FSBM
FACULTE DES SCIENCES BEN M'SIK
UNIVERSITE HASSAN II DE CASABLANCA

Master d'excellence en Intelligence Artificielle &
Master Big Data et Data Science

Module : Bio-Informatique

Clustering et Analyse de l'Expression Génique

Rédigé par :

FADLI Nouhaila CHAABAN Malika

ICHOU Nohaila BOUSSAID Salama

Encadré par :

Pr. Bnamri Ichrak

Année Universitaire : 2025/2026

Dédicace :

Nous dédions ce projet : À nos chers parents, Qui n’ont jamais cessé de formuler des prières pour nous, de nous soutenir, de croire en nous et de nous encourager à chaque étape de notre parcours. À nos familles, Pour leur amour inconditionnel, leur patience et leur réconfort dans les moments de doute comme dans les moments de joie. À nos amies et amis, Pour leur précieuse aide, leurs conseils et leur soutien moral durant les périodes intenses de travail. À notre binôme, Pour la collaboration, l’écoute, la complicité et l’esprit d’équipe qui ont marqué cette belle aventure. À nos professeurs et à l’administration, Pour leur accompagnement, leur dévouement et la qualité de leur enseignement tout au long de notre formation. À toutes les personnes qui, de près ou de loin, ont contribué à la réussite de ce travail, Nous exprimons notre profonde gratitude.

Remerciements :

Nous remercions tout d'abord Dieu le tout puissant de nous avoir donné la volonté et la persévérance pour réaliser ce travail. Nous tenons à saisir cette occasion et adresser nos profonds remerciements et nos profondes reconnaissances à Pr. Bnamri Ichrak, notre encadrant de mémoire, pour ses précieux conseils et son orientation ficelée tout au long de notre recherche. Nous adressons nos remerciements les plus sincères à toutes les personnes qui nous ont soutenus et qui ont contribué de près ou de loin à l'élaboration de ce mémoire ainsi qu'à la réussite de cette belle année universitaire. Nous remercions également l'ensemble de nos enseignants pour leur dévouement, leur patience et leur contribution à notre formation. Enfin, nous exprimons toute notre reconnaissance à nos proches et amis, pour leur soutien constant et leurs encouragements tout au long de ce parcours. Merci à tous et à toutes.

Résumé :

Dans le cadre de ce projet, nous faisons la conception d'un modèle Machine Learning qui permet le clustering et l'analyse des expressions géniques, c'est-à-dire le clustering des gènes selon leur fonctionnement génétique. En effet le clustering est l'une des techniques d'apprentissage non superviser qui permet au modèle d'apprendre d'après les données en découvrant des patterns ou des structures récurrentes dans les données.

Cela nous amène à avoir un modèle Machine Learning qui prend en entrée un jeu de données contenant l'expression génétique humain, pour donner par la suite en sortie ces données représentées sous forme de groupes ou clusters, où chaque cluster va contenir les gènes similaires selon leur fonctionnement dans le corps humain.

L'objectif de ce projet n'est pas uniquement d'organiser les données sous forme de clusters, mais aussi de rendre un jeu de données interprétable, par conséquent permet à ceux qui l'utilisent de prendre des décisions en fonction des résultats obtenus.

Abstract:

As part of this project, we design a Machine Learning model that enables the clustering and analysis of gene expression data, that is, grouping genes according to their genetic functions. Indeed, clustering is one of the unsupervised learning techniques that allows the model to learn directly from data by discovering patterns or recurrent structures within it.

This leads to the development of a Machine Learning model that takes as input a dataset containing human gene expression data and produces as output a representation of these data in the form of groups or clusters, where each cluster contains genes that are similar in terms of their function within the human body.

The objective of this project is not only to organize the data into clusters, but also to make the dataset interpretable, thereby enabling users to make decisions based on the obtained results.

Table des matières :

Table des figures :

Liste des tableaux :

Introduction générale :

1. Contexte du projet :

Dans le cadre du projet de fin de module de la Bio-Informatique nous faisons la conception d'un modèle du Machine Learning qui permet le clustering et l'analyse des expressions géniques. Ce projet se situe dans le domaine de l'Intelligence Artificielle (IA) ou plus spécifiquement Machine Learning pour résoudre un problème du domaine de la biologie.

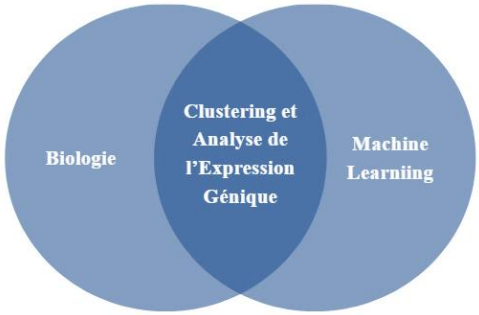

Figure 1: Représentation du contexte du projet

Le Machine Learning est un sous-domaine de l'Intelligence Artificielle, qui permet aux systèmes informatiques d'apprendre d'après leur expérience avec les données, sans être explicitement programmés.

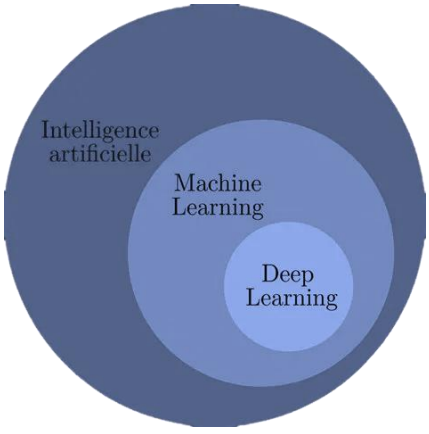

Figure 2: L'Intelligence Artificielle et ces sous-ensembles

Il y a trois types de Machine Learning : apprentissage supervisé, apprentissage non supervisé et apprentissage par renforcement, ces types sont utilisés selon leurs objectifs d'entraînement. Et puisque notre objectif c'est de faire le clustering, donc le type d'apprentissage qui nous concerne c'est l'apprentissage non supervisé.

L'apprentissage non supervisé, contrairement à l'apprentissage supervisé, est un type d'apprentissage automatique où le système s'entraîne sur des données sans avoir le résultat attendu, le système doit lui-même comprendre les données et chercher les relations et les structures récurrentes dans les données. Parmi les types de ce type d'apprentissage nous avons le clustering, ce type vise à regrouper les données sous forme de groupes ou clusters selon la similarité des caractéristiques.

La conception de ce modèle va nous permettre de visualiser tous les comportements des gènes qui existent dans le jeu de données où nous pouvons distinguer ceux qui fonctionnent de manière normale et ceux qui ont un comportement intéressant, ce qui nous amène à faire d'autres recherches et d'autres tests pour comprendre la raison derrière un tel comportement.

FSBM
FACULTE DES SCIENCES BEN M'SIK
UNIVERSITE HASSAN II DE CASABLANCA

Figure 3: Le logo de la Faculté des Sciences Ben M'sick

2. Organisme d'accueil :

La Faculté des Sciences Ben M'Sick (FSBM), composante de l'Université Hassan II de Casablanca, se veut un espace d'excellence scientifique et d'ouverture. Chaque année, elle accueille des milliers d'étudiants issus de toutes les régions du Royaume, ainsi qu'un nombre croissant d'étudiants internationaux, preuve de son rayonnement et de son attractivité. La mission de la Faculté des Sciences Ben M'Sick s'articule autour d'un engagement fondamental : contribuer pleinement à la production, au partage et à la valorisation du savoir, tout en veillant à répondre aux besoins actuels et émergents de la société.
Pour concrétiser cette ambition, la FSBM s'attache à :

• Assurer une formation de haut niveau, innovante et adaptée aux standards internationaux.
• Stimuler une recherche scientifique créative, éthique et porteuse de solutions face aux grands défis nationaux et globaux.
• Consolider une gouvernance moderne, transparente et participative, garante d'une gestion efficace et durable de ses ressources.

3. Problématiques / besoins :

Le problème actuel réside au niveau des données, en effet les données de l'expression génique sont caractérisées par leur volume important et leur forte complexité. Dans ce cas l'analyse manuelle ou traditionnelle sera très difficile.

Cette situation engendre le manque d'automatisation au niveau de l'analyse de données ainsi qu'une difficulté à détecter des structures cachées. Il existe donc un besoin au niveau du regroupement automatique des gènes, la chose qui va affecter la visualisation leur comportements d'où une décision perturbée puisqu'elle a été basée sur les résultats obtenue.

4. Objectifs :

Notre objectif principal est de créer un système qui est capable de faire le clustering des gènes selon leur fonctionnement biologique.

Pour atteindre cet objectif, nous devons réaliser plusieurs objectifs spécifiques :

- **Préparation des données, ce qui inclut le nettoyage et le prétraitement,**
- **Modélisation, incluant l'entraînement, la validation et le test du modèle Machine Learning**
- **Déploiement, ce qui signifie la création d'une interface pour rendre le modèle utilisable**

5. Résultats attendu et impact du projet :

Dans le cadre de ce projet de clustering et analyse des gènes, les résultats attendus sont, dans un premier temps, la mise en place d'un modèle Machine Learning fonctionnel qui fait le clustering avec un taux d'erreur minimal. Ainsi qu'avoir une représentation et visualisation claire des groupes ou clusters des gènes, pour faciliter l'analyse et l'interprétation des résultats.

Ces résultats vont nous aider au niveau de prise de décision. Et avec un système automatisé nous pouvons gagner du temps pour faire d'autres recherches et analyses.

6. Structure du rapport :

Ce rapport est structuré comme suit :

Le premier chapitre représente le cahier de charges et déroulement du projet, incluant les problématiques, besoins et objectifs, ainsi que le déroulement du projet.

Le deuxième chapitre représente l'étude bibliographique, incluant des Définissions des standards et des concepts de base, Documentation nécessaire et les projets similaires ou en relation à l'objectif.

Le troisième chapitre représente le vu conceptuel :

- Architecture du système ;
- Composants du système ;
- Diagrammes UML ;
- Choix des modèles AI et justification des choix ;
- Choix des jeux de données d'entrainements ;
- Architecture réseau ;
- Modèles BI.

Le Quatrième chapitre représente l'implémentation des résultats :

- Environnement software et hardware ;
- Résultats ;
- Interprétation des résultats ;
- Discussion et critiques.

Chapitre 1 : Cahier de charges et déroulement du projet :

1. Introduction :

Ce premier chapitre pose les fondations de notre projet en présentant son contexte général ainsi que les enjeux qui y sont liés. Nous expliquons pourquoi l'analyse des données d'expression génique est devenue un domaine incontournable en bio-informatique, en soulignant les défis posés par la complexité et la quantité importante de ces données. Ensuite, nous présentons l'environnement dans lequel ce travail s'inscrit, en détaillant l'organisme d'accueil et la nature académique du projet. Nous exposons également les principales problématiques auxquelles nous faisons face, ainsi que les objectifs que nous nous sommes fixés pour y répondre. Enfin, ce chapitre décrit l'organisation pratique du projet, en précisant les étapes de son déroulement, la répartition des rôles entre les membres de l'équipe et la méthode de gestion adoptée. Cette introduction globale vise à donner une vision claire et structurée du projet, afin de mieux comprendre la suite du travail.

2. Cahier de charges :

Problématiques / Besoins :

La bio-informatique, c'est un domaine passionnant mais super complexe, surtout quand on parle d'analyser des masses de données d'expression génique venant de trucs comme les microarrays ou le séquençage single-cell. Le but du clustering des gènes (ou classification), c'est de regrouper ceux qui ont des profils d'expression similaires pour en déduire des fonctions biologiques communes, genre des voies de signalisation ou des réponses à des stress. Mais y'a des obstacles partout :

- Les données sont souvent tordues : non linéaires, avec beaucoup de bruit, et parfois en forme d'arbre, comme dans les expériences sur *Caenorhabditis elegans* (C. elegans). Par exemple, les cellules ont des durées de vie variables d'un embryon à l'autre, ce qui rend les séries temporelles non alignées – imagine des mesures toutes les 1,5 minutes qui ne correspondent pas entre gènes.

- Hétérogénéité des cellules : Dans les données single-cell, chaque cellule est différente, donc un simple clustering ne suffit pas ; il faut du biclustering pour capter des patterns locaux, seulement dans certains sous-groupes de cellules.
- Limites des méthodes classiques : Des algos comme K-means regardent juste les valeurs brutes et zappent les dynamiques de changement (dérivées). Ils flanchent face au bruit, aux dimensions élevées ou aux structures arborescentes.
- Besoins des biologistes : On veut des outils pour décrypter des données d'imagerie 4D confocale, pour mieux comprendre le développement embryonnaire chez C. elegans – cet organisme transparent avec une lignée cellulaire fixe est idéal pour étudier les divisions parent-enfant et les expressions dynamiques.

Ces défis nous poussent à créer un modèle hybride, avec des coefficients de Fourier pour lisser et estimer les changements, plus un clustering statistique pour des résultats fiables et bio-sensés.

Objectifs :

Objectif général :

L'objectif principal de ce projet est de mettre en œuvre et d'évaluer des techniques de clustering appliquées aux données d'expression génique afin d'identifier des groupes de gènes présentant des comportements similaires.

Objectifs spécifiques :

- Comprendre les fondements biologiques de l'expression génique.
- Étudier les principales méthodes de clustering utilisées en bio-informatique.
- Prétraiter des données d'expression génique issues d'organismes biologiques (humain ou souris).
- Appliquer des algorithmes de clustering pour regrouper les gènes.
- Analyser et interpréter les résultats obtenus d'un point de vue biologique.

3. Déroulement du projet :

Le projet s'est déroulé selon une succession d'étapes méthodiques et complémentaires, permettant une progression cohérente vers les objectifs fixés. Dans un premier temps, une étude théorique approfondie a été menée afin de comprendre les concepts biologiques fondamentaux liés à l'expression génique, ainsi que les différentes méthodes de clustering utilisées en bio-

informatique. Cette phase a permis d'établir une base solide de connaissances nécessaires pour aborder l'analyse des données.

Ensuite, une recherche bibliographique ciblée a été réalisée afin d'identifier et d'examiner les travaux scientifiques existants relatifs au clustering des gènes. Cette revue critique a contribué à situer le projet dans le contexte actuel de la recherche et à sélectionner les approches les plus adaptées pour la suite des travaux.

La troisième étape a concerné la sélection et la préparation des jeux de données d'expression génique. Cette phase essentielle a consisté à collecter des données pertinentes issues d'organismes modèles, à effectuer leur nettoyage et prétraitement, notamment la gestion des valeurs manquantes et le filtrage des données bruitées, afin d'assurer leur qualité pour l'analyse.

Par la suite, les algorithmes de clustering ont été implémentés à l'aide d'outils informatiques spécifiques. Cette phase a impliqué la programmation, la configuration des paramètres des algorithmes et la mise en place des traitements nécessaires pour appliquer ces méthodes sur les données prétraitées.

Une fois les clusters obtenus, une analyse approfondie des résultats a été réalisée, visant à interpréter biologiquement les groupes de gènes identifiés. Cette étape a permis de mettre en lumière les relations fonctionnelles potentielles entre gènes co-exprimés et de vérifier la pertinence biologique des clusters formés.

Enfin, le projet s'est conclu par la rédaction du rapport final, document détaillé regroupant l'ensemble des travaux réalisés, les analyses effectuées, ainsi que les conclusions et perspectives. Cette étape a également inclus la préparation de la présentation destinée à la soutenance.

GANTT :

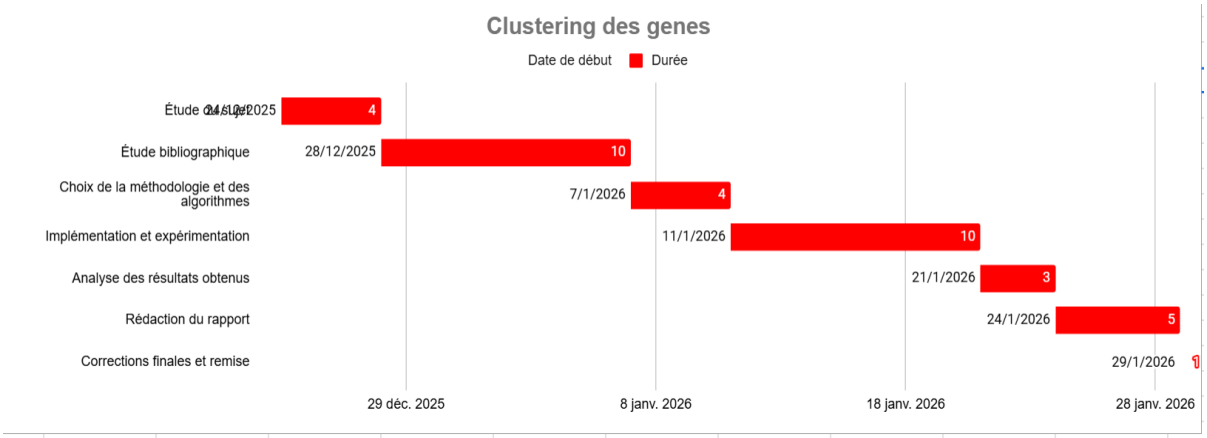

Figure 4: Diagramme de GANTT

Distribution de rôles :

Le projet est réalisé par une équipe composée de quatre étudiantes inscrites en Master Bio-informatique. Afin d'assurer une organisation efficace du travail, les tâches ont été réparties en deux binômes complémentaires, chacun responsable d'un volet précis du projet.

Le premier binôme est chargé des aspects liés à la documentation et à la communication scientifique. Ses missions comprennent la rédaction du rapport académique ainsi que la préparation de la présentation PowerPoint. À l'intérieur de ce binôme, le travail a été réparti de manière équilibrée : une étudiante s'est principalement consacrée à la rédaction des chapitres introductifs et théoriques (chapitres 1 et 2), tandis que l'autre étudiante a pris en charge la rédaction des chapitres suivants (chapitres 3 et 4) et une partie de la présentation orale.

Le second binôme est responsable des aspects techniques et pratiques du projet. Il s'occupe de la recherche et de la préparation des jeux de données biologiques, ainsi que du développement de la plateforme ou de l'application utilisée pour l'analyse des données. De la même manière, les tâches ont été réparties équitablement entre les deux étudiantes, chacune contribuant à la fois à la gestion des données et au développement de la solution informatique.

Une collaboration étroite a été maintenue entre les deux binômes tout au long du projet afin de garantir la cohérence entre la partie théorique, la partie pratique et les résultats obtenus. Des réunions régulières de coordination ont été organisées à chaque étape clé du projet afin de discuter de l'avancement des travaux, de valider les choix méthodologiques et de prendre

collectivement les décisions importantes. Ainsi, bien que les tâches aient été réparties pour des raisons d'organisation, l'ensemble du projet a été réalisé de manière collaborative, avec l'implication active des quatre étudiantes dans chaque phase du travail. L'encadrant académique assure le suivi pédagogique, oriente le travail et valide les différentes étapes du projet.

Méthodologie de gestion :

Pour garantir le bon déroulement du projet et assurer la qualité des résultats, une méthodologie de gestion rigoureuse a été adoptée. Cette méthodologie repose principalement sur une approche **itérative et incrémentale**, qui se caractérise par :

- **Découpage du projet en étapes claires et distinctes** : Chaque phase du projet est définie avec des objectifs précis, ce qui facilite l'organisation du travail et la gestion du temps.

- **Validation progressive** : Après chaque étape, les résultats obtenus sont évalués et validés avant de passer à la suivante. Cela permet de s'assurer que le projet avance dans la bonne direction et de corriger rapidement les erreurs ou problèmes détectés.

- **Flexibilité et adaptation** : En cas de difficulté ou de résultat insatisfaisant, il est possible de revenir sur une étape précédente pour revoir les choix méthodologiques ou techniques. Cette flexibilité évite d'accumuler des erreurs jusqu'à la fin du projet.

- **Collaboration et communication régulières** : Les membres de l'équipe se réunissent régulièrement pour faire le point sur l'avancement, partager les difficultés rencontrées et ajuster les plans si nécessaire. Ces échanges favorisent la cohérence et l'efficacité du travail collectif.

En résumé, cette méthodologie permet de maîtriser la complexité du projet en avançant étape par étape, avec un contrôle continu de la qualité des travaux réalisés. Elle offre ainsi un cadre structuré mais souple, favorisant la réussite du projet tout en limitant les risques d'échec ou de retard.

4. Conclusion :

En résumé, ce chapitre a permis de poser les bases essentielles du projet en clarifiant le contexte, les besoins, ainsi que les objectifs poursuivis. La structuration rigoureuse du

déroulement, la planification détaillée et la répartition claire des rôles témoignent de l'organisation méthodique mise en place pour garantir la réussite du projet. La méthodologie de gestion adoptée, fondée sur une approche itérative et collaborative, vise à assurer un suivi constant de l'avancement tout en permettant des ajustements adaptés. Ces éléments constituent un socle solide pour aborder les phases ultérieures du projet avec confiance et rigueur.

Chapitre 2 : Etude bibliographique :

1. Introduction :

Ce deuxième chapitre est consacré à l'étude bibliographique, qui constitue une étape essentielle pour bien comprendre les concepts et les méthodes liés à notre projet. Nous commencerons par définir les notions fondamentales en bio-informatique, notamment celles qui concernent les gènes et l'expression génique, ainsi que les caractéristiques des données biologiques que nous allons analyser. Ensuite, nous présenterons les principales techniques de clustering, qui sont au cœur de notre démarche analytique, en expliquant leurs principes et leurs spécificités. Enfin, nous passerons en revue les travaux et projets similaires déjà réalisés dans ce domaine, afin de situer notre projet dans le contexte scientifique actuel et d'en souligner la pertinence. Cette étude théorique nous permettra ainsi de poser des bases solides pour la suite de notre travail pratique.

2. Définissions des standards et des concepts de base :

Le gène :

Un gène est une séquence spécifique d'ADN qui contient l'information nécessaire à la synthèse d'une molécule fonctionnelle, principalement une protéine. Chaque gène joue un rôle particulier dans le fonctionnement de la cellule. La variation dans l'expression des gènes, c'est-à-dire la quantité de protéines produites, influence directement les caractéristiques biologiques et physiologiques d'un organisme.

L'expression génique :

L'expression génique est le processus par lequel l'information génétique contenue dans un gène est transcrite en ARN puis traduite en protéine. Ce processus est régulé et peut varier selon le type de cellule, les conditions environnementales, ou le stade de développement. Mesurer l'expression génique permet de comprendre comment les gènes s'activent ou se désactivent dans différents contextes biologiques.

Données d'expression génique :

Les données d'expression génique résultent de mesures quantitatives des niveaux d'expression des gènes dans des échantillons biologiques variés. Elles sont souvent organisées sous forme de matrices où chaque ligne correspond à un gène et chaque colonne à un échantillon ou une condition expérimentale. Ces matrices sont généralement très volumineuses et peuvent contenir du bruit ou des valeurs manquantes, ce qui complique leur analyse.

Technologies d'acquisition des données :

Les deux technologies principales pour obtenir ces données sont les microarrays et le RNA-Seq. Les **microarrays** utilisent des sondes fixées sur une puce pour détecter la présence de fragments d'ARN, tandis que le **RNA-Seq** séquence directement les ARN présents dans un échantillon, offrant une meilleure précision et une plus grande capacité à détecter des niveaux faibles d'expression ou des variantes de transcription.

3. Méthodes de clustering utilisées en bio-informatique

Clustering :

Le clustering est une méthode d'apprentissage non supervisé qui permet de regrouper des objets similaires (des gènes) sans connaissance préalable des catégories. L'objectif est de diviser un ensemble de données en groupes (clusters) homogènes, où les éléments à l'intérieur d'un même groupe sont plus proches les uns des autres qu'à ceux d'autres groupes. Cette méthode est particulièrement utile pour analyser les données d'expression génique afin de découvrir des patterns ou modules biologiques.

K-means :

K-means est un algorithme de clustering simple et rapide qui partitionne les données en un nombre fixe (K) de clusters. Il fonctionne en attribuant chaque point au cluster dont le centre est le plus proche, puis en recalculant les centres jusqu'à convergence. Sa simplicité est un avantage, mais il nécessite de spécifier le nombre de clusters au préalable, et les résultats peuvent varier selon l'initialisation des centres, ce qui peut parfois conduire à des regroupements sous-optimaux.

Clustering hiérarchique :

Cette méthode construit une hiérarchie de clusters sous forme d'un dendrogramme qui illustre comment les données sont regroupées à différents niveaux. Il existe deux approches principales : agglomérative (fusion progressive des clusters) et divisive (division progressive). Le clustering hiérarchique est apprécié en bio-informatique car il offre une visualisation intuitive des relations entre gènes et ne nécessite pas de fixer le nombre de clusters au départ.

Bi clustering :

Le biclustering regroupe simultanément des sous-ensembles de gènes et de conditions où ces gènes présentent des comportements similaires. Contrairement au clustering classique qui se base uniquement sur une dimension (gènes ou échantillons), le biclustering permet de détecter des motifs locaux dans les données, ce qui est crucial en biologie puisque l'expression d'un gène peut être similaire à un groupe de gènes seulement dans certaines conditions spécifiques. Cette méthode est donc adaptée pour révéler des modules biologiques condition-dépendants.

4. Modèles, statistiques et enquêtes existantes :

Pour mener à bien la classification des gènes à partir de données d'expression génique, il est essentiel de s'appuyer sur des méthodes statistiques et des modèles d'intelligence artificielle adaptés à la complexité biologique des données.

D'un point de vue statistique, la validation des clusters obtenus passe par des tests rigoureux. L'analyse d'enrichissement fonctionnel, notamment via la base de données Gene Ontology (GO), permet de vérifier si les groupes de gènes identifiés partagent des fonctions biologiques ou interviennent dans des mêmes voies métaboliques. Cette étape garantit que le regroupement ne résulte pas uniquement d'une convergence mathématique, mais traduit une réalité biologique cohérente et pertinente. Ces tests sont donc indispensables pour assurer la fiabilité des clusters issus des algorithmes.

Du côté des modèles d'intelligence artificielle, la nature complexe et à haute dimensionnalité des données d'expression nécessite des approches avancées. Les algorithmes génétiques, qui utilisent des mécanismes d'optimisation inspirés de la sélection naturelle,

permettent par exemple d'explorer efficacement l'espace des solutions pour identifier des clusters optimaux en prenant en compte plusieurs critères, tels que la similarité des profils d'expression et la pénalité pour chevauchement entre groupes. Par ailleurs, les modèles probabilistes, comme les mélanges gaussiens, facilitent la modélisation de la variabilité naturelle des données et améliorent la classification en capturant la distribution sous-jacente des profils d'expression. Des techniques d'apprentissage profond, notamment via des réseaux de neurones, peuvent également être envisagées pour extraire des structures complexes dans les données, surtout lorsque celles-ci sont bruitées ou comportent des patterns dynamiques.

Enfin, une revue des enquêtes existantes souligne l'importance d'intégrer à la fois des méthodes statistiques robustes et des modèles d'intelligence artificielle pour améliorer la qualité et la pertinence biologique des classifications. Cette combinaison méthodologique est essentielle pour répondre aux défis posés par la variabilité et la dimensionnalité des données d'expression génique.

Autres projets similaires en relation avec l'objectif

Étude sur le biclustering appliqué aux données de cellules uniques de *Caenorhabditis elegans* [1]

Cette étude propose un modèle de biclustering innovant pour analyser des données d'expression génique issues de cellules uniques, en particulier des séries temporelles arborescentes chez *C. elegans*. L'originalité du travail réside dans l'intégration de plusieurs critères biologiques et statistiques dans une fonction objectif optimisée par un algorithme génétique. Parmi ces critères figurent la corrélation de Pearson entre gènes, la taille des clusters, ainsi qu'une pénalité pour éviter les chevauchements excessifs entre biclusters. Les résultats montrent que ce modèle permet d'identifier des groupes de gènes cohérents avec les lignées cellulaires et les fonctions biologiques spécifiques, validés par des analyses d'enrichissement génétique. Cette approche illustre l'importance de combiner plusieurs paramètres pour améliorer la précision de la classification dans des contextes biologiques complexes, ce qui est directement applicable à notre projet de classification des gènes à partir de données d'expression.

Cluster Locator : outil d'analyse de la localisation des gènes en grappes [2]

Cluster Locator est un outil en ligne permettant d'analyser la distribution spatiale des gènes sur un génome, en identifiant des grappes selon un seuil d'écart maximal entre gènes adjacents. Ce projet apporte une dimension complémentaire à la classification fonctionnelle des gènes en introduisant une analyse spatiale. L'outil fournit également des tests statistiques rigoureux, comme le test de Kolmogorov-Smirnov, pour évaluer la significativité des grappes observées par rapport à une distribution aléatoire. Cette capacité à associer analyse statistique et visualisation interactive permet d'explorer l'organisation génomique et ses liens potentiels avec la régulation de l'expression génique. Dans le cadre de notre projet, l'intégration de cette perspective spatiale pourrait enrichir la compréhension des relations entre groupes de gènes et leur localisation, renforçant ainsi l'interprétation biologique des clusters obtenus.

Méthode de clustering basée sur les coefficients de Fourier dérivés pour l'analyse des patterns dynamiques d'expression génique [3]

Cette méthode innovante utilise la représentation des profils d'expression génique par séries de Fourier, en se concentrant particulièrement sur les coefficients dérivés pour capturer les changements dynamiques dans les données temporelles. Le modèle de clustering repose sur une distribution normale multivariée appliquée aux coefficients, permettant de regrouper les gènes selon leurs patterns de variation dans le temps. Comparée aux méthodes classiques comme K-means, cette approche offre une meilleure sensibilité pour détecter des groupes biologiquement cohérents, même dans des données bruitées ou avec un nombre limité de points temporels. Les applications sur des données microarray réelles ont montré que les clusters identifiés correspondent à des processus biologiques spécifiques, validés par des analyses d'enrichissement GO. Ce type de méthode est particulièrement pertinent pour notre projet si nous travaillons avec des séries temporelles d'expression, car elle permet de mieux saisir la dynamique des régulations génétiques.

5. Conclusion

Ce chapitre a présenté les méthodes statistiques et d'intelligence artificielle adaptées à la classification des gènes, ainsi que les travaux récents en lien avec notre projet. La revue des études similaires a montré l'importance de combiner plusieurs critères biologiques et statistiques pour obtenir des clusters fiables et pertinents. Ces éléments théoriques et pratiques serviront de base solide pour guider le choix des méthodes et orienter notre travail vers une classification efficace et biologiquement cohérente des gènes.

Chapitre 3 : Vu conceptuel

1. Introduction :

Dans ce chapitre nous allons représenter les premières étapes pour commencer la conception de notre projet de clustering des gènes, incluant l'architecture, et les composants de système et les détails relative au choix des composants principaux. Ainsi que la modélisation du système et son comportement.

2. Architecture du système :

La conception de ce projet a nécessité l'adoption d'une architecture permettant de crée un modèle fonctionnel et de produire des résultats interprétables pour la prise de décision.

Cette architecture repose sur plusieurs étapes, chacune correspondant à un composant essentiel du système :

Le jeu de données :

La plus importante étape dans ce projet c'est d'avoir des données de bonne qualité et qui répond à nos besoins, puisqu'au niveau de Machine Learning, la qualité de données affecte directement le modèle. C'est pour cette raison, avoir un jeu de données qui répond aux besoins n'est pas suffisant, ce qui nous amène à l'étape suivante ; qui est le **nettoyage des données**.

Modélisation et évaluation :

Incluant le choix de l'algorithme de Machine Learning d'après plusieurs tests avec plusieurs algorithmes en se basant sur les résultats de l'évaluation de chaque algorithme.

Déploiement :

Cette étape valorise tous l'effort exécuté dans les étapes précédentes, où elle rendre le modèle utilisable par les utilisateurs ciblé, à travée l'interface de communication avec le modèle et voir les visualisations.

3. Composants du système :

Ce système se compose des composants essentiels suivant :

Jeu de données :

Le jeu de données utilisé dans ce projet se compose de 70620 lignes et 12 colonnes

Colonne	Description
tax_id	Identifiant taxonomique unique de l'espèce (ex. : 9606 pour l'humain, 10090 pour la souris).
GeneID	Identifiant unique et stable attribué à chaque gène dans la base de données NCBI.
Symbol	Symbole officiel du gène (ex. : *A1BG*, *NAT1*). Il s'agit du nom court le plus utilisé par les chercheurs.
chromosome	Numéro ou lettre du chromosome sur lequel le gène est localisé (ex. : 19, X, MT pour mitochondrial).
map_location	Position précise du gène sur le chromosome, exprimée sous forme de bande cytogénétique (ex. : 19q13.43).
description	Description textuelle complète du gène, incluant son nom long ou sa fonction biologique connue.
type_of_gene	Classification du gène (ex. : *protein-coding*, *tRNA*, *pseudo*, *rRNA*). Cette information est utile pour filtrer les données.
Symbol_from_nomenclature_authority	Symbole officiel validé par une autorité de nomenclature (ex. : HGNC pour les gènes humains).
Full_name_from_nomenclature_authority	Nom complet officiel du gène fourni par l'autorité de nomenclature.
Nomenclature_status	Statut de la nomenclature (ex. : **O** pour *Official*).
Other_designations	Autres noms ou synonymes du gène utilisés historiquement ou dans d'autres bases de données, séparés par des barres verticales (`
Modification_date	Date de la dernière mise à jour de la fiche du gène dans la base de données (format AAAAMMJJ).

Tableau 1: colonnes du jeu de données

Nettoyage et prétraitement :

1. Suppression des colonnes non pertinentes :

Afin de réduire la dimensionalité et garder uniquement les informations utiles, certaines colonnes ont été supprimé.

Colonnes supprimées :

- `LocusTag`
- `Synonyms`
- `dbXrefs`
- `Feature_type`

2. Normalisation des noms de colonnes :

Suppression du caractère spécial `#` dans les noms de colonnes afin d'éviter les problèmes de manipulation et assurer la cohérence des noms.

3. Gestion des valeurs manquantes explicites :

Remplacement du caractère `"-"` par des valeurs manquantes (`NaN`), afin de standardiser les données.

4. Suppression des doublons :

- Suppression des lignes dupliquées en se basant sur la colonne `GeneID`.
- Conservation d'une seule occurrence par gène.

5. Analyse des valeurs manquantes :

Calcul du pourcentage de valeurs manquantes par colonne. Afin d'avoir une idée sur le nombre des valeurs manquantes.

6. Suppression des lignes incomplètes critiques :

Suppression des lignes où la colonne `map_location` est manquante.

7. Suppression des colonnes à fort taux de valeurs manquantes :

- `Symbol_from_nomenclature_authority`
- `Full_name_from_nomenclature_authority`
- `Nomenclature_status`
- `Other_designations`

8. Nettoyage et transformation de données spécifiques :

- Conversion de la colonne `chromosome` en format numérique lorsque cela est possible.
- Gestion des valeurs non convertibles comme valeurs manquantes.

9. Imputation des valeurs manquantes :

Imputation des valeurs manquantes restantes par la **médiane** (après sélection des features numériques).

10.Réduction de dimensionnalité :

t-SNE (*t-Distributed Stochastic Neighbor Embedding*Algorithmes de Machine Learning):

L'algorithme t-SNE est une technique de réduction de dimension pour la visualisation des données.

Il s'agit d'une méthode non linéaire permettant de représenter un ensemble de points d'un espace à grande dimension dans un espace de deux ou trois dimensions. Les données peuvent ensuite être visualisées sous la forme d'un nuage de points. L'algorithme t-SNE tente de trouver une configuration optimale selon un critère de théorie de l'information afin de conserver la proximité entre les points pendant la transformation : deux points qui sont proches dans l'espace d'origine doivent être proches dans l'espace de faible dimension.

EDA (Exploratory Data Analysis):

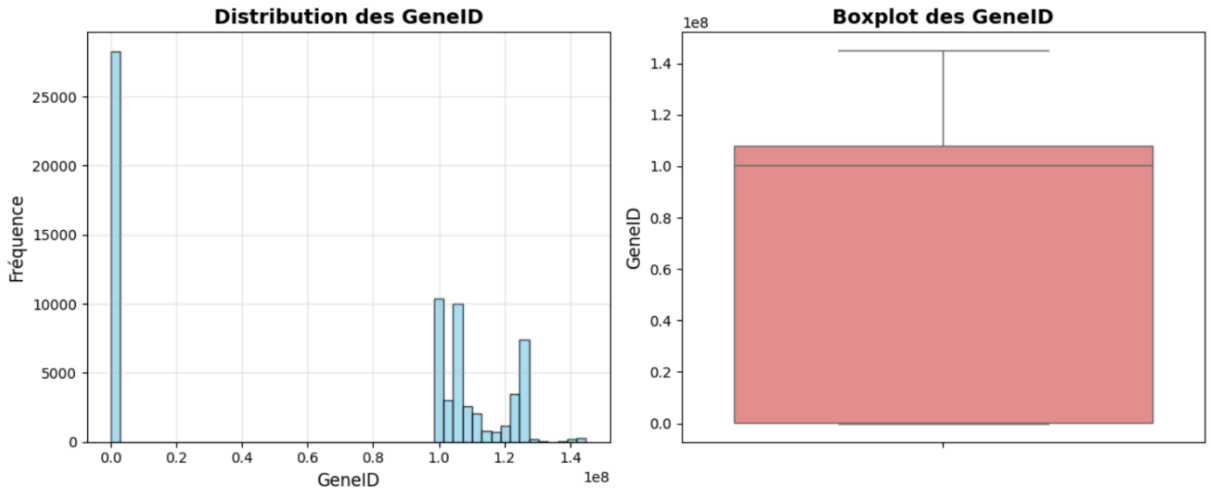

Figure 5: distribution des gènes

La figure représente la distribution des données de la colonne GeneID, la chose qui représente la variété des données, ainsi que le BoxPlot. Ces résultats représentent une distribution qui n'est pas symétrique puisque la moyenne est clairement décalé.

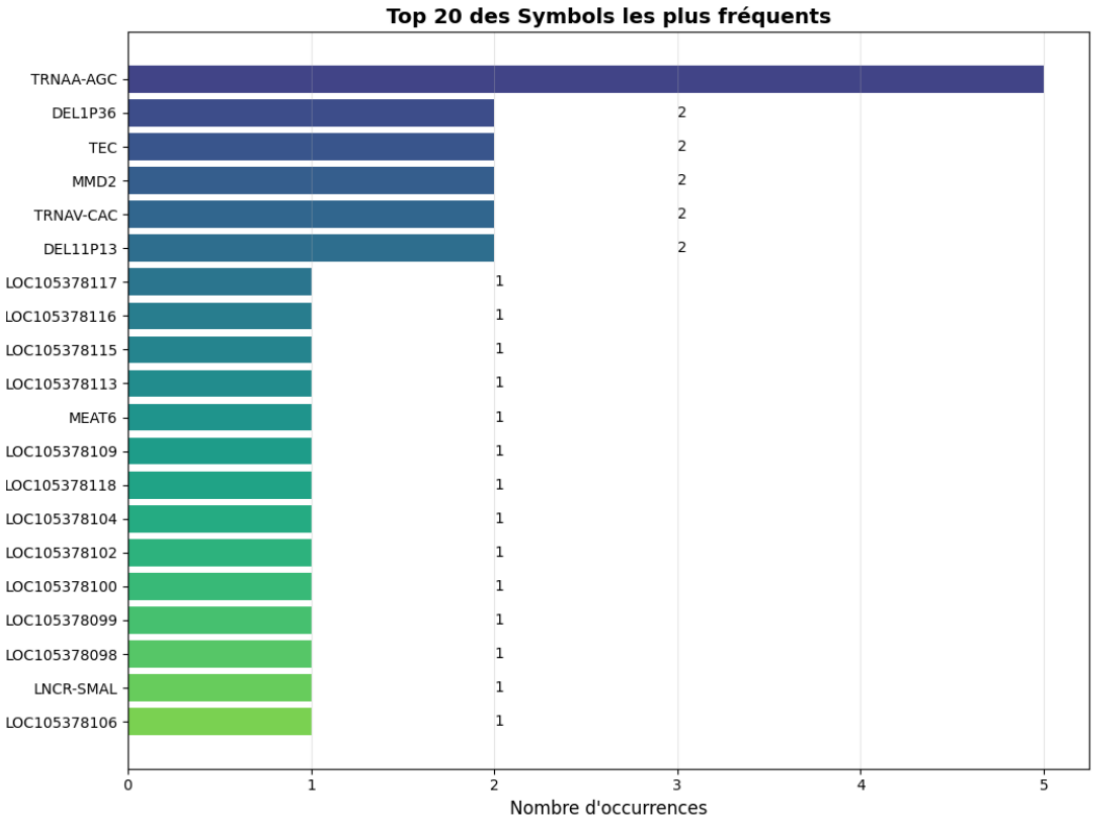

Figure 6; Les symboles les plus fréquents

Cette figure ci-dessus représente la diverssité des données avec des taux important d'occurance

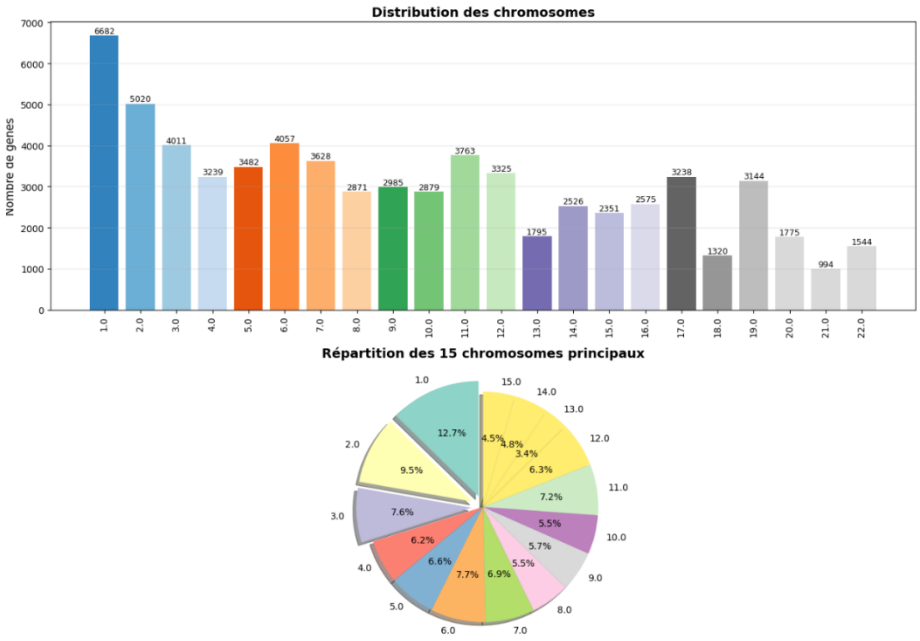

Figure 7: Distribution des chromosomes

La figure ci-dessus représente la distribution des chromosomes selon leurs types, ce qui indique la variété au niveau de données, ainsi que la représentation des chromosomes qui sont représenté de manière importante.

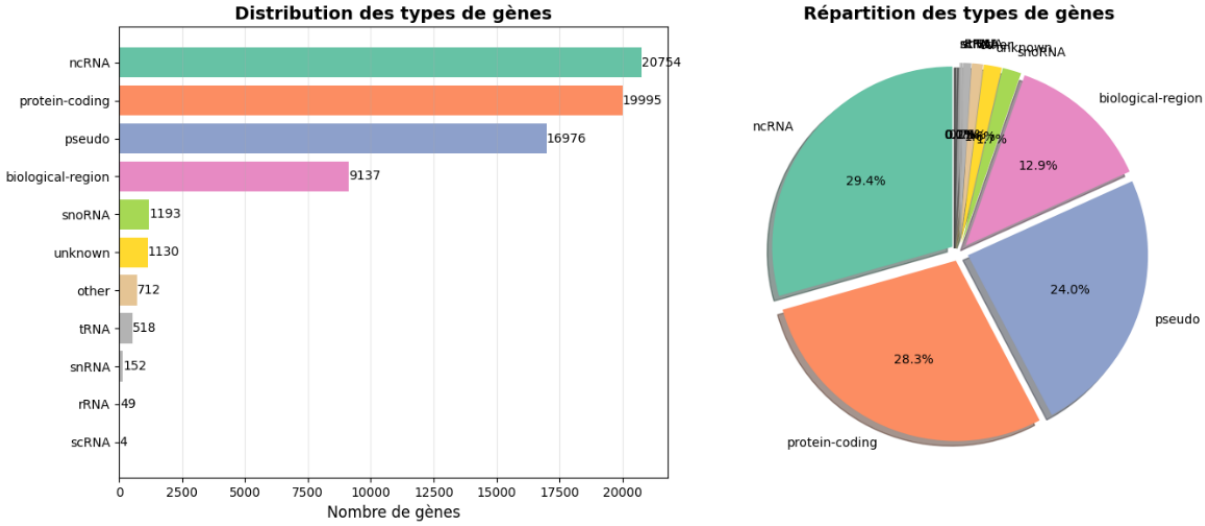

Figure 8: Distribution des types des gènes

La figure ci-dessus représente la distribution des types de gènes en se basant sur le nombre d'occurrence de chaque gènes

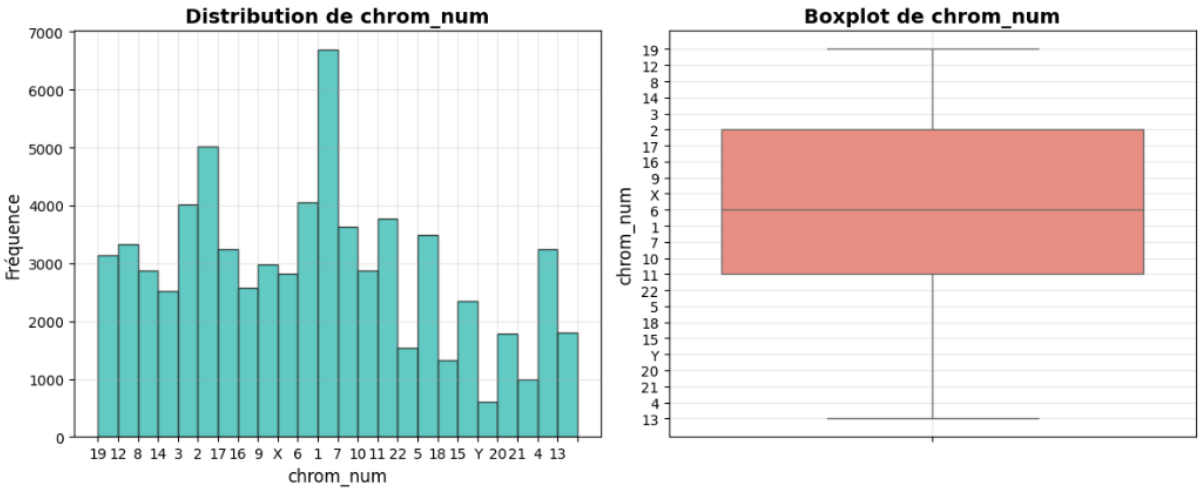

Figure 9: Distribution des chrom_num

Cette représentation, montre une distribution symétrique et équilibrée de la distribution des données de la colonne chom_num.

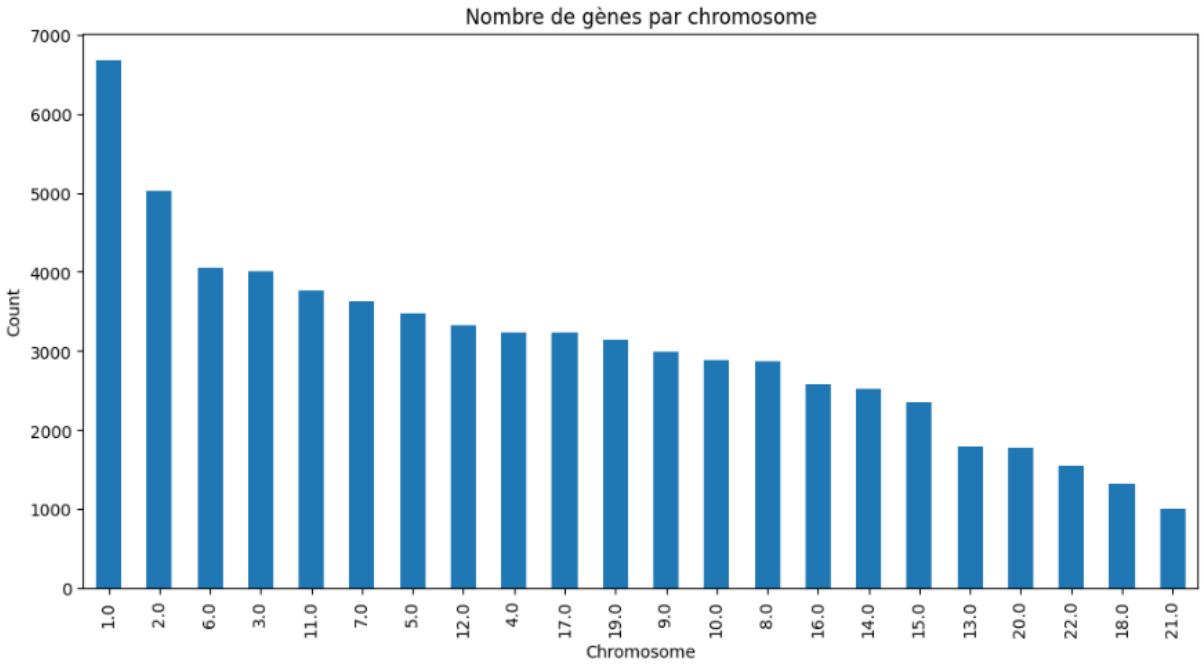

Figure 10: Nombre de gènes par chromosome

Cette figure représente la distribution des gènes par rapport aux chromosomes, ce qui représente une distribution déséquilibré des gènes, la chose qui montre que des chromosomes sont plus favorable pour la présence des gènes que d'autres.

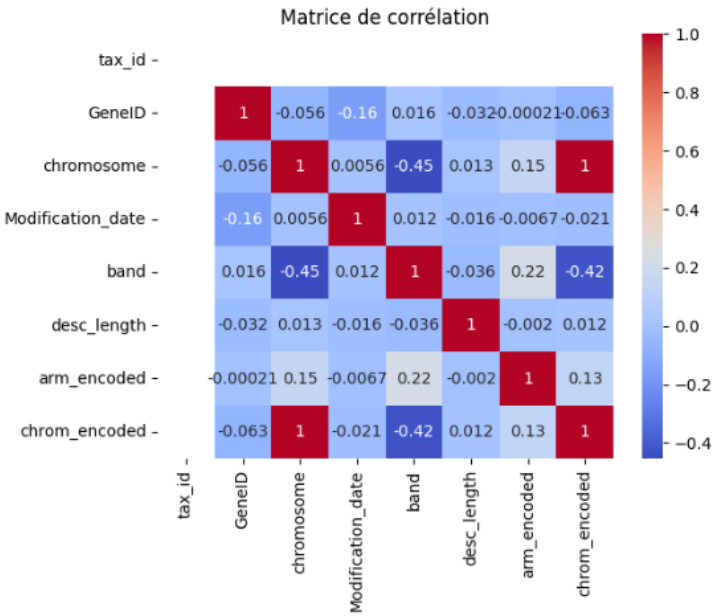

Figure 11: matrice de corrélation

La figure ci-dessus représente la relation des colonnes entre eux, c'est-à-dire la corrélation des données entre eux, plus la valeur est proche de 1 plus la corrélation est forte, et plus la couleurs est proche du rouge plus la corrélation est forte.

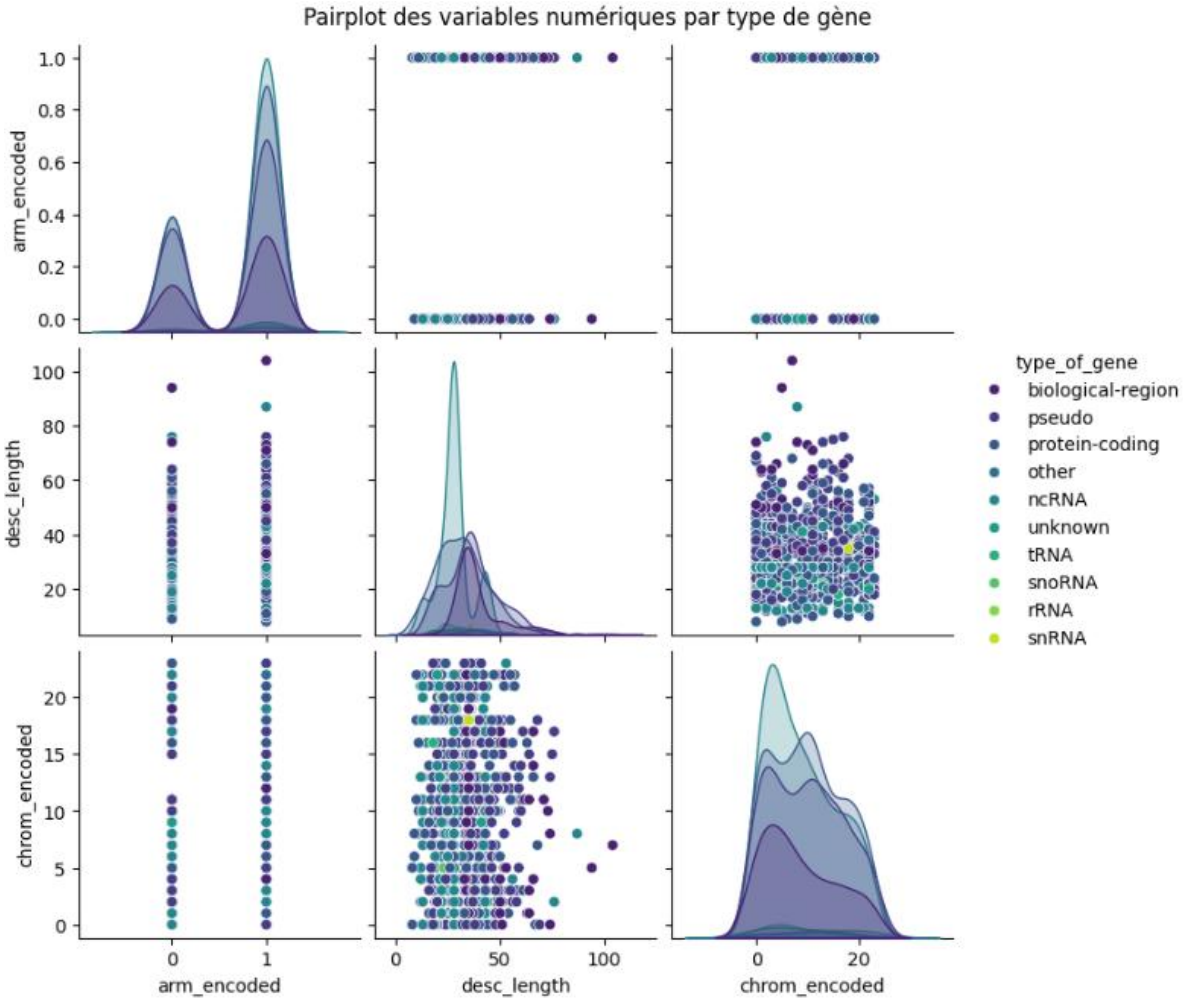

Figure 12: Pairplot des variables numériques par type de gène

La figure ci-dessus représente la distribution des données des colonnes numériques, ce qui montre une forte variabilité et distribution des données ainsi qu'une variabilité au niveau de la représentation.

Algorithmes de Machine Learning :

Algorithme de Kmeans :

La technique de clustering de Kmeans est très simple, son algorithme de base est décrit comme suit :

- Choisir k comme nombre de cluster

- Spécifier les centroides de chaque cluster.
- Calcule de distance par rapport aux centroides.
- Grouper les objets en se basant sur la distance minimale.
- Et pour détérminer le nombres k, nous utilisons la Gap Statistic.

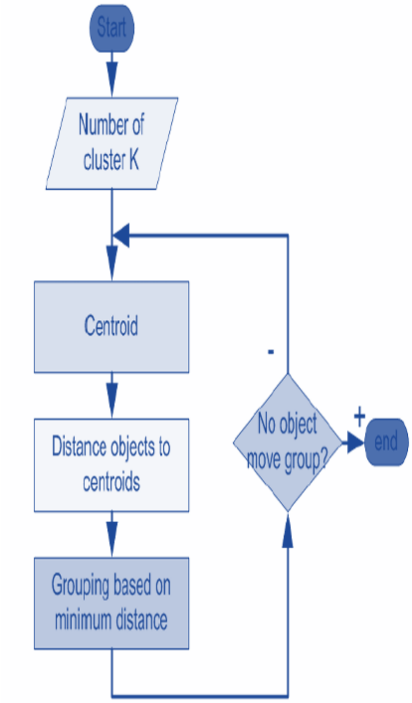

Figure 13: Algorithme Kmeans

Algorithme de DBScan :

DBSCAN permet l'identification des clusters de formes arbitraires et le bruit dans une base de données spatiale. Cet algorithme requiert seulement deux paramètres d'entrée afin que l'utilisateur puisse spécifier une valeur appropriée. .On fixe Eps, le rayon du voisinage à étudier et MinPts, le nombre minimum de points qui doivent être contenus dans le voisinage pour considérer la zone comme dense.

L'idée clé du clustering basé sur la densité est que pour chaque point d'un cluster, son voisinage pour un rayon donné Eps doit contenir un nombre minimum de points MinPts. Ainsi, le cardinal de son voisinage doit dépasser un certain seuil (considéré comme objet principal).

Algoritme de GaussianMixture :

Un Gaussian Mixture Model est un algorithme de clustering probabiliste qui suppose que les données sont générées par un mélange de plusieurs distributions gaussiennes.

Contrairement à K-Means, un point n'appartient pas strictement à un seul cluster, mais à chaque cluster avec une certaine probabilité.

Evaluation :

Le **Silhouette Score** est une mesure d'évaluation de la qualité d'un clustering. Il indique à quel point les clusters sont bien séparés et cohérents.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

- **a(i)** : distance moyenne entre *i* et les points de son cluster
- **b(i)** : distance moyenne entre *i* et le cluster le plus proche

Déploiement :

Pour faire le déploiement de ce projet nous avons utilisé l'outils streamlit qui est une 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. C'est un outil très utilisé par les ingénieurs en machine learning et les data scientists pour présenter leur travail de manière efficace. En résumé, Streamlit permet de transformer des scripts Python en applications web attrayantes et fonctionnelles.

4. Diagrammes UML :

Notre système de clustering et d'analyse de l'expression génique, est représenté dans cinq cas d'utilisation au niveau du diagramme de classe. Ce diagramme représente comment le système va comporter face à l'utilisateur.

En effet l'utilisateur va charger les données et les envoyer au système, le fait de les envoyer déclenche le nettoyage et le clustering des données puis la visualisation. Par conséquent l'utilisateur est capable de faire ces interprétations.

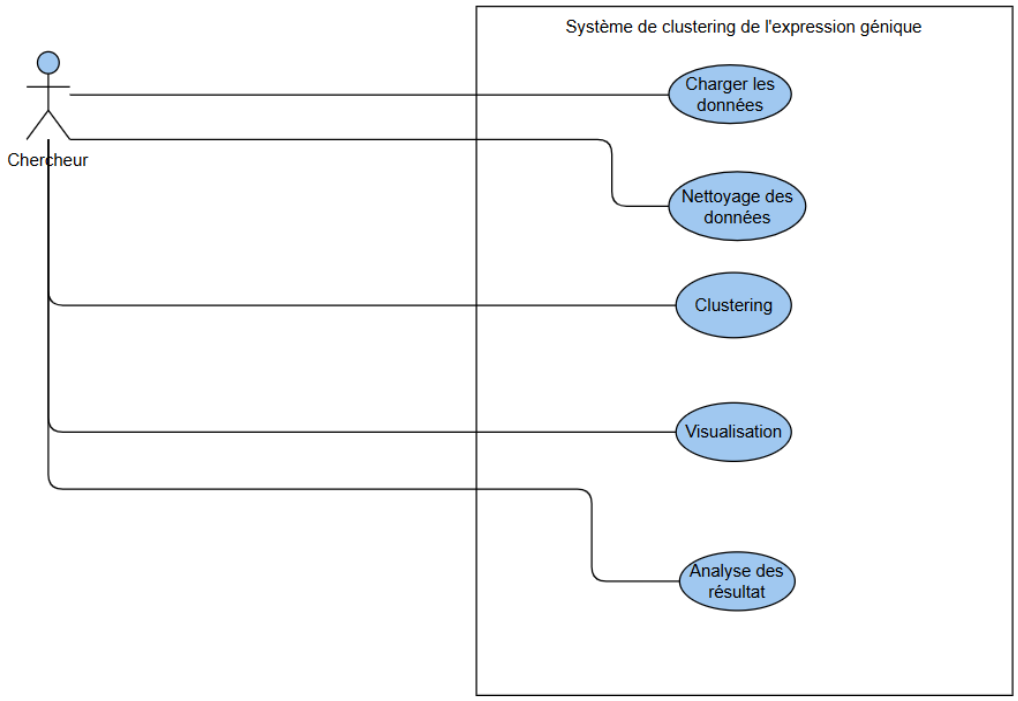

Figure 14: Diagramme de cas d'utilisation

5. Choix des modèles AI et justification des choix :

Nous avons expérimenté plusieurs algorithmes de clustering, notamment DBSCAN, Gaussian Mixture Models (GMM) et **K-means**.

À l'issue de ces expérimentations, l'algorithme K-means a été retenu, en raison de ses bonnes performances en matière de regroupement des données. K-means repose sur la notion de proximité entre les points, ce qui permet de former des clusters cohérents et bien séparés.

De plus, sa flexibilité constitue un avantage important, notamment grâce à la possibilité de déterminer le nombre optimal de clusters à l'aide de méthodes d'évaluation telles que la Gap Statistic.

6. Choix des datasets d'entrainements :

Le jeu de données a été choisi en se basant sur ces caractéristiques, car elles fournissent des informations pertinentes pour la phase d'entraînement du modèle. De plus, ces données sont fiables, puisqu'elles proviennent de la base de données NCBI, une source de référence reconnue dans le domaine de la bioinformatique.

7. Architecture réseau :

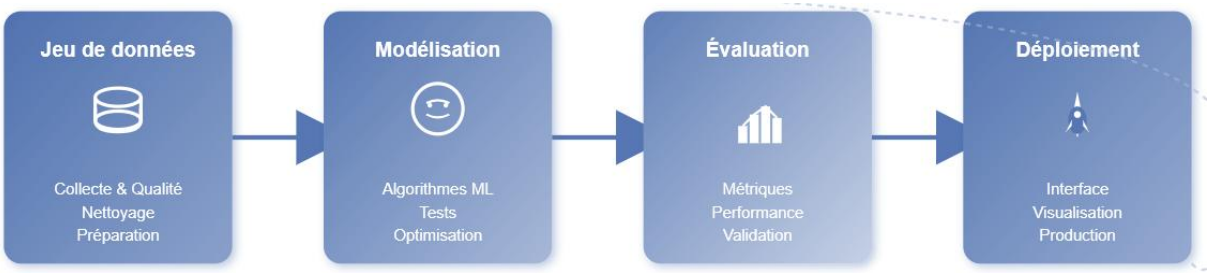

Figure 15: Architecture de système

L'architecture de notre système avec la relation entre eux, en effet les données sont prises d'après NCBI, un jeu de données qui contient des caractéristiques important pour l'entrainement du modèle pour rendre notre projet réaliste, puis le nettoyage et l'analyse des données pour valoriser les données, où nous avons utilisé plusieurs technique et visualisations, ainsi que la modélisation pour atteindre l'objectifs principale de ce projet, et pour rendre le projet et le modèle utilisable et compréhensible nous avons fait la conception de l'interface utilisateur avec Streamlit.

8. Conclusion :

Dans le cadre de ce chapitre nous avons discuté et expliquer les composants de notre projet de clustering des gènes, avec l'architecture qui illustre la relation entre tous les composants. En effet chaque composant représente un rôle important et indispensable pour réaliser ce projet, depuis les données vers le déploiement. Tout cela afin d'avoir un système fonctionnel qui répond aux besoins et aux objectifs.

Chapitre 4 : Implémentation et résultats :

1. Introduction :

L'évaluation complète de notre système de synthèse de clustering et analyse de l'expression génique. L'évaluation s'articule autour de l'analyse des performances obtenues et de la validation de l'efficacité de notre approche développée. L'objectif de cette évaluation est de mesurer la performance du système dans le contexte spécifique, d'analyser les résultats obtenus, et d'identifier les points forts ainsi que les axes d'amélioration pour optimiser le fonctionnement du système.

2. Environnement software et hardware :

Dans le cadre de ce projet de clustering des gènes nous avons utilisé plusieurs outils dans différents environnements :

Environnement software :

Google colab : Colab (ou "Colaboratory") permet a l'utilisateur d'écrire et d'exécuter du code Python dans votre navigateur.

Figure 16: VSCode

Python : durant ce projet le langage utiliser pour développer est Python grâce à son syntaxe simple et au bibliothèques disponibles.

Figure 17: Python

Pandas : Pandas est une bibliothèque Python utilisé pour manipuler et analyser des données tabulaires sous forme de DataFrame (lignes et colonnes) ou Series (colonne unique).

Figure 18: Pandas

Scikit-Learn : est une bibliothèque open-source pour l'apprentissage automatique en Python. Elle est construite sur les bibliothèques NumPy, SciPy et Matplotlib, qui sont des outils populaires pour le calcul numérique et le calcul scientifique en Python.

Figure 19: Sikit-Learn

Environnement hardware :

- **Processeur :** Intel® Core™ i5-8365U CPU @ 1.60 GHz (1.90 GHz)
- **Mémoire RAM installée :** 8,00 Go (7,69 Go utilisable)

- **Carte graphique :** Intel® UHD Graphics 620 (128 MB)
- **Stockage:** 238 GB SSD SAMSUNG MZVLB256HBHQ-000L7
- **Type du système :** Système d'exploitation Windows 64 bits, processeur x64

3. Résultats :

Après l'entrainement du modèle sur les données d'entrainement de l'expression génique, nous avons obtenu les résultats suivants :

- Un modèle fonctionnel permettant le clustering des gènes ;
- Visualisation des clusters des gènes ;
- Capacité d'interpréter les résultats et de prendre des décisions.

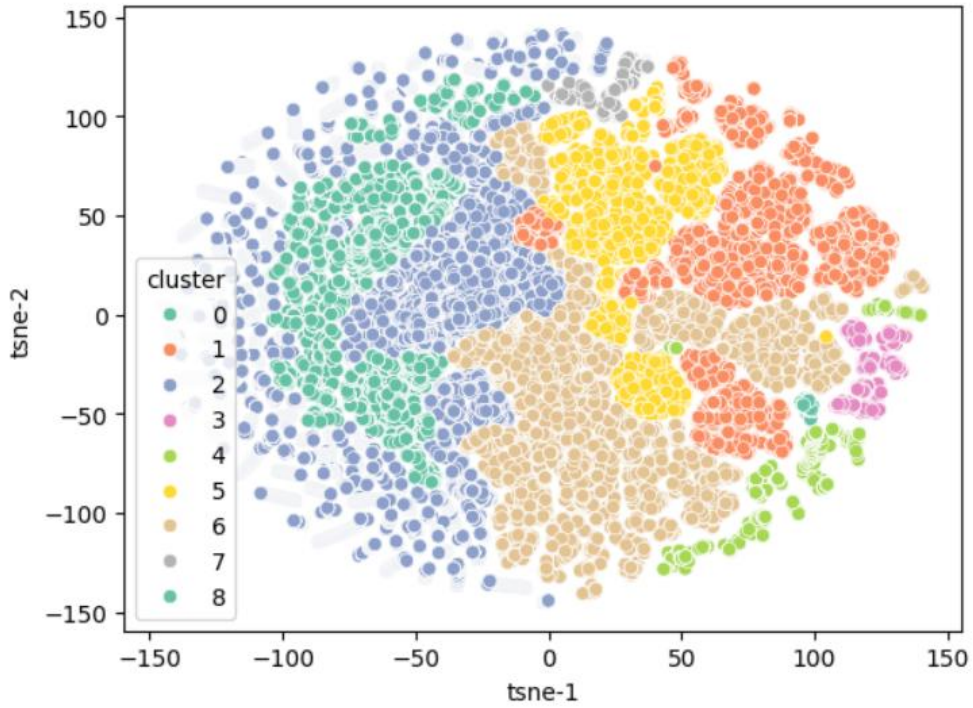

Figure 20: Visualisation des clusters

- D'après l'évaluation le modèle est capable de généraliser.
  Évaluation du modèle de clustering : **Silhouette Score** : 0.326 (Moyenne séparation)

Par conséquent, le système est fonctionnel et il répond aux objectifs cités au préalable.

Parmi nos objectifs nous avons l'objectif de visualiser les données à l'aide d'une interface utilisateur qui permet de charger les données et avoir les résultats attendus.

Figure 21: Première interface

Figure 22: première interface avec les descriptions des clusters

Les deux figure ci-dessus représente l'interface utilisateur représentant l'identification de neuf clusters avec les détails de chacun. En effet c'est la partie de guide de clusters qui aides les utilisateurs d'interprète et de prendre des décisions.

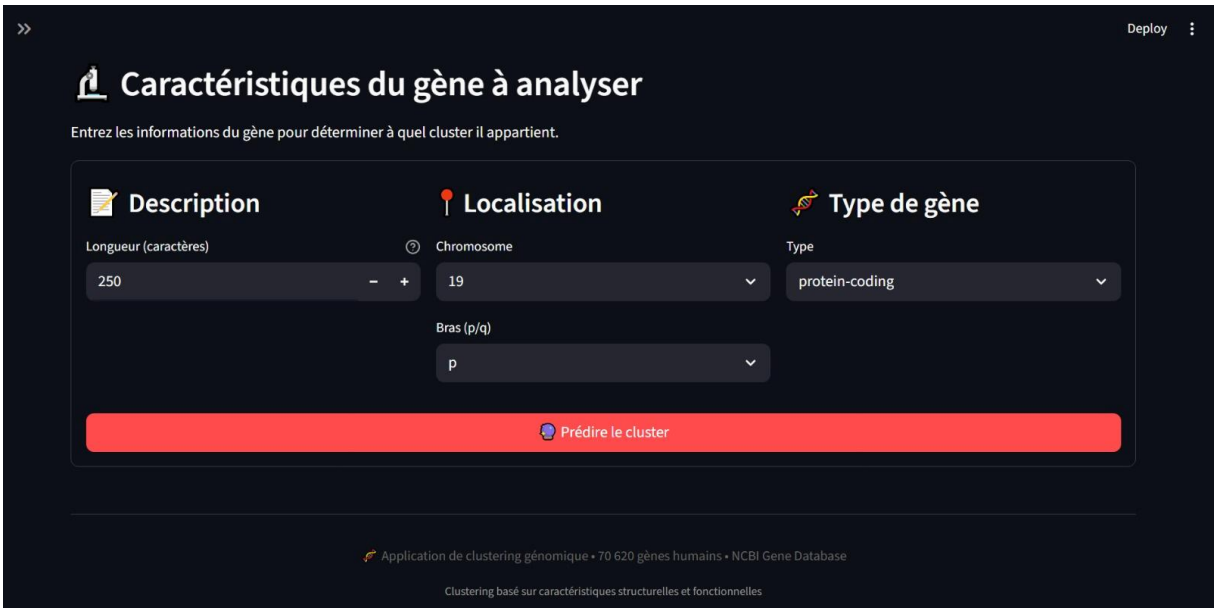

Figure 23: interface de saisie

Cette figure représente l'interface d'interaction avec l'utilisateur qui va charger les données de gène pour pouvoir savoir dans quel clusters il appartient

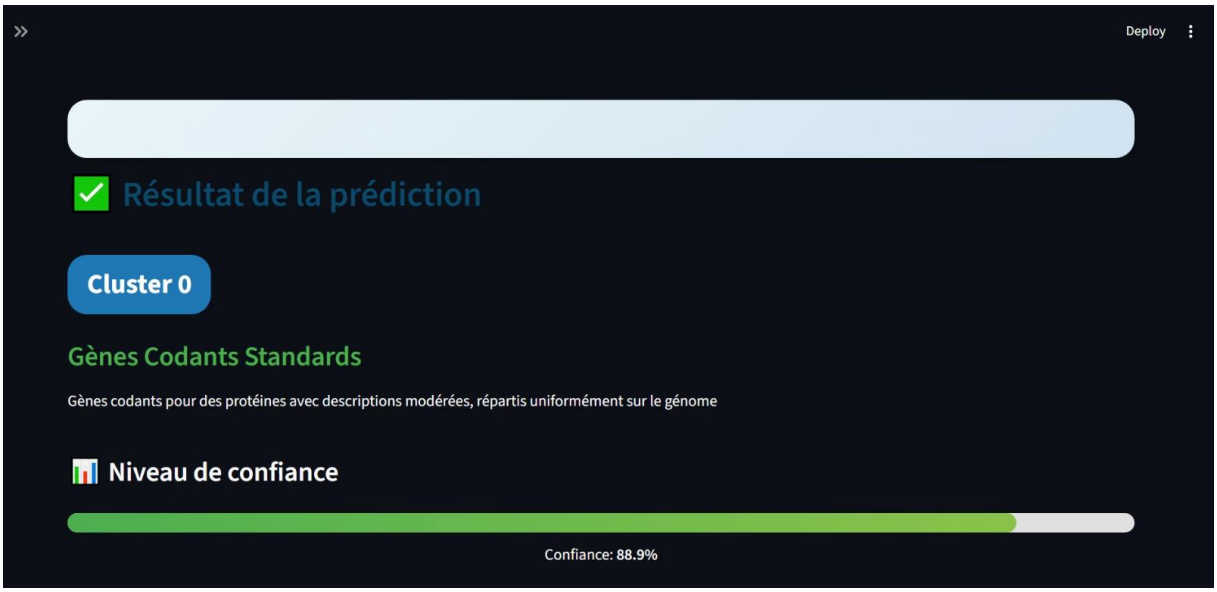

Figure 24: première partie de l'interface des résultats

Figure 25: deuxième partie de l'interface des résultats

Les deux figures ci-dessus, représente les résultats et les interprétations en se basant sur les informations données par l'utilisateur.

Ces résultats montrent que le système est fonctionnel et qu'il répond aux objectifs et besoins

4. Discussion et critiques :

Notre projet de clustering et analyse de l'expression génique, est fonctionnel, où il répond aux objectifs et aux besoins, à savoir le clustering et la visualisation.

Parmi les points forts de notre systèmes, l'interface simple a navigué et utilisable, ainsi que la simplicité du modèle où il suffit uniquement de charger les données.

Même si le modèle fonctionne bien, mais il y a des limitations. En effet ce modèle repose sur la qualité des données pour données des bons résultats, si les données contiennent des erreurs ou ne sont pas compatible avec le modèle, les résultats vont montrer cette incompatibilité.

Pour avoir un système fonctionnel dans différentes situations, nous pouvons l'améliorer par un modèle plus performant.

5. Conclusion :

Dans ce chapitre nous avons détailler les résultats obtenus en les comparant avec les objectifs et les besoins cité, et comme nous avons déjà expliquer, le système est fonctionnel puisque le modèle donne un score acceptable, et l'interface est simple et informative, où elle donne les visualisations des clusters juste après le chargement de données.

Un système fonctionnel ne signifie pas que tout est parfait, mais il y a toujours un espace pour améliorer le système vers une version plus performante.

Conclusion générale et perspectives :

Ce projet a permis de développer un modèle machine Learning permettant l'analyse et le clustering des gènes avec des données de NCBI et un modèle Kmeans avec une interface interactive streamlit .

L'architecture de notre système avec la relation entre eux, en effet les données sont prises d'après NCBI, un jeu de données qui contient des caractéristiques important pour l'entrainement du modèle pour rendre notre projet réaliste, puis le nettoyage et l'analyse des données pour valoriser les données, où nous avons utilisé plusieurs technique et visualisations, ainsi que la modélisation pour atteindre l'objectifs principale de ce projet, et pour rendre le projet et le modèle utilisable et compréhensible nous avons fait la conception de l'interface utilisateur avec Streamlit.

Parmi les points forts de notre systèmes, l'interface simple a navigué et utilisable, ainsi que la simplicité du modèle où il suffit uniquement de charger les données.

Même si le modèle fonctionne bien, mais il y a des limitations. En effet ce modèle repose sur la qualité des données pour données des bons résultats, si les données contiennent des erreurs ou ne sont pas compatible avec le modèle, les résultats vont montrer cette incompatibilité.

Pour avoir un système fonctionnel dans différentes situations, nous pouvons l'améliorer par un modèle plus performant.

Bibliographie

[1] X. Y. Y. W. D. Z. a. J. H. Qi Guan, «Biclustering analysis on tree-shaped time-series single cell gene expression data of Caenorhabditis elegans,» *BIOINFORMATICS*, p. 15, 2024.

[2] P. S. L. L. A. R. C. Flavio Pazos Obrego, «Cluster Locator, online analysis and visualization of gene clustering,» *BIOINFORMATICS*, p. 3, 2018.

[3] a. H. K. Jaehee Kim, «Clustering of change patterns using Fourier coefficients,» *BIOINFORMATICS*, p. 8, 2008.