

[Trimestre 2]

V- Prédiction du risque de crédit à base de descripteurs issus de la modélisation des données en graphes

La plupart des travaux sur le *credit scoring* considèrent uniquement les attributs de description des prêts (caractéristiques de l'emprunteur, somme à prêter, ...), et ne s'attardent pas particulièrement sur la modélisation explicite des relations entre les emprunteurs.

Ceci peut être une limite, car les individus aux caractéristiques communes peuvent avoir les mêmes comportements de prêts et donc à partir des comportements connus d'un ensemble d'individus similaires à un individu cible, on peut déduire le comportement de ce dernier.

V-1- Définition et positionnement dans le projet

V-1-1- Définition

Graphe :

Un graphe est une structure de données qui permet de modéliser les relations entre des entités. La modélisation d'un graphe repose sur deux notions, celle de nœud et celle d'arc. La notion de nœud est associée aux entités qui sont en relation et celle d'arc est associée à la nature de la relation d'une entité (nœud) avec une autre.

Descripteurs extraits des graphes :

Après la construction d'un graphe, il est possible d'extraire des variétés de descripteurs liés soit aux nœuds et donc aux entités, soit à la relation entre les nœuds, et même à la topologie du graphe.

- **Descripteurs d'un nœud :** on peut citer les mesures de centralité qui estiment à quel point le nœud est incontournable dans la navigation dans le graphe. C'est le cas par exemple du PageRank [15] où initialement on attribue le même poids à chaque nœud, puis chaque nœud diffuse son poids à tous ses voisins directs proportionnellement aux poids des relations avec ses voisins. Le processus est répété jusqu'à ce que les poids des nœuds ne changent plus, ou alors jusqu'à ce qu'un nombre maximum d'étapes de diffusion soit atteint. Les nœuds aux poids les plus grands, sont les plus importants.
- **Descripteurs d'une relation entre deux nœuds :** on peut parler des mesures qui décrivent la relation entre deux nœuds à partir du nombre et de la longueur des plus courts chemins entre ces nœuds.
- **Descripteurs de la topologie du graphe :** il est possible d'extraire des communautés dans un graphe (sous-ensemble de nœuds densément connectés entre eux et faiblement connectés au reste du graphe).

Le procédé qui consiste à construire un graphe et à calculer des descripteurs, permet d'apporter de nouvelles informations pour enrichir la description des entités considérées en entrée d'un problème abordé.

V-1-2- Positionnement dans le projet

Dans cette section du projet, il est question de modéliser les données d'une base de prêts bancaires par des graphes dont la définition des nœuds et des arcs est suffisamment pertinente pour que les nouveaux descripteurs extraits de ces graphes contribuent fortement à la décision des modèles d'apprentissage automatique pour la prédiction du risque de crédit.

V-1- Lecture sur l'usage des descripteurs issus de la modélisation des données en graphes, pour la prédiction du risque de crédit

Durant ce trimestre, la lecture de l'article « Multilayer network analysis for improved credit risk prediction » de Óskarsdóttir et al. [16], a été le point d'entrée pour l'usage des descripteurs issus de la modélisation en graphe, pour enrichir les données d'apprentissage des modèles classique de prédiction du risque de crédit.

En effet, les auteurs utilisent la modélisation par des graphes multicouches (multilayer network) où un emprunteur a autant de nœuds qu'il y a de dimensions qui le caractérisent, et dans chaque dimension, il est relié à des attributs qui le définissent suivant cette dimension. Ainsi, plus les emprunteurs sont similaires, plus ils sont proches dans le graphe multicouche. Une dimension peut par exemple être la localisation géographique ou encore le type d'activité exercé.

Les auteurs s'appuient sur l'idée selon laquelle, les prêts des emprunteurs qui ont un grand nombre de caractéristiques en commun (suivant l'ensemble des dimensions) doivent avoir de grandes probabilités d'être de la même classe.

Ainsi, il se pose les difficultés suivantes :

- Comment établir les relations entre les emprunteurs ?
- Comment déduire des caractéristiques à exploiter à partir de la nouvelle représentation ?
- Comment prédire la classe d'un prêt ?

V-1-1- Processus de construction du graphe multicouche

Pour établir les relations entre les emprunteurs, les auteurs construisent un graphe multicouche.

Pour construire un tel graphe, on fixe les dimensions considérées et les attributs associés à chacune de ces dimensions.

Dans le cas de l'article qui s'attarde sur les prêts agricoles, les deux dimensions choisis sont : la localisation géographique et les produits vendus par les agriculteurs.

Les attributs de la dimension localisation géographique peuvent être le district, l'arrondissement etc, et concernant la dimension produit, les attributs peuvent être les différents produits répertoriés.

Dans le graphe multicouche :

- chaque emprunteur a autant de nœuds qu'il y a de couches considérées
- les nœuds de chaque emprunteur sont tous reliés les uns aux autres
- chaque attribut d'une dimension a un nœud associé

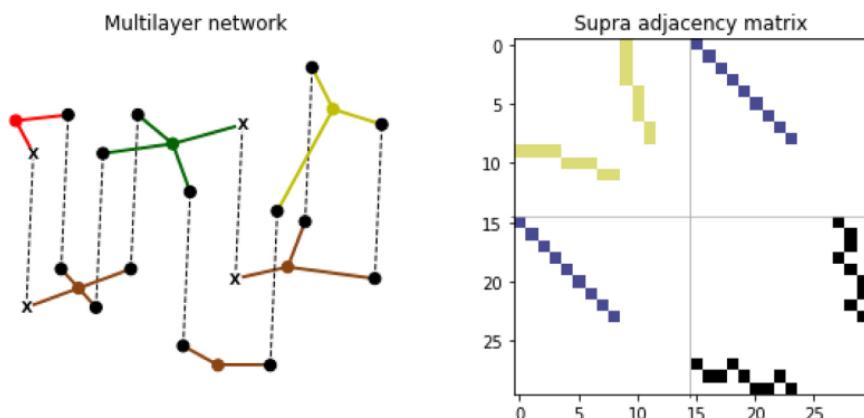
- si un emprunteur est décrit par un attribut dans une dimension donnée, alors le nœud emprunteur de cette dimension est relié au nœud attribut associé
- la navigation d'une couche à une autre se fait en passant par les nœuds emprunteurs des différentes couches

Construction de la matrice d'un graphe multicouche

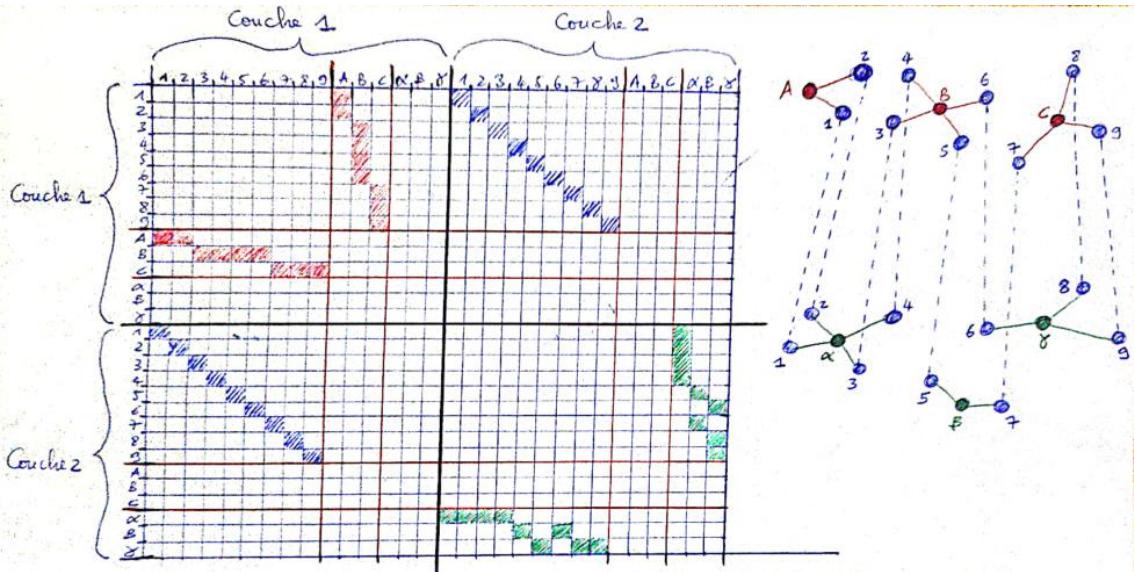
Un graphe multicouche M, ayant N nœuds, et L couches, correspond à une représentation de dimension $N \times N \times L \times L$, ceci peut être résumé en une matrice carrée $(N \times L) \times (N \times L)$.

Dans l'article, les auteurs considèrent deux dimensions pour décrire les emprunteurs dans le graphe multicouche, à savoir la localité et les produits vendus par ces derniers.

Considérons un cas où nous avons 4 emprunteurs (des fermiers), 2 localités (localité des fermiers) et 3 produits (produits agricoles vendus par les fermiers). Dans ce cas de figure, nous avons 3 couches (Emprunteur, Localité et Produit), et nous avons 9 nœuds (4 nœuds emprunteurs + 2 nœuds localités + 3 nœuds produits), et donc la matrice carrée qui permet de représenter le graphe multicouche est de taille $(9 \times 3) \times (9 \times 3)$



A gauche on a un exemple de graphe multicouche à deux couches, et à droite on a la représentation de ce graphe sous forme de matrice. Ce graphe contient 9 emprunteurs (nœuds noirs), 3 nœuds de la dimension localité des emprunteurs (nœuds marron) et 3 nœuds de la dimension produits (vert, rouge, jaune). Les relations inter-couches sont matérialisées par des traits interrompus et existent uniquement entre les nœuds emprunteurs qui représentent le même emprunteur dans les différentes couches. Les relations intra-couche sont matérialisés par les autres types de trait (trait marron dans la dimension localité et les autres couleurs dans la dimension produit).



Couche 11 → Relations entre les emprunteurs et les attributs de la première dimension

Couche 22 → Relations entre les emprunteurs et les utilisateurs de la seconde dimension

Couche 12 → Relations inter-dimensionnelles

Couche 21

Emprunteurs = {1, 2, 3, 4, 5, 6, 7, 8, 9} - 9

Produits = {A, B, C} - 3

localités = {x, y, z} - 3

$N = 9 + 3 + 3$ $L = 2$ (pour 2 dimensions) Matrice $(N \times L) \times (N \times L)$

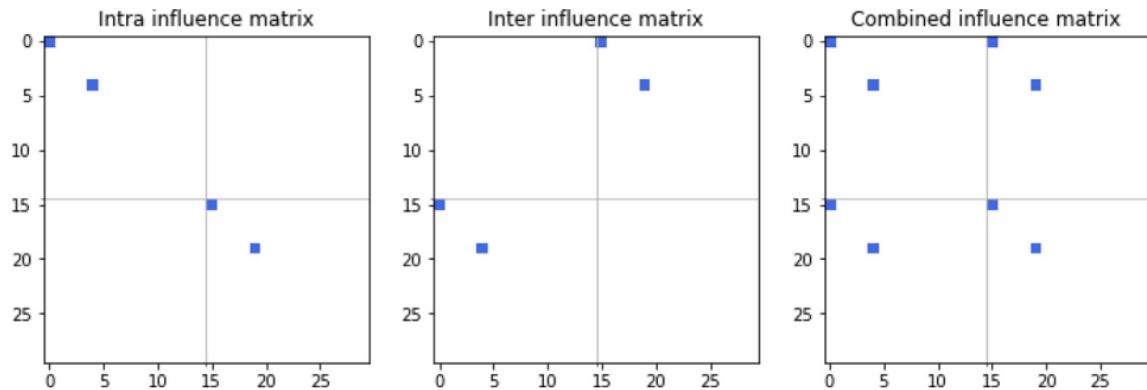
Représente le même graphe multicouche précédent, mais avec des détails supplémentaires sur le procédé de construction du graphe et de la matrice associée. Un emprunteur (fermier) est relié à sa localité et aux produits agricoles qu'il commercialise.

V-1-2- Calcul des nouvelles caractéristiques à exploiter à partir du graphe

Lorsque le graphe multicouche est construit, les nouveaux descripteurs du prêt sont calculés suite à des applications du PageRank Personnalisé sur le graphe résultat.

Les auteurs proposent 03 façons différentes de calculer les nouveaux descripteurs :

- **Intra-influence** : le PageRank Personnalisé est initialisé de manière à favoriser les relations intra-couche dans le processus de diffusion.
- **Inter-influence** : le PageRank Personnalisé est initialisé de manière à favoriser les relations inter-couche dans le processus de diffusion.
- **Influence-combinée** : le PageRank Personnalisé ne favorise pas un type de relation.



Les trois scénarios considérés par les auteurs pour calculer les nouveaux descripteurs des prêts par l'application du PageRank Personnalisé sur le graphe multicouche.

Notons que d'autres descripteurs sont considérés dans le graphe multicouche, à l'exemple de :

- Nombre de nœuds emprunteurs qui vendent les mêmes produits que l'emprunteur cible
- Nombre de nœuds emprunteurs défaillants qui vendent les mêmes produits que l'emprunteur cible
- Nombre de nœuds emprunteurs de la même localité que l'emprunteur cible
- Nombre de nœuds emprunteurs défaillants de la même localité que l'emprunteur cible
- Nombre d'emprunteurs de la même localité que l'emprunteur cible et qui vendent les mêmes produits que lui
- Nombre d'emprunteurs défaillants de la même localité que l'emprunteur cible et qui vendent les mêmes produits que lui

V-1-3- Prédiction de la classe du prêt

Pour procéder à la prédiction du risque de crédit avec les modèles classiques d'apprentissage automatique, les descripteurs présents dans le jeu de données, et les nouveaux descripteurs extraits des graphes, sont utilisés comme données d'apprentissage des modèles classiques choisis (Régression logistique et XGBoost) pour la prédiction des risques de crédit.

Une fois que ces modèles sont construits, ces derniers sont utilisés pour prédire les classes des prêts du jeu de test. Dans l'article, les comparaisons des performances des modèles avant et après l'insertion des nouveaux descripteurs, montrent que les nouveaux descripteurs améliorent la qualité des prédictions.

Par ailleurs, les analyses sur l'explicabilité de ces modèles ont montré que les nouveaux descripteurs étaient parmi ceux qui contribuent le plus à la prise de décision des modèles de Régression logistique et XGBoost.

[Trimestre 3]

V- Prédiction du risque de crédit à base de descripteurs issus de la modélisation des données en graphes

Le trimestre dernier nous avons procédé à la lecture d'un article sur l'extraction de nouveaux descripteurs issus des graphes « Multilayer network analysis for improved credit risk prediction ». Les concepts appris de cette lecture ont été implémenté ce trimestre pour la prédiction de risque de crédit bancaire sur le jeu de données de Afriland First Bank.

V-1- Positionnement dans le projet

Dans cette section du projet, il est question de modéliser les données d'une base de prêts bancaires par des graphes dont la définition des nœuds et des arcs est suffisamment pertinente pour que les nouveaux descripteurs extraits de ces graphes contribuent fortement à la décision des modèles d'apprentissage automatique pour la prédiction du risque de crédit.

V-2- Attributs extraits du graphe multicouches

Nous considérons le résumé de l'article « Multilayer network analysis for improved credit risk prediction » de Óskarsdóttir et al. [6], ainsi que les concepts présentés dans le rapport du précédent trimestre.

Pour expérimenter les concepts appris, il est nécessaire de choisir des attributs qualitatifs qui vont représenter les dimensions (couches) du graphe multicouche. Nous avons donc recensé les attributs catégoriels du jeu de données AFB de Afriland First Bank.

Lorsqu'on ignore la classe des prêts, les attributs catégoriels de ce jeu de données sont :

- **Type / Motif** : le type de prêt ou motif du prêt bancaire
- **Fonction** : le métier ou l'occupation de l'emprunteur
- **Civilité** : civilité de l'emprunteur (Monsieur, Madame, Mademoiselle)
- **Statut matrimonial** : statut matrimonial de l'emprunteur (Célibataire, Marié, Divorcé)

Après avoir recensé les attributs catégoriels, nous avons choisi d'implémenter trois graphes multicouches à deux couches. Le premier graphe multicouche nommé ici **MLN1** est construit à partir des attributs **Fonction & Civilité**, le second graphe **MLN2** est construit à partir des attributs **Fonction & Statut-Matrimonial** et enfin, le troisième graphe **MLN3** est construit à partir des attributs **Fonction & Motif**.

Les attributs extraits des ces différents graphes multicouches sont énumérés dans les sous-sections qui suivent.

V-2-1- Attributs du graphe MLN1 – Fonction & Civilité

- **MLN_fonction_degré** : nombre d'emprunteurs avec la même fonction
- **MLN_civilité_degré** : nombre d'emprunteurs avec la même civilité
- **MLN_fonction_et_civilité_degré** : nombre d'emprunteurs avec la même fonction et civilité
- **MLN_bipart_intra_fonction_civilité** : score PageRank de la valeur maximal de civilité ou fonction issue du PageRank avec initialisation uniquement des nœuds intra de la couche de civilité et fonction

- **MLN_bipart_inter_fonction_civilité:** score PageRank maximal du PageRank avec initialisation uniquement des nœuds inter de la couche de civilité et fonction
- **MLN_bipart_combine_fonction_civilité:** score PageRank maximal du PageRank avec initialisation de tous les nœuds du graph
- **MLN_bipart_intra_fonction_max:** score PageRank maximal de la fonction du PageRank avec initialisation uniquement des nœuds intra
- **MLN_bipart_inter_fonction_max:** score PageRank maximal de la fonction du PageRank avec initialisation uniquement des nœuds inter
- **MLN_bipart_combine_fonction_max:** score PageRank maximal de la fonction du PageRank avec initialisation uniquement des nœuds intra
- **MLN_bipart_intra_civilité_max:** score PageRank maximal de la civilité du PageRank avec initialisation uniquement des nœuds intra
- **MLN_bipart_inter_civilité_max:** score PageRank maximal de la civilité du PageRank avec initialisation uniquement des nœuds inter
- **MLN_bipart_combine_civilité_max:** score PageRank maximal de la civilité du PageRank

V-2-2- Attributs du graphe MLN2 – Fonction & Statut-Matrimonial

- **MLN_fonction_degré :** nombre d'emprunteurs avec la même fonction
- **MLN_stit_matrim_degré:** nombre d'emprunteurs avec la même situation matrimoniale
- **MLN_fonction_et_sit_matrim_degré:** nombre d'emprunteurs avec la même fonction et situation matrimoniale
- **MLN_bipart_intra_fonction_sit_matrim:** score PageRank maximal du PageRank avec initialisation uniquement des nœuds intra de la couche de civilité et situation matrimoniale
- **MLN_bipart_inter_fonction_sit_matrim:** score PageRank maximal du PageRank avec initialisation uniquement des nœuds inter de la couche de situation matrimoniale et fonction
- **MLN_bipart_combine_fonction_sit_matrim:** score PageRank maximal du PageRank avec initialisation de tous les nœuds du graph
- **MLN_bipart_intra_fonction_max:** score PageRank maximal de la fonction du PageRank avec initialisation uniquement des nœuds intra
- **MLN_bipart_inter_fonction_max:** score PageRank maximal de la fonction du PageRank avec initialisation uniquement des nœuds inter
- **MLN_bipart_combine_fonction_max:** score PageRank maximal de la fonction du PageRank avec initialisation uniquement des nœuds intra
- **MLN_bipart_intra_sit_matrim_max:** score PageRank maximal de la situation matrimoniale du PageRank avec initialisation uniquement des nœuds intra
- **MLN_bipart_inter_sit_matrim_max:** score PageRank maximal de la situation matrimoniale du PageRank avec initialisation uniquement des nœuds inter
- **MLN_bipart_combine_sit_matrim_max:** score PageRank maximal de la situation matrimoniale du PageRank

V-2-3- Attributs du graphe MLN3 – Fonction & Motif

- **MLN_fonction_degré** : nombre d'emprunteurs avec la même fonction
- **MLN_motif_degré**: nombre d'emprunteurs avec le même motif de prêt
- **MLN_fonction_et_motif_degré**: nombre d'emprunteurs avec la même fonction et motif de prêt
- **MLN_bipart_intra_fonction_motif**: score PageRank maximal du PageRank avec initialisation uniquement des nœuds intra de la couche de civilité et motif
- **MLN_bipart_inter_fonction_motif**: score PageRank maximal du PageRank avec initialisation uniquement des nœuds inter de la couche de motif et fonction
- **MLN_bipart_combine_fonction_motif**: score PageRank maximal du PageRank avec initialisation de tous les nœuds du graph
- **MLN_bipart_intra_fonction_max**: score PageRank maximal de la fonction du PageRank avec initialisation uniquement des nœuds intra
- **MLN_bipart_inter_fonction_max**: score PageRank maximal de la fonction du PageRank avec initialisation uniquement des nœuds inter
- **MLN_bipart_combine_fonction_max**: score PageRank maximal de la fonction du PageRank
- **MLN_bipart_intra_motif_max**: score PageRank maximal du motif du PageRank avec initialisation uniquement des nœuds intra
- **MLN_bipart_inter_motif_max**: score PageRank maximal du motif du PageRank avec initialisation uniquement des nœuds inter
- **MLN_bipart_combine_motif_max**: score PageRank maximal du motif du PageRank

V-3- Mise en œuvre : intégration des attributs extraits des graphes multicouches dans le processus de prédiction du risque de crédit

Nous avons considéré cinq modèles classiques de l'apprentissage automatique pour la prédiction du risque de crédit : Arbre de décision, Forêt Aléatoire d'arbre de décision (*Random Forest*), Régression logistique, XGBoost et SVM.

Pour chacun des graphes multicouches considérés (MLN1, MLN2, et MLN3), chaque modèle de prédiction est appliqué 04 fois. Chacune des applications du modèle diffère de l'autre par l'ensemble d'attributs descripteurs considérés :

- **Classic** : les attributs considérés sont tous ceux fournis avec le jeu de données
- **Classic + MLNi** : on considère tous les attributs du jeu de données et on ajoute les douze autres attributs extraits du graphe multicouches MLNi
- **Classic – AMLNi** : on considère une partie des attributs fournis avec le jeu de données. Ceux qui sont liés aux dimensions du graphe multicouches MLNi sont ignorés. Par exemple, pour le cas MLN1, les attributs relatifs à Fonction et à Civilités seront complètement écartés de la phase d'apprentissage.
- **Classic – AMLNi + MLNi** : on écarte les attributs relatifs aux dimensions du graphe multicouche MLNi, et on intègre les douze attributs extraits du graphe multicouche MLNi.

Nous pouvons ainsi évaluer l'impact des attributs choisis dans le graphe multicouches associés à leur représentation standard fourni dans le jeu de données (Classic + MLNi), sans

leur représentation standard (Classic – AMLNi + MLNi), et enfin évaluer l'impact de leur absence des données d'apprentissage des modèles (Classic – MLNi).

V-3-1- Résultats des expérimentations

Le tableau ci-dessous présente l'ensemble des résultats obtenus pour les cinq différents modèles de prédiction SVM, XGBoost, Arbre de décision, Régression Logistique et Forêt Aléatoire, avec les trois graphes multicouches considérés (MLN1, MLN2, MLN3) et suivant les métriques Exactitude, Précision, Rappel et F1-score.

		Accuracy	Precision	Recall	F1-score
SVM	classic	0.9577	0.9518	0.9577	0.9577
	classic + MLN1	0.9576	0.9517	0.9576	0.9576
	classic + MLN2	0.9577	0.952	0.9577	0.9577
	classic + MLN3	0.9576	0.9515	0.9576	0.9576
	classic - AMLN1	0.9577	0.9516	0.9577	0.9577
	classic - AMLN2	0.9577	0.9516	0.9577	0.9577
	classic - AMLN3	0.8153	0.9348	0.8154	0.8118
	classic - AMLN1 + MLN1	0.9575	0.9515	0.9575	0.9575
	classic - AMLN2 + MLN2	0.9575	0.9513	0.9575	0.9575
	classic - AMLN3 + MLN3	0.9553	0.9481	0.9553	0.9553
XGBOOST	classic	0.9954	0.9961	0.9954	0.9954
	classic + MLN1	0.9962	0.9977	0.9962	0.9962
	classic + MLN2	0.9956	0.997	0.9956	0.9956
	classic + MLN3	0.9961	0.997	0.9961	0.9961
	classic - AMLN1	0.9947	0.9954	0.9947	0.9947
	classic - AMLN2	0.9952	0.9961	0.9952	0.9952
	classic - AMLN3	0.9558	0.9813	0.9558	0.9558
	classic - AMLN1 + MLN1	0.9953	0.9959	0.9953	0.9953
	classic - AMLN2 + MLN2	0.9947	0.9956	0.9947	0.9947
	classic - AMLN3 + MLN3	0.9959	0.997	0.9959	0.9959
DECISION TREE	classic	0.9959	0.9988	0.9959	0.9959
	classic + MLN1	0.9945	0.9984	0.9945	0.9945
	classic + MLN2	0.9948	0.9984	0.9948	0.9948
	classic + MLN3	0.9932	0.9986	0.9932	0.9932
	classic - AMLN1	0.9947	0.9986	0.9947	0.9947
	classic - AMLN2	0.9948	0.9988	0.9948	0.9948
	classic - AMLN3	0.968	0.9871	0.968	0.9679
	classic - AMLN1 + MLN1	0.9949	0.9986	0.9949	0.9949
	classic - AMLN2 + MLN2	0.9961	0.9986	0.9961	0.9961
	classic - AMLN3 + MLN3	0.9949	0.9984	0.9949	0.9949
LOGISTIC REGRESSION	classic	0.9588	0.9638	0.9588	0.9588
	classic + MLN1	0.959	0.9638	0.959	0.959
	classic + MLN2	0.959	0.964	0.959	0.959
	classic + MLN3	0.9584	0.9636	0.9584	0.9584
	classic - AMLN1	0.9584	0.9649	0.9584	0.9584
	classic - AMLN2	0.9584	0.9649	0.9584	0.9584
	classic - AMLN3	0.8163	0.9394	0.8164	0.8126
	classic - AMLN1 + MLN1	0.9591	0.9649	0.9591	0.9591
	classic - AMLN2 + MLN2	0.9582	0.9646	0.9582	0.9582
	classic - AMLN3 + MLN3	0.9593	0.9602	0.9593	0.9593
RANDOM FOREST	classic	0.996	0.9993	0.996	0.996
	classic + MLN1	0.9952	0.9991	0.9952	0.9952
	classic + MLN2	0.9958	0.9993	0.9958	0.9958
	classic + MLN3	0.9951	0.9993	0.9951	0.9951
	classic - AMLN1	0.9963	0.9993	0.9963	0.9963
	classic - AMLN2	0.9962	0.9995	0.9962	0.9962
	classic - AMLN3	0.966	0.9918	0.966	0.966
	classic - AMLN1 + MLN1	0.9955	0.9991	0.9955	0.9955
	classic - AMLN2 + MLN2	0.9959	0.9993	0.9959	0.9959
	classic - AMLN3 + MLN3	0.9959	0.9995	0.9959	0.9959

Récapitulatif des résultats avec les différents graphes multicouches

Lorsqu'on s'attarde sur le classement des modèles de prédiction, le modèle de prédiction associé à la meilleure performance est la Forêt Aléatoire (Random Forest), suivi d'Arbre de décision et de XGboost. La Régression Logistique et SVM ferme ce classement.

En s'intéressant aux meilleures performances obtenues avec les attributs extraits des graphes multicouches, on constate que le graphe multicouche MLN1 est le meilleur pour le modèles XGBoost, le graphe MLN2 est meilleur pour SVM et Arbre de décision et enfin le graphe MLN3 est le meilleur pour Régression Logistique et Forêt Aléatoire.

Si on considère uniquement les cas de figure où chaque modèle est associé au graphe multicouches qui lui correspond le mieux, on fait les remarques suivantes :

- **SVM** : il ne faut pas considérer les attributs issus du graphe multicouches
- **XGBoost** : les meilleures performances sont atteintes lorsqu'on considère à la fois les attributs extraits du graphe multicouches ainsi que leur forme classique dans le jeu de données. Et la forme classique de ces attributs a plus d'impact que les attributs extraits du graphe multicouches.
- **Arbre de décision** : les meilleures performances sont atteintes lorsqu'on considère uniquement les attributs issus du graphe multicouches et qu'on ignore ces attributs dans leur représentation classique.
- **Régression Logistique** : les meilleures performances sont atteintes lorsqu'on considère uniquement les attributs issus du graphe multicouches et qu'on ignore ces attributs dans leur représentation classique.
- **Forêt Aléatoire** : en considérant la Précision comme métrique d'évaluation, la meilleure performance est atteinte lorsqu'on considère uniquement les attributs issus du graphe multicouches et qu'on ignore ces attributs dans leur représentation classique.

D'après les résultats obtenus, la considération des attributs extraits des graphes multicouches (**Classic + MLNi** et **Classic – AMLNi + MLNi**) permet l'amélioration des performances des modèles Arbre de décision, XGBoost, Random Forest et Régression Logistique. L'unique modèle pour lequel il n'y a pas d'amélioration mais un statu quo c'est le modèle SVM.

V-3-2- Evaluation de la contribution des attributs aux des modèles de prédiction

Suite à l'appréciation des résultats des modèles de prédiction, nous nous sommes intéressés aux contributions des différents attributs pour l'obtention des résultats considérés, notamment pour les cas de figure associés aux meilleurs graphes multicouches. Les graphiques des cinq pages suivantes illustrent ces informations (une page par modèle de prédiction).

Sur chaque graphique, plus une barre est longue et orientée vers la droite, plus l'attribut associé à cette barre contribue positivement aux décisions du modèle de prédiction. Par contre plus une barre est étirée vers la gauche, plus l'attribut contribue négativement aux décisions du modèle de prédiction.

Sur les graphiques, les barres peuvent avoir trois couleurs possibles :

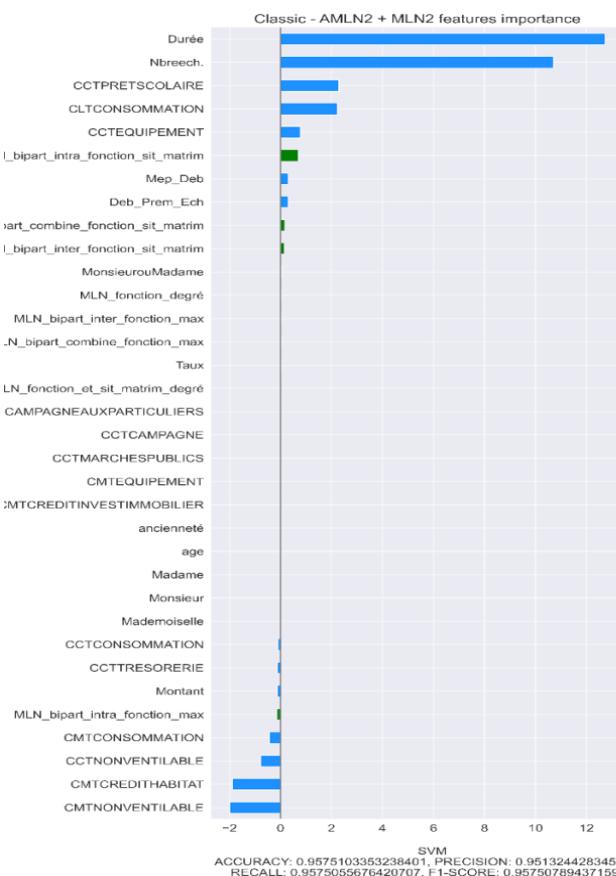
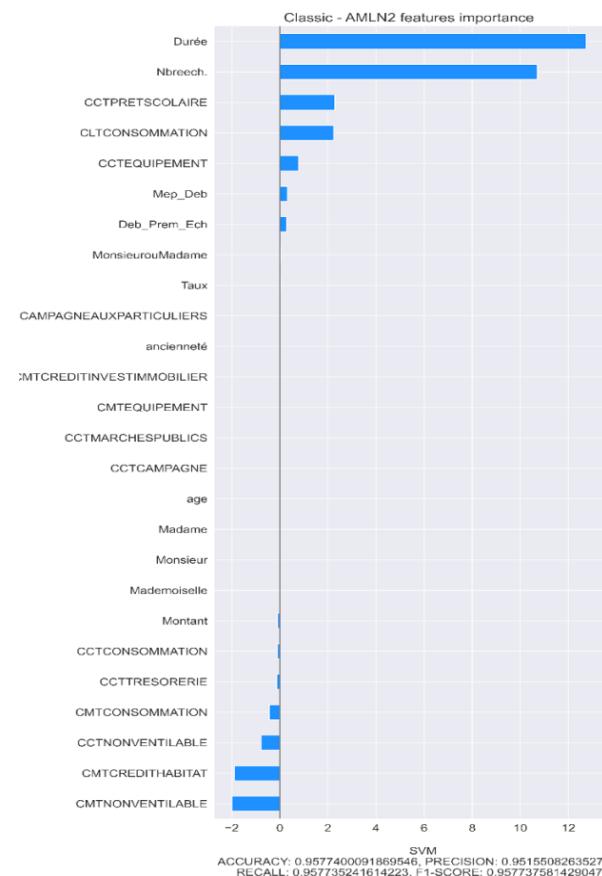
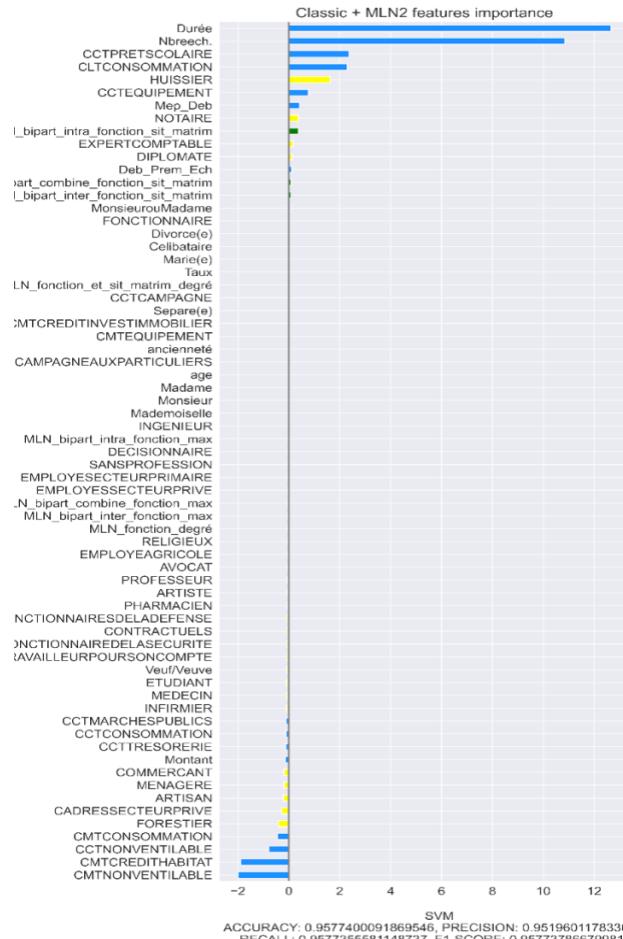
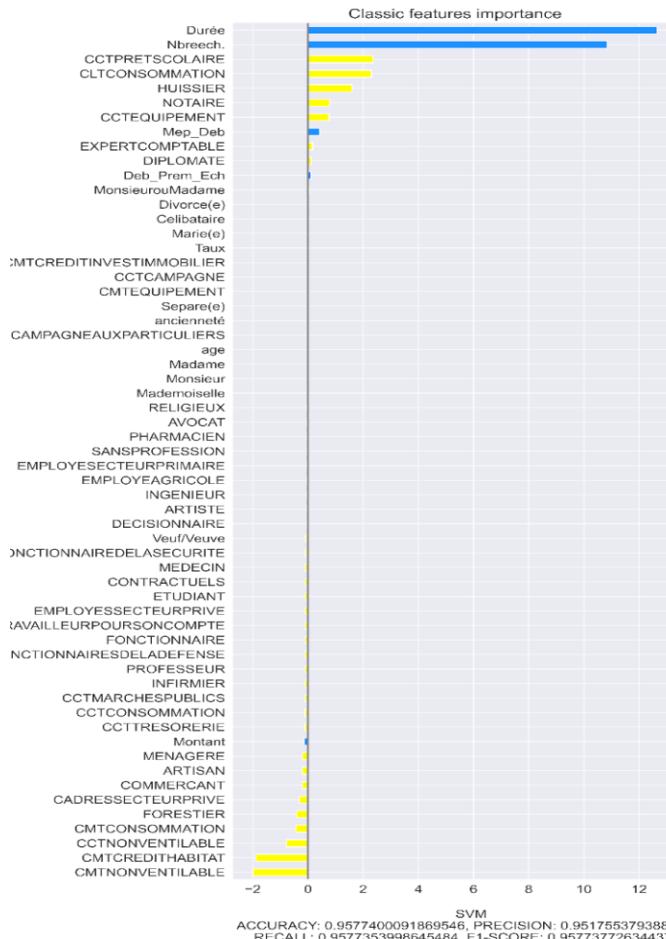
- **Les barres bleus** ; correspondent aux attributs classiques fournis dans le jeu de données, mais qui sont différents des attributs considérés pour le graphe multicouches.
- **Les barres jaunes** : correspondent aux attributs classiques fournis dans le jeu de données et qui sont considérés pour la construction du graphe multicouches.
- **Les barres vertes** : correspondent aux nouveaux descripteurs issus du graphe multicouches correspondant.

En observant les graphiques des résultats des combinaisons de type Classic + MLNi, on constate que pour les modèles de prédiction Forêt Aléatoire (MLN_bipart_intra_Motif_max), XGboost (MLN_Fonction_et_Civilité_Degrée) et Arbre de Décision (MLN_bipart_intra_Fonction_Sit_Matrim), les descripteurs issus du graphe multicouches sont mieux classés que les attributs associés sous leur formes classique

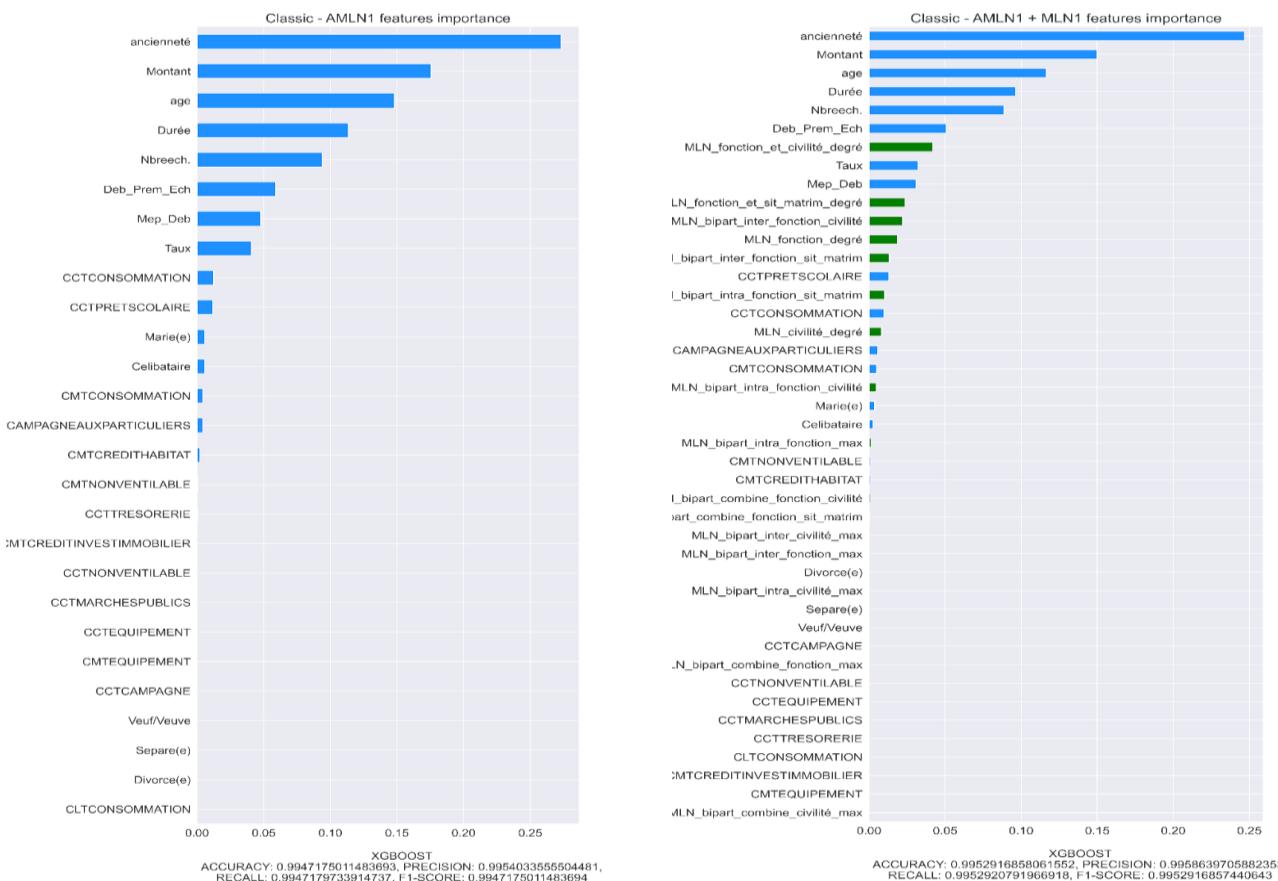
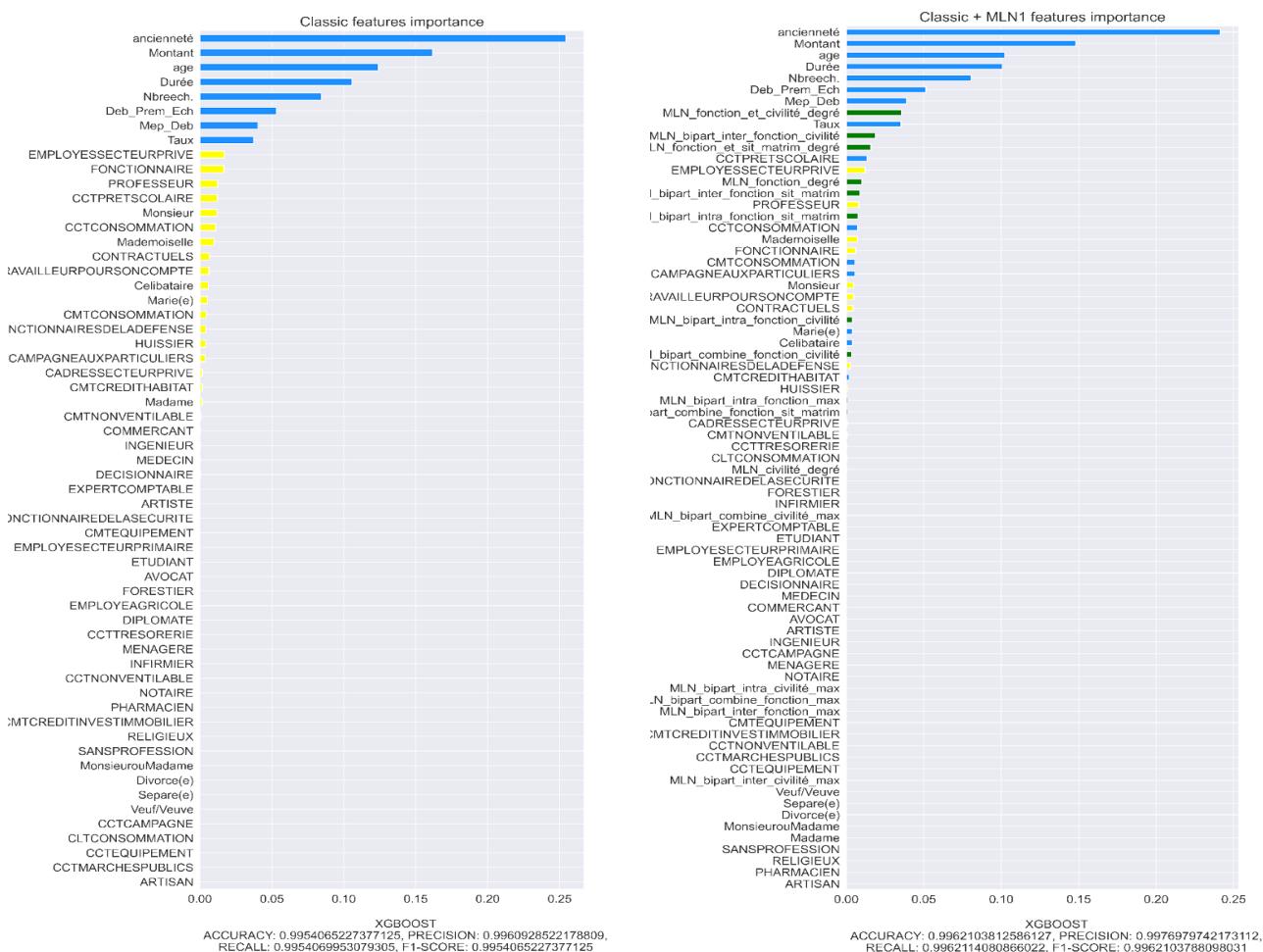
provenant du jeu de données. Par contre, la tendance est inversée pour les cas de SVM et la Régression Logistique où les descripteurs classiques sont mieux classés que ceux extraits des graphes multicouche.

Ce constat renforce la pertinence des descripteurs extraits des graphes multicouches, car ces derniers contribuent beaucoup plus pour les modèles de prédiction associés aux plus grandes performances. Ce qui est davantage renforcé par le cas de la Forêt Aléatoire qui est le modèle le plus performant de tous, et dont l'attribut `MLN_bipart_intra_Motif_max` extrait des graphes multicouches est celui qui contribue le plus aux prises de décisions.

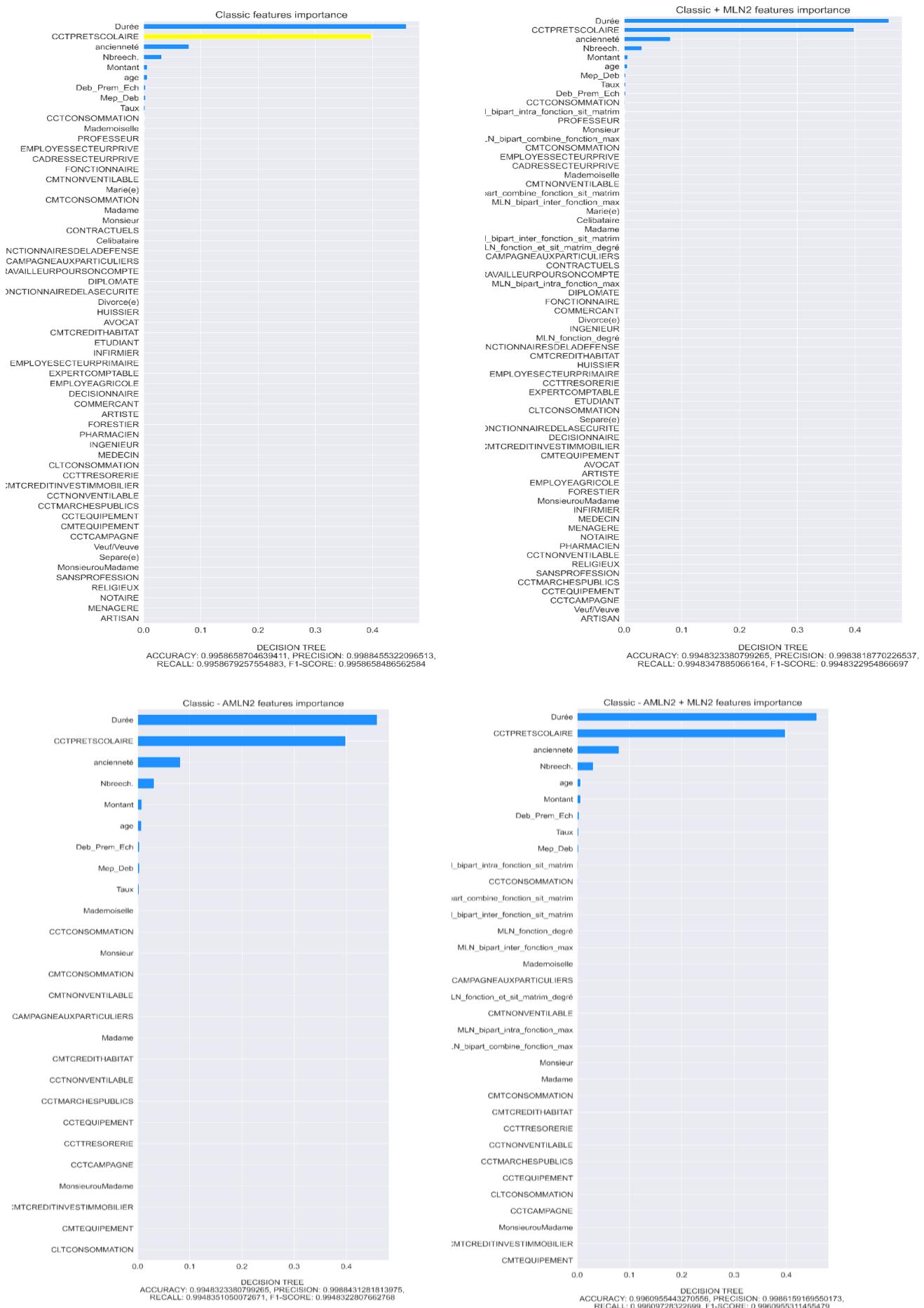
Cas du modèle SVM – meilleur graphe multicouche MLN2



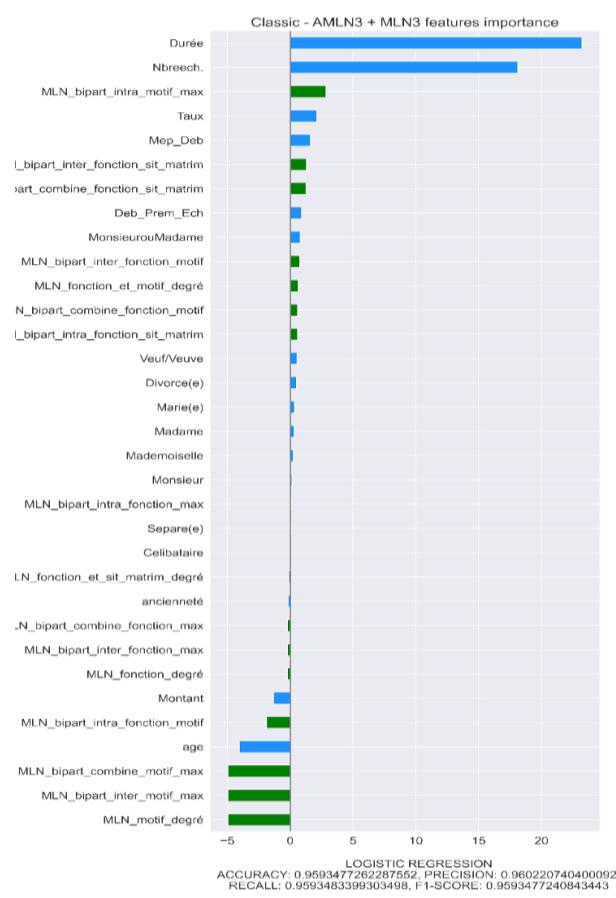
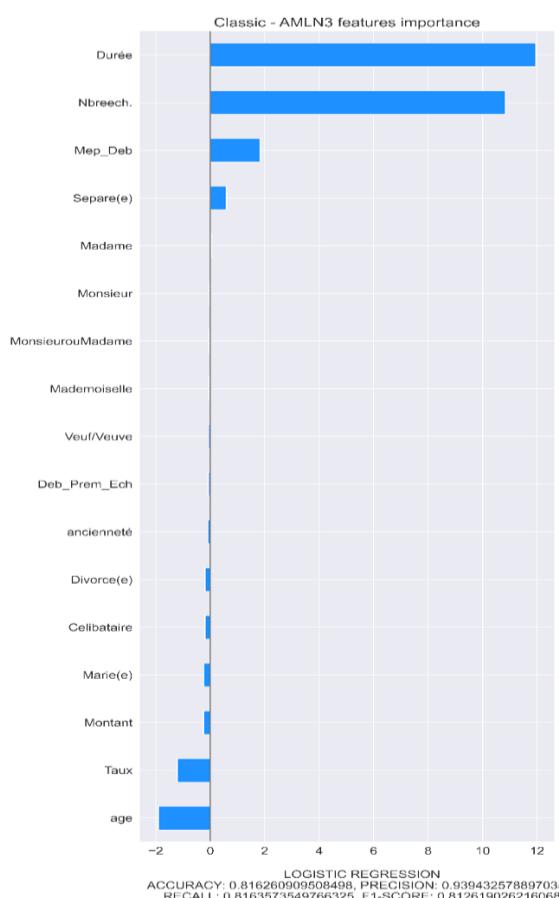
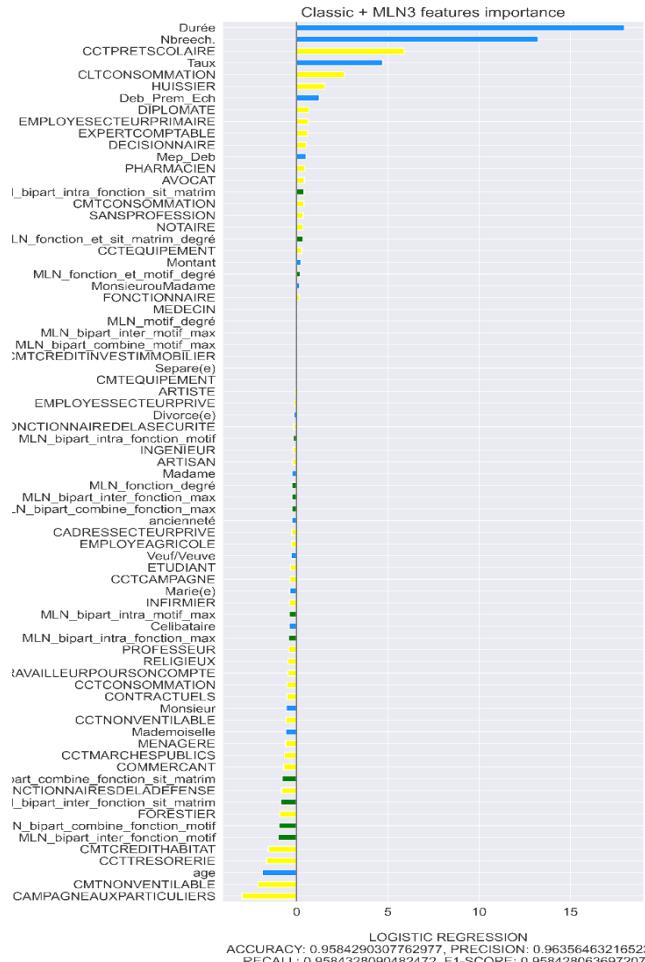
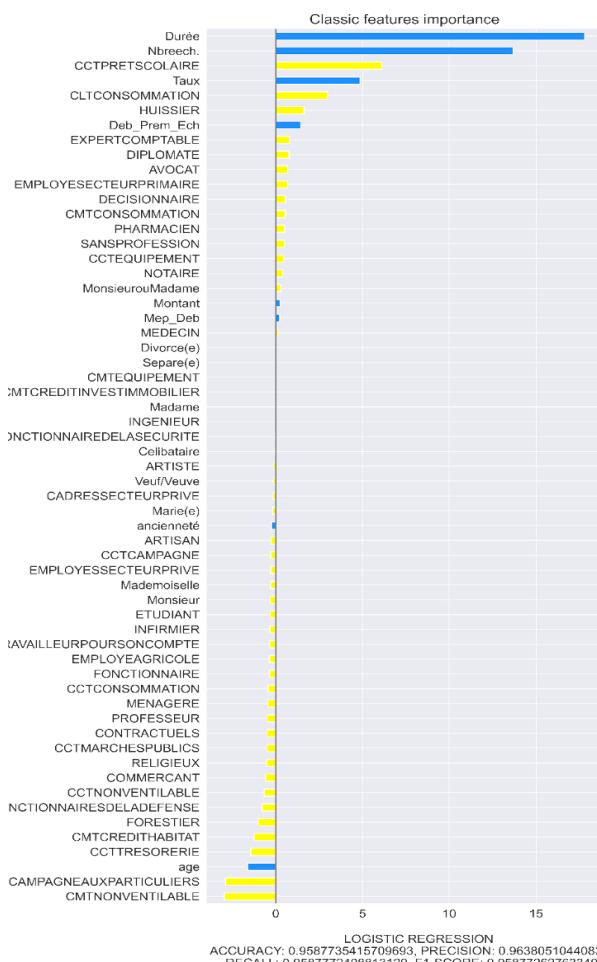
Cas du modèle XGBoost – meilleur graphe multicouche MLN1



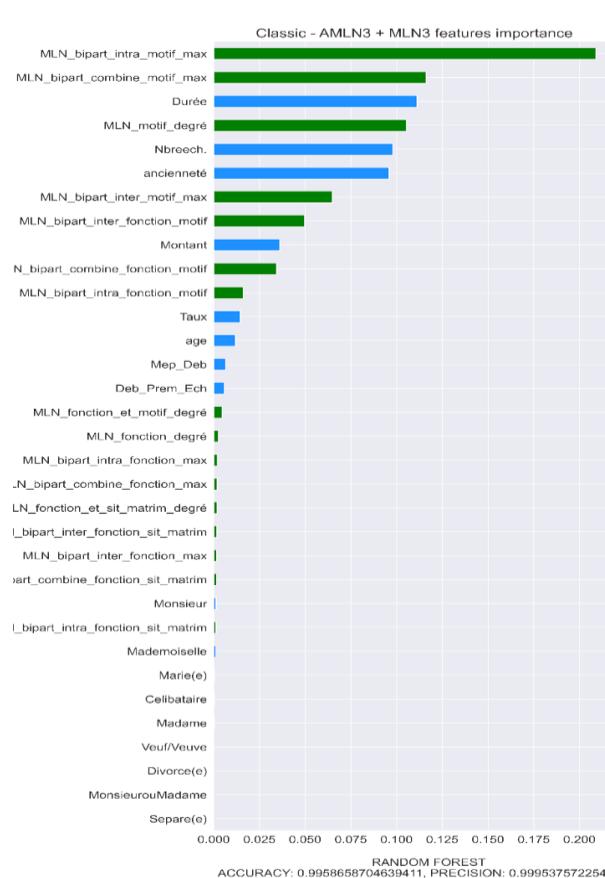
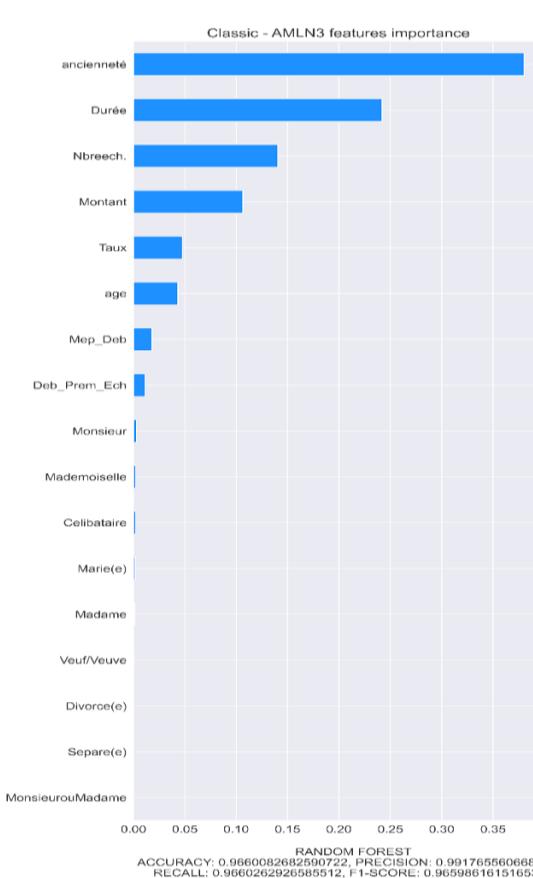
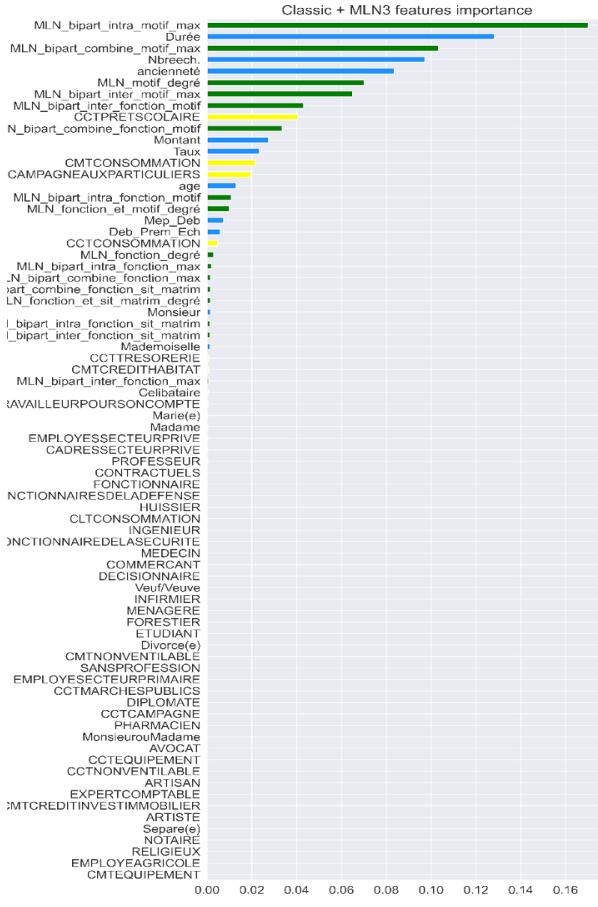
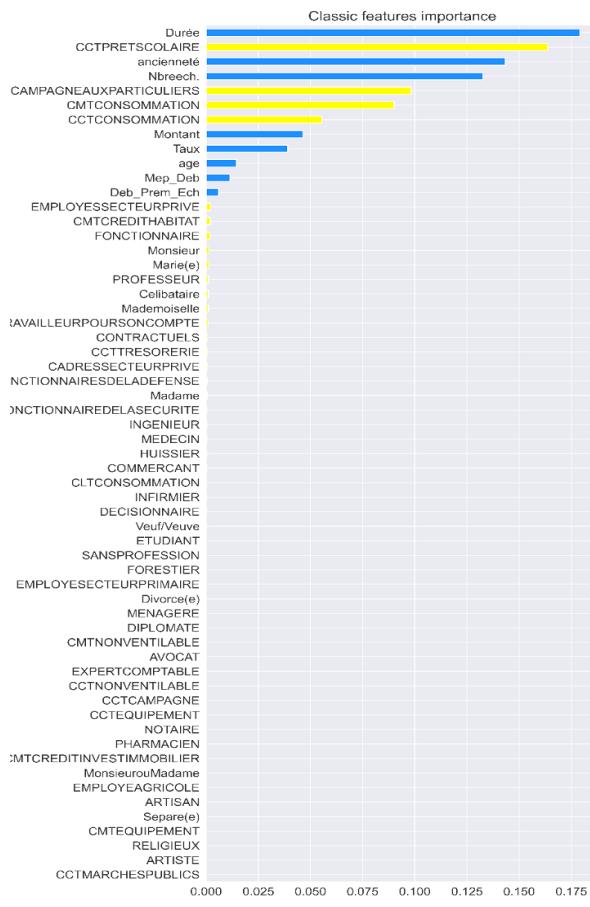
Cas du modèle Arbre de Décision (Decision Tree) – meilleur graphe multicouche MLN2



Cas du modèle Régression Logistique – meilleur graphe multicouche MLN3



Cas du modèle Forêt Aléatoire (Random Forest) – meilleur graphe multicouche MLN3



V-4- Conclusion et travaux futurs

Au terme de ce trimestre, l'extraction des descripteurs issus de la modélisation en graphe a été implémenté sur le jeu de données Afriland First Bank et son impact a été évalué pour les modèles de prédiction SVM, XGBoost, Arbre de Décision, Régression Logistique et enfin Forêt Aléatoire.

Nous avons une fois de plus fait le constat que les modèles de prédiction qui reposent sur les arbres de décisions sont les plus performants pour la tâche de prédiction du risque de crédit sur ce jeu de données. Par ailleurs, les descripteurs extraits des graphes multicouches contribuent mieux aux décisions prises par ces modèles comparés à la forme classique des attributs sur lesquels repose la construction du graphe multicouches. Autrement dit, les nouveaux descripteurs issus des graphes multicouches permettent d'améliorer les performances des meilleurs modèles de prédiction.

Durant le prochain trimestre, nous envisageons la construction des graphes multicouches avec des dimensions qui vont au-delà d'une taille de deux, et nous pourrons également explorer d'autres modélisations en graphes d'une base de données des prêts bancaires dans le but d'extraire de nouveaux descripteurs des prêts.

Suivant les dernières consignes du Professeur, nous développons des axes probables d'extension des travaux sur les graphes multicouches:

- 1- Construire un graphe multicouches avec toutes les variables qualitatives (sans sélectionner arbitrairement 2 comme les auteurs)
- 2- Appliquer un PageRank personnalisé pour chaque exemple pour lequel on doit faire une prédiction, contrairement aux auteurs qui applique un PageRank personnalisé global dont les résultats sont répercuté sur tous les exemples (En d'autre termes, personnaliser l'exécution du PageRank pour chaque exemple du jeu de données)
- 3- Évaluer l'intérêt de considérer qu'une variable qualitative doit être considérée dans sa forme initiale ou alors dans sa forme numérique en passant par le graphe multicouches.
(pour cela nous avons construit des graphes multicouches avec une seule couche, et donc un graphe multicouches pour chacun des attributs)
- 4- Faire ce travail sur plusieurs jeux de données.

Titre (correspondant à l'objectif fixé au sein du projet et en cohérence avec les rapports précédents)

0) Résumé

1) Introduction

2) Revue de la littérature sur le sujet (cette revue doit être très soignée et compréhensible par une personne de niveau M1 en informatique))

3) Présentation détaillée du modèle qu'on a analysé et approfondi

4) Contribution (éventuellement)

- Problème spécifique
- idée et positionnement par rapport à l'état de l'art
- modèle (présentation, propriétés, , ...)
- expérimentations et interprétation

5) Conclusion

Prière de ne pas faire du remplissage et de rédiger avec en tête le fait que le lecteur n'est pas spécialiste du sujet mais n'est PAS DUPE. On doit adopter pour la forme du rapport, la rigueur d'un article scientifique.