Prédiction de l'environnement d'un quartier

Stage de Master 2

Nelly Barret
Université Claude Bernard Lyon 1

Encadrée par Fabien Duchateau et Franck Favetta

26 juin 2020

Contexte du stage

Le projet Home in Love

- Objectif : aider à la recherche immobilière dans le cadre de la mobilité professionnelle
- Projet pluridisciplinaire débuté en 2017

Contexte scientifique

- Comment qualifier simplement l'environnement d'un quartier ?

Objectif du stage

- Prédire l'environnement d'un quartier par apprentissage supervisé

État de l'art

- **Recommandation de logements** [Yuan et al., 2013]
  - 3 critères : localisation, prix, unité urbaine
  - Utilisation d'une ontologie et du *case-based reasoning*

- **Recommandation de quartiers** [Liu et al., 2014]
  - Prise en compte du voisinage
  - Similarité entre quartiers (matrice) et régions (*clustering*)

- **Comparaison manuelle de quartiers** : datafrance.info
  - 5 critères : éducation, santé, services, commerces, loisirs
  - Classement des communes et visualisation cartographique

Positionnement

Recommandation de quartiers basée sur une expertise sociologique

---

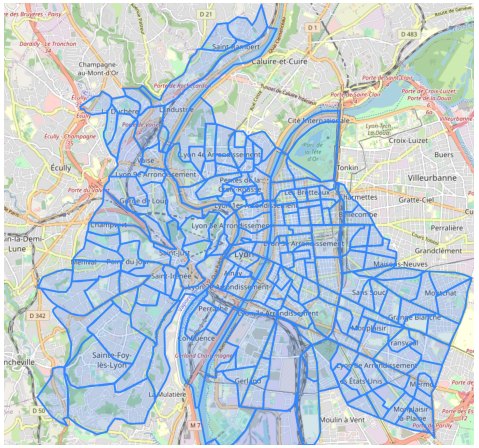
Toward a user-oriented recommendation system for real estate websites.

Exploiting geographical neighborhood characteristics for location recommendation.

Prérequis

Quartier [Humain-Lamoure, 2007]

- Selon l'INSEE, un IRIS → zone de 2000 à 5000 habitants
- En France, 50 000 IRIS



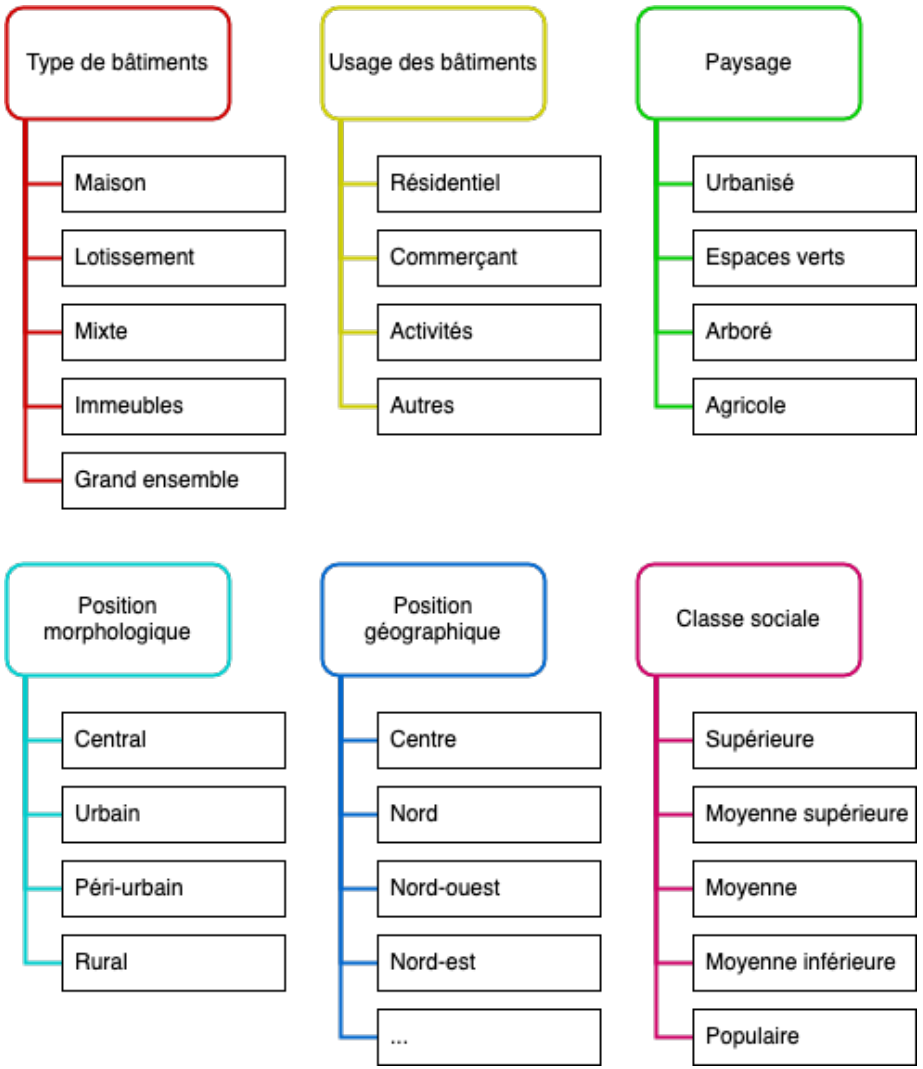
Indicateurs INSEE

- 641 indicateurs
- Nombre de restaurants, population entre 18 et 25 ans, ...

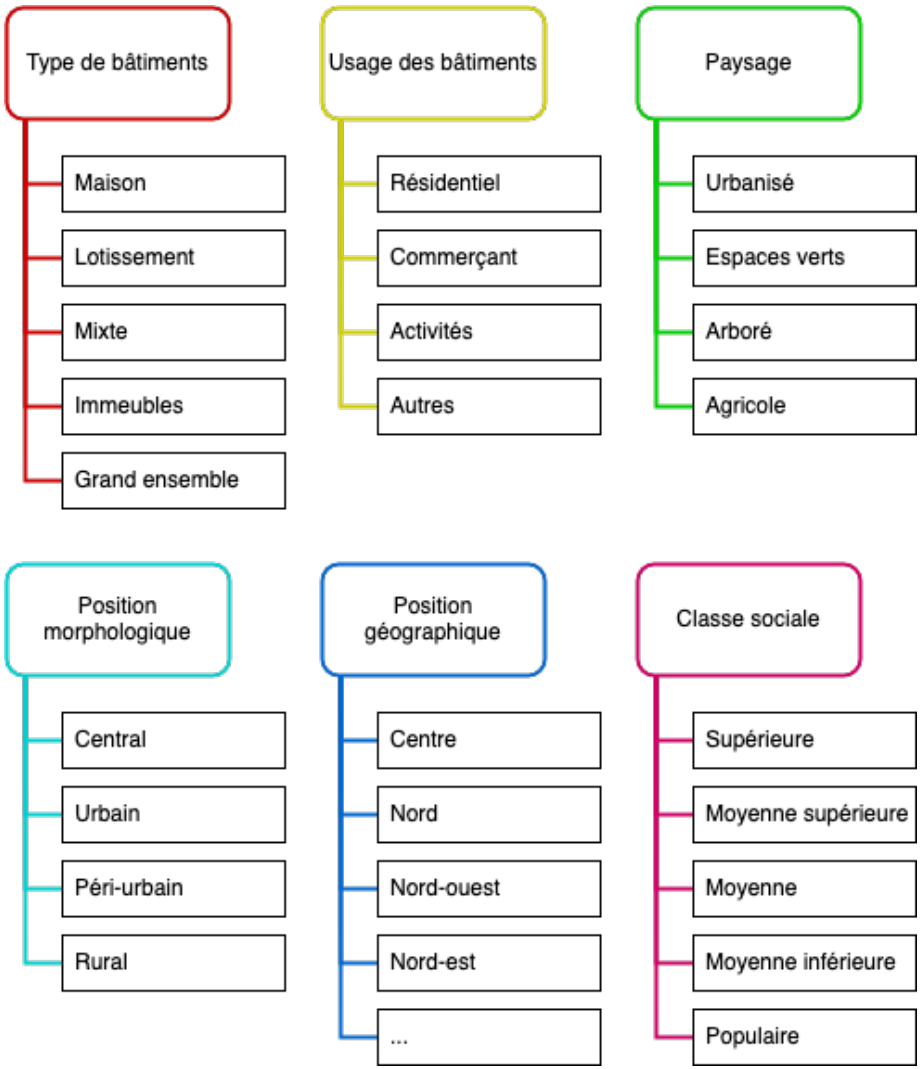
Variables d'environnement

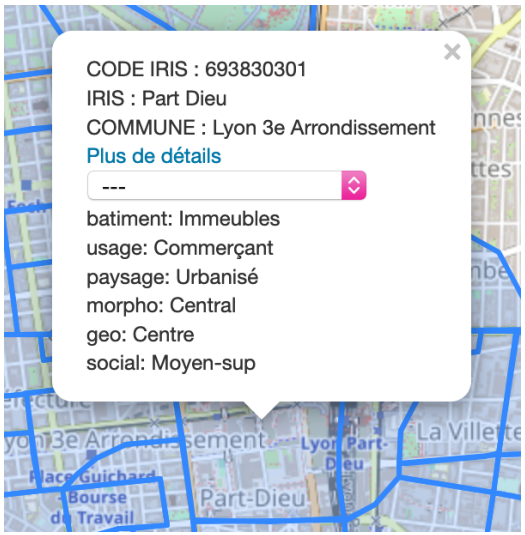
- Résultat d'une analyse qualitative des sociologues sur 300 IRIS
- Description simplifiée d'un quartier par 6 critères

Variables d'environnement

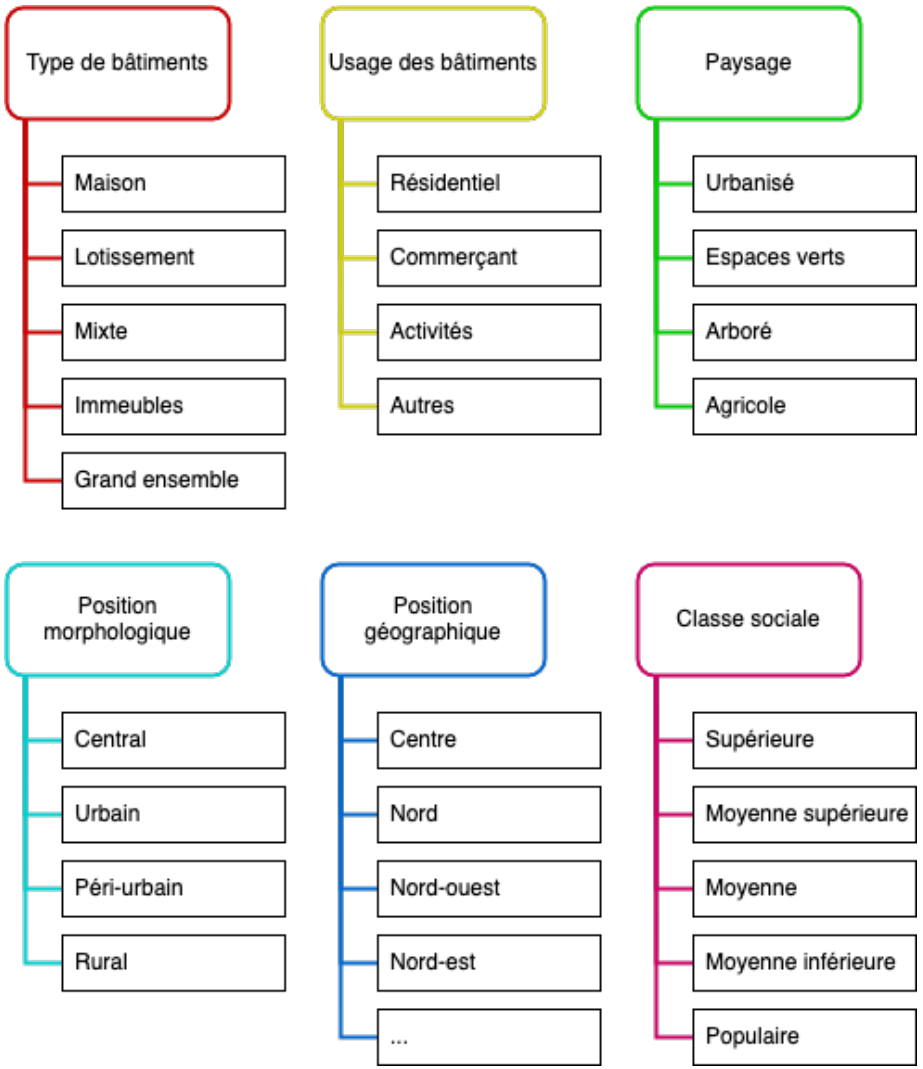


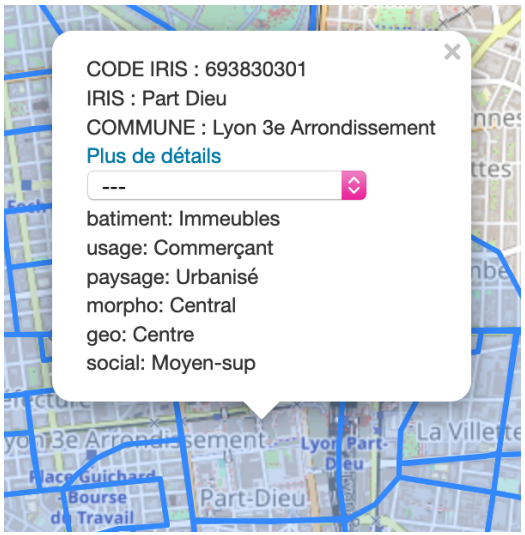
Variables d'environnement



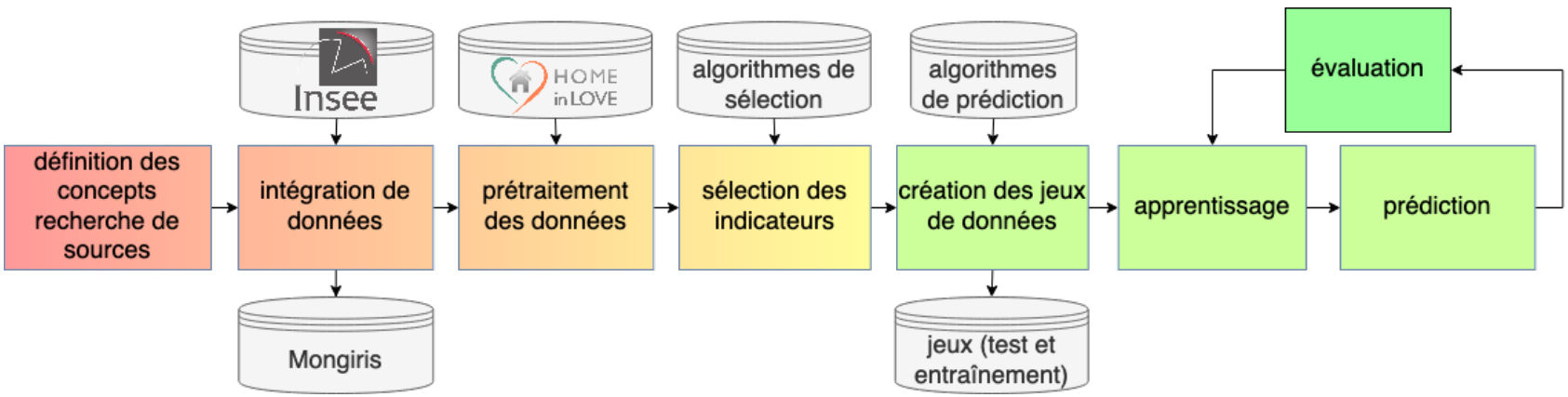


Variables d'environnement





L'approche Predihood



Pré-traitement des données

IRIS expertisés

- Transformation des adresses en IRIS
- Traitement des valeurs inconnues : médiane des valeurs
- Traitement des valeurs erronées : semi-automatique

Indicateurs INSEE

- Traitement des valeurs inconnues : médiane des valeurs
- Normalisation : densité de population

Représentativité des IRIS expertisés

Variable	**Catégorie**	**Expertise**	**France**
Position morphologique	rural	5%	68%
Paysage	rural agricole	17%	68%
Classe sociale	moyenne moyenne$^+$	82%	71%
Type de bâtiments	collectif	68%	44%
Position géographique	nord, sud, ...	équitablement répartie	
Usage des bâtiments	nécessite une analyse particulière		

Tableau : L'analyse des IRIS expertisés montre que la plupart des variables d'environnement comportent un biais.

Sélection des indicateurs

Objectif : générer des listes d'indicateurs utiles à la prédiction

Étapes :

- Filtrage des indicateurs
  - 17 sont descriptifs (code postal, code IRIS...)
  - 208 sont trop spécifiques (nombre de courts de tennis couverts)
  - 59 sont non renseignés
- Sélection d'un sous-ensemble parmi 363 indicateurs
  - Combinaison d'algorithmes
  - Approche alternative : distribution des indicateurs

Résultat : plusieurs listes de k indicateurs pour chaque variable d'environnement v, notée L_v^k

Sélection par combinaison d'algorithmes

- Matrice de corrélation : supprimer les indicateurs 100% corrélés
- Algorithmes RF et ET : classer les indicateurs par importance
- Prise en compte de la diversité des catégories d'indicateurs

Algorithme 1 : Sélection des indicateurs pertinents pour la prédiction

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

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

Interface cartographique

predihood

A tool for visualizing IRIS

203 iris found for query lyon.

Minimal zoom level to display IRIS automatically

12 (actual zoom level = **15**)

Search by IRIS code

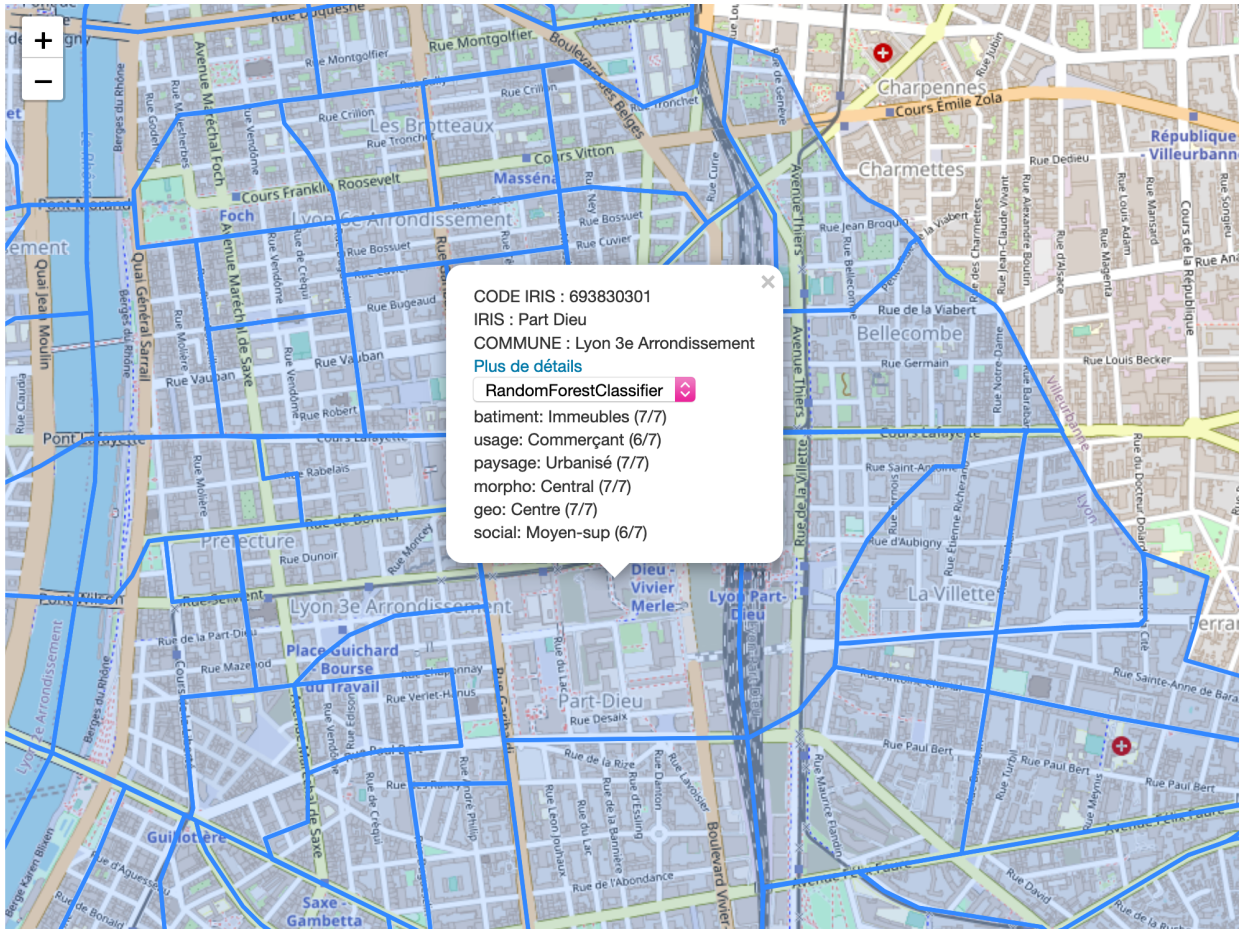
740560104

Search by IRIS name or city

lyon

Clear

LIRIS LABEX IMU UNIVERSITÉ DE LYON



Interface cartographique

predihood

A tool for visualizing IRIS

6 iris found for query miribel.

Minimal zoom level to display IRIS automatically

12 (actual zoom level = **13**)

Search by IRIS code

740560104

Search by IRIS name or city

miribel

Clear

LIRIS — LABEX IMU UNIVERSITÉ DE LYON

Interface de paramétrage

Validation expérimentale

Protocole

Niveau national

- Sélection de 5 algorithmes : Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbours (KNN), Support Vector Classification (SVC) et AdaBoost (AB)
- Calcul de la précision pour chaque variable selon les 7 listes et les 5 algorithmes

Niveau communal

- Analyse qualitative des résultats prédits pour Lyon
- Effectuée par les sociologues

Résultats au niveau national

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

Tableau : Qualité de prédiction pour la variable *usage*

- Les listes améliorent la précision pour presque tous les algorithmes

Résultats au niveau national

Variable d'environnement	$\mathcal{I}$	L^k
Type de bâtiments	57%	60% (L^{20})
Usage des bâtiments	64%	65% (L^{50})
Paysage	61%	63% (L^{20})
Classe sociale	51%	52%(L^{40})
Position géographique	34%	34% ($\mathcal{I}$)
Position morphologique	60%	61% ($L^{20,30,40}$)

Tableau : Qualité de prédiction pour l'algorithme Random Forest

- Random Forest est le meilleur algorithme
- La sélection permet une meilleure explicabilité des résultats

Conclusion et perspectives

Approche Predihood

- Algorithmes pour la prédiction de l'environnement (sélection des indicateurs et représentativité des données)
- Interfaces pour la visualisation des quartiers et le paramétrage d'algorithmes

Perspectives

- Sélectionner les indicateurs selon leur distribution
- Augmenter le jeu de données (avec les sociologues)
- Calculer la position géographique
- Intégrer de nouvelles sources de données (e.g. points d'intérêt)

Merci de votre attention !