2023-2024

RAPPORT – PROJET HACKATHON

HAJJOU Doha

ALBEKBASHY Rahma

SOMMAIRE

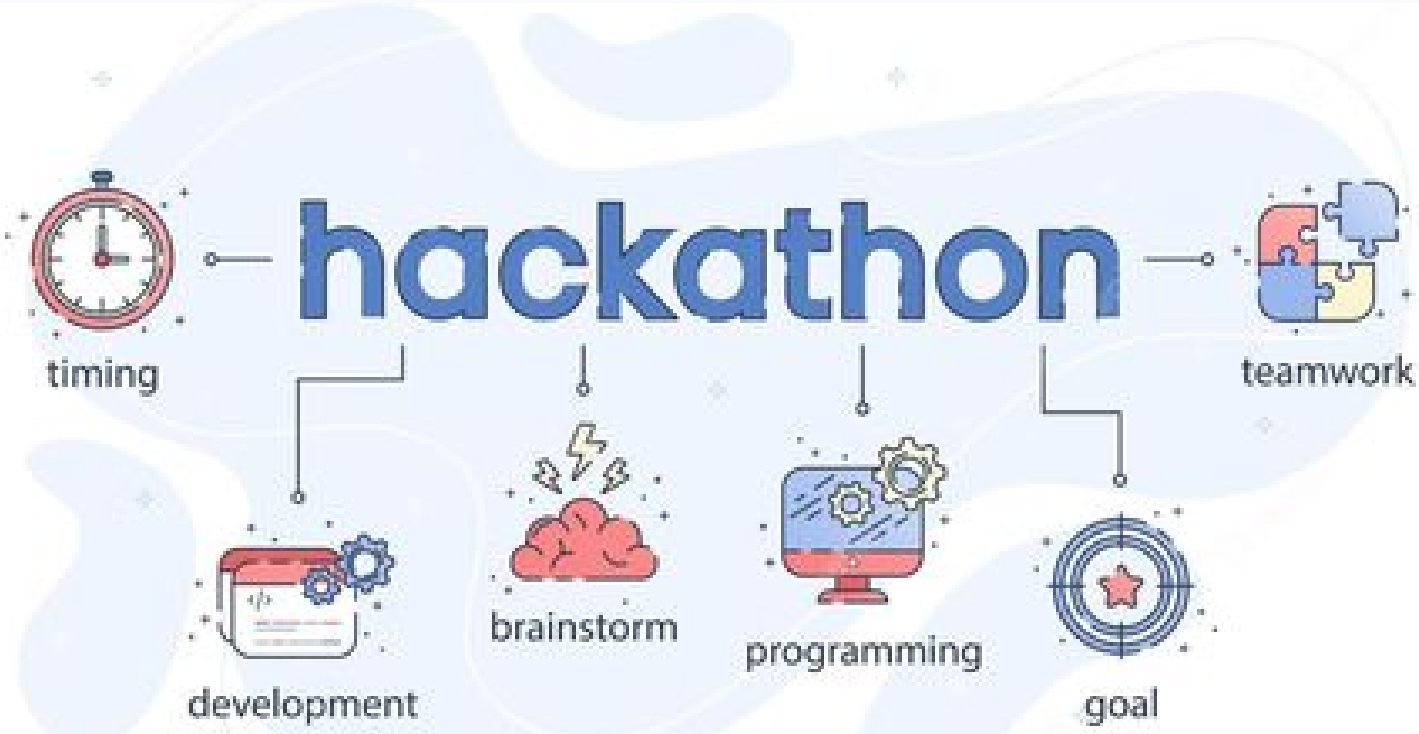
INTRODUCTION

Bienvenue dans cette passionnante aventure intellectuelle qu'est notre Hackathon dédié à l'étude de cas "Environnement logiciel professionnel reconstitué". Ce projet novateur vise à offrir une expérience immersive et pratique dans le domaine de l'informatique, mettant particulièrement l'accent sur deux aspects cruciaux : l'environnement logiciel professionnel et l'évaluation des compétences acquises.

Dans la première phase de notre Hackathon, nous plongerons dans le monde captivant des outils informatiques. Notre objectif est de permettre la manipulation de divers types d'outils, allant des instruments d'extraction, de transformation et de stockage de données aux logiciels spécialisés dans le traitement de données massives. Cette approche pratique nous offrira une compréhension approfondie des technologies actuelles et renforcera nos compétences dans ces domaines essentiels.

La seconde étape de notre Hackathon, tout aussi cruciale, se concentrera sur l'évaluation des compétences acquises. Nous aurons l'opportunité de présenter notre travail au cours d'une soutenance orale et ce document écrit servira à consolider nos connaissances, à présenter les analyses et à démontrer notre compréhension des concepts abordés tout au long du Hackathon.

En résumé, notre projet Hackathon nous offre une opportunité unique d'explorer, d'apprendre et de mettre en pratique des compétences dans un environnement stimulant et axé sur le monde professionnel.



CONTEXT

Le jeu de données intitulé "Étudiants" compile diverses informations relatives à la performance académique des étudiants.

Il inclut des données démographiques telles que l'âge et le sexe de l'étudiant. La catégorie 'Type de Lycée' classe le type d'établissement fréquenté, tandis que la colonne 'Bourse' indique si l'étudiant bénéficie d'une aide financière.

Des détails sur le 'Travail Supplémentaire' et la participation à des 'Activités Sportives' fournissent des perspectives sur les engagements parascolaires.
Chaque entrée est distinguée de manière unique par un identifiant ('Id').

L'attribut 'Transport' décrit le mode de déplacement de chaque étudiant. Les aspects académiques sont saisis à travers les 'Heures d'Étude Hebdomadaires', l'assiduité, ainsi que les évaluations de la 'Lecture', des 'Notes' et de l''Écoute en Classe'.

La synthèse de ces éléments se reflète dans la colonne des 'Notes', offrant une vision globale de la performance étudiante.

Ce jeu de données représente une ressource précieuse pour explorer les dynamiques variées qui influent sur les résultats académiques.

VISUALISATION DES DONNEES

L'objectif principal de cette visualisation des données est de fournir une compréhension visuelle des caractéristiques du jeu de données, de mettre en évidence des tendances, des disparités ou des relations entre différentes variables. Cela facilite l'interprétation des données et peut aider à prendre des décisions éclairées en fonction de ces observations.

- **Diagramme en Barres (Âge des étudiants, Type de lycée, Répartition par genre, Présence en classe, Répartition des notes, Travail supplémentaire) :**

  - **Type de Visualisation :** Le diagramme en barres représente des catégories distinctes sur l'axe des x et utilise des barres verticales pour indiquer la fréquence ou la distribution de chaque catégorie.
  - **Objectif :** Cela permet de visualiser la distribution des données catégoriques, telles que le type de lycée, le genre, la présence en classe, les notes, et la participation au travail supplémentaire. On peut facilement comparer la fréquence des différentes catégories et observer des tendances ou des disparités.

  Âge des étudiants :
  - Utilisation d'un diagramme en barres pour représenter l'âge des étudiants.
  - Chaque barre représente un étudiant identifié par son 'Id'.
  - L'axe des x affiche les identifiants d'étudiants, et l'axe des y affiche l'âge correspondant.

  Type de lycée :
  - Utilisation d'un diagramme en barres pour représenter la répartition des étudiants en fonction du type de lycée fréquenté.
  - L'axe des x affiche les catégories de type de lycée, et l'axe des y affiche le nombre d'étudiants.

  Répartition par genre :
  - Utilisation d'un diagramme en barres pour montrer la répartition des étudiants par genre (Male ou Female).
  - L'axe des x affiche les catégories de genre, et l'axe des y affiche le nombre d'étudiants.

  Présence en classe :
  - Utilisation d'un diagramme en barres pour représenter la fréquence de la présence en classe.
  - Les catégories peuvent être, par exemple, 'Present' ou 'Absent'.

  Répartition des notes :
  - Utilisation d'un diagramme en barres pour représenter la distribution des notes des étudiants.
  - Les catégories de notes sont affichées sur l'axe des x, et le nombre d'étudiants ayant chaque note est affiché sur l'axe des y.

  Travail supplémentaire :
  - Utilisation d'un diagramme en barres pour montrer la répartition des étudiants en fonction de leur engagement dans un travail supplémentaire.
  - Les catégories peuvent inclure 'Oui' ou 'Non'

- **Diagramme Circulaire (Participation à des activités sportives) :**
  - **Type de Visualisation** : Le diagramme circulaire, également appelé camembert, représente la distribution des données en pourcentages autour d'un cercle.
  - **Objectif** : Il est utile pour illustrer la proportion des valeurs relatives à une variable binaire, comme la participation ou non à des activités sportives. Il offre une vue rapide des proportions relatives des catégories.

    Participation à des activités sportives :
    - Utilisation d'un diagramme circulaire (camembert) pour montrer la répartition des étudiants selon leur participation à des activités sportives (True ou False).
    - La taille des parts du camembert est déterminée par la fréquence de chaque catégorie.

- **Histogramme (Heures d'étude hebdomadaires) :**
  - **Type de Visualisation :** L'histogramme représente la distribution des données continues en regroupant les valeurs en intervalles sur l'axe des x et en affichant la fréquence sur l'axe des y.
  - **Objectif :** Il permet d'observer la répartition des heures d'étude hebdomadaires de manière continue. On peut identifier des tendances de concentration ou de dispersion dans les données.

    Heures d'étude hebdomadaires :
    - Utilisation d'un histogramme pour illustrer la distribution des heures d'étude hebdomadaires.
    - Le nombre d'heures est regroupé en 20 intervalles (bins).
    - Les axes x et y sont étiquetés, et le graphique comporte un titre.

NETTAOYEGE DES DONNEES

Cette étape est cruciale dans l'analyse de données, car des données propres et fiables sont essentielles pour obtenir des résultats précis et significatifs. Voici une explication détaillée des principales étapes du nettoyage des données :

- **Suppression des Valeurs Manquantes :** Identification et gestion des données manquantes, que ce soit en les supprimant, en les remplaçant par des valeurs appropriées ou en utilisant des techniques d'imputation.

- **Gestion des Duplicatas :** Détection et élimination des enregistrements en double pour éviter toute redondance et assurer l'exactitude des analyses.

- **Correction des Erreurs de Saisie :** Identification et correction des erreurs de saisie, telles que les fautes de frappe, les incohérences ou les valeurs aberrantes, pour garantir la cohérence des données.

- **Normalisation des Formats :** Uniformisation des formats de données pour faciliter la comparaison et l'analyse. Par exemple, la conversion de scholarship en floattante pas string.

- **Gestion des Incohérences :** Identification et résolution des incohérences logiques ou des contradictions dans les données, assurant ainsi la fiabilité des informations.

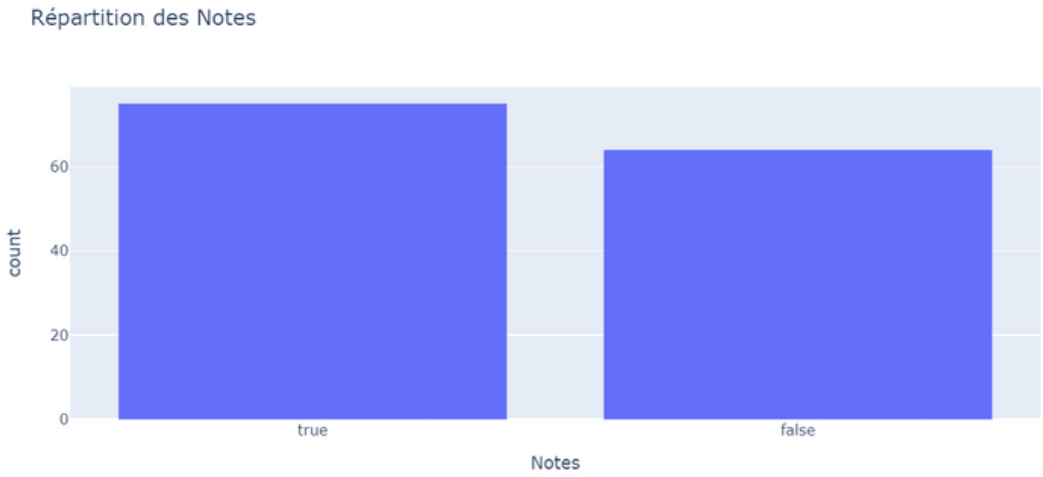
- **Transformation des Données :** Modification ou conversion des données selon les besoins de l'analyse, comme la création de nouvelles variables, la standardisation des unités, ou l'agrégation de catégories.

- **Vérification de la Cohérence avec le Domaine :** Validation des données par rapport à la connaissance du domaine, en s'assurant qu'elles correspondent aux attentes et aux règles spécifiques au contexte.

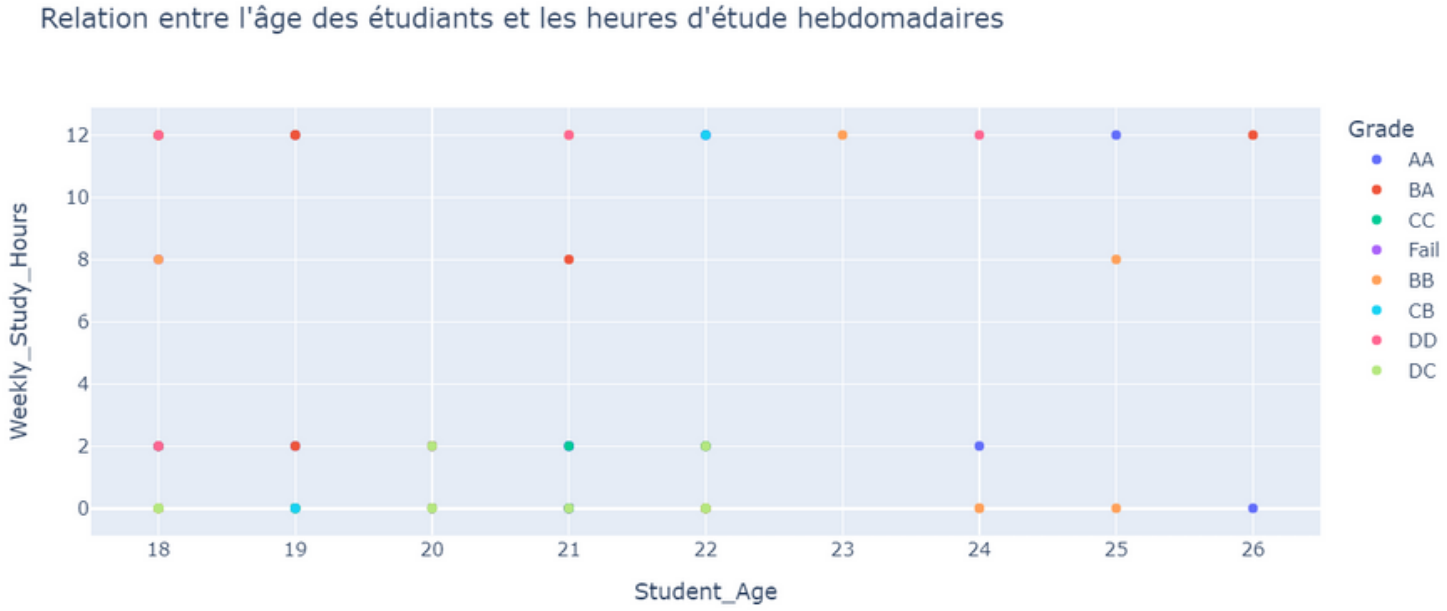
- **Séparation des Données :** Division des données en ensembles d'apprentissage, de validation et de test pour évaluer et améliorer les performances des modèles d'apprentissage automatique.

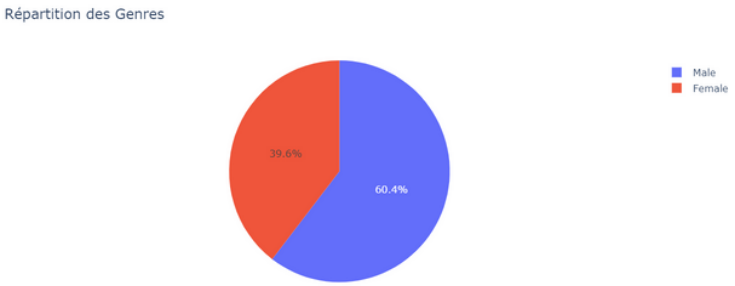
En résumé, le nettoyage des données vise à garantir la qualité, la cohérence et la fiabilité des données, créant ainsi une base solide pour des analyses précises et des résultats pertinents dans le domaine de l'apprentissage automatique, de l'exploration de données ou d'autres applications analytiques.

COMMUNICATION INFOGRAPHIQUE VISUELLE



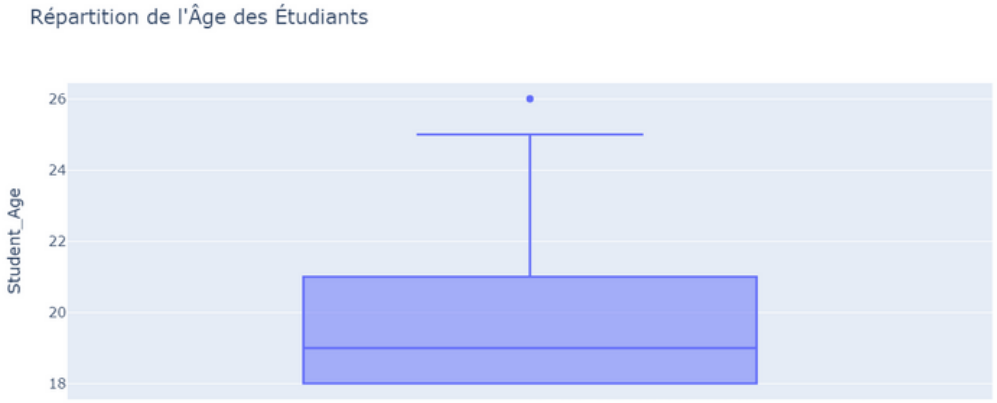
On a considéré les élèves qui ont eu 6 comme 'False' vue qu'il valide pas et comme vous pouvez voir la plupart de la classe a validé



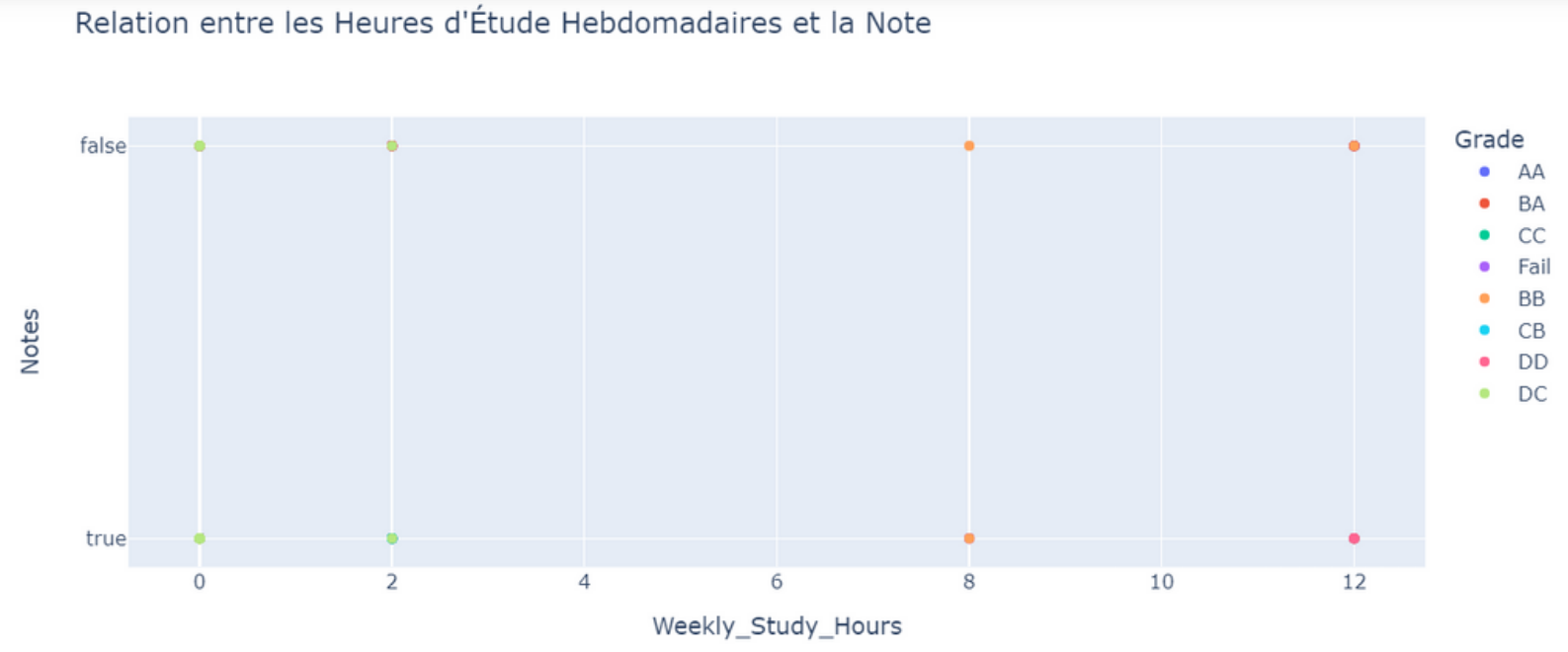


Analyse de la Répartition des Genres:
On peut voir quela plupart de la classe est des hommes

COMMUNICATION INFOGRAPHIQUE VISUELLE



Analyse de la Répartition de l'Âge des Étudiants :
On a bien noté que la plupart de la classe est de 19 ans et le plus agée est de 26 ans



Analyse de la Relation entre les Heures d'Étude Hebdomadaires et la Note :
on peut voir les notes des étudiants selon leur heures de travail par semaine

MODELE MACHINE LEARNING

Le Random Forest Classifier est un algorithme d'apprentissage automatique qui appartient à la famille des ensembles d'arbres de décision. Il combine plusieurs modèles d'arbres de décision pour créer un modèle plus robuste et généralisable.

```
Précision du modèle : 32.14%

Il semble que la précision du modèle RandomForestClassifier ne soit pas satisfaisante. Essayons un autre modèle, comme le Support Vector Classifier (SVC).
```

L**e modèle SVC (Support Vector Classifier)**, également connu sous le nom de SVM (Support Vector Machine) dans le contexte de la classification, est un algorithme d'apprentissage automatique qui s'appuie sur des concepts de géométrie euclidienne et d'optimisation pour séparer des classes dans un espace de caractéristiques.

```
Précision du modèle SVC : 39.29%
```

Le modèle de régression logistique est spécifiquement conçu pour des problèmes de classification, et non pour des tâches de régression. Cette conception est basée sur la nature de la fonction logistique utilisée dans le modèle. La fonction logistique est relativement moins sensible aux valeurs extrêmes que certaines autres fonctions utilisées dans des modèles de régression, ce qui rend le modèle plus robuste dans le contexte de la classification.

```
Précision du modèle de régression logistique : 25.00%
```

Le Gradient Boosting est une technique d'ensemble utilisée en apprentissage automatique pour améliorer la précision des modèles prédictifs. C'est une méthode de boosting, ce qui signifie qu'elle construit un modèle fort en combinant plusieurs modèles plus faibles. Le Gradient Boosting est particulièrement puissant pour les problèmes de régression et de classification.

```
Précision du modèle Gradient Boosting : 46.43%

Avec le Gradient Boosting ca augmente scikit-learn pour un model simple
```

CONCLUSION

À l'issue de notre Hackathon dédié à l'étude de cas "Environnement logiciel professionnel reconstitué", nous clôturons un chapitre intense et instructif.

Cette expérience, bien au-delà d'une simple exploration d'outils informatiques, a représenté une plongée approfondie dans le processus de nettoyage des données, de visualisation et de prédiction par le biais de modèles de machine learning.

Le nettoyage des données a posé les bases, soulignant l'importance cruciale de commencer avec des données fiables. La visualisation a ensuite transformé ces données brutes en insights visuels, offrant un moyen puissant de communiquer nos découvertes de manière claire et persuasive.

Enfin, l'application de modèles de machine learning a couronné notre Hackathon en démontrant notre capacité à transformer des données en prédictions significatives.
Au-delà des compétences techniques, la collaboration et la communication au sein de l'équipe ont été des éléments clés de notre succès. Chacun a apporté une contribution unique, enrichissant ainsi l'expérience collective.

En conclusion, ce Hackathon représente bien plus qu'une simple série de tâches techniques. Les compétences acquises dans le nettoyage des données, la visualisation et les prédictions via des modèles de machine learning sont désormais des atouts précieux pour nos futures entreprises. Ce projet restera une étape marquante, attestant de notre capacité à exceller dans des scénarios professionnels simulés, aussi complexes qu'enrichissants.