

1. Pulizia e Preparazione dei Dati

- Gestione dei Valori Mancanti: Abbiamo optato per l'imputazione dei valori mancanti utilizzando la mediana, che è meno sensibile agli outlier rispetto alla media. Questa scelta è stata guidata dalla presenza di distribuzioni asimmetriche e dalla presenza di valori estremi.
- Eliminazione delle Colonne Non Necessarie: Colonne come 'sfid', 'dt_rif', e 'zone', che sono identificatori unici o categorie geografiche troppo dettagliate, sono state eliminate. Questo perché non contribuiscono in modo significativo alla capacità predittiva del modello e potrebbero introdurre rumore.

2. Esplorazione dei Dati

- Analisi Statistica: Abbiamo esaminato le distribuzioni delle variabili numeriche attraverso istogrammi e boxplot, evidenziando la presenza di skewness e outliers.
- Visualizzazione dei Dati: Le visualizzazioni hanno facilitato la comprensione delle relazioni tra le variabili e la distribuzione di ciascuna di esse. Hanno anche aiutato a identificare anomalie e pattern nei dati.

3. Feature Engineering

- Codifica delle Variabili Categoricali: Utilizzando la tecnica di one-hot encoding, abbiamo trasformato le variabili categoriche in formati numerici, rendendo i dati compatibili con gli algoritmi di machine learning.
- Normalizzazione/Standardizzazione: Abbiamo applicato la scalatura RobustScaler per normalizzare le variabili numeriche, mitigando l'influenza degli outlier e standardizzando i range dei dati.

4. Selezione del Modello

- Scegliere un Modello di Partenza: Per il nostro problema di classificazione, abbiamo selezionato il modello SVC (Support Vector Classifier) con kernel lineare, appropriato per dati di dimensioni moderate e capace di gestire bene spazi caratteristici ad alta dimensione grazie alla trasformazione delle caratteristiche.

5. Iterazione e Valutazione

- Esperimenta con Diversi Modelli: Abbiamo deciso di concentrarci sul fine-tuning del SVC lineare, ottimizzandolo e valutandolo per garantire la migliore prestazione possibile.

6. Fine-tuning e Valutazione Finale

- Fine-tuning dei Parametri: Abbiamo affinato il modello basandoci su una valutazione iniziale e su iterazioni successive per massimizzare l'accuratezza.
- Valutazione del Modello sul Set di Test: Il modello è stato valutato su un set di dati di test separato per assicurare che le prestazioni osservate siano generalizzabili a nuovi dati.