Projet modèles statistiques

'Diabetes Dataset'

Yassmine TAII TAHIRI, Hadil CHTIOUI,
Omar El Farouk CHTIOUI, Nada BOUSSETTA

16 Mai 2025

Table des matières

1 Introduction

1.1 Contexte / Description du jeu de données

Ce projet s'inscrit dans le cadre de la matière Modèles statistiques, dont l'objectif est de mettre en pratique les techniques statistiques explorées en cours à travers l'analyse d'un jeu de données réel.

Le dataset regroupe des informations médicales et démographiques sur environ 100 000 individus. Il contient 16 variables :

- **year** : Année de collecte des données.
- **gender** : Sexe de l'individu (masculin ou féminin ou autre).
- **age** : Âge de la personne en années.
- **location** : Région géographique (son pays).
- **race :AfricanAmerican** : Indicateur binaire pour l'appartenance à la population afro-américaine.
- **race :Asian** : Indicateur binaire pour l'appartenance à la population asiatique.
- **race :Caucasian** : Indicateur binaire pour l'appartenance à la population caucasienne.
- **race :Hispanic** : Indicateur binaire pour l'appartenance à la population hispanique.
- **race :Other** : Indicateur binaire pour toute autre origine ethnique non spécifiée.
- **hypertension** : Présence (1) ou absence (0) d'hypertension artérielle.
- **heart_disease** : Présence (1) ou absence (0) de maladie cardiaque.
- **smoking_history** : Antécédents tabagiques de la personne (ex : jamais fumé, ancien fumeur, fumeur actuel).
- **bmi** : Indice de masse corporelle, indicateur de corpulence.
- **hbA1c_level** : Taux d'HbA1c sanguin, (hémoglobine glyquée, mesure du pourcentage d'hémoglobine dans le sang qui est liée au glucose) reflétant la glycémie moyenne sur plusieurs mois.
- **blood_glucose_level** : Taux de glucose sanguin à un instant donné.
- **diabetes** : Variable cible indiquant si l'individu est diabétique (1) ou non (0).

1.1.1 Obtention du jeu de données

Le jeu de données utilisé dans ce projet a été téléchargé depuis Kaggle, une plateforme populaire de partage de jeux de données et de concours de science des données. Le fichier est au format .csv et a été importé localement dans notre environnement R.

Voici le lien :
https://www.kaggle.com/datasets/priyamchoksi/100000-diabetes-clinical-dataset/code

1.2 Description de la question

L'objectif principal de cette étude est de répondre à la question suivante :

Quels sont les facteurs les plus associés à la présence de diabète chez un individu ?

En d'autres termes, nous chercherons à comprendre quelles variables influencent le plus le statut diabétique.

2 Méthodologie

2.1 Procédures de nettoyage des données

Avant de procéder à toute analyse, nous avons effectué plusieurs opérations de prétraitement :

- Suppression des doublons : élimination des lignes identiques dans le jeu de données afin de garantir l'unicité des observations.
- Gestion des valeurs manquantes : identification des variables contenant des valeurs manquantes. Certaines observations imprécises ou inutilisables (par exemple `No Info` dans la variable `smoking_history`) ont été exclues pour ne pas fausser l'analyse.
- Filtrage des valeurs aberrantes ou peu représentatives : par exemple, les individus identifiés avec le genre "Other" ont été retirés, car ils représentaient une minorité marginale sans cas de diabète, ce qui pourrait biaiser l'analyse.

2.2 Workflow scientifique

Notre méthodologie s'est articulée en plusieurs étapes successives :

- Analyse exploratoire globale : identification des variables clés et compréhension de la structure du dataset.
- Étude univariée : analyse de la variable cible (`diabetes`) en examinant sa distribution.
- Étude bivariée : exploration de la relation entre le diabète et les autres variables (âge, sexe, BMI, hypertension, etc.).
- Matrice de corrélation : évaluation statistique des liens entre variables numériques, en particulier celles associées au diabète.
- Analyse graphique croisée : production de visualisations pour explorer les interactions entre plusieurs facteurs chez les individus diabétiques.

2.3 Choix de représentation des données

— **Variables numériques :** Les colonnes telles que `age`, `BMI`, `blood_glucose_level`, `HbA1c_level`, etc., étaient déjà en format numérique. Aucun traitement particulier n'a été nécessaire pour permettre leur utilisation dans les calculs statistiques ou les visualisations (comme les boxplots et diagrammes de dispersion).

— **Variables catégorielles :** Les colonnes comme `gender`, `smoking_history`, `race`, `hypertension`, `heart_disease` et `diabetes` contiennent des valeurs textuelles ou binaires. Ces colonnes ont été utilisées telles quelles dans les visualisations (ex : diagrammes en barres), sans transformation explicite en facteur. Par exemple, les colonnes `hypertension` et `heart_disease` prennent les valeurs 0 ou 1, indiquant respectivement l'absence ou la présence de ces conditions chez une personne. Elles ont été utilisées pour explorer leur lien éventuel avec la variable `diabetes` (également codée en 0 ou 1).

— **Visualisations :** Les graphiques ont été réalisés avec la librairie `ggplot2`, ce qui a permis une représentation claire et efficace des données, sans nécessiter de transformations supplémentaires. Par exemple, des boxplots (documentation R) ont été utilisés pour comparer les distributions numériques selon la présence ou non du diabète.

3 Analyse en programmation littérale

L'analyse exploratoire des données a été réalisée en langage `R`, en utilisant un style de programmation littérale via un document RMarkdown. Cela a permis d'intégrer de manière fluide le code, les sorties graphiques et les commentaires d'interprétation dans un seul document reproductible.

Les visualisations ont été réalisées avec la librairie `ggplot2`, en générant notamment :

— des histogrammes et boxplots pour visualiser la distribution des variables numériques ;

— des diagrammes en barres pour explorer la répartition des classes (ex. : diabète selon le genre ou l'historique de tabagisme).

L'utilisation de la programmation littérale a permis de commenter directement chaque graphique et chaque étape du traitement dans un flux clair et cohérent.

4 Conclusion

L'analyse exploratoire des données réalisée en programmation littérale a permis de mettre en évidence plusieurs relations intéressantes entre certaines variables et la présence de diabète. On observe notamment que des variables telles que le taux de glucose sanguin (`blood_glucose_level`) et le taux d'HbA1c (`HbA1c_level`) sont significativement plus élevés chez les individus diabétiques, ainsi que l'age, l'hypertention et la BMI, ce qui confirme leur rôle central dans le diagnostic de la maladie.

Grâce à l'utilisation de RMarkdown, nous avons pu combiner efficacement code, commentaires et visualisations dans un même document. Cela a permis une meilleure traçabilité des étapes de nettoyage, d'exploration et de visualisation des données.

5 Références

— Dataset source : https://www.kaggle.com/datasets/priyamchoksi/100000-diabetes-clinical-dataset/code

— Recherche sur les facteurs favorisant la diabète : `https://sante.gouv.fr/IMG/pdf/facteurs_et_marqueurs_de_risque_diabete.pdf`

— Documentation R : `https://www.r-project.org/other-docs.html`