Digital Bridge RDC

Exploration et nettoyage des données

December 2022

Ingénierie des fonctionnalités

- Analyse exploratoire des données (EDA)
  - Données numériques
  - Données catégorielles
  - Valeurs manquantes
  - Valeurs abberantes
- Data cleaning:
  - Gestion des valeurs manquantes
  - Gestion des jeux de données déséquilibrés
  - Trouver les valeurs aberrantes
  - Mise à l'échelle des données (standardisation, normalisation)

Sélection des fonctionnalités

- Basé sur la corrélation, par exemple l'analyse en composantes principales (ACP)
- K-voisin
- Chi-carré
- Algorithme génétique

Lecture dans les données

- Vous pouvez utiliser la fonction Pandas `read_csv()` pour lire un fichier CSV,
- Vous pouvez utiliser `read_excel()` pour lire un fichier Excel.
- Vous pouvez également utiliser la fonction `read_sql()` pour lire les données d'une base de données.

Inspection des données

- Une fois que vous avez lu les données, il est important d'explorer et de comprendre les données.
- Vous pouvez utiliser des fonctions telles que head(), tail(), info() et describe() pour avoir une idée des données et identifier les problèmes potentiels.

Gestion des valeurs manquantes

- Les valeurs manquantes peuvent causer des problèmes lors de l'analyse et de la modélisation des données.
- Vous pouvez utiliser des fonctions telles que isnull() et notnull() pour identifier les valeurs manquantes, et vous pouvez utiliser fillna() pour remplir les valeurs manquantes avec une valeur spécifiée.

Gestion des valeurs aberrantes

- Les valeurs aberrantes peuvent également être un problème lors de l'analyse des données.
- Vous pouvez utiliser la fonction boxplot() pour identifier les valeurs aberrantes potentielles et utiliser des fonctions telles que clip() ou truncate() pour les supprimer.

Gestion des doublons

- Les lignes dupliquées peuvent également causer des problèmes lors de l'analyse des données.
- Vous pouvez utiliser la fonction duplicated() pour identifier les lignes en double et la fonction drop_duplicates() pour les supprimer.

Normalisation et mise à l'échelle des données

- Dans certains cas, il peut être nécessaire de normaliser ou de mettre à l'échelle les données pour s'assurer que toutes les variables sont à la même échelle.
- Vous pouvez utiliser des fonctions telles que `minmax_scale()` ou `StandardScaler()` pour normaliser ou mettre à l'échelle les données.

Enregistrement des données nettoyées

- Une fois que vous avez nettoyé les données, vous voudrez probablement les enregistrer pour une utilisation ultérieure.
- Vous pouvez utiliser la fonction to_csv() pour enregistrer les données en tant que fichier CSV, ou vous pouvez utiliser to_excel() pour enregistrer les données en tant que fichier Excel.

bibliothèques d’exploration des données et nettoyage

Pandas-profiling bibliothèque

- pandas-profiling est une bibliothèque permettant de générer des rapports détaillés sur les caractéristiques d'un Pandas DataFrame.
- Il est souvent utilisé pour l'exploration de données en science des données, car il fournit un moyen rapide et facile de comprendre la structure, la distribution et les relations des variables dans un jeu de données.
- À l'aide du profilage des pandas, vous pouvez générer un rapport de profilage qui inclut des statistiques récapitulatives, des corrélations et des visualisations de la distribution et des relations des variables.
- Cela peut vous aider à identifier les tendances, les modèles et les problèmes potentiels dans les données qui peuvent devoir être traités avant de créer un modèle.
- Pour générer un rapport de profilage à l'aide de pandas-profiling, vous pouvez utiliser le pandas_profiling. ProfileReport et passez un DataFrame Pandas en tant qu'argument.
- Documentation: https://pandas-profiling.ydata.ai/docs/master/index.html

Bibliothèque d'exploration de données Sweetviz

- Sweetviz est une bibliothèque Python qui peut être utilisée pour l'analyse exploratoire des données (EDA) en science des données. Il permet aux utilisateurs de générer rapidement des visualisations et des résumés statistiques de leurs ensembles de données, ce qui peut les aider à mieux comprendre les relations entre les différentes variables et à identifier des modèles ou des tendances dans les données.
- L'une des principales caractéristiques de Sweetviz est sa capacité à comparer deux ensembles de données, tels qu'un ensemble de données de formation et un jeu de données de test, ou deux versions différentes du même jeu de données. Cela peut être utile pour identifier les différences entre les ensembles de données, telles que les changements dans la distribution de certaines variables ou la présence de valeurs manquantes.
- Sweetviz fournit également un certain nombre de graphiques et de graphiques prédéfinis, notamment des nuages de points, des graphiques à barres et des histogrammes, qui peuvent aider les utilisateurs à visualiser la distribution des différentes variables et les relations entre elles. En outre, Sweetviz offre une gamme d'options de personnalisation, permettant aux utilisateurs d'ajuster l'apparence et la disposition de leurs parcelles et graphiques en fonction de leurs besoins.
- Dans l'ensemble, Sweetviz est un outil utile pour l'exploration des données et peut aider les scientifiques et les analystes de données à mieux comprendre leurs ensembles de données et à informer leurs efforts d'analyse et de modélisation des données.

Bibliothèque de nettoyage de données Klib

- Klib est une bibliothèque Python qui fournit un certain nombre de fonctions utilitaires pour le nettoyage et le prétraitement des données. Il peut être particulièrement utile pour les scientifiques et les analystes de données travaillant avec des ensembles de données volumineux et complexes pouvant contenir des erreurs, des incohérences ou des valeurs manquantes.
- L'une des principales caractéristiques de Klib est sa capacité à gérer les données manquantes. Il fournit des fonctions permettant d'identifier et d'imputer les valeurs manquantes, par exemple en remplaçant les valeurs manquantes par la moyenne ou la médiane des données disponibles. Klib inclut également des fonctions de détection et de traitement des valeurs aberrantes, ce qui peut aider à améliorer la précision des analyses et de la modélisation des données.
- Klib fournit également des fonctions de formatage et de normalisation des données, par exemple en mettant à l'échelle des variables numériques sur une plage commune ou en codant des variables catégorielles sous forme de valeurs numériques. Cela peut être utile pour préparer des données pour des algorithmes d'apprentissage automatique, qui peuvent nécessiter que les données d'entrée soient dans un format spécifique.
- Dans l'ensemble, Klib est un outil utile pour le nettoyage et le prétraitement des données, et peut aider les scientifiques et les analystes de données à s'assurer que leurs données sont propres, cohérentes et prêtes pour l'analyse et la modélisation.
- Documentation: https://klib.readthedocs.io/en/latest/