Sujet Examen
prédire la réussite scolaire

L’échec scolaire constitue un défi majeur pour les systèmes éducatifs, ayant des répercussions directes sur les trajectoires sociales, économiques et professionnelles des élèves. La capacité à identifier précocement les élèves à risque de non-validation de leur année peut permettre de déclencher des actions ciblées d'accompagnement pédagogique, psychologique ou social.

Grâce à l’essor de l’intelligence artificielle et du machine learning, il est désormais possible d’analyser des données issues de la scolarité des élèves pour anticiper les situations d’échec. Ces approches s’appuient sur des données académiques, sociales et comportementales, et permettent de concevoir des outils d’aide à la décision pour les équipes éducatives.

Pour illustrer cela, deux bases de données publiques issues d’écoles portugaises vous sont données. Elles contiennent des informations sur les élèves, leur environnement familial, leurs performances scolaires sur les deux premiers trimestres et d’autres variables contextuelles. L’enjeu consiste à exploiter ces données pour **prédire la réussite ou non d’un élève** au troisième trimestre, tout en respectant les contraintes réglementaires et éthiques liées à l’usage de données personnelles sensibles.

Sujet du projet :

Construire et déployer un modèle d'IA prédictif du passage d'un élève au dernier trimestre

Dans ce projet, vous incarnerez un·e professionnel·le de la Data missionné·e pour concevoir une solution prédictive visant à estimer si un·e élève est en situation de réussite ou d'échec à la fin de son année scolaire. Le but est de produire une solution IA robuste, explicable, conforme aux bonnes pratiques d'ingénierie des données et d'éthique.

Vous disposez de deux jeux de données distincts issus d'établissements scolaires portugais, décrivant chacun le parcours d'élèves suivant soit un cursus en mathématiques (`student-mat.csv`), soit un cursus en portugais (`student-por.csv`). Ces jeux de données présentent des caractéristiques sociodémographiques, académiques et familiales pour chaque élève, ainsi que leurs notes trimestrielles (G1, G2, G3).

Le jeu de données final résulte de la **fusion des deux jeux**, en **respectant les bonnes pratiques** de traitement des doublons, d'identification de la provenance, et de nettoyage de la qualité des données. Une colonne `source` sera utilisée pour indiquer si l'observation provient du fichier mathématiques ou portugais.

https://archive.ics.uci.edu/dataset/320/student+performance

Objectif principal

L'objectif est de **prédire si un élève réussira ou non son année**, à partir des informations connues en amont du dernier trimestre. Cette tâche peut être abordée :

- **En régression**, en prédisant la note finale `G3` sur 20.
- **En classification**, en définissant une cible binaire (`1` si `G3 ≥ 10`, `0` sinon).

Dans chaque cas, vous devrez **entraîner plusieurs modèles**, comparer leurs performances avec des **métriques appropriées** (RMSE et R^2 pour la régression ; accuracy et F1-score pour la classification), et analyser la robustesse de vos résultats grâce à la **validation croisée**.

Analyse comparée des scénarios

Un point essentiel de votre mission sera de **comparer les performances obtenues selon différents scénarios d'entraînement**, afin d'estimer l'impact des données sensibles ou indirectement révélatrices de biais :

1. **Scénario 1 – Toutes les variables disponibles**
   Vous utilisez toutes les colonnes disponibles, y compris les informations personnelles (genre, statut familial, métier des parents, etc.), ainsi que les notes des deux premiers trimestres.
2. **Scénario 2 – Sans variables sensibles**
   Vous retirez les variables que vous considérez comme éthiquement discutables en vous justifiant
3. **Scénario 3 – Sans variables sensibles + sans G2**
   Vous retirez en plus la note du second trimestre (`G2`), souvent très corrélée à `G3`, pour simuler un cas d'usage où cette note n'est pas encore disponible.

4. **Scénario 4 – Sans variables sensibles + sans G1 et G2**
   Ce scénario le plus épuré vous permet de tester la faisabilité d'un modèle basé uniquement sur des données contextuelles, sociales et comportementales disponibles avant toute note.

Votre objectif est de **documenter et commenter l'impact de chaque scénario sur les performances des modèles**, et d'**argumenter sur le choix du modèle et des données** à utiliser dans un contexte réel, en tenant compte des enjeux **éthiques, légaux et de robustesse**.

Phase d'industrialisation –

Mise en production d'une application IA

Une fois les modèles de prédiction sélectionnés et validés, vous passerez à la phase d'**industrialisation de la solution IA**, en concevant une application complète accessible à un utilisateur non technique. Cette application sera composée de deux éléments principaux : une **API de prédiction** et une **interface utilisateur graphique**.

L'**API** permettra de mettre à disposition le modèle d'IA sous forme de service, que l'on pourra interroger via des requêtes HTTP POST. Elle devra inclure un mécanisme de chargement du modèle entraîné, et exposer au minimum une route pour prédire la réussite scolaire d'un élève à partir de ses caractéristiques, une route d'entraînement monitoré et une route pour mesurer la santé de votre API. Vous veillerez à **journaliser chaque requête**, en sauvegardant les entrées, les sorties, les dates d'inférence, et idéalement l'utilisateur (ou ID de session). Une attention particulière sera portée à la **gestion des erreurs**, à la **validation des données entrantes**, et à la **robustesse de l'API**.

L'**interface graphique** devra permettre à un utilisateur (enseignant, conseiller pédagogique, etc.) de renseigner les caractéristiques d'un élève via un formulaire simple, d'obtenir une prédiction claire et interprétable, et de consulter un historique des inférences si possible.

Par ailleurs, vous intégrerez des outils de **suivi de modèle** avec MLflow pour tracer les versions du modèle, ses hyperparamètres, ses performances, et permettre un **versionning contrôlé**. Enfin, un **monitoring de l'API** (temps de réponse, erreurs, uptime kuma) pourra être mis en place à l'aide de solutions comme Prometheus + Grafana ou simplement via des logs et alertes automatisées.

Le projet devra intégrer une démarche **CI/CD** complète, en utilisant **GitHub Actions** pour automatiser les étapes clés du cycle de vie du code : tests, linting, entraînement du modèle et déploiement. À chaque mise à jour de la branche principale, une image **Docker** de l'API sera construite automatiquement et déployée sur l'infrastructure cible, garantissant un déploiement reproductible, traçable et conforme aux bonnes pratiques d'ingénierie IA.

Cette phase vise à valider votre capacité à **transformer un modèle de data science en une solution exploitable en environnement réel**, dans le respect des contraintes de fiabilité, d'éthique et de gouvernance des données.

Données personnelles et familiales

Colonne	Description
`school`	École fréquentée : `GP` (Gabriel Pereira) ou `MS` (Mousinho da Silveira)
`sex`	Sexe de l'élève : `F` (fille) ou `M` (garçon)
`age`	Âge de l'élève (entre 15 et 22 ans)
`address`	Type d'habitation : `U` (urbain) ou `R` (rural)
`famsize`	Taille de la famille : `LE3` (≤3) ou `GT3` (>3) personnes
`Pstatus`	Statut marital des parents : `T` (ensemble) ou `A` (séparés)
`Medu`	Niveau d'éducation de la mère (0 à 4)
`Fedu`	Niveau d'éducation du père (0 à 4)
`Mjob`	Profession de la mère (enseignant, santé, services, à domicile, autre)

`Fjob`	Profession du père (idem `Mjob`)
`reason`	Raison du choix de l'école : proximité, réputation, etc.
`guardian`	Tuteur/tutrice principal(e) : mère, père ou autre

Contexte scolaire et extra-scolaire

Colonne	Description
`traveltime`	Temps de trajet domicile-école (1= <15 min, 4= >1h)
`studytime`	Temps d'étude hebdo (1= <2h, 4= >10h)
`failures`	Nb d'échecs passés (1 si jamais redoublé, 4 si 3 redoublements ou plus)
`schoolsup`	Soutien scolaire supplémentaire (yes/no)
`famsup`	Soutien familial (yes/no)
`paid`	Cours payants en maths (yes/no)
`activities`	Activités extrascolaires (yes/no)

`nursery`	Fréquentation d'une maternelle (yes/no)
`higher`	Souhait de poursuivre des études supérieures (yes/no)
`internet`	Accès à internet à la maison (yes/no)
`romantic`	En couple ou non (yes/no)

Vie quotidienne

Colonne	Description
`famrel`	Qualité des relations familiales (1= très mauvaise, 5= excellente)
`freetime`	Temps libre après l'école (1 à 5)
`goout`	Fréquence des sorties avec des amis (1 à 5)
`Dalc`	Consommation d'alcool en semaine (1= très faible, 5= très élevée)
`Walc`	Consommation d'alcool le week-end (1 à 5)

`health`	État de santé (1= très mauvais, 5= excellent)
`absences`	Nombre d'absences scolaires

Résultats scolaires

Colonne	Description
`G1`	Note au premier trimestre (0 à 20)
`G2`	Note au second trimestre (0 à 20)
`G3`	Note finale (0 à 20) — *variable cible*
`source`	Source du fichier : `mat` ou `por` (mathématiques ou portugais)