Rapport des atéliers de Machine Learning

Master Modélisation et Sciences de Données

Pierjos Francis COLERE MBOUKOU
Youssef HOURRI

M. Ali IDRI
Ali.IDRI@um6p.ma

Table des matières

1. Introduction

Les campagnes marketing représentent la voie royale pour communiquer avec ses clients, informer un marché cible du lancement d'un produit, de l'expansion d'un service ou d'une initiative de marque. Elles véhiculent l'identité de la marque, son style, ses valeurs. Pour cette raison, elles ne doivent pas être conçues à la légère. Elles comprennnent aussi souvent des incitations à l'engagement du public. La stratégie promotionnelle est l'un des éléments d'un plan marketing global conçu pour mobiliser les clients existants et atteindre l'objectif plus large de la campagne, à savoir attirer de nouveaux clients. Chaque client a des besoins particuliers, des attentes qui lui sont propres, des comportements différents. C'est la raison pour laquelle il faut s'adresser à ses clients de manière différenciée.

Aujourd'hui, les spécialistes du marketing utilisent le Machine Learning (ML) pour trouver des tendances, structures dans les activités des clients sur un support de communication donnné. Cela permet, ainsi, de prédire le comportement ultérieur des clients, d'optimiser les offres et de prendre rapidement des décisions en fonction des données de clients.

Dans ce document, nous préparons, traitons avant tout les données puis évaluons et comparons les performances des techniques de Machine Learning afin de classifier les clients en potentiel acheteur ou non d'un produit proposé par l'entreprise en se basant sur les données historiques de ces clients. Il s'agit, dans ce cas, d'une classification binaire. KNN (K-Nearest Neighbors, en français K plus proches voisins), Decision Trees (Arbres de décision) et SVM (Support-vector Machines, en français machines à vecteurs de support) sont les techniques ou modèles à évaluer et comparer sur l'ensemble de données BOGO [1]. Comme métriques d'évaluations empiriques, nous avons utilisé quatre critères de performance de classification (Accuracy, Recall ou Sensibilité, Precision et Score F1) pour évaluer ces modèles/techniques et la méthode de vote Borda Count pour classer et choisir le modèle le plus performant.

Dans cette étude, nous allons répondre aux questions :

— (Q1) : Quelle est la performance globale des techniques de Machine Learning dans la classification de données de clients ?

— (Q2) : Existe-t-il une technique de Machine Learning qui surpasse nettement les autres ?

Le reste du travail est organisé comme suit. La section 2 sera consacrée à la description de matériels utilisés. Nous présenterons, dans la section 3, la méthodologie suivie pour évaluer les modèles de Machine Learning. Les résultats et discussions sont présentés dans la section 4. Enfin, la section 5 conclut et s'ouvre sur de futurs travaux.

2. Description de matériels

2.1. Datasets

Dans ce projet, on a travaillé sur une base de données BOGO [1]. Cet ensemble de données contient 6400 données des clients et montre les informations brèves de ces derniers. 54,606 des clients n'achètent pas (c'est-à-dire classe 0) et 9,394 seulement achètent (classe 1). Les variables constituant cet ensemble de données sont décrites, explicitées ci-dessous :

- recency : nombre de mois depuis le dernier achat,
- history : valeur (somme) des achats historiques,
- used_discount : indique si le client a utilisé une remise avant,
- used_bogo : indique si le client a utilisé une offre de type "Buy One Get One",
- zip_code : classe du code postal en tant que Suburbain ou Urbain ou Rural
- is_referral : indique si le client a été acquis à partir d'un canal de référence,
- channel : canal utilisé (téléphone ou web ou multicanal),
- offer : offres envoyées au client (Discount ou But One Get One ou No Offer),
- Conversion : Conversion des clients (achat ou non du produit)

Voici un aperçu de l'ensemble de données :

	recency	history	used_discount	used_bogo	zip_code	is_referral	channel	offer	conversion
0	10	142.44	1	0	Surburban	0	Phone	Buy One Get One	0
1	6	329.08	1	1	Rural	1	Web	No Offer	0
2	7	180.65	0	1	Surburban	1	Web	Buy One Get One	0
3	9	675.83	1	0	Rural	1	Web	Discount	0
4	2	45.34	1	0	Urban	0	Web	Buy One Get One	0

FIGURE 2 – Aperçu du dataset BOGO

2.2. Techniques utilisées

2.2.1. K-Nearest Neighbors (KNN)

Nous avons utilisé l'algorithme K-Nearest Neighbors (KNN) ou K plus proche voisins pour prédire la conversion des clients. Souvent utilisé pour les problèmes prédictifs de classification, KNN est un algorithme d'apprentissage paresseux car il n'apprend pas une fonction à partir des données d'apprentissage, mais "mémorise" l'ensemble des données d'apprentissage, et non paramétrique (seul k doit être fixé).

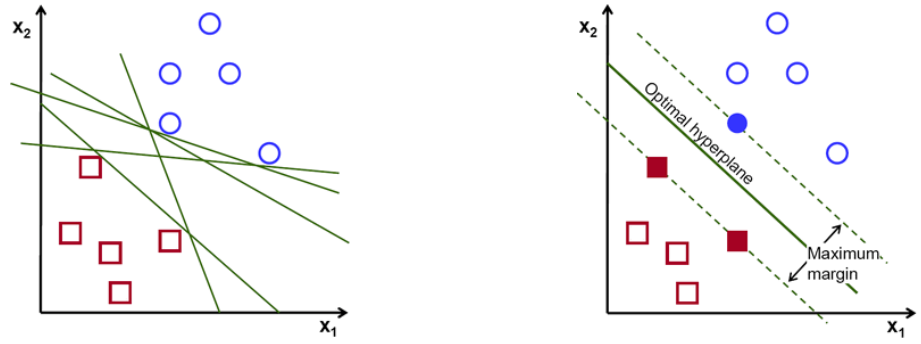
Il utilise la similarité des caractéristiques pour prédire la classe d'appartenance d'une nouvelle donnée en regardant quelle est la classe majoritaire des k données voisines les plus proches (d'où le nom de l'algorithme). Les métriques de similarité les plus souvent choisies sont la distance usuelle dite euclidienne.

2.2.2. Decision Trees (Arbres de décision)

Un arbre de décision est un modèle très simple. Etant donnée plusieurs caractéristiques (variables), la décision se commence par une de ces caractéristiques ; si ce n'ai pas suffisant, on utilise une autre, ainsi de suite. Il existe plusieurs algorithmes automatiques pour construire les arbres de décision : ID3, CART, etc. Dans ce travail, nous avons utilisé ID3. Ce dernier utilise la fonction entropie et le gain d'information pour décider quelle est la meilleure caractéristique.

2.2.3. Support-vector Machines (SVM)

Parmi d'autres algorithmes, nous avons utilisé SVM (Support Vector Machine) qui a pour objectif de trouver un hyperplan dans un espace à N dimensions (où N désigne le nombre de variables) qui classifie distinctement les données. Pour séparer les deux classes, de nombreux hyperplans peuvent être choisis. L'objectif est de trouver un plan qui présente la marge maximale, c'est-à-dire la distance maximale entre les données des deux classes. La maximisation de la distance de la marge fournit un certain renforcement, de sorte que les futures données peuvent être classées avec plus de confiance comme le montrent les figures ci-après.



(a) Possibles hyperplans du SVM **(b)** Meilleur hyperplan

FIGURE 3 – Illustration de l'algorithme SVM

2.2.4. Borda Count

La méthode du Borda Count est un outil simple qui est utilisé dans les élections et la prise de décision dans diverses situations. Dans cette technique, des points sont attribués aux candidats en fonction du critère et de leur classement. 1 point pour le dernier choix, 2 points pour l'avant-dernier choix, et ainsi de suite jusqu'à ce que l'on arrive à attrubyés de points à tous les candidats. Dans notre cas, les candidats et critères sont respectivement les différents modèles et les métriques d'évaluation. Cette méthode est qualifié de système de vote par consensus, car elle est adoptée pour s'assurer que l'on ne favorise pas un critère de performance particulier plutôt qu'un autre. [2]

Nous avons utilisé cette technique pour trouver le modèle Machine Learning le plus performant se basant sur les quatres critères d'évaluation.

2.2.5. Cross-validation

Nous évaluons les modèles à l'aide d'une méthode de validation croisée (cross-validation) appelée k-Fold (k=5 dans notre cas). Cette méthode divise l'ensemble des données en 5 sous-ensembles (folds) de façon à ce que chaque sous-ensemble contienne approximativement le même pourcentage d'échantillons de chaque classe cible que l'ensemble de données. Ensuite, elle choisit 4 sous-ensembles qui constituent l'ensemble d'entraînement. Le sous-ensemble restant constitue l'ensemble de test. Les modèles ont sont donc entrainés sur l'ensemble d'entraînement. Ensuite, on valide sur l'ensemble de test et enregistrons le résultat de la validation. Ce processus est répété 5 fois. Enfin, le score final n'est autre que la moyenne des résultats obtenus.

FIGURE 4 – Cross-validation 5-fold.

2.3. Mesures de performance

Comme mentionné dans les sections précédentes, nous avons utilisé quatre mesures pour évaluer la performance des modèles : accuracy, recall, precision, and F1 score. Ces métriques populaires sont définies de façon ci-dessous :

$$Accuracy = \frac{TP + TN}{TN + TP + FP + FN} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$F1 - score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{4}$$

Dans ces formules :

— TP (True Positive) signifie que l'individu malin est identifié comme malin.
— FP (False Positive) : un cas bénin est identifié comme malin.
— TN (True Negative) : un cas bénin est identifié comme bénin.
— FN (False Negative) : un cas malin est identifié comme bénin.

Tout d'abord, l'accuracy est la mesure de tous les cas correctement identifiés. La précision est une mesure qui quantifie le nombre de prédictions malignes correctes. On doit donc chercher à minimiser les cas bénins qui sont identifiés comme malins. Par contre, le recall est une métrique mesurant le nombre de prédictions malignes correctes faites parmi toutes les prédictions possibles ; on réduit le nombre de cas bénins étiquetés comme malins. Enfin, le score F1 est la moyenne pondérée de la Précision et du Recall. Par conséquent, il tient compte à la fois des faux positifs et des faux négatifs.

2.4. Pré-traitements (préparation) de données

2.4.1. Traitement des données aberrantes (outliers)

Une valeur aberrante est une donnée observée pour une variable qui semble anormale au regard des valeurs dont on dispose pour les autres observations de l'échantillon.

D'après la boîte à moustaches (boxplot) de history ci-dessous, nous remarquons qu'on est en présence d'une variable contenant des données aberrantes. Après une profonde analyse, ces données représentent 5.61% de l'ensemble de données. Le pourcentage étant petit, il est préférable de les supprimer du dataset pour faciliter les tâches à venir.

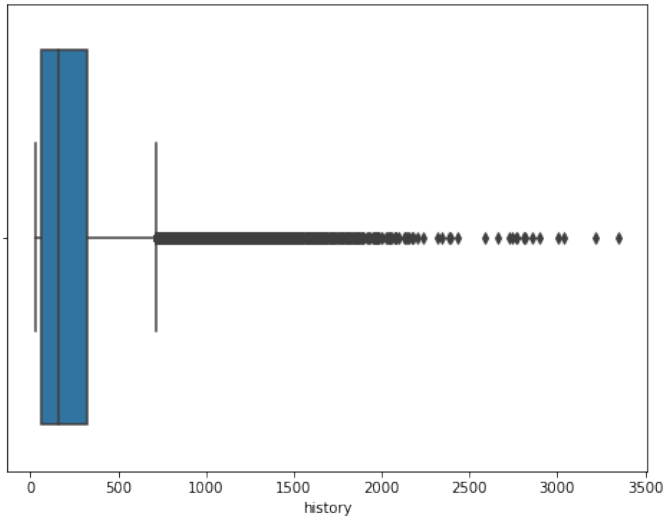


FIGURE 5 – Boîte à moustaches de la variable history

2.4.2. Encodage des données

Le dataset contient trois (3) variables qualitatives : zip_code, channel et offer. En d'autres termes, elles prennent des valeurs appelées catégories, modalités ou niveaux qui n'ont pas de sens quantitatif. La présence de ces variables dans les données complique généralement l'apprentissage. En effet, la plupart des algorithmes d'apprentissage automatique prennent des valeurs numériques en entrée. Ainsi, il faut transformer ces modalités en données numériques. Pour ce faire, nous faisons appel à la méthode de transformation appelée LabelEncoder qui consiste à coder les modalités cibles avec une valeur comprise entre 0 et n_mod - 1, où n'est autre que le nombre de modalités de la variable.

2.4.3. Augmentation de données

L'augmentation des données, plus précisement de classes, est principalement utilisée pour éviter le risque de sur-apprentissage (overfitting). Comme nous pouvons le remarquer, le dataset BOGO est déséquilibré en terme de classes. En effet, plus de 85% des données appartiennent à la classe 0 et seulement 15% pour la classe 1. L'un des problèmes des données déséquilibrées dans la classification est la mauvaise performance du modèle sur l'ensemble de test ou de validation. Les classes doivent donc être équilibrées. Pour ce faire, nous avons sur-échantillonné la classe minoritaire avec la méthode SMOTE (Synthetic Minority Oversampling Technique). SMOTE fonctionne en sélectionnant des exemples de données qui sont proches dans l'espace des variables, en traçant une ligne entre les exemples dans l'espace des caractéristiques et en dessinant un nouvel échantillon à un point le long de cette ligne. Cette méthode permet ainsi d'équilibrer notre dataset BOGO : il y a autant de données de classe 0 que celles de classe 1. Soit 51731 données pour chaque classe.

2.4.4. Normalisation de données

La normalisation est une méthode de prétraitement des données qui permet de réduire la complexité des modèles. C'est également un préalable à l'application de certains algorithmes (KNN par exemple). Dans le dataset utilisé ici, les variables ne sont pas sur une même échelle. Comme illustration, la variable history varie entre 29,990 et 3345, 930. Par contre, la variable recency a des valeurs entre 1 et 12. Il serait ainsi indispensable de les normaliser. À cet effet, nous utilisons la normalisation MinMax. L'idée est la suivante, on ramène toutes les valeurs de la variable entre 0 et 1, tout en conservant les distances entre les valeurs. La formule utilisée est :

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{5}$$

2.4.5. Sélection des variables (Features Selection)

Features Selection est une technique qui vise à réduire le nombre de variables de l'ensemble de données en sélectionnant celles qui sont pertinentes et en éliminant celles

qui ne sont pas pertinentes et redondantes.

2.4.5.1 Information Gain (Gain d'Information)

Information Gain est l'une de méthodes de filtrage qui consiste à calculer la réduction de l'entropie résultant de la transformation d'un ensemble de données. Cette méthode est utilisée pour la sélection des variables en évaluant le gain d'information de chaque variable dans le contexte de la variable cible. En appliquant ce filtre sur notre ensemble de données, l'importance ou contribution de chaque variable à l'explication de la variable cible peut être visualisée dans la figure ci-après.

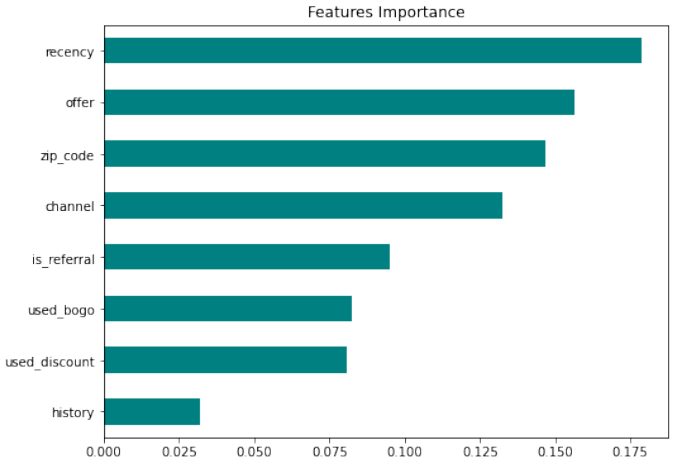


FIGURE 6 – Résultats de la méthode Information Gain.

2.4.5.2 Exhaustive Feature Selection

Il s'agit d'une méthode de wrappers (encapsulation). Cette méthode consiste à réaliser une évaluation par force brute de chaque sous-ensemble de variables. En d'autres termes, elle essaie toutes les combinaisons possibles de variables et renvoie le sous-ensemble le plus performant. L'application de cette méthode sur le dataset BOGO et la métrique Accuracy donne les résultats des 6 meilleures combinaisons de variables :

On remarque bien que l'on peut réduire la dimension de données de 8 à 5 variables. De plus, recency, zip_code, is_referral, channel et offer donnent un meilleur score en terme d'accuracy (86.37%).

TABLE 1 – Résultats de Exhaustive Feature Selection.

Variables	Score (%)
recency, zip_code, is_referral, channel, offer	86.37
used_discount, zip_code, is_referral, channel, offer	86.35
recency, used_discount, zip_code, is_referral, offer	86.23
recency, used_bogo, zip_code, is_referral, offer	86.18
used_discount', used_bogo, zip_code, is_referral, channel, offer	86.15
recency, used_bogo, zip_code, is_referral, channel	85.98

2.5. Abréviations

Afin de faciliter la lecture des noms des techniques et modèles de Machine Learning, nous abrégeons le nom de chaque technique comme le montre le tableau 2.

TABLE 2 – Abréviations de techniques.

Abréviation	Signification
GS	GridSearch (recherche des hyperparamètres optimaux).
FS	Sélection des variables (Features Selection) avec la méthode de gain d'information.
Wrapper	Sélection des variables avec la méthode de wrappers (Exhaustive Feature Selection)
FS20	Utilisation de 20% des variables après Features Selection.
FS40	Utilisation de 40% des variables après Features Selection.
FS50	Utilisation de 50% des variables après Features Selection.
FS70	Utilisation de 70% des variables après Features Selection.
KNN	L'algorithme KNN par défaut
DT	L'algorithme d'arbre de décision implémenté par défaut
SVM	L'algorithme SVM par défaut
+	Association ou ajout. Par exemple, KNN + GS indique que KNN est optimisé.

3. Expérience design

La méthodologie que nous avons suivie pour réaliser toutes les évaluations empiriques est illustrée à la figure 7. Elle se compose de cinq étapes qui sont décrites ci-dessous :

(1) Acquisition des données.

(2) Préparation de données (outliers, encodage, augmentation et normalisation).

(3) Features Selection (Filtre : Information Gain, Wrapper : Exhaustive Feature Selection).

(4) Évaluer l'accuracy, precision, recall, score f1 de chaque modèle (KNN, SVM, Arbre de décision). Optimiser ces modèles en cherchant les meilleures valeurs de leurs hyperparamètres avec GridSearch.

(5) Classer les modèles en utilisant la méthode Borda Count à partir des quatre métriques (accuracy, precision , recall, et score F1). Enfin, sélectionner les meilleurs modèles de Machine Learning.

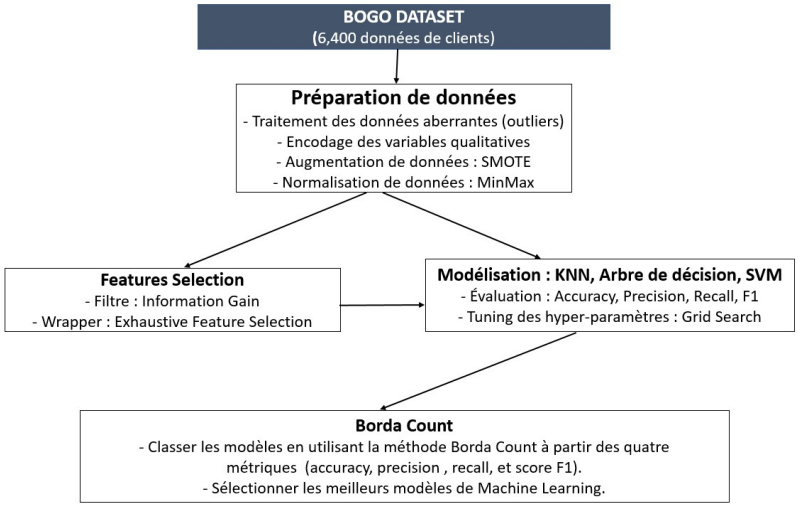


FIGURE 7 – Expérience design.

4. Analyse des résultats et discussion

Dans cette partie, nous montrons et discutons les résultats des évaluations empiriques des techniques de Machine Learning sur le jeux de données BOGO. Ces résultats sont résumés dans le tableau 3. Rappelons que les performances des techniques ML ont été évaluées à l'aide de quatre critères (accuracy, precision, recall et score F1).

D'après ces résultats, on peut dire que ces modèles (techniques) arrivent à généraliser et classifier les données de clients du dataset BOGO car les métriques sont supérieures à

la moyenne (50%). D'autre part, la reduction de dimension de données influence significativement et positivement les performances du modèle KNN. En effet, les performances de KNN + Wrapper + GS et KNN + FS70 + GS sont meilleures par rapport à celles de KNN + GS. De même, pour le SVM. En effectuant une sélection de variables (caractéristiques) par filtre ou wrapper, il y a une amélioration des valeurs de métriques comparé au SVM par défaut.

Ceci montre qu'il existe des variables moins importantes. En d'autres mots, des variables qui ne contribuent pas significativement au fait d'achéter ou non le produit. La corrélation ou le fait que des variables sont liées entre elles peut, d'une manière ou d'une autre, peut justifier, expliquer ces résultats. Par exemple, used_bogo et used_discount sont négativement corrélées (-0,81). Mais ces hypothèses restent à vérifier et valider.

Enfin, la méthode Borda Count appliquée aux critères d'évaluations du tableau 3, a révélé que que l'arbre de décision optimisé (DT + GS) dépasse les autres techniques et est donc classé premier, suivi de (KNN + Wrapper + GS), c'est-à-dire KNN optimisé et appliqué au meilleur sous-ensemble de variables trouvé en section 2.4.5.2. KNN + GS et DT occupent la quatrième (4e) place et sont très proche du troisième (3e) meilleur modèle KNN + FS70 + GS. En outre, le tableau 4 résume les résultats en terme de classement et score de la méthode Borda sur l'ensemble de données BOGO.

En résumé, DT + GS et KNN + Wrapper + GS sont les meilleurs modèles qui arrivent à généraliser et classifier les clients en potentiel achéteurs ou pas.

TABLE 3 – Résultats de la modélisation.

Technique	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
KNN	80,24	88,04	69,62	75,34
KNN + GS	84,75	87,07	82,06	81,78
KNN + FS20 + GS	79,80	1,00	59,59	70,90
KNN + FS40 + GS	84,44	1,00	68,88	77,16
KNN + FS50 + GS	85,91	97,85	73,78	79,22
KNN + FS70 + GS	85,95	97,79	73,88	79,30
KNN + Wrapper + GS	86,37	97,72	74,86	79,79
DT	84,84	86,32	83,75	81.49
DT + GS	86,56	90,19	81,97	82,38
SVM	72,69	77,97	63,34	69,90
SVM + GS	82,05	99,68	64,03	77, 97
SVM + Wrapper + GS	81,83	99,02	64,37	78,02

TABLE 4 – Résultats de Borda Count.

Rang	Technique	Score
1	DT + GS	40
2	KNN + Wrapper + GS	36
3	KNN + FS70 + GS	34
4	KNN + GS	33
4	DT	33
6	KNN + FS50 + GS	32
7	SVM + GS	24
7	SVM + Wrapper + GS	24
9	KNN	17
10	KNN + FS40 + GS	16
11	KNN + FS20 + GS	6
11	SVM	6

5. Conclusion et Perspectives

Ce travail présente et discute les résultats d'une étude empirique comparative des techniques de Machine Learning pour la classification des clients d'une entreprise susceptibles d'acheter ou non un produit par le biais de campagnes marketing. Dans ce document, pour évaluer ces modèles de ML, nous avons utilisé quatre critères de performance et la méthode Borda Count pour classer ces techniques en utilisant le dataset BOGO.

(Q1) : Quelle est la performance globale des techniques de Machine Learning dans la classification de données de clients ?

En termes d'accuracy, precision, recall et F1, nous avons observé que ces modèles donnaient les meilleurs résultats. Les valeurs de ces métriques sont supérieures à la moyenne (50%).

(Q2) : Existe-t-il une technique de Machine Learning qui surpasse nettement les autres ?

Enfin, d'après la méthode Borda Count l'arbre de décision optimisé (DT + GS) surpasse les autres techniques (score de 40) et est classé en premier. KNN + Wrapper + GS vient en deuxième position avec un score 36. Par conséquent, nous concluons que DT + GS est la meilleure technique pour cette classification.

Comme futurs trauvaux, nous allons chercher à comprendre pourquoi la feature selection donne de bons résultats. D'une part, nous allons confronter les résultats de sélection des variables (Information Gain et Exhaustive Feature Selection) aux méthodes intégrées (Embedded methods) et à la méthode de corrélation. D'autres part, nous chercherons à améliorer les performances de modèles, ou utiliser d'autres algorithmes de Machine Learning.

Références

[1] Marketing Promotion Campaign Uplift Modelling, Davin Wijaya, Kaggle.

[2] Emerson, "The original Borda count and partial voting", Soc. Choice Welfare, vol. 40, no. 2, pp.353–358, 2013. https ://doi.org/10.1007/BF02464423