10/08/23

ANALISI DEI DATI INTRODUZIONE

master.D

Campari Mirko

master.D

L'ANALISI DEI DATI È IL PROCESSO DI ESAMINARE, PULIRE, TRASFORMARE E INTERPRETARE DATI AL FINE DI ESTRARRE INFORMAZIONI SIGNIFICATIVE E TRARRE CONCLUSIONI UTILI.

QUESTO PROCESSO COINVOLGE L'UTILIZZO DI VARIE TECNICHE STATISTICHE, ALGORITMI DI MACHINE LEARNING E STRUMENTI SOFTWARE PER COMPRENDERE I MODELLI, LE TENDENZE E LE RELAZIONI PRESENTI NEI DATI.

L'ANALISI DEI DATI È FONDAMENTALE IN MOLTEPLICI CAMPI, TRA CUI SCIENZE SOCIALI, BUSINESS, RICERCA SCIENTIFICA, FINANZA, MEDICINA E MOLTI ALTRI, POICHÉ AIUTA A PRENDERE DECISIONI INFORMATE E A FORMULARE STRATEGIE BASATE SUI DATI.

master.D

PYTHON È UNO DEI LINGUAGGI DI PROGRAMMAZIONE PIÙ POPOLARI E AMPIAMENTE UTILIZZATI PER L'ANALISI DEI DATI. CI SONO DIVERSE RAGIONI PER CUI PYTHON È PREFERITO IN QUESTO CONTESTO:

- SINTASSI LEGGIBILE E INTUITIVA: PYTHON È NOTO PER LA SUA SINTASSI SEMPLICE E LEGGIBILE, CHE LO RENDE MOLTO ADATTO SIA PER I PRINCIPIANTI CHE PER GLI ESPERTI. QUESTO FACILITA L'ANALISI E LA COMPRENSIONE DEL CODICE, RENDENDO PIÙ AGEVOLE IL LAVORO CON I DATI.

master.D

- AMPIA GAMMA DI LIBRERIE: PYTHON DISPONE DI NUMEROSE LIBRERIE SPECIALIZZATE PER L'ANALISI DEI DATI, COME NUMPY (PER CALCOLI SCIENTIFICI), PANDAS (PER LA MANIPOLAZIONE DEI DATI), MATPLOTLIB E SEABORN (PER LA VISUALIZZAZIONE), SCIKIT-LEARN (PER L'APPRENDIMENTO AUTOMATICO) E MOLTE ALTRE.

- QUESTE LIBRERIE SEMPLIFICANO COMPITI COMPLESSI, RIDUCENDO IL TEMPO NECESSARIO PER IMPLEMENTARE ALGORITMI E ANALISI.

master.D

- INTEGRAZIONE CON ALTRE TECNOLOGIE: PYTHON PUÒ ESSERE FACILMENTE INTEGRATO CON ALTRE TECNOLOGIE E STRUMENTI, INCLUSI DATABASE, CLOUD COMPUTING E BIG DATA FRAMEWORK.

- QUESTA FLESSIBILITÀ CONSENTE DI LAVORARE CON UNA VASTA GAMMA DI DATI PROVENIENTI DA DIVERSE FONTI.

master.D

- ACCESSO AI DATI: PYTHON FORNISCE LIBRERIE PER LEGGERE E SCRIVERE DIVERSI FORMATI DI DATI, COME CSV, EXCEL, JSON E MOLTO ALTRO. QUESTO SEMPLIFICA L'IMPORTAZIONE E L'ESPORTAZIONE DEI DATI DURANTE L'ANALISI.

master.D

IN SINTESI, PYTHON È SCELTO PER L'ANALISI DEI DATI A CAUSA DELLA SUA FACILITÀ D'USO, DELLE SUE LIBRERIE SPECIALIZZATE E DELL'AMPIO SUPPORTO DELLA COMUNITÀ, CHE LO RENDONO UNO STRUMENTO POTENTE PER ESPLORARE, ELABORARE E TRARRE CONOSCENZE DA INSIEMI DI DATI COMPLESSI.

master.D

ECCO ALCUNI PASSAGGI CHIAVE COINVOLTI NELL'ANALISI DEI DATI:

- RACCOLTA DEI DATI: RACCOGLIERE DATI DA VARIE FONTI, COME DATABASE, FILE, SENSORI O ALTRI STRUMENTI DI RILEVAMENTO.
- PULIZIA DEI DATI: ELIMINARE ERRORI, VALORI MANCANTI O DATI INCONSISTENTI CHE POTREBBERO COMPROMETTERE L'ACCURATEZZA DELL'ANALISI.

master.D

- ESPLORAZIONE DEI DATI: ANALIZZARE STATISTICHE DESCRITTIVE, COME MEDIE, DEVIAZIONI STANDARD E DISTRIBUZIONI, PER OTTENERE UNA VISIONE INIZIALE DEI DATI.

- VISUALIZZAZIONE DEI DATI: CREARE GRAFICI E VISUALIZZAZIONI PER RAPPRESENTARE I DATI IN MODO VISIVO, FACILITANDO LA COMPRENSIONE DEI MODELLI E DELLE RELAZIONI.

master.D

- TRASFORMAZIONE DEI DATI: APPLICARE OPERAZIONI DI MANIPOLAZIONE DEI DATI, COME AGGREGAZIONI, NORMALIZZAZIONI O TRASFORMAZIONI, PER PREPARARE I DATI PER ANALISI PIÙ AVANZATE.

- ANALISI STATISTICA: UTILIZZARE METODI STATISTICI PER IDENTIFICARE MODELLI, TENDENZE O RELAZIONI NEI DATI. QUESTO PUÒ INCLUDERE L'USO DI TEST IPOTETICI, REGRESSIONE E ALTRE TECNICHE.

master.D

- APPRENDIMENTO AUTOMATICO: IN ALCUNI CASI, UTILIZZARE ALGORITMI DI MACHINE LEARNING PER CREARE MODELLI PREDITTIVI O CLASSIFICATORI CHE POSSONO ESTRARRE INFORMAZIONI NASCOSTE NEI DATI.

- INTERPRETAZIONE E COMUNICAZIONE: TRARRE CONCLUSIONI DAI RISULTATI DELL'ANALISI E COMUNICARLE IN MODO CHIARO E COMPRENSIBILE AI DESTINATARI. QUESTO PUÒ COINVOLGERE LA CREAZIONE DI REPORT, PRESENTAZIONI O VISUALIZZAZIONI INTERATTIVE.

master.D

PYTHON PER L'ANALISI DEI DATI

PYTHON È DIVENTATO UNO DEI LINGUAGGI PREFERITI PER L'ANALISI DEI DATI GRAZIE ALLA SUA VERSATILITÀ E ALLA VASTA GAMMA DI LIBRERIE.

LIBRERIE ESSENZIALI PER L'ANALISI DEI DATI IN PYTHON:

NUMPY: PER LA GESTIONE EFFICIENTE DI ARRAY MULTIDIMENSIONALI E OPERAZIONI MATEMATICHE.

PANDAS: PER LA MANIPOLAZIONE E L'ANALISI DI DATI TABULARI.

MATPLOTLIB: PER LA CREAZIONE DI GRAFICI E VISUALIZZAZIONI.

master.D

NUMPY:

PER LA GESTIONE EFFICIENTE DI ARRAY MULTIDIMENSIONALI E OPERAZIONI MATEMATICHE

NUMPY (NUMERICAL PYTHON) È UNA LIBRERIA FONDAMENTALE PER L'ANALISI DEI DATI E IL CALCOLO SCIENTIFICO IN PYTHON. LE SUE CARATTERISTICHE PRINCIPALI INCLUDONO:

- ARRAY MULTIDIMENSIONALI: NUMPY INTRODUCE UN NUOVO TIPO DI DATO CHIAMATO "ARRAY", CHE CONSENTE DI RAPPRESENTARE E MANIPOLARE DATI MULTIDIMENSIONALI (COME MATRICI O TENSORI) IN MODO EFFICIENTE. QUESTO È ESSENZIALE PER MOLTI CALCOLI SCIENTIFICI E ANALITICI.

- OPERAZIONI MATEMATICHE: NUMPY OFFRE UNA VASTA GAMMA DI FUNZIONI MATEMATICHE E OPERAZIONI DI BASE CHE POSSONO ESSERE ESEGUITE DIRETTAMENTE SUGLI ARRAY. QUESTO SEMPLIFICA NOTEVOLMENTE IL CALCOLO DI OPERAZIONI COMPLESSE SU GRANDI QUANTITÀ DI DATI.

- BROADCASTING: IL BROADCASTING È UNA CARATTERISTICA DI NUMPY CHE CONSENTE DI ESEGUIRE OPERAZIONI SU ARRAY DI FORME DIVERSE IN MODO AUTOMATICO E COERENTE, EVITANDO LA NECESSITÀ DI CREARE LOOP ESPLICITI.

- EFFICIENZA: NUMPY È IMPLEMENTATO IN C E OFFRE PERFORMANCE OTTIMIZZATE PER OPERAZIONI MATEMATICHE E SCIENTIFICHE. INOLTRE, CONSENTE L'INTEGRAZIONE SEMPLICE CON ALTRI LINGUAGGI DI PROGRAMMAZIONE COME C E FORTRAN.

master.d

```
# Importare le librerie
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# --- NumPy ---

# Creare un array multidimensionale
array = np.array([[1, 2, 3], [4, 5, 6]])

# Eseguire operazioni matematiche sugli array
sum_array = np.sum(array)
mean_array = np.mean(array)

# Broadcasting: eseguire operazioni su array con forme diverse
scaled_array = array * 2
```

master.D

PANDAS:

PER LA MANIPOLAZIONE E L'ANALISI DI DATI TABULARI

PANDAS È UNA LIBRERIA FOCALIZZATA SULLA MANIPOLAZIONE, ANALISI E PULIZIA DI DATI TABULARI, COME QUELLI IN FORMATO CSV, EXCEL O DATABASE. LE SUE CARATTERISTICHE INCLUDONO:

- DATAFRAME: IL CUORE DI PANDAS È IL CONCETTO DI "DATAFRAME", UNA STRUTTURA DATI BIDIMENSIONALE CHE PUÒ CONTENERE DATI ETEROGENEI ORGANIZZATI IN RIGHE E COLONNE. I DATAFRAME CONSENTONO LA MANIPOLAZIONE E L'ANALISI DEI DATI IN MODO FLESSIBILE.
- OPERAZIONI DI SELEZIONE E FILTRAGGIO: PANDAS OFFRE UN'AMPIA GAMMA DI FUNZIONI PER SELEZIONARE, FILTRARE E MODIFICARE I DATI ALL'INTERNO DEI DATAFRAME.
- GESTIONE DEI DATI MANCANTI: PANDAS FORNISCE STRUMENTI PER GESTIRE I VALORI MANCANTI NEI DATI, CONSENTENDO DI EFFETTUARE ANALISI PIÙ ACCURATE E COMPLETE.
- AGGREGAZIONE E RAGGRUPPAMENTO: È POSSIBILE AGGREGARE E RAGGRUPPARE I DATI IN BASE A DIVERSE CONDIZIONI, CONSENTENDO ANALISI STATISTICHE E AGGREGAZIONI PERSONALIZZATE.
- INTEGRAZIONE CON ALTRE LIBRERIE: PANDAS È SPESSO UTILIZZATO IN COMBINAZIONE CON NUMPY E ALTRE LIBRERIE DI ANALISI DEI DATI PER OTTENERE ANALISI COMPLESSE E INFORMATIVE.

master.D

```
# Importare le librerie
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# --- Pandas ---

# Creare un DataFrame da un dizionario
data = {'Nome': ['Alice', 'Bob', 'Charlie'],
        'Età': [25, 30, 22]}
df = pd.DataFrame(data)

# Selezione delle righe in base a condizioni
young_people = df[df['Età'] < 30]

# Modifica dei dati nel DataFrame
df.loc[0, 'Età'] = 26

# Calcolare statistiche sui dati
mean_age = df['Età'].mean()
```

master.D

MATPLOTLIB:

PER LA CREAZIONE DI GRAFICI E VISUALIZZAZIONI

MATPLOTLIB È UNA LIBRERIA AMPIAMENTE UTILIZZATA PER CREARE GRAFICI E VISUALIZZAZIONI IN PYTHON. LE SUE CARATTERISTICHE INCLUDONO:

- VARIE TIPOLOGIE DI GRAFICI: MATPLOTLIB OFFRE UN'AMPIA GAMMA DI GRAFICI TRA CUI SCEGLIERE, INCLUSI ISTOGRAMMI, SCATTER PLOT, GRAFICI A BARRE, GRAFICI A DISPERSIONE E ALTRO ANCORA.
- PERSONALIZZAZIONE: È POSSIBILE PERSONALIZZARE TUTTI GLI ASPETTI DEI GRAFICI, TRA CUI COLORI, ETICHETTE DEGLI ASSI, TITOLI E STILI DI LINEA.
- VISUALIZZAZIONI INTERATTIVE: MATPLOTLIB SUPPORTA INTERATTIVITÀ ATTRAVERSO DIVERSE BACKEND E PUÒ ESSERE UTILIZZATO ANCHE IN COMBINAZIONE CON LIBRERIE COME JUPYTER NOTEBOOK PER CREARE GRAFICI INTERATTIVI.
- SUPPORTO PER LA PUBBLICAZIONE: I GRAFICI CREATI CON MATPLOTLIB POSSONO ESSERE ESPORTATI IN DIVERSI FORMATI, COME IMMAGINI O PDF, RENDENDOLI ADATTI PER PUBBLICAZIONI ACCADEMICHE E RAPPORTI.
- INTEGRAZIONE CON ALTRE LIBRERIE: MATPLOTLIB È SPESSO UTILIZZATO INSIEME A PANDAS PER VISUALIZZARE I RISULTATI DELLE ANALISI IN MODO CHIARO E COMPRENSIBILE.

```
# Importare le librerie
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# --- Matplotlib ---

# Creare un grafico a dispersione
x = np.array([1, 2, 3, 4, 5])
y = np.array([10, 7, 15, 8, 12])
plt.scatter(x, y)
plt.title('Grafico a Dispersione')
plt.xlabel('X')
plt.ylabel('Y')

# Mostrare il grafico
plt.show()
```

master.D

Oltre a NumPy, Pandas e Matplotlib, ci sono molte altre librerie e tecniche che Python utilizza nell'analisi dei dati. Ecco alcuni esempi:

SciPy: Estensione di NumPy che fornisce funzioni aggiuntive per l'ottimizzazione, l'interpolazione, l'integrazione numerica, il calcolo delle statistiche e molto altro.

Scikit-learn: Una libreria di machine learning che offre strumenti per il clustering, la classificazione, la regressione e l'elaborazione di dati complessi.

Seaborn: Una libreria per la visualizzazione statistica basata su Matplotlib. Offre stili di grafici preimpostati e funzioni per visualizzazioni statistiche avanzate.

tatsmodels: Libreria per l'analisi delle statistiche e l'adattamento di modelli di regressione, serie temporali e altri tipi di analisi.

NLTK (Natural Language Toolkit): Libreria per l'elaborazione del linguaggio naturale, utilizzata per l'analisi testuale e l'elaborazione di testi.

Beautiful Soup e Scrapy: Librerie utilizzate per il web scraping, cioè l'estrazione di dati da pagine web.

NetworkX: Libreria per l'analisi delle reti e dei grafi, spesso utilizzata per studiare relazioni complesse tra entità.

TensorFlow e PyTorch: Framework di machine learning e deep learning che consentono di creare e addestrare modelli neurali complessi.

Pymc3 e Stan: Librerie per l'analisi bayesiana, utilizzate per modellare incertezza e complessità nelle analisi.

master.D

Hadoop e Spark: Framework per il calcolo distribuito e l'elaborazione di big data.

SQLAlchemy: Libreria per l'interazione con database SQL attraverso Python.

Principal Component Analysis (PCA): Una tecnica di riduzione della dimensionalità che semplifica i dati mantenendo le caratteristiche più importanti.

Time Series Analysis: Approccio all'analisi dei dati temporali per identificare modelli e tendenze nel tempo.

Natural Language Processing (NLP): Un campo dell'IA che si concentra sull'elaborazione del linguaggio naturale, utilizzato per analizzare e comprendere testi.