



Università di Parma

Dipartimento di Ingegneria e Architettura

Introduzione all'Intelligenza Artificiale

Big Data & Business Intelligence

A.A. 2022/2023

Corso di «Introduzione all'Intelligenza Artificiale» Corso di «Big Data & Business Intelligence»

Data Science & modelli predittivi

Monica Mordonini (monica.mordonini@unipr.it)



SOMMARIO

Sommario



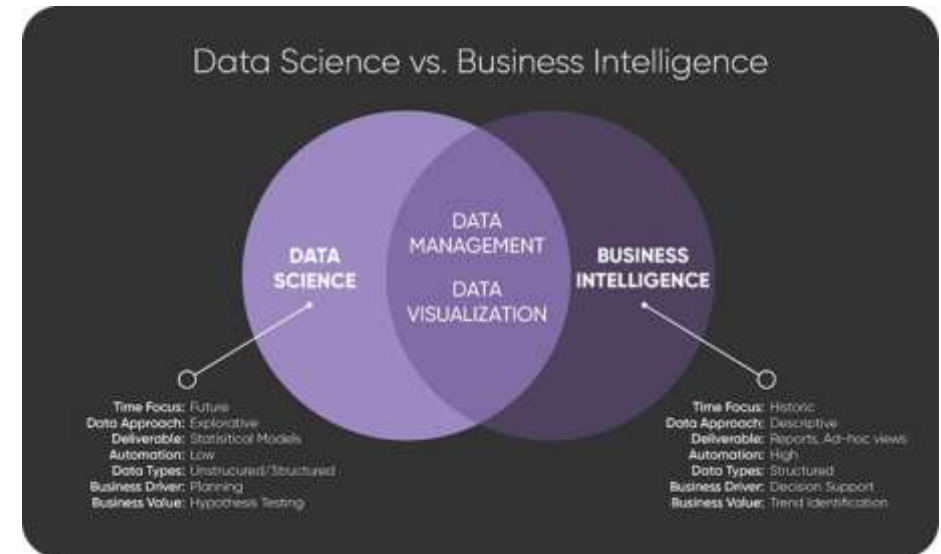
- ✓ Data Science
 - ❑ Introduzione
 - ❑ Una metodologia
 - ❑ Preparazione di un dataset

- ✓ I modelli predittivi
 - ❑ Introduzione
 - ❑ Algoritmi



Bibliografia /credits

- ❑ A. Rezzani (2017). Big Data Analytics. Il manuale del data scientist. Maggioli Editore (Apogeo Education).
 - ❑ S. Ozdemir. Data Science: guida ai principi e alle tecniche base della scienza dei dati. Apogeo.
-

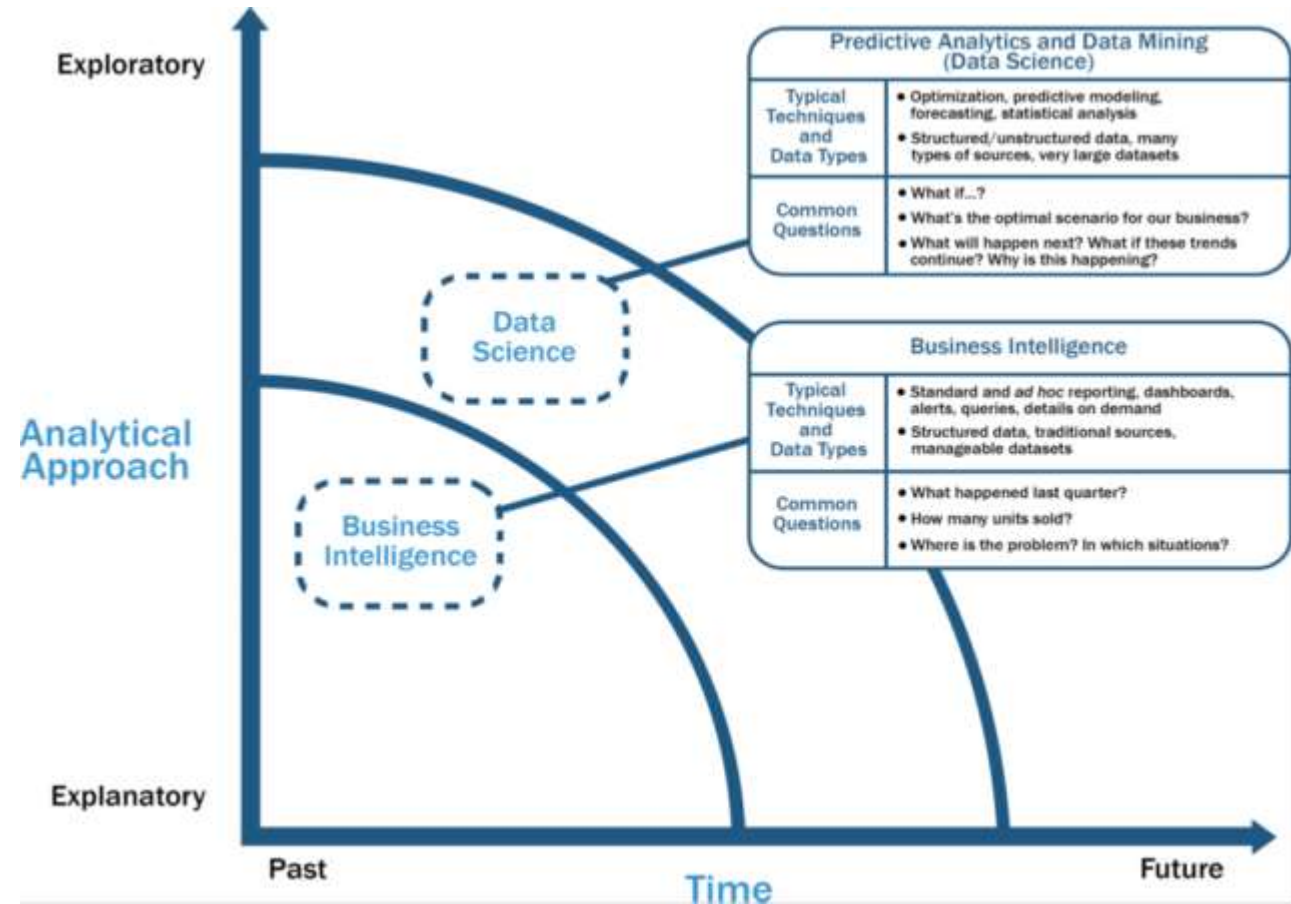


BUSINESS INTELLIGENCE VS DATA SCIENCE

Data monetization

Entrambe cercano di ottenere una «monetizzazione dei dati»

- Ossia la creazione del valore da un *approccio analitico ai dati*





DATA SCIENCE

E UNA METODOLOGIA PER L'ANALISI DEI DATI

Data Science / Scienza dei dati

- E' l'insieme di principi metodologici (basati sul metodo scientifico) e tecniche multidisciplinari volto a interpretare ed estrarre conoscenza dai dati attraverso la relativa fase di analisi da parte di un esperto

➤ **Metodo scientifico**: un processo strutturato, a passi distinti, che se adottato correttamente, preserva l'integrità dei risultati.

- ✓ Il termine "**Scienza dei dati**" è stato introdotto per la prima volta nel 1974 dall'informatico Peter Naur per contrapporlo al concetto più limitativo di informatica (la scienza che si occupa del trattamento dell'informazione mediante procedure automatizzate);
- ma è stata riconosciuta come disciplina a sé stante (quindi non più una branca di informatica e statistica) soltanto nel 2001, quando William Cleveland ne delineò i campi di competenza, elencando sei diverse aree: ricerca multidisciplinare, modelli, elaborazione dati, pedagogia, valutazione degli strumenti e teoria (*Data science: An action plan for expanding the technical areas of the field of statistics*)

I cinque passi della Data Science

- ❑ I cinque passi fondamentali per la scienza dei dati sono i seguenti:
 1. Porre una domanda interessante.
 2. Ottenere i dati.
 3. Esplorare i dati.
 4. Creare un modello per i dati.
 5. Comunicare e presentare i risultati.

I cinque passi della scienza dei dati

1 - *Porre una domanda interessante*

Ci vuole creatività e attenzione al problema

Si possono porre domande indipendentemente dal fatto che si pensi che esistano i dati per rispondere a tali domande: inutile porsi il problema prima di iniziare a cercare i dati.

2 - *Ottenere i dati*

Dove si possono trovare i dati che potrebbero rispondere alla domanda? Sono all'interno dell'azienda? Pubblici o Privati? Costano?

3 - *Esplorare i dati*

Dapprima si cataloga i tipi di dati a disposizione e a manipolarli, ottenendo una maggior conoscenza sul dominio del problema e sul tipo di risposte che i dati in possesso possono fornire.

Prevede la capacità di riconoscere i diversi tipi di dati, trasformare i tipi di dati e usare il codice per migliorare la qualità dell'intero dataset, per prepararlo per la fase di modellazione.

4 - *Creare un modello di dati*

5 - *Comunicare e presentare i risultati*

I cinque passi della scienza dei dati

1 - *Porre una domanda interessante*

2 - *Ottenere i dati*

3 - *Esplorare i dati*

4 - *Creare un modello di dati*

Si fa uso di modelli statistici e di machine learning.

Importante che a questo livello non solo si confrontano e si selezionano i modelli, ma si stabiliscono anche le metriche matematiche di convalida per valutare i modelli e la loro efficacia.

5 - *Comunicare e presentare i risultati*

Questo è un passo molto importante.

Da ricordare che la capacità di concludere la ricerca e presentare i risultati in una forma chiara e comprensibile è molto più difficile di quanto si possa immaginare.

I cinque passi della scienza dei dati

□ I cinque passi fondamentali per la scienza dei dati sono i seguenti:

1. Porre una domanda interessante
 - (obiettivo del business, *ma non solo* spesso si deve proporlo)
2. Ottenere i dati.
3. Esplorare i dati.
4. Creare un modello per i dati.
5. Comunicare e presentare i risultati.

*Questi sono i passi più
«tecnici» di un processo di
scienza dei dati*

*Questo rappresenta la parte di
algoritmi di AI e statistici e la loro
validazione*

1- Porre una domanda interessante - *Business understanding*

- Risulta spesso difficile trovare forti competenze di business e altrettanto forti competenze analitiche, statistiche e informatiche nella stessa persona; per questo, nella realizzazione di questa fase, ma anche nelle successive, è consigliabile la costituzione di un team con più figure professionali.
 - Gli obiettivi costituiscono i benefici del progetto, perciò dovrebbero essere valutati anche dal punto di vista economico.
 - Contemporaneamente dovrebbero essere individuati i costi di progetto (costi del personale interno e/o esterno, costi del software e dell'hardware).
- La documentazione dovrebbe riguardare:
 - Elenco delle risorse disponibili (personale, dati, tecnologie hardware e software).
 - La terminologia di business e la terminologia legata alla predictive analytics. Quest'ultima ha lo scopo di chiarire il significato di termini altrimenti oscuri a molti attori aziendali.
 - Obiettivi e vincoli.
 - Eventuali fattori di rischio.
 - Costi e benefici.
 - Piano di progetto, che include la durata delle varie fasi e le risorse coinvolte.

2- Ottenere i dati- *Data understanding*

- Occorre identificare quali sono i dati rilevanti per la creazione del modello, creando un report che evidenzi le caratteristiche delle fonti dei dati e i criteri di scelta.
- L'attività di esplorazione e descrizione dei dati deve essere completata con la verifica della qualità.
- Devono essere indentificati i dati mancanti e le situazioni anomale (i cosiddetti outliers, ecc.).
- La documentazione da produrre in questa fase è la seguente:
 - Elenco delle fonti dati.
 - Descrizione dei dati.
 - Elenco delle problematiche relative alla qualità dei dati.
 - Descrizione delle attività di esplorazione dei dati (statistiche di base, istogrammi, presenza di valori nulli, ...)

3- Esplorare i dati- *Data preparation*

- *Alla preparazione dei dati è da imputare la maggior parte del tempo speso per un progetto di analisi predittiva*
 - Da essa dipende la qualità del modello utilizzato per l'analisi
- La preparazione del dataset da utilizzare per la costruzione del modello di data mining prevede un'attività di pulizia dei dati
 - qualora la verifica della qualità avesse evidenziato problemi.
- È quasi sempre necessaria l'attività di *feature engineering*
 - cioè la creazione delle variabili di input a partire dai dati originali, che difficilmente potranno essere utilizzati così come lo sono nei modelli predittivi.
- Esiste poi un *problema legato ai valori mancanti*
 - che è da gestire al momento della preparazione dei dati

3- Esplorare i dati- *Data preparation*

- Fa parte della preparazione dei dati la suddivisione del dataset a disposizione in più parti.
 - Una di esse, il *training set*, è utilizzato per la fase di training dell'algoritmo, mentre la valutazione delle performance predittive dell'algoritmo avviene sul *test set*.
- E la relativa documentazione, quali:
 - Motivazioni per l'inclusione o l'esclusione di dati.
 - Descrizione delle operazioni di pulizia dei dati.
 - Descrizione delle attività di trasformazione (aggregazioni, normalizzazioni, ...).

3- Esplorare i dati- *Data preparation* **Domande di base**

- ❑ *I dati sono organizzati oppure no?*

si presentano in una struttura a righe e colonne?

- ❑ *Che cosa rappresenta ogni riga?*
- ❑ *Che cosa rappresenta ogni colonna?*

Si deve identificare il livello di dati di ogni colonna, se si tratta di dati quantitativi o qualitativi e così via.

- Questa suddivisione in categorie potrebbe poi cambiare con il procedere dell'analisi.

- ❑ *Esistono punti dei dati mancanti?*

I dati non sono perfetti.

Ci possono essere dati mancanti e l'esperto

dei dati deve saper prendere delle decisioni sul modo in cui gestire queste discrepanze.

- ❑ *Dobbiamo svolgere trasformazioni sulle colonne?*

A seconda del livello/tipo di dati di ogni colonna, potremmo dover svolgere determinati tipi di trasformazioni.

- Es, in generale, per le attività di modellazione desidereremmo che ogni colonna fosse numerica.

E infine che cosa possiamo inferire dalle valutazioni statistiche inferenziali preliminari?

3- Esplorare i dati- *Data preparation*

Dataset pubblico, reso disponibile dal sito di recensioni di ristoranti Yelp.

- 1- I dati sono organizzati?
- 2 - Cosa rappresenta ogni riga?
- 3 - Ogni colonna?
- 4 - Quali dati sono al livello nominale? Ordinale?
- 5- Esistono dati mancanti?
- 6- Si devono fare delle trasformazioni sulle colonne?

Es: cambio di scala su alcuni dei dati quantitativi? o creazione di variabili fittizie per le variabili qualitative?

A livello di dati nominali:

- Abbiamo un numero ragionevole di elementi univoci?
- La colonna è di puro testo?
- La colonna è del tutto univoca fra tutte le righe?

	business_id	date	review_id	stars	text	type	user_id	cool	useful	funny
0	9yKzy9PApeiPPOUJEtnvkg	2011-01-26	fWKvX83p0-ka4JS3dc6E5A	5	My wife took me here on my birthday for breakf...	review	rLtI8ZkDX5vH5nAx9C3q5Q	2	5	0
1	ZRJwVLyzEJq1VAihDhYiow	2011-07-27	IjZ33sJrzXqU-0X6U8NwyA	5	I have no idea why some people give bad review...	review	0a2KyEL0d3Yb1V6aivbluQ	0	0	0
2	6oRAC4uyJCSjI1X0WZpVSA	2012-06-14	IESLBzqUCLdSzSqm0eCSxQ	4	love the gyro plate. Rice is so good and I als...	review	0hT2KtLiobPvh6cDC8JQg	0	1	0
3	_1QQZuf4zZOyFCvXc0o6Vg	2010-05-27	G-WvGalSbqqaMHINnByodA	5	Rosie, Dakota, and I LOVE Chaparral Dog Park!...	review	uZetI9T0NcROGOyFfughhg	1	2	0
4	6ozycU1RpktNG2-1BroVtw	2012-01-05	1uJFq2r5QfJG_6ExMRCaGw	5	General Manager Scott Petello is a good egg!!!...	review	vYmM4KtsC8ZfQBg-j5MWkw	0	0	0

3- Esplorare i dati- *Data preparation*

Dataset pubblico, reso disponibile dal sito di recensioni di ristoranti Yelp.

E ancora per una Colonna Nominale

quanti valori sono presenti?

quanti valori univoci sono presenti?

il nome dell'elemento più comune nel dataset?

quanto spesso compare nel dataset l'elemento più comune?

	business_id	date	review_id	stars	text	type	user_id	cool	useful	funny
0	9yKzy9PApeiPPOUJEtnvkg	2011-01-26	fWKvX83p0-ka4JS3dc6E5A	5	My wife took me here on my birthday for breakf...	review	rLtl8ZkDX5vH5nAx9C3q5Q	2	5	0
1	ZRJwVLyzEJq1VAihDhYiow	2011-07-27	ljZ33sJrzXqU-0X6U8NwyA	5	I have no idea why some people give bad review...	review	0a2KyEL0d3Yb1V6aivbluQ	0	0	0
2	6oRAC4uyJCSjI1X0WZpVSA	2012-06-14	IESLBzqUCLdSzSqm0eCSxQ	4	love the gyro plate. Rice is so good and I als...	review	0ht2KtflLiobPvh6cDC8JQg	0	1	0
3	_1QQZuf4zZOyFCvXc0o6Vg	2010-05-27	G-WvGalSbqqaMHInNByodA	5	Rosie, Dakota, and I LOVE Chaparral Dog Park!!...	review	uZetl9T0NcROGOyFfughhg	1	2	0
4	6ozycU1RpktNG2-1BroVtw	2012-01-05	1uJFq2r5QfJG_6ExMRCaGw	5	General Manager Scott Petello is a good egg!!!!...	review	vYmM4KTsC8ZfQBg-j5MWkw	0	0	0

Avere una colonna business_id di 1000 valori equivale a dire che vi sono 1000 attività diverse? Vi sono righe «duplicate»? Per esempio c'è lo stesso identico commento? Cioè due celle sono uguali?

3- Esplorare i dati- *Data preparation*

Dataset pubblico, reso disponibile dal sito di recensioni di ristoranti Yelp.

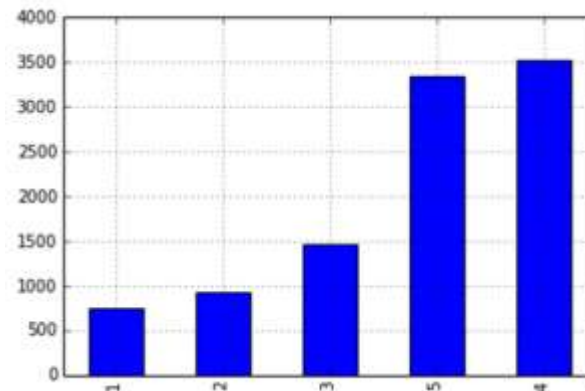
E per una Colonna Ordinale?

Es: La colonna stars

In questo caso non ha molto senso parlare di media aritmetica ma invece quale è la valutazione più comune: 4 o 3 stelle?

Possiamo anche tracciare questi dati per conoscerne meglio l'aspetto....

	business_id	date	review_id	stars	text	type	user_id	cool	useful	funny
0	9yKzy9PApeIPOUJEtnvkg	2011-01-26	fWKvX83p0-ka4JS3dc6E5A	5	My wife took me here on my birthday for breakf...	review	rLtI8ZkDX5vH5nAx9C3q5Q	2	5	0
1	ZRJwVLyzEJq1VAihDhYiow	2011-07-27	IjZ33sJrzXqU-0X6U8NwyA	5	I have no idea why some people give bad review...	review	0a2KyEL0d3Yb1V6aivbluQ	0	0	0
2	6oRAC4uyJCsJI1X0WZpVSA	2012-06-14	IESLBzqUCLdSzSqm0eCSxQ	4	love the gyro plate. Rice is so good and I als...	review	0ht2KtfllobPvh6cDC8JQg	0	1	0
3	_1QQZuf4zZOyFCvXc0o6Vg	2010-05-27	G-WvGalSbqqaMHIInByodA	5	Rosie, Dakota, and I LOVE Chaparral Dog Park!!...	review	uZetI9T0NcROGOyFfughhg	1	2	0
4	6ozycU1RpktNG2-1BroVtw	2012-01-05	1uJFq2r5QfJG_6ExMRCaGw	5	General Manager Scott Petello is a good egg!!!...	review	vYmM4KtsC8ZfQBg-j5MWkw	0	0	0



Sembra che la gente preferisca dare valutazioni positive ...

3- Esplorare i dati- *Data preparation*

Dataset pubblico, titanic

	Survived	Pclass	Name	Sex	Age
0	0	3	Braund, Mr. Owen Harris	male	22
1	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38
2	1	3	Heikkinen, Miss. Laina	female	26
3	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35
4	0	3	Allen, Mr. William Henry	male	35

√Dataset strutturato

=>Ogni riga sembra rappresentare un unico passeggero imbarcato

=>Per le colonne:

1. Survived: una variabile binaria che indica se il passeggero è sopravvissuto. Nominale
2. Pclass: la classe in cui viaggiava il passeggero. Ordinale
3. Name: il nome del passeggero. Nominale
4. Sex: il genere del passeggero. Nominale. E si può trasformare in valore numerico booleano
5. Age: Livello qualitativo o quantitativo? quantitativo, al livello dei rapporti (c'è uno zero naturale)

3- Esplorare i dati- *Data preparation*

Dataset pubblico, titanic

Potremmo decidere di contare le celle valorizzate per ogni colonna ma Age ha solo 714 valori, quindi sono presenti dei valori mancanti...

	Survived	Pclass	Sex	Age
count	891.000000	891.000000	891.000000	714.000000

In questo caso si può:

- eliminare la riga con il valore mancante;
- cercare di completarlo

Nel primo caso si corre il rischio di perdere dati

Nel secondo caso potremmo completarlo inserendo il valore medio delle età.

Oppure il valore medio dell'età femminile o maschile a seconda se il passeggero con età mancante è uomo o donna...

4- Creare un modello per i dati- *Modeling e Evaluation*

- Vi sono molti algoritmi e più algoritmi possono adattarsi ad un determinata analisi
 - Si deve identificare quelli più corretti per il problema prima di proseguire con la creazione del modello.
 - Quest'ultima consiste nell'impostazione dei parametri dell'algoritmo e nella loro calibrazione sulla base dei dati.
- La fase di modellazione comporta, oltre alla scelta dell'algoritmo, anche il *training* dell'algoritmo stesso cioè l'operazione tramite la quale l'algoritmo impara dai dati.
 - Nel caso di *algoritmi supervisionati* (di classificazione o regressione) l'algoritmo tenta di ricavare dal dataset di training le relazioni tra le variabili di input e la variabile di output.
 - Negli *algoritmi non supervisionati* la fase di training riguarda l'estrazione di regole di associazione o la creazione di raggruppamenti che includano elementi del dataset simili tra loro

4- Creare un modello per i dati- *Modeling e Evaluation*

- ✓ La valutazione del modello avviene utilizzando un dataset di test, diverso dal dataset su cui è avvenuto il training del modello.
 - ✓ Ciò vale in particolare per gli algoritmi supervisionati, per i quali è possibile costruire delle metriche basate sulle prediction effettuate su dati per i quali si conosce il valore della variabile target.
 - Il test è significativo solo se è eseguito su dati che l'algoritmo non ha mai trattato in precedenza (cioè nella fase di training).
 - ✓ Per la valutazione degli algoritmi non supervisionati, in particolare quelli di clustering, vi sono alcune metriche da calcolare e su cui basare la decisione di rivedere il modello o di considerarlo pronto per un utilizzo da parte degli utenti finali.
- Se i risultati delle operazioni di test e valutazione non sono soddisfacenti, occorrerà riconsiderare l'attività di modellazione oppure l'attività di preparazione dei dati.

4- Creare un modello per i dati- *Modeling e Evaluation*

- ❑ Le fasi di modellazione e di valutazione sono tipicamente eseguite più volte al fine di trovare l'algoritmo ottimale e il set di parametri che garantisce la miglior performance predittiva.
- ❑ In questa fase la documentazione riguarda:
 - ✓ la descrizione dell'algoritmo utilizzato nella costruzione del modello
 - ✓ le metriche di valutazione per ciascun modello
 - ✓ le eventuali variazioni effettuate ai parametri nelle diverse iterazioni.

5- Comunicare i dati

- ❑ Lo scopo è quello di prendere i risultati ottenuti e spiegarli in modo coerente e comprensibile, in modo che chiunque, indipendentemente dalla sua competenza nel “maneggiare” i dati, sia in grado di comprendere e utilizzare i nostri risultati.
- ❑ La capacità di condurre esperimenti e manipolare i dati in un linguaggio di programmazione non è sufficiente per impiegare la scienza dei dati in modo pratico e applicato.
- ❑ *Questo perché l'efficacia della scienza dei dati dipende molto da come viene impiegata all'atto pratico.*

Un famoso esempio di cattiva gestione della distribuzione dei risultati è il caso dell'abate Gregor Johann Mendel.

Ampiamente riconosciuto come uno dei fondatori della moderna genetica, i suoi risultati (comprensivi di dati e grafici) sono stati presi in considerazione solo dopo la sua morte.

L'abate Mendel li aveva anche inviati a Charles Darwin, che li ignorò sostanzialmente perché erano stati pubblicati solo da un oscuro giornale di Brno, in Moravia.

5-Comunicare i dati

- ❑ Per fare questo si deve:
 - ✓ Identificare i metodi di presentazione efficaci (e quelli inefficaci).
 - ✓ Riconoscere quando i grafici hanno lo scopo di “ingannare” il pubblico.
 - ✓ Riuscire a distinguere la causalità dalla correlazione.
 - ✓ Costruire grafici di grande impatto, per offrire contenuti preziosi.

5-Comunicare i dati

Identificare i metodi di presentazione efficaci (e quelli inefficaci)

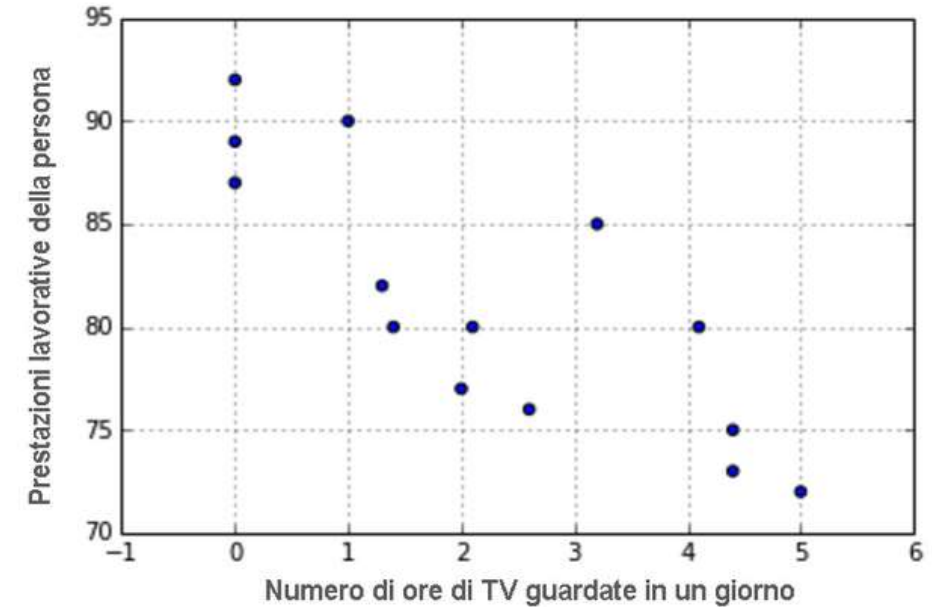
- ❑ L'obiettivo principale della visualizzazione dei dati è quello di comunicare rapidamente i dati al lettore, specificando le possibili tendenze, relazioni e molto altro ancora.
- ❑ L'ideale è che il lettore non debba dedicare più di 5 o 6 secondi ad acquisire una determinata visualizzazione.
- ❑ Osserviamo quattro tipici esempi di grafici:
 - grafici a dispersione,
 - grafici a linee,
 - diagrammi a barre,
 - istogrammi e
 - grafici a scatola e baffi.

5-Comunicare i dati

Grafico a dispersione

- Occorre predisporre due assi quantitativi e usare i punti dei dati per rappresentare le osservazioni.
- L'obiettivo principale di un grafico a dispersione è quello di evidenziare le relazioni esistenti fra due variabili e, se possibile, rivelare una correlazione.
- Ogni punto di un grafico a dispersione rappresenta una singola osservazione
- la sua posizione è un risultato che dice dove si colloca l'osservazione rispetto a ogni variabile.

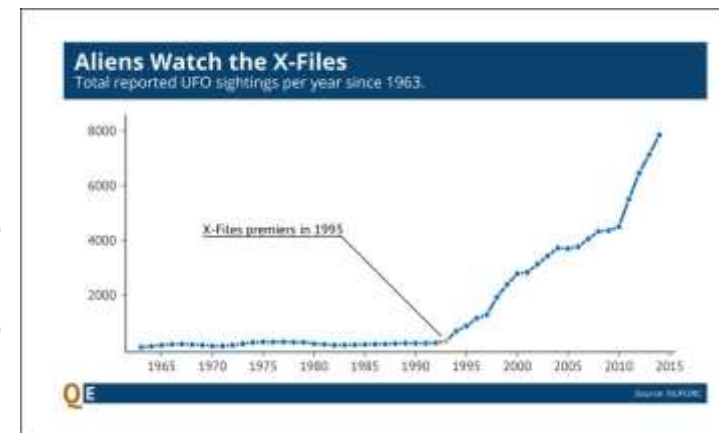
- *Questo grafico a dispersione sembra mostrare una relazione: ciò implica che più TV guardiamo al giorno, più questo pregiudica le nostre prestazioni lavorative.*
- *Questo potrebbe non avere elementi di causalità. Un grafico a dispersione può solo aiutare a rilevare una correlazione o un'associazione, ma non una causalità.*



5-Comunicare i dati

Grafico a linee

- Usa delle linee per connettere i punti dei dati e, normalmente, rappresenta sull'asse x il trascorrere del tempo.
- I grafici a linee sono un modo molto utilizzato per mostrare le variazioni nelle variabili con il trascorrere del tempo.
- Come il grafico a dispersione, il grafico a linee viene usato per tracciare delle variabili quantitative



<http://www.questionable-economics.com/what-do-we-know-about-aliens>.

Sembra evidente che, subito dopo il 1993, anno di uscita della prima stagione di X-Files, il numero di avvistamenti di UFO si è impennato.

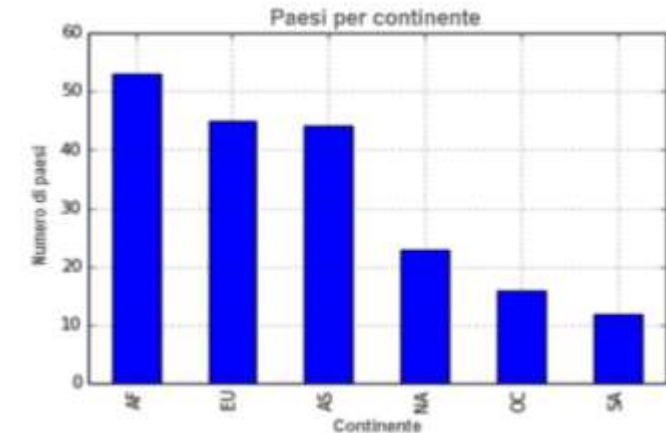
Questo grafico tenta di evidenziare la variazione nel prezzo del gas tracciando tre punti nell'arco del tempo

Però le distanze temporali NON sono affatto uguali: i primi due punti sono separati da un anno, mentre gli ultimi due punti sono separati solo da 7 giorni.

5-Comunicare i dati

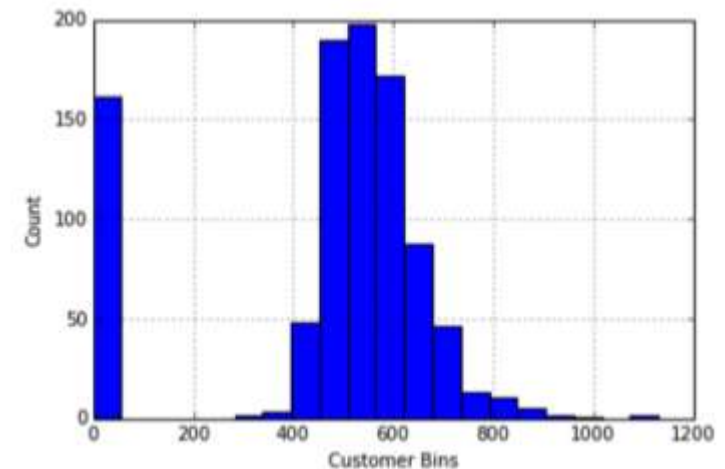
Diagrammi a barre

- Si utilizzano quando si deve tentare di confrontare le variabili di vari gruppi.
- In genere quella sull'asse x è una variabile categorica, mentre quella sull'asse y è quantitativa
- Es: tracciare il numero di paesi per continente.



Istogrammi

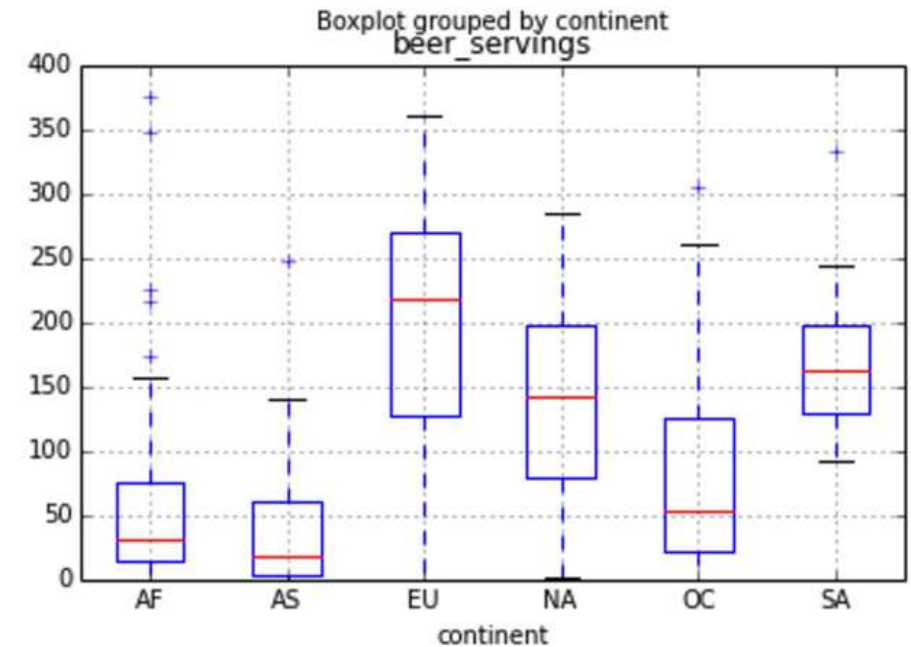
- Gli istogrammi mostrano la distribuzione di frequenza di un'unica variabile quantitativa, raggruppando i dati, per intervallo in lotti equidistanti e tracciando il conteggio delle osservazioni in ogni gruppo.
- Un istogramma è in pratica un grafico a barre in cui l'asse x è un gruppo (intervallo) di valori e l'asse y è un conteggio.
- L'asse x è categorico, per il fatto che ogni categoria comprende un determinato intervallo di valori;



5-Comunicare i dati

Grafici box-plot (a scatola e baffi)

- Vengono usati per mostrare una distribuzione di valori.
- Vengono creati tracciando cinque diversi valori:
 - il valore minimo;
 - il primo quartile (il valore che separa il 25 per cento più basso dei valori da tutti gli altri);
 - la mediana;
 - il terzo quartile (il valore che separa il 25 per cento più alto dei valori da tutti gli altri);
 - il valore massimo.

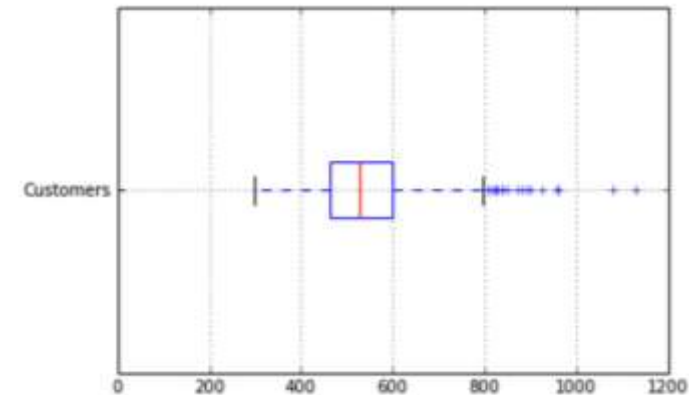
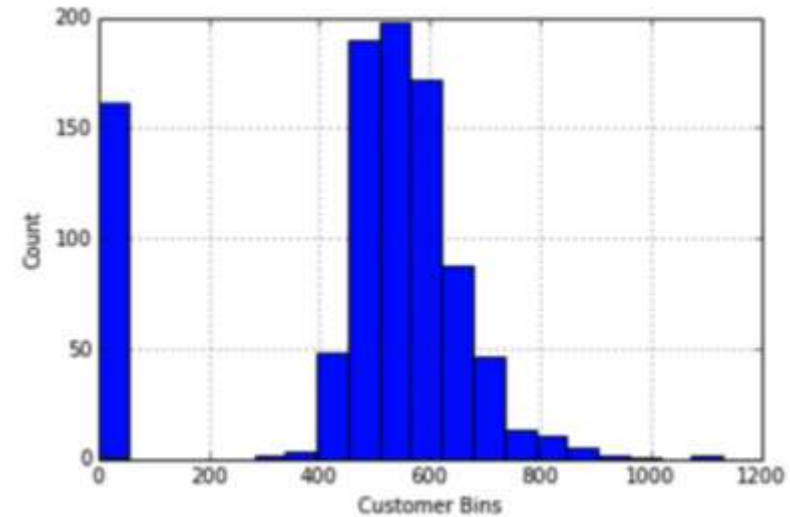


I grafici a scatola e baffi hanno la capacità di mostrare i valori anomali molto meglio di un istogramma. Questo perché nel grafico sono indicati anche il minimo e il massimo.

5-Comunicare i dati

Grafici box-plot e istogramma a confronto

- Il grafico a scatola è molto più rapido nel fornirci il centro dei dati, con la mediana tracciata in rosso
- l'istogramma è più efficace nel mostrarci la dispersione dei dati e dove si collocano prevalentemente i gruppi



5-Comunicare i dati - *Deployment*

- In un'azienda la comunicazione dei dati equivale alla fase di deployment e quindi l'inizio dell'utilizzo reale del modello da parte degli utenti aziendali.
 - E quindi oltre alla presentazione del modello e dell'analisi ci deve essere anche un piano di messa in produzione del modello (risorse, azioni da compiere).
- Un altro aspetto da considerare è il **monitoraggio delle performance** predittive
 - passaggio necessario per ottenere un **feedback** ed eventualmente per operare una revisione del processo di predictive analytics.
- Il feedback consiste nell'analisi dei risultati ottenuti tramite l'applicazione reale del modello.
- Tali risultati possono condurre a due possibili azioni alternative:
 - la nuova applicazione del modello nel periodo futuro
 - la revisione del modello, poiché i risultati attesi non sono coerenti con le aspettative.



DATA SCIENCE

E LA PREPARAZIONE DI UN DATASET

La preparazione di un dataset

- ❑ I cinque passi fondamentali per la scienza dei dati sono i seguenti:
 1. Porre una domanda interessante.
 2. Ottenere i dati.
 - 3. Esplorare i dati.**
 4. Creare un modello per i dati.
 5. Comunicare e presentare i risultati.

La preparazione di un dataset

- ✓ *E' quasi impossibile trovare un dataset immediatamente utilizzabile per un certo tipo di analisi predittiva.*
 - ***Forse la fase più importante all'interno di un processo di predictive analytics***
- **Feature engineering** : attività con cui si creano variabili a partire dai dati grezzi
 - Es. ricerca di comportamenti di un cliente a partire da dati quali numero di transazioni in un certo arco di tempo, numero collegamenti al sito web , etc...
 - Si devono aggregare dati per sintetizzare variabili «utili» al problema
- **Trasformazioni sui dati** per renderli adatti a certi algoritmi:
 - Gli algoritmi di ML vogliono variabili numeriche e normalizzate
- **Definizione della modalità corretta per gestione di dati mancanti**
- **Individuazione** dei valori anomali (gli **outliers**)
- **Riduzione della dimensionalità** (cioè del numero di variabili) senza diminuire la capacità predittiva,
 - cioè scelta delle variabili più discriminanti

La preparazione di un dataset: Le variabili

- ✓ *Le variabili sono gli elementi che formano il dataset e che costituiscono gli attributi delle entità da analizzare.*

Esistono variabili di input (*predictors o features*) e la variabile di output (*obiettivo o target*).

- ✓ I modelli non supervisionati, come il clustering, non hanno il target, anzi il loro obiettivo è proprio la previsione di una variabile di output

Le variabili hanno delle caratteristiche:

- ✓ *alfanumeriche (o stringhe), numeriche* e
- ✓ *date* : rilevanti in analisi di serie storiche, ma che spesso sono ricondotte a variabili numeriche temporali (una sorta di *timestamp* più o meno aggregato) o che, tramite operazioni di trasformazione, danno luogo a nuove variabili
 - ✓ Es: costruzione di nuove variabili avendo i dati dei bonifici bancari con la data in cui sono stati fatti: si può tenere in considerazione le variabili costruite quali il numero dei bonifici ultimi 3 mesi, 6 mesi,...

La preparazione di un dataset: *Le variabili*

Le variabili hanno delle caratteristiche:

- ❑ ***continue*** (o quantitative) e ***categoriche*** (o qualitative)

- ❑ ***Variabili continue*** sono suddivisibili in quelle con scala di misura a intervalli (***interval variables***) e quelle a scala di misura a rapporti (***ratio variables***).
 - Nel primo caso è possibile realizzare confronti per differenza tra i valori, ma non esiste un valore che indica l'assenza del fenomeno.
 - Per esempio, nelle temperature misurate con la scala Celsius lo zero indica una temperatura convenzionale, ma non l'assenza di temperatura.

La preparazione di un dataset: *Le variabili*

- **Variabili categoriche** possono ulteriormente essere suddivise in:
 - **Nominali**: contengono due o più categorie che però non possiedono alcun ordinamento intrinseco
 - Tipo di abitazione: monofamigliare, bilocale, loft ...in realtà non esprimono alcun ordine: un appartamento di prestigio in certe zone può costare molto di più di una villa in campagna.
 - **Dicotomiche** presentano solo due categorie (sì o no, 0 o 1)
 - **Ordinali** contengono più categorie che possono essere ordinate
 - per esempio la variabile rating con i valori AAA, AA, A, BBB, BB, B, ecc. contiene valori che hanno un preciso ordinamento.

Oss.: Nelle variabili ordinali non è detto che gli intervalli tra i valori espressi da ciascuna categoria e quella precedente siano di ugual grandezza.

Per le analisi statistiche e predittive, spesso si assume che lo siano, attribuendo valori numerici ai livelli in modo che essi siano equidistanti, tuttavia occorre prestare molta attenzione

La preparazione di un dataset: *Esplorazione dei dati*

Analisi univariata

➤ *Una variabile per volta*

➤ *Identificate le fonti e recuperati i dati grezzi serve un'analisi esplorativa che ne metta in evidenza le caratteristiche principali*

✓ *Variabili continue*

- Si utilizzano le tecniche statistiche di base per il calcolo degli indicatori di tendenza centrale (media, mediana, moda, ma anche minimo e massimo) e quelli di dispersione (deviazione standard, varianza, quartili).
- A livello visuale possiamo rappresentare i valori delle variabili continue attraverso *istogrammi o box plot*.

✓ *Variabili categoriche*

- Si calcolano i conteggi per ogni categoria sia in valore assoluto sia in percentuale sul totale
- Si visualizzano i risultati tramite un *grafico a barre*

La preparazione di un dataset: *Esplorazione dei dati*

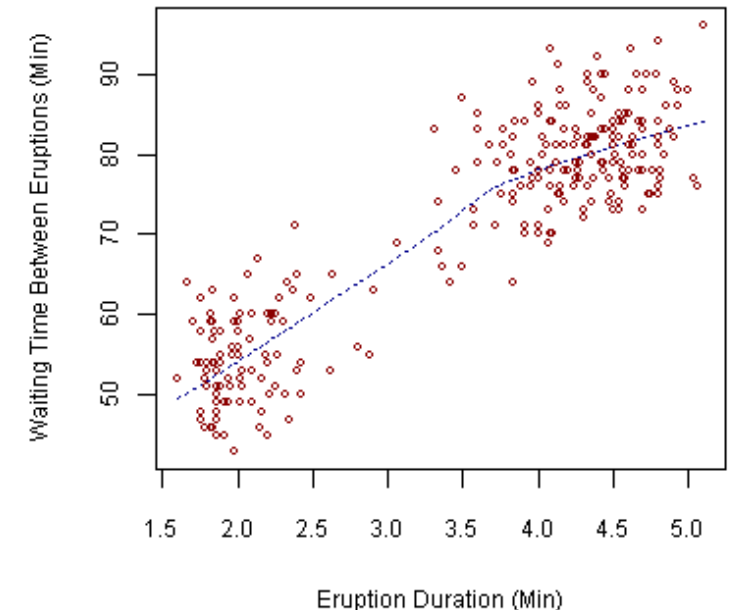
Analisi bivariata

- *Due variabili per volta*
 - *Eseguita su tutte le possibili coppie di variabili (diverse)*
1. una variabile continua confrontata con un'altra variabile continua
 2. una variabile continua confrontata con un'altra variabile categorica
 3. una variabile categorica confrontata con un'altra variabile categorica

La preparazione di un dataset: *Esplorazione dei dati*

- ❑ *una variabile continua confrontata con un'altra variabile continua*
 - Calcolo della correlazione lineare con:
 - 1 (perfetta correlazione *negativa*) ,
 - 0 (assenza di correlazione),
 - 1 (perfetta correlazione *positiva*)
 - Si possono visualizzare con i grafi a dispersione
 - Lo scopo non è dare una misura dell'intensità della relazione
 - ma evidenziarne lo schema
 - sulle x sta la variabile indipendente,
 - sulle y sta la variabile misurata

Old Faithful Eruptions



Tempo di attesa tra le eruzioni e durata delle eruzioni dell'[Old Faithful Geyser](#) nel [Yellowstone National Park](#), [Wyoming](#), USA. Il grafico suggerisce due tipi di eruzioni: corta attesa e corta durata e lunga attesa e lunga durata.

La preparazione di un dataset: *Esplorazione dei dati*

- una variabile continua confrontata con un'altra variabile categorica
 - Il confronto avviene sul piano visuale tramite la creazione di box-plot per ciascun livello della variabile categorica
 - Si cerca significatività statistica nella differenza tra le medie di gruppi di valori, presi due a due, utilizzando uno z-Test per variabili dicotomiche
 - oppure tramite un'analisi della varianza (ANOVA, dall'inglese Analysis of Variance) per variabili con più categorie

z-test è un test statistico di tipo parametrico con lo scopo di verificare se il valore medio di una distribuzione si discosta significativamente da un certo valore di riferimento.

ANOVA è un insieme di tecniche statistiche facenti parte della statistica inferenziale che permettono di confrontare due o più gruppi di dati confrontando la variabilità interna a questi gruppi con la variabilità tra i gruppi.

Preparazione dei dati: *Operazioni sulle variabili*

Così come i dati , le variabili devono essere manipolate per ottenere variabili significative e consistenti.

- ↪ Alcune trasformazioni sono vere e proprie operazioni di modellazione e, come tali, hanno un impatto sulle performance predittive del modello.
- ↪ Altre trasformazioni che avvengono riga per riga, applicando una funzione al singolo valore, senza impiegare parametri ricavati dal dataset non hanno questo problema
 - *Es. l'applicazione della funzione logaritmo per mitigare l'asimmetria della distribuzione di una variabile*

Preparazione dei dati: *Operazioni sulle variabili*

Trasformazioni con impatto sul modello

- ❑ Creazione di variabili dai dati grezzi.
- ❑ Cambiamento dei valori delle variabili che avvengono utilizzando parametri ricavati dal dataset e che quindi sono determinati da più righe dello stesso.
 - Es, tutte le trasformazioni che includono il calcolo di medie, deviazioni standard, minimi, massimi o conteggi.
- ❑ Selezione delle variabili da utilizzare.
 - Quì si decide quali sono le variabili dipendenti da cui scaturisce la prediction.
- ❑ Azioni per contrastare gli effetti della presenza di classi sbilanciate, nei problemi di classificazione.
- ❑ Trattamento dei valori mancanti.
- ❑ Codifica di variabili categoriche finalizzata alla trasformazione in variabili continue.

Preparazione dei dati: *Operazioni sulle variabili*

Trasformazioni con impatto sul modello

- *Queste operazioni devono essere fatte esclusivamente sulla base del training set*
 - ❑ I parametri utilizzati per le trasformazioni (ovvero medie, deviazioni standard, conteggi, ecc.) devono essere successivamente utilizzati anche per trasformare i dati del test set o dei nuovi dati su cui effettuare la prediction.
- ✓ *Si potrebbe pensare che tali valori, se calcolati su tutti i dati disponibili (training set + test set + validation set), siano più significativi.*
 - *Tuttavia ciò renderebbe la valutazione dell'algoritmo meno affidabile.*
 - *Lo scopo del modello predittivo è proprio la sua applicazione ai nuovi dati, sconosciuti al momento del training.*
 - ✓ *È errato anche l'approccio con cui si operano trasformazioni separate sui sottoinsiemi di dati, lavorando quindi sul training set con parametri calcolati su di esso e sul test set con parametri calcolati su quest'ultimo.*
 - *In questo caso si può andare incontro ad inconsistenze dovute alla diversità dei parametri*

Preparazione dei dati: *Operazioni sulle variabili*

Creazione di variabili o *feature engineering*

➤ *dipende dal contesto di business e dal problema che si sta analizzando.*

↗ richiede forti competenze sia di business, sia nel campo dell'elaborazione dei dati ed è un'attività con un forte impatto sul modello predittivo

- ❑ **Aggregazione:** Le aggregazioni avvengono raggruppando i dati per l'entità di riferimento (es: *il cliente*) ed effettuando i calcoli su tutti i dati, oppure su sottoinsiemi determinati dalla tipologia di transazione, o evento, da un intervallo di date o combinando più criteri (es: *conteggio bonifici effettuati negli ultimi 6 mesi*).
- ❑ **Scomposizione:** in alcuni casi una variabile può essere suddivisa in due o più variabili. Es.: una variabile «data» potrebbe essere scomposta in anno, mese e giorno, creando tre variabili distinte.
- ❑ **Applicazione di funzioni:** le variabili possono essere ricavate utilizzando calcoli o funzioni che impiegano come input una o più colonne del dataset originale.

Preparazione dei dati: *Operazioni sulle variabili*

Trasformazioni dei valori

Normalizzazione (o Rescaling)

- ❑ Gli algoritmi basati sul calcolo della distanza tra punti nello spazio multidimensionale beneficiano particolarmente della normalizzazione dei valori ad un intervallo (es $[-1, +1]$, $[0, 1]$).
 - In caso contrario gli algoritmi finirebbero per dare un peso eccessivo ai valori più grandi

Di solito si usa per normalizzare :

- ❑ I valori di minimo-massimo
- ❑ La deviazione standard

Preparazione dei dati: *Operazioni sulle variabili*

❑ *Normalizzazione (o Rescaling) con minimo-massimo*

$$ValNorm_i = \frac{Val_i - Min(Val)}{Max(Val) - Min(Val)}$$

- ✓ Non introduce distorsioni nei dati
- ✓ Non garantisce che, in seguito, non vi siano dati il cui minimo o massimo valore sia fuori dal range utilizzato per la normalizzazione che vanno gestiti:
 - Non compiere nessuna azione, se si usa un algoritmo in grado di gestire i valori fuori range.
 - Prevedere in anticipo abbastanza “spazio” per gli eventuali valori fuori range, utilizzando un valore superiore al massimo al posto di $Max(Val)$ e inferiore al minimo al posto di $Min(Val)$.
 - Effettuare un clipping dei valori che eccedono gli estremi, ponendoli pari al minimo, se sono inferiori ad esso, oppure ponendoli pari al massimo, se sono superiori a quest'ultimo.

Preparazione dei dati: *Operazioni sulle variabili*

Normalizzazione (o Rescaling) con deviazione standard

$$ValNorm_i = \frac{Val_i - Media(Val)}{DevStd(Val)}$$

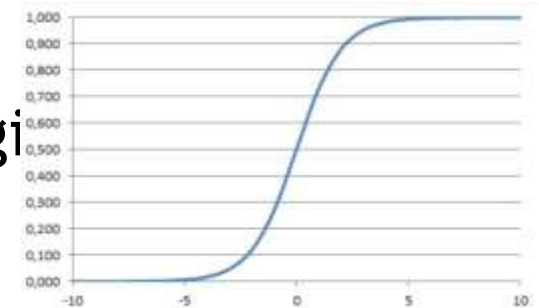
- ❑ il valore normalizzato è dato dalla distanza del valore originale e la media dei valori, espressa in termini di numero di deviazioni standard
- ❑ Non introduce distorsioni nei dati
- ❑ Per le variabili discrete la normalizzazione può avvenire con la stessa formula nella quale, però, la media è pari alla probabilità p dello stato considerato, mentre la deviazione standard è pari a $p*(1-p)$

Preparazione dei dati: *Operazioni sulle variabili*

□ *Normalizzazione (o Rescaling) con funzione logistica*

$$ValNorm_i = \frac{1}{1+e^{-val_i}}$$

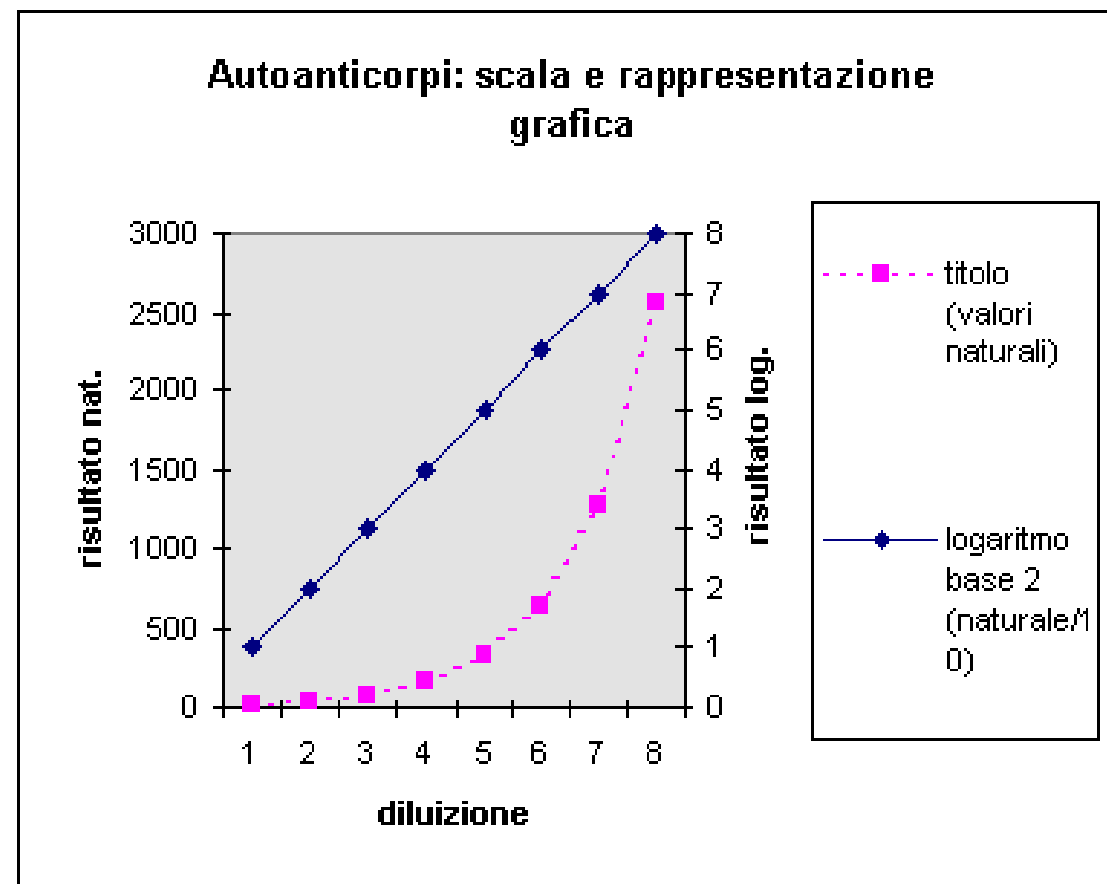
- il valore normalizzato è ottenuto tramite la funzione logi
- Introduce distorsioni nei dati
 - schiaccia i valori nell'intervallo (0,+1)



Preparazione dei dati: *Operazioni sulle variabili*

Simmetria della distribuzione

- I valori che presentano distribuzioni simmetriche sono preferibili a valori con distribuzione con evidenti asimmetrie a destra (alcuni valori molto più grandi rispetto agli altri) o a sinistra (alcuni valori estremamente piccoli se confrontati con la maggior parte degli altri).
- Per far fronte a questi problemi è possibile applicare
 1. nel primo caso una funzione logaritmo, radice quadrata o radice cubica,
 2. nel secondo caso il quadrato, il cubo o l'esponenziale.



Preparazione dei dati: *Operazioni sulle variabili*

Riduzione del numero di livelli nelle variabili categoriche

- Un elevato numero di livelli può rendere inefficiente un algoritmo

Si può :

- ❑ Applicare delle regole di business per raggruppare più livelli
 - in genere molto efficace
- ❑ Utilizzare un livello gerarchico superiore
 - Es.: edificio invece di appartamento
- ❑ Utilizzare delle frequenze di ogni categoria per raggruppare livelli con frequenze simili.
 - creazione di una categoria per i valori poco frequenti e lasciandole altre categorie inalterate.

Preparazione dei dati: *Operazioni sulle variabili*

Discretizzazione o binning

- Da una variabile continua si ricava una variabile categorica attraverso la creazione di fasce di valori (bins)
 - Tecnica che riduce molto il livello di informazione racchiuso nei dati

Utile quando:

- una variabile continua possiede dei livelli standard che sono significativi per un dato problema predittivo, oppure in generale se la distribuzione dei casi si concentra attorno a specifici valori, determinando di fatto una suddivisione in livelli
 - Es.: l'età che in certi casi può essere ridotta a variabile dicotomica (maggiore o no)

Preparazione dei dati: *Operazioni sulle variabili*

Codifica di variabili categoriche

- ❑ **Ordinal Encoding** attribuiamo un numero progressivo alle categorie che costituiscono una data variabile
- ❑ **One-Hot Encoding** si creano n colonne quante sono le categorie e per ogni categoria viene messo a 1 il valore della colonna corrispondente al dato originale

Variabile originale	basso	medio	alto
basso	1	0	0

- ❑ **Binary encoding** si codificano i dati in numero ordinale e poi si converte questo numero in binario

Variabile originale	Ordinal Encoding	Codifica binaria	Nuova Var 1	Nuova Var 2
basso	1	01	0	1
medio	2	10	1	0
alto	3	11	1	1

Preparazione dei dati: *Operazioni sulle variabili*

Conversione di intervalli in numeri

- ❑ Quando nel dataset originale potremmo avere variabili categoriche che rappresentano intervalli di valori numerici.
 - Es: sono le fasce d'età
- ❑ In questi casi si possono creare variabili numeriche:
- ❑ Creazione di due variabili numeriche, una con il limite inferiore dell'intervallo e una con il limite superiore
- ❑ Creazione di una variabile numerica utilizzando un valore di sintesi (media, moda, mediana) di ogni intervallo
 - sempre che sia possibile calcolarlo dai dati di base.

Preparazione dei dati: *Operazioni sulle variabili*

Selezione delle variabili

- ❑ La riduzione delle variabili sembra contrastare con le operazioni di *feature engineering*.
- ❑ Ma mentre le operazioni di feature engineering sono rivolte a creare una serie di variabili che si ritengono importanti.
- ❑ **Le variabili individuate in un problema potrebbero essere svariate decine o anche centinaia e l'utilizzo di tutte le variabili potrebbe avere alcuni effetti indesiderati:**
 - Tempi di training del modello molto lunghi
 - Anche se si deve tenere presente che anche queste operazioni hanno un costo
 - Modelli complessi e difficili da interpretare.
 - Possibilità di overfitting
 - un adattamento eccessivo del modello ai dati di training, con conseguente perdita di generalità e quindi di capacità predittiva.

Preparazione dei dati: *Operazioni sulle variabili*

Selezione delle variabili

Filtri statistici semplici

- ❑ Esclusione delle variabili che hanno un livello di varianza inferiore ad una certa soglia
 - Idea: bassa varianza => bassa capacità predittiva
- ❑ La valutazione di misure legate alla correlazione tra ciascuna delle variabili di input e la variabile obiettivo (per problemi di classificazione o regressione)
 - Idea: il mantenimento delle variabili che abbiano una forte relazione con la variabile target
- ❑ Ricerca di variabili di input fortemente legate tra loro. In tal caso è possibile mantenere una sola variabile per ogni gruppo di variabili correlate.
 - Attenzione fra correlazione e causalità. Si cercano le variabili che sono dipendenti con qualche significato sul dominio

Preparazione dei dati: *Operazioni sulle variabili*

Selezione delle variabili

Metodi iterativi

- ❑ efficaci nella riduzione delle feature, ma lente su dataset molto grandi
- ❑ **Forward selection**
 - creazione di modelli ai quali viene aggiunta progressivamente la feature che porta al più alto miglioramento delle performance predittive.
 - Il processo si ferma quando l'aggiunta di una qualsiasi feature non migliorano più le performance di un modello.
- ❑ **Backward elimination**
 - Si parte dal modello che contiene tutte le k feature e si creano creati altri modelli utilizzando solo $k-1$ feature, tralasciano ogni volta una variabile diversa.
 - Si seleziona l'insieme di $k-1$ variabili che ha generato il modello con il minor incremento

Preparazione dei dati: *Operazioni sulle variabili*

Selezione delle variabili

Random Forest

- ❑ E' un algoritmo di classificazione basato su alberi decisionali
- ❑ L'algoritmo produce come output, oltre alla prediction anche il livello di importanza delle variabili di input
 - Se una variabile compare raramente o ad un livello basso nei vari modelli di insiemi di alberi creati dall'algoritmo, molto probabilmente ha uno scarso impatto sulla capacità predittiva.
 - Se una variabile compare spesso nei vari modelli ed è presente nei livelli più alti degli alberi decisionali, allora avrà una grande importanza dal punto di vista delle performance del modello.

Preparazione dei dati: *Operazioni sulle variabili*

Selezione delle variabili

Principal Component Analysis (PCA)

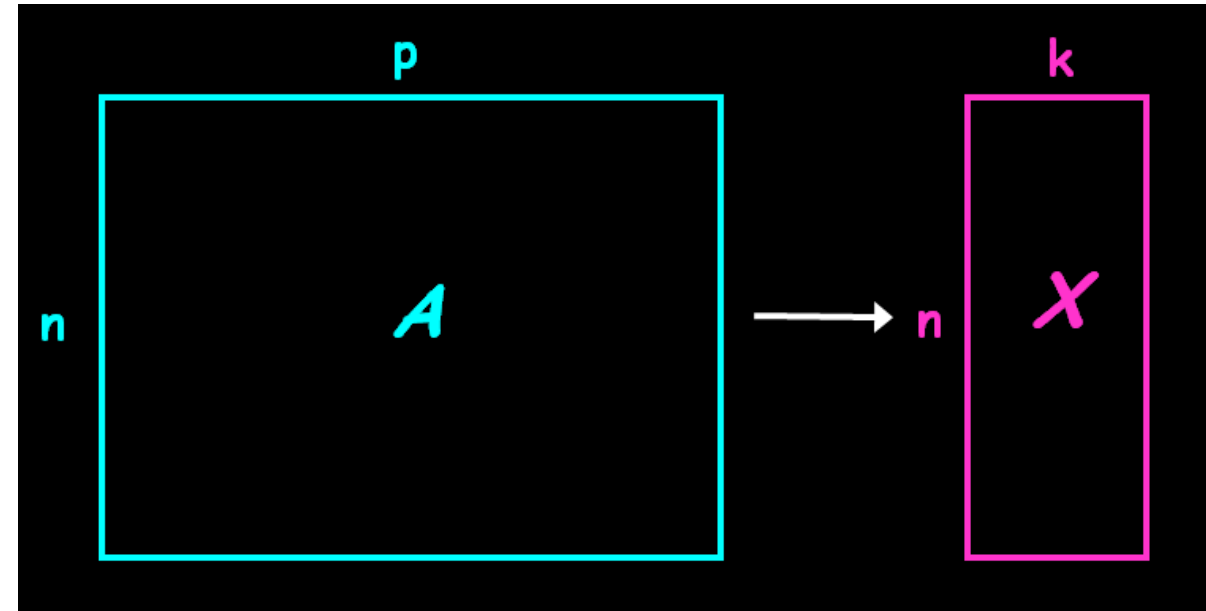
- Un'elevata correlazione tra le variabili è indice di ridondanza nei dati.
- Le variabili più importanti sono quelle che esprimono una varianza più alta
- ❑ La PCA trasforma i dati originali in un nuovo insieme di variabili che conservano però l'informazione contenuta nel dataset originale, ma che non presentano più la ridondanza.
- ❑ In termini più formali si trasformano v variabili quantitative in k (con $k < v$) combinazioni lineari ordinate in base alla variabilità da esse spiegata (in ordine decrescente).
- ❑ *Dal punto di vista matematico la PCA è una trasformazione lineare ortogonale che muta i dati in un nuovo sistema di coordinate facendo in modo che la parte più grande della varianza ricada sulla prima coordinata (prima componente principale), e che via via parti sempre più piccole di varianza siano attribuite alle altre coordinate (seconda, terza, ecc. componenti principali).*

Preparazione dei dati: *Operazioni sulle variabili*

Selezione delle variabili

Principal Component Analysis (PCA)

- probably the most widely-used and well-known of the “standard” multivariate methods
- invented by Pearson (1901) and Hotelling (1933)
- first applied in ecology by Goodall (1954) under the name “factor analysis” (“principal factor analysis” is a synonym of PCA)
- summarization of data with many (p) variables by a smaller set of (k) derived (synthetic, composite) variables.



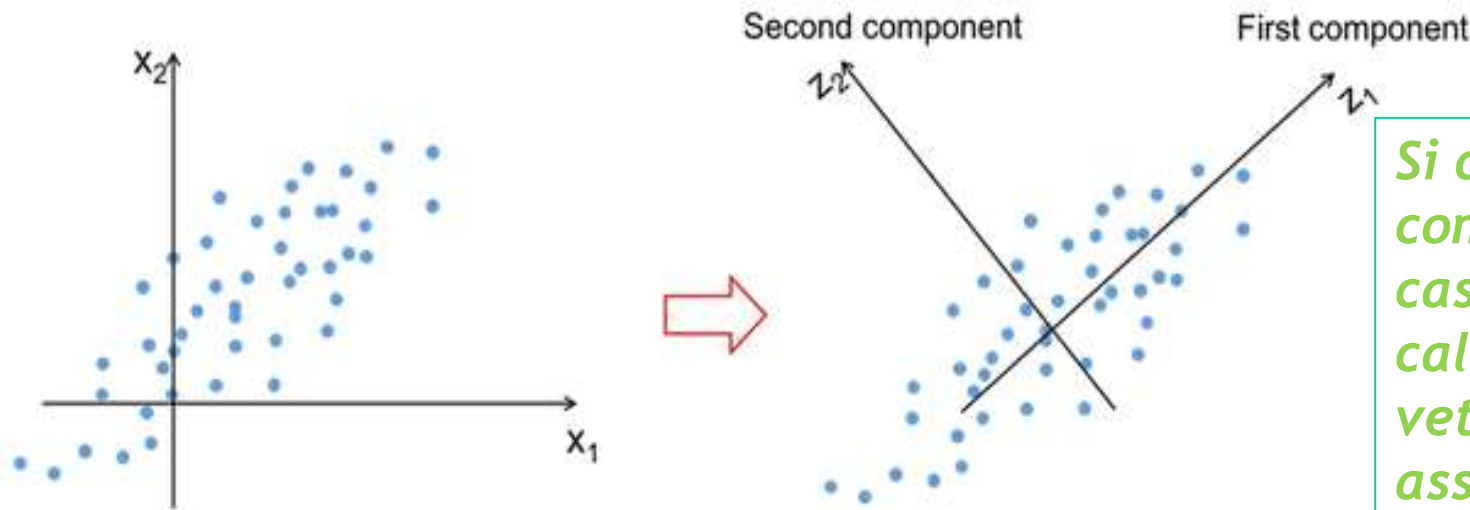
<https://www.mii.lt/zilinskas/uploads/.../lectures/...pca/PCA1.ppt>

Preparazione dei dati: *Operazioni sulle variabili*

Selezione delle variabili

□ *Principal Component Analysis (PCA)*

Se le variabili sono indipendenti, l'applicazione della PCA non è produttiva.



Si dimostra che per determinare le componenti principali di un campione casuale multivariato, è necessario calcolare gli autovalori e gli autovettori della matrice di covarianza associata al campione.

The left graph is the data measured in the original basis. By shifting and rotating (linear transformation), we obtain a new basis spanned by the two principal components (right graph).

©jxchen.net

Preparazione dei dati: *Operazioni sulle variabili*

Selezione delle variabili

Principal Component Analysis (PCA)

- La matrice di covarianza C (*espressa in termini di varianze e covarianze dai valori medi*), è costituita da tutte le covarianze tra le n variabili del dataset standardizzate se espresse in unità di misura diverse
- Si può anche utilizzare una matrice di correlazione R al posto della matrice di covarianza, ottenendo quindi una misura già standardizzata (*espressa in deviazioni standard e un coefficiente di correlazione*)

matrice di correlazione, una matrice quadrata cui termini sono dati da

$$[(X_i - E(X_i))(X_j - E(X_j))].$$

$$\mathbf{R} = \begin{pmatrix} 1 & \rho(X_1, X_2) & \cdots & \rho(X_1, X_n) \\ \rho(X_1, X_2) & 1 & \cdots & \rho(X_2, X_n) \\ \cdots & \cdots & \cdots & \cdots \\ \rho(X_1, X_n) & \rho(X_2, X_n) & \cdots & 1 \end{pmatrix}$$

i termini diagonali (quando i è diverso j) danno le varianze, mentre gli altri danno le covarianze

Preparazione dei dati: *Operazioni sulle variabili*

Selezione delle variabili

□ *Principal Component Analysis (PCA)*

- Si dimostra che per determinare le componenti principali di un campione casuale multivariato, è necessario calcolare gli autovalori e gli autovettori della matrice di covarianza associata al campione.
- coordinates of each object i on the k^{th} principal axis, known as the **scores** on PC k , are computed as

$$z_{ki} = u_{1k}x_{1i} + u_{2k}x_{2i} + \cdots + u_{pk}x_{pi}$$

where Z is the $n \times k$ matrix of **PC scores**, X is the $n \times p$ **centered data matrix** and U is the $p \times k$ **matrix of eigenvectors**.

Preparazione dei dati: *Operazioni sulle variabili*

Selezione delle variabili

- *Principal Component Analysis (PCA)*

- Una prima valutazione del risultato ottenuto calcolando la matrice delle componenti principali P, può essere affrontata calcolando la quota di variabilità mantenuta dai nuovi valori, rispetto ai dati originali:

$$\text{Quota di variabilità mantenuta} = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^v \lambda_i}$$

- Dove k è il numero di componenti principali calcolate, v è il numero di variabili originali e λ è la varianza (che nel caso della componente principale è il corrispondente autovalore).

Preparazione dei dati: *Operazioni sulle variabili*

Selezione delle variabili

- ❑ *Principal Component Analysis (PCA)*

- ❑ Un'altra misura interessante è la correlazione tra le componenti principali e le variabili originali.
- ❑ Essa permette di stabilire da cosa dipendono i valori delle componenti principali.

$$\text{Corr}(p_j, x_i) = \sqrt{\lambda_j} a_{ij}$$

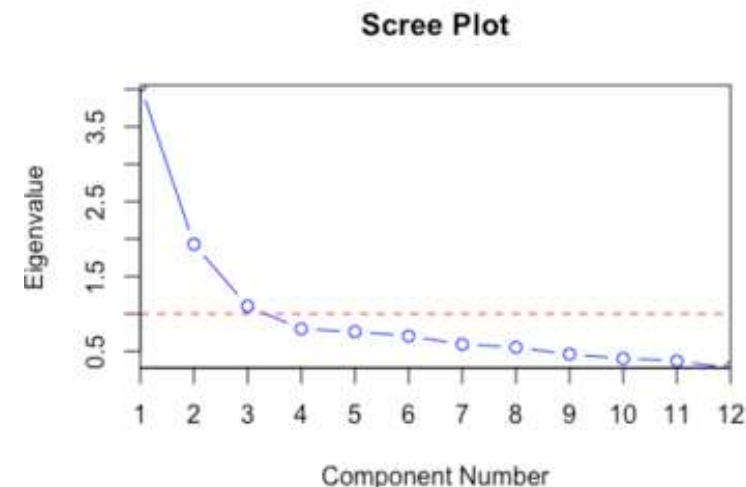
- ❑ Quindi la correlazione dipende dal fattore a_{ij} , ovvero dal peso (*loading*), cioè dall'autovalore, sia per quanto riguarda il segno, sia per quanto riguarda l'ampiezza
- ❑ In base alla variabilità che si vuol conservare, è possibile scegliere le prime n componenti principali da utilizzare come nuovo dataset

Preparazione dei dati: *Operazioni sulle variabili*



Selezione delle variabili

- ❑ *Principal Component Analysis (PCA)*
- ❑ Una tecnica visuale per decidere quante componenti principali mantenere nel nuovo dataset è rappresentata dal **grafico screeplot**, che rappresenta sull'asse orizzontale il numero corrispondente agli autovalori, ordinati in modo decrescente e sull'asse verticale il livello dell'autovalore stesso.
- ❑ Il numero di componenti principali corrisponde al numero di autovalori che vi sono prima che la curva subisca un appiattimento.



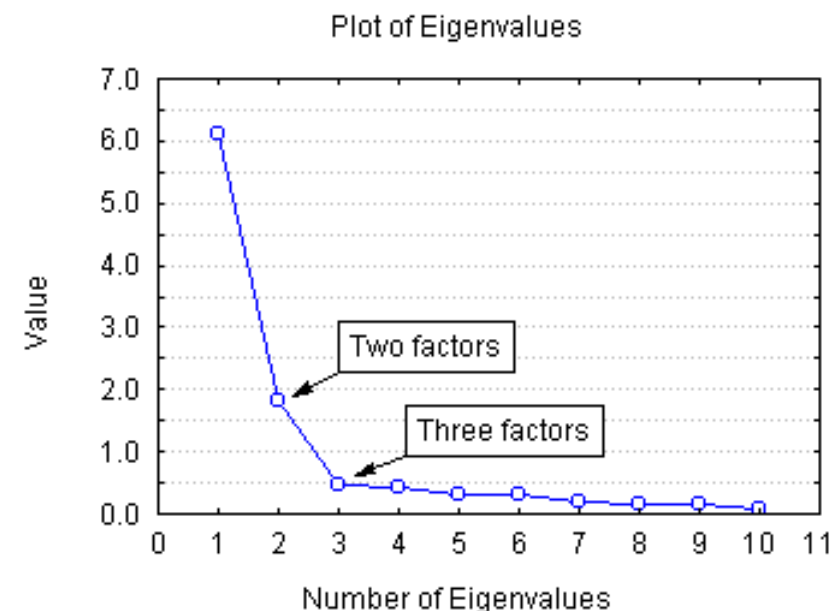
la linea rossa segna il livello 1 degli autovalori e rappresenta una soglia di significatività.

Preparazione dei dati: *Operazioni sulle variabili*

Selezione delle variabili

❑ *Principal Component Analysis (PCA)*

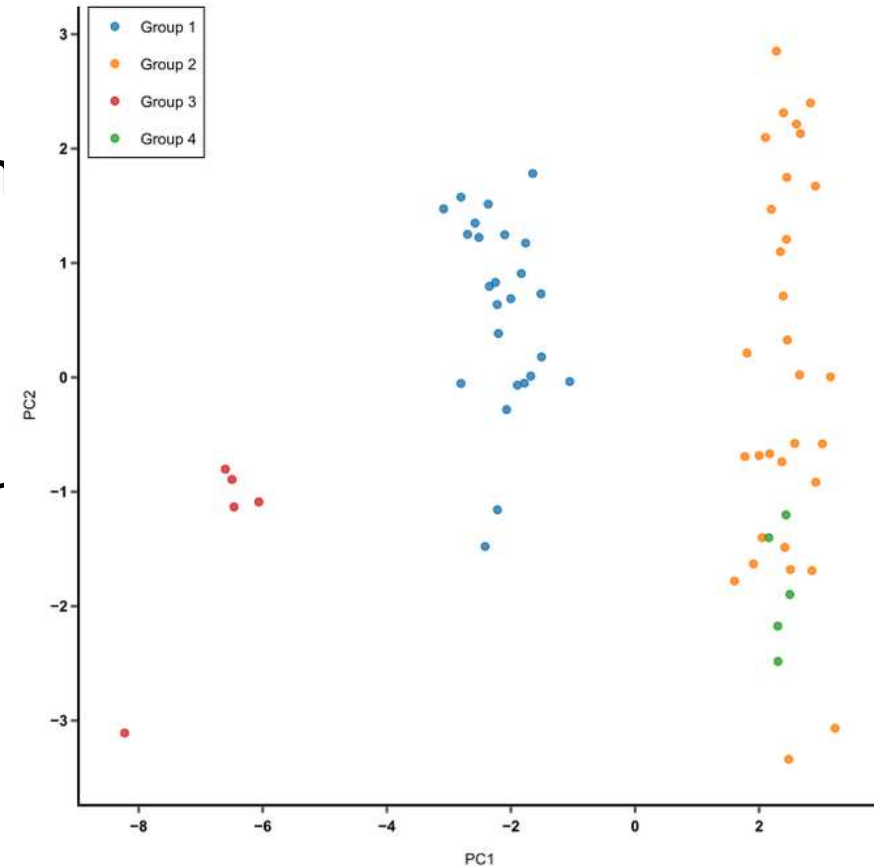
- ❑ Lo scree test (test del «ghiaione») consiste nel trovare il luogo in cui la decisa diminuzione degli autovalori sembra livellata a destra del grafico.
- ❑ A destra di questo punto, presumibilmente, si trova solo "*factorial scree*" - «scree» è il termine geologico che si riferisce ai detriti che si accumulano nella parte inferiore di un pendio roccioso.
- ❑ Pertanto, non deve essere trattenuto più del numero di fattori a sinistra di questo punto.



Preparazione dei dati: *Operazioni sulle variabili*

Selezione delle variabili

- ❑ *Principal Component Analysis (PCA)*
- ❑ Ancora possiamo usare dei **grafici biplot** per vedere graficamente le informazioni su entrambi i campioni e le variabili di una matrice di dati della PCA
- ❑ I biplots sono un tipo di grafico esplorativo utilizzato nelle statistiche, una generalizzazione del semplice diagramma a dispersione a due variabili



Preparazione dei dati: *Operazioni sulle variabili*

Selezione delle variabili

- ❑ *Principal Component Analysis (PCA)*
- ❑ È opportuno notare come la tecnica della PCA sia da utilizzare con le variabili quantitative o qualitative binarie, mentre non è corretto applicarla a variabili categoriche.
- ❑ Per le variabili categoriche è possibile utilizzare una tecnica differente, chiamata *Multiple Correspondence Analysis*.
- ❑ Un procedimento per realizzare la MCA si basa sulla complete disjunctive table, ovvero una matrice dove le righe sono gli individui (osservazioni) e le colonne sono le variabili, o meglio, indicatori che esprimono le diverse categorie presenti nelle variabili
- ❑ In pratica si trasformano le variabili qualitative in una serie di variabili binarie e quindi si può applicare la PCA

Preparazione dei dati: *Operazioni sulle variabili*

Selezione delle variabili

□ *Feature Hashing*

- Utilizzato per ridurre le variabili di un dataset ad un numero predefinito
- Utile nei casi in cui vi sia un'elevata numerosità di feature o addirittura quando non vi è un limite al loro numero.
 - Es.: analisi del testo, nella quale ogni parola di un documento costituisce una singola variabile. Dataset di questo tipo raggiungono facilmente le migliaia di features, ma solo alcune di esse sono valorizzate per un dato documento (non tutti i documenti contengono tutte le possibili parole).
 - La tecnica utilizza una funzione di hashing in grado di produrre un numero intero.
 - Per ogni riga del dataset la funzione è applicata a ciascuna variabile ottenendo per ognuna di esse un valore intero, che è diviso per il numero di feature desiderato.

Nel linguaggio matematico e informatico, l'hash è una funzione non invertibile che mappa una stringa di lunghezza arbitraria in una stringa di lunghezza predefinita

Preparazione dei dati: *Operazioni sulle variabili*

Trattamento dei valori mancanti

- *La costruzione di un dataset che abbia tutte le variabili complete è spesso impossibile, oppure è possibile soltanto selezionando una quantità molto limitata di dati.*

Tipologie di valori mancanti

1. MCAR, *Missing Completely at Random*: i dati di una variabile V sono mancanti indipendentemente sia dalle altre variabili, sia dai valori di V stessa.
2. MAR, *Missing at Random*: i dati di un variabile V sono mancanti indipendentemente dai valori di V. Tuttavia i dati mancanti hanno una dipendenza dalle variabili.
 - Es.: considerando dati demografici, le persone che svolgono un determinato tipo di lavoro potrebbero essere meno propense a dichiarare il loro reddito in un'intervista.
3. NMAR, *Missing not at Random*: i dati mancanti della variabile V dipendono dalla variabile stessa.
 - Es.: le persone con reddito elevato sono meno propense a dichiarare il loro reddito in un'intervista.

Preparazione dei dati: *Operazioni sulle variabili*

Trattamento dei valori mancanti

❑ *Tipologie di valori mancanti*

- ❑ Nel primo caso (MCAR), l'analisi effettuata sul sottoinsieme dei dati che presentano tutti i valori, porta agli stessi risultati che si avrebbero con il dataset intero e completo, visto che i dati sono mancanti in modo puramente casuale.
- ❑ Quindi, nel caso di MCAR, l'eliminazione delle righe incomplete è un'ipotesi percorribile
 - posto che non costituiscano una percentuale troppo grande del dataset
- ❑ Invece l'eliminazione delle righe incomplete nei casi MAR e NMAR significherebbe la riduzione dell'informazione contenuta nel dataset.
 - Es.: nel caso del reddito, se il livello del reddito impedisce ad alcune persone di rivelarlo, ovviamente dobbiamo considerare questo fenomeno.

Preparazione dei dati: *Operazioni sulle variabili*

Trattamento dei valori mancanti

❑ *Tipologie di valori mancanti*


- ❑ L'identificazione del meccanismo per cui i dati di una variabile sono mancanti non è semplice
- ❑ Utilizzo della conoscenza del business e del problema oggetto di analisi, per capire se vi sia una dipendenza dei dati mancanti dalla variabile stessa (NMAR)
- ❑ Creazione di una variabile binaria (0 = dato presente, 1 = dato mancante) e *t-test* o *test chi-quadro* tra la variabile binaria e le altre variabili per stabilire se vi sia qualche dipendenza e stabilire se i dati sono MCAR o MAR
- ❑ Test di Little per cercare dati MCAR.
 - L'ipotesi nulla del test MCAR di Little indica che i dati sono completamente mancanti in modo casuale (MCAR). Si basa sulla funzione asintotica di chi-quadro

Preparazione dei dati: *Operazioni sulle variabili*

Test del chi-quadro

- riguarda il confronto di due percentuali ottenute in un esperimento, allo scopo di verificare se la differenza fra esse è dovuta al caso oppure no, cioè se è «statisticamente significativa».
- Es: confrontare l'efficacia di un nuovo antibiotico con un antibiotico già in uso nella terapia di una malattia del cane

Tabella 1. Dati ottenuti.



Trattamento ↓	Esito →	guariti	non-guariti	totale
xmicina		52 ^a	10 ^b	62
streptomicina		40 ^c	21 ^d	61
totale		92	31	123

Dei 62 cani trattati con xmicina, ne sono guariti 52 (84%)
 $52/62 = 0.84$

Dei 61 cani trattati con streptomicina, ne sono guariti 40 (66%)
 $40/61 = 0.66$

- I dati grezzi indicano che la xmicina è più efficace della streptomicina.
- Però la superiorità di xmicina potrebbe essere dovuta al caso...


http://www.quadernodiepidemiologia.it/epi/assoc/chi_qua.htm

Preparazione dei dati: *Operazioni sulle variabili*

Test del chi-quadro

- Hp che NON esistano differenze nell'efficacia dei due trattamenti.
- Che probabilità c'è di osservare - in uno studio di dimensioni simili a questo - differenze nell'efficacia dei due antibiotici uguali o superiori a quelle osservate?
- *quanto i dati ottenuti si discostano dai dati che «sarebbe lecito attendersi se i trattamenti avessero la stessa efficacia e se i dati fossero influenzati soltanto dalla variazione casuale»?*

Tabella 1. Dati ottenuti.



Treatment ↓	Outcome →	guariti	non-guariti	totale
xmicina		52 ^a	10 ^b	62
streptomicina		40 ^c	21 ^d	61
totale		92	31	123

Dei 62 cani trattati con xmicina, ne sono guariti 52 (84%)
 $52/62 = 0.84$

Dei 61 cani trattati con streptomicina, ne sono guariti 40 (66%)
 $40/61 = 0.66$

Preparazione dei dati: *Operazioni sulle variabili*

Test del chi-quadro

- I dati dimostrano che complessivamente (cioè indipendentemente dal tipo di antibiotico) il trattamento è risultato efficace nel 74.8% (52+40=92 animali su 123 trattati).
- Applicando questa percentuale di successo (74.8%) a ciascuno dei due gruppi di cani in esame si può ricavare

Tabella 2. Dati attesi

Trattamento ↓ Esito →	guariti	non-guariti	<i>totale</i>
xmicina	46.37 a	15.63 b	62.00
streptomicina	45.63 c	15.37 d	61.00
<i>totale</i>	92.00	31.00	123.00

che illustra la situazione che ti saresti aspettato se i due antibiotici avessero avuto la stessa efficacia.

Preparazione dei dati: *Operazioni sulle variabili*

Test del chi-quadro

- Il valore del chi-quadrato quantifica la differenza fra i dati osservati e quelli attesi, ed è la somma delle quattro celle per ciascuna delle quali si calcola il valore della frazione:

$$\frac{(\text{dato osservato} - \text{dato atteso})^2}{\text{dato atteso}}$$

$$\chi^2 = \frac{(52 - 46.37)^2}{46.37} + \frac{(10 - 15.63)^2}{15.63} + \frac{(40 - 45.63)^2}{45.63} + \frac{(21 - 15.37)^2}{15.37} = 5.46$$

Tabella 1. Dati ottenuti.

Trattamento ↓ Esito →	guariti	non-guariti	totale
xmicina	52 a	10 b	62
streptomicina	40 c	21 d	61
totale	92	31	123

Tabella 2. Dati attesi

Trattamento ↓ Esito →	guariti	non-guariti	totale
xmicina	46.37 a	15.63 b	62.00
streptomicina	45.63 c	15.37 d	61.00
totale	92.00	31.00	123.00

Il chi-quadrato aumenta con l'aumentare della differenza dei dati posti a raffronto. Se supera certi valori prefissati, la differenza viene ritenuta significativa.

Preparazione dei dati: *Operazioni sulle variabili*

Test del chi-quadro

- Si deve quindi confrontare il valore ottenuto con la Tabella dei valori di chi-quadrato

Il grado di libertà è uguale a (numero di righe-1)*(numero di colonne-1).

Quindi: $(2-1) * (2-1) = 1$ grado di libertà.



Confrontando il valore chi-quadrato 5.46 con quelli tabulati, esso è >3.841 e <6.635 . Ciò consente di ritenere che la differenza fra i due gruppi sia significativa al livello di probabilità 5% ma non al livello di probabilità 1%.

Tabella dei valori di χ^2

Gradi di libertà	Probabilità	
	5%	1%
1	3.841	6.635
2	5.991	9.210
3	7.815	11.345
4	9.488	13.277
5	11.070	15.086
6	12.592	16.812
7	14.067	18.475
ecc.

La differenza tra animali trattati con xmicina e quelli trattati con streptomicina è **statisticamente significativa al livello di probabilità 5%**.

Preparazione dei dati: *Operazioni sulle variabili*

- ❑ Il livello di **significatività** (o livello α) è una soglia che determina se un risultato di uno studio possa essere considerato statisticamente significativo dopo aver svolto i test statistici pianificati.
 - Il livello di significatività viene posto molto spesso al 5% (o 0,05).
- ❑ ***Rappresenta la probabilità che l'ipotesi nulla possa essere respinta quando è vera.***
 - Es: un livello di significatività dello 0,05 indica un rischio del 5% di concludere che esiste una differenza tra i risultati dello studio e l'ipotesi nulla quando non vi è alcuna differenza effettiva.
- ❑ La probabilità che un risultato sia prodotto dal caso più che da un intervento in studio, se l'ipotesi nulla è vera (cioè, se non vi è una effettiva differenza), è nota come “**valore p**”.
- ❑ Un risultato è quindi statisticamente significativo se porta a un valore **p** uguale o inferiore al livello di significatività dichiarato
 - e quindi non sarà considerato un prodotto del caso.
- ❑ Solitamente, tale risultato viene scritto nella forma $p \leq \alpha$ (pe, $p \leq 0,05$)

Preparazione dei dati: *Operazioni sulle variabili*

Test del chi-quadro

- Si deve quindi confrontare il valore ottenuto con la Tabella dei valori di chi-quadrato

In base ai risultati del test del chi-quadrato, l'affermazione «xmicina è più efficace di streptomicina» ha il 95% di probabilità di essere vera (e quindi ha il 5% di probabilità di essere falsa).



In base ai risultati ottenuti, xmicina è risultata più attiva di streptomicina ($P < 0.05$)» dove il valore P indica la probabilità di respingere una ipotesi zero vera.

Tabella dei valori di χ^2

Gradi di libertà	Probabilità	
	5%	1%
1	3.841	6.635
2	5.991	9.210
3	7.815	11.345
4	9.488	13.277
5	11.070	15.086
6	12.592	16.812
7	14.067	18.475
ecc.

Preparazione dei dati: *Operazioni sulle variabili*

Gestione dei valori mancanti

Tipologie di valori mancanti

- ❑ Eliminazione della variabile, nel caso in cui i valori mancanti superino una certa percentuale sul totale (per esempio, 50%).
- ❑ Eliminazione della riga dal dataset.
 - *Questa tecnica non ha effetti negativi nel caso di dati MCAR.*
- ❑ Sostituzione del valore nullo con la media, con la moda o con un altro valore costante.
 - *Questo approccio riduce la variabilità e indebolisce la covarianza e la correlazione tra i dati, visto che non ne tiene assolutamente conto.*
- ❑ Sostituzione del valore nullo con un valore che non alteri la deviazione standard.
 - *Non si riduce la variabilità, ma ci sono sempre effetti negativi sulla covarianza e sulla correlazione.*
- ❑ Riempimento dei valori mancanti in base a tecniche di correlazione o regressione

Preparazione dei dati: *Operazioni sulle variabili*

Gestione dei valori mancanti

Tipologie di valori mancanti

- ❑ Tutte le tecniche di sostituzione presentano in misura diversa il medesimo problema:
- ❑ ***Il valore, o i valori con cui si sostituiscono i dati mancanti non corrispondono alla realtà e ciò può introdurre distorsioni nelle analisi***

Preparazione dei dati: *Operazioni sulle variabili*

Gli outliers

- ❑ sono valori che, secondo un dato criterio, possono essere definiti anomali, distanti dalle altre osservazioni
- ❑ Possono avere un effetto anche importante sul modello e sulla sua capacità predittiva.
 - Dipende dagli algoritmi utilizzati
- ❑ Nei problemi di ***anomaly detection*** sono lo scopo del modello predittivo
- ❑ Tecniche di identificazione degli outliers
 - Tecniche univariate basate sui valori estremi.
 - Tecniche multivariate basate sui valori estremi.
 - Modelli lineari.
 - Metodi di prossimità

Preparazione dei dati: *Operazioni sulle variabili*

Gli outliers

- Tecniche univariate basate sui valori estremi.
- Si analizza una variabile per volta: valori troppo piccoli o troppo grandi sono considerati outliers.
- Una tecnica semplice consiste nel calcolo dello *z-score*, ovvero del numero di deviazioni standard rispetto al quale un'osservazione si discosta dalla media

$$zscore = \frac{(x - \bar{x})}{\sigma}$$

- Stabilendo un valore massimo di standard deviation (convenzionalmente il valore è 2), si può stabilire che un'osservazione è anomala se il suo *z-score* supera la soglia

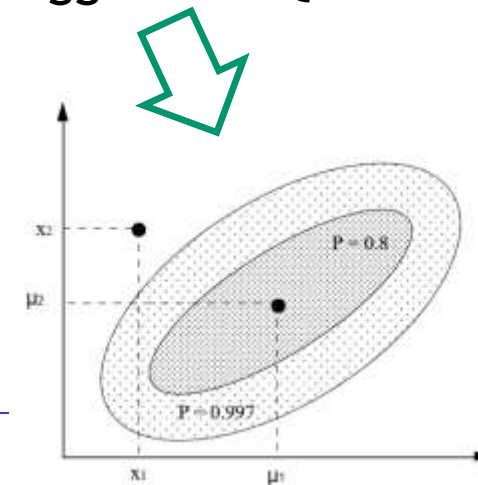
$$m_zscore = 0.6745 \frac{(x_i - \bar{x})}{MAD}$$

Preparazione dei dati: *Operazioni sulle variabili*

Gli outliers

- Tecniche multivariate basate sui valori estremi.
- *In un dataset con più variabili occorre determinare gli outliers con tecniche che tengano conto dell'intero spazio dimensionale*
- In alcuni algoritmi di clustering si usa il calcolo della *distanza di Mahalanobis* che può essere derivata dalla distribuzione normale multivariata
- Per la determinazione degli outliers, si definisce una grandezza Q , rappresentata dal quantile della distribuzione chi quadro con n gradi di libertà (per esempio, il quantile 97.5%) e dichiarare outliers tutti i punti che hanno il quadrato della distanza di Mahalanobis maggiore di Q .

• *Distanza di Mahalanobis* è basata sulle correlazioni tra variabili attraverso le quali differenti pattern possono essere identificati ed analizzati. Differisce dalla distanza euclidea proprio in quanto tiene conto di queste correlazioni.



Preparazione dei dati: *Operazioni sulle variabili*

Gli outliers

– Modelli lineari.

- Attraverso i modelli lineari i dati sono rappresentati in uno spazio dimensionale inferiore, utilizzando la correlazione e le combinazioni lineari tra le variabili
- Uno dei metodi che fa parte di questa categoria è basato sul calcolo della PCA.
- L'idea che sta dietro all'utilizzo della PCA riguarda la possibilità che i dati anomali si adattino molto meno dei dati normali al nuovo sotto spazio formato dalle prime n componenti principali, quindi la perdita di informazione sarà, per gli outlier, maggiore.

Preparazione dei dati: *Operazioni sulle variabili*

Gli outliers

- Metodi di prossimità.
- Nei metodi di prossimità un punto è considerato outlier se è isolato rispetto ad altri punti
- Es: Metodi basati su cluster.
- Utilizzano un algoritmo di clustering per raggruppare i punti. Successivamente sono calcolate le distanze tra ogni punto ed il centro di ciascun cluster. I punti sono ordinati per distanza decrescente e una certa percentuale dei punti più distanti è considerata outlier.



Preparazione dei dati: *Operazioni sulle variabili*

Gli outliers

– Trattamento degli outlier.

- ❑ Eliminazione dell'intero caso (la riga del dataset) che contiene l'outlier.
- ❑ Sostituzione con un valore ritenuto normale, che potrebbe essere calcolato con le stesse metodologie applicate per i valori mancanti.
- ❑ Non fare nulla se la sostituzione peggiora le caratteristiche del modello ottenuto senza fare un'analisi degli outliers

Preparazione dei dati: *Operazioni sulle variabili*

Classi sbilanciate

E' il caso in cui una delle classi presenta un numero di elementi estremamente basso.

- L'effetto negativo dello sbilanciamento consiste spesso in prediction il cui risultato è sempre la classe maggioritaria;
- tuttavia quasi sempre si è interessati alla previsione della classe minoritaria, che quindi non sarebbe mai prodotta come output.

A livello di modellazione, ciò che si può fare riguarda il test di diversi algoritmi per capire se almeno uno di essi sia in grado di cogliere il pattern sottostante alla classe minoritaria.

Si può anche intervenire nella fase di preparazione dei dati con tecniche di

- ❑ *Oversampling*
- ❑ *Undersampling*
- ❑ *Synthetic Samples*

Preparazione dei dati: *Operazioni sulle variabili*

Classi sbilanciate

❑ *Oversampling*

- I punti relativi alla classe minoritaria sono inclusi nel dataset più volte in modo da aumentare la loro presenza in rapporto le altre classi (campionamento con ripetizione)
- può portare ad un eccessivo adattamento dell'algoritmo ai dati di training (*overfitting*), esprimendo una capacità predittiva bassa sui nuovi dati.

❑ *Undersampling*

- Nel caso di campioni molto grandi è possibile lasciare inalterati i dati della classe minoritaria e diminuire il numero di dati delle altre classi
- potrebbe causare perdita di informazione per quanto riguarda la classe maggioritaria

❑ *Synthetic Samples*

- creazione di valori "sintetici", ovvero creati ad hoc dall'analista tramite algoritmi

Preparazione dei dati: *Operazioni sulle variabili*

Errori comuni nella preparazione dei dati

- ✓ Utilizzo di variabili anacronistiche.
- Un errore che spesso si compie consiste nell'utilizzo di variabili che contengono in realtà delle informazioni sulla variabile target, che invece dovrebbero essere sconosciute nell'istante temporale in cui si effettua la prediction.
- La preparazione dei dati e il training di un modello predittivo avvengono su un set di dati del passato poiché, in particolare per i modelli supervisionati, tale dataset deve contenere anche la variabile target.
- Tuttavia ci deve essere un gap temporale tra le variabili di input e la variabile target, che si realizza considerando gli input fino ad un certo istante $T1$ ed i valori del target ad un istante $T2 > T1$.

Preparazione dei dati: *Operazioni sulle variabili*

Errori comuni nella preparazione dei dati

- ✓ Effettuare la modellazione su un campione troppo piccolo.
- Nei campioni troppo piccoli alcune variabili, in particolare quelle che hanno numerosi possibili valori, saranno sicuramente rappresentate in modo inappropriato.
- Per garantire un apprendimento adeguato al modello il dataset deve essere sufficientemente grande.
- *Si utilizzano regole empiriche e conoscenze sul dominio*
- ✓ Scelta errata dei casi su cui effettuare il training e il testing del modello.
- L'identificazione in modo non casuale degli elementi del dataset di training e di test dell'algoritmo porta a sicure distorsioni nei risultati.

Preparazione dei dati: *Operazioni sulle variabili*

Errori comuni nella preparazione dei dati

- ✓ Selezione delle variabili.
- ❑ Questo processo può introdurre delle distorsioni che minano l'efficacia predittiva dei modelli.
- ❑ Vi sono casi in cui l'analista ha il desiderio conscio o inconscio di provare una certa ipotesi e, per questo, la selezione di variabili è effettuata in modo in tale da comprovare l'ipotesi.
- ❑ Anche senza una precisa volontà, si può ricadere in una selezione errata delle variabili, se essa è effettuata in modo puramente soggettivo senza avere la conoscenza del business o della specifica problematica e senza utilizzare le adeguate tecniche statistiche.

Preparazione dei dati: *Operazioni sulle variabili*

Errori comuni nella preparazione dei dati

- ✓ Non tenere in considerazione gli outliers e la gestione dei valori mancanti
- ❑ In alcuni casi la presenza di valori estremi può portare a risultati completamente inaccurati.
- ❑ Per questo motivo, l'identificazione e il trattamento degli outlier deve essere preso in considerazione in un processo di analisi predittiva.
- ❑ In modo simile anche trattamento errato di valori mancanti causa distorsioni nel modello