Catalogue de Projets MLOps

Classification supervisée avec datasets Kaggle

Formation Machine Learning & MLOps

Contexte général

Ce document présente un ensemble de projets pédagogiques destinés aux apprenants en **Machine Learning et MLOps**. Chaque projet repose sur un jeu de données réel disponible sur la plateforme **Kaggle** et vise à mettre en pratique :

— L'analyse exploratoire des données (EDA)

— La modélisation supervisée

— Les bonnes pratiques de structuration de projet

— Le versionnement du code avec GitHub

Contraintes techniques

— Langage : **Python**

— Plateforme de versionnement : **GitHub**

— Source des données : **Kaggle**

— Travail individuel ou en binôme

1 Projet 1 – Détection de fraude bancaire

Description

Les transactions frauduleuses représentent un enjeu majeur pour les institutions financières. L'objectif est de concevoir un modèle capable d'identifier automatiquement les transactions suspectes.

Dataset

— **Credit Card Fraud Detection**

— Lien : `https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud`

Richesse pour l'EDA

— Fort déséquilibre des classes

— Variables issues de PCA

— Analyse temporelle

— Détection d'anomalies

Travail attendu

— Analyse exploratoire complète

— Gestion du déséquilibre des classes

— Comparaison de plusieurs modèles

— Analyse des faux négatifs

2 Projet 2 – Prédiction du churn client

Description

La perte de clients (churn) est un problème critique dans les télécommunications. Le projet vise à prédire quels clients sont susceptibles de quitter l'opérateur.

Dataset

— **Telco Customer Churn**

— Lien : `https://www.kaggle.com/datasets/blastchar/telco-customer-churn`

Richesse pour l'EDA

— Variables numériques et catégorielles

— Analyse comportementale

— Étude des contrats et paiements

Travail attendu

— Feature engineering

— Encodage catégoriel

— Modèles interprétables

3 Projet 3 – Détection de défauts industriels

Description

Dans un contexte industriel, détecter les produits défectueux en amont permet de réduire les coûts de production et d'améliorer la qualité.

Dataset

— **Bosch Production Line Performance**

— Lien : `https://www.kaggle.com/c/bosch-production-line-performance`

Richesse pour l'EDA

— Plus de 900 variables

— Données bruitées et manquantes

— Problématique de haute dimension

Travail attendu

— Sélection de variables

— Modèles robustes

— Analyse des performances

4 Projet 4 – Diagnostic médical assisté

Description

L'objectif est d'aider au diagnostic médical à partir de données cliniques tout en tenant compte des enjeux d'interprétabilité.

Datasets

— Heart Disease : `https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset`

Richesse pour l'EDA

— Analyse démographique

— Corrélations cliniques

— Étude des faux négatifs

5 Projet 5 – Détection de fake news

Description

La propagation de fausses informations est un enjeu sociétal majeur. Ce projet vise à classifier automatiquement des articles de presse.

Dataset

— **Fake News Detection**

— Lien : `https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset`

Richesse pour l'EDA

— Analyse lexicale

— Longueur des textes

— Distribution des classes

6 Projet 6 – Scoring de risque de crédit

Description

Les banques évaluent le risque de défaut avant d'accorder un crédit. L'objectif est de prédire la probabilité de non-remboursement.

Dataset

— **Home Credit Default Risk**

— Lien : `https://www.kaggle.com/c/home-credit-default-risk`

Richesse pour l'EDA

— Variables socio-économiques

— Valeurs manquantes complexes

— Analyse des biais

Structure GitHub attendue

```
project-mlops/
 data/
 notebooks/
```

```
src/
requirements.txt
README.md
.gitignore
```