Aix*Marseille université

Classification de la gravité des accidents de la route

Présenté par :

BATTAL IKRAM

DAOUIRI SARA

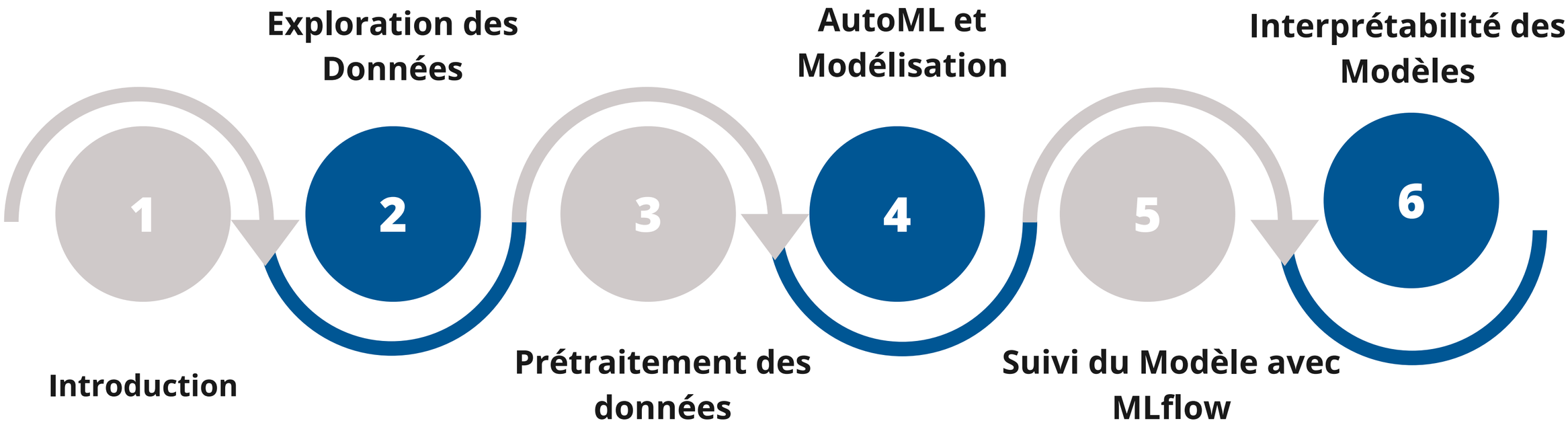
BENHABBACH DOHA

ECH-CHARFI CHAIMAE

Encadré par :

Feda Almuhisen

Plan


1
Introduction
2
Exploration des Données
3
Prétraitement des données
4
AutoML et Modélisation
5
Suivi du Modèle avec MLflow
6
Interprétabilité des Modèles

01

Introduction

01 Introduction

Contexte du Projet

Problème

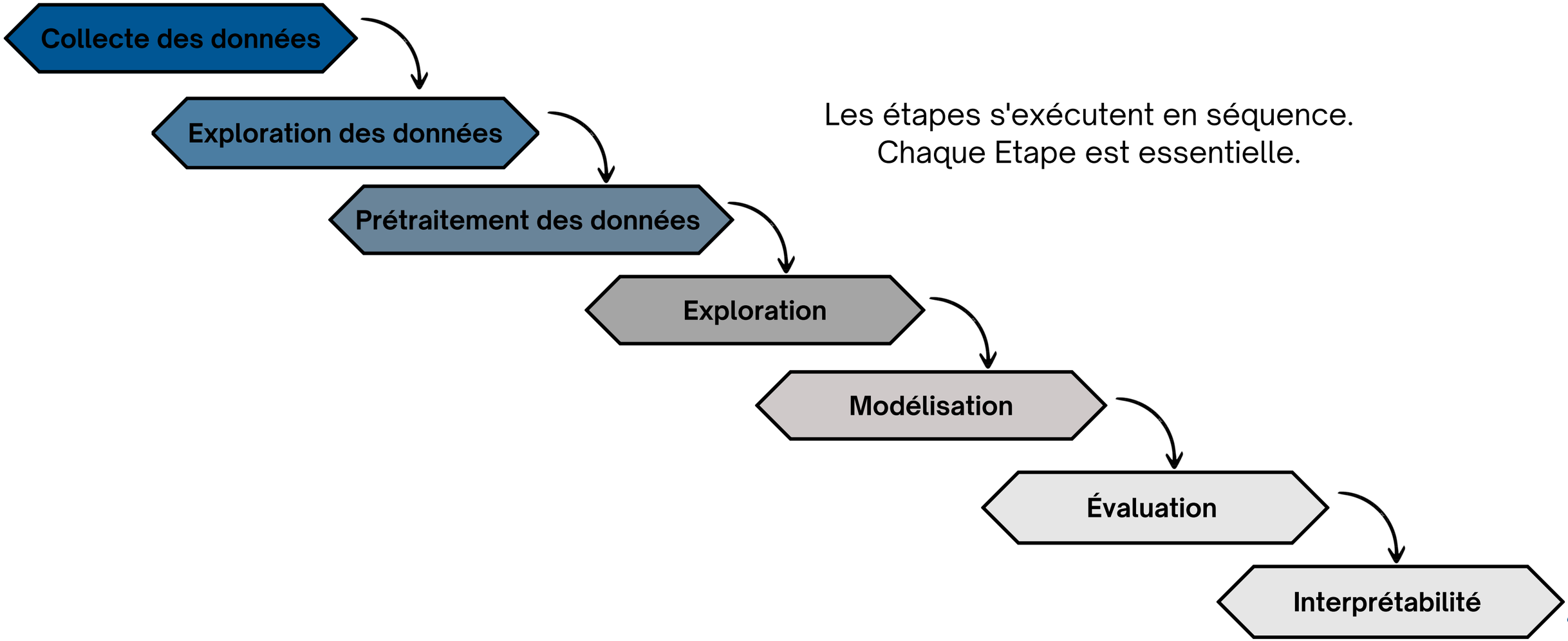
Source des données

Objectif du projet

Méthodologie

01 Introduction

Méthodologie du Projet



Les étapes s'exécutent en séquence.
Chaque Etape est essentielle.

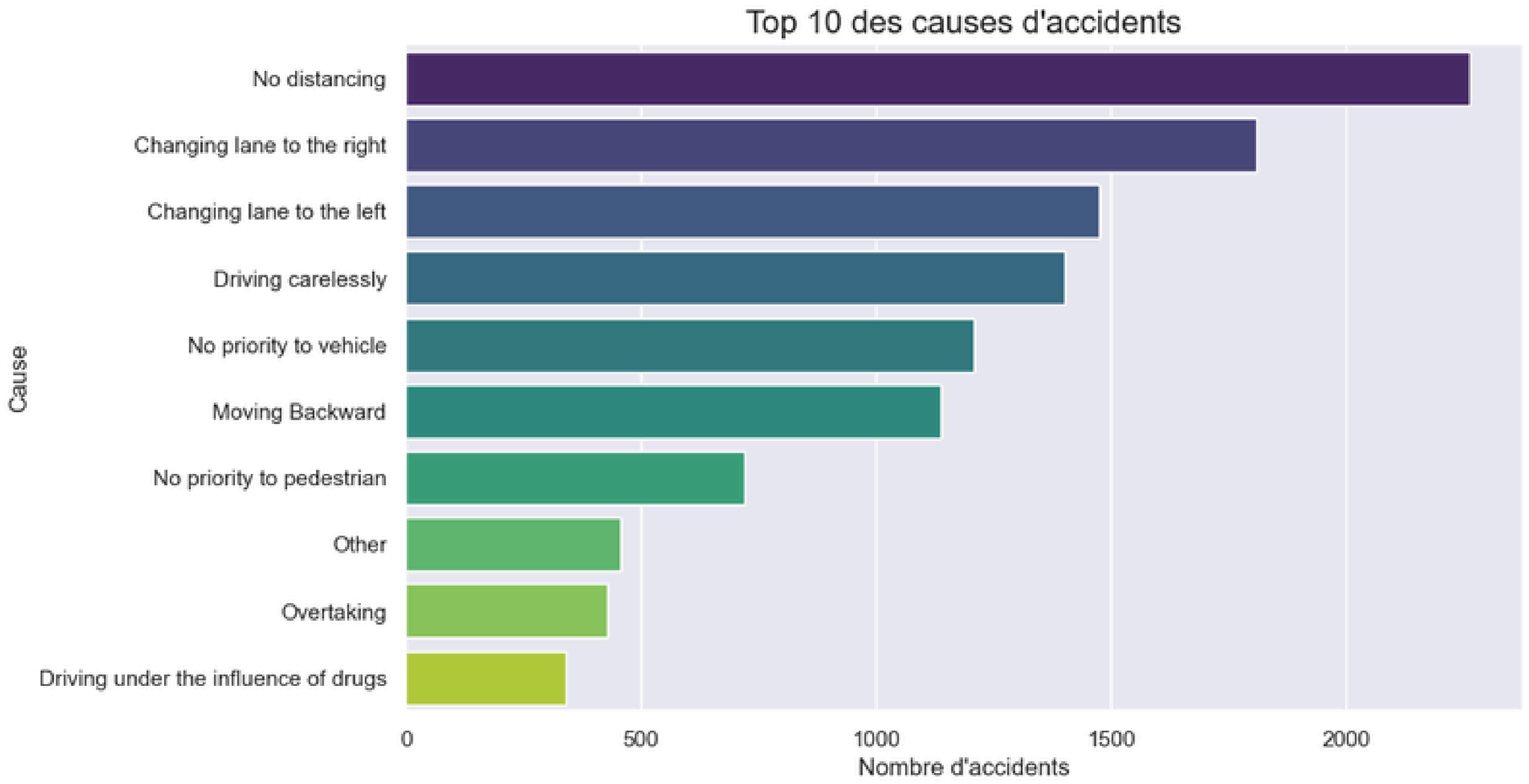
02 Exploration des Données

02 Exploration des Données

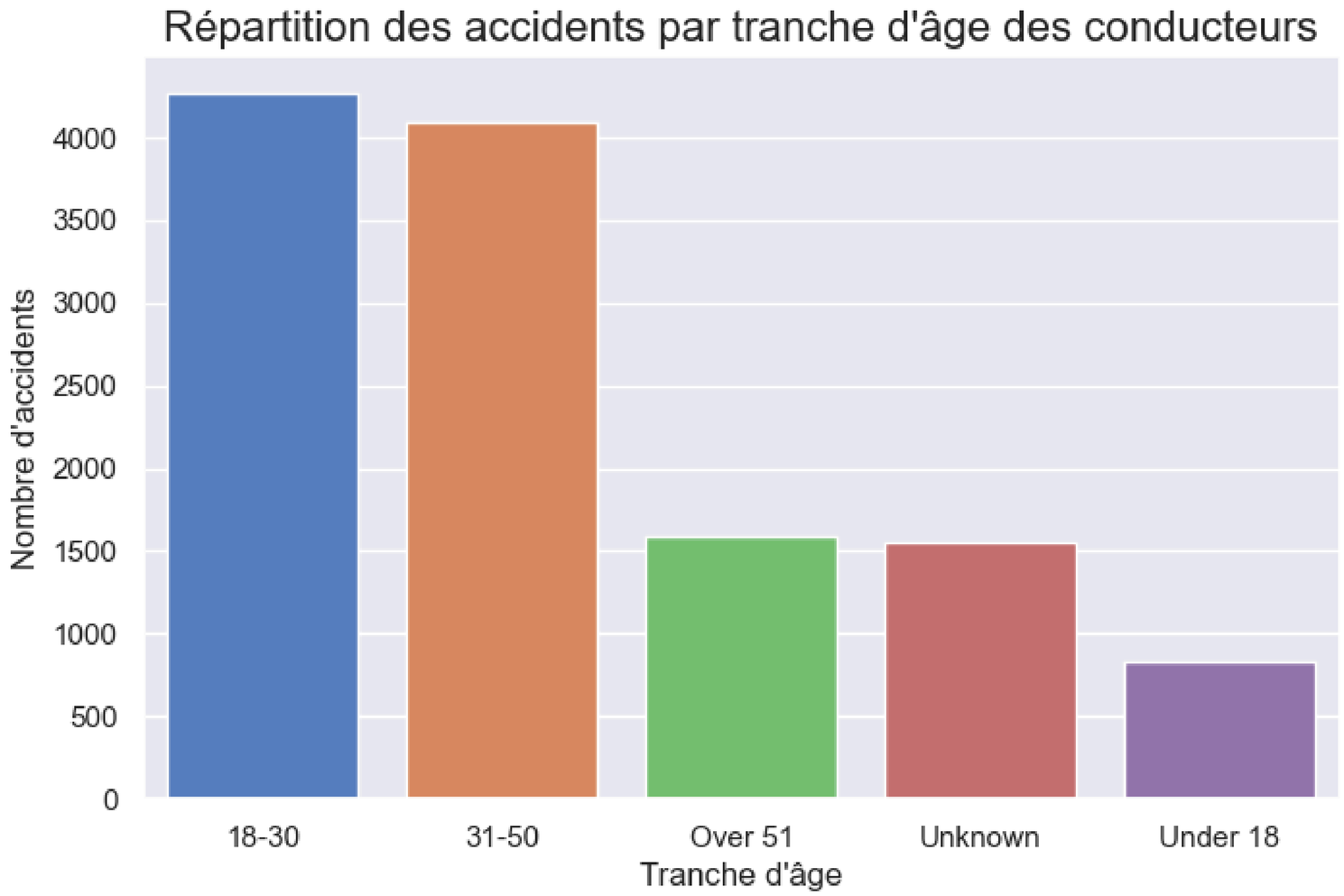
Les causes des accidents

L'âge des conducteurs

BoxPlot

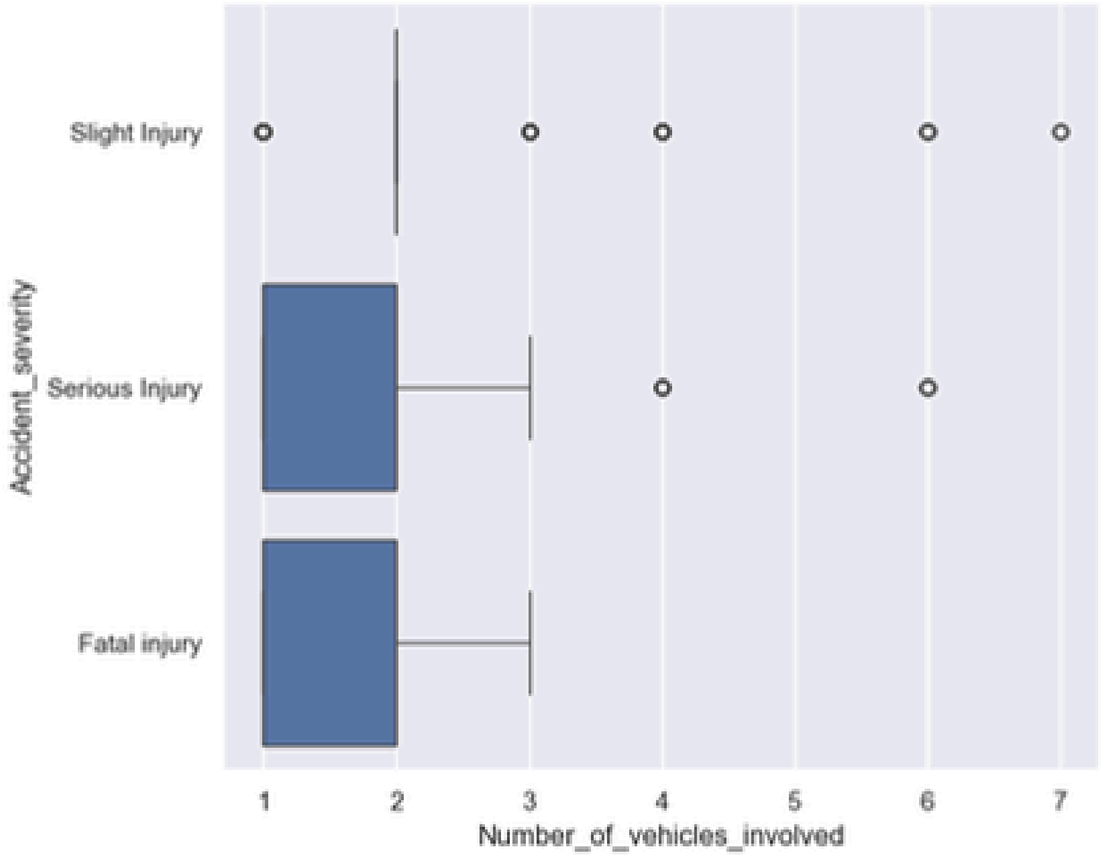


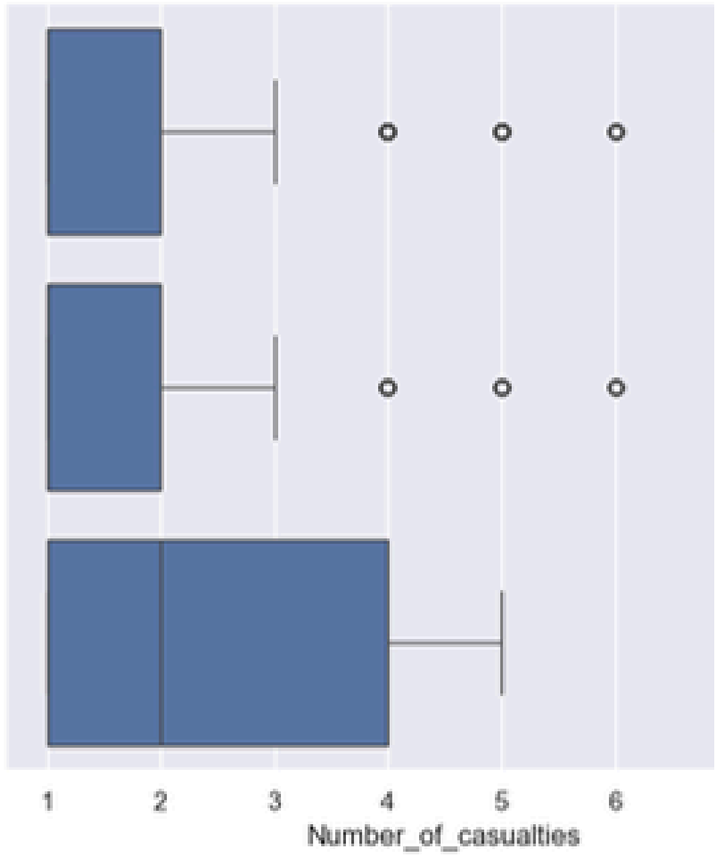
02 Exploration des Données



02 Exploration des Données

Les causes des accidents | L'âge des conducteurs | Boxplot





03 Prétraitement des données

03 Prétraitement des données

Standardisation | Valeurs manquantes | Encodage | Matrice de corrélation

Données → **Définition du format standard** → **Application des transformations**

```
[7]  data['Driving_experience'].unique()
```

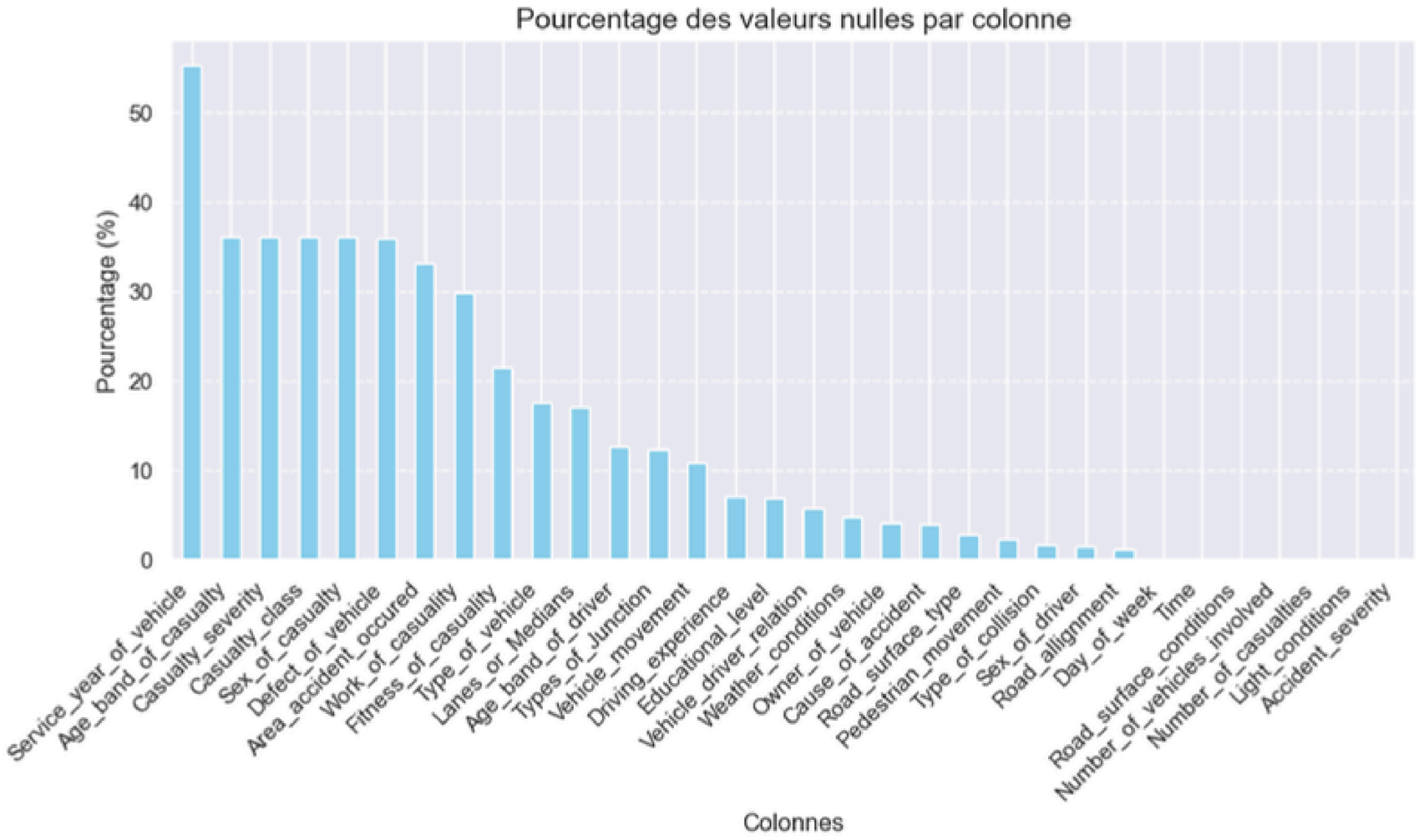
```
array(['1-2yr', 'Above 10yr', '5-10yr', '2-5yr', nan, 'No Licence',
       'Below 1yr', 'unknown'], dtype=object)
```

```
data['Work_of_casuality'].unique()
```

```
array([nan, 'Driver', 'Other', 'Unemployed', 'Employee', 'Self-employed',
       'Student', 'Unknown'], dtype=object)
```

03 Prétraitement des données

Standardisation | **Valeurs manquantes** | Encodage | Matrice de corrélation



03 Prétraitement des données

Standardisation | Valeurs manquantes | **Encodage** | Matrice de corrélation

Colonne Time | Colonne Ordinale | Colonne Non Ordinale

Time au format HH:MM:SS

↓

Extraction des composants : heures, minutes, secondes

↓

(heures * 3600) + (minutes * 60) + secondes]

↓

Temps en secondes

	Time
0	61320
1	61320
2	61320
3	3960
4	3960
...	...
12311	58500
12312	64800
12313	50100
12314	50100
12315	50100

12316 rows × 1 columns

03 Prétraitement des données

Standardisation | Valeurs manquantes | **Encodage** | Matrice de corrélation

Colonne Time | **Colonnes Ordinales** | Colonnes Non Ordinales

Colonne “age band driver”	**Nouvelle Valeur**
Under 18	1
18-30	2
31-50	3
Over 51	4

03 Prétraitement des données

Standardisation | Valeurs manquantes | **Encodage** | Matrice de corrélation

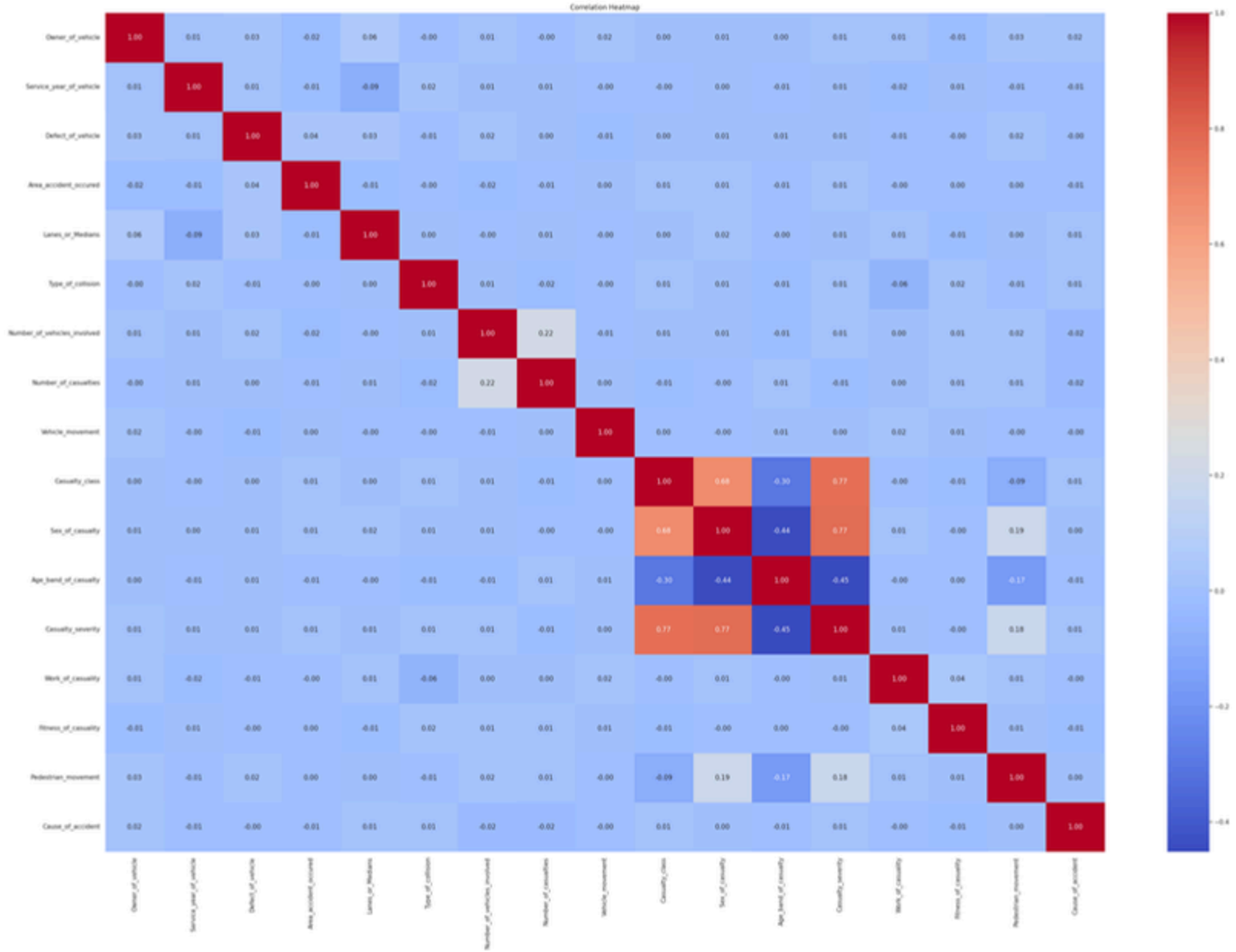
Colonne Time | Colonne Ordinale | **Colonne Non Ordinale**

LabelEncoder

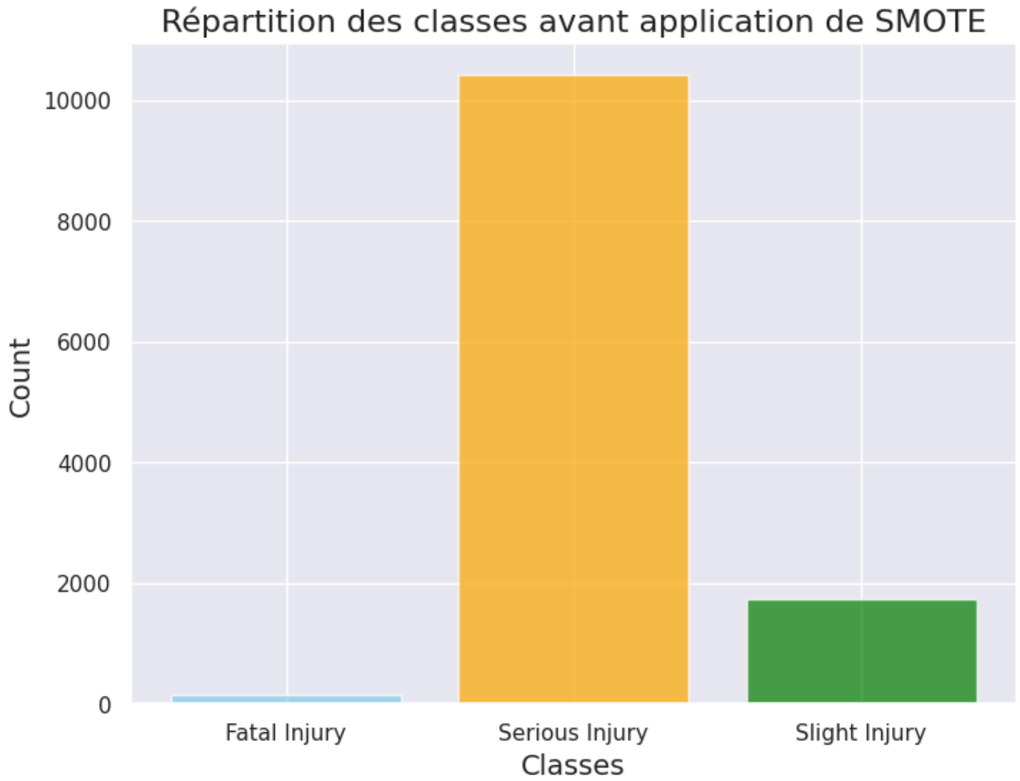
Colonne “Owner_of_vehicle”	**Nouvelle Valeur**
Owner	0
Governmental	1
Organization	2

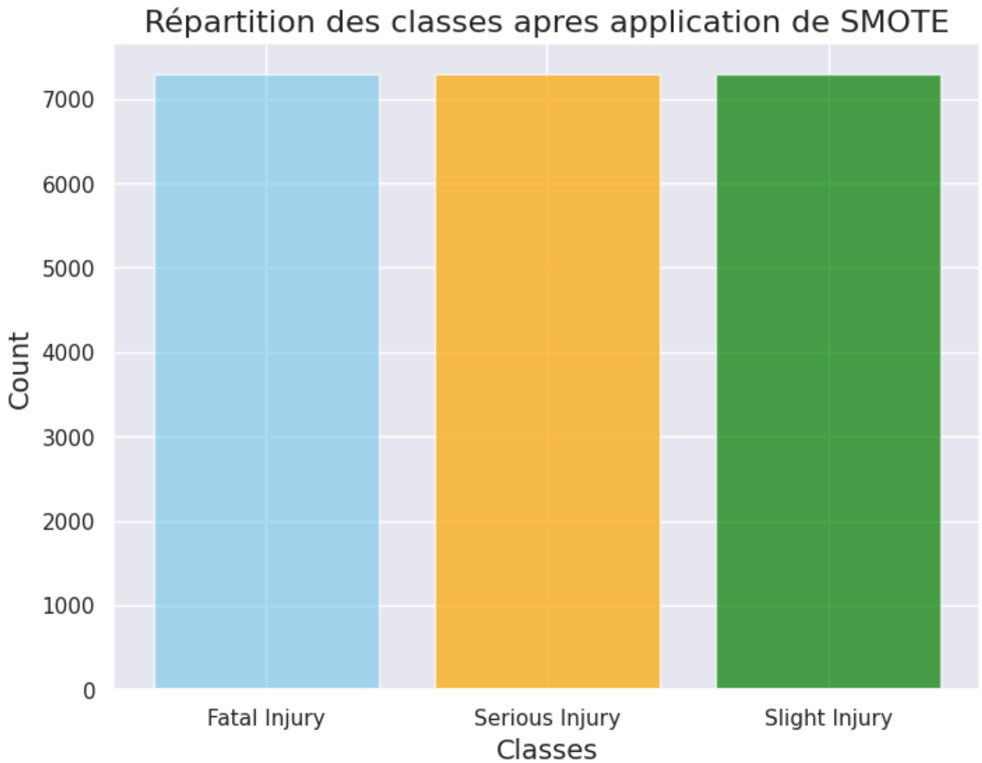
03 Prétraitement des données

Standardisation | Valeurs manquantes | Encodage | **Matrice de corrélation**



Problème des classes déséquilibrées





04 AutoML et Modélisation

04 Exploration AutoML

- prétraitement de données

Données :

Nombre d'observations : 12 316
Nombre de caractéristiques : 25
Données avec valeurs manquantes : 8.2%
Division des données : Entraînement (8621) / Test (3695)

Prétraitement :

Imputation des valeurs manquantes :
Numériques : Moyenne
Catégorielles : Mode

0	Session id	123
1	Target	Accident_severity
2	Target type	Multiclass
3	Original data shape	(12316, 25)
4	Transformed data shape	(12316, 25)
5	Transformed train set shape	(8621, 25)
6	Transformed test set shape	(3695, 25)
7	Numeric features	24
8	Rows with missing values	8.2%
9	Preprocess	True
10	Imputation type	simple
11	Numeric imputation	mean
12	Categorical imputation	mode
13	Fold Generator	StratifiedKFold
14	Fold Number	10
15	CPU Jobs	-1
16	Use GPU	False
17	Log Experiment	False
18	Experiment Name	clf-default-name
19	USI	4ee6

04 Exploration AutoML

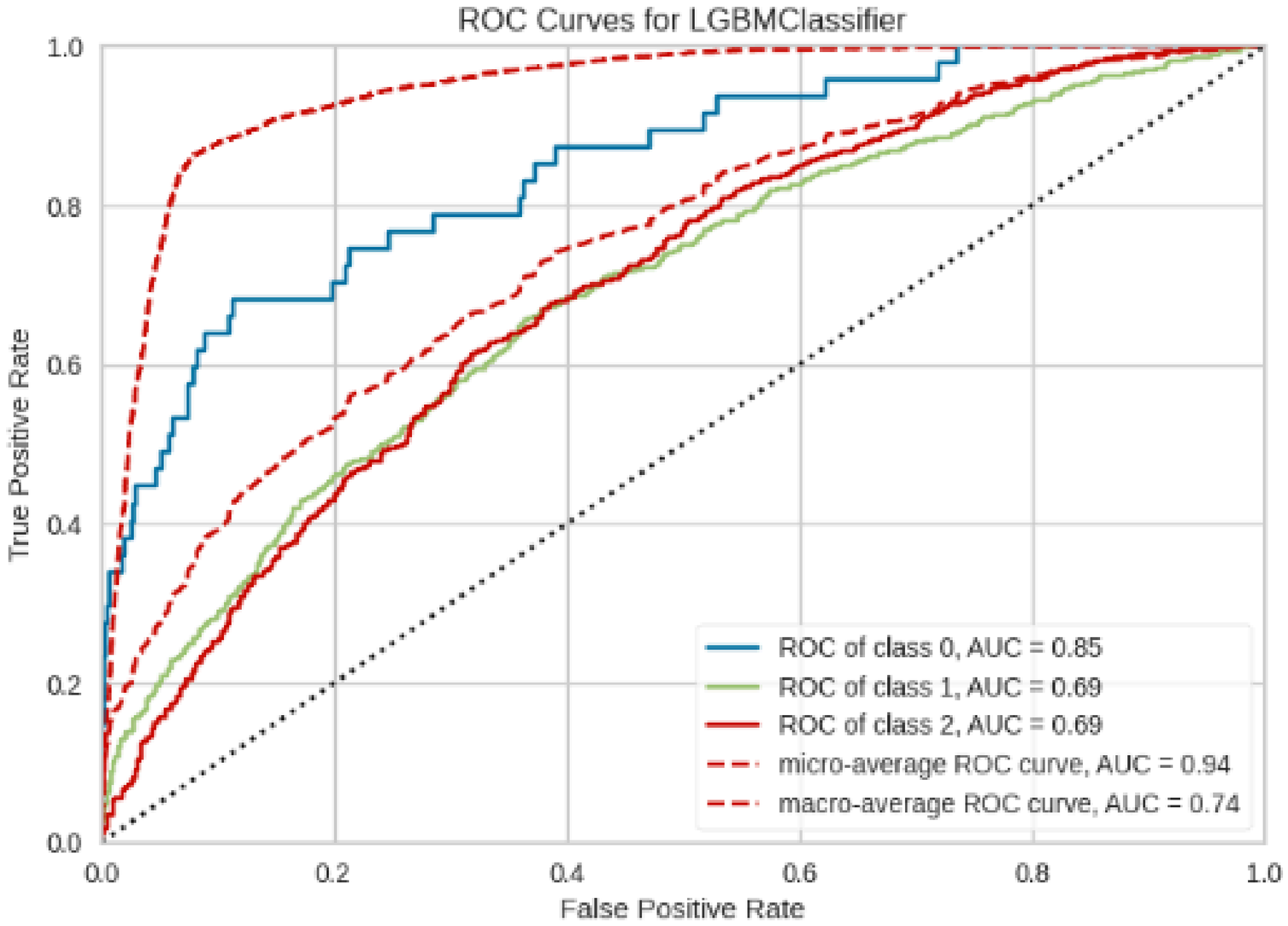
Comparaison des modèles

1- Meilleur modèle : **LightGBM**
Accuracy: **85.41%**
Precison : 82.94%
Recall : 85.41%
2- Deuxième modèle : **Random Forest**
Accuracy: **85.22%**
Precision : 82.77%
3- Troisième modèle : **XGBoost**
Accuracy: **85.05%**
Precision : 81.34%

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.8541	0.6918	0.8541	0.8294	0.8018	0.1298	0.2183	5.0490
rf	Random Forest Classifier	0.8522	0.6694	0.8522	0.8277	0.7940	0.0907	0.1860	1.3350
et	Extra Trees Classifier	0.8505	0.6524	0.8505	0.8229	0.7904	0.0729	0.1603	1.2580
xgboost	Extreme Gradient Boosting	0.8505	0.6982	0.8505	0.8134	0.8042	0.1461	0.2096	0.5820
gbc	Gradient Boosting Classifier	0.8496	0.0000	0.8496	0.8167	0.7906	0.0772	0.1558	2.9910
lr	Logistic Regression	0.8456	0.0000	0.8456	0.7151	0.7749	0.0000	0.0000	3.1570
ridge	Ridge Classifier	0.8456	0.0000	0.8456	0.7151	0.7749	0.0000	0.0000	0.0440
dummy	Dummy Classifier	0.8456	0.5000	0.8456	0.7151	0.7749	0.0000	0.0000	0.0370
ada	Ada Boost Classifier	0.8455	0.0000	0.8455	0.7578	0.7774	0.0168	0.0566	0.3440
lda	Linear Discriminant Analysis	0.8438	0.0000	0.8438	0.7148	0.7739	-0.0020	-0.0085	0.0500
nb	Naive Bayes	0.8402	0.5821	0.8402	0.7145	0.7723	-0.0046	-0.0128	0.0460
knn	K Neighbors Classifier	0.8291	0.6541	0.8291	0.7805	0.7945	0.1290	0.1488	0.1450
svm	SVM - Linear Kernel	0.7625	0.0000	0.7625	0.6437	0.6975	0.0000	0.0000	0.3270
dt	Decision Tree Classifier	0.7609	0.5841	0.7609	0.7770	0.7684	0.1563	0.1571	0.0890
qda	Quadratic Discriminant Analysis	0.3682	0.0000	0.3682	0.5345	0.3986	0.0174	0.0216	0.0490

04 Exploration AutoML

Courbe ROC AUC



Processus de modélisation

1

Test des modèles

Random Forest
Xgboost
LightGBM

2

Optimisation des hyperparamètres

GridSearch
Cross Validation

3

Evaluation des modèles

Courbe ROC AUC
Matrice de confusion

4

Choix du modèle

Meilleure Accuracy

Choix des hyperparamètres

Random Forest :

- **n_estimator :** 500
- **max_depth :** 20
- **min_samples_split :** 10

LightGBM et Xgboost :

- **n_estimators** : 500
- **max_depth** : 20
- **learning_rate** : 0.1
- **subsample :** 1.0

04 Modélisation

Test des modèles

Evaluation

Modèle	F1-Score	Accuracy	Recall	Precision
Random Forest	0,79	0,82	0,82	0,78
Xgboost	0,80	0,83	0,83	0.79
LightGBM	0,80	0.83	0,83	0,79

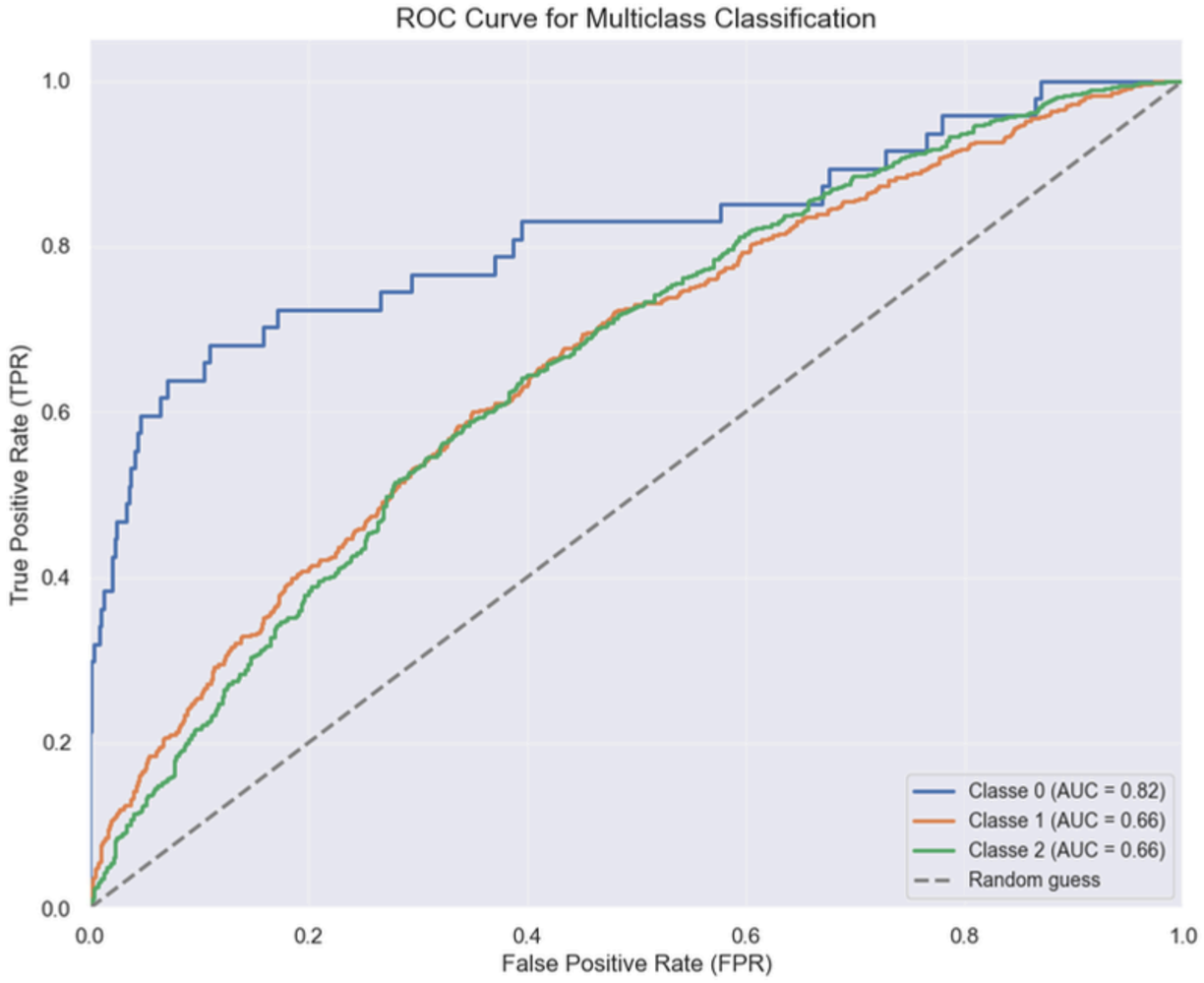
	precision	recall	f1-score
0	0.50	0.11	0.18
1	0.38	0.17	0.24
2	0.86	0.95	0.91

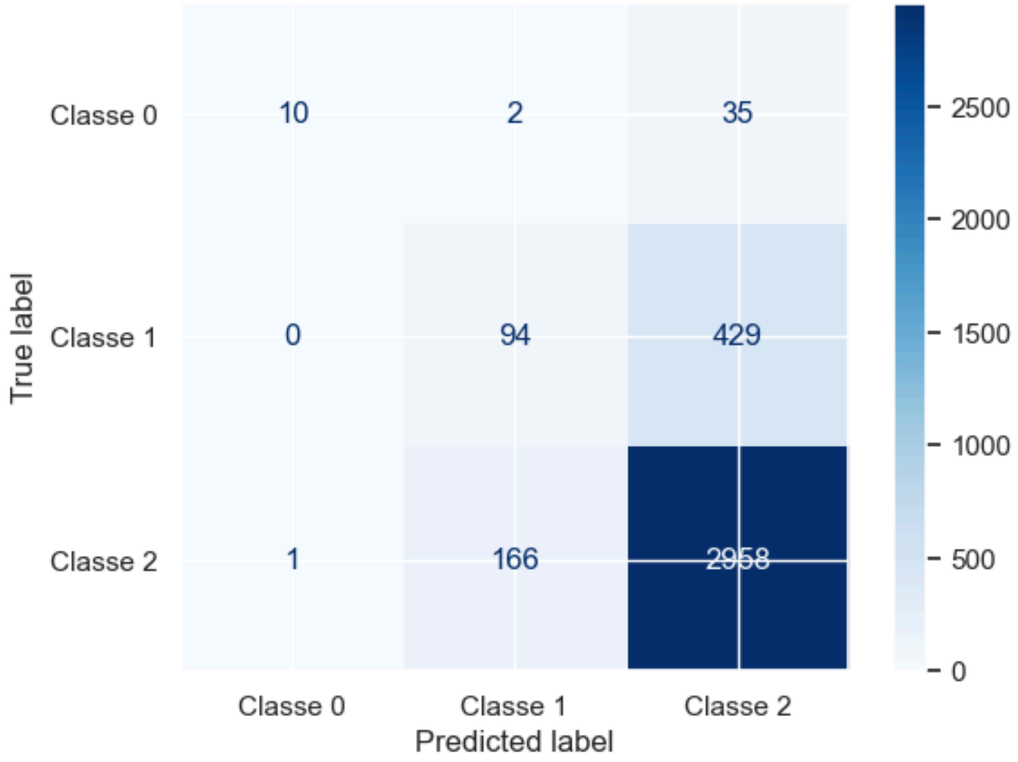
	precision	recall	f1-score
0	0.91	0.21	0.34
1	0.36	0.18	0.24
2	0.86	0.95	0.90

04 Modélisation

Test des modèles

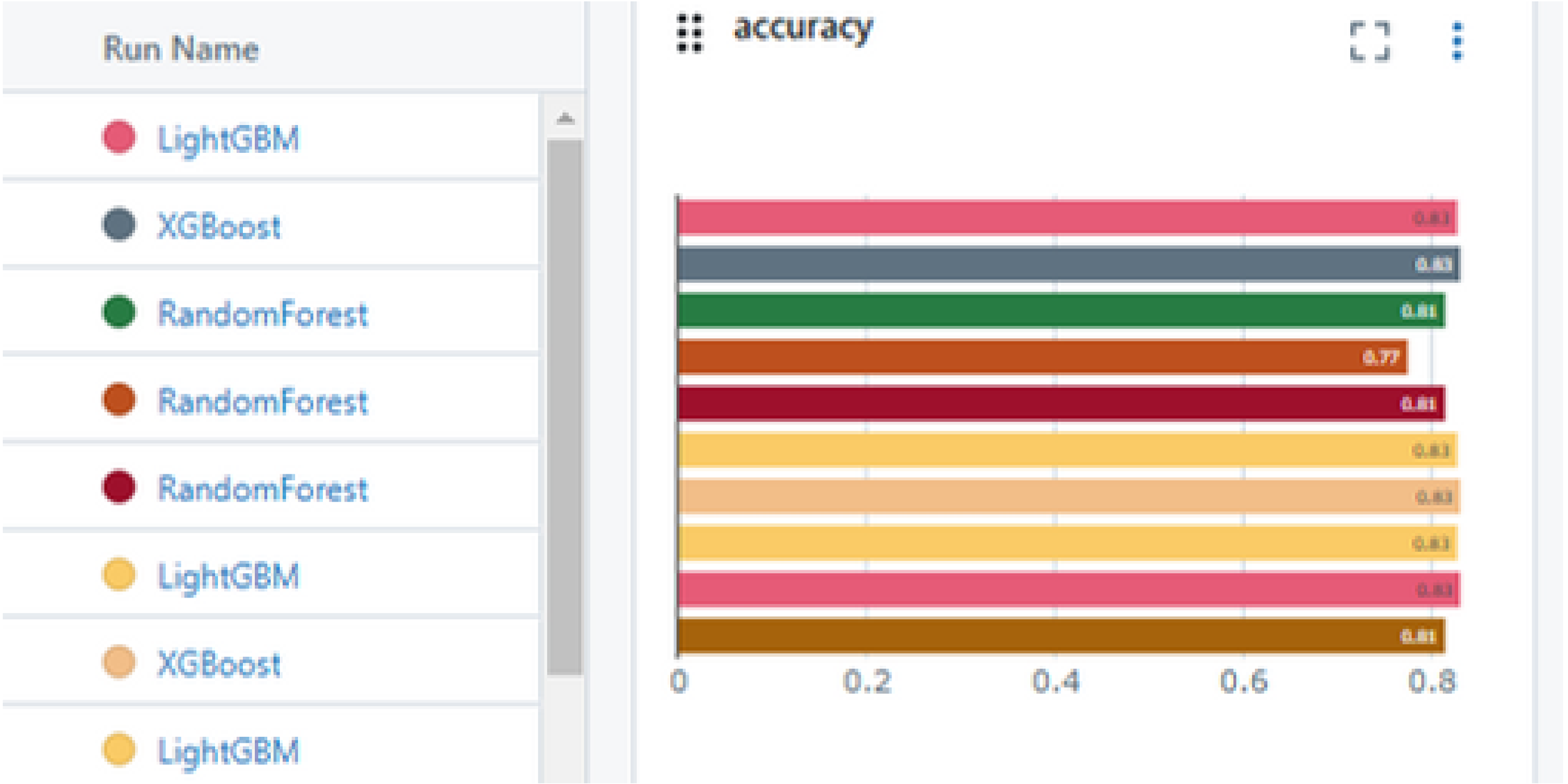
Evaluation





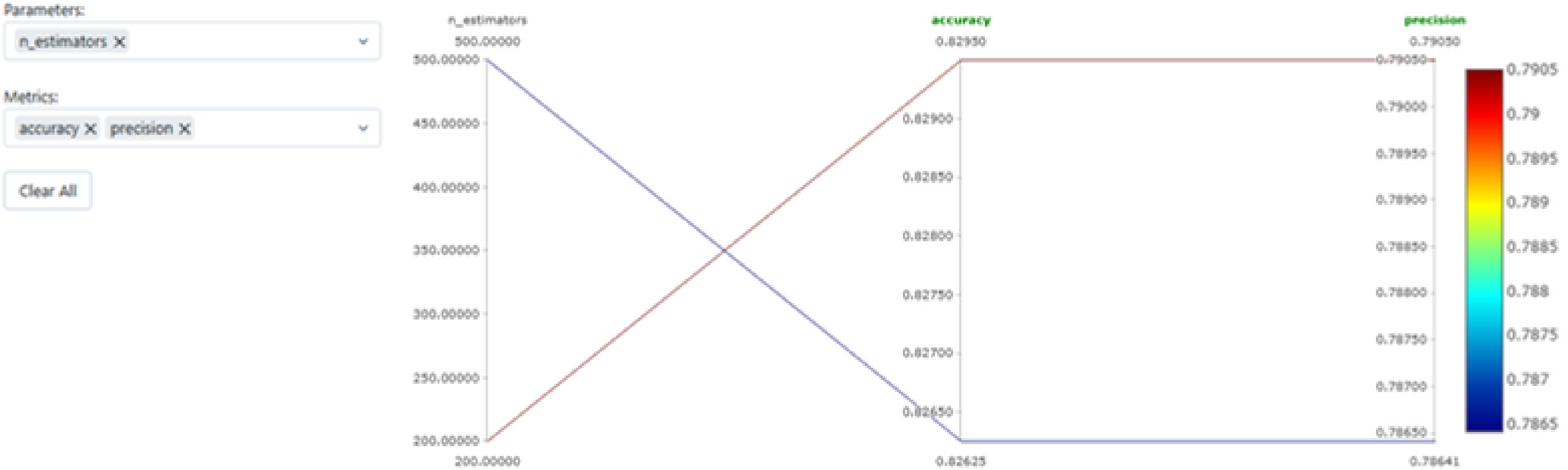
05 Suivi des Modèles avec MLflow

05 Suivi des Modèles avec MLflow



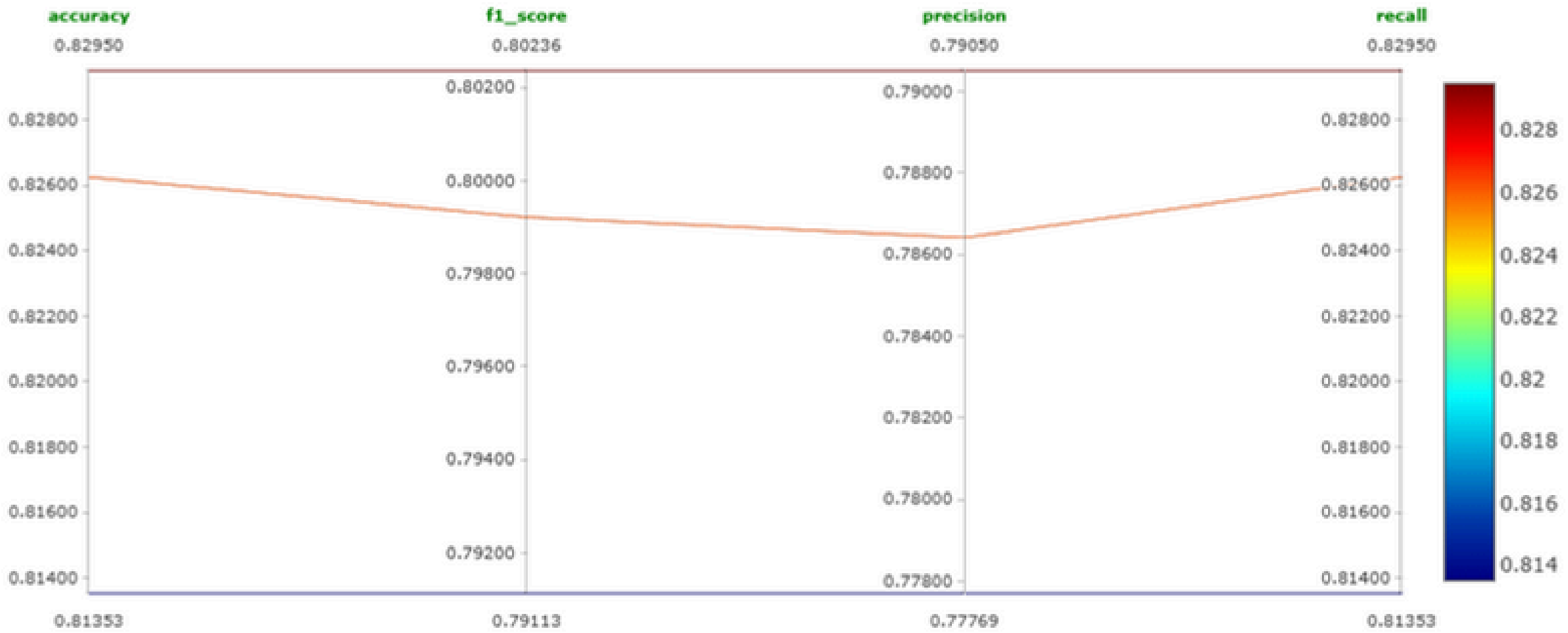
05 Suivi des Modèles avec MLflow

Comparaison des runs



05 Suivi des Modèles avec MLflow

Comparaison des runs



05 Suivi des Modèles avec MLflow

Déploiement du modèle en API REST locale

- MLflow models serve
- Bibliothèque requests

Résultat :

```
else:
    print(f"{response.status_code}: {response.text}")
```

```
L'accident est classé en : Serious Injury
```

```
data = {
    'Time': 61320,
    'Day_of_week': 'Monday',
    'Age_band_of_driver': '18-30',
    'Educational_level': 'Above high school',
    'Driving_experience': '2-5yr',
    'Light_conditions': 'Daylight',
    'Sex_of_driver': 'Male',
    'Vehicle_driver_relation': 'Employee',
    'Type_of_vehicle': 'Automobile',
    'Owner_of_vehicle': 'Owner',
    'Service_year_of_vehicle': 'Above 10yr',
    'Defect_of_vehicle': 'No defect',
    'Area_accident_occured': 'Residential areas',
    'Lanes_or_Medians': 'Undivided Two way',
    'Road_allignment': 'Flat terrain',
    'Types_of_Junction': 'No junction',
    'Road_surface_type': 'Asphalt roads',
    'Road_surface_conditions': 'Dry',
    'Weather_conditions': 'Normal',
    'Type_of_collision': 'Vehicle with vehicle collision',
    'Number_of_vehicles_involved': 2,
    'Number_of_casualties': 2,
    'Vehicle_movement': 'Going straight',
    'Casualty_class': 'Driver or rider',
    'Sex_of_casualty': 'Male',
    'Age_band_of_casualty': '31-50',
    'Casualty_severity': 'Slight Injury',
    'Work_of_casuality': 'Driver',
    'Fitness_of_casuality': 'Normal',
    'Pedestrian_movement': 'Not a Pedestrian',
    'Cause_of_accident': 'Moving Backward'
}


# Encodage des données
encoded_data = encode_data(data)

# Préparation des données pour l'API
request_data = {
    "dataframe_records": [encoded_data]
}

url = "http://127.0.0.1:1234/invocations"
headers = {"Content-Type": "application/json"}
```

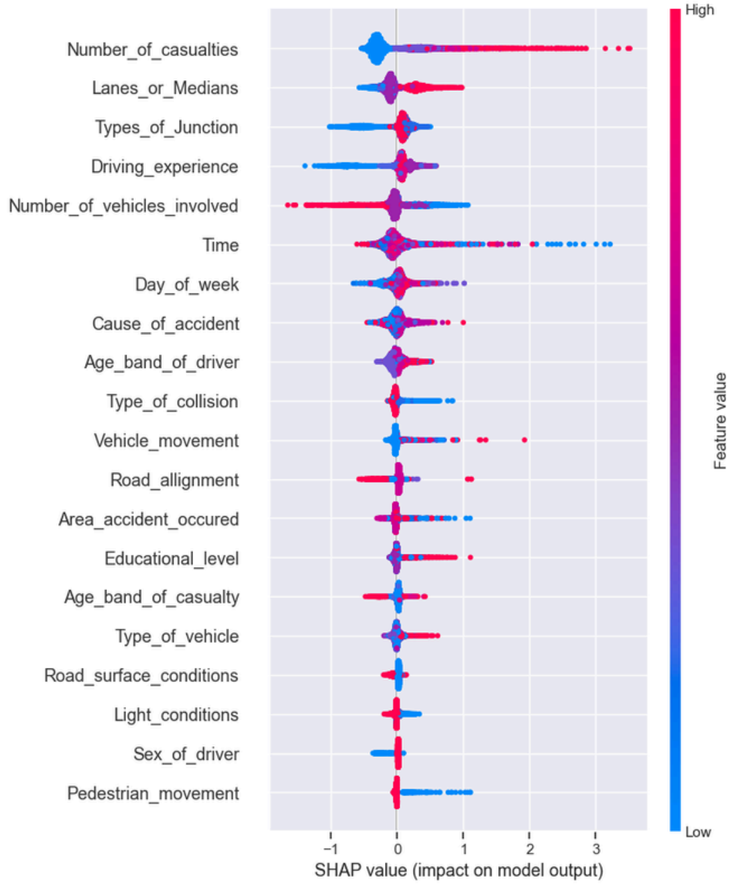
06

Interprétabilité du Modèle

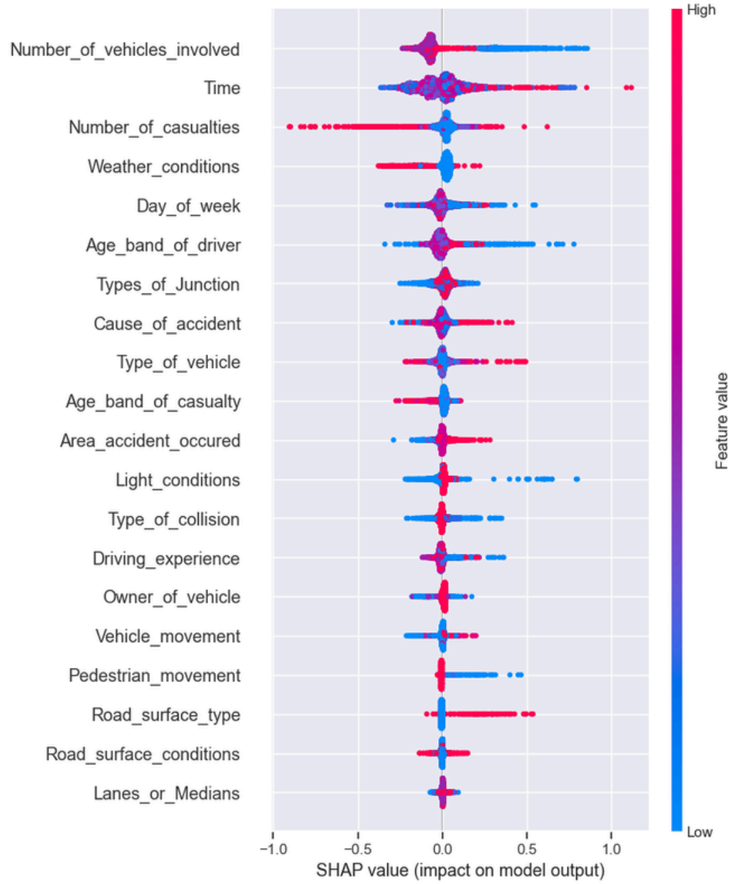
06 Interprétabilité du Modèle

Explication du modèle avec SHAP

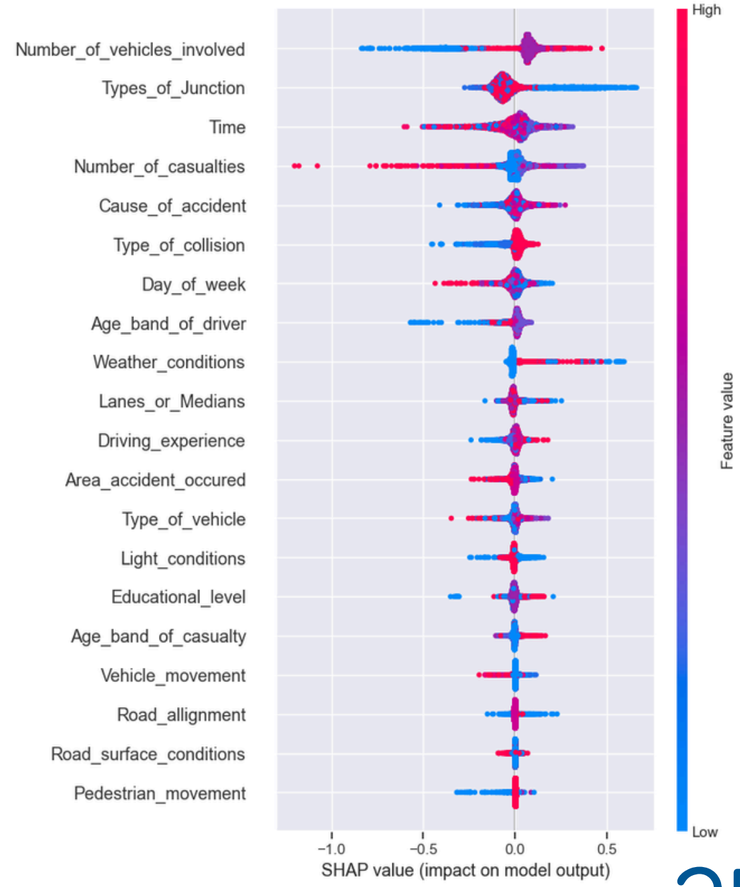
Classe 0 : Fatal Injury



Classe 1 : Serious Injury



Classe 2 : Slight Injury



06 Interprétabilité du Modèle

Explication du modèle avec LIME

Feature	Value
Number_of_casualties	4.00
Pedestrian_movement	5.00
Road_allignment	5.00
Light_conditions	2.00
Age_band_of_casualty	0.00
Road_surface_type	0.00
Type_of_collision	6.00
Lanes_or_Medians	2.00
Vehicle_movement	2.00
Age_band_of_driver	2.00

07 Conclusion

Merci de votre attention