

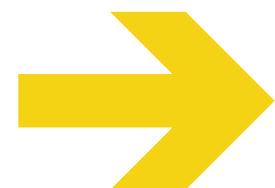
DATA SCIENCE

Analisi della soddisfazione dei
passeggeri di una compagnia aerea

PROGETTO FINALE PAOLO ACONE



INDICE



OBIETTIVO

DATASET

PULIZIA E PREPROCESSING

ANALISI ESPLORATIVA

FEATURE SELECTION

RIDUZIONE FEATURE SET

MODELLAZIONE

OTTIMIZZAZIONE IPERPARAMETRI

PERFORMANCE MODELLI E TEST FINALE

CONSIDERAZIONI FINALI

PROPOSTE PER LA COMPAGNIA

CONTRIBUTO ALLA SOSTENIBILITÀ

CONCLUSIONI

OBIETTIVO DEL PROGETTO

L'**obiettivo principale** è costruire un modello predittivo capace di classificare correttamente il livello di soddisfazione dei passeggeri.

Il **secondo obiettivo** è analizzare le caratteristiche che più influenzano la soddisfazione per fornire insight pratici alla compagnia.

L'**intento finale** è suggerire miglioramenti concreti nei servizi, soprattutto nei voli a corto raggio, con un occhio alla sostenibilità e all'efficienza operativa.



DATASET

Il dataset contiene dati anonimi su passeggeri, includendo informazioni personali (es. genere, età), dettagli del volo (classe, distanza) e valutazioni sui servizi (es. Wi-Fi, intrattenimento, comfort, puntualità).

La variabile target è la soddisfazione (satisfied vs neutral or dissatisfied).

Presenti sia variabili numeriche che categoriche.

PULIZIA & PREPROCESSING

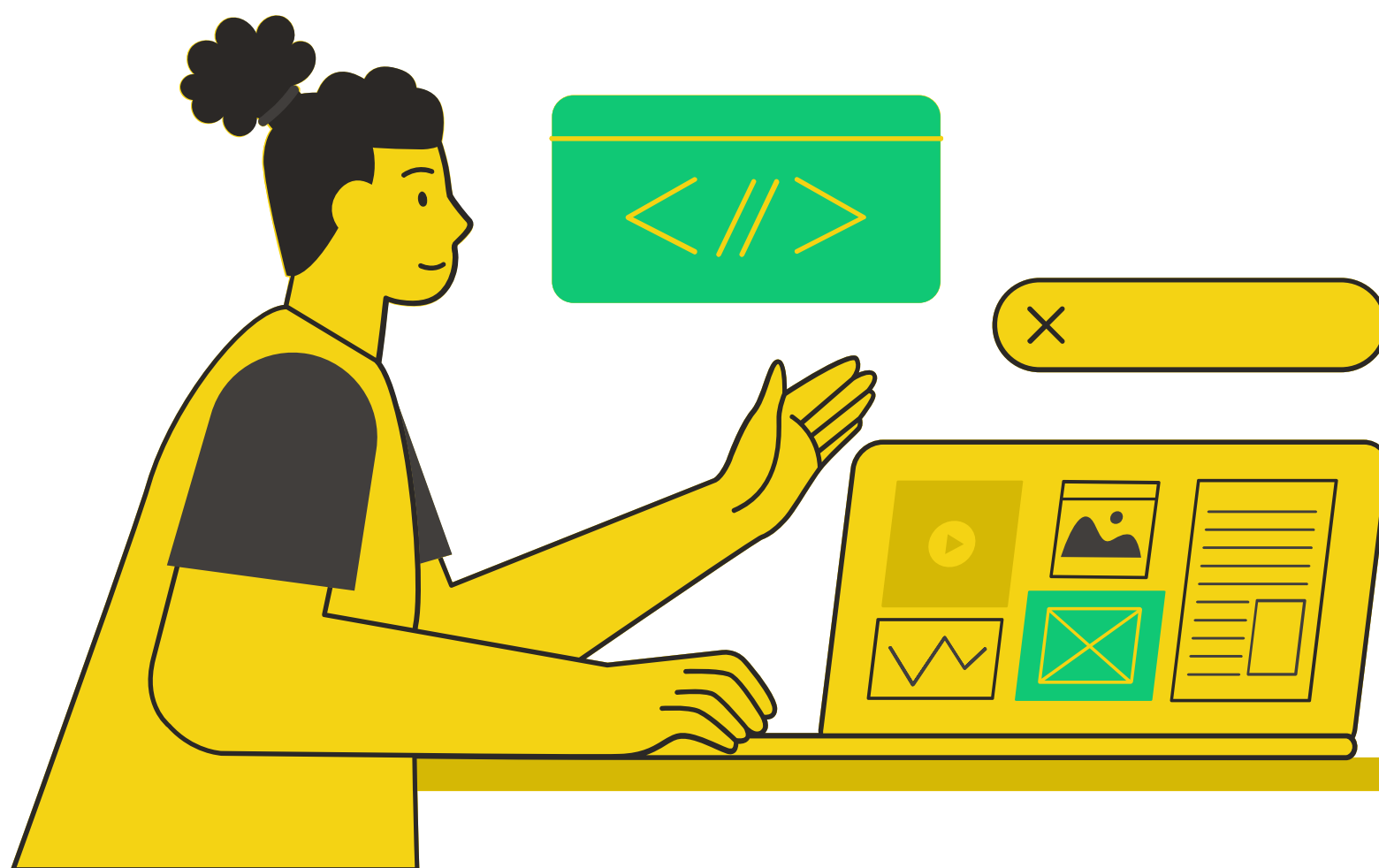
Sono stati rimossi i record con valori anomali.

Le variabili categoriche sono state codificate:

Label Encoding per variabili binarie (es. Gender)

One-Hot Encoding per quelle con più di due categorie (es. Class)

Le colonne del dataset di test sono state riallineate a quelle del train per garantire compatibilità durante la predizione.

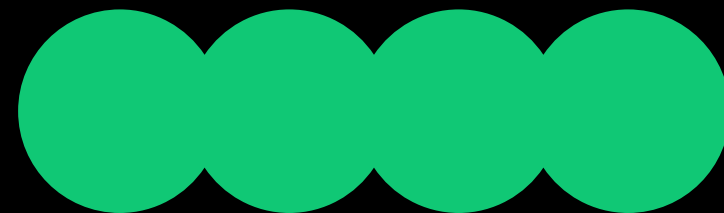


ANALISI ESPLORATIVA (EDA)

L'EDA ha evidenziato come le distribuzioni del train set siano rappresentative del dataset intero.

I passeggeri più insoddisfatti viaggiano in Economy, non sono fidelizzati e viaggiano per motivi personali.

Riscontrata una forte correlazione tra ritardo alla partenza e all'arrivo, sintomo di possibile ridondanza informativa.



FEATURE SELECTION

Sono stati applicati diversi metodi di selezione delle feature:

Correlazione lineare con il target

Chi-quadrato per variabili categoriche

T-test per confronti di media

Mutual Information per la dipendenza non lineare

Le variabili risultate più rilevanti in più test sono:

'Type of Travel',

'Flight Distance',

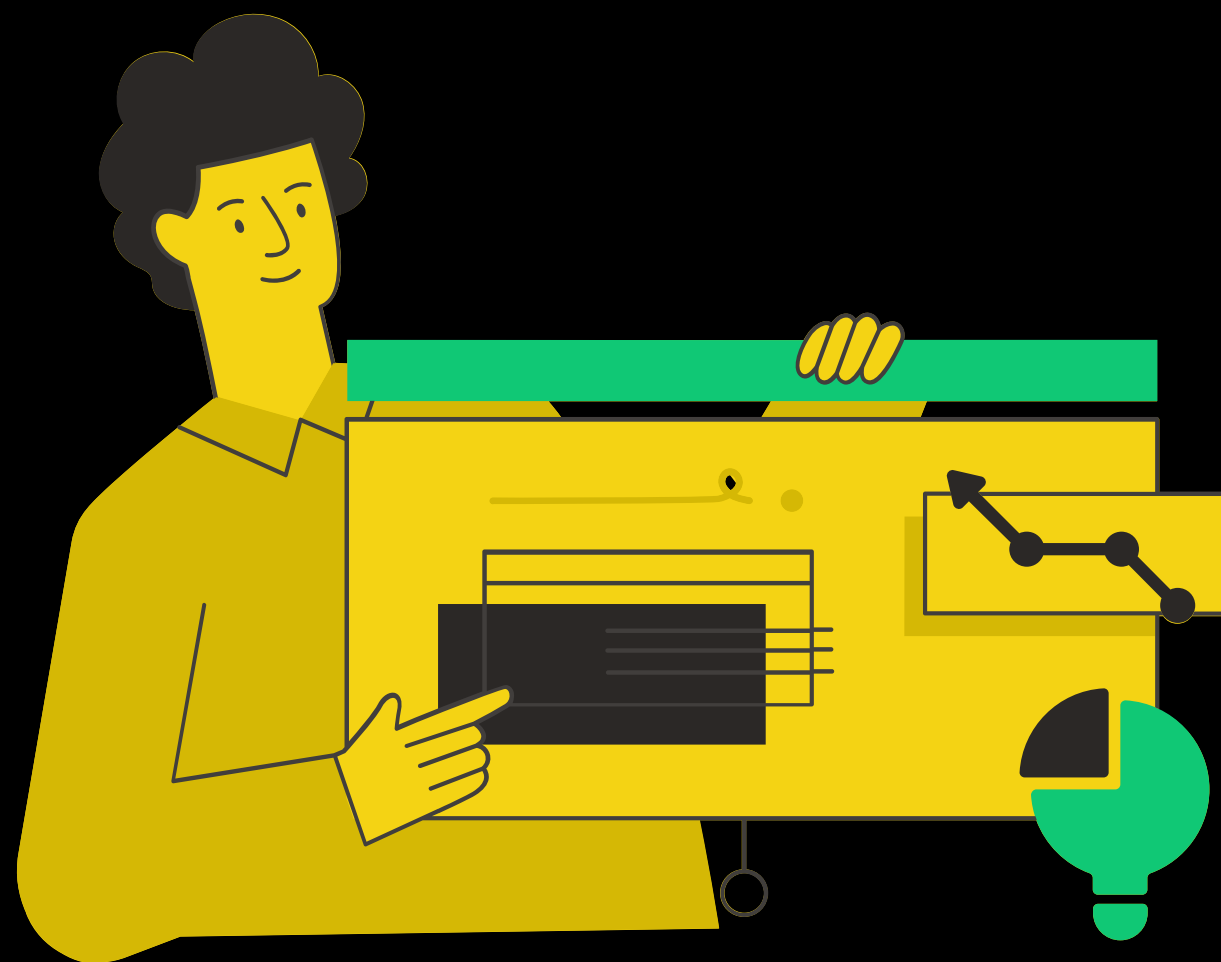
'Inflight wifi service',

'Online boarding',

'Leg room service',

'On-board service'

RIDUZIONE FEATURE SET



A seguito dell'identificazione delle feature più influenti, è stato creato un dataset ridotto contenente unicamente queste variabili.

L'obiettivo di addestrare gli stessi modelli anche su questo dataset compresso è **duplice**:

Riduzione del costo computazionale → Un numero inferiore di feature implica tempi di esecuzione potenzialmente inferiori durante l'addestramento e l'inferenza:

Miglioramento dell'interpretabilità e della semplicità del modello → Un modello basato su un numero ristretto di feature chiave è intrinsecamente più facile da comprendere e interpretare, facilitando l'identificazione dei fattori trainanti della soddisfazione (o dell'impatto ambientale) e semplificando la sua implementazione.

MODELLAZIONE

È stato definito un baseline model con Logistic Regression per avere un punto di riferimento semplice ma valido.

Successivamente si sono testati modelli più avanzati come Random Forest e AdaBoost.

Per confrontare un approccio diverso è stato scelto K-Nearest Neighbors (KNN).

La metrica di valutazione selezionata per questo progetto è l'accuratezza in virtù della distribuzione quasi perfettamente bilanciata tra le classi della variabile target. Inoltre, data l'assenza di una priorità specifica nella minimizzazione di falsi positivi o falsi negativi, l'accuratezza rappresenta una metrica sintetica e appropriata per valutare le prestazioni generali del modello.

È stata adottata una strategia di Stratified K-Fold Cross-Validation per garantire una valutazione robusta e rappresentativa.

È stato condotto uno spot checking su tutti i modelli, sia con dataset completo che con feature selezionate.

Random Forest e K-Nearest neighbors hanno mostrato le migliori performance preliminari.

OTTIMIZZAZIONE IPERPARAMETRI



Per i modelli promettenti (Random Forest e KNN) è stata avviata una fase di ottimizzazione con Random Search.

È stato definito uno spazio di ricerca iniziale e successivamente uno spazio ristretto attorno ai valori ottimali individuati.

Questa doppia iterazione ha permesso una messa a punto precisa e un incremento delle performance predittive.

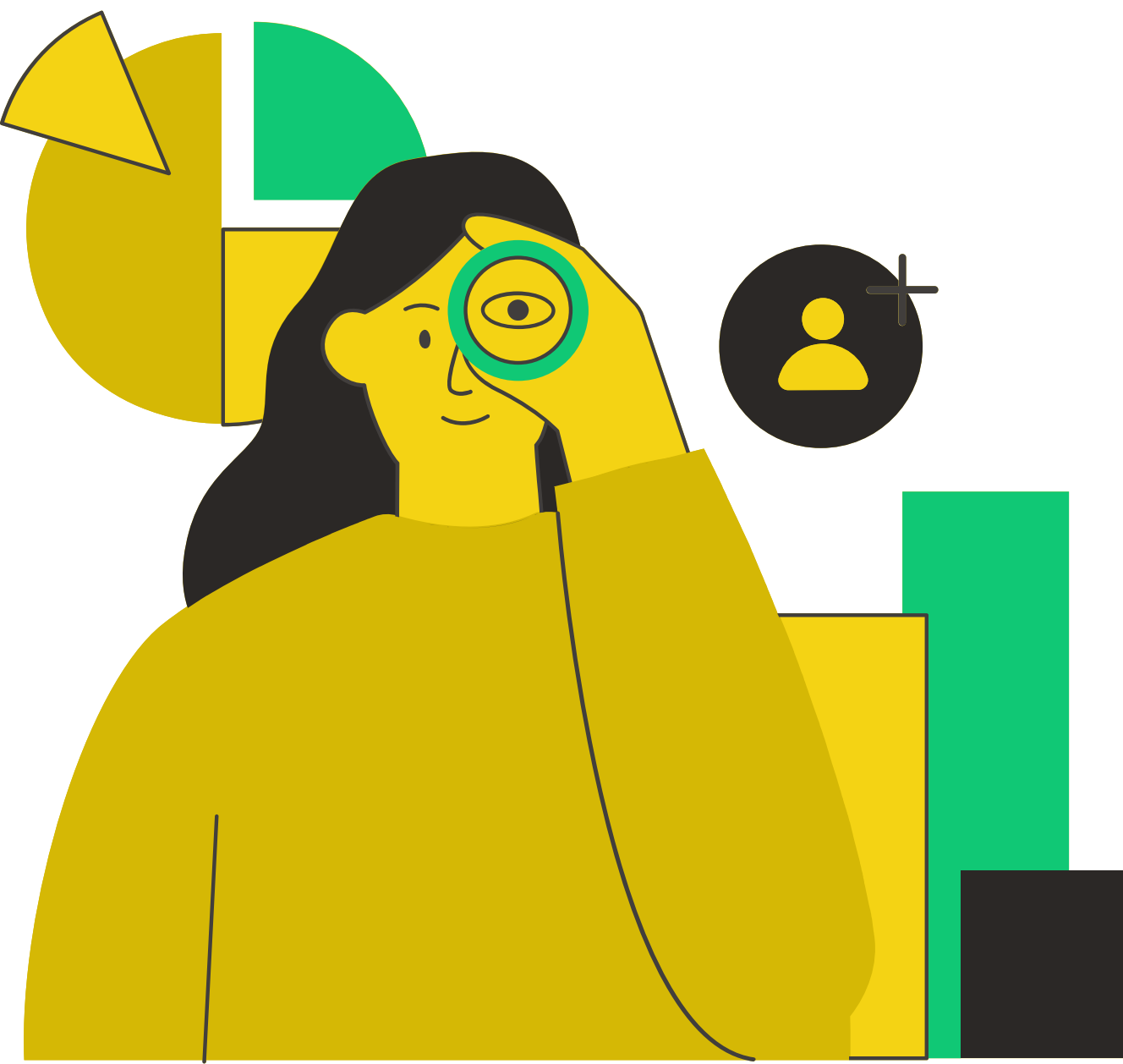
PERFORMANCE MODELLI E TEST FINALE

Il modello vincitore è Random Forest su dataset completo, con un'accuratezza del 96% sul test set.

La riduzione delle feature ha comportato solo una lieve perdita di accuratezza, migliorando l'interpretabilità.

Questi risultati confermano la capacità del modello di generalizzare bene anche su nuovi dati.

CONSIDERAZIONI FINALI



Clienti più soddisfatti sono quelli in classe business, su voli lunghi e fidelizzati.

I servizi maggiormente apprezzati sono: comfort del posto, intrattenimento, spazio per le gambe e imbarco online.

I voli brevi mostrano livelli inferiori di soddisfazione: mancano molti dei servizi presenti nei voli lunghi.

Mettendo insieme queste informazioni si deduce come la quantità e la qualità dei servizi offerti influisce più di qualsiasi altra cosa sulla soddisfazione dei clienti

PROPOSTE

Migliorare i servizi sui voli a corto raggio: introdurre o potenziare Wi-Fi, intrattenimento e boarding smart.

Offrire esperienze più curate per i clienti occasionali (non fidelizzati).


Ridurre i ritardi agendo sull'organizzazione logistica, con un impatto positivo sia sulla soddisfazione sia sulla sostenibilità ambientale.

Investire in comfort anche nei voli brevi: sedili più comodi e maggiore spazio per le gambe.

Potenziare i servizi delle classi Economy ed Eco Plus, dove c'è margine di miglioramento.



CONTRIBUTO ALLA SOSTENIBILITÀ

A large, vibrant green abstract shape, resembling a stylized 'X' or a series of overlapping curves, occupies the top right portion of the page.

L'uso di servizi digitali (check-in online, intrattenimento digitale), oltre che a velocizzare il processo e a offrire un servizio che i clienti apprezzano, riduce l'uso di carta e materiali fisici.

I modelli predittivi aiutano ad allocare risorse dove producono maggiore soddisfazione, evitando sprechi.

Comunicazione trasparente sugli sforzi sostenibili aumenta la soddisfazione dei clienti attenti all'ambiente.

CONCLUSIONE

Il Machine Learning si è dimostrato un valido strumento per comprendere le esigenze dei passeggeri e guidare decisioni basate sui dati.

Concentrarsi sui segmenti meno soddisfatti rappresenta la maggiore opportunità di miglioramento.

L'obiettivo di migliorare la customer satisfaction può andare di pari passo con quello ambientale.

L'innovazione nei servizi, se progettata in ottica efficiente e mirata, valorizza sia il cliente che la sostenibilità.

NOTEBOOK COMPLETO

GRAZIE

