Single-cell RNA-seq classification

Aline Gabriel
Murielle Majum
Alimatou Traore

Projet Data Camp
M2 Data-Science 2023-2024

June 7, 2024

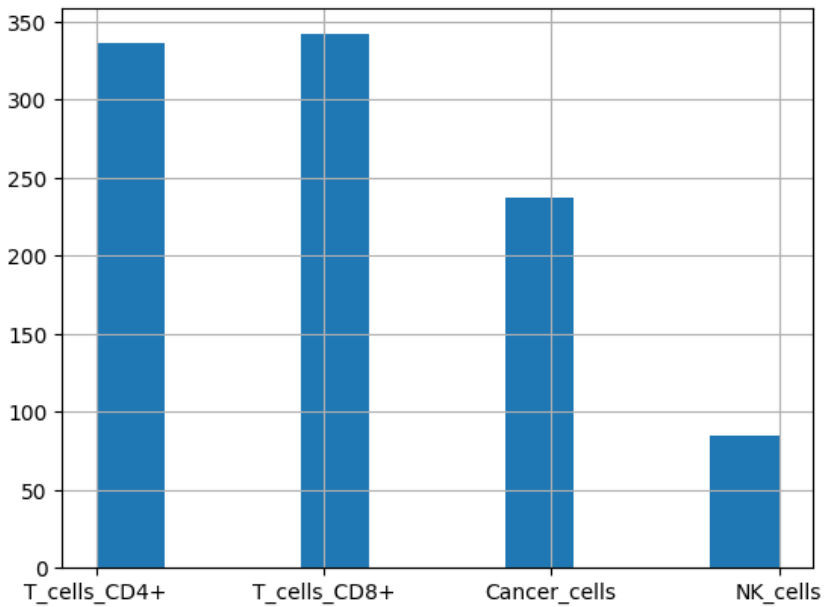
Table of Contents

Introduction

Classification des types cellulaires à partir de données RNA-seq.

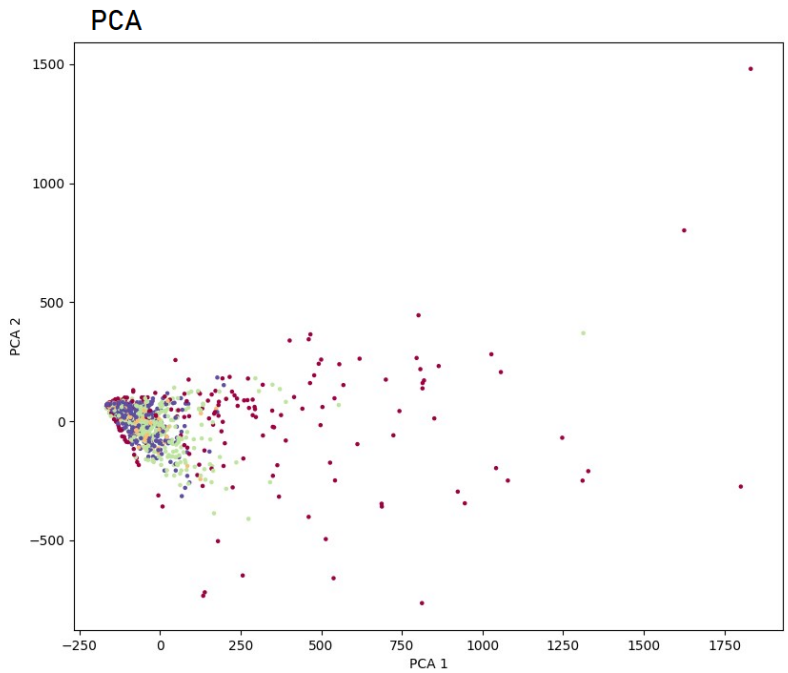
Interets :

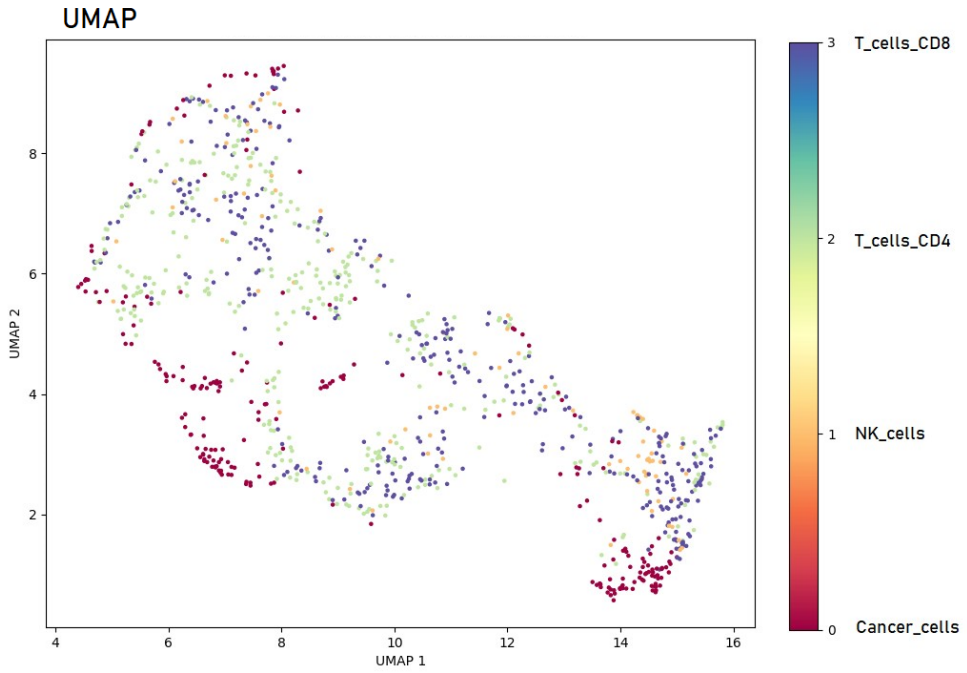
- Comprendre la diversité des cellules et leurs caractéristiques.
- Comprendre les mécanismes biologiques, diagnostiquer les maladies et développer des traitements médicaux ciblés.



Visualisation des données

Visualisation du `X_train` :



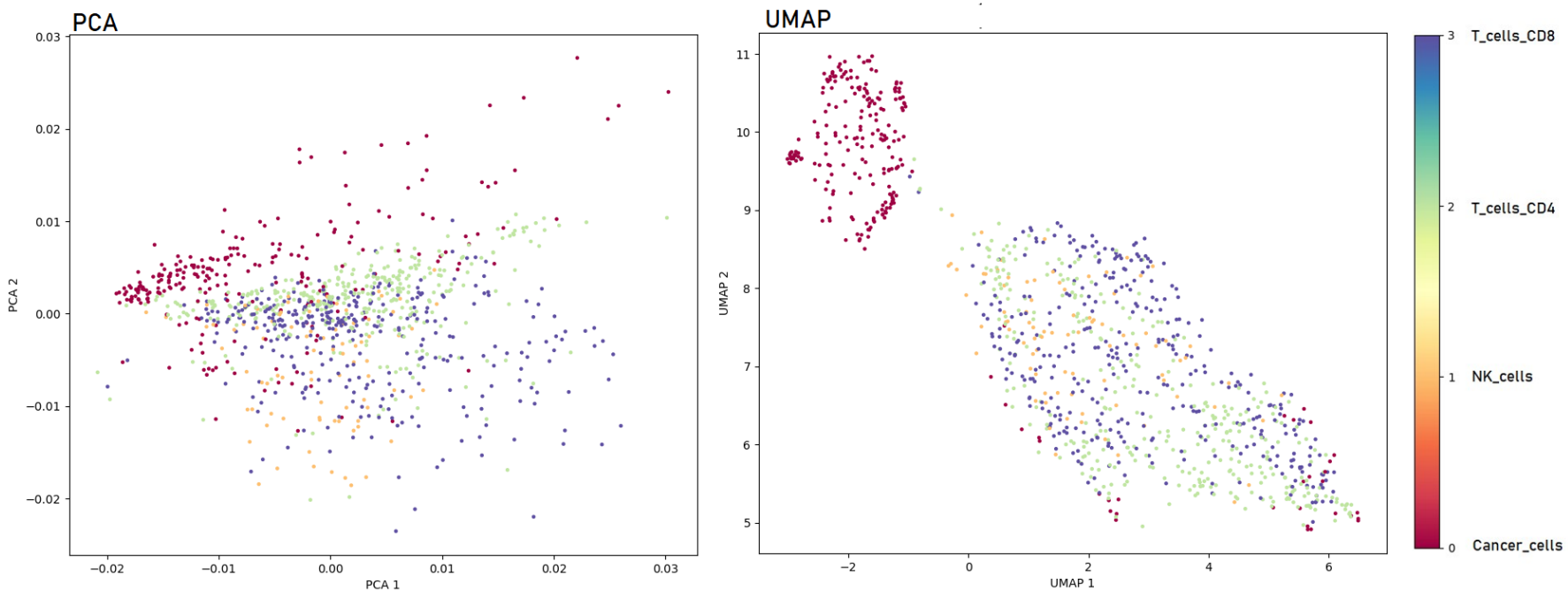


Intérêt du prétraitement des données

Visualisation du `X_train` avec un prétraitement :

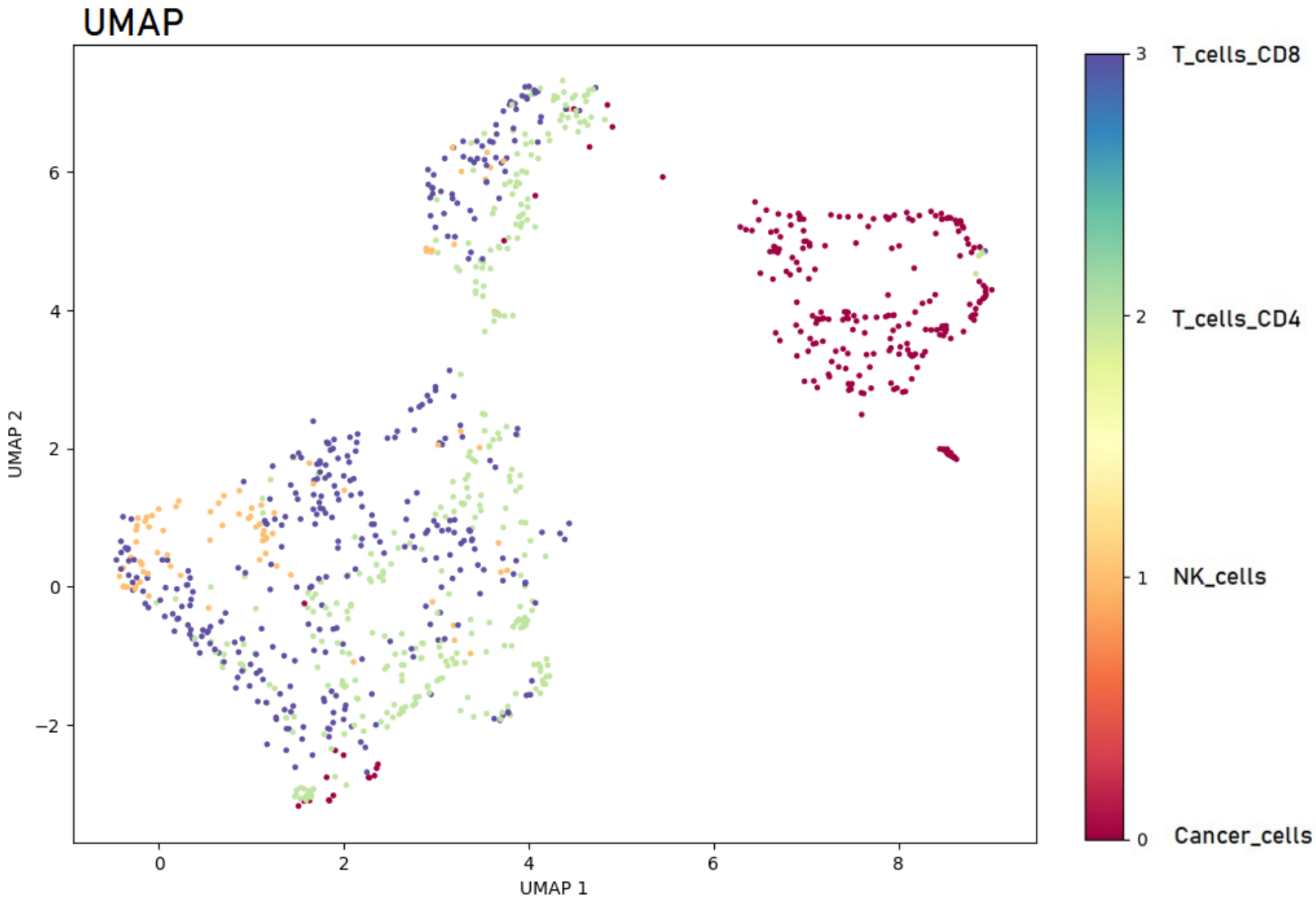
```
# log
X = sc.pp.log1p(X)

# normalize each row
X = X / X.sum(axis=1)[:, np.newaxis]
```



Visualisation du `X_train` avec un prétraitement et PCA :

```
pca = PCA(n_components=100)
X_train_pca = pca.fit_transform(X)
```



Modèle

- Utilisation de modèles de base (KNN, Bagging, AdaBoost, Gradient Boosting, MLP, Random Forest, SVM, ...)
- Mélange de modèles: Stacking (avec MLP en métamodele)

Prétraitement et sélection de variables

Utilisation de plusieurs méthodes pour sélectionner les variables:

- PCA dans un premier temps
- Lasso
- Random Forest

Coeur du modèle

```
base_models = [
            ('rf', make_pipeline(
                SelectFromModel(Lasso(alpha=0.06283860093330873)),
                RandomForestClassifier(random_state=42, n_estimators= 96, max_features=15))),
            ('mlp', make_pipeline(
                SelectFromModel(RandomForestClassifier(n_estimators=74, max_features=13,
                random_state=42)),
                MLPClassifier(hidden_layer_sizes=(256, 256), activation='relu', solver='adam',
                max_iter=865)))
]

meta_model = MLPClassifier(hidden_layer_sizes=(64, 64),
                    activation='relu', solver='adam', max_iter=1187)

self.pipe = make_pipeline(StandardScaler(),
                    StackingClassifier(estimators=base_models, final_estimator=meta_model))
```

Résultats

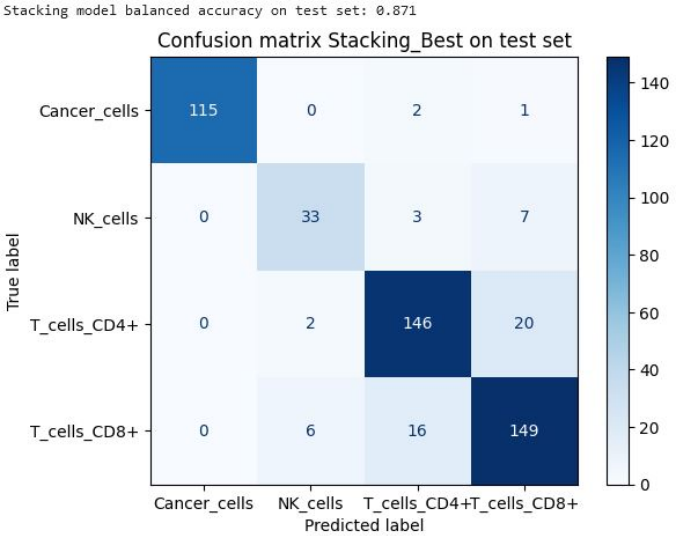


Figure: Best confusion matrix

```
Bagged scores
-------------------------
        score  bal_acc
valid    0.87
test     0.88
```

Figure: Best Score en local

	Train	Test
Accuracy	1.0	0.87
Recall	1.0	0.87
Precision	1.0	0.8766051284384088
F1-score	1.0	0.8759017601139958
Balanced_accuracy	1.0	0.871
Time (after CV)	62.3 +- 2.0	0.5 +- 0.08

Conclusion

Ce projet nous a permis de :

- **Comprendre les Données RNA-seq :** Acquérir une compréhension approfondie des caractéristiques et des nuances des données RNA-seq sur cellules uniques.
- **Comprendre l'Importance du Prétraitement des Données**
- **Faire des Choix en Matière de Modèle de Classification :** Sélectionner judicieusement un modèle de classification adapté à notre type de données.