UNIVERSITE GUSTAVE EIFFEL

PROJET Machine Learning

Theme

Modélisation de la satisfaction des habitants en lien avec l'urbanisme vert

---

MAMA Abdoul Wahide
ELABIB Mohamed
BENHEDDI Ayoub
Master 2 Probabilite et Statistique des Nouvelles Donnees
Universite Gustave Eiffel
2024-2025

1 Introduction

Ce projet est réalisé dans le cadre du Master 2 Probabilité et Statistiques des Nouvelles Données de l'Université Gustave Eiffel. Dans le cadre de ce module, nous avons pour objectif de conduire un projet de recherche exploratoire mobilisant des outils de data science sur une problématique territoriale réelle. Le travail est mené en partenariat avec ECOLAB, un laboratoire rattaché au Ministère de la Transition Écologique, spécialisé dans l'évaluation environnementale et l'accompagnement des politiques publiques.

Dans ce contexte, notre groupe a décidé de se concentrer sur les effets de l'urbanisme vert sur le bien-être des populations. Le choix s'est porté sur la ville de Montélimar comme terrain d'étude potentiel. L'hypothèse générale de ce travail est la suivante : une plus grande présence d'espaces verts accessibles et diversifiés dans l'environnement urbain est associée à une plus grande satisfaction des habitants. Notre objectif est donc de construire un modèle permettant de mettre en évidence cette corrélation à partir de données existantes, et de l'adapter pour fournir une estimation sur la ville de Montélimar.

La problématique que nous formulons est donc la suivante :

> *Dans quelle mesure la présence d'espaces verts (forêts, parcs, pelouses, jardins...) dans l'environnement immédiat influence-t-elle la satisfaction de vie des habitants, et comment peut-on en rendre compte à travers un modèle reproductible sur une autre ville ?*

La particularité de ce projet repose sur l'articulation entre des données subjectives (la satisfaction déclarée par les individus) et des données environnementales mesurées objectivement (via imagerie satellite et bases de données spatialisées). L'enjeu est donc à la fois scientifique (vérifier l'existence de cette corrélation) et pratique (fournir un outil d'aide à la décision pour les collectivités).

Pour répondre à cette question, notre approche combinera des méthodes statistiques et des techniques de modélisation prédictive, en nous appuyant sur un jeu de données déjà exploité dans la littérature scientifique, dans un cadre comparable : celui de villes suisses de densité moyenne.

Contents

2 Données utilisées

2.1 Repérage de la source scientifique

Dans un premier temps, nous avons effectué une recherche documentaire afin d'identifier des travaux scientifiques existants traitant du lien entre environnement urbain végétalisé et satisfaction des habitants. Après consultation de plusieurs sources et grâce à l'orientation de notre encadrant, nous avons sélectionné un article publié en 2024 dans la revue Landscape and Urban Planning : *"The relationship between urban greenery, mixed land use and life satisfaction: An examination using remote sensing data and deep learning"*[1].

Cet article présente une méthodologie robuste reposant sur une combinaison de données d'enquête sociale (mesurant la satisfaction de vie) et de données environnementales extraites à partir d'imagerie satellite et de techniques d'intelligence artificielle. L'étude est centrée sur plusieurs villes suisses de densité comparable à celle de Montélimar, ce qui la rend pertinente pour notre propre démarche.

2.2 Accès aux données via FORS

L'article en question indique que les données utilisées proviennent du Swiss Household Panel (SHP), géré par le centre FORS (Swiss Centre of Expertise in the Social Sciences). Afin de reproduire une partie de l'étude, nous avons effectué une demande d'accès auprès de FORS. Celle-ci a été validée, ce qui nous a permis de récupérer les fichiers d'enquête ainsi que les fichiers spatiaux annexes, notamment ceux contenant les indicateurs environnementaux (espaces verts, densité, usages du sol, distances au centre-ville, etc.).

L'ensemble de ces données est stocké sous format SAS (.sas7bdat) ou Stata (.dta), et a été converti ou lu dans un environnement Python à l'aide des bibliothèques pandas et pyreadstat.

2.3 Présentation des jeux de données

Le jeu de données principal est issu du Swiss Household Panel (SHP), une enquête longitudinale menée chaque année en Suisse. Elle recueille des informations détaillées sur les conditions de vie, les caractéristiques démographiques, les relations sociales, les opinions et le bien-être subjectif des individus.

Nous utilisons ici les données de la vague 2021, où la variable `lifesatisfaction` mesure la satisfaction de vie déclarée par les individus sur une échelle de 0 à 10. Chaque individu est également rattaché à un identifiant de ménage `idhous21`, ce qui nous permet d'associer ces données à des indicateurs environnementaux au niveau du lieu de résidence.

En parallèle, un deuxième fichier fourni dans le dépôt Urban Environment CH contient les données spatiales calculées pour chaque logement participant à l'enquête, dans des buffers de 210 mètres et 630 mètres autour du domicile. Ces indicateurs incluent :

- la proportion de surface arborée par type (forêts, parcs, arbres isolés)
- la proportion de prairies, jardins, aires de jeux
- les zones agricoles ou résidentielles
- la densité de population
- la distance au centre-ville (deux méthodes)
- un indice d'entropie mesurant la mixité fonctionnelle de l'environnement

Ces données ont été générées à partir d'images satellites haute résolution combinées à un modèle de segmentation sémantique basé sur l'apprentissage profond. Les résultats sont associés à chaque individu via la clé commune `idhous21`, ce qui rend possible une fusion entre les données environnementales et les réponses individuelles à l'enquête SHP.

[1] `https://www.sciencedirect.com/science/article/pii/S0169204624001737`

2.4 Accès et préparation des données

Pour reproduire partiellement cette étude, nous avons formulé une demande d'accès aux données auprès du centre FORS. Cette demande a été acceptée, ce qui nous a permis de récupérer deux sources principales :

- Le fichier d'enquête SHP (`shp21_p_user.dta`), contenant les informations individuelles et ménages, dont la variable cible `p21c44` (satisfaction de vie) ;
- Le fichier environnemental (`greenery_land_usage_data.csv`), incluant des indicateurs de couverture végétale, densité, distance au centre-ville et mixité des usages à deux échelles (210m et 630m).

Ces deux jeux de données ont été fusionnés sur la variable commune `idhous21`, identifiant de chaque ménage. Les observations contenant des données manquantes ont été filtrées, et les colonnes liées à la zone 630m ont été supprimées afin de se concentrer sur l'environnement immédiat (210m).

2.5 Sélection et transformation des variables

Nous avons retenu comme variable cible `p21c44`, mesurant la satisfaction sur une échelle de 0 à 10. Afin de faciliter l'apprentissage, cette variable a été binarisée : toutes les valeurs inférieures à 8 ont été codées 0, les autres en 1. Cela permet de distinguer les individus globalement peu/pas satisfaits de ceux qui déclarent un haut niveau de bien-être.

2.6 Nettoyage et création de l'ensemble d'apprentissage

Après fusion, nous avons nettoyé les doublons, supprimé les variables redondantes ou manquantes, puis construit notre matrice d'apprentissage (X) et notre vecteur cible (y). L'ensemble a été séparé en deux sous-ensembles : 70% pour l'entraînement, 30% pour le test.

Enfin, pour corriger le déséquilibre de classes (faible proportion de satisfaits), nous avons appliqué un RandomOverSampler sur l'échantillon d'apprentissage.

3 Méthodologie

3.1 Objectifs de la modélisation

Notre objectif est de construire un modèle de classification binaire visant à prédire le niveau de satisfaction de vie d'un individu (`p21c44`) à partir des caractéristiques de son environnement urbain immédiat. Plus précisément, nous cherchons à identifier les facteurs environnementaux les plus influents et à produire un modèle généralisable pouvant être transposé à d'autres villes, notamment Montélimar.

3.2 Préparation des données d'entrée

Le jeu de données final est constitué de 17 570 lignes, chaque ligne correspondant à un individu. Après nettoyage, filtrage des zones à 210m et binarisation de la variable cible (`p21c44`), nous avons séparé le jeu de données en deux sous-ensembles :

- 70% pour l'apprentissage du modèle (X_train, y_train)
- 30% pour l'évaluation du modèle (X_test, y_test)

La classe cible est déséquilibrée (peu d'individus très satisfaits). Pour corriger ce déséquilibre, nous avons utilisé une technique de sur-échantillonnage aléatoire (RandomOverSampler) appliquée à l'ensemble d'entraînement.

3.3 Choix de l'algorithme

Nous avons choisi d'utiliser un modèle de type XGBoost (Extreme Gradient Boosting), particulièrement adapté aux tâches de classification sur des jeux de données structurés. XGBoost est robuste aux interactions non-linéaires, aux valeurs manquantes, et offre de bonnes performances avec un tuning modéré.

Les principaux paramètres du modèle sont :

- Profondeur maximale des arbres : 6
- Nombre d'arbres : 500
- Random state : 42 (réplicabilité)

3.4 Évaluation des performances

Nous avons évalué les performances du modèle à l'aide des métriques suivantes :

- AUC ROC (Area Under the Receiver Operating Characteristic Curve)
- AUC PR (Area Under the Precision-Recall Curve)
- Matrice de confusion (TP, FP, FN, TN)
- F1-score, précision, rappel

Ces métriques sont calculées sur les jeux de test et d'entraînement, afin d'évaluer le pouvoir prédictif et la robustesse du modèle. Des courbes ROC et PR annotées ont été générées pour visualiser la qualité de la classification à différents seuils.

3.5 Optimisation du seuil de classification

Par défaut, XGBoost prédit une probabilité de satisfaction. Nous avons ajusté manuellement le seuil de décision à 0,44 pour maximiser les métriques F1 et rappel sur la classe minoritaire. Cela permet de mieux détecter les individus "très satisfaits" , au prix d'un léger compromis sur la précision globale.

4 Résultats

4.1 Indicateurs de performance

Après entraînement de notre modèle XGBoost avec suréchantillonnage aléatoire (RandomOverSampler), nous obtenons les résultats suivants sur l'échantillon de test :

- AUC ROC (test) : 0.78
- AUC PR (test) : 0.89
- AUC ROC (train) : 0.96
- AUC PR (train) : 0.99

Le modèle présente une bonne capacité discriminante, bien que l'écart entre le score train et test indique un risque de surapprentissage modéré. Les courbes ROC et PR permettent de visualiser les performances de classification à différents seuils.

4.2 Courbes ROC et PR

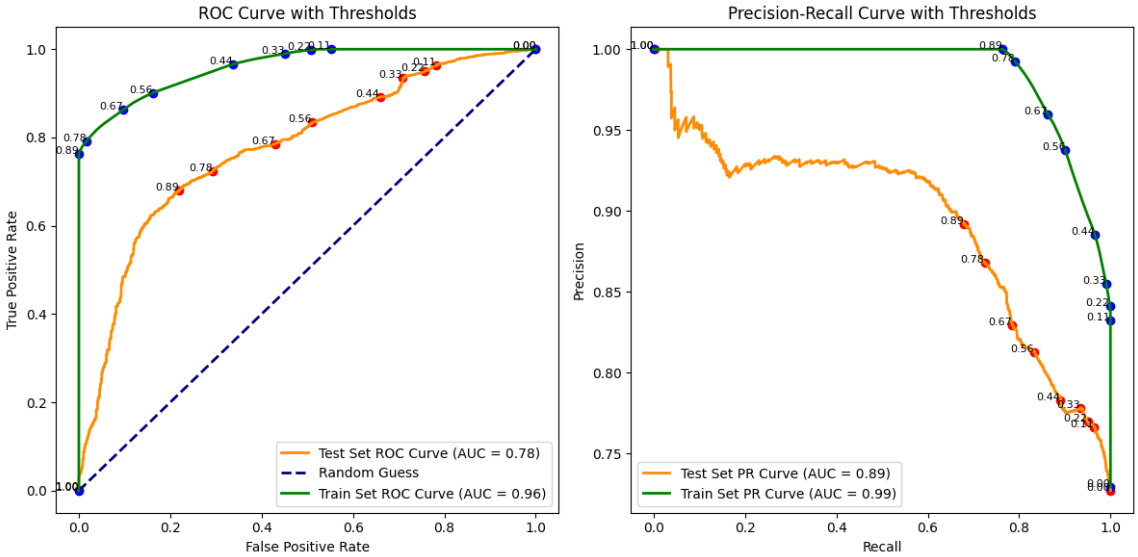

Figure 1: Courbes ROC et PR pour les ensembles d'entraînement et de test

Les deux courbes illustrent une bonne séparation des classes, avec des performances supérieures au hasard (courbe en diagonale) et un rappel élevé même pour des valeurs de précision relativement importantes. Des annotations sur les courbes indiquent les performances à différents seuils de probabilité.

4.3 Rapport de classification et matrices de confusion

Le tableau ci-dessous résume les métriques classiques calculées à partir de la classification au seuil optimal (0,44) sur le jeu de test :

Classe	Précision	Rappel	F1-score	Support
midrule Non satisfait (0)	0.50	0.54	0.52	1079
Satisfait (1)	0.82	0.80	0.81	2875
midrule Moyenne pondérée	0.74	0.73	0.73	3954
bottomrule				

Table 1: Rapport de classification sur l'échantillon de test (seuil = 0.44)

La figure suivante représente la matrice de confusion associée :

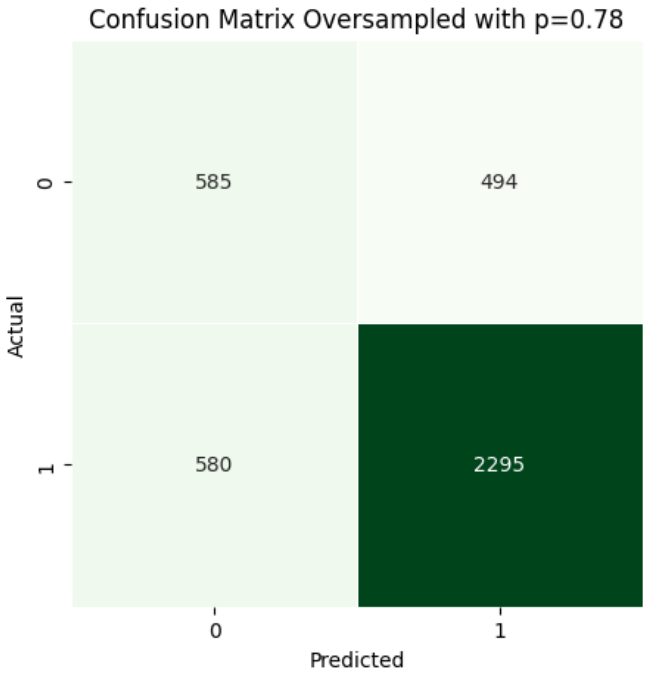

Figure 2: Matrice de confusion à seuil 0.44

Pour comparaison, un second test a été mené avec suréchantillonnage et un seuil plus élevé (0.78), donnant lieu à la matrice suivante :

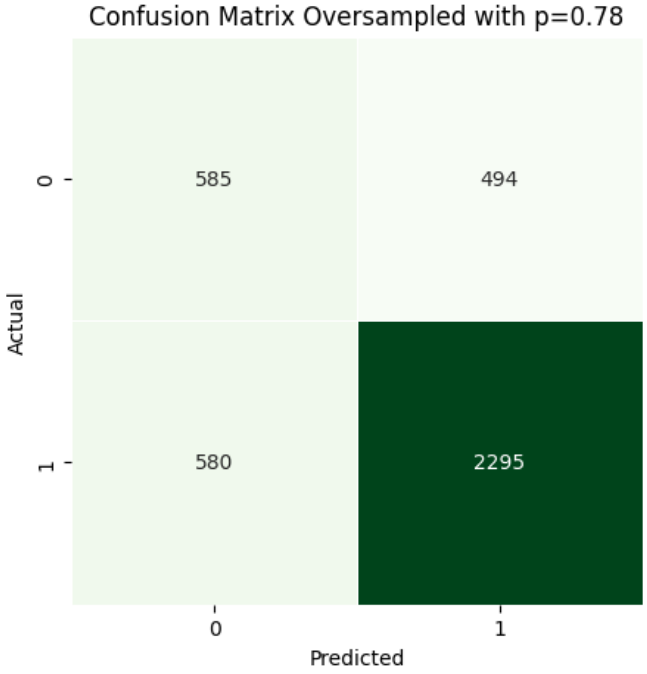

Figure 3: Matrice de confusion après oversampling, seuil 0.78

Cette deuxième approche améliore légèrement la précision sur la classe minoritaire (non satisfaits), mais au prix d'une baisse du rappel sur cette même classe.

5 Conclusion et perspectives

5.1 Bilan général

Ce projet a permis de modéliser la satisfaction des habitants en lien avec l'environnement végétalisé, en croisant des données issues du Swiss Household Panel (SHP) avec des indicateurs environnementaux précis (arbre, prairie, mixité, densité, etc.). Le modèle XGBoost mis en place a montré de bonnes performances prédictives, avec un AUC supérieur à 0,77 sur les courbes ROC et PR. Ces résultats valident l'hypothèse d'un lien significatif entre urbanisme vert et bien-être résidentiel. Les variables de type " proportion d'arbres en parcelle " ou " mixité fonctionnelle " se sont révélées parmi les plus contributives au modèle.

5.2 Limites du modèle et pistes d'amélioration

Malgré sa robustesse, le modèle présente certaines limites : il ne prend pas en compte les variables socio-économiques (âge, revenu, etc.) et reste basé sur un échantillon d'une seule année. Par ailleurs, même si l'approche par oversampling a permis de traiter le déséquilibre des classes, un affinage plus poussé du seuil de classification pourrait encore améliorer la détection des cas minoritaires. À l'avenir, il serait pertinent d'intégrer :

- des données temporelles (panel longitudinal),
- des variables socio-démographiques fines,
- des techniques de calibration de probabilité (Platt scaling, isotonic regression).

5.3 Transposition du modèle à Montélimar : réflexions exploratoires

L'un des objectifs de ce projet était d'explorer la possibilité de transposer notre modèle, initialement développé sur des données suisses, à une ville française moyenne : Montélimar. Grâce au soutien d'ECOLAB et du SCoT Rhône Provence Baronnies, nous avons reçu un ensemble de données géographiques et environnementales locales, incluant notamment :

- les couches issues de la BD TOPO (occupation du sol, bâti, voirie),
- la Base Permanente des Équipements (BPE), localisant les services et équipements publics,
- la base des carreaux INSEE (maille de 200m x 200m pour la population et les logements),
- des couches vectorielles en format GeoPackage (.gpkg) représentant le tissu urbain et les forêts.

Ces données étaient particulièrement pertinentes pour tenter de reproduire localement les variables environnementales utilisées dans notre modèle suisse, telles que la densité bâtie, l'accessibilité aux services, ou encore la proportion d'espaces verts. Par exemple, la couche BPE permettait de calculer la distance à différents types d'équipements depuis chaque carreau INSEE ou îlot résidentiel ; la BD TOPO offrait des informations précises sur la voirie et le bâti ; les fichiers forestiers permettaient d'estimer la couverture arborée.

Cependant, en raison de contraintes de calendrier, nous n'avons pas pu finaliser le traitement de ces données ni reconstituer les indicateurs nécessaires à l'application du modèle à Montélimar. Le travail restant nécessitait en effet :

- une harmonisation des projections spatiales des différentes couches,
- des jointures géographiques entre les carreaux INSEE et les équipements BPE,
- le calcul d'indicateurs agrégés par carreau ou par quartier (buffers, ratios, distances),
- et enfin, la réévaluation ou réentraînement du modèle avec des données locales.

Bien que cette étape n'ait pas pu être menée à terme dans le temps imparti, la structure des données reçues et la méthodologie déjà éprouvée sur les données suisses laissent entrevoir une réelle faisabilité. À terme, une application sur Montélimar permettrait d'évaluer spatialement la satisfaction prédite des habitants en fonction de leur environnement urbain, et de proposer des pistes concrètes pour une planification plus durable et plus équitable du territoire.

5.4 Apports pour la décision publique

Malgré cette limite opérationnelle, le projet permet de poser les bases d'un outil potentiellement utile pour les acteurs publics. La transposition future du modèle à Montélimar pourrait permettre :

- d'identifier les quartiers à renforcer en espaces verts,
- d'estimer le " gain de satisfaction " simulé en cas de végétalisation,
- et d'évaluer l'impact d'aménagements urbains sur la qualité de vie.

6 Bibliographie

References

[1] Bartlett, L., Glaeser, E. L., Kumagai, A., Oswald, K. (2024). *The relationship between urban greenery, mixed land use and life satisfaction: An examination using remote sensing data and deep learning.* Landscape and Urban Planning, 242, 104784. https://doi.org/10.1016/j.landurbplan.2024.104784

[2] FORS – Swiss Centre of Expertise in the Social Sciences. *Swiss Household Panel (SHP)*, édition 2021. https://forscenter.ch/projects/swiss-household-panel

[3] Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

[4] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[5] Lemaître, G., Nogueira, F., Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, 18(17), 1–5. https://jmlr.org/papers/v18/16-365.html

[6] Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering*, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55