**UNIVERSITÀ POLITECNICA DELLE MARCHE**
**FACOLTÀ DI INGEGNERIA**
Dipartimento di Ingegneria dell'Informazione
Corso di Laurea in Ingegneria Informatica e dell'Automazione


**TESI DI LAUREA**

# Progettazione e implementazione di una campagna di data analytics relativa agli incidenti stradali in una metropoli

# Design and implementation of a data analytics campaign related to traffic accidents in a metropolis

Relatore

Prof. Domenico Ursino

Candidato

Riccardo Angelini

Correlatore

Luca Virgili

**ANNO ACCADEMICO 2023-2024**

*La scienza è fatta di dati come una casa è fatta di pietre.*
*Ma un ammasso di dati non è scienza più di quanto un mucchio di pietre sia una vera casa.*

Henry Poincaré

**Sommario**

In questi ultimi anni la Big Data Analytics ha assunto una rilevanza sempre maggiore dato l'enorme volume di dati raccolto dalle imprese. In un contesto aziendale, i dati sono sottoposti a una serie di processi per generare informazioni utili al fine di migliorare il processo decisionale. Tuttavia, anche al di fuori dell'azienda, la loro importanza non è da sottovalutare. In questa tesi, sono stati analizzati i dati relativi agli incidenti stradali nella metropoli di New York dal 2014 sino ad oggi. Dopo aver approfondito il dataset iniziale, sono state svolte operazioni di ETL tramite il software Power BI. Successivamente, si è passati all'Exploratory Data Analysis, per estarre modelli o pattern dai dati. Attraverso l'attività di Data Visualization, sono stati realizzati dei report per informare il lettore dei raccolti dati. Infine, è stata utilizzata la tecnica del Clustering per identificare gruppi omogenei e tendenze all'interno dei dati.

**Keyword**: Big Data Analytics, Extract Trasform and Load, Power BI, Exploratory Data Analysis, Pattern, Data Visualization, Data Mining, Clustering

# Indice

# Elenco delle figure

# Elenco delle tabelle

# Introduzione

Al giorno d'oggi l'importanza dei dati ha raggiunto livelli senza precedenti. Secondo il giornalista Vincenzo Cosenza, i dati sono divenuti il quarto fattore produttivo, dopo i classici terra, lavoro e capitale. La raccolta e l'analisi dei dati ormai rappresenta un bisogno per l'azienda; disporre di informazioni affidabili può tradursi in un vantaggio competitivo significativo. Tuttavia, una delle principali problematiche è rappresentata dalla grande quantità di dati che non possono essere convertiti in informazioni, conosciuti con il termine di rumore. Così come un mucchio di pietre non costituisce una casa senza un progetto, il processo di analisi dei dati richiede metodo e rigore affinchè i dati possano ottenere un valore effettivo.

A ciò si aggiunge lo sviluppo di tecnologie alla base dell'Intelligenza Artificiale (IA), come il Machine Learning e il Data Mining, che permettono di migliorare e di automatizzare il processo di gestione e di analisi dei dati. Lo scienziato statunitense Arthur Lee Samuel, pioniere dell'IA, affermò che il Machine Learning è un campo di studi che dà ai computer la capacità di apprendere senza essere programmati. Dunque, i nuovi algoritmi di apprendimento automatico permettono ai computer di imparare dai dati e prendere decisioni formate. Il termine "Data Mining", invece, si riferisce al processo di scoperta di pattern o modelli nascosti all'interno di una grande mole di dati. Identificare relazioni significative o tendenze, tramite tecniche di Data Mining, potrebbe rivelare preziose informazioni capaci di ottimizzare il processo di decision-making.

Dunque, il connubio tra Big Data e Intelligenza Artificale sta rivoluzionando l'azienda e la sua visione. Come ha dichiarato Andrew Ng, co-fondatore di Google Brain: "L'intelligenza artificiale è la nuova elettricità. Così come l'elettricità ha trasformato quasi tutto 100 anni fa, oggi ho difficoltà a pensare a un'industria che non verrà trasformata dall'IA nei prossimi anni". La capacità di estrarre valore dai dati è oramai uno strumento di innovazione per imporsi nel mercato globale. Tuttavia, è bene sapere che i campi di applicazione di queste nuove tecnologie sono numerosi, dai più complessi ai più semplici. Un esempio concreto è l'uso dell'IA nella gestione del traffico urbano. Infatti, attraverso sensori e telecamere distribuite in tutta la città, è possibile raccogliere informazioni sulla circolazione stradale dei veicoli. A questo punto, algoritmi di Machine Learning e analisi predittive propongono soluzioni a sostegno delle varie problematiche stradali, come congestionamenti del traffico, aree ad alto rischio di incidente e inquinamento atmosferico dovuto agli scarichi. Di conseguenza, queste nuove tecnologie apportano benefici non solo alla crescita aziendale, migliorando l'efficienza e i ricavi dei processi produttivi, ma possono contribuire a migliorare la qualità della vita attraverso nuove applicazioni. L'IA e i Big Data stanno aprendo nuove frontiere ad un futuro più innovativo.

In questo elaborato vengono vengono analizzati i dati relativi agli incidenti nella città

metropolitana di New York. Ogni riga del dataset rappresenta un evento di arresto anomalo segnalato dalla polizia. Inizialmente viene approfondito il dataset Motor Vehicle Collisions-Crashes, esaminandone le colonne e i metadati. Successivamente, vengono eseguite le tre operazioni che compongono l'ETL, tramite il software Power BI, in cui i dati vengono estratti, puliti, trasformati ed, infine, caricati. Dopodichè, viene esaminato il processo di Exploratory Data Analysis (EDA), con un focus particolare sulle tecniche di data Visualization sfruttate per la creazione dei report. Nella parte finale, viene presentata la tecnica di Data Mining del Clustering svolta in maniera autonoma da Power BI. Questa permette di suddividere i dati in gruppi omogenei al fine di trovare pattern nascosti o identificare comportamenti anomali.

La presente tesi è composta da sei capitoli strutturati come di seguito specificato:

- Nel Capitolo 1 sarà presentata la disciplina della Data Analytics, accompagnata da una introduzione al mondo dei Big Data. Dopo aver esaminato nel dettaglio le 4 categorie di Data Analytics, verrà descritto l'intero ciclo di Big Data Analytics, composto da 9 stadi.

- Nel Capitolo 2 saranno illustrate le varie tipologie di dati utilizzati in un ambiente di Big Data. Infine, verrà esposto il dataset Motor Vehicle Collisions-Crashes utilizzato per la campagna di Data Analytics, approfondendo nel dettaglio la sua provenienza e i metadati.

- Nel Capitolo 3, inizialmente, vengono descritte nel dettaglio le tre fasi che compongono l'attività di ETL. Successivamente, dopo aver presentato l'ecosistema Power BI, verranno approfondite le operazioni di Extract, Transform e Load eseguite sul dataset.

- Nel Capitolo 4 verranno presentati il concetto di Exploratory Data Analysis e, in particolare, gli strumenti di Data Visualization impiegati per la creazione di report. Infine, analizzeremo nel dettaglio le operazioni di EDA svolte sul dataset e i report realizzati.

- Nel Capitolo 5, inizialmente, viene introdotto il concetto di pattern e di Data Mining. Dopo aver approfondito alcune tecniche di Data Mining, verrà descritta nel dettaglio la tecnica del clustering eseguita su Power BI.

- Nel Capitolo 6 verrà discusso il lavoro svolto con delle considerazioni pratiche sull'attività di analisi. Infini saranno delineati alcuni possibili sviluppi futuri.

CAPITOLO 1

# Introduzione alla Data Analytics

*In questo primo capitolo verrà presentata la disciplina della Data Analytics. Si partirà con una breve introduzione al mondo dei Big Data, illustrando il modello delle 5V che ne definisce le caratteristiche distintive. Successivamente, verrà introdotta la Data Analytics, sottolineando la differenza con la DataAnalysis e fornendo una descrizione dettagliata delle 4 categorie di Data Analytics. Inoltre, verrà esaminato l'intero ciclo di Big Data Analytics, riportando nel dettaglio i 9 stadi da cui è formato.*

## 1.1 Big Data e dataset

"Siamo nell'epoca del Data as a Product, quindi il dato è un prodotto economico, è la nuova moneta. I pagamenti sono le condivisioni di dati e non è un caso se le aziende più capitalizzate al mondo siano quelle che li trattano", Nel futuro, dunque, vincerà chi saprà raccogliere i dati, leggerli e analizzarli, intendendoli come asset a lungo termine. Si scrive Big Data e si legge come una delle evoluzioni più profonde e pervasive del mondo digitale; può essere visto come una conseguenza a un'evoluzione massiva degli usi e delle abitudini della gente. Ogni volta che usiamo un computer, accendiamo lo smartphone o apriamo una app sul tablet, sempre e comunque lasciamo una nostra impronta digitale fatta di dati. Parlare di Big Data (Figura 1.1) non vuol dire parlare soltanto di grandi moli di dati; la trasformazione in atto è più profonda. Cambia il processo di raccolta e gestione dei dati, si evolvono le tecnologie a supporto del ciclo di vita del dato e si sviluppano nuove competenze per la valorizzazione del dato. Dunque, i progressi nella scienza computazionale permettono di far fronte a dataset (gruppi di dati correlati) che evolvono anch'essi in termini di complessità e grandezza. I dati presenti in un ambiente di Big Data generalmente si accumulano o e si ammassano all'interno dell'impresa. I risultati ottenuti attraverso l'elaborazione dei Big Data possono portare a una grande capacità di analisi e a grandi benefici, quali:

- l'ottimizzazione delle attività operative;
- l'estrazione di informazioni strategiche;
- l'identificazione di nuovi mercati;
- le predizioni accurate;
- la ricerca di difetti o frodi;
- registrazioni più dettagliate;

- il miglioramento delle attività di decision-making;
- scoperte di tipo scientifico.


**Figura 1.1:** Introduzione al mondo dei Big Data

### 1.1.1 5V dei Big Data

Nel 2001, Doug Laney, allora vicepresidente e Service Director dell'azienda Meta Group, definì le caratteristiche dei Big Data secondo il modello delle 3V: Volume, Varietà, Velocità. Ad oggi il paradigma di Laney si è ampliato sino ad arrivare a 5V, arricchendosi di altre due componenti quali, Veridicità e Variabilità (Figura 1.2). Questi cinque principi continuano a essere fondamentali nel definire la natura distintiva dei Big Data. Nelle prossime sottosezioni daremo uno sguardo più approfondito a ciascuno di essi.

### 1.1.2 Volume

L'importanza del volume dei dati è predominante per il successo di un'azienda, maggiore sarà il volume e maggiori saranno le informazioni che concorrono alla creazione di valore per l'azienda stessa. La quantità di dati trattati cresce in maniera esponenziale, ciò richiede un notevole investimento in infrastrutture capaci di archiviare grandi quantità di dati, nonché capacità di elaborazione sempre più elevate.

### 1.1.3 Velocità

Negli ambienti dei Big Data, i dati possono arrivare ad alta velocità, e dataset enormi possono accumularsi entro periodi di tempi molto piccoli . Il concetto di velocità è correlato al processo di: raccolta, elaborazione e fruizione dell'informazione ottenuta. Al termine di questo processo il dato diviene un asset dell'azienda che può essere monetizzato e la velocità è intesa come "tempo di vita" del dato.

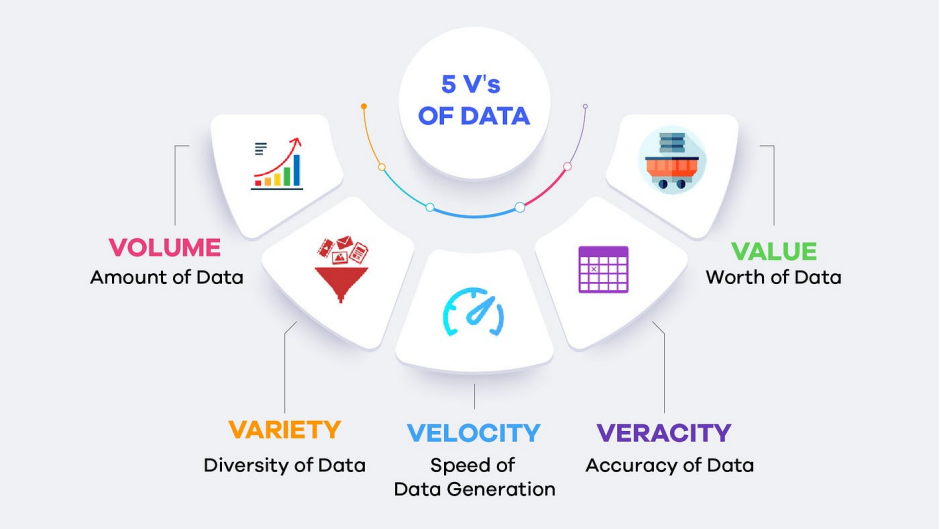


**Figura 1.2:** Illustrazione delle 5V dei Big Data

### 1.1.4 Varietà

Una stessa tipologia di dato può assumere una grande varietà di formati che devono essere supportati dalle soluzioni di Big Data.

Le fonti di dati possono essere sia interne che esterne; per far fronte a questa eterogeneità dei dati all'interno dell'azienda, la varietà di traduce in una sfida in termini di integrazione, trasformazione, elaborazione e memorizzazione dei dati.

### 1.1.5 Veracità

"Bad data is worse than no data".

La veracità si riferisce alla qualità, integrità e accuratezza dei dati raccolti. L'inattendibilità di quest'ultimi porterebbe ad un'analisi inutile; a tal proposito si affiancano attività di data processing per effettuare pulizia e rimuovere i rumori.

Dati che non possono essere convertiti in informazioni prendono il nome di rumore e, di fatto, sono senza valore; la presenza di questi dipende dalla sorgente, oltre che dal tipo di dato.

### 1.1.6 Valore

Abbiamo già detto che, una volta convertito il dato in informazione, questa rappresenta un asset per l'azienda che può essere venduto come un bene o utilizzato nei processi di supporto alle decisioni.

Il valore che assume come generatore di profitto dipende dalle altre caratteristiche, in particolare:

- maggiore è la veracità del dato e superiore sarà il suo valore;
- maggiore è il tempo del processo di trasformazione del dato e minore sarà il suo valore.

## 1.2 Data Analytics e Data Analysis

La Data Analytics è un termine ampio che racchiude la gestione dell'intero ciclo di vita dei dati, ossia una serie di soluzioni per la raccolta, l'elaborazione, la validazione e l'analisi dei

dati al fine di identificare informazioni utili a supporto delle strategie aziendali. L'avvento dell'Intelligenza Artificiale (IA), con le sue capacità di apprendimento automatico, permettono di analizzare ed interpretare i Big Data in modo più rapido ed efficiente; dunque, l'evoluzione di questi metodi di analisi permettono di sfruttare tecnologie distribuite, altamente scalabili per effettuare data Analysis.

La Data Analysis è una parte del ciclo di vita della Big Data Analytics che si occupa di esaminare grandi quantità di dati grezzi e non strutturati al fine di estrarre informazioni. I risultati prodotti potrebbero far emergere pattern o relazioni ne consegue un arricchimento de dati a disposizione e un notevole potenziamento nel processo di decision-making aziendale.

Ci sono quattro categorie generali di Data Analytics che illustreremo nella prossima sezione.

## 1.3 Categorie di Data Analytics

Le categorie di Data Analytics si differenziano sulla base dei risultati ottenuti e delle diverse tecniche e degli algoritmi di analisi adottati.

In figura 1.3 figura si può notare come la complessità ed il costo dei vari livelli di analisi cresce in base al valore dei risultati generati. Nelle prossime sottosezioni esaminiamo più in dettagli tali categorie.

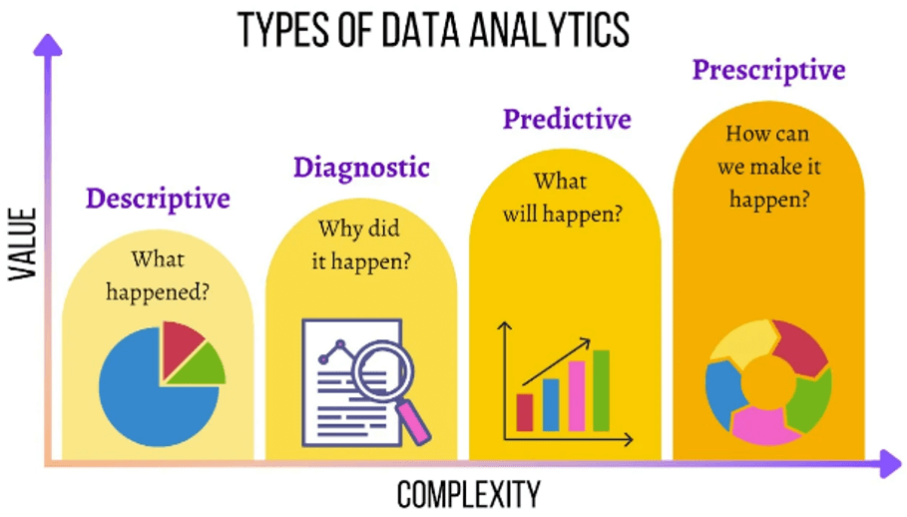


**Figura 1.3:** Tipologie di Data Analytics in relazione alla complessità e al valore

### 1.3.1 Descriptive Analytics

Da come ci suggerisce il nome, questa prima categoria si occupa di sintetizzare e descrivere i dati grezzi, contestualizzandoli, così da renderli interpretabili per l'essere umano.

In particolare, si analizzano eventi passati formulando domande come:

- Quale è stato il volume delle vendite negli ultimi 12 mesi?
- Quale è stato il fatturato degli scorsi anni?

La maggior parte dei risultati prodotti tramite la Data Analytics deriva da questa categoria; tuttavia, l'analisi descrittiva fornisce risultati di più basso valore in quanto, per il suo svolgimento, richiede una competenza ed un insieme di skill relativamente semplice.

I sistemi di dashboard e di reporting sono solitamente utilizzati per comunicare quanto prodotto da questa tipologia di analisi.

### 1.3.2 Diagnostic Analytics

L'analisi diagnostica ha lo scopo di capire l'origine di determinati eventi e la causa scatenante di certi comportamenti; di conseguenza le domande che si pongono gli analisti è sempre sul perché determinati eventi avvengono.

I risultati dell'analisi diagnostica vengono visti tramite tool di visualizzazione interattiva ed hanno un valore maggiore rispetto a quelli dell'analisi descrittiva, poiché questa tipologia di analisi richiede un insieme di skill e di query più avanzate. Oltre a ciò, questa categoria di analisi richiede una collezione di dati da sorgenti multiple e in una struttura che consenta di effettuare analisi di tipo drill-down o roll-up.

### 1.3.3 Predictive Analytics

L'analisi predittiva è il processo di utilizzo dei dati, algoritmi statistici e tecniche di machine learning per prevedere risultati futuri. L'utilizzo di dati storici o di eventi passati consente di allenare i modelli con i quali è possibile creare delle relazioni con le informazioni a disposizione; ciò pone le basi per generare delle predizioni.

Questo processo implica la gestione di ampi dataset che incorporano dati provenienti sia da fonti esterne che interne. Come evidente, l'analisi predittiva richiede un livello di competenza notevolmente avanzato, poiché i risultati ottenuti hanno un valore superiore rispetto alle analisi effettuate in precedenza.

### 1.3.4 Prescriptive Analytics

L'analisi prescrittiva fornisce suggerimenti in merito alle azioni migliori da intraprendere per raggiungere gli obiettivi di business; alla base dell'analisi prescrittiva vi sono i risultati dell'analisi predittiva.

Le domande che si pongono di frequente in questa analisi sono le seguenti:

- Tra questi tre farmaci, quale fornisce i migliori risultati?
- Qual è il momento migliore per vendere un certo stock?

I risultati derivanti da quest'analisi hanno, sicuramente, un impatto maggiore all'interno e all'esterno dell'azienda, in quanto, potrebbero tradursi in un vantaggio competitivo nella situazione di mercato o nella mitigazione di un rischio.

## 1.4 Big Data nell'azienda

L'azienda opera come un sistema a strati basato su una struttura piramidale; al vertice si trova il livello strategico formato dai dirigenti, segue il livello tattico o manageriale, che pilota l'organizzazione secondo le strategie aziendali; alla base della piramide si trova il livello operativo che esegue le operazioni quotidiane.

I Big Data hanno collegamenti con la struttura aziendale ad ogni livello organizzativo; anch'essi vengono rappresentati secondo una struttura piramidale che descrive l'entrata del dato grezzo in azienda e le successive trasformazioni sino ad arrivare alla cima.

Nel primo strato operativo si esamina cosa sta succedendo nell'azienda e viene generato un contesto con il quale si converte il dato in informazione aggiungendo del valore.

Nello strato manageriale si esamina l'informazione secondo le lenti della performance aziendale concentrandosi su come condurre le attività; qui viene dato un significato all'informazione.

A livello strategico, la conoscenza acquisita fino a questo punto può evolvere in saggezza, sfruttando le intuizioni umane. Questa trasformazione sarà particolarmente influente nel processo decisionale per la definizione della strategia aziendale. Tale informazione può essere ulteriormente arricchita per rispondere a domande cruciali riguardanti il motivo per cui l'azienda sta operando al livello attuale.

## 1.5 Ciclo di vita della Big Data Analytics

Il ciclo di vita della Big Data analytics può essere suddiviso in nove fasi, che vedremo in dettaglio nel seguito.

### 1.5.1 Business Case Evaluation

Il primo stadio di ogni ciclo di vita di Big Data Analytics prevede la definizione di un business case che esponga chiaramente le motivazioni e gli obiettivi per effettuare l'analisi; ciò aiuta a determinare le risorse di business necessarie per l'espletamento delle attività da affrontare.

La prima fase, inoltre, serve a definire se il problema da affrontare è inerente ai Big Data, ossia bisogna considerare se la sfida in questione si allinea ad una o più delle caratteristiche dei Big Data quali la vastità dei dati (volume), la rapidità con cui vengono generati e richiedono analisi (velocità) e la diversità dei tipi di informazioni (varietà).

### 1.5.2 Data Identification

Il secondo stadio è dedito all'identificazione dei dataset e delle fonti necessarie per il progetto di analisi.

Maggiore è la diversità delle fonti e maggiore sarà la probabilità di scoprire pattern o correlazioni inaspettate.

Soprattutto quando gli obiettivi di ricerca non sono definiti chiaramente, è vantaggioso cercare il maggior numero possibile di sorgenti di dati correlate. I dataset richiesti e le loro sorgenti possono essere interni o esterni all'impresa, in particolare:

- Per quanto riguarda le fonti interne, si procede compilando un elenco dei dataset disponibili provenienti da risorse interne, che viene quindi confrontato con un insieme predefinito di dataset.
- Per le fonti esterne si procede compilando un elenco di possibili fornitori di dati di terze parti, come i mercati dei dati o i dataset aperti al pubblico.

### 1.5.3 Data Acquisition and Filtering

Una volta identificate le sorgenti dei dati si può passare al terzo stadio che prevede l'acquisizione ed il successivo filtraggio di tutti i dati corrotti o che non siano utili agli obiettivi fissati.

Solitamente i dati non strutturati provenienti da fonti esterne sono irrilevanti (rumore), mentre quelli che contengono record con valori mancanti o senza senso sono da definire "corrotti".

Prima di procedere con l'operazione di filtraggio è usuale fare una copia di backup del dataset originale, in quanto i dati successivamente filtrati potrebbero servire per un'analisi differente.

Sia i dati interni che quelli esterni devono essere resi persistenti una volta che gli stessi sono generati o entrano nel confine aziendale.

### 1.5.4 Data Extraction

Alcuni dati individuati per l'analisi potrebbero presentarsi in un formato che non è immediatamente compatibile con l'ambiente dei Big Data. Questa situazione è più probabile quando si tratta di dati provenienti da fonti esterne che possono essere eterogenee.

Questo stadio si occupa proprio dell'estrazione e della trasformazione di questi dati in un formato utilizzabile dall'infrastruttura dei Big Data per l'analisi. Il processo in questione dipende dal tipo di analisi che si vuole condurre e dalla capacità della soluzione di Big Data utilizzata.

### 1.5.5 Data Validation and Cleansing

Lo stadio di Data Validation and Cleansing ha il compito di stabilire le regole di validazione e di rimuovere tutti i dati definiti come non validi.

Questo poiché l'input dei dati nelle analisi dei Big Data può essere non strutturato e i dati che appaiono essere non validi potrebbero ,da una parte, falsificare o inquinare i risultati dell'analisi e ,dall'altra, essere indicatori di pattern e trend nascosti.

È importante precisare come dataset differenti possano restituire dati ridondanti; questo può volgere a nostro favore quando la ridondanza permette di riempire dati validi mancanti o assemblare parametri di validazione in dataset interconnessi.

La provenienza dei dati in questo stadio gioca un ruolo assai importante per determinare l'accuratezza e la qualità di essi.

### 1.5.6 Data Aggregation and Representation

Come detto in precedenza, i dati possono essere distribuiti su più dataset; è, dunque, necessaria una operazione di integrazione dei dataset al fine di ottenere una visione unificata.

L'esecuzione di questa attività può risultare complicata a causa di due aspetti:

- *Struttura dei dati*: il formato può essere lo stesso, ma la struttura, quindi il modello, potrebbe essere diverso
- *Semantica*: un valore può essere identificato in due o più modi differenti, pure rappresentando sempre lo stesso dato.

I considerevoli volumi gestiti dalle soluzioni di Big Data possono rendere oneroso questo processo, sia in termini di tempo che di risorse; inoltre, è essenziale adottare un metodo di riconciliazione dei dati o identificare il dataset che rappresenta il valore corretto. Tutto ciò

richiede l'implementazione di una logica complessa che viene eseguita automaticamente senza l'intervento umano.

### 1.5.7 Data Analysis

Arrivati al settimo stadio inizia la reale analisi dei dati. La complessità di questo stadio dipende dall'obiettivo che si vuole raggiungere; in alcune situazioni può trattarsi di una semplice operazione di interrogazione di un dataset. D'altra parte, potrebbe richiedere la combinazione di tecniche di data mining o tecniche statistiche complesse allo scopo di generare pattern e anomalie o un modello statistico. È possibile distinguere due tipi di analisi dei dati:

- *Esplorativa*; prevede un approccio induttivo correlato al Data Mining; non vengono effettuate ipotesi ma si raggiunge una conclusione esplorando i dati. Il problema di questo metodo è che non fornisce risposte definitive, ma orienta alla scoperta di pattern e anomalie.
- *Confermativa*; si basa su un approccio deduttivo, ossia formulando ipotesi preliminari, sulla base delle quali i dati vengono analizzati per confermarle o respingerle.

### 1.5.8 Data Visualization

La fase di Data Visualization si concentra sull'utilizzo di tecniche e strumenti di visualizzazione dei dati per rappresentare graficamente i risultati dell'analisi, agevolando così una comprensione efficace da parte degli utenti aziendali.

I risultati di questo stadio sono molto importanti poiché, una volta che i dati vengono comunicati agli utenti, questi ultimi sono incentivati alla scoperta di risposte a domande che potrebbero non essersi ancora poste.

Dunque, è fondamentale utilizzare la tecnica di visualizzazione più adatta che non porti ad una interpretazione errata dei dati, in quanto gli stessi dati possono essere rappresentati in modo differente.

### 1.5.9 Utilization of Analysis Result

L'ultimo stadio è dedicato a determinare come e dove i dati elaborati dall'analisi possono essere ulteriormente impiegati.

I risultati di analisi possono produrre "modelli", rappresentati come equazioni matematiche o insiemi di regole, che racchiudono nuove intuizioni sulla natura dei pattern e delle relazioni nei dati.

Le aree comuni che vengono esplorate durante questa fase sono:

- *Input per sistemi enterprise*, integrazione automatica o manuale dei risultati dell'analisi nei sistemi aziendali per migliorarne comportamenti e performance.
- *Ottimizzazione dei Business Process*: i pattern identificati, le correlazioni e le anomalie vengono usati per il raffinamento del business process; inoltre, i modelli portano ad opportunità per migliorare la logica.
- *Alert*: i risultati dell'analisi possono essere impiegati come input per alert esistenti oppure rappresentare la base per la creazione di nuovi alert.

CAPITOLO 2

# Descrizione dei dati di riferimento

*In questo capitolo verranno illustrati i dati di riferimento. Inizialmente si descriveranno le varie tipologie di dati utilizzati in un ambiente di Big Data, tra cui dati strutturati, non strutturati, semi-strutturati e metadati. Successivamente, sarà presentato il dataset Motor Vehicle Collisions-Crashes, impiegato per la campagna di Data Analytics. In particolare, si esamineranno la provenienza del dataset e gli attributi a disposizione.*

## 2.1 Varietà dei formati coinvolti nei Big Data

Per effettuare campagne di Data Analytics, l'azienda ha bisogno di dati, che possono essere raccolti da varie sorgenti. Un esempio è rappresentato dalle piattaforme social, come Instagram e Tik Tok, che vengono utilizzate giornalmente da milioni di utenti. I dati utilizzati in un ambiente di Big Data possono essere generati:

- dall'uomo e la sua interazione con i sistemi digitali; ne sono un esempio i dati strutturati e testuali;
- dalle macchine; in questo caso essi vengono prodotti da programmi e dispositivi come, ad esempio, dati relativi ai sensori, web log e dati sugli utilizzi degli apparecchi.

Come evidenziato precedentemente, dati ottenuti dalle aziende derivano da diverse fonti e si possono presentare in svariati formati. I tipi di dati fondamentali, che si riferiscono all'organizzazione interna e vengono processati da soluzioni Big Data, sono:

- dati strutturati;
- dati non strutturati;
- dati semi-strutturati.

Oltre a questi, si aggiunge un altro tipo di dati di grande rilevanza nell'ambito dei Big Data: i metadati.

Nelle prossime sottosezioni daremo uno sguardo più approfondito ad ogni tipologia di dato.

### 2.1.1 Dati strutturati

I dati strutturati hanno un formato predefinito e vengono rappresentati in tabelle; alcuni esempi banali possono essere le transazioni bancarie, le fatture ed i record dei clienti.

Questa tipologia di dato è, spesso, memorizzata in database relazionali ,al fine di catturare le relazioni tra le diverse entità.

Generalmente questo tipo di dato viene prodotto dai sistemi informativi aziendali quali CRM (Customer Relationship Management) ed ERP (Enterprise Resource Planning).

L'elaborazione e la memorizzazione di dati strutturati non richiedono considerazioni speciali, grazie all'abbondanza di tool e database.

### 2.1.2 Dati non strutturati

I dati non strutturati sono i più comuni all'interno di un'azienda; essi rappresentano, infatti, circa l'80% del totale. La particolarità di questo dato è che non si conforma ad uno schema o ad un modello, al contrario di quanto avviene per i dati strutturati. Esempi di dati non strutturati sono i video, gli audio e le immagini; i relativi formati sono di tipo testuale e binario, solitamente registrati in file che sono auto-contenuti e non relazionali. I file di testo possono contenere tweet o post di un blog, mentre i file binari rappresentano file multimediali. I file di tipo testuale e binario hanno una struttura definita dal formato del file stesso; eppure, ciò che li rende dati non strutturati è il formato dei dati contenuti in essi. Questa tipologia di dato non può essere interrogata tramite SQL, ma è possibile utilizzare un database non relazionale (noSQL) per memorizzare dati non strutturati a fianco di dati strutturati .

### 2.1.3 Dati semi-strutturati

I dati semi-strutturati sono gerarchici ed hanno una struttura definita, ma sono, per natura, non relazionali. I formati più comuni per questo tipo di dato sono l'XML ed il JSON; inoltre, grazie alla loro natura testuale ed alla corrisponenza ad una struttura, sono più facili da elaborare rispetto ai dati non strutturati. Se il formato di partenza non è di tipo testuale possono sorgere delle richieste di pre-processing e di memorizzazioni speciali; un esempio può essere la validazione di un file XML. Alcuni esempi di fonti di dati semi-strutturati comprendono i file EDI (Electronic Data Interchange), un sistema che permette alle aziende lo scambio di documenti di business, in un formato elettronico predefinito.

### 2.1.4 Metadati

Questa tipologia di dati viene generata in gran parte dalla macchina che restituisce informazioni sulla struttura e le caratteristiche di un dataset. Il tracking dei metadati è di fondamentale importanza nel ciclo di vita dei Big Data, in quanto fornisce informazioni sulla provenienza e il "peedigre" (originalità, qualità e storia) dei dati.

Esempi tipici di metadati sono:

- tag XML che identificano l'autore e la data di creazione di un documento ;
- attributi che definiscono la dimensione di un file.

## 2.2 Motor Vehicle Collisions-Crashes

In questo paragrafo analizzeremo il dataset utilizzato per effettuare la campagna di Data Analytics. Per la ricerca del nostro dataset è stata utilizzata la piattaforma di Data Science Kaggle, fondata nel 2010, in cui avvengono competizioni per la costruzione di modelli

predittivi e analitici. Il dataset selezionato, denominato Motor Vehicle Collisions-Crashes (Figura 2.1), descrive nel dettaglio le collisioni di veicoli a motore segnalate dalla polizia nell'area metropolitana di New York a partire dal 2012.

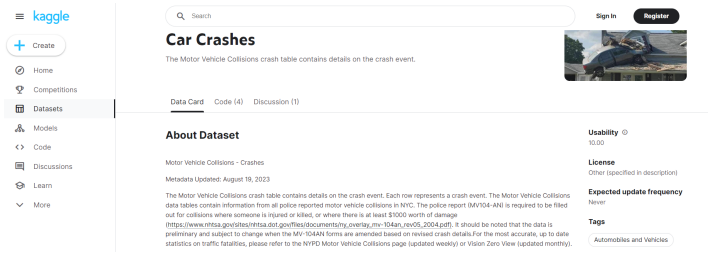


**Figura 2.1:** Schermata di Kaggle che introduce il dataset Motor Vehicle Collisions-Crashes

### 2.2.1 Provenienza e metadati

La sorgente primaria dei nostri dati è il governo degli Stati Uniti, ed in particolare, il sito web Data.gov (Figura 2.2). Questa piattaforma fornisce l'accesso a una vasta gamma di set di dati pubblicati dalle agenzie di tutto il governo federale.


**Figura 2.2:** Schermata iniziale del sito web Data.gov

La Figura 2.3 presenta una schermata di Data.gov contenente le informazioni originali sul set di dati, mostrando anche i vari formati in cui è possibile scaricarlo.

I dati sugli incidenti sono preliminari e soggetti a modifiche in base ai moduli (MV-104AN) compilati dalla polizia per le collisioni in cui si registrano vittime o feriti, o in cui si verificano danni di almeno 1000 dollari.

Grazie allo sviluppo dei sistemi di gestione del traffico stradale, gli agenti di polizia inseriscono elettronicamente, tramite computer o cellulare, tutti i campi dell'MV-104AN che vengono, poi, arichiviati nel magazzino dei dati del dipartimento. La raccolta di grandi quantità di dati sugli incidenti permette di condurre analisi dettagliate sulla sicurezza del traffico, al fine di ridurre il numero di vittime della strada e migliorare la sicurezza stradale a livello cittadino.

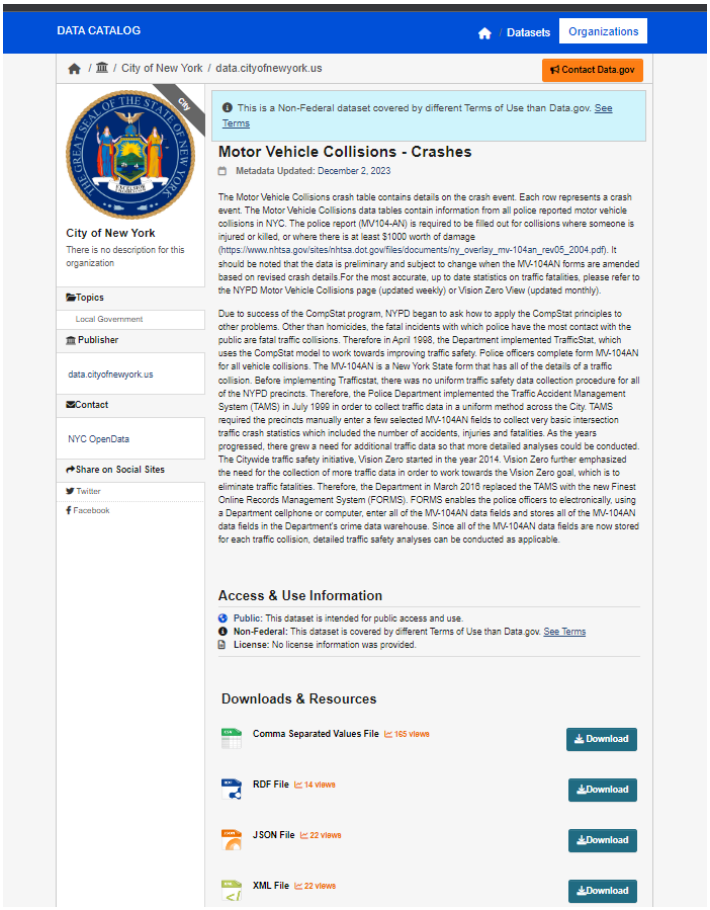


**Figura 2.3:** Schermata del sito Data.gov contenente le informazioni originali sul set di dati

### 2.2.2 Dettagli dataset

Il dataset è composto da 29 colonne e più di 1000 righe. Di seguito, La Figura 2.4 mostra una schermata contenente un'anteprima dei dati, in cui si evidenziano le prime righe e colonne del set di dati in esame.

Ogni riga rappresenta un incidente descritto dai seguenti attributi :

- *Crash date*: data dell'incidente nel formato mm/dd/yyyy;
- *Crash time*: orario in cui è avvenuto il sinistro stradale;
- *Borough*: quartiere di New York in cui è avvenuta la collisione;
- *Latitude*: latitudine dell'incidente secondo il formato DMM, che si basa sui gradi e minuti decimali;
- *Longitude*: longitudine dell'incidente secondo il formato DMM, che si basa sui gradi e minuti decimali;
- *On street name*: strada principale lungo la quale è avvenuta la collisione;
- *Cross street name*: incrocio in cui è avvenuto l'incidente;
- *Off street name*: edficio di fronte al quale è avvenuto lo scontro;
- *Cyclist killed*: numero di ciclisti deceduti nell'incidente;

Motor_Vehicle_Collisions_-_Crashes.csv (429.58 MB)

Detail Compact Column

| CRASH DA... | CRASH TI... | BOROUGH | LATITUDE | LONGITUDE | ON STREE... | CROSS ST... | OFF STRE... | NUMBER ... | NUMBER ... |
|---|---|---|---|---|---|---|---|---|---|
| 09/11/2021 | 2:39 | | | | WHITESTONE EXPRESSWAY | 20 AVENUE | | 2 | 0 |
| 03/26/2022 | 11:45 | | | | QUEENSBORO BRIDGE UPPER | | | 1 | 0 |
| 06/29/2022 | 6:55 | | | | THROGS NECK BRIDGE | | | 0 | 0 |
| 09/11/2021 | 9:35 | BROOKLYN | 40.667202 | -73.8665 | | | 1211 LORING AVENUE | 0 | 0 |
| 12/14/2021 | 8:13 | BROOKLYN | 40.683304 | -73.917274 | SARATOGA AVENUE | DECATUR STREET | | 0 | 0 |
| 04/14/2021 | 12:47 | | | | MAJOR DEEGAN EXPRESSWAY RAMP | | | 0 | 0 |
| 12/14/2021 | 17:05 | | 40.709183 | -73.956825 | BROOKLYN QUEENS EXPRESSWAY | | | 0 | 0 |
| 12/14/2021 | 8:17 | BRONX | 40.86816 | -73.83148 | | | 344 BAYCHESTER AVENUE | 2 | 0 |
| 12/14/2021 | 21:10 | BROOKLYN | 40.67172 | -73.8971 | | | 2047 PITKIN AVENUE | 0 | 0 |
| 12/14/2021 | 14:58 | MANHATTAN | 40.75144 | -73.97397 | 3 AVENUE | EAST 43 STREET | | 0 | 0 |
| 12/13/2021 | 0:34 | | 40.701275 | -73.88887 | MYRTLE AVENUE | | | 0 | 0 |
| 12/14/2021 | 16:50 | QUEENS | 40.675884 | -73.75577 | SPRINGFIELD BOULEVARD | EAST GATE PLAZA | | 0 | 0 |
| 12/14/2021 | 8:30 | | | | broadway | west 80 street -west 81 street | | 0 | 0 |
| 12/14/2021 | 0:59 | | 40.59662 | -74.00231 | BELT PARKWAY | | | 0 | 0 |
| 12/14/2021 | 23:10 | QUEENS | 40.66684 | -73.78941 | NORTH CONDUIT AVENUE | 150 STREET | | 2 | 0 |

**Figura 2.4:** Schermata contenente una porzione del dataset Motor Vehicle Collisions-Crashes

- *Cyclist injured*: numero di ciclisti feriti nell'incidente;
- *Motorist killed*: numero di motociclisti deceduti nell'incidente;
- *Motorist injured*: numero di motociclisti feriti nell'incidente;
- *Pedestrian killed*: numero di pedoni deceduti nell'incidente;
- *Pedestrian injured*: numero di pedoni feriti nell'incidente;
- *Persons Killed*: numero totale di vittime dell'incidente, formato dalle tre categorie precedenti (ciclisti, motociclisti, pedoni);
- *Persons injured*: numero totale di feriti dell'incidente, formato dalle tre categorie precedenti (ciclisti, motociclisti, pedoni);
- *Collision id*: identificatore univoco dell'incidente;
- *Contributing factor vehicle 1* : causa dell'incidente del veicolo;
- *Contributing factor vehicle 2*: causa dell'incidente del secondo veicolo, coinvolto nel sinistro stradale;
- *Vehicle type code 1*: tipologia di veicolo coinvolto nell'incidente;
- *Vehicle type code 2*: tipologia del secondo veicolo coinvolto nel sinistro stradale.

# CAPITOLO 3

## Attività di Extraction, Transformation and Loading

*In questo capitolo verrà introdotto il concetto di ETL e la sua importanza all'interno dell'azienda; dopodichè, verranno approfondite ognuna delle tre fasi che lo compongono: Extract, Transform e Load. Successivamente verrà esaminato l'ecosistema Power BI, utilizzato per eseguire operazioni di ETL. In particolare, ci soffermeremo sul flusso di azioni in Power BI e su alcuni dei suoi componenti fondamentali. Infine, saranno esaminate in dettaglio le operazioni di ETL eseguite sul dataset Motor Vehicle Collisions-Crashes.*

## 3.1 ETL: Extract, Transform e Load

L'ETL, acronimo di Extract Transform e Load, si riferisce al processo di raccolta dei dati provenienti da qualsiasi tipologia di sorgente. Questa attività viene svolta prima della Data Analysis e prevede, inoltre, l'organizzazione e l'integrazione dei dati all'interno di un unico repository. Il termine "Data Integration" fa riferimento a tutte quelle azioni necessarie a unificare diverse sorgenti informative per fornire all'utente una visione unificata di quei dati. Dunque, il processo di ETL si propone di ricavare un pacchetto di dati puliti e accessibili, i quali assumono una rilevanza fondamentale ai fini dell'analisi dei dati, contribuendo a migliorare la Business Intelligent (BI). Quest'ultima si può definire come un insieme di modelli, strumenti e tecnologie utilizzate per trasformare i dati in informazioni utili per l'azienda, al fine di migliorare il processo decisionale. Nelle prossime sottosezioni daremo uno sguardo più approfondito ai tre processi che compongono l'ETL.

### 3.1.1 Extract

La prima fase del processo di ETL è l'estrazione o *Extraction*. Questa si occupa della raccolta di dati grezzi da una qualsiasi fonte, come ad esempio:

- database relazionali SQL;
- comuni file di testo o XML;
- report su anomalie;
- database noSQL.

Una volta estratti, i dati vengono inseriti nell'*area di staging*, una zona di archiviazione temporanea che si interpone tra le sorgenti dei dati e le destinazioni. Le aree di staging

sono spesso transitorie, in quanto i loro i contenuti vengono cancellati una volta completata l'estrazione. Infine, i dati confluiscono nel repository finale o destinazione, che spesso consiste in Data Lake o Data Warehouse. Gli aspetti che distinguono le due soluzioni di storage finale sono diversi; in particolare, un Data Lake viene definito come un enorme insieme di dati grezzi, strutturati e non, il cui scopo non è ancora definito. Un Data Warehouse è un repository di dati strutturati e filtrati, già elaborati per una finalità specifica.

### 3.1.2 Transform

Nella fase di trasformazione i dati raccolti vengono sottoposti ad una serie di operazioni per renderli utilizzabili secondo le esigenze specifiche dell'utente. Le principali operazioni a cui vengono sottoposti i dati sono:

- *Pulizia dei dati o data cleansing*: i dati vengono esaminati al fine di individuare ed eliminare errori, duplicati, inconsistenze o altre anomalie.
- *Normalizzazione dei dati*: nella fase di estrazione i dati possono essere memorizzati in formati differenti. Il processo di trasformazione si occupa di normalizzare questi dati, così da ottenere un formato uniforme e coerente.
- *Filtraggio dei dati*: per soddisfare esigenze specifiche dell'utente, si possono utilizzare tecniche o procedure di filtraggio, escludendo o includendo righe o colonne specifiche.
- *Join e merge*: prevede l'accoppiamento di dati appartenenti a differenti tabelle.

### 3.1.3 Load

L'ultima fase del processo di ETL prevede il caricamento dei dati trasformati presso l'archivio di destinazione. Durante questa fase, i dati possono essere caricati in modi distinti, quali:

- *Caricamento completo*: prevede la riscrittura completa dei dati, ovvero tutti i dati dell'origine vengono trasferiti nel sistema di destinazione. Generalmente, questa tipologia di caricamento avviene la prima volta che si trasferiscono i dati da un sistema di origine ad un sistema di destinazione.
- *Caricamento incrementale*: viene adottato quando è necessario ridurre al minimo il trasferimento dei dati. Questo metodo implica il caricamento esclusivo dei dati che sono stati aggiunti o modificati dall'ultima operazione ETL, garantendo, così, che vengano caricati solo i record aggiunti.

## 3.2 Power BI

Power BI è una piattaforma di Business Intelligence e analytics sviluppata da Microsoft che consente agli utenti di analizzare, visualizzare, e condividere dati aziendali in modo efficace. Power BI offre una vasta gamma di strumenti software e app per creare report interattivi e dashboard; grazie a questi, gli utenti aziendali possono ottenere rapidamente una panoramica dei dati e delle tendenze, nonché individuare i problemi e prendere decisioni formate. Il flusso di azioni in Power BI segue un processo ben definito che solitamente comprende:

- *Connessione ai dati*: gli utenti possono connettersi a una vasta gamma di sorgenti di dati, come database, fogli di calcolo e file CSV.

- *Trasformazione dei dati*: i dati vengono estratti dalla fonte e caricati in Power BI per l'analisi. Durante questa fase è possibile pulire, filtrare e manipolare i dati per creare un modello dei dati.
- *Creazione degli oggetti visivi*: gli utenti creano le prime rappresentazioni grafiche di dati mediante oggetti visivi come grafici a barre, grafici a torta, tabelle e mappe.
- *Creazione dei report*: gli utenti raccolgono gli oggetti visivi in una o più pagine di report.
- *Condivisione dei report*: una volta che il report è stato ultimato è possibile condividerlo con altri utenti tramite il servizio Power BI.

Tra le componenti dell'ecosistema Power BI da noi utilizzate vi è *Power BI Desktop*, che consente la progettazione dei report e la visualizzazione dei dati. Inoltre, per la manipolazione dei dati, sono stati ampiamente utilizzati l'editor di query *Power Query* ed il *linguaggio DAX* (Data Analysis eXpression) per la creazione di misure personalizzate. Nella prossima sezione daremo uno sguardo più approfondito ai processi di ETL effettuati sul dataset Motor Vehicle Collisions-Crashes.

## 3.3 ETL su Motor Vehicle Collisions-Crashes

In questa sezione illustreremo nel dettaglio le tre fasi di estrazione, trasformazione e caricamento applicate al nostro dataset.

### 3.3.1 Extract

La prima fase prevede l'estrazione dei dati dalla nostra sorgente. Una volta avviato Power BI desktop viene visualizzata la schermata iniziale (*Figura 3.1*), in cui sono visibili le voci *Recupera Dati* e *Origini Recenti*. Per connettersi ai dati selezionare la prima voce in alto; altrimenti è possibile farlo successivamente nella toolbar di Power BI Desktop (*Figura 3.2*).

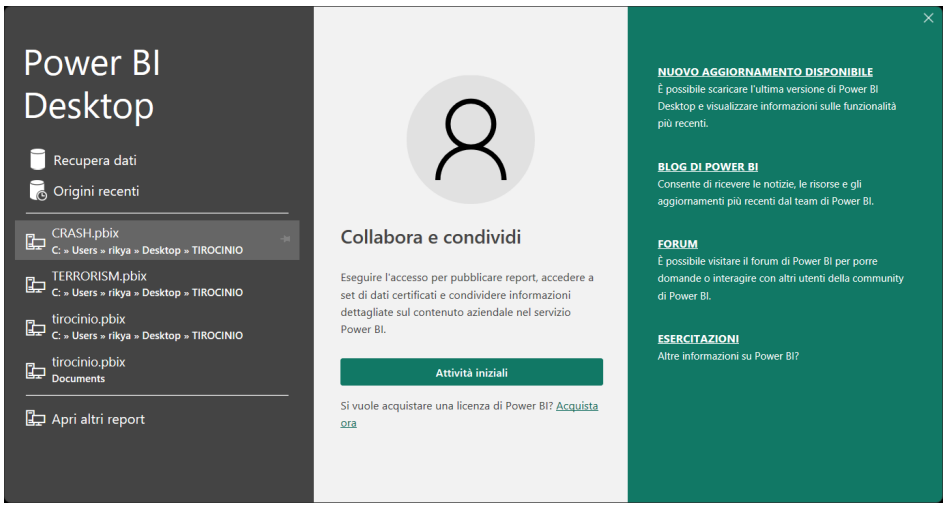


**Figura 3.1:** Schermata iniziale di Power BI

La schermata di recupero dei dati, visibile nella *Figura 3.3*, permette di scegliere una tipologia di sorgente, compatibile con Power BI, da cui estrarremo i dati. Nel nostro caso Kaggle mette a disposizione il dataset nel formato csv; quindi, selezioniamo testo/csv e, successivamente, procediamo con *Connetti* per instaurare la connessione tra la sorgente e

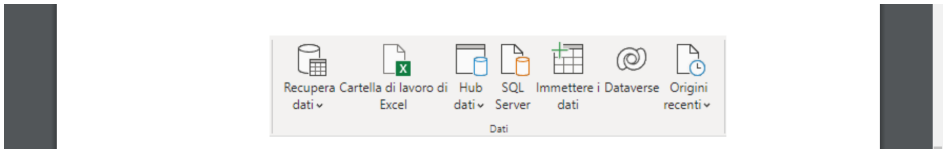


**Figura 3.2:** Toolbar Power BI Desktop

Power BI. Dopodichè, verrà aperta una schermata in cui sarà visibile un'anteprima dei nostri dati, assieme alle voci *Carica* e *Trasforma*. La prima voce esegue il caricamento dei dati, senza apportare alcuna modifica; la seconda procede con l'avvio dell'editor di Power Query per la modifica della forma dei dati. Nel nostro caso selezioneremo la voce "Trasforma" per passare alla seconda fase di ETL.

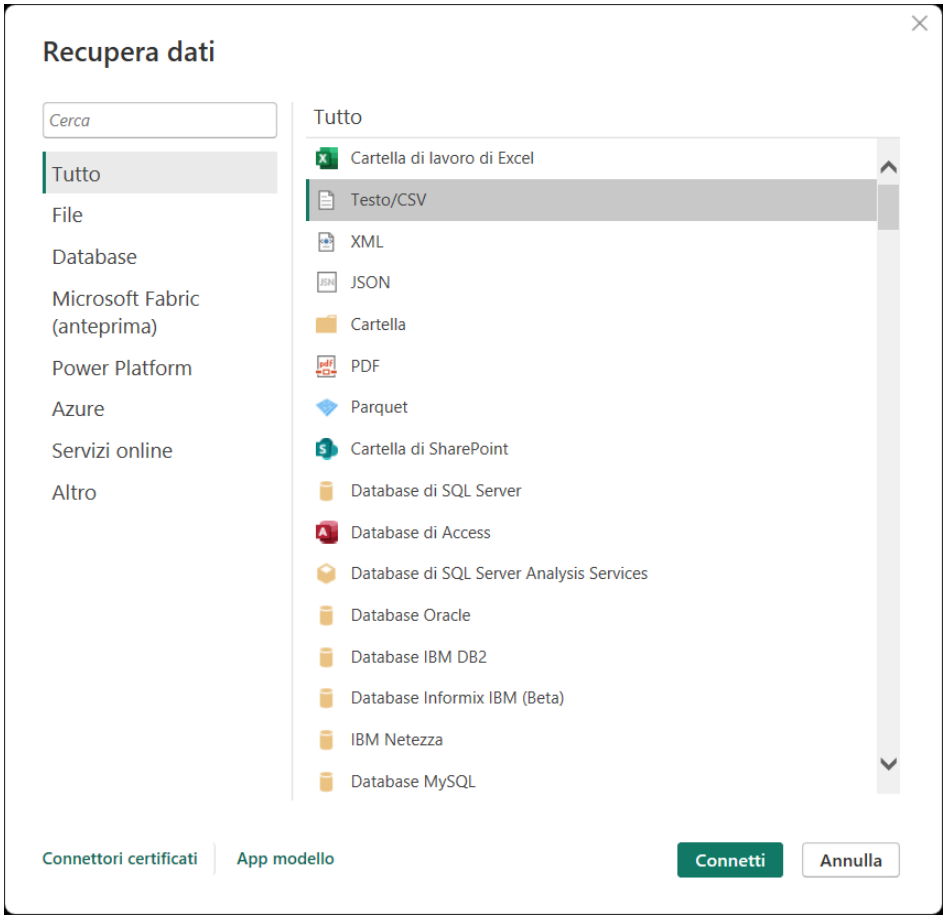


**Figura 3.3:** Schermata di Power BI per la scelta della tipologia di sorgente dei dati

### 3.3.2 Transform

La fase di trasformazione prevede l'utilizzo dell'editor di query Power Query e delle sue funzionalità per esplorare e pulire i dati. Ogni operazione di modifica effettuata viene salvata in ordine cronologico nella colonna "Passaggi Applicati"; selezionando i passaggi si può vedere l'effetto che ha sui dati nell'editor di Power Query, come mostrato in *Figura 3.4*.

Il primo step prevede la rimozione dalla tabella di tutte le colonne superflue. Procediamo selezionando l'intestazione della colonna che vogliamo eliminare e poi utilizziamo il pulsante *Rimuovi colonna* presente nella toolbar. Come possiamo vedere nella *Figura 3.5*, il

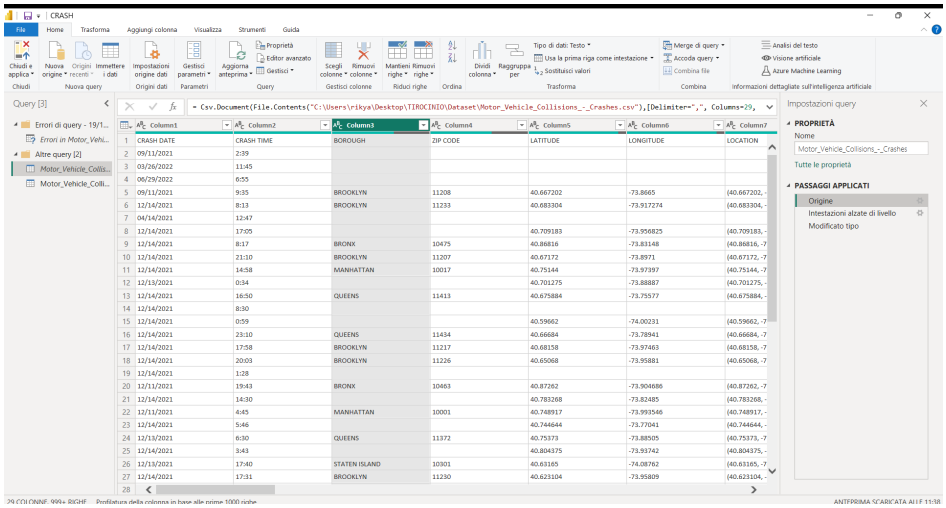

**Figura 3.4:** Visualizzazione degli effetti delle trasformazioni sui dati

passaggio denominato “Rimosse Colonne” elenca tutte le colonne eliminate dallo script *Table.RemoveColumns*, quali:

- *zip code*;
- *contributing factor vehicle 3*;
- *contributing factor vehicle 4*;
- *contributing factor vehicle 5*;
- *vehicle type code 3*;
- *vehicle type code 4*;
- *vehicle type code 5*;
- *off street name*.

.

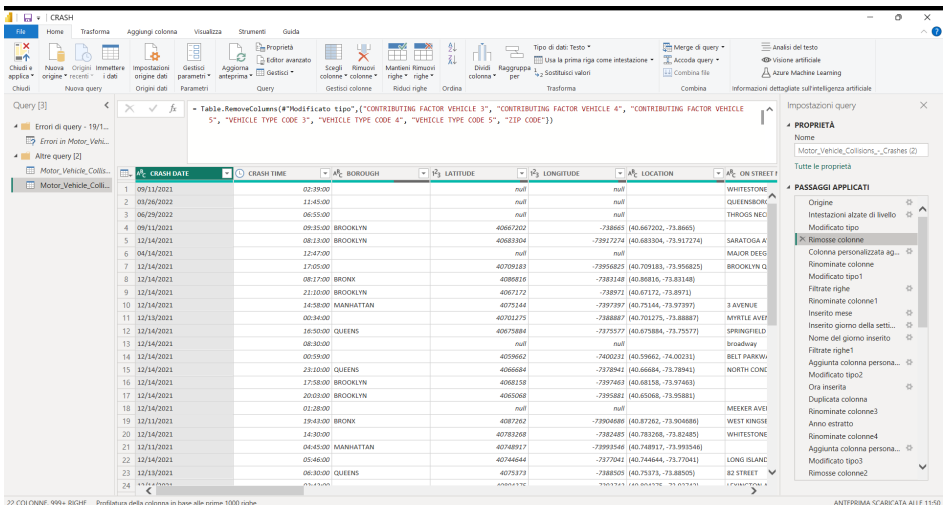

**Figura 3.5:** Rimozione delle colonne superflue

A questo punto possiamo iniziare a modificare le singole colonne. Notiamo che i dati nella prima colonna Crash Date presentano un formato errato del tipo: mm/dd/yyyy. Procediamo con la creazione di una nuova colonna denominata *NewDate*, che inverte la posizione dei mesi con quella dei giorni, grazie allo script *Text.Combine* mostrato nella *Figura 3.6*. Nel passaggio successivo si converte il tipo di dati da testuale a data. Infine, concludiamo il perfezionamento dei dati temporali con la creazione di una colonna che contiene l'ora esatta in cui si verifica l'incidente. Per quanto riguarda l'orario dell'incidente è definito nella colonna *Crash Time* nel formato hh:mm:ss. Utilizziamo la funzione *Time.hour* che permette di estrarre l'ora da un valore di tipo data/ora.

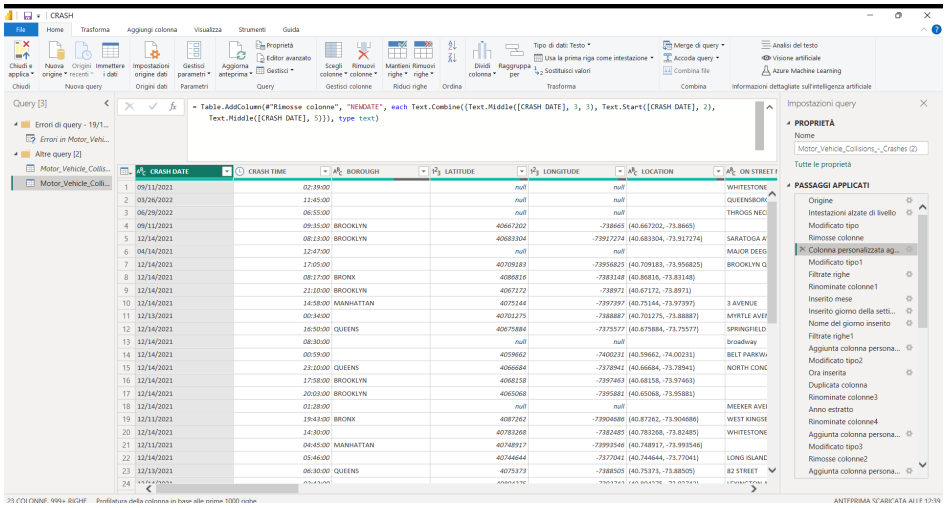

**Figura 3.6:** Creazione della Nuova Colonna "NewDate"

Un'altra operazione da effettuare prevede la creazione di nuove colonne, denominate:

- *Crash*;
- *Cyclist Crash*;
- *Motorist Crash*;
- *Pedestrian Crash*.

La colonna *Crash* rappresenta l'avvenimento di un incidente, quindi sarà composta esclusivamente da 1. La utilizzeremo in seguito per il calcolo delle rimanenti colonne. Come mostrato in *Figura 3.7* creiamo *Cyclist Crash* sommando gli elementi delle due colonne *Cyclist Injured* e *Cyclist killed*, contenenti, rispettivamente, il numero di ciclisti feriti ed uccisi nell'incidente. A questo punto procediamo creando una nuova colonna *Personalizzato* (*Figura 3.8*), utilizzando la funzione *List.min*, che restituisce il valore minimo tra gli elementi delle colonne *Cyclist Crash* e *Crash*. In questo modo, otteniamo righe con valore 1 quando si verificano incidenti che coinvolgono ciclisti e, d'altra parte, righe di 0 quando non sono coinvolti. Dopodichè, eliminiamo la colonna *Cyclist Crash*, poichè verrà sostituita da *Personalizzato* che prenderà il suo nome. Questo procedimento viene svolto in maniera analoga per ogni categoria.

Infine, l'ultimo problema da risolvere riguarda la Location degli incidenti. Il formato iniziale delle coordinate, latitudine e longitudine, non è supportato da Power BI. Per risolvere questo problema, procediamo con la creazione di due nuove colonne utilizzando la funzione *Text.combine*. In questo modo, inseriamo il formato (DMM) che si basa sui gradi e minuti decimali (*Figura 3.9*). Una volta ottenute le colonne corrette, che restituiscono la latitudine e

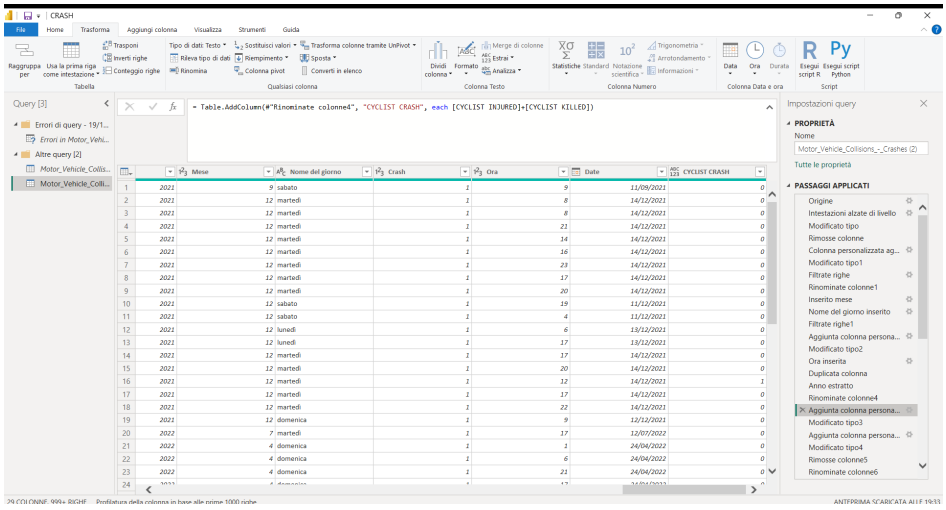

**Figura 3.7:** Creazione della colonna *Cyclist Crash*

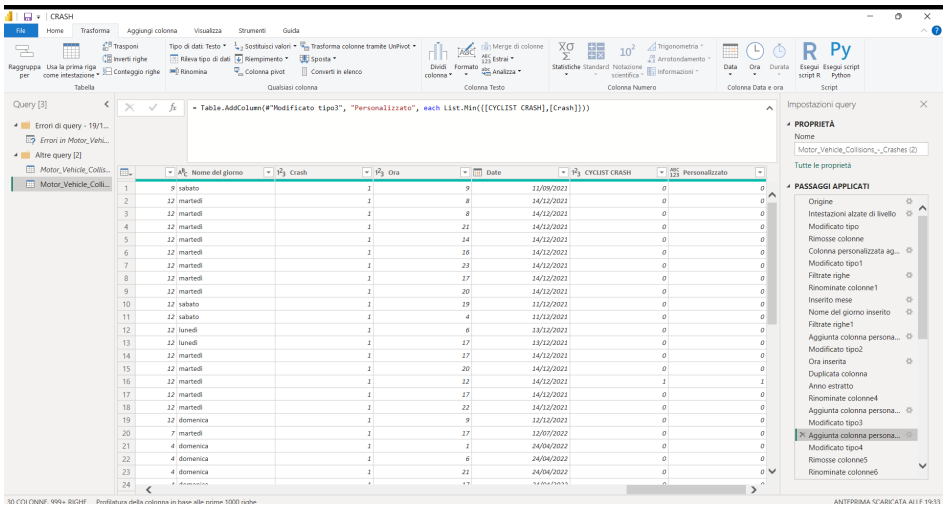

**Figura 3.8:** Creazione della colonna *Personalizzato* per *Cyclist Crash*

la longitudine di un incidente, è possibile rimuovere quelle precedenti. Successivamente, si esegue il merge tra le colonne *On street* e *Off street*. Inseriamo i nomi delle due strade separati dal simbolo & all'interno della nuova colonna *Intersection*; ciò restituisce l'incrocio esatto in cui avviene l'incidente. La *Figura 3.10* mostra i dati geografici a disposizione nelle tre colonne appena create.

### 3.3.3 Load

Terminata la fase di trasformazione, passiamo a quella di caricamento, dove i dati vengono trasferiti dall'editor di Power Query a Power BI Desktop. Il caricamento dei dati può essere eseguito selezionando la voce *Chiudi e Applica* presente nella toolbar di Power Query. In alternativa, nella schermata di Power BI Desktop apparirà una nuova sezione che consente di applicare o rimuovere le modifiche effettuate.

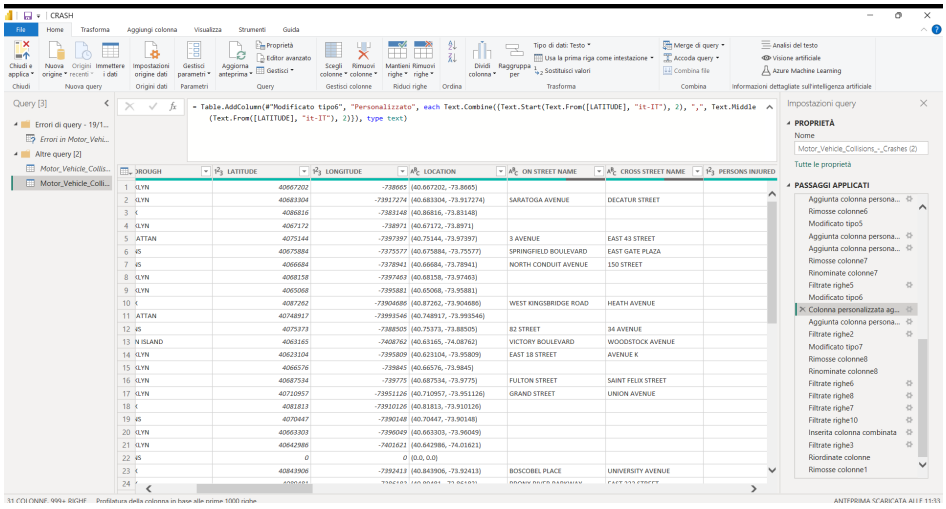

**Figura 3.9:** Creazione della colonna *Personalizzato* per Latitude

| 1.2 LATITUDINE | 1.2 LONGITUDINE | Intersection |
|---|---|---|
| 40,75144 | -73,97397 | 3 AVENUE & EAST 43 STREET |
| 40,675884 | -73,75577 | SPRINGFIELD BOULEVARD & EAST GATE PLAZA |
| 40,66684 | -73,78941 | NORTH CONDUIT AVENUE & 150 STREET |
| 40,710957 | -73,951126 | GRAND STREET & UNION AVENUE |
| 40,843906 | -73,92413 | BOSCOBEL PLACE & UNIVERSITY AVENUE |
| 40,89481 | -73,86183 | BRONX RIVER PARKWAY & EAST 233 STREET |
| 40,861862 | -73,91275 | MAJOR DEEGAN EXPRESSWAY & WEST FORDHA... |
| 40,767242 | -73,986206 | WEST 56 STREET & 9 AVENUE |
| 40,692356 | -73,94282 | THROOP AVENUE & DE KALB AVENUE |
| 40,65011 | -73,930214 | UTICA AVENUE & SNYDER AVENUE |
| 40,745235 | -73,937706 | THOMSON AVENUE & SKILLMAN AVENUE |
| 40,758705 | -73,93793 | 21 STREET & 37 AVENUE |
| 40,695156 | -73,845406 | JAMAICA AVENUE & 102 STREET |
| 40,748158 | -73,97033 | 1 AVENUE & EAST 41 STREET |
| 40,59207 | -73,96299 | EAST 7 STREET & CRAWFORD AVENUE |
| 40,653023 | -73,73895 | 149 AVENUE & HUXLEY STREET |
| 40,665375 | -73,934235 | CROWN STREET & SCHENECTADY AVENUE |
| 40,776237 | -73,943825 | EAST END AVENUE & EAST 88 STREET |
| 40,640835 | -73,98967 | 12 AVENUE & 41 STREET |
| 40,754295 | -73,93946 | 23 STREET & 40 AVENUE |
| 40,638523 | -73,92607 | KINGS HIGHWAY & FARRAGUT ROAD |
| 40,637833 | -74,08193 | CORSON AVENUE & WESTERVELT AVENUE |
| 40,797836 | -73,96946 | WEST 101 STREET & BROADWAY |

**Figura 3.10:** Colonne *Latitude*, *Longitude* e *Intersection*

CAPITOLO 4

# Analisi Esplorativa dei Dati

*In questo capitolo introdurremo il concetto di EDA e la sua importanza nel campo della Data Science. Successivamente approfondiremo il ruolo fondamentale della Data Visualizaton nell'EDA, soffermandoci sugli strumenti offerti da Power BI per la creazione di report interattivi. Infine, esamineremo il processo di EDA svolto sul nostro dataset, ponendo particolare attenzione sulle operazioni di Data Visualization e sul report realizzato.*

## 4.1 Che cos'è l'Analisi Esplorativa dei Dati

L'Exploratory Data Analysis (EDA) è una tecnica utilizzata nel campo della Data Science per analizzare e approfondire la conoscenza di un dataset. L'obiettivo principale è di individuare, tramite una serie di analisi preliminari, una serie di fattori come:

- anomalie nei dati o *outlier*;
- modelli all'interno dei dati;
- pattern e relazioni interessanti tra variabili.

Questa è una fase molto importante poiché può portare alla luce informazioni rilevanti per lo studio del set di dati. In statistica, il termine "outlier" fa riferimento ad un valore anomalo o estremo, che dunque si discosta dalla normale distribuzione dei dati (o dal resto del dataset). Durante l'analisi dei dati è importante gestire in modo corretto gli outlier, poiché potrebbero influenzare negativamente la validità dei risultati.

Nell'ambito dell'analisi esplorativa dei dati, esistono diverse tipologie di analisi che possono essere utilizzate. Tra le principali abbiamo:

- *Analisi Univariata*: si concentra sull'esplorazione di una singola variabile all'interno del dataset, senza considerare le relazioni con le altre. Vengono utilizzate tecniche statistiche e di visualizzazione per comprendere le distribuzioni, le tendenze e le caratteristiche di una variabile.
- *Analisi Multivariata*: si esaminano due o più variabili contemporaneamente, al fine di identificare associazioni, correlazioni o dipendenze tra queste.
- *Analisi Temporale*: questa tipologia si concentra sull'analisi di dati che variano nel tempo. L'obiettivo è identificare trend, stagionalità, cicli e altre tendenze temporali nei dati.

- *Analisi Spaziale*: l'analisi spaziale esamina la distribuzione geografica dei dati e le relazioni spaziali tra le variabili. Questo tipo di analisi è comune nei dati geospaziali e può rivelare pattern o cluster spaziali.

### 4.1.1 Data Visualization

La visualizzazione dei dati è uno degli strumenti più potenti dell'EDA che consente agli analisti di esplorare, comprendere e comunicare i dati. La rappresentazione visiva dei dati aiuta ad estrapolare informazioni che non sono ottenibili dai semplici dati grezzi. L'utilizzo di elementi visivi, come diagrammi, grafici e mappe, è di fondamentale importanza, perché permette di comunicare ed interpretare grandi quantità di dati, in modo intuitivo e rapido, con una singola immagine. La visualizzazione dei dati è uno strumento molto importante che aiuta il processo di analisi esplorativa dei dati, poichè:

- *Semplifica la complessità*: i dati possono essere complessi e con numerose variabili. La visualizzazione semplifica questa complessità presentando le informazioni in un formato visivo facile da comprendere.
- *Consente il riconoscimento di modelli*: le visualizzazioni semplificano l'identificazione di modelli e relazioni all'interno dei dati.
- *Migliora la comunicazione*: le rappresentazioni visive dei dati sono più accessibili e coinvolgenti, rendendo più semplice trasmettere informazioni.
- *Consente il rilevamento delle anomalie*: le visualizzazioni evidenziano più rapidamente outlier.
- *Accresce l'efficienza in termini di tempo*: le visualizzazioni forniscono una rapida panoramica dei dati, risparmiando tempo rispetto all'ispezione manuale dei dati grezzi.

Power BI offre una serie di strumenti e tool che semplificano il processo di Data Visualization e permettono la creazione di report interattivi. Quando si caricano i dati in Power BI Desktop per la prima volta viene visualizzata un'area di disegno vuota, assieme ai riquadri: *Filtri*, *Visualizzazione* e *Campi*. Dopo avere importato i dati, è possibile trascinare i campi nell'area di disegno del report per creare oggetti visivi. Un oggetto visivo è una rappresentazione grafica dei dati; una raccolta di oggetti visivi viene chiamata report. Nel riquadro *Visualizzazioni* si seleziona la tipologia di data visualization. Le diverse tipologie sono fondamentali per presentare i dati in modo che evidenzino pattern, correlazioni, distribuzioni e tendenze. La scelta dipende dalla natura dei dati che si vogliono esplorare e dall'obiettivo dell'analisi.

Gli oggetti visivi più utilizzati sono:

- *Istogrammi.*

  Gli istogrammi forniscono una rappresentazione grafica della distribuzione di una singola variabile continua. Questo tipo di visualizzazione è rappresentata da un diagramma a barre che consente di mostrare la tendenza centrale e la diffusione dei dati, ovvero, la media e la varianza.

- *Grafici a dispersione.*

  È una tecnica di visualizzazione fondamentale che permette di esaminare la relazione tra due variabili continue. Il grafico è rappresentato da uno spazio cartesiano in cui vengono riportate le due variabili sugli assi. Questo tipo di visualizzazione è molto utile per identificare modelli come tendenze, cluster o valori anomali.

- G*rafici a barre e a colonne.*

  Questi grafici sono utilizzati per confrontare quantità relative a diverse categorie. Le barre possono essere disposte sia orizzontalmente che verticalmente, facilitando il confronto visivo tra categorie. Simili ai grafici a barre, ma disposti verticalmente, i grafici a colonne sono efficaci per mostrare cambiamenti di dati relativi ad un periodo di tempo per diverse categorie.

- *Grafici a linee e ad aree.*

  I grafici a linee e ad aree sono ideali per visualizzare tendenze o modelli nei dati nel corso del tempo. Vengono utilizzati per evidenziare la variazione di una o più variabili nel tempo. I grafici a linee sono rappresentati da una serie di punti collegati da segmenti, in aggiunta, i grafici ad aree riempiono lo spazio sottostante alle linee di collegamento con un colore.

- *Mappe.*

  Le mappe sono usate per rappresentare dati geografici, combinando informazioni visive con dati numerici. Possono includere mappe coropletiche, dove le aree sono colorate o ombreggiate in proporzione a una statistica, o mappe a bolle, dove le dimensioni di una bolla rappresentano la grandezza di un dato in una specifica località.

- *Grafici ad anello e a torta.*

  Questi grafici forniscono un modo intuitivo per visualizzare e confrontare le dimensioni di diverse categorie o gruppi di un insieme di dati. Mostrano proporzioni e percentuali tra categorie. Sono efficaci per dati con un numero limitato di categorie che compongono un intero. Tuttavia, possono diventare difficili da interpretare quando le categorie sono molte.

### 4.1.2 Filtri

Un filtro dei dati è un tipo di visualizzazione molto utile che consente all'utente di adattare i criteri di analisi secondo determinate specifiche o esigenze. In particolare, Power BI desktop permette di impostare tre diversi livelli di filtri all'interno del report:

- *Livello di pagina*: all'interno del riquadro *Dati* del report, scegliere il campo da aggiungere come filtro a livello di pagina. Successivamente, trascinare il campo selezionato all'interno del riquadro *Filtri* sotto la voce "Filtri in questa pagina".

- *Livello del report*: nel riquadro *Dati* del report, selezionare il campo che si desidera aggiungere come nuovo filtro a livello di report e trascinarlo nell'area *Filtri* sotto la voce "Filtri in tutte le pagine".

- *Livello visivo*: una volta selezionato un oggetto visivo, i suoi campi vengono visualizzati nel riquadro *Filtri* sotto la nuova voce "Filtri in questo oggetto visivo". A quel punto è possibile aggiungere nuovi campi da filtrare o agire sui campi inziali, modificando esclusivamente l'oggetto visivo in questione.

All' interno del riquadro *Filtri* è possibile accedere alla scheda filtro di un campo che mette a disposizione ulteriori sottofiltri. Questi permettono all'utente di effettuare analisi specifiche selezionando manualmente i valori desiderati. I principali tipi di sottofiltri sono i seguenti:

- *Filtro avanzato*: filtra i dati applicando delle condizioni ben precise come "maggiore di", "minore di", "uguale a". Viene maggiormente utilizzato quando si vogliono impostare dei criteri complessi di filtraggio.
- *Filtro di base*: consente di selezionare rapidamente valori specifici da un elenco predefinito, senza doverli inserire manualmente.
- *Filtro Primi N*: filtra i primi N valori in base ad una misura specifica. Consente di identificare rapidamente i valori più rilevanti all'interno di un set di dati.

### 4.1.3 Formule DAX

Il linguaggio DAX (Data Analysis eXpressions) è uno strumento utile delle analisi dei dati che viene utilizzato dall'utente per creare delle misure personalizzate e colonne calcolate. Queste misure calcolano un risultato da una formula di espressione DAX che può includere funzioni, operatori, valori per eseguire query e calcoli sui dati a disposizione. DAX include una libreria che offre un'enorme flessibilità nella creazione di misure per il calcolo dei risultati e per qualsiasi esigenza di analisi. Nelle formule vanno sempre specificate tabelle e colonne (se il nome della tabella contiene spazi, bisogna scriverlo tra apici). Le colonne e le misure sono racchiuse tra parentesi quadre. In Power BI Desktop le misure vengono create selezionando i comandi "Nuova misura" o "Misura rapida", presenti nella toolbar. Una volta create vengono visualizzate nell'elenco *Campi* con un'icona a forma di calcolatrice. Inoltre, è possibile assegnare un nome alle misure desiderate. Una funzione DAX fa sempre riferimento a una colonna completa o una tabella. Se si desidera utilizzare solo valori specifici di una tabella o colonna, è possibile aggiungere filtri alla formula.

## 4.2 EDA sul dataset "Motor Vehicle Collisions-Crashes"

In questa sezione approfondiremo nel dettaglio le tecniche e gli strumenti impiegati per condurre l'Analisi Esplorativa dei Dati sul dataset "Motor Vehicle Collisions– Crashes".

### 4.2.1 Definizione categorie utenti della stada

Il dataset contiene informazioni dettagliate sugli incidenti stradali che coinvolgono veicoli a motore. In particolare, per ogni incidente vengono esposti il numero di feriti e di vittime per le seguenti categorie di utenti:

- *ciclisti*;
- *motociclisti*;
- *pedoni*;
- *persone*.

Quest'ultimo indica la somma totale dei soggetti rimasti coinvolti nell'incidente. A questo punto l'analisi si concentra sulle prime tre categorie, esplorando specifiche come la letalità, le cause principali degli incidenti e i veicoli maggiormente coinvolti. Per effettuare analisi di questo tipo sono necessarie delle colonne che identificano gli incidenti che coinvolgono queste categorie. Nel capitolo precedente abbiamo illustrato la creazione delle colonne *Cyclist Crash*, *Motorist Crash* e *Pedestrian Crash*. Queste contengono righe di 0 quando l'incidente non include nessuna vittima e nessun ferito che rientra tra le categorie di cui sopra, mentre contengono righe di 1 quando la categoria corrispondente presenta degli 1 nella colonna

dei feriti o delle vittime. Possiamo avere nella stessa riga anche tre 1, se l'incidente riguarda tutti i soggetti. Infine, le ultime pagine del report sono dedicate all'analisi dei tre veicoli: *bicicletta*, *moto* e *sedan*. Con quest'ultimo termine di intende la classica berlina, che è il veicolo con il maggior numero di incidenti. Basandoci sulle informazioni precedenti cerchiamo di effettuare analisi approfondite sui rischi che apportano alla circolazione questi veicoli.

### 4.2.2 Scopo dell'analisi

Nella sezione precedente abbiamo illustrato le categorie di utenti e veicoli soggetti all'analisi. Ora possiamo definire gli obiettivi di questa campagna di Data Analytics. Inizialmente, si propone un'analisi generale degli incidenti verificatisi nell'area metropolitana di New York. Successivamente, si effettuano analisi dettagliate sui vari gruppi, esaminando le tendenze sia dal punto di vista spaziale che temporale. In sintesi, l'obiettivo è cercare di comunicare ai lettori, attraverso i dati raccolti e analizzati, che i ciclisti sono esposti a dei rischi inferiori rispetto alle altre categorie. Inoltre, si cerca di evidenziare l'utilizzo della bicicletta come mezzo di trasporto, poichè apporta una maggior sicurezza alla circolazione, per se stessi e per gli altri utenti della strada.

### 4.2.3 Tassi e percentuali (DAX)

In questa sottosezione analizziamo le misure create con Power BI Desktop per l'analisi esplorativa dei dati. Per ogni gruppo di utenti della strada sono stati calcolate tre misure personalizzate:

- *Percentuale di incidente*: indica la percentuale di incidenti di quella categoria. La *Figura 4.1* mostra come il tasso percentuale è calcolato dividendo la somma dei *Cyclist Crash* per la somma totale dei *Crash*.
- *Tasso di mortalità*: indica la percentuale di letalità di un incidente con feriti (quanti dei feriti sono rimasti uccisi nell'incidente. Il tasso percentuale è dato dal rapporto tra la somma di *Cyclist Killed* diviso la somma di *Cyclist Killed* più la somma di *Cyclist Injured* (*Figura 4.2*).
- *Media annuale degli incidenti*: indica la media annuale degli incidenti della categoria in questione.


**Figura 4.1:** Percentuale di incidente dei ciclisti


**Figura 4.2:** Tasso di mortalità dei ciclisti

La situazione è analoga per quanto riguarda i veicoli; le misure calcolate sono le seguenti:

- *Tasso di mortalità*.
- *Media annuale degli incidenti*.
- *Percentuale di incidenti con feriti*: definisce la percentuale di essere coinvolti in un incidente che presenta dei feriti di qualsiasi categoria.

I risultati di queste formule vengono confrontati al fine di estrarre informazioni, trovare tendenze, modelli e pattern.

### 4.2.4 Data Visualization ed esplorazione delle distribuzioni

In questa sottosezione ci occuperemo di mostrare e descrivere le diverse tipologie di Data Visualization adottate per l'analisi esplorativa dei dati del dataset Motor Vehicle Collisions – Crashes. La *Figura 4.3* mostra la prima pagina del report che serve ad introdurre le caratteristiche generali del dataset. Il primo oggetto visivo che incontriamo è la scheda, presente sia nella parte superiore che nella parte inferiore. Le schede mostrano la somma degli incidenti, delle vittime e dei feriti. Grazie a questi dati è possibile notare come la categoria dei motociclisti ha un numero maggiore di incidenti e di vittime rispetto agli altri. D'altra parte, la categoria dei pedoni predomina per numero di uccisioni, nonostante gli incidenti siano inferiori a quelli dei motociclisti. In basso a destra è presente una mappa dei quartieri della città, in cui sono raffigurate delle bolle. La dimensione di queste variano in base al numero di incidenti. Pertanto, la mappa serve ad evidenziare i quartieri in cui avvengono più incidenti.

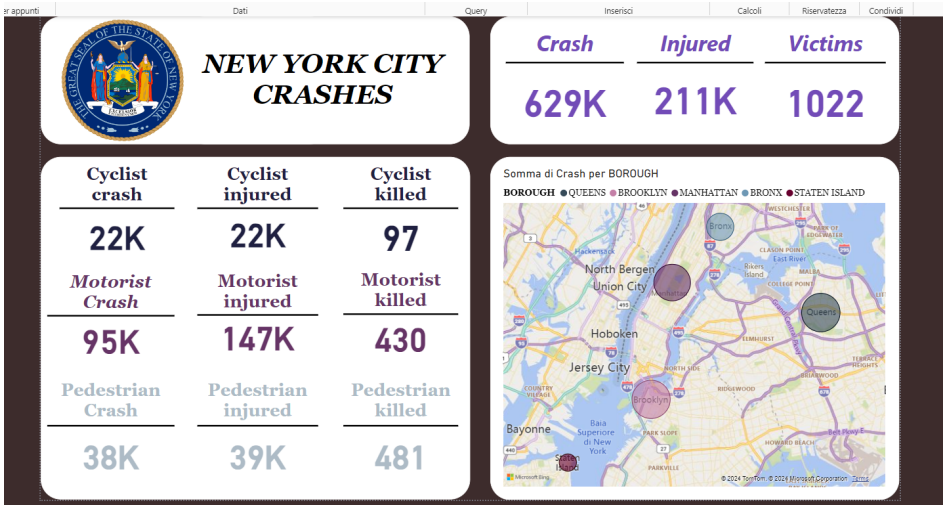


**Figura 4.3:** Prima pagina del report

Nella seconda pagina del report, mostrata in*Figura 4.4*, troviamo: in alto a destra una scheda con il tasso di mortalità e la media annuale degli incidenti. In basso a sinistra, un grafico a barre disposte orizzontalmente, che mostra le 10 strade di New York dove si verificano più incidenti. Tra queste spiccano vie famose come *Fifth Avenue*, *Broadway* e *Park Avenue*. In basso a destra sono presenti due ulteriori grafici. Il primo dall'alto è un grafico a colonne che mostra la percentuale di mortalità per ogni quartiere di New York. In testa vi è *Staten Island* con un tasso superiore allo 0.6%, seguito poi da *Manhattan* e *Brooklyn*. Sotto di questo troviamo un grafico a linee che raffigura la tendenza della media annuale degli incidenti dal 2012 al 2023. Vediamo che nel 2015 è stato registrato un picco massimo con una media di oltre 80.000 incidenti. Negli ultimi anni, invece, la media si è dimezzata.

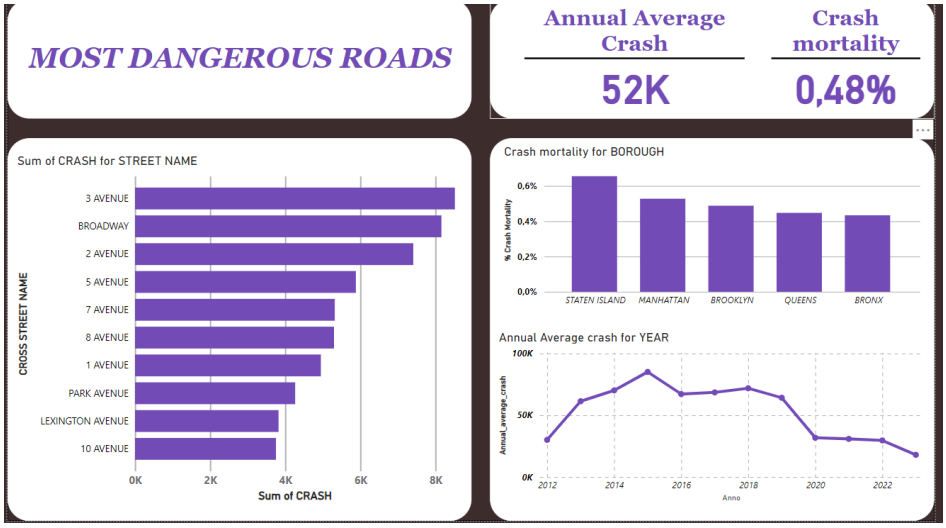


**Figura 4.4:** Seconda pagina del report

### 4.2.5 Report sulle categorie

In questa sottosezione analizzeremo l'organizzazione e il contenuto delle pagine che compongono il report relativo alle catgorie. La prima pagina è strutturata in modo che in alto vi è il titolo che identifica la categoria, seguito da un grafico a barre orizzontali che evidenzia i veicoli maggiormente coinvolti negli incidenti di quel gruppo. Ad esempio, nella *Figura 4.5*, che rappresenta il report relativo ai ciclisti, la *berlina* risulta essere il primo veicolo, con circa 5.000 incidenti. Nella seconda pagina troviamo, in alto, due schede che mostrano la somma totale dei feriti ed il tasso di mortalità del gruppo. Nella parte inferiore troviamo un altro grafico a barre orizzontali che mostra le principali cause di incidenti. Dalla *Figura 4.6*, che raffigura sempre il gruppo dei ciclisti, si può notare come la disattenzione del conducente risulta essere il motivo primario, avendo provocato più di 6.0000 incidenti.

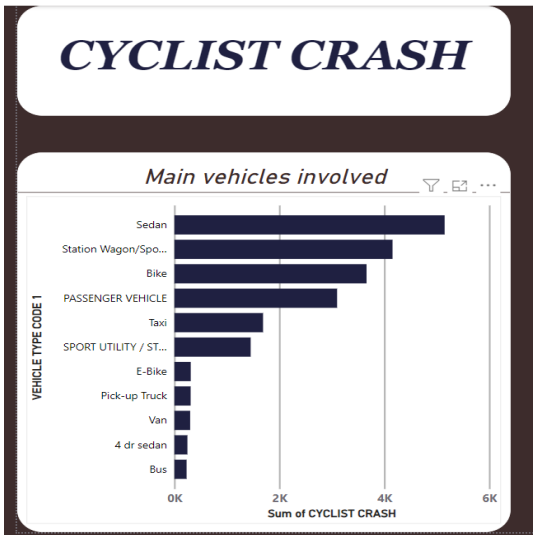


**Figura 4.5:** Prima pagina del report dei ciclisti

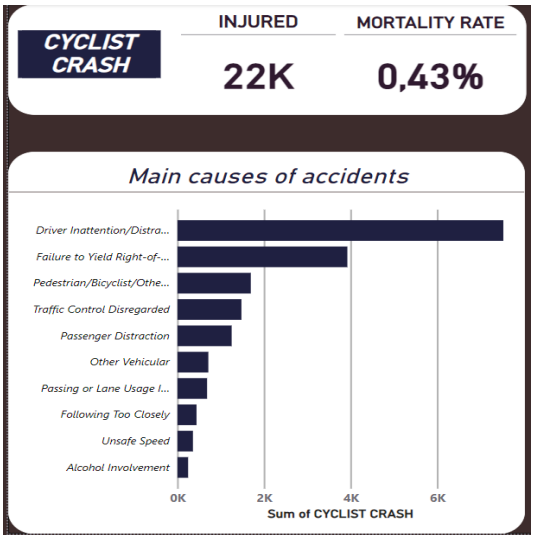


**Figura 4.6:** Seconda pagina report dei ciclisti

### 4.2.6 Analisi temporale e spaziale

Nel report sono presenti diversi tipi di visualizzazioni che permettono di effettuare analisi temporali e spaziali. La *Figura 4.7* presenta, nella parte superiore, un filtro dei dati che consente di selezionare uno o più quartieri per filtrare i valori visualizzati nella pagina. Nella parte inferiore troviamo una "mappa a bolle" che mostra i luoghi degli incidenti dei ciclisti suddivisi in base alla strada. La dimensione delle bolle varia in base al numero di incidenti avvenuti in quel luogo. La posizione viene calcolata effettuando una media della latitudine e della longitudine, prendendo poi come località la colonna Intersection.

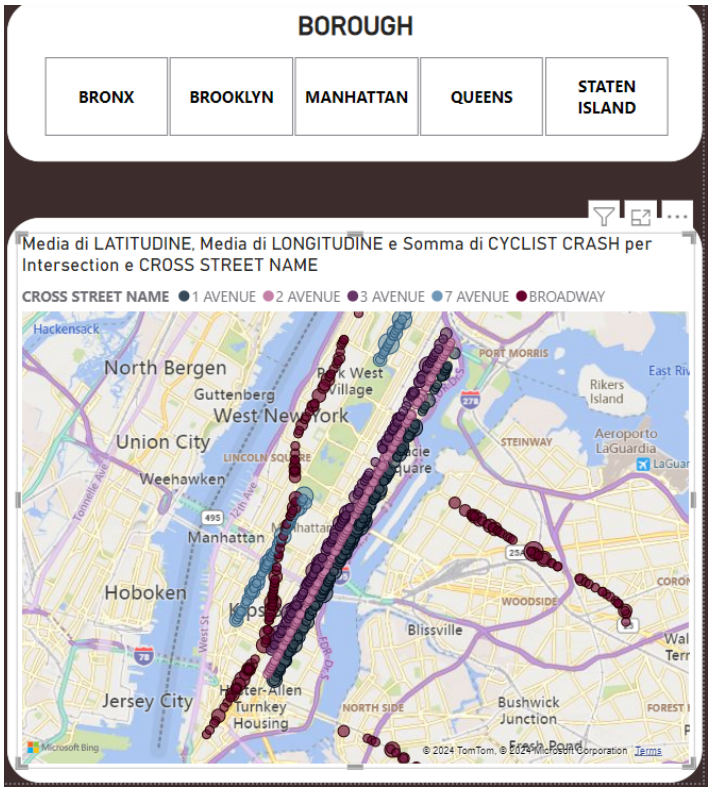


**Figura 4.7:** Analisi spaziale nel report dei ciclisti

Nella *Figura 4.8*, sulla parte sinistra, è presente un grafico a linee che mostra l'andamento della media annuale degli incidenti a partire dal 2014. La legenda definisce il colore delle linee per ogni categoria. Nella parte destra troviamo un grafico a nastri che mostra la variazione della percentuale di incidenti a partire dal 2018. Il colore indicato per le categorie è lo stesso definito nella legenda del grafico precedente; in questo modo associamo la categoria ad un singolo colore facilitando l'approccio visivo. Entrambi i grafici evidenziano che la categoria dei motociclisti è soggetta a un numero maggiore di incidenti rispetto alle altre categorie.

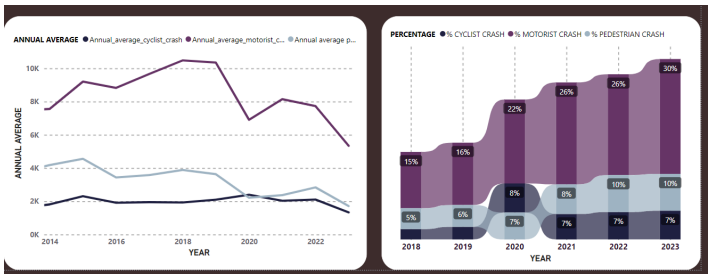


**Figura 4.8:** Analisi temporale sulle categorie

# CAPITOLO 5

## Estrazione di pattern di conoscenza complessi

*In questo capitolo verranno presentate le operazioni eseguite per estrarre modelli o schemi ricorrenti all'interno del dataset. Si partirà con l'introduzione del concetto di pattern, del quale verrà data una definizione esplorandone le caratteristiche. Successivamente, verrà descritto nel dettaglio il processo di Data Mining, ponendo particolare attenzione alle tecniche più utilizzate. Infine, esamineremo nel dettaglio i grafici a dispersione delle tre categorie di utenti della strada, osservando la tecnica del clustering utilizzata sui nostri dati in Power BI.*

## 5.1 Pattern

Pattern è un termine inglese che significa "modello". Nell'ambito dell'Intelligenza Artificiale e del Data Mining viene utilizzato per descrivere un modello o uno schema ricorrente. In generale, indica la ripetizione di una determinata sequenza all'interno di un insieme di dati grezzi. In sintesi, il pattern indica un'informazione nascosta all'interno di una grande mole di dati. Le caratteristiche comuni che identificano un pattern sono:

- *Rilevanza*: il pattern deve garantire una certa utilità affinché l'utente possa intraprendere azioni di conseguenza. Alcuni pattern vengono utilizzati per fare previsioni sul comportamento futuro dei dati o del fenomeno in esame.
- *Interpretabilità*: deve essere comprensibile dall'utente dal punto di vista sintattico e semantico.
- *Ripetitività e consistenza*: deve essere ripetitivo all'interno dei dati o del fenomeno in esame, manifestandosi più volte ed esprimendo un modello ricorrente. Tuttavia, potrebbe anche rappresentare un modello eccezionale.
- *Unicità*: il pattern deve presentare una caratteristica distintiva che lo rende precedentemente sconosciuto.

## 5.2 Data Mining

L'estrazione dei dati, o data mining, rappresenta l'insieme delle tecniche e delle metodologie che hanno per oggetto l'estrazione di informazioni utili da grandi quantità di dati grezzi, come "Data Warehouse" e "Data Lake". In sostanza, può essere definito come un processo di analisi eseguito su basi di dati, preceduto da altre fasi di trasformazione e filtraggio dei dati.

Mentre gli esseri umani sono in grado di riconoscere pattern e relazioni tra i dati, le macchine possiedono la capacità di processare grandi quantità di dati in tempi brevi. Se l'abilità umana fosse combinata con la velocità di calcolo degli elaboratori, questi ultimi sarebbero in grado di elaborare enormi quantità di dati con un intervento minimo, o addirittura nullo, da parte dell'uomo. Questo è il concetto alla base del Machine Learning, o apprendimento automatico, che viene anche chiamato Data Mining. L'apprendimento automatico è strettamente correlato al Data Mining, infatti, la scoperta di pattern è paragonabile all'apprendimento del sistema di Data Mining. L'apprendimento automatico si occupa di creare algoritmi che siano in grado di apprendere da un insieme di dati e fare delle predizioni costruendo un modello in modo induttivo. Questa pratica consiste nel trasformare grandi insiemi di dati grezzi in conoscenza (KDD – Knowledge Discovery in Data) attraverso l'identificazione di modelli e schemi ricorrenti. Il KDD è il processo, solitamente interattivo e iterativo, di ricerca, estrazione ed interpretazione di pattern dai dati. Questo prevede l'applicazione di algoritmi e tecniche di Data Mining e la comprensione dei pattern generati da tali algoritmi. I principali passi della sequenza di KDD, mostrati in *Figura 5.1*, sono:

- Identificazione dell'obiettivo che si vuole raggiungere.
- Preselezione dei dati validi.
- Pulizia dei dati e pre-elaborazione: prevede la separazione fra dati utili e inutili, il trattamento dei campi incompleti o vuoti ed infine la selezione delle informazioni fondamentali per il modello di riferimento.
- Trasformazione: si verifica il formato con il quale sono rappresentati i dati; se questo non è supportato, i dati devono essere convertiti.
- Data mining: viene selezionato il software più adatto, il quale scandaglia il Data Warehouse in modo selettivo per fornire la risposta cercata. Il Data Mining solitamente si compone di più sottopassaggi, anche ripetuti diverse volte, per affinare la procedura e verificare man mano i risultati raggiunti.
- Interpretazione dei risultati: se il processo non è andato a buon fine si procede con la reiterazione del passo precedente e, talvolta, anche di altri.
- Visualizzazione dei risultati in un formato comprensibile.

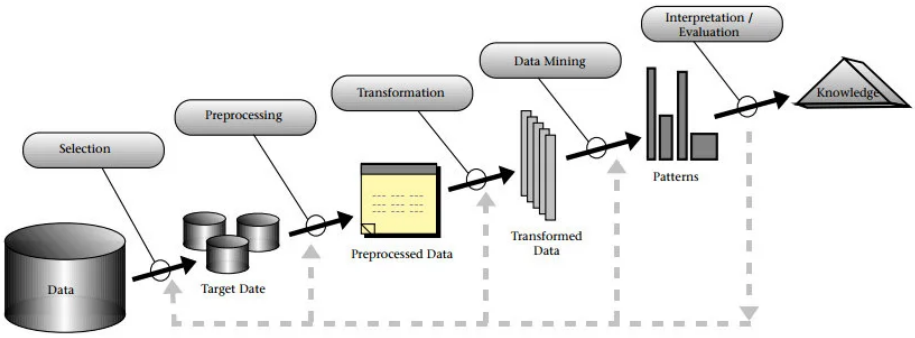


**Figura 5.1:** Sequenza di fasi del processo di KDD

Tuttavia, occorre tenere presente che il processo di Data Mining è sempre sottoposto al rischio di identificare relazioni causali errate, che, in seguito, si rivelano inesistenti. Tutte

le maggiori piattaforme della Big Data Analytics utilizzano oggi il data mining. Il processo di analisi può essere semi-automatico o interamente automatico e da questo si ricavano dei pattern. Pertanto, il data mining offre gli strumenti per sfruttare appieno Big Data e trasformarli in conoscenza fruibile. I modelli, o "task", di Data Mining possono essere suddivisi in due categorie:

- *Data Mining Descrittivo*: questo tipo di modello si occupa di descrivere e comprendere i dati storici. Lo scopo è quello di estrarre informazioni utili per identificare pattern, strutture o relazioni interessanti presenti nei dati. Il Data Mining Descrittivo utilizza diverse tecniche, come il clustering, l'analisi delle associazioni e l'analisi delle sequenze.
- *Data Mining Predittivo*: questo tipo di task utilizza i dati storici per fare previsioni o classificazioni su eventi futuri o su nuovi dati. L'obiettivo principale è quello di costruire modelli predittivi che possano essere utilizzati per prendere decisioni informate o per anticipare comportamenti futuri. I campi di applicazione del Data Mining Predittivo sono molteplici. Tecniche come la regressione, la classificazione e le reti neurali possono predire il prezzo delle azioni o diagnosticare una malattia basandosi sui sintomi. In sintesi, questo task si basa sullo studio di modelli matematici e statistici al fine di prevedere o stimare dei risultati e scenari che hanno luogo in futuro.

Il termine "task" nel contesto del Data Mining si riferisce alle attività specifiche che vengono eseguite per ottenere determinati risultati o obiettivi. La scelta del task dipende dal tipo di dati a disposizione e dalla tipologia di pattern che si desidera ottenere. Nella prossima sottosezione daremo uno sguardo più approfondito alle principali tecniche di Data Mining utilizzate all'interno dei task.

### 5.2.1 Tecniche di Data Mining

Il Data Mining include una vasta gamma di tecniche e algoritmi, ciascuno progettato per eseguire specifiche operazioni di analisi e scoperta sui dati. Al giorno d'oggi sono numerose le tecniche di Data Mining basate su diversi campi dell'apprendimento come il Machine Learning, la matematica e l'analisi statistica. Tra le tecniche più utilizzate ci sono:

- *Analisi delle associazioni*: questa tecnica utilizzata nel Data Mining Descrittivo, consente di determinare delle regole di implicazione logica all'interno dei dati. In questo modo è possibile stabilire le associazioni tra diverse variabili.
- *Classificazione*: questa tecnica di Data Mining Predittivo prevede l'organizzazione dei dati in classi. Queste ultime sono predeterminate sulla base di un modello di ordinamento.
- *Alberi decisionali*: sono particolari classificatori che, mediante una struttura ad albero, permettono di identificare in ordine di importanza le cause che portano al verificarsi di un evento. Nel Data Mining gli alberi di decisione sono utilizzati per diversi task come la segmentazione, la classificazione, la regressione e le serie storiche.
- *Clustering*: questa tecnica, tipica del Data Mining Descrittivo, prevede il raggruppamento degli elementi di un insieme, a seconda delle loro caratteristiche, in classi, o "cluster", non assegnate a priori. Viene utilizzato per identificare gruppi omogenei e modelli/tendenze all'interno dei dati.
- *Serie temporali*: permettono l'individuazione di pattern ricorrenti o atipici nei dati raccolti sequenzialmente nel tempo. Solitamente vengono utilizzate per effettuare analisi di tendenza e analisi predittive.

- *Regressione*: questa tecnica, ampiamente utilizzata nel mining predittivo, modella la relazione tra una o più variabili predittore indipendenti e una variabile risposta dipendente. La regressione può essere applicata per risolvere sia problemi lineari che problemi non lineari.

I principali task di Data Mining sono i seguenti:

- *descrizione di classi e concetti*;
- *scoperta di regole associative*;
- *classificazione e predizione*;
- *clustering*;
- *analisi degli "outlier"*;
- *analisi evolutive*;
- *web mining*;
- *Social Network Analysis*.

Sono disponibili diversi linguaggi di programmazione come *R* e *Python* per eseguire operazioni di Data Mining. Questi offrono una serie di strumenti e librerie, come "scikit-learn" per Python e "arules" per R. Inoltre, sono disponibili strumenti commerciali e pacchetti open-source che offrono un'interfaccia intuitiva per eseguire analisi dei dati e costruire modelli predittivi. Tra i più famosi abbiamo: RapidMiner, Oracle Data Miner, Microsoft SQL Server e SPSS Clementine.

## 5.3 Clustering in Power BI

In Power BI, il clustering è una tecnica utilizzata per raggruppare insieme dati simili in base alle loro caratteristiche comuni. Questa tecnica viene eseguita automaticamente dal software e può essere utile per identificare pattern e segmentare i dati in gruppi significativi per l'analisi. Inizialmente bisogna selezionare l'icona di grafico a dispersione nella sezione *Visualizzazione*. Successivamente, è necessario trascinare e rilasciare i due parametri che si desidera confrontare nel riquadro *Campi*; da qui verrà generato un semplice grafico a dispersione. Questo è formato da due assi cartesiani e da una serie di punti in corrispondenza dei valori assunti dai dati. Selezionando l'icona a tre punti presente nell'angolo a destra della visualizzazione, troveremo la voce *Trova automaticamente i cluster*. Una volta cliccata, si aprirà una schermata nella quale sarà possibile assegnare un nome e scegliere il numero dei cluster da creare. I grafici a dispersione mostrano il rapporto tra due variabili continue, rappresentando una variabile sull'asse x e l'altra sull'asse y. I grafici a dispersione più comuni sono:

- *Relazione crescente*;
- *Relazione decrescente*;
- *Nessuna relazione*;
- *Relazione curva*;
- *Presenza di Outlier*.

I grafici a "bolle", creati per ogni categoria di utenti della strada, presentano sull'asse delle ascisse la somma dei feriti e su quello delle ordinate la somma delle vittime. I cluster generati automaticamente vengono aggiunti al campo Dati con il nome "Main Causes of" seguito dalla categoria a cui si riferiscono. Questi si ottengono inserendo come valore la colonna "Contributing Factor Vehicle"; in questo modo possiamo raggruppare in cluster le cause principali di incidenti in base al numero di vittime e di feriti.

I cluster creati per la categoria dei ciclisti sono tre e li distinguiamo in base al colore delle bolle. Come possiamo vedere in Figura 5.2, la retta generata dall'algoritmo di clustering ha una pendenza crescente. Questo ci suggerisce che i cluster tendono a discostarsi l'uno dall'altro, mentre i dati in essi contenuti sono più omogenei e simili tra loro. La causa principale di incidente per i ciclisti è indicata come Driver Inattention, o disattenzione stradale. Questo potrebbe rappresentare un outlier in quanto si discosta notevolmente dagli altri e forma un unico cluster da solo. Il grafico dei pedoni, rappresentato in Figura 5.3, risulta molto simile a quello precedentemente descritto. In particolare, qui vengono creati esclusivamente due cluster. Ciò è dovuto dalla chiara separazione dei due gruppi di dati.

Infine, in Figura 5.4, troviamo il grafico dei ciclisti. Si può notare come la retta presenti una tendenza più orizzontale rispetto alle altre. Questa situazione potrebbe essere dovuta ai Cluster 1 e 3 che sono stati generati. In particolare, dati come Traffic Control Disgregated e Unsafe Speed per il Cluster 1, e Driver Inattention per il Cluster 3, tendono a distribuirsi, rispettivamente, lungo l'asse delle y e delle x. Ciò indica delle situazioni atipiche, in quanto non è possibile derivare una relazione tra le variabili rappresentate sull'asse delle x e delle y che mostrino una tendenza crescente come nei grafici precedenti. Pertanto, se al crescere dei feriti non aumenta il numero delle vittime, questa relazione potrebbe essere influenzata da ulteriori fattori o variabili da tenere in considerazione.

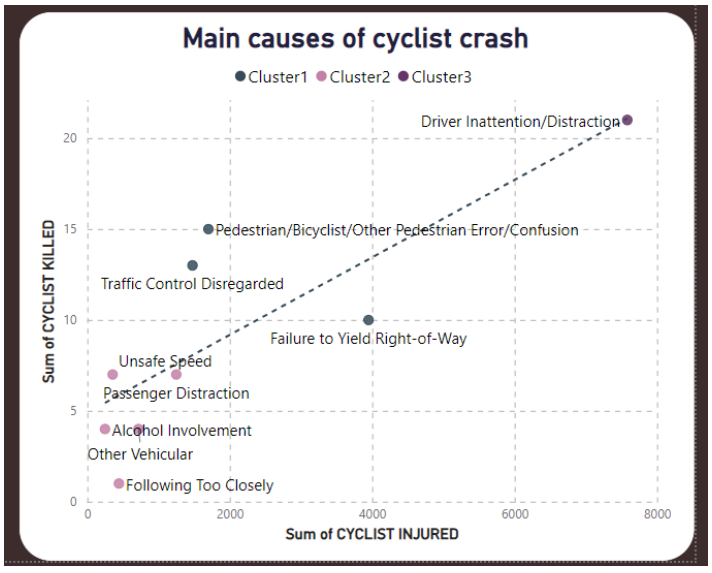


**Figura 5.2:** Grafico a bolle *Cyclist Crash*

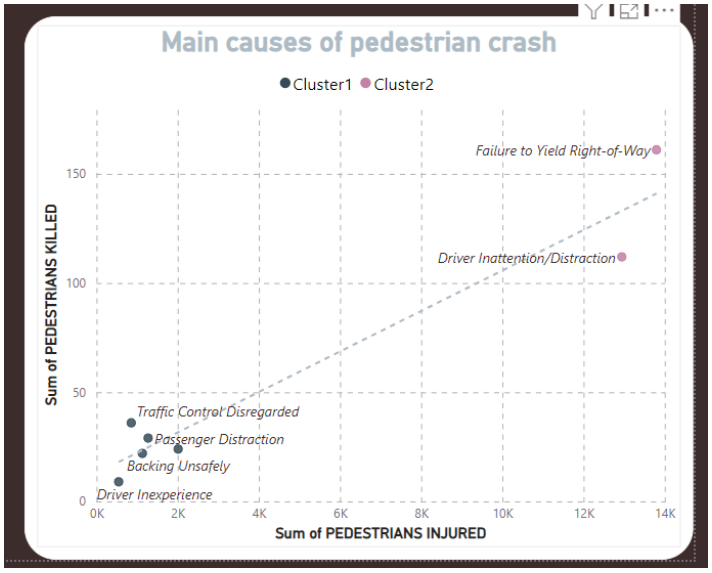


**Figura 5.3:** Grafico a bolle *Pedestrian Crash*

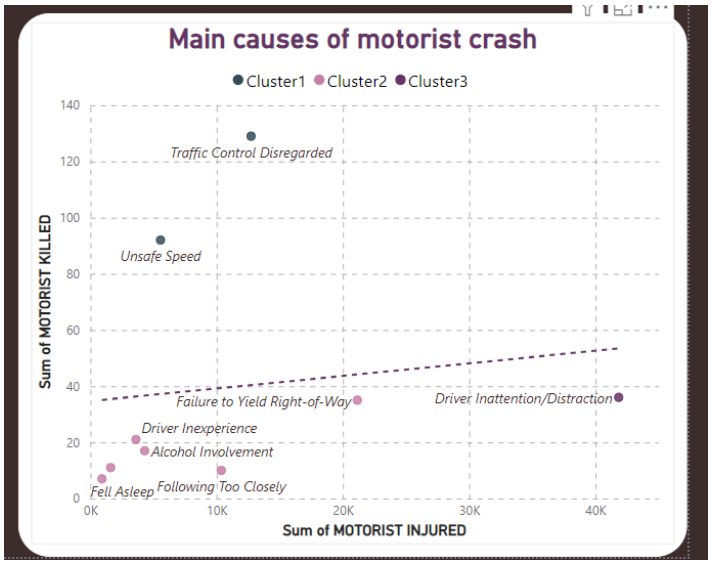


**Figura 5.4:** Grafico a bolle *Motorist Crash*

# CAPITOLO 6

# Discussione in merito al lavoro svolto

*In questo capitolo conclusivo saranno discussi i punti di maggior interesse emersi nel corso dell'analisi. Inizialmente verrà trattata l'analisi dei dati dal punto di vista pratico, approfondendo i software utilizzati e le difficoltà incontrate. Successivamente, saranno descritti i risultati ottenuti dalla campagna di Data Analytics, seguiti da alcune possibilità per ampliarla ulteriormente.*

## 6.1 Considerazioni pratiche sull'analisi dei dati

Gli strumenti utilizzati per l'implementazione di questa campagna di Data Analytics sono molteplici. Kaggle, gestito da Google LLC, si è rivelato una risorsa preziosa per la ricerca e l'accesso a dataset di alta qualità. La piattaforma conta una community di oltre 15 milione di utenti, i quali possono aumentare il proprio livello grazie ad un sistema di progressione. In questo modo si motiva l'utente a migliorare le proprie competenze e, allo stesso tempo, a contribuire allo sviluppo della piattaforma. Kaggle facilita la ricerca del dataset appropriato grazie a diverse funzionalità di filtraggio. Con queste è possibile selezionare un set di dati in base a parametri specifici, come la dimensione del file, il rating e i tag. Ciò garantisce la selezione di un set di dati aggiornati e affidabili. I motivi principali che hanno portato alla scelta del dataset *Motor Vehicle Collisions - Crashes* sono due: un numero consistente di colonne e l'obiettivo della campagna. Il primo è stato essenziale per effettuare svariate operazioni di analisi. La sicurezza stradale, oggetto della campagna, è un argomento di grande rilevanza e attualità, che ha reso il dataset particolarmente interessante da approfondire.

### 6.1.1 Power BI

Il pacchetto software sviluppato da Microsoft, Power BI, viene riconosciuto come leader tra le piattaforme di analisi dei dati e Business Intelligence. Uno dei suoi maggiori punti di forza è la varietà dei servizi e delle applicazioni offerte come Power BI Desktop, Power BI Service e Power BI Mobile. In particolare, Power BI Desktop è stato determinante nel processo di analisi dei dati. Gli strumenti messi a disposizione per svolgere attività di ETL hanno consentito di lavorare in maniera rapida ed efficiente. Inoltre, l'editor di Power Query, grazie alle numerose funzionalità e all'interfaccia altamente intuitiva, garantisce un'esperienza utente fluida e dinamica riducendo la necessità di scrivere codice complesso.

### 6.1.2 Sfide e difficoltà incontrate

Durante il processo di analisi si sono presentate diverse difficoltà che sono state abilmente superate grazie all'efficacia del software, alle risorse disponibili e a solide capacità di problem solving. Una delle principali sfide è stata l'attività di trasformazione del dataset, che ha richiesto un tempo considerevole. Alcune colonne iniziali presentavano un formato dei dati non utilizzabile, come, ad esempio: *Crash Date*, *Latitude* e *Longitude*. Tuttavia, grazie alle funzionalità di Power Query e al linguaggio M, con poche righe di codice, è stato possibile adattare il formato per poter lavorare accuratamente con i dati di localizzazione. Inoltre, a causa della poca chiarezza nelle indicazioni delle strade nelle colonne *On street*, *Off Street* e *Cross Street*, è stato necessario un approfondimento sulla geografia e sulla toponomastica della città di New York.

## 6.2 Analisi dei risultati della campagna di Data Analytics

I dati raccolti durante la campagna sono rappresentati utilizzando i vari strumenti di data visualization presenti in Power BI. Gli oggetti visivi sono stati selezionati in modo da rendere la lettura dei report accattivante ed evidente, facilitando la comunicazione delle informazioni. L'obiettivo iniziale della campagna era promuovere l'uso della bicicletta come mezzo di trasporto sicuro. Grazie ai dati raccolti e alle visualizzazioni create in Power BI, i lettori sono stati informati dei pericoli associati alla guida di determinati veicoli, evidenziando i rischi a cui sono esposti i vari utenti della strada. Dai dati è emerso che i pedoni hanno il tasso di letalità più elevato, mentre i motociclisti registrano il maggior numero di incidenti e feriti. Per ogni categoria di utente si può affermare che la berlina risulta il veicolo maggiormente coinvolto in un sinistro e la disattenzione stradale è la principale causa di incidenti. Tra i mezzi di trasporto, la bicicletta e la berlina sono le opzioni più sicure, con un tasso di letalità pressoché uguale. Tuttavia, è importante fare alcune considerazioni. Il numero di incidenti che coinvolgono berline è di gran lunga superiore a quello delle biciclette; ciò vuol dire che il numero delle vittime è significativamente inferiore rispetto a quello dei feriti. Se immaginiamo un sinistro stradale in cui sono coinvolti un'auto, una bici ed una moto, è evidente che l'autovettura è sottoposta ad un rischio inferiore, data la sua struttura, senza considerare la dinamica dell'incidente. D'altra parte, i ciclisti ed i motociclisti sono molto più vulnerabili. Quindi considerando la pericolosità che apporta il veicolo agli altri utenti della strada la bicicletta risulta tra i mezzi di trasporto più sicuri sia per i conducenti che per gli altri utenti della strada.

### 6.2.1 Possibili sviluppi e prospettive future

Nell'era dei Big Data, l'analisi avanzata dei dati assume una rilevanza sempre più fondamentale per affrontare le sfide moderne. L'integrazione con l'Intelligenza Artificiale sta trasformando il processo di analisi ed interpretazione dei dati a supporto del decision-making o processo decisionale. L'uso di tecnologie avanzate, come il Machine Learning e l'analisi predittiva, offrono innumerevoli opportunità a sostegno della sicurezza stradale. Ad esempio, queste tecnologie possono fornire previsioni su incidenti basandosi su variabili come traffico, condizioni meteorologiche e comportamento dei conducenti o utenti della strada. Un ulteriore sviluppo potrebbe essere l'implementazione di sensori o dispositivi che monitorano il traffico in tempo reale, permettendo interventi tempestivi per prevenire possibili incidenti, o nella peggiore delle ipotesi, fornire assistenza immediata. Una prospettiva futura per ampliare il contenuto di questa campagna riguarda l'inquinamento prodotto dai mezzi di trasporto. L'analisi delle emissioni di Co2 correlate al traffico, soprattutto nelle aree urbane, è una problematica ormai diffusa. A tal proposito, i Big Data possono essere utilizzati per

monitorare i livelli di inquinamento ed elaborare soluzioni per migliorare la qualità dell'aria. In conclusione, l'integrazione tra Big Data e Machine Learning rappresenta un'opportunità per far fronte alle sfide attuali. Lo sviluppo di queste tecnologie richiederà un impegno costante e ingenti investimenti al fine di migliorare gli strumenti di analisi e l'accuratezza dei modelli predittivi.

# Conclusioni

Questa tesi ha illustrato nel dettaglio le procedure per lo svolgimento di una campagna di Data Analytics. I dati relativi agli eventi di arresto anomalo della città di New York sono stati raccolti dalla polizia e, successivamente, resi pubblici nel sito del governo americano. Il dataset è stato accuratamente selezionato dalla piattaforma Kaggle, dove è disponibile in diversi formati. Successivamente, una volta ottenuti i dati, questi sono stati caricati nel software Power BI, dove è stato avviato il processo di ETL. Dopo l'estrazione dei dati grezzi sono state svolte operazioni di data cleaning, di merging delle colonne e di normalizzazione dei dati. A questo punto, i dati trasformati sono stati caricati nel sistema di destinazione di Power BI.

Una volta terminato l'intero processo di ETL, sono stati utilizzati strumenti di Analisi Esplorativa dei Dati per approfondire la conoscenza del dataset. Tra i più importanti, la Data Visualization è stata fondamentale per la stesura dei report attraverso l'utilizzo di elementi visivi come diagrammi, grafici e mappe. Il linguggio DAX, strumento di Power BI Desktop, si è rivelato molto utile per la creazione di colonne personalizzate e per il calcolo delle percentuali. Infine, è stata realizzata la procedura di Clustering automatico, disponibile su Power BI, sui grafici a bolle creati per ogni categoria di utente della strada. In questo modo abbiamo raggruppato in Cluster le cause principali di incidenti in base al numero di vittime e di feriti. Questa tecnica di Data Mining Descrittivo è stata efficace per la scoperta di outlier.

L'esposizione di questa tesi si conclude, infine, con una discussione in merito al lavoro svolto. Inizialmente sono stati portati avanti delle considerazioni sul pacchetto software Power BI e tutti gli strumenti utilizzati messi a disposizione. Poi, c'è stato un breve riepilogo delle sfide e delle difficoltà che si sono palesate durante l'intero processo di analisi. Successivamente, sono stati esposti i risultati ottenuti dalla campagna di Data Analytics, che ha visto la conferma della bicicletta come mezzo di trasporto più sicuro per i guidatori e per gli altri utenti della strada.

Le prospettive future, derivanti dallo sviluppo di tecnologie avanzate basate sull'Intelligenza Artificiale, sono ricche di potenziale. Negli ultimi anni, i progressi nell'apprendimento automatico hanno accelerato l'integrazione dell'IA in diversi settori, tra cui i trasporti e la mobilità. Un esempio è la gestione intelligente del traffico che è già stata implementata con successo in diverse città, come Taichung, Vienna, York o Roma. Il software PTV Optima combina tecniche di Machine Learning con la modellazione dinamica del traffico; in questo modo, offre agli utenti previsioni affidabili sul traffico fino a 60 minuti in anticipo, in grado di identificare gli scenari migliori per prevenire gli ingorghi nel traffico e le chiusure stradali.

# Bibliografia


AHMED, S., HOSSAIN, M. A., RAY, S. K., BHUIYAN, M. M. I. e SABUJ, S. R. (2023), «A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance», *Transportation Research Interdisciplinary Perspectives*.

ANDERSON, C. (2008), «The End of Theory: The Data Deluge Makes the Scientific Method Obsolete», *Wired*, URL `https://www.wired.com/2008/06/pb-theory/`.

BELLINI, M. (2017), «Big data, cosa sono e come sono utili alle aziende», *BigData4Innovation*.

DROTOS, A. (2024), «8 Types of Data Analytics to Improve Decision-Making», *DataCamp*, URL `https://www.datacamp.com/blog/types-of-data-analytics-to-improve-decision-making`.

FANTINI, F. e NARAYANDAS, D. (2023), «Analytics for Marketers», *Harvard Business Review*, URL `https://hbr.org/2023/05/analytics-for-marketers`.

GIANOTTI, A. (2021), «Le mappe di Open Street Map, i Big data di Istat e gli incidenti stradali», *Il sole 24 ore*.

HONG, M. S., SUN, W., ANTHONY, B. W. e BRAATZ, R. D. (2022), «Teaching Process Data Analytics and Machine Learning at MIT», Rap. tecn., Massachusetts Institute of Technology, Boston.

JOSHI, A., KHOSRAVY, M. e GUPTA, N. (2020), *Machine Learning for Predictive Analysis: Proceedings of Ictis*, Springer Nature.

KIMBALL, R. e CASERTA, J. (2004), *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*, Wiley.

MAURO, A. (2019), *Big Data Analytics. Analizzare e interpretare dati con il machine learning*, Apogeo.

MÜLLER, A. e GUIDO, S. (2016), *Introduction to Machine Learning with Python: A Guide for Data Scientists*, O'Reilly Media.

SALVAGGIO, A. (2023), *Business Intelligence con Microsoft Power BI. Guida completa per l'analisi e la visualizzazione dei dati*, Edizioni LSWR.

SCITOVSKI, R., SABO, K., MARTÍNEZ-ÁLVAREZ, F. e ŠIME UNGAR (2021), *Cluster Analysis and Applications*, Springer Cham.

TAN, P.-N., STEINBACH, M., KARPATNE, A. e KUMAR, V. (2006), *Introduction to Data Mining*, Addison-Wesley, Boston, MA.

THOMAS, W. S. (2020), «Power BI: An analytical view», *Journal of Accountancy*, URL https://www.journalofaccountancy.com/issues/2020/mar/microsoft-power-bi-data-excel.html.

## Siti web consultati

- Microsoft Learn: che cos'è power bi? – https://learn.microsoft.com/it-it/power-bi/fundamentals/power-bi-overview
- Kaggle – www.kaggle.com
- Power BI: What is Data Visualization? – https://powerbi.microsoft.com/en-us/data-visualization/
- Motor Vehicle Collisions - Crashes – https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes
- Data Mining: strumenti e tecniche di data mining –https://www.intelligenzaartificiale.it/data-mining
- Wikipedia – www.wikipedia.org

# Ringraziamenti

La maggior parte della persone si chiede spesso cosa sia più importante fra il traguardo o il viaggio stesso. La risposta è semplice, nessuno dei due. Le persone di cui ti circondi durante il tuo percorso lo rendono speciale ed unico. Così vorrei dedicare la parte finale del mio primo traguardo a tutti coloro che mi sono stati vicini. In primis ci tengo a ringraziare di cuore la mia famiglia, che ha dovuto sopportare tutte le mie sfuriate dovute agli esami; vi sarò sempre immensamente grato di avermi sostenuto dall'inizio alla fine.

Un ringraziamento speciale và alla mia casa, Corso Mazzini 8. Un luogo testimone di innumerevoli momenti di studio e contemplazione. A tutti i miei ex-coinquilini, Andrea, Matteo, Rip, Cheste, Stefano. Rise e Marco un pò di meno.

Un sentito ringraziamento al il mio professore Domenico Ursino, per avermi guidato con tanta pazienza e dedizione nello svolgemiento di questa tesi. Vorrei anche ringraziare il Dott. Luca Virgili per la sua professionalità e il suo supporto.

Infine il più importante. A mia nonna, che mi ha chiamato prima e dopo ogni esame che ho dato. Spesso ero sempre arrabbiato per una prova andata male, ma a prescindere dal risultato, eri lì ad ascoltarmi e a darmi la forza di andare avanti.

Da te, 3 anni fa.