BIG DATA ANALYTICS

TECNICHE DI ANALISI

Prof.ssa Romina Eramo

Università degli Studi di Teramo

Dipartimento di Scienze della Comunicazione

reramo@unite.it

Quattro modi di analizzare i dati

Ogni tipologia di analytics risponde a una domanda diversa.

01

DESCRITTIVA

Che cosa è successo?

Report, KPI, dashboard

02

DIAGNOSTICA

Perché è successo?

Confronti, ipotesi, correlazioni

03

PREDITTIVA

Che cosa potrebbe succedere?

Modelli, probabilità, previsioni

04

PRESCRITTIVA

Che cosa conviene fare?

Decisioni, vincoli, ottimizzazione

Dai dati alla conoscenza, dalla conoscenza a decisioni motivate

Il caso guida: promuovere un evento culturale

Una campagna su più canali: da visualizzazioni e clic fino alle iscrizioni.

OBIETTIVO DELLA CAMPAGNA

600 iscrizioni

in tre settimane, con un budget definito e canali digitali differenti.

CANALI
Instagram · Ricerca · Newsletter

DATI
Impression · Clic · Iscrizioni

DOMANDA
Quale combinazione funziona meglio?

Dati sintetici per esercitarsi

Analisi descrittiva: leggere i risultati

Una dashboard permette di misurare l’andamento complessivo e per canale.

IMPRESSION	CLIC	ISCRIZIONI	CONVERSIONE / CLIC
93.000	4.800	550	11,5%

ISCRIZIONI PER CANALE



OBIETTIVO

550 / 600

50 iscrizioni sotto il target

KPI = Key Performance Indicator. Tasso di conversione = iscrizioni / clic × 100.

Esplorare i dati con la Business Intelligence

Le operazioni OLAP cambiano il livello di dettaglio e la prospettiva di analisi.

DRILL-DOWN

Da regione a provincia e comune

Aumenta il dettaglio

ROLL-UP

Da comune a provincia e regione

Aggrega i dati

SLICING

Seleziona un periodo o una dimensione

Ritaglia una vista

DICING

Incrocia più criteri di filtro

Crea un sottoinsieme

PIVOTING

Scambia righe e colonne

Cambia prospettiva

DRILL-THROUGH

Dall’aggregato ai record sorgente

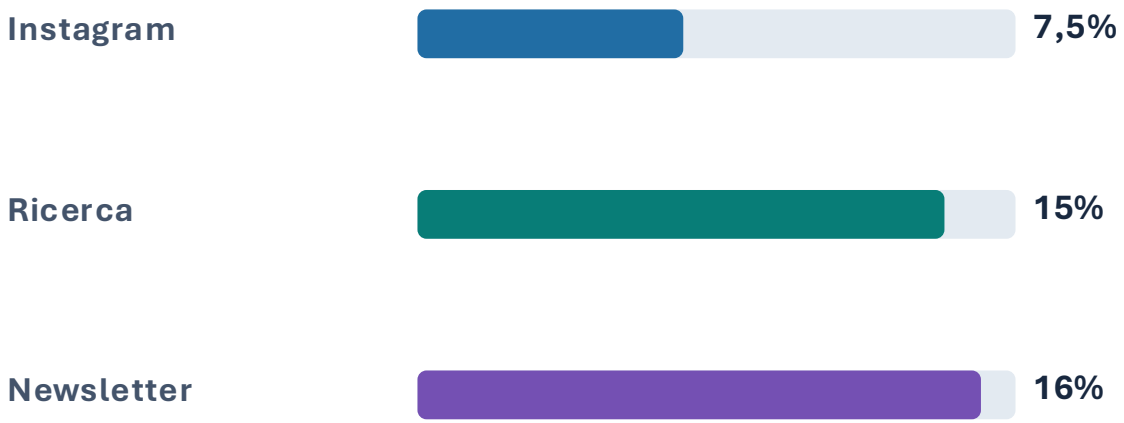
Accede ai dati originali

OLAP = Online Analytical Processing · Tecnologie correlate: data warehouse, dashboard e report.

Analisi diagnostica: cercare spiegazioni

Confrontare i tassi di conversione è diverso dal confrontare solo i volumi.

ISCRIZIONI / CLIC PER CANALE



IPOTESI DA VERIFICARE

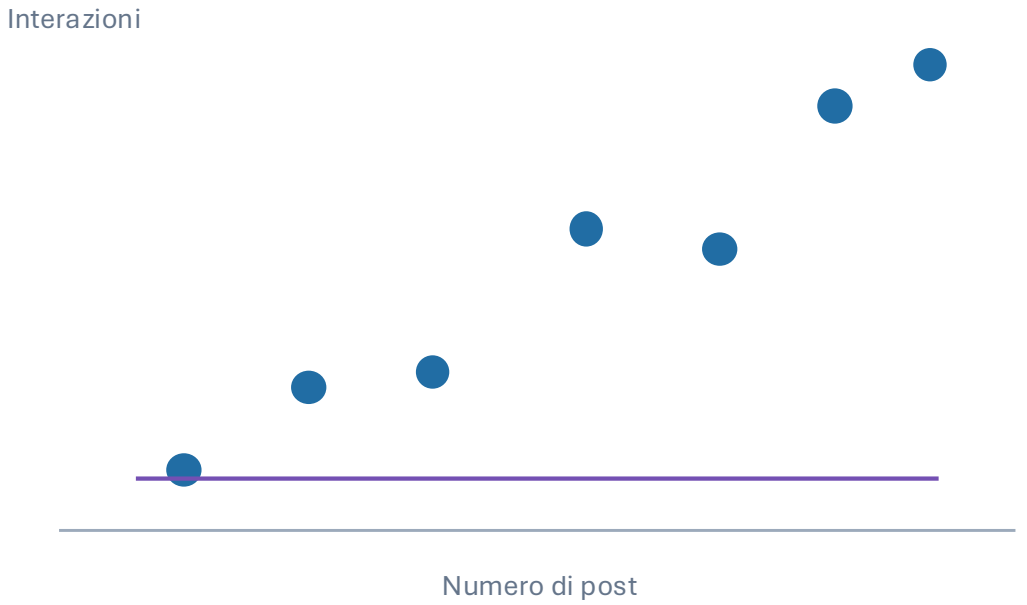
- Pubblici con intenzioni diverse
- Creatività e CTA differenti
- Orari di esposizione
- Percorso di iscrizione

Differenza osservata ≠ spiegazione dimostrata: servono controlli, segmentazioni e test.

Correlazione non significa causalità

Una relazione statistica non è, da sola, una prova di causa ed effetto.

ESEMPIO: POST PUBBLICATI E INTERAZIONI



QUALI ALTRE VARIABILI CONTANO?

- **Dimensione del pubblico**
  Più follower possono aumentare entrambi i valori.
- **Qualità dei contenuti**
  Un messaggio più rilevante favorisce interazioni.
- **Tempo e contesto**
  Eventi, stagionalità e investimenti influenzano i dati.

Esempio schematico con valori sintetici; non implica una relazione causale.

Machine Learning: supervisionato e non supervisionato

Gli algoritmi apprendono regolarità dai dati, ma non tutti richiedono una variabile target.

SUPERVISIONATO

Si conosce il risultato corretto negli esempi storici: il modello cerca di stimarlo per nuove osservazioni.

- CLASSIFICAZIONE
- REGRESSIONE

NON SUPERVISIONATO

Non c’è un risultato target da imparare: si cercano gruppi, strutture e regolarità nei dati.

- CLUSTERING
- ASSOCIAZIONI

Le regole di associazione e il clustering non sono di per sé previsioni di eventi futuri.

Che cosa significa Machine Learning?

Dal programma con regole esplicite a un modello che apprende regolarità dai dati.

PROGRAMMAZIONE TRADIZIONALE

Regole scritte a mano

↓

Il software applica le regole definite da chi lo sviluppa.

MACHINE LEARNING

Esempi storici + risultati

↓

L'algoritmo stima un modello dai dati per trattare nuovi casi.

L'essere umano definisce obiettivo, dati, criteri di valutazione e limiti d'uso.

Quali dati usa un modello?

Ogni riga rappresenta un'osservazione; le feature descrivono il caso e il target è il risultato noto.

Utente	Visite sito	Clic su email	Interazioni social	Iscritto?
A	4	1	3	SI
B	1	0	0	NO
C	3	1	2	SI
D	2	0	1	NO
E	5	1	2	SI

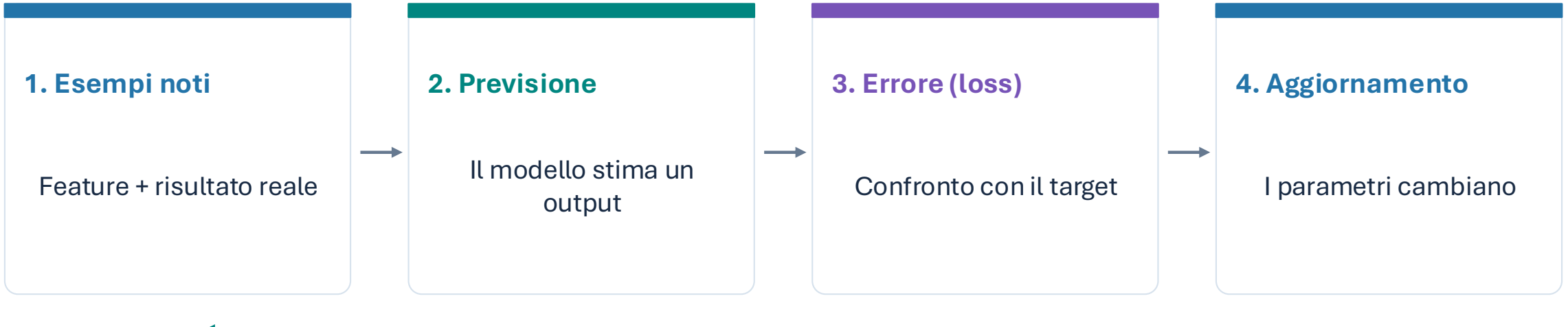
FEATURE (INPUT)

TARGET

In previsione non possiamo usare informazioni disponibili solo dopo l'iscrizione.

Addestramento: come il modello impara

Training significa modificare progressivamente i parametri del modello per ridurre una misura di errore.



Il ciclo si ripete su molti esempi e spesso per più epoche (passaggi sui dati).

Esempio: se il target reale è 1 e la probabilità prevista è 0,30, l'errore guida l'aggiornamento.

Training, validation e test: ruoli distinti

La separazione impedisce di confondere la capacità di ricordare i dati con quella di generalizzare.

Esempio illustrativo: 1.000 interazioni storiche



TRAINING

Per apprendere i parametri del modello.

Il modello vede questi esempi durante il fit.

VALIDATION

Per scegliere modello e iperparametri.

Guida le scelte di sviluppo.

TEST

Per stimare la prestazione finale.

Resta separato fino alla valutazione.

La divisione 70/15/15 è solo un esempio, non una regola universale.

Training e inferenza: due fasi diverse

Il modello è costruito sui dati storici; poi viene applicato a osservazioni mai viste.

FASE 1: ADDESTRAMENTO

Dati storici con etichette

↓

Algoritmo di apprendimento

Risultato: MODELLO addestrato

FASE 2: INFERENZA

Nuovo utente: feature disponibili

↓

Modello già addestrato

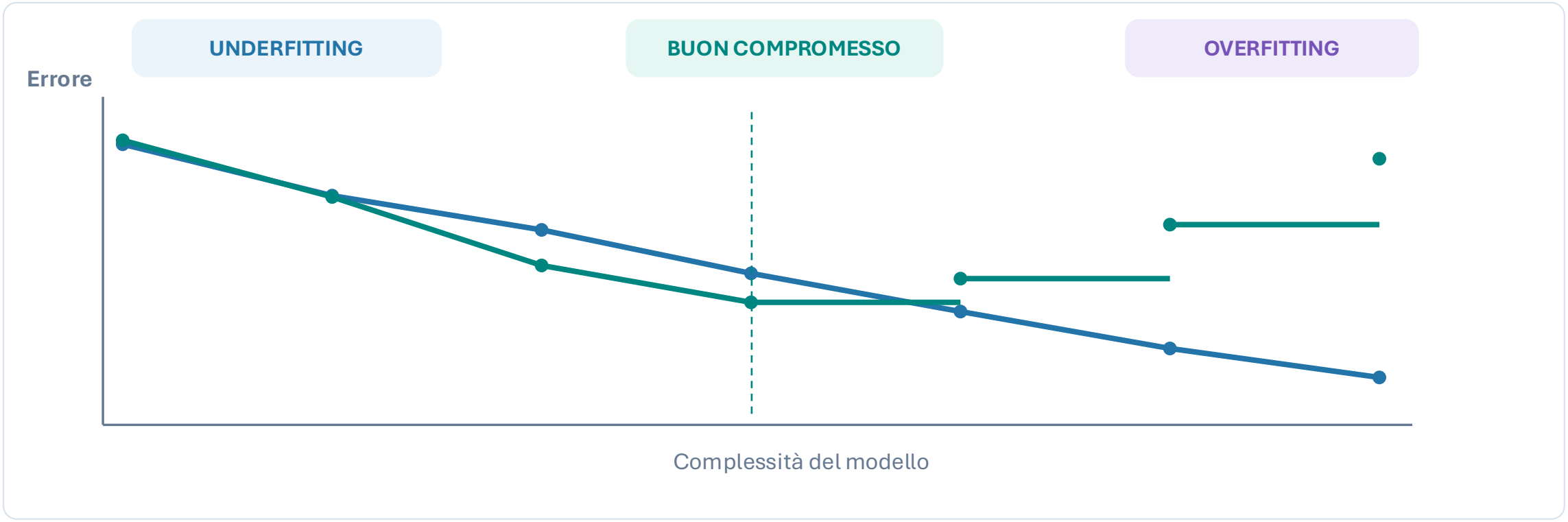
Probabilità stimata = 0,72

Soglia 0,60: classe prevista SI

Una previsione non è una certezza né una decisione automatica da eseguire.

Generalizzazione: underfitting e overfitting

Un buon modello deve apprendere i pattern utili, non memorizzare il rumore del training set.



Errore sul training

Errore sulla validation

Curve schematiche, non dati reali.

Cross-validation: valutare più volte

La k-fold cross-validation riduce la dipendenza da una sola divisione training/validation.

Esempio: 5-fold cross-validation sul set di sviluppo

Giro 1	VALIDATION	TRAINING	TRAINING	TRAINING	TRAINING
Giro 2	TRAINING	VALIDATION	TRAINING	TRAINING	TRAINING
Giro 3	TRAINING	TRAINING	VALIDATION	TRAINING	TRAINING
Giro 4	TRAINING	TRAINING	TRAINING	VALIDATION	TRAINING
Giro 5	TRAINING	TRAINING	TRAINING	TRAINING	VALIDATION

Si confrontano le metriche nei diversi giri e se ne sintetizzano i risultati.

Il TEST finale resta separato. Per serie temporali si usa una validazione rispettosa del tempo.

Analisi predittiva: dai dati a una probabilità

Un modello collega caratteristiche osservate e risultati storici.

01

INPUT

Visite · Clic · Interessi

→

02

MODELLO

Apprende pattern dai casi storici

→

03

OUTPUT

P(iscrizione) = 70%

Un valore del 70% è una stima del modello, non la certezza che un utente si iscriverà.

L’esempio è illustrativo: in un caso reale il modello andrebbe addestrato e validato su dati adeguati.

Quale tecnica risponde a quale domanda?

La scelta dipende dal tipo di risultato richiesto e dalla struttura dei dati.

CLASSIFICAZIONE

L’utente si iscriverà?

Regressione logistica, alberi, Random Forest

REGRESSIONE

Quante iscrizioni stimiamo?

Regressione lineare e modelli non lineari

CLUSTERING

Quali utenti si somigliano?

K-Means e clustering gerarchico

ASSOCIAZIONI

Quali interessi ricorrono insieme?

Apriori; supporto, confidenza e lift

SERIE TEMPORALI

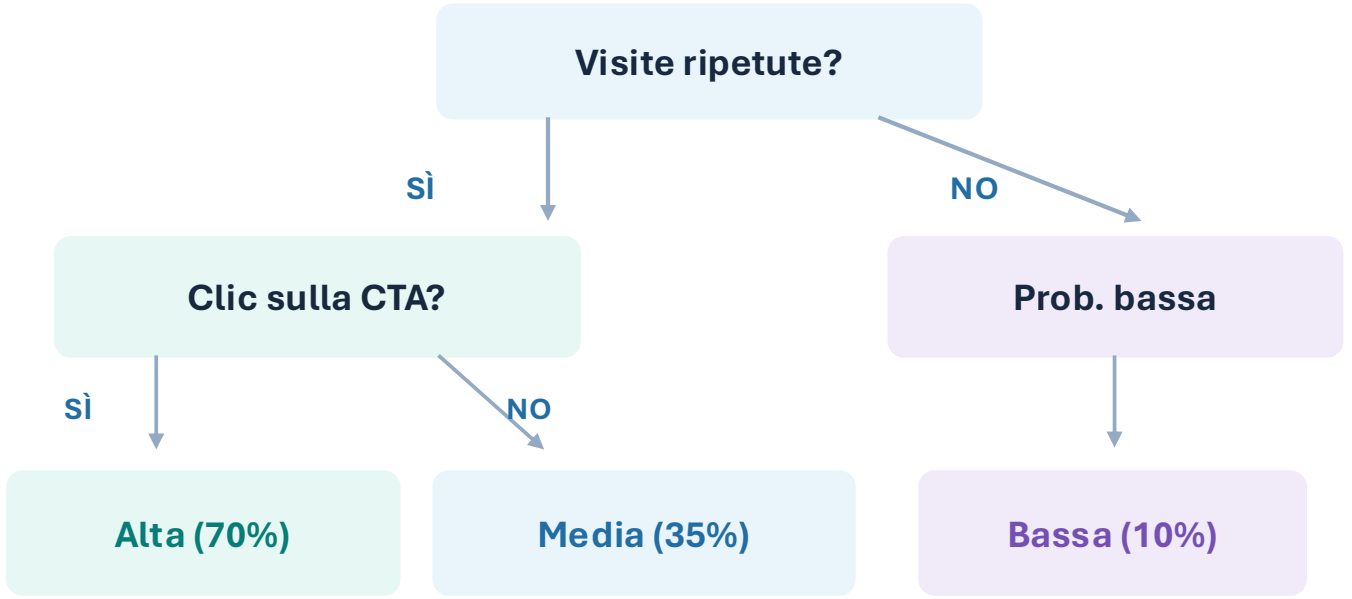
Come evolverà il traffico?

Trend, stagionalità e forecasting

Alcuni strumenti producono previsioni, altri descrivono pattern o raggruppamenti.

Classificazione: chi potrebbe iscriversi?

Un albero decisionale rappresenta una sequenza di condizioni e un esito stimato.



COME INTERPRETARLO

Le condizioni dividono i casi in gruppi. Ogni foglia associa una stima o una classe.

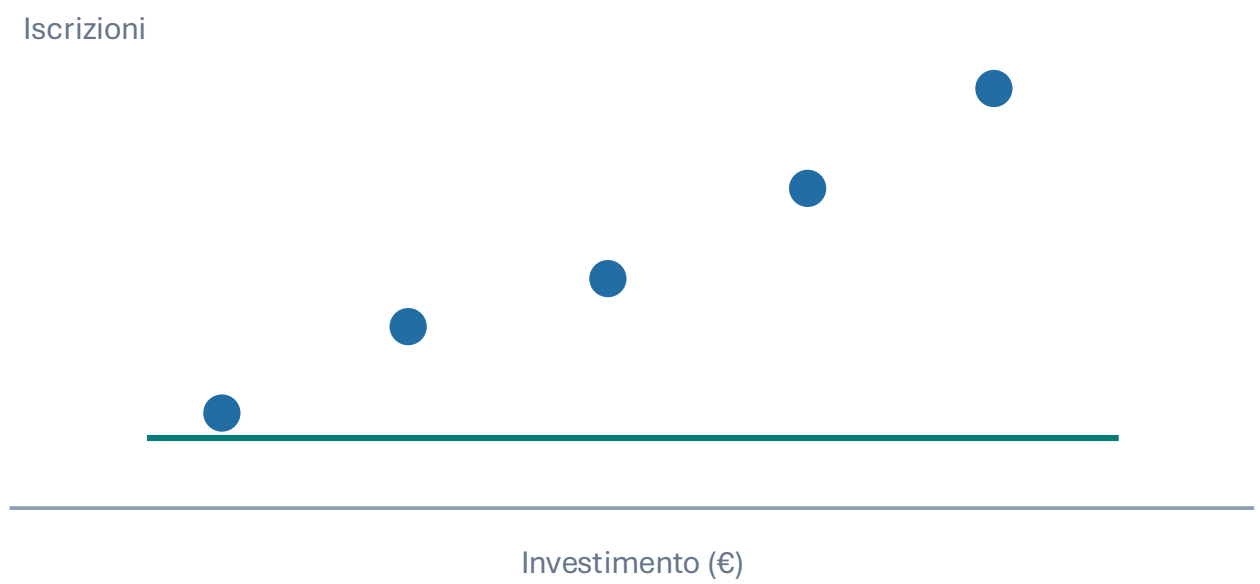
ESEMPIO DIDATTICO

In un modello reale le condizioni, i valori di probabilità e le prestazioni dipendono dal dataset.

Regressione: stimare un valore numerico

Dalle osservazioni si ricava una relazione utile a stimare un risultato.

SPESA PUBBLICITARIA E ISCRIZIONI



MODELLO DI REGRESSIONE

Input numerico

Budget destinato alle inserzioni

Output numerico

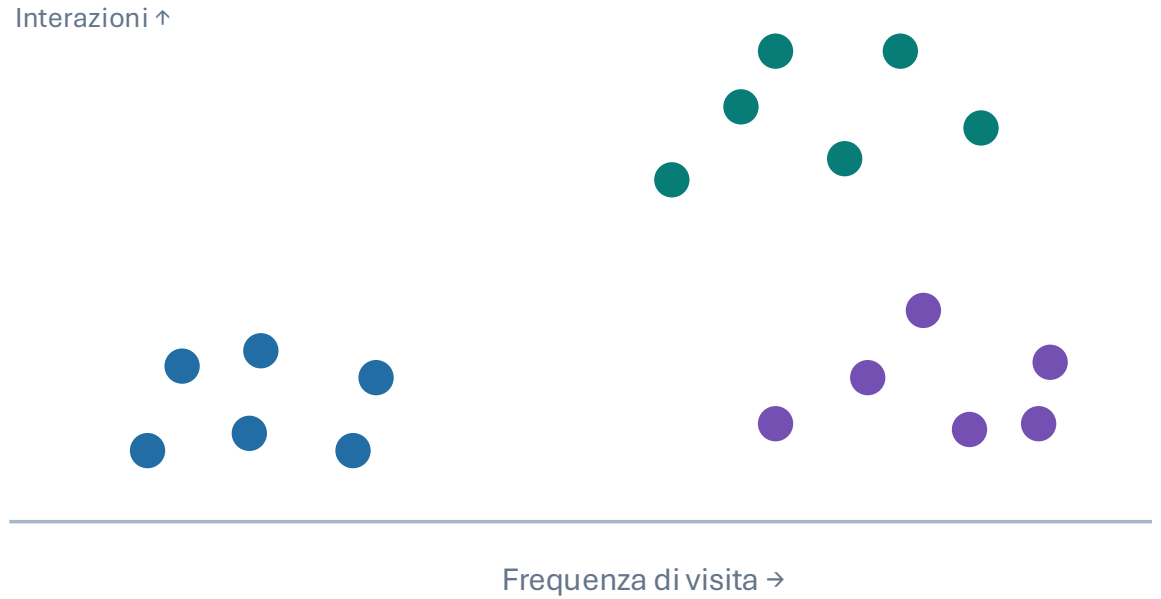
Numero atteso di iscrizioni

Relazione illustrativa, non dimostrazione di causalità. I modelli possono usare più variabili.

Clustering: segmentare il pubblico

Un algoritmo raggruppa osservazioni simili, senza etichette predefinite.

FREQUENZA DI VISITA × INTERAZIONI



OCCASIONALI
Bassa frequenza e interazione

FEDELI
Visite frequenti e coinvolgimento

ESPLORATORI
Molte visite, interazione variabile

I nomi dei segmenti sono interpretazioni dell'analista: vanno verificati sui dati reali.

Associazioni e sistemi di raccomandazione

Le co-occorrenze aiutano a esplorare interessi e combinazioni ricorrenti.

ESEMPIO SU 100 UTENTI

A → B

A = interesse per i workshop
B = interesse per i talk

40%

SUPPORTO

40 su 100 interessati a entrambi

80%

CONFIDENZA

40 su 50 interessati ad A anche a B

1,6

LIFT

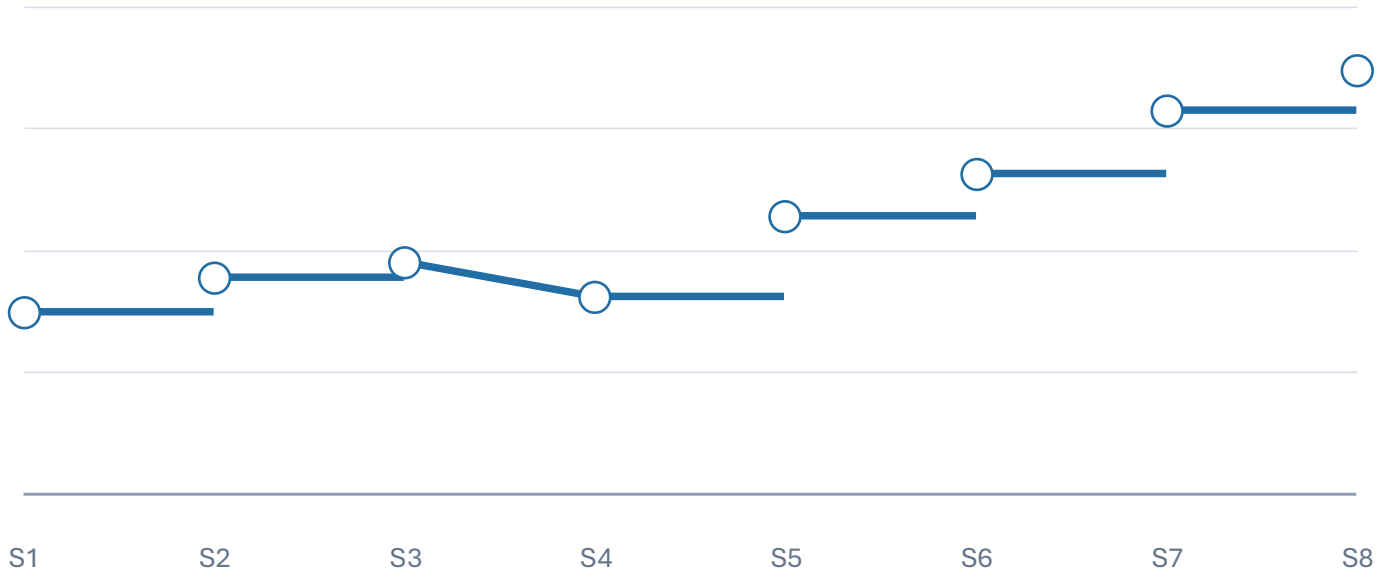
L’associazione supera la frequenza attesa

Dati didattici: 50 utenti interessati ad A; 50 a B; 40 a entrambi. Co-occorrenza ≠ causalità.

Serie temporali: trend e stagionalità

La successione temporale dei dati contiene informazioni preziose per le previsioni.

VISITE AL SITO NELLE ULTIME 8 SETTIMANE



COSA OSSERVARE?

- **Trend**
  Crescita o riduzione di lungo periodo.
- **Stagionalità**
  Pattern che si ripetono nel tempo.

Per valutare un forecast, il test deve rispettare la sequenza temporale dei dati.

Costruire un modello predittivo: la pipeline

L’algoritmo è solo una fase di un processo più ampio, iterativo e verificabile.

01 DOMANDA

Definire target e obiettivo

02 DATI

Raccogliere e controllare qualità

03 PREPARAZIONE

Pulizia e costruzione delle feature

04 SUDDIVISIONE

Separare training e test

05 TRAINING

Addestrare e selezionare il modello

06 VALUTAZIONE

Misurare, usare e monitorare

Attenzione al data leakage: il test non deve influenzare preprocessing e addestramento.

Valutare una classificazione

La matrice di confusione distingue predizioni corrette e diversi tipi di errore.

MATRICE DI CONFUSIONE — 200 CASI

REALTÀ	PREDIZIONE: ISCRITTO	PREDIZIONE: NON ISCRITTO
ISCRITTO	30 — Vero positivo	20 — Falso negativo
NON ISCRITTO	10 — Falso positivo	140 — Vero negativo

METRICHE SUL CASO

85% — **ACCURATEZZA**

170 predizioni corrette su 200

75% — **PRECISION**

30 vere iscrizioni su 40 segnalate

60% — **RECALL**

30 iscritti individuati su 50 reali

Esempio ipotetico. Altre metriche: F1-score, ROC-AUC, PR-AUC. La metrica dipende dall’obiettivo.

Quali metriche per quale problema?

Non esiste una metrica migliore per tutti i problemi.

CLASSIFICAZIONE

Accuratezza · Precision · Recall · F1

Le diverse forme di errore hanno un peso diverso.

REGRESSIONE

MAE · RMSE

Quanto i valori previsti si discostano dai valori osservati?

CLUSTERING

Silhouette score

Quanto sono compatti e separati i gruppi individuati?

Un modello utile deve anche generalizzare, essere monitorato ed essere coerente con l’uso previsto.

F1-score: precision e recall insieme

Riprendiamo la classificazione delle iscrizioni: 30 veri positivi, 10 falsi positivi e 20 falsi negativi.

DALLO STESSO CASO DELLA SLIDE 17

PRECISION	30 / (30 + 10)	75%
RECALL	30 / (30 + 20)	60%
F1-SCORE	**2 × (0,75 × 0,60) / (0,75 + 0,60)**	**66,7%**

COME INTERPRETARLO?

66,7%

È una media armonica: un valore basso di precision o recall riduce l'F1.

ROC-AUC e PR-AUC

Valutano più soglie di decisione; non si ricavano da una sola matrice di confusione.

Esempio numerico ipotetico. F1 vicino a 1 indica migliore equilibrio tra precision e recall.

Regressione: calcolare MAE e RMSE

Esempio: un modello prevede le visite settimanali alla pagina di un evento culturale.

DATI OSSERVATI E PREVISIONI

SETT.	REALTÀ	PREVISIONE	|ERRORE|	ERRORE²
S1	80	75	5	25
S2	100	110	10	100
S3	120	115	5	25
S4	**140**	**180**	**40**	**1.600**

Un errore di 40 visite pesa molto di più nell'RMSE.

ERRORE MEDIO

MAE = 15

(5 + 10 + 5 + 40) / 4 = 15 visite

RMSE ≈ 20,9

√[(25 + 100 + 25 + 1600) / 4]

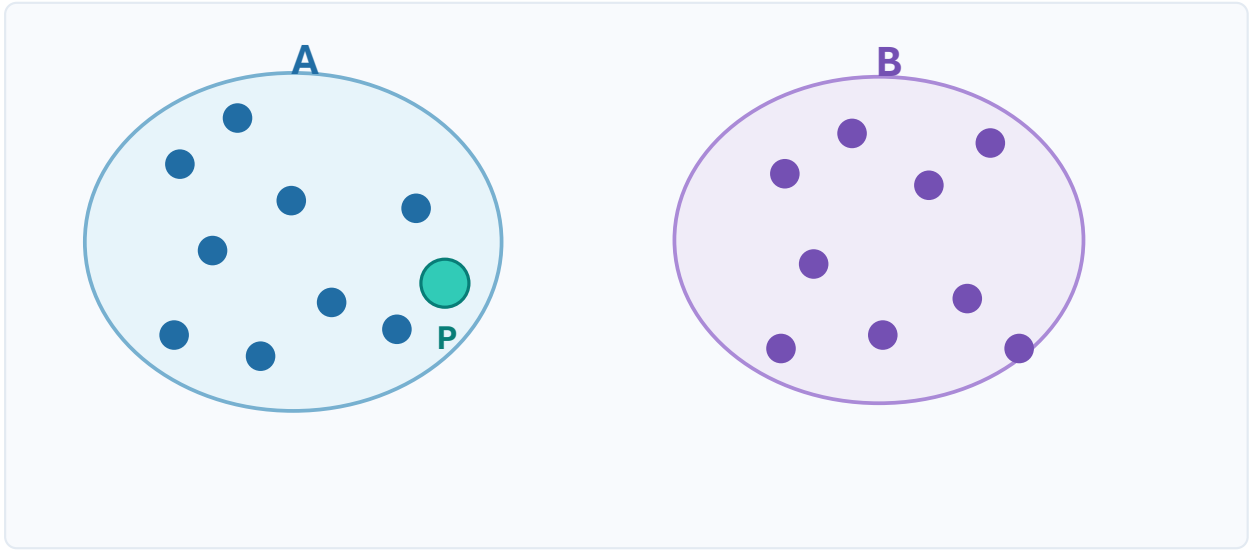
Entrambi in visite; più basso è meglio.

Dati illustrativi. MAE (Mean Absolute Error) media gli errori assoluti; RMSE (Root Mean Squared Error) penalizza maggiormente quelli più grandi.

Clustering: interpretare il silhouette score

Quanto i segmenti di pubblico sono compatti al loro interno e distinti fra loro?

GRUPPI DI UTENTI PER INTERESSI



Schema illustrativo: P appartiene al cluster A.

ESEMPIO PER UN PUNTO P

a = 2 Distanza media da A

b = 5 Distanza media da B

s = (5 − 2) / 5

= +0,60

+1 gruppi ben separati

0 gruppi sovrapposti

< 0 possibile assegnazione errata

Distanze numeriche ipotetiche; il valore complessivo è la media sui punti. Non basta per stabilire l'utilità dei cluster.

Analisi prescrittiva: dalla stima all’azione

Predire un risultato è diverso dal scegliere la decisione migliore.

01

MODELLO PREDITTIVO

PREVISIONE

La ricerca potrebbe generare molte iscrizioni.

02

REGOLE DEL PROBLEMA

VINCOLI

Budget, capacità dei canali, target e tempi.

03

OTTIMIZZAZIONE / SIMULAZIONE

DECISIONE

Come distribuire le risorse fra i canali?

La prescrizione richiede obiettivi, alternative, vincoli e valutazione delle conseguenze.

Simulazione: dove investire 5.000 €?

Confrontiamo tre allocazioni possibili e le conversioni attese.

SCENARIO	INSTAGRAM	RICERCA	NEWSLETTER	ISCRIZIONI ATTESE
A	2.500 €	1.500 €	1.000 €	≈ 638
B	2.000 €	2.000 €	1.000 €	650
C	3.000 €	1.000 €	1.000 €	625

ASSUNZIONI DIDATTICHE Costo per iscrizione stimato: Instagram 10 € · Ricerca 8 € · Newsletter 5 €

Tutti i risultati sono simulazioni; nel problema reale vanno inclusi capienza, saturazione e incertezza delle stime.

Dove la Analytics aiuta la comunicazione

La stessa famiglia di tecniche risponde a bisogni organizzativi diversi.

CRM E CHURN

Prevedere l’abbandono e supportare azioni di fidelizzazione.

SEGMENTAZIONE

Identificare pubblici con interessi o comportamenti simili.

CAMPAIGN TARGETING

Stimare risposte e allocare contatti o investimenti.

RICERCA DI ANOMALIE

Individuare pattern insoliti nel traffico, nei clic o nelle transazioni.

Marketing, CRM e social analytics: il valore dipende dalla domanda e dai dati disponibili.

Social e contenuti: analizzare il linguaggio

Sentiment analysis e topic modeling osservano aspetti diversi di testi non strutturati.

SENTIMENT ANALYSIS

«Evento organizzato benissimo!»	POSITIVO
«La prenotazione è confusa.»	NEGATIVO
«L’evento inizia alle 18.»	NEUTRO

TOPIC MODELING

Individua gruppi tematici ricorrenti, senza classificarli come positivi o negativi.

PROGRAMMA	relatori · workshop · orari
LOGISTICA	sede · trasporti · accesso
ESPERIENZA	accoglienza · organizzazione · attese

Limiti: sarcasmo, ironia, contesto, errori di classificazione, rappresentatività e privacy.

Tecnologie e strumenti per un primo laboratorio

Scegliere lo strumento in base alla domanda, non soltanto alla sua popolarità.

ORANGE DATA MINING

Machine Learning visuale
Workflow a widget; clustering e classificazione.

KNIME

Pipeline analitiche visuali
Preparazione, trasformazione, modelli e report.

GOOGLE ANALYTICS 4

Web e audience analytics
Traffico, eventi, funnel e conversioni.

DATA STUDIO

Dashboard e comunicazione
Visualizzazioni, report interattivi e condivisione.

Per iniziare: Orange per gli algoritmi · GA4 / Data Studio per la comunicazione

Esercitazione: dai numeri alle decisioni

Applicare le quattro domande della Data Analytics al caso della campagna.

01 DESCRIVI

Quali KPI sono sopra o sotto il target?

02 DIAGNOSTICA

Perché i canali potrebbero comportarsi diversamente?

03 PREVEDI

Quali dati useresti per stimare nuove iscrizioni?

04 DECIDI

Come distribuiresti un nuovo budget di 5.000 €?

Il valore nasce dall’unione di dati affidabili, metodi appropriati e interpretazione critica.