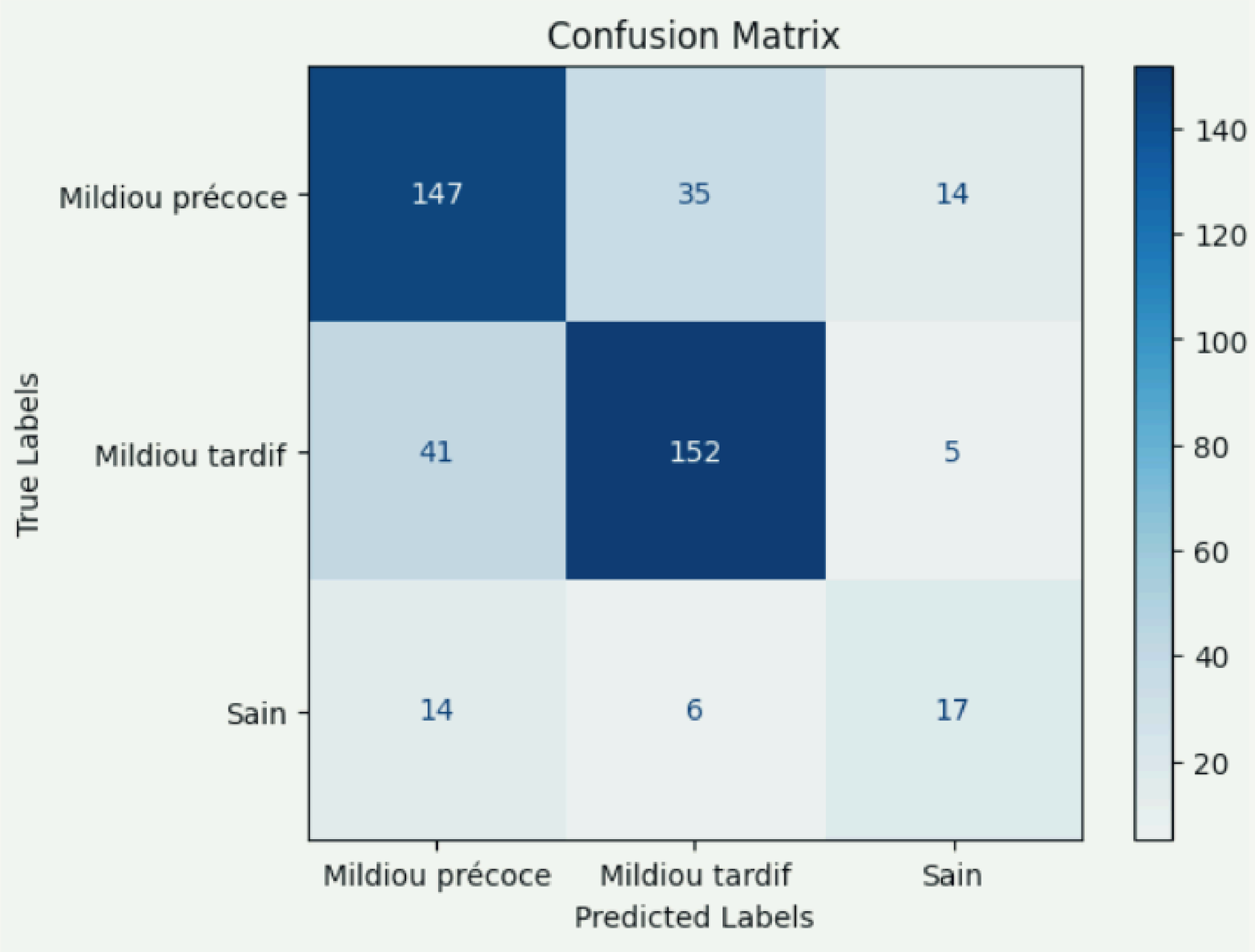
Choix des hyperparamètres

- max_depth=10 : Limite la profondeur de l'arbre à 10 niveaux. Cela empêche l'arbre de devenir trop complexe et de surapprendre sur les données d'entraînement
- min_samples_split=2 : C'est le nombre minimum d'échantillons requis pour diviser un nœud. Avec une valeur de 2, on autorise le modèle à diviser un nœud dès qu'il y a au moins 2 échantillons.
- min_samples_leaf=1 : spécifie le nombre minimum d'échantillons dans une feuille. on autorise des feuilles avec un seul échantillon.
- criterion='gini' : l'objectif est de minimiser l'impureté (la diversité des classes) dans les nœuds.

Résultats

Decision Tree Accuracy: 0.73

	precision	recall	f1-score	support
0	0.73	0.75	0.74	196
1	0.79	0.77	0.78	198
2	0.47	0.46	0.47	37
accuracy			0.73	431
macro avg	0.66	0.66	0.66	431
weighted avg	0.73	0.73	0.73	431


Confusion Matrix
Mildiou précoce
Mildiou tardif
Sain
True Labels
147
35
14
41
152
5
14
6
17
Mildiou précoce
Mildiou tardif
Sain
Predicted Labels
140
120
100
80
60
40
20

CONCLUSION POUR LES DECISION TREE

DecisionTreeClassifier donne des résultats acceptables sur les classes majoritaires, il a des difficultés à traiter la classe minoritaire (classe 2)

Random Forest

Objectifs de la recherche

1

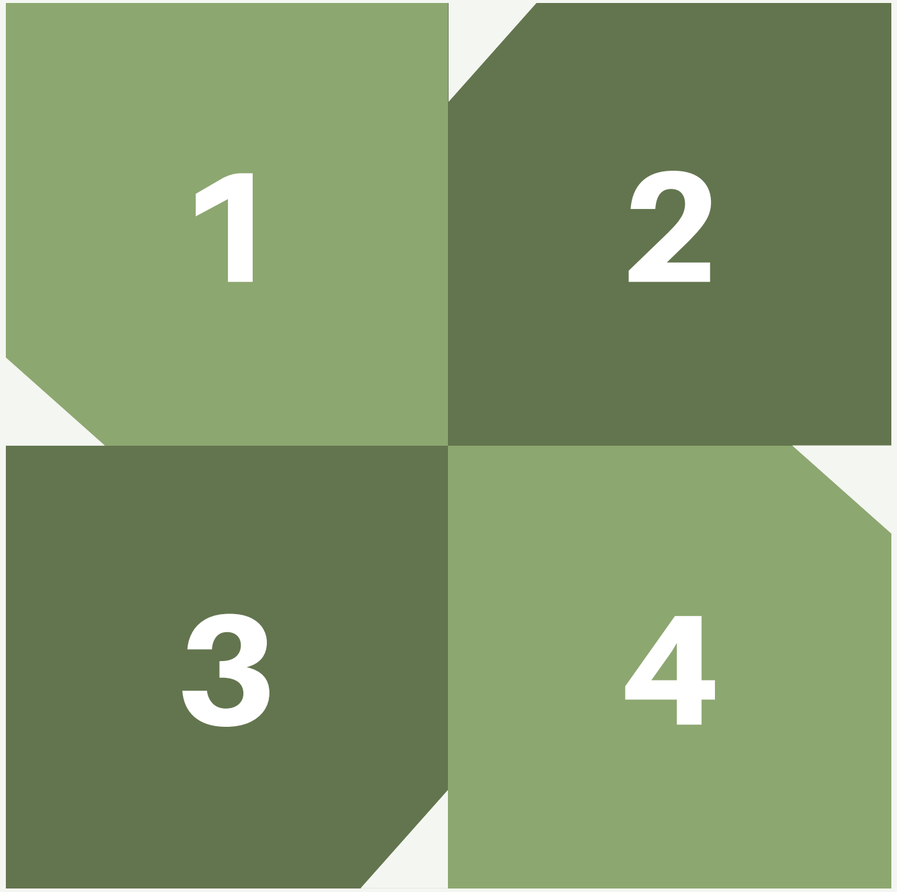
Choix de la méthode

2

Choix des hyperparamètres

Choix de la méthode

Random Forest est également très résistant au bruit dans les données



Contrairement à un modèle de décision unique (comme le DecisionTreeClassifier), le Random Forest utilise plusieurs arbres, ce qui réduit la variance du modèle.

Le modèle Random Forest est connu pour ses performances solides sur les problèmes de classification grâce à la combinaison de plusieurs arbres de décision.

Mieux équipé pour traiter des classes déséquilibrées que des modèles plus simples, grâce à sa capacité à effectuer un échantillonnage aléatoire lors de la construction de chaque arbre.

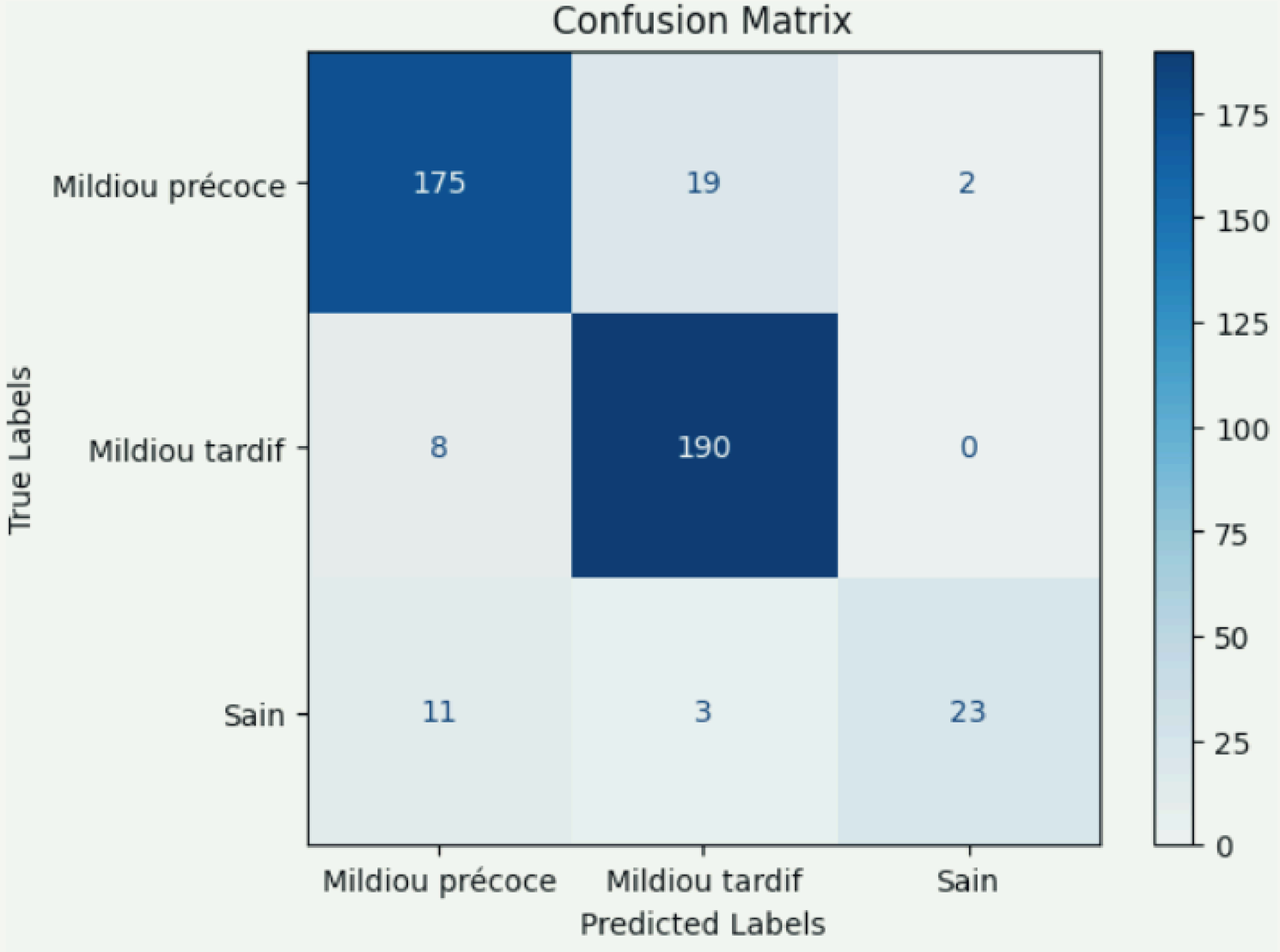
Choix des hyperparamètres

- n_estimators=100 : 100 arbres de décision se qui améliore la stabilité des prédictions
- max_features='sqrt' : Pour chaque division dans un arbre, seulement la racine carrée du nombre total de caractéristiques sera considérée. Cela permet d'accélérer le calcul .

Résultats

Random Forest Accuracy: 0.90

	precision	recall	f1-score	support
0	0.90	0.89	0.90	196
1	0.90	0.96	0.93	198
2	0.92	0.62	0.74	37
accuracy			0.90	431
macro avg	0.91	0.82	0.86	431
weighted avg	0.90	0.90	0.90	431


Confusion Matrix
Mildiou précoce
Mildiou tardif
Sain
True Labels
175
19
2
8
190
0
11
3
23
Mildiou précoce
Mildiou tardif
Sain
Predicted Labels
175
150
125
100
75
50
25
0

CONCLUSION POUR LES DECISION TREE

Avec une accuracy de 90% et des précisions et F1-scores élevés pour les classes 0 et 1, le modèle Random Forest se comporte très bien pour les deux premières classes, qui ont plus de données d'entraînement. Le modèle a plus de difficultés à bien identifier la classe 2 (62% de rappel)

Modèle CNN

1

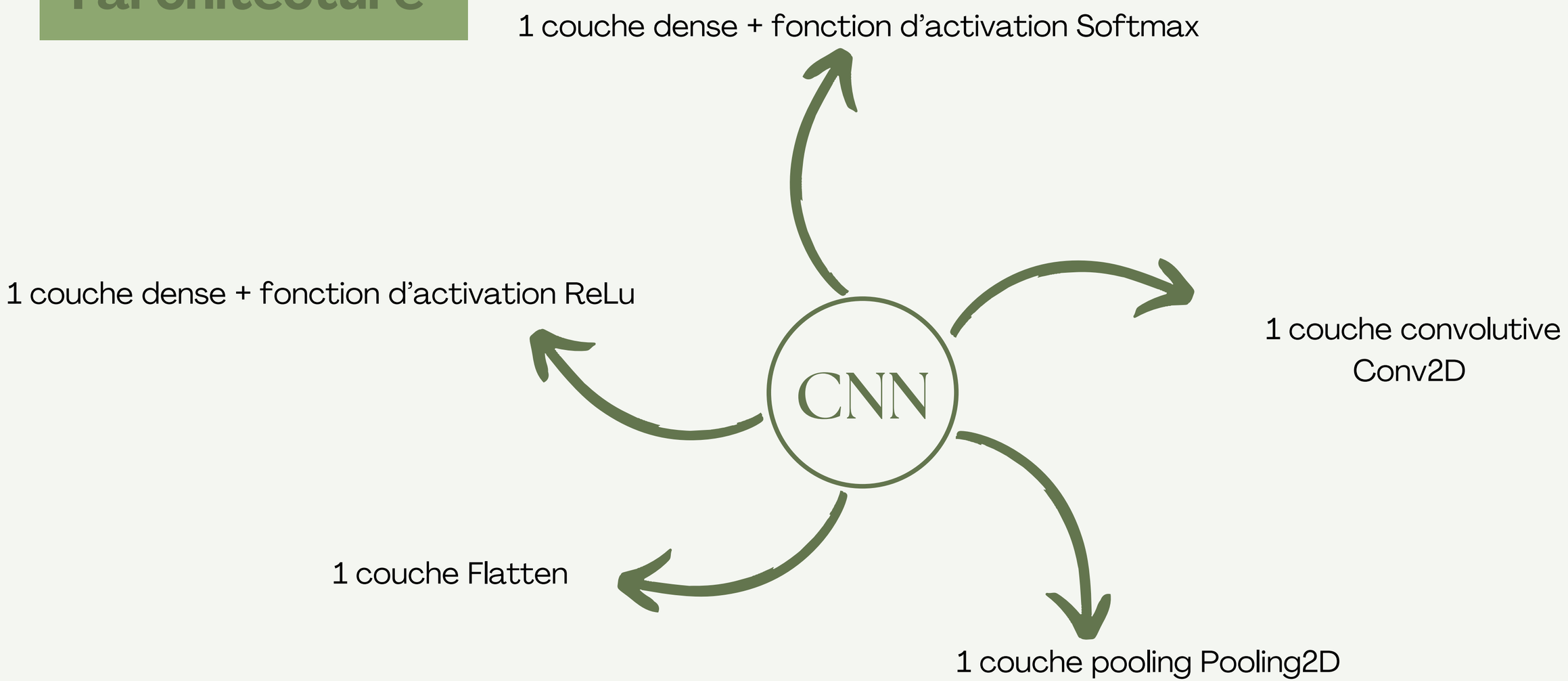
Choix de l’architecture

2

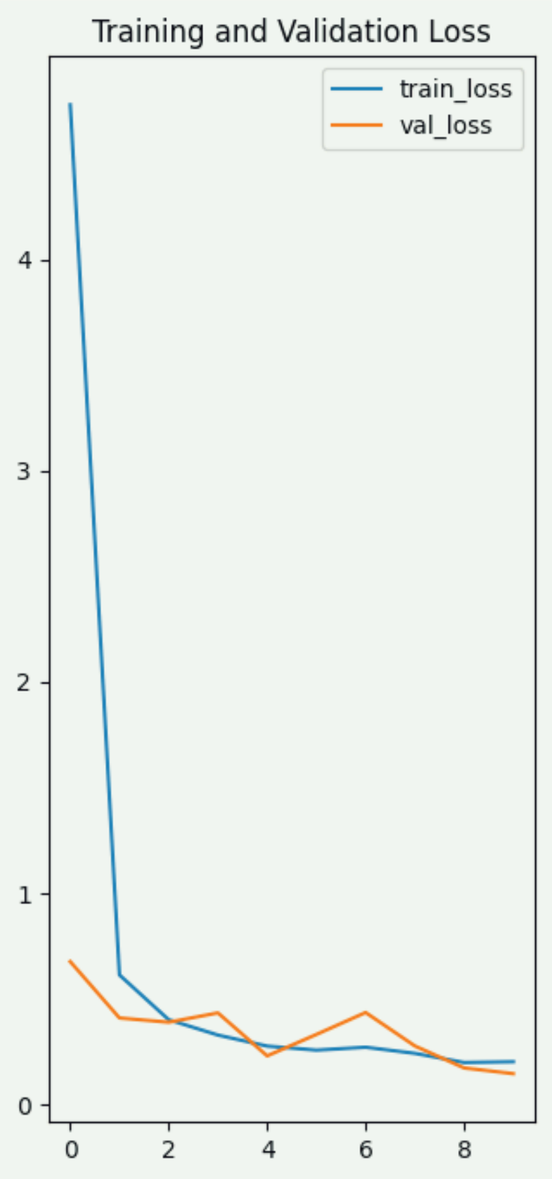
Analyse des performances du modèle

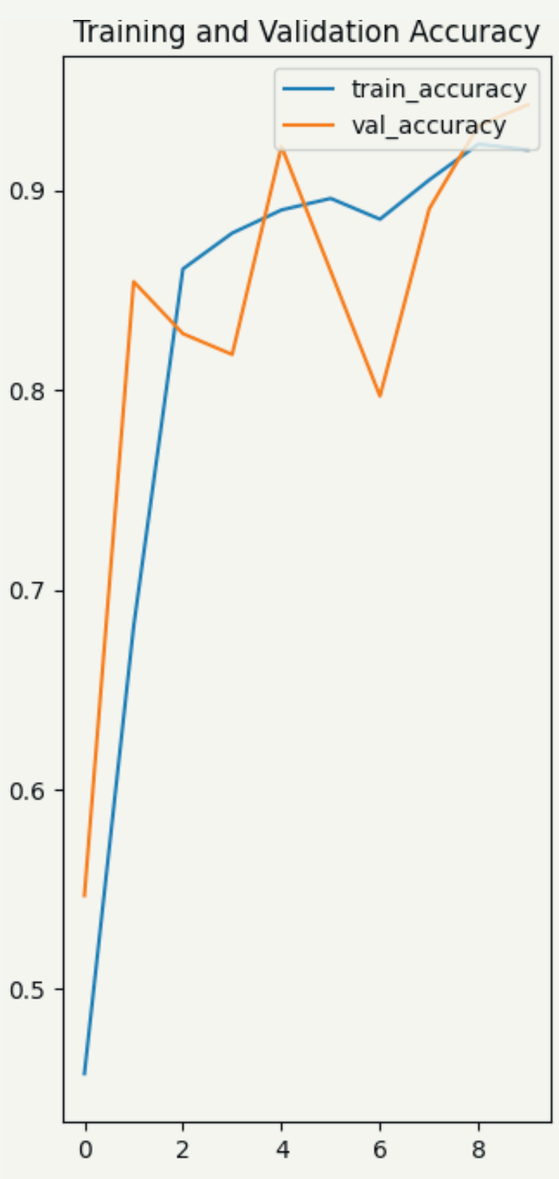
Choix de l'architecture

Modèle CNN



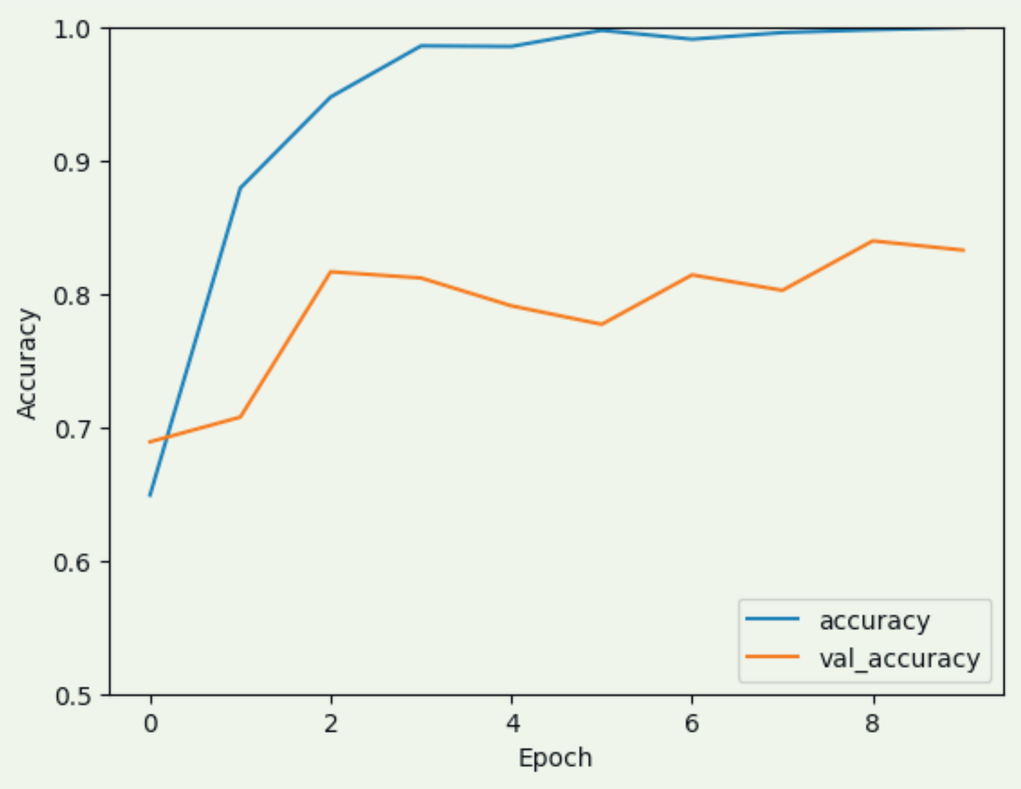
Classes déséquilibrées

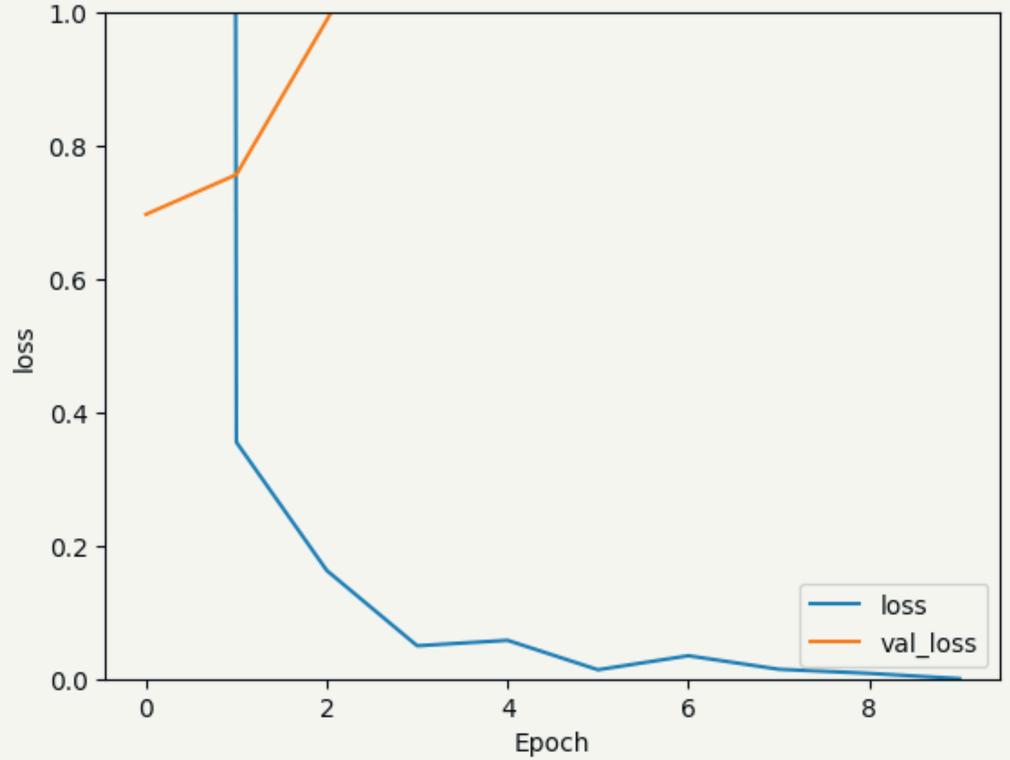




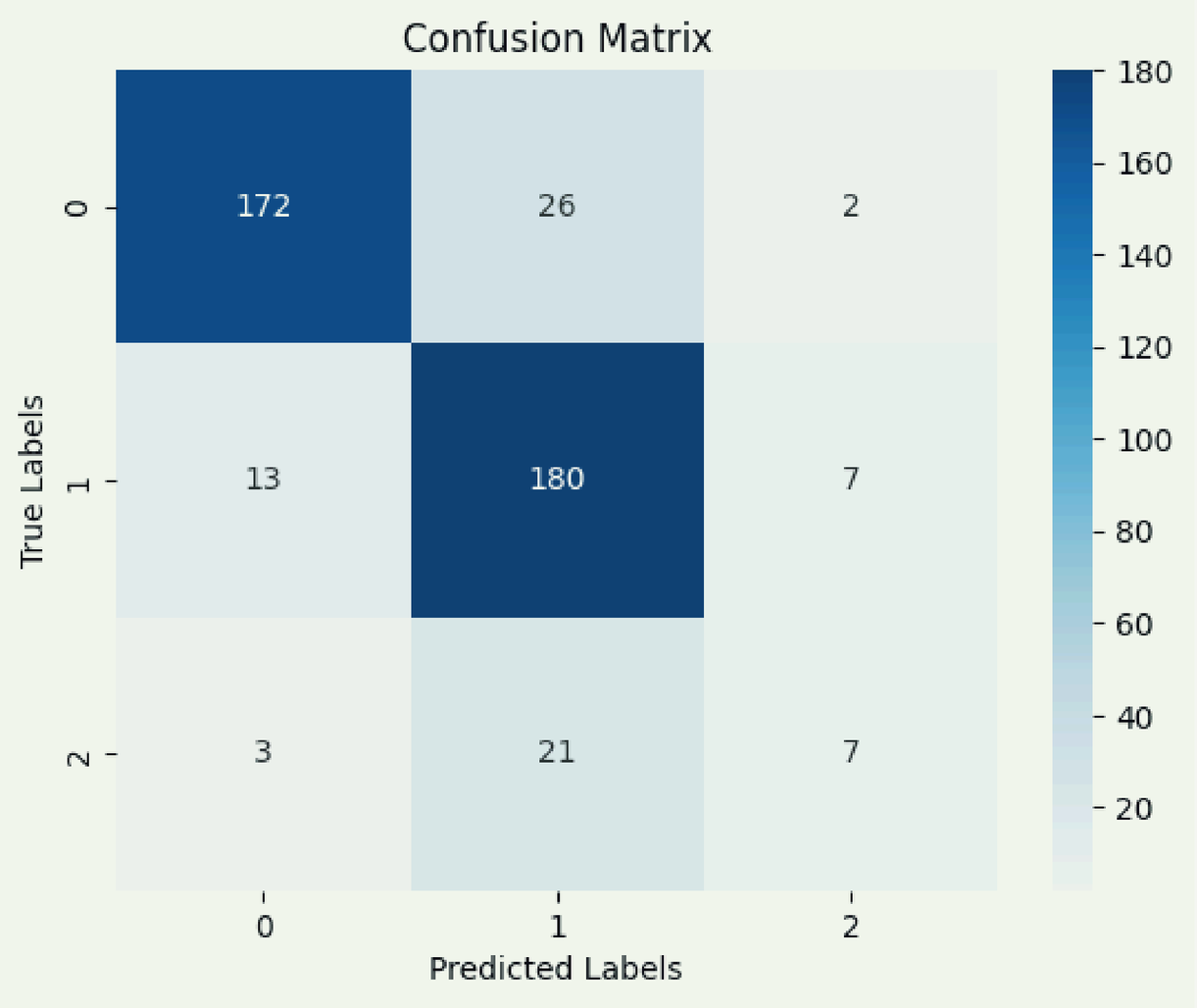
Accuracy : 0.95
Loss: 0.1332

Classes rééquilibrées





Classes rééquilibrées



Accuracy : 0.83

Discussion

Limites

Temps de calcul important, nécessité d'utiliser Google Colab

Nombre de données insuffisant pour l'optimisation des méthodes de Deep Learning

Perspectives

Construction d'un modèle hiérarchique entre les classes

Conclusion

INSTITUT AGRO RENNES-ANGERS

Merci !

Yasmine BOUCHIBTI
Leslie CIETERS
Meryem GRIMAJ