

CLUSTERING ET ANALYSE DE L'EXPRESSION GÉNIQUE

MODULE: BIO-INFORMATIQUE

Présenter par:

FADLI NOUHAILA
ICHOU NOUHAILA

CHAABAN MALIKA
BOUSSAID SALAMA

PLAN



Problématique



Jeu de données



Prétraitement



Modélisation



Déploiement



Conclusion



PROBLÉMATIQUE

OBJECTIF

Conception d'un modèle Machine Learning permettant le Clustering et l'Analyse des gènes



DONÉES D'EXPRESSION GÉNIQUE



VOLUME IMPORTANT



FORTE COMPLEXITÉ



TRAITEMENT MANUELLE



TRAITEMENT AUTOMATIQUE



MACHINE LEARNING

JEU DE DONNÉES

Source : NCBI (National Center for Biotechnology Information)

Nombre de gènes : 193 860 gènes

Nombre de caractéristiques : 25 features



TAX_ID

GENEID

SYMBOL

CHROMOSOME

MAP_LOCATION

DESCRIPTION

TYPE_OF_GEN
E

NOMENCLATUR
E_STATUS

JEU DE DONNÉES

TYPES DE GÈNES INCLUS

PROTEIN-CODING : GÈNES CODANTS POUR
DES PROTÉINES

NCRNA : ARN NON CODANTS

PSEUDO : PSEUDOGÈNES

RRNA, TRNA : ARN RIBOSOMAux ET DE
TRANSFERT

CARACTÉRISTIQUES ENCODÉES

- **ARM_ENCODED**
- **CHROM_ENCODED**
- **TYPE_BIOLOGICAL-REGION**
- **TYPE_NCRNA**
- **TYPE_PROTEIN-CODING**

PRÉTRAITEMENT

GESTION DES VALEURS
MANQUANTES

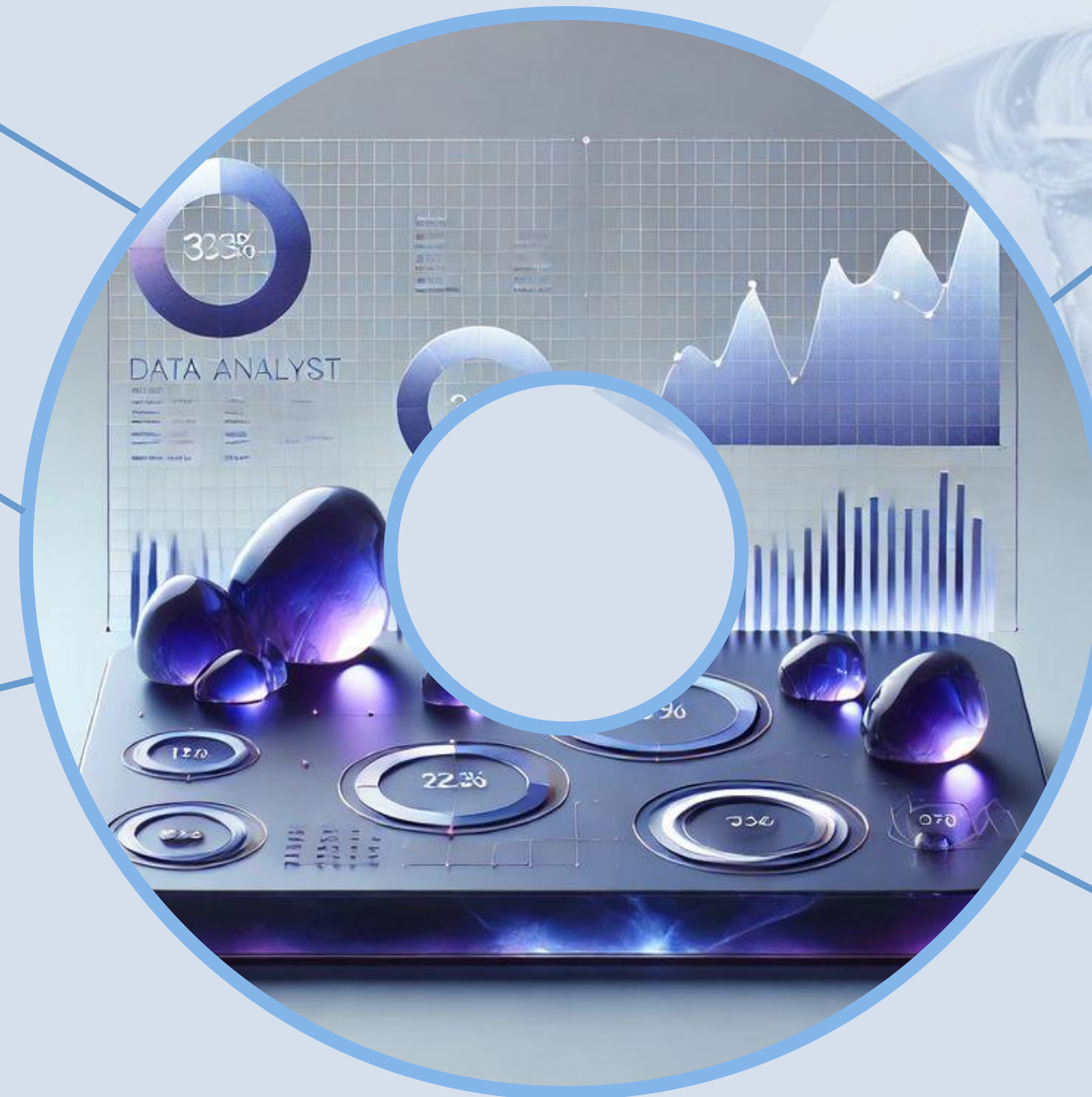
NORMALISATION

ENCODAGE

GÉSTION DES VALEURS
ABÉRENTES

PCA

T-SNE



MODÉLISATION

ALGORITHMES DE CLUSTERING IMPLÉMENTÉS



K

K-Means

PARTITIONNEMENT EN K
CLUSTERS

- MINIMISATION DE LA VARIANCE INTRA-CLUSTER
- ALGORITHME ITÉRATIF ET RAPIDE
- ADAPTÉ AUX GRANDS DATASETS

D

DBSCAN

DENSITY-BASED SPATIAL
CLUSTERING

- DÉTECTION AUTOMATIQUE DU NOMBRE DE CLUSTERS
- IDENTIFICATION DES POINTS ABERRANTS
- BASÉ SUR LA DENSITÉ SPATIALE

G

GMM

GAUSSIAN MIXTURE MODEL

- MODÈLE PROBABILISTE DE CLUSTERING
- CLUSTERS DE FORMES ELLIPTIQUES
- SOFT CLUSTERING (PROBABILITÉS D'APPARTENANCE)

MODÉLISATION

PIPELINE DE CLUSTERING

- Chargement des données CSV (193 860 gènes)
- Encodage des variables catégorielles (LabelEncoder)
- Normalisation des features (StandardScaler)
- Application des algorithmes de clustering
- Évaluation des performances (métriques)
- Sauvegarde des modèles (.pkl)

Bibliothèques & Outils Utilisés

Python : Langage principal

scikit-learn : Algorithmes de Machine Learning

pandas : Manipulation de données

numpy : Calcul numérique

matplotlib : Visualisation

seaborn : Visualisation avancée

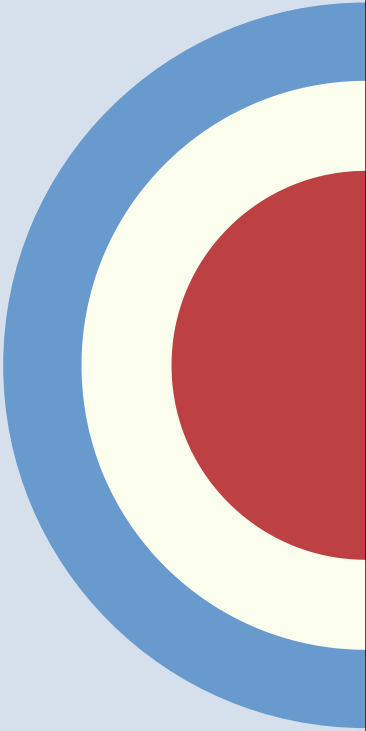
joblib : Sauvegarde des modèles

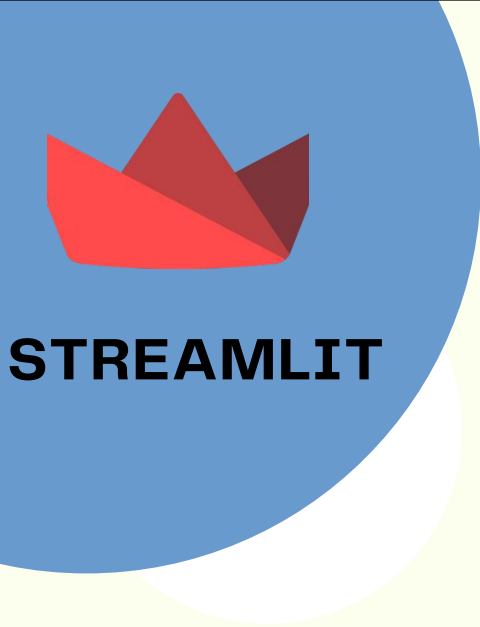
DEPLOIEMENT



STREAMLIT

bibliothèque open-source en Python conçue pour faciliter la création rapide d'applications web interactives, principalement pour la visualisation de données et le déploiement de modèles d'apprentissage automatique





DEPLOIEMENT

INTERFACE DE GUIDE DES CLUSTERS

Mode Debug

Prédiction de Cluster de Gènes

Guide des Clusters

Voici les 9 types de clusters identifiés dans notre modèle. Chaque cluster regroupe des gènes avec des caractéristiques similaires.

Cluster 0: Gènes Codants Standards

Cluster 1: ARN Non-Codants Régulateurs

Cluster 2: Gènes Richement Annotés

Cluster 3: Pseudogènes

Cluster 4: ARN Structuraux (rRNA, tRNA)

Cluster 5: Gènes du Chromosome X

Cluster 6: Régions Biologiques

Cluster 7: Gènes Mitochondriaux

Cluster 8: Gènes Spécialisés Rares

Mode Debug

Guide des Clusters

Voici les 9 types de clusters identifiés dans notre modèle. Chaque cluster regroupe des gènes avec des caractéristiques similaires.

Cluster 0: Gènes Codants Standards

Gènes codants pour des protéines avec descriptions modérées, répartis uniformément sur le génome

Caractéristiques principales:

- Majoritairement des gènes protein-coding
- Descriptions de longueur moyenne (200-400 caractères)

Cluster 1: ARN Non-Codants Régulateurs

ARN non-codants (lncRNA, miRNA) avec descriptions courtes, rôles régulateurs

Caractéristiques principales:

- Principalement lncRNA et miRNA
- Descriptions très courtes (<150 caractères)

Cluster 2: Gènes Richement Annotés

Gènes avec descriptions très détaillées, souvent liés à des maladies

Caractéristiques principales:

- Descriptions très longues (>600 caractères)
- Forte association avec pathologies humaines

Cluster 3: Pseudogènes


Cluster 4: ARN Structuraux

Cluster 5: Gènes du

DEPLOIEMENT


INTERFACE DE SAISIE DES DONNÉES

>>
Deploy
⋮



Caractéristiques du gène à analyser


Entrez les informations du gène pour déterminer à quel cluster il appartient.



Description

Longueur (caractères)

250
-
+




Localisation

Chromosome

19
▼

Bras (p/q)


p
▼




Type de gène

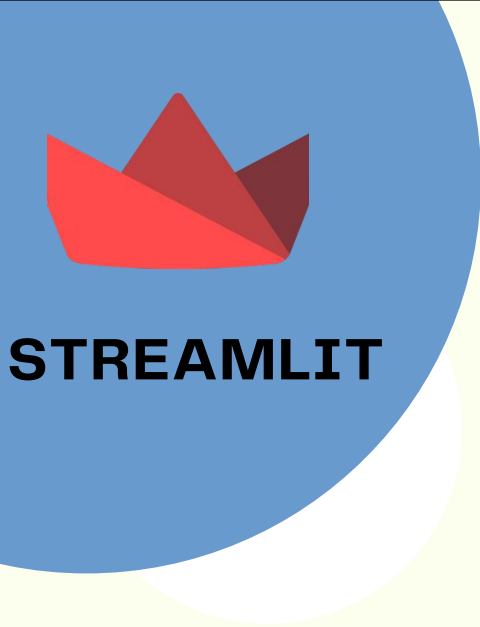
Type

protein-coding
▼


Prédire le cluster


Application de clustering génomique • 70 620 gènes humains • NCBI Gene Database

Clustering basé sur caractéristiques structurelles et fonctionnelles



DEPLOIEMENT

INTERFACE DES RÉSULTATS

>>

Deploy ⋮

✓ Résultat de la prédiction

Cluster 0

Gènes Codants Standards

Gènes codants pour des protéines avec descriptions modérées, répartis uniformément sur le génome

Niveau de confiance

Confiance: 88.9%

>>

Deploy ⋮

📊 Caractéristiques du cluster

- 🔥 Majoritairement des gènes protein-coding
- 📏 Descriptions de longueur moyenne (200-400 caractères)
- 📍 Distribution équilibrée sur tous les chromosomes
- ⚙️ Fonctions cellulaires générales et métabolisme

📖 Exemples de gènes connus

- ACTB - Actine beta (cytosquelette)
- GAPDH - Glycéraldéhyde-3-phosphate déshydrogénase
- TUBB - Tubuline beta

💡 Utilité

En classant votre gène dans le Cluster 0, vous pouvez:

- Explorer des gènes similaires bien caractérisés
- Formuler des hypothèses sur sa fonction
- Identifier des collaborations possibles

🎯 Rôle biologique

Fonctions housekeeping et métabolisme de base

> 📊 Comparaison avec les autres clusters

> 🔧 Détails techniques

CONCLUSION

Le clustering des gènes s'est révélé être un outil puissant pour explorer les relations fonctionnelles dans les données génomiques complexes. La mise en œuvre d'un pipeline d'analyse complet a permis de valider la pertinence de l'approche méthodologique adoptée. Ces travaux ouvrent la voie à des applications concrètes en recherche biomédicale et en médecine personnalisée.



MERCI

