Obesity Predictor

Francesco Ferrara 0512109789

January 16, 2024

INDICE

Performance	La misura di performance dell'agente è la sua capacità di categorizzare correttamente lo stato del paziente. Questa verrà valutata tramite le metriche di valutazione quali precisione, accuratezza e recall.
Environment	L'ambiente in cui l'agente opera è quello dei dati clinici riguardo l'obesità, inclusi parametri come il peso e l'età. L'ambiente è: • **Completamente osservabile:** in quanto si ha accesso a tutte le informazioni in ogni momento; • **Stocastico:** in quanto le diagnosi possono essere influenzate da fattori con una certa variabilità; • **Sequenziale:** in quanto le azioni che la persona compie come la dieta o lo sport, hanno ripercussioni nel tempo; • **Dinamico:** in quanto la persona può evolvere durante il tempo; • **Discreto:**le variabili assumono valori in un determinato intervallo; • **A singolo agente:** in quanto l'unico agente che opera in questo ambiente è quello in oggetto.
Actuators	Gli attuatori dell'agente sono i modelli di machine learning addestrati.
Sensors	Il sensore dell'agente è rappresentato dalla tastiera (tramite essa l'agente riceve gli input) e dal dataset che contiene le informazioni relative alle caratteristiche.

3 CRISP-DM

Il modello che seguirò durante lo sviluppo del progetto è il CRISP-DM. Il CRISP-DM (Cross-Industry Standard Process for Data Mining) è un modello non sequenziale che rappresenta il ciclo di vita di progetti basati su intelligenza artificiale e data science.

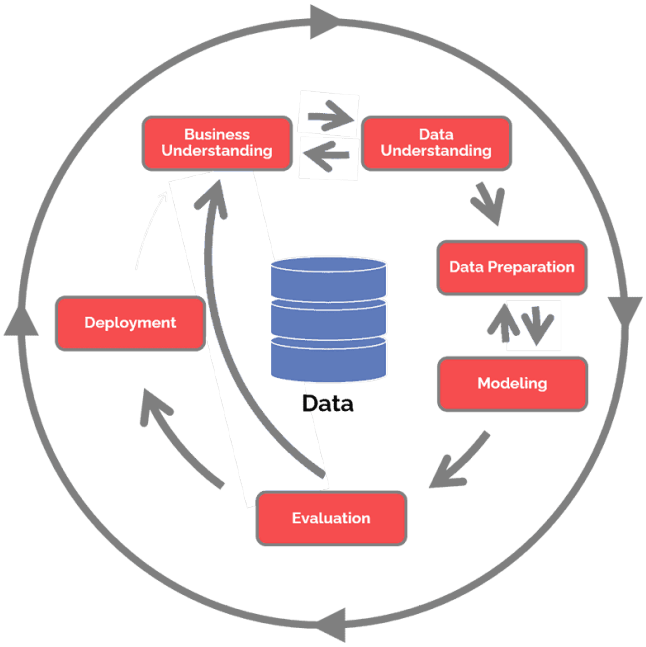

Figure 1: Modello CRISP-DM

Di seguito vengono spiegate nel dettaglio le varie fasi.

4 Business Understanding

La prima fase è quella di raccolta dei requisiti e di definizione degli obiettivi di business che si vogliono raggiungere (ossia cosa deve fare il machine learner) e dei cosiddetti business success criteria, ovvero i criteri secondo i quali potremo accertare che il sistema costruito è in linea con gli obiettivi di business.

Inoltre bisogna determinare la disponibilità delle risorse, stimare i rischi e condurre una analisi costi-benefici. Oltre a ciò, vanno definiti gli obiettivi tecnici che si intendono raggiungere e le tecnologie ed i tool necessari agli obiettivi.

Andiamo quindi a definire i vari step:

• **Obiettivi di business**: Il machine learner da progettare avrà l'obiettivo di stimare o predire il valore della variabile dipendente: nel nostro caso dovrà predire lo stato di salute della persona, ossia se è obesa oppure no.

• **Risorse disponibili**: Per raggiungere l'obiettivo ci serviremo di un dataset contenente i dati relativi all'obesità.

• **Stima dei rischi**:Il problema in questione non è di facile risoluzione ma il dataset preso in esame risulta essere abbastanza completo.

• **Tecnologie e tool**: utilizzerò Kaggle per l'acquisizione del dataset, Python come linguaggio in quanto è il più adatto per l'analisi di dataset e Google Colab perchè ha integrato Jupyter Notebook per la creazione e visualizzazione delle azioni svolte sul dataset.

5 Data Understanding

La seconda fase consiste nell'identificazione, collezione e analisi dei dataset che possono portare al raggiungimento degli obiettivi definiti precedentemente. Inoltre andremo a descrivere e visualizzare i dati e ne controlleremo la qualità.

5.1 Identificazione del dataset

Per affrontare il problema in esame userò Kaggle, una piattaforma online che contiene migliaia di dataset su diversi ambiti.

In particolare, il dataset scelto è il seguente: https://www.kaggle.com/datasets/tathagatbanerjee/obesity-dataset-uci-ml?select=ObesityDataSet_raw_and_data_sinthetic.csv

Questo dataset è stato scelto rispetto ad altri due, che riguardavano la stessa analisi, in quanto più completo.

5.2 Analisi del dataset

Il dataset in questione si presenta ben strutturato ed è formato da 2111 righe e 17 colonne. A questo punto si può passare alla sua descrizione dettagliata.

5.2.1 Descrizione dei dati

I dati presentano le seguenti caratteristiche:

1. **Gender**: indica se il soggetto è maschio(Male) o femmina(Female);
2. **Age**: indica l'età dei partecipanti;
3. **Height**: indica l'altezza dei partecipanti in metri;
4. **Weight**: indica il peso dei partecipanti in kg;
5. **family_history_with_overweight**: è yes se un familiare del partecipante era o è obeso, no altrimenti;
6. **FAVC**: è yes se il soggetto mangia frequentemente cibi ad alto quantitativo calorico, no altrimenti;
7. **FCVC**: indica se il soggetto mangia verdure durante i pasti (1 sta per Never, 2 per Sometimes, 3 per Always);
8. **NCP**: rappresenta quanti pasti principali il soggetto ha durante la giornata, in una scala tra 1 e 4;
9. **CAEC**: indica se il soggetto mangia tra i pasti (no, Sometimes, Frequently, Always);

10. **SMOKE**: yes se il soggetto fuma, no altrimenti;
11. **CH2O**: indica quanta acqua beve il soggetto in una scala da 1 a 3;
12. **SCC**: yes se il soggetto monitora le calorie, no altrimenti;
13. **FAF**: indica quanto spesso il soggetto svolge attività fisica in una scala da 0 a 3.
14. **TUE**: indica quanto tempo il soggetto passa davanti a device elettronici in una scala da 0 a 2.
15. **CALC**: indica quanto spesso il soggetto beve (no, Sometimes, Frequently, Always);
16. **MTRANS**: indica quale mezzo di trasporto viene usato di solito dal soggetto(Automobile, Motorbike, Public_Transportation, Walking, Bike);
17. **NObeyesdad**: indica la condizione del soggetto (Normal_Weight, Overweight_Level_I, Overweight_Level_II, Obesity_Type_I, Insufficient_Weight, Obesity_Type_II, Obesity_Type_III);

La variabile dipendente scelta è **NObeyesdad**.

5.2.2 Data Quality

In questa sezione verranno esaminate le questioni relative alla presenza di dati duplicati e dati mancanti.

Grazie al metodo duplicated() di Python è possibile vedere che il dataset contiene 24 duplicati, ossia l'1,14% del totale. Dato che il numero è molto basso ho deciso di rimuoverli.
Per quanto riguarda i dati mancanti, grazie a Kaggle è possibile che non ce ne sono.

Oltre a questo ho notato che alcune colonne presentano dati non interi: ad esempio nella colonna dell'età alcune celle sono 20.9, 22.5 ecc. Credo quindi che sia opportuno correggere questi valori e approssimarli ad un intero.
Analizzando poi la variabile dipendente, essa presenta la seguente distribuzione delle etichette:

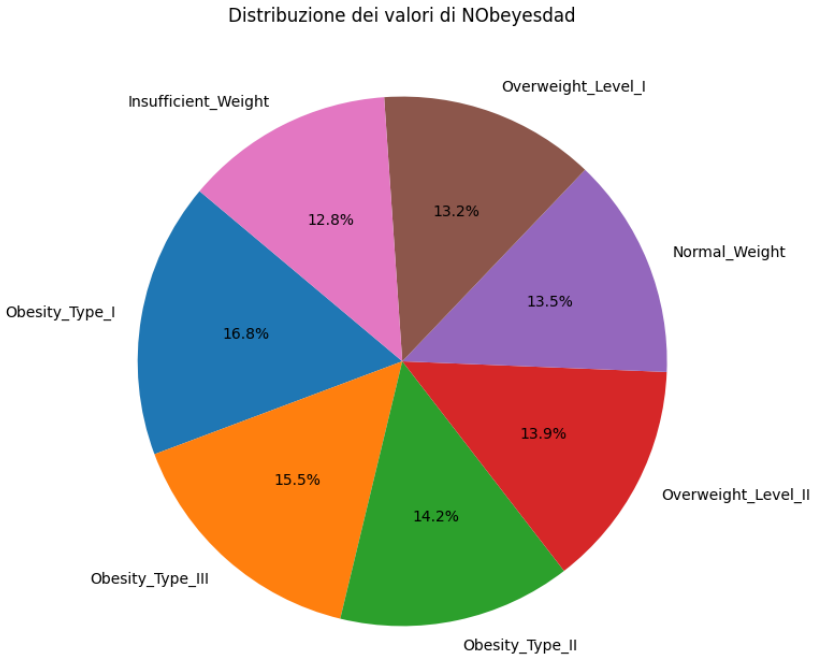

Figure 2: Enter Caption

Sebbene la distribuzione sia già piuttosto bilanciata, credo sia opportuno cercare di migliorare la situazione tramite tecniche di undersampling o oversampling nelle fasi successive.

Andrà fatta un'analisi riguardo il loro numero: si può decidere di accorpare le etichette dell'Obesità e quelle relative all'Overweight, e quindi considerare solo 4 etichette totali.

6 Data Preparation

In questa fase del CRISP-DM verranno preparati i dati in modo che possano essere utilizzati da un algoritmo di Machine Learning. Le quattro operazioni che saranno effettuate sono:

1. **Data Cleaning:** verrà effettuata la pulizia dei dati, ossia rimozione di valori nulli e di duplicati ed eventuale data imputation;
2. **Feature Scaling:** verrà effettuata la trasformazione, normalizzazione e scalatura dei dati;
3. **Feature Selection:** verranno selezionate le caratteristiche più rilevanti del problema in esame;
4. **Data Balancing:** verranno applicate tecniche di oversampling o undersampling se necessario al bilanciamento del dataset.

6.1 Data Cleaning

In fase di Data Understanding è stata riscontrata la presenza di alcuni dati duplicati, quindi si procederà alla loro rimozione.

Non sarà poi necessaria la rimozione di valori null in quanto è già stata verificata la loro assenza su Kaggle.

6.2 Feature Scaling

Alcune colonne come Age, Weight e Height presentano valori in R: sebbene questa scelta sia corretta per variabili come il peso e l'altezza, perde la sua rilevanza per un concetto come l'età.

Ho quindi ritenuto opportuno rendere interi i valori di Age e Weight, e approssimare alla seconda cifra decimale i valori di Height.

C'è poi lo stesso problema con le colonne CH2O, FAF, TUE, FCVC e NCP: siccome rappresentano valori in scala andrò a renderle intere.

A questo punto ci troviamo con tutti i dati normalizzati ma non tutti hanno lo stesso dominio di valori: alcune colonne presentano dati numerici, mentre altre variabili categoriche.

Dato che alcuni modelli di apprendimento potrebbero incontrare problemi con variabili categoriche, trasformo tutte le feature in variabili numeriche. In particolare, le feature **Gender, family_history_with_overweight, FAVC, SMOKE ed SCC** avranno valori binari 0 ed 1, mentre le feature **CAEC, CALC ed MTRANS** avranno valori da 0 a 3 per le prime due e da 0 a 4 per MTRANS.

A questo punto si può passare allo step successivo.

6.3 Feature Selection

6.4 Data Balancing

7 Data Modeling

8 Evaluation

9 Deployment

10 Bibliografia

https://www.sciencedirect.com/science/article/pii/S2352340919306985