RAPPORT MODULE

SCIENCE DES DONNÉES 4

ETUDE DE L'IMPACT DES FACTEURS SOCIAUX SUR LES MALADIES EN FRANCE MÉTROPOLITAINE

HAKIRI Siwar, BENRAMDANE Rayane

Département MIASHS, UFR 6 Informatique, Mathématique et Statistique
Université Paul Valéry, Montpellier 3

Avril 2025

Remerciements

Nos plus sincères remerciements vont à Mme PATEL Namrata qui nous a encadré tout au long de ce rapport. Les remerciements vont également a Mme BRINGAY Sandra et Monsieur Théodore MICHEL PICQUE pour nous avoir accompagné et guidé tout au long de ce projet.

10/04/2025

Résumé

Durant cette étude de projet, nous avons décidé de nous intéresser à l'influence de l'environnement social sur les maladies en France. Pour ce faire nous avons dans un premier temps recherché des bases de données relatives aux facteurs sociaux. Nous avons ensuite séléctionné les facteurs nous semblant les plus pertinents et nous avons gardé un data contenant les taux de chomage, pauvreté, logements sociaux pour chaque département entre 2018 et 2023. Nous avons également décidé de chercher une base de données qui caractérise le nombre de sportifs par département par année en France pour analyser les tendeances d'activités physiques.

Une fois ces facteurs réunis, nous avons séléctionné une base de donnée relative aux pathologies en France et qui affiche chaque maladie en France par département ainsi que plusieurs facteurs intéressants à analyser (comme la prévalence). Après la réunification de ces 2 datas, nous avons démarré la partie Visualisation des Données pour avoir une approche plus schématique de notre problématique. Ainsi pour chaque facteur social nous avons modélisé une figure pour voir les tendeances du facteur social selon les zones en France. Et pour chaque figure nous la comparions avec un diagramme en barre représentant le taux de chaque pathologie en France par zone pour pouvoir observer d'éventuels liens. Le but principal de notre étude consiste a se questionner sur l'impact des facteurs sociaux dans la prévalence des pathologies dans le territoire métropolitain. Cette problématique nous a orienté vers des modèles de Machine Learning appuyant notre travail précédent de prévisualisation des données grâce aux figures.

Nous avons adopté une approche basée sur des algorithmes de Machine Learning afin de voir s'il était possible de prédire ou d'observer des tendeances de maladies plus fréquentes dans certaines zone en fonction des facteurs de l'environnement social. En effet nous avons cherché à prédire la maladie la plus fréquente grâce a des bases de données remodifiées que nous vous détaillerons dans la suite du rapport. Puis nous avons par la suite réaliser une ACP afin de faciliter l'étude de notre problème, accompagnée d'un clustering par départements (en fonction des similitudes des valeurs de facteurs sociaux). Et suite à cela, nous avons réalise une "HeatMap" afin d'observer la prévalence moyenne de chaque pathologie dans chaque cluster pour tenter observer d'éventuels corrélations dans l'étude de notre problématique. Enfin nous concluerons sur ce rapport en résumant nos résultats obtenus et tentant de répondre a notre problématique tout en expliquant les limites et difficultés rencontrées lors de cette étude.

Table des matières

Chapitre 1

Description de la collecte et du nettoyage des données

1.1 Collecte des données

Dans le cadre de ce projet, nous avons utilisé une base de données intitulée effectif de patients par pathologie, sexe, classe d'âge et territoire (département, région) disponible sur le portail de données de l'Assurance Maladie .Cette base fournit des informations détaillées sur le nombre de patients pris en charge pour une large variété de pathologies, qu'elles soient aiguës ou chroniques. Les données permettent d'analyser la répartition des affections à l'échelle nationale, en fonction de différentes pathologies, du sexe, de l'âge et de la localisation géographique par département. La base comprend plusieurs colonnes essentielles pour l'analyse des pathologies et des populations concernées. Parmi ces colonnes, nous retrouvons :

annee : L'année des données, permettant d'observer l'évolution des prises en charge des pathologies au fil du temps.

dept : Le code des départements, facilitant une analyse encore plus précise à l'échelle départementale.

patho_niv1 : Le premier niveau de classification des pathologies, regroupant des affections générales telles que "Cancer", "Maladies cardiovasculaires", ou "Maladies respiratoires".

patho_niv2 : Le second niveau, qui spécifie davantage ces pathologies. Par exemple, sous la catégorie "Cancer", on peut retrouver des pathologies plus précises comme "Cancer broncho-pulmonaire", "Cancer colorectal", "Cancer du sein", etc.

patho_niv3 : Un troisième niveau de classification encore plus spécifique, qui permet de distinguer des sous-groupes comme "Autre cancer du poumon" ou "Autre forme de cancer colorectal". **ntop** : Le nombre total de patients atteints de chaque pathologie dans une zone

géographique donnée (département)

npop : Le nombre total de la population dans la zone géographique étudiée, ce qui permet de calculer la prévalence des pathologies en rapportant le nombre de patients au nombre total d'habitants.

prev : La prévalence des pathologies, représentant la proportion de personnes atteintes par une pathologie donnée dans la population totale d'un département ou d'une région. Afin de

compléter nos données relatives à la santé, nous avons intégré une base de données socio-économiques intitulée « Logements et logements sociaux dans les départements », disponible sur la plateforme publique data.gouv.fr. Cette base propose un aperçu global de la situation du logement dans chaque département français, en mettant en évidence non seulement le nombre total de logements et la proportion de logements sociaux, mais également plusieurs indicateurs clés liés aux conditions de vie, tels que le taux de pauvreté et le taux de chômage.
Cette base de données comprend plusieurs variables clés, parmi lesquelles :

code_departement : code INSEE permettant le croisement avec d'autres jeux de données **nom_departement** : nom du département concerné ; **Nombre de logements** : volume total

de logements présents sur le territoire ;

Nombre de logements sociaux : nombre de logements à vocation sociale, indicateur central de l'offre en habitat accessible ;

Taux de logements sociaux (en %) : part des logements sociaux dans l'ensemble du parc immobilier ;

Taux de pauvreté (en %) : pourcentage de la population vivant en dessous du seuil de pauvreté, révélateur du niveau de précarité locale ;

Taux de chômage au T4 (en %) : taux de chômage observé au quatrième trimestre, reflétant la situation du marché de l'emploi.

Les données disponibles au format CSV ont permis d'approfondir notre analyse de l'apparition des maladies chroniques dans les départements, en se basant sur l'effectif de patients par pathologie. En associant ces données avec des indicateurs socio-économiques tels que le taux de pauvreté, le chômage et la proportion de logements sociaux, nous avons pu identifier les variations territoriales dans la prévalence des différentes pathologies. Cette analyse a ainsi révélé comment les conditions socio-économiques peuvent affecter la répartition des maladies chroniques à travers les départements. La troisième base utilisée dans ce projet concerne le nombre de licenciés sportifs par département, disponible sur le site Observatoire des territoires. Cette base fournit des informations détaillées sur le nombre de licenciés dans différentes disciplines sportives, répartis par département en France. Ces données sont particulièrement utiles pour analyser la pratique du sport à l'échelle locale et évaluer son influence sur la santé publique et d'autres facteurs socio-économiques.
La base de données contient les colonnes suivantes :

codgeo : Le code géographique unique attribué à chaque département, permettant une identification précise des zones géographiques.

libgeo : Le nom du département, offrant une identification claire et compréhensible des territoires.

an : L'année des données, permettant de suivre l'évolution des indicateurs au fil du temps.

nb_licsport : Le nombre total de licenciés sportifs dans chaque département pour l'année donnée. Cette colonne renseigne sur la pratique sportive locale et son ampleur.

Ces données sont essentielles pour étudier la répartition de la pratique sportive à l'échelle nationale et peuvent être croisées avec d'autres variables pour analyser les impacts sur la santé, les inégalités d'accès au sport, ou encore la relation entre la pratique sportive et des facteurs socio-économiques dans chaque département.

La quatrième base utilisée dans ce projet provient du fichier Finess – Extraction du fichier des établissements de santé, disponible sur le site data.gouv.fr. Cette base regroupe des informations concernant les établissements de santé en France, qu'il s'agisse d'hôpitaux, de cliniques privées, de centres de soins ou d'autres structures spécialisées. Elle permet de localiser ces établissements par département, ce qui est essentiel pour évaluer la répartition des services de santé sur le territoire et leur accessibilité pour les populations.

La base contient principalement les colonnes suivantes :

Code : Le code unique d'identification de chaque établissement de santé, utilisé pour distinguer chaque structure.

Departement : Le nom du département où l'établissement est situé. Cette colonne permet de localiser géographiquement chaque établissement dans le pays. **etab** : Le nom de

l'établissement de santé, permettant d'identifier les différents types de structures de soins présentes dans chaque département.

Ces données sont utiles pour analyser la répartition et l'accessibilité des soins dans les différents départements, en lien avec d'autres indicateurs comme la prévalence des maladies, les taux de pauvreté, ou encore la pratique sportive, afin de mieux comprendre l'impact de l'accessibilité aux soins sur la santé publique.

La derniere base utilisée dans ce projet provient de l'INSEE et concerne la répartition des catégories socio-professionnelles par département, disponible sur le site INSEE. Cette base fournit des informations sur la part des différentes catégories socioprofessionnelles dans chaque département français, ce qui permet d'analyser la structure professionnelle de la population à l'échelle départementale. Ces données sont particulièrement pertinentes pour comprendre les différences socio-économiques entre les départements et leur impact potentiel sur d'autres variables comme la santé, la pauvreté ou l'accès aux soins.

La base contient les colonnes suivantes :

code : Le code du département, permettant d'identifier chaque région géographique de manière unique.

departement : Le nom du département, facilitant la lecture des données.

Part des agriculteurs exploitants : La part de la population active travaillant dans le secteur agricole, en pourcentage.

Part des artisans : La proportion de la population active exerçant une activité artisanale.

Part des cadres, professions intellect. supérieures : La part de la population active travaillant dans des postes de direction ou des professions intellectuelles supérieures. **Part des**

professions interméd. : La proportion de personnes occupant des professions intermédiaires, telles que les techniciens ou les agents de maîtrise.

Part des employés : La part de la population active travaillant dans des métiers employés, souvent dans les services.

Part des ouvriers : La proportion de la population active travaillant dans des métiers manuels ou industriels.

Autres : La part restante de la population active, incluant d'autres catégories professionnelles non spécifiées dans les autres colonnes.

Ces données permettent d'examiner la répartition des catégories socio-professionnelles entre les départements, offrant ainsi une vision de la structure économique de chaque région. Elles peuvent être croisées avec des indicateurs de santé et des facteurs sociaux, afin d'analyser comment les caractéristiques socio-professionnelles influencent la répartition des maladies et les disparités sociales au sein des différentes departements.

1.2 Pré-traitement des données

Avant de procéder aux analyses statistiques et aux visualisations, il est essentiel d'effectuer un travail de prétraitement des données. Le prétraitement vise notamment à corriger les erreurs éventuelles, harmoniser les formats, gérer les valeurs manquantes, normaliser les noms de colonnes, et créer des correspondances entre les jeux de données via des identifiants communs (comme le code département). Dans ce chapitre, nous détaillerons les différentes étapes de préparation appliquées à chacune des bases. La première base traitée dans le cadre de ce projet est `effectifs_filtered_sorted7.csv`. Nous avons appliqué plusieurs opérations de prétraitement afin d'assurer la qualité et la cohérence des données en vue de leur exploitation. Tout d'abord, le fichier a été importé en testant plusieurs encodages (`utf-8` puis `ISO-8859-1`) pour éviter toute erreur de décodage. Le séparateur de champs utilisé était le point-virgule (;), mais une vérification a également été effectuée avec la virgule (,) pour s'assurer de la bonne structure du fichier. Ensuite, les noms des colonnes ont été nettoyés, notamment par la suppression des espaces superflus, afin de faciliter leur manipulation dans les outils d'analyse. Une attention particulière a été portée à la colonne `patho_niv2`, dont l'existence a été vérifiée. Les différentes modalités qu'elle contient ont été explorées à l'aide d'une extraction des valeurs uniques pour mieux comprendre le contenu de la base. Les lignes contenant des valeurs manquantes dans cette colonne ont été supprimées, afin de ne conserver que les données exploitables. De plus, certaines modalités non pertinentes ont été écartées, notamment celles faisant référence à des hospitalisations générales, des soins non spécifiques ou encore des maternités. Ce filtrage a permis de recentrer l'analyse uniquement sur les pathologies identifiées. À l'issue de ce nettoyage, le fichier traité a été sauvegardé sous le nom `effectifs_cleaned1.csv`, en conservant un séparateur point-virgule, garantissant ainsi sa compatibilité avec les outils d'analyse utilisés par la suite. Le processus de nettoyage des

données de la deuxième base « taux_logement.csv » a principalement consisté à traiter les valeurs manquantes, afin de garantir la fiabilité et la pertinence des analyses. Le fichier a été chargé à l'aide de la bibliothèque `Pandas`, puis une vérification a été effectuée sur la colonne « Taux de pauvreté (en %) », jugée essentielle pour l'étude. Les lignes contenant des valeurs manquantes (`NaN`) dans cette colonne ont été supprimées à l'aide de la fonction `dropna()`, afin d'éliminer les observations incomplètes susceptibles de biaiser les résultats. Une fois cette opération réalisée, les données nettoyées ont été enregistrées sous un nouveau nom, `taux_logement_nettoye.csv`, permettant ainsi de préserver l'intégrité du fichier original. Un aperçu des premières lignes du fichier final a été affiché pour vérifier que le traitement s'était déroulé correctement, et le nombre total de lignes supprimées a été comptabilisé pour évaluer l'impact du nettoyage. Lors de l'inspection initiale du fichier `licsport (2).xlsx`, il a été observé que les noms des colonnes ne figuraient qu'à la cinquième ligne du document. Par conséquent, lors du chargement avec la bibliothèque `Pandas`, la ligne correspondante a été définie comme en-tête à l'aide de l'option `header=4`. Cette étape a permis de structurer correctement le jeu de données. Ensuite, un filtrage a été appliqué afin de supprimer toutes les lignes contenant des valeurs manquantes dans la colonne `nb_licsport`, représentant le nombre de licenciés. Cette opération visait à écarter les observations incomplètes qui pourraient compromettre la précision des analyses. Enfin, la version nettoyée du fichier a été enregistrée sous le nom `licsport_nettoye.csv`. Le fichier `etablissementsante.csv` a été chargé en conservant la première ligne comme en-tête (`header=0`), et le type de toutes les colonnes a été

forcé en chaîne de caractères (`dtype=str`) afin d'éviter toute conversion automatique pouvant entraîner des pertes d'information. Une première vérification a permis d'identifier les noms des colonnes, facilitant la localisation de celle contenant les données sur les établissements. La colonne pertinente, contenant les informations sur les établissements de santé, a été repérée dynamiquement en recherchant un intitulé contenant le mot-clé « etab », rendant le processus adaptable à d'éventuelles variations de format. Une fois cette colonne identifiée, un filtrage a été appliqué pour supprimer les lignes comportant des valeurs manquantes ou des espaces vides, garantissant ainsi l'intégration uniquement d'entrées exploitables. Les lignes entièrement vides ont également été éliminées afin d'obtenir une structure de fichier propre. Le jeu de données ainsi nettoyé a été sauvegardé sous le nom `etablissementsante_clean.csv`, avec un séparateur point-virgule (;) afin d'assurer sa compatibilité avec d'autres outils d'analyse.

Pour la dernière base utilisée, le fichier a été lu sans en-tête, afin d'en analyser la structure et de repérer la ligne contenant les noms des colonnes. Il a été observé que ces noms figuraient à la quatrième ligne (index 3 en Python), ce qui a permis de recharger correctement les données en définissant cette ligne comme en-tête (`header=4`).

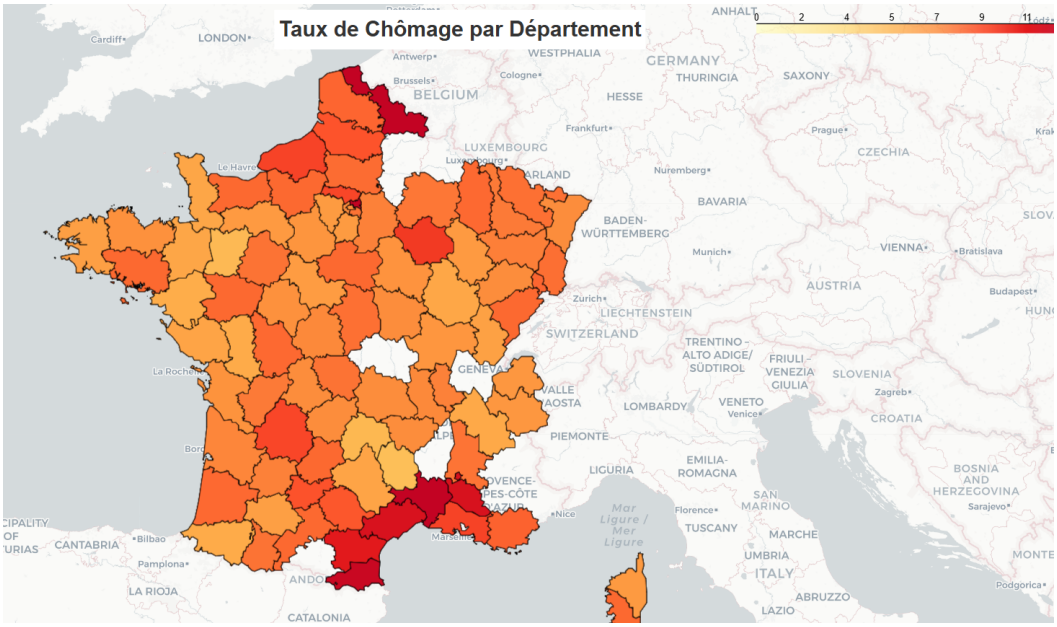
Par la suite, un nettoyage des noms de colonnes a été réalisé afin de supprimer les espaces superflus et caractères indésirables, assurant ainsi une meilleure lisibilité et facilitant les traitements ultérieurs. Les lignes entièrement vides ont été éliminées à l'aide de la fonction `dropna(how='all')`, garantissant une structure propre et sans doublons inutiles.

Enfin, le fichier nettoyé a été enregistré sous le nom `categorie.csv`, avec un séparateur point-virgule (;).

Chapitre 2

Visualisation des données

2.1 Présentation des figures et intepretations



Sur cette figure, nous avons cherché à visualiser le taux de chômage moyen en France entre 2018 et 2022.On peut s'apercevoir que le Nord et le Sud sont les régions les plus touchées. Dans le Nord, la désindustrialisation a impliqué de nombreuses difficultés sociales qui ont perduré dans le temps, tandis que le Sud, fortement dépendant du tourisme et des emplois saisonniers, subit une instabilité professionnelle. En effet en fonction des périodes le secteur de l'emploi est très aléatoire. Ce facteur économique joue un rôle clé dans l'état de santé des populations, en limitant l'accès aux soins et en aggravant les inégalités sanitaires.

En mettant en parallèle ces données avec la figure représentant les pathologies , on observe que les maladies chroniques (cancers, maladies cardiovasculaires, affections respiratoires et troubles psychiatriques) sont plus fréquentes dans les zones à fort chômage. L'insécurité financière pousse à retarder les consultations médicales, ce qui entraîne des diagnostics tardifs et des prises en charge plus difficiles. De plus, les conditions de vie souvent misérables dans certains cas, avec des logements moins bien isolés ou insalubres, favorisent les maladies respiratoires.

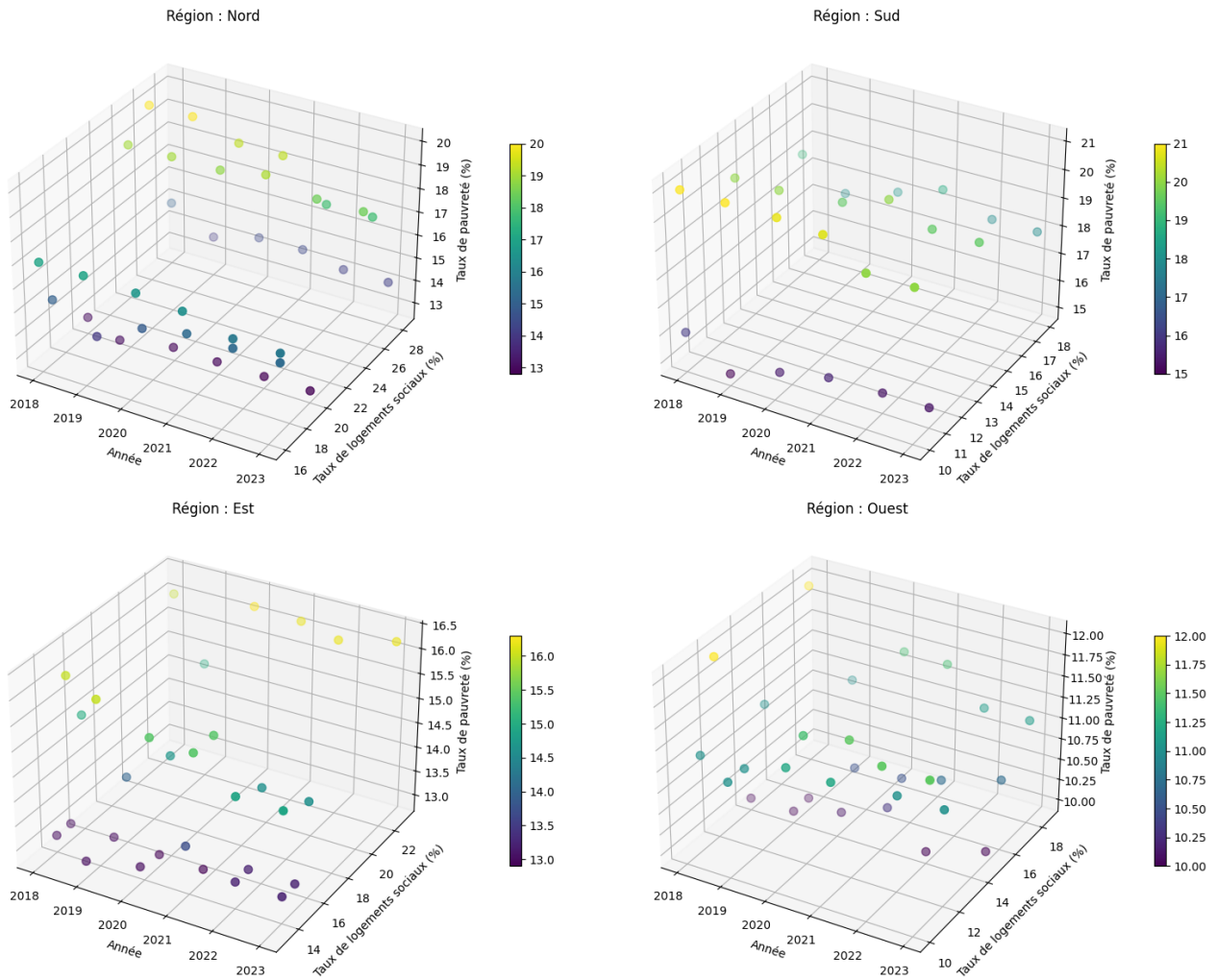
Nous pouvons également observer que le lien entre chômage et santé mentale est également impactant. Cela peut être lié au stress qui est impacté par l'absence de travail et la précarité

financière augmente les risques de dépression, d'anxiété et d'addictions (tabac, alcool), ce qui contribue à un état de santé globalement plus dégradé dans ces régions. Ce cercle vicieux, où la maladie freine le retour à l'emploi et où le chômage aggrave les pathologies, est particulièrement marqué dans le Nord*.

Bien que le Sud connaisse aussi un fort taux de chômage, certaines différences atténuent l'impact sur la santé : On pourrait expliquer cela par un climat plus "temperé", qui réduit les maladies respiratoires, et une alimentation plus méditerranéenne, bénéfique pour la santé cardiovasculaire. Cependant, la précarité dans certaines zones urbaines du Sud se traduit par une prévalence élevée du diabète et des troubles psychiatriques.

Toutes ces analyses et hypothèses nous montrent l'importance d'une approche intégrée, alliant politique de santé et emploi, pour réduire les inégalités régionales et améliorer la prise en charge des populations vulnérables.

— Figure 2 : Analyse de l'évolution des facteurs sociaux par année et par zone géographique



Dans cette figure nous pouvons constater que le Nord et le Sud sont les régions où le taux de pauvreté est le plus élevé, atteignant 20-21% dans certains départements. Ces valeurs indiquent une situation persistante de pauvreté, qui ne semble pas vraiment s'améliorer au fil des années. En

ce qui concerne ,L'Est affiche un taux intermédiaire, oscillant entre 13% et 16%, avec une légère tendance à la baisse dans certains départements. L'Ouest, quant à lui, présente les taux les plus faibles, variant entre 10% et 12%, et semble être la région la moins touchée par la pauvreté.

Plusieurs raisons expliquent pourquoi la pauvreté est plus marquée dans le Nord et le Sud. Ces régions pourraient être marquées par des déficites au niveau du développement importants entre les départements, où certains territoires disposent de moins de ressources et d'infrastructures de soutien aux populations en difficulté. De plus, elles pourraient connaître une forte concentration de ménages vivant sous le seuil de pauvreté, avec un accès limité aux dispositifs d'aide et aux opportunités d'amélioration des conditions de vie.

Nous savons également que les territoires du Nord et du Sud sont souventés confrontés a des hausses de prix et des couts élevés, rendant plus difficile pour les personnes modestes de couvrir leurs besoins. En effet, étant donné que nous avons choisi de mettre la région parisienne dans les départements du Nord, cela influe grandement la répartition des pathologies. Dans certaines zones, les logements sont plus coûteux et les services de base moins accessibles, accentuant les difficultés sociales et rendant la sortie de la pauvreté plus compliquée.

Enfin, il est possible que ces régions connaissent une pauvreté intergénérationnelle, où les familles en difficulté depuis plusieurs générations ont plus de mal à améliorer leur situation. La transmission des inégalités joue alors un rôle majeur, empêchant la baisse du taux de pauvreté sur le long terme.

En conclusion, le Nord et le Sud restent les régions les plus touchées par la pauvreté en raison d'inégalités territoriales, d'une concentration plus élevée de ménages précaires, d'un coût de la vie parfois plus élevé et d'une pauvreté persistante sur plusieurs générations. À l'inverse, l'Ouest affiche un taux plus faible et plus stable, tandis que l'Est montre une tendance à l'amélioration.

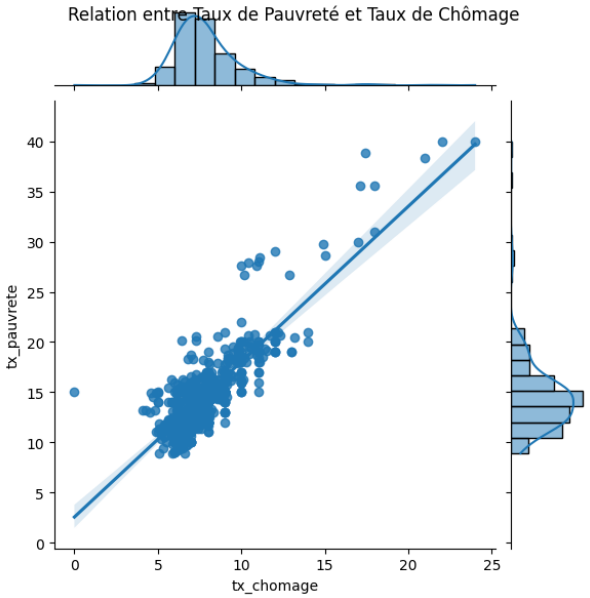


FIGURE 2.1: Corrélation le taux de pauvreté et le taux de chômage

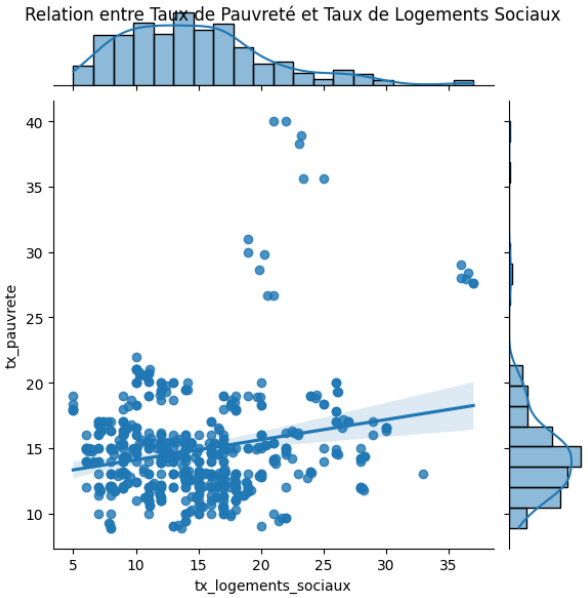


FIGURE 2.2: Corrélation entre le taux de pauvreté et le taux de logements sociaux

Le lien fort entre taux de chômage et taux de pauvreté : une cause structurelle Le premier graphique montre une corrélation très forte entre le taux de chômage et le taux de pauvreté. Cette relation s'explique par le fait que l'absence d'emploi prive les ménages de ressources financières stables, augmentant ainsi la précarité. Dans les territoires où le chômage est structurellement élevé, notamment en raison de la disparition d'industries ou d'un marché du travail peu dynamique, les habitants ont plus de difficultés à sortir de la pauvreté.

La relation entre taux de logements sociaux et taux de pauvreté : un effet de concentration Le deuxième graphique met en évidence un lien plus faible mais existant entre taux de logements sociaux et taux de pauvreté. Contrairement au chômage, les logements sociaux ne créent pas directement la pauvreté, mais ils ont tendance à être construits dans des zones où la pauvreté est déjà élevée. Cela peut donner l'impression que plus un territoire compte de logements sociaux, plus la pauvreté y est importante, alors qu'en réalité ces logements sont une conséquence plutôt qu'une cause. Une autre hypothèse possible est que les logements sociaux attirent principalement des populations aux revenus modestes, qui ne peuvent pas accéder toujours à des zones plus "favorisées". Cela entraîne une concentration de la précarité dans certaines zones, parfois accompagnée de difficultés sociales supplémentaires (isolement, moindre mixité sociale, accès limité aux opportunités économiques). Dans certains cas, cette concentration peut créer un effet de "zones isolées", où la pauvreté se transmet de génération en génération sans véritable ascenseur social.

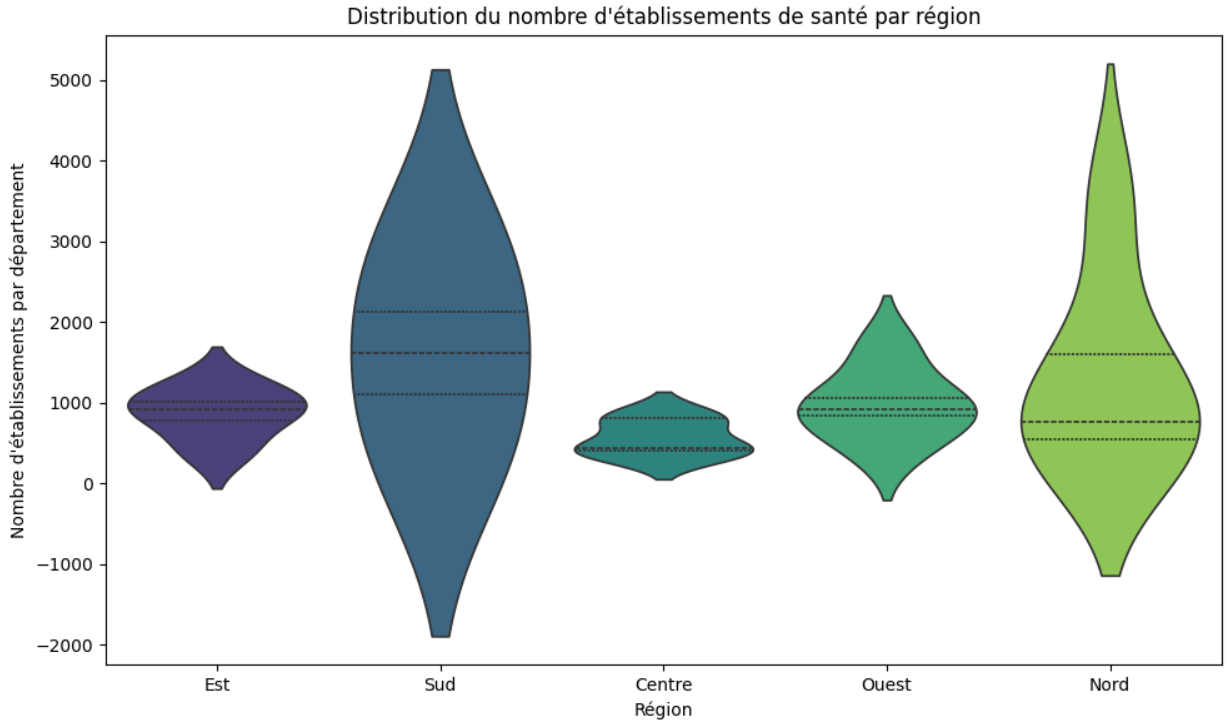


FIGURE 2.3: Distribution du nombre d'établissements de santé par région

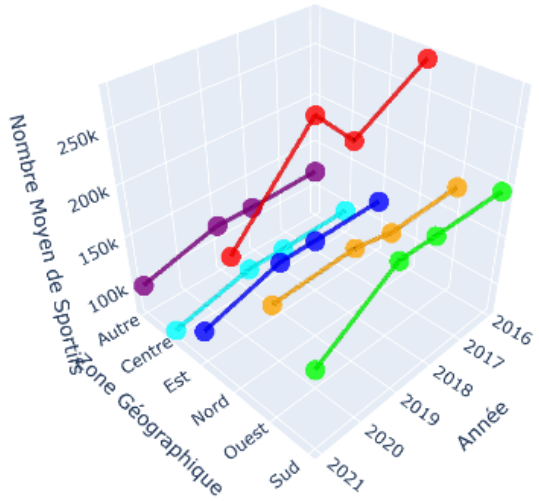
Ce graphique en violon présente la distribution du nombre d'établissements de santé par département au sein de cinq régions : Est, Sud, Centre, Ouest et Nord. La région Sud se distingue par une forte dispersion et un nombre élevé d'établissements, suggérant une grande variabilité entre les départements, potentiellement due à une forte densité de population et à la présence de grands centres urbains. La région Nord présente également une distribution étalée et un nombre élevé d'établissements, possiblement en raison de la présence de grandes villes. Les régions Est et Ouest montrent des distributions similaires et moins dispersées, indiquant une répartition plus homogène des établissements. En revanche, la région Centre se caractérise par une faible dispersion et un nombre plus bas d'établissements, ce qui pourrait s'expliquer par une densité de population plus faible et une centralisation des services de santé. En résumé, la distribution du nombre d'établissements de santé varie considérablement d'une région à l'autre, reflétant probablement des différences en termes de densité de population, d'urbanisation et de politiques de santé régionales.

En comparant cette figure avec la répartion et distribtuions des pathologies en France on peut hypothétiquer le fait qu'il existe un lien entre la densité médicale, le nombre d'hôpitaux et le nombre de maladies dans une région. Le fait que le Nord et le Sud possèdent un grand nombre d'hôpitaux peut être une conséquence directe d'une plus forte prévalence des maladies dans ces zones. Ces régions, marquées par des taux de chômage élevés et une précarité accrue, connaissent une population plus vulnérable, nécessitant une prise en charge médicale plus importante. La forte concentration d'hôpitaux dans ces territoires ne signifie pas forcément un meilleur accès aux soins, mais plutôt une réponse à une demande élevée liée à des conditions de vie plus difficiles, un environnement plus pollué et une prévalence accrue de maladies chroniques.

À l'inverse, l'Ouest et le Centre de la France, où l'on trouve moins d'hôpitaux, enregistrent souvent de meilleures conditions de vie : une pollution moindre, un climat plus clément, une alimentation plus équilibrée et une population globalement en meilleure santé. Cela signifie que la demande en infrastructures hospitalières y est plus faible, ce qui peut expliquer une densité médicale plus réduite. Cependant, un faible nombre d'hôpitaux peut aussi être un frein à l'accès aux soins, obligeant les habitants de ces régions à parcourir de longues distances pour consulter, ce qui peut retarder les diagnostics et le traitement de certaines pathologies.

Ainsi, la répartition des hôpitaux en France reflète à la fois la prévalence des maladies et l'accessibilité aux soins, mettant en lumière l'importance d'adapter les infrastructures médicales aux besoins réels des populations.

FIGURE 2.4: Cube 3D Evolution du nombre de sportifs par zone au fil des années



La figure en 3D montre l'évolution du nombre moyen de sportifs en France, répartis selon les départements (Nord, Sud, Est, Ouest) et les années. Chaque couleur représente une région, et on observe une baisse progressive du taux de sportifs au fil des années. Cette diminution est plus marquée dans le Nord, ce qui pourrait s'expliquer par plusieurs facteurs. Le COVID a sans doute joué un rôle majeur dans cette baisse, avec les confinements et la fermeture des équipements sportifs affectant plus durement certaines régions, notamment celles du Nord, où les conditions climatiques rendent la pratique extérieure plus difficile pendant l'hiver (étant donné qu'on a décide de classer Paris (75) comme département du Nord). La sédentarisation croissante et le manque d'infrastructures accessibles peuvent aussi avoir un impact plus important dans ces zones. De plus, la pression urbaine est plus forte dans le Nord, avec une densité de population plus élevée et des contraintes d'espace, ce qui peut limiter les opportunités de pratique sportive régulière. Par ailleurs, la mise en place du télétravail et des modes de vie plus restreins dans de mêmes lieux "fixes", combiné à une augmentation du temps passé sur les écrans, réduit la pratique de l'activité physique. Enfin, la précarité économique, notamment dans les régions les plus touchées par le chômage, limite l'accès aux loisirs sportifs, souvent coûteux. En revanche, les départements du Sud ou de l'Ouest ont une baisse moins marquée, probablement grâce à un climat plus favorable, une plus grande accessibilité à des espaces extérieurs et une plus faible densité de population, qui facilite l'engagement dans des activités physiques. En résumé, la figure montre une baisse généralisée de la pratique sportive, accentuée dans le Nord par des facteurs géographiques, économiques et sociaux. La crise du COVID-19, associée à des conditions climatiques et urbaines difficiles dans ces régions, explique cette pente plus accentuée.

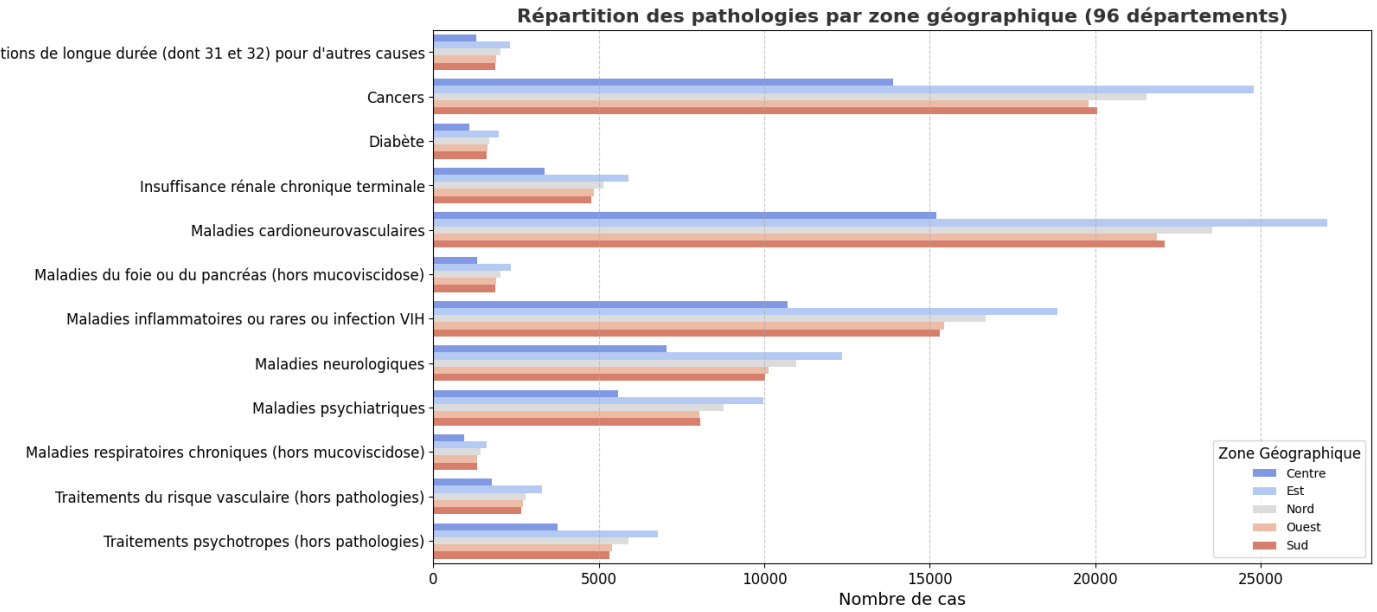


FIGURE 2.5: Répartition des pathologies par zone géographique

Dans cette figure présente nous avons chercher a représenter la répartition des pathologies de longue durée par zone géographique en France, couvrant 96 départements. Les pathologies incluent des affections graves et chroniques comme les cancers, le diabète, l'insuffisance rénale chronique terminale, les maladies cardioneurovasculaires, et d'autres conditions telles que les maladies inflammatoires, neurologiques, psychiatriques, et respiratoires chroniques. Les données, réparties entre les zones Centre, Est, Nord, Ouest, et Sud, montrent des variations significatives.

Régions de l'Est et du Nord : Ces zones sont souvent les plus touchées par les maladies, probablement en raison de facteurs socio-économiques (précarité, chômage), environnementaux (pollution industrielle, climat plus rude), et de modes de vie (alimentation riche, tabagisme). Ces facteurs contribuent à une prévalence plus élevée de maladies chroniques comme les cancers, les maladies cardioneurovasculaires, et les maladies respiratoires. Précédemment nous avons étudié une figure 3D sur la pauvreté qui confirme que cette région est également parmi les plus précaires. On pourrait y'voir un potentiel lien sur le fait que la pauvreté limite l'accès à une alimentation équilibrée, aux soins médicaux et à la prévention, favorisant l'apparition et l'aggravation des maladies*. Centre : Cette zone est généralement en dessous des autres en termes de nombre de cas, ce qui pourrait s'expliquer par une meilleure accessibilité aux soins, une population moins dense, et des conditions de vie plus favorables (moins de pollution, habitudes alimentaires plus équilibrées).

Pour ce qui est du Sud de la France, Contrairement au Nord, cette zone bénéficie d'un climat plus clément et d'un accès plus facile aux activités physiques (moins de freins liés aux conditions météorologiques) bien que la pauvreté y'est également élévée comme vu précedemment.

Impact du COVID-19 :

La pandémie a exacerbé les conditions préexistantes, en particulier dans les zones déjà vulnérables comme l'Est et le Nord, où les systèmes de santé ont été fortement sollicités. Les retards de diagnostics et de traitements ont probablement aggravé les maladies de longues durées dans ces régions.

Traitements et prévention :

Les différences régionales dans l'accès aux soins et la qualité des infrastructures de santé jouent un rôle clé dans les disparités observées. Le Centre, bénéficiant peut-être d'un meilleur accès aux soins,

montre des chiffres plus bas, tandis que l'Est et le Nord, confrontés à des défis socio-économiques et environnementaux, sont plus touchés.

En résumé, les disparités régionales reflètent des différences socio-économiques, environnementales, et d'accès aux soins, soulignant la nécessité de politiques de santé adaptées pour réduire ces inégalités.

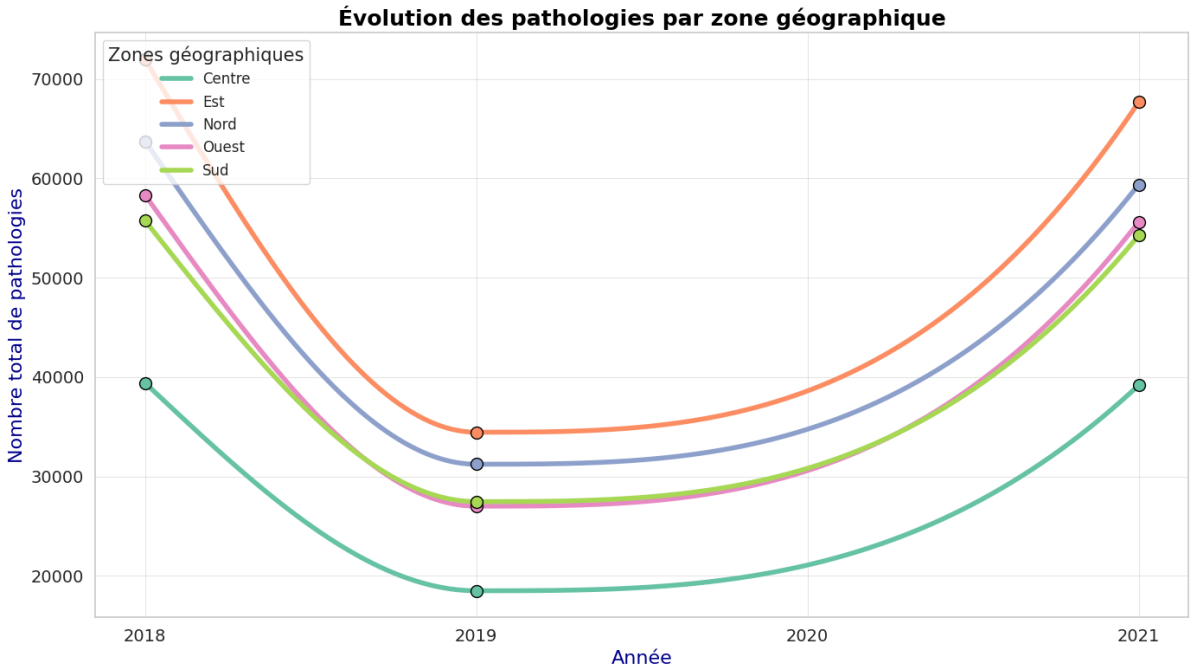


FIGURE 2.6: Evolution des pathologies par zones géographiques en France entre 2018 et 2021

Désormais nous nous sommes intéressés à l'évolution du nombre de pathologies en France de 2018 à 2021, avec chaque courbe représentant une région différente (Nord, Sud, Est, Ouest, Centre). Chaque région est associée à une couleur distincte, permettant de visualiser les tendances spécifiques à chaque zone géographique.

De 2018 à 2019 : On observe une baisse générale du nombre de pathologies dans toutes les régions. Cette diminution pourrait être liée à des campagnes de prévention efficaces, une meilleure prise en charge médicale, ou des facteurs environnementaux favorables. Cette diminution pourrait refléter des efforts de santé publique réussis, tels que des campagnes de sensibilisation, des améliorations dans les traitements, ou une meilleure gestion des maladies chroniques. Par exemple, des programmes de prévention du diabète ou des maladies cardiovasculaires pourraient avoir porté leurs fruits

À partir de 2019 : Les courbes repartent à la hausse, avec une augmentation notable en 2020. Cette tendance est visible dans toutes les régions, bien que certaines zones (comme le Nord ou l'Est) montrent une progression plus marquée que d'autres (comme le Sud ou le Centre).

Néanmoins, la cause principle de cette augmentation soudaine semble être l'Impact du COVID-19 : La pandémie a eu un effet significatif sur la santé publique. Les confinements et les perturbations des services de santé ont probablement retardé les diagnostics et les traitements, exacerbant les pathologies existantes. De plus, les patients atteints de maladies chroniques ont été plus vulnérables aux complications du COVID-19, ce qui a pu augmenter le nombre de cas recensés.

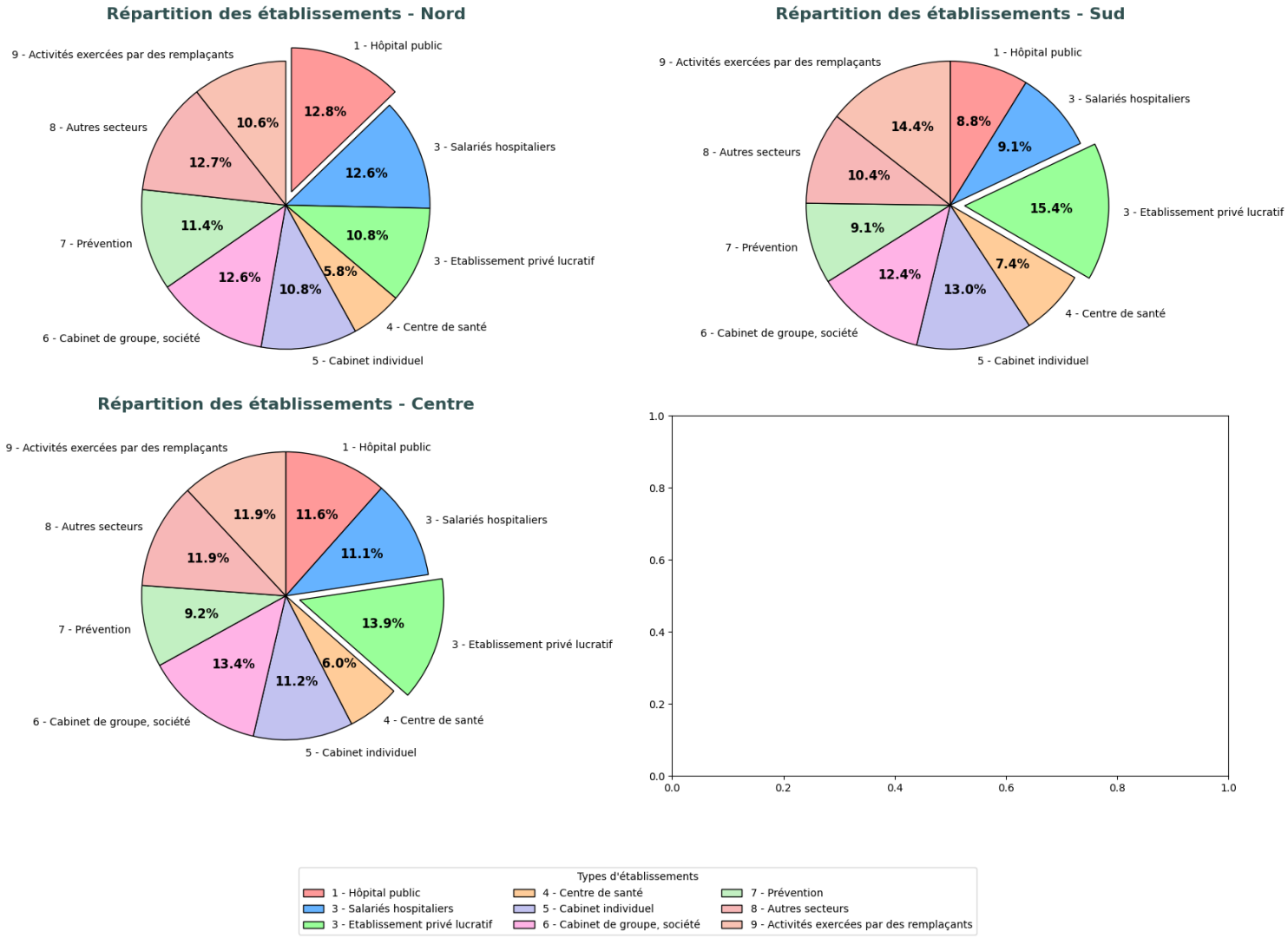
Ou surtout comme on peut voir avec la figure traitant des licensiés sportifs on peut constater une nette diminution des sportifs en France notamment accentuée après la pandémie ce qui pourrait potentiellement expliquer la réhausse des malaides chroniques. En effet, l'activité phyisque étant un facteur essentiel au bon fonctionnement du coprs humain.***

La crise sanitaire a également eu un impact sur la santé mentale, avec une augmentation des maladies psychiatriques et des troubles anxio-dépressifs, ce qui pourrait expliquer en partie la hausse des pathologies.

Les régions comme le Nord et l'Est, souvent confrontées à des défis socio-économiques et environnementaux plus importants (chômage, pollution, précarité), montrent une hausse plus marquée des pathologies. Cela souligne l'importance de cibler ces zones avec des politiques de santé adaptées.

Le Sud et le Centre, bien que touchés par la hausse, semblent moins affectés, probablement en raison de conditions de vie plus favorables (climat, alimentation, accès aux soins)

FIGURE 2.7: Répartition des services des établissements de santé par zone géographique en France



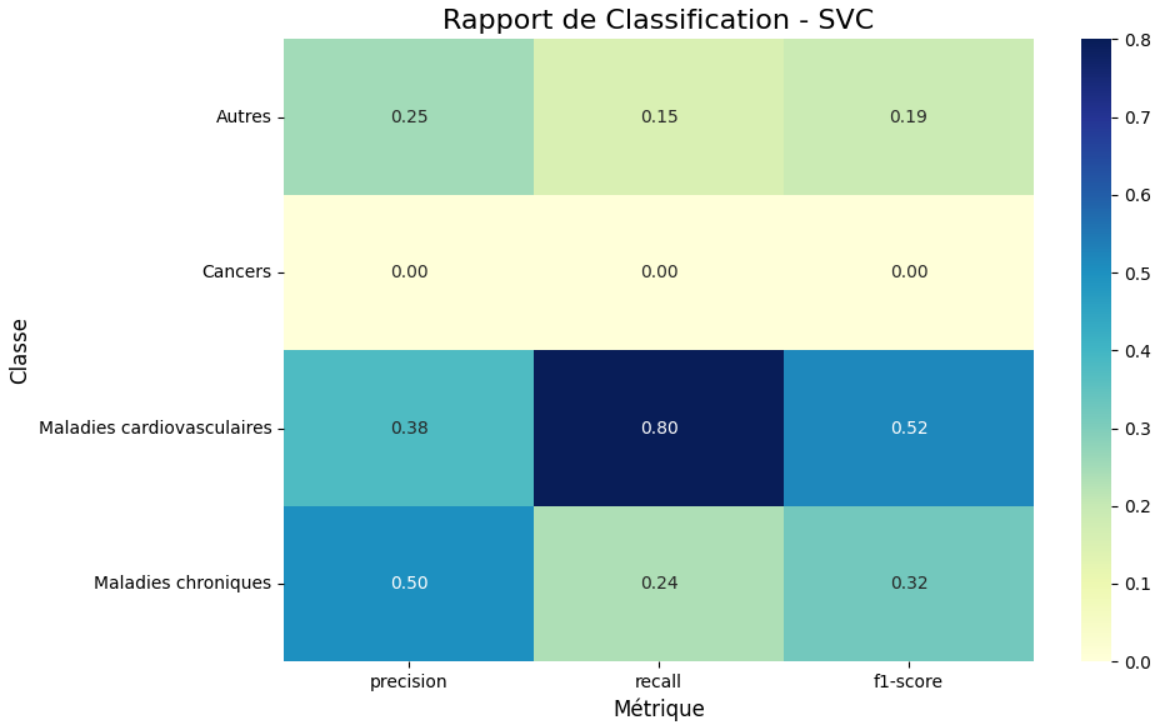
Chapitre 3

Prédiction de la maladie la plus fréquente dans un département grâce à de la classification

3.1 Premiere étude de modèles de machine Learning

Dans le cadre de notre projet, l'objectif principal était de prédire la maladie la plus fréquente dans chaque département à partir d'un ensemble de variables socio-économiques.Tout d 'abord, nous avons fusionné plusieurs bases de données en utlilisant le "code département"comme clé commune, afin de créer une base consolidée. Cette base comprenait des informations telles que le taux de chômage, le taux de pauvreté, le taux de logements sociaux, le nombre de licenciés sportifs ainsi que le nombre et le type d'établissements de santé (regroupés en trois catégories : Hôpitaux et Centres Hospitaliers, Établissements spécialisés, et Autres établissements médicaux et services). Par ailleurs, à partir de la base des effectifs de pathologies, nous avons identifié la maladie chronique la plus fréquente dans chaque département , que nous avons ajoutée comme variable cible dans notre jeu de données.Pour de faciliter l'apprentissage automatique, nous avons regroupé les maladies en catégories cliniques plus larges (comme Maladies vasculaires, Cancers, Troubles neurodégénératifs, etc.), ce qui permet une meilleure généralisation du modèle. Enfin ,les variables catégorielles ont été encodées à l'aide du LabelEncoder de scikit-learn.

Ensuite ,nous avons entraîné un modèle de classification SVC (Support Vector Classifier), en divisant les données en un ensemble d'entraînement (80%) et un ensemble de test (20%). Le modèle a été entraîné pour prédire la catégorie de la maladie dominante en se basant sur les caractéristiques socio-économiques du département . Enfin, nous avons évalué la performance du modèle en utilisant des métriques classiques telles que le rapport de classification et le score de précision (accuracy), pour estimer la capacité du modèle à prédire correctement sur des départements qui n'ont pas encore été observés.



RESULTAT GL0BAL : 0.37

Cependant, les résultats obtenus avec ce premier modèle SVC n'ont pas été satisfaisants, avec une précision globale de 0,38. La classe 0 présente des scores relativement faibles (précision : 0,25, rappel : 0,15), tandis que la classe 1, avec seulement 8 échantillons, affiche des scores nuls. La classe 2 montre des performances un peu meilleures (précision : 0,38, rappel : 0,80), bien que la précision reste limitée. La classe 3 montre un déséquilibre entre précision (0,50) et rappel (0,24). Les moyennes macro et pondérée confirment une faible capacité de généralisation. Ces résultats suggèrent que la taille réduite du jeu de données (290 lignes) limite l'apprentissage et la modélisation.

Il est possible que l'utilisation d'un jeu de données plus large améliore la performance du modèle. Afin de résoudre ce problème, nous avons décidé de dupliquer de manière raisonnée le jeu de données : au lieu de garder uniquement une seule maladie dominante par département, nous avons élargi notre cible en listant l'ensemble des maladies apparues dans chaque département. Cela nous a permis d'augmenter la quantité de données disponibles tout en assurant une cohérence des informations explicatives.

Cette nouvelle approche nous a permis d'enrichir et de mieux définir le problème, en associant à des caractéristiques socio-économiques chaque association département-maladie. De plus, pour améliorer la pertinence temporelle de notre modèle, nous avons pris en compte l'aspect chronologique dans notre analyse. Pour cela, nous avons exploité les données disponibles sur plusieurs années et avons calculé des moyennes temporelles pour les variables telles que le taux de pauvreté, le taux de chômage, le nombre de logements sociaux ou encore le nombre de licenciés sportifs. L'objectif était de refléter précisément l'évolution des conditions sociales dans chaque département, afin de fournir au modèle des données plus stables et représentatives pour expliquer de manière plus approfondie la présence de certaines pathologies sur le long terme.

Dans la phase de modélisation, nous avons regroupé les différentes pathologies en grandes catégories, telles que Cancers, Maladies chroniques, Maladies vasculaires ou Autres, afin de mieux structurer la classification. Ensuite, nous avons normalisé et intégré les variables explicatives (comme les taux de pauvreté, de chômage, le nombre d'hôpitaux, etc.) dans un modèle de type SVM linéaire

(LinearSVC). Compte tenu du déséquilibre important entre les classes, nous avons utilisé la méthode SMOTE pour rééchantillonner les données d'entraînement et équilibrer les classes minoritaires.

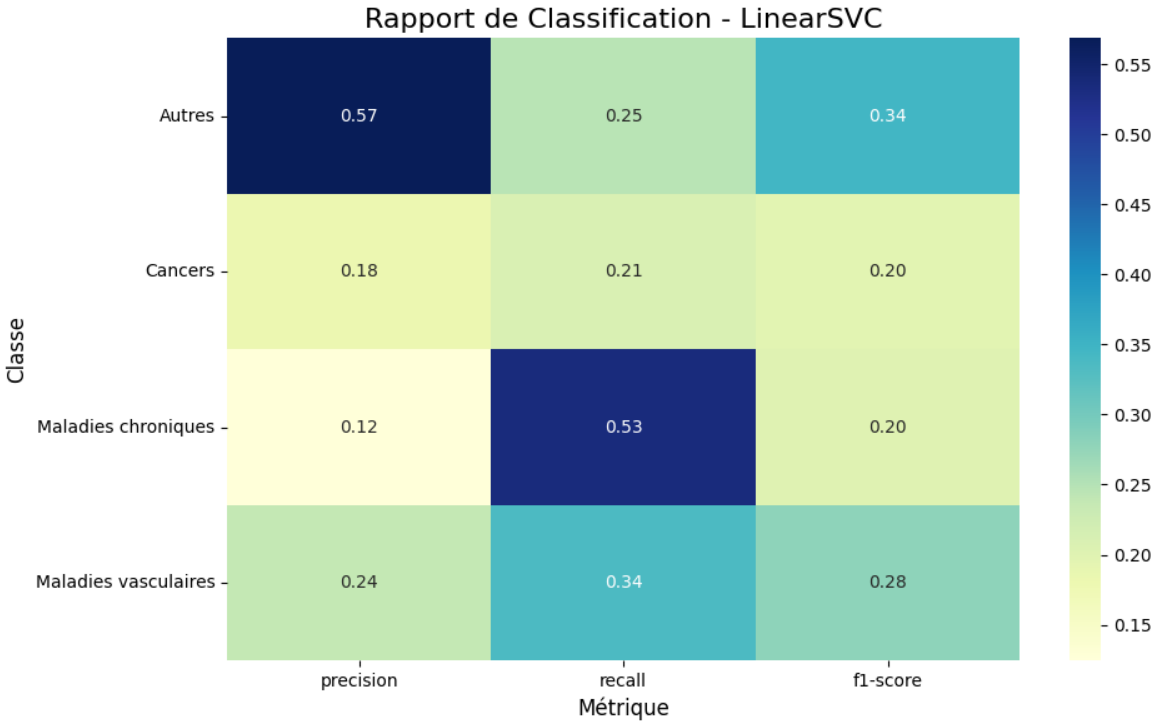


FIGURE 3.1: Matrice de confusion deuxième base

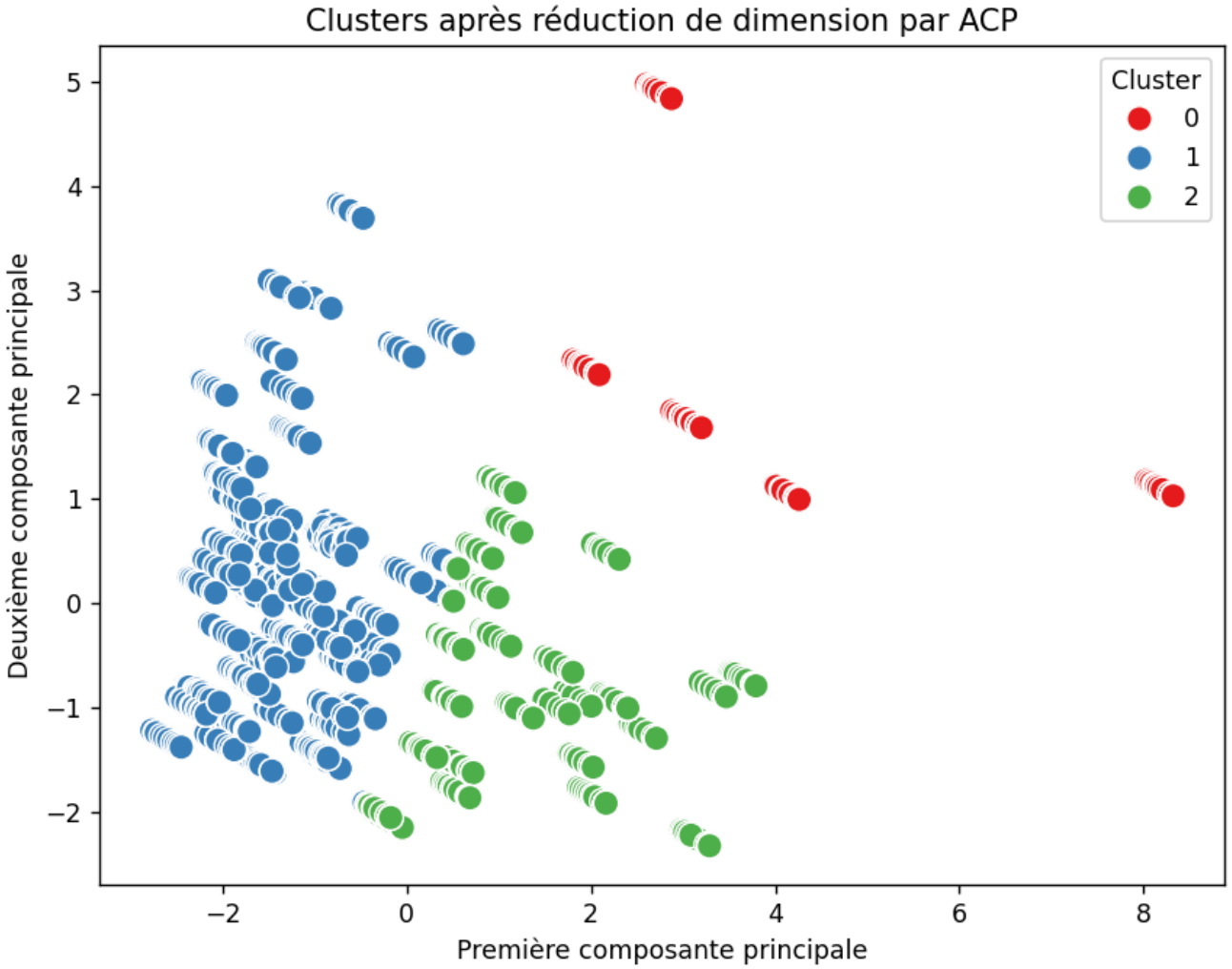
Malgré l'enrichissement de notre base de données, à la fois par la duplication structurée des lignes et par l'intégration d'informations chronologiques moyennées, les performances de notre modèle de machine learning restent modestes. En effet le résultat global du modèle a même diminué avec une valeur de **0.27**. En détails, on constate, une précision globale (accuracy) de 0.276 . En effet Les scores F1 par classe révèlent des disparités importantes : bien que la classe "Autres" obtienne un F1-score de 0.34, les performances sont nettement plus faibles pour les classes "Cancers" (0.20), "Maladies chroniques" (0.20) et "Maladies vasculaires" (0.28). Ces résultats montrent que le modèle a du mal à distinguer correctement entre les différentes catégories, notamment celles représentant des pathologies spécifiques. La classe "Maladies chroniques", malgré un bon rappel (0.53), souffre d'une très faible précision (0.12), ce qui signifiequ'il y a un grand nombre de faux positifs. Par ailleurs, les moyennes macro et pondérées des scores F1 (0.26 et 0.29, respectivement) confirment le déséquilibre du modèle, qui semble privilégier certaines classes majoritaires comme "Autres" au détriment des classes minoritaires.Il est donc nécessaire de procéder à des ajustements notamment par un rééquilibrage des classes ou une approche plus adaptée aux jeux de données déséquilibrés.

Il existe plusieurs facteurs qui peuvent expliquer cette situation.Tout d'abord, il est essentiel de noter que l'efficacité d'un modèle d'apprentissage automatique dépend non seulement de la quantité de données, mais surtout de la qualité et de la pertinence des variables utilisées. Dans notre cas, malgré la construction d'une base relativement volumineuse (plus de 500 000 lignes), les informations rassemblées restent principalement contextuelles : taux de chômage, taux de pauvreté, nombre d'établissements de santé, nombre de sportifs, etc. Ces variables, bien qu'utiles pour caractériser

l'environnement socio-économique d'un département , n'ont pas de lien direct ou spécifique avec l'apparition d'une maladie particulière. Elles agissent en tnat que facteurs indirects ou de fond, ce qui complique l'apprentissage de relations claires entre les entrées et la cible à prédire. En outre , le regroupement des maladies en grandes catégories, bien qu'important pour simplifier l'analyse, mais cela peut egalement réduire la capacité du modèle à saisir les nuances spécifiques de certaines pathologies. Deux départements peuvent présenter des indicateurs sociaux très proches, tout en ayant des profils de morbidité très différents, en raison de facteurs absents de notre base, comme les habitudes de vie, l'environnement, l'accès réel aux soins ou les politiques locales de santé. Le manque de données cliniques, comportementales ou environnementales précises a probablement limité la précision des prédictions De plus, la duplication des lignes, bien qu'elle ait permis d'augmenter artificiellement le nombre d'observations, n'a pas ajouté de diversité réelle à la base. En absence de nouvelles variables explicatives ou sources d'information, le modèle est resté exposé aux mêmes corrélations faibles, accentuées par la redondance des données. De plus,en calculant la moyenne des variables sociales sur plusieurs années on risque d'atténuer des évolutions importantes ou masquer des ruptures significatives, ce qui peut rendre certaines relations moins détectables pour l'algorithme.

Face à ces difficultés, nous avons pensé qu'une approche différente pourrait améliorer les résultats. En particulier, nous avons envisagé d'utiliser l'Analyse en Composantes Principales (ACP) pour regrouper les variables socio-économiques et de santé pertinentes. Lobjectif de l'ACP serait de réduire la complexité des données tout en conservant l'essentiel des informations clés , ce qui pourrait faciliter la compréhension des liens entre les variables.Ensuite, nous avons proposé d'explorer l'utilisation du clustering pour identifier des groupes de départements présentant des profils socio-économiques similaires. En identifiant ces clusters et en examinant la prévalence des différentes pathologies dans chaque groupe, nous pourrions mieux comprendre les facteurs qui influent sur la répartition des maladies. Cette approche permettrait d'affiner notre modèle, en se concentrant sur des groupes plus homogènes, et potentiellement d'améliorer les prédictions.

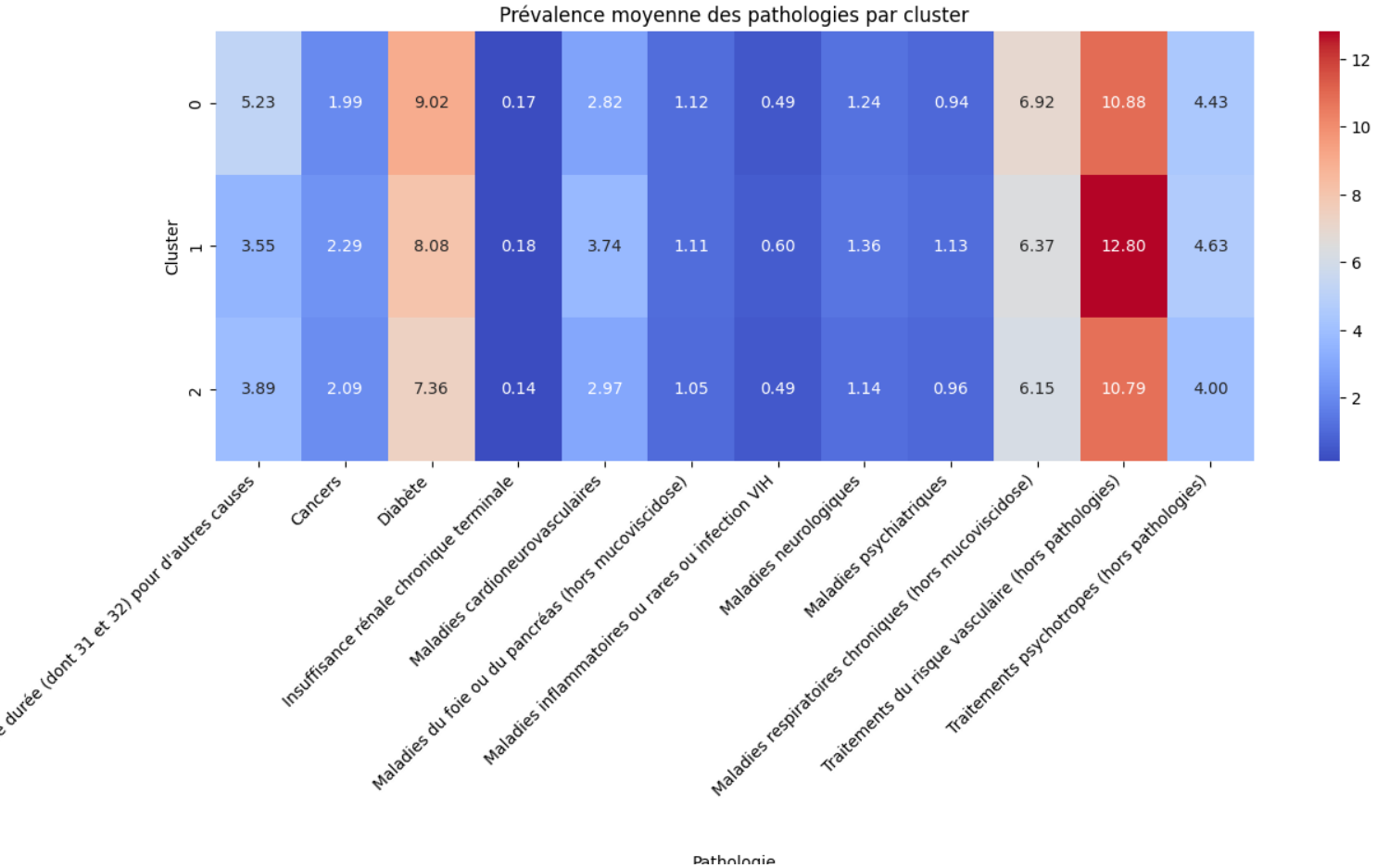
3.2 Etude et analyse des résultats obtenus grâce à une ACP + algorithme de clustering



Pour cette partie, nous nous sommes fixés comme objectif d'analyser la prévalence des groupements de maladie par clusters de départements. Etant donné que dans notre base de données relative a cette partie de Machine Learning nous avons 7 variables expliquatives nous avons décidé de réaliser une ACP (Analyse Composante Principales) afin de réduire le nombre de variables tout en gardant un maximum d'informations sur notre problématique. Pour ce faire nous avons dans un premier temps importe la nouvelle base de données relative à cette étude, puis nous avons selectionné comme variables explicatives celles exipliquées précedemment (tx_chomage, tx_pauvreté, nb_hopitaux. . ..). On a par la suite standardisé chaque variable afin de les disposer sur une même échelle pour accompagner notre modèle K-Means et l'ACP. Ensuite grâce à la fonction « pca », nous avons appliqué l'ACP en gardant 2 composantes principales en ayant pour but de résumé l'information des variables sur ces 2 axes pour avoir des clusters pour représentatifs. En effet, nous avions auparavant décidé de garder toutes les variables mais nous avions obtenu des clusters très confus avec beaucoup trop de variables pour tirer suffisamment d'intérprétations.

Une fois l'ACP réalisée, nous avons utilisé un algorithme de Machine Learning intitulé « K-Means » pour regrouper les données en 3 clusters, et puis nous avons visualisé les clusters après réduction de l'ACP. Une fois cette figure réalisée, nous avons illustré la distribution de la prévalence moyenne des pathologies grâce à une HeatMap qui affiche la valeur de chaque prévalence pour chaque groupement de maladie a chaque cluster.

Pour avoir une idée de la classification des departements dans les clusters, nous avons affiché sur Python le poids de chaque variables dans chaque axe principal. On observe que PC1 est majoritairement représenté grâce aux variables nombre d'hopitaux (0.45), établissements de santé (0.50), et nombre de sportifs (0.50) tandis que PC2 est représenté majoritairement par les variables tx_chomage et tx_pauvreté (respectivement 0.67, 0.7). On observe dans un premier temps le cluster 0 représenté par les départements Bouches du Rhone, Nord, Calais, Seine Saint Denis. On comprend que ces départements ont pu être regroupés notamment grâce à leur très forte précarité sociale. En effet la majorité de ces départements possède une valeur élévée de la 2nde composante principale, mais qu'ils ont tout de même un PC1 moyen (donc bien équipé en terme d'infrastructures de santé). Le cluster 1 quant à lui représente majoritairement des départements ruraux et montagneux (Jura, Aveyron, Cantal...) qui sont caractérisés par un PC1 faible donc peu d'équipements de santé, moins de services hospitaliers et une valeur de PC2 assez moyenne ce qui traduit majoritairement des taux de chomage ou pauvreté assez modéres (très aléatoires selon les zones). Enfin le cluster 2 représente globalement le reste des départements surtout urbains et périphériques. Ces derniers sont poussées par une valeur de PC1 élévé, donc bien entourés d'établissements de santé spécialisés ou hopitaux, et une valeur de PC2 qui reste assez similaire au cluster 1 donc assez modéré a faible (tout dépend encore une fois des zones des départements). Ce sont des zones assez attrayantes ou la mobilité et le dynamisme sont fortement représentés mais ce ne sont pas non plus des endroits ou la précarité sociale est urgente comme certains départements du cluster 0.



Maintenant, nous nous sommes intéressés à analyser la HeatMap pour essayer de répondre plus objectivement à notre problématique. On observe que le cluster 0 possèdes des valeurs de prévalence moyenne supérieures aux autres clusters pour 5 groupement de maladies (« Autres Causes », « Diabète », « Maladies respiratoires chroniques », « Maladies neurologiques ») sur 11 pathologies ce qui est assez élevé. On pourrait supposer qu'il semble exister un éventuel lien entre les départements fortement touchés par le chomage et la pauvreté et certains types de maladies spécifiques. Ce cluster affiche par exemple des prévalences assez élevées pour divers pathologies graves comme le diabète (9.02) et des maladies respiratoires (10.88), ce qui peut s'expliquer , des habitudes alimentaires déséquilibrées et des conditions sociales de logement dégradées. Néanmoins la prévalence des troubles psychiatriques semble assez équilibrée au vus des valeurs dans les autres clusters, mais cela n'empêche pas que la précarité sociale reste une urgence dans certaines zones "défavorisées" de ces cluster. En ce qui concerne le cluster 1, il possède lui aussi des prévalences élévés par rapport aux autres clusters et assez proches bien que ces départements de ce cluster ne rencontrent pas spécifiquement de précarité sociale et possèdent une densité médicale assez élevée En effet on peut voir cela notamment pour la prévalence des pathologies respiratoires (12.80) ou celle des maladies cardiovasculaires. Ainsi il ne semble pas vraiment y'avoir de lien dans le cadre de notre problématique. Etant donné que les variables taux_chomage, de pauvreté ou de logements sociaux ne semblent pas expliquer la cause de cette forte prévalence pour certaines pathologies, on pourrait expliquer cela tout simplement par vieillissement en milieu rural avec des maladies rencontrés dans ces zones-ci lié à l'âge et surtout pas assez deservies en équipement hospitalier (notamment par exemple la diagonale du vide). L'isolement peut parfois paraitre comme un problème occupant parfois une place "moins prédominante" au sein des autorités publiques et pourtant on y retrouve parfois de nombreux problèmes comme expliquées précedemment qui peuvent influencer l'apparition ou la prévention de certaines pathologies. Et enfin le cluster 2 se distingue par une bonne santé globale (comme nous l'avons vu dans les résultats de l'ACP) mais qui reste tout de même exposée a certaines maladies (du probablement à du stress rencontré ou certains facteurs que l'on a pas pris en compte dans notre étude). Le fait que la prévalence globale des maladies semble être moins dominante peut etre expliquée par le fait ces villes peuvent être « favorisées » mieux équipées en infrastructures de santé ce qui implique une meilleur prévention globale et donc moins de maladies sérieusement lourdes (Diabète, Cancers...). Néanmoins on peut notifier le fait que d'avoir des prévalences faibles n'implique pas automatiquement une absence de risques.

Chapitre 4

Conclusion, perspectives, limites et difficultées rencontrées

Durant tout ce rapport, nous nous sommes focalisés sur l'étude des liens entre les facteurs relatifs à l'environnement social et la prévalence des diverses pathologies. Grâce aux figures et à la partie de prévisualisation, nous avons pu constater qu'il semblait y avoir des liens entre certains facteurs et types de pathologies dans diverses zones en France. Ce fut le cas, par exemple, de la forte prévalence de cancers dans les départements du Nord et du Sud, qui étaient principalement associés à des taux de pauvreté et de chômage élevés. Bien que nous ayons, dans certains cas, eu des résultats cohérents, cela ne garantit en aucun cas l'existence de liens sûrs entre les facteurs sociaux et les maladies. En effet, nous avons vu grâce à la visualisation que le nombre de cas de certaines pathologies était totalement indépendant du poids de chaque facteur social dans un département associé. Cela nous amène directement à discuter les limites de cette étude. Tout d'abord, nous pouvons souligner le fait que nous ne disposions pas d'un nombre suffisant de variables affectant l'environnement social pour mener une analyse approfondie. Il a été très difficile de recueillir une multitude de facteurs sur des années cohérentes et sur l'ensemble du territoire français. De plus, la base de données sur les pathologies présentait des valeurs de prévalence très irrégulières d'un département à l'autre, parfois même d'une année à l'autre, ce qui a pu fortement influencer la qualité de nos résultats.

Ces limites nous amènent à évoquer les difficultés rencontrées durant ce projet. Notamment, il fut complexe d'obtenir des figures pertinentes en combinant la base des pathologies avec celle des facteurs sociaux. Cela explique pourquoi nous avons préféré séparer nos visualisations en deux parties : une dédiée aux pathologies par zone géographique, et une autre aux facteurs sociaux, afin de tenter d'observer visuellement des corrélations.

Concernant la partie Machine Learning, les résultats ont également été difficiles à interpréter de manière fiable. Comme nous l'avons mentionné, les variables liées à l'environnement social n'ont pas de lien direct ou évident avec la prévalence des pathologies. Il nous a donc semblé plus judicieux d'avoir recours à une ACP, suivie d'un algorithme de clustering (KMeans), afin d'obtenir une représentation plus claire des profils territoriaux associés aux différentes pathologies observées. Bien que nous ayons choisi un sujet d'étude difficile à analyser en terme de projection des donnéees brutes, ce projet nous a été bénéfique sur divers points notamment pour chaque difficulté rencontrée nous avons su trouver une solution alteranative pour pouvoir répondre du mieux possible à notre problématique que ce soit pour la partie Machine Learning ou Visualisation des Données.

Bibliographie

ChatGpt (correction orthographique du rapport + aide Python réalisation de figures)

Lien GitHub codes : Lien GitHub

https://data.ameli.fr/explore/dataset/effectifs/export/

https://www.observatoire-des-territoires.gouv.fr/nombre-de-licencies-sportifs

https://www.data.gouv.fr/fr/datasets/logements-et-logements-sociaux-dans-les-departements-1/

https://www.data.gouv.fr/fr/datasets/finess-extraction-du-fichier-des-etablissements/