Aix Marseille Université

2024-2025
Master 2 Informatique (SID)

Classification de la gravité des accidents de la route

Etudiantes :
Ikram BATTAL
Doha BENHABBACH
Sara DAOUIRI
Chaimae ECH-CHARFI

Enseignante :
Feda Almuhisen

1 Introduction

Les accidents de la route causent chaque année 1,35 million de décès et des millions de blessures graves dans le monde, selon l'OMS. Comprendre les facteurs influençant leur gravité est essentiel pour limiter ces impacts. Ce projet vise à développer un modèle prédictif classant la gravité des accidents en trois catégories : léger, grave et mortel. Le jeu de données utilisé comprend 12 316 accidents et 32 caractéristiques collectées entre 2017 et 2020 par les départements de police d'Addis-Abeba, couvrant des informations sur les conducteurs, les véhicules et les conditions environnementales.

2 Exploration des Données

2.1 Chargement et Exploration du Jeu de Données

L'analyse initiale des données met en évidence des relations significatives entre certaines caractéristiques. Par exemple, `Casualty_class` est fortement corrélé à `Casualty_severity` (0.77) et à `Age_band_of_casualty` (0.75), ce qui suggère que l'âge et la classe des victimes jouent un rôle clé dans la gravité des accidents. En revanche, des variables comme `Time` ou `Day_of_week` présentent des corrélations faibles avec les autres caractéristiques, indiquant qu'elles ont probablement un impact limité sur la gravité des accidents. Ces observations fournissent une base solide pour identifier les caractéristiques pertinentes pour la modélisation.

2.2 Analyse des Caractéristiques

L'étude des relations linéaires entre les caractéristiques révèle des tendances intéressantes. Par exemple, `Sex_of_casualty` et `Casualty_class` affichent une forte corrélation positive (0.68), tandis que `Pedestrian_movement` montre une faible corrélation négative (-0.08) avec `Casualty_class`. Ces résultats suggèrent que certaines variables, comme `Age_band_of_casualty` et `Number_of_casualties`, pourraient avoir un rôle prépondérant dans la gravité des accidents, tandis que d'autres, comme `Day_of_week`, semblent avoir une influence moindre.

3 Prétraitement des Données

3.1 Standardisation

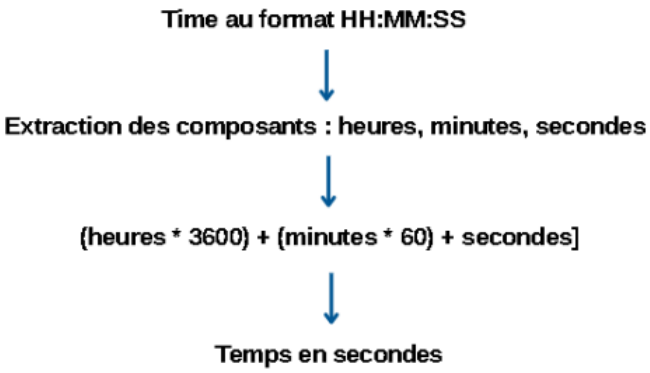
La standardisation permet de garantir que les différentes variables ont une échelle de valeurs cohérente. Par exemple, des valeurs comme `"other"` et `"Other"` peuvent être standardisées en `"order"` pour assurer une cohérence dans les catégories.

3.2 Traitement des valeurs manquantes

Le pourcentage des valeurs nulles est élevé (plus de 35%) comme `Service_year_of_vehicle` et `Age_band_of_casuality`. Cela peut indiquer des informations manquantes pour ces variables, ce qui peut nuire à la qualité du modèle de prédiction. Il est donc préférable de les supprimer.

3.3 Encodage de la colonne `Time`

La colonne `Time` doit être encodée pour être utilisée efficacement dans les modèles. La méthode d'encodage n'est pas détaillée ici.



3.4 Encodage des variables catégoriques

3.4.1 Variables ordinales

Les variables ordinales ont un ordre ou un classement entre les différentes catégories. Elles doivent être encodées avec des valeurs numériques représentant leur ordre. Par exemple : `Age_band_of_driver` : `'Under 18'` = 1, `'18-30'` = 2, `'31-50'` = 3, `'Over 51'` = 4.

3.4.2 Variables non ordinales

Les variables non ordinales n'ont pas d'ordre particulier entre les catégories. Par exemple, `Day_of_week` peut être transformé en 7 nouvelles variables numériques : `"Monday"` : 0, `"Tuesday"` : 1, ..., `"Sunday"` : 7.

4 Optimisation avec AutoML

4.1 Présentation d'AutoML

4.1.1 Concepts

AutoML (Automated Machine Learning) désigne un ensemble de techniques permettant d'automatiser le processus de modélisation des données..

4.1.2 Outils AutoML utilisés

L'outil choisi pour ce projet est **PyCaret**, un framework Python open-source conçu pour simplifier l'application des algorithmes de machine learning.

4.2 Application sur les données d'accidents

4.2.1 Résultats obtenus avec AutoML

Après avoir configuré le pipeline avec `setup()`, les résultats montrent que : Les données initiales contiennent 24 variables numériques et un taux de valeurs manquantes de 8.2%, gérées automatiquement.

Les modèles testés avec `compare_models()` sont :

— **LightGBM (Light Gradient Boosting Machine)** : Meilleur modèle avec une précision (Accuracy) de 85.30%.
— **XGBoost** : Légèrement inférieur avec une précision de 85.28%.
— **Random Forest** et **Extra Trees** : Performances proches avec une précision d'environ 85.12%.

4.2.2 Visualisations clés générées avec AutoML

— **Confusion Matrix** : Montre la distribution des prédictions correctes et erronées pour chaque classe (slight, serious, fatal).
— **AUC Plot** : Évalue la capacité des modèles à distinguer entre les classes.
— **Feature Importance** : Identifie les variables les plus influentes pour les décisions du modèle, offrant un aperçu des facteurs critiques des accidents.

Modélisation

1.1 Optimisation des Hyperparamètres

Dans cette étude, plusieurs algorithmes de classification ont été testés, accompagnés d'une recherche d'hyperparamètres pour optimiser leurs performances. `GridSearchCV` a permis de tester systématiquement diverses combinaisons d'hyperparamètres clés pour chaque modèle. Par exemple, pour *Random Forest*, des hyperparamètres tels que le nombre d'arbres (`n_estimators`), la profondeur maximale des arbres (`max_depth`) et le nombre minimal d'échantillons pour diviser un nœud (`min_samples_split`) ont été explorés. Pour les modèles de gradient boosting comme *XGBoost* et *LightGBM*, des hyperparamètres comme le taux d'apprentissage (`learning_rate`), la profondeur des arbres et la fraction d'échantillonnage (`subsample`) ont été ajustés pour maximiser la précision tout en évitant le surapprentissage.

Évaluation des Modèles

Ce tableau résume les performances des algorithmes testés et montre que *LightGBM* a obtenu une précision (*accuracy*) légèrement supérieure par rapport aux autres modèles :

Modèle	F1-Score	Accuracy	Recall	Precision
Random Forest	0,80	0,85	0,85	0,86
XGBoost	0,80	0,85	0,85	0,84
LightGBM	0,80	0,86	0,86	0,85

Table 1 – Performances des Algorithmes de Classification

Courbes ROC AUC

Ces courbes ROC AUC résument la capacité des algorithmes à distinguer entre les différentes classes de la variable cible. Par exemple, dans le cas de *XGBoost*, les AUC pour les classes 0, 1 et 2 sont respectivement de 0,86, 0,68 et 0,69, tandis que pour *LightGBM*, les AUC pour les classes 0, 1 et 2 sont de 0,81, 0,70 et 0,70. Ces résultats illustrent les performances relatives des différents algorithmes dans ce contexte multiclasses.

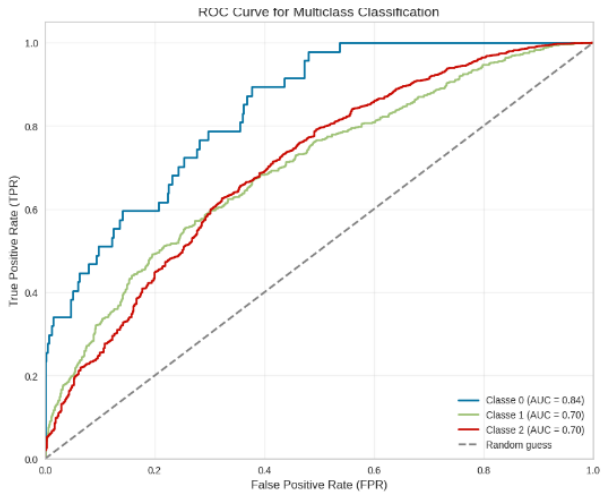


FIGURE 1 – Courbe ROC AUC pour XGBoost

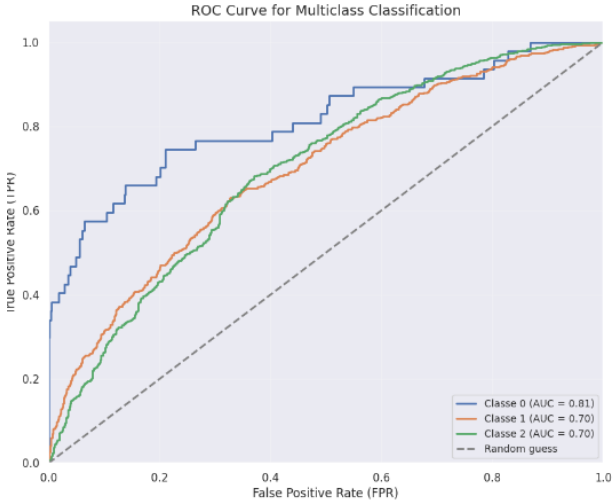


FIGURE 2 – Courbe ROC AUC pour LightGBM

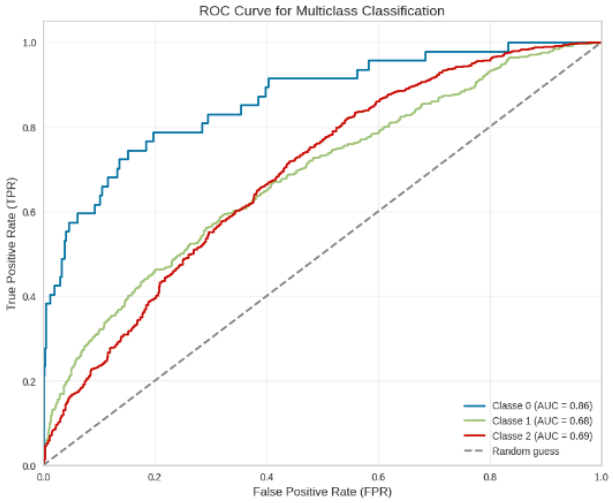
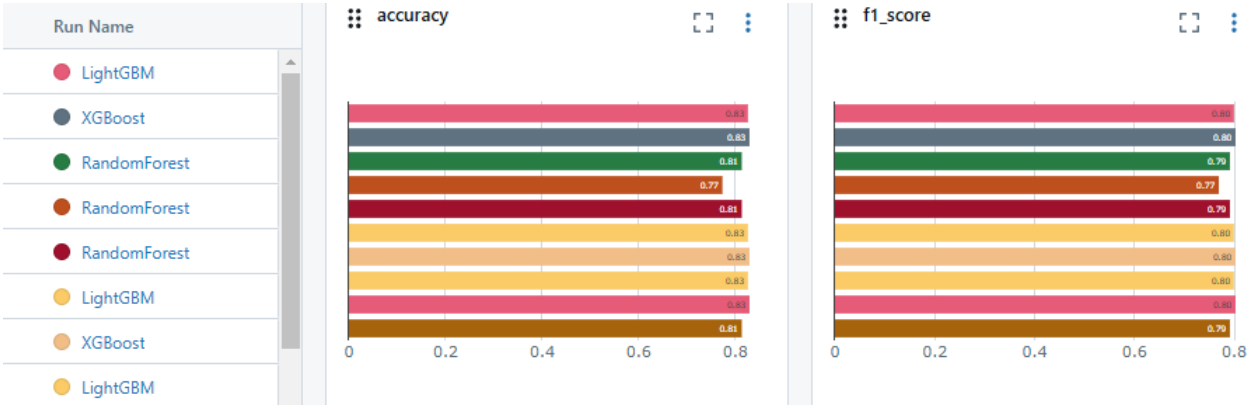


FIGURE 3 – Courbe ROC AUC pour Random Forest

5 Expérimentation avec MLOps

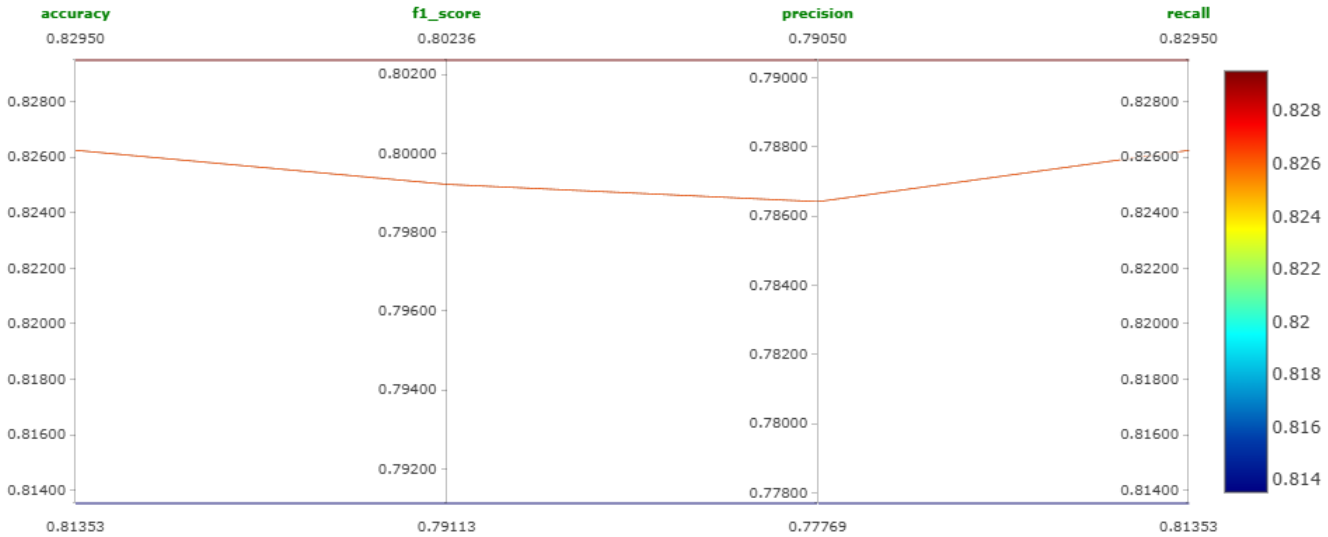
5.1 Suivi des expérimentations avec MLflow

L'intégration de MLflow, un outil de gestion du cycle de vie de l'apprentissage automatique, a permis de consigner et de suivre les performances des modèles de classification développés, notamment Random Forest, XGBoost et LightGBM. Au cours des exécutions avec MLflow, les hyperparamètres de chaque modèle ont été enregistrés, fournissant une trace détaillée des configurations utilisées lors de l'entraînement. Les performances des modèles ont ensuite été évaluées sur les ensembles de test et de validation à l'aide des métriques de précision (*accuracy*), du score F1 (*F1-score*).



5.2 Comparaison des runs

Les différentes exécutions (*runs*) enregistrées dans MLflow ont ensuite été comparées afin d'identifier les modèles et configurations les plus efficaces.



XGBoost
LightGBM
Random Forest

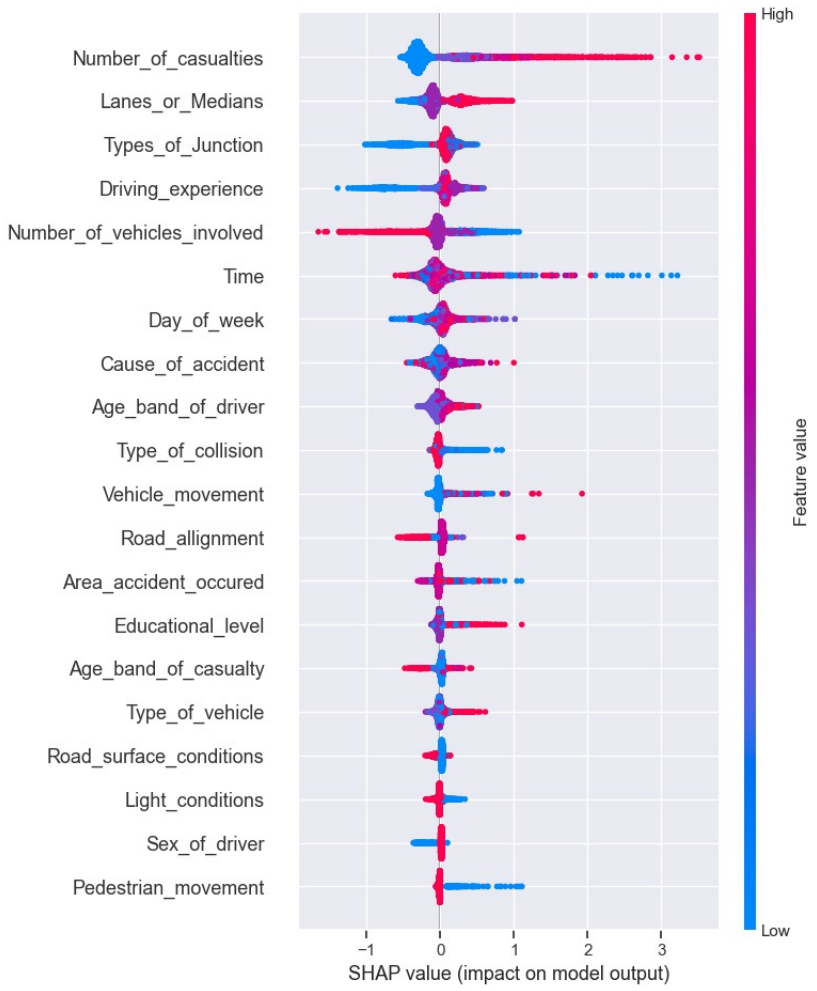
Dans ce graphique, les modèles XGBoost et LightGBM présentent des performances très proches, avec des scores d'accuracy similaires, autour de 82,4 %. Ces résultats suggèrent que les deux modèles sont efficaces pour classer correctement les échantillons.

En ce qui concerne les autres métriques, telles que la précision et le rappel, XGBoost et LightGBM continuent de montrer des résultats compétitifs, se positionnant de manière quasi identique. Cela reflète une performance robuste dans la capacité à bien détecter les classes positives tout en maintenant un bon équilibre global entre les différentes catégories.

En revanche, le modèle RandomForest montre des résultats légèrement inférieurs par rapport à XGBoost et LightGBM, tant en accuracy qu'en précision et rappel, ce qui indique qu'il est légèrement moins performant pour cette tâche de classification.

En résumé, XGBoost et LightGBM se révèlent être des modèles très compétitifs et performants, avec des résultats similaires, ce qui laisse penser que le choix entre les deux peut dépendre de critères spécifiques à l'application, tels que la vitesse d'entraînement ou des considérations liées aux ressources.

5.3 Explication des modèles avec SHAP



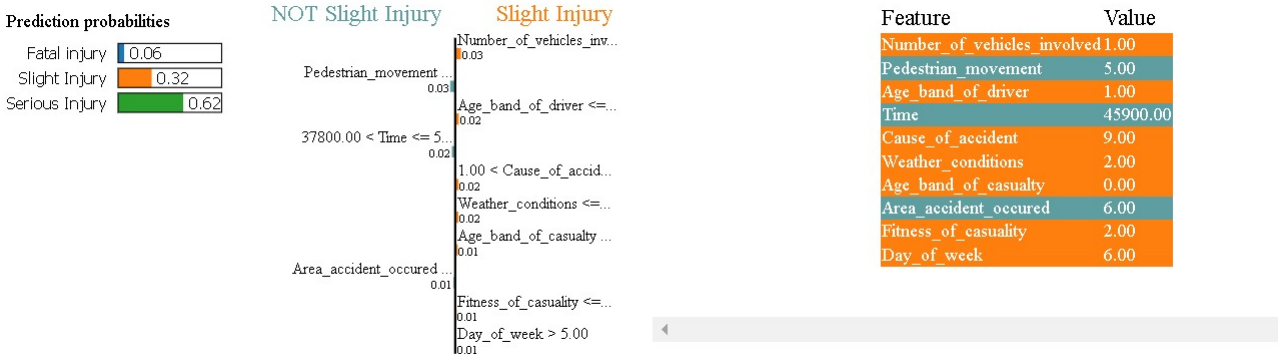
Le graphique de résumé des valeurs SHAP illustre l'impact global des caractéristiques sur l'ensemble des prédictions. Chaque point sur ce graphique représente une observation spécifique, et la position horizontale indique l'importance de la caractéristique (valeurs SHAP).

Les caractéristiques sont classées par ordre d'importance globale, celles en haut ayant le plus grand impact. La couleur des points reflète les valeurs des caractéristiques :

— Rouge : valeurs élevées de la caractéristique
— Bleu : valeurs faibles de la caractéristique

Par exemple, si la caractéristique `Number_of_casualties` se trouve en haut avec de nombreux points rouges à droite, cela indique que des valeurs élevées de cette caractéristique augmentent la probabilité d'accidents graves.

5.4 Explication des modèles avec LIME



Le modèle prédit une probabilité élevée de `Serious Injury` :

- **Serious Injury** : 62% (dominant, barre verte),
- **Slight Injury** : 32% (barre orange),
- **Fatal Injury** : 6% (faible probabilité).

Les variables les plus significatives pour cette prédiction incluent :

- `Number_of_vehicles_involved`,
- `Age_band_of_driver`,
- `Cause_of_accident`,
- `Conditions météorologiques et Mouvement des piétons`.

Les variables les moins influentes pour cette classe, ou dont l'impact est faible ou opposé, sont :

- `Time`,
- `Pedestrian mouvement`,
- `Area accident occured`.