zero11

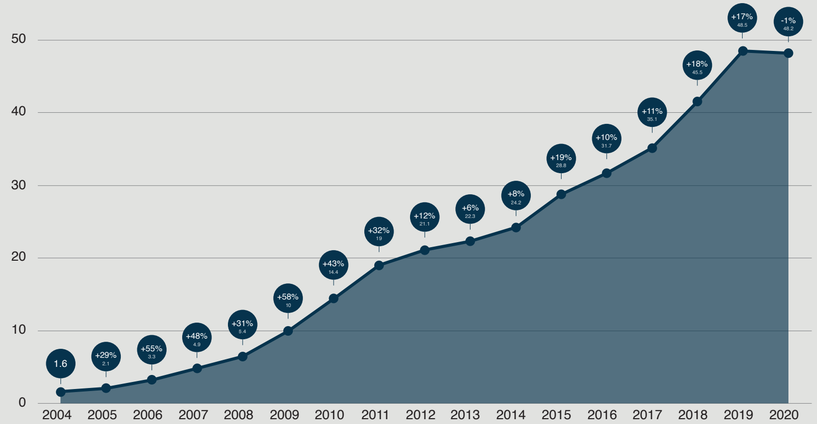
Politecnico di Torino

Analisi del comportamento dei clienti di un e-commerce per prevedere le intenzioni d'acquisto

Candidato: **Simone Sinagra**
Relatore: **Luigi De Russis**
Tutor aziendale: **Michele Sonnessa**

E-commerce e Customer eXperience (CX)

- Record del fatturato pari a 48.5 miliardi di € nel 2019
- Compound Annual Growth Rate del 25.5% nel periodo 2004-2019
- Entrata nel mercato digitale di 10.467 nuove imprese nel 2020



Variazione % del fatturato e-commerce in miliardi di €

Fonte dati: Casaleggio Associati - Report 2021

Customer Journey

Customer eXperience: esperienza complessiva dei clienti che si relazionano con l'azienda, percorrendo il cosiddetto **Customer Journey**

Come migliorare la CX?

Utilizzo di **contenuti dinamici**

Per poter offrire una CX più rilevante si possono mostrare contenuti che si adattano dinamicamente in base al cliente:

- Azioni compiute dal cliente durante la visita al sito
- Caratteristiche demografiche del cliente o del dispositivo utilizzato
- Interazioni passate del cliente con il sito

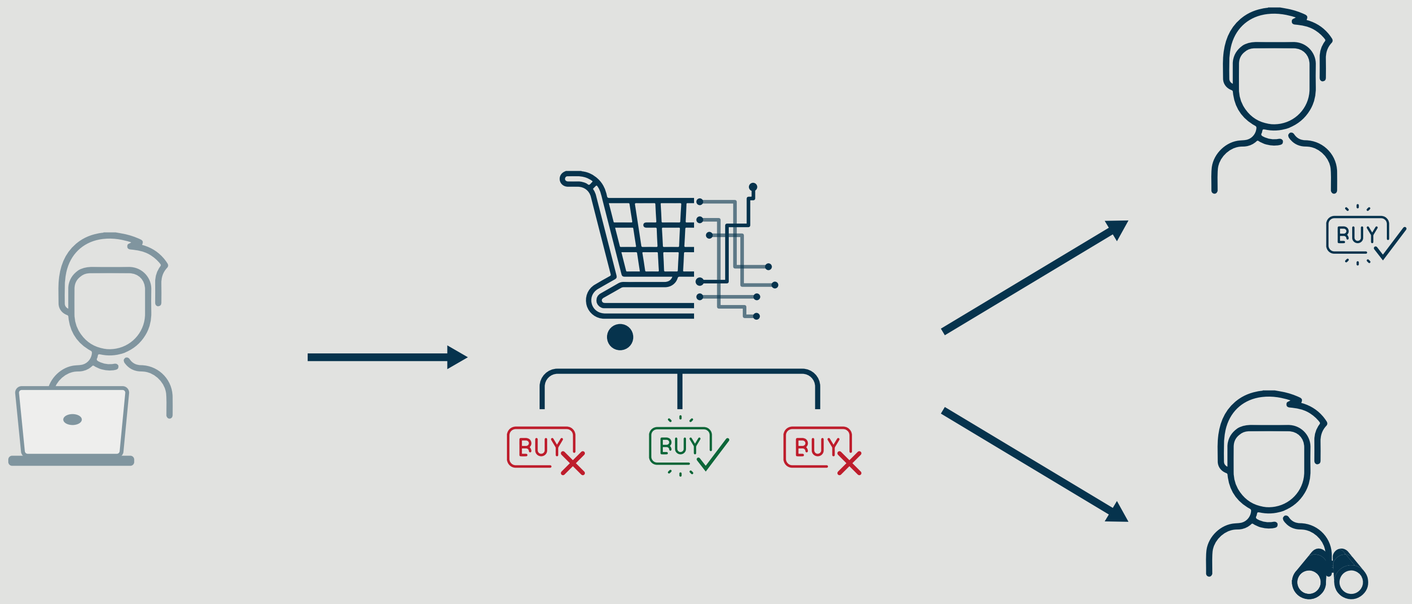
Big Data

- Strumenti di monitoraggio raccolgono i dati delle azioni dei clienti dell’e-commerce
- Aumento della potenza di calcolo e dello spazio di archiviazione

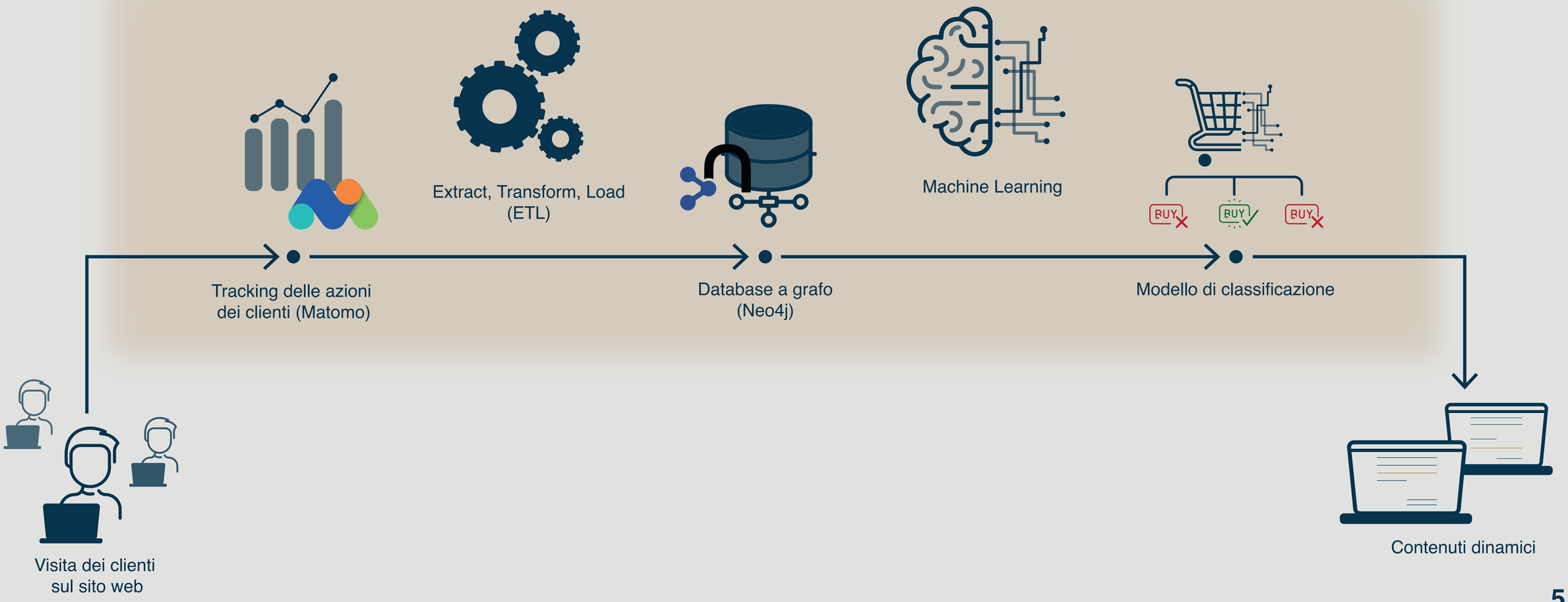
Analisi efficiente di grandi volumi di dati per poter estrarre informazioni che aggiungono valore al business aziendale

Obiettivo della Tesi

Creare un modello predittivo che permetta di distinguere le visite dei clienti intenzionati ad acquistare dalle visite dei clienti che non effe-tuano alcun acquisto



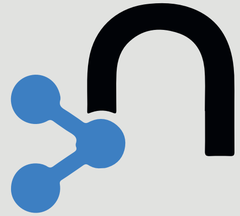
Processo di sviluppo



Tecnologie utilizzate

Python

- Linguaggio non compilato: richiede un interprete Python che permetta di convertire il codice in bytecode
- Ben fornita di librerie per la manipolazione dei dati e degli algoritmi di machine learning (Pandas, Scikit-learn e Matplotlib)



Neo4j

- Database open source NoSQL a grafo
- Struttura a nodi e archi che rappresentano entità e relazioni
- Linguaggio di query (DDL, DML, DQL) Cypher

Database a grafo

- Flessibilità
- Prestazioni eccellenti per strutture altamente connesse
- Scalabilità orizzontale

Extract, Transform, Load (ETL)

Estrazione dei dati

- Richiesta HTTP delle visite del sito all'API Matomo
- Libreria Python Requests
- Formato JSON

Preparazione dei dati

- Visualizzazione e manipolazione dei dati ottenuti
- Libreria Python Pandas
- DataFrame (struttura dati tabulare bidimensionale)

Caricamento dei dati

- Collegamento alla base di dati Neo4j
- Community drivers di Neo4j (package Python neo4j)
- Base di dati a grafo

Estrazione dei dati

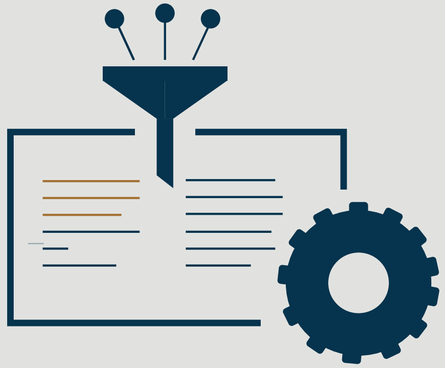
I dati ottenuti, in formato JSON, rappresentano le visite dei clienti sul sito.

Una visita contiene 103 attributi, tra cui l’attributo “actionDetails”.

Ad ogni visita è associato un numero variabile di azioni che contengono un massimo di 36 attributi.

Le azioni possono essere di 7 tipi: action, search, event, ecommerceOrder, ecommerceAbandonedCart, outlink, download

Preparazione dei dati



1. Riduzione della dimensionalità del DataFrame:
   - selezione degli attributi più importanti
   - filtraggio delle visite non rilevanti (es. visite senza azioni)

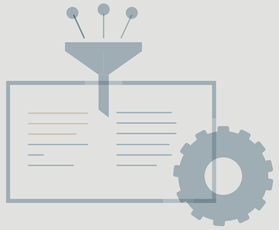
2. Classificazione delle pagine web associate alle azioni di una visita

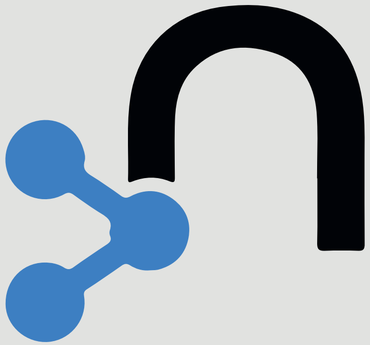
DataFrame delle visite

	# Righe	# Colonne	Memoria [MB]
DataFrame iniziale	30.184	103	~ 23.9
DataFrame finale	18.905	22	~ 21.4

DataFrame delle azioni

	# Righe	# Colonne	Memoria [MB]
DataFrame iniziale	213.116	36	~ 58.5
DataFrame finale	200.440	14	~ 55.4





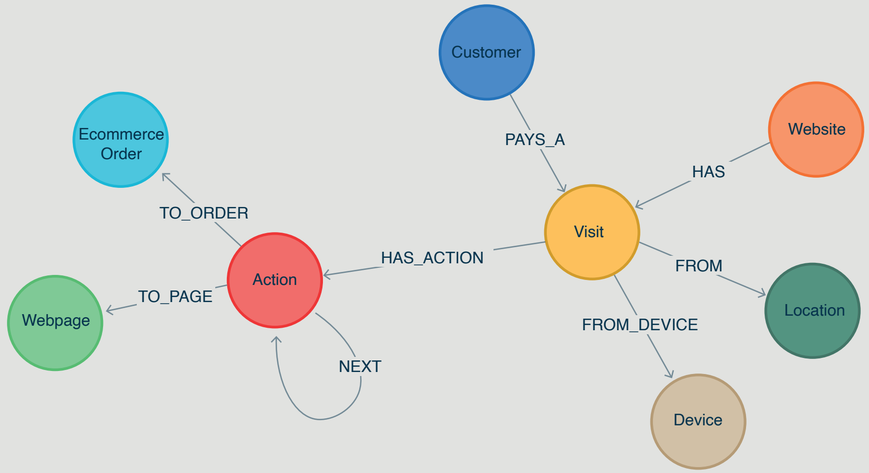
Caricamento dei dati

Salvataggio a blocchi per ottenere prestazioni migliori in quanto si riutilizza una sola sessione per inserire un insieme di visite

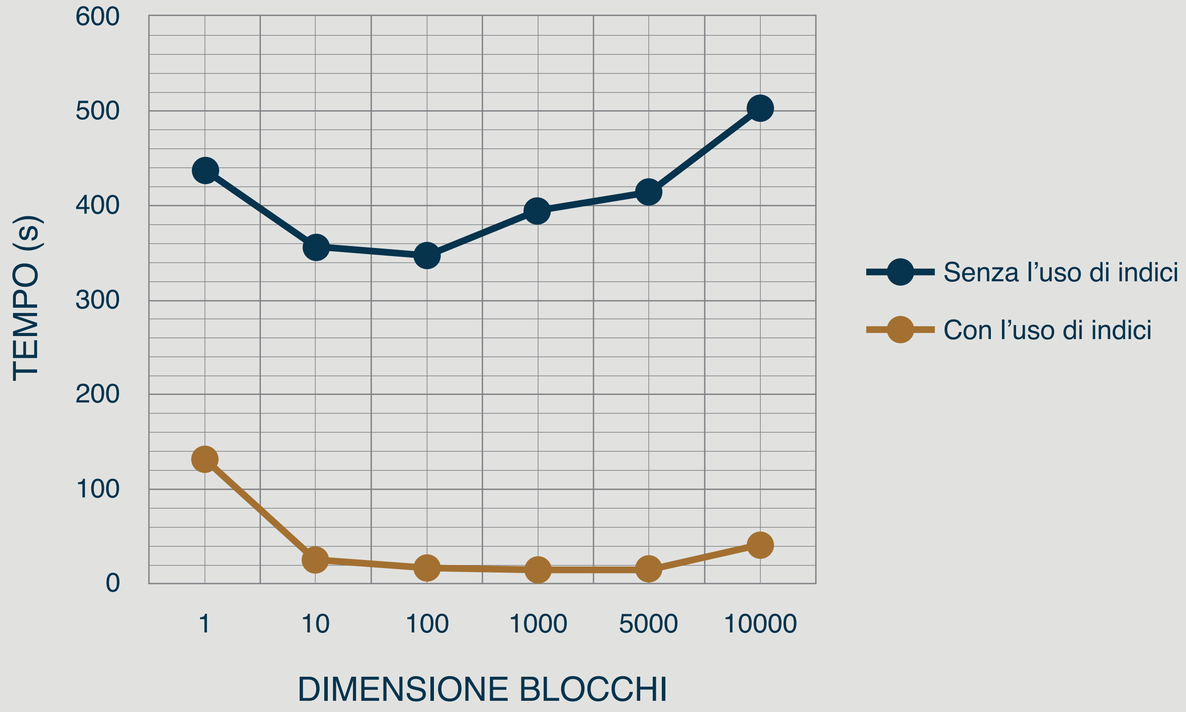
Tempi di inserimento brevi

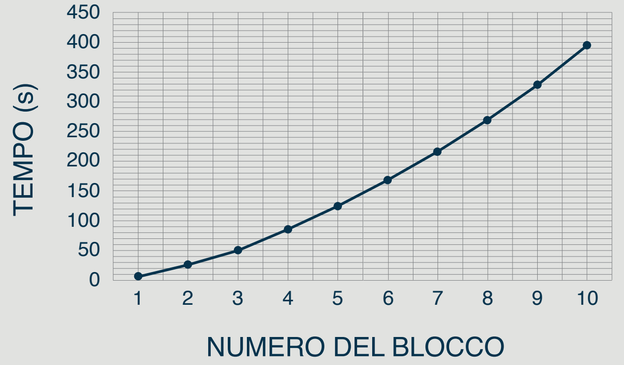
Elevata complessità della query cypher

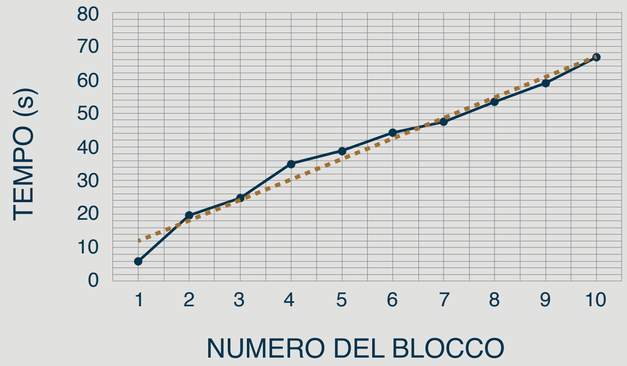
Definizione di indici e vincoli di integrità alle chiavi primarie dei nodi



Valutazione dell’algoritmo ETL







Analisi descrittiva del dataset

Il dataset in analisi contiene i dati di navigazione dei clienti su un sito B2B

- 23 giugno 2021 - 23 agosto 2021
- 152.424 visite totali
- 76.290 visite autenticate
- 3.291 visite con acquisti

	Con acquisti	Senza acquisti	TOTALE
Autenticate	3.291	72.999	76.290
Non autenticate	0	76.134	76.134
TOTALE	3.291	149.133	152.424

Lunghezza delle visite (numero di azioni)

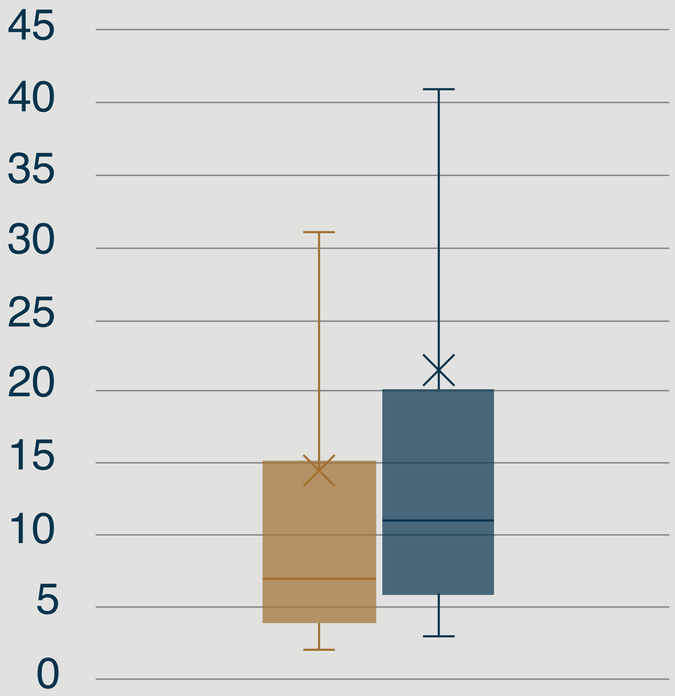
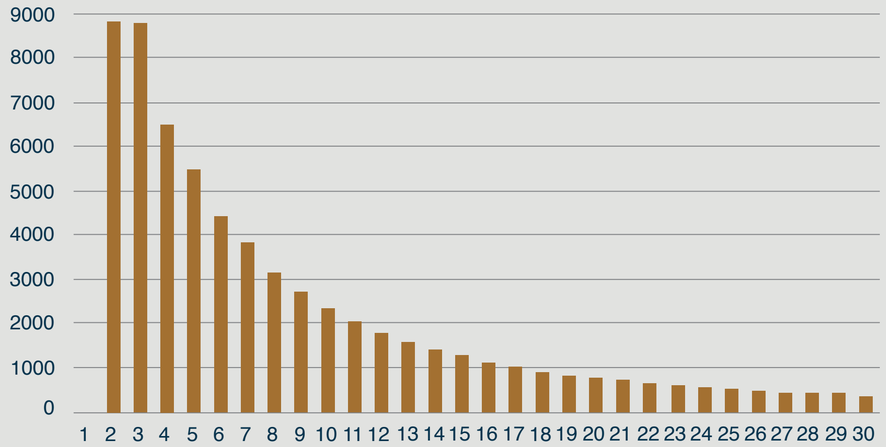


Diagramma a scatola e baffi della distribuzione dei valori di lunghezza delle visite

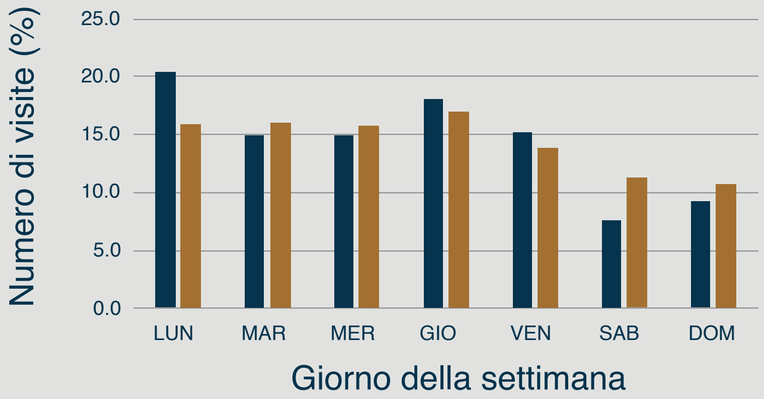


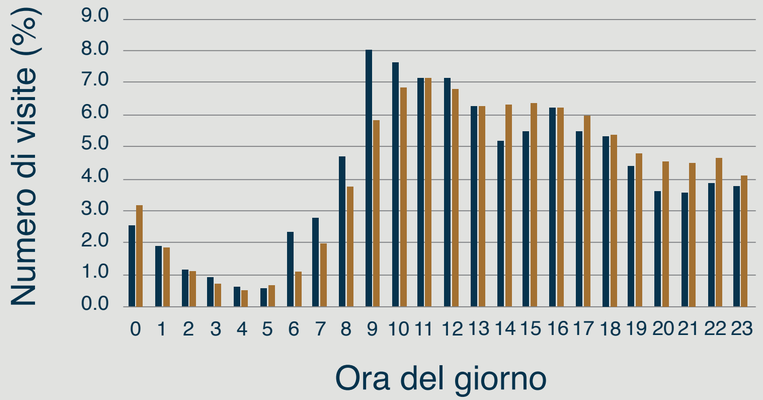
Visite senza acquisti

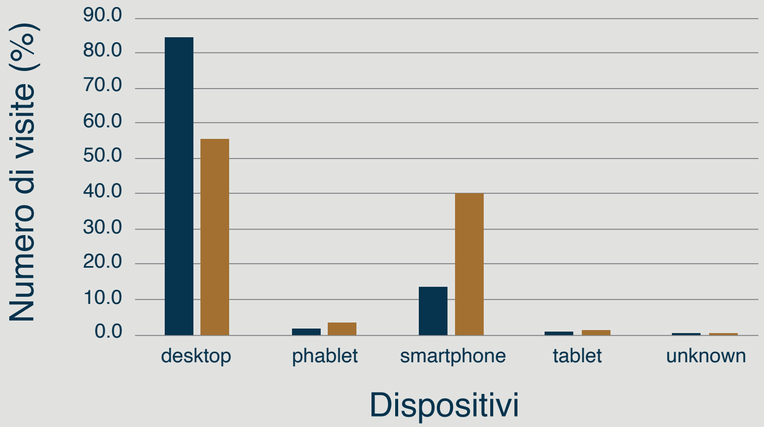
Visite con acquisti

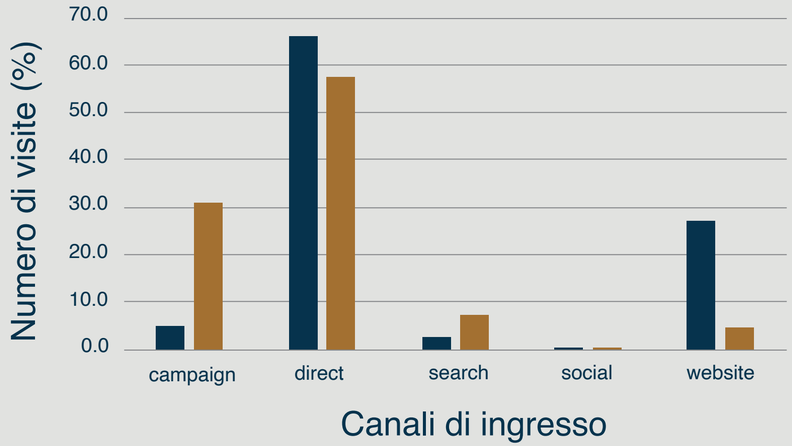
Distribuzioni delle variabili indipendenti (esogena)

Impatto sulla variabile dipendente (endogena)









Visite senza acquisti

Visite con acquisti

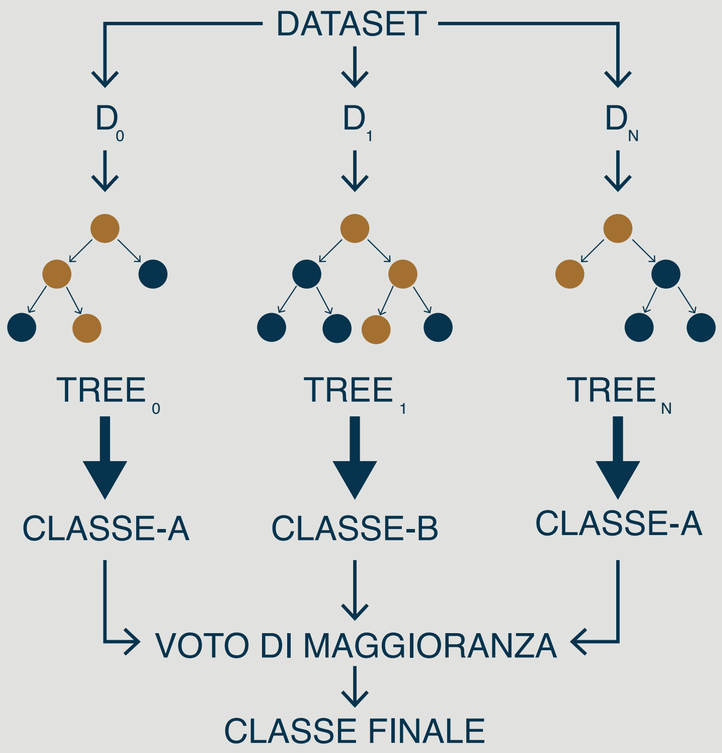
Pre-processing

- Filtraggio
- Feature Engineering
- Feature Selection
- Gestione dei valori mancanti
- Gestione delle variabili categoriche

	Dataset iniziale	Dataset finale
Visite totali (A)	152.424	29.157
Visite con acquisti (B)	3.291	2.010
	↓	↓
rapporto % tra le visite con acquisti e le visite totali (B/A)	2,16%	6,89%

Modello di classificazione Random Forest

Tecnica di apprendimento d'insieme



Accuratezza maggiore degli alberi decisionali

Robustezza al rumore ed eventuali outliers

Efficienza nella generazione del modello

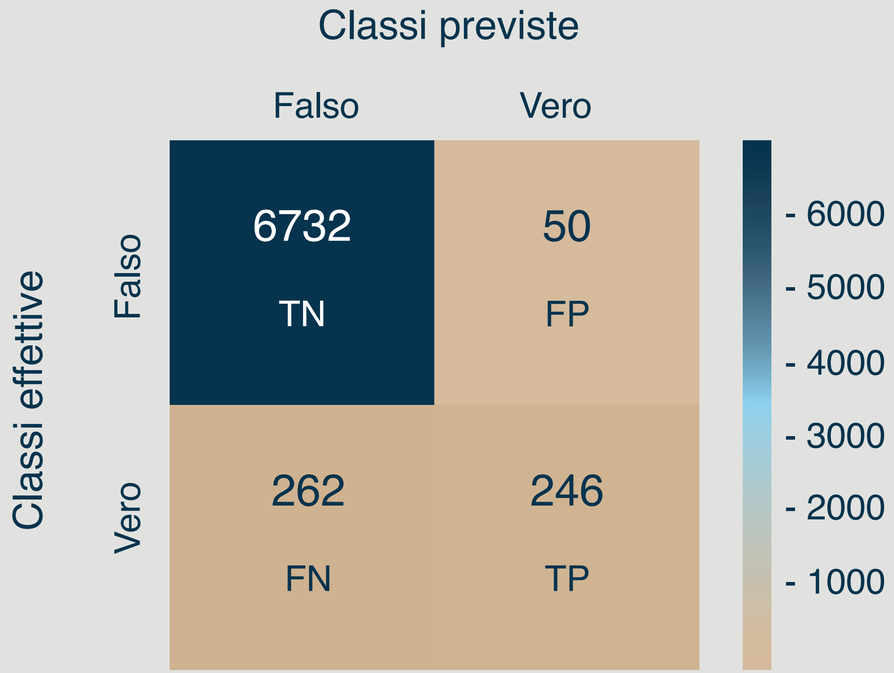
Scalabilità del training set e del numero di attributi

Scarsa interpretabilità

Non incrementabile

Risultati del modello di classificazione

I risultati sono stati ottenuti utilizzando il 75% di dati per il train e 25% di dati per il test



$$\text{Accuratezza} = \frac{TP + TN}{TP + TN + FP + FN} = 95{,}78\%$$

$$F_1 = 2 \cdot \frac{\text{Precisione} \cdot \text{Richiamo}}{\text{Precisione} + \text{Richiamo}} = 61{,}88\%$$

$$\text{Precisione} = \frac{TP}{TP + FP} \qquad \text{Richiamo} = \frac{TP}{TP + FN}$$

Valutazioni del modello di classificazione

Classificazione di una visita dopo un numero predefinito di azioni eseguite

Accuratezza del modello

~50% delle visite dei compratori (classe di minoranza) è classificato correttamente

Necessità di dover attendere un numero predefinito di azioni eseguite per poter classificare una visita

Sviluppi futuri

- Utilizzo e valutazione dell’efficacia del modello predittivo
- Miglioramento dell’algoritmo di classificazione aggiungendo nuove feature per il training
- Sviluppo di altri algoritmi per classificare la visita ad ogni azione del cliente

GRAZIE PER L’ATTENZIONE!

?

DOMANDE