**Diabète**

**Intro**

Bonjour,

Je suis Patrice DANGER et je vais vous parler d’un sujet qui concerne une grande partie de la planète.

**Projet**

Bonjour,

Je suis Patrice DANGER et je vais vous parler d’un sujet qui concerne une grande partie de la planète.

**Raisons**

Commençons par les éléments qui donnent du sens à ce projet

Ce projet aborde une problématique de santé mondiale majeure touchant plus de 463 millions de personnes âgées de 20 à 80 ans.

Cette ampleur souligne l'urgence et la pertinence de cette analyse, alors que les cas sont en forte augmentation depuis 25 ans.

Le diabète est la 7ème cause de décès dans le monde selon l’OMS.

Cette réalité met en lumière l'impératif de mieux comprendre les facteurs sous-jacents, de développer des méthodes de prévention efficaces et d'améliorer les stratégies de prise en charge.

**Problématique**

Comment pourrait-on prédire ou identifier précocement des personnes atteintes ou à risque de développer le diabète ?

**Opportunités**

Mon objectif est de permette d'identifier précocement les personnes à risque de diabète.

Comment : Par la construction d’un modèle de détection du diabète robuste et fiable, qui apprendra dans le temps pour s'adapter aux nouvelles données et informations disponibles.

Ce modèle aura plusieurs applications :

**La Prévention**

**Le Suivi**

Ce modèle sera bénéfique pour divers acteurs :

**Les Professionnels de la santé**

**Pour les Entreprises MEDTECH**

**Méthodologie**

Tout d’abord, j'ai recherché une base de données sur le Diabète sur le site Kaggle.

J'ai sélectionné un Dataset possédant une taille significative pour mes analyses.

Je suis passé à l’exploration des données, ce qui m’a permis d’identifier la variable cible : le "diabète"

Ensuite, j'ai exploré les relations entre la cible et les autres variables afin de comprendre les dynamiques sous-jacentes.

J'ai également cherché à détecter d'éventuels déséquilibres dans mon jeu de données pour m'assurer de la représentativité des données.

J’ai procédé aux traitements des données afin de les normaliser et de les standardiser.

Puis j’ai procédé à un équilibrage des valeurs afin de minimiser les biais potentiels et d'assurer la robustesse de mes analyses.

Ensuite, j'ai segmenté le Dataset en deux groupes distincts, pour faciliter les analyses ultérieures.

Aux vues des données à prédire, j'ai sélectionné et créé plusieurs modèles de Machine Learning.

Enfin, j'ai évalué les performances de chaque modèle pour déterminer leur efficacité dans la prédiction du diabète.

Sur la base des scores obtenus, j'ai choisi le modèle le plus performant pour poursuivre mon analyse.

**Dataset**

Le Dataset se compose de 9 colonnes et de 100 000 lignes

Lors de l’affichage des données dans Python, j'ai remarqué un déséquilibre des valeurs sur la variable cible.

Ce déséquilibre souligne l'importance de prendre des mesures appropriées pour garantir la représentativité des données dans les analyses.

**EDA**

L'objectif de cette étape est de découvrir et de comprendre les caractéristiques essentielles des données, afin d'en retirer des informations pertinentes pour l’analyse.

**Matrice**

J'ai réalisé une matrice de corrélation pour explorer les relations entre les différentes variables de notre Dataset.

Cette matrice met en avant l'importance des facteurs suivants :

* Le glucose
* Le taux d’hémoglobine
* L’âge
* L’indice de masse corporelle

**Facteurs discriminants**

À la suite de mon analyse de la matrice de corrélation, j’ai élaboré des graphiques comparatifs qui permettent de visualiser les valeurs des différents facteurs en fonction des personnes positives ou négatives au diabète.

Ces graphiques sont conçus pour offrir une perspective visuelle claire sur les différences significatives entre les deux groupes de personnes, en mettant en évidence les variations dans les valeurs des facteurs tels que l'âge, la glycémie, l'HbA1c (Hémoglobine) et l'indice de masse corporelle (BMI).

Facteurs non discriminants

Pour donner suite à mon analyse, j’ai identifié certains facteurs qui semblent ne pas avoir d'impact significatif sur la prévalence du diabète.

* L’hypertension
* La présence de maladie cardiaque
* Les genres masculins et féminins
* Les antécédents de tabagisme

**Preprocessing**

Le prétraitement des données a été une étape essentielle de mon processus d'analyse.

Voici les principales actions que j'ai entreprises :

* J'ai commencé par convertir les valeurs non numériques en valeurs numériques.
* Ensuite, j'ai segmenté le Dataset en deux groupes distincts
* Pour résoudre le déséquilibre des valeurs identifiées précédemment dans la colonne 'diabète', j'ai utilisé la technique SMOTE,

**SMOTE**

Cette méthode consiste à générer synthétiquement de nouveaux exemples de la classe minoritaire afin de rétablir un équilibre plus adéquat entre les classes.

Cela garantit que notre modèle ne soit pas biaisé par la sur-représentation d'une classe par rapport à l'autre.

**Modèles de ML**

J’ai décidé d'utiliser plusieurs modèles de machine learning afin d'évaluer et de comparer leur performance dans la prédiction de la variable cible.

Voici les modèles que j'ai sélectionnés :

* Logistic Regression
* Decision Tree
* Random Forest
* Gradient Boosting Classifier

Tous ces modèles ont été choisis en raison de leur adaptation naturelle aux données binaires, ce qui les rend appropriés pour notre variable cible, le diabète.

**Résultats de ML**

Après avoir évalué les performances des différents modèles de Machine Learning, j’ai examiné quatre métriques clés :

l’Accuracy, la Précision, le Recall et le F1 Score.

Ces métriques permettent d'avoir une vision globale de la capacité de nos modèles.

Il ressort que le modèle Random Forest est celui qui a obtenu les meilleures performances sur l'ensemble de ces métriques.

Son score élevé en fait le choix le plus optimal pour notre problème de prédiction du diabète.

Voyons maintenant plus en détail ses résultats.

**Random Forest**

Dans le cadran en haut à droit :

Les cas où le modèle à prédit un test négatif et les sujets étaient positif, mais dans une proportion faible.

**Relations variables & valeurs cibles**

En examinant le graphique des features, j’ai identifié les variables les plus impactantes sur la performance du modèle.

Voici mes principales observations :

La variable (Hémoglobine) HbA1c se démarque comme LA Feature la plus influente sur la prédiction du modèle, représentant près de 41%.

Cette constatation souligne l'importance cruciale de cette mesure dans la détection du diabète.

La glycémie représente environ 26% dans la prédiction du modèle, ce qui en fait le second indicateur clé pour le diagnostic du diabète.

L'âge et l'indice de masse corporelle (BMI) suivent en termes d'impact sur le modèle.

Ils contribuent tous deux à la prédiction du diabète, mettant en évidence la prise en compte de ces facteurs dans l'évaluation du risque de diabète.

Les autres Features semblent avoir une importance moindre sur la prédiction du modèle, étant considérées comme plus négligeables dans leur contribution à la performance globale.

**Conclusion**

En conclusion.

Nous avons exploré et analysé un ensemble de données complexe, en utilisant des techniques de prétraitement, des modèles de Machine Learning et des méthodes d'évaluation pour mieux comprendre les facteurs influant sur la présence du diabète.

Les résultats de l’analyse ont révélé que le modèle Random Forest s'est avéré être le plus performant, offrant une précision élevée dans la prédiction du diabète.

De plus, nous avons identifié les variables les plus impactantes sur le modèle, notamment l’HbA1c et la glycémie, soulignant l'importance cruciale de ces mesures dans l'évaluation du risque de diabète.

**Recommandations**

Pour aller plus loin dans cette étude, voici quelques suggestions :

Nous pourrions explorer des techniques d'optimisation supplémentaires pour améliorer encore la performance du modèle Random Forest, telles que l'ajustement des hyperparamètres (Taille des nœuds, nombres d’arbres,).

**L’exploration des autres Features :** Bien que nous ayons identifié les Features les plus impactantes, une analyse plus approfondie des autres variables pourrait révéler des relations subtiles ou des interactions complexes qui pourraient également contribuer à la prédiction du diabète.

**Validation Externe :** Il serait bénéfique de valider notre modèle sur des ensembles de données externes pour évaluer sa applicabilité et sa robustesse dans différents contextes et populations.

**L’intégration de Données Additionnelles :** telles que des données génétiques ou des mesures de style de vie, pourrait enrichir notre analyse et améliorer la précision de notre modèle.

En poursuivant ces axes de recherche, nous pourrons approfondir notre compréhension de la détection précoce du diabète et développer des outils plus efficaces pour la prévention et la gestion de cette maladie chronique.