Objectif :

Appliquer un algorithme d'apprentissage automatique sur un ensemble de données réelles.

Données :

Vous avez 3 jeux de données à votre disposition. Vous devez choisir un ensemble de données parmi les 3 ou vous pouvez travailler avec votre propre ensemble de données si vous voulez travailler sur un sujet en particulier.

Méthodologie :

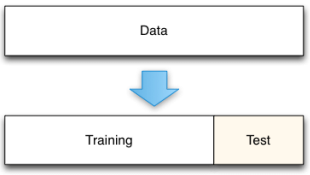
Prétraitement et chargement des données :

- Chargez l'ensemble de données avec lequel vous avez choisi de travailler.
- L'une des premières étapes du processus d'exploration des données, lorsque l'objectif final est de prédire le résultat, consiste à créer des visualisations qui permettent de connaître le résultat, puis de découvrir les relations entre les attributs et le résultat.
- Vous pouvez utiliser de nombreux outils de visualisation des données tels que les diagrammes à barres, les histogrammes, les boxplots, les matrices de corrélation, les diagrammes de paires et les analyses PCA.
- Choisissez les variables qui vous semblent utiles pour votre prédiction.
- Prenez soin d'éliminer les caractéristiques corrélées.
- N'hésitez pas à créer de nouvelles fonctionnalités en combinant d'autres fonctionnalités. Si vous pensez qu'elles ont un impact négatif sur vos prédictions, supprimez certaines valeurs aberrantes.

Engineering dataset :

Divisez votre ensemble de données en 2 parties :

- ensemble d'entraînement
- ensemble de test



Si vous pensez que c'est utile, appliquez la normalisation sur votre ensemble de données. Veillez à appliquer la transformation inverse sur vos prédictions à la fin pour comparer votre résultat avec la vraie sortie.

Définir le modèle :

Choisissez un algorithme d'apprentissage automatique et adaptez-le aux données. Pour ce modèle, essayez d'utiliser la méthode **Grid Search** pour optimiser les hyperparamètres.

Évaluer les modèles :

Évaluez les modèles sur l'ensemble de test et conservez le meilleur modèle. Faites une visualisation de certaines prédictions et comparez-la aux véritables prédictions avec le meilleur modèle.

diabetes.csv :

Ce jeu de données provient du National Institute of Diabetes and Digestive and Kidney Diseases.

L'objectif de ce jeu de données est de prédire de manière diagnostique si un patient est diabétique ou non, en fonction de certaines mesures diagnostiques incluses dans le jeu de données. Plusieurs contraintes ont été imposées à la sélection de ces instances à partir d'une base de données plus importantes. En particulier, tous les patients sont des femmes âgées d'au moins 21 ans et d'origine indienne Pima.

L'ensemble de données se compose de plusieurs variables prédictives médicales et d'une variable cible, le `Outcome`. Les variables prédictives comprennent le nombre de grossesses que la patiente a eues, son IMC, son taux d'insuline, son âge, etc.

Pouvez-vous construire un modèle d'apprentissage automatique pour prédire avec précision si les patients de l'ensemble de données sont diabétiques ou non ?

Variable de sortie / Output variable : Outcome (0 ou 1)

house.csv :

Ce jeu de données contient les prix de vente des maisons pour la ville de King, qui se situe à Seattle.

Il comprend les maisons vendues entre mai 2014 et mai 2015. C'est un excellent jeu de données pour évaluer les modèles de régression simples permettant de prédire le prix d'une maison en fonction de ses caractéristiques.

Variable de sortie : Price (variable continue, le prix)

wine.csv :

L'ensemble de données concerne les vins rouges et blancs du Portugal "Vinho Verde".

Pour des raisons de confidentialité et de logistique, seules les variables physico-chimiques (les entrées) et sensorielles (les sorties) sont disponibles (par exemple, il n'y a pas de données sur les types de raisins, la marque de vin, le prix de vente du vin, etc.). Les classes sont ordonnées et non équilibrées (par exemple, il y a beaucoup plus de vins normaux que d'excellents ou de mauvais vins).

Utilisez l'apprentissage automatique pour déterminer les propriétés physico chimiques qui font qu'un vin est "bon" !

Variable de sortie : la qualité (score entre 0 et 10)

Comment obtenir une bonne note ? 

- Commentez chaque bloc de code et justifiez chaque choix. Une visualisation qui n'est pas commentée est considérée comme non faite.
  ex : Le box plot montre 3 valeurs aberrantes, nous le supprimons de notre jeu de données pour le reste de l'étude.
- Il est plus important d'avoir de mauvaises prédictions et d'avoir une bonne méthodologie et une bonne justification que le contraire.

Conseils : 

- Utiliser la bibliothèque python sklearn pour l'apprentissage automatique : https://scikit learn.org/stable/

Environnement de travail :

Vous travaillerez avec les cahiers google colab. Une fois le projet réalisé, téléchargez le notebook au format .PDF, créez une restitution de votre travail à l'aide du notebook qui vous servira lors de la soutenance ou bien créez une restitution sous forme de slides.

Vous transférerez l'ensemble de votre travail dans un dossier compressé, en respectant le nom du dossier par **nom_prenom** à mon adresse mail : kattin.dassance@hupi.fr. Prenez soin d'exécuter chaque cellule avant de le partager avec moi afin que je n'aie pas à exécuter votre code à nouveau. Je veux donc un fichier au format PDF avec chaque ligne de code exécutée avant afin de voir tous les logs et impressions des graphiques.

La date limite est le 18 décembre 2024.

Bonne chance ! :)

Lexique / Informations des bases de données :

diabetes.csv

- Pregnancies: To express the Number of pregnancies
- Glucose: To express the Glucose level in blood
- BloodPressure: To express the Blood pressure measurement
- SkinThickness: To express the thickness of the skin
- Insulin: To express the Insulin level in blood
- BMI: To express the Body mass index
- DiabetesPedigreeFunction: To express the Diabetes percentage
- Age: To express the age
- **Outcome**: To express the final result 1 is Yes and 0 is No

house.csv

- id - Unique ID for each home sold
- date - Date of the home sale
- **price** - Price of each home sold
- bedrooms - Number of bedrooms
- bathrooms - Number of bathrooms, where .5 accounts for a room with a toilet but no shower
- sqft_living - Square footage of the apartments interior living space
- sqft_lot - Square footage of the land space
- floors - Number of floors
- waterfront - A dummy variable for whether the apartment was overlooking the waterfront or not
- view - An index from 0 to 4 of how good the view of the property was
- condition - An index from 1 to 5 on the condition of the apartment,
- grade - An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.
- sqft_above - The square footage of the interior housing space that is above ground level
- sqft_basement - The square footage of the interior housing space that is below ground level
- yr_built - The year the house was initially built
- yr_renovated - The year of the house's last renovation
- zipcode - What zipcode area the house is in
- lat - Lattitude
- long - Longitude
- sqft_living15 - The square footage of interior housing living space for the nearest 15 neighbors
- sqft_lot15 - The square footage of the land lots of the nearest 15 neighbors

wine.csv

- fixed acidity

- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol
- **Output variable (based on sensory data): quality (score between 0 and 10)**