



IUT CLERMONT AUVERGNE

Aurillac - Clermont-Ferrand - Le Puy-en-Velay
Montluçon - Moulins - Vichy



Science des
données
Aurillac



RAPPORT D'ACTIVITE POUR LA SAE 4-01 : Expliquer ou prédire une variable quantitative à partir de plusieurs facteurs

ETUDIANTS :

OLAYEMI JOSIAH
AGUIDA OTINIEL
FOTSO ERWAN
KOCOVI AUREL

ENCADRANT:

M. PAUL MARIE
M. MIGDAL

ANNEE UNIVERSITAIRE 2025-2026

JEUDI, 17 FEVRIER 2025

INTRODUCTION

Le projet **Cancer-DNASeq** s'inscrit dans le cadre de la SAE 4-01 croisée EMS/VCOD du département Science des données. Il prend appui sur un scénario réaliste simulant une collaboration entre une organisation non gouvernementale spécialisée dans l'analyse des données relatives au VIH et la Commission européenne, dans un contexte de centralisation et d'exploitation de données génomiques issues de plusieurs pays.

Dans ce cadre, l'ONG AidsSupport se voit confier la mission de structurer, analyser et interpréter des données hétérogènes portant sur des séquences ADN du VIH, avec pour objectif d'améliorer le suivi de la dynamique de propagation du virus et d'explorer l'existence de liens indirects avec certaines pathologies, notamment le cancer du sein. Les données mobilisées dans ce projet proviennent de sources multiples, sont transmises progressivement et nécessitent un travail rigoureux de mise en cohérence avant toute exploitation analytique.

Le travail présenté dans ce rapport correspond à la phase initiale de ce projet. Il se concentre sur l'organisation du travail en équipe, la collecte et la centralisation des données, ainsi que sur la mise en place des premières analyses. Ce rapport a pour vocation de rendre compte de la démarche adoptée, des choix méthodologiques effectués et des enseignements tirés de cette première phase de travail, dans un contexte proche des conditions rencontrées en milieu professionnel.

1. Description des activités réalisées et résultats obtenus

Les activités réalisées dans le cadre de ce projet se sont déroulées en plusieurs étapes successives. Celles-ci correspondent à l'évolution progressive des données mises à disposition et à l'organisation du travail au sein de l'équipe.

Cette partie décrit les principales actions menées, depuis la collecte et la centralisation des données jusqu'aux premières analyses effectuées, en mettant en évidence les résultats obtenus ainsi que les difficultés rencontrées.

1.1 Centralisation et structuration des données (Jour 1)

Lors de la première journée de travail, les tâches ont été réparties entre les deux sous-équipes afin de progresser en parallèle sur la collecte et la compréhension des données.

L'équipe VCOD s'est chargée de la récupération des données complémentaires via l'API du serveur, tandis que l'équipe EMS a analysé les fichiers déposés sur la plateforme Moodle afin d'en évaluer le contenu et la structure. Ces travaux préliminaires ont permis d'identifier les variables disponibles et d'orienter les choix méthodologiques.

À l'issue de cette phase initiale, l'équipe s'est réunie pour définir un cahier des charges commun et formaliser les missions suivantes :

- **Etude de l'évolution de la contamination au VIH chez les femmes atteintes d'un cancer, selon le temps et les pays considérés.**
- **Conception d'un algorithme de détection de motifs génétiques spécifiques (amorces) dans les séquences ADN, en tenant compte de potentielles mutations.**
- **Mise en place d'un modèle linéaire auto-régressif pour prédire l'évolution du nombre de cas d'un trimestre à l'autre, à l'échelle des pays ou de l'Europe.**
- **Etude du lien entre le cancer du sein et la présence de certains facteurs comme le statut de fumeur ou la consommation d'alcool.**

La seconde partie de la journée a été consacrée à la centralisation des données collectées via l'API. Les différents jeux de données ont été concaténés afin de constituer un fichier CSV unique, facilitant les analyses ultérieures et les travaux de modélisation prédictive.

1.2 Définition du cahier des charges et passage à l'exécution (Jour 2)

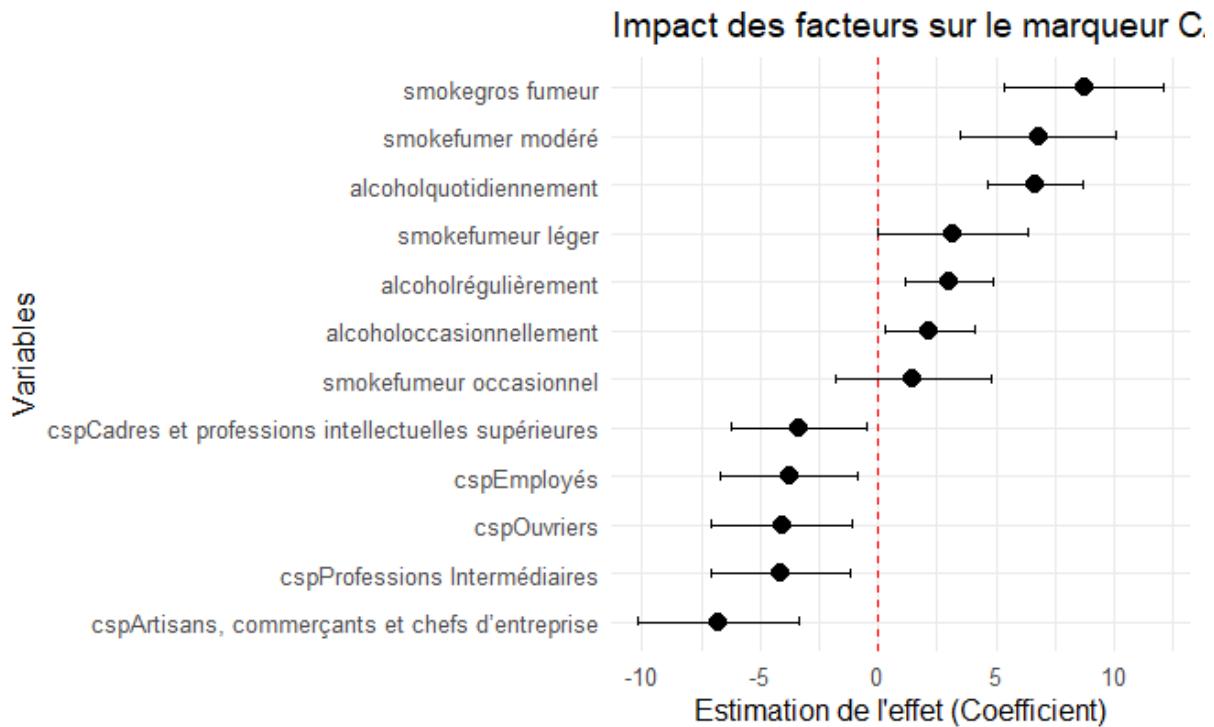
La deuxième journée a marqué le passage à une phase plus exécutive du projet. Dans un premier temps, l'équipe s'est réunie afin d'établir un cahier des charges détaillant l'ensemble des tâches à réaliser. Ce document a permis de structurer le travail et de fixer des objectifs clairs.

Enfin, l'équipe est passée à l'action en développant les différents outils et analyses nécessaires. Au terme de la journée, le présent rapport a pu être rédigé pour faire le point des avancées effectuées dans notre démarche.

2. Résultats et interprétations au terme des deux jours de travail

L'analyse du lien entre le cancer du sein et certains facteurs de mode de vie met en évidence des effets variables selon le statut de fumeur et la consommation d'alcool. Les résultats

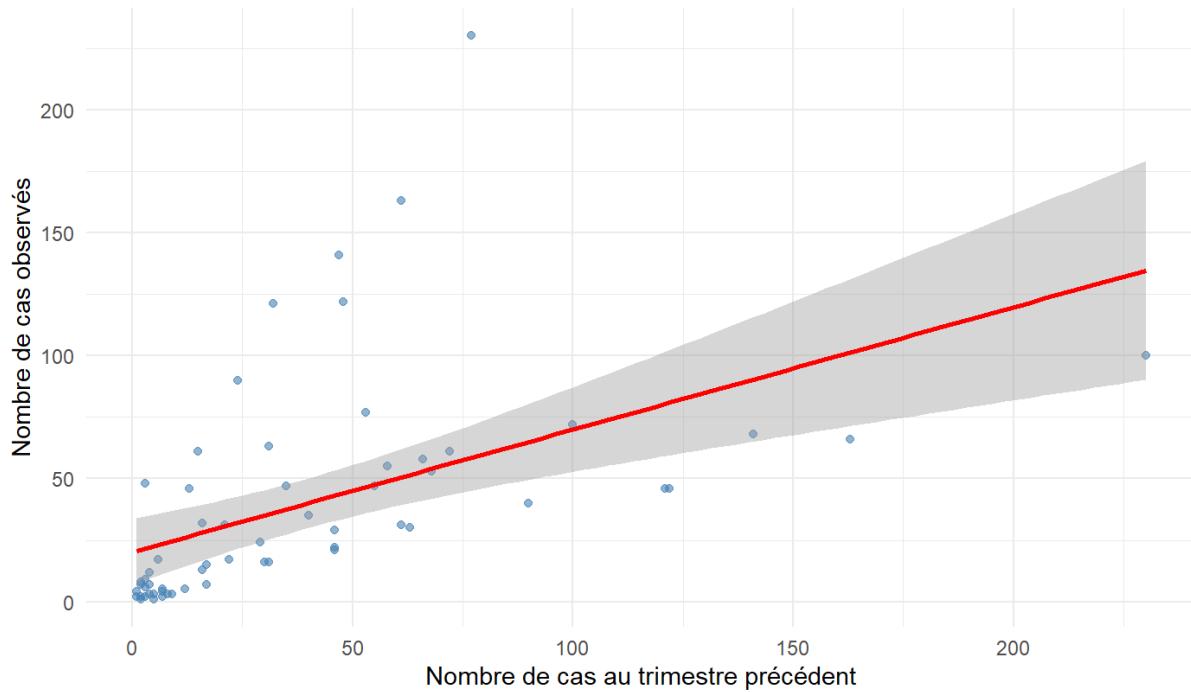
suggèrent que le tabagisme modéré à intensif ainsi qu'une consommation régulière ou quotidienne d'alcool sont associés à des effets positifs plus marqués sur le marqueur étudié, tandis que les statuts occasionnels ou légers présentent des effets plus faibles.



Certaines catégories socioprofessionnelles apparaissent quant à elles associées à des effets négatifs, indiquant un rôle potentiel des facteurs socio-économiques. Ces résultats doivent toutefois être interprétés avec prudence, dans un cadre exploratoire.

L'analyse auto-régressive met en évidence une relation positive entre le nombre de cas observés à un trimestre donné et celui du trimestre précédent. Le nuage de points et la droite de régression suggèrent que les valeurs passées constituent un bon indicateur de l'évolution à court terme, confirmant la pertinence d'un modèle linéaire auto-régressif pour ce type de données.

Relation entre les cas du trimestre T et T-1



Toutefois, la dispersion observée, notamment pour les valeurs élevées, indique une variabilité non négligeable qui limite la précision des prédictions. Ces résultats montrent que ce modèle fournit une première approximation utile, mais qu'il gagnerait à être complété par d'autres variables explicatives ou des modèles plus élaborés.

3. Difficultés rencontrées et solutions apportées

Plusieurs difficultés ont été rencontrées tout au long du projet. La principale concernait l'hétérogénéité des données et leur arrivée progressive, qui a nécessité des ajustements réguliers du traitement mis en place. De plus, certaines analyses se sont révélées complexes en raison de la structure des données et de leur volume.

Pour faire face à ces difficultés, l'équipe a adopté une approche itérative, combinant nettoyage des données, adaptation des algorithmes et échanges réguliers entre les membres du groupe. Cette méthodologie a permis de surmonter les obstacles rencontrés et d'assurer la continuité du travail.

CONCLUSION

Le travail réalisé dans le cadre du projet *Cancer-DNASEq* a permis de mettre en place une chaîne complète de traitement et d'analyse de données ADN du VIH. L'organisation en sous-équipes a favorisé une répartition efficace des tâches et une bonne gestion du temps, tandis que la mise en commun des compétences a renforcé la qualité des résultats obtenus.

Ce projet a également permis de mobiliser et de développer des compétences variées, telles que le travail en équipe, la gestion de données complexes, l'analyse statistique et la programmation. Les résultats obtenus constituent une base solide pour la poursuite du projet.

Les perspectives incluent l'amélioration des algorithmes existants, l'intégration de nouvelles données et l'approfondissement des analyses statistiques, notamment en vue d'une restitution plus interactive des résultats.