UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES ET INFORMATIQUES
APPLIQUÉES (3799)

PAR
SYLVAIN MARCOUX

INDICE DE VITALITÉ POUR LA VILLE DE SHAWINIGAN
UNE APPROCHE DE PRÉVISION INTELLIGENTE DE LA VITALITÉ DES QUARTIERS

JUIN 2025

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE :

Jean-Sébastien Dessureault, Directeur de mémoire,
Professeur au département de mathématiques et informatique.

M. François Meunier,
Professeur au département de mathématiques et informatique.

M. Pierre-Olivier Parisé,
Professeur au département de mathématiques et informatique.

Résumé

L'intelligence artificielle offre aujourd'hui des outils puissants pour l'analyse territoriale et la prise de décision municipale. Ce mémoire explore l'apport des méthodes d'apprentissage automatique et de leurs techniques explicables dans le contexte de la planification urbaine, à travers une application concrète à la ville de Shawinigan. Divers algorithmes dont K-Nearest Neighbors, Random Forest, Multi-Layer Perceptron et K-Means sont mobilisés pour construire des modèles robustes, interprétables et applicables à des données d'une ville. L'objectif principal est de démontrer comment ces approches permettent de traiter un grand volume de données, mais aussi d'en extraire des connaissances explicables. Pour valider cette approche, deux indices ont été développés : l'Indice de Vitalité Actuel, fondé sur 13 indicateurs socio-économiques, et l'Indice de Vitalité à Long Terme, basé sur l'évolution de quatre variables clés sur 15 ans. Ces indices servent de cas d'application pour illustrer les capacités des modèles d'apprentissage automatique en matière d'imputation, de classification, de prédiction et d'explication. Les résultats sont visualisés à l'aide de cartes interactives et de séries temporelles, et validés par des experts du territoire. L'analyse des valeurs SHAP permet de rendre ces modèles interprétables, renforçant ainsi leur utilité pour les décideurs. Ce travail met en lumière le potentiel de l'IA explicable pour guider des politiques urbaines fondées sur des données, et ouvre la voie à une planification municipale plus transparente, efficace et durable.

Abstract

Artificial intelligence now offers powerful tools for territorial analysis and municipal decision-making. This thesis explores the contribution of machine learning methods and their explainability techniques in the context of urban planning, through a practical application in the city of Shawinigan. Various algorithms, including K-Nearest Neighbors, Random Forest, Multi-Layer Perceptron, and k-means, are used to build robust, interpretable models applicable to municipal data. The main objective is to demonstrate how these approaches can handle large volumes of data while also extracting explainable insights.

To validate this approach, two indices were developed : the Current Vitality Index, based on 13 socio-economic indicators, and the Long-Term Vitality Index, built on the evolution of four key variables over 15 years. These indices serve as case studies to illustrate the capabilities of machine learning models in imputation, classification, prediction, and explanation. The results are visualized using interactive maps and time series and validated by territorial experts. The use of SHAP values makes the models interpretable, thereby enhancing their usefulness for decision-makers. This work highlights the potential of explainable AI to inform data-driven urban policies and paves the way for more transparent, effective, and sustainable municipal planning.

Avant-propos

Je souhaite exprimer ma plus profonde reconnaissance à mon directeur de recherche, le Professeur Jean-Sébastien Dessureault, pour son soutien indéfectible et la confiance qu'il m'a témoignée tout au long de mon parcours à l'UQTR. Sa disponibilité, sa rigueur intellectuelle et ses conseils éclairés ont joué un rôle déterminant dans la réalisation de ce document. Je lui suis infiniment reconnaissant pour les connaissances précieuses et les expériences enrichissantes qu'il m'a transmises au cours de ma maîtrise, ainsi que pour le soutien financier qui a facilité mes recherches.

Je dédie ce mémoire à ma conjointe, Mélissa Paré, à mes enfants, Xavier et Hanaé, à ma maman, et à tous ceux qui, comme moi, chérissent la vie et en profitent pleinement.

Quand je pense à tous les livres qu'il me reste à lire, j'ai la certitude d'être encore heureux. Jules Renard(1864-1910)

Table des matières

Liste des tableaux

Table des figures

Liste des abréviations

— **AA** : Apprentissage automatique
— **AD** : Aires de diffusion
— **ACP** : Analyse en composantes principales
— **EQM** : Erreurs quadratiques moyennes
— **IA** : Intelligence artificielle
— **IDMS** : Indice de défavorisation sociale et matérielle
— **IVA** : Indice de Vitalité Actuel
— **IVL** : Indice de Vitalité à Long Terme
— **INSPQ** : Institut national de santé publique du Québec
— **KNN** : *K-Nearest Neighbors*
— **MLP** : *Multi-Layer Perceptron*
— **RF** : *Random Forest*
— **RLS** : Réseau local de service
— **SHAP** : *SHapley Additive Explanations*
— **SVM** : *Support Vector Machine*

Chapitre 1

Introduction

Dans un contexte de transformations urbaines, les outils traditionnels ont des difficultés à offrir une lecture fine et prédictive des dynamiques locales. Ce mémoire explore le potentiel de l'intelligence artificielle (IA) et des méthodes d'explicabilité pour mieux comprendre la vitalité des quartiers. L'introduction qui suit expose les fondements de cette démarche.

1.1 Mise en contexte

L'IA a connu une grande évolution depuis les années 1950. Au début, elle était centrée sur la résolution de problèmes logiques et mathématiques. Ensuite, l'IA s'est progressivement diversifiée pour intégrer des domaines tels que l'apprentissage automatique (AA), le traitement du langage naturel et la vision par ordinateur. Les avancées technologiques et la croissance exponentielle des données disponibles ont permis l'émergence de nouvelles méthodes d'analyse et de prédiction. Aujourd'hui, l'IA s'impose comme un outil incontournable dans divers secteurs, dont la planifica-

tion urbaine. Les centres urbains au Québec, à l'image de la ville de Shawinigan, sont en pleine transformation sous l'effet de multiples facteurs sociaux, démographiques et économiques. Parmi ces facteurs, les politiques publiques ont joué un rôle important : des décisions comme les fusions municipales ou les subventions à l'industrie technologique apportées par le gouvernement canadien influencent directement le développement urbain et économique. [1]

En parallèle, d'autres dynamiques, comme la crise du logement [2] ou l'augmentation de l'itinérance [3], redéfinissent les enjeux sociaux et économiques auxquels les villes doivent faire face. Ces changements peuvent entraîner des disparités entre les différents quartiers, nécessitant de nouvelles stratégies d'analyse pour comprendre les impacts et anticiper les besoins futurs.

Pour répondre à ces défis, ce travail mobilise différentes techniques d'AA, telles que les forêts aléatoires (RF) pour l'identification des caractéristiques explicatives les plus influentes, l'algorithme K-means pour le regroupement des aires de diffusion (AD) en différents secteurs, et les réseaux de neurones *Multi-Layer Perceptron* (MLP) ou la régression linéaire pour la prédiction de tendances futures. Ces méthodes permettent d'exploiter le plein potentiel des données disponibles. Le processus inclut également des étapes de traitement, telles que l'imputation des données manquantes par l'algorithme *K-Nearest Neighbors* (KNN) et la normalisation selon la méthode Min-Max, afin d'assurer la comparabilité des résultats.

Par ailleurs, l'explicabilité des modèles est assurée par l'utilisation des valeurs SHAP (*SHapley Additive exPlanations*), qui permettent de quantifier la contribution de chaque caractéristique aux prédictions. Ces résultats sont représentés sous forme de diagrammes en violon, offrant une visualisation intuitive pour les urbanistes et

1. https ://www.canada.ca/fr/developpement-economique-regions-quebec/nouvelles/2024/07/le-gouvernement-du-canada-investit-dans-linnovation-en-soutenant-le-digihub-shawinigan.html
2. https ://www.lacsq.org/actualite/crise-du-logement-portrait-dune-dure-realite/
3. https ://www.lapresse.ca/actualites/2023-09-13/l-itinerance-a-bondi-de-44-en-cinq-ans-au-quebec.php

les décideurs. Cette intégration d'outils d'explication contribue à une meilleure transparence des modèles, tout en renforçant leur acceptabilité. Conscients des enjeux éthiques liés à l'IA, nous avons privilégié une approche axée sur les huit principes de l'*Institute for Ethical AI & Machine Learning*[4]. Ces principes fondamentaux permettent de garantir que l'IA, dans cette étude, demeure socialement acceptable.

1.2 Problématique

Dans ce contexte urbain en mutation, les quartiers constituent des unités fondamentales où s'exercent les dynamiques sociales et économiques. Ils sont le lieu où se construisent les communautés, mais aussi où se manifestent les inégalités urbaines. L'évolution des politiques locales, provinciales et nationales, ainsi que les changements démographiques, influencent directement la vitalité de ces quartiers. Ces répercussions peuvent conduire à des inégalités dans l'attribution des ressources et des investissements, favorisant parfois la dévitalisation de certains secteurs. Pour assurer un développement harmonieux et équitable, il est nécessaire de disposer d'un outil permettant de suivre et d'évaluer l'impact de ces changements sur la vie des citoyens et l'évolution des quartiers. Cet outil offrirait une base solide pour éclairer les décisions politiques et orienter les priorités en matière de développement urbain.

Or, les méthodes traditionnelles d'analyse urbaine comme les système d'information géographique ne permettent pas toujours de détecter de façon fine et proactive les signaux de transformation des quartiers. Dans cette optique, l'intégration de l'IA et des méthodes d'AA, combinée à des techniques d'explicabilité comme les valeurs de Shapley, ouvre de nouvelles perspectives. En traitant de grands volumes de données sociales, économiques et géographiques, ces approches permettent non seulement d'identifier les facteurs les plus influents, mais aussi de modéliser l'évolution potentielle de

4. https ://ethical.institute/index.html

la vitalité des quartiers à court et à long terme. Ainsi, le développement d'indices dynamiques de vitalité urbaine, basés sur ces outils, pourrait permettre aux décideurs municipaux de mieux cibler les interventions, de réduire les inégalités territoriales et de favoriser une planification plus juste et durable.

1.3 Objectifs du mémoire

En définissant et ensuite en utilisant diverses méthodes d'IA telles que les algorithmes de classification, de régression, de regroupement non supervisé, ainsi que des outils d'explicabilité comme les valeurs SHAP, ce mémoire a pour objectif de développer un outil de suivi de la vitalité des quartiers de la ville de Shawinigan. Deux indicateurs distincts seront créés : l'Indice de Vitalité Actuel (IVA) et l'Indice de Vitalité à Long Terme (IVL), afin d'évaluer les dynamiques urbaines à court et à long terme à partir de données sociales, économiques et démographiques.

L'indice IVA fournira une évaluation à jour de la situation actuelle des quartiers en s'appuyant sur les données les plus récentes. Il permettra d'avoir une « photo » précise et immédiate de la vitalité de chaque secteur, avec une analyse fine basée sur une multitude de caractéristiques disponibles pour l'année en cours.

Le deuxième indice, nommé IVL, sera élaboré à partir des données disponibles sur une période plus étendue. Il comportera moins de caractéristiques, en raison de la rareté des données historiques, mais offrira une vision évolutive de la vitalité des quartiers. Il permettra de suivre les tendances à long terme et de mieux anticiper les besoins d'aménagement futurs.

Le chapitre deux sera consacré à la revue de littérature, le chapitre trois portera sur les notions d'apprentissage automatique et d'explicabilité, le chapitre quatre

présentera la méthodologie ainsi que les résultats liés à l'indice de vitalité de la ville de Shawinigan, et le chapitre cinq conclura le mémoire.

1.4 Conclusion

Finalement, face aux enjeux croissants de planification urbaine, l'IA offre des outils puissants pour mieux comprendre les dynamiques locales et orienter les décisions publiques. En combinant des techniques d'AA à des méthodes d'explicabilité, ce mémoire propose une approche méthodique, transparente et éthique. L'objectif est de développer des indices représentant la vitalité urbaine, capables de refléter les réalités actuelles et d'en prévoir l'évolution.

Chapitre 2

Revue de Littérature

L'analyse des données des quartiers urbains de la ville de Shawinigan nécessite l'utilisation de méthodes d'apprentissage afin de tirer un maximum d'informations à partir de données souvent complexes et incomplètes. Dans cette optique, plusieurs approches en science des données sont mobilisées pour appuyer l'étude de la vitalité d'un territoire.

L'analyse en composantes principales (ACP) permet de réduire la dimensionnalité des jeux de données tout en conservant l'essentiel de l'information, ce qui facilite la visualisation et l'interprétation. La méthode d'agrégation k-means, quant à elle, est utilisée pour regrouper des unités territoriales selon leurs caractéristiques similaires, menant à une classification utile pour l'analyse des dynamiques urbaines. Afin de prédire la vitalité à plus long terme, le XGB Regressor et la régression linéaire se révèlent particulièrement performants lorsqu'ils sont entraînés à partir des indicateurs des années 2006, 2011, 2016 et 2021 pour estimer la valeur projetée de 2026. Pour mieux comprendre les décisions de ces modèles, l'approche par les valeurs de Shapley est utilisée afin de déterminer l'importance relative des caractéristiques. En amont, la technique d'imputation KNN permet de compléter les données manquantes avec

rigueur. Cela évite d'exclure des données, ce qui permet de préserver la représentativité de l'ensemble des informations recueillies. L'objectif final est de construire un indice de vitalité de la ville qui soit à la fois fidèle à la réalité du terrain et ancré dans une démarche analytique rigoureuse. Cette revue de littérature présente donc les méthodes mobilisées, leurs fondements, ainsi que leur pertinence dans le cadre d'une étude sur la vitalité des quartiers de la ville de Shawinigan.

2.1 Les différentes méthodes

2.1.1 Random Forest

L'algorithme RF a été introduit par Leo Breiman en 2001 [1]. Breiman a longtemps défendu l'idée que les modèles statistiques complexes ne sont pas toujours la meilleure solution. En effet, il a mis en avant l'importance de modèles plus simples, qui restent suffisamment performants tout en étant plus compréhensibles et interprétables que d'autres approches. Les forêts aléatoires trouvent leur origine dans les arbres de décision. Toutefois, ces derniers présentent certains biais à l'apprentissage, notamment une forte tendance au surapprentissage lorsqu'ils sont utilisés seuls. L'approche des forêts aléatoires atténue ces limites en faisant un *pooling* de prédictions d'un grand nombre d'arbres construits à partir d'échantillons aléatoires des données. De plus, l'auteur nous informe que les algorithmes de type RF peuvent offrir de bonnes performances de prédiction. Cependant, selon lui, il est difficile de déterminer les paramètres optimaux pour garantir une bonne performance. L'auteur recommande, entre autres, d'augmenter le nombre de variables aléatoires que l'algorithme considère à chaque nœud des arbres générés.

La sélection de caractéristiques par RF est une utilisation classique de l'algorithme et elle est étudiée dans plusieurs articles, notamment dans [2], [3] et [4]. Par exemple, Jaiswal [2] applique l'algorithme RF à la sélection de caractéristiques ainsi qu'à la classification et à la régression, afin de réaliser une étude comparative de cet algorithme selon différentes perspectives. Les auteurs abordent également les concepts de calcul de l'importance des caractéristiques, tant par permutation que par l'indice de Gini. L'indice de Gini mesure l'inégalité dans la répartition d'une variable au sein d'une population.

Par ailleurs, Genuer [3] cite plusieurs autres méthodes de sélection de caractéristiques, notamment les méthodes *Wrapper*, les méthodes fondées sur les scores Support Vector Machine (SVM), ainsi que celles basées sur les scores CART. La sensibilité aux variables n (taille de l'échantillon) et p (nombre de caractéristiques prédictives) dans le calcul de l'importance est aussi abordée dans un tableau comparatif regroupant plusieurs combinaisons de ces paramètres. L'expérimentation montre que l'importance des caractéristiques devient instable lorsque p est élevé. Une conclusion notable est que, lorsque des caractéristiques bruitées sont introduites, leur importance mesurée est nulle. Les auteurs distinguent deux objectifs pour la sélection de caractéristiques : (1) identifier les caractéristiques importantes fortement liées à la variable cible dans un objectif d'interprétation, et (2) effectuer une réduction de dimension suffisante pour permettre une bonne prédiction de la caractéristique cible. Ils proposent une méthodologie en deux étapes fondée sur l'importance des caractéristiques calculée avec RF. Dans un objectif d'interprétation, il s'agit de construire une collection de modèles RF en utilisant les k premières caractéristiques les plus importantes et de sélectionner celles incluses dans le modèle menant à la plus faible erreur *out-of-bag* (OOB). Pour la prédiction, en partant des caractéristiques sélectionnées pour l'interprétation, on construit une séquence ascendante de modèles RF en ajoutant et testant les caractéristiques une à une. Les caractéristiques du dernier modèle sont alors retenues.

L'étude menée par les auteurs de [4] révèle une mesure naïve de l'importance des caractéristiques basée sur les RF, qui consiste simplement à compter le nombre de fois où chaque caractéristique est sélectionnée par l'ensemble des arbres individuels de l'algorithme. Comme alternative, ils proposent d'utiliser la fonction `cforest`, disponible en langage R. Contrairement à RF, la fonction `cforest` crée des forêts aléatoires non pas à partir d'arbres de classification CART fondés sur le critère de séparation de Gini, mais à partir d'arbres de classification non biaisés, basés sur une inférence conditionnelle. Le travail des auteurs consiste à comparer les résultats obtenus avec RF et `cforest`. Ils concluent que le nombre de catégories d'une caractéristique influence son importance, et qu'il est déconseillé d'utiliser des caractéristiques présentant un grand nombre de catégories conjointement avec d'autres n'en présentant que peu.

2.1.2 XGB Regressor

Pour appuyer les résultats obtenus avec l'approche RF, l'algorithme eXtreme Gradient Boosting (XGBoost) est utilisé. Le XGB Regressor est un algorithme d'AA basé sur le principe du boosting d'arbres de décision. Le boosting est une technique d'apprentissage automatique qui combine plusieurs modèles faibles et un modèle fort pour créer un modèle global plus performant. Il est largement reconnu pour son efficacité, sa rapidité et ses performances dans le traitement de grandes quantités de données. Il a été récompensé à de nombreuses reprises dans le domaine de la modélisation prédictive, comme le mentionnent T. Chen [5] dans l'article XGBoost : *A Scalable Tree Boosting System*. Selon l'auteur, le système XGBoost fonctionne plus de dix fois plus rapidement que les solutions populaires existantes. Toujours selon T. Chen, l'algorithme XGBoost se compare avantageusement à d'autres algorithmes de Tree Boosting, tels que pGBRT, Spark MLlib, H2O, scikit-learn et GBM de R. En effet, l'algorithme propose différentes méthodes de construction d'arbres que les autres n'utilisent pas. Par exemple, la méthode Exact Greedy Algorithm, un algorithme de recherche de séparation qui énumère toutes les séparations possibles sur l'ensemble

des caractéristiques. Les auteurs soulignent toutefois que cette méthode peut être coûteuse en calcul, notamment pour les variables continues.

2.1.3 KNN Imputation

Selon [6], les méthodes d'imputation des données manquantes visent à remplacer les valeurs manquantes par des estimations basées sur les données disponibles. Selon la théorie de Rubin (1976), mentionnée par [7], il existe trois types de mécanismes expliquant l'absence de données : lorsque les données sont dites manquantes de façon **complètement** aléatoire, cette absence n'est donc liée à aucune autre caractéristique et totalement indépendante. Si les données sont manquantes de façon aléatoire seulement, cela implique que les valeurs manquantes dépendent d'autres caractéristiques observées. Enfin, si les données sont manquantes de façon non aléatoire, cela suggère que l'absence est liée aux autres données observées, rendant ainsi l'imputation plus complexe. Les auteurs de [6] expliquent différentes méthodes d'imputation courantes, on trouve la substitution par un autre cas, l'imputation par la moyenne ou le mode, ainsi que les techniques de *hot deck* et *cold deck*. La méthode *hot deck* remplace les valeurs manquantes en utilisant des données similaires regroupées en groupes, tandis que la méthode *cold deck* utilise des sources externes. Les modèles prédictifs, qui utilisent les autres caractéristiques pour prédire les valeurs manquantes, offrent une approche plus avancée, mais peuvent souffrir d'un biais en raison de la régularité des valeurs prédites. Bien que ces méthodes soient utiles, leur précision dépend des relations entre les caractéristiques et la structure des données.

Un des premiers regroupements d'auteurs à utiliser l'algorithme KNN pour l'imputation de données est [8]. Il a évalué trois méthodes : (1) *Singular Value Decomposition* (SVD), (2) *KNNimpute* et (3) Moyenne par ligne. Sur la base des résultats de leur étude, ils recommandent la méthode d'imputation KNN. L'article [7] fait aussi une comparaison entre différentes méthodes. Par exemple, l'étude compare les algorithmes

Regression, EM, MICE, KNN, Cluster, RF et CART à l'aide d'un jeu de données issu d'une cohorte portant sur les maladies cardiovasculaires au sein de la population du sud du Xinjiang, en Chine. L'étude porte sur 38 variables réparties en cinq catégories : informations personnelles, examens physiques, questionnaires, résultats biochimiques en laboratoire et indicateurs de santé. La méthodologie choisie par les chercheurs consiste à utiliser une approche d'apprentissage supervisée, via un SVM, pour construire un modèle d'apprentissage visant à prédire les événements cardiovasculaires, puis à comparer différentes méthodes d'imputation des données manquantes en évaluant leur performance à l'aide de trois métriques : l'erreur absolue moyenne, la racine carrée de l'erreur quadratique moyenne et l'aire sous la courbe. L'erreur absolue moyenne est une métrique utilisée pour évaluer la performance d'un modèle de régression. Les méthodes d'imputation sont d'abord appliquées pour combler les valeurs manquantes, puis les données imputées sont comparées au jeu de données complet à l'aide des métriques sélectionnées. Des modèles de prédiction basés sur les SVM sont ensuite construits pour chacun des jeux de données traités, incluant le jeu de données complet. Les performances de chaque modèle sont calculées, et les résultats des comparaisons permettent d'identifier la méthode d'imputation la plus performante selon les critères de l'étude. Les résultats montrent que la méthode KNN présente la plus faible erreur absolue moyenne ainsi que la plus faible erreur quadratique moyenne parmi toutes les méthodes utilisées. La méthode RF arrive en deuxième position avec également de bons résultats. Les auteurs [7] et [8] discutent des limites de leur étude. Par exemple, celle-ci n'a pas évalué d'autres approches pour le traitement des valeurs manquantes, telles que l'imputation par *hot-deck* ou par réseaux de neurones. De plus, les auteurs rappellent l'importance des données réelles et soulignent qu'il est crucial de minimiser le nombre de données manquantes.

Dans leur article *Nearest neighbor imputation algorithms : a critical evaluation*, L. Beretta et A. Santaniello [9] offrent une évaluation approfondie des méthodes d'imputation basées sur les plus proches voisins comme un plus proche voisin (1NN), kNN ou voisins pondérés (wkNN). Les auteurs analysent les forces et les faiblesses de ces

approches en les comparant à d'autres techniques populaires comme la moyenne. Dans la méthodologie employée, des simulations ont été réalisées sur des jeux de données synthétiques et réels, dans lesquels différents modèles de valeurs manquantes ont été introduits. L'objectif était d'évaluer la performance de trois variantes de l'imputation par plus proche voisin. Ces méthodes ont été testées avec différents ensembles de données : avec l'ensemble de données complet, avec une sélection de caractéristiques avec ReliefF, avec du bagging, ou avec une sélection aléatoire des caractéristiques. Dans les résultats, ils soulignent que le principal avantage de KNN réside dans sa capacité à préserver la structure des données. L'étude montre que le choix du nombre de voisins k influence grandement la qualité de l'imputation. L'article propose des recommandations pratiques pour le choix de k selon le type de jeux de données. Cette contribution est donc précieuse pour éclairer l'utilisation de KNN dans un cadre rigoureux, notamment pour des études comme celle portant sur la vitalité des quartiers urbains.

2.1.4 k-means

L'algorithme k-means est une méthode de regroupement d'apprentissage non supervisée. L'algorithme k-means a été introduit par James MacQueen en 1967 [10]. L'auteur de l'article [11] donne une définition du regroupement : étant donnée une représentation de n objets, il s'agit de trouver k groupes en se basant sur une mesure de similarité, de manière à ce que les similarités entre les objets d'un même groupe soient élevées, tandis que les similarités entre les objets de groupes différents soient faibles. Selon [11], les principales applications du regroupement concernent, en premier lieu, la mise en évidence de la structure des données. Ensuite, il permet l'établissement d'un regroupement naturel, visant à identifier un degré de similarité entre les objets. Finalement, le regroupement peut être utilisé comme outil de compression des données, en les organisant et en les résumant à l'aide de prototypes de groupes. Des variantes de k-means existent, comme G-Means, conçu par [12]. Il s'agit d'un algorithme de

regroupement qui étend k-means en remédiant à ses limitations concernant la forme et la distribution des groupes. En appliquant un test d'hypothèse statistique sur les points de chaque groupe, G-Means vérifie s'ils suivent une distribution normale. Si ce n'est pas le cas, le groupe est subdivisé en deux, et les nouveaux centres sont ajustés de manière itérative jusqu'à ce que tous les groupes suivent une distribution normale.

L'auteur de l'article [13] propose une autre variante de k-means avec l'algorithme FC-Kmeans, une solution au problème posé par les algorithmes de regroupement classiques, qui ne permettent pas de fixer certains centres pendant le processus de regroupement. En effet, ces méthodes considèrent que tous les centroïdes peuvent être ajustés, ce qui limite leur utilité dans des contextes réels où certains centres sont déjà établis. Pour pallier cette limitation, les auteurs introduisent deux algorithmes, FC-Kmeans et FC-Kmeans 2, qui permettent de maintenir fixe une partie des centroïdes tout en optimisant l'emplacement des centres restants. L'objectif de leur étude est de comparer ces deux approches au k-means standard, en utilisant la métrique du *Silhouette Index* pour évaluer la cohérence des regroupements. Les performances des algorithmes sont ensuite analysées en fonction du pourcentage de centroïdes fixés. Pour conclure, les auteurs de l'article [13] démontrent que pour certaines applications pratiques, comme la logistique ou la planification urbaine, il est essentiel de pouvoir faire du regroupement avec des centroïdes partiellement fixes, ce que les approches classiques ne permettent pas encore de faire efficacement.

L'auteur de l'article [14] souligne quant à lui les limites du k-means, notamment sa sensibilité à l'initialisation des centroïdes. L'étude propose une comparaison entre différentes variantes de l'algorithme, en testant plusieurs stratégies d'initialisation, tant déterministes que stochastiques. Tout d'abord, l'approche classique d'initialisation consiste à choisir les centres au hasard. Ensuite, d'autres méthodes d'initialisation existent comme k-means++, qui utilise une approche probabiliste pour choisir les centres le plus loin possible des autres. Aussi, la variante Maximin permet de choisir des centres en maximisant la distance entre les centres initiaux, tout en choi-

sissant le premier centre au hasard. Une autre méthode est le processus Kaufman qui commence par sélectionner comme premier centroïde l'observation la plus centrale, c'est-à-dire celle qui minimise la somme des distances vers toutes les autres observations. Ensuite, les centres suivants sont choisis de manière à obtenir un point qui est le plus éloigné possible des centres actuels, de manière à couvrir l'espace des données de façon plus équilibrée. De son côté, Density k-means++ est une amélioration de la méthode d'initialisation k-means++. L'idée est d'utiliser la densité des données pour guider le choix des centroïdes initiaux. Au lieu de se baser uniquement sur la distance entre les points, cette approche cherche à initier les regroupements dans les zones où les données sont plus concentrées. Les résultats des auteurs de l'article [14] montrent que les approches déterministes se distinguent généralement par leur efficacité, mais certaines méthodes stochastiques, si elles sont répétées suffisamment, peuvent aussi offrir de très bons résultats. En considérant le temps d'exécution, les auteurs concluent que les techniques d'initialisation déterministes représentent un bon compromis entre performance et simplicité, ce qui renforce leur intérêt dans des contextes appliqués comme dans notre étude de la vitalité d'une ville.

2.1.5 Analyse en composantes principales

Introduite pour la première fois par Karl Pearson en 1901, puis reprise et formalisée [15] en 1933, l'ACP constitue une avancée majeure dans la réduction de la dimensionnalité. Le but de l'ACP est de transformer des données de haute dimension en un espace de plus faible dimension, tout en maximisant la variance des données dans ce nouvel espace. En optant pour une ACP, il devient possible de représenter l'information de manière condensée à travers quelques axes principaux. En effet, l'ACP s'avère être une méthode puissante pour simplifier la visualisation des données et mettre en lumière des relations complexes.

Selon [16], une discipline dans laquelle l'ACP a été largement utilisée est celle des sciences de l'atmosphère. Plusieurs articles traitent de nombreux aspects de l'ACP dans le contexte de la météorologie et de l'océanographie, comme c'est le cas avec [17]. Les auteurs de l'article [17] soulignent également, dans leur conclusion, que l'ACP est un outil d'analyse descriptive des données à la fois largement utilisé et polyvalent. Elle possède de nombreuses adaptations qui la rendent utile dans une grande variété de situations, de types de données et de disciplines. Parmi ces adaptations, les auteurs de l'article [17] mentionnent notamment des variantes de l'ACP pour les données binaires, ordinales, compositionnelles, discrètes, symboliques, ainsi que pour les séries temporelles. L'article [18] mentionne également qu'il existe des limites pratiques quant à la taille et à la dimensionnalité des données pouvant être traitées. Dans plusieurs applications de l'ACP, la taille et la dimensionnalité des ensembles de données peuvent être très élevées, ce qui entraîne des problèmes de calcul.

2.1.6 Valeurs de SHAPLEY

Pour expliquer les différents algorithmes de cette étude, les valeurs de Shapley (SHAP) seront utilisées. Elles tirent leur nom de Lloyd Shapley [19], qui a introduit ce concept en 1952 dans le cadre de la théorie des jeux. Basées sur cette théorie, les valeurs de Shapley reposent sur une formule combinatoire finie (3.13) permettant de répartir de manière unique l'excédent total généré par une coalition entre tous ses membres. L'article [20] mentionne que les valeurs de Shapley, bien qu'utiles en IA explicable (XAI) pour attribuer l'importance des caractéristiques, sont complexes à calculer. Des méthodes d'approximation ont donc été développées afin d'en faciliter l'utilisation. Néanmoins, les valeurs de Shapley constituent une méthode originale et équitable d'explicabilité, en déterminant combien chaque participant devrait recevoir dans un jeu coopératif. En effet, elles sont les seules règles d'allocation des gains qui respectent les quatre propriétés suivantes : l'efficacité, où la somme des gains individuels est égale à la valeur totale générée par la coalition complète ; la symétrie, qui assure

que des joueurs ayant un impact identique reçoivent un gain identique ; la linéarité, selon laquelle la combinaison de deux jeux produit une attribution équivalente à la somme des attributions de chaque jeu pris séparément ; et la nullité, où un joueur qui n'apporte aucune contribution ne reçoit aucun gain.

La méthodologie des auteurs de l'article [20] consiste tout d'abord à utiliser l'ACP afin de réduire le nombre de caractéristiques dans le jeu de données « Vinho Verde » wine. Une réduction à deux dimensions a été effectuée afin de visualiser l'ensemble des vins sur un même graphique. Ensuite, la méthode k-means est utilisée pour effectuer des regroupements. Les résultats peuvent être influencés par l'initialisation et le choix préliminaire du nombre de groupes. Une boucle a donc été effectuée de 2 à 11 groupes, en calculant l'inertie intra-groupe, le score de silhouette et l'indice de Davies-Bouldin pour chaque configuration. L'inertie, qui mesure la consistance des groupes, a longtemps été utilisée comme critère d'évaluation. L'inertie mesure la distance totale entre les points de données et leurs centroïdes de groupe. Toutefois, elle tend à toujours diminuer à mesure que le nombre de groupes augmente, ce qui rend difficile le choix optimal de k. Pour cette raison, l'indice silhouette est souvent préféré. Les résultats des auteurs montrent que le choix de k = 3 a été retenu, en raison d'un indice de Davies-Bouldin élevé. Finalement, la bibliothèque SHAP est utilisée afin de mieux comprendre les caractéristiques des groupes dans le jeu de données « Vinho Verde » wine. En examinant le classement des caractéristiques ayant le plus d'influence sur le résultat du regroupement et la façon dont ces caractéristiques contribuent à la formation des groupes, les auteurs [20] sont en mesure de fournir une explication des différents regroupements obtenus. De ce fait, l'article présente des résultats d'interprétation concernant les caractéristiques des vins. Par exemple, les vins ayant des valeurs élevées de densité de dioxyde de soufre total et de dioxyde de soufre libre sont plus susceptibles d'être affectés au groupe 0. Les valeurs SHAP montrent que ces deux caractéristiques ont un impact significatif sur l'affectation des échantillons à ce groupe. Néanmoins, M. Louhichi met en garde les utilisateurs quant à l'utilisation des valeurs de Shapley, car elles ne fournissent aucune mesure concernant la précision ou

la fiabilité des explications de l'algorithme. Sans surprise, les auteurs [20] conseillent de valider les résultats en tenant compte d'autres sources.

Dans l'article *Shapley value : from cooperative game to explainable artificial intelligence* [21], les auteurs proposent une revue détaillée des approches inspirées des valeurs de Shapley pour expliquer les décisions des modèles d'AA. Les auteurs [21] présentent un cadre de classification en trois dimensions, basé sur la nature de la valeur de Shapley, le type de remplacement des caractéristiques, ainsi que la méthode utilisée pour l'approximation (spécifique au modèle ou agnostique). Un lien de visualisation est fourni pour illustrer la répartition de ces méthodes dans un espace tridimensionnel. Les auteurs [21] notent que les méthodes d'approximation de la valeur de Shapley basées sur l'échantillonnage marginal figurent parmi les plus couramment utilisées, selon leur graphique. Cet article met en évidence les nombreuses applications des valeurs de Shapley, tant pour la sélection de caractéristiques que pour l'explication locale et globale des modèles. Comme d'autres auteurs, ils discutent des limites actuelles de cette méthode, notamment les coûts computationnels élevés. Néanmoins, selon eux, la valeur de Shapley demeure la méthode la plus complète pour l'attribution des caractéristiques dans les recherches actuelles.

2.1.7 Indice de Vitalité

J. Jacobs [22] auteur du livre *Dark Age Ahead : Author of The Death and Life of Great American Cities*, fut parmi les premiers à s'intéresser à la vitalité urbaine, mettant en avant des éléments comme la densité urbaine, la diversité des usages des terres et l'architecture. Jacobs critique notamment l'urbanisme moderne pour ses développements à grande échelle, qui privilégient l'expansion structurelle au détriment des besoins et des interactions humaines. Dans sa méthodologie, elle accorde la priorité à l'observation empirique plutôt qu'aux grands modèles théoriques.

Aussi, l'article [23] et l'article [24] définissent la vitalité comme l'aptitude d'un organisme à survivre ou à fonctionner efficacement. Ainsi, un indice de vitalité urbaine représente la capacité d'un centre urbain à être viable et à répondre aux besoins de la communauté, tout en améliorant la qualité de vie de ses résidents. L'indice proposé par [23] se structure autour des dimensions de bien-être, satisfaction, description sociale et dimension spatiale, chacune mesurée selon la moyenne des quartiles des variables associées.

Par ailleurs, l'article [25] intègrent huit variables, incluant des aspects économiques et sociaux tels que les indices de défavorisation matérielle et sociale, l'état des logements, la valeur des propriétés, le taux d'inoccupation, la valeur des permis de rénovation et les loyers. Ils utilisent un algorithme génétique et k-means pour regrouper 135 AD en 10 groupes, tout en employant l'algorithme RF pour évaluer l'importance des caractéristiques.

En complément, un indice de vitalité métropolitaine élaboré par [26] comprend 24 indicateurs, notamment le nombre d'entreprises de moins de 20 employés et les espaces dédiés aux activités physiques. Les zones étudiées ont été classées de 1 à 381 pour chaque indicateur, et ces classements ont ensuite été cumulés. Finalement, les valeurs de l'Indice de vitalité métropolitaine ont été recalibrées dans une plage de 80 à 120, les scores plus élevés reflétant une meilleure performance. De plus, l'article [27] propose un indice de vitalité communautaire, basé sur l'évolution des tendances des données au fil du temps pour des variables telles que le nombre de crimes contre la propriété, le bénévolat et les crimes violents.

Par analogie avec un organisme vivant l'article [28] examine des indicateurs de croissance, de diversité et de mobilité, incluant l'augmentation de la proportion de jeunes, la diversité économique et raciale, ainsi que la proximité des services. La normalisation MinMax est utilisée pour les indicateurs, tandis que la pondération des caractéristiques est effectuée à l'aide de la méthode de l'entropie, en tenant aussi

compte des scores attribués par des experts en urbanisme. En outre, les corrélations entre les différentes dimensions de la vitalité urbaine sont analysées pour mieux comprendre leurs interrelations.

L'article [29] offre une revue exhaustive sur les villes intelligentes et l'IA, en citant Singapour et Zurich parmi les dix premières villes intelligentes. Quatre-vingts pour cent des articles sur l'urbanisation et le machine learning utilisent le clustering ou le regroupement, comme décrit dans l'article [30], avec des algorithmes populaires tels que k-means, les cartes auto-organisées et DBSCAN. Les recherches récentes emploient aussi des algorithmes non supervisés, comme en témoignent les travaux de [24] et de [25].

Pour évaluer l'efficacité des regroupements effectués par des méthodes d'AA, l'article [31] a introduit le Silhouette Score, une métrique qui mesure la qualité du regroupement des points au sein de leurs groupes respectifs. L'algorithme k-means, utilisé pour la classification des zones, a été initialement développé par [32]. Cette méthode de regroupement non supervisée est réputée pour sa simplicité, son efficacité et sa rapidité, puisqu'elle minimise la somme des distances au carré entre les points de données et leurs centres de groupes respectifs, appelés « centroïdes ». Les centres initiaux sont habituellement choisis au hasard, mais il existe certaines méthodes pour les sélectionner, comme mentionné dans l'article [13]. De plus, l'imputation des données est effectuée à l'aide de l'algorithme KNN, introduit par [33] en 1967.

2.2 Conclusion

Cette revue de littérature expose la diversité des méthodes mobilisées pour évaluer la vitalité urbaine. Les travaux fondateurs de Jane Jacobs [22] ont ouvert la voie à une réflexion profonde sur l'importance de l'environnement urbain dans le bien-

être collectif, en soulignant l'importance des interactions humaines et de la mixité fonctionnelle des quartiers. Depuis, de nombreux travaux ont affiné cette notion en intégrant des indicateurs économiques, sociaux, spatiaux, et en s'appuyant de plus en plus sur des données quantitatives complexes.

L'émergence de l'IA et des sciences des données a permis d'explorer ces dynamiques sous de nouveaux angles. Des techniques telles que l'ACP permettent de réduire la dimensionnalité des jeux de données pour en faciliter l'analyse. Les algorithmes de regroupement, comme le k-means, ont été adaptés pour intégrer des contraintes spécifiques au contexte urbain, par exemple en initialisant certains centroïdes à des emplacements existants afin d'améliorer la cohérence des regroupements.

La qualité des données demeure un enjeu central, la méthode d'imputation KNN permet de traiter les valeurs manquantes sans introduire de biais majeurs. Dans une perspective prédictive, les modèles comme XGB Regressor se montrent particulièrement efficaces pour anticiper l'évolution de la vitalité des quartiers. Pour en extraire des informations explicables et utiles, les valeurs de SHAP apportent un éclairage essentiel sur l'importance relative des caractéristiques, rendant les modèles plus transparents et exploitables pour les urbanistes. L'ensemble de ces approches combinées forme une base solide pour concevoir une méthodologie utile à la construction d'un indice de vitalité urbaine adapté à la réalité de la ville de Shawinigan. Le chapitre suivant propose un tour d'horizon théorique des différentes techniques d'AA utilisées dans la méthodologie permettant d'estimer cet indice de vitalité urbain.

Chapitre 3

Notions d'apprentissage automatique et explicabilité.

L'AA permet aux ordinateurs d'apprendre à partir de données, sans être explicitement programmés. Ces méthodes sont de plus en plus utilisées pour détecter des tendances, faire des prédictions ou automatiser des décisions complexes. Par contre, la complexité de certains algorithmes rend leurs résultats difficiles à interpréter ou à expliquer. Alors, l'explicabilité devient importante pour comprendre et justifier les décisions des modèles. Ce chapitre présente les fondements de ces deux concepts.

3.1 Méthodes d'apprentissage automatique

3.1.1 Random Forest

RF est un algorithme supervisé, basé sur une collection d'arbres de décision comme par exemple celui de la figure 3.1. Cette collection d'arbres de décision est représentée

par le modèle $h(x;\theta_k)$, qui effectue une prédiction sur une entrée x, où θ_k désigne les paramètres spécifiques à l'arbre k. Ces paramètres définissent la structure de l'arbre ainsi que ses règles de décision. L'algorithme peut être utilisé aussi bien pour la régression que pour la classification. Dans le cas d'une régression, la prédiction finale est obtenue en faisant la moyenne des prédictions des différents arbres, comme indiqué dans l'équation (3.1).

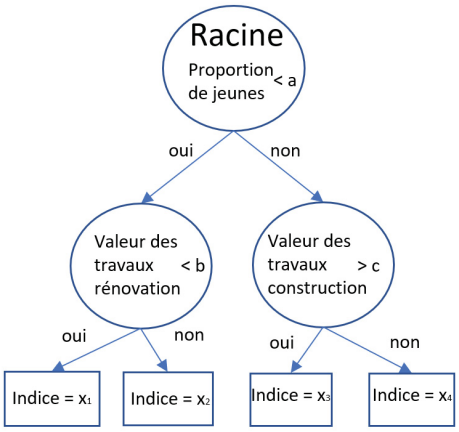


FIGURE 3.1 – Exemple d'arbre de décision

$$\hat{y} = \frac{1}{K}\sum_{k=1}^{K} h(x;\theta_k) \tag{3.1}$$

En classification, c'est le vote majoritaire parmi les arbres de la forêt qui détermine la classe prédite. Chaque arbre k de la forêt est entraîné à l'aide d'un échantillonnage avec remise à partir de l'ensemble de données d'origine.

Quand $t \to \infty$, la loi des grands nombres nous assure que l'expression (3.2) converge et que la moyenne des prédictions de la forêt aléatoire tend vers une valeur limite :

$$\lim_{t\to\infty} E_{X,Y}\left[(Y - \bar{h}(X))^2\right] = E_{X,Y}\left[(Y - E_\theta h(X;\theta))^2\right] = PE^* \tag{3.2}$$

À l'aide de l'algorithme RF, il est aussi possible d'établir un classement des caractéristiques par ordre d'importance. Chaque fois qu'une meilleure division est réalisée au sein d'un arbre, l'impureté des données (mesurée par l'indice de Gini) est réduite. Cette réduction est déterminée en comparant l'impureté avant et après la division. J. K. Jaiswal et R. Samikannu [2], traitent de la sélection des caractéristiques les plus importantes avec l'algorithme de régression RF. Selon ces auteurs, la régression par forêts aléatoires génère des arbres pour des valeurs numériques, contrairement à la classification qui s'applique à des variables qualitatives. Par conséquent, les valeurs de sortie sont également numériques. De plus, les données d'entraînement sont sélectionnées indépendamment à partir de l'ensemble de données initial. Les auteurs mentionnent également que l'erreur quadratique moyenne pour toute caractéristique prédictive X est estimée par : $E_{X,y}(y - X)^2$ pour la variable de classe y. Pour ces auteurs, il est possible de calculer l'importance des caractéristiques en utilisant les données de test. En supposant que le nombre de votes en faveur de la classe correcte est x, et que, pour les caractéristiques test permutées, il est y, la différence moyenne entre x et y sur l'ensemble des arbres de la forêt est considérée comme le score brut d'importance de la caractéristique x.

On peut également estimer l'importance des caractéristiques à l'aide de l'indice de Gini. Celui-ci mesure l'hétérogénéité des données à un certain point de l'arbre. Une valeur plus faible est donc préférable, car elle indique une meilleure séparation des classes. Dans un arbre de décision, l'indice de Gini du nœud parent est toujours plus élevé que celui des deux nœuds enfants après une division, ce qui reflète une amélioration de la pureté des nœuds. L'article [34] considère que RF estime l'importance d'une caractéristique en observant dans quelle mesure l'erreur de prédiction augmente lorsque les données de cette variable sont permutées, tandis que toutes les autres restent inchangées. Les calculs nécessaires sont effectués arbre par arbre, au fur et à mesure de la construction de la forêt aléatoire. Il est essentiel de noter, comme le soulignent [4], que l'importance mesurée à l'aide de l'indice de Gini dans les forêts aléatoires n'est pas équitable. En effet, cette mesure tend à favoriser

les caractéristiques prédictives comportant un grand nombre de catégories ou les variables continues. À l'inverse, l'importance calculée par permutation est généralement considérée comme un indicateur plus fiable et moins biaisé.

Comme Breiman [35] l'a décrit, l'algorithme RF dispose d'une routine intégrée pour gérer les valeurs manquantes. Cette procédure repose sur la pondération des valeurs observées d'une caractéristique à l'aide des proximités entre observations, calculées par l'algorithme après l'entraînement sur un jeu de données préalablement imputé par la moyenne. Concrètement, l'algorithme commence par remplacer les valeurs manquantes par la moyenne ou le mode selon le type de variable. Un modèle RF est ensuite entraîné, ce qui permet de calculer la proximité entre les observations. La mesure de proximité est la fréquence d'apparition commune dans les mêmes feuilles des arbres. Finalement, pour chaque valeur manquante, une estimation est obtenue en calculant une moyenne pondérée des valeurs des observations les plus proches. D'autres méthodes d'imputation basée sur RF, telles que RF MICE élaborées par [36] et MissForest conçu par [37], semblent donner d'excellents résultats.

3.1.2 KNN imputation

KNN imputation est une méthode d'imputation des valeurs manquantes basée sur l'idée de similarité entre les observations. Elle permet d'estimer des valeurs manquantes en se basant sur les valeurs des voisins les plus proches. Pour chaque valeur manquante, l'imputeur KNN calcule la distance euclidienne, métrique de distance par défaut, entre les valeurs manquantes et tous les autres points de données du jeu de données à l'aide de l'expression (3.3) :

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}, \tag{3.3}$$

où $d(x_i, x_j)$ représente la distance entre les observations x_i et les x_j. Les valeurs x_{ik} et x_{jk} sont les valeurs des caractéristiques des observations i et j pour la k-ème caractéristique. Ensuite, p est le nombre total de caractéristiques dans les données et $\sum_{k=1}^{p}$ représente la somme de toutes les caractéristiques k pour calculer la distance totale.

Dans l'étape suivante, les k voisins les plus proches de la valeur manquante sont identifiés. La valeur manquante est imputée en utilisant la moyenne des k voisins les plus proches identifiés l'équation (3.4). Lorsque l'ensemble de données est multivarié, l'imputeur KNN prend en compte toutes les caractéristiques disponibles.

$$\hat{x}_{ik} = \frac{1}{k} \sum_{j \in \{j_1, j_2, ..., j_k\}} x_{jk} \tag{3.4}$$

Où $\hat{x}_{ik}$ est la valeur imputée pour l'observation x_i et la caractéristique k, Les x_{jk} représentent les valeurs des caractéristiques des k voisins les plus proches pour la caractéristique. Les indices $(j_1, j_2, ..., j_k)$ représentent les k voisins les plus proches.

Il est aussi possible qu'un voisin ait des valeurs manquantes pour certaines caractéristiques. Alors, l'imputation de ce voisin doit être recalculée en fonction de ses propres voisins.

3.1.3 k-means

D'après [11], le but du regroupement de données est de découvrir les regroupements naturels d'un ensemble de motifs, de points ou d'objets. En effet, il s'agit de trouver K groupes en se basant sur une mesure de similarité. Ensuite, l'algorithme s'assure que les similarités entre les objets d'un même groupe soient élevée et que les similarités entre

les objets de groupes différents soient faibles. k-means est un algorithme populaire de regroupement non supervisé introduit en 1967 par MacQueen [10]. Il est utilisé pour regrouper un ensemble de données en un certain nombre de groupes basés sur des caractéristiques similaires. L'idée de l'algorithme k-means est de diviser les données en k groupes, où k est le nombre de groupes spécifié par l'utilisateur. Les k points, qui seront les centroïdes initiaux des k groupes, sont habituellement choisis au hasard. Il est également possible de choisir les centroïdes selon certains critères pour rendre les groupes plus significatifs. Dans notre étude de cas de la ville de Shawinigan, les centroïdes sont choisis pour avoir trois groupes significatifs : commercial, résidentiel et urbain.

Le calcul de la distance entre x_i et le centroïde c_k est donné par (3.5) :

$$d(x_i, c_k) = \|x_i - c_k\|. \tag{3.5}$$

L'assignation des x_i, au groupe k dont le centroïde est le plus proche, est donnée par l'équation (3.6) :

$$c_k = \arg\min_k \sum_{i=1}^{n} \|x_i - c_k\|^2, \tag{3.6}$$

où n est le nombre de données, i est l'index des points et k l'index des groupes. Les centroïdes sont mis à jour avec l'expression (3.7), et chaque point x est assigné au groupe le plus proche. Finalement, il s'agit de répéter la mise à jour des centroïdes jusqu'à l'obtention de la convergence, c'est-à-dire jusqu'à ce que la position des centroïdes ne change plus ou que les changements deviennent négligeables par l'équation :

$$c_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i. \tag{3.7}$$

3.1.4 ACP

La méthode d'ACP permet de réduire la dimensionnalité des données. Une fois cette réduction effectuée, l'ACP s'avère être un outil puissant pour la visualisation des données et l'interprétation des relations complexes entre les caractéristiques. Ci-dessous, un aperçu des principales étapes de cette méthode. Tout d'abord, il faut centrer les données en calculant la cote Z de chaque donnée en utilisant l'équation (3.8) :

$$Z = \frac{X - \mu}{\sigma}, \tag{3.8}$$

où Z est la valeur normalisée, X est la valeur observée, μ est la moyenne et σ est l'écart-type. Ensuite, il s'agit de calculer la matrice des covariances entre toutes les paires de caractéristiques avec l'expression (3.9) où x_i et y_i représentent les valeurs individuelles des caractéristiques X et Y, $\bar{x}$ et $\bar{y}$ sont leurs moyennes respectives et enfin n est le nombre d'observations :

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}. \tag{3.9}$$

Calculer la matrice des corrélations avec l'expression (3.10) :

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{3.10}$$

où x_i et y_i sont les valeurs individuelles des caractéristiques X et Y, $\bar{x}$ et $\bar{y}$ sont leurs moyennes respectives, et n est le nombre total d'observations. L'expression r_{XY} représente le coefficient de corrélation linéaire entre les caractéristiques X et Y.

Calculer les valeurs propres avec l'équation (3.11) :

$$\det(A - \lambda I) = 0, \tag{3.11}$$

où A est une matrice carrée, λ représente une valeur propre, I est la matrice identité de même dimension que A, et det désigne le déterminant.

Pour le choix du nombre de composantes principales, il faut sélectionner un nombre de composantes à l'aide d'une des méthodes suivantes : obtenir plus de 80% d'explication, garder les $\lambda > 1$ ou garder les λ précédant le coude de la courbe des valeurs propres. Ensuite, on peut afficher les projections sur deux ou trois axes pour interpréter les données. En plus de la réduction de dimensionnalité, l'ACP permet de mesurer la contribution des caractéristiques aux différentes composantes, ce qui représente la qualité de représentation des caractéristiques sur le graphique graphique à deux axes (CP1 et CP2) de l'ACP. Elle est calculée en élevant les coordonnées au carré. Par la suite, les contributions en pourcentage des caractéristiques aux composantes principales sont calculées comme dans l'équation (3.12) :

$$\text{Contribution}_{ij} = \frac{\cos^2_{ij}}{\sum_{k=1}^{p} \cos^2_{kj}} \times 100\%, \tag{3.12}$$

où $\cos^2_{ij}$ est le carré du cosinus entre la donnée i et l'axe factoriel j, p est le nombre total de données, et Contribution_{ij} représente la contribution relative de l'individu i à l'axe j.

3.1.5 XGBoost

XGBoost est un algorithme d'AA basé sur le boosting d'arbres de décision, introduit dans la communauté scientifique par T. Chen [5]. Depuis, il est devenu l'un des outils les plus performants et les plus populaires, notamment dans les compétitions Kaggle. XGBoost repose sur la technique du gradient boosting, qui consiste à corriger les erreurs des prédictions précédentes en ajoutant itérativement de nouveaux arbres de décision. Il peut être utilisé à la fois pour des tâches de classification ou de régression.

3.2 Éthique en apprentissage automatique

Le présent travail de recherche vise à explorer le domaine du machine learning sous l'angle de l'éthique et de l'explicabilité. Avec l'émergence croissante de ces techniques de prise de décision, il est désormais crucial d'adopter une approche vigilante et exemplaire, en respectant des principes éthiques fondamentaux pour garantir que l'IA demeure socialement acceptable. L'approche utilisée dans ce travail provient de l'organisme The Institute for Ethical AI & Machine Learning. Cet institut est aussi un centre de recherche européen qui développe des plateformes pour supporter le développement responsable de l'IA. Voici les principes de l'organisme qui sont respectés dans le présent travail de recherche.

1. **Équité** : L'équité est respectée dans cette étude, car toutes les AD sont prises en compte, assurant qu'aucune donnée n'est écartée ou favorisée. Alors, toutes les données sont traitées de manière égale.
2. **Transparence** : Les données et les algorithmes utilisés sont entièrement accessibles et expliqués à l'utilisateur, garantissant ainsi une meilleure transparence. En effet, les résultats sont interprétés de manière claire pour que chaque utilisateur comprenne le processus de décision.

3. **Responsabilité** : Les auteurs de l'étude s'engagent à assumer pleinement la responsabilité des résultats produits par les systèmes et à fournir des moyens de corriger toute erreur éventuelle.
4. **Confidentialité** : Les données utilisées dans ce travail sont du domaine public et en plus elles sont regroupées en aire de diffusion (AD). Alors, aucune donnée personnelle n'est divulguée dans cette étude, respectant ainsi strictement la confidentialité des informations sensibles.
5. **Durabilité** : L'utilisation de la librairie CodeCarbone permet de mesurer l'empreinte carbone des algorithmes, contribuant ainsi à une approche écoresponsable dans l'usage de l'IA.
6. **Collaboration** : Ce travail est conçu en partenariat avec des urbanistes de la ville de Shawinigan, démontrant ainsi un esprit de collaboration pour garantir la robustesse des résultats.
7. **Innovation responsable** : L'étude sur la vitalité des quartiers de Shawinigan incarne une innovation responsable, visant à améliorer la qualité de vie des citoyens tout en tenant compte des enjeux sociaux et éthiques.

En somme, cet étude sur la Vitalité des secteurs de la ville de Shawinigan respecte les huit principes de l'*Institute for Ethical AI & Machine Learning*.

3.3 Explicabilité en machine learning

La AD de la figure 3.2 illustre le compromis entre l'interprétabilité et la précision.

Les auteurs [38] expliquent que l'interprétabilité d'un modèle fait référence au degré auquel les prédictions d'un modèle d'AA peuvent être comprises par des êtres humains. On parle alors de méthode ante-hoc, comme pour la régression linéaire, qui est une méthode interprétable, car il est déjà possible de l'expliquer à l'aide du modèle. Par contre, quand il faut une méthode supplémentaire pour faire la lumière

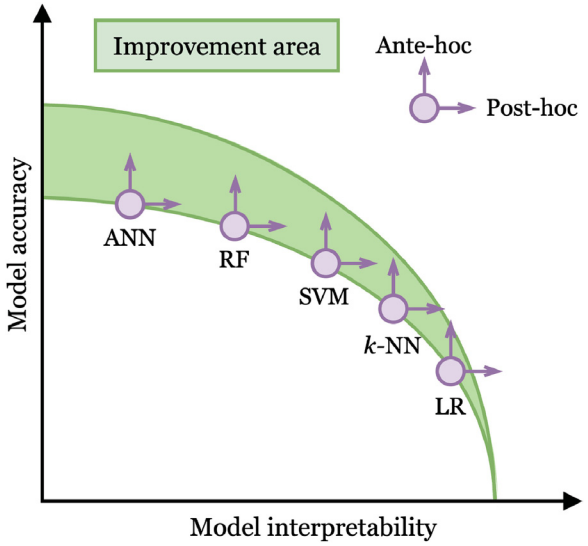


FIGURE 3.2 – Le compromis entre interprétabilité et précision.

Source : https://www.sciencedirect.com/science/article/pii/S2666792423000021#fig0001

sur la méthode, on parle de méthode post-hoc et donc d'explicabilité. Par exemple, les réseaux de neurones sont des méthodes post-hoc. En effet, si on augmente le nombre de couches cachées dans un réseau de neurones artificiels, on peut souvent améliorer sa précision, mais cela réduit l'interprétabilité du modèle. Le modèle devient alors plus complexe, ce qui rend les résultats plus difficiles à interpréter pour les utilisateurs. Alors, pour expliquer les résultats de ces algorithmes, il faudra utiliser des méthodes post-hoc comme les valeurs SHAP.

L'explicabilité, ou la transparence comme cité plus haut, en machine learning est un domaine important pour rendre l'IA socialement acceptable. Certaines librairies, comme XAI, sont spécialisée dans le domaine. La Figure 3.3 présente les différentes étapes pour une AA explicable (XAi). Les grandes étapes sont l'Analyse des données, l'Évaluation du modèle et la Supervision continue du modèle.

Ce travail tente de suivre ce processus méthodique. En effet, l'analyse des données est abordée dans la section 4.2.1, où les résultats préliminaires sont explorés. Le modèle de classification des AD, basé sur l'algorithme k-means, est évalué de plusieurs manières. Tout d'abord, la qualité du regroupement est mesurée à l'aide du

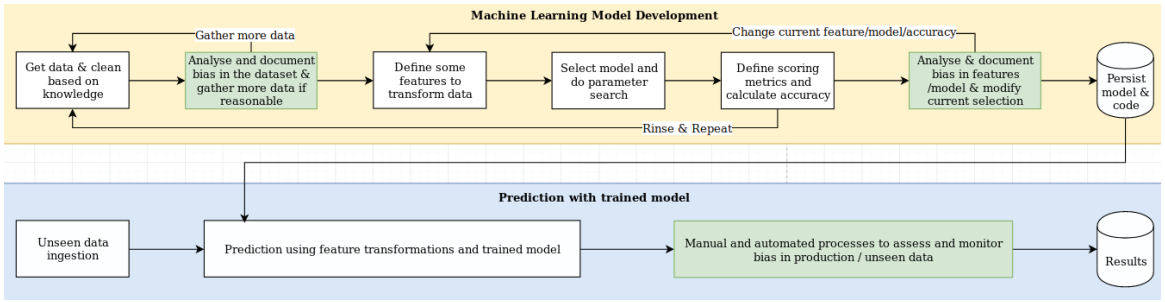


FIGURE 3.3 – Étapes pour l'explicabilité en AA
Source :`https://github.com/EthicalML/XAI`

score silhouette, qui permet de vérifier la cohésion et la séparation des groupes. Ensuite, les valeurs de Shapley sont utilisées pour fournir une interprétation des facteurs influençant le regroupement, renforçant ainsi l'aspect explicable du modèle. Enfin, une supervision continue est assurée grâce à l'interface utilisateur prévue dans cette étude, qui permettra aux utilisateurs de visualiser les résultats, même après des modifications concernant les données ou le modèle.

En résumé, ce travail de recherche s'inscrit dans une démarche éthique et responsable dans le domaine du machine learning, en se basant sur les principes établis par l'Institute for Ethical AI & Machine Learning et l'utilisation de la librairie EthicalML/XAI. L'analyse minutieuse des données, la transparence des processus, et la responsabilité dans l'application des modèles témoignent de l'engagement de l'auteur envers une utilisation éthique de l'IA et en contribuant à une société où l'IA est non seulement innovante, mais aussi juste et responsable.

3.4 Explicabilité avec valeurs SHAP

Les valeurs SHAP utilisées dans cette étude, permettent de mesurer la contribution de chaque variable au classement des AD. En d'autres termes, basée sur la théorie des jeux, la valeur de Shapley considère les prédicteurs comme des joueurs, un groupe de

prédicteurs comme une coalition et la prédiction du modèle comme le paiement du jeu.

La formule (3.13), qui est une version de la formule originale tirée de la thèse *Algorithms to Estimate Shapley Value Feature Attributions* de H. Chen [39], permet de calculer ces valeurs :

$$\phi_i(v) = \sum_{S \subseteq D \setminus \{i\}} \frac{|S|! \cdot (|D| - |S| - 1)!}{|D|!} \left[v(S \cup \{i\}) - v(S) \right], \tag{3.13}$$

où ϕ_i est la valeur de Shapley de la variable i, représentant sa contribution à la prédiction, $D = \{2006, 2011, 2016, 2021\}$, $S \subseteq D \setminus \{i\}$ est un sous-ensemble des caractéristiques qui n'inclut pas la variable i, $v(S)$ est la Valeur de la fonction prédictive lorsque seule la combinaison de caractéristiques dans S est utilisée, $v(S \cup \{i\})$ est la valeur de la fonction prédictive lorsque la variable i est ajoutée à l'ensemble S et $\frac{|S|! \cdot (|D| - |S| - 1)!}{|D|!}$ est le poids de chaque sous-ensemble S.

Calcul d'une valeur SHAP

Dans le but de bien comprendre les valeurs de SHAP, il est possible pour un petit nombres de données, de faire un calcul complet de valeur SHAP à l'aide de l'équation (3.13). Pour ce faire, nous ferons le calcul pour $\phi(2006)$, ce qui représente l'apport de l'année 2006 à la prédiction de l'IVL pour l'année 2026 de notre étude. Donc, pour l'année 2006, l'équation (3.13) devient :

$$\phi(2006) = \sum_{S \subseteq \{2011, 2016, 2021\}} \frac{|S|!(3 - |S| - 1)!}{3!} \left(v(S \cup \{2006\}) - v(S) \right) \tag{3.14}$$

Ensuite, la Table 3.1 montre les prédictions fait avec l'algorithme MLP. Un MLP de type régresseur a été utilisé avec 5 couches cachées, un nombre d'itérations d'apprentissage maximal de 5000, et ayant un $random_state$ de 1. Le paramètre $random_state$ est la graine aléatoire choisi, cela rend les résultats reproductibles. Les prédictions sont le résultat de la fonction v de l'équation (3.13). Les coalitions de joueurs (années) sont les combinaisons possibles de toutes les années. Les prédiction 2026 dans la Table 3.1 représentent la prédiction des différentes coalitions.

Coalition	Prédiction 2026
$v((2006))$	0.3247
$v((2011))$	0.3261
$v((2016))$	0.3274
$v((2021))$	0.3287
$v((2006, 2011))$	0.3254
$v((2006, 2016))$	0.3261
$v((2006, 2021))$	0.3267
$v((2006, 2011, 2016))$	0.3261
$v((2006, 2011, 2021))$	0.3265
$v((2006, 2016, 2021))$	0.3269
$v((2011, 2016))$	0.3274
$v((2006, 2011, 2016, 2021))$	0.3267

TABLE 3.1 – Valeurs de la fonction prédictive v pour les coalitions

Détails du calcul des poids

Les poids des contributions marginales sont calculés en fonction de la taille des sous-ensembles non vides $S \subseteq \{2011, 2016, 2021\}$.

Les poids sont donnés par : Poids de $S = \frac{(n-|S|)!(|S|)!}{(n+1)!}$, où n est le nombre total d'années (3 dans ce cas). Comme on ne considère qu'un sous-ensemble réduit de D ($D = \{2011, 2016, 2021\}$, donc n=3). Alors, $(n+1)!$ au dénominateur correspond à $|D|!$ dans la formule (3.13).

$S = \emptyset :$ $\frac{3!0!}{4!} = \frac{1}{4}$ $S = \{2011, 2016\} :$ $\frac{1!2!}{4!} = \frac{1}{24}$

$S = \{2011\} :$ $\frac{2!1!}{4!} = \frac{1}{12}$ $S = \{2011, 2021\} :$ $\frac{1!2!}{4!} = \frac{1}{24}$

$S = \{2016\} :$ $\frac{2!1!}{4!} = \frac{1}{12}$ $S = \{2016, 2021\} :$ $\frac{1!2!}{4!} = \frac{1}{24}$

$S = \{2021\} :$ $\frac{2!1!}{4!} = \frac{1}{12}$ $S = \{2011, 2016, 2021\} :$ $\frac{0!3!}{4!} = \frac{1}{24}$

Contributions marginales pondérées

Les contributions marginales pour chaque sous-ensemble $S \subseteq \{2011, 2016, 2021\}$ sont :

$S = \emptyset :$ $v(\{2006\}) - v(\emptyset) = 0.3247356 - 0 = 0.3247356$ (poids $= \frac{1}{4}$)

$S = \{2011\} :$ $v(\{2006, 2011\}) - v(\{2011\}) = 0.3253970 - 0.3260585 = -0.0006615$ (poids $= \frac{1}{12}$)

$S = \{2016\} :$ $v(\{2006, 2016\}) - v(\{2016\}) = 0.3260585 - 0.3273814 = -0.0013229$ (poids $= \frac{1}{12}$)

$S = \{2021\} :$ $v(\{2006, 2021\}) - v(\{2021\}) = 0.3267199 - 0.3287043 = -0.0019844$ (poids $= \frac{1}{12}$)

$S = \{2011, 2016\}$: $v(\{2006, 2011, 2016\}) - v(\{2011, 2016\}) = 0.3260585 - 0.3267199 = -0.0013229$ (poids $= \frac{1}{24}$)

$S = \{2011, 2021\}$: $v(\{2006, 2011, 2021\}) - v(\{2011, 2021\}) = 0.3264994 - 0.3273814 = -0.0015434$ (poids $= \frac{1}{24}$)

$S = \{2016, 2021\}$: $v(\{2006, 2016, 2021\}) - v(\{2016, 2021\}) = 0.3269404 - 0.3280428 = -0.0017639$ (poids $= \frac{1}{24}$)

$S = \{2011, 2016, 2021\}$: $v(\{2006, 2011, 2016, 2021\}) - v(\{2011, 2016, 2021\}) = 0.3267199 - 0.3273814 = -0.0019844$ (poids $= \frac{1}{24}$)

Somme des contributions pondérées

En additionnant les contributions pondérées, nous obtenons :

$$\phi(2006) = \frac{1}{4}(0.3247356) + \frac{1}{12}(-0.0006615) + \frac{1}{12}(-0.0013229) + \frac{1}{12}(-0.0019844)$$

$$+\frac{1}{24}(-0.0013229) + \frac{1}{24}(-0.0015434) + \frac{1}{24}(-0.0017639) + \frac{1}{24}(-0.0019844)$$

Après calcul, nous obtenons :

$$\phi(2006) \approx 0.0803019$$

Interprétation de $\phi(2006)$

La valeur de $\phi(2006)$ indique que l'ajout des données de l'année 2006 à l'ensemble des autres années, 2011, 2016, et 2021, augmente la prédiction d'environ 8.03 %. Cela signifie que l'ajout des données de 2006 a une influence sur les résultats. Dans le contexte de cette étude, une valeur de Shapley positive suggère que l'inclusion de 2006 contribue de manière constructive au modèle, en apportant des informations qui améliorent la précision des prédictions.

Pour aller plus loin, les valeurs ϕ calculées pour chaque année permettent de comparer leur influence relative sur la prédiction de 2026. À cette fin, un algorithme Python spécifique a été conçu, reposant sur la formule (3.13), sans recourir à la librairie *scikit-learn*. La Table 3.2 résume ces valeurs SHAP, révélant que chaque année contribue de façon relativement similaire à l'indice prévisionnel à long terme.

Année	Valeur de Shapley
2006	0.0803
2011	0.0813
2016	0.0821
2021	0.0829

TABLE 3.2 – Valeur de Shapley par année

3.5 Conclusion

Ce chapitre a proposé un survol des principales méthodes d'AA utilisées dans le cadre de cette recherche. Aussi, l'explication des modèles, notamment par l'entre-

mise des valeurs de Shapley, a permis de rendre ces approches plus transparentes et interprétables, ce qui constitue un enjeu central en IA. Par ailleurs, les dimensions éthiques liées à l'utilisation de ces algorithmes ont également été abordées, soulignant l'importance de la rigueur méthodologique et de la responsabilité sociale dans l'usage de ces outils. Le chapitre suivant s'appuie sur ces fondements théoriques et techniques pour proposer une méthodologie structurée, visant à développer un outil d'analyse de la vitalité urbaine à la fois rigoureux, reproductible et ancré dans des données réelles. Il détaille le processus d'acquisition, de traitement et de modélisation des données, notamment à l'aide de l'algorithme k-means, en vue de construire deux indicateurs : l'IVA et l'IVL.

Chapitre 4

Indice de vitalité de la ville de Shawinigan

Ce chapitre présente la méthodologie utilisée pour construire des indices de vitalité de la ville de Shawinigan à partir de données publiques accessibles à l'échelle des AD. Il détaille la sélection des indicateurs, les sources utilisées, ainsi que les étapes de traitement des données. Aussi, ce chapitre présente les résultats et leurs analyses. L'objectif est de développer un indice permettant de mieux saisir les dynamiques locales.

4.1 Méthodologie

4.1.1 Présentation des données

Les données utilisées dans cette étude proviennent de différentes sources telles que les recensements de Statistique Canada, le gouvernement du Québec et les rapports

de la ville de Shawinigan. Ces données couvrent plusieurs aspects socio-économiques, démographiques et immobiliers, y compris le prix des loyers, les frais de logement pour les propriétaires et les locataires, ainsi que les indices de défavorisation. Concernant l'uniformité des données, il est important, pour que l'analyse soit possible, que toutes les données soient regroupées selon les AD de Statistique Canada. Selon le gouvernement fédéral, une AD est une petite unité géographique relativement stable formée d'un ou de plusieurs îlots de diffusion avoisinants dont la population moyenne est d'environ 500 habitants. Cette unité de mesure a été choisie, car c'est la plus petite unité de mesure du gouvernement fédéral. De plus, il est possible d'obtenir plusieurs données selon cette unité géographique, facilitant ainsi les analyses à l'échelle locale.

La figure 4.1 présente les frontières des 87 AD de la ville de Shawinigan.

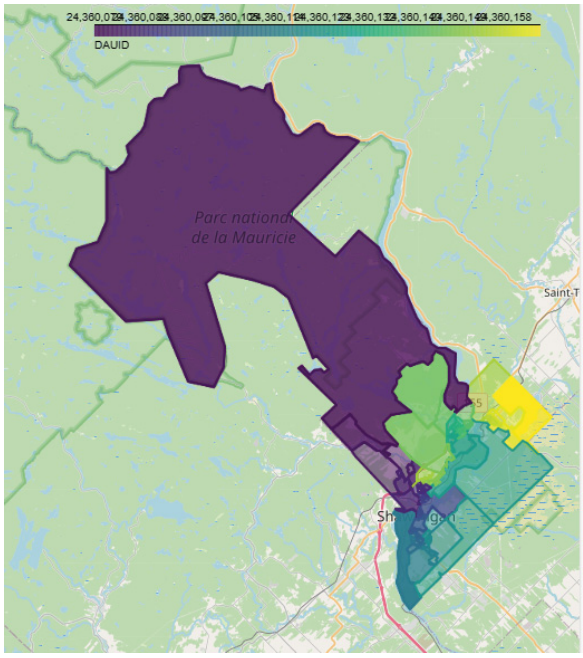


FIGURE 4.1 – Les 87 AD de la ville de Shawinigan

Les données recueillies se répartissent en quatre grandes catégories. La première, les données socio-économiques, issues du recensement de Statistique Canada, incluent des variables telles que le taux de logements inoccupés, la proportion de jeunes et les frais de logement pour locataires et propriétaires. La deuxième, les données immobilières, fournies par la ville de Shawinigan, englobent la valeur des travaux de

rénovation et de construction, la valeur des immeubles et le nombre de bâtiments vacants. Aussi, l'Indice de défavorisation correspond à l'Indice de défavorisation sociale et matérielle (IDMS) fourni par l'Institut national de santé publique du Québec (INSPQ). Cet indice peut être réparti en quartiles par rapport à la moyenne du réseau local de services (RLS) du Centre-de-la-Mauricie (CSSS de l'Énergie). Les quartiles sociaux et matériels vont de 1 (AD favorisées) à 3 (AD défavorisées), les deux quartiles centraux étant regroupés pour constituer une catégorie représentant 50 % de la population. Cet indice est constitué de plusieurs indicateurs. Ces indicateurs permettent d'évaluer différents aspects socio-économiques d'une population. La proportion de personnes de 15 ans et plus n'ayant aucun certificat ou diplôme d'études secondaires reflète le niveau de scolarisation et l'accès à l'éducation. La proportion de personnes de 15 ans et plus occupant un emploi est un indicateur clé du taux d'activité et de l'intégration au marché du travail. Le revenu moyen des personnes de 15 ans et plus, incluant diverses sources de revenu, donne une mesure du niveau de vie général. La proportion de personnes de 15 ans et plus vivant seules dans leur ménage renseigne sur les dynamiques sociales et l'isolement potentiel. La proportion de personnes de 15 ans et plus dont l'état matrimonial est soit séparé, divorcé ou veuf permet d'observer les tendances en matière de structure familiale. Enfin, la proportion de familles monoparentales est un indicateur important des réalités socioéconomiques touchant les familles. Ensuite, une ACP est réalisée sur ces données, révélant deux composantes principales : la composante sociale et la composante matérielle. La dernière catégorie concerne les données supplémentaires, une procédure a été développée pour obtenir le nombre de commerces par AD à l'aide de l'API *Google Places*. Aussi, les tables 4.1, 4.2 et 4.3 présentent en détail les variables utilisées dans cette étude, leurs indicateurs ainsi que leur source.

4.1.2 Les variables choisies

Valeur des bâtiments			
Donnée	**Caractéristique**	**Indicateur**	**Source**
Valeur des travaux de constructions effectués	**1** : Valeur des travaux de constructions	Une valeur plus basse que la moyenne indique une dévitalisation	Shawinigan, 2024
Valeur des immeubles de 1 logement ou plus	**2** : Valeur des immeubles	Une valeur plus basse que la moyenne indique une dévitalisation	Shawinigan, 2024
Nombre de logements occupés et Nombre de logements total	**3** : Taux d'inoccupation des logements	Un taux d'inoccupation plus haut que la moyenne indique une dévitalisation	Recensement, 2021
Nombre de bâtiments vacants	**4** : Nombre de bâtiments vacants	Une valeur plus haute que la moyenne indique une dévitalisation	Shawinigan, 2024

TABLE 4.1 – Description des caractéristiques concernant la valeur des bâtiments

Facteurs socio-économiques et démographiques			
Donnée	**Caractéristique**	**Indicateur**	**Source**
Indice de défavorisation matérielle	**5** : Indice de défavorisation matérielle	Un indice élevé indique une dévitalisation	INSPQ, 2024
Indice de défavorisation social	**6** : Indice de défavorisation social	Un indice élevé indique une dévitalisation	INSPQ, 2024
Proportion de jeunes de 0-14 ans	**7** : Proportion de jeunes de 0-14 ans	Une proportion plus petite que dans la ville indique une dévitalisation	Recensement, 2021
Nombre de commerces	**8** : Nombre de commerces	Un nombre de commerces élevé indique une bonne vitalité	Google Places API, 2024
Frais de logement locataire	**9** : Frais de logement moyen locataire	Une valeur plus basse que la moyenne de la ville indique une dévitalisation	Recensement, 2021
Frais de logement propriétaire	**10** : Frais de logement moyen propriétaire	Une valeur plus basse que la moyenne de la ville indique une dévitalisation	Recensement, 2021

TABLE 4.2 – Description des caractéristiques socio-économiques et démographiques

État des bâtiments			
Donnée	**Caractéristique**	**Indicateur**	**Source**
Nombre de logements nécessitant des réparations mineures	**11** : Nombre de logements nécessitant des réparations mineures	Un nombre plus élevé indique une dévitalisation	Recensement, 2021
Nombre de logements nécessitant des réparations majeures	**12** : Nombre de logements nécessitant des réparations majeures	Un nombre plus élevé indique une dévitalisation	Recensement, 2021
Valeur des travaux de rénovation effectués	**13** : Valeur des travaux de rénovation	Une valeur plus basse que la moyenne indique une dévitalisation	Shawinigan, 2024

TABLE 4.3 – Description des caractéristiques concernant l'état des bâtiments

4.1.3 Acquisition des données

Les données ont été collectées à partir de plusieurs sources publiques et privées pour obtenir une vue d'ensemble des dynamiques sociales, économiques, démographiques à Shawinigan. Les paragraphes suivants présentent les étapes clés de l'acquisition des données.

Recensements de Statistiques Canada : Les données socio-économiques et démographiques telles que les frais de logement, le taux de logements inoccupés, la proportion de jeunes et le nombre de logements nécessitants des réparations majeures et mineures ont été extraites des recensements de 2006 à 2021.

Ville de Shawinigan : La ville a fourni des données détaillées concernant les travaux de rénovation et de construction, la valeur des immeubles, ainsi que le nombre de bâtiments vacants. Ces informations, collectées pour l'année 2024, sont stockées dans des fichiers au format CSV, incluant l'adresse des bâtiments et les détails relatifs aux travaux ou à la valeur des immeubles. Par ailleurs, un fichier CSV distinct répertorie les adresses des bâtiments vacants. Afin de géolocaliser ces adresses, une fonction Python appelée `geoloc_adresse` a été développée. Cette fonction repose sur un processus de géolocalisation permettant d'associer chaque adresse aux AD correspondantes. Une clé fournie par Google est nécessaire, et des frais sont appliqués pour chaque requête dépassant un quota préétabli. L'intégration de l'API *Google Cloud* dans le programme a permis d'extraire les coordonnées géographiques des adresses fournies par la ville de Shawinigan et de les assigner aux AD adéquates, assurant ainsi une correspondance précise entre les adresses et les divisions géographiques locales. Le code de cette procédure se retrouve à l'Annexe A. Toutes les données ont ensuite été fusionnées dans un fichier `GeoPandas` contenant les coordonnées géographiques des AD ainsi que les indicateurs pertinents pour chacune de ces divisions.

Institut national de santé publique du Québec : L'institut publie un IDMS pour chaque AD du Québec.

Google : Une fonction Python a été développée pour utiliser l'API *Google Places* afin de géolocaliser les commerces au sein des AD. Cette méthode permet d'obtenir une mesure précise de la vitalité commerciale dans chaque zone. Elle utilise comme paramètre un fichier JSON contenant la géométrie des AD, essentiel pour associer chaque adresse de commerce à la zone correspondante. Un autre paramètre clé est

le type de commerce que l'API Google doit rechercher. La table 4.4 présente les différentes catégories de commerces considérées dans cette procédure. Enfin, la fonction `geol_commerce` génère un fichier CSV répertoriant le nombre de commerces dans chaque AD. Le code source de cette procédure se trouve en Annexe A.

Catégories
Bank, *Pharmacy*, *Jewelry Store*, *Bus Station*, *Car Repair*, *Restaurant*, *Hospital*, *Laundry*, *Clothing Store*, *Convenience Store*, *Hostel*, *Gym*, *Movie Theater*, *Pet Store*, *Dentist*, *Park*, *Library*, *Post Office*, *Department Store*, *Drugstore*, *Museum*, *School*, *Tourist Attraction*, *Electronics Store*, *Furniture Store*, *Supermarket*, *Church*, *Art Gallery*, *Hair Care*, *Hardware Store*, *Things to Do*, *Mosque*, *Bakery*, *Home Goods Store*, *Gas Station*, *Cafe*, *Synagogue*, *Beauty Salon*, *Bicycle Store*, *Shopping Mall*, *Bar*, *Amusement Park*, *Book Store*, *Store*

TABLE 4.4 – Liste des catégories de commerces considérées

4.1.4 Prétraitement des données

Le prétraitement des données a comporté plusieurs étapes, détaillées dans les paragraphes suivants.

Traitement des valeurs manquantes : Les données manquantes ont été gérées à l'aide de l'imputation `KNNImputer`. Cette méthode repose sur l'imputation par les k plus proches voisins, où les valeurs manquantes sont remplacées par des estimations calculées à partir des valeurs basées sur des valeurs non-manquantes de leurs voisins les plus similaires. Dans notre cas, le nombre de voisins choisi comme paramètre est fixé à cinq. Par ailleurs, pour certaines caractéristiques spécifiques (comme le nombre de bâtiments vacants), les valeurs manquantes ont été directement remplacées par des zéros, car dans ces situations, l'absence de données indique simplement qu'aucun bâtiment vacant n'était présent, et donc la valeur correcte pour cette caractéristique est effectivement zéro.

Inversion : Certaines caractéristiques ont été inversées à l'aide de l'équation (4.1) pour garantir la cohérence des indicateurs de vitalité. En effet, une valeur élevée d'un indicateur montre une meilleure vitalité, tandis qu'une valeur basse montre une dévitalisation. L'inversion est effectuée avec l'équation (4.1) :

$$x_{inverse} = x_{max} - x \tag{4.1}$$

L'interface permettra de modifier les caractéristiques sélectionnées pour l'inversion. Par exemple, dans ce travail, la caractéristique *Taux d'inoccupation* a été inversée, mais cette décision pourra être ajustée via l'interface, offrant ainsi une flexibilité dans le calcul de l'indice. Cela permet d'adapter le modèle en fonction des préférences de l'utilisateur, offrant une personnalisation pour refléter les priorités de la ville de Shawinigan.

Normalisation : Toutes les données ont été normalisées à l'aide de la méthode `MinMax` pour rendre les caractéristiques comparables entre elles. Ainsi, toutes les valeurs des caractéristiques seront situées dans l'intervalle de 0 à 1. La normalisation est calculée avec l'expression (4.2).

$$MinMax(x) = \frac{x - min(X)}{max(X) - min(X)} \tag{4.2}$$

Calculs par AD : Les valeurs moyennes de toutes les caractéristiques ont été calculées pour chaque AD.

4.1.5 Diagramme de méthodologie

La figure 4.2 montre un diagramme de la méthodologie utilisée. Dans cette figure, en rouge, on retrouve les grands axes de traitement de la méthodologie et la méthode d'AA utilisée. La boîte rouge de prétraitement représente l'étape de prétraitement, incluant l'imputation, la normalisation et la transformation inverse des données. La deuxième boîte rouge représente le regroupement des AD en trois districts à l'aide de l'algorithme k-means. La boîte IVA correspond au calcul de l'IVA, mettant en évidence les caractéristiques les plus importantes grâce à l'algorithme de forêt aléatoire. La boîte IVL représente le calcul de l'IVL pour les trois districts ainsi que la prédiction de 2026 en utilisant soit un perceptron multicouche (MLP), soit une régression linéaire. Enfin, le centre de la figure 4.2 illustre l'étape d'explicabilité et les résultats du processus, incluant des cartes et des informations précieuses pour les urbanistes. Les différentes sources de données entrent dans ces zones rouges pour être traitées. Ensuite, les différents affichages comme les graphiques et les tables résultantes sont acheminés aux urbanistes.

4.1.6 Calcul de l'Indice de Vitalité Actuel

L'IVA de l'AD k pour la ville de Shawinigan se calcule comme présenté par l'équation (4.3) :

$$IVA_k = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{4.3}$$

où IVA_k est l'IVA pour l'AD k, n est le nombre de caractéristiques, x_i est la valeur de la caractéristique i pour l'AD k et k est l'AD.

Dans le cadre de cette étude, le calcul de l'IVA inclut 13 caractéristiques (n) réparties sur 87 AD (k). Pour visualiser les différentes valeurs de cet indice, une

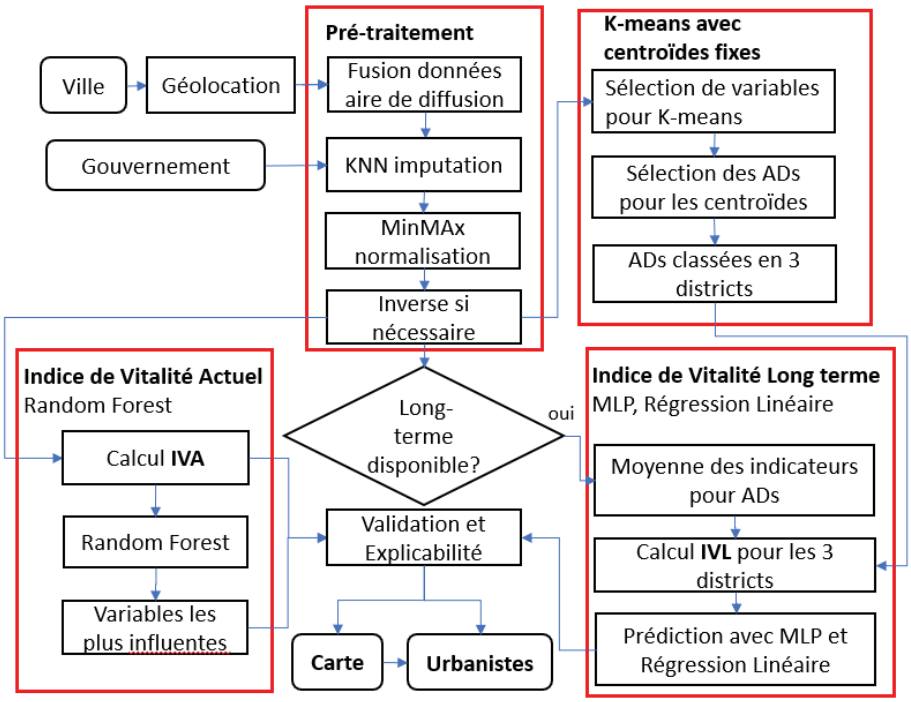


FIGURE 4.2 – Diagramme de méthodologie pour l'IVA et l'IVL

procédure appelée `graph_indice` a été développée. Cette procédure génère un histogramme, permettant ainsi une représentation graphique des données. L'histogramme offre également la possibilité d'évaluer visuellement si l'indice suit une distribution particulière, telle qu'une loi normale. En complément, la procédure `shapiro_wilk` a été mise en place pour effectuer un test d'hypothèse statistique, vérifiant si la distribution de l'indice suit effectivement une loi normale. Si la p-value obtenue est supérieure à 5 %, cela signifie que l'on n'a pas suffisamment de preuves pour rejeter l'hypothèse de normalité, suggérant que l'indice suit probablement une distribution normale. Aussi, différents calculs statistiques comme la moyenne et l'écart-type de l'IVA sont obtenus.

Pour visualiser sur une carte de la ville de Shawinigan les frontières des AD en fonction de leur IVA, une procédure Python a été développée à l'aide de la librairie `geopandas`. Cette carte interactive permet de visualiser les valeurs des indicateurs d'une AD sélectionnée. De plus, les AD sont colorées (heatmap) en fonction de la valeur de l'IVA. Le choix de la variable pour la coloration peut être changé grâce à

l'interface utilisateur. La figure 4.3 présente une carte interactive utilisant un code couleur basé sur l'IVA. Comme l'indique la légende, une couleur pâle (jaune) représente une faible vitalité, tandis qu'une couleur foncée (mauve) correspond à une forte vitalité. Il est possible d'afficher les valeurs des indicateurs en sélectionnant une AD.

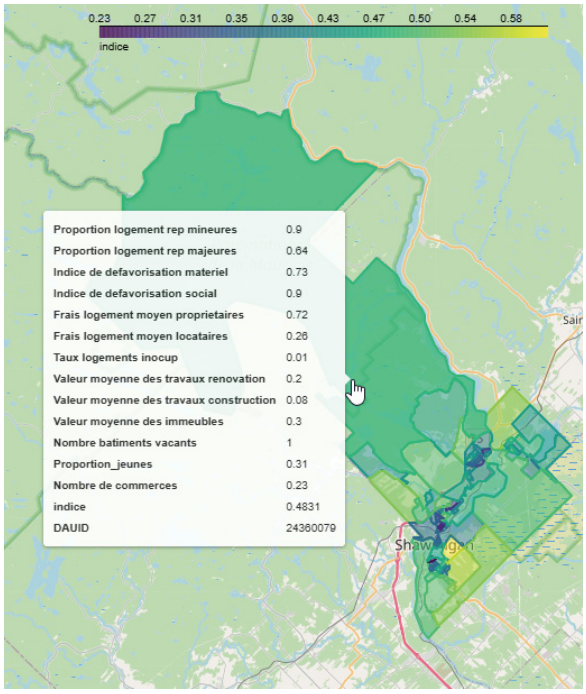


FIGURE 4.3 – Carte interactive avec les indicateurs de chaque AD

Une autre procédure a également été mise en place pour visualiser les frontières des AD à faible indice par rapport à celles à indice élevé. Les AD dont l'IVA est inférieur à un certain seuil sont affichées en rouge, tandis que celles dont l'indice dépasse un certain seuil sont colorées en vert. Ces seuils sont configurables via l'interface, ce qui permet à l'utilisateur de les ajuster en fonction des besoins de l'analyse. Le code source de ces procédures est présenté en Annexe A.

4.1.7 Les caractéristiques les plus significatives

Une question cruciale est de déterminer quelle caractéristique influence le plus l'IVA. Identifier cette caractéristique permettrait aux décideurs de la ville de Shawinigan de cibler leurs investissements dans les AD où l'indice est problématique. Pour ce faire, l'algorithme de régression `RandomForestRegressor`, avec l'IVA comme variable cible, est utilisé. L'algorithme `RandomForestRegressor` de la librairie *scikit-learn* est employé avec le paramètre `n_estimators=500`. Ce paramètre, qui définit le nombre d'arbres dans la forêt, a été fixé à 500 après avoir constaté des changements mineurs dans le classement des caractéristiques à partir de `n_estimators=400`. Finalement, les caractéristiques ont été classées par ordre d'importance et présentées visuellement sous forme d'un diagramme à bâtons, afin de communiquer ces résultats aux urbanistes. Afin d'expliquer les résultats de l'algorithme `RandomForestRegressor` concernant les caractéristiques les plus influentes, les valeurs SHAP seront calculées à partir de l'équation (3.13) et représentées sous forme de graphiques en « violon ». En plus, pour valider les résultats obtenus avec l'algorithme `RandomForestRegressor`, nous utiliserons également l'algorithme `XGBRegressor` de la librairie `xgboost` afin de mesurer l'importance des caractéristiques.

4.1.8 Classement des AD

Pour obtenir un IVL exploitable et permettre une analyse approfondie, il a été nécessaire de faire des regroupements parmi les 87 AD. L'algorithme k-means sera utilisé pour effectuer ce regroupement. L'algorithme k-means, initialement créé par S. Lloyd [32], est une méthode non supervisée de classification. L'algorithme est basé sur la minimisation de la somme des distances au carré entre les données et les centres. Le paramètre `init` de l'algorithme permet de choisir les centroïdes de départ. En effet, afin d'obtenir un classement plus pertinent, les centroïdes sélectionnés correspondront

aux caractéristiques des trois groupes : **Urbain**, **Résidentiel** et **Commercial**. Ce classement facilitera la comparaison et l'interprétation des tendances sur plusieurs années dans ces différents types de secteurs. Ensuite, en collaboration avec les urbanistes de la ville de Shawinigan, les caractéristiques *Densité de la population au kilomètre carré*, *Proportion de jeunes*, *Nombre de commerces* et *Valeur moyenne des immeubles* ont été sélectionnées afin d'optimiser la procédure. Cette réduction de dimensionnalité a été appliquée sur les treize caractéristiques initiales afin d'obtenir un classement cohérent et un bon indice silhouette. Les caractéristiques retenues ont été choisies pour différencier clairement les trois secteurs définis pour les regroupements. Selon les urbanistes, l'indicateur *Densité de la population* permet de caractériser les quartiers urbains, tandis que la caractéristique *Proportion de jeunes* aide à distinguer les quartiers résidentiels. De plus, les caractéristiques *Nombre de commerces* et *Valeur moyenne des immeubles* facilitent le classement des AD dans le secteur commercial.

Pour le choix des centroïdes de départ, une AD correspondant le plus fidèlement possible à chacun des trois groupes a été sélectionnée. De plus, le paramètre `n_clusters` est initialisé à trois en référence aux trois secteurs. La table 4.5 montre les AD choisies ainsi que leur secteur correspondant.

Secteur	**AD modèle**
Urbain	24360101
Résidentiel	24360123
Commercial	24360090

TABLE 4.5 – Choix des AD comme centroïdes

La figure 4.4 montre l'AD modèle du secteur **Urbain**. Le choix de cette AD comme centroïde pour ce groupe est défini en fonction de la densité élevée de la population dans cette agglomération. En effet, les quartiers urbains contiennent habituellement un grand nombre de personnes au kilomètre carré. L'AD 24360101 contient la densité de population la plus élevée de la ville ; c'est donc un bon choix de modèle pour ce secteur.

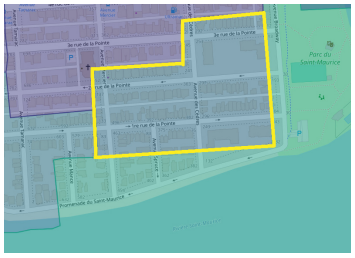


FIGURE 4.4 – **Frontière de l'AD 24360101**

La figure 4.5 montre, en jaune, l'AD modèle du secteur **Résidentiel**. Le choix de cette AD comme centroïde pour ce groupe est défini en fonction du nombre élevé de personnes dont l'âge est inférieur ou égal à 14 ans. Les quartiers résidentiels contiennent beaucoup de familles, alors la proportion de jeunes devrait être grande. L'AD 24360123 est celle qui contient la plus grande proportion de jeunes.

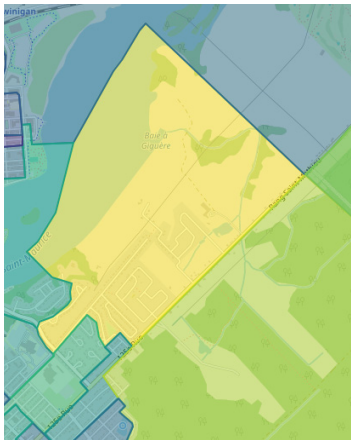


FIGURE 4.5 – Frontière de l'AD 24360123

La figure 4.1.8 montre, en vert pâle, l'AD modèle pour le secteur **Commercial**. Le choix de cette AD comme modèle pour ce groupe est défini en fonction du nombre élevé de commerces. L'aire 24360090 contient le plus grand nombre de commerces de la ville de Shawinigan. En effet, c'est dans cette partie de la ville que se situe le centre commercial *La Plaza de la Mauricie*.

[!htbp] 

Les AD contenues dans les trois secteurs seront utilisées pour calculer l'IVL de ces trois regroupements. Pour évaluer la cohésion et la qualité du regroupement, nous utiliserons l'indice Silhouette. Conçu par P.J. Rousseeuw [31], cet indice mesure la manière dont les points sont bien regroupés au sein de leurs groupes. Il évalue la cohésion en comparant la distance moyenne d'un point à tous les autres points de son propre groupe avec la distance moyenne de ce point aux points du groupe le plus proche. Chaque point se voit attribuer un indice Silhouette individuel, qui reflète dans quelle mesure il est bien assigné à son groupe.

Un indice Silhouette proche de 1 indique que le point est bien regroupé, tandis qu'un indice proche de 0 signifie qu'il se trouve à la frontière entre deux groupes. Un indice négatif, en revanche, suggère que le point pourrait être mal classé, c'est-à-dire qu'il est plus proche d'un autre groupe que de celui auquel il a été assigné. L'indice global de Silhouette correspond à la moyenne des indices individuels de l'ensemble des points. Plus cette moyenne est élevée, meilleure est la séparation entre les groupes.

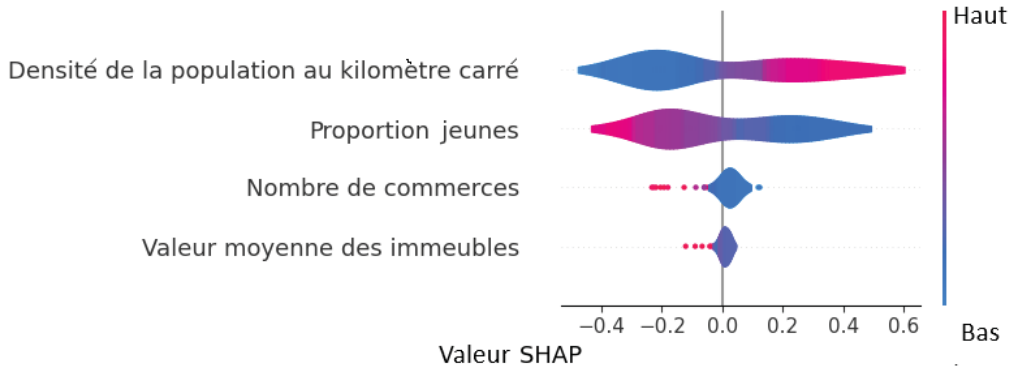
4.1.9 Explicabilité de l'algorithme k-means

Le classement des AD, présenté à la section 4.1.8, sera justifié à l'aide des valeurs SHAP. Un graphique distinct sera généré pour chaque secteur défini à la section 4.1.8, à savoir : Urbain, Résidentiel et Commercial. Ces types de secteurs ont été choisis

par les urbanistes de la ville de Shawinigan. En ce qui concerne l'interprétation d'un graphique des valeurs SHAP, tel que montré à la figure 4.1.9, il est important de noter que les points rouges correspondent à des valeurs élevées des caractéristiques, tandis que les points bleus représentent des valeurs plus faibles. Les valeurs positives, situées à droite de l'axe principal, indiquent les contributions qui tendent à classer les observations dans le groupe représenté sur le graphique, dans ce cas-ci, le groupe Urbain.

De plus, un graphique de type nuage de points sera généré pour appuyer l'analyse des valeurs SHAP. Les nuages de points de cette figure illustrent les relations entre chaque paire de variables, avec une coloration des points selon le secteur. Cela permet d'identifier, à travers les regroupements observés, quelles caractéristiques influencent ces regroupements. Sur la diagonale, les histogrammes montrent les distributions individuelles des variables. Dans les nuages de points, l'attention sera portée sur l'identification de groupes distincts associés à un même secteur. Pour les histogrammes, nous chercherons à observer une séparation nette entre les courbes, ce qui pourrait indiquer des différences significatives entre les secteurs. Finalement, des graphiques radars seront produits pour vérifier la cohérence interne de chaque regroupement.

Ainsi, à l'aide de ces outils, les décideurs de la ville pourront mieux comprendre les facteurs influençant le classement généré par l'algorithme et, par conséquent, mieux appréhender les raisons pour lesquelles l'algorithme RF a produit ces résultats. Grâce à cette analyse, ils seront en mesure d'identifier les caractéristiques qui impactent le classement et de prendre des décisions plus éclairées en fonction de ces informations.



[!htbp]

4.1.10 Indice de Vitalité à Long Terme

Pour la ville de Shawinigan, il est essentiel de pouvoir suivre l'évolution de la vitalité des AD sur une période prolongée. Les données des recensements de 2006 à 2021 nous fournissent des informations sur plusieurs indicateurs couvrant une période de quinze ans. Étant donné que la ville est divisée en 87 AD, il n'est pas optimal de les suivre individuellement, car une telle analyse deviendrait complexe. Il est donc préférable de regrouper ces aires afin de simplifier l'interprétation des résultats. Les groupes identifiés sont les secteurs **Urbain**, **Résidentiel** et **Commercial**. L'algorithme k-means a été utilisé pour regrouper les AD selon ces trois secteurs distincts. La méthodologie employée pour ce regroupement est détaillée dans la section 4.1.8. Par la suite, pour chacun de ces trois secteurs, des prévisions pourront être réalisées pour les années à venir à l'aide d'un MLP ou d'une régression linéaire.

Pour concevoir un IVL, toutes les caractéristiques de l'IVA n'ont pas pu être utilisées en raison de limitations dans leur disponibilité. Ainsi, les caractéristiques retenues pour cette évaluation à long terme sont présentées dans la table 4.6 :

Variable	**Source**
Indice de défavorisation sociale	INSPQ, 2021, 2016, 2011, 2006
Indice de défavorisation matérielle	INSPQ, 2021, 2016, 2011, 2006
Valeur moyenne des rénovations	Ville de Shawinigan, 2021, 2016, 2011, 2006
Valeur moyenne des travaux de construction	Ville de Shawinigan, 2021, 2016, 2011, 2006

TABLE 4.6 – Variables sélectionnées pour l'IVL

Comme pour l'IVA, un certain traitement a été effectué sur ces indicateurs, présenté dans la table 4.7. En effet, il faut normaliser toutes les données pour que les caractéristiques soient sur la même échelle. La méthode utilisée est MinMax, ce qui permet d'avoir toutes les valeurs entre 0 et 1. De plus, il est nécessaire d'inverser les caractéristiques *Indice de défavorisation matérielle* et *Indice de défavorisation sociale*, car une valeur de défavorisation forte signifie une vitalité faible de l'AD. Aussi, certaines données sont manquantes pour les caractéristiques *Indice de défavorisation matérielle* et *Indice de défavorisation sociale*. Il est donc nécessaire de combler ces valeurs à l'aide de l'algorithme `KNNimputer` de la librairie *scikit-learn*.

Traitement	**caractéristiques**
MinMax	Toutes les caractéristiques
Inversion	Indice de défavorisation matérielle et sociale
Imputation	Indice de défavorisation matérielle et sociale

TABLE 4.7 – Traitement des caractéristiques pour l'IVL

Pour permettre un calcul d'indice, une moyenne pour chacune des quatre caractéristiques de la table 4.6 est calculée pour chaque AD et pour chaque année. Les années considérées sont celles des recensements fédéraux : 2006, 2011, 2016 et 2021. Ensuite, une moyenne est effectuée pour les AD pour les secteurs Urbain, Résidentiel et Commercial. La formule (4.4) est utilisée pour ce calcul :

$$IVL_s = \frac{1}{N} \sum_{i=1}^{N} IVA_i, \tag{4.4}$$

où IVL_s est l'indice à long terme moyen pour le secteur s, N est le nombre d'AD dans le secteur s, IVA_i est la valeur de l'indice pour l'AD i et s est le secteur où $s \in \{1, 2, 3\}$. Dans le cadre de cette étude, le calcul de l'IVL comporte quatre caractéristiques. De plus, les 87 AD sont regroupées en trois secteurs s. Les trois secteurs obtiennent ainsi un indice moyen $Indice_s$ pour les années 2006, 2011, 2016 et 2021. C'est l'IVL pour la ville de Shawinigan. Ensuite, la méthode neuronale d'IA Perceptron multicouche (MLP) sera utilisée pour effectuer une régression pour l'année 2026. L'algorithme `MLPRegressor` de la librairie *scikit-learn* est utilisé. L'argument `hidden_layer_sizes(50, 25)` est utilisé pour configurer le réseau avec deux couches cachées. La première couche cachée contient 50 neurones, et la deuxième en contient 25. Pour valider les prédictions, la méthode des moindres carrés sera employée comme métrique d'évaluation. Les modèles MLP, RF et Régression linéaire seront évalués. Elle est calculée selon la formule (4.5). Les calculs sont effectués à l'aide de la fonction `mean_squared_error` de la librairie `sklearn.metrics`.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{4.5}$$

où MSE est l'erreur quadratique moyenne, n est le nombre total d'observations, y_i est la valeur réelle pour l'observation i, $\hat{y}_i$ est la valeur prédite pour l'observation i et $\sum_{i=1}^{n}$ est la somme des erreurs au carré pour chaque observation.

4.2 Résultats

4.2.1 Analyse statistique

La figure 4.6 représente la matrice des corrélations des données pour l'année 2021.

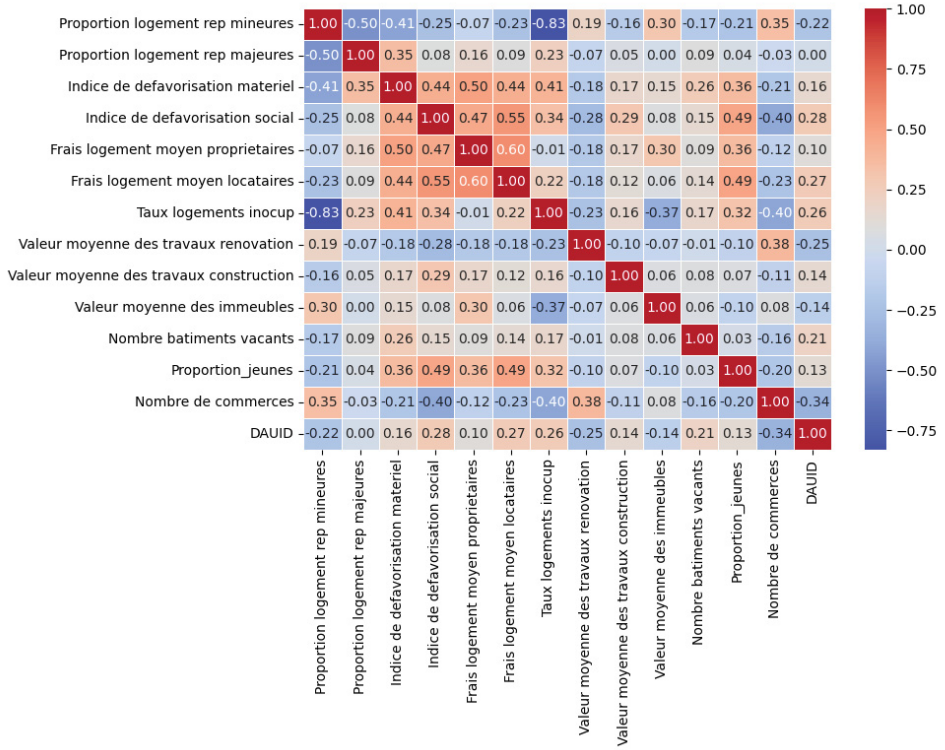


FIGURE 4.6 – Matrice des corrélations des 13 variables

Les histogrammes des figures 4.7 et 4.8 illustrent la répartition des différentes variables. Aussi, les figures 4.9, 4.10 et 4.11 sont des diagrammes à batons permettant de mettre en évidence certains biais.

La figure 4.12 montre des graphiques de type « boîtes à moustache » pour chacune des treize variables analysées.

4.2.2 Analyse en composantes principales

L'ACP est une méthode de réduction de la dimensionnalité qui permet de simplifier un jeu de données tout en préservant l'essentiel de l'information. Elle repose sur la transformation des variables initiales, souvent corrélées entre elles, en un ensemble de nouvelles variables non corrélées appelées « composantes principales ». En projetant

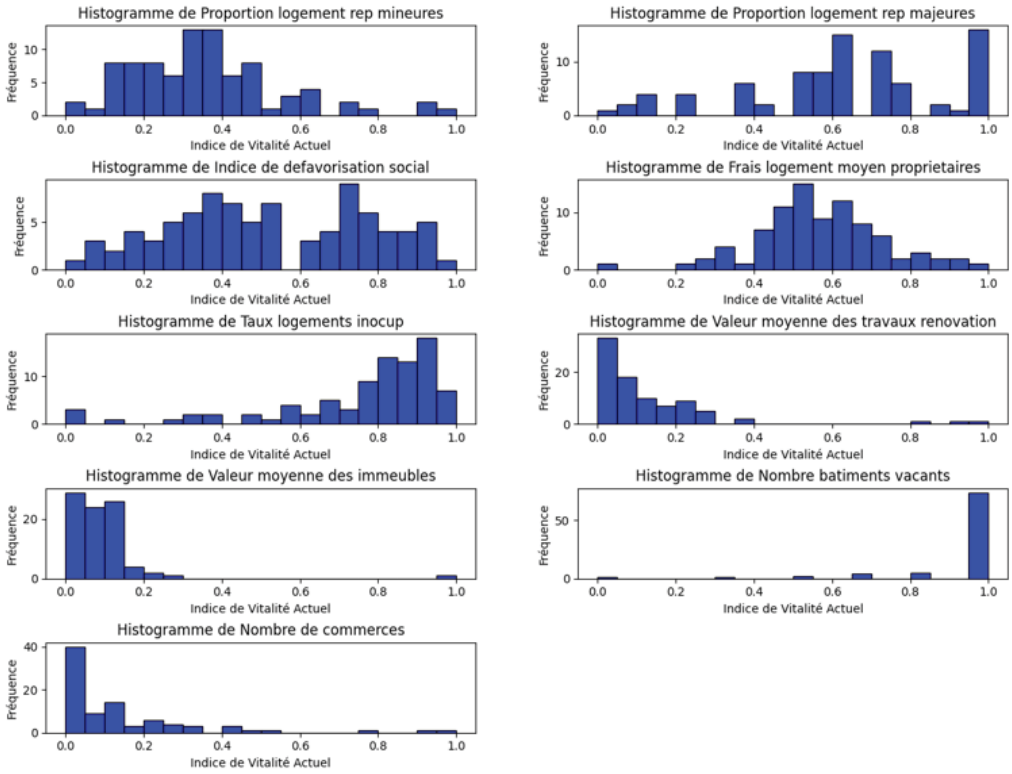


FIGURE 4.7 – Distribution des 13 variables

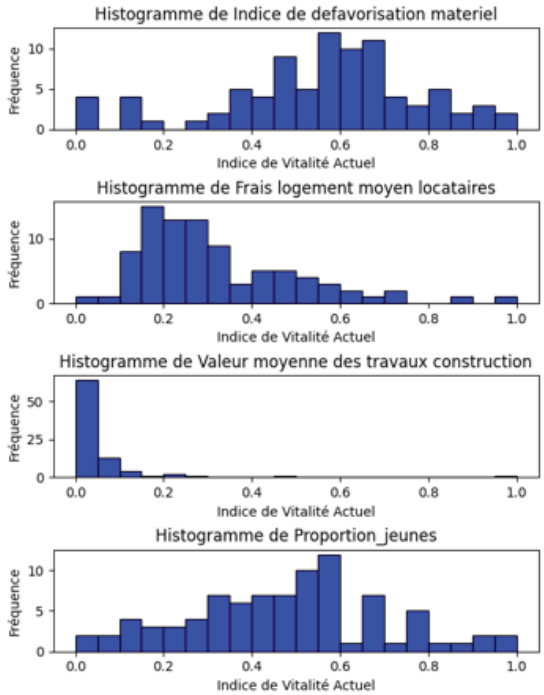


FIGURE 4.8 – Distribution des 13 variables (suite)

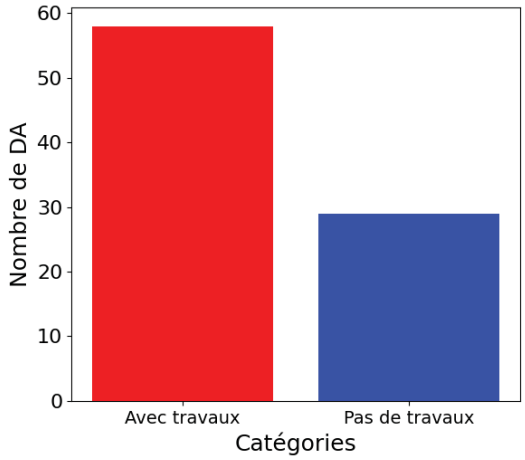


FIGURE 4.9 – Comparaison des travaux de construction selon les secteurs

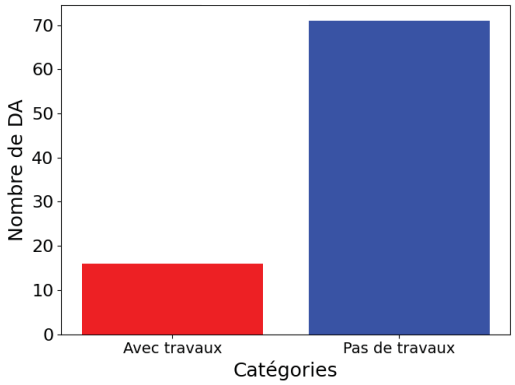


FIGURE 4.10 – Comparaison des travaux de rénovation selon les secteurs

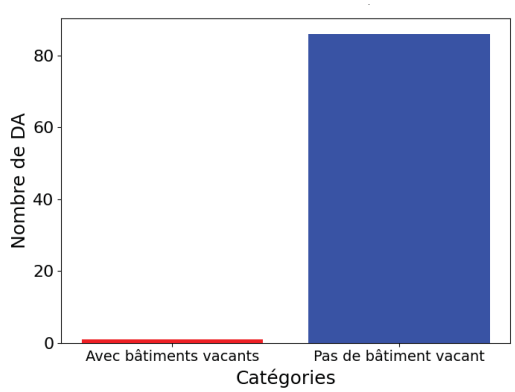


FIGURE 4.11 – Comparaison du nombre de bâtiments vacants selon les secteurs

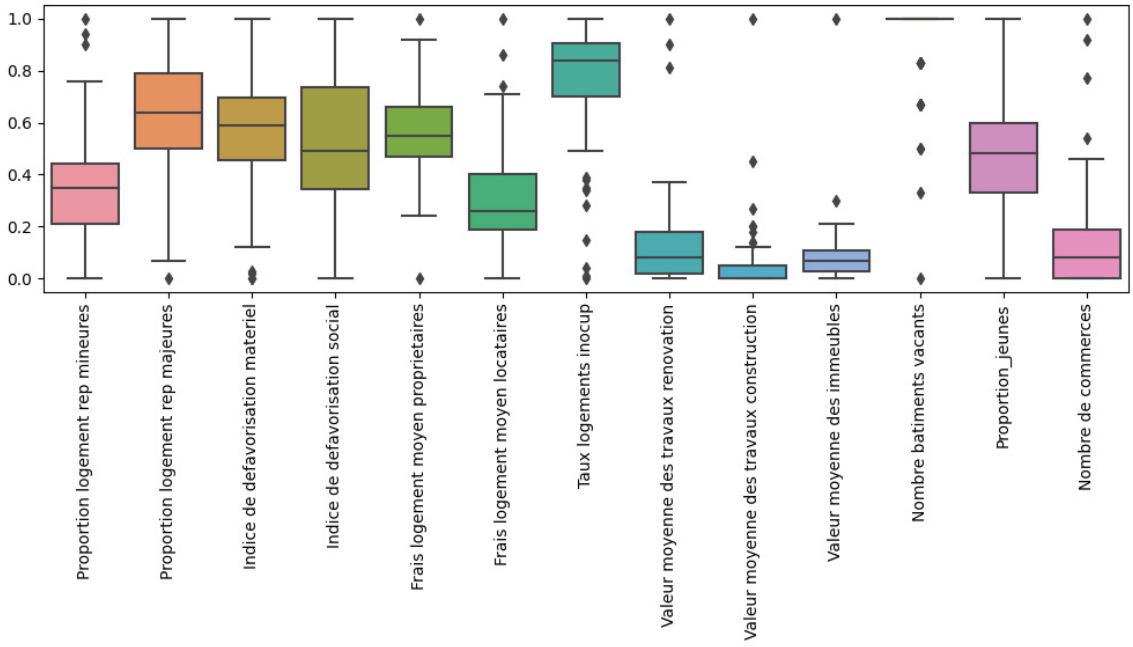


FIGURE 4.12 – Boîtes à moustache des variables de l'IVA

les données initiales sur ces deux composantes principales, l'ACP facilite la visualisation et l'interprétation des relations entre les variables et les observations. Ainsi, elle fournit une vue simplifiée des données tout en conservant une certaine structure globale, ce qui est particulièrement utile pour l'analyse. Les valeurs propres obtenues dans le cadre de l'ACP sont associées aux différentes composantes principales et représentent la part de la variance totale expliquée par chacune de ces composantes. Les valeurs propres illustrées dans la figure 4.13 permettent d'évaluer le nombre de composantes principales à retenir pour capturer l'essentiel de la structure des données.

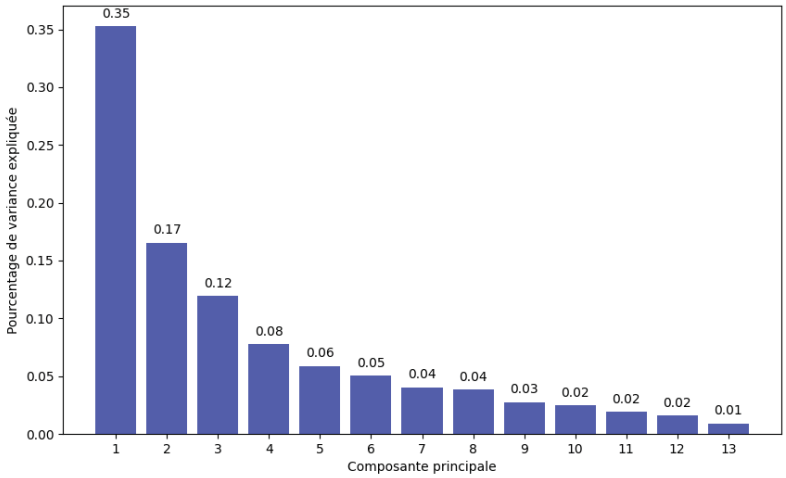


FIGURE 4.13 – Pourcentage d'explication des composantes principales

En se basant sur le critère de Kaiser (1961), seules les composantes ayant des valeurs propres supérieures à 1 sont géréralement conservées, car elles expliquent une proportion significative de la variance totale. On remarque que, si l'on retient les deux premières composantes, il est possible de représenter les données des AD de la ville de Shawinigan sur deux axes. La figure 4.14 montre cette représentation.

La figure 4.15 représente le graphique circulaire de L'ACP. Dans ce graphique, chaque flèche représente une variable, et la direction et la longueur de la flèche indiquent la relation entre la variable et les axes principaux. Les variables qui sont proches d'un axe ont une forte corrélation avec cet axe principal, tandis que celles qui sont plus éloignées ont une corrélation plus faible.

Dans la représentation des données des AD de la ville de Shawinigan sur deux axes à l'aide l'ACP, la figure 4.16 montre, en vert, une zone de bonne vitalité.

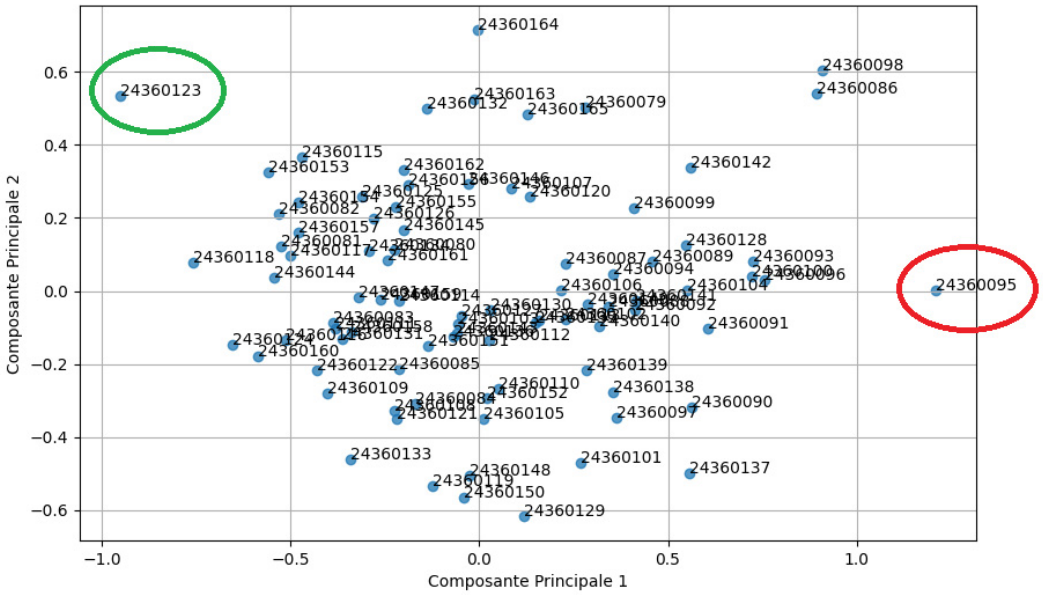


FIGURE 4.14 – Projection des données selon les deux premiers axes

En ne gardant que les deux premières composantes. Il est possible d'établir la contribution des caractéristiques à ces composantes. Cette contribution est présentée dans la table 4.8.

4.2.3 Indice de Vitalité Actuel

La figure 4.17 montre un diagramme à bâtons représentant l'IVA, en valeur continue, pour chaque AD. Pour examiner la distribution de l'IVA, il est approprié de construire un histogramme. En choisissant de grouper en huit classes, la figure 4.18 est ainsi obtenue.

Les données concernant l'indice de vitalité sous la forme d'un histogramme, figure 4.18, suggèrent une distribution normale. Pour vérifier cette hypothèse, un test de Shapiro Wilk est effectué sur les résultats du calcul de l'indice. À l'aide de la librairie `scipy.stats`, une p-value de 0.99 est obtenue, ce qui est supérieur à 0.05. Alors, on ne

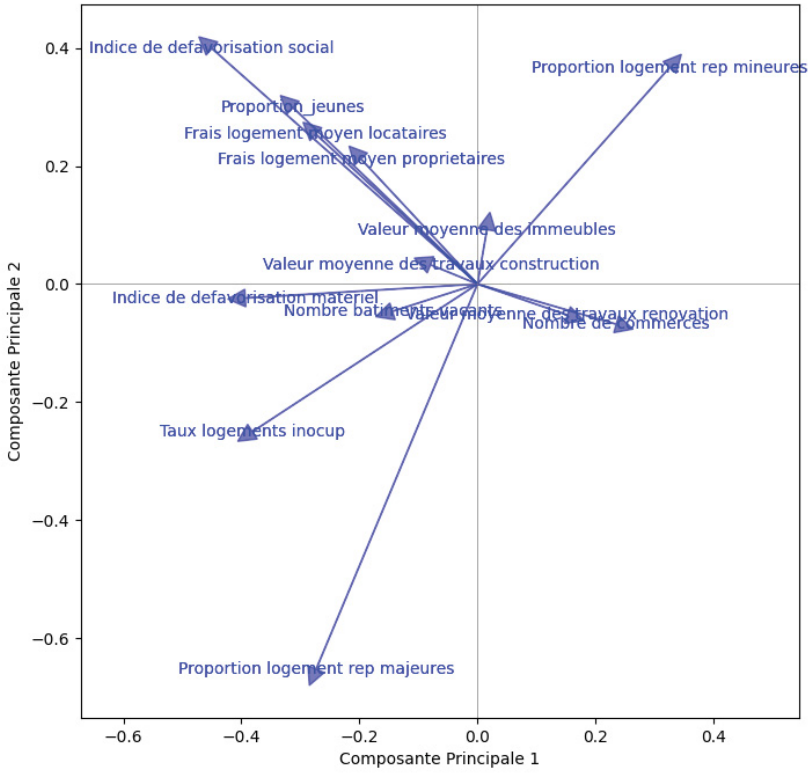


FIGURE 4.15 – Graphique circulaire de l'ACP

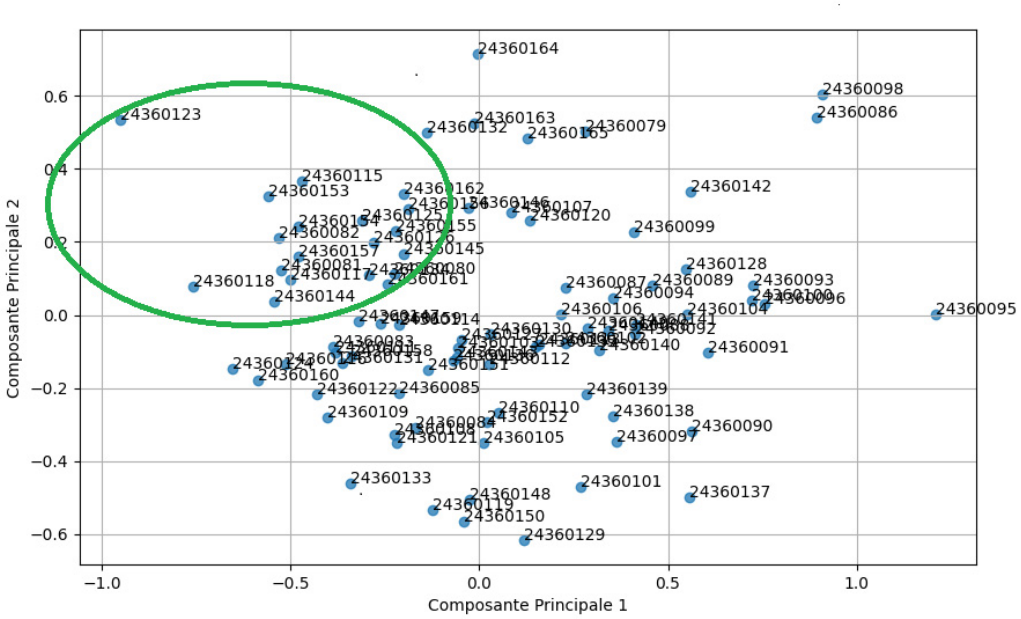


FIGURE 4.16 – Zone de bonne vitalité

TABLE 4.8 – Contribution des caractéristiques aux deux premières composantes

Variable	**Composante 1**	**Composante 2**
Proportion logement réparations mineures	0.3259	0.3670
Proportion logement réparations majeures	-0.2783	-0.6639
Indice de défavorisation matérielle	-0.3992	-0.0317
Indice de défavorisation social	-0.4523	0.3972
Frais logement moyen propriétaires	-0.2034	0.2039
Frais logement moyen locataires	-0.2806	0.2469
Taux logements inoccupés	-0.3765	-0.2438
Valeur moyenne des travaux de rénovation	0.1547	-0.0517
Valeur moyenne des travaux de construction	-0.0799	0.0324
Valeur moyenne des immeubles	0.0158	0.0903
Nombre bâtiments vacants	-0.0685	-0.0165
Proportion jeunes	-0.3178	0.2940
Nombre de commerces	0.2329	-0.0726

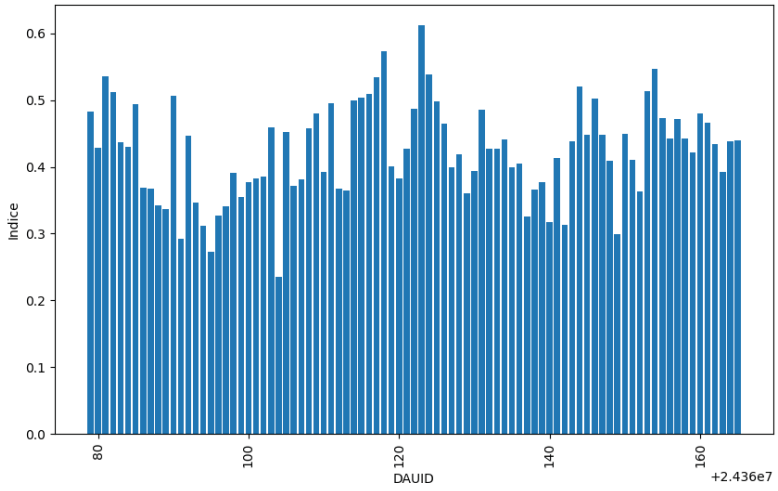


FIGURE 4.17 – Indice de vitalité pour chaque AD

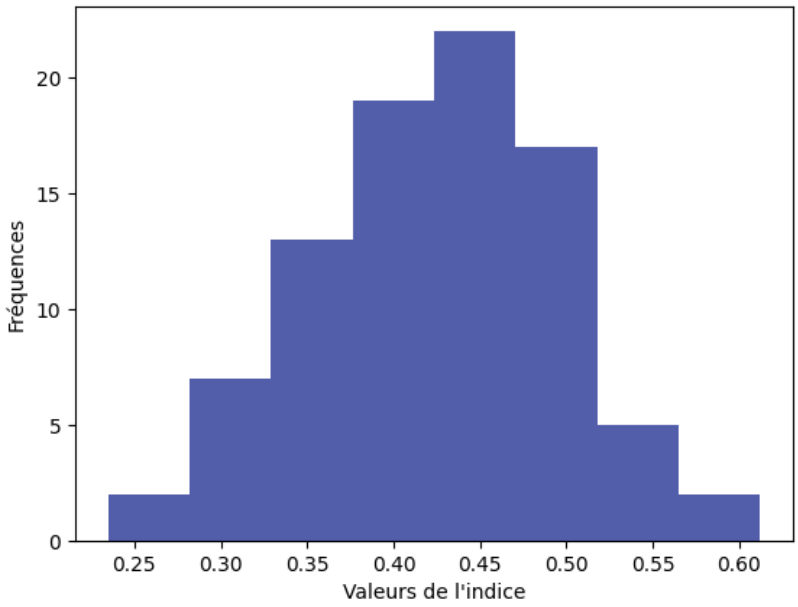


FIGURE 4.18 – Distribution de l'IVA

peut rejeter l'hypothèse selon laquelle l'indice de vitalité est distribué normalement. Les données concernant l'indice de vitalité sont donc distribuées normalement avec une moyenne de 0.4240 et un écart type de 0.0716.

Pour étudier la vitalité des AD, il est possible de voir les endroits où la vitalité est la plus grande ainsi que les endroits où elle est la plus petite. Ainsi, les cinq plus grandes valeurs de l'indice et les cinq plus petites sont présentées dans la table 4.9.

TABLE 4.9 – Valeurs extrêmes d'IVA et les AD correspondantes

AD	**Indices IVA extrêmes**
24360104	0.2346
24360095	0.2723
24360091	0.2923
24360149	0.2992
24360094	0.3115
24360081	0.5362
24360124	0.5392
24360154	0.5469
24360118	0.5731
24360123	0.6123

Il est important de voir où sont situées ces AD. Pour le permettre, une procédure a été créée et elle sera utilisée pour visualiser ces endroits. En effet, les cinq AD ayant le meilleur indice sont en vert et les cinq AD ayant l'indice le plus faible sont en rouge sur la figure 4.19. En guise de complément d'information, la table 4.10, présente les indicateurs pour l'AD 24360104 ayant le plus faible IVA.

L'AD 24360104 représenté sur la figure 4.20 est le quartier avec le plus faible IVA. Les variables **Valeur moyenne des travaux construction**, **Nombre de commerces**, **Valeur moyenne des immeubles** et **Nombre batiments vacants** sont très faibles ou égale à zéro. À l'autre extrémité du spectre des indices de vitalité se trouve l'AD 24360123, figure 4.21, avec un indice favorable de 0.6123.

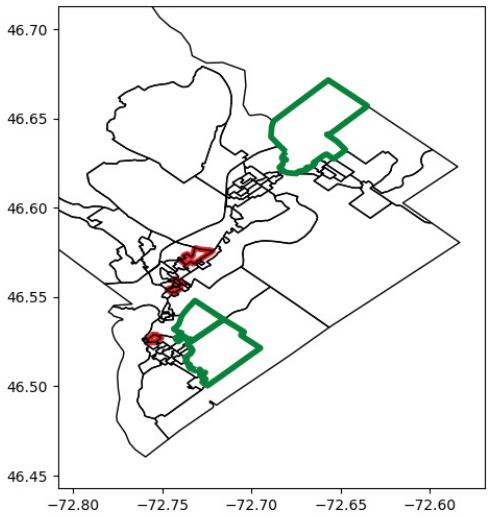


FIGURE 4.19 – Carte des AD avec des IVA élevés et faibles

TABLE 4.10 – table des indicateurs pour l'AD 24360104

Variable	**Indice**
Proportion logement réparations mineures	0.41
Proportion logement réparations majeures	0.43
Indice de défavorisation matérielle	0.47
Indice de défavorisation social	0.20
Frais logement moyen propriétaires	0.54
Frais logement moyen locataires	0.19
Taux logements inoccupé	0.55
Valeur moyenne des travaux rénovation	0.06
Valeur moyenne des travaux construction	0.00
Valeur moyenne des immeubles	0.01
Nombre bâtiments vacants	0
Proportion jeunes	0.19
Nombre de commerces	0
IVA	0.2346

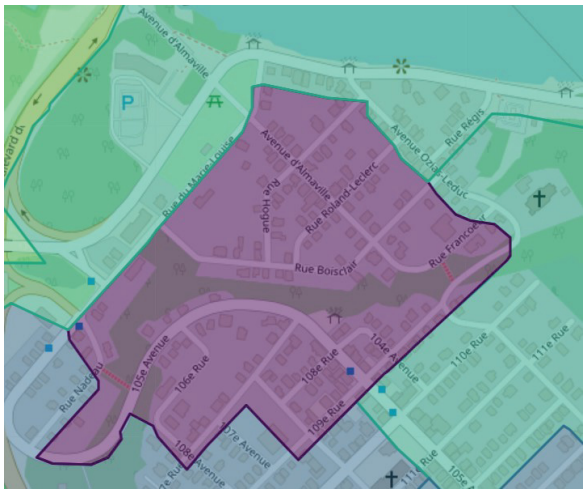


FIGURE 4.20 – Frontières de l'AD 24360104

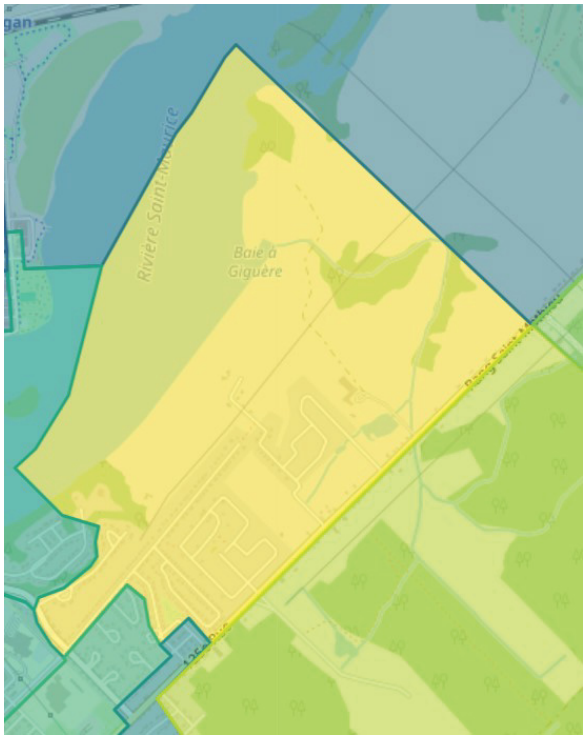


FIGURE 4.21 – Frontières de l'AD 24360123

4.2.4 Importance des caractéristiques

La table 4.11 ainsi que la figure 4.22 présentent l'importance des caractéristiques calculée à l'aide de la méthode RF. Cette importance est déterminée en permutant les valeurs d'une variable et en mesurant la dégradation de la performance du modèle qui en résulte. Pour ce travail, l'algorithme `RandomForestRegressor` a été utilisé. La variable cible de cet algorithme, l'IVA, a été sélectionnée. L'algorithme `RandomForestregressor` de la librairie *scikit-learn* peut être employé avec le paramètre `n_estimators=500`. Ce paramètre, qui correspond au nombre d'arbres, a été ajusté à 500 après avoir constaté des changements mineurs dans le classement des caractéristiques à partir de `n_estimators=400`. Pour valider ces résultats et identifier les caractéristiques la plus influentes, l'algorithme XGBoost a également été employé. Comme l'illustre la figure 4.23, les caractéristiques **Indice de défavorisation matérielle** et **Indice de défavorisation social** se démarquent encore une fois comme étant les plus importantes.

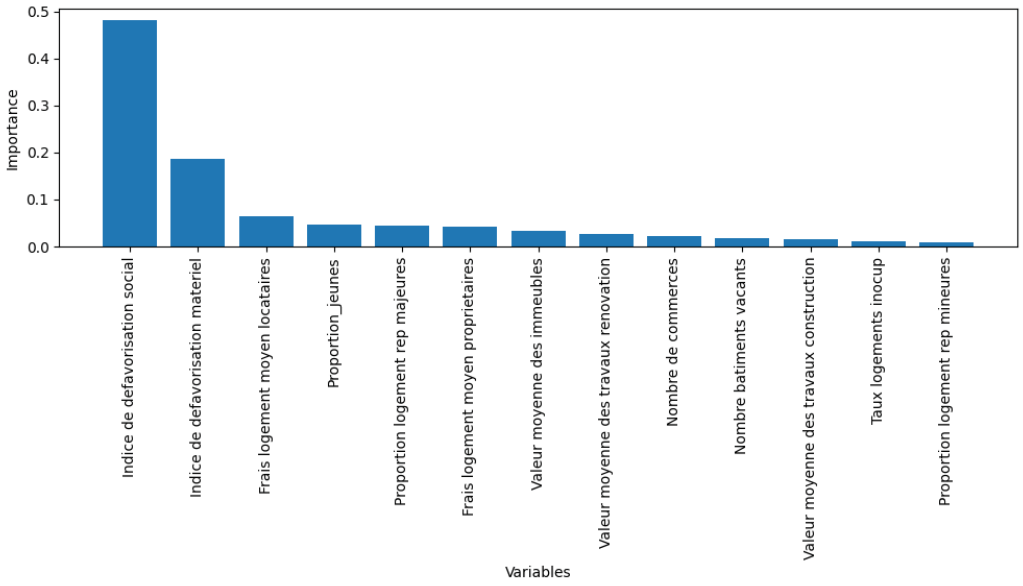


FIGURE 4.22 – Importance des caractéristiques avec RF

TABLE 4.11 – Importance des caractéristiques avec RF

Variable	**Importance**
Indice de défavorisation sociale	0.4808
Indice de défavorisation matérielle	0.1864
Frais logement moyen locataires	0.0637
Proportion jeunes	0.0477
Proportion logements réparations majeures	0.0453
Frais logement moyen propriétaires	0.0422
Valeur moyenne des immeubles	0.0328
Valeur moyenne travaux rénovation	0.0259
Nombre de commerces	0.0234
Nombre de bâtiments vacants	0.0171
Valeur moyenne travaux construction	0.0158
Taux logements inoccupés	0.0109
Proportion logements réparations mineures	0.0081

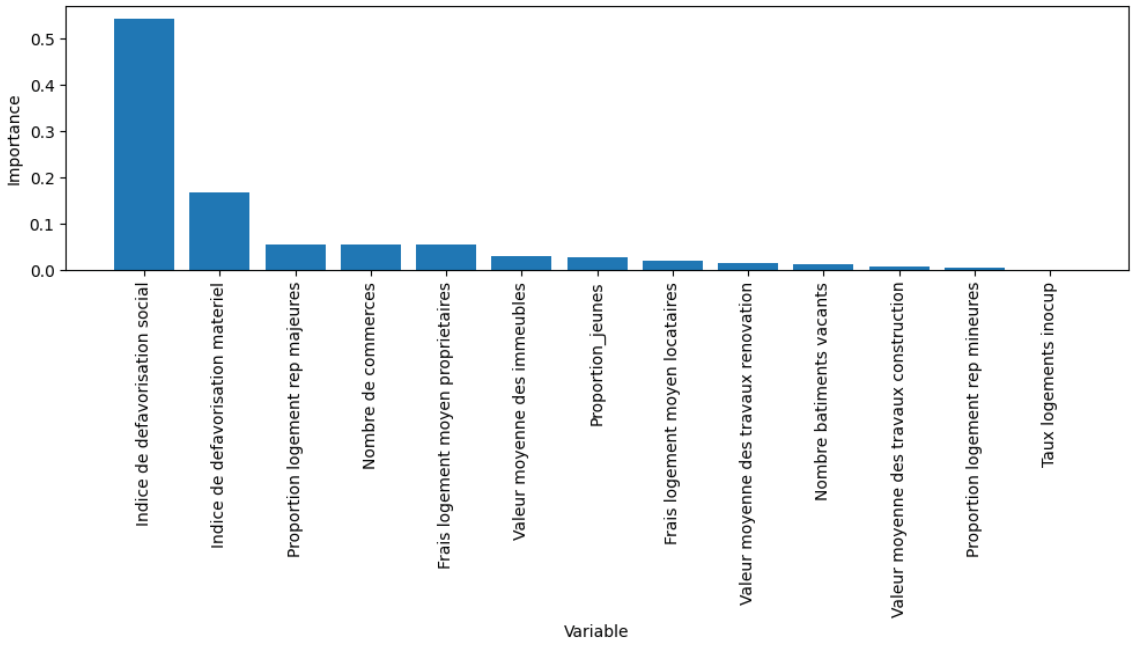


FIGURE 4.23 – Importance des caractéristiques avec XGBoost

4.2.5 Valeurs SHAP

La figure 4.24 illustre l'utilisation des valeurs SHAP pour évaluer l'importance des différents indicateurs.

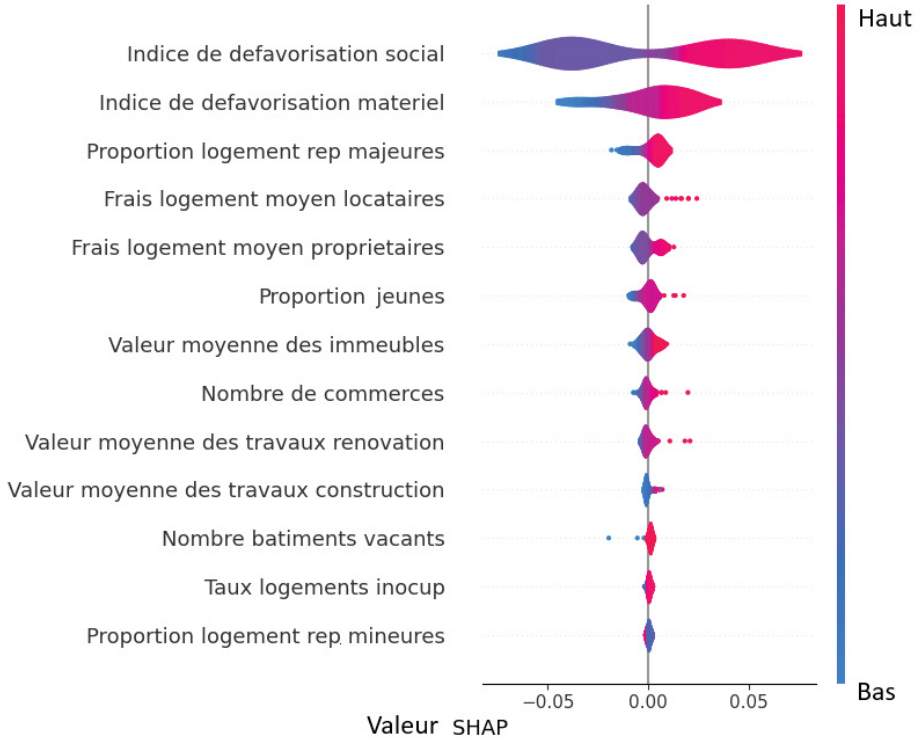


FIGURE 4.24 – Valeurs SHAP pour RF

4.2.6 Répartition des AD selon l'Indice de vitalité

La figure 4.25 montre la carte des AD en fonction de leur IVA. La couleur de chaque AD reflète la valeur de son indice : plus la teinte se rapproche du mauve, plus l'Indice de Vitalité est faible. À l'inverse, une teinte tendant vers le jaune indique une vitalité plus élevée, comme le montre l'échelle située en haut de la carte. En premier lieu, on peut remarquer que les quartiers centraux sont en mauve, ce qui indique une faible vitalité.

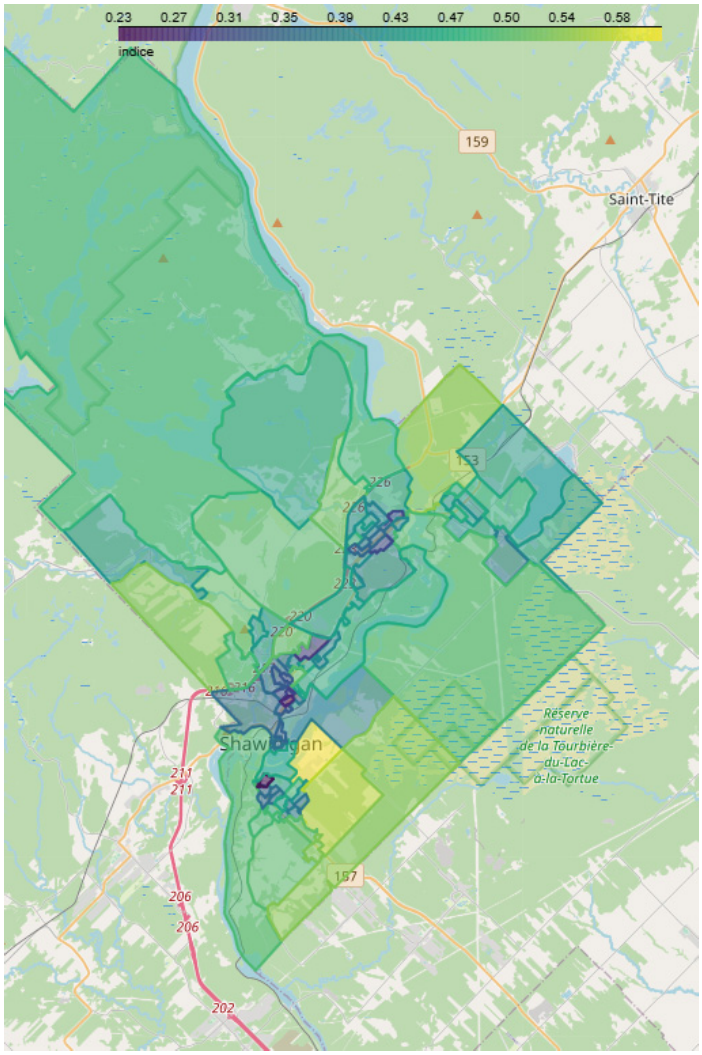


FIGURE 4.25 – Carte de l'IVA en 2021

4.2.7 Classement avec k-means

Avec l'algorithme k-means, il est possible de classer les AD. Pour le classement des AD de la ville de Shawinigan, il a été décidé de fixer les centroïdes de départ afin d'obtenir une classification alignée avec les besoins de la ville. L'objectif est de créer des groupes ayant une signification concrète pour les urbanistes, c'est-à-dire de former des regroupements de quartiers en fonction de critères pertinents. Ainsi, les centroïdes ont été choisis pour diviser les AD en trois groupes, classés selon les types de quartiers : Urbain, Résidentiel et Commercial. Les centroïdes de la méthode k-means ayant été choisis en tenant compte de ces types de quartiers, il a fallu sélectionner des AD modèles représentant le plus fidèlement possible ces différentes catégories, comme mentionné dans la table 4.5. La table 4.12 présente les résultats du classement effectué à l'aide de l'algorithme k-means. Cette table permet d'observer les numéros des AD regroupés par type de quartier, montrant ainsi à quel groupe chaque AD appartient. De plus, la figure 4.26 illustre la répartition des AD sur la carte de la ville de Shawinigan. Dans cette figure, les unités sur les axes représentent la latitude et la longitude.

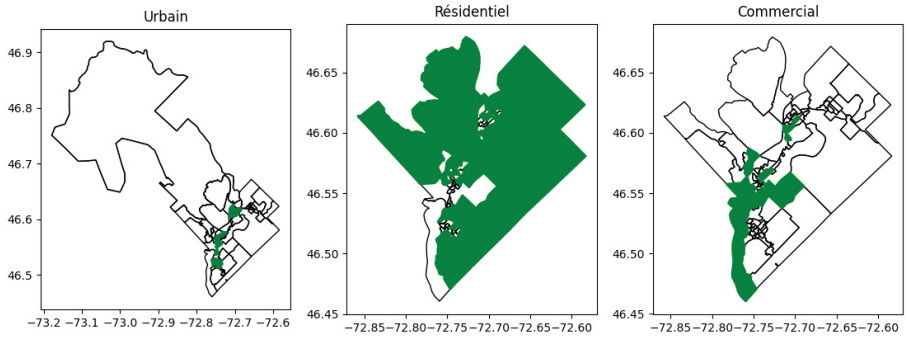


FIGURE 4.26 – Cartes des AD contenues dans chaque secteur

TABLE 4.12 – Les AD contenues dans chaque secteur

Type de secteur	**AD**
Urbain	24360113, 24360137, 24360093, 24360131, 24360101, 24360152, 24360119, 24360134, 24360105, 24360097, 24360096, 24360104, 24360091, 24360110, 24360102, 24360130, 24360121, 24360088, 24360107, 24360092, 24360094, 24360140, 24360112, 24360142, 24360133, 24360120, 24360106, 24360141, 24360149, 24360108, 24360129, 24360099, 24360089, 24360095
Résidentiel	24360132, 24360155, 24360082, 24360154, 24360146, 24360143, 24360136, 24360123, 24360164, 24360162, 24360083, 24360151, 24360158, 24360159, 24360157, 24360087, 24360127, 24360165, 24360145, 24360122, 24360139, 24360163, 24360080, 24360081, 24360109, 24360125, 24360115, 24360111, 24360118, 24360135, 24360116, 24360103, 24360156, 24360144, 24360160, 24360126, 24360124, 24360148, 24360153, 24360161, 24360117, 24360147, 24360085
Commercial	24360084, 24360098, 24360114, 24360150, 24360128, 24360086, 24360079, 24360100, 24360090, 24360138

4.2.8 Explicabilité du classement k-means avec les valeurs SHAP

Les figures 4.27 et 4.28 illustrent l'explication du classement des AD à l'aide des valeurs SHAP pour les secteurs urbain et commercial.

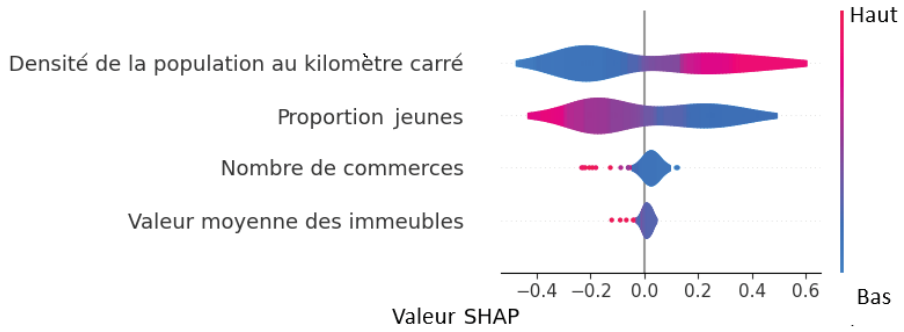


FIGURE 4.27 – Valeurs SHAP pour secteur Urbain

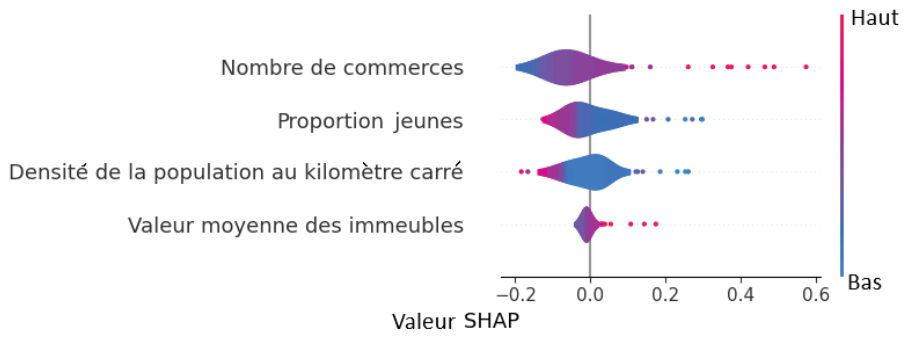


FIGURE 4.28 – Valeurs SHAP pour secteur Commercial

Dans la figure 4.29, les valeurs SHAP moyennes montrent que les variables **Proportion de jeunes** et **Densité de population** sont les plus influentes pour les catégories Urbain et Résidentiel. En ce qui concerne la catégorie Commercial, c'est la variable **Nombre de commerces** qui joue le rôle prédominant. Enfin, l'indicateur **Valeur moyenne des immeubles** n'a que peu d'influence sur le classement global.

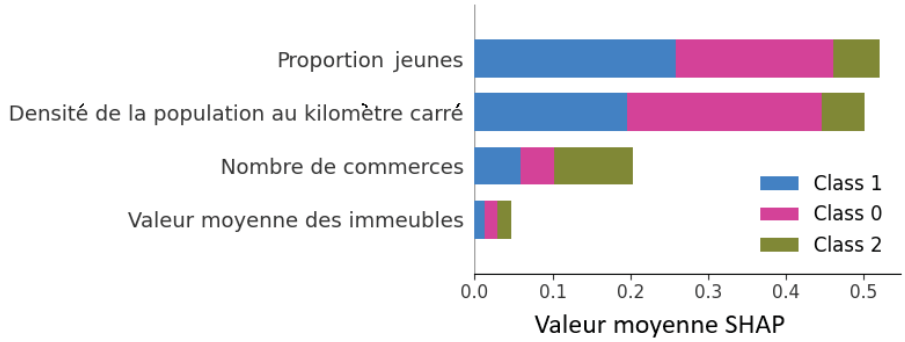


FIGURE 4.29 – Valeurs SHAP moyennes pour l'IVL

Dans la figure 4.29 : Classe 0 représente la catégorie **Urbain**, Classe 1 représente la catégorie **Résidentiel** et Classe 2 représente la catégorie **Commercial**.

4.2.9 Explicabilité du classement k-means avec des nuages de points

Dans la figure 4.30, chaque nuage de points, représentant une paire de variables, est coloré selon le secteur d'appartenance : Urbain, Résidentiel ou Commercial.

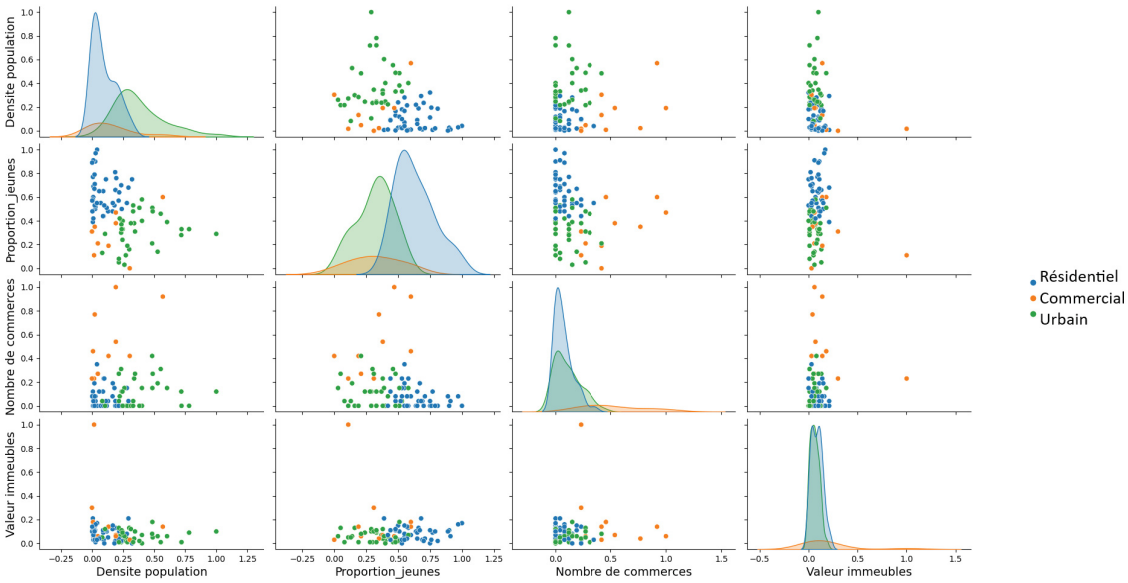


FIGURE 4.30 – Nuage de points de paires de variables de l'IVL

4.2.10 Consistance des regroupements

Les figures 4.31, 4.32 et 4.33 présentent un graphique radar pour chacun des regroupements.

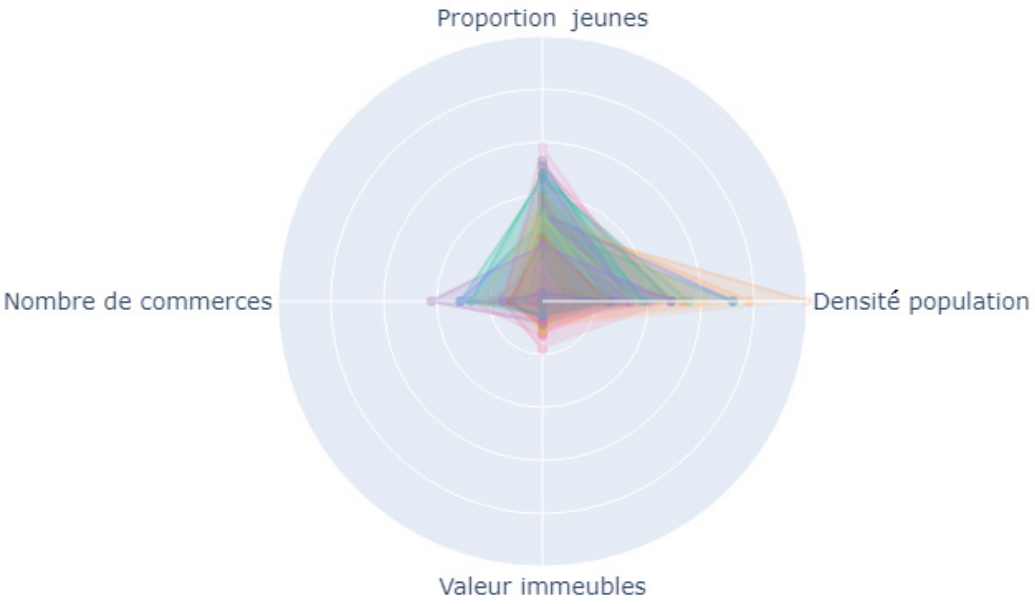


FIGURE 4.31 – Graphique radar - Urbain

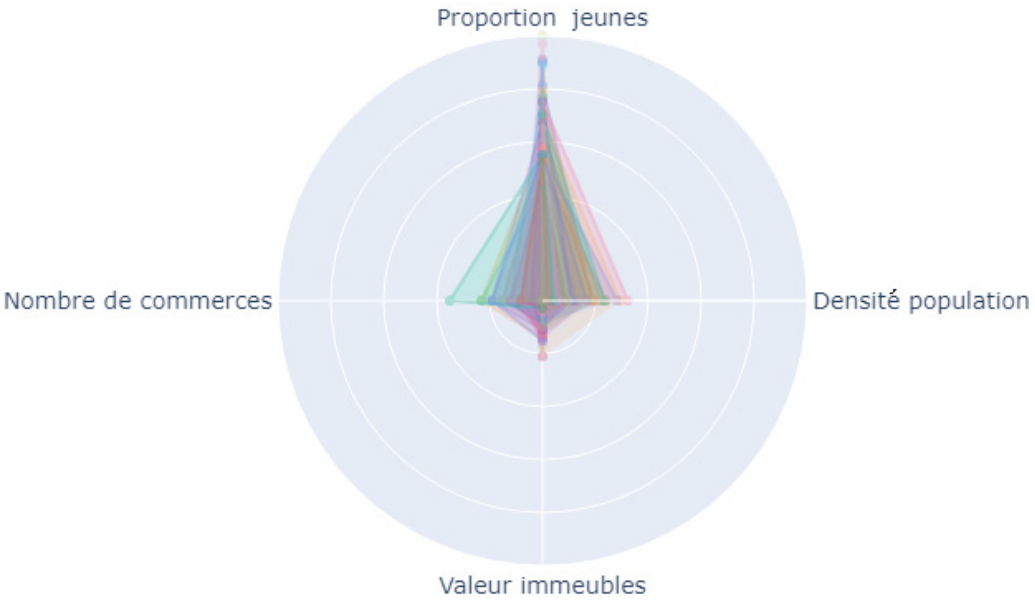


FIGURE 4.32 – Graphique radar - Résidentiel

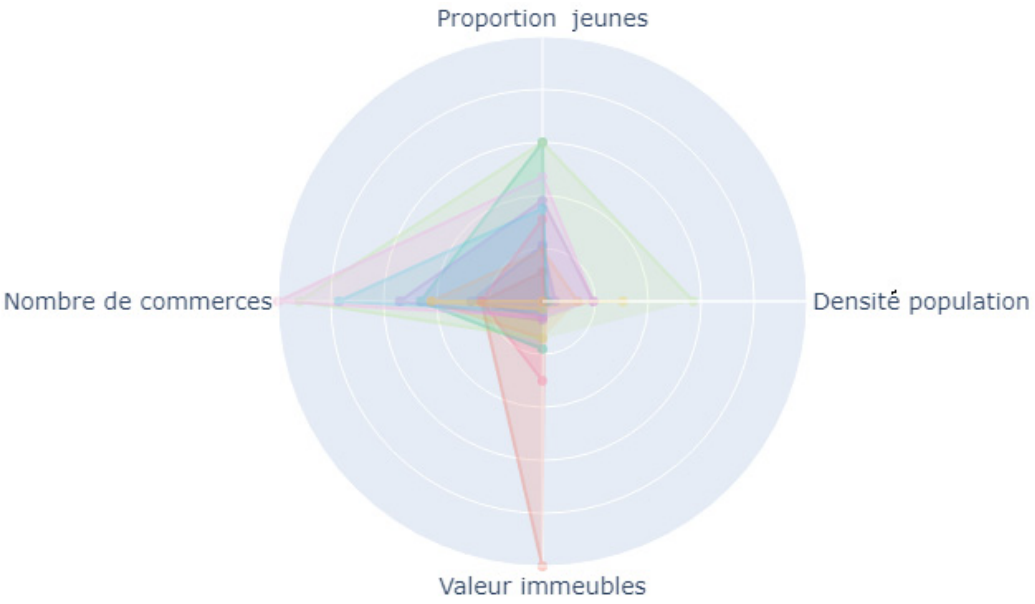


FIGURE 4.33 – Graphique radar - Commercial

La figure 4.34 montre que l'indice de Silhouette moyen obtenu par l'algorithme k-means est de 0.3229.

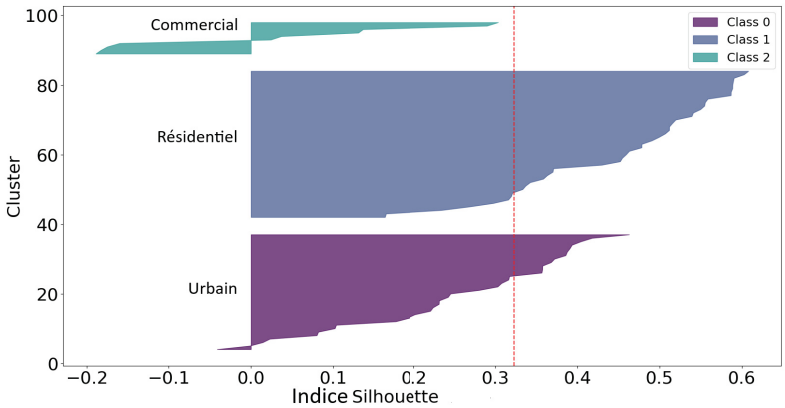


FIGURE 4.34 – Indices Silhouettes

4.2.11 Indice de Vitalité à Long Terme

À l'aide de la méthodologie décrite à la section 4.1.10, un IVL est conçu. Alors, à partir des groupes formés dans la table 4.12 selon les types de quartier, il est possible de suivre l'évolution de l'IVL dans le temps pour les regroupements formés par l'algorithme k-means dans la table 4.13.

TABLE 4.13 – table des moyennes par cluster et des prédictions pour 2026

Cluster	**Moyenne 2006**	**Moyenne 2011**	**Moyenne 2016**	**Moyenne 2021**	**Prédiction 2026**
Urbain	0.309	0.306	0.307	0.243	0.300
Résidentiel	0.401	0.421	0.451	0.369	0.399
Commercial	0.307	0.369	0.409	0.299	0.345

La figure 4.35 montre l'évolution de l'IVL pour les années 2006 à 2021, réparti entre les trois types de quartiers : Urbain, Résidentiel et Commercial. En outre, la figure présente également, en pointillés, les prévisions pour l'année 2026 obtenues grâce à un modèle de régression basé sur une régression linéaire ou sur un MLP. Le MLP est une méthode d'IA couramment utilisée pour la prédiction. Ce modèle a été entraîné sur les données historiques pour estimer la tendance future de l'IVL. Dans ce cas-ci, comme le montre la table 4.15, l'erreur quadratique moyenne étant plus faible pour la régression linéaire, ce sera donc le modèle choisi.

La métrique employée pour évaluer les prédictions est basée sur la méthode des moindres carrés. Les résultats obtenus pour chaque secteur sont présentés dans la table 4.14.

Les figures 4.36, 4.37 et 4.38 présentent les valeurs SHAP pour chaque type d'AD dans le cadre du modèle de régression linéaire utilisé pour estimer l'IVL en 2026.

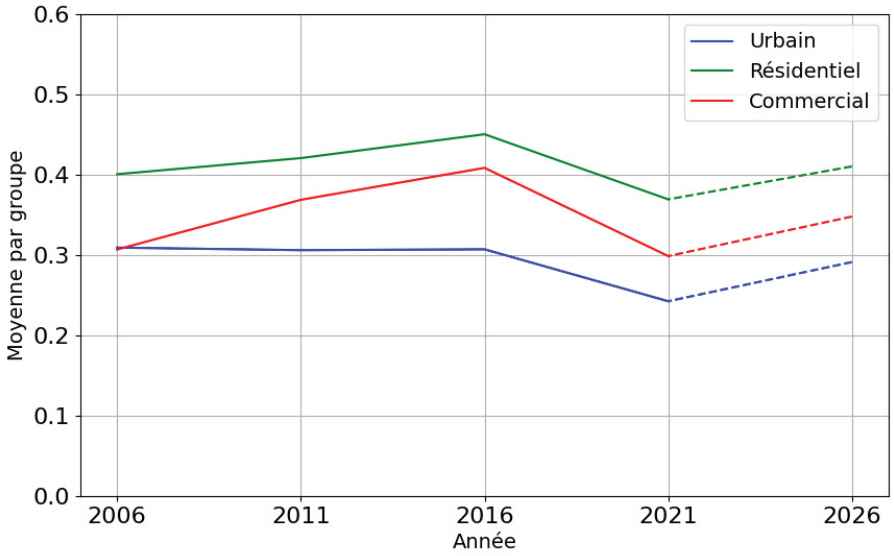


FIGURE 4.35 – Comparaison de l'IVL entre les 3 groupes

TABLE 4.14 – Erreur quadratique moyenne pour chaque cluster sur l'ensemble de test

Cluster	**Erreur Quadratique Moyenne**
Urbain	0.002186951
Résidentiel	0.002275352
Commercial	0.002342466

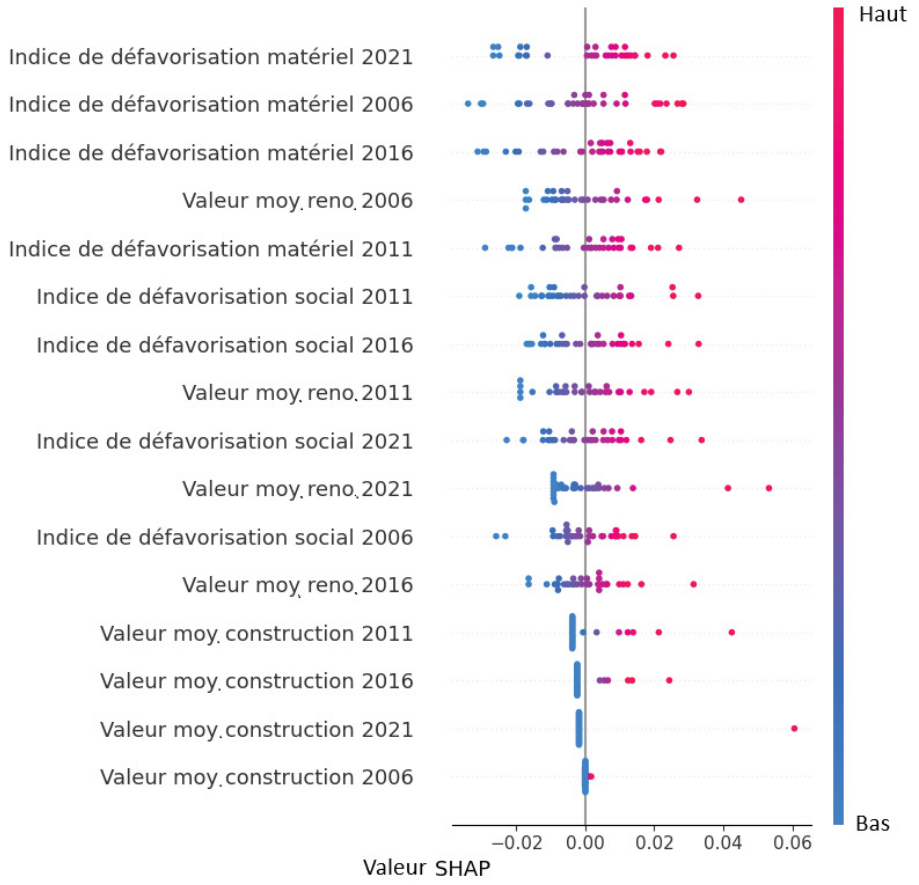


FIGURE 4.36 – Valeurs SHAP pour le modèle de régression linéaire pour le groupe Urbain

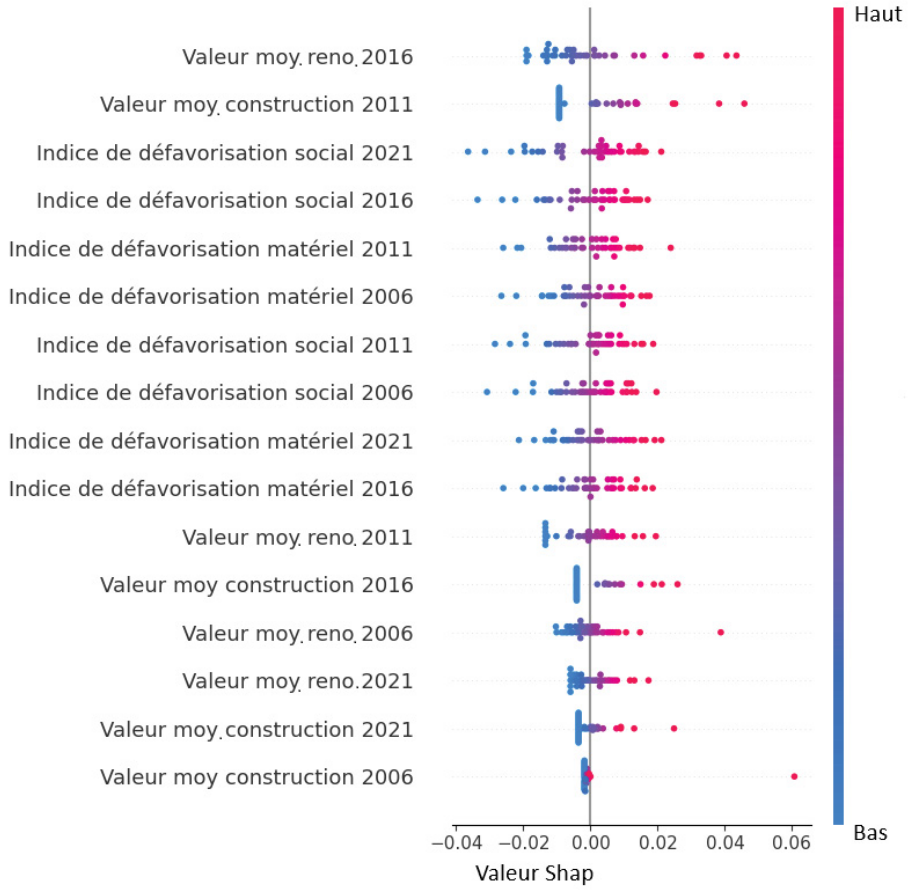


FIGURE 4.37 – Valeurs SHAP pour le modèle de régression linéaire pour le groupe Résidentiel

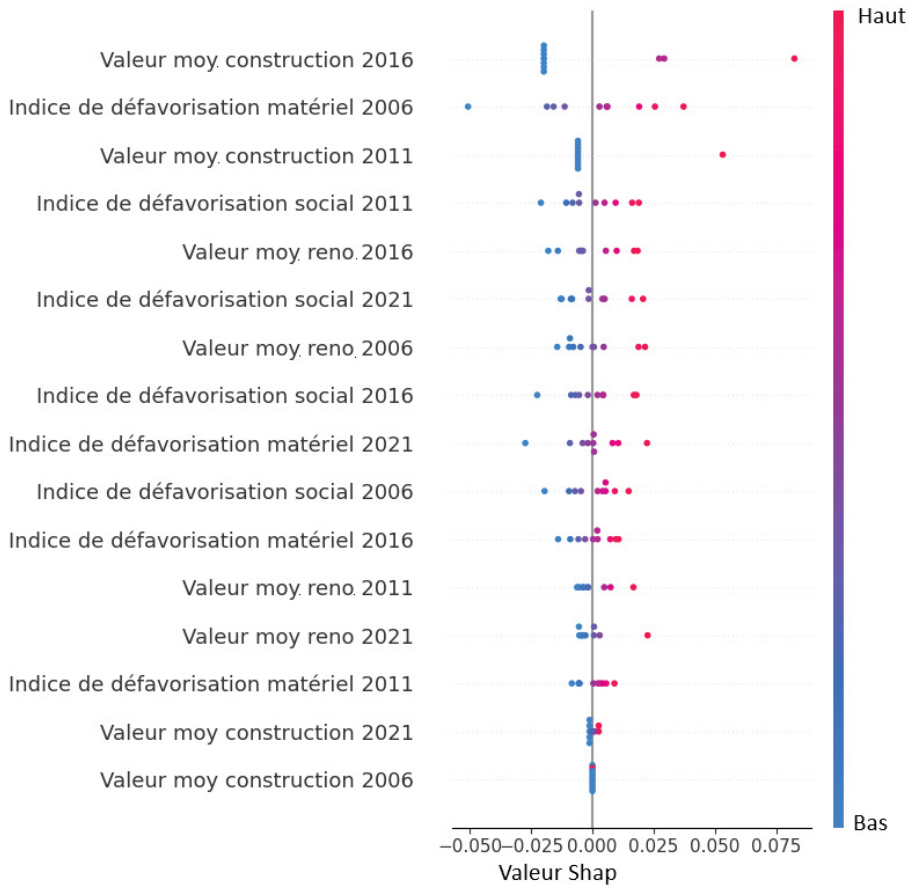


FIGURE 4.38 – Valeurs SHAP pour le modèle de régression linéaire pour le groupe Commercial

Les figures 4.39, 4.40 et 4.41 montrent l'évolution des indices de défavorisation pour les trois secteurs.

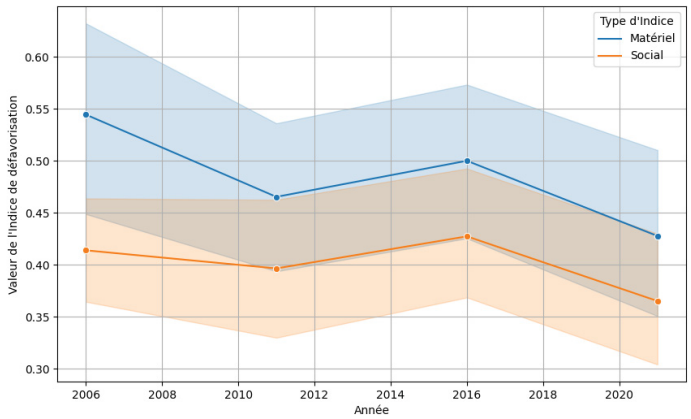


FIGURE 4.39 – Indice de défavorisation – Secteur Urbain

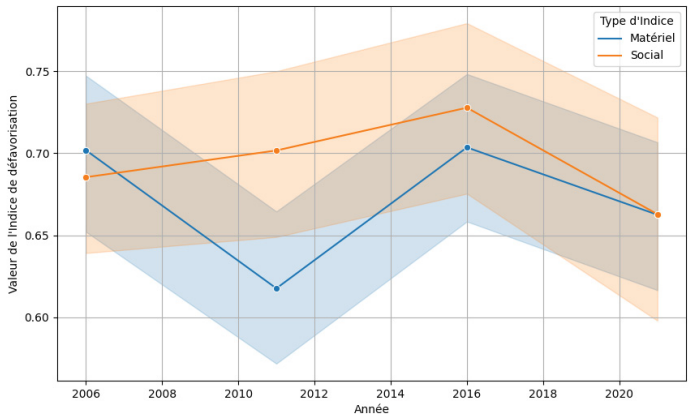


FIGURE 4.40 – Indice de défavorisation – Secteur Résidentiel

La table 4.15 présente les erreurs quadratiques moyennes (EQM) pour chaque modèle. Parmi les trois modèles testés, la régression linéaire se distingue par l'erreur quadratique moyenne la plus faible, indiquant qu'elle fournit les prédictions les plus précises pour cette analyse.

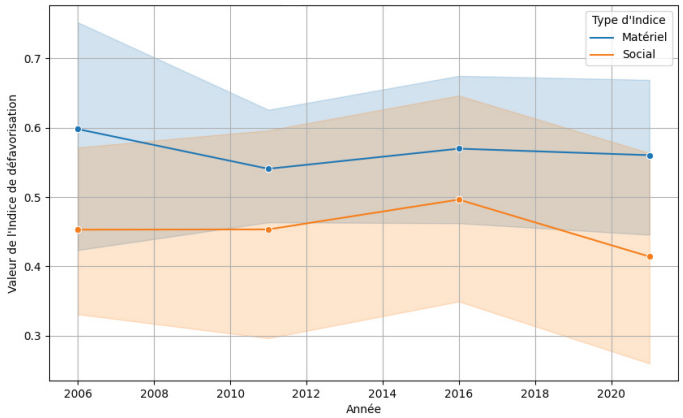


FIGURE 4.41 – Indice de défavorisation – Secteur Commercial

TABLE 4.15 – Erreurs quadratiques moyennes des modèles par cluster

Cluster	**MLP**	**RF**	**Régression Linéaire**
Urbain	0.0022	0.0017	0.0000
Résidentiel	0.0023	0.0005	0.0000
Commercial	0.0023	0.0004	0.0002

La table 4.16 présente les prédictions obtenues pour l'année 2026 concernant l'IVL avec les trois modèles.

TABLE 4.16 – Moyennes des prédictions par modèle pour chaque cluster

Cluster	**MLP**	**RF**	**Régression Lin.**
Urbain	0.2870	0.2897	0.2913
Résidentiel	0.4094	0.4091	0.4104
Commercial	0.3517	0.3489	0.3480

4.3 Discussions

Nous débuterons la discussion avec une analyse statistique des données brutes obtenues. Dans la figure 4.6 illustrant la matrice des corrélations des données pour l'année 2021, on observe que la majorité des coefficients de corrélation sont faibles, ne dépassant que rarement 0,5 en valeur absolue. Toutefois, une corrélation très forte et négative se distingue entre **La proportion de logements nécessitant des réparations mineures** et **Le taux de logements inoccupés**. Une interprétation possible de cette relation est que, dans les quartiers où le taux d'inoccupation est bas, les plaintes concernant des réparations mineures pourraient être plus fréquentes. En d'autres termes, si les logements sont occupés, les demandes de réparations mineures peuvent augmenter, tandis qu'un taux d'inoccupation élevé signifie qu'il y a moins de plaintes liées aux réparations, car les logements vacants ne génèrent pas de telles plaintes.

Les histogrammes des figures 4.7 et 4.8 illustrent la répartition des différentes variables et il apparaît clairement que les variables construites à partir des données fournies par la ville et géolocalisées ne suivent pas une distribution normale, ce qui peut être observé directement à partir des graphiques sans qu'il soit nécessaire d'effectuer un test de Shapiro-Wilk. Par exemple, les variables telles que la **Valeur moyenne des travaux de rénovation**, la **Valeur moyenne des travaux de construction**, et le **Nombre de bâtiments vacants** affichent des distributions nettement éloignées de la loi normale, indiquant la présence d'un biais. Ce biais est en partie attribuable au fait qu'un grand nombre d'AD à Shawinigan ne présentent aucun travaux de construction, et que très peu d'entre elles ont des travaux de rénovation, ce qui crée une surreprésentation de ces aires dans les données. Les figures 4.9 et 4.10 mettent en évidence ce déséquilibre. Concernant le **Nombre de bâtiments vacants**, la figure 4.11 montre clairement que les AD sans bâtiments vacants sont largement surreprésentées.

Dans les graphiques en boîte à moustache (voir figure 4.12), il ressort clairement que les variables **Valeur moyenne des travaux de construction**, **Valeur moyenne des immeubles** et **Nombre de bâtiments vacants** ne présentent pas une distribution homogène. En particulier, la variable **Nombre de bâtiments vacants** est marquée par des valeurs aberrantes, ce qui s'explique par un volume de données insuffisant. Ces informations, fournies par la Ville de Shawinigan, sont néanmoins jugées fiables. Il convient également de souligner que les bâtiments vacants ne sont pas présents dans toutes les AD. En effet, certaines AD n'en comptent aucun, tandis que d'autres en abritent plusieurs. Par exemple, l'AD 24360104 en dénombre sept, tandis que l'AD 24360165 en recense quatre. De plus, 57 AD ne comportent aucun bâtiment vacant. Dans le cadre de cette étude, puisque le nombre de bâtiments vacants constitue un indicateur significatif de la vitalité, cette variable est conservée. Pour les autres variables, la médiane est située au centre de presque toutes les boîtes, indiquant une distribution symétrique, donc sans biais. Enfin, la longueur des boîtes reste constante pour toutes les variables.

Concernant l'analyse de ACP, malgré un faible pourcentage d'explicabilité (52%) pour les deux premières composantes principales, une interprétation demeure possible. Par exemple, les AD 24360095 et 24360123 se trouvent aux extrémités du nuage de points (voir figure 4.14). En effet, ces deux quartiers présentent des caractéristiques très différentes. L'AD 24360095, en rouge, est fortement défavorisée, avec un indice de défavorisation sociale et un indice de défavorisation matérielle plus élevés que les autres. En revanche, l'AD 24360123, en vert, est très favorisée, présentant l'indice de défavorisation sociale le plus bas de la ville. En combinant les informations des figures 4.14 et 4.15, il est possible d'identifier des regroupements d'AD. Il est important de noter que certaines variables ont été inversées : ainsi, pour chaque variable, une valeur plus élevée est associée à une meilleure situation pour le quartier. Par exemple, les AD 24360098 et 24360086, situées dans le premier quadrant supérieur, affichent une valeur élevée pour la variable proportion de logements ayant besoin de réparations mineures. En raison de l'inversion de cette variable, cela signifie que ces logements n'ont

pas besoin de réparations mineures. Cependant, cela n'exclut pas la possibilité que des réparations majeures soient nécessaires, comme c'est le cas pour l'AD 24360086. Dans le deuxième quadrant, où l'on retrouve plusieurs vecteurs influents (voir figure 4.15), se situent de nombreuses AD susceptibles de bien performer en termes de vitalité. L'AD 24360123, par exemple, est la plus performante et représente un quartier favorisé, avec une bonne vitalité. La figure 4.16 met en évidence une zone qui pourrait inclure des AD à forte vitalité, indiquant un potentiel de développement positif pour ces quartiers. Enfin, les AD situées dans le troisième quadrant regroupent des logements ne nécessitant pas de réparations majeures et affichant un faible taux d'inoccupation. Il s'agit également de zones matériellement favorisées. Pour la contribution des caractéristiques aux deux premières composantes (voir table 4.8), il convient de noter que l'Indice de défavorisation sociale, avec -0.4524 pour la composante 1, et la Proportion de logements nécessitant des réparations majeures, avec -0.6639 pour la composante 2, sont les caractéristiques qui présentent les plus fortes contributions. Comme il est difficile, dans notre cas, d'attribuer une signification claire aux deux composantes, l'ACP ne sera utilisée qu'en combinant les informations issues des figures 4.14 et 4.15, comme présenté précédemment.

En somme, l'analyse statistique et l'ACP ont permis de dégager des tendances intéressantes concernant les AD de Shawinigan. La matrice de corrélations a révélé des relations significatives entre certaines variables, notamment le lien entre l'entretien des logements et le taux d'inoccupation. De plus, bien que les deux premières composantes de l'ACP n'expliquent que 52% de la variance totale, elles offrent une vue d'ensemble utile pour différencier les quartiers. Certaines AD, comme l'AD 24360090 (défavorisée) et l'AD 24360123 (favorisée), se démarquent nettement en raison de leurs caractéristiques socio-économiques opposées. Les variables liées à la défavorisation sociale et matérielle ainsi qu'à l'état des logements se révèlent particulièrement influentes dans cette classification. Les regroupements observés dans les figures 4.14 et 4.15 permettent d'identifier des zones à forte vitalité et des quartiers nécessitant potentiellement des interventions spécifiques. L'ACP a ainsi fourni une base utile pour mieux

comprendre les dynamiques locales, bien que l'interprétation des résultats demeure limitée par la faible explicabilité des premières composantes. Néanmoins, en combinant les informations visuelles et statistiques, cette analyse offre des pistes pertinentes pour orienter les actions de revitalisation urbaine dans la ville de Shawinigan.

Nous procéderons maintenant à l'analyse des résultats de l'IVA. Tout d'abord, la figure 4.17 permet de constater qu'il existe une bonne variabilité dans les valeurs de l'IVA. De plus, aucune AD ne présente un indice supérieur à 0,7 ni inférieur à 0,2. À l'aide de la carte 4.19, on remarque que les quartiers centraux sont nettement ceux où la vitalité est la plus faible. L'AD 24360104, avec un indice de vitalité de 0,2346, est celle qui affiche l'IVA le plus faible. La table 4.10 présente les caractéristiques de cette zone : l'**Indice de défavorisation sociale** y est élevé, tout comme les **valeurs moyennes des travaux de rénovation et de construction**, ainsi que la **valeur moyenne des immeubles**, qui sont très faibles, ce qui peut indiquer un sous-financement dans l'entretien des bâtiments. En effet, le nombre de bâtiments vacants dans ce quartier est le plus élevé de toute la ville. Par ailleurs, le nombre de jeunes et de commerces y est relativement bas, ce qui illustre encore une fois une faible vitalité. D'un autre côté, l'AD 24360123, avec un indice favorable de 0.6123, représente un écart exceptionnel de 2,63 écarts-types par rapport à la moyenne. Comme mentionné précédemment, l'Indice de Vitalité suit une distribution normale. Ainsi, dans ces conditions, la probabilité d'observer une AD avec un indice supérieur à 0.6123 est inférieure à 0,4%.

Dans le cadre de l'analyse des données de la ville de Shawinigan pour l'année 2021 concernant l'importance des indicateurs, la variable **Indice de défavorisation sociale** se distingue comme la plus significative. En effet, comme le montre la table 4.11, cette variable contribue à réduire l'erreur de prédiction globale de l'algorithme de près de 50%. En deuxième position, l'**Indice de défavorisation matérielle** se révèle également crucial, entraînant une réduction de l'erreur totale d'environ 20%. Les autres caractéristiques présentent une importance nettement inférieure, avec des

réductions d'erreur allant de 1% à 6% environ. Comme le montre la figure 4.24, les caractéristiques **Indice de défavorisation matérielle** et **Indice de défavorisation sociale** se distinguent à nouveau comme les plus influentes. L'interprétation de ce graphique des valeurs SHAP suit la méthodologie décrite au chapitre 3 de cet ouvrage. En effet, des valeurs élevées de l'**Indice de défavorisation matérielle** sont associées à une augmentation de l'IVA. Il en va de même pour des valeurs élevées de l'**Indice de défavorisation sociale**. Toutefois, étant donné que ces deux indices sont inversés, il est important de comprendre que ce sont en réalité des valeurs plus faibles qui favorisent une augmentation de l'IVA. En somme, les résultats concernant l'IVA, combinés à l'analyse des caractéristiques les plus influentes, permettent d'identifier rapidement les AD problématiques et aident la ville à agir sur les bons indicateurs.

Ensuite, l'analyse du classement des AD en trois secteurs à l'aide de l'algorithme k-means permet un suivi dans le temps de l'IVL. Les AD correspondant aux quartiers urbains sont bien représentées, comme le montre leur disposition sur la figure 4.26 Cependant, le groupe commercial présente quelques anomalies en raison de la grande superficie de certaines AD. Par exemple, l'AD 24360114 est l'une des plus grandes AD de Shawinigan. Elle inclut non seulement des zones urbaines centrales, mais aussi des commerces et des zones résidentielles. Bien qu'elle se classe au cinquième rang en termes de nombre de commerces, sa densité de population figure parmi les plus faibles de la ville, un critère pourtant utilisé pour classer les AD dans le groupe urbain. De plus, la proportion de jeunes y est élevée, ce qui est également un indicateur associé aux zones résidentielles. En résumé, cela démontre que la classification selon les critères choisis peut être difficile pour certaines AD, comme l'AD 24360114, qui présente des caractéristiques variées rendant son appartenance à un seul groupe plus complexe. Enfin, les AD classées dans le groupe Résidentiel semblent regrouper celles qui ne sont ni urbaines ni commerciales, ce qui est cohérent avec les caractéristiques de ce groupe.

L'interprétation des valeurs SHAP pour le classement k-means se fera tel que prescrit dans la méthodologie à la section 4.1.9. Les figures 4.27, **??** et 4.28 illustrent l'explication du regroupement des AD à l'aide des valeurs SHAP. Pour les AD classées dans la catégorie Urbain, les grandes valeurs de densité de population combinées à une faible proportion de jeunes jouent un rôle déterminant. En revanche, une forte proportion de jeunes et une faible densité de population sont des critères clés pour le classement dans la catégorie Résidentiel. Quant à la catégorie Commercial, elle est plus difficile à caractériser, car aucune tendance marquante ne se dégage. Toutefois, la variable liée au nombre de commerces se révèle être le facteur le plus influent pour cette catégorie. Cette analyse démontre que les caractéristiques démographiques et commerciales influencent de manière significative le classement des AD. Cependant, la catégorie **Commercial** est plus difficile à cerner que les autres, peut-être en raison des caractéristiques des AD de la ville de Shawinigan.

Dans la figure 4.30, on observe que la paire de variables **Proportion de jeunes** et **Densité de la population** permet de distinguer clairement les secteurs résidentiels et urbains. En effet, les secteurs urbains se caractérisent par une densité de population plus élevée, tandis que les secteurs résidentiels affichent souvent une proportion plus importante de jeunes. De plus, les nuages de points intégrant la variable **Nombre de commerces** montrent une séparation nette du secteur commercial par rapport aux autres secteurs. Toutefois, la variable **Valeur moyenne des immeubles** ne semble pas permettre une distinction claire entre les secteurs dans ce graphique. Ensuite, en examinant la diagonale de la figure 4.30, il est évident que les variables **Densité de la population** et **Proportion de jeunes** jouent un rôle déterminant dans la distinction entre les secteurs Résidentiel et Urbain. Cette observation renforce nos affirmations précédentes concernant les caractéristiques spécifiques de ces secteurs. Ainsi, la forte densité de la population est souvent associée à des zones urbaines, tandis qu'une proportion élevée de jeunes se retrouve fréquemment dans les secteurs résidentiels, corroborant ainsi les tendances observées dans nos analyses antérieures. Une analyse visuelle des graphiques faites à partir des figures 4.31, 4.32 et 4.33, révèle que les

groupes Urbain et Résidentiel montrent une grande cohérence interne. En revanche, le groupe Commercial regroupe des AD présentant des caractéristiques plus variées, ce qui corrobore notre analyse précédente basée sur la table 4.12.

Comme le montre la figure 4.34, l'indice de silhouette moyen obtenu par l'algorithme k-means est de 0,3229. Bien que cet indice ne soit pas exceptionnellement élevé ni proche de 1, il indique une cohésion modérée entre les points au sein des regroupements, ainsi qu'une séparation partielle entre les groupes. C'est le groupe Commercial qui obtient l'indice de silhouette le plus faible. Cela pourrait être dû au fait que ce groupe n'est pas bien séparé dans l'espace, rendant sa distinction avec les autres groupes plus difficile. Le choix des caractéristiques pourrait améliorer cette séparation, mais il est difficile d'obtenir de nouvelles informations couvrant une période aussi longue que quinze ans. Il est important de noter que, malgré ce résultat, les groupes obtenus correspondent bien aux trois types de quartiers définis par la ville, ce qui suggère que les centroïdes initiaux choisis pour le regroupement étaient pertinents. De même, les urbanistes de la ville de Shawinigan ont confirmés que les regroupements étaient cohérents avec la réalité de la ville. Ces résultats confirment que ces groupes pourront être utilisés pour illustrer les tendances à long termes.

L'analyse de l'IVL se fera tout d'abord avec la figure 4.35, on observe une baisse généralisée dans les trois groupes entre 2016 et 2021. Cette diminution pourrait être liée à la pandémie mondiale de la COVID-19, qui a affecté plusieurs aspects économiques et sociaux des quartiers, entraînant une diminution de leur vitalité.

En comparant les erreurs quadratiques moyennes des différents modèles de prédiction pour l'année 2026 (table 4.15), on remarque que c'est le modèle le plus simple, la régression linéaire, qui est le plus précis. Les résultats de cette régression linéaire permettent d'anticiper une stabilisation ou une éventuelle reprise pour l'année 2026, bien que les tendances restent incertaines en fonction des conditions économiques et sociales à venir. Cette approche combinée, basée à la fois sur des observations

historiques et des prévisions par IA, offre une vue d'ensemble robuste de l'évolution passée et future de la vitalité des quartiers de la ville de Shawinigan.

Les graphiques des valeurs SHAP pour chaque secteur (voir les figures 4.36, 4.37 et 4.38) montrent que pour toutes les variables et tous les groupes, de grandes valeurs d'une variable entraînent une augmentation de l'IVL, ce qui est conforme à notre méthodologie concernant l'inversion des données (voir section 4.1.4). On peut également souligner que la variable ayant l'impact le plus significatif sur la prévision de l'IVL en 2026 pour le secteur urbain est l'**Indice de défavorisation matérielle**. En effet, en rouge, de grandes valeurs de défavorisation augmentent l'IVL. Toutefois, il est important de rappeler que l'indice de défavorisation est inversé : ainsi, c'est lorsque la défavorisation est faible que l'IVL augmente. D'un autre côté, les variables représentant la valeur moyenne des constructions pour les années 2006 à 2016 semble être les moins influentes. Pour les trois secteurs, les caractéristiques valeur moyenne des constructions 2006 et 2021 sont les moins influentes. Aussi, la table 4.14 montrent que les valeurs d'erreur quadratique moyenne pour chaque secteur sont faibles, même avec la normalisation des données. Cela nous permet de conclure que le modèle de régression linéaire généré est fiable et performant.

Un aspect intriguant ressort de l'analyse des valeurs SHAP de l'algorithme MLP : l'**Indice de défavorisation sociale** de 2006 se distingue en affichant une influence inversée par rapport aux autres caractéristiques. En effet, selon la figure **??**, des valeurs élevées de cet indice sont associées à une diminution de la prévision pour l'indice de vitalité de 2026. Cette tendance est contre-intuitive, car une valeur élevée de cet indice, qui est inversé, devrait normalement être associée à une augmentation de la prévision. Pour approfondir cette analyse, il pourrait être pertinent de comparer le comportement de cette caractéristique à travers toutes les années disponibles en s'appuyant sur les figures 4.39, 4.40 et 4.41. La figure 4.6 montre une corrélation entre l'**Indice de défavorisation sociale** et l'**Indice de défavorisation matérielle**, indiquant que ces deux caractéristiques évoluent de manière similaire au fil du temps.

Aucun élément dans cette figure ne suggère que l'**Indice de défavorisation sociale** devrait se comporter de manière inverse par rapport aux autres caractéristiques dans les prévisions. À ce stade de l'analyse, cette inversion reste difficile à interpréter. Les autres modèles prédictifs, comme la régression linéaire et les RF, ne montrent pas cette inversion. De plus, étant donné que la régression linéaire présente une erreur quadratique moyenne plus faible, cela confirme notre choix de modèle.

4.3.1 Conclusion

Cette analyse statistique des quartiers de Shawinigan en 2021 a permis de mieux comprendre les dynamiques sociales et économiques à l'échelle locale. L'élaboration de l'IVA a fourni une mesure instantanée utile pour évaluer la qualité de vie dans chaque secteur, tandis que l'IVL nous montre l'évolution de la vitalité des 15 dernières années. Les disparités observées entre les quartiers urbains et résidentiels soulignent l'importance d'adapter les interventions urbaines à la vitalité propres de chaque quartier. Les résultats de cette étude constituent une base solide pour orienter les politiques locales de revitalisation.

Chapitre 5

Conclusion

Ce dernier chapitre présente la conclusion de cette étude, en synthétisant les principaux résultats obtenus sur la vitalité des quartiers de Shawinigan. Il résume les contributions méthodologiques et empiriques de la recherche. Aussi, il propose des recommandations concrètes ainsi que des pistes pour les travaux futurs.

5.1 Résumé des contributions principales de la recherche

Ce mémoire a exploré la vitalité des quartiers de Shawinigan à travers l'analyse de divers indicateurs, notamment l'IVA et l'IVL. En regroupant les 87 AD en trois catégories (Urbain, Résidentiel, Commercial), nous avons pu évaluer de manière précise les dynamiques socio-économiques en jeu. L'utilisation de l'algorithme K-Means a permis d'identifier des groupes significatifs qui éclairent les disparités entre les différents quartiers.

En somme, cette étude a permis de mieux comprendre les dynamiques de vitalité urbaine à Shawinigan en mobilisant des indicateurs quantitatifs et des techniques d'analyse avancées. La combinaison de l'IVA et de l'IVL a offert une double lecture de la réalité : d'une part, une photographie instantanée des quartiers selon leur situation sociale et économique, et d'autre part, une perspective évolutive permettant d'anticiper les trajectoires de transformation.

L'analyse des résultats a permis de mieux comprendre la dynamique de la vitalité des quartiers de Shawinigan. Tout d'abord, l'IVA révèle une bonne variabilité entre les AD, confirmant la pertinence de l'indicateur développé. Le classement des AD en trois secteurs (urbain, résidentiel et commercial) à l'aide de l'algorithme K-means a permis de structurer l'analyse à long terme. Bien que certaines AD, comme la 24360114, présentent des caractéristiques hybrides compliquant leur classement, les regroupements ont pu être validés par les urbaniste de la ville de Shawinigan.

Les résultats des valeurs SHAP ont également mis en évidence le rôle important de l'IDMS dans l'explication de l'IVA. À l'inverse, les caractéristiques liées à la valeur moyenne des travaux de construction semblent avoir une influence marginale. En ce qui concerne l'IVL, les caractéristiques les plus influentes varient selon les secteurs, mais elles demeurent cohérentes avec les caractéristiques propres à chaque groupe.

5.2 Implications théoriques et pratiques

Les résultats de cette étude soulignent également l'importance de la vitalité dans le développement urbain. Des quartiers dynamiques favorisent non seulement le bien-être des résidents, mais contribuent également à l'attractivité économique de la ville. Nous avons observé que les zones résidentielles présentent des signes de vitalité plus marqués que les quartiers urbains et commerciaux. Cela peut être lié à un IDMS plus

faible dans ces zones résidentielles, ce qui se traduit par des conditions de vie plus favorables. De plus, ces quartiers bénéficient de valeurs moyennes plus élevées en matière de rénovations et de constructions, indiquant une activité soutenue dans l'entretien et l'amélioration de ces infrastructures. Ces investissements peuvent renforcer l'attrait des quartiers résidentiels, stimulant la vitalité et encourageant le renouvellement immobilier, créant un environnement propice à la qualité de vie et à la croissance économique de ces quartiers. Cependant, il est crucial de reconnaître les défis que certaines zones doivent surmonter. Les AD dans le regroupement urbains appellent à une intervention stratégique des autorités locales pour revitaliser ces zones. La mise en œuvre de politiques ciblées, basées sur les données recueillies, pourrait permettre de stimuler la vitalité de ces quartiers et d'améliorer la qualité de vie de leurs habitants.

Mais au-delà des résultats descriptifs, ce mémoire démontre aussi la puissance des techniques d'AA pour appuyer la prise de décision en aménagement du territoire. En effet, les algorithmes mobilisés dans cette étude, tels que KNN pour l'imputation des données manquantes, ont permis de reconstituer un jeu de données complet et cohérent, essentiel à la robustesse des analyses subséquentes. Grâce à KNN, il a été possible de préserver la variabilité naturelle du jeu de données sans introduire de biais systématique. Ensuite, RF et XGBoost ont été mobilisés pour l'identification des caractéristiques les plus influentes sur la vitalité des quartiers. Leur capacité à attribuer une importance à chaque caractéristique a facilité une lecture plus claire des facteurs clés influençant l'indice de vitalité, orientant ainsi les recommandations aux urbanistes. Aussi, K-Means, utilisé pour le regroupement des AD, a permis de distinguer les différents types de quartiers selon leurs caractéristiques socio-économiques. Grâce à cette méthode de regroupement non supervisée, trois grands ensembles, urbain, résidentiel et commercial, ont été identifiés, chacun présentant des dynamiques de vitalité distinctes.

En plus, en utilisant un MLP et la régression linéaire, il a été possible d'effectuer des prévisions sur l'évolution de la vitalité à l'horizon 2026 (voir figure 4.35).

Ces modèles ont permis de projeter la vitalité à venir dans les différents types de regroupements, en tenant compte des évolutions passées et actuelles. Les comparaisons entre modèles prédictifs ont démontré que la régression linéaire, malgré sa simplicité, offre les meilleures performances, confirmant son choix pour les prévisions de 2026. Les tendances observées, notamment la baisse généralisée de l'IVL entre 2016 et 2021, pourraient s'expliquer en partie par l'impact de la pandémie de Covid-19. Les prévisions pour 2026 laissent entrevoir une possible stabilisation. Ces modèles ne se contentent pas seulement de générer des résultats : ils permettent également une meilleure compréhension des mécanismes sous-jacents grâce à l'explicabilité. L'usage des valeurs SHAP a ainsi offert une lecture précise de l'influence de chaque caractéristique dans les décisions des modèles, ce qui est essentiel pour légitimer les recommandations auprès des urbanistes, élus et citoyens. Ce souci de transparence rend l'IA plus accessible et pertinente dans un contexte de gouvernance municipale. Loin de remplacer l'expertise humaine, elle la complète, en fournissant une base objective et reproductible pour éclairer des choix souvent complexes. Ce mémoire ouvre donc la voie à une planification urbaine plus fine, fondée sur des données robustes, une modélisation transparente, et une volonté de réduire les inégalités territoriales. En intégrant l'IA dans les outils d'aide à la décision, il devient possible de faire évoluer la ville selon des critères scientifiques. En plus, cette recherche ouvre la voie à d'autres études sur la dynamique des quartiers, en intégrant des facteurs tels que l'environnement, l'accès aux transports et la participation citoyenne. À l'avenir, une approche plus interdisciplinaire pourrait enrichir la compréhension des mécanismes de la vitalité des quartiers et de leur impact sur la société.

5.3 Suggestions pour les recherches futures

Nous avons rencontré certaines limitations en matière d'accessibilité aux données, en raison de la taille réduite de la ville étudiée. Contrairement aux grandes villes ca-

nadiennes, les plus petites disposent souvent de données moins complètes et moins facilement accessibles, ce qui a posé des défis pour cette analyse. Par ailleurs, une recherche approfondie sur les indicateurs disponibles sur une période de 15 ans pourrait être envisagée afin d'améliorer l'IVL. La Ville de Shawinigan devrait également envisager de collecter le plus de données possible, dans le but de les rendre accessibles et de les intégrer aux indices de vitalité (IVA et IVL). En complément, il serait pertinent de développer une interface pour cette étude sur la vitalité des quartiers. Celle-ci permettrait à l'utilisateur de personnaliser les indicateurs et les méthodes d'AA utilisés. Les choix pourraient être appuyés par des indicateurs de performance et des visualisations graphiques, renforçant ainsi la compréhension des décisions prises. L'objectif de cette interface serait d'offrir à l'utilisateur une capacité de décision éclairée, fondée sur des informations explicatives relatives aux méthodes appliquées.

En conclusion, la vitalité des quartiers de Shawinigan est un sujet complexe qui nécessite une attention continue. Les résultats de ce mémoire fournissent une base solide pour des actions futures, visant à promouvoir des secteurs résilients et florissants. En tant qu'outils de planification urbaine, l'IVA et l'IVL peuvent éclairer les décisions politiques et mobiliser les ressources nécessaires pour garantir le développement durable de Shawinigan.

Bibliographie

[1] L. Breiman, "Random Forests," *Machine Learning*, t. 45, nº 1, p. 5-32, 2001. doi : 10.1023/a:1010933404324. adresse : https://link.springer.com/article/10.1023/A:1010933404324.

[2] J. K. Jaiswal et R. Samikannu, "Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression," in *2017 World Congress on Computing and Communication Technologies (WCCCT)*, fév. 2017, p. 65-68. doi : 10.1109/WCCCT.2016.25. adresse : https://ieeexplore.ieee.org/abstract/document/8074494 (visité le 02/04/2025).

[3] R. Genuer, J.-M. Poggi et C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, t. 31, nº 14, p. 2225-2236, 15 oct. 2010, issn : 0167-8655. doi : 10.1016/j.patrec.2010.03.014. adresse : https://www.sciencedirect.com/science/article/pii/S0167865510000954 (visité le 07/04/2025).

[4] C. Strobl, A.-L. Boulesteix, A. Zeileis et T. Hothorn, "Bias in random forest variable importance measures : illustrations, sources and a solution," *BMC Bioinformatics*, t. 8, nº 1, p. 25, 25 jan. 2007, issn : 1471-2105. doi : 10.1186/1471-2105-8-25. adresse : https://doi.org/10.1186/1471-2105-8-25 (visité le 07/04/2025).

[5] T. Chen et C. Guestrin, "XGBoost : A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining, 13 août 2016, p. 785-794. DOI : 10.1145/2939672.2939785. arXiv : 1603.02754[cs]. adresse : http://arxiv.org/abs/1603.02754 (visité le 08/04/2024).

[6] G. E. Batista et M. C. Monard, "A study of K-nearest neighbour as an imputation method.," *His*, t. 87, n° 251, p. 48, 2002. adresse : https://www.academia.edu/download/71559365/A_Study_of_K-Nearest_Neighbour_as_an_Imp20211006-13950-t4clbz.pdf (visité le 08/04/2025).

[7] J. Li, S. Guo, R. Ma et al., "Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets," *BMC Medical Research Methodology*, t. 24, n° 1, p. 41, 16 fév. 2024, ISSN : 1471-2288. DOI : 10.1186/s12874-024-02173-x. adresse : https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-024-02173-x (visité le 16/04/2025).

[8] O. Troyanskaya, M. Cantor, G. Sherlock et al., "Missing value estimation methods for DNA microarrays," *Bioinformatics*, t. 17, n° 6, p. 520-525, 1er juin 2001, ISSN : 1367-4803. DOI : 10.1093/bioinformatics/17.6.520. adresse : https://doi.org/10.1093/bioinformatics/17.6.520 (visité le 08/04/2025).

[9] L. Beretta et A. Santaniello, "Nearest neighbor imputation algorithms : a critical evaluation," *BMC Medical Informatics and Decision Making*, t. 16, n° 3, p. 74, 25 juill. 2016, ISSN : 1472-6947. DOI : 10.1186/s12911-016-0318-z. adresse : https://doi.org/10.1186/s12911-016-0318-z (visité le 09/04/2025).

[10] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Statistics*, t. 5, University of California press, 1967, p. 281-298. adresse : https://projecteuclid.org/ebook/Download?urlId=bsmsp/1200512992&isFullBook=false (visité le 09/04/2025).

[11] A. K. Jain, "Data clustering : 50 years beyond K-means," *Pattern recognition letters*, t. 31, n° 8, p. 651-666, 2010, Publisher : Elsevier. adresse : https://www.sciencedirect.com/science/article/pii/S0167865509002323 (visité le 11/04/2025).

[12] G. Hamerly et C. Elkan, "Learning the k in k-means," in *Advances in Neural Information Processing Systems*, t. 16, MIT Press, 2003. adresse : https://proceedings.neurips.cc/paper/2003/hash/234833147b97bb6aed53a8f4f1c7a7d8-Abstract.html (visité le 10/04/2025).

[13] M. Ay, L. Özbakir, S. Kulluk, B. Gülmez, G. Öztürk et S. Özer, "FC-Kmeans : Fixed-centered K-means algorithm," *Expert Systems with Applications*, t. 211, p. 118 656, 1er jan. 2023, issn : 0957-4174. doi : 10.1016/j.eswa.2022.118656. adresse : https://www.sciencedirect.com/science/article/pii/S0957417422016979 (visité le 21/05/2024).

[14] A. Vouros, S. Langdell, M. Croucher et E. Vasilaki, "An empirical comparison between stochastic and deterministic centroid initialisation for k-means variations," *Machine Learning*, t. 110, n° 8, p. 1975-2003, août 2021, issn : 0885-6125, 1573-0565. doi : 10.1007/s10994-021-06021-7. adresse : https://link.springer.com/10.1007/s10994-021-06021-7 (visité le 21/04/2025).

[15] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, t. 24, n° 6, p. 417, 1933, Publisher : Warwick & York. adresse : https://psycnet.apa.org/record/1934-00645-001 (visité le 14/04/2025).

[16] I. T. Jolliffe et J. Cadima, "Principal component analysis : a review and recent developments," *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, t. 374, n° 2065, p. 20 150 202, 13 avr. 2016, Publisher : Royal Society. doi : 10.1098/rsta.2015.0202. adresse : https://royalsocietypublishing.org/doi/full/10.1098/rsta.2015.0202 (visité le 14/04/2025).

[17] E. N. Lorenz, "Empirical orthogonal functions and statistical weather prediction," *(No Title)*, 1956. adresse : https://cir.nii.ac.jp/crid/1370004237455296143 (visité le 14/04/2025).

[18] M. Greenacre, P. J. F. Groenen, T. Hastie, A. Markos et E. Tuzhilina, "Principal component analysis," *Principal Component Analysis*, 2023.

[19] L. S. Shapley. "A value for n-person games," SciSpace - Paper. Issue : 28 Pages : 307-317 Publisher : RAND Corporation. (18 mars 1952), adresse : https://typeset.io/papers/a-value-for-n-person-games-58w5vcviy3 (visité le 09/10/2024).

[20] M. Louhichi, R. Nesmaoui, M. Mbarek et M. Lazaar, "Shapley values for explaining the black box nature of machine learning model clustering," *Procedia Computer Science*, t. 220, p. 806-811, 2023, issn : 18770509. doi : 10.1016/j.procs.2023.03.107. adresse : https://linkinghub.elsevier.com/retrieve/pii/S1877050923006427 (visité le 26/11/2023).

[21] M. Li, H. Sun, Y. Huang et H. Chen, "Shapley value : from cooperative game to explainable artificial intelligence," t. 4, n° 1, p. 2, 2024, Num Pages : 2 Place : Singapore, Netherlands Publisher : Springer Nature B.V. doi : 10.1007/s43684-023-00060-8. adresse : https://www.proquest.com/docview/2924116820/abstract/2E58929DD0244663PQ/1 (visité le 23/04/2025).

[22] J. Jacobs, *Dark Age Ahead : Author of The Death and Life of Great American Cities*. Random House of Canada, 25 juin 2010, 258 p., Google-Books-ID : QHHMgYaeurIC, isbn : 978-0-307-36963-5.

[23] J. E. Drewes et M. Van Aswegen, "Determining the vitality of urban centres," **presentedat**The Sustainable World, 30 sept. 2010, p. 15-25. doi : 10.2495/SW100021. adresse : http://library.witpress.com/viewpaper.asp?pcode=SW10-002-1 (visité le 24/01/2024).

[24] X. Li, S. Cheng, Z. Lv, H. Song, T. Jia et N. Lu, "Data analytics of urban fabric metrics for smart cities," *Future Generation Computer Systems*, t. 107, p. 871-882, 1er juin 2020, issn : 0167-739X. doi : 10.1016/j.future.2018.

02.017. adresse : https://www.sciencedirect.com/science/article/pii/S0167739X1731227X (visité le 18/03/2024).

[25] J.-S. Dessureault, J. Simard et D. Massicotte, *Unsupervised Machine learning methods for city vitality index*, 29 jan. 2021. doi : 10.48550/arXiv.2012.12082. arXiv : 2012.12082[cs]. adresse : http://arxiv.org/abs/2012.12082 (visité le 31/01/2024).

[26] M. Haynes, A. Hook, G. Chiodi Grensing et H. Ecklund, "Analysis of duluth-superior and peer metropolitan areas," University of Minnesota Duluth, Report, 2018, Accepted : 2019-06-12T16 :39 :07Z. adresse : http://conservancy.umn.edu/handle/11299/203562 (visité le 18/03/2024).

[27] K. Scott. "Katherine scott — canadian index of wellbeing." (1er oct. 2010), adresse : https://uwaterloo.ca/canadian-index-wellbeing/people-profiles/katherine-scott (visité le 18/03/2024).

[28] H. Liu, P. Gou et J. Xiong, "Vital triangle : A new concept to evaluate urban vitality," *Computers, Environment and Urban Systems*, t. 98, p. 101 886, 1er déc. 2022, issn : 0198-9715. doi : 10.1016/j.compenvurbsys.2022.101886. adresse : https://www.sciencedirect.com/science/article/pii/S0198971522001302 (visité le 18/03/2024).

[29] H. M. K. K. M. B. Herath et M. Mittal, "Adoption of artificial intelligence in smart cities : A comprehensive review," *International Journal of Information Management Data Insights*, t. 2, no 1, p. 100 076, 1er avr. 2022, issn : 2667-0968. doi : 10.1016/j.jjimei.2022.100076. adresse : https://www.sciencedirect.com/science/article/pii/S2667096822000192 (visité le 27/02/2024).

[30] J. Wang et F. Biljecki, "Unsupervised machine learning in urban studies : A systematic review of applications," *Cities*, t. 129, p. 103 925, 1er oct. 2022, issn : 0264-2751. doi : 10.1016/j.cities.2022.103925. adresse : https://www.sciencedirect.com/science/article/pii/S026427512200364X (visité le 18/03/2024).

[31] P. J. Rousseeuw, "Silhouettes : A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, t. 20, p. 53-65, 1er nov. 1987, issn : 0377-0427. doi : 10.1016/0377-0427(87)90125-7. adresse : https://www.sciencedirect.com/science/article/pii/0377042787901257 (visité le 16/09/2024).

[32] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, t. 28, no 2, p. 129-137, mars 1982, Conference Name : IEEE Transactions on Information Theory, issn : 1557-9654. doi : 10.1109/TIT.1982.1056489. adresse : https://ieeexplore.ieee.org/document/1056489/?arnumber=1056489 (visité le 16/09/2024).

[33] T. Cover et P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, t. 13, no 1, p. 21-27, jan. 1967, Conference Name : IEEE Transactions on Information Theory, issn : 1557-9654. doi : 10.1109/TIT.1967.1053964. adresse : https://ieeexplore.ieee.org/document/1053964/?arnumber=1053964 (visité le 20/01/2025).

[34] A. Liaw et M. Wiener, "Classification and regression by randomForest," t. 2, 2002.

[35] L. Breiman, "Random forests," *Machine Learning*, t. 45, no 1, p. 5-32, 1er oct. 2001, issn : 1573-0565. doi : 10.1023/A:1010933404324. adresse : https://doi.org/10.1023/A:1010933404324 (visité le 22/01/2025).

[36] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas et H. Hemingway, "Comparison of random forest and parametric imputation models for imputing missing data using MICE : a CALIBER study," *American Journal of Epidemiology*, t. 179, no 6, p. 764-774, 15 mars 2014, issn : 1476-6256, 0002-9262. doi : 10.1093/aje/kwt312. adresse : https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwt312 (visité le 09/04/2025).

[37] D. J. Stekhoven et P. Bühlmann, "MissForest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, t. 28, no 1, p. 112-118, 1er jan. 2012, issn : 1367-4803. doi : 10.1093/bioinformatics/btr597.

adresse : https://doi.org/10.1093/bioinformatics/btr597 (visité le 09/04/2025).

[38] Z. Chen, F. Xiao, F. Guo et J. Yan, "Interpretable machine learning for building energy management : A state-of-the-art review," *Advances in Applied Energy*, t. 9, p. 100 123, 1er fév. 2023, ISSN : 2666-7924. DOI : 10.1016/j.adapen.2023.100123. adresse : https://www.sciencedirect.com/science/article/pii/S2666792423000021 (visité le 19/05/2025).

[39] H. Chen, "Algorithms to Estimate Shapley Value Feature Attributions," ISBN : 9798351441597, thèse de doct., University of Washington, United States – Washington, 2022, 178 p. adresse : https://www.proquest.com/docview/2730286591/abstract/6018971119424E91PQ/1 (visité le 09/10/2024).

Annexe A

Code python

```
def geoloc_adresse(self, annee="", geopandas_file="",
   adresse_file="", filename="", variable=""):
    self.geopandas_file = geopandas_file
    self.filename = filename
    self.adresse_file = adresse_file
    self.variable = variable
    self.annee = annee

    # read gpd file 'resultats_2021.json'
    regions = gpd.read_file("datasets/"+ annee +'/'+
       geopandas_file, crs='EPSG:4326')
    # read csv file "datasets/2021-reno-residentielles.csv"
    renovations_df = pd.read_csv('datasets/' +annee+ '/'+ self.
       adresse_file,encoding="cp1252")

    if self.adresse_file != "Batiments-vacants.csv":
        renovations_df["No civique de l'emplacement"] =
           renovations_df["No civique de l'emplacement"].astype
           (str)
```

```
15          renovations_df["No civique de l'emplacement"].fillna('')
16          renovations_df["Voie publique de l'emplacement"].fillna('')
17          renovations_df["Description code ancienne ville"].fillna('Shawinigan', inplace=True)
18          renovations_df['Adresse_complete'] = renovations_df["No civique de l'emplacement"] + ' ' + renovations_df["Voie publique de l'emplacement"] + ' ' + renovations_df["Description code ancienne ville"] + '  ' + "quebec" + ' ' + "canada"
19
20      if self.adresse_file == "Batiments-vacants.csv":
21          renovations_df['Adresse_complete'] = renovations_df["Emplacement"] + ' ' +renovations_df["Secteur"]
22
23      coords = []
24      noms_adresses = []
25
26      geolocator = GoogleV3(api_key='VOTRE_CLE_API')
27
28      with open('datasets/'+annee+'/' + filename, 'w', newline='') as csvfile:
29          writer = csv.writer(csvfile)
30          if self.adresse_file != "Batiments-vacants.csv":
31              writer.writerow(['Adresse', 'Latitude', 'Longitude', variable, 'DAUID'])
32          else:
33              writer.writerow(['Adresse', 'Latitude', 'Longitude','DAUID'])
34
35      DAUID = None
```

```
36      for index, row in renovations_df.iterrows():
37          adresse = row['Adresse_complete']
38          if self.adresse_file != "Batiments-vacants.csv":
39              valeur_travaux = row[variable]
40
41          try:
42              time.sleep(0.5)
43              location = geolocator.geocode(adresse)
44              if location:
45                  coords.append((location.latitude, location.
                        longitude))
46                  noms_adresses.append(adresse)
47                  point = Point(location.longitude, location.
                        latitude)
48
49                  region_correspondante = None
50                  for index, region in regions.iterrows():
51                      if region['geometry'].contains(point):
52                          DAUID = region['DAUID']
53                          break
54
55                  renovations_df['DAUID'] = DAUID
56
57                  with open('datasets/'+annee+'/' + filename, 'a'
                        , newline='') as csvfile:
58                      writer = csv.writer(csvfile)
59                      if self.adresse_file != "Batiments-vacants.
                            csv":
60                          writer.writerow([adresse, location.
                                latitude, location.longitude,
                                valeur_travaux, DAUID])
61                      else:
```

```
                        writer.writerow([adresse, location.
                            latitude, location.longitude,  DAUID
                            ])

        except (GeocoderTimedOut, GeocoderUnavailable) as e:
            print("Erreur de connexion:", e)
            print("Tentative de reconnexion dans 1 secondes..."
                )
            time.sleep(1)

def group_DA(self, filename=""):
    self.filename = filename
    resultats_geocodage = pd.read_csv('datasets/' + self.
        filename , encoding="cp1252")
    resultats_geocodage = resultats_geocodage.groupby('DAUID').
        sum().reset_index()
    return resultats_geocodage
```

```
    def geol_commerce(self, geopandas_file="", filename="",
        annee=""):
        self.geopandas_file = geopandas_file
        self.filename = filename
        self.annee =annee

        regions = gpd.read_file("datasets/" + annee + '/' +
            geopandas_file, crs='EPSG:4326')

        geolocator = GoogleV3(api_key='
            AIzaSyCuOrD_pMYPwHa3HIAL6SIxdY070eqqncw')

        categories = ['bank','restaurant','hostel','park','
            museum',
```

```
11                      'supermarket','things to do','cafe','bar'
                           ,'pharmacy',
12                      'store','hospital','gym','museum',
13                        'library','school','church','mosque','
                             synagogue','shopping_mall',
14                      'gas_station', 'amusement_park' ,'
                           art_gallery', 'bakery', 'beauty_salon'
                           ,
15                      'bicycle_store', 'book_store' , '
                           bus_station', 'car_repair', '
                           clothing_store',
16                      'convenience_store', 'dentist', '
                           department_store', 'drugstore',
17                      'electronics_store' , 'furniture_store',
                           'gym','hair_care','hardware_store'
18                      ,'home_goods_store','jewelry_store','
                           laundry','movie_theater','pet_store',
19                      'post_office','tourist_attraction']
20
21
22          url = 'https://maps.googleapis.com/maps/api/place/
                nearbysearch/json'
23          # Liste pour stocker les informations sur les lieux
24          places_data = []
25          # Parametres de la requete de recherche
26          for category in categories:
27              params = {
28                  'key': 'AIzaSyCuOrD_pMYPwHa3HIAL6SIxdY070eqqncw
                        ',
29                  'location': '46.56675000,-72.74913000',  #
                        Coordonn es g ographiques du centre de la
                        zone de recherche
```

```
30                'radius': '15000',
31                'types': category
32            }
33        # Envoi de la requete GET a l'API Places
34            response = requests.get(url, params=params)
35
36        # Analyse de la reponse JSON
37            data = response.json()
38        # Initialisation des listes de coordonnees
39            latitudes = []
40            longitudes = []
41
42
43            if 'results' in data:
44                for result in data['results']:
45                    location = result.get('geometry', {}).get('
                          location', {})
46                    latitude = location.get('lat', None)
47                    longitude = location.get('lng', None)
48                    name = result.get('name', 'N/A')
49                    if latitude is not None and longitude is
                          not None:
50
51                        point = Point(longitude, latitude)
52
53                        DAUID = None
54                        print(f"{name} ({category}): {latitude
                              }, {longitude}")
55
56                        for index, region in regions.iterrows()
                              :
```

```
                            if region['geometry'].contains(
                                point):
                                DAUID = region['DAUID']
                                break

                        if DAUID:
                            print("Le point se trouve dans la
                                region avec DAUID :", DAUID)
                        else:
                            print("Aucune region correspondante
                                trouvee pour le point.")

                        places_data.append({'Nom': name, 'DAUID
                            ': DAUID})

        places_df = pd.DataFrame(places_data)
        # Supprimer les doublons des noms des lieux
        places_df = places_df.drop_duplicates(subset=['Nom'])
        places_df = places_df.reset_index(drop=True)
        business_count = places_df.groupby('DAUID').size().
            reset_index(name='Nombre de commerces')
            business_count.to_csv('datasets/2021/'+ filename ,
            index=False)
        print(business_count)
```

```

def graph_couleur_indice(low, high,Shawinigan_2021_gdf=''):

    red_filtered_gdf = Shawinigan_2021_gdf[Shawinigan_2021_gdf[
        'indice'] <= low]
    yellow_filtered_gdf = Shawinigan_2021_gdf[
```

```
6         (Shawinigan_2021_gdf['indice'] > low) & (
              Shawinigan_2021_gdf['indice'] < high)]
7     green_filtered_gdf = Shawinigan_2021_gdf[
          Shawinigan_2021_gdf['indice'] >= high]
8
9     fig, ax = plt.subplots(figsize=(10, 6))
10    red_filtered_gdf.plot(edgecolor='red', facecolor='none', ax
          =ax, linewidth=4)  # Tous les polygones sauf celui avec
          DAUID==24360079
11    yellow_filtered_gdf.plot(edgecolor='black', facecolor='none
          ', ax=ax)  # Le polygone avec DAUID==24360079
12    green_filtered_gdf.plot(edgecolor='green', facecolor='none'
          , ax=ax, linewidth=4)
13    plt.show()
```

Annexe B

Article

District Vitality Index Using Machine Learning Methods for Urban Planners

Sylvain Marcoux
sylvain.marcoux@cegeptr.qc.ca
Université du Québec à Trois-Rivières
Trois-Rivières, Québec, Canada

Jean-Sébastien Dessureault
Université du Québec à Trois-Rivières
Trois-Rivières, Canada
jean-sebastien.dessureault@uqtr.ca

ABSTRACT

City leaders face critical decisions regarding budget allocation and investment priorities. How can they identify which city districts require revitalization? To address this challenge, a Current Vitality Index (CVI) and a Long-Term Vitality Index (LVI) are proposed. These indexes are based on a carefully curated set of indicators. Missing data is handled using K-Nearest Neighbors (KNN) imputation, while Random Forest (RF) is employed to identify the most reliable and significant features. Additionally, k-means clustering is utilized to generate meaningful data groupings for enhanced monitoring of Long-Term Vitality. Current vitality is visualized through an interactive map, while Long-Term Vitality is tracked over 15 years with predictions made using Multilayer Perceptron (MLP) or Linear Regression (LR). The results, approved by urban planners, are already promising and helpful, with the potential for further improvement as more data becomes available. This paper proposes leveraging machine learning methods to optimize urban planning and enhance citizens' quality of life.

CCS CONCEPTS

• **Computing methodologies → Classification and regression trees**.

KEYWORDS

Supervised learning methods, Unsupervised learning methods, Explainability, Smart city, Urbanisation

ACM Reference Format:
Sylvain Marcoux and Jean-Sébastien Dessureault. 2025. District Vitality Index Using Machine Learning Methods for Urban Planners. In *Proceedings of (District Vitality Index Using Machine Learning methods).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In a modern urban context, city districts play a central role as fundamental units of urban life. They are places where social bonds are formed, communities thrive, and the challenges and opportunities of urban life manifest. However, cities face numerous ongoing economic and social challenges, making it difficult to address all of these issues and make informed decisions. Cities face challenges such as rising rents and a low rental vacancy rate, which contribute to a housing crisis, as highlighted by [14]. Additionally, insufficient information on city indicators often leads to inequalities in resource allocation and investments between city districts, contributing to a decline in vitality in certain urban areas over time [1], regarding city revitalization projects. In this context, it becomes imperative to have a tool to monitor and evaluate the impact of these changes on the lives of citizens and city districts. Evaluating the vitality of urban districts is essential for better understanding local needs and improving the quality of life of urban residents.

The remainder of this paper is organized as follows: Section 2 details some research reviews, Section 3 presents the proposed methodology, Section 4 presents the results, Section 5 discusses the results and their implications, and Section 6 concludes the research.

2 LITERATURE REVIEW

J. Jacobs, author of *The Death and Life of Great American Cities* [9], was among the first to focus on urban vitality, highlighting elements such as urban density, land use diversity, and urban planning. Jacobs criticizes modern city planning for large-scale developments that prioritize structural expansion over the needs and scale of human interaction. In her methodology, she prioritizes empirical observation over large theoretical models. Also, J. E. Drewes and M. Van Aswegen [6], and X. Li and al. [10], define vitality as the ability of an entity to survive or function effectively. Thus, an urban Vitality Index represents the capacity of an urban center to remain viable, meet community needs, and enhance residents' quality of life. Drewes and Van Aswegen's index is constructed using normative scores for welfare, satisfaction, social dynamics, and spatial distribution, calculated based on the average quartiles of the associated variables. Meanwhile, Dessureault et al. [5] include eight variables, such as material and social deprivation indices, housing condition, property values, vacancy rates, renovation permit values, and rents. They use a genetic algorithm and k-means to cluster 135 areas into 10 clusters, while also using RF for weighting features. On the other hand, Haynes, Hook, Chiodi Grensing, and Ecklund [7] propose a Metropolitan Vitality Index (MVI) with 24 indicators, including the number of small businesses and physical activity spaces. The studied areas were ranked from 1 to 381 for each indicator, and those rankings were then summed. Finally, the MVI values were recalculated within a range of 80 to 120, with higher scores indicating better performance. K. Scott [16] suggests a Community Vitality Index based on the direction of data trends over time for variables such as property crimes, volunteerism, and violent crimes. Analogously, Liu, Gou, and Xiong [11] examine

District Vitality Index Using Machine Learning methods, March 25,2025, Québec, Canada

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

[1] https://ici.radio-canada.ca/plusieurs_2019

growth, diversity, and mobility indicators, such as youth proportion, economic and racial diversity, and service proximity. MinMax standardization is applied to the indicators, and feature weighting is performed using the entropy method, along with scores provided by urban planning scholars. Additionally, correlations between different dimensions of urban vitality are analyzed. Herath and Mittal [8] provide a comprehensive review of smart cities and artificial intelligence (AI), citing Singapore and Zurich are among the top ten smart cities. Eighty percent of studies on urbanization and machine learning use clustering, as detailed by Wang and Biljecki [17], with popular algorithms such as k-means, self-organizing maps, and DBSCAN. Unsupervised algorithms are also used, as demonstrated by Dessureault [5] and Li [10]. To evaluate the effectiveness of clustering, P.J. Rousseeuw introduced the Silhouette Score [15], a metric that measures how well points are grouped within their respective clusters. The k-means algorithm, used for the classification of areas, was originally developed by S. Lloyd [12]. This unsupervised clustering method is renowned for its simplicity, efficiency, and speed, as it minimizes the sum of squared distances between data points and their respective cluster centers, called 'centroids'. Additionally, data imputation is performed using K-Nearest Neighbors, introduced by Thomas Cover and Peter Hart in 1967 [4]. Furthermore, feature weighting is done using the RF algorithm, introduced by Leo Breiman in 2001 [2].

3 METHODOLOGY

In this paper, two indices are developed. On the one hand, the CVI allows us to assess the vitality of city districts at the present moment using 13 variables. On the other hand, the LVI allows us to monitor four variables over 15 years. Figure 1 illustrates the architecture of the framework. The preprocessing box represents the preprocessing stage, including imputation, normalization, and inverse transformation of the data. The second box represents the Kmeans clustering of DAs into three districts. The CVI box corresponds to the CVI calculation, highlighting the most important features using RF. The LVI box represents the LVI calculation for the three districts and the 2026 prediction using either MLP or LR. The center of Figure 1 presents the explainability stage and the outputs of the process, including maps and valuable insights for urban planners.

3.1 Current Vitality Index

The CVI is made up of 13 variables that come from two main sources: the Government of Canada and the City of Shawinigan. Data are first obtained in groupings called dissemination areas (DAs). A dissemination area (DA) is a small and relatively stable geographic unit composed of one or more adjacent dissemination blocks. DAs cover all the territory of Canada. These groups (DAs) can be visualized in Figure 3. The indicators included in the Index are: *Proportion of apartments that need minor and major repairs*, *Material and Social Deprivation Index*, *Average housing costs for owners*, *Average renter housing costs*, *Unoccupied housing rate*, *Average value of renovations*, *Average value of construction*, *Average value of buildings*, *Number of vacant buildings*, *Proportion of young people*, *Number of businesses*. A Google Geolocation API has been utilized to obtain

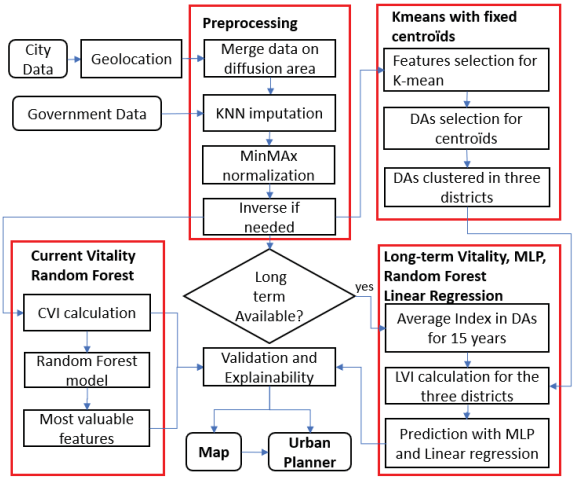


Figure 1: Method's architecture

Number of businesses. Additionally, for *The average cost of renovation and construction*, this API was essential for linking addresses to their corresponding dissemination areas. For data preprocessing, a MinMax scaling function was applied. The specific preprocessing steps for each feature are detailed in Table 1.

Treatments	Indicators
MinMax	All of them
Inversion	Material and Social deprivation Index Unoccupied housing rate Number of vacant buildings Prop. of apartments that need minor repairs Prop. of apartments that need major repairs
Imputation	Material and Social deprivation Index Average tenant housing costs

Table 1: Feature's preprocessing

The CVI for dissemination area k is calculated with (1)

$$CVI_k = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

Where CVI_k is the CVI for the dissemination area k, n is the number of variables, x_i is the value of variable i for DA k and k represents the DA. In this study, the calculation of the CVI includes 13 variables (n) across 87 dissemination areas (k). To visualize the boundaries of dissemination areas (DAs) on a map of the city based on their CVI, a Python procedure was developed using the `Geopandas` library. This interactive map allows users to visualize the values of indicators for a selected dissemination area. Additionally, the DAs are color-coded (heat map) based on the CVI values (Figure 3). Another procedure has also been implemented to visualize the boundaries of DAs with low indices compared to those with high indices.

3.2 Variable Importance in Vitality Index

A critical question is determining which variable has the greatest influence on the Current Vitality Index. Identifying this variable would enable city policymakers to strategically invest in areas where improvements in this feature could enhance vitality scores. The RF algorithm was used to rank variables by importance. RF is a supervised learning algorithm based on decision trees. Each optimal split within a tree reduces data impurity, measured by the Gini index. The reduction in impurity is calculated by comparing impurity before and after the split. The target variable for this algorithm, the CVI, was selected. The `RandomForestRegressor` implementation from the *scikit-learn* library was used with the parameter `n_estimators=500`, wich represents the number of decision trees. The variables were ranked by importance and visually presented in a bar chart to communicate the results to urban planners. To validate the results obtained with `RandomForestRegressor`, the `XGBRegressor` algorithm from the `xgboost` library was also employed to assess variable importance. This algorithm, renowned for its efficiency, speed, and performance in handling large datasets, is widely recognized in predictive modeling competitions, as highlighted by T. Chen and C. Guestrin in their work, "XGBoost: A Scalable Tree Boosting System" [3].

3.3 Long-Term Vitality Index

To derive an exploitable LVI and enable an in-depth analysis, it was necessary to group the 87 dissemination areas. The `K-means` unsupervised algorithm from the *scikit-learn* library was used for this clustering. While initial cluster centers are typically chosen randomly, alternative methods exist for their selection, as highlighted in the article *FC-KMeans: Fixed-centred k-means algorithm* [1]. In this study, the `init` parameter was used to define these initial centroids. Specifically, the chosen centroids correspond to the characteristics of three districts: Urban, Residential, and Commercial. This classification facilitates the comparison and interpretation of trends across several years within these different types of districts. In agreement with urban planners, the following variables were used to optimize the clustering procedure: *Population density per square kilometer*, *Proportion of young people*, *Number of businesses*, and *Average building value*. This dimensionality reduction from the initial 13 variables ensured coherent clustering and a strong Silhouette score. The selected variables were chosen, in accordance with urban planner, to clearly differentiate the three defined districts. The *Population density* indicator characterizes urban city districts, while the *Proportion of young people* helps distinguish residential areas. Finally, the *Number of businesses* and *Average building value* variables facilitate clustering dissemination areas within the commercial sector. For the selection of initial centroids, a dissemination area that most accurately represents each of the three groups was chosen. To assess the cohesion and quality of the clustering, the Silhouette Index was used. The dissemination areas within the three districts were used to calculate the LVI for these groupings. An average of all indicators within each sector was computed to derive three Long-Term Vitality Indices, one for each sector, for the following years: 2006, 2011, 2016, 2021. A MLP type neural network, with two hidden layers, and LR are used to perform regression for the next census year, 2026. The `MLPRegressor` and `LinearRegression` algorithms from the scikit-learn library are employed for this task. To evaluate the quality of the predictions, the least squares method is applied, and the mean_squared_error function from the sklearn.metrics library is used to calculate the error.

The LVI for sector s is calculated with (2)

$$LVI_s = \frac{1}{nbrAD_s} \sum_{i=1}^{nbrAD_s} CVI_i \tag{2}$$

Where LVI_s is the LVI for sector s, $nbrAD_s$ is the number of dissemination areas within sector s, CVI_i is the CVI of dissemination area i and
$s \in$ {Urban, Commercial, Residential}

3.4 Explainability

To further explain the results of the `RandomForestRegressor` regarding the most influential variables in the CVI, SHAP (Shapley Additive Explanations) values were computed and represented using violin plots, as shown, for example, in Figure 2. In this figure, the positive red SHAP values indicate that the feature is positively correlated with higher CVI values in the regression model's predictions [13]. SHAP is a post-hoc method that explains models through an independent, model-agnostic process external to the model itself. In contrast, ante-hoc methods are inherently interpretable, meaning their interpretability is intrinsic to the model's structure and decision-making process. The clustering of dissemination areas used in the LVI will also be validated and explained using SHAP values. Additionally, radar charts will be generated to assess the internal consistency of each grouping (Figures 4). An interpretation of all these values will be automatically generated through a Python procedure and included in a PDF file. This file will feature both explanatory charts and interpretative text. The charts will include violin plots and horizontal bar plots for SHAP values. A distinct chart will be generated for each defined sector: Urban, Residential, and Commercial. Using these tools, city decision-makers will gain a clearer understanding of the factors influencing the ranking generated by the RF algorithm and the clustering of dissemination areas with k-means. This analysis will help them identify key variables impacting the rankings and enable more informed decision-making based on these insights.

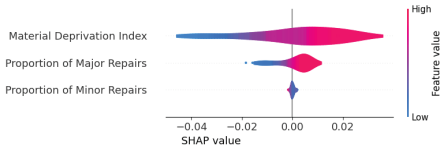


Figure 2: SHAP impact on model output magnitude

4 RESULTS

4.1 Current Vitality Index

To analyze the distribution of the CVI, a histogram was created by dividing the data into eight classes. The resulting histogram indicates that the CVI follows a normal distribution. Figure 3 represents an interactive heat map based on the CVI. Users can visualize the

values of the indicators by selecting a dissemination area. In Figure 3, dissemination area 24360085 is selected.

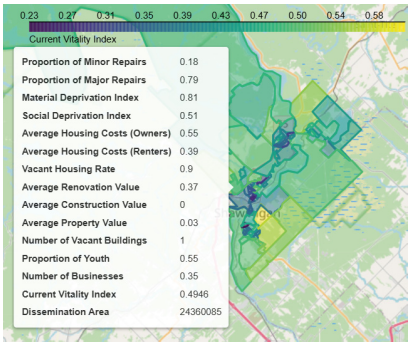


Figure 3: Heat Map of Current Vitality Index

4.2 Long-Term Vitality Index

The clustering of dissemination areas using the K-means algorithm with the three fixed centroids for the LVI can be analyzed through stacked radar charts. Figure 4 presents the radar chart for the *Residential* cluster. A visual analysis reveals that this cluster exhibits a high degree of internal consistency. In contrast, the *Commercial* cluster consists of dissemination areas with more varied characteristics. With this information, urban planners can validate the three clusters.

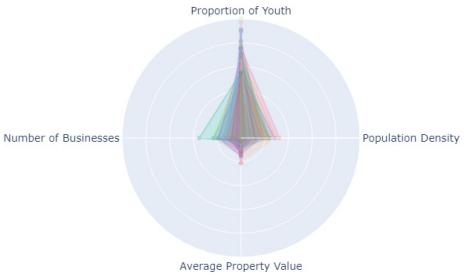


Figure 4: Residential cluster of K-means

As shown in Figure 5, the average Silhouette score obtained by the K-means algorithm is 0.3229 as shown by the dotted red line. While this index is not exceptionally high or close to 1, it indicates moderate cohesion between points within the clusters and partial separation between the clusters. The three colors represent the clusters, and the varying bandwidths indicate that some clusters include more data than others. Additionally, negative values in the first cluster suggest that those points may be misclassified.

SHAP values are used to improve the explainability of the k-means clustering results. In Figure 6, the average SHAP values show that the variables *Proportion of Youth* and *Population Density* are the most influential in the Urban (Class 0) and Residential (Class 1) categories. In contrast, the variable *Number of Businesses* plays

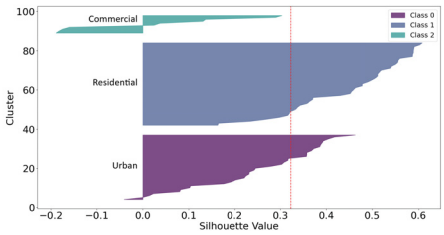


Figure 5: Clustering consistency using Silhouette score

a key role in defining the Commercial (Class 2) category. Finally, the *Average Property Value* indicator has minimal influence on the overall clustering.

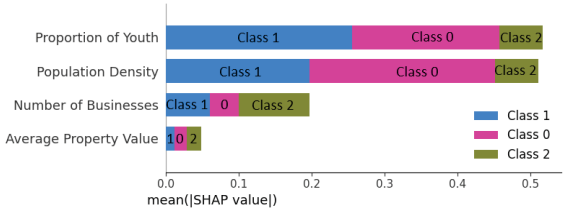


Figure 6: SHAP Average impact on model output magnitude

Figure 7 illustrates the evolution of the LVI from 2006 to 2021, segmented by the three neighborhood clusters: Urban, Residential, and Commercial. The figure also includes projections for 2026, shown as dotted lines, which are obtained using a LR model. Table 2 presents the Mean Squared Errors (MSE) for each predictive model. Among the three models evaluated, LR emerges as the most accurate, achieving the lowest MSE.

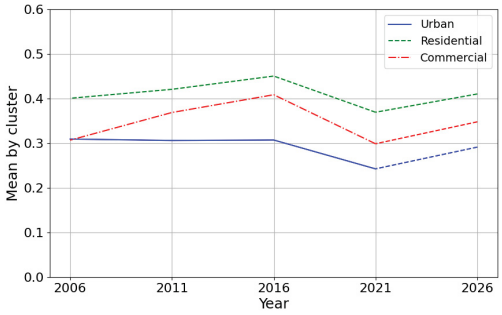


Figure 7: Time series of the LVI

5 DISCUSSION

It can be challenging to obtain very high values when we have more than two dimensions in our data. This justifies the dimensionality reduction to 3, rather than using all of our features. However,

Cluster	MLP Regressor	RF	LR
Urban	0.0022	0.0017	1.3e-33
Residential	0.0023	0.00052	2.8e-32
Commercial	0.0023	0.00036	0.00017

Table 2: Mean Squared Errors of the Models by Cluster

the obtained clusters align well with the three types of neighborhoods defined for the city: Urban, Residential, and Commercial. This suggests that the initial centroids chosen for the clustering were relevant for this classification. Furthermore, urban planners in Shawinigan have confirmed that the dissemination areas within the clusters correspond closely to the city's realities. These findings indicate that the three clusters can effectively illustrate long-term trends. Analyzing the clustering using SHAP values highlights the significant impact of demographic characteristics on the classification of dissemination areas. However, the Commercial cluster remains more challenging to delineate compared to the others. A general decline in the LVI is observed in all three clusters between 2016 and 2021. This decrease may be related to the global COVID-19 pandemic, which impacted the clusters' various economic and social aspects, reducing their vitality. The LR results suggest a potential stabilization or possible recovery by 2026, although the trends remain uncertain and depend on future economic and social conditions. This combined approach, which takes advantage of both historical data and AI-based predictions, provides a comprehensive overview of the past and future evolution of neighborhood vitality in Shawinigan. Table 3 highlights the enhancements made over the Vitality Index (VI) of Dessureault et al. [5]. Urban planners utilizing the two Vitality Indexes introduced here can analyze urban vitality more effectively using the interactive maps. Furthermore, the clustering approach, which employs fixed centroids, results in significant clusters validated by urban planners.

Aspects	Dessureault Vitality Index (VI)	CVI and LVI
Imputation with RF	X	✓
Identification of Significant Clusters	X	✓
Control Over Regression Models	X	✓
Control Over Indicators	X	✓
Interactive Map Features	X	✓

Table 3: Comparison of improvements over Dessureault et al. [5]

6 CONCLUSION

This paper examines the vitality of Shawinigan's city districts using two key indicators: the CVI and the LVI. By classifying 87 dissemination areas into Urban, Residential, and Commercial categories through k-means clustering, significant socio-economic dynamics were identified. These indexes are designed to monitor vitality and identify priority areas for investment within the city. In summary, our study found that residential areas exhibited greater vitality, supported by favorable socioeconomic conditions and active investments in renovations and construction. In contrast, urban and commercial zones require targeted interventions to enhance quality of life and drive revitalization. These findings underscore the effectiveness of monitoring vitality through these indexes, which serve as essential tools for achieving this paper objective. We encountered limitations in data accessibility due to the smaller size of the city studied. Unlike larger cities in Canada, smaller ones often lack comprehensive and readily available data, which posed challenges for this analysis. In addition, extensive research on the indicators available over a 15-year period could be conducted to improve the LVI. This research provides a solid foundation for intelligent and sustainable urban planning, aiming to significantly improve citizens' quality of life.

ACKNOWLEDGMENTS

The authors thank Mr. Marc-Antoine Langlois, Division Manager - Planning and Sustainable Development Spatial Planning Department, for his invaluable guidance and insightful discussions throughout this research.

REFERENCES

[1] Merhad Ay, Lale Özbakır, Sinem Kulluk, Burak Gülmez, Güney Öztürk, and Sertay Özer. 2023. FC-Kmeans: Fixed-centered K-means algorithm. *Expert Systems with Applications* 211 (Jan. 2023), 118656. https://doi.org/10.1016/j.eswa.2022.118656
[2] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (Oct. 2001), 5–32. https://doi.org/10.1023/A:1010933404324
[3] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794. https://doi.org/10.1145/2939672.2939785 arXiv:1603.02754 [cs].
[4] T. Cover and P. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13, 1 (Jan. 1967), 21–27. https://doi.org/10.1109/TIT.1967.1053964 Conference Name: IEEE Transactions on Information Theory.
[5] Jean-Sébastien Dessureault, Jonathan Simard, and Daniel Massicotte. 2021. Unsupervised Machine learning methods for city vitality index. https://doi.org/10.48550/arXiv.2012.12082 arXiv:2012.12082 [cs].
[6] J. E. Drewes and M. Van Aswegen. 2010. Determining the vitality of urban centres. 15–25. https://doi.org/10.2495/SW100021
[7] Monica Haynes, Alexander Hook, Gina Chiodi Grensing, and Hattie Ecklund. 2018. *Analysis of Duluth-Superior and Peer Metropolitan Areas.* Report. University of Minnesota Duluth. http://conservancy.umn.edu/handle/11299/203562 Accepted: 2019-06-12T16:39:07Z.
[8] H. M. K. K. M. B. Herath and Mamta Mittal. 2022. Adoption of artificial intelligence in smart cities: A comprehensive review. *International Journal of Information Management Data Insights* 2, 1 (April 2022), 100076. https://doi.org/10.1016/j.jjimei.2022.100076
[9] Jane Jacobs. 1993. *The Death and Life of Great American Cities.* Vintage Books.
[10] Xin Li, Shidan Cheng, Zhihan Lv, Houbing Song, Tao Jia, and Ning Lu. 2020. Data analytics of urban fabric metrics for smart cities. *Future Generation Computer Systems* 107 (June 2020), 871–882. https://doi.org/10.1016/j.future.2018.02.017
[11] Haimeng Liu, Peng Gou, and Jieyang Xiong. 2022. Vital triangle: A new concept to evaluate urban vitality. *Computers, Environment and Urban Systems* 98 (Dec. 2022), 101886. https://doi.org/10.1016/j.compenvurbsys.2022.101886
[12] S. Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28, 2 (March 1982), 129–137. https://doi.org/10.1109/TIT.1982.1056489 Conference Name: IEEE Transactions on Information Theory.
[13] Mouad Louhichi, Redwane Nesmaoui, Marwan Mbarek, and Mohamed Lazaar. 2023. Shapley Values for Explaining the Black Box Nature of Machine Learning Model Clustering. *Procedia Computer Science* 220 (Jan. 2023), 806–811. https://doi.org/10.1016/j.procs.2023.03.107
[14] Minh Nguyen. 2023. Crise du logement : portrait d'une dure réalité. https://www.lacsq.org/actualite/crise-du-logement-portrait-dune-dure-realite/
[15] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (Nov. 1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7
[16] Katherine Scott. 2010. Katherine Scott | Canadian Index of Wellbeing. https://uwaterloo.ca/canadian-index-wellbeing/people-profiles/katherine-scott
[17] Jing Wang and Filip Biljecki. 2022. Unsupervised machine learning in urban studies: A systematic review of applications. *Cities* 129 (Oct. 2022), 103925. https://doi.org/10.1016/j.cities.2022.103925