GECKO SPORT

DIMMI CHE SPORTIVO SEI E TI DIRO' COSA INDOSSARE

Elisa Amadori
Pasquale Arpino
Giulia Caci
Rosa Gargiulo
Raffaela Tomasetig

INDICE

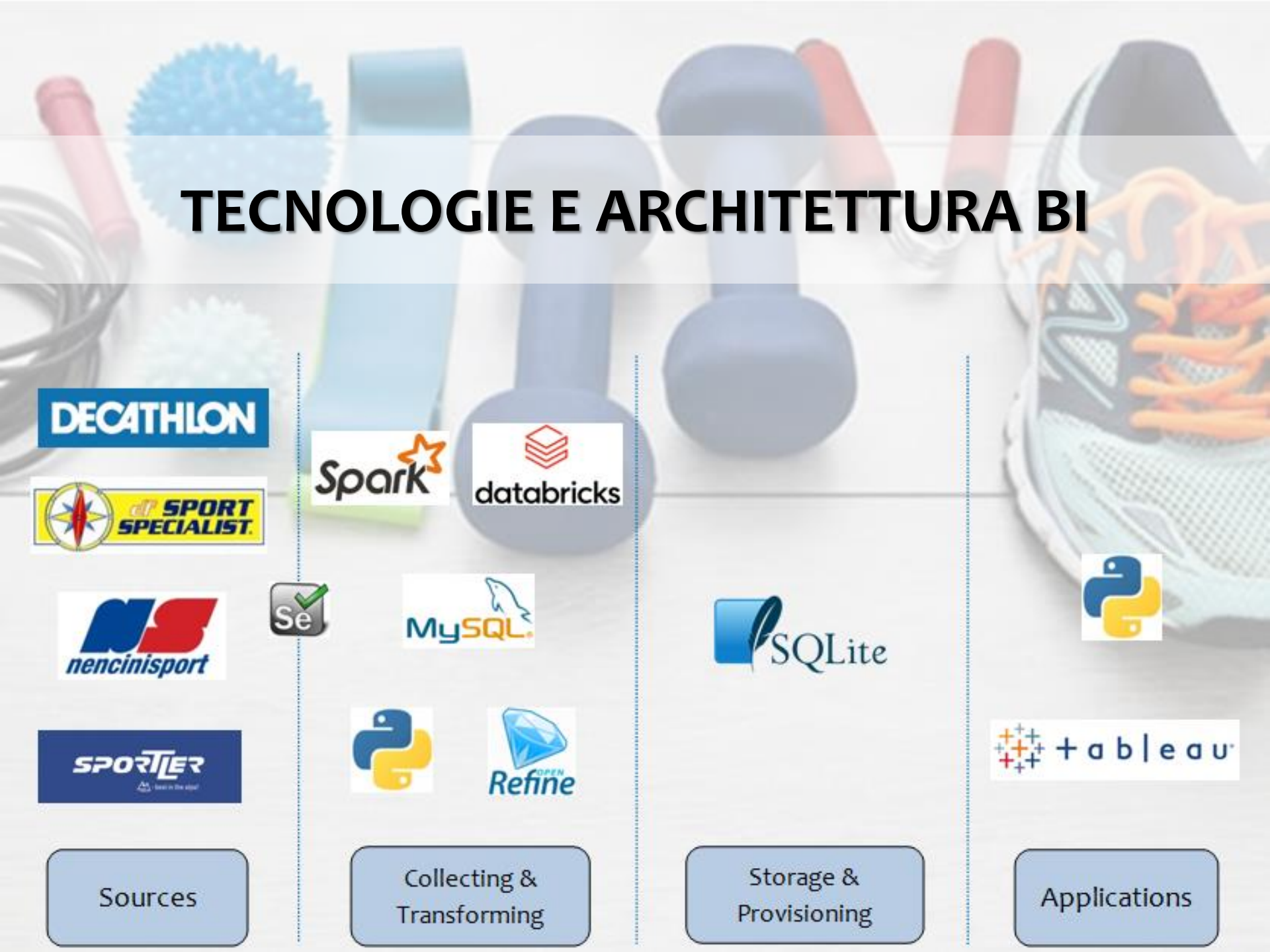
OBIETTIVI

Creazione di una **interfaccia** che raccoglie prodotti di abbigliamento sportivo da più store on-line e propone diverse combinazioni di outfit personalizzate per utente

STAKEHOLDERS

- **Store sportivi:** facilita la vendita on-line raggiungendo futuri potenziali clienti
- **Utenti sportivi:** facilita e velocizza l'acquisto, offrendo una vasta gamma di articoli scelti tra più negozi contemporaneamente

TECNOLOGIE E ARCHITETTURA BI



SCRAPING

COSA?
Running per uomo e donna

QUALI DATI?

- Nome prodotto
- Codice prodotto
- Url
- Descrizione prodotto
- Brand
- Tipo prodotto
- Colore
- Prezzo
- Genere
- Immagine

DATA PROFILING – Caratteristiche dataset

NENCINI

- 428 articoli
- 9 tipologie diverse di abbigliamento incorporate nel nome prodotto
- Brand incorporato nel nome prodotto

DECATHLON

- 267 articoli
- 26 tipologie diverse di abbigliamento incorporate nel nome prodotto
- Genere incorporato nel nome prodotto

DATA PROFILING – Caratteristiche dataset

SPORTLER

- 1051 articoli
- 17 tipologie diverse di abbigliamento incorporate nel nome prodotto
- Stesso articolo disponibile in più colori salvato in un'unica stringa di testo

SPORT SPECIALIST

- 863 articoli
- 17 tipologie diverse di abbigliamento incorporate nel nome prodotto
- Brand incorporato nel nome prodotto
- Genere incorporato nel nome prodotto

ETL

- "Genere", "Brand" e "Tipo" di prodotto ricavate dal nome dell'articolo e splittate su nuove colonne
- Individuate le principali tipologie di abbigliamento e clusterizzate in tre principali: **maglia**, **pantaloni** e **scarpe**

- Prezzi trasformati in numero con il . come separatore
- Scelta dei criteri di uniformità (traduzione in italiano, lower case letter)
- Aggiunta di "Sito" che indica lo store da cui provengono i dati
- Scelta del colore predominante per articoli con colore multiplo (es. [Nero/Giallo] → [Nero])

- Pulizia della descrizione prodotto da elenchi puntati, andare a capo (\n), tab (\t)

ETL – Data augmentation

I prodotti che presentavano più varianti di colori sono stati divisi in più articoli

Column	id	url	Name	Tipo_dettaglio	colour	Genere	tipo
761	10837478	https://w	Cloudflow	scarpe	Black-Red-Yellow	uomo	scarpe

Column	id	url	Name	Tipo_dettagli	colour	Genere	tipo
761	10837478	https://w	Cloudflow	scarpe	Black	uomo	scarpe
761	10837478	https://w	Cloudflow	scarpe	Red	uomo	scarpe
761	10837478	https://w	Cloudflow	scarpe	Yellow	uomo	scarpe

Esempio di Data augmentation

SITO	N. articoli prima split colore	N. articoli dopo split colore	Delta record
Sportler	1051	1448	397
Decathlon	267	466	199
Nencini	428	428	0
Sport specialist	863	863	0
TOTALE	**2609**	**3205**	**596**

Al termine del processo di Data augmentation, il DB guadagna quasi 600 record da Sportler e Decathlon

DB unificato

I DB dei 4 siti sono stati uniti su Databricks utilizzando SQL, ottenendo così un'unica grande tabella completamente denormalizzata

databricks

Sito	Id_prodotto	Url	Nome	Tipo	Tipo_specifico	Brand	Descrizione	Genere	Colore	Prezzo	Immagine
nencini	283181	https://www.nen	Asics Nimbus St	vario	calzini	asics	Affronta maratone, mezze maratone e 10 km indossando una calza ideat Ottieni una calza con ammortizzazione ispirata alla nostra scarpa di live Mantieni una buona igiene del piede con una calza che crea il clima perf Corri all'insegna del comfort con il tallone e le punte dei piedi protette cor Colore come da immagine RIDUCI	donna	rosa	9.9	img_283181.jpg
nencini	435879	https://www.nen	New Balance 86	scarpe	scarpe	new balance	Realizzate per il runner che cerca stabilità, le nostre scarpe da corsa Fre DETTAGLI PRODOTTO L'ammortizzazione dell'intersuola in FRESH FOAM è progettata per assi Tomaia in mesh tecnica per una calzata che sembra realizzata appositan Materiale sintetico/mesh leggero Il pannello mediale aiuta a controllare la pronazione Suola esterna in gomma soffiata assicura durabilità Il design Ultra Heel avvolge il tallone assicurando una calzata aderente e Ammortiizzazione in base ai dati per comfort su lunghe distanze Stabilità comprovata RIDUCI	uomo	grigio	112.0	img_435879.jpg
nencini	382322	https://www.nen	Asics Pant	pantaloni	pantaloni	asics	Affronta il tuo prossimo allenamento con comfort e stile, grazie a questi p I pantaloni hanno una linea elegante con inserti modellati sul ginocchio e Questi pantaloni sono anche una scelta eccellente per la corsa all'aperto Tessuto ad asciugatura rapida anti-umidità Struttura in tessuto misto Inserti modellati per le ginocchia Tasca laminata per custodire in sicurezza i dispositivi Apertura a zip sulla parte posteriore delle gambe Cordino regolabile Logo ASICS riflettente Strisce riflettenti intorno alle ginocchia per una migliore visibilità RIDUCI	uomo	nero	42.0	img_382322.jpg
nencini	279610	https://www.nen	Bunf Jacket Tecl	maglia	giacca	bunf	Giacca unisex antivento e idrorepellente. Cappuccio elastico fisso. Zip intera. Due tasche con zip. Slim fit. 100% poliestere. RIDUCI	uomo	verde	22.5	img_279610.jpg

CLUSTER UTILIZZATI

SOCIAL

Utilizzando il numero di interazioni di # e @ per brand degli ultimi 6 mesi su Twitter, sono stati classificati i brand come “Social SI” e “Social NO”

ECOLOGICO

Utilizzando parole chiave come “recycled” “ecological” “natex” sono stati clusterizzati i prodotti come “Ecologico SI” ed “Ecologico NO”

ECCENTRICO

Partendo da assunzioni personali e dai colori degli articoli, sono stati classificati i prodotti “Eccentrico SI” ed “Eccentrico NO”

TECNICO

Utilizzando un modello di ML TF-IDF a partire da un glossario condiviso, sono stati clusterizzati i prodotti come “Tecnico SI” e “Tecnico NO”

CLUSTER SOCIAL

- Utilizzata libreria GetOldTweets
- Query considerate: "#Brand" + "@Brand"
- Periodo : dal 01/01/2020 al 31/07/2020
- Eliminati dai risultati i brand dai significati maggiormente ambigui: es. Castelli, Scott
- Assunta come soglia di "Social SI" un numero di interazioni nel periodo >100

Brand	N. interazioni	Social
Nike	2989	SI
Adidas	1687	SI
Puma	1045	SI
Dainese	549	SI
Patagonia	477	SI
Under Armour	276	SI
Columbia	253	SI
New Balance	222	SI
Reebok	155	SI
Asics	140	SI
Brooks	125	SI
Mizuno	112	SI
The North Face	112	SI
Karpos	85	NO
Diadora	82	NO

Soglia 100 interactions

CLUSTER ECOLOGICO

- Le descrizioni dei prodotti presenti sugli store sono state utilizzate per definire il cluster "Ecologico SI"/"Ecologico NO" dei capi d'abbigliamento

- I **Key terms** che indicano l'appartenza o meno al cluster "Ecologico SI"/"Ecologico NO" sono stati desunti dalla descrizione dei materiali dei prodotto o all'interno delle descrizioni generali

I principali termini utilizzati e loro derivati sono stati:

- **ECOLOGICO SI:**

"100% Riciclabile","Recycled", "Ecological", "Ambientale", "Sustainable", "Natex", "Etichetta ambientale", "100% Vegan"

- **ECOLOGICO NO:**

la non presenza dei suddetti termini

CLUSTER ECCENTRICO

Per definire la caratteristica “Eccentrico SI”/”Eccentrico NO” di un capo d’abbigliamento è stato eseguito dal team di progetto un sondaggio interno sulla personale percezione, in termini di “sobrio” ed “eccentrico”, di ciascun colore. La maggioranza di risultato ha permesso di classificare i colori come **eccentrico** e **non eccentrico**

CLUSTER TECNICO - Glossario

Le descrizioni dei prodotti presenti sugli store sono state utilizzate per definire il cluster “Tecnico SI”/ “Tecnico NO” del capo d’abbigliamento

- All’interno delle descrizioni sono presenti termini chiave che riescono a discriminare l’appartenza o meno al cluster in oggetto
- Per classificare i capi sportivi è stato costruito un algoritmo di ML supervisionato utilizzando un glossario di termini definito sulla base di fonti online e delle conoscenze dell’esperto di dominio

Le principali parole chiave e loro derivate sono state:

- **Tecnico SI**: “Performance”, “Gara”, “Prestazioni”, “Runner esperto”, “Tessuto tecnico”, “Frequenza d’uso: regolare”, “Ammortizzazione CloudTec”, “Tessuto T-Hexagon”…
- **Tecnico NO:** “Principiante”, “Primi passi”, “Camminata”, “Jogging occasionale”, “Brevi distanze”…

CLUSTER TECNICO – Machine Learning

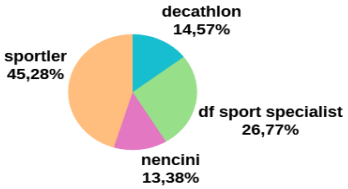
- Utilizzo di TF-IDF e Naive Bayes
- 1367 articoli di partenza classificati a mano di cui **881 TECNICO SI** e **486 TECNICO NO**
- **80%** dei dati utilizzati per il **training**, **20%** dati utilizzati per il **test**
- Eliminazione di stop words e parole con lunghezza inferiore a 3
- Sono stati considerati gli n-grams fino a 3 parole

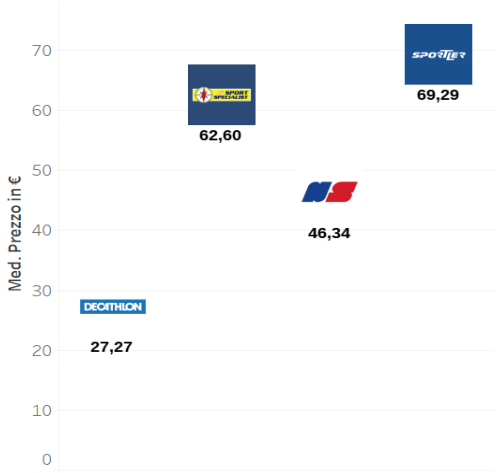
PARAMETRI DI FUNZIONAMENTO

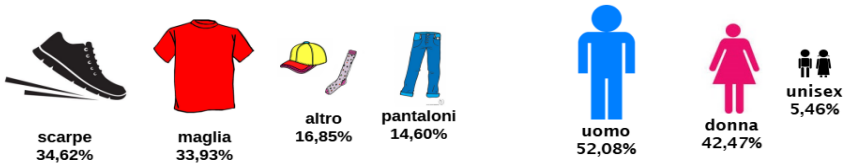
Precisione = 0.8687

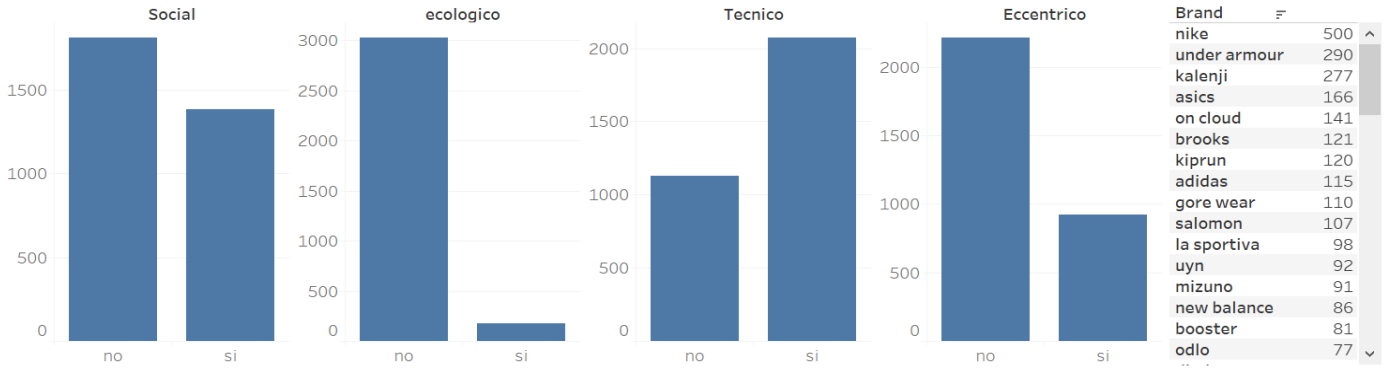
Recall = 0.8686

F = 0.8668









Brand	
nike	500
under armour	290
kalenji	277
asics	166
on cloud	141
brooks	121
kiprun	120
adidas	115
gore wear	110
salomon	107
la sportiva	98
uyn	92
mizuno	91
new balance	86
booster	81
odlo	77

VISUALIZZAZIONE Dashboard

LINK TABLEAU DASHBOARD:

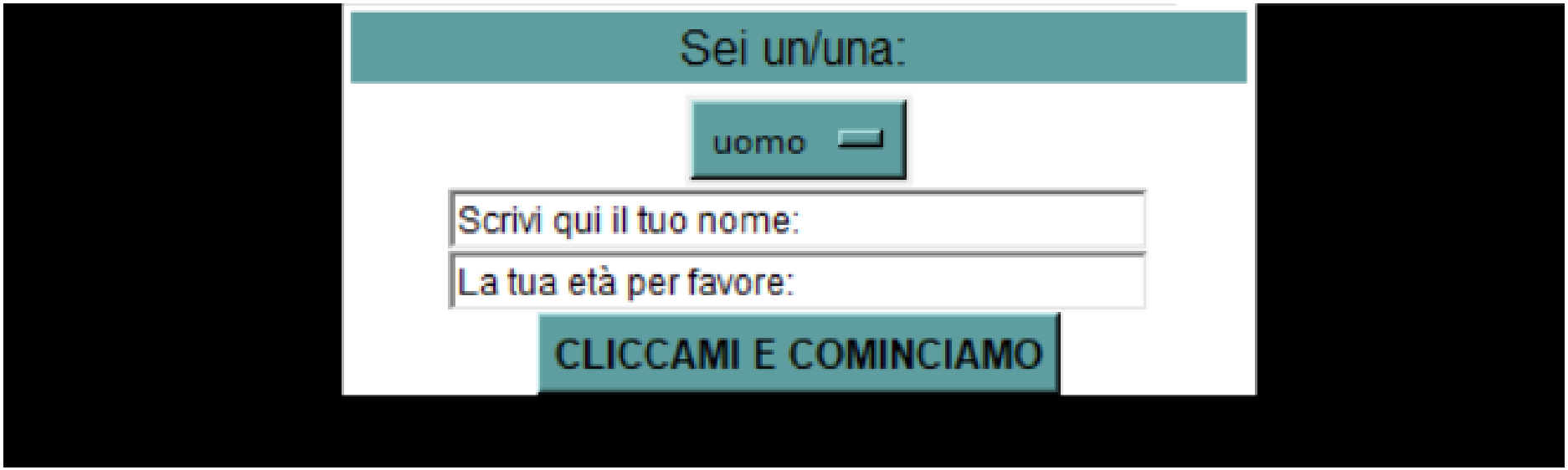
https://public.tableau.com/profile/giulia1439#!/vizhome/sportviewok/Dashboard5

GUI with Python 1/4

Un'interfaccia utente che in modo interattivo pone delle domande all'utente in modo da assegnarli il medesimo cluster degli items presenti nella factual table

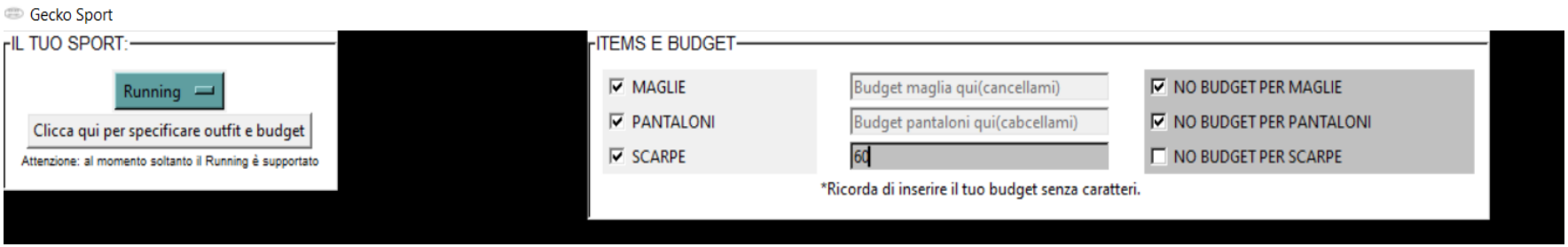
Caratteristiche

- Possibilità di impostare un budget per ogni categoria di item
- Risposte alle domande utilizzate come filtro della Fact Table
- Limit 10 dei risultati, order by Prezzo Desc ed estrazione random degli item in modo da non presentare prodotti dello stesso store in modo consecutivo



GUI with Python 2/4

Per cominciare viene chiesto all'utente per quale disciplina sportiva sta cercando abbigliamento (al momento solo Running disponibile)

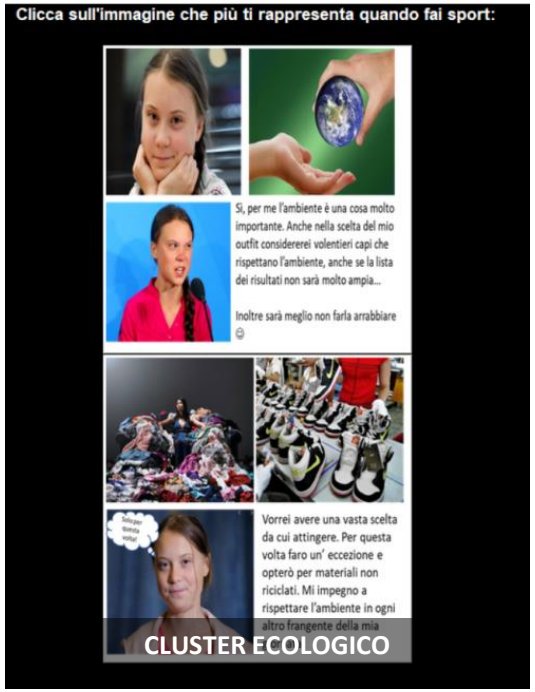


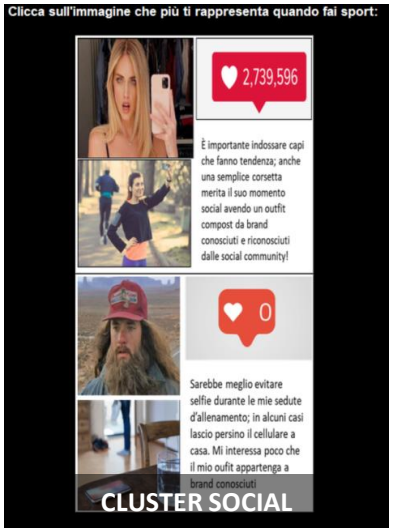
Dopo aver effettuato la scelta gli viene chiesto quali categorie di indumenti sta cercando. Di default è presente un flag sull'intero Outfit (maglia+pantaloni+scarpe)

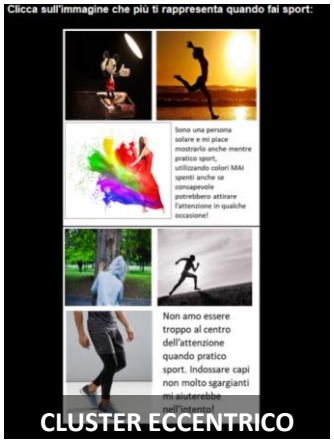
Infine è possibile impostare o meno la presenza di un budget max per ogni categoria

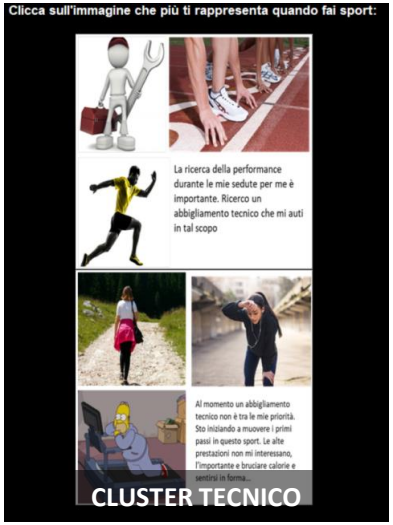
GUI with Python 3/4

È arrivato il momento di clusterizzare l'utente ponendogli delle domande le cui risposte verranno immagazzinate ed utilizzate successivamente come filtro nel DB SQL integrato dei vari siti









GUI with Python 4/4

In base alle risposte fornite in precedenza, vengono presentati gli indumenti con le caratteristiche migliori per l'utente.

È possibile creare visivamente il proprio outfit avendo i box delle 3 categorie (maglia,pantaloni e scarpe) presentati in maniera logica e sovrapposta.

Inoltre per ogni categoria, cliccando su "**Back**" o "**Next**", è consentito muoversi dinamicamente tra le varie scelte avendo il costo totale dell'oufit che si modifica di conseguenza.

Cliccando su "**Ho Finito**" vengono visualizzati i link che una volta selezionati rimandano alle pagine dei siti dove acquistare i capi.

Obiettivi raggiunti e criticità

OBIETTIVI RAGGIUNTI

- Obiettivo minimo di proporre all'utente un'interfaccia con almeno uno sport selezionabile e la scelta di un outfit che si avvicina alle sue caratteristiche
- Database unificato con un numero sufficiente di record clusterizzati provenienti da vari siti tale da fornire all'utente la percezione di un servizio a valore aggiunto che va oltre il semplice acquisto online dai singoli store

CRITICITÀ

- Data cleaning ed integrazione di database con strutture, formati ed informazioni diverse tra loro
- Definizione e ricerca di un glossario di possibili parole chiave per classificare i prodotti in "Tecnico SI"/"Tecnicno NO" con un algoritmo di ML
- Gestione di prodotti disponibili in colori differenti e definizione di quali colori risultassero nel cluster "Eccentrico SI"/"Eccentrico NO"
- Numero di prodotti ecosostenibili basso che limita il numero di combinazioni possibili con gli altri Cluster

Cosa migliorare? Sviluppi futuri

- Definire un utilizzo secondario dei dati a disposizione per creare delle sottocategorie per tipologie prodotto: ad es. Categoria “Maglia”, sottocategorie “Canotta”, “Giacca”, “Top” ecc.
- Integrare nell’interfaccia utente ulteriori discipline sportive oltre al Running
- In base alla disciplina sportiva selezionata, definire differenti categorie e sottocategorie di prodotti associabili a quello sport; ad es. Nuoto con categorie “Cuffietta”, “Costume”, “Ciabatte”
- Utilizzare un numero maggiore di Store da cui fare Scraping
- Definire una Pipeline per automatizzare il Data Ingestion, ETL, ML e Application

FINE

GRAZIE PER L'ATTENZIONE

Elisa Amadori
Pasquale Arpino
Giulia Caci
Rosa Gargiulo
Raffaela Tomasetig

GITHUB

Qui il link ai file e la presentazione del progetto:

https://github.com/Pasq9219/Gecko_sport_academic_project