Abdallah Remmide

Serigne Fall

**Rapport : SAE 1.02** - Ecriture et Lecture de fichiers de données

Pour commencer, nous avons chargé les deux fichiers CSV dans des DataFrames. Cela nous a permis de manipuler facilement les données et de préparer les ajustements nécessaires. En parallèle, nous avons mis en place un journal pour noter toutes les modifications, ce qui nous aide à garder une trace précise de notre travail.

Ensuite, nous avons traité les lignes avec des informations manquantes dans des colonnes essentielles comme « Diagnosis » et « TreatmentCost ». Plutôt que de conserver ces lignes incomplètes, nous avons choisi de les supprimer pour garantir la fiabilité des résultats à venir. Après cela, nous avons pris le temps de nettoyer la colonne « Diagnosis » en la mettant en majuscules et en enlevant les espaces inutiles, ce qui facilite grandement les comparaisons.

En ce qui concerne les dates, nous avons normalisé les colonnes « DOB » (date de naissance) et « LastVisit » (dernière visite) pour les convertir dans un format standard « DD/MM/YYYY ». Cette cohérence est essentielle pour pouvoir exploiter les données correctement.

Pour les valeurs manquantes dans certaines colonnes, nous avons adopté des solutions simples. Par exemple, les diagnostics manquants ont été remplacés par « Inconnu », tandis que les coûts des traitements ont été remplis avec la moyenne calculée. De cette manière, nous avons évité de perdre trop d'informations tout en conservant des données fiables. De plus, pour la colonne « LastVisit », nous avons ajouté une date par défaut (« 31/12/2024 ») lorsque l'information était absente.

Ensuite, nous avons ajouté deux colonnes importantes. La première, « Age », a été calculée à partir de la date de naissance et permet d'avoir une vision claire de l'âge des patients. La deuxième, « Statut », indique si un patient est considéré comme « Sain » ou

« Malade » en fonction de son diagnostic. Ces colonnes nous permettent de mieux structurer l'information et d'améliorer la lisibilité des données.

Une fois que ces préparations ont été terminées, nous avons réuni les deux jeux de données en utilisant la colonne « PatientID » pour effectuer une fusion. Cette opération a permis de regrouper toutes les informations disponibles sur chaque patient. Les doublons issus de cette fusion ont été gérés en combinant les colonnes similaires, comme les noms, dates de naissance et diagnostics. Enfin, nous avons supprimé les colonnes en doublon pour ne garder que les informations essentielles.

Pour finir, nous avons sauvegardé les fichiers nettoyés et fusionnés. Cette sauvegarde nous garantit que tout le travail réalisé est conservé.

En conclusion, toutes ces étapes nous ont permis de transformer des données brutes et parfois incohérentes en un jeu de données complet et prêt à l'emploi.