Challenge on data

Accenture - Università della Calabria

accenture

Indice

Introduzione

La Challenge descritta nel presente documento ha l'obiettivo di proporre un caso d'uso afferente al mondo dell'Advanced Data Analytics, delineando il percorso che i dati dovranno attraversare, dalla sorgente al modello di visualizzazione finale.

Nello specifico, la Challenge può essere sintetizzata dai seguenti step, che verranno meglio descritti nel seguito del documento:

1. Individuazione del/dei dataset sorgente e scaricamento a partire dalla fonte
2. Analisi preliminare dei dati per valutarne la qualità
3. Pulizia e trasformazione dei dati contenuti nel dataset
4. Analisi sui dati
5. Sviluppo di un modello basato su Machine-Learning
6. Esposizione dei risultati mediante dashboard

Dataset e tecnologie

Dataset: Open Data Regione Calabria

Il Dataset scelto per la challenge è il Parco Veicoli circolanti in Regione Calabria, aggiornato al 2024 con una profondità storica dal 1998 al 2022.

Il file da utilizzare è quello contenuto nell'archivio "Parco_Veicoli_2024.zip" scaricabile dal link sopra.

Il dataset conterrà le seguenti colonne:

- Progressivo: numero progressivo assegnato al veicolo
- Tipo_Targa: tipologia di targa (non è presente un'anagrafica)
- Utilizzo: tipologia di utilizzo (non è presente un'anagrafica)
- Specialità
- Rimorchiabilità: possibilità di agganciare un rimorchio al veicolo
- Cilindrata: cilindrata del motore espressa in cc
- Alimentazione: tipo di alimentazione (non è presente un'anagrafica)
- Portata: portata del veicolo espressa in KG
- Num_Posti: numero di posti
- Tipo_Impianto: tipologia impianto Gas/Metano (non è presente un'anagrafica)
- Disistallazione_Impianto: colonna correlata alla precedente
- Kw: potenza del veicolo
- Anno_Immatricolazione: anno di prima immatricolazione
- Num_Assi: numero assi, ovvero coppie di ruote
- Euro: classe di emissioni
- Co2: quantitativo di CO2 immessa nell'ambiente
- Peso_Complessivo: peso del veicolo
- Cap_Residenza_Propr: CAP di residenza del proprietario
- Prov_Residenza: Provincia di residenza del proprietario

Tecnologie: Python e librerie open-source

Per lo sviluppo della Challenge si suggerisce l'uso delle seguenti tecnologie, assumendo che il linguaggio di programmazione sia **Python**:

Purpose	Tecnologia
Ambiente di sviluppo	Un IDE (ad es. Pycharm), o in alternativa di un notebook editor (es. Jupyter, Google Colab), la scelta dipende dalle proprie preferenze. Vista la natura dei task, è suggerito l'uso di un notebook editor.
Data Analysis e Data Transformation/Cleaning	Librerie per la rappresentazione e manipolazione dei dati in un formato table-like (es. Pandas)
Modello di analisi	Librerie per l'addestramento di modelli di machine-learning. (es. Scikit-learn)
Visualizzazione dei dati	Librerie per la creazione di visualizzazioni (es. Plotly matplotlib)

Svolgimento della Challenge

Lo svolgimento della Challenge prevede il completamento dei vari step di seguito descritti, nell'ordine corretto.

STEP 1 - Ambiente di sviluppo e dataset

Il primo step prevede la configurazione di un ambiente di sviluppo, tra quelli suggeriti nel paragrafo "Tecnologie".

Nel caso in cui si scelga un notebook editor online (es. Google Colab, Jupyter Lab), potrebbero essere applicati limiti di dimensione massima dei file da caricare e/o limiti in termini di CPU/RAM assegnate all'ambiente:

- Nel primo caso, procedere allo split del file sorgente in più file di dimensione minore;
- Nel secondo caso, estrarre dal file sorgente un sotto-insieme di righe, per ridurre il carico sulle risorse (in particolare, durante l'apprendimento del modello di analisi).

In tutti i casi, occorrerà installare le librerie riportate nel paragrafo "Tecnologie" (e ogni altra libreria aggiuntiva o sostitutiva) mediante comando "pip install".

Una volta preparato l'ambiente, occorrerà importare il dataset (file CSV), prestando attenzione ai caratteri di separazione e delimitazione del testo, nonché alla presenza dell'header.

STEP 2 - Analisi preliminare dei dati

Una volta caricato il dataset in una struttura dati in Python (ad es. Pandas), occorrerà eseguire un'analisi preliminare dei dati, volta a ottenere le informazioni aggregate e valutare eventuali step di pulizia dei dati.

Una buona pratica è quella di rinominare le colonne utilizzando una convenzione. In questo caso, si suggerisce l'uso della "snake_notation", ovvero "nome_colonna".

Alcuni esempi di operazioni di analisi (eseguirne almeno 3):

- Numero di record totale.
- Numero di record per singola colonna (es. anno immatricolazione) o combinazione di esse (es. anno immatricolazione, provincia, classe emissioni euro), ordinato in maniera decrescente.
- Valori unici della colonna "provincia di residenza".
- Valori unici della colonna "anno di immatricolazione", verificando l'eventuale presenza di buchi temporali nel periodo di osservazione (1998-2022).
- Cilindrata media delle auto per provincia e anno di immatricolazione.
- Numero di auto ad alimentazione elettrica ("ELE").

STEP 3 - Pulizia e trasformazione dei dati

Di seguito sono elencate le operazioni di trasformazione/pulizia da effettuare sul dataframe, al fine di preparare quello che sarà l'input dello step di addestramento di un modello di analisi. I nomi di colonne riportati sono solo indicativi, poiché è richiesto che tutti i dataframe lavorati adottino la convenzione snake nei nomi delle colonne.

1. Rimuovere dal dataframe i record che presentano valore nullo (0) o vuoto (null) in almeno una delle colonne: "Provincia Residenza", "Alimentazione", "Anno immatricolazione", "CO2", ma…

ATTENZIONE; il dataset contiene anche record che fanno riferimento ad auto elettriche (“alimentazione”=”ELE”), con CO2 nulla (0) o vuota (null). Tali record **non devono essere rimossi** dal dataframe.

2. Selezionare le colonne “Provincia Residenza”, “Anno immatricolazione”, “CO2” del dataframe.
3. Sostituire i “null” nella colonna CO2 col valore 0, in modo da facilitare l’apprendimento del modello. È importante osservare che se le precedenti operazioni saranno svolte correttamente, allora i record con valore di CO2 “null” saranno tutti e soli quelli afferenti ad auto elettriche.
4. Aggregare i dati, in particolare calcolando la media di CO2 per “Provincia Residenza” e “Anno immatricolazione”. Tale media dovrà essere convertita in intero, facendo un arrotondamento a zero cifre decimali.

Il dataframe risultante dovrebbe presentarsi come (i dati sono solo esemplificativi):

provincia_residenza	anno_immatricolazione	co2_media
CS	2021	132
CS	2022	56
…	…	…

STEP 4 - Addestramento di un modello di analisi e calcolo delle predizioni

L’obiettivo di questo step è quello di addestrare un semplice modello di regressione, utilizzando la libreria scikit-learn, al fine di prevedere la C02 media, per ciascuna provincia di residenza, per l’anno successivo al massimo anno di osservazione nei dati (ovvero, per il 2023).

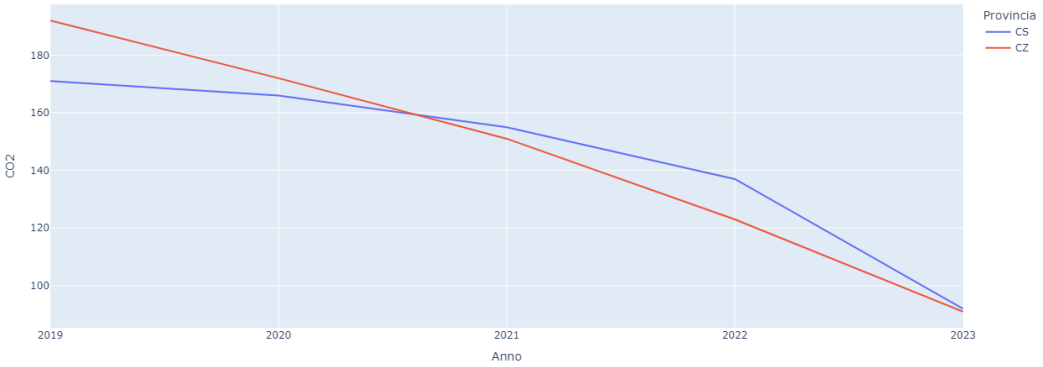
A tal fine, si suggerisce il seguente work-flow:

1. Il dataframe in input è quello in output allo Step 3.
2. Inizializzare un dizionario vuoto che conterrà per ogni chiave (provincia), la predizione della CO2 media per l’anno di immatricolazione 2023.
3. Ottenere i valori distinti della colonna “provincia_residenza”, memorizzandoli in una struttura di appoggio.
4. Iterare sulle varie province. Per ciascuna iterazione:
    a. Filtrare i dati del dataframe di input sulla provincia iterata.
    b. Suddividere il dataframe risultante in training set (80%) e test set (20%).
    c. Utilizzare il training set per l’addestramento del modello.
    d. Utilizzare il test set per la valutazione dell’accuratezza (RMSE).
    e. Calcolare la predizione per l’anno 2023.
    f. Memorizzare la predizione relativa alla particolare provincia nel dizionario di appoggio.
5. Il dizionario è l’output del presente step.

STEP 5 - Visualizzazione dei risultati delle predizioni

Lo step finale prevede la creazione di una visualizzazione che mostri con chiarezza, in un line plot, l’andamento dell’emissione di CO2 media per anno di immatricolazione e provincia di residenza. A tal fine, è suggerito l’uso di una libreria come Plotly.

Di seguito, a titolo solo esemplificativo, un possibile output di questo STEP:



Più nello specifico, il line plot sarà caratterizzato da:

- Una linea per ciascuna provincia di residenza, associata ad un colore, riportato in una legenda.
- L'asse x contiene gli anni di immatricolazione dal meno recente a quello su cui è effettuata la predizione, ovvero il 2023 (valutare di mostrare solo gli ultimi 5-6 anni, per migliorare la leggibilità).
- L'asse y fa riferimento alle emissioni medie di CO2.
- **OPZIONALE**: se possibile, evidenziare con un artefatto visivo (es. un label) la predizione dell'anno 2023.

È importante notare che l'input del line plot deve essere il dataframe in output allo step 3, con in aggiunta le righe corrispondenti alle predizioni memorizzate nel dizionario in output allo step 4. A tal fine, si richiede una soluzione che preveda il minor numero di righe di codice.

STEP 6 - Nuova predizione [NECESSARIO PER SPECIALISTICA/OPZIONALE PER TRIENNALE]

L'obiettivo di questo step è quello di addestrare un secondo modello, proposto dal gruppo di lavoro sulla base dei dati utilizzati, per predire un'altra variabile di interesse (da proporre) sempre mediante la libreria scikit-learn.
In maniera analoga al modello precedente i risultati dovranno essere visualizzati e commentati.