

Pipeline per analisi dati

1. Introduzione

L'analisi dati è un processo fondamentale per trasformare dati grezzi, apparentemente inutili, in informazioni preziose per decisioni ponderate.

Per garantire il successo di una corretta analisi dati, bisogna prima strutturare una pipeline la quale possa garantire efficienza e una replicabilità.

Una pipeline all'interno del mondo dei dati, rappresenta una serie di fasi sequenziali, che guidano il flusso di lavoro dalla definizione del problema fino alla sperimentazione finale.

Questo documento andrà ad esplorare tutte le fasi necessarie per garantire un pipeline ben strutturata.

2. Fasi della pipeline

2.1. Definizione del problema

La definizione del problema è forse la fase più importante della pipeline, bisogna comprendere quale problema si andrà ad affrontare attraverso modelli di machine learning. Lo scopo è quindi definire le informazioni che si vogliono estrarre dai dati e attraverso quale strategia.

La definizione del problema non finisce con il concetto di business che si vuole applicare, ma comprende anche come verrà affrontato. Bisognerà capire dunque che tipo di machine learning verrà applicato.

2.1.1. Supervisioned learning

I modelli di supervised learning, vengono così chiamati perché i dati sono strutturati in un certo modo. La suddivisione prevede la parte denominata "Data" e la parte chiamata "Label". Il concetto è semplice, i modelli supervisionati cercano tramite le informazioni presenti nella porzione "Data" a trovare dei pattern per predire delle "etichette". La parte supervisionata avviene durante il training del modello, in cui se la previsione della "label" dovesse essere sbagliata, il modello cercherà di correggersi. Le categorie principali per problemi supervisionati di machine learning sono:

- **Classification:** Viene utilizzata per i problemi che necessitano un'assegnazione di etichetta binaria. Un esempio pratico può essere la classificazione delle mail in spam o non spam;
- **Regression:** Utilizzata per problemi che necessitano un valore continuo. La differenza principale con la "Classification" è appunto l'output che deve essere una variabile continua e non discreta. Un esempio è la predizione del valore di una casa attraverso l'immissione di diversi parametri;
- **Raccomandation:** "Raccomandation" è un problema che prevede la raccomandazione di prodotti o servizi a degli utenti basandosi sullo storico dei dati o dei comportamenti.

2.1.2. Unsupervised learning

L'approccio ai modelli di Unsupervised learning avviene quando disponiamo dei dati ma non delle "etichette". Un problema di machine learning non supervisionato è la segmentazione della clientela, si cerca quindi di raggruppare i diversi utenti per diversi parametri di somiglianza all'interno di cluster. Questo permetterà di assegnare delle etichette in base ai cluster formati.

2.2. Raccolta e Gestione dei dati

La raccolta e gestione dei dati è un altro passaggio chiaramente fondamentale per la pipeline, senza dati non si può analizzare niente. La raccolta dei dati si suddivide principalmente in due

categorie, stream di dati costanti che possono derivare da diverse fonti per defluire in un data warehouse o data lake, oppure in dati più statici che non hanno un flusso continuo come file CSV o xlsx.

Bisogna fare anche una distinzione dei dati raccolti e di quale categoria questi facciano parte. E' possibile ottenere dati strutturati, facilmente comprensibili e analizzabili, hanno una struttura ben definita, possono essere disposti in maniera tabellare e utilizzano tipi di dato ben definiti (interi, date, booleani, float...).

Dati invece non strutturati possono essere come immagini, audio, linguaggio testuale. Categoria di dati che è più difficile disporre all'interno di una tabella o database.

2.3. Pulizia e Preprocessing dei dati

2.4. Esplorazione e Analisi dei dati

2.5. Ingegnerizzazione delle caratteristiche

2.6. Sviluppo e Addestramento dei modelli

2.7. Valutazione e Validazione dei modelli

2.8. Documentazione e Comunicazione