

Groupe : DJIEUNANG NOUMBO NELIE Mabelle, GABRIEL-ATANGANA MBOA Bryan , TAMBAT
Tresor Megane

du 30 Septembre au 16 Octobre 2024

UQO - Introduction à la Science de données - projet Automne 2024

Analyse des indicateurs de santé liés au diabète basé sur nos données afin de trouver les causes du diabète et apprendre à entraîner un modèle pour prédire si quelqu'un a du diabète ou non se basant sur les ‘feature’ d’entraînement



Table des matières

UQO - Introduction à la Science de données - projet Automne 2024	1
Table des matières	2
Introduction	3
Contexte	3
Objectifs du Projet	4
Nomenclature des variables	5
Étape 1 : Chargement des données	8
Étape 2 : Nettoyage	9
Étape 3 : Analyse	9
Etudes de quelques cas spécifiques	11
Étape 4 : Création d'un modèles de machine learning	15
Conclusions potentielles que nous pouvons tirer	18
Conclusions potentielles que nous ne pouvons pas tirer avec certitude	19

Introduction

Le diabète est une maladie chronique qui affecte des millions de personnes à travers le monde, et particulièrement aux Etats-Unis d'Amérique. Cette affection, caractérisée par des niveaux élevés de glucose dans le sang, peut entraîner de graves complications, notamment des maladies cardio-vasculaires, des lésions nerveuses et des problèmes rénaux. Au fil des ans, le diabète est devenu un problème de santé publique majeur, exacerbé par des facteurs tels que l'obésité, un mode de vie sédentaire et des habitudes alimentaires peu saines.

Contexte

Le taux de diabète continue d'augmenter dans le monde et en particulier aux Etats-Unis, soulignant l'urgence de comprendre les déterminants de cette épidémie. Malgré les progrès réalisés dans le diagnostic et la gestion du diabète, de nombreux patients restent mal informés sur la maladie, ce qui complique la prévention et le traitement. De plus, les inégalités en matière de santé, telles que l'accès inégal aux soins médicaux et aux ressources éducatives, aggravent la situation, en particulier dans les communautés vulnérables.

Objectifs du Projet

Ce projet vise à analyser les indicateurs de santé liés au diabète basé sur nos données qui proviennent du dataset "***diabetes_binary_5050split_health_indicators_BRFSS2015.csv***" (source : <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>), en mettant l'accent sur les données démographiques, les habitudes de vie, et l'accès aux soins de santé.

Les objectifs spécifiques incluent :

1. Collecter et analyser des données sur la prévalence du diabète, les facteurs de risque associés et les complications liées à la maladie.
2. Effectuer des analyses statistiques descriptives et des visualisations.
3. Identifier les disparités en matière de santé liées au diabète..
4. Supprimer les variables qui ne permettent pas d'expliquer la variable cible, et repérer celle qui impacte plus ou moins la maladie.
5. Implémenter un modèle qui nous permettra de prédire si un individu est diabétique ou pas.

En abordant ces objectifs, ce projet espère contribuer à une meilleure compréhension et à une gestion plus efficace du diabète.

Nomenclature des variables

1- Diabetes_012 : Indique la présence ou l'absence de diabète (par exemple, 0 = pas de diabète, 1 = pré diabète, 2 = diabète).

2- Hypertension (HighBP) : Indique si l'individu souffre d'hypertension (Oui/Non).

3- Cholestérol élevé (HighChol) : Indique si l'individu a un taux de cholestérol élevé (Oui/Non).

4- Vérification du cholestérol (CholCheck) : Indique si l'individu a fait vérifier son cholestérol récemment (5 dernières années) (Oui/Non).

5- IMC (BMI) : Indice de Masse Corporelle, une mesure de la graisse corporelle basée sur la taille et le poids.

6- Fumeur (Smoker) : Indique si l'individu fume actuellement ou a déjà fumé dans le passé (Oui/Non).

7- AVC (Stroke) : Indique si l'individu a déjà eu un accident vasculaire cérébral (Oui/Non).

8- Maladie cardiaque ou crise cardiaque (HeartDiseaseorAttack) : Indique si l'individu a déjà eu une maladie cardiaque ou une crise cardiaque (Oui/Non).

9- Activité physique (PhysActivity) : Indique si l'individu pratique une activité physique régulière (Oui/Non).

10- Fruits : Indique la fréquence de consommation de fruits (peut être binaire ou basée sur la fréquence).

11- Légumes (Veggies) : Indique la fréquence de consommation de légumes (semblable à Fruits).

12- Consommation d'alcool excessive (HvyAlcoholConsump) : indicateur de consommation excessive d'alcool (Oui/Non ou basé sur la fréquence).

13- Couverture de soins de santé (AnyHealthcare) : Indique si l'individu dispose d'une couverture de soins de santé (Oui/Non).

14- Pas de médecin à cause du coût (NoDocbcCost) : Indique si l'individu a évité de voir un médecin en raison des coûts (Oui/Non).

15- Santé générale (GenHlth) : Auto-évaluation de l'état de santé général (souvent sur une échelle de 1 à 5, 1 = excellent, 5 = mauvais).

16- Santé mentale (MentHlth) : Nombre de jours durant lesquels l'individu a ressenti une mauvaise santé mentale au cours du mois précédent.

17- Santé physique (PhysHlth) : Nombre de jours durant lesquels l'individu a ressenti une mauvaise santé physique au cours du mois précédent.

18- Difficulté à marcher (DiffWalk) : Indique si l'individu a des difficultés à marcher (Oui/Non).

19- Sexe (Sex) : Sexe de l'individu (Homme/Femme).

20- Âge (Age) : Âge de l'individu (peut être continu ou catégorisé en tranches d'âge).

21- Niveau d'éducation (Education) : Niveau d'éducation atteint (par exemple, école secondaire, études supérieures).

22- Revenu (Income) : Niveau de revenu de l'individu (peut être catégorisé ou continu).

Étape 1 : Chargement des données

Pour commencer notre étude il était premièrement question de trouver un dataset et de le charger, ce que nous avons accompli avec la fonction “`read_csv()`” de pandas. par la suite nous avons trouvé bon de regarder les premier données du dataset pour nous assurer qu'il est bien chargé, et de voir toutes les colonnes qui, puis d'avoir un bref résumé de nom donné en utilisant les commandes `dataframe.head()`, `dataframe.columns` , `dataframe.describe()`. Nous avons pu voir à la sortie une table avec les information générique sur notre dataset. Le tableau de sortie à présenter des statistiques descriptives des différentes colonnes du dataset. Des informations telle que:

count, le nombre d'observations non nulles pour chaque colonne. Nous avons obtenu un count de 70692. Nous n'avions donc pas de valeurs vides étant donné que notre dataset était constitué de 70692 lignes.

Les autres valeurs telles que **mean** (*La moyenne des valeurs pour chaque colonne*), **std**(*L'écart-type, qui mesure la dispersion des valeurs autour de la moyenne*), **min**(*La valeur minimale observée pour chaque colonne*), **25%**(*Le premier quartile, qui représente la valeur en dessous de laquelle se trouvent 25% des observations.*) , **50%**(*La médiane ou le deuxième quartile, qui représente la valeur en dessous de laquelle se trouvent 50% des observations*), **75%**(*Le troisième quartile, qui représente la valeur en dessous de laquelle se trouvent 75% des observations*), **max** (*La valeur maximale observée pour chaque colonne*) était majoritairement entre 1 et 0 étant donné que le dataset est basé sur des données tirée d'un questionnaire avec des oui et des non. Les seules valeur qui n'étaient pas comprises entre 1 et 0 étaient les colonnes **BMI**(Indice de Masse corporelle) , **GenHlth** (Auto-évaluation de l'état de santé général sur une échelle de 1 à 5. 1 pour excellent et 5 pour mauvais), **MentHlth** (Nombre de jours durant lesquels l'individu a ressenti une mauvaise santé mentale au cours du mois

précédent), ***PhysHlth*** (Nombre de jours durant lesquels l'individu a ressenti une mauvaise santé physique au cours du mois précédent), ***Age*** , ***Education*** et ***Income*** (Niveau de revenu de l'individu)

Ces statistiques nous ont permis d'avoir une vue d'ensemble des différentes variables dans le dataset, ce qui est essentiel pour comprendre les données avant de procéder à des analyses plus approfondies.

Étape 2 : Nettoyage

Pour le nettoyage des données, nous avons vérifié qu'il n'y a pas de valeurs nulles dans le dataset. Pour cela nous avons utilisé la commande ***dataframe.isna().any()*** qui retourne une série avec les noms des colonnes et “True” indiquant qu'il existe des valeurs nulles ou “False” indiquant qu'il n'en existe pas

Nous pouvions également normaliser les données afin que toutes les données soient comprises entre 0 et 1.

Étape 3 : Analyse

Nous avons commencé notre analyse par générer une matrice de corrélation pour identifier les relations linéaires entre les variables. Ici nous voulons vérifier s'il y a une corrélation linéaire entre nos variables. Ceci nous permettra de voir si nos variables sont linéairement corrélées.

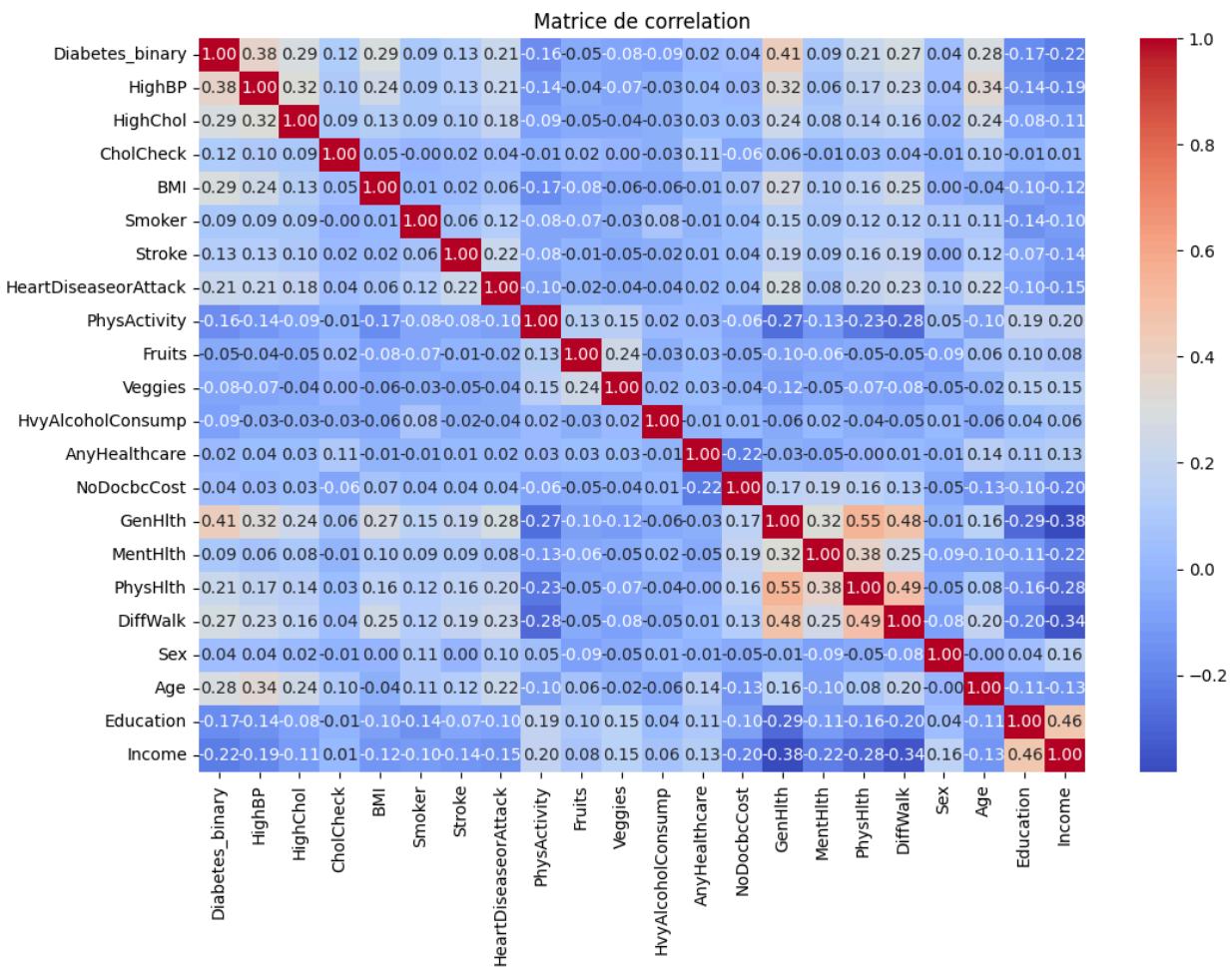


figure 1 : matrice de corrélation

En examinant le graphique, on peut constater une relation positive entre les difficultés de marche (**DiffWalk**) d'une personne et son état de santé physique (**PhysHlth**). Cela signifie que plus une personne éprouve des difficultés à marcher, plus elle est susceptible d'avoir une mauvaise santé physique.

Nous constatons également que la santé physique (**PhysHlth**) a une corrélation linéaire positive avec la santé générale (**GenHlth**). Si quelqu'un se sent bien physiquement, généralement il se sent bien.

Nous remarquons également une légère corrélation linéaire entre le nombre de jours qu'un individu a ressenti du stress, de la dépression ou un problème émotionnel (**MentHlth**) et le nombre de jours où cet individu a ressenti un problème physique (**PhysHlth**).

Légère corrélation linéaire entre **diabetes_012** et **High BP**. Une corrélation positive de 0.38 est observée entre le diabète (**Diabetes_binary**) et l'hypertension (**HighBP**). Cela suggère que les personnes atteintes de diabète sont plus susceptibles de souffrir d'hypertension.

Etudes de quelques cas spécifiques

Nous avons fait une étude statistique plus poussée pour voir comment les variables étaient reliées à la variables en affichant des graphiques pour une visualisation plus claire.

1.Taux de diabète en fonctions de la consommation des fruits, légumes, alcools et fumé

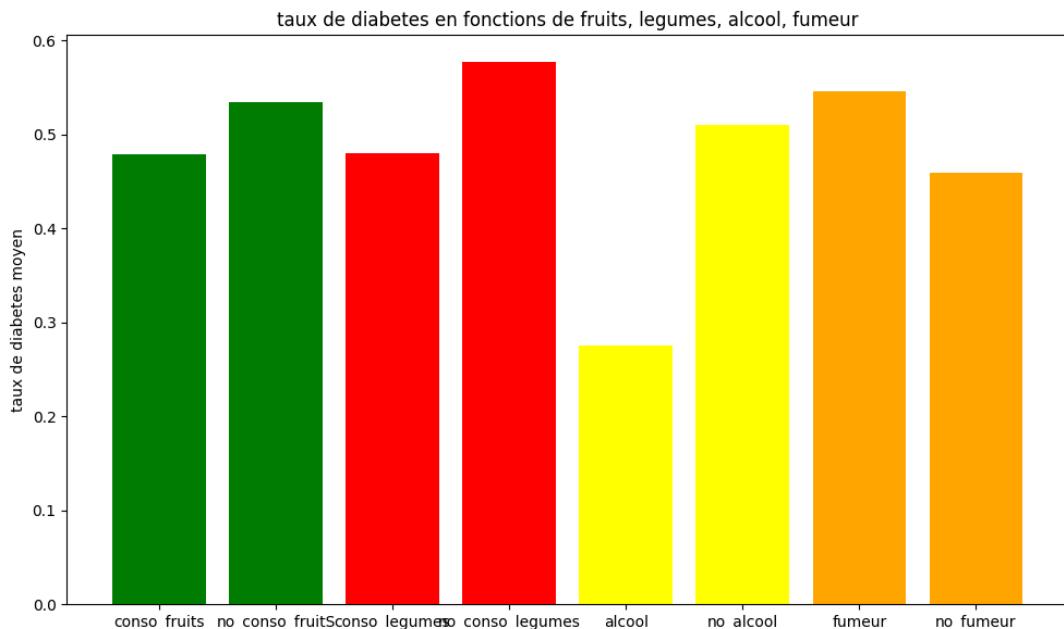


figure 1 : Taux de diabète en fonctions de la consommation des fruits, légumes, alcools et fumé

Le graphique montre le taux moyen de diabète dans différents groupes basés sur leurs habitudes alimentaires.

Commentaires

Les personnes qui consomment des fruits, des légumes et qui sont non fumeur ont un taux de diabète plus faible comparée à celles qui font l'inverse. Nous ne déduisons pas que ces variables ont un effet sur le diabète car la différence n'est pas très significative. rappelons nous que nous avons affaire à une partie de la population. Les alcooliques ont un taux de diabètes moins élevé contrairement à ceux qui n'en consomment pas. Dans notre DataSet, nous comprenons que le tabagisme n'a pas un grand effet sur notre population donc ne peut pas être classifié comme déclencheur du diabète de manière directe.

2.Taux de diabète en fonctions des maladies tels que l'hypertension, le cholestérol, l'avc et les maladie cardiaques

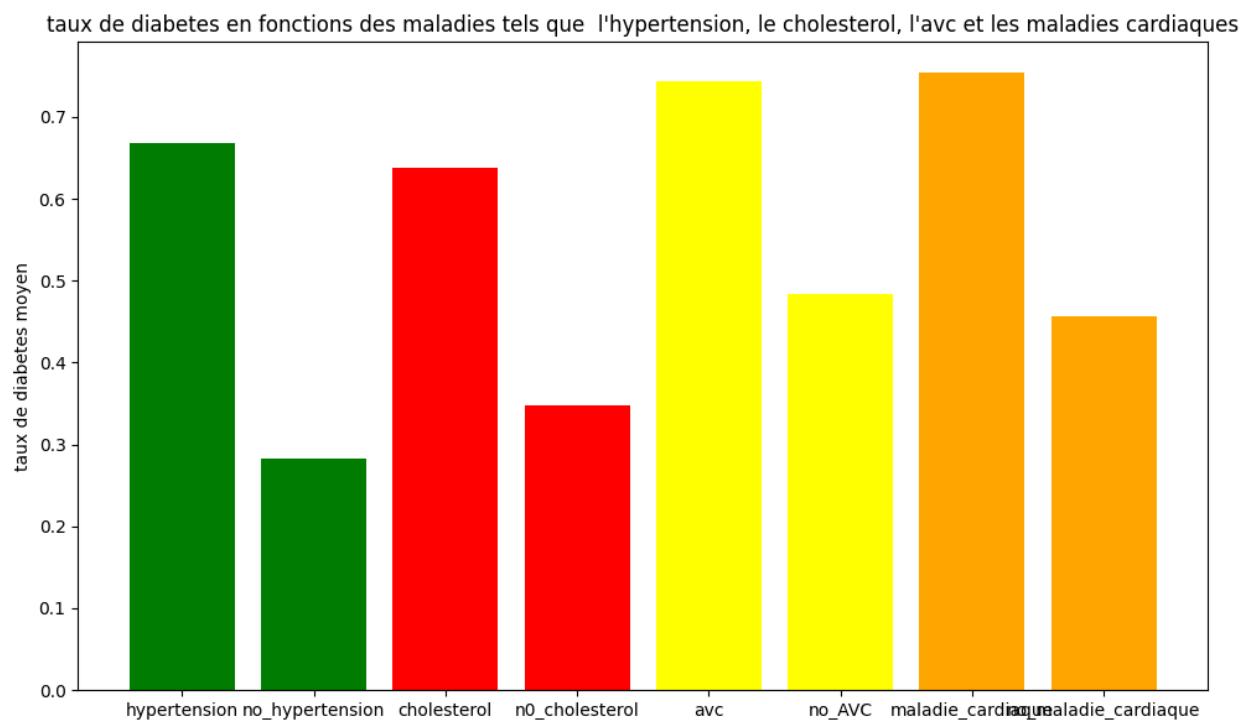


figure 2 : Taux de diabète en fonctions des maladies tels que l'hypertension, le cholestérol, l'avc et les maladie cardiaques

Commentaires

Ici, nous avons un graphique en fonction des maladies particulières qui attaquent une partie de cette population.

- Nous constatons que les maladies telles que l'hypertension, le cholestérol, l'AVC, les maladies cardiaques sont des facteurs à risque importants pour le diabète.
- Nous constatons que la grande partie des personnes atteintes de ces maladies souffre également de diabète.

nous mettons un point sur ceux ayant l'hypertension car elle est considérablement plus élevée.

3.Taux de diabète en fonctions de l'Age, de l'IMC, de l'activité physique, et du sexe de l'individue

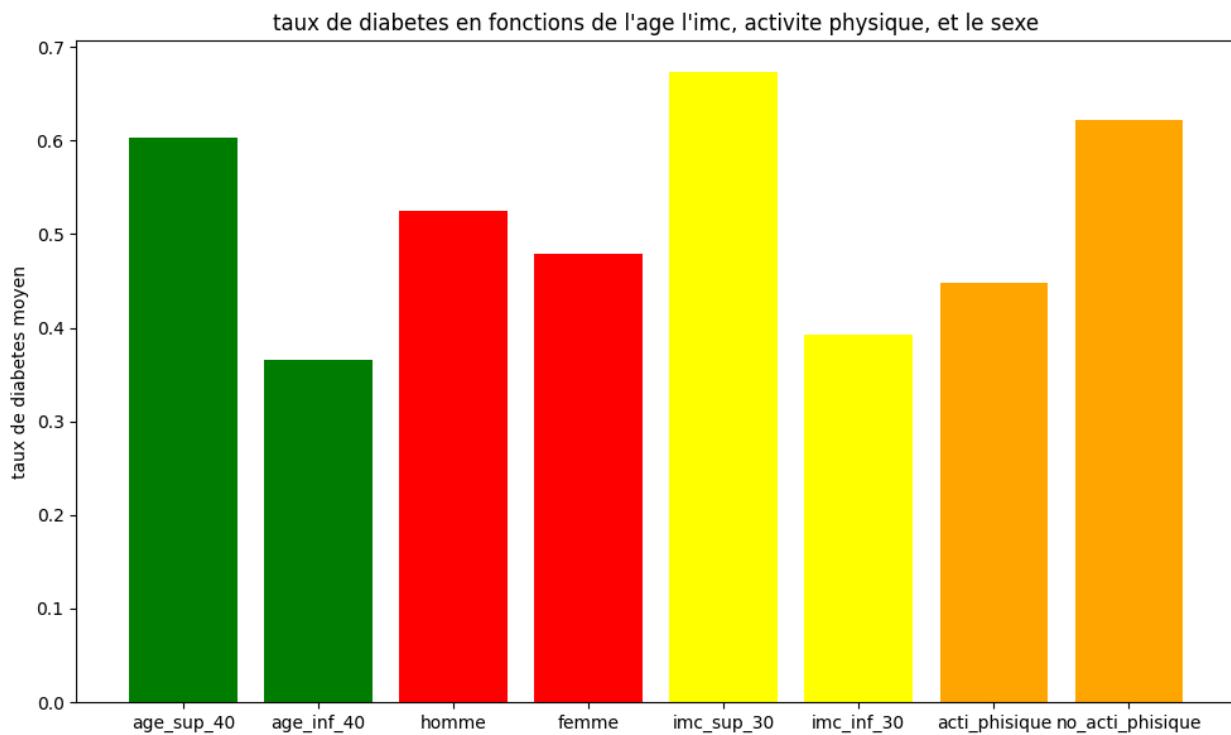


figure 3: Taux de diabète en fonction de l'âge, de l'IMC, de l'activité physique, et du sexe de l'individu

Commentaires

ici nous avons plusieurs remarques:

- le diabète est plus fréquent chez les personnes ayant un âge supérieur à 40.
- fréquent chez les hommes mais reste également assez significatif chez les femmes.
- plus fréquent chez les individus ayant une masse corporelle supérieure à 30.
- présent chez ceux n'effectuant pas suffisamment d'activité sportive.

4.Taux de diabète en fonction de l'assurance médicale du revenu et de l'éducation

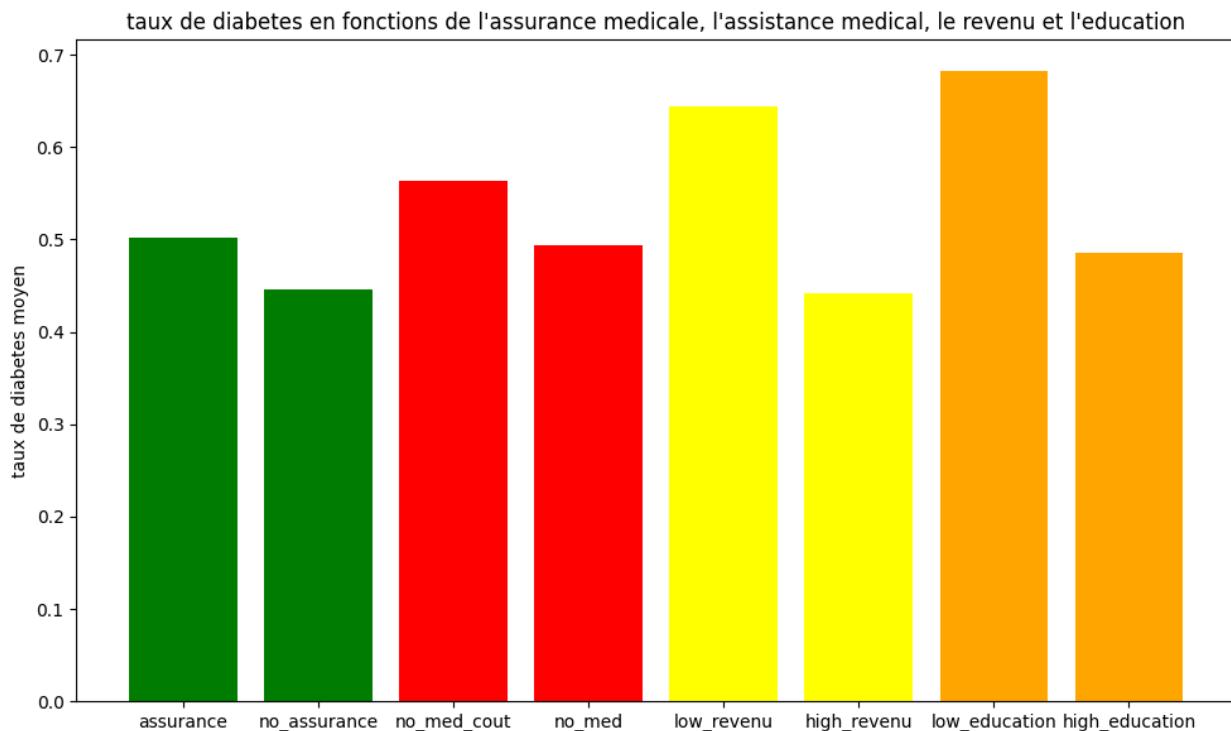


figure 4: Taux de diabète en fonction de l'âge, de l'IMC, de l'activité physique, et du sexe de l'individu

Commentaires

Ici nous constatons que l'effet d'avoir une assurance ou de ne pas l'avoir n'impacte pas considérablement sur le taux de diabète.

le fait de ne pas suivre un médecin à cause du coût ou pour une autre raison, d'avoir des revenus bas ou même par manque d'éducation pourrait avoir des impacts sur le taux de diabète. "ceci reste à étudier..."

Étape 4 : Création d'un modèles de machine learning

Étant donné que la valeur que nous voulons prédire c'est **Diabetes_binary** notre objectif est de supprimer les variables qui ne permettent pas d'expliquer la variable cible pour avoir un modèle plus performant. La matrice de corrélation nous permet de voir la relation linéaire entre les différentes variables, nous pouvons utiliser le module feature de scikit learn pour voir les variables essentiel pour déterminer si quelqu'un a le diabète ou pas.

Pour cela nous avons utilisé une fonction de la librairie Scikit Learn pour la sélection des “features” appelée **mutual_info_regression** qui permet de calculer l’information mutuelle. L’information mutuelle décrit les relations en termes d’incertitude. L’information mutuelle entre deux variables est une mesure de l’ampleur à laquelle la connaissance de l’une réduit l’incertitude concernant l’autre. Par exemple, si l’on connaît la valeur d’une feature, dans quelle mesure serions-nous plus confiants quant à la cible ?

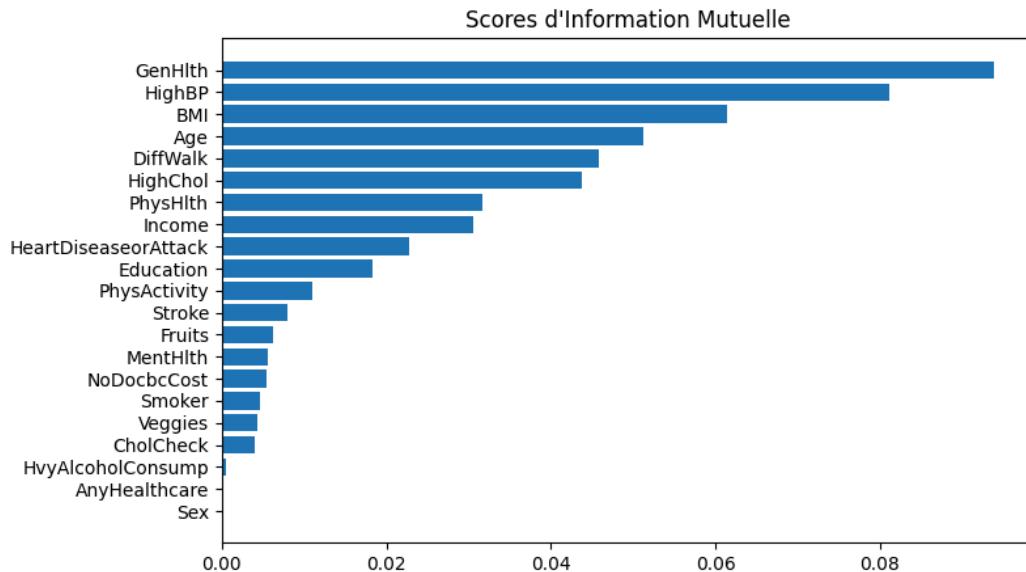


figure 5: information mutuel en rapport à la variable Diabetes_binary

Nous pouvons tiré du graphique que les variables HvyAlcoholConsump, AnyHealthcare et le sexe de l'individu ont un score d'information mutuelle faible. Ceci nous dit que c'est variable ne permettent pas d'expliquer si quelqu'un a le diabète ou pas selon nos données. Pour l'entraînement nous avons utilisé les librairies suivantes.

import seaborn as sns : pour la visualisation

import matplotlib.pyplot as plt : pour afficher les visuels

from xgboost import XGB Classifier : Un classe de la librairie xgboost permettant de créer un modèle de classification

from sklearn.model_selection import train_test_split : pour séparer les données d'entraînement, de validation et de test.

from sklearn.metrics import confusion_matrix, accuracy_score : pour évaluer notre modèle

Après l'entraînement du modèle nous avons fait une matrice de confusion pour évaluer notre modèle.

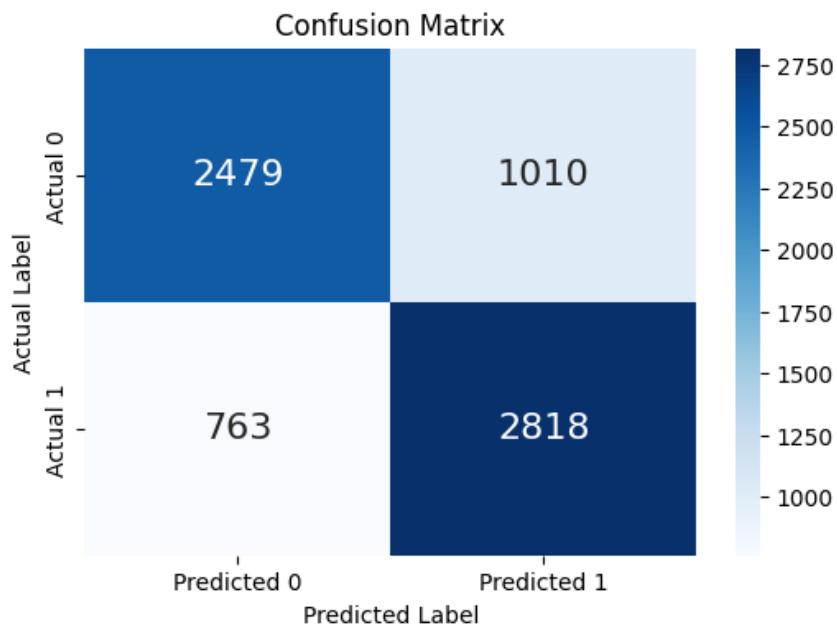


figure 6: Matrice de confusion

Résumé

Précision globale (Accuracy): 0.75

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

figure 7: Formule pour calculer la précision globale

Ce chiffre nous permet de savoir que notre modèle a trouvé la bonne réponse dans 75% des cas.

Vrai négatif (TN): 2479

Ce chiffre Consiste du nombre de fois que le modèle à réussi à prédire une personne non diabétique correctement

Faux positif (FP): 1010

Pour 1010 individus faisant partie des données de test, le modèle les a classés comme diabétiques alors qu'ils ne le sont pas.

Faux Négatif (FN): 763

Pour 763 individus faisant partie des données de test, le modèle les a classés comme non diabétiques alors qu'ils étaient diabétiques.

Vrai Positif (TP): 2818

Ce chiffre Consiste du nombre de fois que le modèle à réussi à prédire une personne diabétique correctement

Conclusions potentielles que nous pouvons tirer

Performance du modèle

Le modèle a atteint une précision globale d'environ 75 % sur l'ensemble de tests. Cela indique que le modèle est raisonnablement performant pour prédire si un individu est diabétique ou non en se basant sur les caractéristiques fournies.

Importantes caractéristiques

Les scores d'information mutuelle indiquent que des caractéristiques telles que **GenHlth, HighBP, BMI, Age** et **DiffWalk** sont parmi les prédicteurs les plus importants du diabète dans cet ensemble de données. Cela suggère que l'état de santé général, la pression artérielle, l'indice de masse corporelle, l'âge et la difficulté à marcher sont des facteurs significatifs pour prédire le diabète.

Matrice de confusion

La matrice de confusion montre que le modèle a une performance relativement équilibrée en termes de vrais positifs (2818) et de vrais négatifs (2479). Cependant, il y a également un nombre notable de faux positifs (1010) et de faux négatifs (763), indiquant des zones où le modèle pourrait être amélioré.

Analyse de relation

La matrice de corrélation révèle certaines relations linéaires entre les variables. Par exemple, il y a une corrélation positive (0.38) entre **Diabetes_binary** et **HighBP**, suggérant que les individus souffrant d'hypertension sont plus susceptibles d'être diabétiques. Cette relation est confirmée par le score d'information mutuel.

Conclusions potentielles que nous ne pouvons pas tirer avec certitude

Généralisabilité

La performance du modèle est basée sur l'ensemble de données spécifique utilisé pour l'entraînement et les tests. Nous ne pouvons pas être certains que le modèle fonctionnera aussi bien sur d'autres ensembles de données ou populations sans validation supplémentaire.

Limites du modèle

La précision de 75 % du modèle indique qu'il reste encore de la marge pour des améliorations. Nous ne pouvons pas conclure que ce modèle est le meilleur possible pour prédire le diabète. Certains hyperparamètres pourraient être ajustés pour le rendre encore meilleur.

Impact des Caractéristiques Supprimées

La suppression de certaines caractéristiques (par exemple, **Veggies**, **AnyHealthcare**, `HvyAlcoholConsump`, etc.) a été basée sur leurs faibles scores d'information mutuelle. Cependant, nous ne pouvons pas être certains que ces caractéristiques n'aient aucun impact sur la prédiction du diabète sans investigations supplémentaires. Elles pourraient encore contribuer de manière non linéaire ou interagir avec d'autres caractéristiques.

Conclusion

Notre objectif était de collecter et traiter les données, d'effectuer une analyse statistique, des visualisations afin de repérer les variables qui impactent sur la maladie. A travers l'analyse statistique, nous avons pu identifier des facteurs critiques liés au diabète tels que l'âge, l'indice de masse corporelle et l'hypertension.

Les facteurs tels que les “veggies”, la consommation d'alcool, l'assurance, le sex ont été exclus lors de l'entraînement du modèle afin de réduire le bruit car ils avaient un score d'information mutuelle bas.

Nous avons par la suite créé et entraîné un modèle de machine learning basique qui s'est avéré raisonnablement performant avec un score de 75%.

Cependant, la majorité de nos données était sous une forme binaire(oui ou non). Cette nature binaire des données peut masquer certaines informations sur la relation entre les variables. Des analyses futures pourraient bénéficier de l'incorporation de variables continues ou de l'élargissement de l'ensemble des données pour inclure des réponses plus détaillées, ce qui permettrait une compréhension plus riche des facteurs influençant le risque de diabète.

Annexes

Code sources :

https://github.com/BryanGabrielAtangana/UQO_data_science_project_2024_aut

Participants :

TAMBAT TRESOR MEGANE TAMT79360604

BRYAN GABRIEL-ATANGANA MBOA GABB78300209

DJIEUNANG NOUMBO NELIE MABELLE DJIN28279107