Report analisi predittiva dei crimini a Chicago

Introduzione

L’obiettivo è di analizzare i dati sui crimini avvenuti a Chicago e costruire un modello predittivo per stimare la probabilità che si verifichi un furto (THEFT) nei diversi quartieri della città.
Il dataset utilizzato è quello pubblico “Crimes in Chicago”, che raccoglie milioni di segnalazioni dal 2001 in poi, descrivendo tipologia del reato, localizzazione geografica, data e altri dettagli utili a comprendere il fenomeno criminale.

Analisi Esplorativa del Dataset

Il dataset originale contiene oltre 7 milioni di record con numerose colonne descrittive: tipo di crimine, data, coordinate geografiche, quartiere (Community Area), descrizione del luogo, ecc.
Per facilitare la fase di sviluppo e ottimizzare le prestazioni, è stato caricato un campione casuale da 100.000 record, mantenendo comunque una distribuzione rappresentativa delle principali tipologie di crimine.

Selezione delle Colonne Rilevanti

Colonna	Rilevanza	Motivazione
ID	No	Identificativo tecnico, non utile per la predizione.
Case Number	No	Solo identificativo, non informativo per il modello.
Date	Sì	Fondamentale per analisi temporali e trend stagionali.
Block	No	Indirizzo parziale, poco informativo rispetto a coordinate e quartiere.
IUCR	No	Codice tecnico, ridondante rispetto a Primary Type.
Primary Type	Sì	Variabile target o predittiva principale (tipologia del crimine).
Description	No	Dettaglio testuale, poco utile per modelli predittivi standard.
Location Description	Sì	Utile per analisi contestuale e spaziale.
Arrest	No	Esito dell’evento, non predittivo del crimine stesso.
Domestic	No	Specifico solo per alcuni crimini.
Beat, District	Sì	Identificano zona geografica, utili per aggregazione e previsione.
Ward, Community Area	Sì	Aggregazione per quartiere/area amministrativa.
FBI Code	No	Codifica ridondante rispetto a Primary Type.
X/Y Coordinate	No	Coordinate tecniche, preferibili latitudine e longitudine.
Year	Sì	Utile per analisi temporali e trend.
Updated On	No	Non rilevante per la predizione.
Latitude, Longitude	Sì	Necessarie per visualizzazione geografica e clustering spaziale.
Location	Sì	Alternativa diretta a latitudine/longitudine per mappe.

Selezione delle colonne

Per costruire un modello predittivo efficace, è stato necessario valutare criticamente le colonne del dataset e scegliere solo quelle più rilevanti.

- Primary Type
- Community Area
- Location Description
- Year, Month, Hour
- Latitude, Longitude

Pulizia dei Dati

Le principali operazioni di pulizia previste sono:

- Rimozione dei duplicati.
- Gestione dei valori nulli: eliminati record con valori mancanti in colonne chiave (Primary Type, Community Area, Location Description, Date, Latitude, Longitude).
- Conversione delle date in formato datetime.
- Creazione di nuove feature temporali: Year, Month, Hour estratte da Date.

Colonne Irrilevanti e Motivazione dell'Esclusione

- ID, Case Number, IUCR, FBI Code: identificativi e codici tecnici non utili per la previsione.
- Description, Arrest, Domestic: dettagli o esiti non predittivi del verificarsi del crimine.
- Block, X/Y Coordinate, Updated On: ridondanti o poco informativi rispetto ad altre colonne scelte.

Campionamento

Per motivi di performance, è stato usato un campione casuale di 100.000 record. Questo ha permesso di:

- Ridurre drasticamente i tempi di elaborazione.
- Mantenere una distribuzione simile a quella del dataset completo.
- Iterare rapidamente tra analisi, visualizzazioni e tuning del modello.

Motivazione delle Scelte Effettuate

- Colonne selezionate: scelte per massimizzare l'informazione predittiva e la chiarezza delle visualizzazioni (tipologia, posizione, tempo).
- Esclusione di colonne: motivata da ridondanza, natura puramente identificativa o mancanza di valore predittivo.
- Label scelta: probabilità che un certo tipo di crimine si verifichi in ciascun quartiere, perché risponde direttamente all'obiettivo di prevenzione e gestione delle risorse.

Descrizione del Modello AI Scelto

- Obiettivo del modello: stimare, per ciascun quartiere (Community Area), la probabilità che si verifichi un furto (THEFT).

- Algoritmo selezionato: Random Forest Classifier.

- Motivazione: è robusto, gestisce bene dati categoriali e numerici, permette di stimare l'importanza delle variabili e riduce il rischio di overfitting rispetto ad altri modelli.

  Feature utilizzate nel modello:

- Community Area (quartiere)
- Location Description (descrizione del luogo)
- Year, Month, Hour (estratte dalla colonna Date)
- Variabile target (label):
  is_theft: vale 1 se il reato è classificato come "THEFT" in "Primary Type", 0 altrimenti.

Discussione dei Risultati Ottenuti

Valutazione del modello:

- Metriche utilizzate: accuratezza, precision, recall, F1-score, ROC-AUC.
- Il modello mostra una recall molto bassa per la classe 'furto', segno che fatica a individuare correttamente questi eventi quando sono meno frequenti nel dataset. Questo deriva anche dallo sbilanciamento delle classi (i furti rappresentano solo circa il 20% dei casi nel campione).

Visualizzazioni:

- Heatmap geografiche → realizzate con Folium per rappresentare su mappa le zone con maggior densità di furti.
- Grafici temporali (andamento mensile, annuale, per fascia oraria) → costruiti con Matplotlib e Seaborn per evidenziare i picchi stagionali o orari.
- Distribuzioni per tipologia di crimine → istogrammi e countplot creati con Seaborn.

Limiti:

- Predizione meno accurata per crimini rari.
- Possibili bias dovuti a dati mancanti o non aggiornati.

Implicazioni pratiche:

- Il modello può supportare le forze dell'ordine nell'allocazione delle risorse.
- Le visualizzazioni aiutano a identificare pattern ricorrenti e zone critiche.

Conclusioni

Il modello sviluppato consente di stimare la probabilità di furti nei quartieri di Chicago, evidenziando pattern spaziali e temporali utili per la prevenzione. Miglioramenti futuri potrebbero includere tecniche di bilanciamento delle classi e l'utilizzo di modelli più complessi per migliorare la recall sui crimini meno frequenti.