

ELLIOT LAGAISE

DEMODAY

2024



Accidents à vélo

Jedha

Janvier 2024

SOMMAIRE

01

INTRODUCTION

02

LE DATASET

03

PROBLÉMATIQUE

04

PowerBi

05

PERSONA

06

MACHINE LEARNING

07

APP, LANDING PAGE

08

CONCLUSION



01



INTRODUCTION

Introduction

Le contexte

Le Demoday est un événement organisé à la fin de chaque promotion par l'école Jedha, lors de celui-ci nous devons présenter à l'oral un projet que nous avons pu réaliser lors des deux semaines précédentes devant un jury, il est composé du directeur, des professeurs et d'autres élèves. Cette évènement est rediffusé sur les réseaux sociaux de l'école ainsi que sur Youtube. Il a pour but d'utiliser l'ensemble des compétences apprises lors de ma formation.

Nous avions deux semaines pour réaliser ce projet, des choses ne sont pas forcément totalement abouties et des partie-pris ont du être fait pour pouvoir continuer notre projet. Il est évident que certains choix ne seraient pas pris dans des cas réels d'entreprise.



Le but de ce projet est aussi aussi l'occasion pour les élèves de choisir une problématique qui peut-être encrée dans le réel mais j'ai choisis de tourner ma problématique de manière humoristique pour que l'oral soit plus ludique.



02

LE DATASET

Le dataset

Explications et cleaning



Provenance du dataset

Le dataset a été récupéré sur le site du gouvernement qui redirige vers un autre site, il est mis à jour régulièrement avec des données presque bien nettoyées.

<https://opendata.koumoul.com/datasets/accidents-velos>

Explications

Le dataset porte sur les accidents de vélo avec d'autres véhicules de 2005 à 2021, il est possible de retrouver des précisions sur les circonstances de l'accident (équipement, véhicule, blessé...). La localisation précise du lieu de l'accident (latitude, longitude, intersection...) ainsi que 4 classes sur l'état du cycliste : indemne, mort, blessé léger ou hospitalisé.

Cleaning

J'ai donc construit une table de correspondance ([lien Google Drive](#)).

Cette table a été construite manuellement en matchant le dataset et la correspondance disponible sur le site, et en utilisant le numéro d'accident comme identifiant unique.

Cela a permis d'obtenir tous les libellés et dans un premier temps d'avoir une meilleure compréhension des données de notre Dataset.

Pour l'utilisation du Dataset avec Power BI, j'ai finalement utilisé une fonction DAX permettant de remplacer les chiffres par les libellés. Exemple :

```
agglomeration =  
SWITCH([agg],  
    1, "hors_agglomeration",  
    2, "en_agglomeration",  
    "")  
)
```

Creation d'une table de correspondance

Modification des types de données (float en date etc..)

Remplacement des données numériques par les champs correspondants



03

PROBLÉMATIQUE

Le dataset

Quel est le meilleur moyen de mourir dans un accident de vélo ?



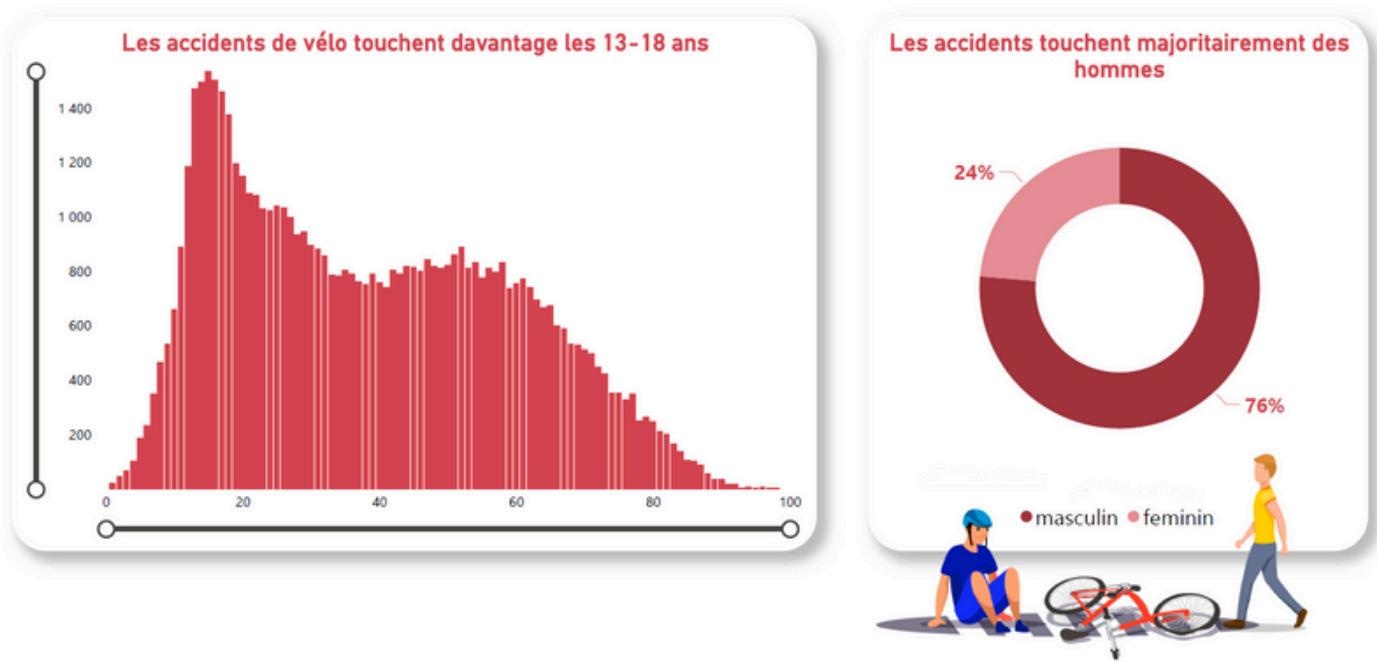
04

POWERBi

PowerBi

Visualisation des données

Dashboard 1



Sur ce premier dashboard, je souhaite effectuer une première analyse de la population concernée.

- En fonction de l'âge

J'ai créé un histogramme présentant la répartition des accidents selon l'âge de la personne.

Ce choix de graphique permet de visualiser la tranche d'âge la plus représentée, en l'occurrence les adolescents (13–18 ans) avec un pic sur l'âge de 15 ans.

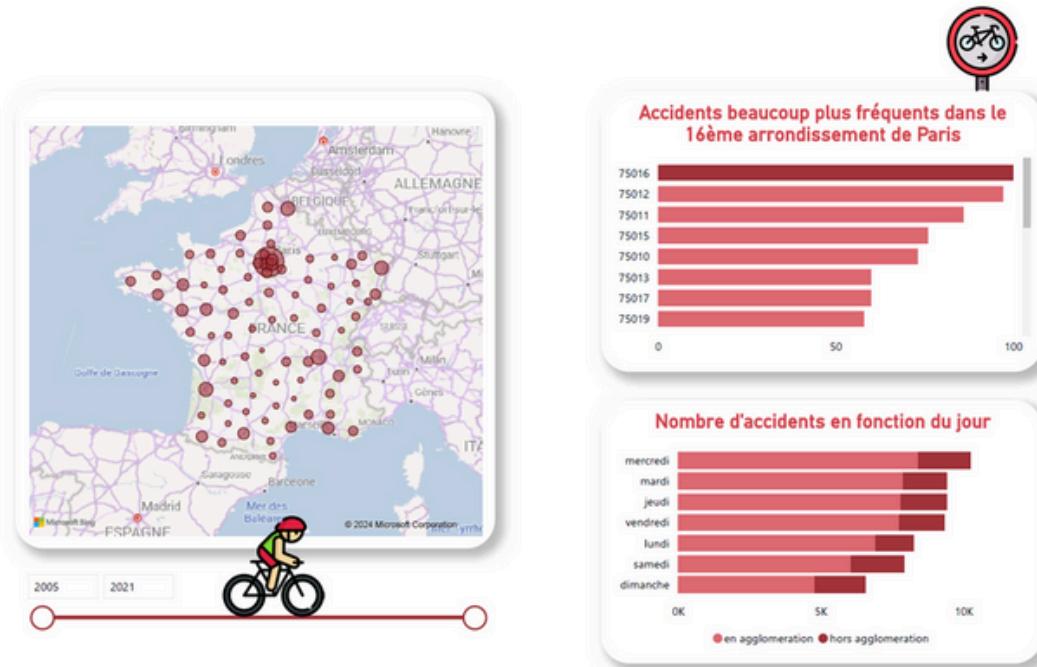
- En fonction du sexe

Avec deux choix possibles (masculin ou féminin), le pie chart était le type de graphique le plus adapté pour visualiser la répartition des accidents entre hommes et femmes.

PowerBi

Visualisation des données

Dashboard 2



Notre 2ème dashboard propose un axe géographique et temporel

- Localisation des accidents

J'ai utilisé un premier graphique de type « Map » pour représenter la densité des accidents. Le graphique prend notamment comme données les départements.

Sur un second graphique, J'ai réalisé un focus sur Paris (département le plus dense en nombre d'accidents). Afin d'identifier les arrondissements les plus à risques, J'ai fait le choix d'un « Stacked Bar Chart », avec des données basées sur les communes (codes postaux).

- Répartition des accidents selon le jour

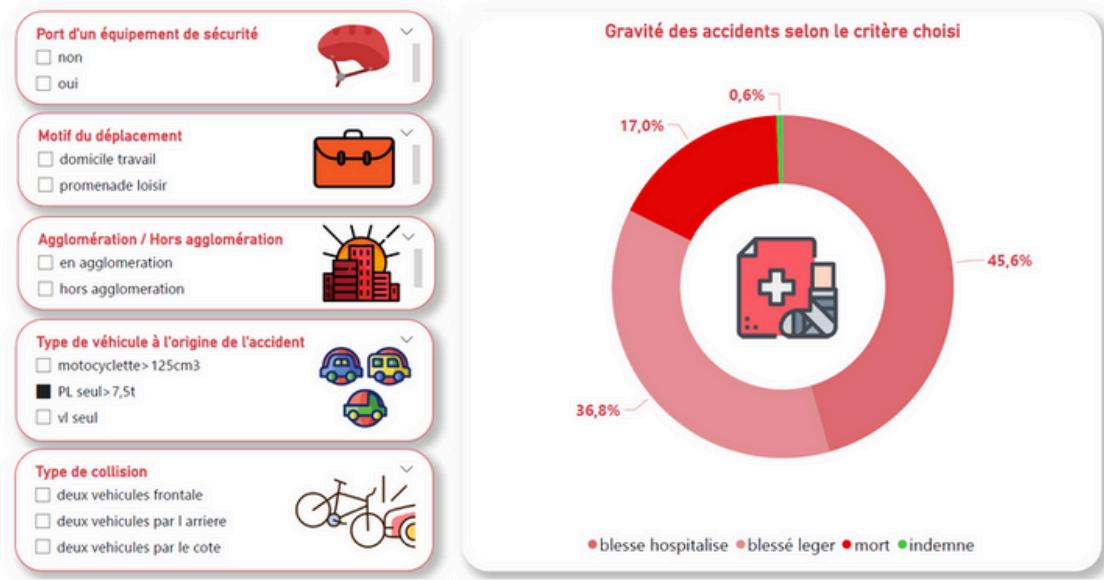
Enfin, le dernier graphique offre une double lecture :

- a.Un classement présentant les jours les plus à risques
- b.Une répartition des accidents selon s'ils surviennent en agglomération ou hors agglomération

PowerBi

Visualisation des données

Dashboard 3



Le dernier dashboard se voulait dynamique avec une représentation de la gravité des accidents évolutive en fonction des principaux critères de notre modèle.

Le choix des cinq critères repose à la fois sur leur impact dans la gravité des accidents et sur les critères mis en avant par le Machine Learning.

Le pie chart est apparu comme le plus parlant pour montrer la gravité des accidents. Le niveau de gravité évolue en fonction des critères choisis dans les cinq « sliders » créés.

05

PERSONA

Persona

Le persona permet de connaitre le profil type de la personnes qui comporte toutes les caractéristiques afin de répondre à notre problématique, dans ce cas précis c'est la personne qui a le plus de chance de faire un accident.

John est un adepte du vélo de 15 ans. Il utilise ce moyen de transport depuis tout petit avec sa soeur et continue de rouler désormais pour aller à l'école, mais surtout il fait du vélo avec ses amis.



Vit en région
Parisienne avec sa
famille.



Il ne porte pas
de casque car il
connaît ce moyen
de transport et
ne voit pas
l'utilité.



Il est un peu
casse-cou et aime
prendre des
risques pour
impressionner ses
amis.



Il sort tous les
mercredis avec ses
amis car il n'a pas
école. Ils se baladent
en ville.

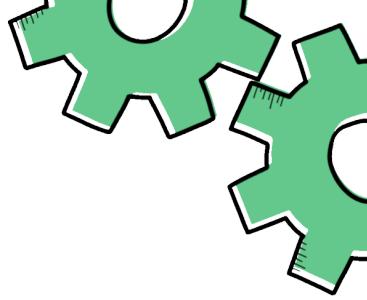


06

MACHiNE
LEARNING

Machine learning

Prédiction de la gravité de blessures



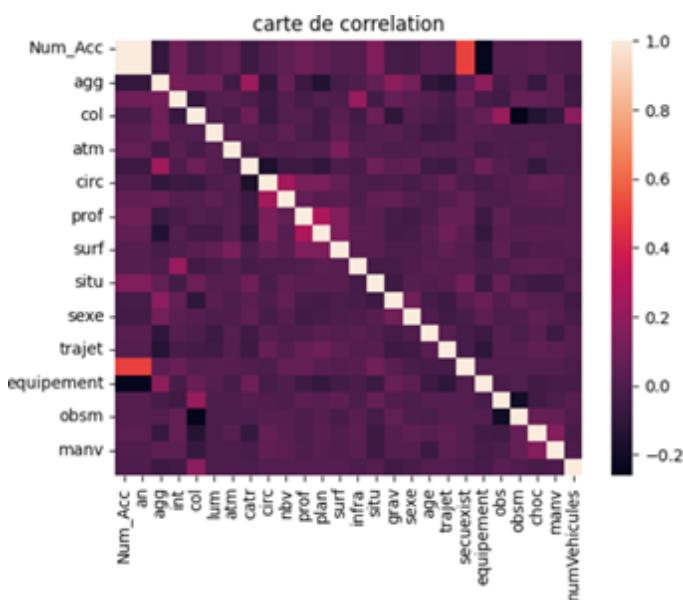
But du machine learning

Le but du modèle est de prédire les critères principaux qui pourraient impacter la gravité de l'accident divisés en 4 catégories (1 = indemne, 2 = tué, 3 = blessé hospitalisé, 4 = blessé léger) Pour ensuite faire une application qui permette de prédire le meilleur moyen de mourir dans un accident. Pour rappel c'est un projet de groupe où nous avons une connaissance du machine learning basique qui ne rentre pas dans notre programme, c'est pour cela que les résultats sont imprécis par manque de temps. Le but est de créer une application qui utilise le model.

Clean du dataset

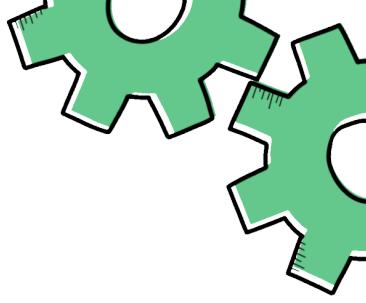
Après avoir fait un cleaning général pour faire du powerBI et de l'EDA sur python, il était aussi nécessaire d'en faire un spécialement pour faire les modèles, notamment numériser les données comme par exemple les jours.

La création des modèles est faite sur Dataiku. Pour réduire un maximum de features qui peuvent ralentir et biaiser la qualité du programme j'ai réalisé une matrice de corrélation sur python :

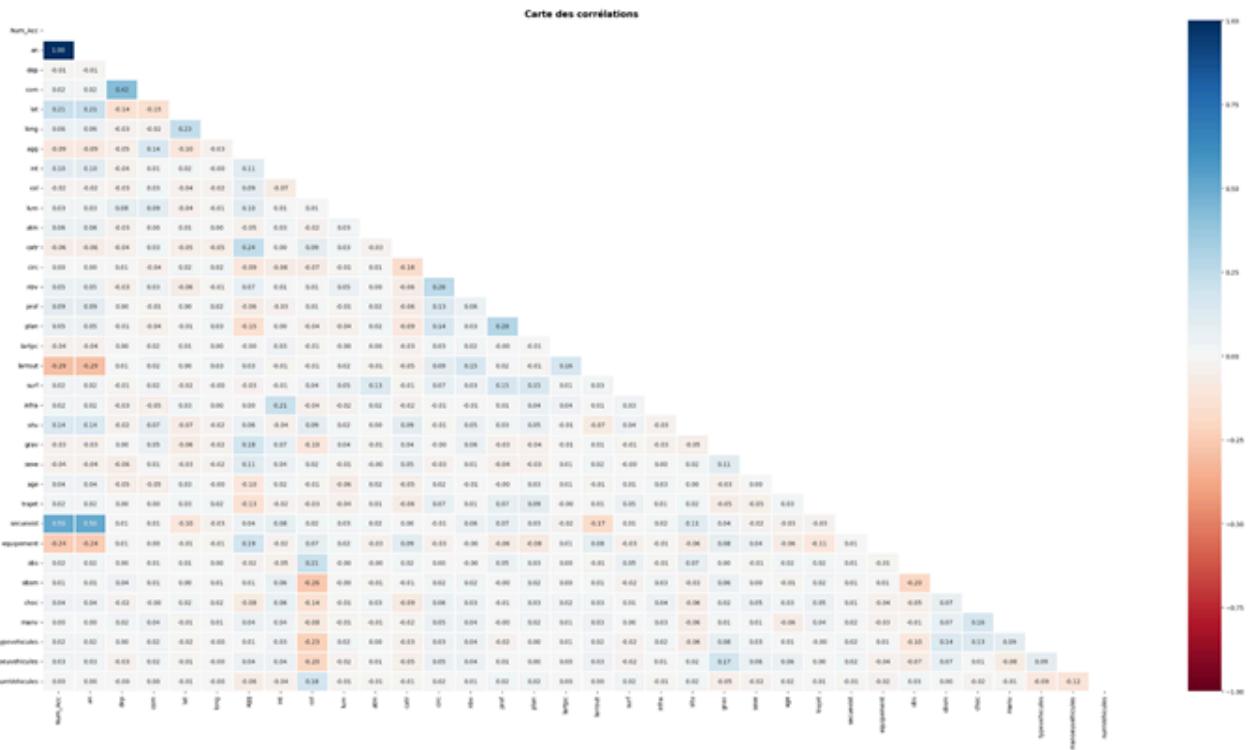


Machine learning

Prédiction de la gravité de blessures



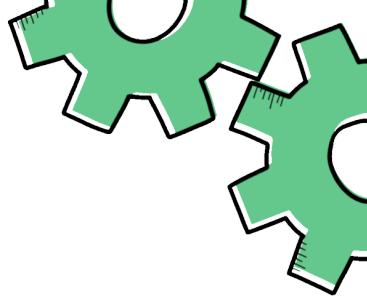
Cette matrice est très peu lisible, la partie gauche ne sert pas et l'écart est à -0,2.
Après certaines modifications voici le résultat :



J'ai pris comme base 0,70 pour me dire que les colonnes étaient corrélées, il n'y en avait donc aucune à supprimer. Même si le graphique nous dit qu'il ne faut pas en supprimer, J'ai réduit de 39 features à 31 car des colonnes n'étaient pas utiles pour notre analyse comme par exemple celles des départements alors que j'ai déjà les communes, des ID de véhicules...

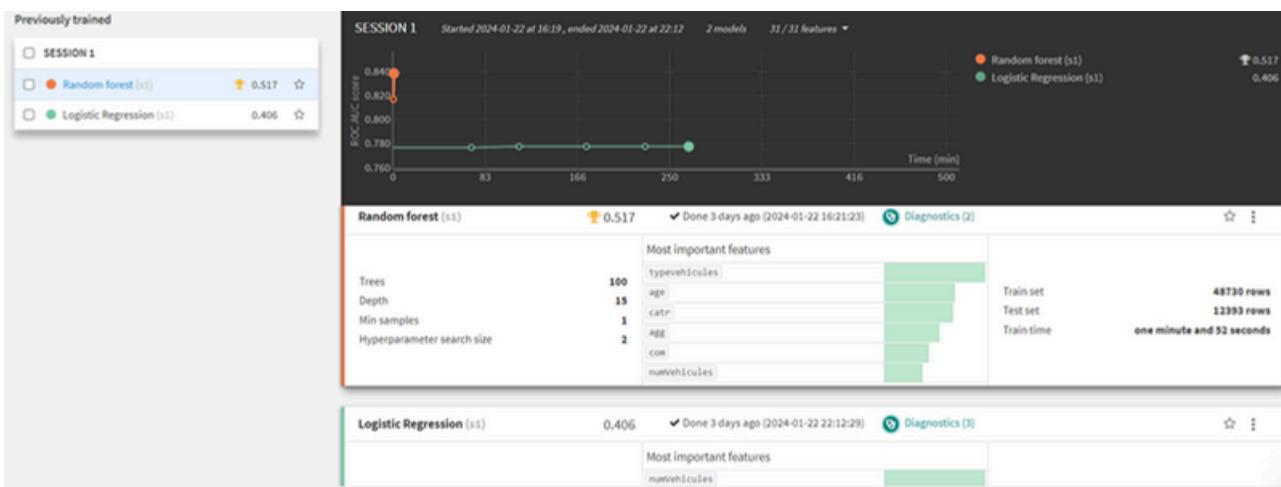
Machine learning

Prédiction de la gravité de blessures



Les différents modèles de machine learning

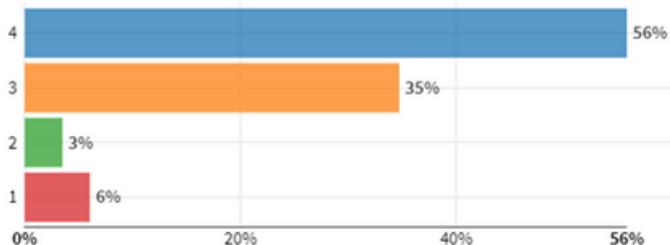
J'ai donc réalisé plusieurs modèles de machine learning en multiclass classification vu que nous avons 4 classes. Après avoir eu les premiers résultats avec des critères qui me semble cohérents, j'ai créé des modèles avec des top 5 et top 10 des critères pour voir si le résultat change. Voici la précision la plus haute :



Nous pouvons voir certains problèmes : le dernier critère n'est pas bon puisqu'il ne sert pas dans l'analyse, le deuxième est la répartition des catégories du dataset qui peut biaiser les résultats que l'on peut analyser avec la matrice de confusion :

Target classes

Proportions of classes computed on a sample using the current sampling settings

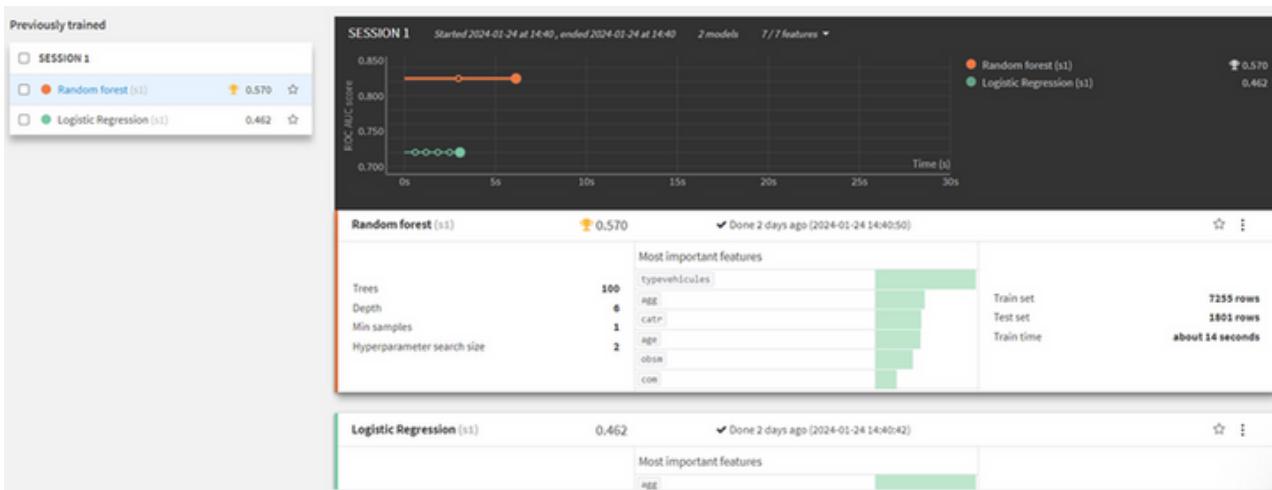


Actual	Predicted				100 %
	4	3	2	1	
4	75 %	17 %	2 %	6 %	100 %
3	37 %	47 %	12 %	3 %	100 %
2	13 %	49 %	37 %	< 1 %	100 %
1	22 %	9 %	< 1 %	68 %	100 %

Machine learning

Prédiction de la gravité des blessures

Les classes avec le moins de données sont les plus difficiles à prédire (cf matrice). Pour régler les problèmes, j'ai donc créé un nouveau dataset sur python qui possède le même nombre de lignes par classe en prenant la classe qui avait le moins de lignes comme référence avec le même principe de top 5 et top 10. Voici le meilleur résultat :



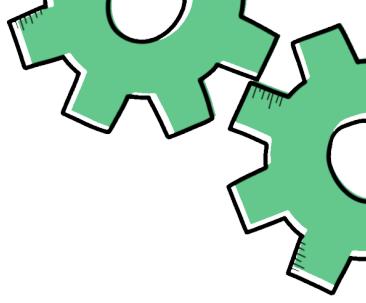
Grâce à cette deuxième analyse, la précision augmente de 0,06%. Avec une meilleure matrice de confusion. J'ai aussi fait un test avec un autre modèle qui est le xgboost avec une modification des hyperparamètres pour un résultat qui est passé à 0,62.

Incohérence avec l'application

Les critères qui ressortent dans le top 5 ne sont pas en cohérence avec notre application. Par exemple, une personne qui veut savoir par quel moyen il peut se suicider facilement ne peut pas savoir à l'avance sur quel véhicule il va tomber. Du coup j'ai dû prendre le problème à l'envers et créer un modèle avec des features qui paraissent cohérentes à rentrer. Même si le modèle n'est qu'à 0,45 et qu'il devient donc vraiment inutilisable pour un cas qui est réel. J'ai pris le choix de l'utiliser pour la cohérence globale du projet et par manque de temps. Évidemment il a des pistes d'amélioration pour le modèle.

Machine learning

Prédiction de la gravité des blessures



Pistes d'amélioration

- Plus de données dans le dataset de base
- Mieux balancer les données
- Voir s'il n'y a pas un autre modèle plus précis
- Regrouper les classes "blessé léger" et "hospitalisé" pour améliorer la précision (qui marche seulement dans notre cas pour le but de l'application)
- Changer les hyperparamètres



OFF

APP ET
LANDING PAGE

Application

Streamlit et Figma

J'ai voulu faire fonctionner le programme de machine learning dans une application, dans celles-ci les utilisateurs pourront rentrer les conditions de leur trajet afin de savoir s'ils ont la possibilité de faire un accident ou non. j'ai décidé de développer l'application via streamlit. Voici le code afin de faire tourner l'application :

```
4 import joblib
5
6 # Title of the app
7 st.title("La roue de l'infortune")
8
9 # Text input
10 sexe = st.selectbox("Votre sexe :", ["Femme", "Homme"])
11 équipement = st.selectbox('Choisissez votre équipement :', ["Casque", "Gilet réfléchissant", "Gants", "Aucun"])
12 trajet = st.selectbox("Quel sera le motif de votre déplacement ? :", ["Domicile-travail", "Promenade-Loisir"])
13 age = st.number_input("Entrez votre âge :", min_value=1, max_value=100)
14 agg = st.selectbox("Choisissez votre lieu d'habitation :", ["Agglomération", "Hors agglomération"])
15 atm = st.selectbox("Quel temps va t'il faire lors de votre trajet ? :", ["Normal", "Pluie", "Neige ou grève", "Brouillard"])
16 jour = st.selectbox("Quel jour allez-vous rouler :", ["Lundi", "Mardi", "Mercredi", "Jeudi", "Vendredi", "Samedi", "Dimanche"])
17
18 user_submit = st.button("Savoir mes chances de mourir")
19
20 # Prepare user data for prediction
21 if user_submit:
22     # Create a DataFrame using user input
23     user_data = pd.DataFrame({
24         "sexe": [1 if sexe == "Homme" else 0],
```

Grâce à ce code voilà la page que j'arrive à sortir :

La roue de l'infortune

Votre sexe :

Femme

Choisissez votre équipement :

Casque

Quel sera le motif de votre déplacement ? :

Domicile-travail

Entrez votre âge :

1

Choisissez votre lieu d'habitation :

Agglomération

Quel temps va t'il faire lors de votre trajet ? :

Normal

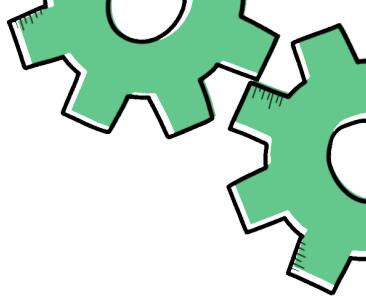
Quel jour allez-vous rouler :

Lundi

Savoir mes chances de mourir

Application

Streamlit et Figma



Proposition d'amélioration

Comme vous pouvez le voir, l'application n'est pas très esthétique, j'ai donc donc voulu pousser le concept à son maximum en proposant une maquette figma. Voici ce que je pourrions proposer si l'application avait été aboutie :

The wireframe illustrates a Streamlit application interface designed with a Figma mockup. It features a sidebar on the left and a main content area on the right.

Left Sidebar:

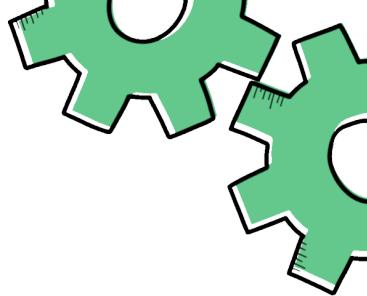
- Logo:** "La roue de l'infortune" (The Fortune Wheel) with a yellow gear icon.
- Section:** "TU VEUX METTRE FIN A TES JOURS?" (Do you want to end your life?) featuring a cartoon character of a man in a blue shirt and red pants.
- Text:** "Entre tes informations et notre roue va déterminer les dégâts de l'impact!" (Enter your information and our wheel will determine the damage of the impact!).
- Button:** "DÉCOUVRIR" (Discover).
- Icon:** A person wearing a headset with a speech bubble, labeled "CONTACT UN CONSEILLER" (Contact a counselor). Subtext: "Un conseiller peut t'aider. Glisse pour découvrir le numéro".
- Image:** An illustration of a tombstone with "R.I.P." and a small mouse running away from it.

Main Content Area:

- Header:** "La roue de l'infortune" with a yellow gear icon.
- Section:** "SEXUE" (Sex) with a dropdown menu.
- Section:** "AGE" (Age) with a dropdown menu.
- Section:** "QUELJOUR ALLEZ-VOUS ROULER ?" (When are you going to roll?) with a dropdown menu.
- Section:** "CHOISISSEZ VOTRE ÉQUIPEMENT" (Choose your equipment) with a dropdown menu.
- Section:** "CHOISISSEZ VOTRE LIEU D'HABITATION" (Choose your place of residence) with a dropdown menu.
- Section:** "QUEL TEMPS VA T'IL FAIRE LOIRS DE VOTRE TRAJET ?" (What will the weather be like during your trip?) with a dropdown menu.
- Section:** "QUEL SERA LE MOTIF DE VOTRE DÉPLACEMENT ?" (What will be the reason for your displacement?) with a dropdown menu.
- Text:** "NOTRE PRÉDICTION N'EST PAS AU POINT. PAS DE REMBOURSEMENT POSSIBLE DÉSOLÉ!" (Our prediction is not yet ready. No refund possible, sorry!).
- Text:** "Adresse email ..."
- Section:** "TON RÉCAP EN PDF" (Your recap in PDF) with an illustration of the "Wheel of Fortune" (La Roue de l'Infortune) showing various figures and symbols.
- Text:** "DÉCOUVRIR" (Discover).

Landing page

Attirer vers l'application



Enfin afin de conclure ce projet, j'ai décidé de créer une landing page afin d'amener les personnes à télécharger l'application que j'ai présenté juste avant. La landing page sera disponible une semaine avant le lancement de notre application afin de permettre au plus grand nombre de personnes de s'inscrire à l'avant-première du téléchargement. J'ai proposé une landing page via Hubspot :

LA ROUE DE L'INFORTUNE

Envie de mourir à vélo ?

Découvrez notre nouvelle future application de détection du danger à vélo, afin d'optimiser votre parcours et d'être certain de mourir.

Inscrivez-vous pour être dans nos futurs clients les plus mortels !

E-mail*

Notre application détectera que vous roulez pour nous donner au sujet de nos produits et services. Vous pouvez nous déconnecter de nos communications à tout moment. Consultez notre politique de confidentialité pour en savoir plus sur nos méthodes d'abonnement, ou pour nous gérer et nous déconnecter de nos communications. Nous ne partagerons pas vos informations avec des tiers.

ENVOYER

E-mail*

Notre application détectera que vous roulez pour nous donner au sujet de nos produits et services. Vous pouvez nous déconnecter de nos communications à tout moment. Consultez notre politique de confidentialité pour en savoir plus sur nos méthodes d'abonnement, ou pour nous gérer et nous déconnecter de nos communications. Nous ne partagerons pas vos informations avec des tiers.

ENVOYER

Créez un compte

Etes-vous le *suicidal biker* par excellence ou roulez vous dans les clous ?

Choisissez votre équipement :

Casque

Choisissez votre lieu d'utilisation :

Agglomération

Quel temps va t'il faire lors de votre trajet ? :

Normal

Rentrez vos informations

Ville ou campagne, équipements que vous portez (ou non !), quel(s) jour(s) roulez vous le plus ...

Découvrez comment vous pourriez mourir

Un camion pourrait vous percuter, vous pourriez mourir de nuit, la voiture viendra de derrière ...

Conclusion

& remerciements

Conclusion

Pour conclure ce dossier, j'ai voulu mettre en avant, via un dataset du gouvernement, la dangerosité des trajets en vélo dans la vie quotidienne, surtout lorsque l'usager ne porte pas d'équipement de protection.

Ce travail était challengeant, autant dans sa durée que dans sa complexité, il est le fruit de deux semaines de réflexion. Je me suis également amusé, en tournant ce sujet sur le ton de l'humour, à la fois pour choquer mais également pour dénoncer.

Je tiens à remercier nos professeurs, qui nous ont accompagnés durant ces deux semaines mais également durant tout notre cursus de formation en data analyst. Merci à vous pour votre disponibilité et pour vos réponses à nos nombreuses questions.

Enfin, merci à l'équipe de Jedha de nous avoir permis de présenter ce merveilleux projet.



ELLIOT LAGAISE



Accidents à vélo

Jedha

Janvier 2024