ISEN
ALL IS DIGITAL!
yncréa

Intelligence Artificielle

Exploration et Analyse du Patrimoine Arboré de la ville de Saint-Quentin

Meliha ACAR, Clarance CLEUZIOU, Agathe DORMOY

DIAGRAMME DE GANTT :

Meliha :
Clarance :
Agathe :

		Lundi		Mardi		Mercredi		Jeudi	
		matin	Apres-midi	Matin	Après-midi	Matin	Après-midi	Matin	Après-midi
Besoin 1	Préparation des données								
	Apprentissage non supervisé								
	Métriques								
	Visualisation								
Besoin 2	Préparation des données								
	Apprentissage non supervisé								
	Métriques								
	Visualisation								
Besoin 3	Préparation des données								
	Apprentissage non supervisé								
	Métriques								
	Visualisation								

Dans le cadre du projet, l'objectif est de concevoir une application pour étudier le patrimoine arboré de Saint-Quentin (Aisne) de manière approfondie. Cette semaine, l'attention se porte sur l'Intelligence Artificielle (IA) avec plusieurs objectifs spécifiques. Tout d'abord, l'accent est mis sur la compréhension détaillée des différentes étapes d'un projet d'apprentissage automatique, de la préparation des données à l'évaluation des modèles, pour assurer le succès de l'application. Ensuite, il est question de maîtriser l'évaluation appropriée des performances des méthodes d'apprentissage automatique, en choisissant les métriques pertinentes et en interprétant correctement les résultats obtenus.Diverses techniques d'apprentissage automatique sont explorées, notamment le clustering pour l'apprentissage non supervisé et la classification/régression pour l'apprentissage supervisé, dans le but d'analyser et de prédire des caractéristiques des arbres.Dans ce contexte, les données sur les arbres sont collectées et modélisées minutieusement, en mettant l'accent sur la prédiction de l'âge des arbres et la création d'une visualisation cartographique interactive pour classifier les arbres selon leur taille.De plus, un modèle d'intelligence artificielle est développé pour prédire l'âge des arbres en utilisant diverses caractéristiques, et un système d'alerte est mis en place pour anticiper les arbres susceptibles d'être déracinés lors de tempêtes, dans le but d'aider à la planification urbaine et à la gestion de l'environnement.

I. <u>Besoin client 1 : Visualisation sur carte</u>

Dans cette section initiale du rapport, l'accent est mis sur la création d'une visualisation cartographique permettant de segmenter les arbres en fonction de leur taille. Cette fonctionnalité permet à l'utilisateur de choisir entre différents nombres de catégories (par exemple, deux catégories pour les petits et grands arbres, ou trois catégories pour les petits, moyens et grands arbres), offrant ainsi une exploration personnalisée et détaillée de la distribution des arbres dans la ville.

a) Préparation des Données :

Dans la phase initiale de préparation des données, deux étapes cruciales ont été entreprises:

La sélection des colonnes pertinentes de la base de données a été effectuée en identifiant les caractéristiques clés telles que la localisation, la hauteur et le diamètre du tronc, afin d'optimiser l'ensemble de données pour les analyses ultérieures.

Dans le cadre du développement d'un modèle de prédiction de l'âge des arbres, les données ont été minutieusement préparées. Cela inclut l'importation des données à partir d'un fichier CSV, la sélection des colonnes pertinentes, la séparation en ensembles de caractéristiques (X) et cible (y), ainsi que la normalisation des caractéristiques pour optimiser les performances du modèle.

Enfin, les données ont été divisées en ensembles d'entraînement et de tests pour évaluer la capacité du modèle à généraliser sur de nouvelles données.

b) Apprentissage Non Supervisé

L'apprentissage non supervisé est une branche de l'apprentissage automatique où les données ne sont pas étiquetées ou catégorisées au préalable.

Dans le cadre de notre analyse, nous avons exploré plusieurs algorithmes de clustering pour segmenter les données sur le patrimoine arboré. Le clustering est une technique d'apprentissage non supervisé qui vise à regrouper des données similaires en clusters ou en groupes.Parmi les modèles utilisés, nous avons inclus le K-Means, agglomerative clustering, la méthode de Ward, le bisecting K-Means, la méthode GMM, le HDBSCAN et DBSCAN.

Le K-Means, un algorithme de clustering classique, divise les données en K clusters en cherchant à minimiser la variance intra-cluster. Agglomerative Clustering, quant à lui, adopte une approche hiérarchique en fusionnant progressivement les clusters voisins pour former un dendrogramme en fonction de leur proximité. La méthode de Ward, faisant partie du regroupement hiérarchique, vise à réduire la variance intra-cluster lors de la fusion des clusters. Le Bisecting K-Means, une variante du K-Means, subdivise les données de manière récursive en deux clusters jusqu'à ce qu'un nombre prédéfini de clusters soit atteint. Le HDBSCAN construit une hiérarchie de clusters en analysant les densités locales des points de données. Le GMM suppose que les données sont générées par un mélange de plusieurs distributions normales (gaussiennes) avec des paramètres inconnus. Après avoir évalué ces méthodes, nous avons constaté que certaines étaient similaires ou ne répondaient pas aux besoins du client. Nous avons donc choisi de conserver uniquement l'agglomerative clustering, le bisecting K-Means et le K-means pour la suite de notre analyse.

Chacun de ces modèles a ses propres avantages et limitations, et le choix dépend des caractéristiques spécifiques de nos données et de nos objectifs d'analyse. Nous avons évalué les performances de chaque algorithme en termes de cohérence intra-cluster, de séparation inter-cluster et de capacité à gérer les données de grande dimension pour déterminer le modèle le plus approprié pour notre besoin.

Dans le contexte de la visualisation sur carte, le clustering revêt une importance particulière. En segmentant les données en clusters, nous pouvons identifier des groupes d'éléments qui partagent des caractéristiques similaires, soit la taille dans cette première étude. Ces clusters sont représentés sur les cartes sous forme de zones colorées, ce qui permet une visualisation intuitive de la répartition spatiale des données. Ainsi, nous pouvons rapidement identifier les zones où certains types d'arbres sont plus abondants, les zones de biodiversité élevée, ou même les zones présentant des anomalies ou des concentrations particulières d'arbres.

c) Métriques pour l'Apprentissage Non Supervisé

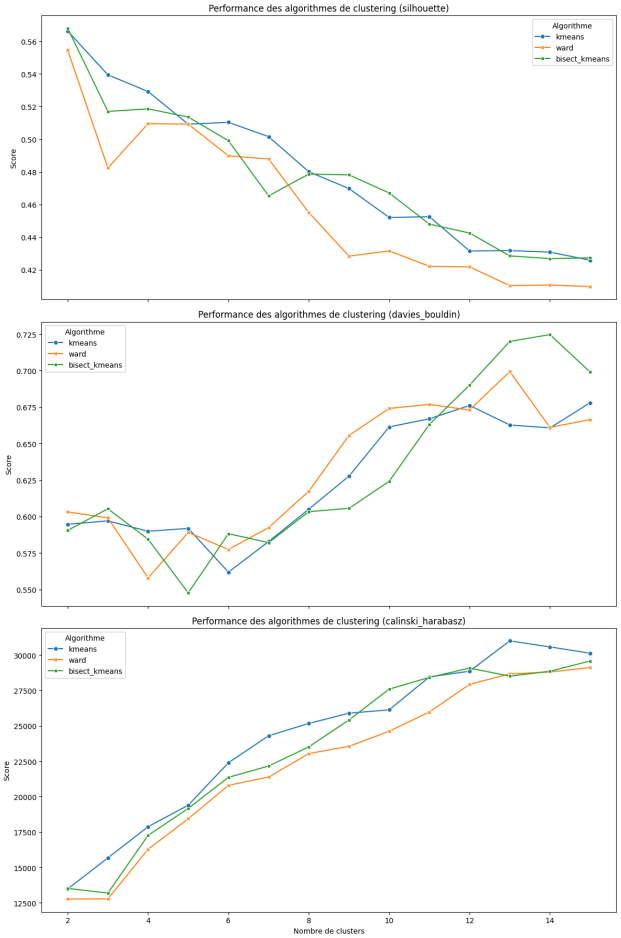
Dans le domaine de l'apprentissage non supervisé, évaluer la qualité des clusters est crucial. Pour cela, plusieurs métriques ont été développées, notamment le silhouette score, l'indice de Davies-Bouldin et l'indice de Calinski-Harabasz.

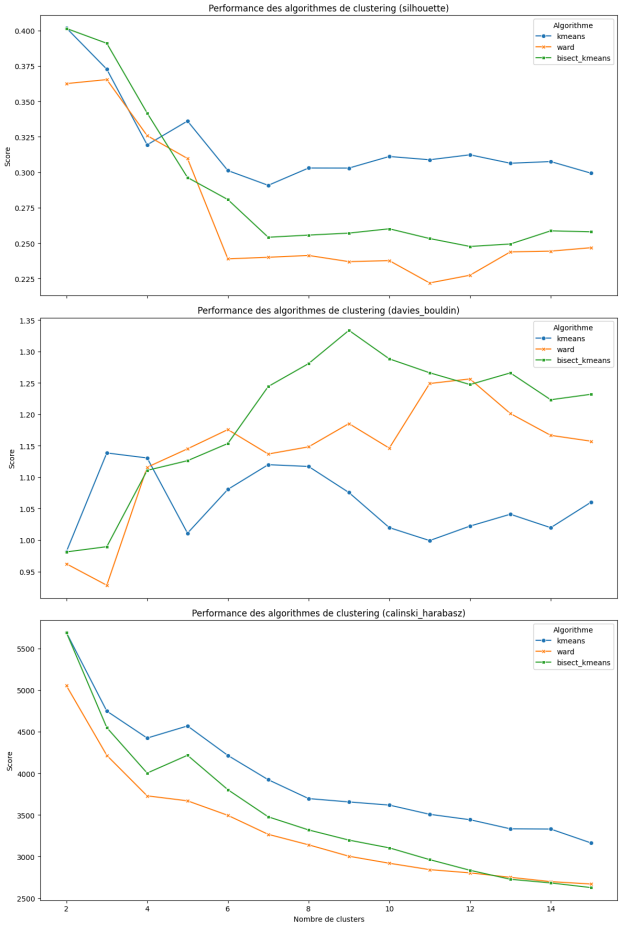
Le silhouette score mesure la cohérence d'un objet avec son propre cluster par rapport aux autres clusters. Un score proche de 1 indique une bonne classification de l'objet dans son propre cluster, ce qui traduit une segmentation précise des données.

L'indice de Davies-Bouldin évalue la séparation entre les clusters en calculant la similarité moyenne entre chaque cluster et son cluster le plus proche. Un indice faible indique une bonne séparation entre les clusters, correspondant à une segmentation claire et bien définie des données.

L'indice de Calinski-Harabasz est une autre métrique évaluant la qualité des clusters. Un score élevé indique à la fois une forte cohésion intra-cluster et une bonne séparation inter-cluster, assurant ainsi une segmentation efficace des données. Utilisé en combinaison avec d'autres métriques, il offre une évaluation complète des performances des algorithmes de clustering.

Ces métriques permettent une évaluation objective de la qualité des clusters générés. En calculant le silhouette score et l'indice de Davies-Bouldin, nous comparons les performances des algorithmes de clustering pour choisir celui produisant les clusters les plus cohérents et les mieux séparés.





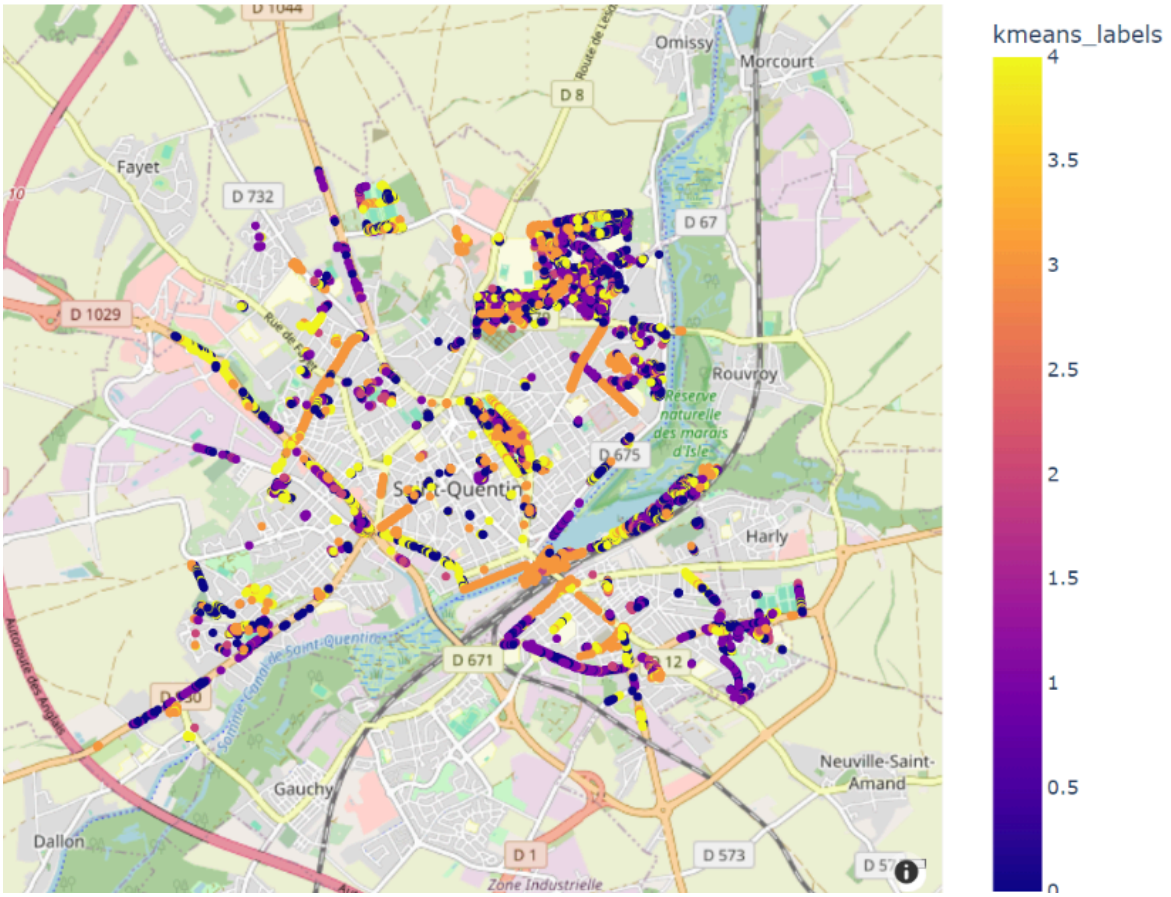
Normalisation avec haut_tot, tronc_diam, haut_tronc

sans Normalisation avec haut_tot, tronc_diam, haut_tronc

C’est graphique montre que la sans normalisation les algorithmes sont bien plus efficaces, peu importe les métriques et les données. Il est également observable que la méthode des K-means semble légèrement plus efficace que les autres, c’est donc que nous choisirons pour ce besoin. Nous n’avons pas pu présenter les autres graphiques ici, mais nous avons également pu rassurer notre choix de variables en essayant différentes possibilités comme par exemple avec seulement la hauteur totale ou la hauteur du tronc ainsi que le diamètre du tronc.

d) Visualisation sur Carte

Voici la carte des arbres de Saint-Quentin réparti selon leur taille en 5 clusters différents.



e) Préparation du Script

Après une analyse approfondie, la décision a été prise de sélectionner k-means sans normalisation comme modèle final pour la tâche de classification. Cette méthode se distingue par sa simplicité, son efficacité et son adaptation à notre jeu de données. En partitionnant les données en clusters distincts, k-means facilite grandement l'analyse et l'interprétation des résultats obtenus.

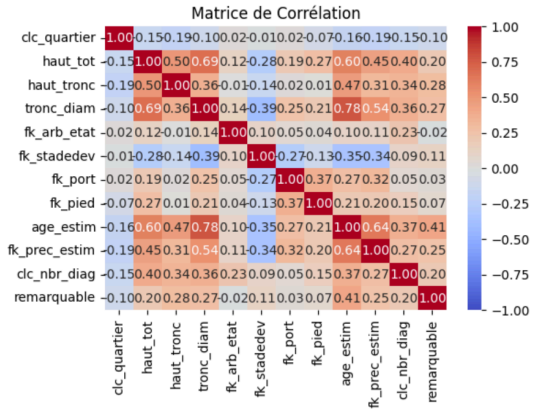
II. Besoin client 2 : Modèle de prédiction de l'âge

Dans cette deuxième phase du projet, l'accent a été mis sur le développement d'un modèle d'intelligence artificielle capable de prédire l'âge des arbres. Cette tâche est cruciale car elle permet d'estimer l'âge des arbres dans différentes parties de la ville de Saint-Quentin, fournissant ainsi des informations précieuses pour la planification urbaine et la gestion de l'environnement.

Pour développer ce modèle, des techniques d'apprentissage automatique supervisé sont utilisées. Cela implique l'utilisation de données d'entrée associées à des étiquettes : les caractéristiques des arbres (telles que la localisation géographique, la hauteur, le diamètre du tronc) constituent les entrées, tandis que l'âge estimé de l'arbre est l'étiquette cible.

a) Préparation des Données

Il était important de commencer cette partie par la réalisation d'une matrice de corrélation, permettant de visualiser la corrélation entre l'âge estimé de l'arbre et les autres variables du tableau. Cette matrice aide à identifier les variables les plus influentes sur l'âge des arbres, facilitant ainsi la sélection des caractéristiques pertinentes pour les modèles de prédiction. Après visualisation, il a été estimé que les variables les plus corrélées avec l'âge étaient : 'haut_tot', 'haut_tronc', 'tronc_diam', 'age_estim', 'remarquable', 'clc_nbr_diag', 'fk_stadedev', 'fk_prec_estim', 'fk_pied', et 'fk_port'.



b) Apprentissage Supervisé pour la classification & Métriques pour la classification

Des algorithmes de classification ont ensuite été mis en œuvre, sélectionnant cinq modèles : le KNN classifier, le support vector machine, le random forest classifier, le gradient boosting, et le neural network. Ces algorithmes ont été choisis pour leur capacité à gérer des données de nature variée et leur popularité dans des tâches similaires.

Après l'entraînement des modèles et l'évaluation des performances, les métriques obtenues incluaient l'accuracy, le F1 score, la précision et le rappel. Cependant, les résultats montraient une accuracy faible allant de 0,3 à 0,6, indiquant une faible performance des modèles. En réponse, l'attention s'est portée sur des algorithmes de régression. Les

modèles choisis incluaient une régression linéaire, un random forest regressor et un KNN regressor. Les résultats obtenus montraient des R^2 allant de 0,7 à 0,8, indiquant de meilleures performances que la classification.

Des grid searches ont été effectués pour optimiser les hyper paramètres critiques. Sur le modèle random forest regressor, les hyperparmatres testés étaient : 'n_estimators': [50, 100, 150, 200],'max_depth': [10, 20, None], 'min_samples_split': [2, 5, 10]. Après un nouvel entraînement des modèles et une normalisation inverse des y, un R^2 de 0,86 avec une RMSE de 0,03, ce qui en faisait un modèle très performant pour la bonne réalisation du besoin du client.

c) Préparation du Script

Pour estimer l'âge d'un arbre, un fichier JSON contenant des informations basées sur les paramètres retenus est utilisé :{


```
    "haut_tot": 10.5,
    "haut_tronc": 5.2,
    "tronc_diam": 3.8,
    "remarquable": 1,
    "clc_nbr_diag": 2,
    "fk_stadedev": 1,
    "fk_prec_estim": 0,
    "fk_pied": 3,
    "fk_port": 2
}
```


Parallèlement, l'entraînement de l'algorithme Random Forest Regressor a été réalisé en optimisant les hyperparamètres pour obtenir les meilleures performances. Ce modèle, étant le plus performant, a été sauvegardé et converti au format .pkl. En combinant le fichier JSON et le fichier .pkl, un script a été développé pour estimer l'âge d'un arbre. Concrètement, il charge un modèle de forêt aléatoire et un scaler, lit les données JSON, encode les colonnes catégorielles, et convertit les types de données si nécessaire. Ensuite, il fait des prédictions en utilisant le modèle, inverse la normalisation des prédictions, et ajoute les âges prédits au DataFrame. Finalement, le DataFrame est converti en JSON pour être retourné. Avec les données du JSON au-dessus, l'âge estimé de l'arbre est de 15,2 ans avec un RMSE de 0,03.

III. Besoin client 3 : Système d'alerte pour les tempêtes

L'objectif de cette partie est de développer un système d'alerte capable de prédire la position des arbres susceptibles d'être déracinés en cas de tempête. Ce système permettrait de prendre des mesures préventives pour minimiser les dégâts et assurer la sécurité publique. Pour atteindre cet objectif, nous avons exploré différentes méthodes de classification supervisée.

a) Préparation des Données

Pour prédire la position des arbres susceptibles d'être déracinés en cas de tempête, une préparation minutieuse des données est essentielle. Une matrice de corrélation a d'abord été générée pour visualiser les relations entre l'état de l'arbre et les autres variables du tableau. Cette analyse a permis d'identifier les variables les plus influentes sur l'âge des arbres, facilitant ainsi la sélection des caractéristiques pertinentes pour les modèles de prédiction. Après examen, il a été observé que les variables les plus corrélées avec l'état de l'arbre étaient : 'haut_tot', 'tronc_diam', 'fk_stadedev', 'age_estim', 'fk_prec_estim' ,'clc_prec_diag'.

Après la sélection des variables, les données sont divisées en deux ensembles : les variables indépendantes (X) et la variable dépendante (y).

Par la suite, les données sont divisées en ensembles d'entraînement et de test pour évaluer les performances des modèles. Cela permet de former les modèles sur un sous-ensemble des données et de tester leur précision sur un autre sous-ensemble non vu pendant l'entraînement.

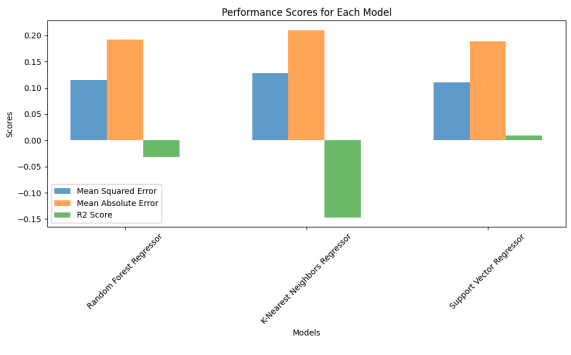
b) Apprentissage Supervisé pour la classification et métriques pour la classification

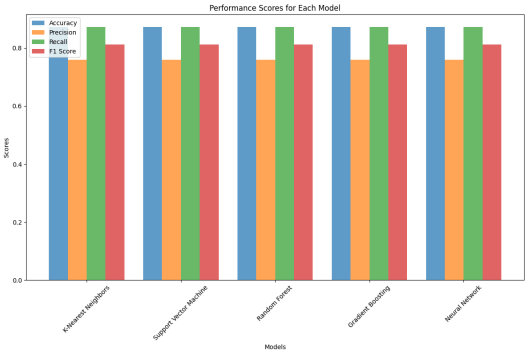
L'étude a débuté vers une approche de classification pour prédire quels arbres devaient être abattus. Les modèles de classification, incluant K-Nearest Neighbors, Support Vector Machine, Random Forest, Gradient Boosting, et Neural Network, ont tous affiché des performances similaires avec une accuracy, une précision, un rappel, et un F1 Score tous égaux à 0.8718. Ces résultats indiquent que les modèles étaient capables de classer de manière efficace les arbres à abattre par rapport aux autres.

En termes de sélection d'algorithme, bien que tous les modèles de classification aient démontré des performances élevées, il a été choisi d'utiliser une méthode de régression afin d'avoir des pourcentages dans le cas où nous aurions eu le temps de réaliser le bonus.

Différents modèles de régression ont été testés pour prédire les arbres à abattre. Les résultats obtenus ont révélé que le Random Forest Regressor a produit un Mean Squared Error (MSE) de 0.1153, un Mean Absolute Error (MAE) de 0.1926, et un R2 Score de -0.0317. En comparaison, le K-Nearest Neighbors Regressor a montré un MSE de 0.1282, un MAE de 0.2103, et un R2 Score de -0.1471, tandis que le Support Vector Regressor a présenté un MSE de 0.1108, un MAE de 0.1885, et un R2 Score de 0.0091. L'utilisation du Gridsearch pour améliorer les hyperparamètres du modèle n'a pas été concluante cette fois-ci avec des résultats proches de ceux obtenus sans.

Les métriques fournissent une évaluation globale des performances des modèles, en tenant compte de la proportion de prédictions correctes, de la capacité à éviter les faux positifs et à détecter les vrais positifs. Deux graphiques ont été générés pour visualiser le modèle adapté à l'étude du client 3. La visualisation confirme que le modèle adapté est bel et bien le Random Forest.

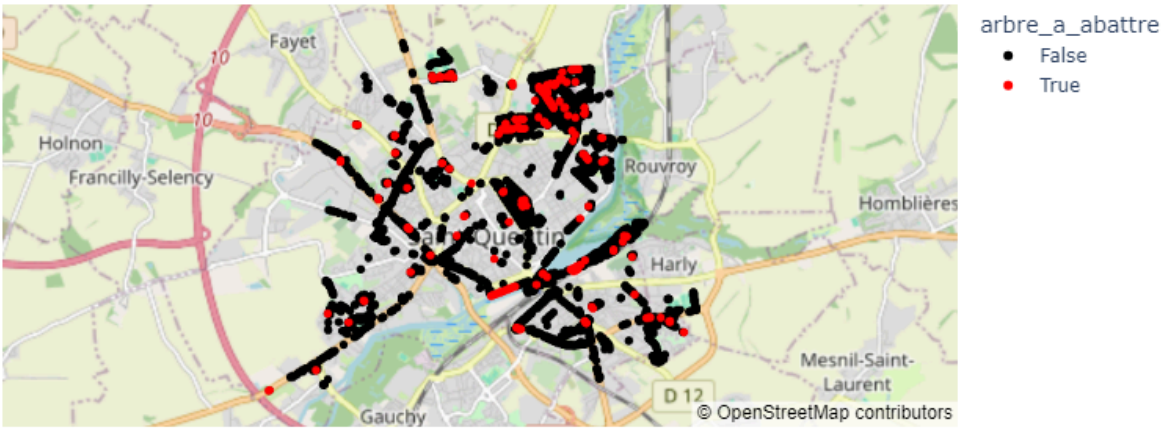




Après avoir entraîné le modèle sur 200 arbres, il a été appliqué sur les 7000 arbres du dataframe initial, fournissant des prédictions. Ces prédictions permettent de générer une carte indiquant les arbres susceptibles d'être abattus en fonction des critères définis par le modèle optimisé.

Par question de facilité de codage, cette carte a été réalisée grâce aux données d'entraînement de l'algorithme de random forest regressor. Si le y prédit était supérieur à 0,5 l'arbre était considéré comme "à abattre" , si celui-ci avait une une valeur inférieure à 0,5, alors il n'avait pas de risque de chuter en cas de tempête. L'utilisation d'un algorithme de classification aurait également été possible ici, avec 1 pour les arbres à abattre et 0 pour les arbres à ne pas abattre.

Carte des arbres à abattre pour risque de chute en cas de tempête



Conclusion :

Le projet a répondu à trois besoins spécifiques en matière de gestion des arbres urbains. Pour le premier besoin, une visualisation sur carte a été créée pour catégoriser les arbres selon leur taille. Après la préparation des données, un apprentissage non supervisé a segmenté les arbres en différentes catégories de taille. Les résultats ont été visualisés sur une carte interactive.

Pour le deuxième besoin, un modèle de prédiction de l'âge des arbres a été développé. Les données ont été préparées et normalisées, puis utilisées pour entraîner plusieurs modèles de machine learning. Le Random Forest Regressor a fourni les meilleures prédictions de l'âge des arbres, évaluées à l'aide de métriques comme l'erreur quadratique moyenne

(RMSE) et le coefficient de détermination (R^2). Un script a été créé pour permettre l'utilisation du modèle en ligne de commande à partir d'un fichier JSON.

Pour le troisième besoin, un système d'alerte pour les tempêtes a été mis en place pour prédire quels arbres pourraient être déracinés. Les données ont été préparées pour inclure les états des arbres, et un modèle supervisé a été utilisé pour classifier ces états. Le Random Forest Regressor a permis de prédire la fragilité des arbres, avec les prédictions visualisées sur une carte. Un script a été développé pour automatiser ce processus

Ce projet a permis d'exploiter divers outils Python utilisés en intelligence artificielle, incluant des modèles de classification et de régression ainsi que des métriques adaptées pour chaque modèle. L'importance de choisir un modèle adéquat pour chaque tâche a été mise en évidence, ainsi que l'optimisation de l'entraînement grâce au Grid Search, permettant de trouver les meilleurs hyperparamètres.

Ce projet a démontré l'efficacité des techniques de machine learning dans la gestion des arbres urbains, fournissant des outils précieux pour la visualisation, la prédiction de l'âge et la prévention des risques liés aux tempêtes. Les scripts développés permettent une utilisation facile et répétée des modèles, rendant les solutions proposées pratiques et accessibles.