Rapport de Projet : Apprentissage Supervisé

BOUCHOUIA Mohammed Lamine

MEKHALDI Rachda Naila

I. Introduction

L'apprentissage supervisé est l'une des techniques d'Intelligence Artificielle en plein essor. Elle est utilisée pour la classification et la prédiction des données. Il existe plusieurs méthodes mettant en œuvre l'Apprentissage Automatique, dont : Le Bayésien Naïf, K-plus Proches Voisins (KNN en anglais), etc.

Dans le cadre de ce projet, nous allons présenté plusieurs techniques d'Apprentissage Automatique que nous appliquons sur différents ensemble de données. Nous avons deux types d'ensembles de données : synthétiques et réelles. En premier, nous exposons les résultats obtenus sur les données synthétiques puis ceux sur les données réelles.

II. Données Synthétiques

Nous avons utilisé 3 data sets synthétiques décrits comme suit :

Data Set	Nombre d'observations	Nombre de variables	Nombre de classes
Flame	**240**	**3**	**2**
Spiral	**312**	**3**	**3**
Aggregation	**788**	**3**	**7**

Table 1: Description des données synthétiques

Nous avons divisé chaque base de données en 80% ensemble d'apprentissage et 20% ensemble de test pour construire et exploiter nos modèles par la suite.

1. Analyse exploratoire : ensemble de données : Flame

Cet ensemble de données est partitionné en deux classes. Nous avons affiché la répartition des observations selon leur labels de classe. Nous avons pu constaté que les données Flame ne sont pas linéairement séparable. D'où, les algorithmes de classification qui se base sur la linéarité des données ne sont pas adapté à cet ensemble de données.

le résultats de l'affichage est le suivant :

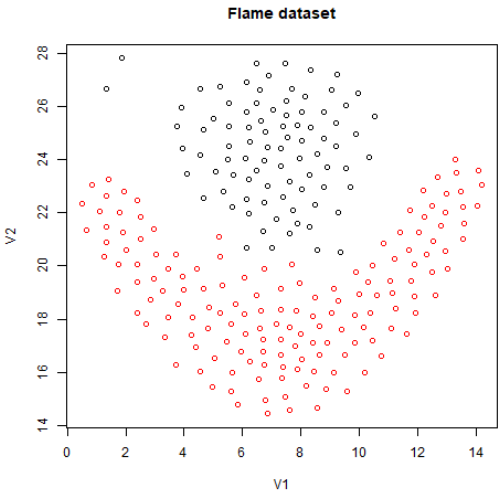

Figure 1: Partitionnement des données Flame

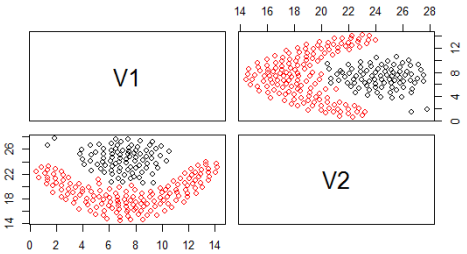

Figure 2: Croisement des variables des données Flame

2. Analyse exploratoire : ensemble de données : Spiral

L'ensemble de données Spiral comporte 312 individus divisé en 3 classes comme nous l'avons mentionné dans le tableau précédent. Les graphes suivants montrent cette classification en 3 classes.

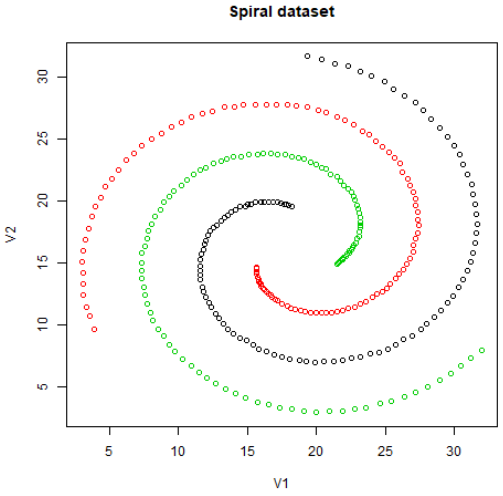

Figure 3: Partitionnement des données Spiral

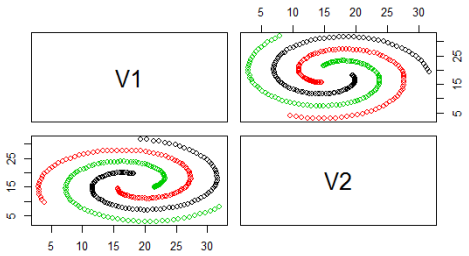

Figure 4: Croisement des variables des données Spiral

Cet ensemble est bien évidemment non linéairement séparable. En effet, les données sont de forme "spirale" comme le nom du data set indique.

3. Analyse exploratoire : ensemble de données : Aggregation

Le dernier ensemble de données synthétique se compose de 788 observations divisé en 7 classes comme nous l'avons mentionné dans le tableau précédent. De la même manière que les data sets précédents, nous avons affiché les données réparties en 7 classes et le croisement des deux variables explicatives. Les données sont non plus linéairement séparable dans ce cas. Les figures suivantes illustrent cela.

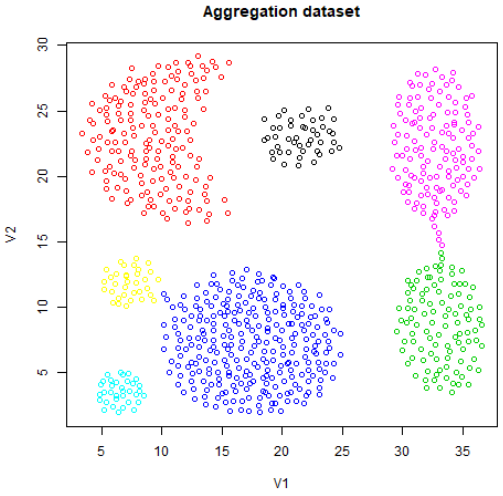

Figure 5: Partitionnement des données Aggregation

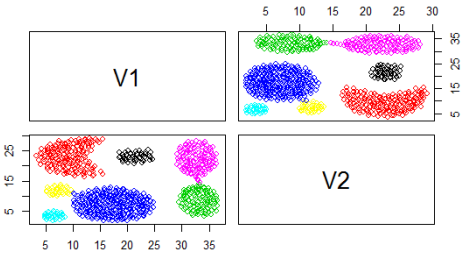

Figure 6: Croisement des variables des données Spiral

4. Résultats et expérimentation

Nous avons appliqué plusieurs techniques d'apprentissage automatique, nous allons présenté dans ce qui suit les résultats obtenus pour chaque méthode et analysé ces résultats.

Le tableau suivant donne les taux d'erreurs obtenus pour chaque data set et pour chaque méthode de classification supervisé :

	Flame	***Spiral***	***Aggregation***
KNN	***0***	***0***	***0***
Linear	***0.0625***	***0.650793650793651***	***0.632911392405063***
Logistic	***0.0625***	***0.0625***	***0.0625***
Bays	***0***	***0.714285714285714***	***0***
Svm	***0***	***0.0793650793650794***	***0***
Random Forest	***0.0208333333333333***	***0.0476190476190476***	***0***

Table 2: Taux d'erreur pour les données synthétiques

Passons maintenant à l'analyse des résultats.

A partir du tableau ci-dessus, nous remarquons ce qui suit :

- L'algorithme de classification KNN est le meilleur classifieur pour tout les ensembles de données avec un taux d'erreur null. En effet, comme cet algorithme se base sur un calcul de distance entre l'ensemble d'apprentissage et l'ensemble de test (prendre les K plus proches voisins selon cette distance puis choisir la classe majoritaire), il ne prend pas en considération la linéarité des données d'où le taux d'erreur obtenu dans ce cas.
- Pour les deux algorithmes régression linéaire et régression logistique, dans le cas de l'ensemble de données Flame, le taux d'erreur est de 0.0625. Les individus étant divisés en deux classes (variable expliquée binaire dans le cas logistique et discrète continue cas linéaire), ce taux d'erreur est beaucoup moins élevé à ceux des deux data set Spiral et Aggregation divisés en 3 et 7 classes respectivement.
- les deux algorithmes : pour le classifier Naïf de Bayes et les SVM non linéaire, le taux d'erreur est égal à 0 pour les data sets Flame et Aggregation alors qu'il dépasse 0.7 pour Spiral.
- le Random Forest est adapté pour tout les data sets ainsi que pour le data set Aggregation.

III. Ensembles de données réels

1. Analyse exploratoire : Données Visa Premier

L'ensemble de données Visa premier contient 1073 observations décrite chacune par 48 variables. Les variables explicatives sont soit quantitatives ou qualitatives. La variable à expliquer est la variable binaire "Possession de la carte visa" représenté par l'attribut cartevp. Ce data set est donc divisé en deux classes.

En analysant le contenu de ce data set, nous avons remarqué les points suivants :

- Il existe des variables avec des valeurs constantes (pour chaque observation, la valeur de cet attribut est la même), ou bien, avec une fréquence très basse (la valeur est apparue une seule fois dans le data set). Ces variables sont : nbimpaye, mteparte et mtbon.
- Les variables departem, sitfamil, codeqlt et nbpaiecb comporte des valeurs manquantes représentées par un point ".".
- L'ensemble de données Visa premier contient des variables dupliquées. Les deux variables cartevpr et cartevp sont les même, de plus c'est la variable expliquée et donc la mettre avec

le ensemble des variables explicatives contredit le principe d'apprentissage automatique. De même, les variables sexer et sexe sont les même.

A partir de ces remarques, nous avons jugé qu'il est inutile de garder les variables avec les valeurs constantes et celles avec des valeurs manquantes d'une part, et éliminé une des variables dupliquée pour ne garder qu'un seul représentant de l'attribut.

Le nombre de variables est donc réduit à 39 variables.

Une fois ces pré-traitements sont achevés, nous avons divisé notre ensemble de données final en un ensemble d'apprentissage et un ensemble de test.

2. Résultats et expérimentation

Pour l'ensemble des méthodes que nous avons implémenté, les résultats obtenus sont résumés dans le tableau suivant :

Méthodes	Naif Bays	SVM linéaire	SVM non linéaire	KNN	Random Forest	ADL
Taux de précision	0.74	0.82	0.82	0.81	0.88	0.5

Table 3: Données Visa Premier

Les résultats obtenus montrent ceux qui suit :

- Nous avons pu utilisé la mesure de précision pour évaluer les méthodes ce-dessus car l'ensemble de données Visa Card n'est pas très déséquilibré.

Classes	Cnon	Coui
Nombre d'instances	741	359

Table 4: Répartition Visa Card

- Le meilleure taux de précision est obtenu en appliquant l'algorithme Random Forest. Ce taux est égal à 0.88 d'où on pourra considéré l'algorithme Random Forest comme un bon classifieur pour le data set Visa Premier.
- les deux versions de l'algorithme SVM ont donné le meme résultats égal à 0.82.

- Les résultats des différents algorithmes KNN, SVM linéaire et SVM non linéaire se rapprochent et dépassent les 80%.
- la méthode ADL a donné un taux égal à 0.5. Cette méthode considéré le problème de classification comme un sous problème dans ce cas, elle pourra être utilisée pour la visualisation des données.
- Nous n'avons pas travaillé avec les fonctions de régression logistique et linéaire car le but principale de ces fonctions est la régression et non pas la classification d'une part, et d'autre part nos variables à expliquer sont de type qualitatives.

IV. Analyse exploratoire : Credit Card Fraud

Cet ensemble de données contient 284 807 observations. Chaque observation représente une transaction, au total nous avons 492 opérations de fraude sur cet ensemble de transactions.

Ce data set est un data set déséquilibré. Le nombre d'instances qui appartiennent à la classe 1 (fraude) est très petit par rapport à celui qui représente les instances de la classe 0 (non fraude).

Le graphe suivant représente cette répartition :

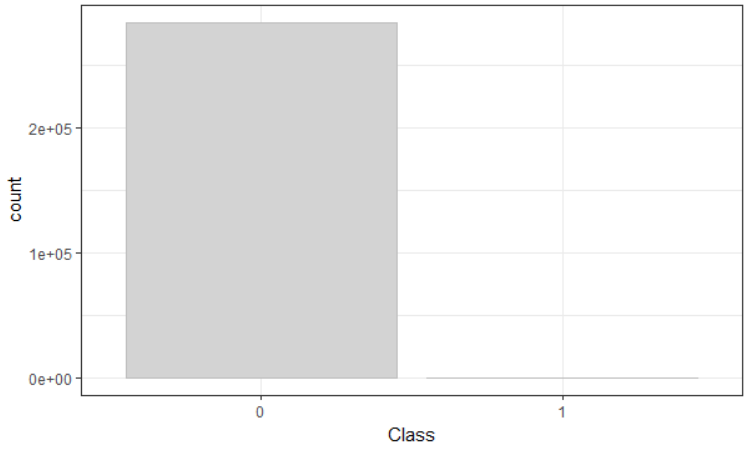

Figure 7: Partitionnement des données Credit Card Fraude

Les résultats obtenus sont apparus dans le tableau suivant :

	KNN	Naïf Bayes	SVM	Random Forest	Kernal SVM
p_value	<2e -16	<2e -16	7.34e-06	0.003	1.917e-05
Kappa	0.105	0.11	0.82	0.86	0.83
Sensibilité	5.691e-02	0.83	0.72	0.80	0.74
Spécificité	1	0.98	0.99	0.99	1

Table 5: Résultats pour l'ensemble de données Credit Card Fraud

Dans ce cas la, le taux de précision n'est pas significatif. En effet, le data set étant non équilibré, on aura toujours un taux de précision élevé sans rien faire.

A partir du tableau ci-dessus, nous avons d'abord pris en compte la valeur de p-value (si c'est supérieur à 0.05 on rejette le résultat). Comme tous les algorithme donne des p-value inférieur à 0.05,nous avons comparé ces algorithme a l'aide des mesures kappa et le couple sensibilité et spécificité.

Nous remarquons que random forest est dominant sur ces mesures et donc est considéré comme le meilleur classifieur pour ce data set aussi.

D'autre part , le classifieur Naïf de Bayes et les SVM donnent de bon résultats sur ces données que ce soit en version simple ou la version kernal.

Nous voulons toujours avoir un compromis entre la sensibilité et la spécificité ce qui est le cas avec cet algorithme.

L'algorithme KNN donne de mauvais résultats nous constatons que la spécificité est égale à 1 , et les autres mesures sont très basses, c'est dû au fait que knn fait un sur-apprentissage sur la classe dominante.

V. Conclusion

Nous avons abordé dans le cadre de ce projet différentes techniques utilisées pour la classification des données. Nous avons les avons appliqué sur des données réelles et des données synthétiques.

Nous avons pu constaté que les résultats obtenus en appliquant une méthode à un autre diffère et cela dépend du data set (le type des données : linéaire, non linéaire, le nombre de classes dans chaque ensemble, etc.

Il est toujours pas évident de classifier des données non équilibrées (dans notre cas détection de fraude). Les taux de précisions obtenus sont souvent non significatifs.

Pour conclure, nous proposons de réaliser une imputation de données Visa premier pour résoudre le problèmes des valeurs manquantes. Pour l'ensemble de données Fraude, attribuer des poids aux classes rare (fraude = 0 par exemple). Le taux de bon classement dans ce cas n'a aucun sens. Les réseaux de neurones peuvent aussi êtres utilisés pour la détection de fraude.