

La fouille de données au service du développement durable

SUZANNE CRESSANT, ARSLANE MEDJAHDI, ALOYSE
PHULPIN

Projet de Fouille de données et extraction de connaissances
UNIVERSITÉ DE LORRAINE
Faculté des Sciences et Technologies de Vandoeuvre-lès-Nancy
Année universitaire 2022-2023

Table des matières

1	Introduction	2
1.1	Présentation des données :	3
2	Défi 1 :	3
2.1	Visualisations préalables :	3
2.2	Gestion des valeurs manquantes, variables qualitatives.	4
2.3	Gestion des valeurs manquantes, variables quantitatives.	5
2.4	Classification uni-label :	6
2.4.1	Équilibrage et poids des variables :	6
2.4.2	Regression logisitique :	6
2.4.3	Classification	7
2.4.4	Réseau de Neurones :	9
2.4.5	Conclusion :	11
2.5	Classification multi-label :	11
2.5.1	Équilibrage et poids des variables :	11
2.5.2	Plusieurs classifications uni-label :	11
2.5.3	Conclusion :	13
2.6	Conclusion défi 1 :	14
3	Défi 2 :	15
3.1	Présentation des Secteurs de Grenoble :	15
3.2	Secteur par secteur :	16
3.3	Préconisation :	18
4	Source externes :	18

1 Introduction

Le but de ce projet de FDEC est de prédire le défaut d'un arbre du parc végétal de Grenoble à partir de différentes informations préalablement disposées dans un tableau excel. Ce projet s'organise autour de deux défis : la prédiction de défaut des arbres et la présentation de l'état global du parc végétal afin de fournir des préconisations pour faciliter son entretien.

1.1 Présentation des données :

Les données se présentent sous forme de tableaux de données : un premier dataframe pour entraîner les modèles, et un second pour les évaluer. Chacun des individus de nos données représente un arbre du parc végétal de Grenoble. De plus, notre dataframe possède 34 variables : 22 variables catégoriques, 7 variables numériques et 5 variables cibles.

2 Défi 1 :

Le premier défi se divise en deux tâches de prédiction : l'une est uni-label et prédit le défaut d'un arbre, l'autre est une prédiction multi-label de la localisation du ou des défaut(s) de l'arbre (la racine, le houppier, le tronc et le collet).

2.1 Visualisations préalables :

On décide tout d'abord de constater la quantité de valeurs manquantes dans la base de données. On distingue sur la figure 1 trois types de profils pour nos variables :

- Aucune valeur manquante ;
- Au maximum 30% de valeurs manquantes ;
- Plus de 70% de valeurs manquantes.

On décide d'adopter deux comportements différents. Dans un premier temps, nous supprimons les variables ayant plus de 70% de valeurs manquantes ; dans un second, nous imputerons les autres.

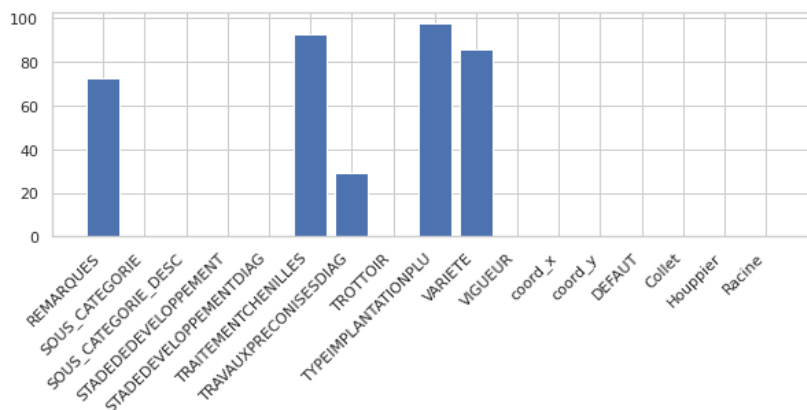


FIGURE 1 – Proportion de valeurs manquantes

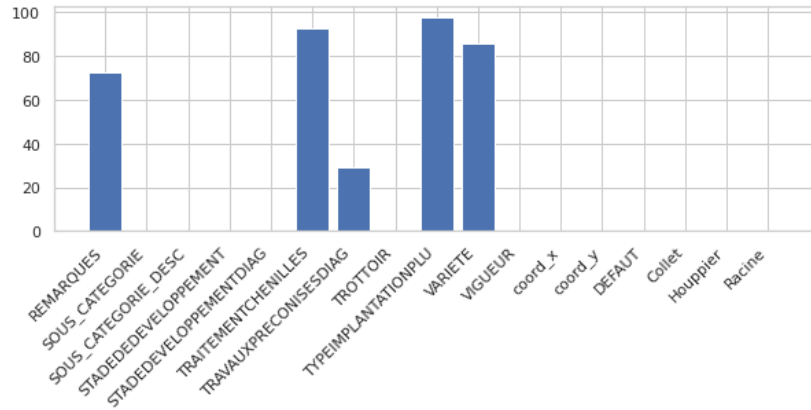


FIGURE 2 – Proportion de valeurs manquantes

2.2 Gestion des valeurs manquantes, variables qualitatives.

On décide d'imputer les variables 'objet' par la modalité la plus fréquente de l'attribut. Ainsi, on affiche ci dessous la répartition des variables converties avec et sans imputation. En fait, la seule imputation visible et introduisant un second biais est 'TRAVAUXPRECONISEDIAG'.

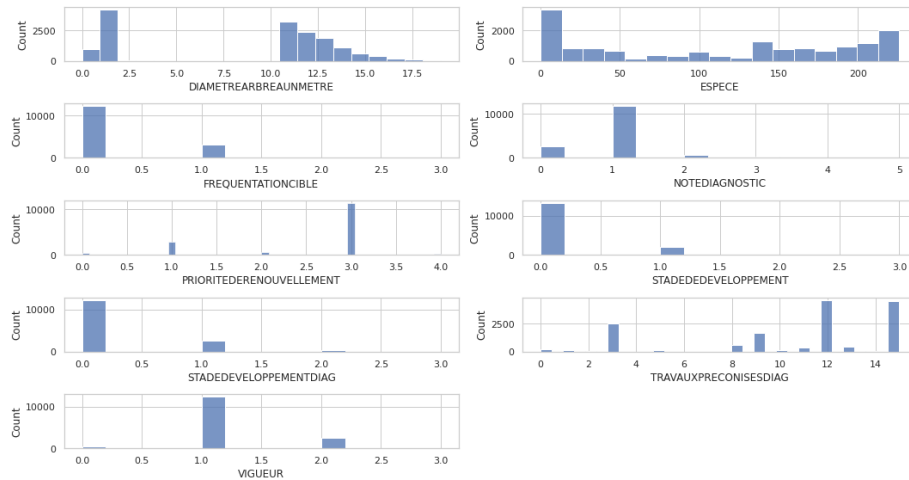


FIGURE 3 – Répartition des données catégoriques avant imputation

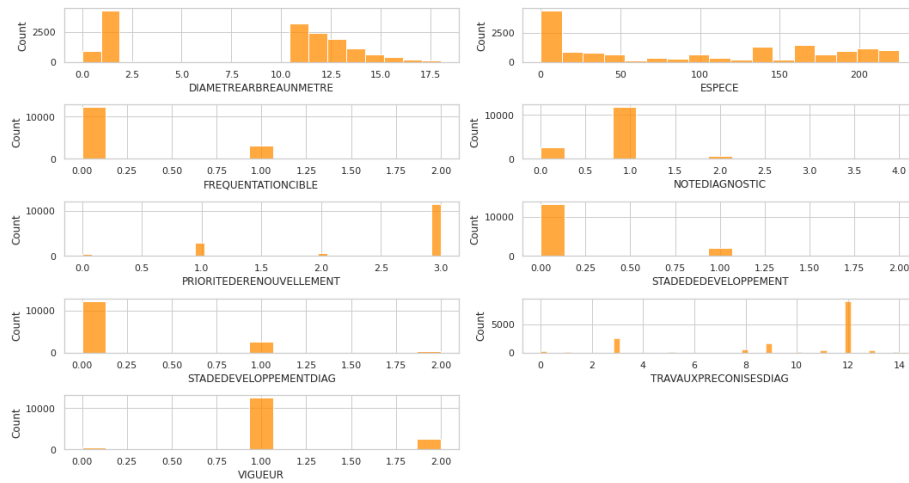


FIGURE 4 – Répartition des données catégoriques après imputation

2.3 Gestion des valeurs manquantes, variables quantitatives.

On dispose de deux variables numériques comportant des valeurs manquantes. Pour chacune d'elles, on se demande quelle est la stratégie d'imputation adaptée. Pour cela, nous en comparerons deux : par la moyenne et par la médiane.

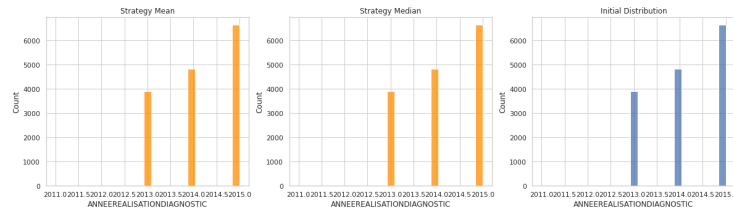


FIGURE 5 –

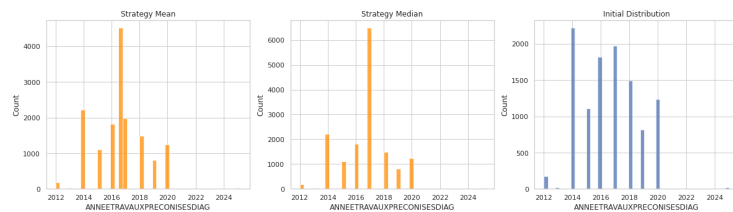


FIGURE 6 –

Nous faisons le choix de l'imputation par la médiane. Tandis qu'il n'existe pas de biais significatif pour la première variable, pour laquelle nous remplaçons peu d'observations, nous en observons un pour 'ANNEETRAVEAUXPRECONISEDIAG', où près de 30% des observations étaient manquantes. Nous en tiendrons compte dans l'analyse des classificateurs.

2.4 Classification uni-label :

2.4.1 Équilibrage et poids des variables :

Comme nous pouvons le constater sur la figure 6, la distribution des classes est inégale. C'est pourquoi nous créons un second tableau de données équilibrées en dupliquant les observations de modalité 1 pour l'attribut 'DEFAULT'. Par ailleurs, on normalise les deux jeux de données. Nous séparons également

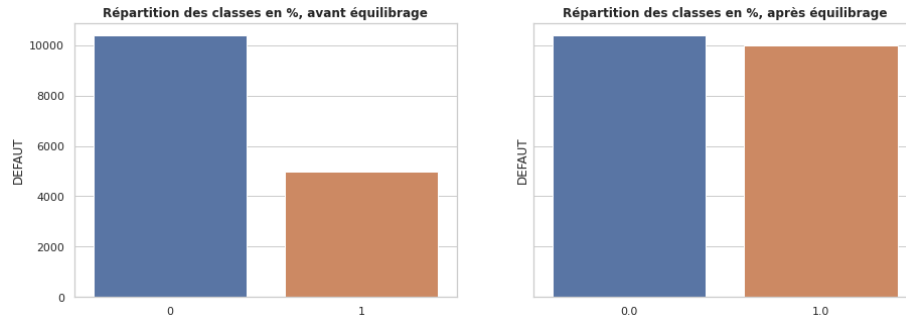


FIGURE 7 –

les jeux de données en ensembles d'apprentissage et de validation.

2.4.2 Regression logisitique :

Notons le modèle 0 la régression logistique. Sur la figure 7, nous pouvons observer, pour la distribution avec et sans équilibrage des classes, les résultats de précision associées à la classe 1. L'exactitude désigne la précision générale du modèle.

Equilibrage	Précision	Rappel	Excactitude	AUC
Sans	0.48	0.72	0.77	0.7999
Avec	0.71	0.74	0.73	0.8012

FIGURE 8 –

2.4.3 Classification

Afin de déterminer le classificateur adapté pour prédire la variable 'DE-FAUT', entraînons différents modèles de classification. Nous testerons ensuite ces derniers afin de les comparer.

Numéro	Modèle	Base	Paramètres
1	Arbre de décision		Profondeur 12
2	Forêt aléatoire	Arbre de décision	Profondeur 16
3	Adaboost	Arbre de décision	profondeur 16
4	Gradient Boosting	Arbre de décision	profondeur 10
5	XGB classifier	Arbre de décision	$\eta = 0.1, \gamma = 0.20$ profondeur max= 10
6	Hard Voting	Modèle 0 à 5	
7	Soft Voting	Modèle 0 à 5	

FIGURE 9 –

Parmi les modèles, certains n'ont pas été exploités en cours. Nous en donnons une description ci-dessous.

- Modèle 5 : Cette méthode séquentielle repose sur la méthode de Boosting du gradient. Le principe est de combiner les résultats issus d'arbres de type CART afin de fournir de meilleures prédictions.
- Modèle 6 : Cette méthode choisit la moyenne des prédictions des différents classificateurs comme la classe sortante
- Modèle 7 : Cette méthode choisit la classe majoritaire parmi les prédictions des différents classificateurs comme la classe sortante.

Avec une distribution inéquitable des classes, nous obtenons les résultats suivants pour la classe 1. Ici, l'exactitude désigne la précision globale.

Modèle	Précision	Rappel	Exactitude	AUC
1	0.71	0.83	0.85	0.8467
2	0.75	0.86	0.88	0.9230
3	0.73	0.80	0.85	0.8983
4	0.75	0.83	0.86	0.9155
5	0.75	0.85	0.87	0.9106
6	0.75	0.86	0.87	
7	0.75	0.85	0.87	0.9144

FIGURE 10 – Sans 926333

Avec une distribution équitable des classes, nous obtenons les résultats suivants.


Modèle	Précision	Rappel	Exactitude	AUC
1	0.74	0.86	0.81	0.9100
2	0.77	0.91	0.84	0.9176
3	0.72	0.89	0.81	0.8986
4	0.76	0.89	0.83	0.9107
5	0.79	0.79	0.86	0.9231
6	0.74	0.92	0.84	
7	0.74	0.91	0.84	0.9147

FIGURE 11 – Avec équilibrage

Nous obtenons notamment les courbes ROC suivantes pour chacun des modèles ci-dessus. On remarque rapidement qu'elles sont semblables avec et sans équilibrage.

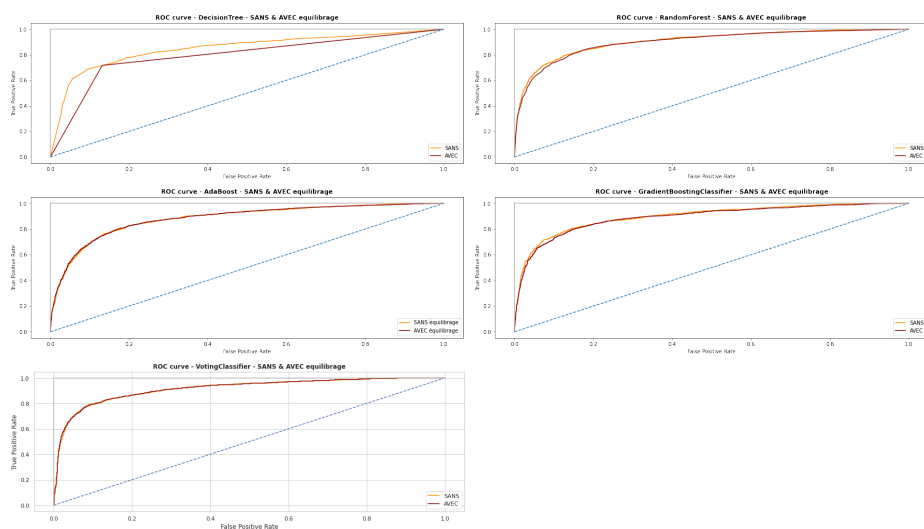
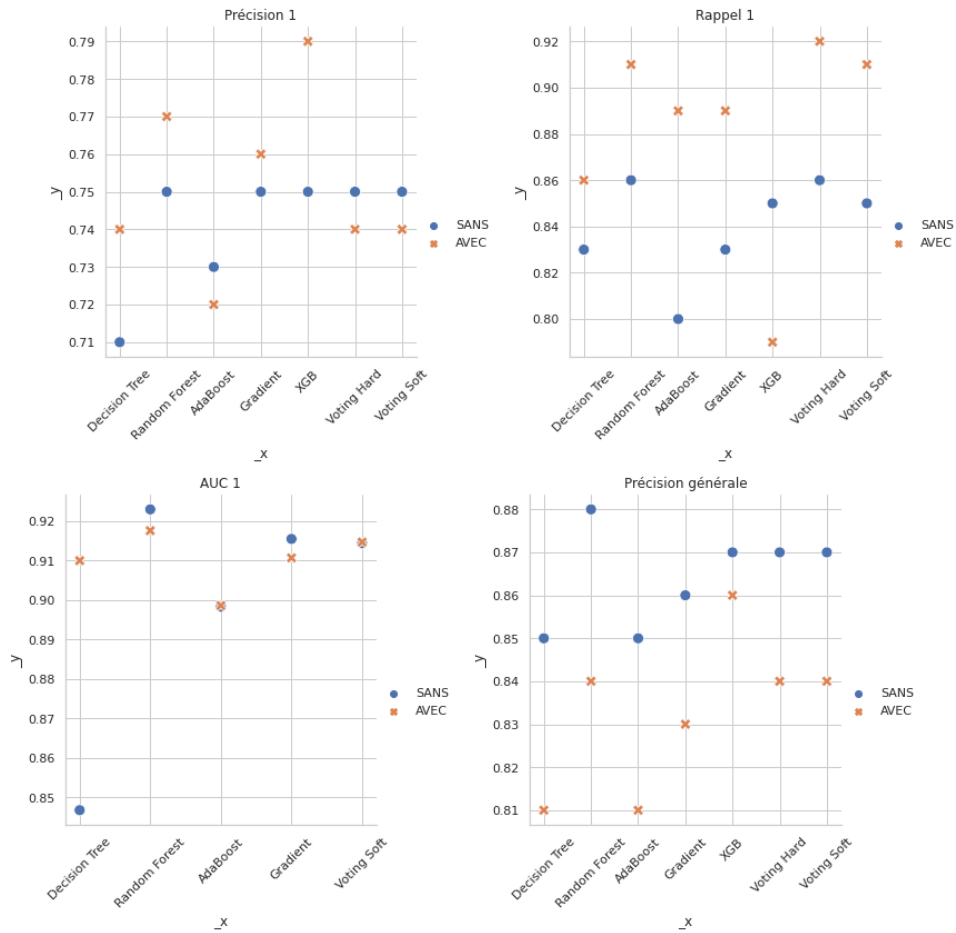


FIGURE 12 – Courbe ROC

Pour conclure, nous affichons une visualisation graphique de ces résultats interprétés.



2.4.4 Réseau de Neurones :

On se propose ici de réaliser un réseau de neurones contenant 2 couches pour réaliser la classification supervisée de notre défi 1.

On a donc :

- Une première couche de 16 neurones et de fonction d'activation sigmoid,
- Une seconde couche d'un seul neurone et de fonction d'activation sigmoid.

Après avoir entraîné notre modèle sur notre ensemble d'apprentissage, nous obtenons les résultats suivants :

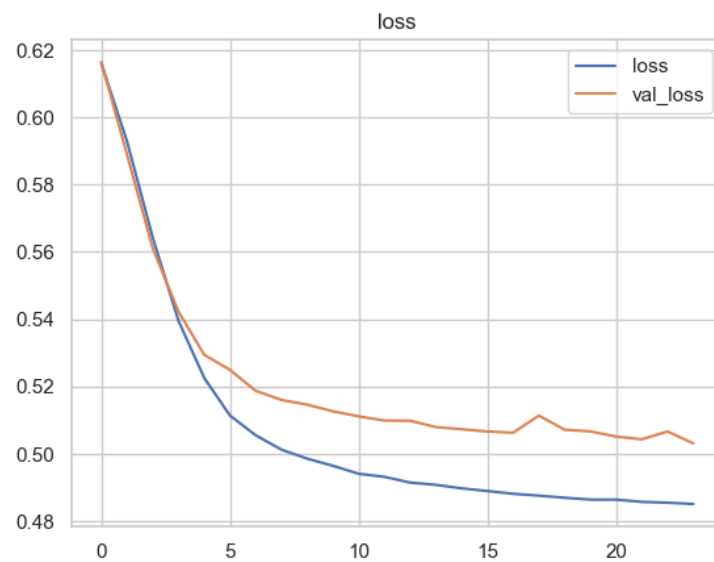


FIGURE 13 – Évolution de la fonction de perte

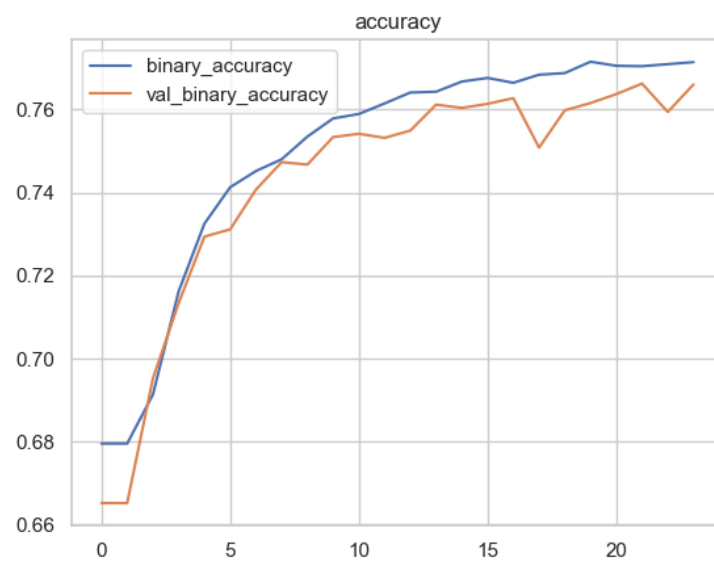


FIGURE 14 – Évolution de accuracy

Ces résultats restent cependant bien en dessous des autres modèles utilisés précédemment.

2.4.5 Conclusion :

Nous faisons le choix de la fiabilité et sélectionnons les modèles au plus haut taux d'AUC pour chacun des jeux de données (original et modifié) -à savoir les modèles 2 et 5, correspondant respectivement à la forêt aléatoire et XGB classifier. Afin d'analyser leurs performances, on effectue une k-folds cross validation, à cet effet on choisit k=11. Voici les résultats moyens obtenus :

Classificateur	Précision	AUC
XGB	0.8251	0.9085
RandomForest	0.9087	0.9696

2.5 Classification multi-label :

2.5.1 Équilibrage et poids des variables :

De la même manière que précédemment, on analyse les répartitions des classes affichées sur la figure 16. On y constate d'énormes écarts de distributions, notamment pour les variables 'Racine' et 'Collet'. Elles représentent chacune moins de 10% des défauts constatés. C'est pourquoi nous dupliquons pour les variables à prédire, hormis 'Tronc', les individus de classe 1 afin de réduire un peu cette disproportion mais sans complètement modifier la distribution des classes.

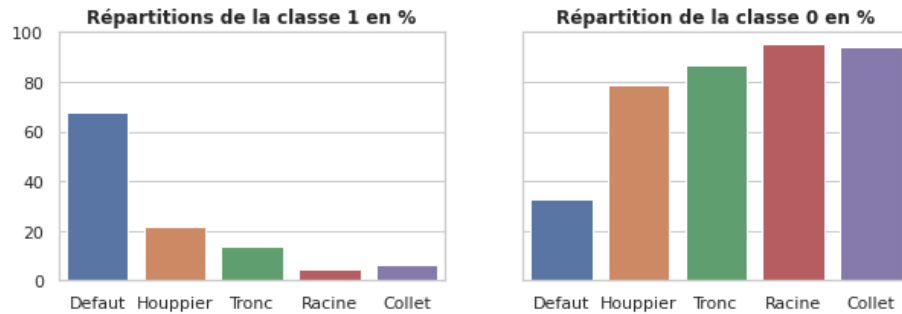


FIGURE 15 –

2.5.2 Plusieurs classifications uni-label :

Dans cette partie, on s'emploie à diviser la tâche de classification multi-label en quatre tâches de classification uni-label. On construit ainsi 4 classificateurs permettant de prédire les variables 'Houppier', 'Collet', 'Racine' et 'Tronc'. Pour cela, nous apprenons différents modèles, après avoir divisé les données en ensembles d'apprentissage et de validation. Nous ferons notamment le choix des modèles dont les résultats étaient significatifs dans la tâche précédente. Nous obtenons les résultats suivants pour la classe 1.


Modèle	Précision	Rappel	Exactitude	AUC
2	0.61	0.69	0.72	0.7837
3	0.57	0.65	0.69	0.7481
4	0.62	0.68	0.72	0.7813
5	0.59	0.67	0.71	0.7542
6	0.58	0.75	0.80	
7	0.61	0.66	0.71	0.7796

FIGURE 16 – Tronc


Modèle	Précision	Rappel	Exactitude	AUC
2	0.80	0.89	0.85	0.9238
3	0.81	0.88	0.85	0.9428
4	0.80	0.88	0.85	0.9324
5	0.81	0.89	0.85	0.9337
6	0.81	0.89	0.86	
7	0.83	0.89	0.87	0.9416

FIGURE 17 – Houppier


Modèle	Précision	Rappel	Exactitude	AUC
2	0.81	0.86	0.90	0.9261
3	0.78	0.84	0.89	0.9077
4	0.81	0.82	0.89	0.9127
5	0.82	0.86	0.9	0.9259
6	0.58	0.75	0.80	
7	0.81	0.86	0.90	0.9254

FIGURE 18 – Collet


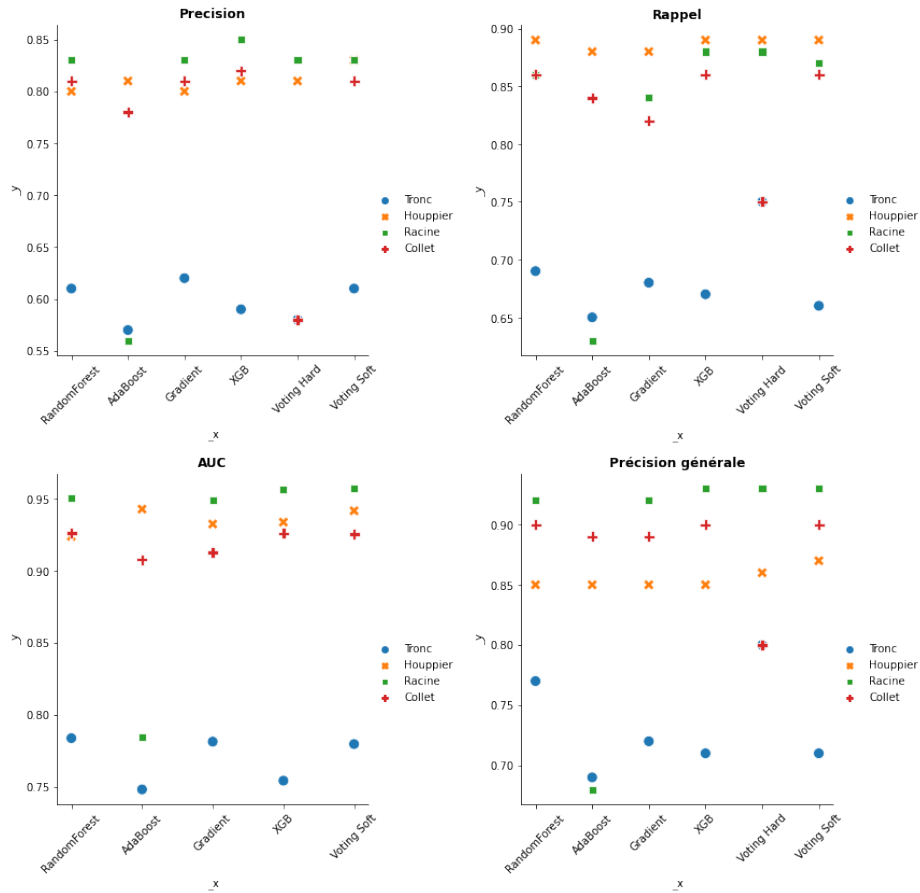
Modèle	Précision	Rappel	Exactitude	AUC
2	0.83	0.86	0.92	0.9504
3	0.56	0.63	0.68	0.7482
4	0.83	0.84	0.92	0.9488
5	0.85	0.88	0.93	0.9567
6	0.83	0.88	0.93	
7	0.83	0.87	0.93	0.9571

FIGURE 19 – Racine

Pour en simplifier l'interprétation, on visualise les différents résultats.



On remarque que quelque soit le classificateur utilisé, il existe une gradation dans les résultats de prédictions entre les 4 labels. En effet, chacun des résultats concernant la variable 'Tronc' est bien moindre que ceux des autres labels. Cela est surprenant sachant qu'il ne s'agissait pas de l'attribut dont la classe 1 était la moins bien représentée.

2.5.3 Conclusion :

Finalement, pour les 4 variables, nous choisissons le Soft VotingClassifier. Afin d'optimiser les performances, on effectue une k-folds cross validation, à cet effet on choisit k=11. Voici les résultats moyens obtenus :

Classificateur	Précision	AUC
Tronc	0.7180	0.7946
Collet	0.9661	0.9900
Houppier	0.9300	0.9860
Racine	0.9751	0.9951
Moyenne	0.8973	0.9414

2.6 Conclusion défi 1 :

Enfin, dans un esprit de critique de notre démarche, nous aurions pu, lors de la partie preprocessing, nous intéresser davantage à certaines variables avec des valeurs manquantes pour en extraire un maximum d'information. Typiquement, la variable 'Remarque' que nous avons d'office rejetée du fait du nombre important de valeurs manquantes, mais aussi du fait qu'elle soit intégralement constituée de texte, difficilement exploitable.

Par la suite, nous aurions pu choisir d'autres stratégies afin d'imputer nos valeurs manquantes : en effet remplacer par la médiane pour les variables numériques et par la modalité la plus fréquente pour les variables catégoriques conduit à avoir certaines classes sur-représentées et introduit donc un biais sur les données. Celui-ci ne se retrouve sûrement pas dans les données de test, ce qui peut conduire à de moins bons résultats concernant nos modèles testés sur ces dernières.

3 Défi 2 :

Le but de ce défi est de visualiser l'état général du Parc de Grenoble, de mieux en prendre connaissance et de fournir des préconisations pour faciliter son entretien.

3.1 Présentation des Secteurs de Grenoble :

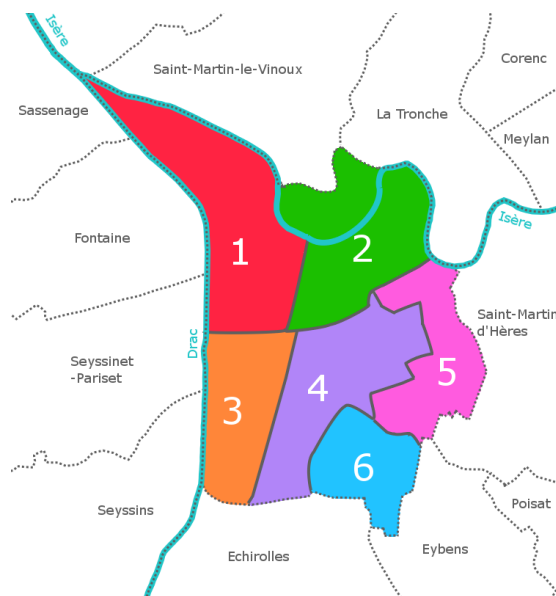


FIGURE 20 –

Avant de s'intéresser à la visualisation de nos données, on cherche à d'avantage connaître la géographie des secteurs de Grenoble. Le secteur 1 correspond au pôle scientifique de Grenoble. Dans le secteur 2 se situe notamment l'hyper centre de Grenoble, ainsi qu'une zone commerciale. Le secteur 4 contient le village olympique construit dans les années 60 lorsque Grenoble a accueilli les Jeux Olympiques d'hiver 1968. Les autres secteurs sont des quartiers urbains classiques.

On retrouve aussi sur le graphique ci-dessous le pourcentage de la population que compte chaque secteur (données de 2014).

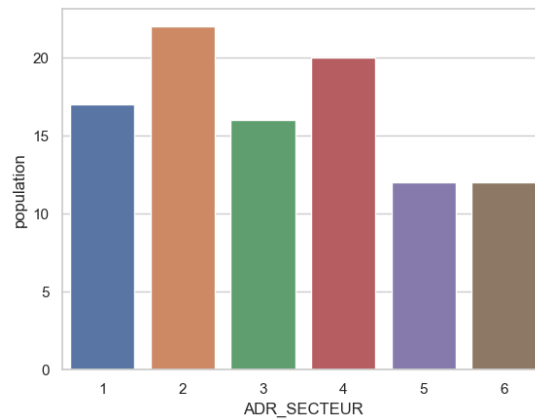


FIGURE 21 –

3.2 Secteur par secteur :

On présente ici le pourcentage des arbres par secteur, puis le pourcentage d'arbres ayant un défaut par secteur. Ensuite, le pourcentage d'arbres qui n'ont pas de défaut par secteur.

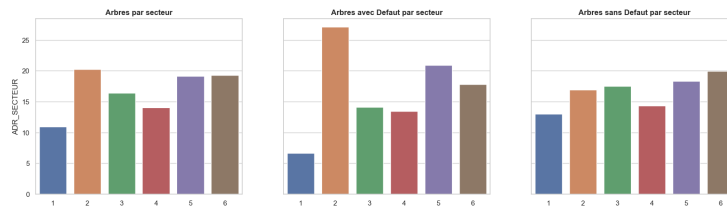


FIGURE 22 –

On se rend alors compte que le secteur 1 compte peu d'arbres (en pourcentage) ayant un défaut, comparé aux autres secteurs. Et inversement, le secteur 2 en compte le plus.

On s'intéresse maintenant au nombre de défauts de chaque arbre en fonction de son secteur.

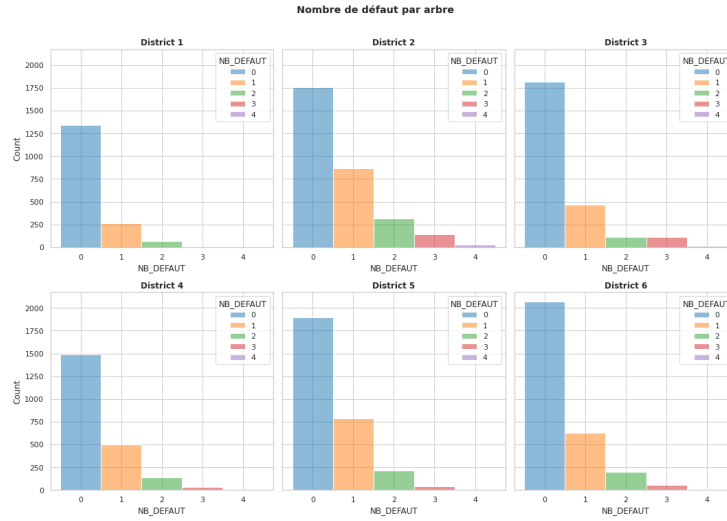


FIGURE 23 –

On remarque alors que c'est dans les secteurs 1 et 4 que l'on retrouve le moins d'arbres ayant 1 défaut ou plus. Et inversement c'est dans les secteurs 2 et 5 que l'on en retrouve le plus grand nombre .

Enfin une visualisation de la fréquentation à côté de l'arbre en fonction de son secteur nous apporte de nouvelles informations. Les arbres ayant la plus forte fréquentation autour d'eux se trouvent dans les secteurs 2 et 5. De plus, il y a peu de fréquentation du côté des arbres du secteur 1, ce qui explique donc en partie le peu d'arbres avec des défauts que ce dernier compte.

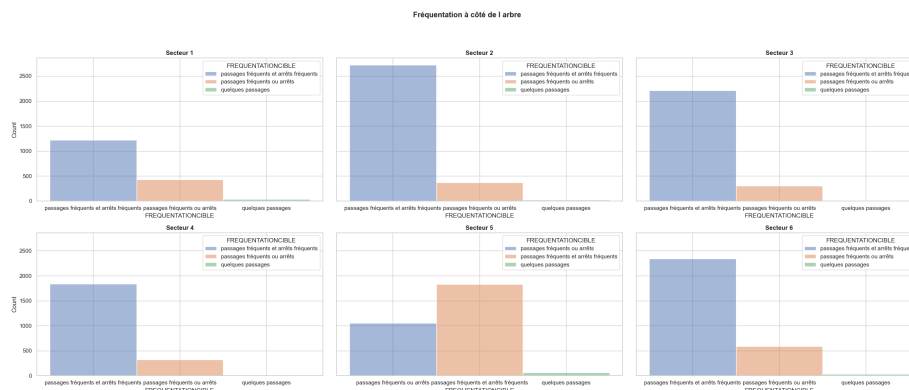


FIGURE 24 –

3.3 Préconisation :

De par nos différents graphiques et informations complémentaires, on remarque donc que plus les secteurs comptant le plus de passage sont aussi ceux qui présente le plus grand nombre d'arbres présentant au moins un défaut. On pourrait donc par exemple, dans le secteur 2 contenant l'hyper centre de Grenoble, interdire la circulation automobile à certaines rues afin de piétonniser une partie du secteur.

4 Source externes :

- <https://www.grenoble.fr/510-fiches-secteurs-de-la-ville-de-grenoble.htm>