Project Overview

Le Projet **PROJET PERSONNEL :** *Multi-Class Prediction of Cirrhosis Outcomes*
Ce projet est issu d'une compétition Kaggle et tente de prédire l'état post expérience de patients atteint d'une cirrhose du foie en fonction de plusieurs variables (métriques biologiques, âge, médicamentation...) grâce à une prédiction multiclasse.
Pour plus d'information et télécharger le dataset, veuillez consulter le lien suivant : https://www.kaggle.com/competitions/playground-series-s3e26

Etape 1 : Réaliser une EDA pour découvrir et mieux comprendre les données.
Etape 2 : Construire et sélectionner l'algorithme le plus adapté et l'optimiser afin de réaliser des inférences (no free lunch theorem method)

Les Objectifs
Utiliser une approche multi-classe pour prédire l'évolution des patients atteints de cirrhose selon 3 labels (décès : D, vivant; C, vivant grâce à une greffe : CL). Les soumissions sont évaluées à l'aide de la LOG LOSS. Chaque identifiant de l'ensemble de test a un unique label. Pour chaque identifiant (patient), il faut prédire une probabilité pour chacun des trois résultats possibles. Il n'est pas nécessaire que la somme des probabilités soumises pour une ligne donnée soit égale à 1.

Livrable
- réaliser un EDA pour s'approprier les données et faire des tests statistiques pour mettre en lumière des résultats statistiquement significatifs
- preprocessing, oversampling, choix du meilleur algorithme, choix des meilleurs métriques et optimisation
- Le fichier doit contenir un en-tête et avoir le format suivant :
identifiant, Statut_C, Statut_CL, Statut_D
7905,0.628084,0.034788,0.337128

lien du projet : https://github.com/Thibaut-Longchamps/Certification-Fullstack/tree/main/Challenge%20kaggle%20multiclass%20pred%20and%20EDA%20projet%20personnel

---

Le Projet **TINDER** *(Analyse exploratoire descriptive et inférentielle de données)*
L'équipe marketing de Tinder cherche à comprendre pourquoi le nombre de matchs diminue sur leur plateforme. Pour ce faire, ils ont organisé des sessions de speed dating où les participants fournissent des informations détaillées sur eux-mêmes, similaires à celles qu'ils partageraient sur leur profil Tinder. Tinder a ensuite collecté les données de ces sessions, chaque ligne représentant un speed dating entre 2 personnes et indiquant si elles ont accepté d'aller à un deuxième rendez-vous ensemble. L'objectif est d'analyser ces données pour identifier les facteurs qui influent sur la décision des participants d'accepter ou non un deuxième rendez-vous, afin d'améliorer l'algorithme de l'app.

Les Objectifs
Utiliser le dataset pour comprendre les raisons pour lesquelles les gens match et acceptent de se revoir pour un deuxième rendez-vous et délivrer quelque recommandations.

Portée de ce projet
1. Des données ont été collectées auprès des participants à des événements expérimentaux de speed dating organisés entre 2002 et 2004. Pendant ces événements, chaque participant a eu des rendez-vous de quatre minutes avec chaque autre participant du sexe opposé. À la fin de chaque rendez-vous, les participants devaient indiquer s'ils étaient intéressés pour revoir leur partenaire. De plus, ils ont été invités à évaluer leur partenaire selon six critères : Attractivité, Sincérité, Intelligence, Amusement, Ambition et Intérêts communs.
2. Les données comprennent également des réponses à des questionnaires remplis par les participants à différents moments du processus. Ces questionnaires couvrent divers aspects tels que les données démographiques, les habitudes de drague, la perception de soi basée sur des attributs clés, les croyances sur ce que les autres trouvent attractif chez un partenaire, ainsi que des informations sur le style de vie.

Livrable
Un notebook contenant : des statistiques descriptives, des visualisations, des légendes et des interprétations sur la façon dont les statistiques et les visualisations sont pertinentes pour expliquer pourquoi les participants acceptent un second rendez-vous.

lien du projet : https://github.com/Thibaut-Longchamps/Certification-Fullstack/tree/main/Analyse%20exploratoire%20descriptive%20et%20inf%C3%A9rentielle%20de%20donn%C3%A9es

---

Description de l'entreprise
AT&T Inc. est une multinationale américaine de télécommunications. C'est la plus grande entreprise de télécommunications au monde en termes de chiffre d'affaires et le troisième fournisseur de services de téléphonie mobile aux USA.

Project **SPAM DETECTOR** *(NLP, Analyse prédictive de données structurées par l'intelligence artificielle)*
L'une des principales difficultés rencontrées par les utilisateurs d'AT&T est l'exposition constante aux messages SPAM.
AT&T a été en mesure de signaler manuellement les messages de spam pendant un certain temps, mais ils sont à la recherche d'un moyen automatisé de détection des spams pour protéger leurs utilisateurs.

Objectifs
Construire un détecteur de spam, qui peut automatiquement signaler les spams en temps réel en se basant uniquement sur le contenu du sms.

Champ d'application du projet
[Télécharger l'ensemble de données](https://full-stack-bigdata-datasets.s3.eu-west-3.amazonaws.com/Deep+Learning/project/spam.csv)

Livrable
livrer un notebook qui exécute un prétraitement et entraîne un ou plusieurs modèles de machine learning afin de prédire la nature spam ou ham du sms. Indiquer clairement les performances obtenues.

lien du projet : https://github.com/Thibaut-Longchamps/Certification-Fullstack/tree/main/Analyse%20pr%C3%A9dictive%20de%20donn%C3%A9es%20non%20structur%C3%A9es%20par%20l'intelligence%20artificielle

---

Description de l'entreprise
Walmart Inc. est une multinationale américaine spécialisée dans la vente au détail, qui exploite une chaîne d'hypermarchés discount aux États-Unis.

Projet **WALMART : prédire les ventes hebdomadaires** *(Analyse prédictive de données structurées par apprentissage automatique)*
Le service marketing de Walmart cherche un modèle de machine learning capable d'estimer les ventes hebdomadaires dans leurs magasins, avec la meilleure précision possible sur les prédictions faites. Un tel modèle les aiderait à mieux comprendre comment les ventes sont influencées par les indicateurs économiques et qui pourrait être utilisé pour planifier de futures campagnes marketing.

Objectifs
Le projet peut être divisé en trois étapes :

- **Partie 1** : réaliser une EDA et tous les prétraitements nécessaires pour préparer les données pour le machine learning
- **Partie 2** : entraîner un modèle de régression linéaire
- **Partie 3** : éviter le surapprentissage en entraînant un modèle de régression régularisé

Portée de ce projet
Pour ce projet, nous avons à disposition un ensemble de données contenant des informations sur les ventes hebdomadaires réalisées par différents magasins Walmart, ainsi que d'autres variables telles que le taux de chômage ou le prix du carburant, qui pourraient être utiles pour prédire le montant des ventes. L'ensemble de données est issu d'un concours Kaggle.

Livrable
Pour mener à bien ce projet, il faudra :
- Créer des visualisations
- Former un modèle de régression linéaire sur l'ensemble de données, qui prédit le montant des ventes hebdos en fonction des autres variables
- Évaluer les performances du modèle en utilisant une métrique pertinente pour les problèmes de régression
- Interpréter les coefficients du modèle pour identifier les caractéristiques importantes pour la prédiction
- Entraîner au moins un modèle avec régularisation (Lasso ou Ridge) pour réduire le surajustement

lien du projet : https://github.com/Thibaut-Longchamps/Certification-Fullstack/tree/main/Analyse%20pr%C3%A9dictive%20de%20donn%C3%A9es%20structur%C3%A9es%20par%20l'intelligence%20artificielle/Walm

---

THE NORTH FACE 🚧 *(Clustering, Analyse prédictive de données non structurées par apprentissage automatique)*
Le service marketing souhaite profiter des solutions de machine learning pour booster les ventes en ligne sur le site : https://www.thenorthface.fr/

Ils ont identifié deux solutions majeures qui pourraient avoir un effet énorme sur les taux de conversion :
- Déployer un système de recommandation qui permettra de suggérer aux utilisateurs des produits complémentaires, similaires aux articles qui les intéressent déjà. Les recommandations pourront être matérialisées par un "vous pourriez également être intéressé par ces produits... " section qui apparaîtrait sur chaque page produit du site Web.
- Améliorer la structure du catalogue de produits grâce à l'extraction de sujets. L'idée est d'utiliser des méthodes non supervisées pour améliorer les catégories existantes : Est-il possible de trouver de nouvelles catégories de produits qui seraient plus adaptées à la navigation sur le site ?

Objectifs 🎯
Le projet peut être découpé en trois étapes :

Etape 1 : Identifier les groupes de produits ayant des descriptions similaires.
Etape 2 : Utiliser les groupes de produits similaires pour créer un algorithme de système de recommandation simple.
Etape 3 : Utiliser des algorithmes de modélisation pour mettre automatiquement en lumière des sujets latents présents dans les descriptions d'éléments.

Dans ce projet, nous travaillons avec un corpus de descriptions d'articles du catalogue de produits de The North Face. Les données peuvent être trouvées ici :

👉👉 https://www.kaggle.com/cclark/product-item-data?select=sample-data.csv

Livrable 📫
- Former au moins un modèle de clustering sur le corpus et afficher des nuages de mots décrivant les clusters
- Développer un code python permettant à un utilisateur de saisir l'identifiant d'un produit qui l'intéresse et obtenir une liste d'articles similaires
- Former au moins un Modèle TruncatedSVD sur le corpus et afficher des nuages de mots décrivant les sujets latents

lien du projet : https://github.com/Thibaut-Longchamps/Certification-Fullstack/tree/main/Analyse%20pr%C3%A9dictive%20de%20donn%C3%A9es%20structur%C3%A9es%20par%20l'intelligence%20artificielle/The%20North%20Face

---

Data Science Weekly

Challenge : CONVERSION RATE 🏆🏆🥇🥇

Description de l'entreprise 📬
www.datascienceweekly.org est une célèbre newsletter créée par des data scientists indépendants qui aborde divers sujets sur de la data science.

Projet Convertion Rate 🚧 *(Analyse prédictive de données structurées par apprentissage automatique)*
Les data scientists qui ont créé la newsletter aimeraient mieux comprendre le comportement des utilisateurs visitant leur site internet. Ils aimeraient savoir s'il est possible de construire un modèle qui prédit si un utilisateur donné s'abonnera à la newsletter, en utilisant seulement quelques informations sur l'utilisateur ; et peut-être découvrir un nouveau levier d'action pour améliorer le taux de conversion de la newsletter. Ils ont conçu un concours visant à construire un modèle permettant de prédire les conversions (c'est-à-dire le moment où un utilisateur s'abonnera à la newsletter). Pour ce faire, ils ont mis à disposition un dataset contenant des données sur le trafic sur leur site Web. Pour évaluer le classement des différentes équipes en compétition, ils ont décidé d'utiliser le f1-score.

Objectifs 🎯
- Partie 1 : réaliser une EDA et les prétraitements nécessaires pour entraîner un modèle de machine learning an utilisant le fichier *data_train.csv*
- Partie 2 : améliorez le f1_score du modèle grâce aux données tests
- Partie 3 : utiliser le modèle pour réaliser des prédictions avec le fichier *data_test.csv,* et transférer les prédictions dans un fichier *.csv* qui sera envoyé à Kaggle (en fait, à votre professeur/TA). Existe-t-il des leviers d'action qui permettraient d'améliorer le taux de conversion de la newsletter ? Quelles recommandations feriez vous à l'équipe ?

Livrable 📫
- Réaliser un EDA
- Former au moins un modèle qui prédit les conversions et évaluer ses performances (f1, matrices de confusion)
- Faire au moins une soumission au classement et donner quelques recommandations pour améliorer le taux de conversion

lien du projet : https://github.com/Thibaut-Longchamps/Certification-Fullstack/tree/main/Analyse%20pr%C3%A9dictive%20de%20donn%C3%A9es%20structur%C3%A9es%20par%20l'intelligence%20artificielle/conversation%2

[Orate]

Description de l'entreprise
Kayak est une plateforme numérique qui permet aux utilisateurs de rechercher, comparer et réserver des vols, des hôtels, des voitures de location...

Projet Kayak *(Contruction et alimentation d'une infrastructure de gestion de donnée)*
L'équipe marketing a besoin d'aide sur un nouveau projet. Après avoir effectué quelques recherches auprès des utilisateurs, l'équipe a découvert que 70 % de leurs utilisateurs qui planifient un voyage aimeraient avoir plus d'informations sur la destination vers laquelle ils se rendent.
De plus, les recherches menées auprès des utilisateurs montrent que les gens ont tendance à se méfier des informations qu'ils lisent s'ils ne connaissent pas la marque qui a produit le contenu. Par conséquent, l'équipe marketing de Kayak souhaite créer une application qui recommandera aux gens où planifier leurs prochaines vacances basée sur des données réelles concernant : la **Météo** et les **hôtels dans la région.** L'application sera capable de recommander les meilleures destinations et hôtels en fonction des variables ci-dessus à tout moment.

Objectifs
Nous ne dispose d'aucune donnée pouvant être utilisée pour créer cette application. Ainsi, le travail consistera à :
- Récupérer les données des destinations
- Obtenir des informations sur les hôtels sur chaque destination
- Stocker toutes les informations ci-dessus dans un lac de données
- Extraire, transformer et charger les données nettoyées de notre datalake vers un entrepôt de données

lien du projet : https://github.com/Thibaut-Longchamps/Certification-Fullstack/tree/main/Contruction%20et%20alimentation%20d'une%20infrastructure%20de%20gestion%20de%20donn%C3%A9e/kayak

---

Scrapper Booking.com

Description de l'entreprise *(Analyse prédictive de données structurées par apprentissage automatique)*
Booking.com est une plateforme de réservation en ligne qui permet aux utilisateurs de rechercher, comparer et réserver des hébergements, des vols, des locations de voitures et d'autres services de voyage dans le monde entier. Étant donné que BookingHoldings ne dispose pas de bases de données agrégées, il sera beaucoup plus rapide d'extraire les données directement de booking.com.

Il est suggéré d'obtenir au moins les informations suivantes :
- le nom de l'Hotel, l'Url vers sa page booking.com, ses coordonnées : latitude et longitude, la note attribuée par les utilisateurs, la description textuelle de l'hôtel
- Créez un data lake avec S3 et stocker le fichier sous forme de fichier csv

ETL
Une fois les données téléchargées sur S3, elles seront extraites et nettoyées à partir d'un data warehouse. Pour ce faire, une base de données SQL sera créée à l'aide d'AWS RDS, permettant de stocker les données extraites de S3.

Livrable
-
un fichier `.csv` dans un bucket S3 contenant des informations enrichies sur la météo et les hôtels de chaque ville française
- une base de données SQL où nous devrions pouvoir obtenir les mêmes données que sur S3
- deux cartes interactives avec le Top-5 des destinations et un Top-20 des hôtels de la région.

lien du projet : https://github.com/Thibaut-Longchamps/Certification-Fullstack/tree/main/Contruction%20et%20alimentation%20d'une%20infrastructure%20de%20gestion%20de%20donn%C3%A9e/booking

---

Description de l'entreprise *(Industrialisation d'un algorithme d'apprentissage automatique et automatisation des processus de décision)*
GetAround est l'Airbnb des voitures. Vous pouvez louer des voitures à n'importe qui pour quelques heures à quelques jours !

Contexte
Lors de la location d'une voiture, les utilisateurs doivent effectuer un enregistrement au début de la location et un paiement à la fin de la location afin:
- D'évaluer l'état de la voiture et informer les autres parties des dommages préexistants ou survenus pendant la location (carburant, kilométrage...)
- Le check-in et le check-out des locations peuvent se faire de trois manières distinctes :

Contrat de location via l'applications : chauffeur et propriétaire se rencontrent et signent tous deux le contrat de location sur le smartphone du propriétaire, ***Connect*** : le conducteur ne rencontre pas le propriétaire et ouvre la voiture avec son smartphone, ***Contrat papier*** : (négligeable)

Projet 🚧
Pour cette étude de cas, nous vous proposons de vous mettre à notre place et de réaliser une analyse que nous avons réalisée en 2017 🔮
Lorsqu'ils utilisent Getaround, les conducteurs réservent des voitures pour une période spécifique, allant d'1h à quelques jours. Ils sont censés ramener la voiture à temps, mais il arrive que les chauffeurs soient en retard au passage en caisse.
Les retours tardifs à la caisse peuvent entraîner des frictions importantes pour le conducteur suivant si la voiture était prévue pour une nouvelle location le même jour, provoquant souvent l'insatisfaction des utilisateurs qui ont dû attendre ou annuler leur location faute de restitution à temps.

Objectifs 🎯
Afin d'atténuer ces problèmes, nous avons décidé de mettre en place un délai minimum entre deux locations. Une voiture ne sera pas affichée dans les résultats de recherche si les heures d'arrivée ou de départ demandées sont trop proches d'une location déjà réservée. Cela résout le problème des départs tardifs, mais peut également nuire aux revenus de Getaround et des propriétaires : nous devons trouver le bon compromis.

Notre chef de produit doit encore décider :
- le seuil : quelle doit être la durée du délai minimum ?
- devrions-nous activer la fonctionnalité pour toutes les voitures ?, uniquement pour les voitures Connect ?

Livrable 📌 **répondre aux questions**
- Quelle part des revenus de notre propriétaire serait potentiellement affectée par cette fonctionnalité ?
- Combien de locations seraient concernées par cette fonctionnalité, en fonction du seuil et de la portée que nous choisissons ?
- À quelle fréquence les conducteurs sont-ils en retard au prochain enregistrement ? Quel impact cela a-t-il sur le prochain conducteur ?
- Combien de cas problématiques résoudra-t-il en fonction du seuil et de la portée choisis ?
- Créer et déployer un tableau de bord qui aidera l'équipe de gestion de produit à répondre aux questions ci-dessus.

De plus, l'équipe Data Science travaille sur l'optimisation des prix. Ils ont rassemblé des données pour suggérer des prix optimaux aux propriétaires de voitures grâce au Machine Learning.

Livrable 📌
- Un tableau de bord en production (accessible via une page web)
- Une API en ligne documentée sur le serveur Heroku (ou tout autre fournisseur de votre choix) contenant au moins un point de terminaison /predict qui respecte la description technique ci-dessus. Nous devrions pouvoir demander le point de terminaison de l'API /predict en utilisant curl :

lien du projet : https://github.com/Thibaut-Longchamps/Certification-Fullstack/tree/main/Industrialisation%20d'un%20algorithme%20d'apprentissage%20automatique%20et%20automatisation%20des%20processus%20de%20d%C3%A9cision/Get%20around

---

 # wildfire-fs-project : Final project of data fullstack JEDHA's bootcamp *(Direction de projets de gestion de données)*

Description:
Wildfire Detection est une application conçue pour aider les utilisateurs à identifier et signaler rapidement les incendies de forêt. Notre application utilise des techniques de pointe en computer vision pour détecter les incendies et la fumée dans les images, fournissant ainsi des infos cruciales aux services d'urgence.

Téléchargement de fichiers : Les utilisateurs peuvent télécharger des images de feux de forêt. ***Entrée caméra*** : Les utilisateurs peuvent utiliser l'appareil photo de leur téléphone pour capturer et analyser des images en temps réel. ***Détection des incendies et des fumées*** : Notre application utilise un modèle YOLOv8 formé sur mesure pour détecter les incendies et la fumée dans une image. ***Géolocalisation*** : Si elle est disponible, l'application affiche les coordonnées GPS de l'image, ce qui aide les services d'urgence à localiser l'incendie. ***Informations météorologiques*** : Les utilisateurs peuvent accéder aux données météorologiques relatives à l'emplacement de l'incendie détecté, y compris la vitesse et la direction du vent. ***Interface conviviale*** : L'interface utilisateur est simple et intuitive

Objectifs 🎯
Cette application accepte en entrée des fichiers au format (png, jpg) et des photos provenant de device. Une fois la sélection confirmée, une carte interactive s'ouvrira avec votre position et des données météorologiques en temps réel pour la zone.

[API with our trained YOLOv8 model] **https://wildfire-project-backend.herokuapp.com == (disabled)**
[Streamlit app with our integrated fire&smoke detector for emergency services] **https://wildfire-project-streamlit.herokuapp.com/ == (disabled)**
link to download the dataset : **https://drive.google.com/drive/folders/1oNRu0h1sXO5HsZAXp5kkHgl7SdTRb4zz?usp=sharing**
link to download the best model PyTorch (best.pt) : **https://drive.google.com/drive/folders/10BBB7h6iRx9Mb9ChAf-tJ4LAD6nj53WE?usp=sharing**

lien du projet : https://github.com/Thibaut-Longchamps/Certification-Fullstack/tree/main/Direction%20de%20projets%20de%20gestion%20de%20donn%C3%A9es