Prédiction de l'environnement d'un quartier

Rapport de stage

Nelly Barret[1]
Encadrée par Fabien Duchateau[2] et Franck Favetta[2]

[1] Université Claude Bernard Lyon 1
prenom.nom@etu.univ-lyon1.fr
[2] LIRIS UMR 5205
prenom.nom@liris.cnrs.fr

Résumé Arriver dans une nouvelle ville suite à une mutation est toujours un défi ! En effet, il est courant d'arriver dans une ville que l'on ne connaît pas et la recherche d'un nouveau lieu de vie devient alors complexe. Proche des transports en commun pour certains, un cadre plus rural pour d'autres, plutôt animé pour les premiers, loin de l'agitation urbaine pour les autres : les critères pour choisir son futur quartier ne manquent pas. Dans ce rapport, nous présentons Predihood, un nouvel outil qui facilite la recherche et le choix d'un quartier.
Mots-clés : Apprentissage automatique, Apprentissage supervisé par classification, Nettoyage de données, Visualisation d'informations, Étude des quartiers

Abstract. Getting to a new city after a job transfer is always a challenge! We often arrive in a city that we don't know, thus finding the perfect living place becomes complex. Nearby public transport on one hand, a rural landscape on the other hand, an animated neighbourhood for some, far from urban hustle and bustle for others: there are many criteria for choosing your future neighbourhood. This report presents Predihood, a new tool which facilitates the search and the choice of a neighbourhood.
Keywords: Machine learning, Supervised learning by classification, Data cleaning, Information visualization, Neighbourhood study

Remerciements

Je souhaite ici remercier toutes les personnes qui ont contribué de près ou de loin à mon stage.

Dans un premier temps, mes remerciements vont vers mes deux maîtres de stage, Messieurs Fabien Duchateau et Franck Favetta, enseignants-chercheurs au LIRIS[3]. Sans eux, mon stage n'aurait pas pu voir le jour et ils ont su m'accompagner tout du long malgré leurs emplois du temps et le confinement dû à la crise sanitaire du Covid 19. C'est donc avec gratitude que je les remercie.

Mes remerciements se tournent également vers la start-up Home in Love avec qui le LIRIS collabore sur un projet éponyme. Pour avoir travaillé avec eux, je remercie tous les membres de ce projet :

- Nelly Duong, co-fondatrice de Home in Love,
- Loïc Bonneval, enseignant-chercheur en sciences sociales à Lyon 2,
- Behnaz Jullien, stagiaire en psychologie à Lyon 2,
- Wissame Laddada, post-doctorante en informatique à Lyon 1,
- ainsi que Ludovic Moncla, maître de conférences en informatique à l'INSA.

Je poursuis avec le LabEx IMU (Laboratoire d'Excellence, Intelligence des Mondes Urbains) qui a financé mon stage ainsi que Marion Nicolas, membre du LabEx, pour ses conseils en lien avec le projet et mon stage.

Bien entendu, mes remerciements vont au personnel du LIRIS pour son accueil, à l'Université Claude Bernard Lyon 1 pour mes cinq années passées ici, ainsi qu'à Madame Salima Hassas sans qui ce Master 2 Intelligence Artificielle ne serait pas le même.

[3] Laboratoire d'InfoRmatique en Imagerie et Systèmes d'information.

1 Introduction

Dans un monde où la quantité de données ne cesse d'augmenter, i.e. l'ère du Big Data, de nouveaux challenges apparaissent à différents niveaux [16]. En effet, les données ainsi que leurs sources sont en pleine explosion : les données sont estimées à plus de 40 zettaoctets en 2020 [10] et les sources ne cessent d'augmenter. De plus, les données produites sont de plus en plus variées : itinéraires routiers, données médicales, recommandation de produits, et tant d'autres. Plusieurs enjeux émergent de ces constats : stockage, gestion, traitement puis exploitation (e.g. la recommandation de contenu ou la génération de nouvelles connaissances). En tant que laboratoire de recherche en informatique, le LIRIS se positionne tout naturellement sur ces questions.

Le LIRIS est un laboratoire d'informatique (UMR 5205). Il dépend du CNRS, de l'INSA Lyon, de l'Université Claude Bernard Lyon 1, de l'Université Lumière Lyon 2 et de l'École Centrale de Lyon. Son organigramme se compose de 14 équipes réparties sur 6 pôles. Je suis attachée à l'équipe Bases de Données (BD), inscrite dans le pôle Sciences des Données. L'objectif de cette équipe est double : la conception de nouveaux modèles pour faire face à l'augmentation de la masse de données et le développement d'outils répondant à ces nouveaux besoins, e.g. l'accès, la diffusion et l'usage de ces données. Les projets entrepris par l'équipe BD sont financés par différentes sources, dont le LabEx IMU qui soutient principalement des projets pluridisciplinaires. Ce LabEx s'intéresse aux enjeux sociétaux autour de la ville, de l'urbain et de la métropolisation. L'un des projets soutenus, qui rassemble des chercheurs en sociologie et en informatique, est le projet Home in Love, dans lequel s'inscrit mon stage.

Étant en deuxième année de Master informatique en Intelligence Artificielle à l'Université Claude Bernard Lyon 1, je dois effectuer un stage de fin de cycle. Cette immersion dure 6 mois, de février à juillet, et a pour objectif de nous faire travailler sur des projets mêlant recherche et développement pour nous former à nos futurs emplois. Comme je souhaite poursuivre en thèse l'année prochaine et ayant effectué mon stage de Licence sur le projet Home in Love avec Fabien Duchateau et Franck Favetta, ces derniers m'ont proposé un sujet de stage sur ce même projet. Cette proposition regroupait toutes mes attentes : de la recherche, du développement, un sujet d'Intelligence Artificielle et un projet pluridisciplinaire.

Le projet Home in Love a pour objectif de faciliter la recherche immobilière, en particulier pour les personnes en mobilité salariale, e.g. lors d'une mutation professionnelle ou lors d'une alternance. Cette aide à la recherche immobilière se caractérise par la recommandation de quartiers et de logements pertinents. Cela nécessite, entre autres, de décrire simplement un quartier et de prendre en compte le profil des utilisateurs, i.e. leurs envies, leurs besoins et le style de vie qu'ils recherchent. Dans ce cadre, mon stage a pour objectif de prédire, par apprentissage supervisé, l'environnement d'un quartier pour aider les utilisateurs qui sont à la recherche d'un nouveau lieu de vie. Les quartiers peuvent être défi-

nis selon des centaines d'indicateurs, par exemple l'INSEE en fournit plus de 600 pour chacun. Seulement, autant d'indicateurs ne permettent pas de recommander efficacement et de manière interprétable ni d'avoir une vision globale de l'environnement d'un quartier. Par exemple connaître le nombre de restaurants et le nombre de supermarchés n'est pas suffisant pour caractériser l'environnement d'un quartier. En revanche savoir si un quartier est commerçant est important lors de la prise de décision dans ce contexte. Il est donc nécessaire de définir un nombre restreint d'indicateurs, que nous appellerons variables d'environnement (Annexe A). Elles ont été définies par les sociologues, sont au nombre de six et possèdent un nombre limité de valeurs. Par exemple, la variable d'environnement *paysage* a pour valeurs *urbanisé*, *espaces verts*, *arboré* ou *agricole* tandis que celle de la *classe sociale* s'étend de *populaire* à *supérieure*. L'approche Predihood proposée pendant mon stage consiste donc à collecter et intégrer des données ou indicateurs sur les quartiers, et à prédire les six variables d'environnement pour un quartier donné. Cette approche sera implémentée dans un outil avec une interface cartographique, qui permettra aux utilisateurs de comparer les quartiers entre eux grâce aux six variables d'environnement. Quatre enjeux émergent de ces objectifs. Le premier sera d'assurer une grande couverture géographique, i.e. de pouvoir prédire les variables d'environnement de n'importe quel quartier en France. Le second concerne la qualité de la prédiction, i.e. l'approche doit prédire correctement les valeurs des variables d'environnement qui serviront notamment de justification lors de l'étape de recommandation de quartiers aux clients Home in Love. Le troisième enjeu est la performance puisque l'approche Predihood doit calculer des résultats au moyen d'algorithmes efficients. Enfin, le dernier enjeu concerne la reproductibilité et la réutilisation, i.e. être capable de reproduire les résultats obtenus et de permettre à la communauté d'ajouter facilement de nouveaux algorithmes de prédiction.

Ce rapport abordera en premier l'état de l'art en Section 2 puis un aperçu de l'approche Predihood en Section 3. Ensuite nous détaillerons le processus de préparation des données en Section 4 puis la prédiction de l'environnement en Section 5. Enfin, nous présenterons la validation expérimentale en Section 6, une discussion en Section 7 avant de conclure et de proposer des perspectives en Section 8.

2 État de l'art

La recommandation de contenu est un domaine très étudié depuis les années 1990. De plus, les besoins sociétaux ont évolué et une tendance se dessine : générer beaucoup de données pour en extraire de l'information intéressante à recommander aux utilisateurs en fonction de leur profil, notamment grâce aux systèmes de recommandation. Par exemple, dans une médiathèque un utilisateur aura le choix parmi une centaine de disques alors que sur Internet, le nombre de musiques proposées est si important qu'il est nécessaire de recommander à l'utilisateur celles qu'il est susceptible d'écouter (e.g. Deezer offrait plus de 53

millions de titres en 2018). Ainsi, les systèmes de recommandation permettent de filtrer l'information pour ne présenter aux utilisateurs que des ressources pertinentes qui pourraient les intéresser [4]. L'enjeu qui nous intéresse dans le cadre du projet Home in Love est le besoin de recommandation de contenu, et plus précisément de quartiers.

Pour prédire l'environnement d'un quartier, il est nécessaire de définir la notion de quartier puis de récolter des données pour ensuite prédire les valeurs des six variables d'environnement.
Dès les années 1970 [11], les notions de voisinage et de quartier ont émergé et sont devenues une problématique commune à plusieurs domaines de recherche. Toutefois, la littérature montre qu'il existe plusieurs freins à une définition complète et réaliste de la notion de voisinage. Il est tout de même possible de s'appuyer sur les définitions proposées par les deux travaux suivants. Le premier [9] montre qu'il existe plusieurs types de quartiers, i.e. ceux définis par les institutions, ceux par les relations sociales et ceux physiquement. De plus, l'article note que la dimension d'un quartier est toute aussi importante que sa définition. Un second article [5], plus récent, appuie ces propos et relève quatre enjeux pour lever les freins à une définition satisfaisante : la description, la délimitation, la comparaison et l'évaluation des différentes notions de quartier. Lorsque la notion de quartier a été définie, le défi suivant relève du cas d'application, par exemple la comparaison ou la recommandation de quartiers, qui sont les deux approches principales pour l'aide à la recherche immobilière.
Dans l'**approche comparative**, les systèmes cartographient les différences pour que les utilisateurs puissent comparer eux-mêmes. Cette approche peut être étayée par plusieurs articles. Le premier [12] s'appuie sur les réseaux sociaux qui sont des sources riches d'informations à propos des lieux de vie d'un quartier, e.g. des tweets dans les centres commerciaux et des *checks-in* dans les restaurants ou les cafés. L'idée de cet article est de mesurer la similarité des quartiers en utilisant l'activité économique de chacun. Pour ce faire, Foursquare met à disposition des millions de points d'enregistrement, *check-in* en anglais, à travers le monde. La distance *Earth-mover* est ensuite utilisée pour détecter les quartiers similaires entre eux, i.e. les quartiers les plus similaires sont ceux qui subissent le plus faible effort de transformation entre eux et le quartier d'origine. Le second article [18] propose d'organiser les espaces urbains en quartiers en utilisant des données de géolocalisation fournies par les réseaux sociaux et les points d'enregistrement Foursquare. Les activités des habitants, e.g. aller au Starbucks, visiter une galerie d'art ou encore faire du sport, sont associées aux catégories Foursquare (plus de 300) pour en déduire les raisons de venir dans un quartier ainsi que les délimitations de celui-ci. De plus, la temporalité est inclue de manière à détecter les heures pleines et creuses de chaque quartier. Enfin, les activités sont catégorisées comme locales ou touristiques à l'aide d'un arbre de décision. Toujours dans une approche comparative, plusieurs interfaces ont été proposées à différentes échelles. L'interface **Better Life Index**, mise à disposition par l'OCDE, propose onze critères dans le but de classer et de comparer une quarantaine de pays par score. Cette plateforme se base principalement sur les statistiques de l'OCDE

et des Nations Unies. Ensuite, **GéoPortail** est une plateforme gouvernementale visant à faciliter la diffusion et la visualisation des données géographiques en France, et permettant entre autres la comparaison manuelle de villes grâce à l'affichage de données thématiques (e.g. les restaurants, les supermarchés et les écoles). GéoPortail se base sur plus de 80 sources dont OpenStreetMap et des sources ministérielles. **DataFrance** est une interface agrégeant des centaines d'indices en cinq catégories, dont l'éducation, les commerces et les loisirs, pour mesurer la qualité de vie d'une commune en France. Enfin, **KelQuartier** est un outil d'aide à la décision pour les particuliers qui souhaitent déménager. Six catégories, dont l'éducation, le logement et les habitants, sont proposées dans lesquelles il y a des indicateurs numériques plus détaillés, e.g. ville fleurie, nombre de commerces tous les 100 mètres et revenus moyens. KelQuartier propose aussi une comparaison avec les communes adjacentes. Leurs sources proviennent de plusieurs centaines d'administrations publiques françaises dont l'INSEE, les ministères et la Direction Générale des Impôts.
L'**approche de recommandation** permet de proposer aux utilisateurs uniquement des lieux qui pourraient les intéresser. Elle peut être étayée par plusieurs articles et interfaces. Un premier article [14] utilise les réseaux sociaux géolocalisés et les points d'enregistrement Foursquare pour la recommandation de quartiers. Cette recommandation, via l'algorithme Instance-Region Neighborhood Matrix Factorization, est proposée à deux niveaux : au niveau du quartier (i.e. un utilisateur est susceptible d'apprécier les quartiers voisins car ils partagent des propriétés similaires) et au niveau de la région e.g. affaires, scolarité et loisirs. Un second article [1] propose de la recommandation de quartiers via des mesures de similarité et des algorithmes de regroupement (*clustering* en anglais). De plus, l'article mentionne un outil, nommé VizLiris, qui apporte une interface cartographique à la recommandation de quartiers en France. Ce prototype propose deux fonctionnalités : la détection des meilleurs quartiers similaires à un quartier d'origine ainsi que le regroupement d'une zone géographique par type de quartiers grâce à des méthodes de *clustering*. Comme VizLiris recommande des quartiers pertinents à partir d'un quartier d'origine, il est nécessaire que ce quartier soit en France (ce qui n'est pas toujours le cas, par exemple avec les habitants étrangers) et que celui-ci plaise à la personne, ce qui n'est pas forcément vrai pour tous les critères. Une seconde approche [17], proposée par une équipe coréenne, montre l'utilisation du raisonnement par cas dans le cadre de la recommandation de logements. Les utilisateurs peuvent émettre trois contraintes : la localisation, le prix et l'unité d'habitation (e.g. le nombre de salles de bains ou la surface de la cuisine). Enfin, le projet Livehoods [7] découpe les villes en fonction de leurs dynamiques sur la base des données générées via les médias sociaux par les habitants.

Le sujet de mon stage apporte une **approche différente** de ces travaux existants sur deux points majeurs. Le premier concerne la description d'un quartier en termes de lieu de vie. C'est une tâche complexe car le lien entre les nombreux indicateurs disponibles (e.g. indicateurs INSEE, points d'intérêt, statistiques) et l'environnement d'un quartier (qui relève du domaine du sensible) n'est pas fais-

able manuellement. Cette subjectivité nécessite des compétences en sociologie, e.g. pour définir l'ambiance ou le paysage, et un cadre pluridisciplinaire. Le second point est la qualification de l'environnement d'un quartier (à travers les variables d'environnement) tandis que les travaux existants caractérisent la qualité de vie (e.g. scolarité des enfants, sécurité, proche des restaurants). La proposition d'un nombre restreint de variables d'environnement permet aussi de simplifier la justification de la recommandation. De nombreux projets, comme Livehoods et Hoodsquare, utilisent les réseaux sociaux géolocalisés pour déterminer par exemple les dynamiques des quartiers. Les données générées par ces réseaux sont complexes à exploiter, notamment car les données ne sont pas disponibles uniformément (en particulier dans les campagnes) et sont soumises à un biais de par la population qui utilise ces réseaux sociaux. C'est pourquoi nous avons choisi une approche pluridisciplinaire avec des sociologues, car c'est un gage de qualité. Enfin, l'outil Predihood sera fonctionnel au niveau national tout en respectant un partitionnement fin de la taille d'un quartier, ce qui est important dans le contexte des mutations professionnelles.

3 Présentation de l'approche Predihood

L'approche Predihood consiste en la simplification de la qualification de l'environnement d'un quartier. Pour ce faire, cette approche propose la définition de six variables d'environnement, la sélection des critères importants pour la prédiction de ces six variables ainsi que le déploiement de l'approche dans un nouvel outil, nommé Predihood. La conception, le développement et le déploiement du projet Predihood suivent le modèle de développement CRISP-DM, *Cross-Industry Standard Process for Data Mining* [15]. C'est un modèle standard open-source qui détaille les processus pour le développement d'outils en data mining. La Figure 1 illustre le déroulement du développement de Predihood (deuxième ligne) en parallèle de celui de ce modèle (première ligne).

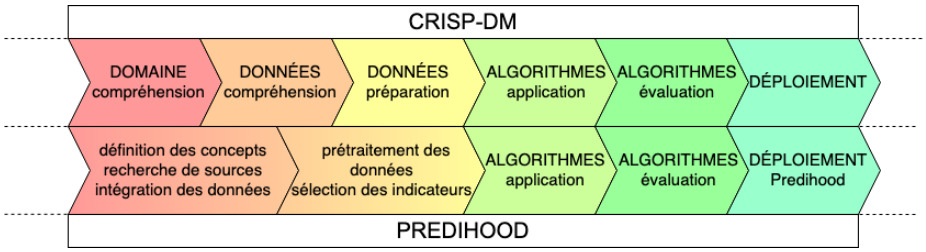

Fig. 1: Processus pour le développement de Predihood en comparaison du modèle CRISP-DM.

Afin de prédire les variables d'environnement dans le but d'aider à la recherche de quartiers avec l'approche Predihood, il est nécessaire de définir les concepts utilisés, de rechercher des sources pertinentes et de les intégrer dans un outil unique.

Ces trois étapes[4] correspondent au premier processus de la seconde ligne de la Figure 1. Une fois l'intégration de données réalisée, il faut préparer les données afin de pouvoir les utiliser dans les étapes suivantes puis sélectionner les indicateurs pertinents pour la prédiction. Ensuite, les algorithmes d'apprentissage supervisés peuvent prédire les six variables à partir de la sélection des indicateurs et des données nettoyées avant d'être évalués grâce à des métriques de performance. Enfin, le dernier processus correspond au déploiement de l'approche dans une interface web nommée Predihood.

La Figure 2 illustre en détail les étapes de l'approche Predihood. Les trois premières étapes, i.e. la définition des concepts, la recherche de sources, l'intégration et le prétraitement de celles-ci sont présentées en Section 4. La sélection des indicateurs, réalisée à partir de la combinaison de trois techniques dédiées, est présentée en Section 5. Les trois étapes suivantes, i.e. la création des jeux de données, l'apprentissage et la prédiction, correspondent au troisième processus du cycle CRISP-DM. Une fois le processus d'apprentissage terminé, les prédictions sont évaluées, comme présenté en Section 6. L'interface Predihood propose une interface cartographique facilitant la comparaison de quartiers et une interface de paramétrage permettant de configurer et tester différents algorithmes.

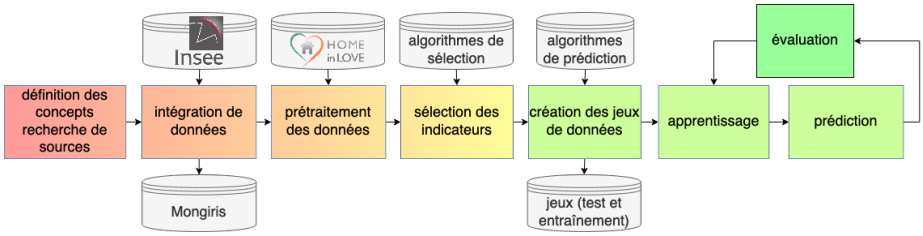

Fig. 2: Illustration des processus de l'approche Predihood.

4 Préparation des données

Dans les travaux de Data Science, il est courant d'estimer que 80% du travail consiste à rechercher, nettoyer, préparer et intégrer les données [8]. Pour l'approche Predihood, il est d'abord nécessaire de définir les concepts de quartier et de variable d'environnement. Ensuite, il faut rechercher des sources qui seront potentiellement utiles à la prédiction. Une fois les sources trouvées, il faut les intégrer de manière à unifier l'accès à ces différentes sources, ce qui correspond au processus d'intégration de données. Enfin, il est nécessaire de préparer les données avant de les utiliser, notamment car celles-ci contiennent des incohérences et des incomplétudes. Ces quatre processus, détaillés ci-après, correspondent au processus général de la préparation de données.

[4] Elles ont été réalisées en amont de mon stage et j'ai contribué à la recherche de sources et à l'intégration des données lors de mon stage de Licence.

4.1 Définition des concepts

Il est nécessaire de choisir une définition adéquate pour les notions de **quartier** et de **variable d'environnement**. L'unité choisie pour le quartier est l'IRIS[5] car elle permet un découpage fin et fiable puisque le fournisseur est l'INSEE[6]. Un IRIS correspond à une zone d'environ 2.000 habitants. En découpant toute la France ainsi, on obtient un maillage d'approximativement 50.000 IRIS (Annexe B). Dans les centres-villes, les IRIS sont de petite taille, e.g. Villeurbanne est découpé en plus de 40 IRIS, tandis qu'une ville en périphérie urbaine telle que les communes des Monts du Lyonnais n'en comporte que quelques-uns. Les lieux éloignés des centres urbains ne comportent qu'un seul IRIS pour toute la commune. Chaque IRIS possède plus de 600 indicateurs bruts[7], e.g. le nombre de restaurants, le nombre de logements construits avant 1950 ou encore le nombre d'habitants par catégorie socio-professionnelle. Afin de qualifier de manière simple l'environnement d'un quartier, six variables d'environnement ont été définies à l'aide des sociologues (Annexe A). Le *type de bâtiments* représente le type majoritaire de bâtiments dans l'IRIS. L'*usage* représente l'activité économique majoritaire. La variable *paysage* définit le caractère naturel, arboré des espaces. Les variables *position morphologique* et *position géographique* représentent respectivement la proximité au centre de l'agglomération contenant l'IRIS et la direction de l'IRIS par rapport à ce centre. Enfin, la *classe sociale* correspond à la richesse des habitants de l'IRIS. Dans la suite, nous utiliserons les termes de quartier et d'IRIS comme synonymes.

4.2 Recherche de sources

Après avoir défini les concepts, il faut constituer une base de données regroupant des données sur les quartiers et potentiellement utiles à la prédiction. Pour peupler cette base de données, il est donc nécessaire de rechercher plusieurs sources pertinentes. Pour l'instant, les trois sources suivantes sont considérées :

1. **IRIS.** Les IRIS sont nos objets d'étude et l'INSEE fournit une grande quantité de données les concernant. Elle met à disposition 647 indicateurs pour chaque IRIS. Parmi ces indicateurs, 17 sont des indicateurs de description (e.g. identifiant, code postal, nom de l'IRIS, ...) et seront utiles pour la visualisation cartographique. Les indicateurs restants permettent de quantifier l'environnement de l'IRIS et sont de différents types : quantités (e.g. nombre de restaurants) et quantités unitaires (e.g. revenu moyen), coefficients (e.g. coefficient de Gini), pourcentages (e.g. pourcentage de chômeurs) ou encore des chaînes de caractères (e.g. le type d'IRIS). Tous ces indicateurs sont livrés dans beaucoup de fichiers Excel dont les jointures ne sont pas toujours évidentes, comme expliqué en Section 4.3.

[5] Ilots Regroupés pour l'Information Statistique.
[6] Institut National de la Statistique et des Études Économiques.
[7] Les indicateurs bruts correspondent aux indicateurs fournis par l'INSEE.

2. **Données spatiales.** Comme les IRIS sont des entités spatiales, il est intéressant d'en exploiter les propriétés. Chaque IRIS est défini par sa géométrie, i.e. un ensemble de points qui trace ses contours. De cette géométrie, il est possible de calculer la surface de chaque IRIS. Cela permettra, par la suite, de calculer la densité de population de celui-ci et ainsi de le comparer aux autres.
3. **Home in Love.** Puisque la prédiction des variables d'environnement est effectuée par apprentissage supervisé, il est nécessaire d'avoir des données reflétant la réalité-terrain, i.e. des données annotées et vérifiées. Ces données sont les dossiers anonymisés des clients fournis par Home in Love. Pour chaque client, l'adresse de l'ancien lieu de vie et celle du nouveau suite à la mobilité sont fournies (Annexe C). Chaque adresse est ensuite mise en correspondance avec l'IRIS auquel elle appartient. Les dossiers sont stockés sous la forme d'un document Excel.

4.3 Intégration des données

Après avoir identifié plusieurs sources de données pertinentes, il faut intégrer les données spatiales dans une base de données unique. Cette base de données, nommée **Mongiris**, intègre de manière unifiée les IRIS et leurs données spatiales sélectionnés lors de l'étape précédente[8]. Mongiris est une base de données orientée document en MongoDB et contient trois collections : *collindic* pour les indicateurs, *collsources* pour les sources de données de l'INSEE et *colliris* pour les IRIS.

1. ***Collindic.*** Elle intègre les 647 indicateurs bruts fournis par l'INSEE. Chacun possède un identifiant, un libellé court (e.g. "P14_POP0610") et un libellé long (e.g. "Pop 6-10 ans en 2014 (princ)"). De plus chaque indicateur possède la liste des fichiers INSEE dans lesquels il apparaît.
2. ***Collsources.*** Elle regroupe les informations des fichiers INSEE intégrés, i.e. leur chemin, leur titre, et leur date de publication.
3. ***Colliris.*** Cette collection intègre les 49 404 IRIS de la France avec leurs données. Ces données correspondent aux indicateurs de description, à la géométrie, aux indicateurs bruts fournis par l'INSEE et à un nouvel indicateur nommé *grouped_ indicators*. La **géométrie** correspond à la liste des points formant le contour de l'IRIS et est disponible via l'IGN[9]. Les sources de l'IGN ont été converties en GeoJSON, format permettant de stocker des données géospatiales au format JSON. Quelques contours ont dû être corrigés, e.g. lors de l'intersection des lignes d'un polygone. Les **indicateurs bruts** sont répartis par thème dans 16 fichiers Excel, e.g. un fichier pour les données démographiques (Annexe D), un pour les loisirs, un pour les

[8] La majorité des données fournies par Home in Love concerne les clients. Ces données ne sont donc pas directement intégrées à Mongiris, qui doit fournir une interface générique pour l'accès aux données géographiques en France.

[9] Institut Géographique National.

commerces. Afin d'avoir une vue unifiée de tous ces indicateurs, il a fallu opérer plusieurs traitements sur ces données en considérant les verrous suivants : différentes interprétations d'un même concept, hétérogénéité des labels, regroupement ou partitionnement de certains IRIS. C'est pourquoi un appariement est nécessaire au niveau des schémas, i.e. des propriétés utilisées par les fournisseurs de données, et au niveau des entités, i.e. des IRIS. Enfin, l'indicateur ***grouped_ indicators***[10], qui permet de représenter un IRIS avec un plus haut niveau d'abstraction, a été calculé pour chaque IRIS.

Le processus d'intégration de données a donc permis de créer Mongiris, une base de données contenant les IRIS avec leurs indicateurs bruts et leurs indicateurs regroupés, les indicateurs INSEE et les sources utilisées. Au total, Mongiris intègre 49 404 IRIS et 647 indicateurs bruts. Une API Python a également été développée pour faciliter l'interaction avec la base de données.

4.4 Prétraitement des données

Il n'est pas possible d'utiliser directement les données fournies par Home in Love ni les indicateurs bruts fournis par l'INSEE sans nettoyage et prétraitement préliminaires. En effet, plusieurs problèmes apparaissent dans ces données, notamment des incohérences et des données manquantes.

Données Home in Love Les dossiers clients fournis par Home in Love sont constitués d'une trentaine de champs (e.g. identifiant d'anonymat, adresse de départ, revenu mensuel, texte relatant du contexte de la mutation) mais seulement deux sont utilisés pendant le prétraitement : l'adresse de départ (ancien lieu de vie) et l'adresse d'arrivée (adresse du nouveau lieu de vie après mutation). Sur les 310 dossiers clients fournis par Home in Love, 11 adresses de départ sont étrangères et 23 ne sont pas renseignées ou non géolocalisables. Ces adresses ne peuvent donc pas être converties en IRIS. Il reste ainsi 276 lieux à expertiser. Ces lieux ont été annotés par les sociologues, i.e. ils ont attribué des valeurs aux six variables d'environnement (Annexe C). Cette expertise manuelle requiert entre une à deux heures par IRIS. Parmi ces IRIS expertisés, il existe des incohérences dont des valeurs mal orthographiées ou incorrectes (e.g. *Oui* pour la classe sociale, *Location* pour le type de bâtiments) pour les variables d'environnement et des noms de colonnes non formatés. Pour remédier au problème des colonnes, celles-ci ont été renommées. Pour résoudre les erreurs dans les valeurs, un script a été développé pour les détecter. Bien qu'une correction automatique aurait été possible, nous avons opté pour une validation par un expert afin de s'assurer de la bonne qualité du jeu de données. Nous présentons les différentes étapes algorithmiques de ce script :

[10] Les indicateurs regroupés ont été défini lors de mon stage de Licence (Annexe E).

1. Récupération de l'ensemble des valeurs utilisées pour chaque variable d'environnement.
2. Construction d'un dictionnaire pour regrouper par similarité les versions de chaque valeur avec comme mesure de similarité Levenshtein (Annexe F).
3. Intervention de l'expert pour valider les valeurs acceptées pour chaque variable d'environnement.
4. Remplacement des valeurs erronées du jeu de données par les valeurs choisies par l'expert lors de l'étape précédente.
5. Traitement manuel des valeurs aberrantes restantes par suppression selon une liste définie manuellement. Par exemple, la valeur *Oui* ne correspond à aucune des valeurs définies par les sociologues et les valeurs *Moyen-sup* et *Moyenne-sup* ont deux opérations de différence et ne sont donc pas détectées par la mesure de similarité qui a un seuil fixé à une différence.

Au total, 49 valeurs ont été corrigées, ce qui correspond à un taux d'erreur de 3%. Pour limiter ces problèmes par la suite, un formulaire Excel a été proposé à Home in Love afin d'uniformiser les saisies.

IRIS Deux méthodes de prétraitement sont appliquées sur les indicateurs des IRIS expertisés : la normalisation et le traitement des valeurs inconnues. Il est important de normaliser les données car cela permet de pouvoir comparer équitablement les IRIS entre eux. Dans notre contexte, deux choix sont possibles pour la normalisation : la population ou la densité de population. La **normalisation** retenue est celle de la densité de population, i.e. le nombre d'habitants divisé par la surface en km^2. Ce choix est motivé par le fait qu'il est important de prendre en compte la surface de l'IRIS et pas uniquement sa population. Par exemple, un IRIS d'une grande agglomération qui a plusieurs dizaines de restaurants est considéré commerçant, tandis qu'un IRIS dans cette même ville ayant seulement 4 ou 5 restaurants ne le sera pas. En revanche un IRIS dans une petite ville qui a 4 ou 5 restaurants est commerçant. La normalisation par la densité de population permet de pallier en partie ce problème. Le **traitement des valeurs inconnues** est essentiel dans le prétraitement des données. Les valeurs non renseignées par l'INSEE doivent être complétées avant d'être utilisées par les algorithmes de prédiction. Il n'est pas possible de remplacer ces valeurs par des zéros, puisque le zéro a une valeur propre, i.e. que l'indicateur a été identifié comme ne contenant pas de donnée tandis qu'une valeur nulle correspond à l'absence de données. Le choix actuellement retenu est celui de la médiane des valeurs de l'indicateur car elle est moins sensible aux valeurs extrêmes que la moyenne. Ce choix comporte un biais lorsqu'une valeur est manquante pour les IRIS ruraux et que la plupart des IRIS ayant une valeur sont urbains par exemple. La correction de ce biais nécessite un travail important qui fait partie des perspectives.

5 Prédiction de l'environnement d'un quartier

Une fois la préparation des données terminée, la prédiction de l'environnement peut être mise en place. Elle regroupe la sélection des indicateurs, l'analyse des données expertisées et enfin la mise en place dans l'outil Predihood.

5.1 Sélection des indicateurs

Les indicateurs INSEE étant très nombreux, il est nécessaire de faire une étape de sélection. Celle-ci permettra de sélectionner un nombre restreint d'indicateurs pertinents pour la prédiction des variables d'environnement. La sélection de variables a plusieurs avantages dont des meilleures performances et l'explicabilité. Cette sélection se base en premier sur le filtrage des indicateurs non utiles à la prédiction puis sur la sélection de sous-ensembles d'indicateurs pertinents.

Filtrage des indicateurs Le filtrage permet de retirer les indicateurs non utiles à la prédiction. Tout d'abord, il y a 17 indicateurs descriptifs tels que le nom de l'IRIS, son identifiant et son code postal. Ces indicateurs seront utiles pour la visualisation dans Predihood mais ne sont pas utiles pour la prédiction des variables d'environnement donc ils sont retirés. Ensuite, certains indicateurs paraissent trop spécifiques pour la prédiction et sont donc aussi retirés. Par exemple, il y a trois indicateurs se rapportant aux courts de tennis : un pour le nombre de terrains de tennis, un second pour le nombre de courts de tennis et un dernier pour le nombre de courts de tennis couverts. Les indicateurs trop spécifiques ont été définis manuellement (Annexe G), et sont au nombre de 208. Enfin, 59 indicateurs nuls, i.e. pour lesquels il n'y a aucune valeur pour aucun IRIS expertisé, sont retirés. Au final, il reste 363 indicateurs, soit 55% des indicateurs fournis par l'INSEE.

Sélection des sous-ensembles Après avoir filtré les indicateurs, il paraît pertinent de sélectionner des sous-ensembles parmi ces 363 indicateurs, notamment pour des raisons d'explicabilité en plus du fait qu'il soit commun de sélectionner un ensemble restreint d'indicateurs. Chaque variable d'environnement ayant ses valeurs, il est nécessaire de sélectionner un sous-ensemble d'indicateurs par variable. En effet, les indicateurs pour la prédiction du paysage d'un quartier ne sont intuitivement pas les mêmes que ceux pour prédire la classe sociale. Il faut aussi définir la taille de ces sous-ensembles. Selon Lillesland et al. [13], la taille des données d'entraînement doit être comprise entre $10p$ et $100p$ où p est le nombre d'indicateurs sélectionnés. Avec 276 IRIS expertisés, la taille du sous-ensemble doit être entre 3 et 27 indicateurs. Afin d'explorer plus de possibilités quant à la validation expérimentale, sept sous-ensembles, que nous appellerons **listes**, seront définis pour chaque variable d'environnement. Ces sous-ensembles sont respectivement de taille 10, 20, 30, 40, 50, 75 et 100 indicateurs chacun. Autrement

dit, chaque liste L_v^k est un sous-ensemble des k indicateurs les plus pertinents pour la variable d'environnement v. Afin de créer ces listes, nous avons construit **un algorithme à partir de trois techniques dédiées** : une matrice de corrélation, les forêts aléatoires (*Random Forest Classifier*) et un méta-estimateur (*Extra Tree Classifier*). L'idée de cette sélection, détaillée dans l'Algorithme 1, est de retirer les indicateurs totalement corrélés grâce à la matrice de corrélation (lignes 1 et 2) et de combiner deux algorithmes qui ordonnent les variables par pertinence pour la prédiction (lignes 5 à 14).

Algorithme 1 : Sélection des indicateurs pertinents pour la prédiction

Entrée : liste d'indicateurs $\mathcal{I}$, liste des variables d'environnement $\mathcal{V}$
Sortie : listes d'indicateurs L_v^k

1 $C \longleftarrow$ matriceCorrelation($\mathcal{I}$).where(corr = 1);
2 $\mathcal{I} \longleftarrow \mathcal{I} - C$;
3 **for** $k \in [10, 20, 30, 40, 50, 75, 100]$ **do**
4 **for** $v \in \mathcal{V}$ **do**
5 $L_v \longleftarrow \emptyset$;
6 $F_v^{ET} \longleftarrow$ top-k(ET.rank_features($\mathcal{I}$), k);
7 $F_v^{RF} \longleftarrow$ top-k(RF.rank_features($\mathcal{I}$), k);
8 $F \longleftarrow F_v^{ET} \cup F_v^{RF}$;
9 **for** $f \in F$ **do**
10 $p_f \longleftarrow$ parent(f);
11 **if** $p_f \in F$ **then**
12 $p_f.score \longleftarrow p_f.score + f.score$;
13 $F \longleftarrow F - \{f\}$;
14 $L_v^k \longleftarrow F$;

La **matrice de corrélation** est une technique permettant de visualiser la dépendance entre plusieurs variables. Appliqué à notre problème, tracer la matrice de corrélation revient à tracer les dépendances entre les indicateurs INSEE filtrés (Figure 3) grâce à la méthode de Spearman. Plus un point est clair, plus les deux indicateurs sont corrélés.

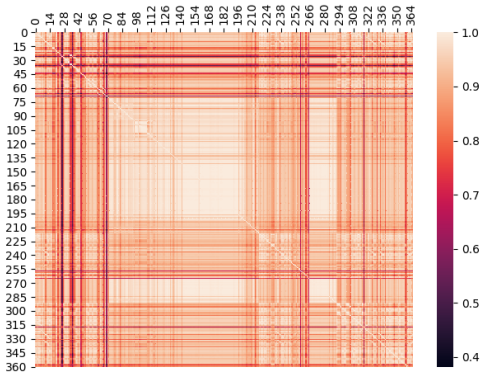

Fig. 3: Matrice de corrélation pour les 363 indicateurs (après le filtrage).

Pour chaque paire d'indicateurs totalement corrélés (i.e. que leur corrélation vaut 1), un seul des deux indicateurs est conservé (celui se trouvant dans la demi matrice inférieure). La matrice de corrélation permet de retirer 40 indicateurs. L'étape suivante est la sélection des indicateurs pertinents sur l'ensemble restant d'indicateurs grâce à la combinaison de méthodes dédiées. Plusieurs méthodes ont été testées (dont l'analyse en composante principale et l'algorithme SelectKbest) et les deux algorithmes retenus sont **Random Forest et Extra Tree Classifier**[11]. Ces deux algorithmes génèrent chacun une liste des indicateurs ordonnés par ordre d'importance pour la prédiction (lignes 6 et 7). Pour les deux listes générées, les k premiers éléments sont sélectionnés où k est la taille de la liste souhaitée, et F résulte de l'union de ces deux listes (ligne 8). Enfin, pour conserver la diversité des indicateurs fournis par l'INSEE, les indicateurs dont le parent a été sélectionné sont intégrés dans leur parent, i.e. que le score du fils est ajouté à celui de son parent et le fils est ensuite retiré (lignes 9 à 13). Par exemple, si les indicateurs du nombre de personnes, du nombre d'hommes et du nombre de femmes en France sont sélectionnés, alors les scores des indicateurs du nombre d'hommes et de femmes seront ajoutés au score du nombre de personnes en France (et ils seront supprimés). Pour ce faire, une hiérarchie de la totalité des indicateurs bruts a été construite manuellement (Annexe H).

5.2 Analyse des IRIS expertisés

Il n'est pas rare que les données disponibles ne soient pas totalement représentatives de la réalité, ce qui crée des jeux de données déséquilibrés. Il est donc important d'analyser les IRIS expertisés. Cette analyse correspond à deux processus : l'**analyse de leur distribution** et l'**analyse de leur représentativité**.

Analyse de la distribution Cette analyse montre l'évolution de l'environnement des clients Home in Love avant et après leur mutation. La **Figure 4** illustre l'évolution du statut des clients. On remarque que les clients deviennent souvent locataires suite à leur mutation, qu'ils l'aient été auparavant ou non. Lorsque l'on arrive dans un nouveau quartier, voire une nouvelle ville, on souhaite généralement découvrir ce nouvel environnement avant de devenir propriétaire. Il est aussi possible que l'on doive conserver le statut de locataire à cause de mutations récurrentes ou si l'on n'a pas les moyens de redevenir propriétaire. La **Figure 5** illustre l'évolution des habitations avant et après la mutation. Elle montre que le nombre de personnes habitant des maisons ou des lotissements diminue au profit de celui des personnes habitant des immeubles. En effet, les clients n'ont pas toujours les moyens d'avoir le même logement qu'avant ou certains jeunes doivent partir de leur maison familiale suite à leur alternance ou première embauche.

[11] Ces deux algorithmes sont efficaces en temps d'exécution, contrairement à l'algorithme RFECV par exemple, et fournissent les indicateurs utiles à la prédiction et pas uniquement le nombre d'indicateurs à sélectionner.

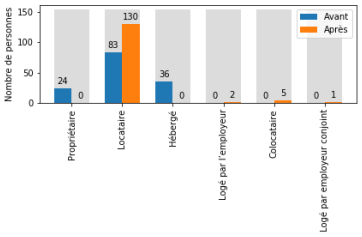

Fig. 4: Statuts des clients.

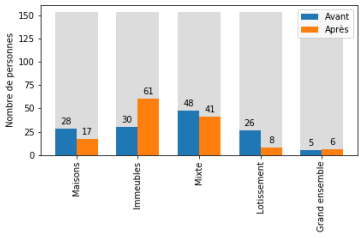

Fig. 5: Types de bâtiments.

La **Figure 6** représente l'évolution de la classe sociale et montre que la majorité des clients se situe dans les classes moyenne et moyenne-supérieure. Intuitivement, les habitants se situant dans la classe populaire ont plutôt des logements de type grand ensemble ou immeuble tandis que les personnes de classes moyenne ou moyenne-supérieure habitent plutôt dans des maisons ou des lotissements. La **Figure 7** illustre l'usage des bâtiments que les clients de Home in Love occupent. Suite à leur mutation, les clients qui habitaient un lieu résidentiel ont tendance à loger dans un lieu commerçant mais ceux qui habitaient dans une zone d'autres activités y restent. Cela se corrèle avec le fait que les personnes en mutation professionnelle ont tendance à devenir locataires de logements collectifs (Figures 4 et 5).

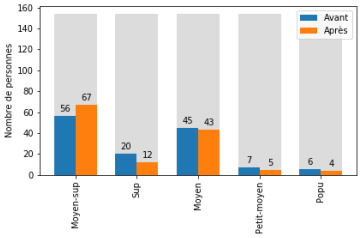

Fig. 6: Classe sociale.

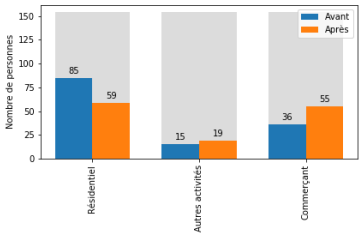

Fig. 7: Usage des bâtiments.

La **Figure 8** illustre l'évolution de la position morphologique où l'on remarque que les quartiers centraux sont en augmentation par rapport aux autres. Les personnes vivant dans des régions rurales déménagent souvent vers des zones plus centrales (*central*, *urbain*, *péri-urbain*) quand elles sont mutées. Enfin, la **Figure 9** illustre l'évolution des paysages des quartiers habités par les clients Home in Love. On observe une diminution des espaces naturels (i.e. *espaces verts*, *arboré* et *agricole*) au profit des zones urbanisées. En effet, les mutations s'effectuent souvent des zones extra-urbaines (i.e. les campagnes) vers les zones intra-urbaines (i.e. les centres-villes).

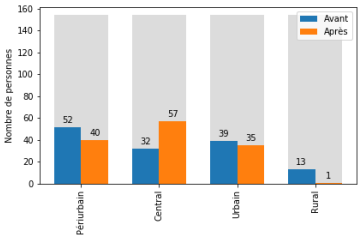

Fig. 8: Position morphologique.

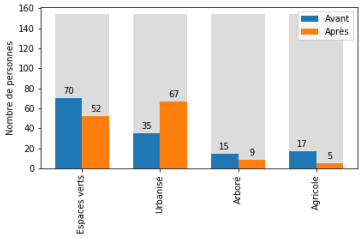

Fig. 9: Paysage.

Analyse de la représentativité Puisque les IRIS expertisés représentent seulement 0.6% des IRIS en France, nous devons analyser la représentativité de ces IRIS. Bien que cette analyse soit faite manuellement, elle permet d'étudier le biais que les variables d'environnement peuvent avoir.

- **Position morphologique.** D'après l'INSEE, 16.100 IRIS ont été construits avec les communes de plus de 10.000 habitants et la plupart des communes de 5.000 à 10.000 habitants. Pour couvrir le reste du territoire, un IRIS a été créé pour chacune des communes restantes. Si l'on considère que ces IRIS restants sont de type *rural*, alors 68% des IRIS en France seraient de type *rural*. Notre jeu de données contient 14 IRIS sur 268 annotés avec ce type, soit environ 5%. Cette différence peut s'expliquer par le fait que, dans le cadre des mutations professionnelles, les personnes ont tendance à quitter les villes rurales pour se rapprocher des centres urbains, notamment pour vivre plus près de leur travail. Cette analyse montre qu'il y a un biais sur la variable de la position morphologique. Elle montre aussi que, puisque notre jeu de données ne compte que 5% d'IRIS ruraux, il est quasi représentatif des IRIS non ruraux.
- **Paysage.** Cette variable est en lien avec la variable de la position morphologique. En effet, les IRIS éloignés des centres urbains auront tendance à être catégorisés *agricoles* voire *ruraux*. Inversement les IRIS des métropoles seront catégorisés soit *urbanisés* soit *espaces verts*. Dans notre jeu de données, 46 IRIS sont annotés *agricoles* ou *ruraux*, soit 17%. Ce pourcentage est aussi très loin des 68% d'IRIS ruraux. Donc la variable du paysage est biaisée, au détriment du paysage rural.
- **Classe sociale.** C'est une variable plutôt difficile à analyser puisque les classes sociales ne sont pas clairement définies. En France, 59% des ménages appartiennent à la classe moyenne, i.e. que leurs revenus sont compris entre 70% et 150% du revenu médian selon l'INSEE [3]. Les Français de classe moyenne sont répartis sur 71% des IRIS. Notre jeu de données contient 82% d'IRIS appartenant à la classe moyenne, donc la variable de la classe sociale n'est que légèrement biaisée.
- **Position géographique.** Elle est équitablement répartie entre les valeurs puisqu'il y a environ 25 IRIS pour chacune. Il est tout de même important de noter que certaines valeurs telles que *centre*, *nord* et *sud*, comptent deux

fois plus d'IRIS. En effet, les populations se concentrent en général dans les zones urbaines, d'où la valeur *centre* plus importante. Les valeurs *nord* et *sud* peuvent s'expliquer par le fait que certaines agglomérations ont parfois leur centre de gravité excentré. De plus amples recherches doivent être menées, notamment avec les sociologues, pour expliquer ces phénomènes.
- **Type de bâtiments.** L'INSEE recense, en 2018, 56% de logements individuels et 44% de logements collectifs[12]. Pour faire le lien avec notre variable d'environnement, les logements individuels sont les *maisons* et les *lotissements* tandis que les logements collectifs sont les *logements mixtes*, les *immeubles* et les *grands ensembles*. Dans notre jeu de données, 183 IRIS sont de type collectif, soit 68%, donc cette variable comporte un biais.
- **Usage des bâtiments.** Cette variable demande une analyse particulière, notamment avec les sociologues, car les données permettant cette analyse sont difficilement exploitables à priori.

Pour conclure, la plupart de nos variables d'environnement sont biaisées (de légèrement à fortement). Cela s'explique notamment par le fait que les personnes en mobilité géographique ont tendance à quitter les zones rurales pour venir en ville et que la location est souvent préférée à l'achat propriétaire. Décrivons maintenant l'outil Predihood qui implémente les propositions décrites dans les sections précédentes.

5.3 L'outil Predihood

Predihood est tout d'abord une approche, mais c'est aussi un outil facilitant la comparaison de quartiers selon six variables d'environnement. Cet outil se présente sous la forme d'une interface cartographique (Annexe B). Il a été développé en Python avec Flask, un framework open-source pour le développement web en Python, et Scikit-Learn, une bibliothèque populaire open-source pour les techniques de *machine learning* en Python. Son diagramme de classes est en Annexe I. Détaillons d'abord les jeux de données utilisés dans Predihood, puis deux cas d'utilisation pour notre outil.

Jeux de données La prédiction des variables d'environnement à partir des indicateurs INSEE est un problème multiclasse (plusieurs valeurs par variable d'environnement) et multilabel (six variables d'environnement). Pour pallier ce verrou, les jeux de données contiennent les six variables d'environnement mais elles sont traitées séparément lors de la prédiction, ce qui permet de réduire ce double problème à un problème multiclasse. Une entrée du jeu de données correspond à un IRIS avec les informations suivantes : le code IRIS, sa surface,

[12] Les logements individuels sont des constructions qui ne comprennent qu'un seul logement et les logements collectifs sont des logements dans un immeuble collectif. Nombre de logements individuels et collectifs en France au 1er janvier 2018.

sa densité de population, la totalité des indicateurs INSEE et les 6 variables d'environnement expertisées. L'API Pyris permet de retrouver le code IRIS à partir d'une adresse (pour transformer les adresses des dossiers en IRIS) et le module area permet de calculer la surface d'un IRIS grâce à sa géométrie.

Cas d'utilisation 1 Alice est commerciale dans le secteur informatique, ce qui l'oblige à se déplacer régulièrement dans toute la France. Elle vient d'obtenir une mission à Lyon pour quelques mois avant de repartir pour Paris. Alice aimerait trouver un quartier urbain, proche des commerces et, si possible, près d'une salle de sport. Elle sait, de par ses amies, que le quartier de la Part-Dieu est central mais elle voudrait tout de même en comparer plusieurs avant de se décider. Avec Predihood, Alice cherche "Lyon" dans la barre de recherche prévue à cet effet. Ensuite elle compare les différents quartiers grâce aux variables d'environnement, et en repère quelques-uns qui pourraient lui plaire. Ainsi elle regarde en détail les informations de deux IRIS : *Part-Dieu* et *Danton-Bir Akeim*. Les indicateurs regroupés (Annexe J) lui indiquent que le premier a beaucoup de commerces et de services (indicateurs *service-divers-prive*, *service-divers-public* et *animation-commerce-nonalimentaire*). Les indicateurs bruts (Annexe J) lui indiquent que le second dispose d'une salle de sport (indicateurs *salles multisports* et *salles de remise en forme*). Alice préfère être proche des commerces et se rendra en vélo à sa salle de sport. Elle privilégie donc l'IRIS de la Part-Dieu pour chercher un logement. L'outil propose en plus un score de confiance, qui correspond au nombre de listes L_v^k qui ont prédit la valeur proposée. Par exemple, toutes les listes ont prédit que l'IRIS de la Part Dieu était composé d'immeubles (score à 7/7).

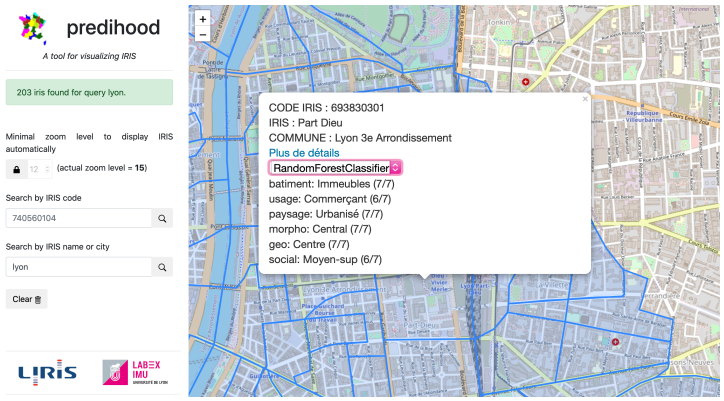

Fig. 10: Alice a saisi la requête "Lyon" et a cliqué sur l'IRIS de la Part-Dieu pour obtenir son environnement.

Cas d'utilisation 2 Bob est enseignant-chercheur en intelligence artificielle à l'Université. Il travaille actuellement sur la création d'un nouvel algorithme

d'apprentissage supervisé ainsi que sur l'amélioration d'algorithmes supervisés existants. Il ajoute tout d'abord à l'interface de paramétrage son nouvel algorithme et teste différentes configurations. Il aimerait de plus tester la fiabilité et la robustesse de ses avancées sur de nouveaux jeux de données. Il les intègre donc dans l'interface de paramétrage de Predihood et teste différentes configurations sur le jeu de données des IRIS expertisés, par exemple le nombre de voisins pour sa version améliorée de KNN. Enfin, Bob enseigne un cours sur les techniques d'apprentissage automatique et souhaiterait que ses étudiants travaillent sur un TP de prise en main des algorithmes de Scikit-Learn. Grâce à Predihood, les étudiants de Bob utilisent l'interface pour tester différentes configurations et comprendre l'influence de chaque paramètre. De plus, la possibilité de conserver les résultats des différents algorithmes exécutés, avec l'export au format Excel[13], leur permet de rendre un rapport de TP avec une partie "expérimentations" assez détaillée.

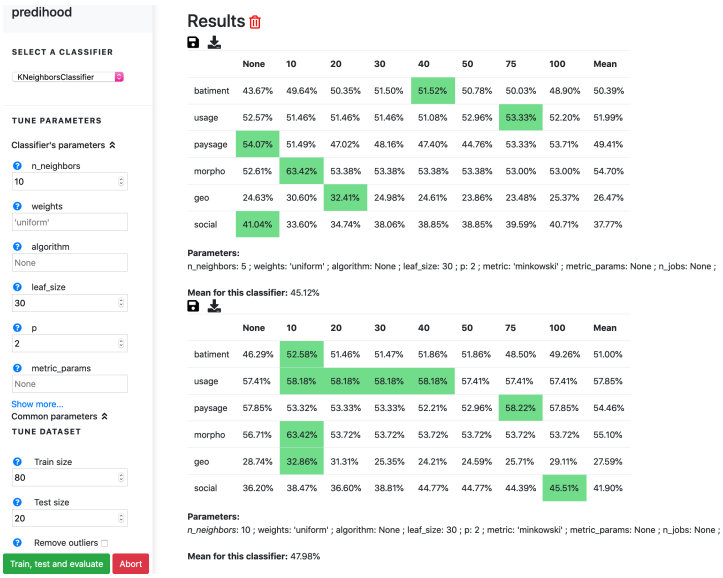

Fig. 11: Paramétrage de l'algorithme KNN dans l'interface Predihood.

Ces deux scénarios illustrent les capacités de Predihood à aider dans la recherche immobilière ainsi que dans le paramétrage générique de classifieurs. Plusieurs perspectives sont envisagées pour l'outil, comme discuté en Section 7.

[13] Le fichier Excel généré correspond au tableau des précision pour chaque variable d'environnement (lignes) et chaque ensemble d'indicateurs (colonnes).

6 Validation expérimentale

Puisque tout approche empirique doit être scientifiquement validée, il est important de vérifier les résultats de notre approche. Pour cela, il est nécessaire de mettre en place un protocole de validation expérimentale notamment pour définir les algorithmes utilisés et leurs paramètres. Deux expérimentations sont ensuite présentées pour illustrer les résultats de l'approche Predihood à l'échelle nationale puis sur le cas de la ville de Lyon.

6.1 Protocole

La validation expérimentale a pour objectif de **vérifier si la prédiction fournit des résultats satisfaisants, et donc de montrer l'intérêt de la sélection de variables sous forme de listes**. Pour ce faire, plusieurs algorithmes ont été testés et nous en avons sélectionné cinq pour ce protocole : Logistic Regression, Random Forest, K-Nearest Neighbours, Support Vector Classification et AdaBoost. Chaque algorithme contient un grand nombre de paramètres, en plus des hyperparamètres, qu'il est important d'ajuster. Grâce à l'interface de paramétrage de Predihood, un bon nombre de configurations ont été testées. Un hyperparamètre très important est la répartition entre les données d'entraînement et celles de test. Le pourcentage retenu est la répartition 80% pour les données d'entraînement et 20% pour les données de test, comme le recommande la littérature [6]. Comme les IRIS expertisés sont déjà peu nombreux, un système de validation croisée a été utilisé pour pallier ce problème puisqu'elle évite de créer les données de validation. Les tableaux de résultats de cette section correspondent aux performances, i.e. à la précision (*accuracy* en anglais), des algorithmes avec leur meilleure configuration.

6.2 Expérimentation 1 : prédiction sur l'ensemble de la France

Cette première expérimentation a pour objectif de **montrer les résultats obtenus à l'échelle nationale**. Les algorithmes ont pour objectif de prédire correctement les six variables d'environnement. Les Tableaux 1 à 6 illustrent la précision (en %) des algorithmes pour chaque variable d'environnement. Ce calcul de précision correspond au nombre de prédictions correctes par rapport au nombre de prédictions totales. L'ensemble $\mathcal{I}$ représente l'ensemble des indicateurs bruts et les listes L_v^k sont celles générées lors de la sélection. Les scores soulignés mettent en valeur le meilleur résultat de chaque algorithme. Les **scores en gras** montrent qu'ils sont meilleurs que la liste $\mathcal{I}$. Les scores en vert correspondent à la meilleure précision obtenue tout algorithme confondu, ce qui permet de mettre en avant les meilleurs algorithmes et la liste qu'ils utilisent.

Le **Tableau 1** illustre les résultats de la prédiction pour le *type de bâtiments*. Les listes permettent de gagner en moyenne quelques pourcents de précision avec

un maximum à 7% (algorithme AdaBoost pour la liste L^{20}) et Random Forest atteint le meilleur score, 60%, avec la liste L^{20}. Le **Tableau 2** montre les résultats pour l'*usage des bâtiments*. Ici aussi, Random Forest obtient le meilleur score, avec une précision de 64.9%. La liste L^{50} gagne 5% de précision avec AdaBoost, même si le score reste moins élevé que celui de Random Forest avec l'ensemble des indicateurs.

	LR	RF	KNN	SVC	AB
$\mathcal{I}$	46.6	57.0	55.2	45.5	36.5
L^{10}	44.3	**59.3**	**57.8**	44.7	**41.7**
L^{20}	**49.2**	**60.0**	**56.3**	43.6	**43.6**
L^{30}	45.1	**58.9**	**55.9**	43.6	32.1
L^{40}	46.2	**59.3**	54.8	43.2	27.6
L^{50}	46.6	**58.9**	54.8	45.5	32.4
L^{75}	44.3	**58.2**	55.2	**45.9**	32.0
L^{100}	43.6	57.0	55.2	45.5	36.5

Table 1: Qualité de prédiction pour la variable *type de bâtiments*.

	LR	RF	KNN	SVC	AB
$\mathcal{I}$	52.9	64.5	59.3	51.1	55.6
L^{10}	52.6	61.2	**63.8**	49.6	**59.6**
L^{20}	**55.9**	64.1	**63.0**	49.6	**56.6**
L^{30}	51.1	61.2	**62.3**	49.6	**60.8**
L^{40}	**57.8**	63.0	**60.8**	49.2	**56.3**
L^{50}	**56.3**	**64.9**	**62.2**	46.6	**61.1**
L^{75}	50.7	63.4	**60.8**	51.1	**58.2**
L^{100}	**53.7**	64.5	59.3	51.1	55.6

Table 2: Qualité de la prédiction pour la variable *usage*.

Le **Tableau 3** décrit les résultats obtenus pour le *paysage*. De même, les listes permettent de gagner plusieurs pourcents de précision, par exemple l'algorithme Random Forest gagne 2% de précision avec la liste L^{20} par rapport à $\mathcal{I}$. Le **Tableau 4** illustre les résultats obtenus pour la *classe sociale*. Random Forest est encore une fois le meilleur algorithme avec une précision de 51.8%. Les résultats sont moins élevés pour cette variable d'environnement, et cela peut s'expliquer par le fait que la limite entre chaque valeur n'est pas évidente (e.g. entre *moyen* et *moyen-sup*) et que l'expertise réalisée par les sociologue reste subjective (e.g. la classe sociale a été estimée en visionnant le quartier en mode street-view).

	LR	RF	KNN	SVC	AB
$\mathcal{I}$	53.7	60.8	59.6	47.7	50.3
L^{10}	48.1	**62.7**	59.6	47.7	**51.8**
L^{20}	51.5	**63.0**	**60.4**	47.7	**52.6**
L^{30}	50.3	60.8	**61.9**	47.7	**52.5**
L^{40}	49.2	**62.7**	**61.5**	47.7	49.2
L^{50}	47.7	**61.5**	**61.1**	47.7	48.1
L^{75}	52.6	**62.3**	59.3	47.7	48.5
L^{100}	**56.3**	60.8	59.6	47.7	50.3

Table 3: Qualité de la prédiction pour la variable *paysage*.

	LR	RF	KNN	SVC	AB
$\mathcal{I}$	44.4	51.1	42.1	45.5	36.5
L^{10}	43.6	46.6	**43.9**	44.7	**41.7**
L^{20}	39.1	46.6	**45.1**	43.6	**43.6**
L^{30}	41.4	49.6	**45.1**	43.6	32.1
L^{40}	39.1	**51.8**	**46.6**	43.2	27.6
L^{50}	42.1	48.1	**44.3**	45.5	32.4
L^{75}	**45.1**	48.1	**44.0**	**45.9**	32.0
L^{100}	40.7	51.1	42.1	45.5	36.5

Table 4: Qualité de prédiction pour la variable *classe sociale*.

Le **Tableau 5** décrit les résultats de la *position morphologique*. L'algorithme AdaBoost gagne jusqu'à 5% de précision, avec la liste L^{40}. Comme pour les résultats précédents, Random Forest est le meilleur avec une précision de 61.2%. Enfin, le **Tableau 6** illustre les résultats de la prédiction de la *position géographique*. Cette variable semble plutôt compliquée à prédire puisqu'il est difficile de calculer l'emplacement d'un quartier par rapport à son agglomération avec le type d'indicateurs que nous fournit l'INSEE. Tous les algorithmes, sauf Random Forest, sont meilleurs que $\mathcal{I}$ avec L^{20}.

	LR	**RF**	**KNN**	**SVC**	**AB**
$\mathcal{I}$	46.6	59.7	58.2	44.7	45.8
L^{10}	**48.5**	**60.0**	**60.8**	44.0	**49.9**
L^{20}	44.0	**61.2**	**58.5**	44.4	**48.5**
L^{30}	39.2	**61.2**	58.2	44.4	**48.8**
L^{40}	33.5	**61.2**	**58.6**	44.4	**50.7**
L^{50}	36.1	59.3	57.4	44.4	**46.2**
L^{75}	41.3	**60.8**	57.1	44.7	**49.2**
L^{100}	43.2	59.7	58.2	44.7	45.8

Table 5: Qualité de prédiction pour la variable *position morphologique*.

	LR	**RF**	**KNN**	**SVC**	**AB**
$\mathcal{I}$	22.0	33.6	27.2	25.0	15.6
L^{10}	**25.3**	29.9	**27.6**	24.6	**21.9**
L^{20}	**26.1**	31.3	**29.5**	**25.3**	**20.1**
L^{30}	**26.1**	31.7	**28.3**	**27.2**	**17.5**
L^{40}	**29.1**	32.8	**28.3**	24.6	**17.1**
L^{50}	**25.0**	32.1	27.2	23.8	**19.0**
L^{75}	**24.6**	32.8	27.2	25.0	**17.9**
L^{100}	**24.6**	33.6	27.2	25.0	15.6

Table 6: Qualité de prédiction pour la variable *position géographique*.

Pour conclure cette première expérimentation, les listes générées par la sélection de variables permettent d'améliorer les résultats. L'algorithme Random Forest reste le meilleur parmi les cinq algorithmes utilisés puisqu'il obtient le meilleur score pour toutes les variables d'environnement.

6.3 Expérimentation 2 : prédiction sur la métropole de Lyon

Cette seconde expérimentation a pour objectif d'**illustrer les performances de Predihood sur la ville de Lyon et ses environs**, une zone géographique que nous connaissons. La Figure 10, qui explique le cas d'Alice, illustre la prédiction des variables d'environnement pour l'IRIS de la Part-Dieu et selon l'algorithme Random Forest. Par rapport à la réalité terrain, le quartier de la Part-Dieu est surtout composé d'immeubles et de peu d'espaces verts malgré les initiatives récentes (e.g. le projet de refonte du centre commercial de la Part-Dieu). Les variables d'environnement *bâtiment* et *paysage* le confirment. Ensuite, ce quartier est un quartier central dans Lyon, comme le confirme la prédiction des variables *position morphologique* et *position géographique*. Enfin, le site meilleursagents.com confirme la classe sociale puisque la valeur moyenne du mètre carré est de 4900 euros à Lyon.

7 Discussion

Dans cette section, nous allons discuter des pistes d'amélioration, notamment celles prévues d'ici la fin de mon stage (31 juillet).

7.1 Sources de données

Dans le cadre de la prédiction de variables, il est souvent intéressant d'utiliser différentes sources de données, que ce soit pour comparer les données ou pour améliorer la prédiction. C'est pourquoi quatre sources pourront être ajoutées à Mongiris :

- **Prix immobiliers.** Intégrer des données de prix pourrait aider à la prédiction des variables d'environnement telles que la classe sociale. Seulement, les prix immobiliers sont rarement libres d'accès, et s'ils le sont c'est à une échelle élevée (e.g. au niveau départemental). Toutefois, l'initiative DVF, qui met à disposition les ventes immobilières récentes, semble intéressante à exploiter en collaboration avec les sociologues. De plus, les données immobilières sont souvent basées sur un découpage en parcelles cadastrales plutôt qu'en IRIS (c'est le cas pour DVF). Un appariement entre les plans cadastraux et les IRIS est donc nécessaire. Le découpage cadastral apporte plus de précision dans les zones rurales. En effet, les communes rurales ne sont souvent pas découpées par l'INSEE, i.e. que la commune correspond à un IRIS, tandis que cette même commune peut être découpée en une dizaine de parcelles cadastrales.
- **Points d'intérêts.** Les points d'intérêts, *Point Of Interest* (*POI*) en anglais, sont des objets du monde réel, comme notre Dame de Fourvière, un supermarché ou encore le Mont Blanc. Les fournisseurs géographiques représentent ces points d'intérêt par des entités, e.g. le fournisseur Geonames représente le POI "Notre Dame de Fourvière" par l'entité 8015555. Les IRIS ne comportant que des indicateurs numériques (e.g. nombre de supermarchés), prendre en compte les points d'intérêt (e.g. nom et type de supermarchés) permettrait d'ajouter de l'information à l'environnement d'un quartier. L'outil GeoAlign [2] permet de collecter les POI de plusieurs fournisseurs de manière unifiée et semi-automatique, ce qui permet une plus grande complétude des POI d'une zone donnée.
- **Offres d'emploi.** De nos jours, il y a pléthore de sites proposant des offres d'emplois (e.g. les moteurs de recherche américains Indeed et Monster, la plateforme gouvernementale PôleEmploi, ...) et des API, telles que Offres d'emploi, sont mises à disposition. Dans le cadre de la recommandation immobilière et de la mutation professionnelle, il peut être intéressant de trouver un quartier voire un logement qui plaît et de trouver un emploi dans une zone géographique proche.
- **Itinéraires.** Les données permettant le calcul d'itinéraires, e.g. les transports en commun tels que les bus et les trains, et les services routiers tels que

Google Maps ou Waze sont une source d'informations permettant de qualifier le paysage d'un IRIS. En effet, il est courant de retrouver une grande offre de transports en commun dans les quartiers urbains, inversement une offre plus réduite voire inexistante est observable pour les quartiers ruraux. De plus, les habitants des villes ont tendance à favoriser les transports en commun (moins coûteux et parfois plus rapides), d'où la réduction des trajets en voiture pour les citadins. En revanche, les habitants en périphérie voire en dehors des centres urbains n'ont pas d'autre choix que de prendre leur voiture. De plus, les services routiers permettant le calcul d'itinéraires peuvent compléter les données de l'INSEE quant aux données routières qui sont encore peu référencées.

Les sources de données ont été définies au début du projet, et leurs données évoluent. Par exemple, l'INSEE met à jour ses jeux de données tous les ans (lors du recensement de la population) ou environ tous les 4 ans (e.g. pour les équipements). Quand les données ne sont pas accessibles par une API (qui garantit la fraîcheur des données), il faut prévoir un mécanisme de mise à jour des données (e.g. un script d'intégration).

7.2 Traitement des valeurs inconnues

Les indicateurs qui n'ont pas de valeur sont complétés par la médiane des valeurs connues de celui-ci. Comme expliqué dans la Section 4.4, ce traitement comporte un biais. La correction de ce biais nécessite un travail supplémentaire dont l'idée générale est de calculer la médiane uniquement avec les IRIS qui sont similaires en termes de variables d'environnement. Par exemple, si un IRIS rural n'a pas de valeur pour l'indicateur *transports en commun*, la médiane sera calculée uniquement avec les valeurs des transports en communs des IRIS ruraux.

7.3 Distribution des indicateurs

Une seconde approche pour la sélection des indicateurs est d'utiliser la distribution des indicateurs dans les IRIS. Ces distributions correspondent aux valeurs des indicateurs normalisés par la densité de population, et sont représentées sous la forme de points reliés. Sur l'axe des abscisses se trouvent les indicateurs (numérotés) et sur l'axe des ordonnées se trouvent les valeurs de chaque indicateur.

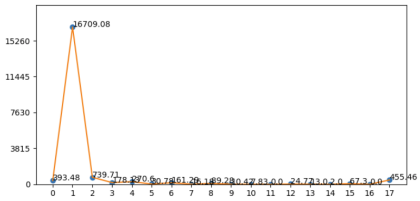

Fig. 12: IRIS de la Doua

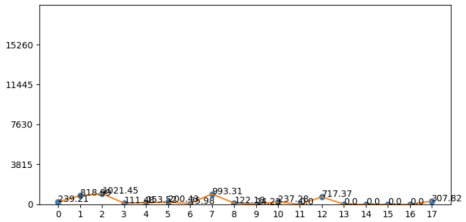

Fig. 13: IRIS de Saint-Cyr au Mont d'or

Les Figures 12 et 13 montrent respectivement la distribution des valeurs des IRIS de la Part-Dieu (forte densité, peu de résidences principales de type maison, peu d'habitations de plus de 120 m^2) et celle de Saint-Cyr au Mont d'Or (densité faible, présence de maisons et d'habitations de plus de 120 m^2) pour un sous-ensemble de 18 indicateurs. Une analyse visuelle de ces deux IRIS permet de repérer que les indicateurs sur la densité (1), le nombre de résidences principales qui sont des maisons (7) et le nombre d'habitations de plus de 120 m^2 (12) semblent utiles à la prédiction. En étendant cette étude à une dizaine d'IRIS exemples, on remarque que certains indicateurs se démarquent souvent dans le même type d'IRIS, e.g. la densité est élevée dans les IRIS urbains, le nombre de pièces est élevé dans les IRIS périurbains et ruraux, la population étrangère est plus importante dans l'Est que dans l'Ouest ou encore les habitations de plus de 120m^2 se trouvent dans l'ouest périurbain ou dans les campagnes. L'étude de la distribution des indicateurs permet donc de proposer deux points de vue : le premier est de repérer les indicateurs discriminants dans le cadre de la sélection de variable et le second est de former des groupes d'IRIS expertisés partageant des valeurs d'indicateurs similaires. L'idée du premier point de vue est de **repérer les indicateurs les plus discriminants grâce à la distribution de leurs indicateurs** pour créer une liste d'indicateurs à utiliser dans les algorithmes de prédiction. Ces indicateurs se repèrent visuellement par les "pics" (Figures 12 et 13). Pour automatiser cette tâche, les indicateurs sélectionnés sont ceux qui ont une valeur supérieure à la moyenne des indicateurs (Algorithme 2). Pour chaque IRIS expertisé, la moyenne de ses indicateurs est calculée (ligne 3) puis tous les indicateurs ayant une valeur supérieure à cette moyenne sont retenus (lignes 4 à 7). Enfin, la liste résultante de cet algorithme correspond à l'union des listes d'indicateurs sélectionnés pour chaque IRIS expertisé. Il est important de noter que cette sélection n'est pas dépendante des variables d'environnement.

L'idée du second point de vue est de **regrouper les IRIS grâce à la distribution de leurs indicateurs**. Intuitivement, des IRIS ayant des distributions similaires d'indicateurs seront catégorisés par la même valeur de variable d'environnement, i.e. qu'ils appartiendront au même groupe. Ce regroupement est effectué par l'algorithme de classification non supervisée K-Means qui permet de séparer les observations d'un jeu de données en k groupes. Appliqué à notre cas, l'algorithme K-Means permet donc de partitionner les IRIS expertisés en k groupes où k est le nombre de valeurs pour la variable d'environnement consid-

Algorithme 2 : Sélection des indicateurs via leur distribution

```
   Entrée : liste d'IRIS expertisés R, liste des indicateurs I
   Sortie : liste d'indicateurs L
1  L ⟵ ∅;
2  for iris ∈ R do
3  |  x̄ ⟵ sum(iris.values) / size(iris.values);
4  |  L' ⟵ ∅;
5  |  for i ∈ I do
6  |  |  if iris[i] > x̄ then
7  |  |  |  L'.append(i);
8  |  L ⟵ L ∪ L'
```

érée. La technique dite "du coude"[14], *elbow method* en anglais, permet de déterminer le meilleur k. Les résultats de cette méthode appliquée à nos iris expertisés montrent qu'il faut entre 3 et 7 groupes pour une variable d'environnement donnée, ce qui donne du crédit aux valeurs définies par les sociologues et inversement. Les résultats de l'algorithme K-Means montrent que les groupes formés correspondent bien à la distribution des IRIS, i.e. que les IRIS qui appartiennent au même groupe ont une distribution similaire.

Il est aussi possible de regrouper les IRIS en comparant la courbe de leur distribution. L'idée est de regrouper dans un même *cluster* les IRIS qui ont des distributions similaires. Pour ce faire, une formule de similarité entre deux courbes a été définie spécifiquement pour le contexte. Cette formule prend en compte la différence absolue entre les deux indicateurs comparés ainsi que la différence d'inclinaison entre les deux droites qui relient les indicateurs considérés et les indicateurs suivants. Une rapide implémentation et quelques tests ont montré que les mesures de similarité des courbes sont toutes très proches (entre 95 et 98% de similarité) et qu'il est donc difficile d'exploiter en l'état cette piste comme aide à la prédiction.

7.4 Calcul de la position géographique

L'idée consiste à calculer la position géographique d'un IRIS au lieu de la prédire, i.e. de chercher la grande ville la plus proche de l'IRIS et de calculer sa direction par rapport à l'iris. La recherche d'une grande ville est effectuée à partir des IRIS ayant le même code postal, avec un nombre minimal d'IRIS pour considérer que c'est une grande ville. Le seuil actuel est fixé à 20 IRIS mais de plus amples expérimentations sont nécessaires pour déterminer cette valeur. Si une grande ville est trouvée, deux points représentatifs sont définis : un pour l'IRIS et un pour la grande ville. Enfin, la direction est calculée par rapport à ces points représentatifs. Plusieurs verrous émergent de cette procédure :

– Quelle est la **définition d'une grande ville**, i.e. comment fixer un bon seuil pour le nombre minimal d'IRIS ?

[14] Cette technique permet d'estimer le nombre de groupes à former en traçant pour plusieurs k la courbe correspondant au rapprochement intra-clusters.

- Comment gérer la **détection simultanée de deux grandes villes** ?
- Comment définir les **points représentatifs** d'une surface (e.g. barycentre, points les plus proches entre les deux surfaces) ?
- Comment gérer les métropoles qui contiennent des **arrondissements** (Lyon, Paris et Marseille) et les grandes communes proches de ces métropoles ?

En plus de ces nombreux questionnements, les expérimentations ont montré que l'annotation des sociologues, i.e. en utilisant des cartes, pouvait être erronée, notamment sur la direction, car l'oeil humain ne considère pas la surface contrairement à un algorithme. L'implémentation de cette méthode montre des résultats similaires, i.e. une précision d'environ 30%, à la prédiction par apprentissage supervisé.

7.5 Taille du jeu de données

Dans le contexte de l'apprentissage supervisé, il est important d'avoir assez de données expertisées, notamment car les petits jeux de données ont tendance à sur-appendre, *overfitting* en anglais, et ont par conséquence du mal à généraliser et prédire de nouvelles données. Un nombre plus important de données expertisées permet de réduire ce problème mais l'expertise des données est très coûteuse en temps. Il est possible de déterminer la taille minimale de ces données avec une formule prenant en compte un niveau de confiance et un intervalle de confiance[15]. Avec un niveau de confiance à 95% et une intervalle de confiance de 5%, on obtient la taille de 381 entrées. Le jeu de données actuel dont nous disposons est composé de 276 IRIS, ce qui reste petit notamment parce que l'expertise requiert aux sociologues entre une à deux heures par IRIS. L'interface Predihood pourrait faciliter leur travail d'expertise et ainsi permettre d'augmenter la taille du jeu de données.

7.6 Améliorations techniques

Quelques améliorations techniques de l'interface sont proposées et prévues d'ici la fin de mon stage (fin juillet) :

- La **sauvegarde des meilleures configurations** d'algorithmes dans Mongiris. Celle-ci permettrait notamment de comparer ses configurations aux meilleures enregistrées, e.g. dans le cadre d'un protocole de validation expérimentale.
- La possibilité d'**ajouter de nouveaux jeux de données**, ce qui permettrait à Predihood de devenir une interface pour Scikit-Learn. Notre interface de paramétrage est déjà générique au niveau du paramétrage, il serait donc intéressant de la rendre générique dans l'ajout de nouveaux jeux de données, e.g. pour tester la robustesse d'un algorithme selon plusieurs jeux de données ayant des caractéristiques différentes.

[15] https://www.surveysystem.com/sscalc.htm

– La **recommandation à l'utilisateur** à partir de ses choix. L'idée ici est de proposer un formulaire à l'utilisateur pour recueillir ses envies à l'aide des six variables d'environnement. L'utilisateur choisirait les critères qui lui plaisent (e.g. quartier commerçant, environnement arboré), ainsi qu'une zone géographique (Lyon par exemple). Predihood sélectionnerait ensuite les quartiers qui correspondent le mieux aux critères dans la zone géographique donnée.

8 Conclusion et perspectives

L'objectif de mon stage était de proposer un outil facilitant la comparaison de quartiers dans le cadre du projet Home in Love. L'objectif principal concernait la prédiction des six variables d'environnement. Cet objectif est atteint et des perspectives sont envisagées quant à celui-ci. Le second objectif concernait l'implémentation d'une interface utilisateur facilitant la comparaison de quartiers. Cet objectif est implémenté dans Predihood. Enfin une interface permettant le paramétrage générique d'algorithmes était souhaitée. Elle est aussi implémentée dans Predihood.

Revenons rapidement sur le déroulement de mon stage. Le mois de février s'est articulé autour du nettoyage des données ainsi que de la lecture et la synthèse d'articles scientifiques pour l'état de l'art. Le mois de mars m'a permis de réaliser et de programmer la sélection des indicateurs. Les mois d'avril et mai m'ont permis de réaliser la validation expérimentale. Enfin, le mois de juin m'a permis de rédiger mon rapport de stage et de me préparer à la soutenance. Le travail de mes trois premiers mois de stage a été récompensé par la soumission et l'acceptation de l'article "Predicting the environment of a neighbourhood: a use-case for France" à la conférence internationale Data 2020. Je présenterai cet article lors de la session des présentations orales sur le thème de la *Data Science* de la conférence. Enfin, depuis le début de mon stage, j'ai présenté mon travail lors des réunions avec les différents membres du projet, ce qui était l'occasion de faire un point sur les avancées des informaticiens, des sociologues et des envies et besoins de Home in Love.

La suite de mon stage me permettra d'intégrer de nouvelles sources pour renforcer la qualité de prédiction des variables d'environnement, notamment avec l'intégration des points d'intérêts. Elle me permettra également d'ajouter les fonctionnalités proposées dans la section de discussion. Si le temps le permet, je pourrai retravailler les propositions telles que le regroupement des indicateurs selon leur distribution et la détermination des indicateurs discriminants, toujours dans l'optique de renforcer le processus de prédiction. Enfin, l'interface cartographique de Predihood permettra d'augmenter le nombre d'IRIS expertisés grâce à notre collaboration avec les sociologues et l'interface de paramétrage pour Scikit-Learn permettra de faire de nouvelles expérimentations avec les propositions citées ci-dessus.

Ces quatre mois et demi de stage sont une très belle expérience professionnelle, tant au niveau humain que technique. J'ai su apprécier l'implication de mes maîtres de stage pour le temps qu'ils m'ont accordé et les nombreux conseils et avis sur mes travaux : présentations orales, documents de suivi, propositions et rapport de stage. Côté technique, j'ai acquis une meilleure autonomie quant à mon travail et aux difficultés rencontrées. J'ai aussi su mettre en pratique les techniques d'intelligence artificielle que j'ai acquises au cours de mon Master 2, notamment avec l'UE sur les techniques d'apprentissage automatique. L'UE sur les connaissances de la recherche a renforcé mes connaissances sur le monde de la recherche et de la publication scientifique, ce qui m'a été utile pour la rédaction de l'article. Enfin, chaque UE a su apporter sa pierre à l'édifice, par exemple vérifier les hypothèses d'un protocole expérimental, imaginer des solutions pour résoudre un problème, chercher efficacement des solutions aux problèmes techniques, et bien d'autres encore. J'ai énormément apprécié ces quatre mois et demi de stage (dont seulement un mois et demi au LIRIS, le restant en télétravail) et cela me conforte dans mon projet professionnel et mon envie de continuer l'aventure avec une thèse de doctorat.

References

1. Barret, N., Duchateau, F., Favetta, F., Miquel, M., Gentil, A., Bonneval, L.: À la recherche du quartier idéal. In: EGC. pp. 429–432 (2019)
2. Barret, N., Duchateau, F., Favetta, F., Moncla, L.: Spatial Entity Matching with GeoAlign (demo paper). In: ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. pp. 580–583 (2019)
3. Bigot, R., Croutte, P., Müller, J., Osier, G.: Les classes moyennes en Europe. Paris, Le CRÉDOC, Cahier de recherche **282** (2011)
4. Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A.: Recommender systems survey. Knowledge-based systems **46**, 109–132 (2013)
5. Bonneval, L., Duchateau, F., Favetta, F., Gentil, A., Jelassi, M., Miquel, M., Moncla, L.: Étude des quartiers: défis et pistes de recherche. In: Conférence Extraction et Gestion de Connaissances, Atelier DigitAl Humanities and cuLtural herItAge: data and knowledge management and analysis (DAHLIA) (2019)
6. Bruce, P., Bruce, A.: Practical statistics for data scientists: 50 essential concepts. " O'Reilly Media, Inc." (2017)
7. Cranshaw, J., Schwartz, R., Hong, J., Sadeh, N.: The livehoods project: Utilizing social media to understand the dynamics of a city. In: Sixth International AAAI Conference on Weblogs and Social Media (2012)
8. Donoho, D.: 50 years of data science. Journal of Computational and Graphical Statistics **26**(4), 745–766 (2017)
9. Guest, A.M., Lee, B.A.: How urbanites define their neighborhoods. Population and Environment **7**(1), 32–56 (1984)
10. Hammad, K.A.I., Fakharaldien, M., Zain, J., Majid, M.: Big data analysis and storage. In: International Conference on Operations Excellence and Service Engineering. pp. 10–11 (2015)
11. Humain-Lamoure, A.L.: Le quartier comme objet en géographie. AUTHIER JY, BACQUE MH, GUERIN-PACE F., Le quartier. Enjeux scientifiques, actions politiques et pratiques sociales, Paris, La Découverte pp. 41–51 (2007)
12. Le Falher, G., Gionis, A., Mathioudakis, M.: Where is the Soho of Rome? Measures and algorithms for finding similar neighborhoods in cities. In: Ninth International AAAI Conference on Web and Social Media (2015)
13. Lillesand, T., Kiefer, R.W., Chipman, J.: Remote Sensing and Image Interpretation. John Wiley & Sons (2015)
14. Liu, Y., Wei, W., Sun, A., Miao, C.: Exploiting geographical neighborhood characteristics for location recommendation. In: Conference on Information and Knowledge Management. pp. 739–748 (2014)
15. Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., García, Á.L., Heredia, I., Malík, P., Hluchỳ, L.: Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. AI Review **52**(1), 77–124 (2019)
16. Qiu, J., Wu, Q., Ding, G., Xu, Y., Feng, S.: A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing **2016**(1), 67 (2016)
17. Yuan, X., Lee, J.H., Kim, S.J., Kim, Y.H.: Toward a user-oriented recommendation system for real estate websites. Information Systems **38**(2), 231–243 (2013)
18. Zhang, A.X., Noulas, A., Scellato, S., Mascolo, C.: Hoodsquare: Modeling and recommending neighborhoods in location-based social networks. In: 2013 International Conference on Social Computing. pp. 69–74. IEEE (2013)

A Variables d'environnement

Les six variables d'environnement résultent d'un travail commun et pluridisciplinaire entre les informaticiens et les sociologues. Ces six variables permettent de définir simplement l'environnement d'un quartier, ce qui facilite aussi la comparaison et la recommandation de ceux-ci.

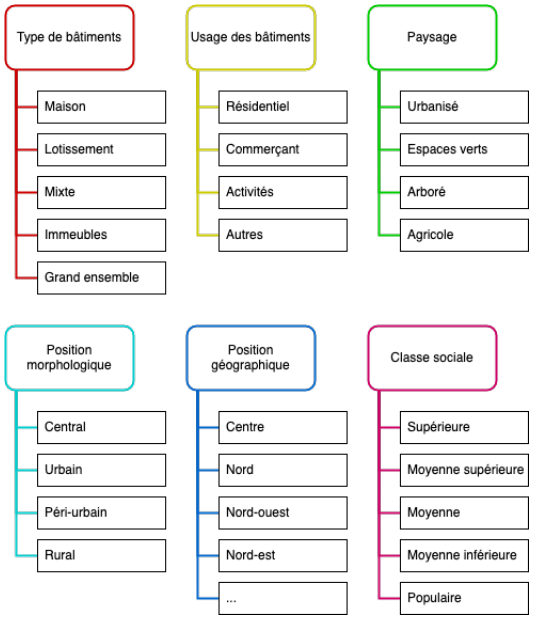

Fig. 14: Liste des six variables d'environnement.

B Visualisation des IRIS

L'outil Predihood permet de visualiser les IRIS sur une carte (Figure 15).

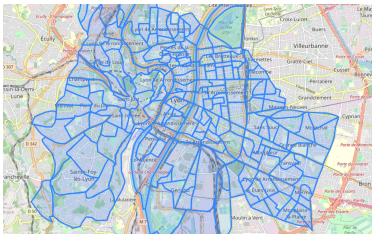

Fig. 15: Visualisation (partielle) d'IRIS pour la requête "Lyon" (d'autres IRIS contenant Lyon dans leur nom ou dans le nom de leur commune sont également affichés par cette requête).

C Fichiers clients Home in Love

Les Figures 16a et 16b illustrent des extraits des dossiers fournis par Home in Love. Pour chaque dossier sont renseignés l'adresse de l'ancien lieu de vie (colonne B) avec l'expertise des six variables d'environnement (colonnes C à H) et l'adresse du nouveau lieu de vie (colonne I) avec de même l'expertise des six variables d'environnement (colonnes J à O).

	A	B	C	D	E	F	G	H
1	num_HiL	Aadresse	Abatiment	Ausage	Apaysage	Amorpho	Ageo	Asocial
2	...	19 Chemin du Grand Roule 69350 La Mulatière	Mixte	Résidentiel	Arboré	Urbain	Ouest Lyon	Sup
3	...	40 boucle Saint-Exupéry 57310 Guenange	Lotissement	Autres activités	Espaces verts	Périurbain	Est Thionville	Moyen
4	...	5 allée des Chênes 72650 Aigné	Lotissement	Résidentiel	Agricole	Périurbain	Nord de Le Mans	Moyen-sup
5	...	9 rue Jean-Baptiste Lebas 59760 Grande-Synthe	Mixte	Résidentiel	Espaces verts	Périurbain	Ouest Dunkerque	Moyen
6	...	9 allée des Lilas 37230 Fondettes	Lotissement	Résidentiel	Agricole	Périurbain	Ouest Tours	Moyen-sup
7	...	27 rue Gauthey 75017 Paris	Immeubles	Commerçant	Urbanisé	Urbain	Nord Paris	Moyen
8	...	4 rue Clos des Vignes 86130 Jauney Clan	Lotissement	Résidentiel	Espaces verts	Périurbain	Nord Poitiers	Petit-moyen
9	...	37 allée de la Rue Basse 87350 Panazol	Maisons	Résidentiel	Arboré	Périurbain	Est Limoges	Petit-moyen
10	...	150 chemin de la Tuilerie 13320 Bouc-Bel-Air	Maisons	Résidentiel	Espaces verts	Périurbain	Nord Marseille	Sup

(a) Exemple d'expertise pour les anciennes adresses des clients Home in Love.

	A	I	J	K	L	M	N	O
1	num_HiL	Nadresse	Nbatiment	Nusage	Npaysage	Nmorpho	Ngeo	Nsocial
2	...	1 allée du Limayrac 31500 Toulouse	Lotissement	Résidentiel	Arboré	Périurbain	Ouest Toulouse	Sup
3	...	52 chemin de Robert 13270 Fos-sur-Mer	Lotissement	Résidentiel	Espaces verts	Périurbain	Nord Fos-sur-Mer	Moyen-sup
4	...	301 La Grange 49770 La Meignanne	Immeubles	Autres activités	Espaces verts	Périurbain	Ouest Angers	Moyen
5	...	10 chemin des Flouriès 81600 Gaillac	Maisons	Résidentiel	Espaces verts	Périurbain	Nord Gaillac	Moyen-sup
6	...	4 chemin du Calvaire 42580 La Tour en Jarez	Lotissement	Résidentiel	Espaces verts	Périurbain	Nord Saint-Étienne	Moyen-sup
7	...	22 rue Jérôme Dulaar 69004 Lyon	Immeubles	Résidentiel	Espaces verts	Urbain	Ouest Lyon	Moyen-sup
8	...	19 rue de la Gare 33320 Eysines	Mixte	Résidentiel	Espaces verts	Périurbain	Nord-Ouest Bordeaux	Moyen
9	...	14 rue André Moinier 63000 Clermont-Ferrand	Immeubles	Commerçant	Urbanisé	Central	Centre Clermont-Ferrand	Moyen
10	...	91 rue Hippolyte Kahn 69100 Villeurbanne	Grand ensemble	Résidentiel	Urbanisé	Central	Centre Villeurbanne	Moyen

(b) Exemple d'expertise pour les nouvelles adresses des clients Home in Love.

Fig. 16: Exemple d'expertise des dossiers fournis par Home in Love.

D Fichiers INSEE

La Figure 17 illustre un extrait de fichier source fourni par l'INSEE. Celui-ci correspond aux données relatives à la démographie, e.g. l'IRIS de la Doua avait une population de 2559 habitants en 2014. Les indicateurs suivants détaillent la répartition de ces habitants par catégorie d'âge. À noter que ce fichier comporte bien plus de colonnes (83 en tout) pour chaque iris.

IRIS	Libellé de l'IRIS	Population en 2014 (princ)	Pop 0-2 ans en 2014 (princ)	Pop 3-5 ans en 2014 (princ)	Pop 6-10 ans en 2014 (princ)	Pop 11-17 ans en 2014 (princ)	Pop 18-24 ans en 2014 (princ)	Pop 25-39 ans en 2014 (princ)
IRIS	**LIBIRIS**	**P14_POP**	**P14_POP0002**	**P14_POP0305**	**P14_POP0610**	**P14_POP1117**	**P14_POP1824**	**P14_POP2539**
692640601	Belleroche	3736	301	211	392	445	345	718
692650000	Ville-sur-Jarnioux (commune non irisée)	831	28	35	70	91	38	137
692660101	Charmettes	3567	168	103	181	177	790	1058
692660102	Charles-Hernu	4908	218	169	220	337	954	1576
692660103	Charpenne-Wilson	5616	174	195	245	352	1095	1418
692660201	Doua	2559	3	0	0	27	2399	117
692660202	Onze-Novembre	2987	107	50	67	78	1208	668
692660301	Tonkin-Sud	4358	242	199	261	274	541	1189
692660302	Espace-Central	3181	188	126	175	191	288	761
692660401	Stalingrad	0	0	0	0	0	0	0
692660402	Tonkin-Ouest	2254	107	95	174	210	195	390
692660403	Tonkin-Nord	2309	102	83	93	151	364	739
692660501	Croix-Luizet-Ouest	3524	32	27	39	117	1963	516
692660502	Croix-Luizet-Est	2382	78	38	44	116	903	578
692660601	Einstein-Salengro	2312	121	118	129	170	332	603
692660701	Buers-Est	2611	121	107	167	204	254	409
692660702	Buers-Nord	3672	160	180	229	225	470	754
692660703	Buers-Sud	2305	130	96	136	145	219	412
692660801	Saint-Jean	4099	223	201	393	481	455	799

Fig. 17: Exemple de fichier source fourni par l'INSEE.

E Indicateurs regroupés

La Figure 18 montre la hiérarchie des indicateurs regroupés.

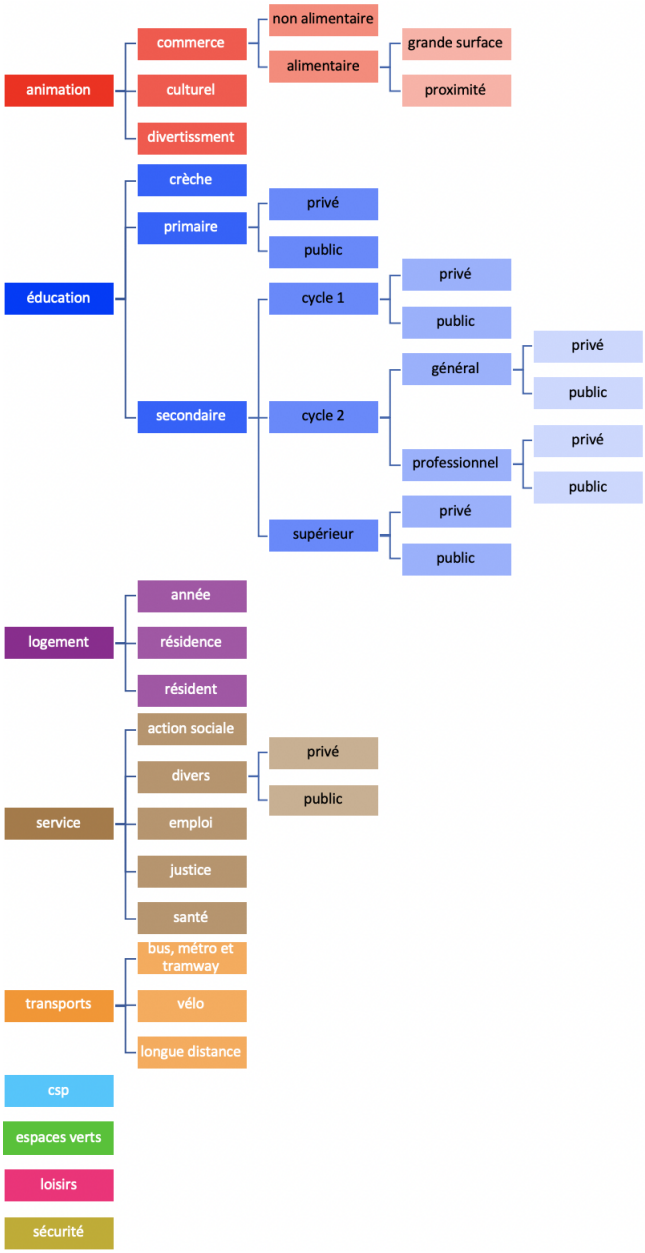

Fig. 18: Hiérarchie des indicateurs regroupés.

F Détection des valeurs erronées

Puisque l'expertise des dossiers clients a été manuelle, certaines valeurs sont erronées. Pour pallier ce problème, un dictionnaire est construit pour regrouper les valeurs par similarité. La mesure de similarité utilisée est celle de Levenshtein avec un seuil d'une opération. Ainsi, deux valeurs sont considérées suffisamment similaires lorsqu'une seule opération (i.e. suppression, ajout ou modification) est nécessaire pour passer de l'une à l'autre. Dans le Tableau 7, on remarque que, pour la variable *bâtiment*, la valeur "Immeubles" contient deux mentions ("Immeuble" et "Immeubles") alors que la valeur "Maisons" est toujours bien orthographiée.

bâtiment	[Immeubles, Immeuble], [Maisons], [Grand ensemble], ...
usage	[Espaces verts, Espaces vert, Espace verts], ...
...	...
social	[Pop, Popu], [Sup], [Oui], [Moyen-sup], [Moyenne-sup], ...

Table 7: Valeurs des variables d'environnement saisies dans les dossiers clients regroupées par similarité.

G Indicateurs spécifiques

Le filtrage des indicateurs a été réalisé à partir d'un fichier Excel défini préalablement à la main. Chaque indicateur a un statut : 0 si l'indicateur est conservé, 1 si l'indicateur est considéré comme trop spécifique et donc retiré des indicateurs potentiellement utiles à la prédiction. La Figure 19 illustre un extrait de ce fichier Excel d'aide au filtrage.

	A	B	C
1	INDICATOR	DESCRIPTION	STATUS
292	NB_F103	Tennis	0
293	NB_F103_NB_AIREJEU	Tennis - nombre de courts	1
294	NB_F103_NB_COU	Tennis avec au moins un court couvert	1
295	NB_F103_NB_ECL	Tennis avec au moins un court éclairé	1
296	NB_F104	Équipement de cyclisme	0
297	NB_F104_NB_AIREJEU	Équipement de cyclisme - nombre de pistes	1
298	NB_F104_NB_COU	Équipement de cyclisme avec au moins une piste couverte	1
299	NB_F104_NB_ECL	Équipement de cyclisme avec au moins une piste éclairée	1

Fig. 19: Exemples d'indicateurs trop spécifiques.

H Hiérarchie des indicateurs

La hiérarchie des indicateurs a permis de créer un arbre généalogique de la totalité des indicateurs INSEE avec comme racines (niveau 1) les indicateurs les plus généraux, e.g. la population, le nombre de ménages ou encore le nombre de logements, et comme descendance les indicateurs ordonnés par spécificité (Figure 20). Au total, la hiérarchie contient cinq niveaux. Elle a été construite manuellement, ce qui a été coûteux en temps mais elle présente deux avantages : elle a été définie en Excel donc elle est facilement exploitable par un programme Python et elle pourrait avoir d'autres utilités comme l'aide à la justification.

	A	B	C	D
1	INDICATOR	LABEL	LEVEL	ANCESTOR
18	C14_ACT1564	Actifs 15-64 ans en 2014 (compl)	2	['P14_POP']
19	C14_ACT1564_CS1	Actifs 15-64 ans Agriculteurs exploitants en 2014 (compl)	3	['C14_ACT1564']
20	C14_ACT1564_CS2	Actifs 15-64 ans Artisans, Comm., Chefs entr. en 2014 (compl)	3	['C14_ACT1564']
21	C14_ACT1564_CS3	Actifs 15-64 ans Cadres, Prof. intel. sup. en 2014 (compl)	3	['C14_ACT1564']
22	C14_ACT1564_CS4	Actifs 15-64 ans Prof. intermédiaires en 2014 (compl)	3	['C14_ACT1564']
23	C14_ACT1564_CS5	Actifs 15-64 ans Employés en 2014 (compl)	3	['C14_ACT1564']
24	C14_ACT1564_CS6	Actifs 15-64 ans Ouvriers en 2014 (compl)	3	['C14_ACT1564']
25	C14_ACTOCC1564	Actifs occupés 15-64 ans en 2014 (compl)	3	['C14_ACT1564']
26	C14_ACTOCC1564_CS1	Actifs occ 15-64 ans Agriculteurs exploitants en 2014 (compl)	4	['C14_ACTOCC1564']
27	C14_ACTOCC1564_CS2	Actifs occ 15-64 ans Artisans, Comm., Chefs entr. en 2014 (co	4	['C14_ACTOCC1564']
28	C14_ACTOCC1564_CS3	Actifs occ 15-64 ans Cadres, Prof. intel. sup. en 2014 (compl)	4	['C14_ACTOCC1564']
29	C14_ACTOCC1564_CS4	Actifs occ 15-64 ans Prof. intermédiaires en 2014 (compl)	4	['C14_ACTOCC1564']
30	C14_ACTOCC1564_CS5	Actifs occ 15-64 ans Employés en 2014 (compl)	4	['C14_ACTOCC1564']
31	C14_ACTOCC1564_CS6	Actifs occ 15-64 ans Ouvriers en 2014 (compl)	4	['C14_ACTOCC1564']

Fig. 20: Exemple de la hiérarchie des indicateurs.

I Diagramme de classes de Predihood

La modélisation de Predihood se veut générique pour faciliter l'ajout de nouveaux algorithmes et est basée sur Scikit-Learn, ce qui permet d'intégrer tous les algorithmes qu'ils proposent. Scikit-Learn est une des références en termes de librairie pour l'apprentissage automatique et les résultats de leurs algorithmes peuvent servir de référence, notamment pour évaluer la qualité d'un nouvel algorithme. Le projet Predihood se base principalement sur trois classes : *Data*, *Dataset* et *Method*. Les objets (fonctions et attributs) précédés d'un signe - sont des éléments internes à la classe qui n'ont pas lieu d'être utilisés lors de l'utilisation des classes. À l'inverse, les objets précédés d'un signe + sont ceux qui sont considérés comme publics et donc accessibles et utiles à part entière.

Fig. 21: Diagramme de classes du projet Predihood.

La modélisation de chaque classe est détaillée ci-dessous :

- **Classe Data.** Elle représente les données sous la forme d'une DataFrame. La classe Data contient aussi des attributs tels que les indicateurs du jeu de données (*indicators*). La méthode *init_all_in_one* permet de générer une instance contenant les données du fichier Excel nettoyé contenant les clients avec la liste des indicateurs dans le jeu de données.
- **Classe Dataset.** Elle contient une instance de type Data afin d'avoir accès aux données ainsi que les jeux d'apprentissage et de test (*X_train*, *Y_train*, *X_test* et *Y_test*) puisque les algorithmes utilisés sont de type supervisés. La méthode *init_all_in_one* permet d'initialiser les données et de créer les jeux de données.
- **Classe Method.** Elle est la classe mère de quatre classes et représente un processus (e.g. le nettoyage des données et la sélection des indicateurs).
  - La classe *MethodCleaning* nettoie les données des IRIS expertisés, i.e. les données Home in Love. La fonction *clean* permet de lancer le processus

de nettoyage et permet à l'expert de choisir les valeurs souhaitées pour les valeurs contenant des fautes.
- La classe *MethodSelection* correspond au processus de sélection des indicateurs. Chaque instance de cette classe correspond à une méthode de sélection, i.e. la sélection avec l'algorithme Random Forest est une instance de cette classe et la sélection avec l'algorithme Extra Tree Classifier en est une seconde. L'attribut *best_ indicators* contient les meilleurs indicateurs pour l'instance considérée.
- La classe *MethodPrediction* permet la prédiction des variables d'environnement d'un IRIS et la prédiction des variables d'environnement pour tout le jeu de données des IRIS expertisés. La fonction *predict* permet notamment de prédire les valeurs des six variables d'environnement de l'IRIS donné en paramètre ou de prédire les valeurs pour tous les IRIS expertisés et ainsi évaluer la qualité de la prédiction avec les jeux d'entraînement et de test.
- Enfin la classe *Chart* permet de modéliser les débuts de réflexion sur l'intégration de la distribution des indicateurs comme sélection d'indicateurs ou comme aide à la prédiction.

J Indicateurs de l'IRIS de la Part-Dieu

Les Figures 22 et 23 décrivent quelques uns des indicateurs pour l'IRIS de la Part-Dieu, Lyon.

Grouped indicators for IRIS 693830301 - Part Dieu

Indicator label	Value
logement-resident	0.0
education-superieur-prive	0.0
animation-commerce-alimentaire-grandesurface	4.0
education-secondaire-cycle1-public	0.0
education-secondaire-cycle2-professionnel-public	0.0
animation-commerce-nonalimentaire	209.0
education-primaire-prive	0.0
espacevert	0.0
service-sante	1.0
service-divers-public	39.0
education-secondaire-cycle2-professionnel-prive	0.0
education-secondaire-cycle2-general-public	0.0
education-creche	0.0
animation-divertissement	102.0
animation-commerce-alimentaire-proximite	7.0
csp	0.0
service-divers-prive	126.0

Fig. 22: Indicateurs regroupés de l'IRIS de la Part-Dieu.

NB_D221	Chirurgien dentiste	7.0
P14_PMEN_ANEM10P	Pop mén emménagés depuis 10 ans ou plus en 2014 (princ)	829.082470571081
UU2010	Unité urbaine	00758
P14_HNSCOL15P	Hommes 15 ans ou plus non scolarisés en 2014 (princ)	854.2754752870591
NB_F104_NB_COU	Équipement de cyclisme avec au moins une piste couverte	0.0
P14_CHOM5564	Chômeurs 55-64 ans en 2014 (princ)	5.64788469277966
P14_F2554	Pop 25-54 ans Femmes en 2014 (princ)	580.7037863504088
NB_F113_NB_AIREJEU	Terrains de grands jeux - nombre de terrains	0.0
NB_F121	Salles multisports (gymnase)	0.0
NB_F120	Salles de remise en forme	0.0
NB_A122	Réseau de proximité pôle emploi	0.0

Fig. 23: Indicateurs bruts de l'IRIS de la Part-Dieu.