**CREDIT CARD FRAUD DETECTION**

**Documentazione Caso di Studio Ingegneria della Conoscenza A.A. 2022/23**

**Gruppo di lavoro:**

* Dafne Spaccavento, [d.spaccavento@studenti.uniba.it](mailto:d.spaccavento@studenti.uniba.it)
* Gaetano Schiralli, [g.schiralli5@studenti.uniba.it](mailto:g.schiralli5@studenti.uniba.it)

**Repository GitHub:**

<https://github.com/Nanuccio01/CREDIT_CARD_FRAUD_DETECTION>

**Indice:**

**1. Definizione dell'Obiettivo e Comprensione dei Dati**

* L'obiettivo
* Analisi Esplorativa dei Dati

**2. Preparazione dei Dati**

**3. Overfitting**

* Concetto e Tecniche utilizzate per mitigare l'Overfitting

**4. Apprendimento Non Supervisionato e Clustering**

* Applicazione del Metodo di Clustering: K-means
* Interpretazione dei Cluster Ottenuti

**5. Classificazione Transazioni**

* Introduzione
* Scelta degli Iperparametri
* KNN
* Alberi di decisione
* Regressione Logistica
* Introduzione al Modello Probabilistico Naive Bayes

**7. Riassunto dei Risultati**

Discussione delle Sfide e Prospettive Future

**1. Definizione dell'Obiettivo e Comprensione dei Dati**

**• L'obiettivo:**

L'obiettivo di questo progetto è sviluppare un modello in grado di individuare le transazioni fraudolente all'interno di un dataset contenente transazioni effettuate con carte di credito. Il dataset in questione contiene transazioni effettuate da titolari di carte di credito europee nel mese di settembre 2013.

**• Analisi Esplorativa dei Dati:**

Il dataset è composto da transazioni che si sono verificate in due giorni, con un totale di 284.807 transazioni. All'interno di queste transazioni, sono presenti 492 casi di frode. Le variabili di input sono tutte numeriche. Il dataset presenta uno sbilanciamento significativo tra le classi, poiché la classe positiva (frodi) costituisce lo 0.172% di tutte le transazioni.

Immagine che contiene schermata, cerchio, diagramma

Descrizione generata automaticamente

* Le feature V1, V2, ..., V28 sono campi float e rappresentano le principali componenti ottenute come risultato di una trasformazione PCA (Principal Component Analysis), perché a causa delle tutele privacy, il nome di questi campi è reso anonimo. Per implementare una trasformazione PCA, le feature devono essere preventivamente scalate. In questo caso, quindi, tutte le feature V1, V2, ..., V28 sono state scalate o almeno è ciò che assumiamo abbiano fatto gli sviluppatori del dataset.
* Le uniche feature che non sono state trasformate con PCA sono 'Time' e 'Amount', entrambe intere. La feature 'Time' rappresenta i secondi trascorsi tra ogni transazione e la prima transazione nel dataset, mentre la feature 'Amount' rappresenta l'importo della transazione. La feature 'Class' è la variabile di risposta e assume il valore 1 in caso di frode e 0 altrimenti.

Successivamente abbiamo scelto di analizzare e descrivere approfonditamente, solo le colonne 'Time', 'Amount' e 'Class' (tralasciando le feature V1, V2, ..., V28), per ottenere una visione dettagliata di alcune delle caratteristiche chiave dei dati che possono avere un impatto significativo sull'analisi delle frodi su carte di credito:

* "Amount" (Importo della Transazione): Si analizza questa colonna per capire la distribuzione degli importi ed il totale delle transazioni nel dataset. L'importo delle transazioni potrebbe variare ampiamente e potrebbe essere utile comprendere se ci sono trend o pattern specifici nella distribuzione degli importi per le transazioni legittime o fraudolente.
* "Time" (Tempo della Transazione): Si esamina questa colonna per identificare qualsiasi modello temporale nei dati. Potrebbe esserci una correlazione tra i momenti delle transazioni e la probabilità di frode. Analizzando questa colonna, si potrebbero individuare intervalli di tempo in cui si verificano più frodi o altre tendenze temporali interessanti, in quanto nel dataset sono presenti più transazioni per ogni secondo analizzato.

Immagine che contiene testo, linea, Diagramma, diagramma

Descrizione generata automaticamenteDa questo grafico possiamo notare che le transazioni fraudolente hanno una distribuzione più uniforme nel tempo rispetto alle transazioni legittime: esse sono distribuite in modo più uniforme nel tempo, al contrario delle transazioni legittime che subiscono un calo durante le ore notturne del fuso orario europeo.

* "Class" (Classe di Transazione - Legittima(0) o Fraudolenta(1)): Si inserisce questa colonna perché rappresenta la variabile target dell'analisi delle frodi. Sebbene la si inserisca solo per capire la distribuzione delle classi nel dataset (già analizzate in precedenza), questo campo è nuovamente analizzato per completezza.

Qui troviamo riportate le statistiche descrittive delle tre features sopra descritte:

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Guardando alla caratteristica "Time", possiamo confermare che i dati contengono 284807 transazioni, distribuite durante 2 giorni consecutivi (o 172792 secondi).

Dopo aver esplorato inizialmente i dati, possiamo affermare a colpo d'occhio di non aver notato nessuna ripetizione o comportamento significativo tra essi.

Per avere un quadro più generale delle correlazioni, quindi, si sono utilizzati i grafici KDE che mostrano la stima delle densità delle caratteristiche per due classi: Classe 0 (transazioni legittime) e Classe 1 (transazioni fraudolente):

* I grafici servono a visualizzare come le distribuzioni delle caratteristiche differiscano tra le due classi; infatti, consentono di vedere la forma approssimativa delle distribuzioni delle caratteristiche per entrambe le classi. Se ci fossero differenze significative nelle distribuzioni tra le classi per una particolare caratteristica, potrebbe indicare che essa è utile per distinguere tra transazioni legittime e fraudolente.Immagine che contiene testo, diagramma, linea, numero

  Descrizione generata automaticamente

Immagine che contiene testo, diagramma, numero, linea

Descrizione generata automaticamente

Per alcune delle caratteristiche possiamo osservare una buona selettività in termini di distribuzione per i due valori della classe:

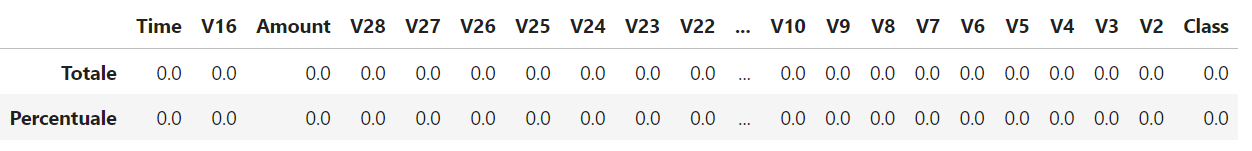
* V4 e V11 hanno distribuzioni chiaramente separate per i valori di Classe 0 e 1.
* V12 e V14 sono parzialmente separati.
* V1, V2, V3 e V10 hanno un profilo piuttosto distinto.
* V25, V26 e V28 hanno profili simili per i due valori della Classe.

In generale, con poche eccezioni (Tempo e Importo), la distribuzione delle caratteristiche per le transazioni legittime (valori di Classe = 0) è centrata intorno a 0, talvolta con una lunga coda su uno dei lati estremi. Allo stesso tempo, le transazioni fraudolente (valori di Classe = 1) hanno una distribuzione asimmetrica (sbilanciata).

Basandoci sui grafici KDE, potremmo fare delle scelte informate su quali feature includere nel modello successivamente. Se alcune caratteristiche avessero distribuzioni molto simili tra le due classi, avremmo potuto decidere di escluderle dal modello, riducendo così la complessità e migliorando le prestazioni.

**2. Preparazione dei Dati**

Gestione dei dati mancanti e non utilizzabili: Una volta caricato il dataset completo si analizza la sua struttura. Si controlla la presenza di valori mancanti nel Dataset. In questo caso nessun problema vien creato, in quanto tutte le colonne presentano dei valori.



Successivamente analizzando il dataset per la ricerca di valori nulli come importo, si è notato che:

* Il numero delle transazioni fraudolente con importo nullo (Amount=0) è: 27
* Il numero delle transazioni legittime con importo nullo (Amount=0) è: 1798

Inizialmente si stava considerando l'eliminazione di questi campi poiché non sembrava esserci un motivo o uno scopo chiaro per la loro presenza, in quanto una transazione non si potrebbe definire valida con un ammontare pari a zero. Ricercando ulteriormente però, si è appurato che esistono sia venditori che effettuano una transazione di prova per verificare gli estremi della carta bancaria, e sia siti web di lotterie o concorsi dove il vincitore effettua una transazione con ammontare pari a zero. Considerando codeste variabili reali quindi, tali righe sono state accettate come valide nel dataset.

Inoltre, in questa fase, andremo a scalare le colonne Time e Amount, per avere valori simili alle altre colonne precedentemente scalate in seguito alla trasformazione PCA.

Immagine che contiene testo, Carattere, schermata, bianco

Descrizione generata automaticamente

**3. Overfitting**

**• Concetto e Tecniche utilizzate per mitigare l'Overfitting:**

Uno dei problemi che potrebbe creare lo sbilanciamento del nostro dataset originale è quello dell'overfitting portando, dunque, il modello di apprendimento automatico a adattarsi troppo ai dati di addestramento. Di conseguenza, il modello potrebbe avere una performance eccezionalmente buona sui dati di addestramento, ma si comporterebbe in maniera pessima su nuovi dati che non ha mai elaborato prima.

Dunque, se utilizzassimo questo dataframe come base per i nostri modelli predittivi e per le analisi, potremmo ottenere errori, in quanto gli algoritmi probabilmente "assumeranno" che la maggior parte delle transazioni non siano truffe (ricordiamo che sono solo lo 0.172% di tutte le transazioni).

Per questo andremo ad utilizzare e a mettere a confronto le seguenti tecniche:

* Creare un bilanciamento delle classi del dataset (sub-Sampling): In questo scenario, il nostro subsample sarà un dataframe con un rapporto 50/50 tra transazioni fraudolente e non fraudolente.
* Utilizzare la Cross-validation stratificata: Abbiamo utilizzato la 5-fold stratified cross validation per garantire che ogni fold della cross validation mantenesse la stessa proporzione tra le classi "legittimo" e "fraudolento" presente nel dataset completo. Questo per contribuire ad evitare che il modello apprendesse troppo dalle transazioni "legittime" e garantiva una valutazione accurata delle prestazioni su entrambe le classi.

**4. Apprendimento Non Supervisionato e Clustering**

**• Applicazione del Metodo di Clustering: K-Means**

In questa fase abbiamo scelto di impiegare l'algoritmo di K-Means per condurre un'analisi di clustering sui dati. L'obiettivo principale di questa fase è identificare possibili sottoclassi o gruppi di transazioni all'interno delle due classi principali, "legittime" e "fraudolente".

Tuttavia, come abbiamo già evidenziato, è importante notare che è già presente una divisione in due classi ben definita e utilizzata come etichetta di apprendimento supervisionato nei nostri modelli. Pertanto, un approccio di K-Means con K=2 (due cluster) non sarebbe appropriato, poiché verosimilmente produrrebbe risultati simili alla suddivisione preesistente tra "legittime" e "fraudolente". Questo non aggiungerebbe valore aggiunto alla nostra analisi.

Invece, intendiamo utilizzare K-Means con un valore di K superiore a 2, consentendo così all'algoritmo di identificare gruppi o cluster all'interno delle due classi principali. Per trovare il k ottimale andremo ad utilizzare il metodo del gomito, provando diversi valori tra 3 e 10.

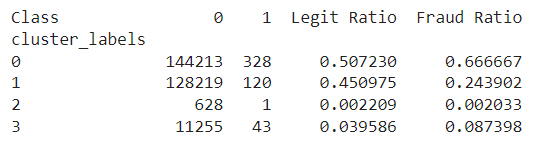
Immagine che contiene testo, schermata, linea, Diagramma

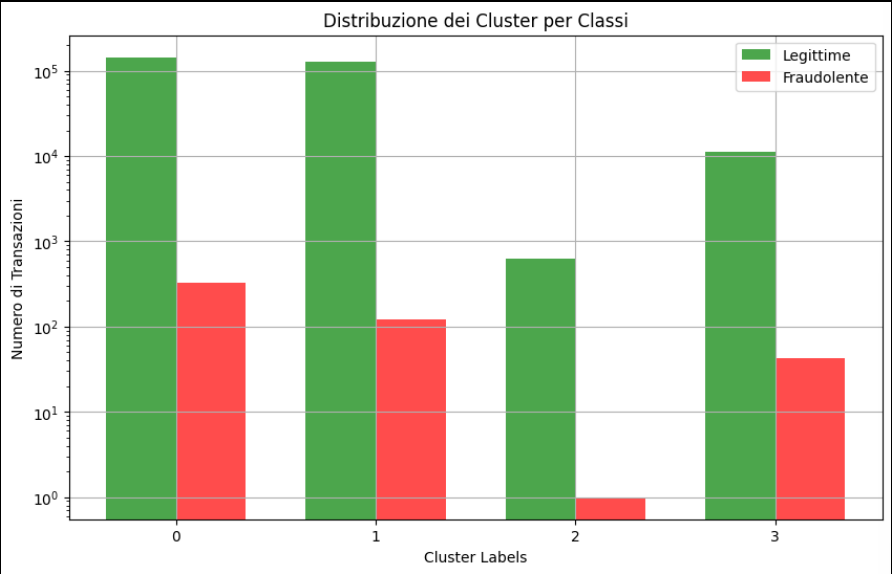
Descrizione generata automaticamente

Nel grafico si evince che sia 4 che 7 potrebbero rappresentare il "gomito" della curva.

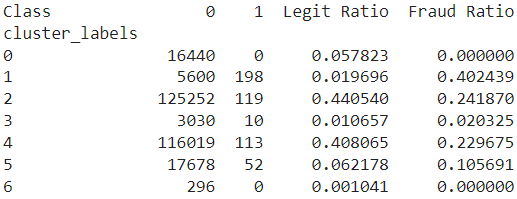
**• Interpretazione dei cluster ottenuti**

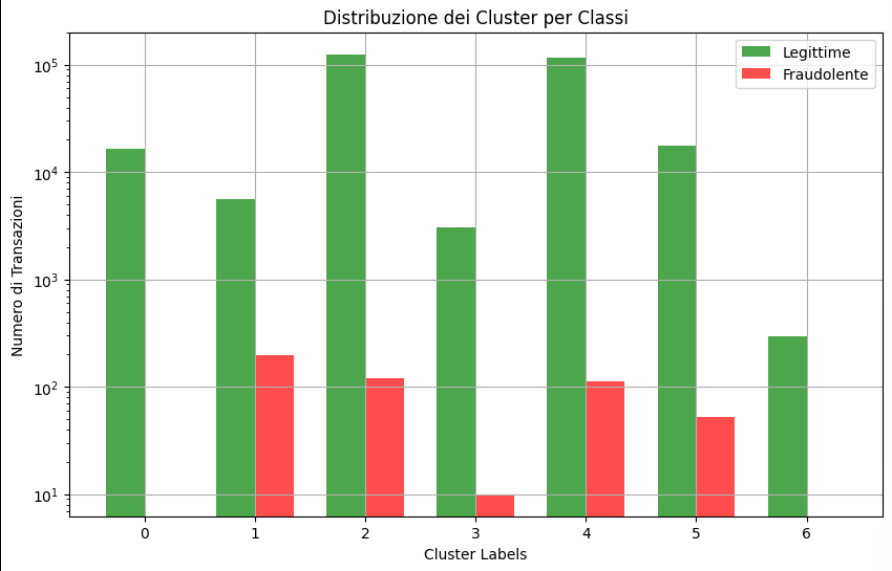
Iniziamo analizzando i risultati utilizzando k=4: Per valutare come le quattro label dei cluster fossero distribuite tra le transazioni legittime e fraudolente, abbiamo calcolato il rapporto tra il numero di transazioni associate a ciascuna label di cluster e il totale delle transazioni legittime e fraudolente nel dataset. Questo rapporto ha fornito una panoramica chiara della distribuzione delle label dei cluster all'interno di entrambe le classi. Tale analisi è risultata inconcludente e non abbiamo identificato eventuali pattern distinti all'interno di ciascuna classe. I rapporti sono pressoché simili, tranne per la label "1" che ha una percentuale più alta nella classe delle transazioni legittime.

****



Procediamo con l'analisi dei risultati utilizzando k=7: Abbiamo dunque ripetuto il medesimo procedimento di analisi delle label dei cluster, ma questa volta abbiamo impostato il valore di k a 7. Questo ci ha permesso di suddividere ulteriormente le transazioni in sottogruppi più specifici. Durante questa analisi, abbiamo notato un pattern significativo nella label 1 del clustering, in quanto si è evidenziata un'alta percentuale di transazioni fraudolente, approssimativamente intorno al 40%. Tale distribuzione è in netto contrasto con la classe delle transazioni legittime, in cui la percentuale di distribuzione per la stessa label è stata di circa il 2%. Tuttavia, è importante sottolineare che nelle altre label dei cluster non sono emerse differenze significative tra le due classi, il che suggerisce che il clustering ha identificato un pattern distintivo principalmente nella label 1, mentre le altre label sembrano essere distribuite in modo simile tra le transazioni legittime e fraudolente.





Secondo la nostra valutazione, per quanto con k=7 si sia evidenziata una differenza evidente nella classe di transazioni fraudolente con label 1, non riteniamo che sia utile inserire le etichette evidenziate nel clustering nella successiva fase di apprendimento supervisionato.

**5. Classificazione Transazioni**

**• Introduzione**

Nelle transazioni, le features di input saranno le colonne numeriche, ad esempio "Time", "Amount" e le features da "V1" a "V28". La feature target sarà la colonna "Class", che indica se una transazione è fraudolenta o legittima.

Esempi di Training e Test: Gli "esempi di training" saranno un sottoinsieme del dataset con tutte le colonne, inclusa la colonna "Class". Questi esempi verranno utilizzati per addestrare il modello.

Gli "esempi di test" saranno un altro sottoinsieme del dataset contenente solo le colonne numeriche (senza la colonna "Class"). Questi esempi verranno utilizzati per testare il modello addestrato e fare previsioni sulle classi delle transazioni.

**• Scelta degli iperparametri**

Per selezionare gli iperparametri ottimali per i nostri modelli di apprendimento abbiamo utilizzato il metodo di ricerca della griglia, noto come "Grid Search". Questo è una tecnica di ricerca esaustiva che ci ha permesso di esplorare un insieme predefinito di combinazioni di iperparametri e determinare quelle che conducono alle prestazioni migliori per i nostri modelli.

Esempi di iperparametri trovati con GridSearch per i diversi classificatori addestrati su tutto il Dataset:

Immagine che contiene testo, schermata, Carattere, documento

Descrizione generata automaticamente

Esempi di iperparametri trovati con GridSearch per i diversi classificatori addestrati sul subsample:

Immagine che contiene testo, schermata, Carattere, documento

Descrizione generata automaticamente

Non si sono ricercati gli iperparametri del Naive Bayes in quanto, esso è noto per la sua semplicità ed è generalmente meno sensibile agli iperparametri rispetto ad altri algoritmi di machine learning.

Non si sono ricercati gli iperparametri di tutti i nostri modelli di apprendimento durante la 5-fold Stratified Cross Validation a causa di un altissimo costo computazionale.

Gli score ed i valori che andremo ad analizzare successivamente corrisponderanno ai modelli addestrati con gli iperparametri trovati, in quanto abbiamo notato un miglioramento significativo rispetto ai modelli addestrati con i parametri di default. Questo miglioramento si osserva in maniera più chiara per gli alberi di decisione.

**• KNN**

Considerando che il nostro dataset contiene una porzione significativa di dati privati i quali non possono essere selezionati o esclusi a causa della mancanza di conoscenza sul loro contenuto, potrebbe essere prudente utilizzare tutte le features disponibili nel contesto dell'algoritmo k-Nearest Neighbors (k-NN). Questo approccio ci permette di sfruttare tutte le informazioni disponibili per effettuare le previsioni, tenendo presente che alcune features potrebbero contenere informazioni utili.

Tuttavia, siamo consapevoli dei rischi associati all'utilizzo di tutte le features, specialmente quando il dataset contiene dati sensibili o potenzialmente rumorosi. L'aumento della dimensionalità dello spazio delle feature potrebbe comportare distorsioni nelle misure delle distanze tra le osservazioni, influenzando le prestazioni del modello k-NN.

Quindi, il K-NN è stato sostanzialmente eseguito sul dataset avente 31 features.

* Features di input: Time, "Amount e le features da "V1" a "V28".
* Features di output: la classe Class.

**• Alberi di decisione**

Ecco alcuni dei principali algoritmi ensemble che abbiamo utilizzato:

* DecisionTreeClassifier: Abbiamo iniziato con il DecisionTreeClassifier, un algoritmo che costruisce un singolo albero decisionale. Questo ci ha permesso di esplorare le potenzialità di base degli alberi decisionali nel nostro caso di rilevamento delle transazioni fraudolente
* RandomForest: Successivamente, abbiamo implementato il RandomForest, un algoritmo ensemble che combina diversi alberi decisionali creati su sottoinsiemi casuali dei dati e delle feature. Questo ha migliorato la generalizzazione e la robustezza del nostro modello, riducendo l'overfitting.
* AdaBoost: Abbiamo anche esaminato l'AdaBoost, un algoritmo che assegna maggior peso agli esempi classificati erroneamente, permettendo agli alberi successivi di concentrarsi sulle aree difficili da classificare. Questo ha portato a un miglioramento delle prestazioni sui casi più complessi.
* GradientBoostingClassifier: Infine, abbiamo sperimentato il GradientBoostingClassifier, che costruisce gli alberi in modo sequenziale, ognuno correggendo gli errori dei precedenti. Questa metodologia ha permesso di ottenere previsioni sempre più accurate, affinando progressivamente il modello.

**• Regressione Logistica**

La regressione logistica è computazionalmente efficiente, il che significa che può essere addestrata su grandi dataset in tempi relativamente brevi. Questo è vantaggioso quando si hanno dataset con molte transazioni, come nel nostro caso.

**• TAN Naive BayesA**