Maîtriser la manipulation et la visualisation des données

1 Description

Dans le monde actuel axé sur les données, la capacité à exploiter la puissance du big data pour obtenir des informations et prendre des décisions est une compétence précieuse. Ce projet offre une opportunité passionnante aux étudiants de plonger dans le monde de l'analyse de données en utilisant Python et des ensembles de données du monde réel. Dans ce projet, les étudiants exploreront le monde multifacette de la science des données en travaillant avec un jeu de données : tabulaires/séries temporelles.

2 Objectifs du projet

Ce projet est conçu pour fournir une compréhension des techniques de science des données automatique, ainsi que de leur applicabilité à différentes modalités de données. Plus de détails :

- **Exploration des données** : Vous apprendrez à naviguer et à comprendre des jeux de données complexes, en acquérant des informations sur les structures, les motifs et les anomalies des données.
- **Prétraitement des données** : Vous découvrirez des techniques pour nettoyer, prétraiter et transformer les données brutes/images en un format utilisable, garantissant la qualité des données.
- **Visualisation** : Vous créerez des visualisations informatives des données pour présenter efficacement vos résultats.
- **Interprétation** : Vous développerez les compétences nécessaires pour interpréter les résultats du modèle et tirer des informations significatives des données.
- **Modélisation d'apprentissage automatique (optionnel)** : Vous appliquerez des algorithmes d'apprentissage automatique pour résoudre deux types de tâches différentes, telles que la classification/régression et le regroupement.

3 Jeu de données

Ce projet concerne la manipulation et la visualisation des données avec Python en utilisant des jeux de données du répertoire de l'apprentissage automatique de "www.bigdata-ai.fraunhofer.de," qui contient une liste de jeux de données volumineux possibles. Dans ce projet, chaque groupe travaillera avec un jeu de données distinct.

3.1 Jeu de données tabulaires ou Séries Temporelles

Le premier jeu de données doit être tabulaire ou séries temporelles, que l'on rencontre couramment dans divers domaines, notamment la finance, les soins de santé et le marketing.

3.2 Jeux de données

Vous pouvez chercher un jeu de données existant dans ce lien. Vous trouverez quelques ensembles de données de l'UCI (tabulaires, séries temporelles et images). Vous avez la liberté de choisir n'importe lequel, à condition qu'il n'ait pas déjà été choisi par un autre groupe.

En effet, la majorité de ces jeux de données sont connus pour leur taille et leur complexité, ce qui en fait d'excellents choix pour les projets impliquant l'analyse de big data et l'apprentissage automatique. D'autre part, l'une des principales distinctions de ce projet réside dans le contraste entre les deux ensembles de données. Alors que les ensembles de données d'images reposent fortement sur les modèles d'apprentissage en profondeur, les données tabulaires/séries temporelles nécessitent que les étudiants utilisent un éventail plus large de techniques d'apprentissage automatique, y compris des algorithmes traditionnels tels que les arbres de décision, les forêts aléatoires (Random Forest) et les machines à vecteurs de support (SVM), etc. Cette approche double vous permettra d'apprécier les forces et les limites des différentes méthodes d'apprentissage automatique et souligne l'importance de choisir les meilleures méthodes pour des types de données et des tâches spécifiques.

4 Tâches à accomplir

Ce qui suit est une série de questions liées à divers aspects du traitement des données.

4.1 Data Loading: Jeu de données tabulaires ou séries temporelles

1. Charger le jeu de données sélectionné (par exemple, dans un DataFrame Pandas en utilisant des fonctions appropriées telles que "***pd.read_csv()***" dans le cas de jeux de données tabulaires/séries temporelles).

2. Effectuer des tâches de prétraitement des données telles que la gestion des valeurs manquantes, les conversions de types de données et le nettoyage des données.

    - Vérifier les valeurs manquantes en utilisant "***df.isna()***" et les traiter en les imputant ou en les supprimant. Assurez-vous d'expliquer les raisons derrière votre décision d'utiliser des techniques d'imputation ou de suppression pour traiter les données manquantes.
    - Convertir les types de données au besoin en utilisant "***df.astype()***".
    - Nettoyer les données manquantes en supprimant les doublons à l'aide de "***df.drop_duplicates()***" et en corrigeant les valeurs incohérentes. Vous devez indiquer quelle technique a été appliquée. Si vous choisissez l'imputation, précisez quelle méthode spécifique vous trouvez la plus appropriée pour votre jeu de données et pourquoi?
    - Créer de nouvelles fonctionnalités ou transformer celles existantes pour améliorer la qualité et la pertinence des données (si possible).

4.2 Analyse exploratoire des données (EDA)

Profilage des données pour obtenir des informations sur leur distribution, leurs relations, leurs statistiques sommaires et les éventuels problèmes de qualité des données (données aberrantes).

Dans cette partie, vous utiliserez Matplotlib, Seaborn ou Plotly pour créer une variété de graphiques, notamment :

- Des graphiques linéaires pour visualiser les tendances au fil du temps.
- Des graphiques de dispersion (scattering) pour identifier les relations entre les variables numériques.
- Des graphiques à barres pour comparer les données catégorielles.
- Des cartes thermiques pour montrer les corrélations entre les variables.

- Créer des visualisations interactives à l'aide de bibliothèques comme Plotly pour améliorer l'expérience utilisateur.

Ensuite, vous devez :

1. Calculer des statistiques sommaires (à l'aide de fonctions comme "***df.describe()***") pour comprendre les tendances centrales de jeu de données et les distributions en utilisant des histogrammes, des tracés de densité (KDE) et d'autres visualisations pour visualiser la distribution des données.

2. Les valeurs aberrantes peuvent avoir un impact significatif sur l'analyse et les performances. Déterminez si les valeurs aberrantes sont des points de données valides ou des erreurs, et gérez-les en conséquence. Vous pouvez gérer les valeurs aberrantes en les visualisant à l'aide de box plots et en décidant de les conserver ou de les supprimer. Les box plot permettent de visualiser les quartiles, la médiane et les valeurs aberrantes dans les données.

3. Calculer les matrices de corrélation et les tracer pour identifier les relations entre les variables numériques. Réalisez également un graphique entre les variables et la variable cible (dans le cas de la régression/classification).

4. Utiliser des techniques appropriées de détection d'anomalies statistiques telles que Z-score pour identifier les valeurs aberrantes/anomalies dans l'ensemble de données. Visualiser et analyser les anomalies détectées. Quelles conclusions pouvez-vous en tirer?

4.3 Manipulation des données

1. Appliquer le regroupement (grouping) et l'agrégation pour calculer des statistiques sommaires pour des catégories spécifiques. Vous pouvez utiliser la fonction ***groupby()*** et des fonctions d'agrégation telles que ***sum()***, ***mean()***, et ***count()*** pour créer des tables de synthèse.

2. Effectuer des opérations de filtrage (à l'aide de l'indexation booléenne basée sur des conditions spécifiques) et des opérations de tri pour extraire des sous-ensembles de données (afin d'obtenir des informations).

3. Appliquer des techniques d'analyse de séries temporelles pour découvrir des tendances temporelles dans le cas d'un jeu de données de séries temporelles. Utiliser des moyennes mobiles (rolling averages) ou d'autres fonctions de séries temporelles pour lisser le bruit dans les données.

4.4 Dérivation d'Informations (IA)

1. Assurer-vous d'avoir traité toutes les valeurs manquantes ou anomalies identifiées lors de l'examen du jeu de données.

2. Effectuer une analyse pour identifier les corrélations au sein d'un jeu de données sélectionné. Dans le cas d'un ensemble de données de séries temporelles, une analyse plus approfondie est nécessaire pour identifier la saisonnalité et les tendances au fil du temps.

3. Interpréter les résultats pour tirer des conclusions significatives.

Bonne chance