Conserver la meilleure précision en réduisant le nombre de dimension

Ambre Collard[1] Clément Hibon[1]
Kerryghan Relot[1] Xavier Coupé[1]
(1) Le Mans Université, LIUM, 72000 Le Mans, France
{Ambre.Collard.Etu, Clement.Hibon.Etu, Kerryghan.Relot.Etu,
Xavier.Coupe.Etu}@univ-lemans.fr

RÉSUMÉ

Lors du traitement de données liées à la sécurité routière, nombre d'analyse sont possible: classement de dangerosité des routes, analyse de la variation de l'accidentogénité au fil du temps ou encore évaluation de la gravité. C'est sur ce dernier point que ce concentre cet article. Ce travail présente la mise en oeuvre de l'optimisation de nos données dans le corpus sur la sécurité routière. Nous avons pour tâche préliminaire de faire un prétraitement des données afin de les uniformiser. Nous avons ensuite réduit le nombre de dimension de nos données en sélectionnant de manière adéquate les caractéristiques ainsi qu'en effectuant une analyse en composante principale. Pour évaluer et quantifier le gain ou la perte d'information, nous avons utilisé le modèle Random Forest tout au long de nos mesures, ceci afin d'avoir une base commune de comparaison sans introduire de variabilité dûe au modèle. Cet article vise à rapporter l'impact des modifications effectuées sur les données sur la précision de notre modèle de prédiction.

Résultats préliminaire

ABSTRACT

Conservation of the best accuracy while reducing number of dimensions.
TODO : translate and write the summary here.

MOTS-CLÉS : Accident, Données, Gravité, Descripteur, Classes, Accident de la route, Prétraitement, ACP, Random Forest

KEYWORDS : Road accident, Gravity, Features, Data, Preprocessing, PCA, Random Forest

1 Introduction

Lors d'un accident corporel beaucoup d'informations sont collectés par les forces de l'ordre pour le caractériser. Un accident corporel est un accident survenu sur une voie ouverte à la circulation publique, impliquant au moins un véhicule et ayant fait au moins une victime ayant nécessité des soins.

Notre questionnement original se portait sur le niveau d'importance de chaque informations dans la caractérisation de la gravité de l'accident. Autrement dis, nous avons tenté de rechercher quels pouvaient être les éléments prédominants – si ils existent – pour déterminer la gravité d'un accident corporel.

Notre travail se portant en priorité sur le prétraitement des données, nous avons donc décidé de ne choisir qu'un seul modèle pour l'entraînement ainsi que l'évaluation, et ceci afin de pouvoir facilement comparer nos résultats. Le modèle que nous avons choisi est RandomForest. Concernant le prétraitement plus particulièrement, nous avons mis en place plusieurs méthodes : Retypage et Nettoyage de données (data cleaning), Sélection de caractéristiques (features selection), Analyse en Composante Principale (PCA)

Après avoir présenté l'état de l'art de le domaine, nous présenterons comment et pourquoi nous avons nettoyé nos données. Ensuite, nous décrirons nos différentes démarches et approches pour la sélection de caractéristiques, avant de finir en explicitant la réalisation de notre analyse en composante principale et comment elle nous a permis de réduire la dimensionnalité de nos données.

2 Travaux connexe

Dans cet article[i], on trouve que RandomForest est le meilleur modèle pour prédire la gravité d'un accident.
Bien qu'il soit allé plus loins que nous dans sa démarche, l'article de recherche par Miaomiao et Yindong[ii] présente une problématique similaire à la notre, ce qui montre que notre approche n'est pas dénuée de sens non plus.
La rédaction de cette partie est à améliorer.

3 Méthodologie

3.1 Données Utilisées

Pour cette étude nous nous sommes basées sur les données d'accidents corporels mis à disposition par le Ministère de l'Intérieur[iii]. Plus particulièrement nous nous sommes concentré sur les donnée pour l'année 2022 afin de ne pas nous éparpiller et d'avoir une analyses basée sur des données récentes. Au moment de nos premières recherches, les données les plus récentes étaient les données pour l'année 2022 ; aussi, depuis 2019 la catégorisation des blessés à changé dans le jeu de données, c'est une raison de plus qui nous a convaincu dans le choix de données récentes, afin de pouvoir nous baser sur des caractéristiques qui ne changent pas selon les années.

3.2 Nettoyage de données

On a d'abord mis toutes les données en entiers 8bits afin d'économiser à la fois de l'espace mémoire et de la puissance de calcul.

[ajouter un peu de précision]

3.3 Sélection de caractéristiques

Afin de sélectionner les caractéristiques à garder nous avons tenté plusieurs approches.

[décrire davantage ce que nous avons fait et pourquoi]

3.4 Analyses en Composante Principale

Enfin, nous avons aussi réaliser une Analyse en Composante Principale, avec des données centrées réduites et des données non-centrées réduites pour comparé l'impact de la normalisation de nos données. Pour ce faire nous avons utilisé le module '*StandardScaler*'.

4 Résultats / Discussion

Lorem Ipsum

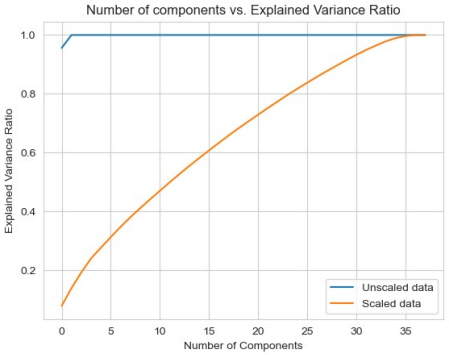

Figure 1: Comparaison de l'information cumulée porté par les n premiers axes

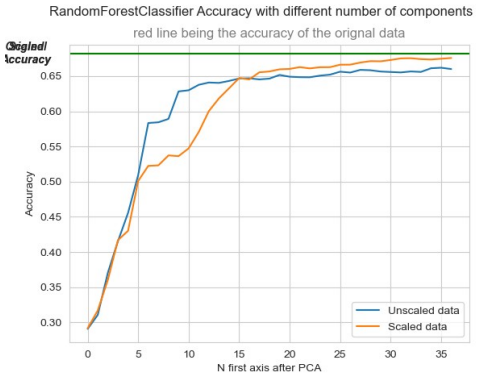

Figure 2: Comparaison de la précision en fonction du nombre de composant choisi

Lorem Ipsum

5 Conclusion

Pour conclure cet article, nous avons pu observer que…

Références

[i] SANTOS K., DIAS J. P., AMADO C. (2021). A literature review of machine learning algorithms for crash injury severity prediction. Elsevier, 254-269. DOI : 10.1016/j.jsr.2021.12.007

[ii] MIAOMIAO Y. & YINDONG S. (2022). Traffic Accident Severity Prediction Based on Random Forest. Sustainability, 14, 1729. DOI : 10.3390/su14031729

[iii] Ministère de l'Intérieur (màj 2024). Bases de données annuelles des accidents corporels de la circulation routière - Années de 2005 à 2023. data.gouv.fr