Compte Rendu de la description et justification de la méthodologie d'évaluation pour le dataset shill Bidding.

Réalisé par : Rachida OUCHENE, Lylia TOUAZI, Koussaila HAMMOUCHE.
GROUPE 1

I.Introduction

Le dataset que nous utilisons est extrait de ce site https://archive.ics.uci.edu/ml/datasets/Shill+BiddingDataset. Le jeu de données a été récolté par Ahmad Alzahrani et Samira Sadaoui des enchères eBay sur le produit iphone7. L'objectif de ce dataset est de développer des modèles de détection et de classification. Dans notre première partie nous avons décrit et analysé le jeu de données, dans cette partie nous allons décrire et justifier les défirent méthodes d'apprentissage automatique a utiliser afin de bien séparer les enchères frauduleuse de celle normal.

II.Méthode:

Dans cette section nous allons présenter les trois méthodes que nous avons choisies pour évaluer notre modèle. Nous avons fixé quelques critères de sélection pour arriver à ce choix. Parmi ces critères nous avons :

- La quantité de données que nous possédons(nous avons 6321 individue)
- La structure de nos données (nous avons des données structurées).
- Normalité des données.(comme nos données ne suivent pas des lois de probabilités normales, donc pour cela nous avons opté pour des modèles non paramétriques).
- Type de problème à traiter : problème de classification.

D'après ces critères et le faite que nous données ne sont pas linéairement séparables nous avons optés pour ces trois modèles:

❖ **k plus proches voisins(k-Nearest Neighbours) :** C'est une méthode d'apprentissage supervisé son fonctionnement peut être assimilé à l'analogie suivante *"dis moi qui sont tes voisins, je te dirais qui tu es"*.parmis ces avantages il exiger aucune hypothèse sur les données et c'est un algorithme simple et facile à mettre en œuvre. Cet algorithme retient en mémoire la totalité de datasets que en lui fournie pour ensuite calculer ces prédictions on calculant la distance entre le nouveau point et les K instances de datasets les plus proches de cette observation. Pour cela Il existe plusieurs fonctions de calcul de distance, notamment, la distance euclidienne, la distance de Manhattan, … etc,comme la plupart de nos données que nous manipule sont quantitatives et du même type (float), la distance

1. euclidienne est un bon candidat dans notre cas qui se calcule de la façon suivante:

$$D_e(x,y) = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}$$

aussi nous devons choisir la valeur de l'hyper-paramètre K (qui correspond au nombre de voisins plus proche).K-NN

❖ **Arbre de décision :**

C'est une méthode d'apprentissage supervisé,nous l'avons choisie parce que il est simple à comprendre et à traiter.l'arbre de décision nous permet *de représenter un ensemble de choix* sous la forme graphique d'un arbre.Chaque nœud intermédiaire réalise un test portant sur une variable dont le résultat indique la branche à suivre dans l'arbre. Pour classer un nouveau cas en suivant le chemin partant de la racine (nœud initial) à une feuille de l'arbre(qui spécifie sa classe) en effectuant les différents tests à chaque nœud. Tel que nous pouvons sélectionner les meilleurs attributs sur lesquels vont porter les tests des nœuds de l'arbre, en plus de ça nous pouvons aussi choisir la hauteur de l'arbre. Les arbres de décision ont pour avantage d'être simple à interpréter, très rapide à entraîner, d'être non paramétrique, et de nécessiter très peu de prétraitement des données aussi il peut résoudre des problème linéaire ou non linéaire comme dans notre cas.

❖ **Support vector machine(SVM) :**

C'est un modèle d'apprentissage supervisé qui analyse les données à des fins de classification. Son but est de trouver un séparateur entre deux classes qui sont au maximum éloignées de n'importe quel point des données d'entraînement.Nous devons bien régler les hyper paramètres suivants: kernel, gamma, et C.Et comme nos données ne sont pas linéaire nous devons mettre kernel = `'RBF'`,`aussi` plus gamma est petit , plus le sigma de la fonction de noyau est grand et moins le classificateur est sensible à la distance entre les points individuels. Nous avons choisi cette méthode car SVM fonctionne bien sur les petits datasets ce qui va nous permettre une grande précision de prédiction.

III.Protocole :

❖ Découpage DataSet en training testing set.

Étant donné une tâche d'apprentissage supervisée, le but est donc d'estimer plusieurs modèles afin de prédire au mieux la variable cible, qui est si une enchère est fraudée ou non. Pour sélectionner le modèle, il faut d'abord mélanger le dataset, ensuite les diviser en deux ensembles de données, comme illustré dans la figure 1: le training set où données sont utilisées pour entraîner et apprendre le modèle, et le testing set qui est réservé pour tester et évaluer le modèle sur des valeurs qui n'ont jamais vu.

Figure 1 : Découpage de datasets en deux parties.

❖ Stratified-K-Fold Cross Validation :

Ajouter un validation set permet de chercher les réglages et les hyper-paramètres du modèle qui donne les meilleures performances tout en gardant les données du testing set. donc ici on aura (training set, validation set, testing set). Le problème maintenant c'est qu'on est pas sur que cette découpe soit la bonne ou non, et pour résoudre ça on va utiliser la technique de validation croisée qui consiste à entraîner puis valider notre modèle sur plusieurs découpes possible du training set. Du coup, lors de l'évaluation de plusieurs modèles on sera sûr d'avoir choisi le modèle qui a en moyen eu une meilleure performance. Mais le problème de la validation croisée c'est qu'on est pas sur que toutes les découpes ou les splits sont équilibrés entre les deux classes ("1", "2").

La méthode validation croisée stratifiée à été choisie pour éviter d'introduire des biais et d'avoir un déséquilibre entre les deux classes dans chaque découpe.

En effet le training set est divisé en k splits dans lesquels à chaque split on aura une petite portion de chacune de nos deux classes. comme illustré sur la figure 2. Donc au final on aura un bon équilibre entre les classes dans les différents splits.

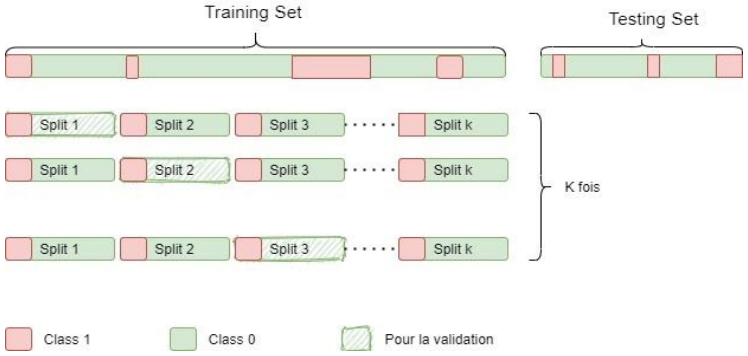

Figure 2: illustration de validation croisée stratifiée.

❖ Mesure d'évaluation:

Comme on a uniquement deux classes Y = {1, 2}, beaucoup de mesures de performance sont décrites par le biais du tableau de contingence suivant appelé matrice de confusion qui est présenté dans le figure 3.

Notre objectif est de prédire si une enchères est une arnaque ou non, donc si on choisit la mesure "précision" seule, qui permet de réduire au maximum le taux de faux positif et donc d'éviter de dire c'est une arnaque alors que c'est faux, on peut avoir facilement une bonne précision en faisant très peu de prédiction positives.

C'est la même chose si on veut choisir le recall seul, on peut facilement avoir un bon recall en prédisant tous positifs.

donc on doit utiliser les deux en même temps, et pour les résumer en un seul nombre on calcule la fonction F1 qui est la moyenne harmonique de la précision et du rappel, qui est défini comme suite :

$$F1 = 2 * \frac{Précision * recall}{précision + recall}$$

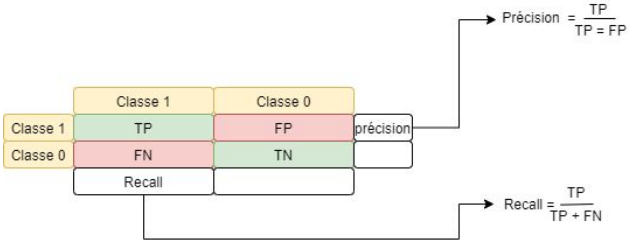

Figure 3 : matrice de confusion

.

IV. Conclusion

Notre but est de choisir le modèle d'apprentissage supervisé performant qui prédira si une enchère est fraudée ou non, donc pour cela nous avons déjà analysé les données. Dans cette partie nous avons proposé trois modèles de classification, les k plus proches voisins KNN, support vector machine SVM et l'arbre de décision, et nous avons expliqué pourquoi avoir fait ces choix. Nous avons aussi décrit le protocole que nous allons mettre en œuvre dans la prochaine partie, pour nous permettre de bien mesurer et comparer les performances entre ces trois modèles.