

ELABORATO DI MACHINE LEARNING & DATA MINING

CLASSIFICAZIONE DEI TIPI DI CRIMINE DI SAN FRANCISCO



Università degli Studi di Brescia

Dipartimento di Ingegneria dell’Informazione

Corso di Laurea Magistrale in Ingegneria Informatica



BETTINELLI LUCA 85947

MARIO MATTEO 86015

MASCIALE MICHELE 85982

# indice dei contenuti

[indice dei contenuti II](#_Toc443663163)

[Capitolo 1 – Descrizione del problema 1](#_Toc443663164)

[1.1 - Introduzione 1](#_Toc443663165)

[1.2 – Struttura del dataset 1](#_Toc443663166)

# Capitolo 1 – Descrizione del problema

## 1.1 - Introduzione

Dal 1934 al 1963, San Francisco era nota come patria di alcuni dei criminali più famigerati al mondo. Oggi la città è famosa più per la sua tecnologia che non per il suo passato criminale. Tuttavia, con la crescente disuguaglianza sociale, carenza di alloggi e proliferazione di mezzi tecnologici, tra cui BART (Bay Area Rapid Transit District), non vi è certamente scarsità di crimine. Dai distretti di Sunset a South of Market, da quelli di Marina a Excelsior, il dataset a nostra disposizione offre quasi 13 anni di notizie di reati da tutti i quartieri di San Francisco. Data l’ora e la posizione in cui si sono verificati i reati, è necessario prevederne la categoria. Si tratta di una delle competizioni accessibili a chiunque sul sito di Kaggle.

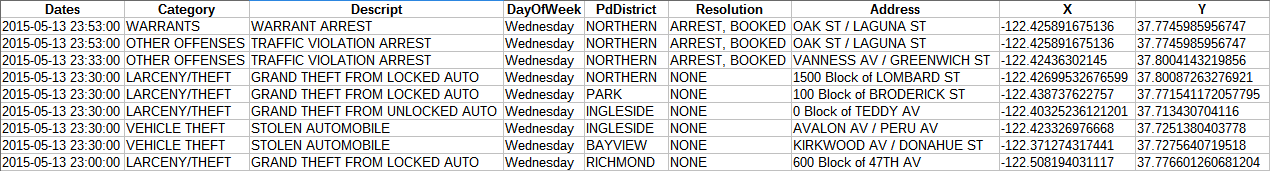
## 1.2 – Struttura del dataset

Il dataset a nostra disposizione è suddiviso in due file in formato CSV: il training set (circa 125 MB) e il test set (circa 89 MB). Questo dataset contiene le informazioni su reati e crimini registrati dal sistema di report del dipartimento di polizia di San Francisco. In particolare sono stati raccolti i dati del periodo che va dal 01/01/2003 al 13/05/2015 e sono stati successivamente distribuiti nei due file a settimane alterne: le settimane dispari appartengono al test set, quelle pari al training set.

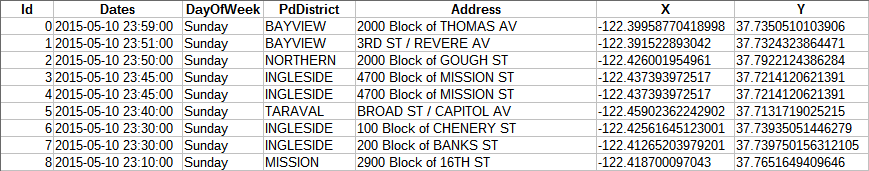
Di seguito l’elenco degli attributi che descrivono un crimine nel dataset:

* **Id**: un numero identificativo auto-incrementale presente solo nel test set;
* **Dates**: una stringa che rappresenta il timestamp, ovvero la data in cui è avvenuto il crimine;
* **DayOfWeek**: una stringa che rappresenta il giorno della settimana in cui è avvenuto il crimine;
* **PdDistrict**: una stringa che rappresenta il nome del distretto nel cui dipartimento di polizia si è registrato il crimine;
* **Address**: una stringa che rappresenta l’indirizzo della via in cui è avvenuto il crimine;
* **X**: un numero che rappresenta la longitudine;
* **Y**: un numero che rappresenta la latitudine;
* **Category**: una stringa che rappresenta la categoria del crimine, presente solo nel training set. Questa è proprio la variabile target da predire;
* **Descript**: una stringa che rappresenta una descrizione dettagliata del crimine, presente solo nel training set;
* **Resolution**: una stringa che rappresenta la modalità con cui il crimine è stato condannato, presente solo nel training set.

In *Figura 1.1* e *Figura 1.2* è riportato un esempio di dataset, estratto da entrambi i file:



*Figura 1.1: Un estratto del training set.*

**

*Figura 1.2: Un estratto del test set.*