



UNIVERSITÀ DI PISA



Università degli Studi di Pisa

Dipartimento di Informatica

Master II livello in BigData Analytics & Artificial Intelligence for Society

“Proactive Theft: Sviluppo di un modello di Machine Learning per l'individuazione di pattern ricorrenti di viaggi relativi a furti”

Il Candidato

Giulia Segurini

Il Relatore

Mirco Nanni

A.A. 2023/ 2024

Ho trovato la mappa del labirinto: checkmate.

Storytelling - Proactive theft: “Sviluppo di un modello di Machine Learning per l'individuazione di pattern ricorrenti di viaggi relativi a furti”

Grazie alla tecnologia innovativa e all'intelligenza artificiale, stiamo finalmente raggiungendo l'obiettivo di una realtà in cui il furto dei veicoli è solo un lontano ricordo. Sono entusiasta di condividere la mia esperienza di tirocinio presso Octo Telematics, un leader nel settore della telematica automobilistica, dove ho potuto contribuire a un progetto all'avanguardia: il sistema di supporto alla prevenzione dei furti di veicoli nell'ambito del progetto SVR: supporto veicoli rubati.

Durante il mio percorso di tirocinio presso Octo Telematics S.P.A., ho avuto l'opportunità di lavorare su diversi aspetti del progetto Drive Ability, focalizzandomi sull'analisi dei dati GPS provenienti da migliaia di veicoli equipaggiati con scatole nere per l'analisi delle abilità di guida. Questi dispositivi non sono solo strumenti di monitoraggio, ma veri e propri sentinelle in grado di rilevare improvvise deviazioni di percorso, accelerazioni brusche e comportamenti di guida irregolari. Ogni dato raccolto è un pezzo del puzzle che ci permette di costruire un modello di machine learning in grado di identificare pattern ricorrenti di furti, un passo fondamentale per prevenire questi crimini.

Ma non è stato tutto facile. Affrontare le sfide del progetto proactive Theft, soprattutto in tempi brevi, è stata un'esperienza formativa. La prima sfida riguardava la qualità dei dati GPS. Anche se i dati erano aggiornati in tempo reale, era necessario un processo di scrematura per garantire la loro affidabilità. La seconda sfida era l'allenamento dei modelli di machine learning: dovevamo assicurarci che fossero interpretabili e reattivi, per permettere interventi umani tempestivi quando necessario. Infine, la terza sfida riguardava il controllo dei falsi positivi e negativi, un aspetto cruciale per garantire l'efficacia del sistema.

Per affrontare queste sfide, ho proposto un modello di machine learning che ha rappresentato un significativo miglioramento nell'identificazione preventiva dei furti. Questo modello non solo riconosce automaticamente i pattern di rischio, ma se interpolato con altri modelli di ensemble potrà fornire anche avvisi in tempo reale, consentendo alle forze dell'ordine e alle compagnie assicuratrici di intervenire rapidamente e ridurre i danni.

La mia esperienza presso Octo Telematics mi ha insegnato che il lavoro di squadra è fondamentale in un ambiente dinamico e veloce come quello della telematica automobilistica. Ogni membro del team ha un ruolo cruciale nel raggiungere obiettivi comuni, e sono grata di aver potuto contribuire a un progetto che non solo migliora la sicurezza dei veicoli, ma offre anche un supporto tangibile ai proprietari, agli erogatori di servizi di noleggio, alle Forze dell'Ordine.

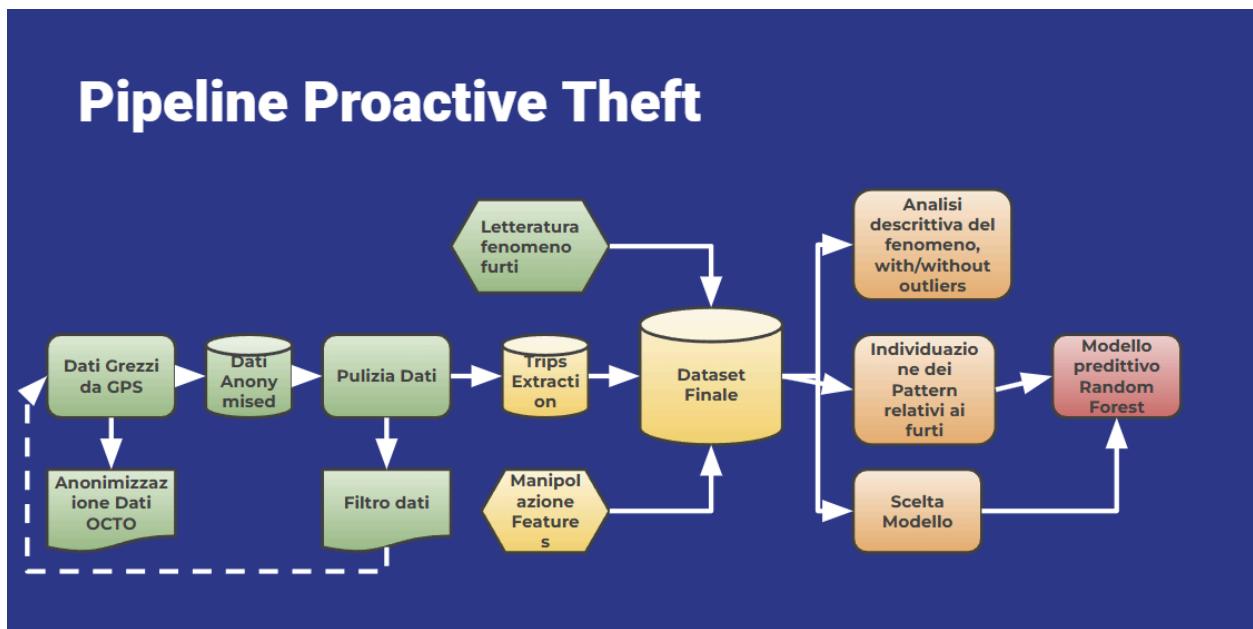
Il mio viaggio nel mondo della telematica automobilistica mi ha aperto gli occhi su come la tecnologia possa trasformarsi in un alleato nella lotta contro il furto dei veicoli. Con progetti come il nostro, il futuro della sicurezza automobilistica appare luminoso e promettente.

Durante le prime settimane mi sono occupata di entrare nel vivo di procedure aziendali relative alle abilità di guida, alla realizzazione di pipeline in python per l' analisi della qualità dei dati DQR e all' affinamento di analisi per un progetto di Sub1Hz Score, riguardante appunto il punteggio ottenuto dai guidatori come forma di incentivo per una guida in maggiore sicurezza per sè e per gli altri.



Con il 10 gennaio segnando l'inizio della mia immersione nel cuore del progetto Productive Test, ho avuto un solo focus: analizzare i dati telematici per estrarre caratteristiche rilevanti legate ai furti e costruire un modello predittivo basato su Decision Tree, che successivamente si sarebbe evoluto in un Random Forest. Questa fase iniziale, soprattutto il preprocessing, ha richiesto un'attenta pianificazione e un approccio metodico, poiché il successo del progetto dipendeva dalla qualità e dall'efficacia del modello che avrei sviluppato.

Durante la pausa natalizia, pur essendo in attesa del dataset, ho approfittato del tempo per impostare tabelle in SQL e gestire i database interni dell'azienda (Jira, Hue, Bitbucket). Quando finalmente ho ricevuto il dataset anonimizzato, composto da circa 2 milioni di righe e 13 colonne, ero pronta a entrare nel vivo dell'analisi.



Il pre-processing è diventato il mio campo di battaglia. Ho iniziato con la pulizia del dataset, eliminando colonne non rilevanti e gestendo valori mancanti e anomalie. La segmentazione dei viaggi è stata cruciale: ho identificato le pause prolungate e creato un identificatore unico per ogni viaggio, denominato "trip ID". Questo passaggio ha permesso di avere una visione chiara e dettagliata dei vari percorsi.

Una volta filtrata la qualità dei dati, mantenendo solo quelli con precisione elevata, ho iniziato a visualizzare i dati attraverso mappe che mostravano la distribuzione dei viaggi, evidenziando in

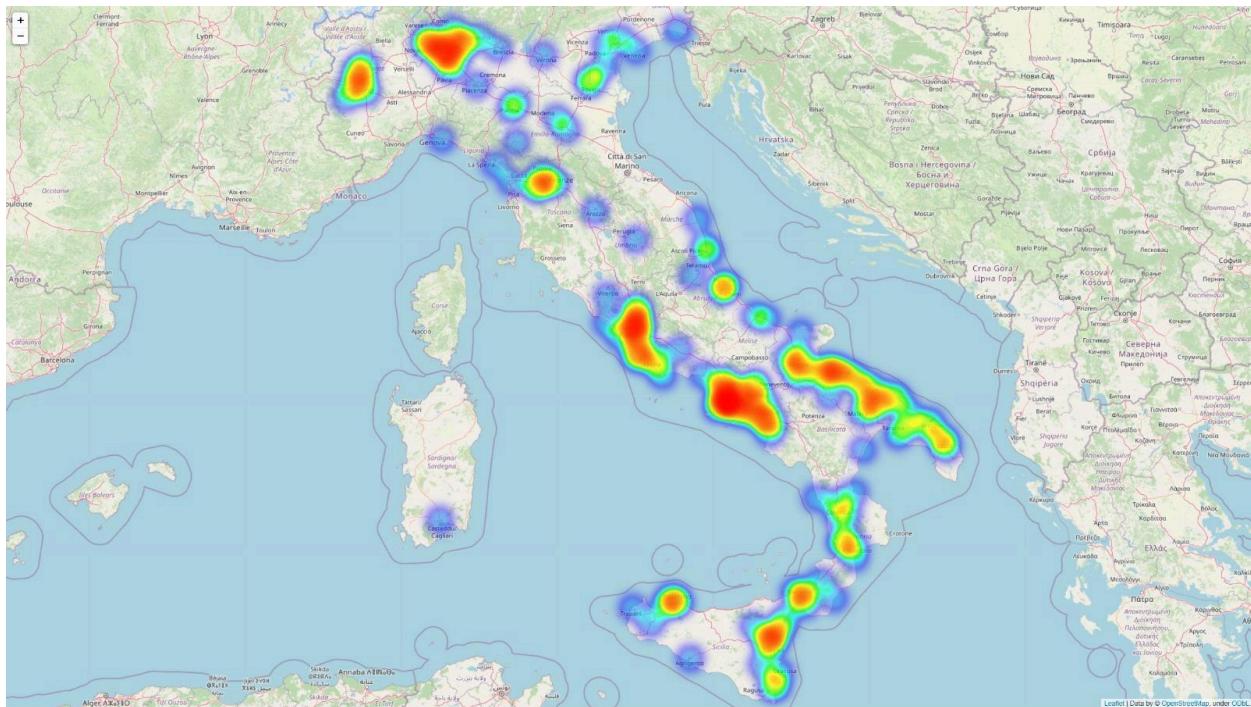
rosso i furti e in blu i percorsi. Questa visualizzazione non solo era informativa, ma anche coinvolgente, dando vita ai numeri e creando una storia visiva del fenomeno.



https://drive.google.com/file/d/1F6Xim5bEu10S7HOfqDwtXgczsZ-X_ZJc/view?usp=sharing

Nella mappa qui sopra rappresentata (versione zoommabile nel link cliccabile) sono rappresentati i tragitti associati a eventi di furto e non furto per diversi clienti. Le linee rette nel grafico possono indicare tragitti effettuati su traghetti o percorsi su acqua, oppure salti di posizione dovuti a traini o spostamenti del veicolo con il quadro spento. Se queste linee sono in azzurro, significa che in quel momento il veicolo non era soggetto a furto.

Viene riportata qui di seguito una mappa di calore furti ed il relativo link cliccabile:



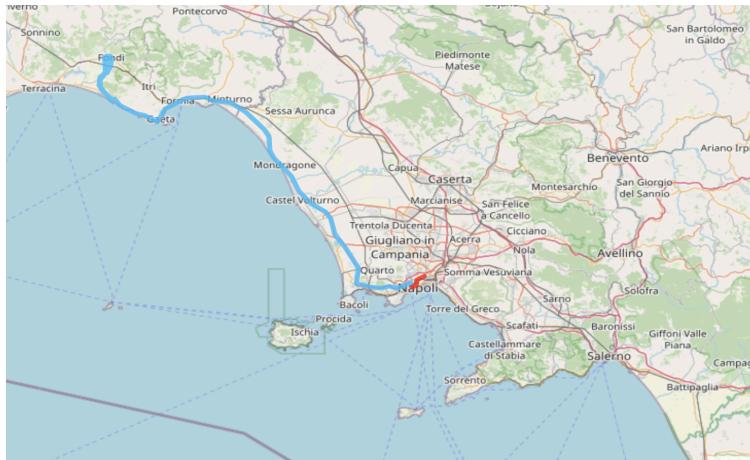
<https://drive.google.com/file/d/1nE-mGmjdG4TsJToaU4aSXWqL82PADRvl/view?usp=sharing>

L'intensità dei colori indica la densità degli eventi di furto registrati:

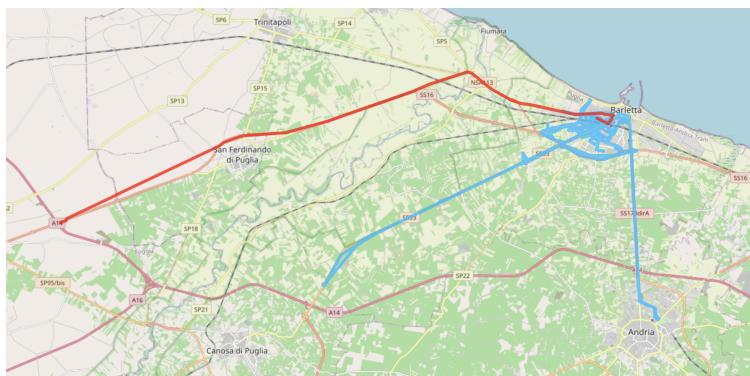
- Rosso e arancione → Aree con il maggior numero di furti segnalati
- Giallo e verde → Zone con una densità media di furti
- Blu → Zone con pochi furti registrati

La mappa mostra una distribuzione concentrata in alcune regioni specifiche, con punti caldi in Lombardia, Emilia-Romagna, Lazio, Campania e Sicilia. Questo suggerisce che i furti potrebbero essere più frequenti in determinate città o lungo specifici corridoi di transito.

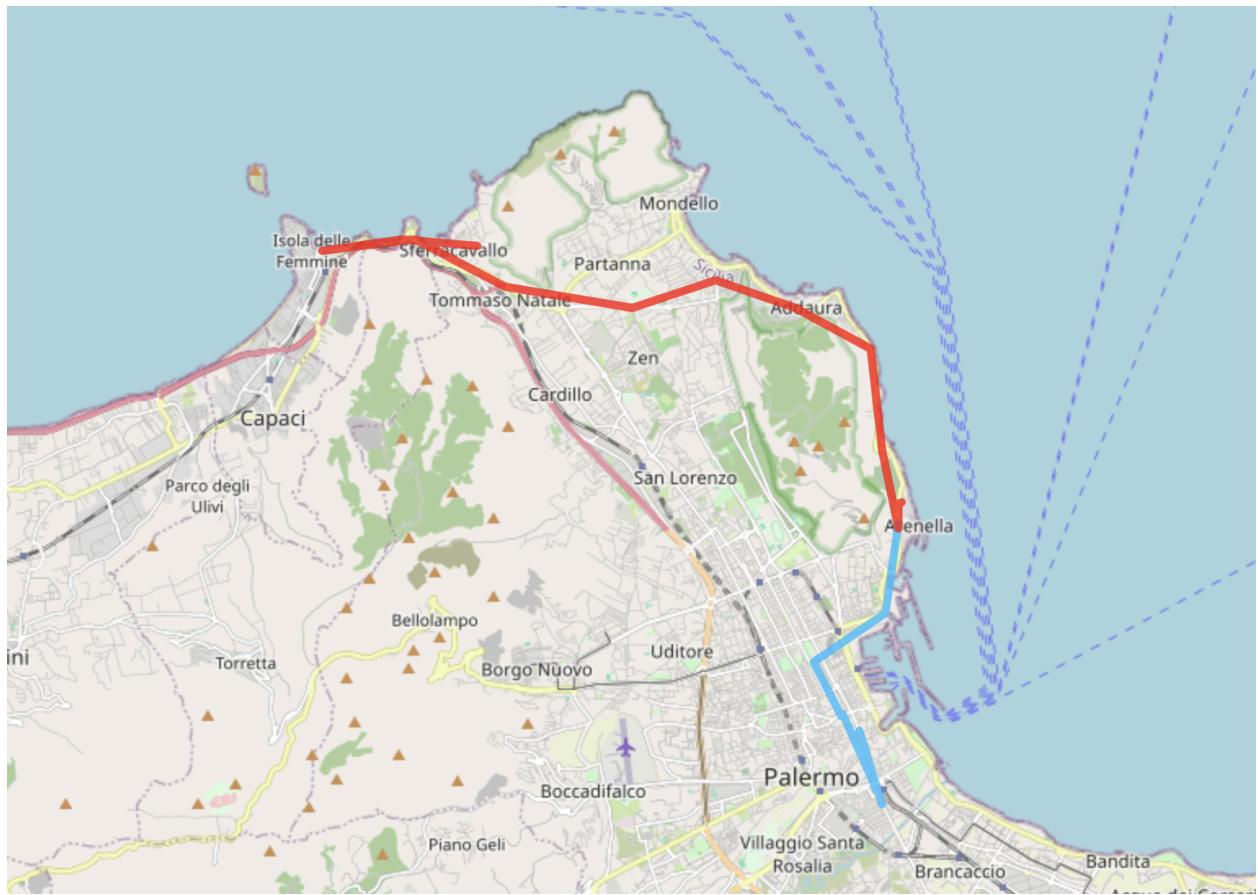
Qui di seguito invece viene riportato un esempio di mappatura di trip di singoli veicoli soggetti a furto: Qui invece immagini relativi a un esempi di trip per singoli veicoli soggetti a furto con relativo link alla mappa cliccabile:



<https://drive.google.com/file/d/1KdOjOsfwXalBrQyN5xwFl5mfZRvlNeP/view?usp=sharing>



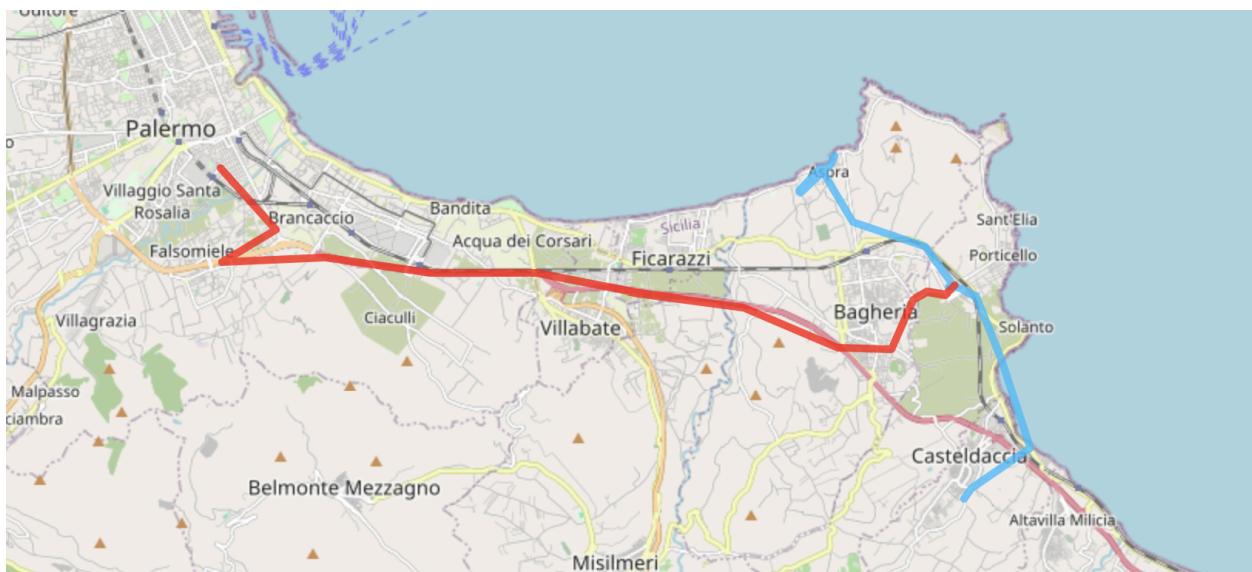
<https://drive.google.com/file/d/1GYZOjWsXgkg0zEbetcaAV23b5ETJmVCL/view?usp=sharing>



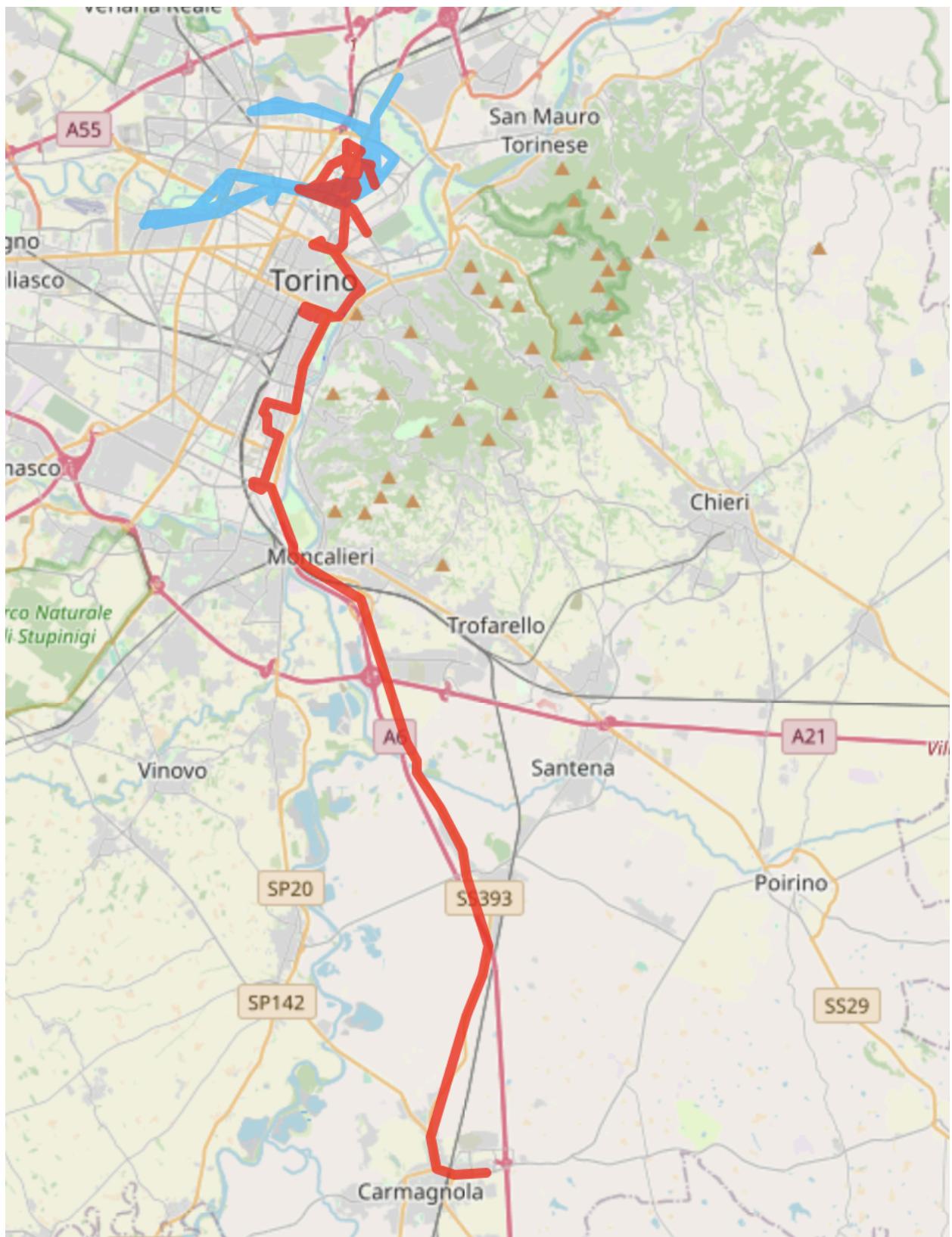
<https://drive.google.com/file/d/1sMnMs71M5FLZ7R10M4DjsKoI-9L9O2YR/view?usp=sharing>



<https://drive.google.com/file/d/1k-ivebme1rFuGgf7ndDTL8LqCE3wcVd/view?usp=sharing>



<https://drive.google.com/file/d/1PhbORAUHX5aEoaTeDDy1AxBlgi-RPIXo/view?usp=sharing>



<https://drive.google.com/file/d/1-cXqxU4omkBAtZOjmWkF8a-4b-pK6ec0/view?usp=sharing>

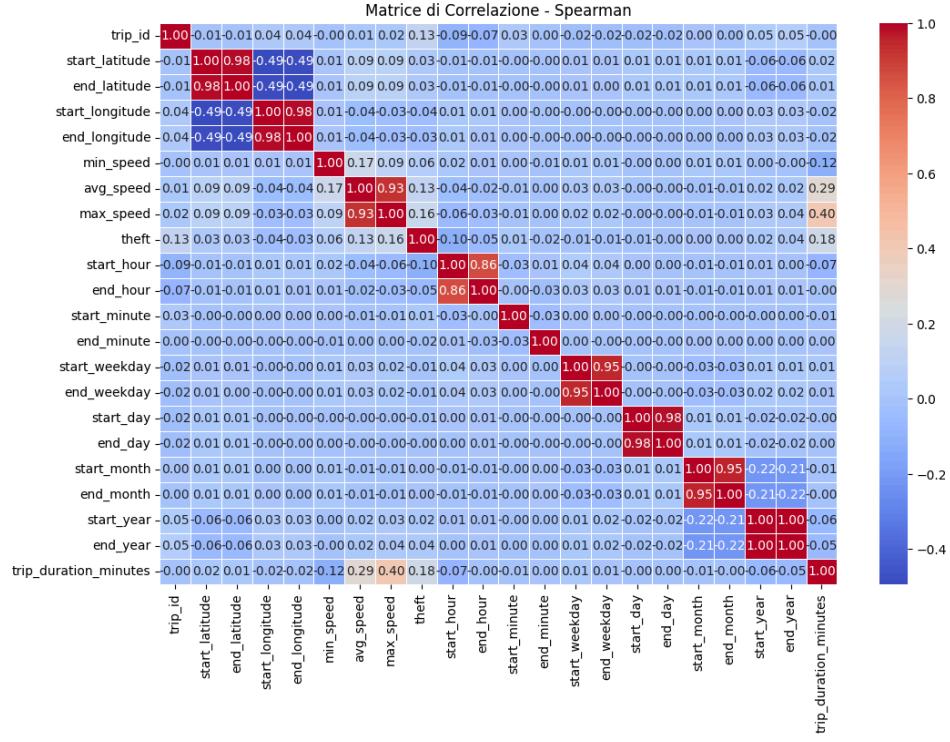
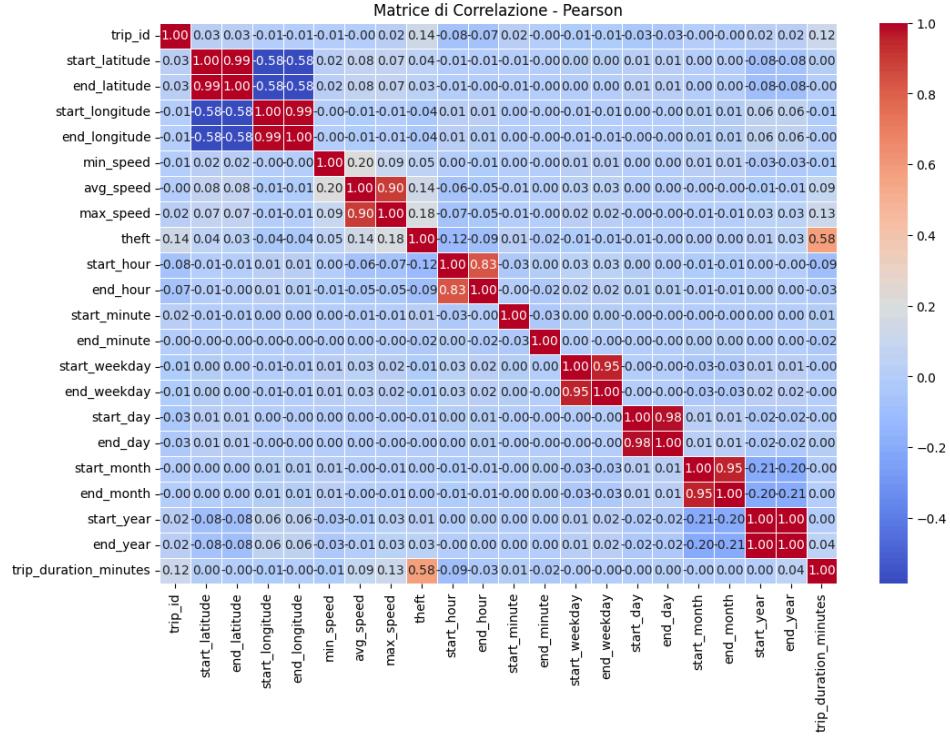
Nota: nelle mappe di esempi di trip non sono ovviamente riportati tutti i trip del veicolo ma solo quelli del furto (in rosso) e quelli precedenti nella giornata (in blu). Inoltre i furti sono un parziale di tutti i furti che avvengono in italia ma a distribuzione è rappresentativa perché i furti sono selezionati randomicamente quindi rappresentano un campione

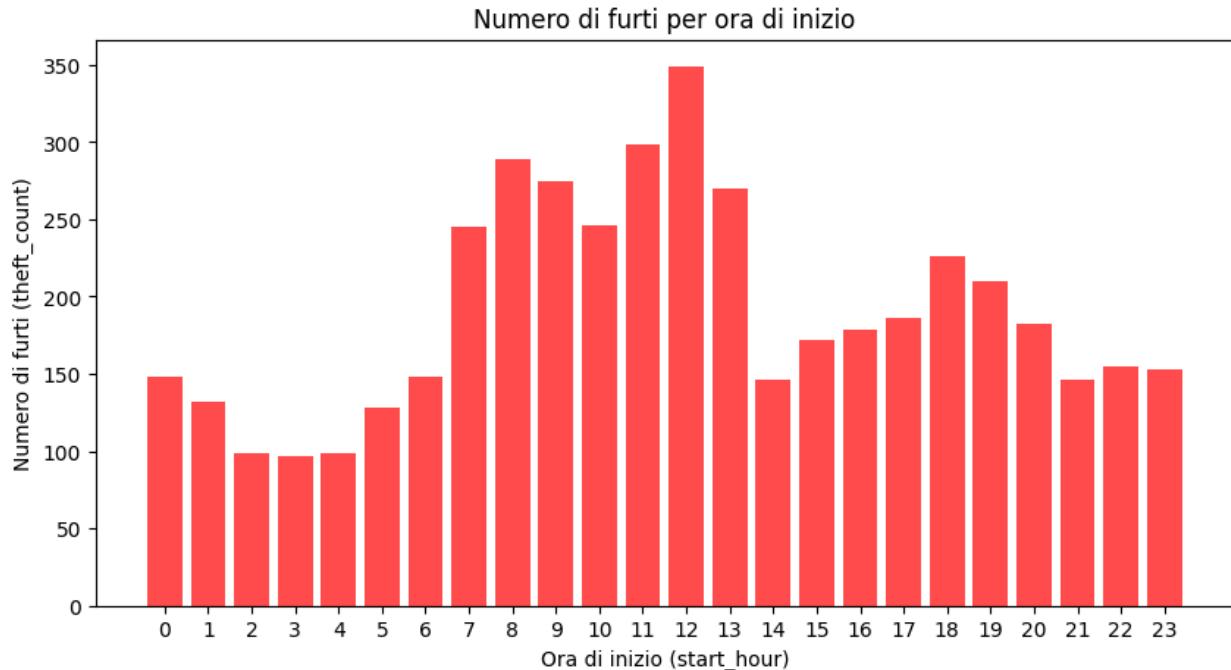
L'analisi esplorativa del dataset preprocessato ha rivelato tendenze significative. I furti variavano in base all'orario, con picchi tra le tre e le cinque del mattino, dove il tasso di furti superava il 40% del totale dei viaggi. Ho trovato anche una correlazione tra velocità elevata e furti, con viaggi che superavano i 150 km/h associati a furti. Questi risultati hanno messo in luce l'importanza di fattori temporali e comportamentali nella previsione dei furti.

Considerazioni relative alle matrici di correlazione: è possibile osservare che le feature sono "debolmente" correlate fra loro, fatta eccezione per la feature target. La bassa correlazione tra le feature implica che il modello non soffrirà di problemi di multicollinearità, questo è positivo per la stabilità delle stime dei coefficienti in modelli come la regressione lineare. Poiché nessuna feature è fortemente ridondante rispetto alle altre, potrebbe essere utile mantenerle tutte. Devo anche vedere la correlazione con il target. Dato che la correlazione tra le feature è bassa, l'eliminazione basata solo sulla correlazione potrebbe non essere utile.

Poichè le matrici di correlazione non mi danno informazioni rilevanti per effettuare la classificazione ho la necessità di utilizzare altri metodi per individuare pattern rilevanti. Per verificare la catturabilità del fenomeno, ho generato l'istogramma riportato di seguito, confermando la possibilità di proseguire l' analisi.

Vengono riportati di seguito i grafici a cui ho fatto riferimento.





Il grafico a istogrammi “Numero di furti per ora di inizio” mostra il numero di furti per ora di inizio di una fase del trip. Si nota che il numero di furti sembra essere particolarmente alto tra le 7:00 e le 14:00, con un picco attorno alle 12:00. Questo potrebbe indicare che i furti avvengono più frequentemente durante le ore lavorative o quando c’è più movimento nelle città.

Il numero di furti è più basso tra la mezzanotte e le prime ore del mattino (0:00 - 5:00). Questo potrebbe essere dovuto al minor numero di persone in circolazione. Tuttavia, è importante mettere in relazione i furti anche con i non furti: ad esempio, sebbene di giorno il numero assoluto di furti sia maggiore, durante la notte la percentuale di furti rispetto ai viaggi totali potrebbe essere più alta.

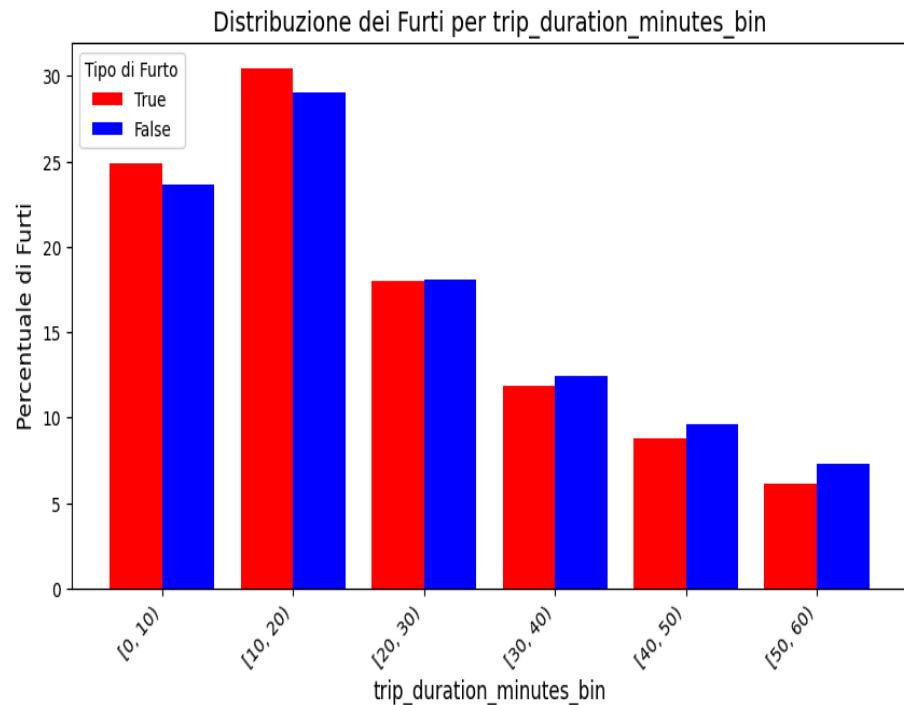
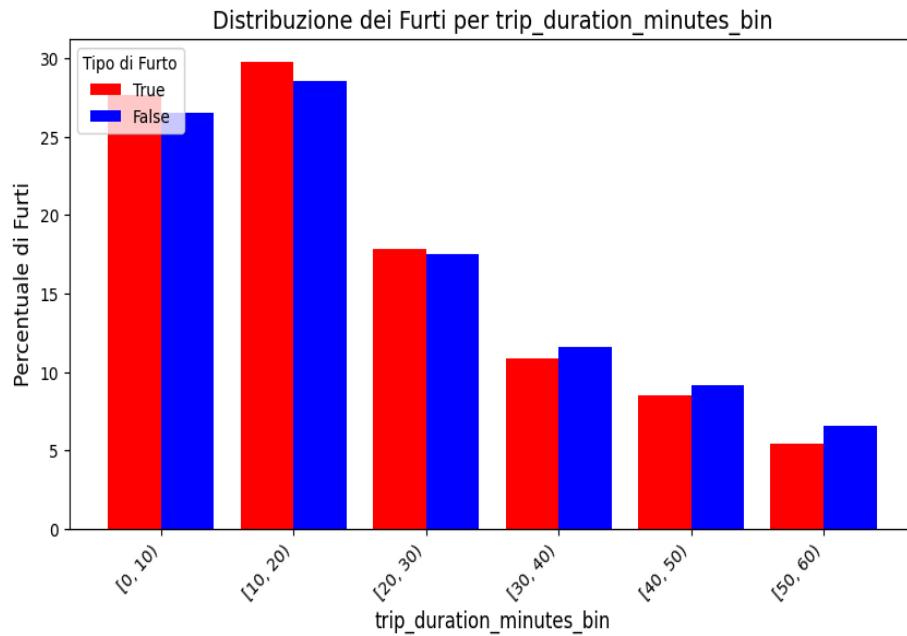
Con il dataset finale, che includeva variabili chiave come velocità, coordinate GPS e durata dei viaggi, mi sono concentrata sull’analisi dei furti, considerando solo il primo furto come riferimento per evitare problemi di cumulazione nelle visualizzazioni. Questa scelta ha permesso di ottenere una visione più chiara della distribuzione dei furti, distinguendo tra furti veri e furti da non considerarsi.

I risultati sono stati illuminanti: le percentuali di furti e non furti hanno mostrato tendenze temporali, evidenziando le ore del giorno più a rischio e permettendo di confrontare le frequenze

dei furti. La fase successiva del mio lavoro è stata l'interpretazione dei pattern, dove ho potuto vedere come i dati raccontassero una storia precisa e dettagliata dei furti.

L' individuazione e l'analisi dei pattern

Proseguendo il mio viaggio nell'analisi dei dati telematici, ho iniziato a mettere a fuoco i pattern che emergevano, prima di trattare gli outliers. Questi comportamenti anomali (gli outliers), sebbene potessero sembrare distorsivi, avevano il potenziale di rivelare informazioni cruciali sui furti. Rimuoverli avrebbe significato sacrificare una parte della varietà informativa del fenomeno.



Considerazioni generali relative ai grafici:

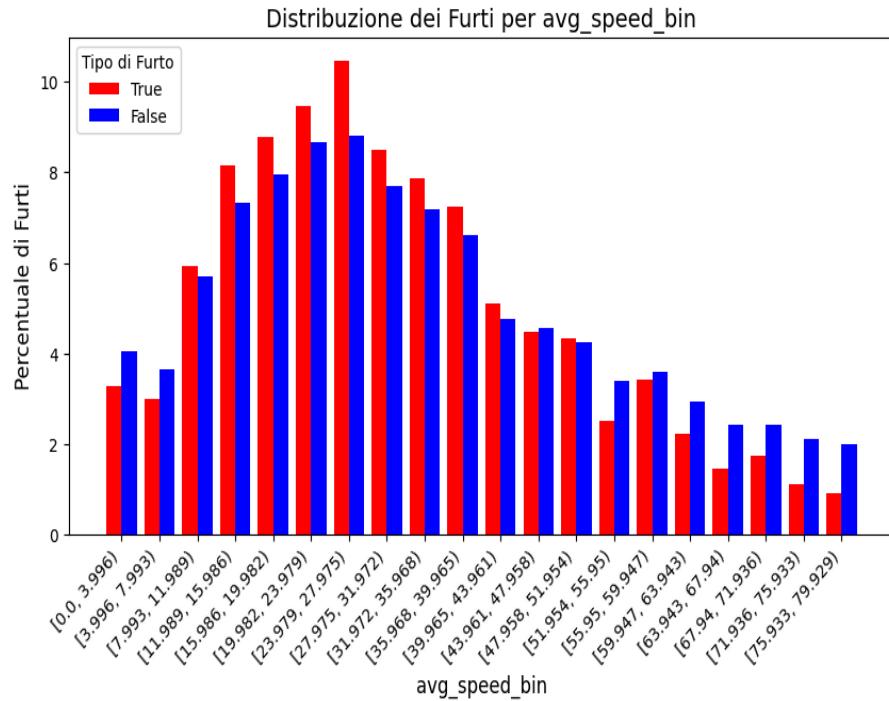
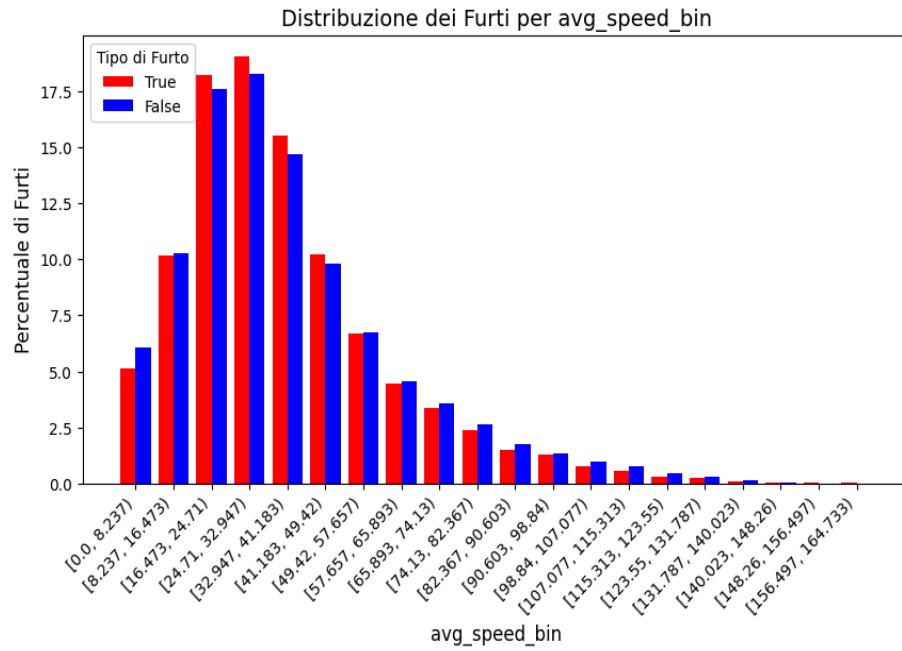
- la maggior parte dei furti (True, in rosso) si concentra nei bin di durata tra 0-10 minuti e 10-20 minuti.
- nei tratti più brevi potrebbe essere più semplice fare appostamenti o individuare il momento giusto per rubare il veicolo. se i tragitti sono brevi, il veicolo potrebbe fermarsi più spesso (ad esempio in soste brevi), offrendo più opportunità per un furto rapido.
- se un veicolo è utilizzato per molti spostamenti brevi, potrebbe essere più esposto a un furto rispetto a un viaggio lungo e ininterrotto, dove il veicolo è sempre in movimento e non si presta a un furto.
- i veicoli rubati vengono utilizzati in modo simile a quelli non rubati in termini di durata dei viaggi.

Considerazioni relative al confronto tra i due grafici:

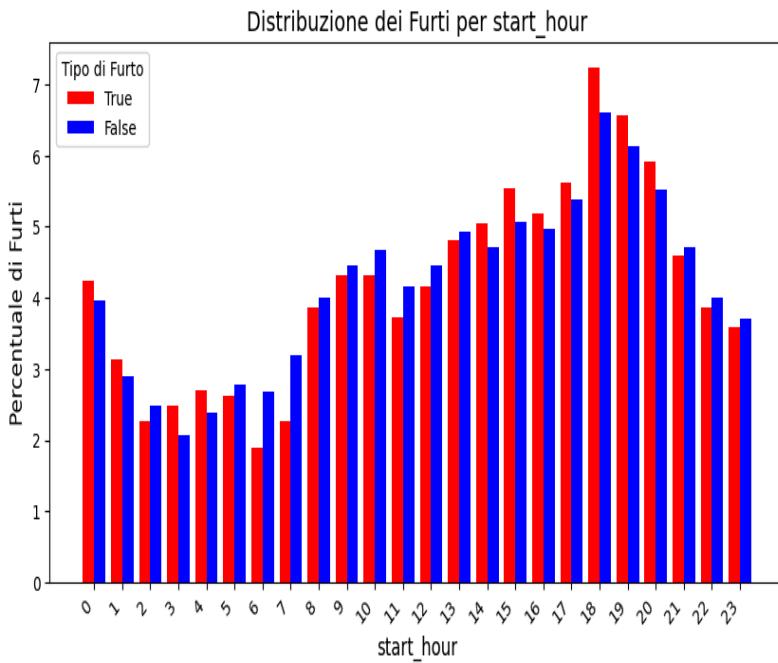
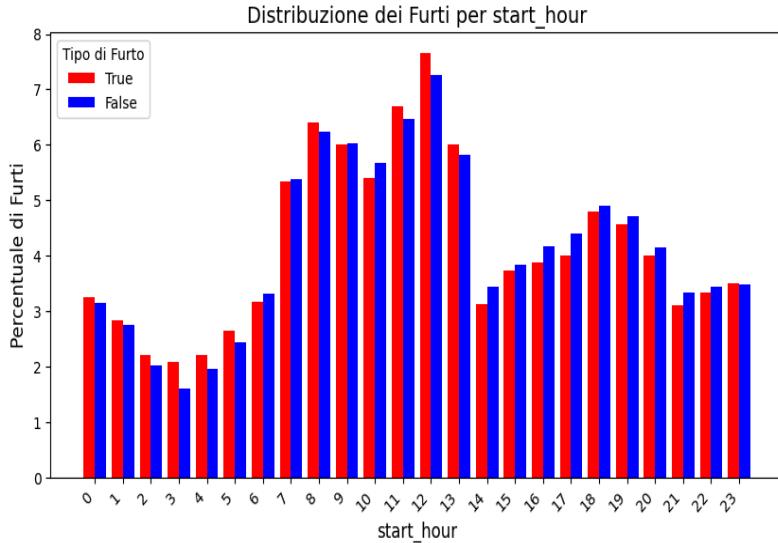
se la rimozione degli outlier avesse avuto un forte impatto, avremmo dovuto vedere cambiamenti più evidenti nella distribuzione. Dato che le due distribuzioni appaiono quasi identiche, possiamo concludere che la presenza di outliers non altera le conclusioni generali relative al fenomeno. Inoltre gli outliers erano quantitativamente pochi e quelli trovati non avevano comunque valori troppo estremi. Questo è confermato anche dal fatto che i range dei bin non cambiano per trip_duration_minutes, a differenza di quanto accade nei grafici successivi.

Nei primi due grafici, ho confrontato la distribuzione dei furti in base alla durata dei viaggi. Prima del trattamento degli outliers, ho notato che circa il 60% dei furti si verificava nei primi 20 minuti di viaggio. Le percentuali di furti veri e falsi si mantenevano vicine, con una leggera predominanza di furti veri nei viaggi più lunghi. Dopo aver applicato il metodo interquartile per rimuovere gli outliers, la distribuzione mostrava una concentrazione maggiore nei viaggi brevi. Questo cambiamento suggeriva che i viaggi di durata prolungata, che superavano i 60 minuti, erano stati eliminati e che ora la maggior parte dei furti si concentrava nei primi intervalli di viaggio.

Tuttavia, la percentuale di furti veri e falsi non mostrava significative variazioni, il che indicava che l'eliminazione degli outliers non aveva alterato in modo sostanziale i dati. Inoltre, analizzando i viaggi senza furto, mi sono resa conto che i viaggi brevi non erano esclusivi dei furti, ma comuni anche nei percorsi normali. Questo ha rivelato che per individuare i furti sulla base della durata, sarebbero state necessarie ulteriori variabili di supporto.

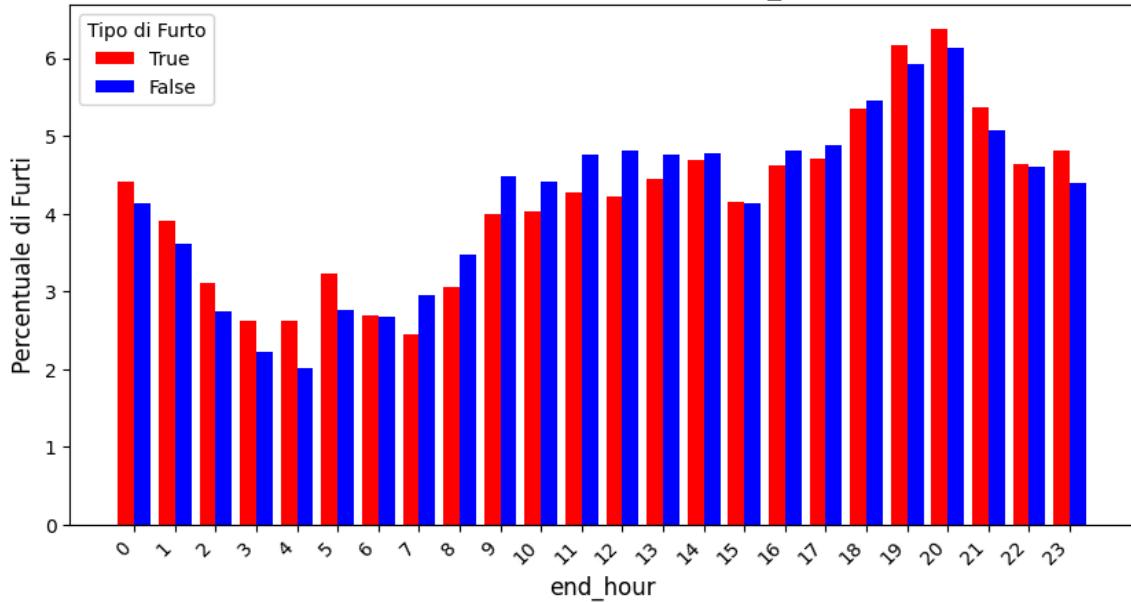


Nei due grafici successivi, ho esaminato la distribuzione dei furti in relazione alla velocità media. Nel primo grafico, prima del trattamento degli outliers, la distribuzione delle velocità mostrava valori che raggiungevano e superavano i 160 km/h, con un picco tra i 40 e i 50 km/h. Tuttavia, c'erano code più lunghe verso velocità elevate, suggerendo la presenza di dati anomali. Dopo aver trattato gli outliers, la gamma delle velocità si era ristretta, con un evidente taglio intorno agli 80 km/h, e il picco era rimasto invariato. Questo indicava che la pulizia dei dati non aveva alterato il comportamento principale, ma aveva ridotto il rumore proveniente da velocità estreme.

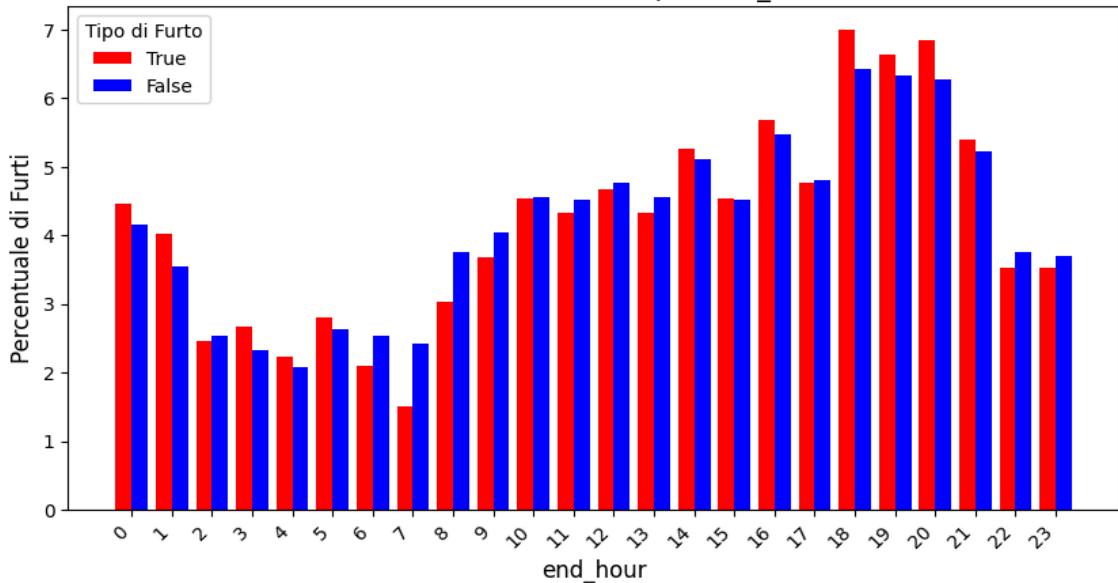


Passando all'analisi della distribuzione dei furti in base all'ora di inizio, il primo grafico mostrava due picchi principali: uno al mattino e l'altro nel pomeriggio. Le percentuali di furti e non furti erano molto simili, suggerendo un comportamento orario omogeneo. Dopo il trattamento degli outliers, la distribuzione era rimasta simile, ma più uniforme, con un picco mattutino leggermente ridotto e il picco serale che si era accentuato. Questo suggeriva che le anomalie potessero influenzare maggiormente i dati mattutini, mentre l'effetto serale era più pronunciato.

Distribuzione dei Furti per end_hour



Distribuzione dei Furti per end_hour



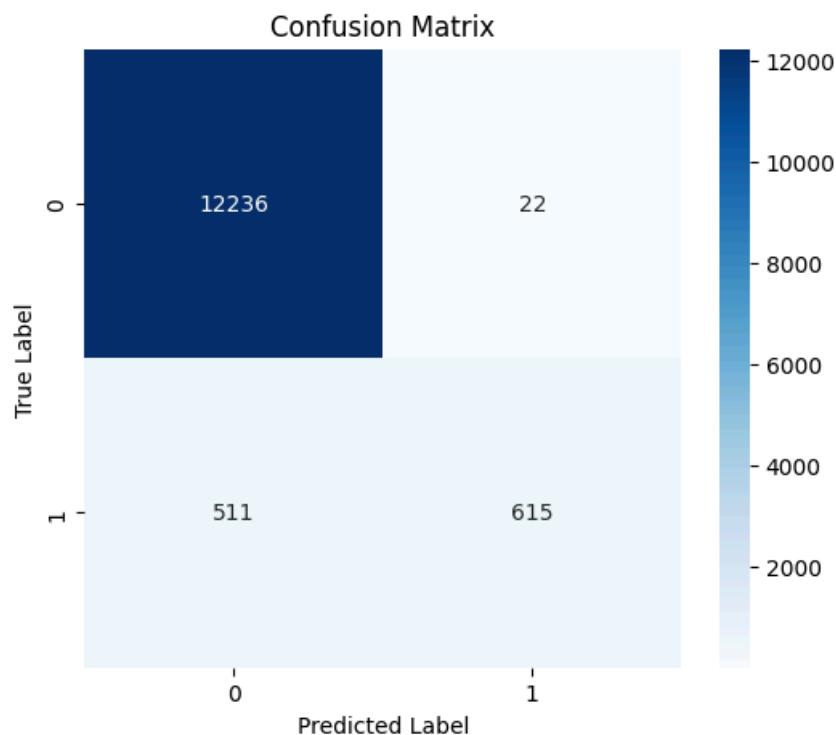
Infine, confrontando la distribuzione dei furti in base all'ora di fine viaggio, notai che prima del trattamento, i furti si verificavano in modo relativamente stabile durante il giorno, con picchi significativi nelle ore serali. Dopo il trattamento degli outliers con metodo IQR, la distribuzione mostrava variazioni più marcate e una riduzione dei valori nelle ore notturne. Il picco serale sembrava accentuato, suggerendo che gli outliers eliminati fossero distribuiti in modo più uniforme nelle ore diurne.

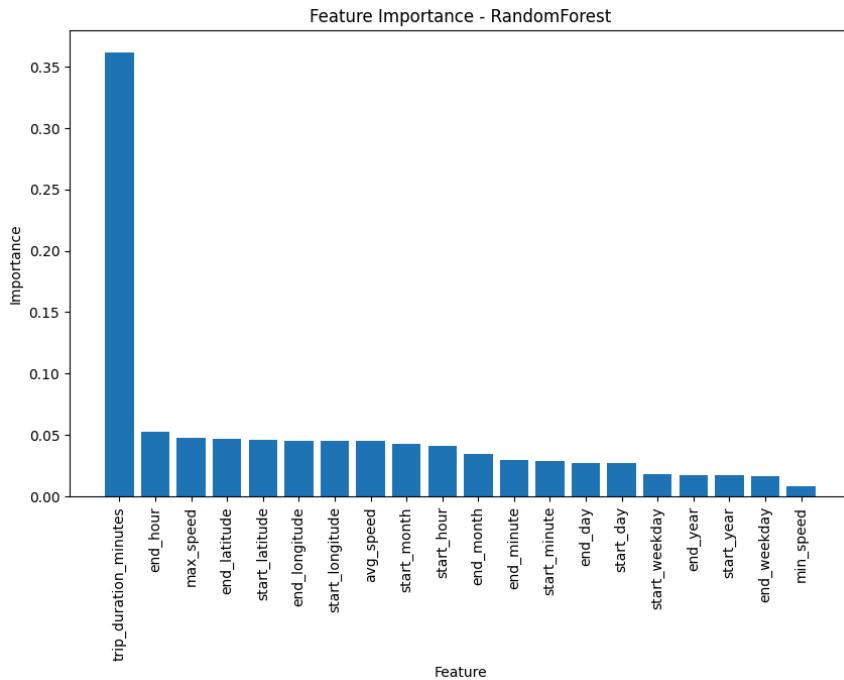
Queste analisi hanno rappresentato un punto di svolta nella mia comprensione del fenomeno dei furti. I pattern emersi, ora più chiari e meno influenzati da anomalie, hanno fornito informazioni preziose per affinare il modello predittivo. Ogni grafico raccontava una storia, e ogni dato si univa a un mosaico complesso di comportamenti umani e dinamiche temporali.

Con queste informazioni, il nostro modello di machine learning poteva essere in grado di identificare e anticipare situazioni di rischio, permettendo interventi tempestivi e mirati. Questa fase finale di analisi ha consolidato la mia convinzione che, attraverso la tecnologia e l'analisi dei dati, potremmo fare passi significativi nella lotta contro il furto dei veicoli, contribuendo a un futuro più sicuro per tutti.

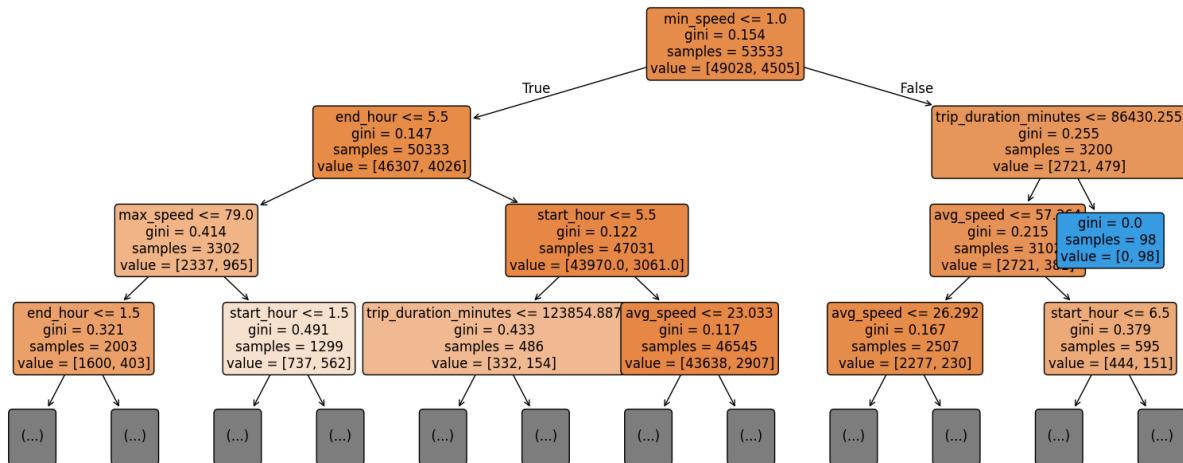
Verso il Futuro: Affinamento del Modello Predittivo per la Prevenzione dei Furti

Dopo aver concluso l'analisi dei pattern e aver individuato il Random Forest con Model selection sul validation e aver lanciato il modello, ho esaminato i risultati della matrice di confusione. Qui è emerso che il modello aveva identificato 12.236 veri negativi, 22 falsi positivi, 511 falsi negativi e 615 veri positivi. Sebbene l'accuratezza del modello fosse impressionante, con un livello del 96%, era chiaro che ci fossero delle difficoltà nel rilevare correttamente i casi positivi, evidenziando un'area di miglioramento.





Analizzando ulteriormente il modello, ho scoperto che le feature più influenti erano la durata del viaggio, l'orario di inizio e fine, e la velocità media. La Random Forest, il modello scelto per questo progetto, aveva mostrato performance eccellenti con una precisione e un richiamo entrambi al 96%, e un F1 score del 95,5%. Questi numeri indicavano che, sebbene il modello fosse robusto, c'era ancora spazio per ottimizzarlo ulteriormente, specialmente nella classificazione dei furti.



La distribuzione dei dati rilevava una netta prevalenza di viaggi non associati ai furti, con l'impurità di Gini che diminuiva man mano che si scendeva nell'albero decisionale. Questo suggeriva che il modello potesse differenziare sempre meglio tra le due classi. Ho notato che i furti sembravano più probabili quando la velocità minima era molto bassa, il che poteva indicare soste prolungate o movimenti sospetti. Al contrario, viaggi con una velocità media molto alta e di lunga durata non sembravano associati ai furti, mentre i viaggi che si concludevano dopo le cinque e mezza, presumibilmente al mattino, erano meno frequentemente associati a furti.

L'analisi dell'albero decisionale ha rivelato pattern significativi, con la velocità minima che si è dimostrata il principale indicatore di viaggi sospetti. Un valore molto basso, ad esempio, inferiore a 55 km/h, era un forte segnale di potenziali furti, specialmente nelle prime ore del mattino. I viaggi che iniziavano prima delle 5:30 presentavano impurità di Gini relativamente alte, suggerendo che questo fosse un intervallo temporale meno frequente per entrambi i tipi di viaggi.

Sviluppi Futuri

Riconoscendo le prestazioni del modello, ho contemplato gli sviluppi futuri. Sebbene avesse dimostrato di essere efficace nella previsione dei furti, era cruciale ridurre il numero di falsi negativi. Le potenziali soluzioni includevano l'adozione di modelli ensemble come XGBoost, oltre a bilanciare il dataset utilizzando tecniche di smoothing. Introducendo feature ingegnerizzate e ottimizzando ulteriormente il modello, avremmo potuto migliorare la sua affidabilità nella rilevazione dei furti nei dati GPS analizzati.

Un altro aspetto fondamentale era l'integrazione del modello con tecnologie emergenti. Ad esempio, l'implementazione di tecniche di Deep Learning, come le reti neurali ricorrenti, avrebbe potuto catturare pattern più complessi, consentendo previsioni più accurate. Inoltre, l'idea di sviluppare un sistema di allerta in tempo reale che utilizzasse il modello per segnalare potenziali furti mentre i veicoli erano in movimento rappresentava un passo avanti significativo.

Sfruttando l'analisi geo-spatiale avanzata e tecniche di clustering, avremmo potuto identificare aree ad alto rischio di furto, ottimizzando le risorse delle forze dell'ordine e delle compagnie assicurative. Infine, migliorare la pipeline di pre-processing potrebbe facilitare una gestione più efficiente dei dati mancanti e una selezione più mirata delle feature informative.

Conclusioni

Con questi sviluppi futuri in mente, il potenziale del nostro modello predittivo si espande. Non solo come strumento per rilevare furti, ma come parte integrante di un sistema di sicurezza automobilistica più ampio, capace di adattarsi e rispondere a un ambiente in continua evoluzione. Il nostro impegno per la sicurezza stradale e la protezione dei beni dei cittadini è solo all'inizio, e sono entusiasta di vedere come la tecnologia e l'innovazione possano contribuire a creare un futuro più sicuro per tutti.

Riferimenti:

- Dati resi disponibili da Octo Telematics SPA
 - Documentazione disponibile su BigData Research Infrastructure
 - Report Tecnico Disponibile qui:
https://docs.google.com/document/d/11nxXjZ3r6rpu-hplzqeqFWfie-f6heb061OB0QK_jc/edit?usp=sharing
-

Appendici

- Link cliccabile al progetto di tesi nel mio Github:
<https://github.com/GiuliaSegurini/Proactive-Theft.git>