Projet du Data-camp
Single-cell RNA-seq classification

Data Saiyentist

M2 Data Science : Santé, Assurance, Finance

Mercredi 26 Avril 2023

Table of Contents

université PARIS-SACLAY

Introduction

Contexte : Jeu de données scRNA-seq

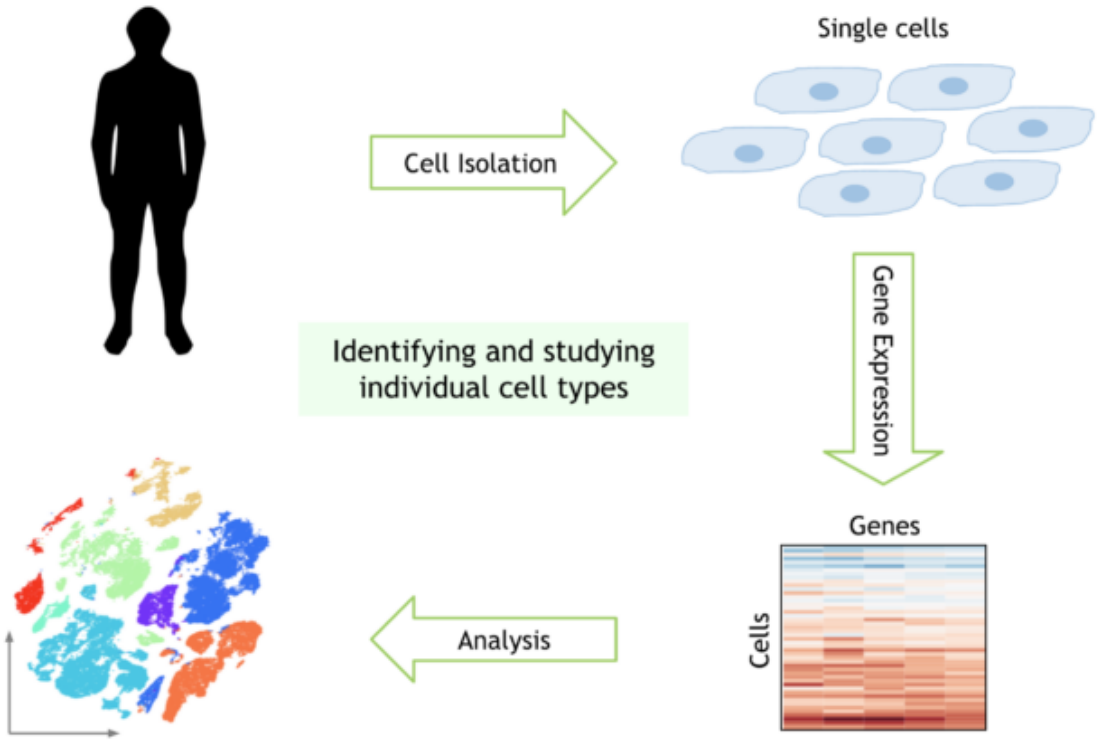


Figure: Workflow for a single-cell RNA sequencing experiment

Jeu de données

- 1500 cellules (1000 entrainements, 500 tests)
- 13551 gènes
- 4 types de cellule :
  1. Cancer-cells
  2. NK-cells
  3. T-cells-CD4+
  4. T-cells-CD8+

Soumission sur RAMP

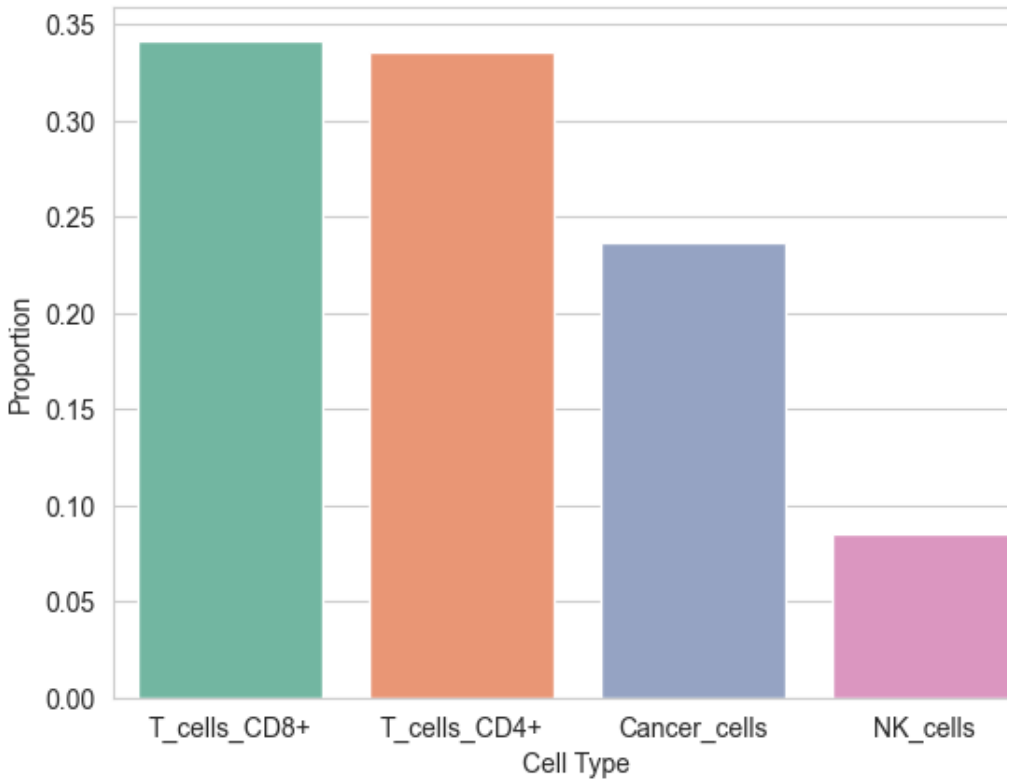
1. *Classifier.py* (prétraitement + pipeline du modèle)
2. 1500 données d'entrainements (public)
3. 1500 donéées de test (privée)

→ Utilisation de Sklearn (datacamp environement)
→ Code disponible sur :

https://github.com/DataSaiyentist/DataCamp_scRNAseq

Exploration des données

Proportion de cellules



1. T-cells-CD4+ 34%
2. T-cells-CD8+ 34%
3. Cancer-cells 24%
4. NK-cells 8%

Figure: Proportion of cell types

→ Jeu de données disproportionné

Exploration des données

Visualisation lignes/colonnes

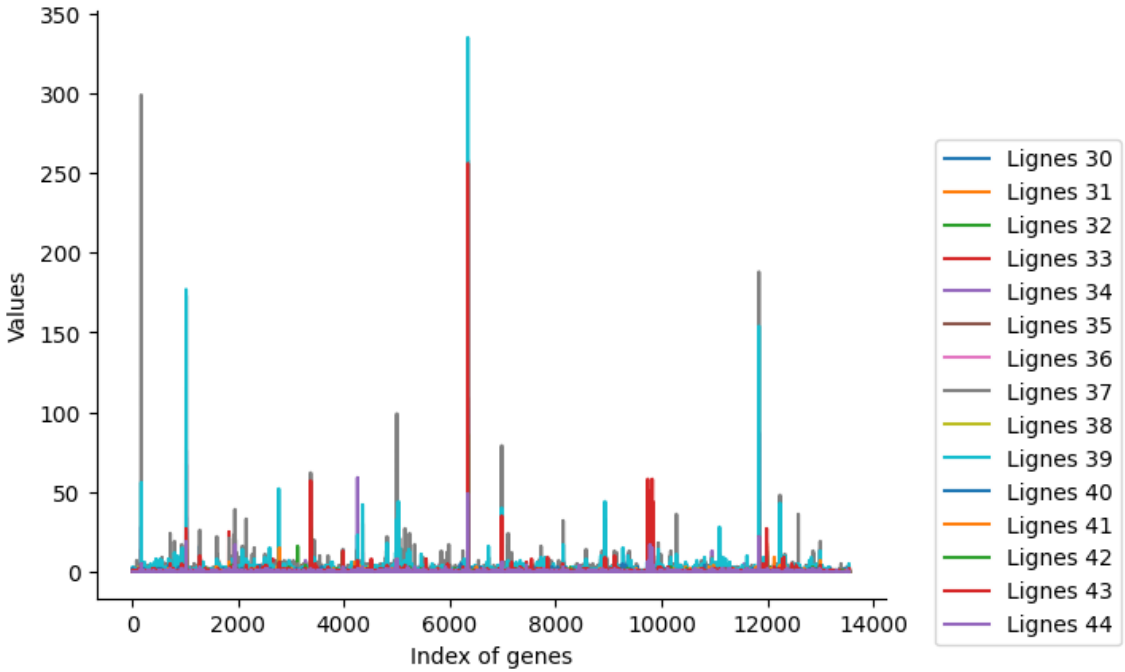


Figure: Visualisation de certaines lignes

→ Difficulté à extraire des tendances

→ Difficulté à distinguer des cellules

Exploration des données

Visualisation des données - Violin plot

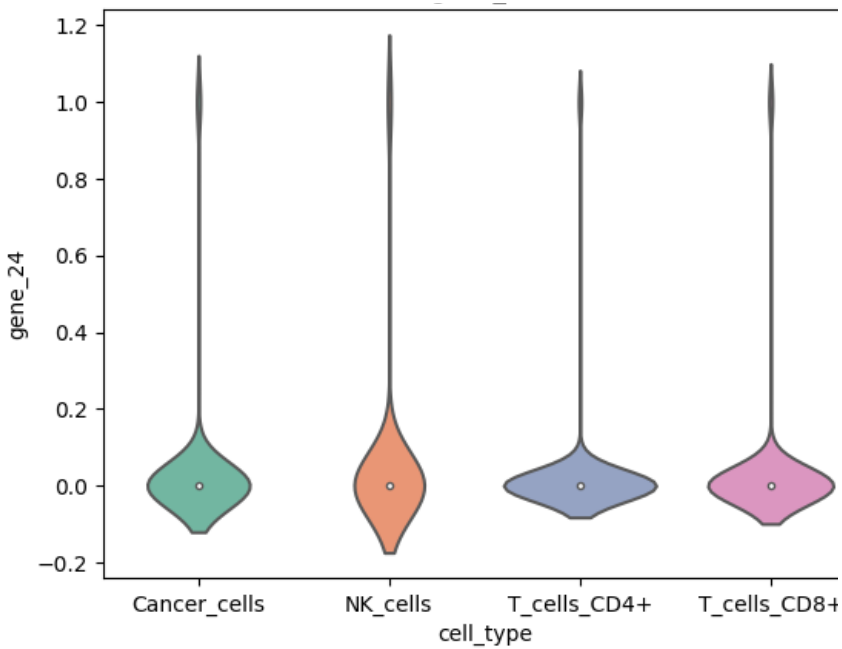


Figure: Visualisation du gene-32

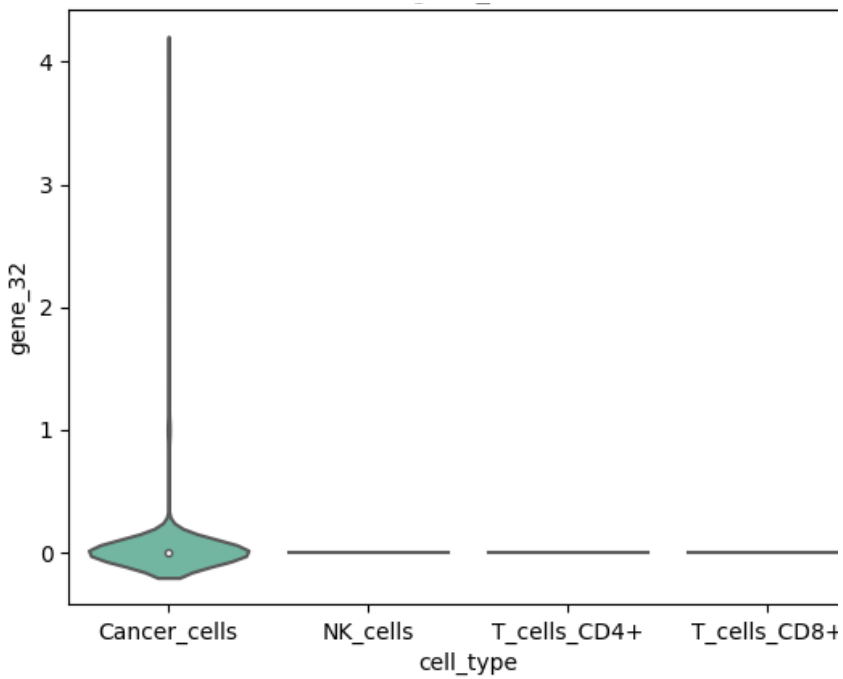


Figure: Visualisation du gene-24

→ Difference remarquable dans les expressions de gènes

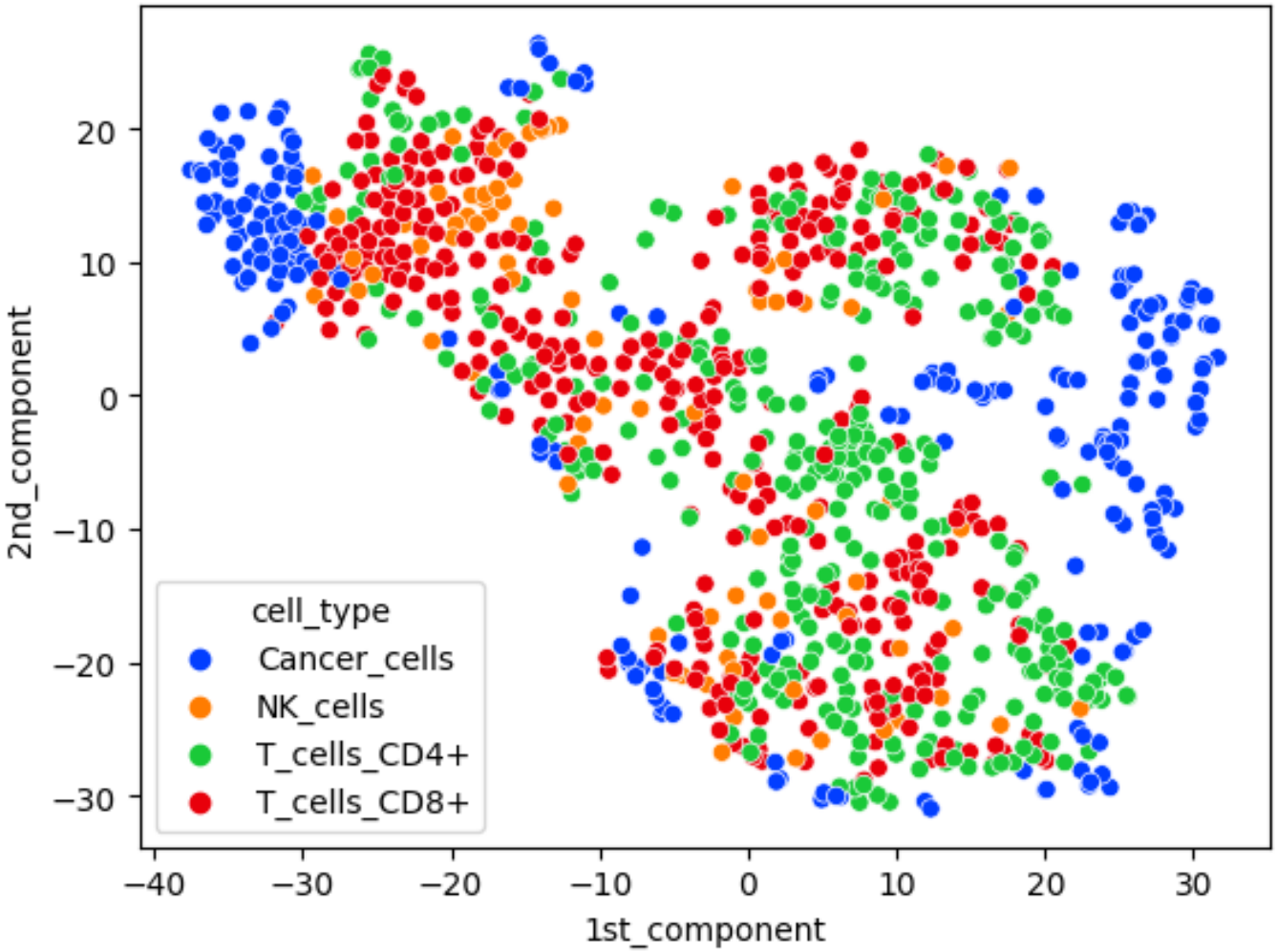


Figure: Visualisation du train avec un t-SNE

Traitement des données

Contrôle qualité

Tests de différents critères sur les cellules :

- le nombre total de gènes qui se sont exprimés
- le nombre total de gènes différents qui se sont exprimés
- le call rate, ie. le pourcentage de gènes exprimés
- le novelty score, ie. le logarithme du call rate

$\rightarrow$ Pas d'augmentation significative des résultats

université PARIS-SACLAY

Méthodes de séléction de variables utilisées

- ACP
- t-SNE (t-Distributed Stochastic Neighbor Embedding)
- SelectKBest
- Réduction de la Variance

$\rightarrow$ Tests de différentes combinaisons en termes de précision pondérée

Modélisation

Modèles de machine learning utilisés

- SVM (Support Vector Machines)
- KNN (K-Nearest Neighbors)
- Random Forest
- Gradient Boosting
- MLP (perceptron multicouche) : 2 couches avec 1000 neurones/couche

Métriques d'évaluation

- précision pondérée (balanced accuracy)
- matrice de confusion
- courbes ROC

Résultats

Visualisation des données de test avec t-SNE

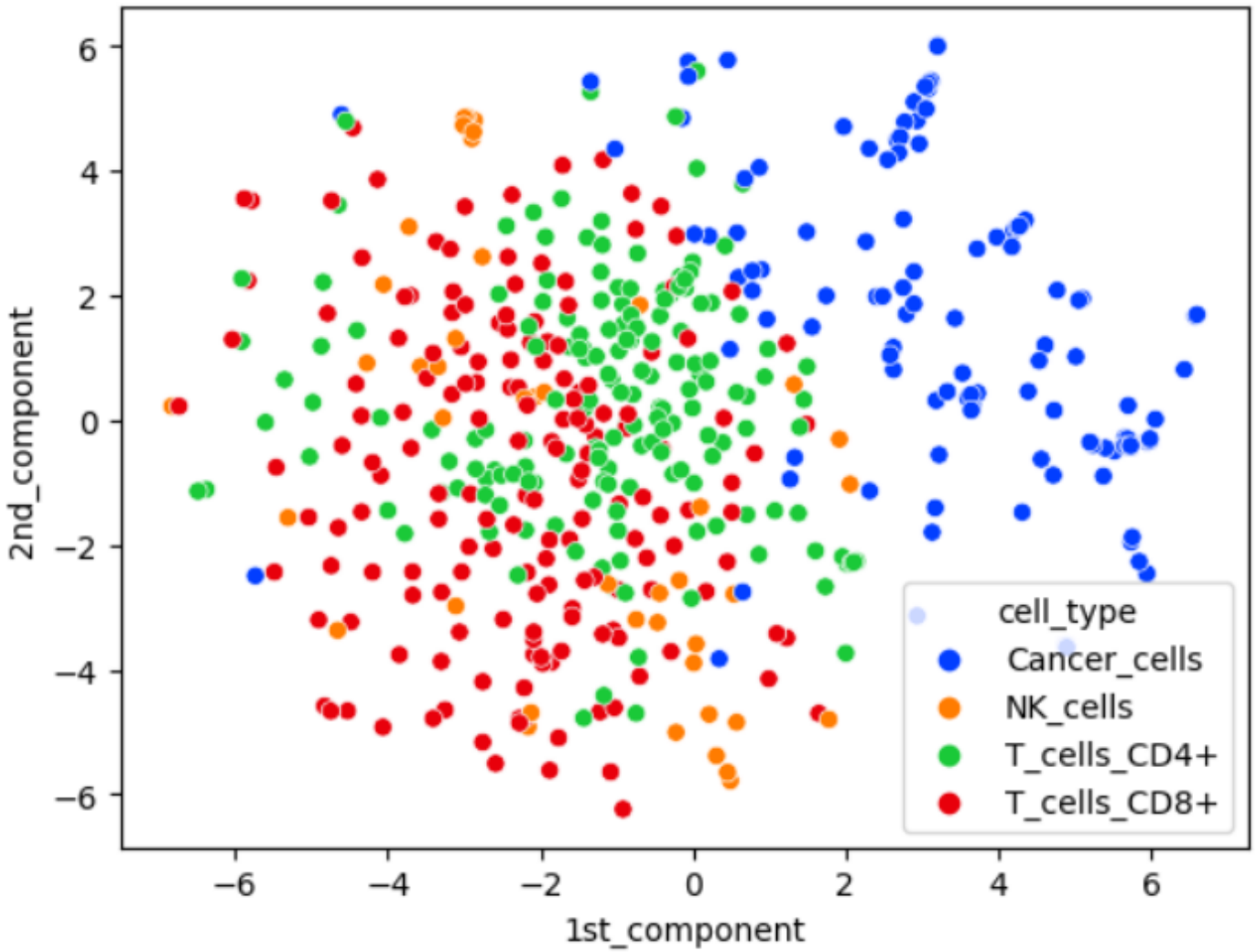


Figure: Visualisation de test avec un t-SNE

Résultats

Visualisation des prédictions avec un t-SNE

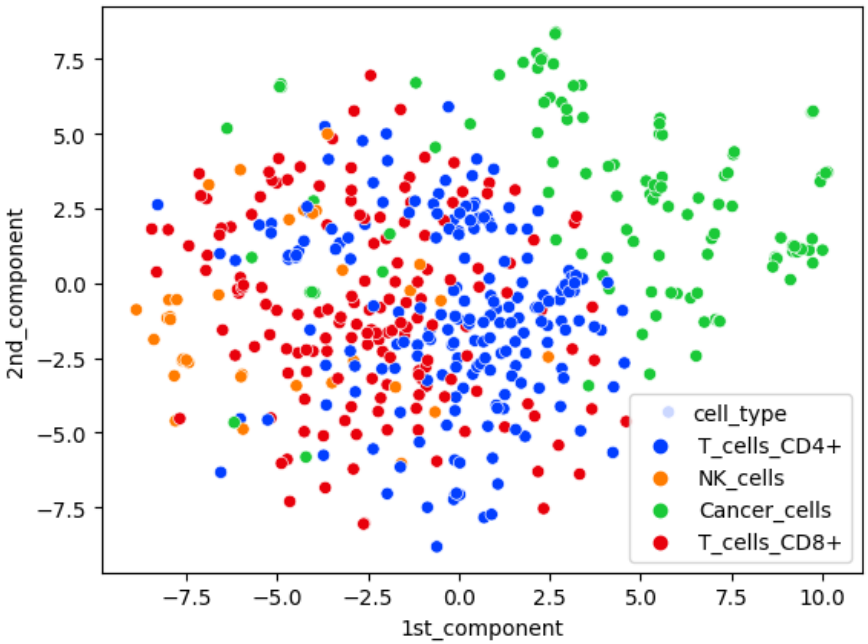


Figure: t-SNE avec SVM

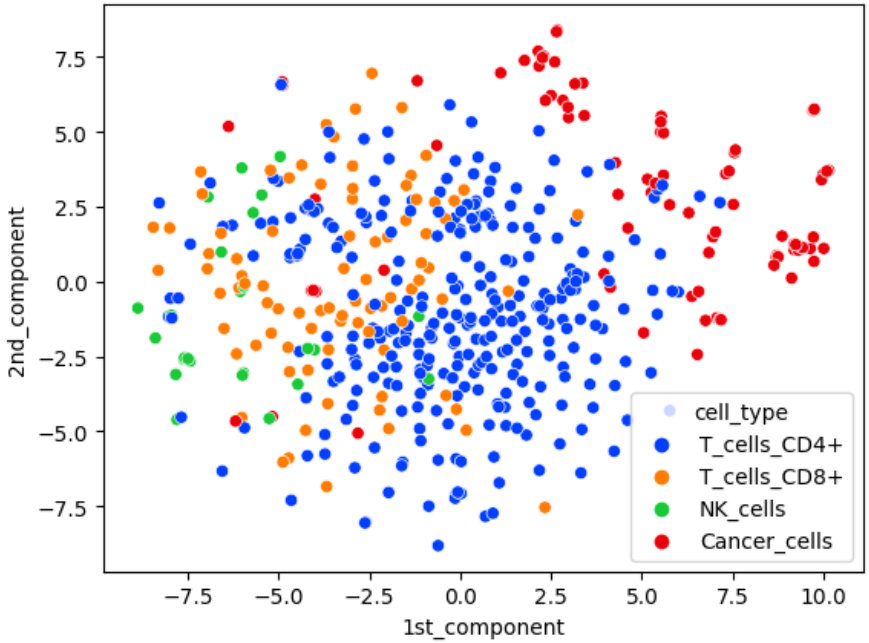


Figure: t-SNE avec KNN

$\rightarrow$ Balanced accuracy $= 0.8$

$\rightarrow$ Balanced accuracy $= 0.61$

université PARIS-SACLAY

Résultats

Visualisation des prédictions avec un t-SNE

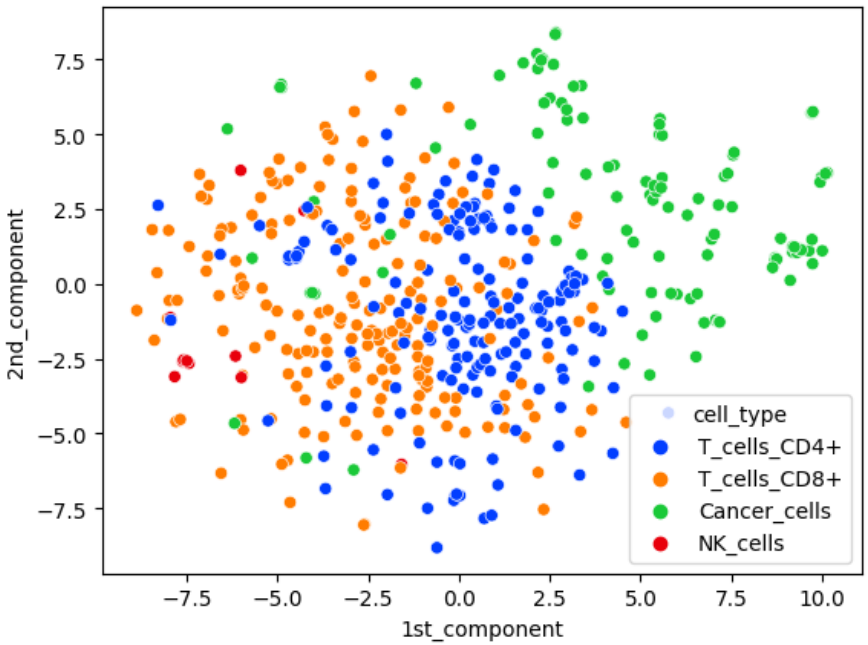


Figure: t-SNE avec RandomForest

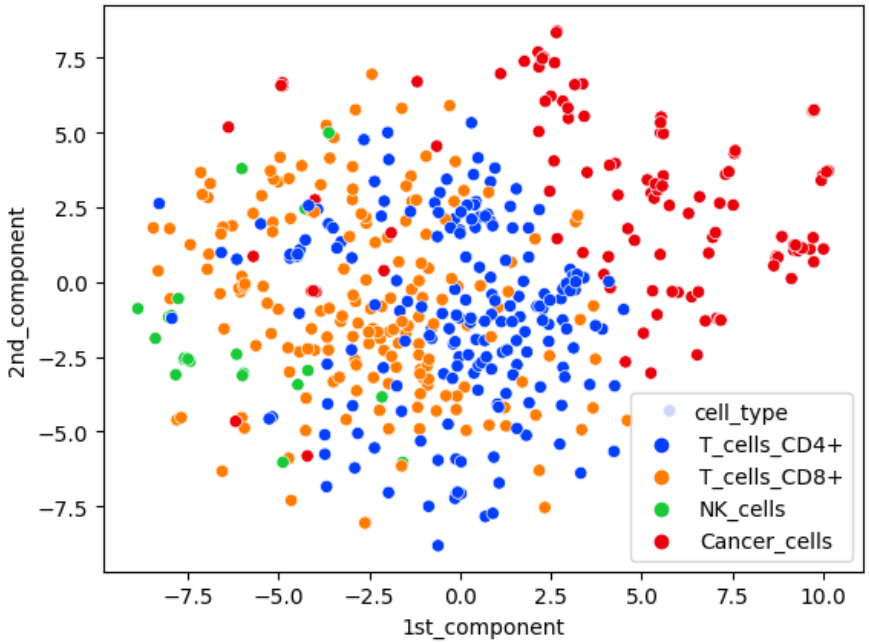


Figure: t-SNE avec GradientBoosting

→ Balanced accuracy = 0.75

→ Balanced accuracy = 0.84

Résultats

Visualisation des prédictions : Matrice de confusion

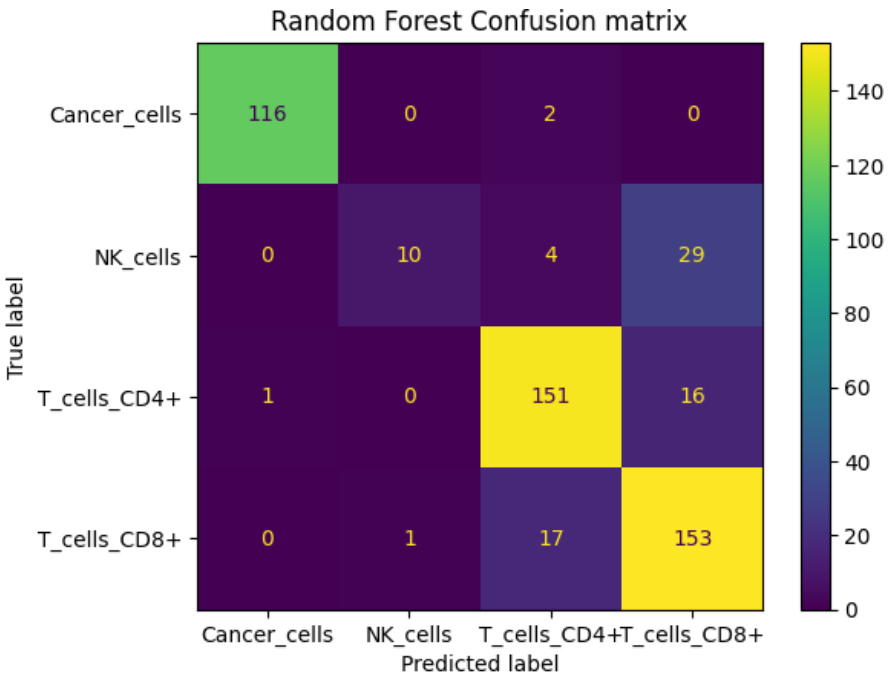


Figure: SVM

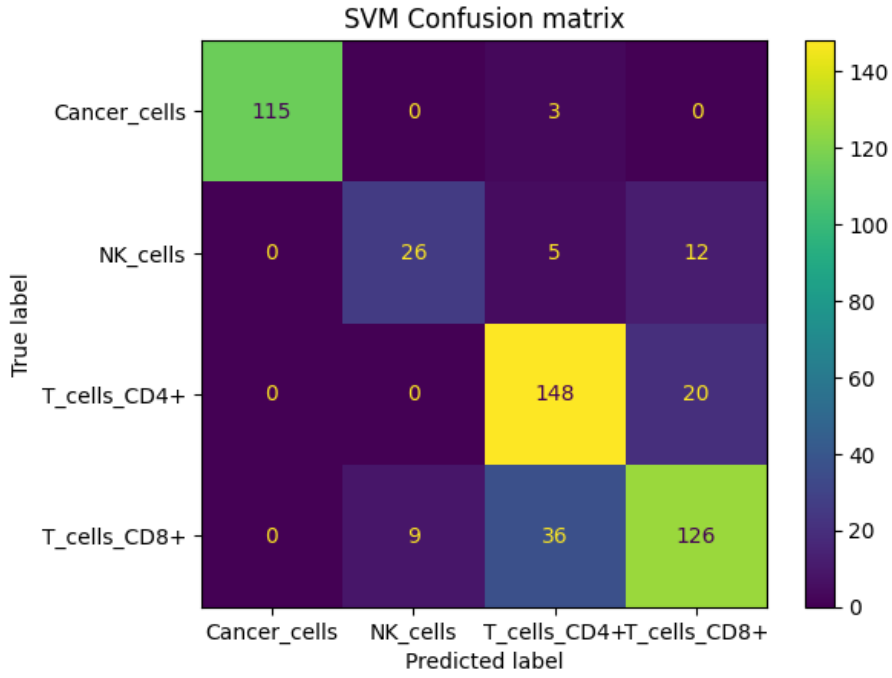


Figure: KNN

Résultats

Visualisation des prédictions : Matrice de confusion

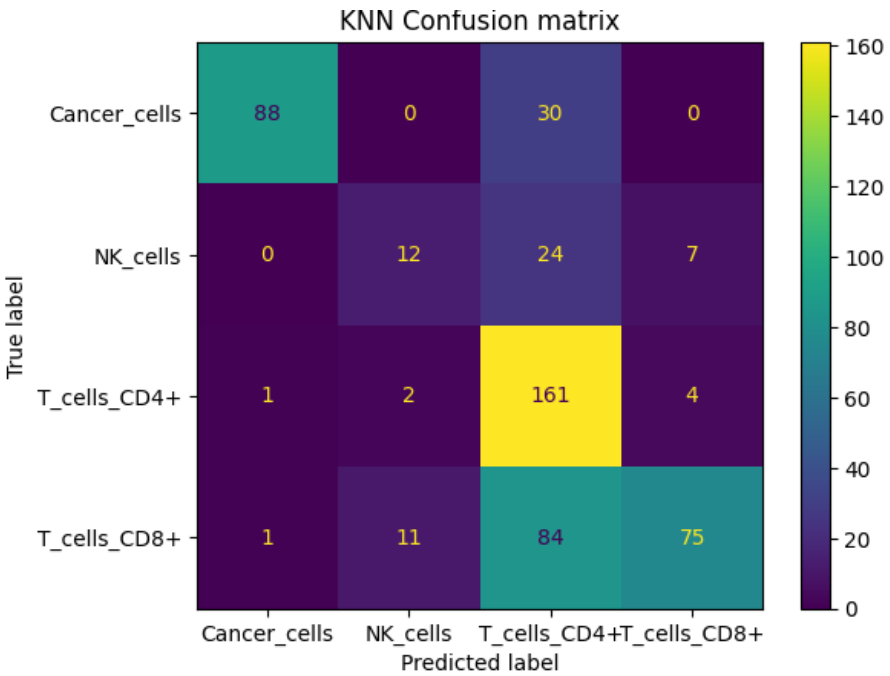


Figure: RandomForest

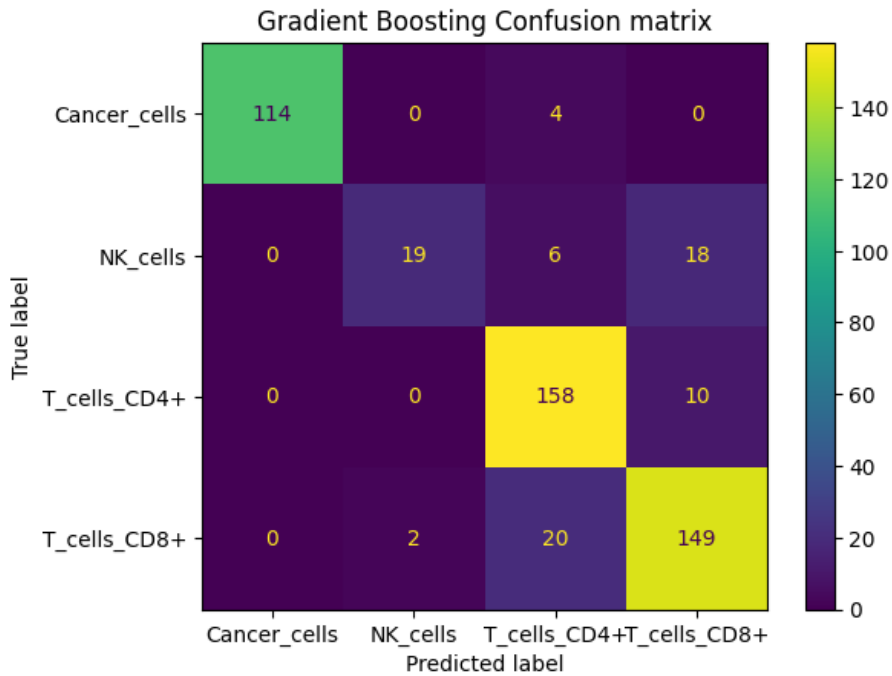


Figure: GradientBoosting

Résultats

Visualisation des prédictions : Courbes ROC

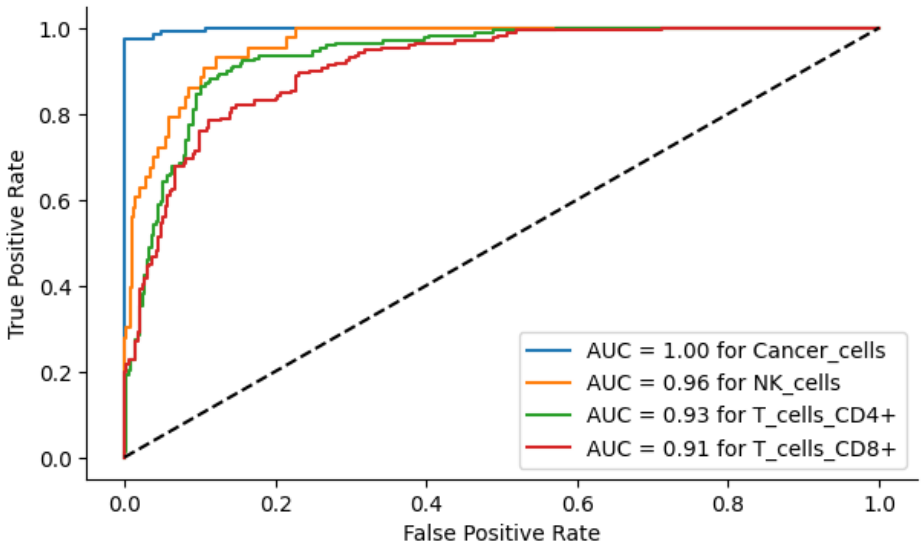


Figure: SVM

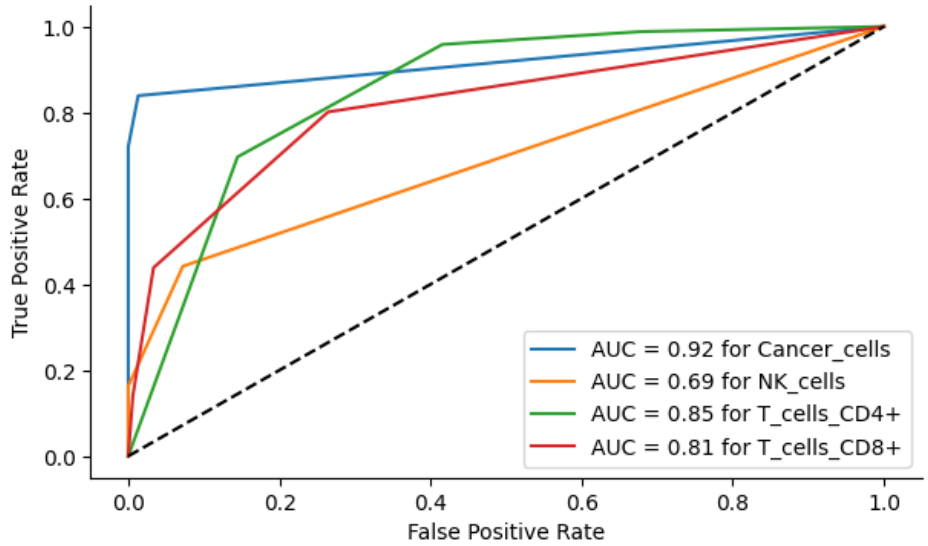


Figure: KNN

Résultats

Visualisation des prédictions : Courbes ROC

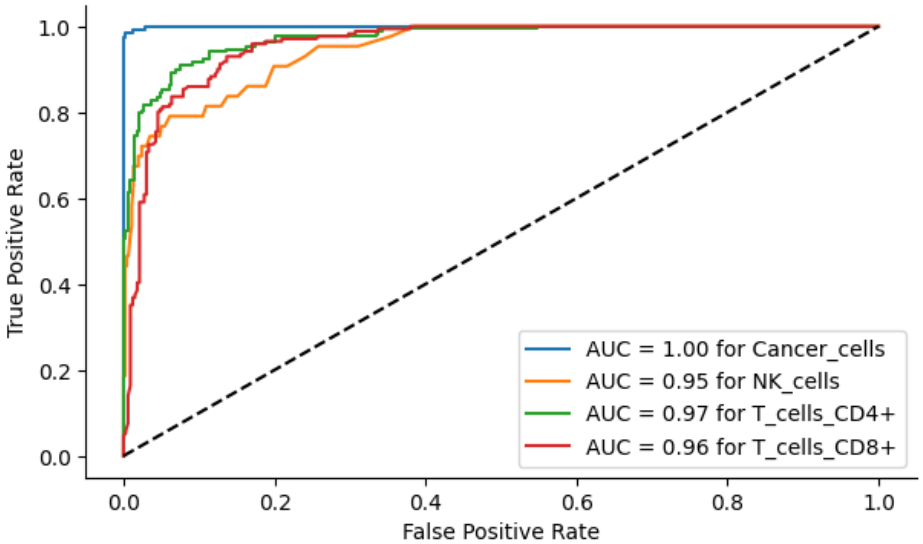


Figure: RandomForest

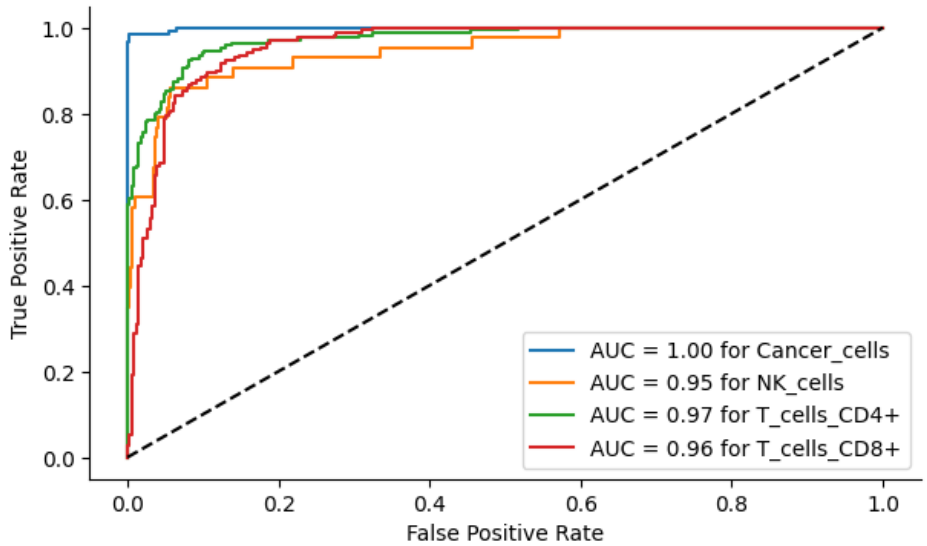


Figure: GradientBoosting

Résultats

Visualisation des prédictions : Meilleur modèle

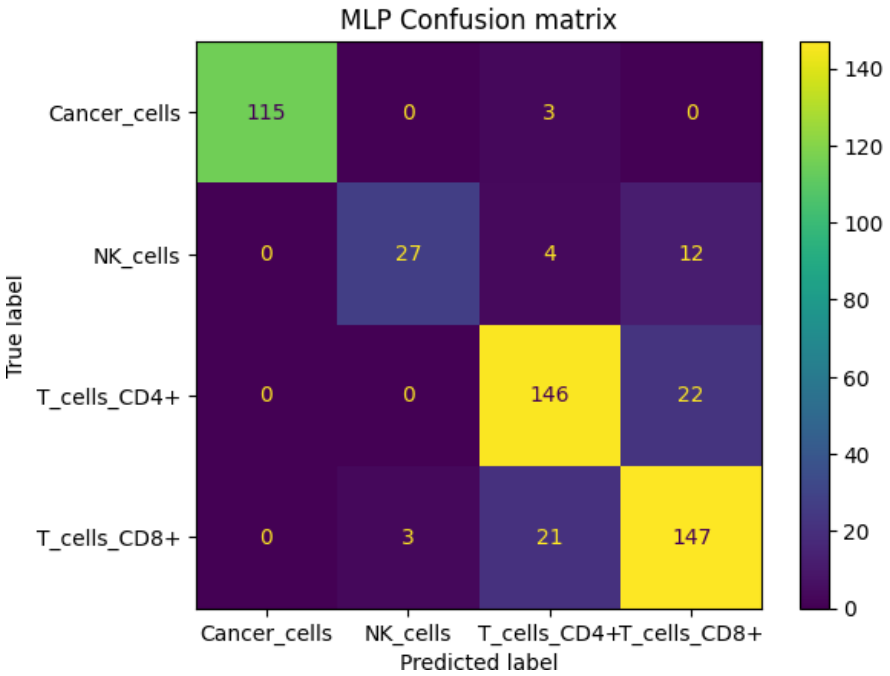


Figure: Matrice de confusion MLP

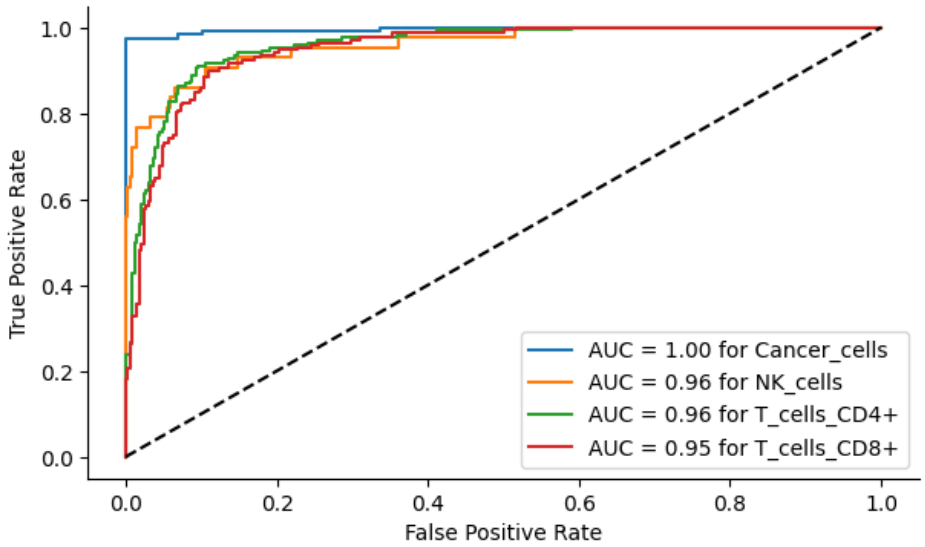


Figure: Courbe ROC MLP

→ Précision pondérée de 0.85

Résultats

Comparaison des modèles

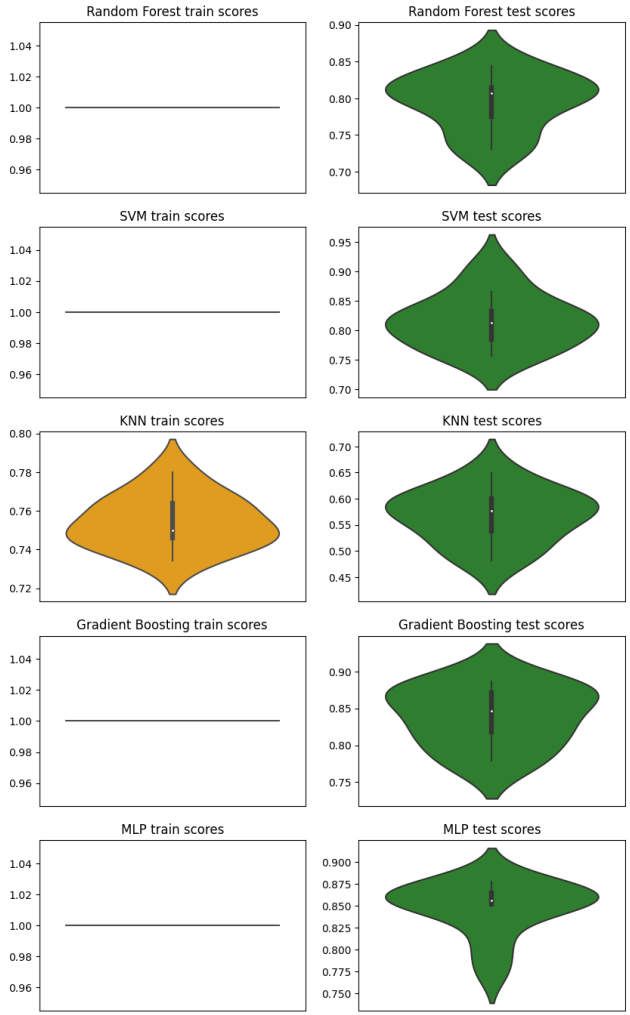


Figure: Validation croisée avec 10 fold stratifiés

Conclusion

- Potentiel du machine learning et ses applications médicales
- Étendre l'étude à un ensemble plus large de types de cellules.
- Perspectives d'amélioration
- Expérience personnelle

université PARIS-SACLAY

Bibliographie

1. Swechha, D. Mendonca, O. Focsa, JJ. Díaz-Mejía, S. Cooper, "scMARK an MNIST like benchmark to evaluate and optimize models for unifying scRNA data", 2021, bioRxiv, doi: https://doi.org/10.1101/2021.12.08.471773
2. R. Hong, Y. Koga, S. Bandyadka, A. Leshchyk, Y. Wang et al. "Comprehensive generation, visualization, and reporting of quality control metrics for single-cell RNA sequencing data", 2022, Nature communication, doi: https://doi.org/10.1038/s41467-022-29212-9
3. L. M. Su, T. Pan, QZ. Chen et al. "Data analysis guidelines for single-cell RNA-seq in biomedical studies and clinical applications", 2022, Military Medical Research, doi: https://doi.org/10.1186/s40779-022-00434-8
4. L. Yu, Y. Cao, JHY. Yang et al. "Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data", 2022, Genome Biolology, doi: https://doi.org/10.1186/s13059-022-02622-0

université PARIS-SACLAY