UNIVERSITÁ DEGLI STUDI DI MILANO-BICOCCA

Facoltà di Economia

Corso di laurea in Scienze Statistiche ed Economiche

MODELLI BAYESIANI

Applicati al calcolo di rischio sulla corruzione e sul finanziamento al terrorismo

Relatori: Prof.ssa Sonia Migliorati

Dott. Mario Turla

Tesi di Laurea di:

Joana Curri

Matr. N. 790625

Anno Accademico 2016/2017

Sommario

INTRODUZIONE

Il percorso formativo svolto nei sei mesi di stage, presso la T3M Innovation s.r.l., società iscritta al registro delle startup innovative della Camera di Commercio di Milano, si è concentrato sulla realizzazione e commercializzazione di soluzioni per la gestione del rischio basato su modelli probabilistici predittivi. L'offerta T3M si concentra sul rischio "reale" con strumenti efficaci per la gestione del rischio preso in considerazione, ed in particolare quello di riciclaggio di denaro, di finanziamento al terrorismo e di corruzione.
La start up nasce dal desiderio di poter offrire al territorio italiano e non solo, una soluzione che mira a soddisfare le esigenze di tutti coloro che sono soggetti alla normativa europea e nazionale in materia, ovvero: intermediari finanziari, professionisti, case d'asta, P.A., etc.
Ogni soluzione, pur essendo progettata in funzione del tipo di cliente e del tipo di attività da esso svolta, si basa su una piattaforma comune altamente flessibile, FARADAY, che consente di gestire differenti tipologie di dati e supportare il calcolo del rischio in vari ambiti.
Durante la prima metà dello stage, si è avuto modo di approfondire le conoscenze in ambito bancario, acquisendo consapevolezza maggiore in terminologie e in movimentazioni bancarie.
L'argomento principale trattato ed ampliato è stato relativo all'ambito del riciclaggio del denaro, il quale comprende al suo interno due temi fondamentali, il finanziamento al terrorismo e la corruzione, entrambi argomenti sui quali sono stati creati classificatori predittivi.
Da un punto di vista applicativo, l'aspetto positivo riscontrato fin dai primi giorni, è stato l'utilizzo di un software, R Studio, che ho avuto già modo di conoscere in maniera piuttosto approfondita durante i tre anni di università grazie ai laboratori didattici, tramite i quali ho acquisito una certa domestichezza nel creare modelli probabilistici.
Quindi dopo una prima parte di studio e apprendimento in materia, mi è stato dato il compito di creare due modelli probabilistici predittivi, tramite l'algoritmo Naive Bayes, con funzione di autoapprendimento e di rilevare le previsioni tramite il modello creato precedentemente.
Più in generale l'idea è stata quella di utilizzare i dati provenienti da alcune banche, applicare i due modelli creati, uno sulla corruzione e l'altro sul finanziamento del

terrorismo, che contribuiscono al calcolo complessivo del rischio del riciclaggio del denaro, e riuscire a determinare le probabilità di rischio per ogni singolo soggetto e cluster. Il dataset che avevo a disposizione era composto da dati reali riguardanti la monetica, ma allo stesso tempo era fittizio, nel senso che non avendo ancora le informazioni reali dalle banche, ho utilizzato tali dati per addestrarmi nell'applicare il classificatore, per poi, in seguito, riuscire a procedere più velocemente alla sperimentazione con dati reali che alcune banche ci avrebbero fornito nei mesi successivi.

La parte nuova che ho riscontrato nel lavorare presso tale società, oltre al classificatore Naive Bayes, del quale però non ho rilevato grossi problemi, è stata la parte di apprendimento automatico del classificatore.

L'apprendimento automatico è stato un mondo nuovo nel quale mi sono cimentata, senza dubbio ho riscontrato delle difficoltà iniziali, ma allo stesso tempo ho reputato tale argomento molto interessante fin dall'inizio sia dal punto di vista dello studio sia dal punto di vista applicativo.

Nelle pagine successive viene introdotto il tema del riciclaggio in maniera sintetizzata, concernente i due fenomeni principali sui quali ho basato la creazione dei modelli predittivi, il finanziamento al terrorismo e la corruzione, ed inoltre vi è un accenno sulla situazione italiana precedentemente a tale fenomeno, e successivamente le motivazioni per cui abbiamo deciso di applicarvi l'inferenza Bayesiana e la Features Selection alle variabili.

Nel primo capitolo viene fatto un quadro generale su riciclaggio del denaro, introducendo le direttive europee a riguardo, e l'evoluzione radicale del quadro normativo comunitario in materia di riciclaggio porta a combattere i rischi connessi al finanziamento del terrorismo e alla corruzione, argomenti che vengono trattati in maniera più approfondita nel primo capitolo.

Nel capitolo successivo, viene introdotto il classificatore Naive Bayes utilizzato nella creazione dei modelli, e tutti i suoi aspetti positivi e negativi, descrivendo in dettaglio il teorema di Bayes, sul quale i classificatori sono basati, ma delineando anche gli altri modelli di classificatori presenti nella statistica, motivando la nostra scelta relativa al classificatore.

In seguito viene descritto l'apprendimento automatico Bayesiano, e le modalità con cui viene utilizzato nel software, ed in ultimo è possibile avere una visione più chiara dei features selection e della modalità con cui è stato applicato.
Mentre nei primi due capitoli è stata fatta una disamina teorica dei temi che ho riscontrato durante il mio periodo di percorso formativo presso l'ente che mi ha ospitato, i capitoli tre e quattro sono dedicati interamente allo sviluppo e alla creazione del progetto.
Inizialmente viene presentato nel dettaglio l'intero sviluppo del progetto, analizzando i dati a disposizione su cui lavorare ed eventualmente eseguire una ripulitura di questi qualora ci fossero informazioni mancanti.
Successivamente vengono motivati e spiegati i passaggi eseguiti nel creare un classificatore adeguato. Dunque si inizia dalla fase di divisione del dataset in training set e test set necessaria per addestrare il modello, e solo in seguito ad un buon risultato di previsione, è possibile riutilizzarlo su dati nuovi sviluppando previsioni con un alto indice di accuratezza.
Successivamente vengono esaminati i risultati sperimentali ottenuti dall'applicazione dei modelli, analizzando la loro bontà nel poter essere applicati su dati nuovi, che non presentano l'etichetta di rischio.
Fase preliminare alla creazione del classificatore risulta essere la Features Selection, ossia una tecnica di riduzione delle variabili esplicative utile per migliorare la performance dell'algoritmo di classificazione, dove vengono selezionate le features rilevanti per la costruzione del modello.

Considerazioni Generali

La storia dell'antiriciclaggio in Italia nasce sotto il profilo repressivo, negli anni settanta e più precisamente nel 1978, per contrastare due fenomeni particolarmente gravi che si verificavano in quegli anni: il rapimento di persona e le rapine a mano armata.
Sotto la spinta di una forte indignazione dell'opinione pubblica la politica era stata indotta ad intervenire varando la prima legge che puniva il riciclaggio in Italia con l'art. 648 bis del codice penale.

Legge che, purtroppo, nasceva già monca poiché non prevedeva il fenomeno dell'"autoriciclaggio", cioè la possibilità di punire una persona che effettuando il reato, impiega il denaro proveniente da questi reati in attività lecite.

Il riciclaggio di denaro è definito in letteratura come un insieme di operazioni atte a dare una veste lecita a capitali di origine illecita, con operazioni che rendono difficile a risalire alla sua origine, tale attività può essere definita come un cancro nella società, poiché entra mascherandosi tramite un soggetto che apparentemente agisce rispettando le regole del mercato ma che in realtà ha uno scopo ben diverso, che è reimmettere nel mercato legale capitali di origine illecita accettando anche una perdita.

Tale reato multiforme e fino a qualche tempo fa inafferrabile è visto come una fattispecie penale di non facile inquadramento, questo poiché i proventi illeciti entrano all'interno dell'economia globale una volta depurati del marchio di origine criminale.

In quegli anni le organizzazioni criminali stavano realizzando i loro massimi profitti perseguendo altri reati ed in modo particolare lo spaccio delle sostanze stupefacenti, fenomeno delittuoso che, tra

l'altro, aveva assunto una preoccupante dimensione internazionale.

Fino al punto che in una convenzione ONU, siglata a Vienna nel 1988, specificamente contro il traffico delle sostanze stupefacenti, induceva la comunità internazionale a prende un impegno per contrastare il riciclaggio del denaro proveniente da tale reato, sollecitando gli stati membri a prevedere nei loro ordinamenti una particolare rilevanza penale a questa attività delittuosa.

L'Italia, spinta dalla gravità del fenomeno, modifica quindi la norma recependo le indicazioni formulate nella convenzione di Vienna, introducendo l'art.23 della legge del 19 marzo 1990, che estende i reati presupposti come antiriciclaggio anche ai delitti per la produzione ed il traffico di sostanze stupefacenti o psicotrope, ma, anche in questo caso, senza introdurre il reato di autoriciclaggio.

L'obiettivo principale delle organizzazioni criminali è sempre stato quello di creare un guadagno da attività illecite. Questo comporta che il denaro guadagnato dovrà essere impiegato in molteplici attività:

- nelle stesse attività illecite, compresa la corruzione
- per mantenere le famiglie degli affiliati che stanno in carcere
- buona parte deve essere riciclato in attività "lecite"

Ne segue che la fase del riciclaggio è essenziale per il ciclo economico dell'economia dell'illegalità. Ma per far questo si ha bisogno di quella zona grigia fatta da professionisti che permette di mettere in contatto i due mondi applicando le loro competenze per la gestione di questo denaro. Essi pensano più o meno consapevolmente di prestare i loro servizi sfruttando le competenze a clienti danarosi, applicando il principio "pecunia non olet", il denaro non ha odore, ma si sbagliano poiché questi clienti non rispettano le regole del mercato e se le cose non vanno secondo le loro aspettative sono pronti a fare pressioni sfruttando tutte le leve in loro mano. Dal ricatto per dei favori ricevuti alla violenza fisica ed è solo in questo momento che si rendono conto di essere totalmente nelle mani di queste organizzazioni criminali.

Se in passato il riciclaggio di denaro sporco da parte delle organizzazioni criminali era ritenuto fonte primaria di inquinamento dei mercati finanziari, oggi l'utilizzo di operazioni di corruzione e del circuito finanziario da parte di organizzazioni terroristiche per autofinanziarsi tramite l'utilizzo di capitali leciti, costituiscono un altrettanto concreto fattore di inquinamento dei mercati e di pericolo per la società.

Ed è proprio la continua e imponente crescita dei profitti illeciti delle organizzazioni criminali che aumenta il bisogno di riciclare il denaro di provenienza illecita in attività lecite, ripulendolo dall'origine criminale che spinge la UE a intervenire legiferando sulla materia.

Motivazioni

Negli ultimi anni è emerso con maggiore chiarezza come nel quadro della lotta contro il riciclaggio un ruolo chiave sia svolto dal contrasto al finanziamento del terrorismo internazionale e dalla corruzione. I numeri che riguardano il riciclaggio nel mondo sono impressionanti, secondo l'ONU riguarderebbe il 2% del PIL mondiale che ammonterebbe a circa 1500 miliardi di dollari, ed in Italia alcune stime lo danno su circa il 10%. In un mondo del genere sembrerebbe che la vera potenza mondiale siano le organizzazioni criminali, capaci di influenzare l'economia a livello mondiale e comprarsi la politica.

Pertanto avere una attenzione particolare sui flussi finanziari non è solo una buona prassi ma un dovere per non far sì che la società degradi in mano all'illegalità.

Come sottolineato dal Consiglio di Sicurezza delle Nazioni Unite gli attentati terroristici dell'ultimo anno hanno reso ancor più evidente la necessità di attuare misure di contrasto al finanziamento al terrorismo, alla corruzione e al riciclaggio di denaro sul piano internazionale, regionale e nazionale. L'Unione europea (UE), in linea con la posizione ONU, ha collocato le misure antiriciclaggio nel quadro della lotta al terrorismo internazionale.
Le parole del Procuratore Nazionale Antimafia aggiunto Pier Luigi Maria Dell'Osso in risposta al problema globale della corruzione, sono chiare:

"Abbiamo gli strumenti: non dobbiamo e non possiamo più sopportare il fenomeno".

Le parole sono nette e riflettono la sua angoscia per un mondo che si trova nelle mani della criminalità da anni oramai, e che purtroppo riscontra difficilmente una via d'uscita, questa causata dall'alto livello di corruzione in cui si trova il nostro paese.
La strada che abbiamo seguito per prevenire le diverse classi di "rischio" sul finanziamento al terrorismo e corruzione è indubbiamente quella probabilistica, per via dell'aleatorietà del rischio, essendo quest'ultimo una situazione che non rispetta i principi fondamentali di tracciabilità e trasparenza.
È ovvio che un approccio deterministico non ha le basi concrete per essere eseguito, in quanto, quando si parla di rischio, ci riferiamo ad un evento irrazionale, che non può essere mai definito certo dal significato stesso della parola.
L'utilizzo dei metodi Bayesiani, che si collocano tra i modelli di classificazione probabilistica, è consono per la classificazione di pattern, basata sull'esperienza. Come nella medicina, anche in questo ambito, per poter classificare un record con una probabilità di rischio è fondamentale avere dei casi reali su cui fare riferimento, per addestrare il nostro classificatore.
In ambito medico il sintomo e il test di laboratorio rappresentano le variabili esplicative, mentre la malattia è la variabile risposta, dove per poter attribuire una classe target a quest'ultima è necessario analizzare i valori presenti nelle esplicative. Ma per poter fare questo procedimento è indubbiamente necessario avere dei casi reali analizzati dove sono presenti i sintomi e i test di laboratorio e soprattutto la relazione con la malattia presentata.

I classificatori Bayesiani ci permettono di calcolare in modo esplicito la probabilità a posteriori $P(Y|Xn)$, che un'osservazione appartenga a una specifica classe target ricavandola sulla base del teorema di Bayes, note le probabilità a priori $P(Y)$ e le probabilità condizionate alle classi P(Xn|Y).
Qui Y è la variabile risposta che rappresenta le cinque differenti classi target di rischio di inoltro delle segnalazioni di operazioni sospette proposte dalla U.I.F.:

$$Y = \{basso, medio\backslash basso, medio, medio\backslash alto, alto\}$$

mentre Xn, sono le n variabili esplicative:

$$Xn = \{X1, X2, \dots, Xn\}$$

le quali presentano valori discreti e soprattutto, seguendo l'ipotesi fondamentale del teorema di Bayes, sono tra loro indipendenti.
La creazione dei due modelli avviene inizialmente analizzando la bontà che questi presentano, e successivamente, una volta raggiunto un buon grado di bontà, è possibile applicare il modello su dati non comprendenti la variabile risposta, e solo a questo punto sarà il modello stesso a fare le nuove previsioni, con un leggero margine di errore.
Lavorando su un dataset ad alta dimensionalità, la fase successiva, riguardante l'applicazione della tecnica del feature selection, ci permette di diminuire la dimensione del dataset, in particolare delle variabili esplicative. Dunque queste, se non presentano un forte legame di correlazione con la variabile risposta, verrebbero eliminate dal dataset. Mentre alle variabili che presentano un alto legame di correlazione con il rischio, vengono assegnati dei pesi in base al livello di legame che mostrano con la classe target.

Capitolo 1 - CORRUZIONE E FINANZIAMENTO AL TERRORISMO

La corruzione

La corruzione è un fenomeno che in ogni tempo e in ogni luogo ha influenzato il nostro vivere ed è quindi un fatto che coinvolge tutti noi da sempre, essa rappresenta il vero male del nostro sistema, un cancro con radici profonde, difficile da scardinare, che si sviluppa con disarmante facilità infettando la parte sana dell'economia.
Dare una definizione di "corruzione" è al tempo stesso un'operazione estremamente semplice ed estremamente complessa, in quanto non esiste una definizione unica, completa ed universalmente accettata, a causa di diverse sensibilità politiche, storiche e sociali.
L'Unione Europea adotta una definizione ampia di corruzione che comprende ogni forma di abuso di potere per guadagno privato e che risulta applicabile sia al settore pubblico che al settore privato.
La corruzione pubblica è un particolare accordo tra due o più soggetti, appartenenti a due diverse categorie, un privato cittadino e un funzionario pubblico, in base al quale il primo devolve al secondo un compenso che non gli spetta, per un atto non dovuto.
Il Codice Penale italiano disciplina la corruzione nella Pubblica Amministrazione, negli articoli dal 318 al 322, mentre l'articolo 2635 del Codice Civile detta le disposizioni relative alla "Corruzione tra privati."
In prima battuta, essa può essere definita come un particolare accordo "pactum sceleris" tra un funzionario pubblico e un soggetto privato mediante il quale il primo accetta dal secondo, per un atto relativo alle proprie mansioni, un compenso che non gli è dovuto.
Infatti alla base dello scambio corruttivo vi è fondamentalmente un incrocio di interessi perseguiti dal corruttore e dal corrotto, interessi che possono essere di varia natura, anche non economici.

La corruzione presenta tre caratteristiche fondamentali:

- *Seriale:* in quanto coloro che sono dediti a questi illeciti tendono a commetterli ogni volta che ne hanno occasione, con ragionevole certezza di impunità.

- *Diffusiva*: in quanto corrotti, corruttori e intermediari, al fine di assicurarsi la realizzazione dei patti illeciti e di evitare di essere scoperti, tendono a coinvolgere altre persone, creando una fitta rete di interrelazioni illecite, fino a che sono gli onesti ad essere esclusi dagli ambienti prevalentemente corrotti.

- *Oscura*: ovvero non denunciato perché sia il corruttore sia il corrotto non hanno interesse a farlo.

Secondo la nota formula di Robert Klitgaard, uno dei maggiori esperti del fenomeno corruttivo a livello mondiale, la corruzione viene espressa con la seguente formula economica:

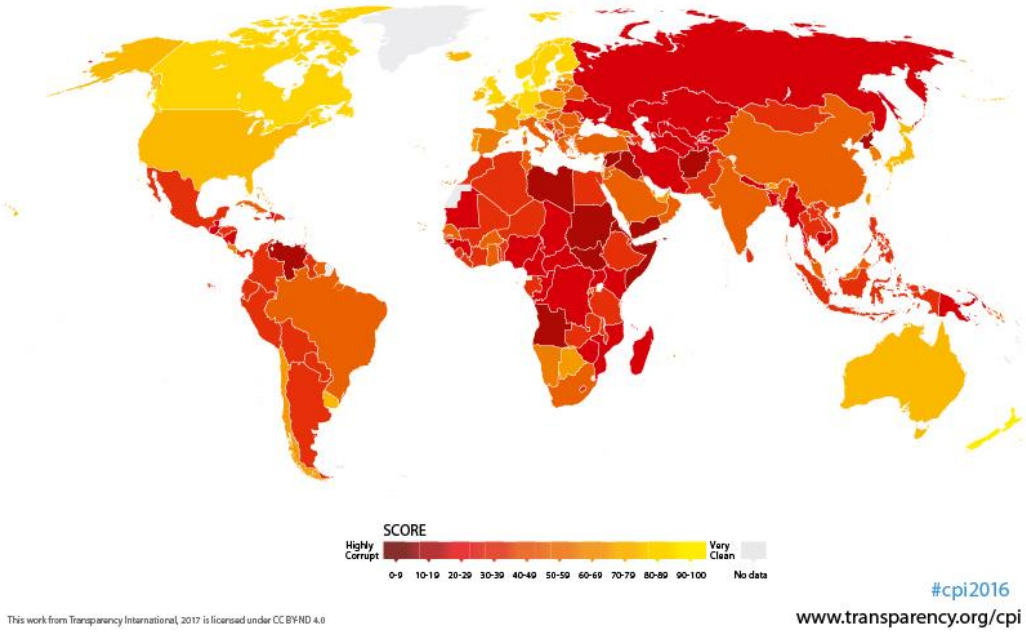
$$C = M + S - R$$

$$dove: C = corruzione,\ \ M = monopolio,\ \ S = segretezza\ e\ \ R = responsabilità$$

In altri termini, il fenomeno corruttivo è dato dalla somma di monopolio più segretezza, meno responsabilità. Ove maggiore sia il monopolio e la segretezza e minore sia la responsabilità e l'obbligo di trasparenza e rendicontazione, maggiore sarà la corruzione.
Secondo lo studioso, la corruzione è tanto più ricorrente quanto più alta è la somma del valore del monopolio, ossia della possibilità che uno o più soggetti dispongano di un bene in via esclusiva, e della segretezza, quindi la procedura di aggiudicazione sia priva di trasparenza.
La responsabilità, nel nostro sistema giuridico, è costituita da molteplici forme (civile, penale, amministrativa, amministrativo-contabile, dirigenziale), a seconda delle regole che vengono infrante.
Se gli elementi monopolio e segretezza assumono valori elevati, efficace strumento per diminuire la propensione alla corruzione può essere rappresentato dall'alta probabilità della certezza di una che il mero inasprimento delle pene edittali.

Dati della corruzione

Nel nostro paese la corruzione è presente in maniera abbastanza pervasiva, ed infatti l'Italia si posiziona al 60° posto nel contesto internazionale secondo l'ultima rilevazione dell'Indice di Percezione della Corruzione (CPI) su 176 paesi pubblicata da Transparency Internazional nel 2016, pur essendoci un lieve miglioramento di posizione rispetto all'anno scorso in cui era 61esima. Il voto assegnato al nostro Paese è di 47 su 100, e ci vede migliorare anche qui, di 3 punti significativi.



In Europa però, la situazione per l'Italia non può dirsi ottimale: si trova infatti come "fanalino di coda", seguita solo da Grecia e Bulgaria, rispettivamente al 69° e 75° posto della classifica mondiale.

Secondo quanto afferma Transparency International Italia, dal 2012, anno dell'approvazione della Legge Severino, l'Italia ha scalato 12 posizioni, passando dalla 72°, all'attuale 60°. Il nostro Paese segna quindi un miglioramento, ma ancora troppo poco per potersi dire soddisfatti.

Misurare correttamente in termini monetari il danno che l'Italia subisce a causa della corruzione è evidentemente difficile, a causa della natura di questo fenomeno.

Nel corso del 2012, molte testate giornalistiche nazionali hanno riportato la cifra di 60 miliardi di euro, circa il 3% del PIL italiano, come stima attendibile dei danni causati dalla corruzione.

Il problema relativo a questo valore è che non è stato ottenuto attraverso un ragionamento corretto: infatti, nel 2004 la Banca Mondiale pubblicò un rapporto in cui teorizzava che la corruzione, a livello mondiale, ammontasse a circa mille miliardi di dollari – il 3% del PIL globale dell'epoca. Questo 3% è stato "applicato" al PIL italiano, generando un risultato pari a 60 miliardi.

In Italia "qualcuno" ha dunque pensato di fare una semplice proporzione per poter affermare che la corruzione nel nostro paese costava ogni anno 60 miliardi, commettendo due grossolani errori: primo, la ricerca si riferiva al costo delle tangenti; secondo, la Banca Mondiale sottolineava come questa percentuale vari da paese a paese.

Ma anche se fosse solo la metà della cifra dichiarata in Italia sull'indice di corruzione, si parlerebbe di un quantitativo di denaro enorme, capace di portare forti disquilibri monetari.

Il peso della corruzione nella società

La diffusione della corruzione comporta danni economici "diretti" di svariati miliardi di euro, di meno agevole quantificazione sono i costi "indiretti": basti pensare ai ritardi nella definizione di pratiche amministrative, al cattivo funzionamento degli apparati pubblici, all'inadeguatezza, se non all'inutilità, di alcune opere pubbliche. In particolare, si assiste a un rialzo straordinario nel costo delle grandi opere.

In una prospettiva più generalizzata, la corruzione compromette la fiducia dei mercati e delle imprese determinando una perdita di competitività: gli operatori scelgono di investire in Paesi che diano, in tal senso, maggiori garanzie.

Al riguardo, è stato calcolato che ogni punto di discesa nella classifica redatta ogni anno dal Transparency International, provoca la perdita del sedici per cento degli investimenti.

Volendo rivolgere uno sguardo al passato, si è avuto modo di constatare che se l'Italia avesse avuto un valore nell'indice di percezione della corruzione di Transparency International al livello di uno dei Paesi meno corrotti il tasso di crescita economica sarebbe stato di oltre il triplo a breve termine e di circa il doppio a lungo termine, dunque è evidente

la presenza di correlazione tra l'indice di corruzione indicato da Transparency e l'incremento del Pil.

Se, poi, si procede verso un'analisi più puntuale dell'impatto della corruzione sulle imprese, si registra che un basso livello di corruzione è solitamente associato ad una regolamentazione che favorisce la nascita di nuove imprese ed un tempo medio contenuto per dare inizio ad un'attività economica, mentre procedure burocratiche eccessivamente lunghe sono generalmente associate ad alti livelli di corruzione. Quanto agli effetti negativi della corruzione, analoghe valutazioni possono essere espresse riguardo al costo del capitale, al livello di competitività e, più in generale, alla qualità del business environment. Uno studio della Banca Mondiale condotto su un ampio numero di Paesi mostra come le imprese costrette a fronteggiare una pubblica amministrazione corrotta e che devono pagare tangenti crescono in media quasi del 25% di meno di imprese che non fronteggiano tale problema. Aspetto ancora più preoccupante è che ad essere più fortemente colpite sono le piccole e medie imprese e le imprese più giovani. Il Rapporto della Banca mondiale, inoltre, dimostra come le piccole imprese abbiano un tasso di crescita delle vendite di più del 40% inferiore rispetto a quelle grandi.

Il terrorismo internazionale

Il terrorismo internazionale è un fenomeno che, sebbene dai confini incerti, sta attualmente rappresentando una sfida globale all'interno della comunità internazionale. La definizione di tale fenomeno rappresenta, ancora oggi, una problematica rilevante per l'intera Comunità internazionale, non esistendo una nozione cristallizzata che possa delimitarne la portata.

Vengono coinvolti molto spesso molteplici Stati, i quali sono chiamati a cooperare tra loro per reprimere tale fenomeno e per tale motivo si rende necessaria una sua definizione generalmente accettata.

Tuttavia, è difficile ottenere consenso su una nozione giuridica di terrorismo internazionale. Ogni Stato, infatti, nel legiferare in materia, definisce il terrorismo in modo diverso, associandolo a reati interni comuni, come l'omicidio, il rapimento, la tortura, rendendo perciò ardua e disomogenea una repressione a livello internazionale.

È necessario, dunque, un quadro comune per far sì che gli Stati possano trovarsi d'accordo sull'obiettivo della loro azione repressiva, tenendo presente che il terrorismo internazionale è una minaccia difficile da sradicare dalla società a causa del forte tasso di eterogeneità dei suoi elementi costitutivi.

Anche se ad oggi non vi è una definizione condivisa di terrorismo, nel diritto internazionale con tale accezione si fa riferimento ad un complesso di attività criminose compiute al fine di provocare panico presso singoli individui, gruppi di persone o un'intera collettività e quindi perseguire finalità ulteriori, generalmente di natura politica.

I diversi strumenti giuridici internazionali di contrasto al fenomeno, susseguitisi nel tempo, pur se apprezzabili come importanti spunti progressivi, dimostrano il loro limite nell'individuazione di una definitiva soluzione definitoria testimoniando, però, al contempo l'impegno manifestato da ogni Stato nella lotta al fenomeno terroristico, un dovere quest'ultimo, accentuatosi dopo l'11 settembre 2001 e che ha dato l'avvio ad una innovativa forma di contrasto basata sull'aggressione dei flussi di finanziamento.

Metodi di finanziamento al terrorismo

È noto oramai che il terrorismo internazionale ha bisogno di fondi per sostenere la propria attività. Con la globalizzazione dei mercati finanziari internazionali, il tema del finanziamento del terrorismo internazionale si è collocato al centro di vivaci dibattiti in seno alle organizzazioni internazionali e regionali. I terroristi sono infatti in grado di affrontare le opportunità fornite dai mercati attraverso attività che sono legali ma che si indirizzano verso obiettivi illegali, quali la pianificazione di un attentato.

In questi termini, si dovrebbe intendere con "finanziamento del terrorismo" la fornitura di denaro per uno scopo terroristico.

Si stima del resto che solo il 50-55% dei flussi finanziari provenienti dalle rimesse degli emigrati passi attraverso canali formali. Una parte rilevante delle risorse trasferite non viene assorbita infatti dal sistema bancario e dai canali ufficiali, ma si avvale di modalità di trasferimento informali o semi-formali, che sfuggono inevitabilmente ad ogni statistica e contabilizzazione.

In questi circuiti informali di trasferimento fondi si annidano peraltro i maggiori rischi, sia sotto il profilo del riciclaggio, sia sotto quello del finanziamento del terrorismo internazionale.

La misurazione di questi flussi internazionali di denaro non è però agevole, anche perché le statistiche ufficiali non conteggiano il flusso di rimesse che passa attraverso canali informali di intermediazione, che vanno dalla consegna personale a mano durante i periodici viaggi nel paese d'origine, all'invio tramite amici e familiari, al ricorso ad organizzazioni professionali di trasferimento finanziario non registrate, come il sistema cinese chop o flyng monet, quello colombiano del black market peso Exchange e i sistemi hawala o hundi, conosciuti in Asia meridionale.

Secondo quanto emerge, ad esempio, dalle ultime indagini della Direzione Nazionale Antimafia, la Guardia di Finanza ha recentemente individuato ben 410 agenzie di *money transfer* abusive in piena attività. In questi esercizi, in prevalenza rivendite di tabacchi, ricevitorie del lotto, phone center e internet point, sono state eseguite 280 mila operazioni per 88 milioni di euro, tutti trasferiti all'estero senza verifiche e controlli.

Si tratta di un vero e proprio "sistema bancario parallelo o alternativo", che rischia di mettere in crisi anche quello legale, essendo stati identificati circa 25 mila punti di raccolta presenti in Italia, dei quali si stima che il 30 per cento, circa 8 mila, sia illegale.

L'ampiezza e pericolosità del fenomeno non deve essere dunque sottovalutata.

Dunque le fonti di reddito principali del terrorismo sono: i sequestri, il contrabbando di armi e il traffico di droga.

I gruppi terroristici internazionali hanno fatto dello spaccio di sostanze stupefacenti un corridoio per sovvenzionare le proprie attività. La base del fenomeno risiede storicamente negli accordi strutturali tra terroristi e narcotrafficanti. I primi si garantiscono introiti eterodiretti, i secondi di una difesa naturale delle loro zone franche che, presidiate dai terroristi, rimangono di difficile tracciabilità per gli Stati.

Nel 2009, Antonio Maria Costa, direttore esecutivo dell'Unodc, L'Ufficio per i narcotici e il crimine delle Nazioni Uniti, dichiarò in merito: "Se non recidiamo il legame tra crimine, droga e terrorismo, il mondo assisterà alla nascita di un ibrido e cioè di organizzazioni terroristiche della criminalità organizzata". La tipologia di criminale che ne è venuta fuori, insomma, è quella del "narcoterrorista".

Facendo riferimento a un paese come l'Afghanistan, dove l'assenza di strutture statali forti e solide è quasi da sempre una prerogativa, inoltre, il suo essere al centro di quella regione nota come Mezzaluna d'oro fa sì che esso sia, ancora oggi, il cuore pulsante della produzione di oppio nella regione, nello specifico, l'oppio rappresenta il 60% del Pil afghano.
Basti pensare che circa il 92% della produzione mondiale di oppiacei avviene proprio in questo crocevia dell'Asia centrale.

Capitolo 2 - IL CLASSIFICATORE NAIVE BAYES

In questo secondo capitolo viene trattato nel dettaglio l'argomento principe della tesina, ossia l'algoritmo vero e proprio: il classificatore Bayesiano Naive. In una prima parte del capitolo, dopo una breve introduzione storica sull'autore, viene introdotto il Teorema di Bayes, su cui l'algoritmo si basa, e vengono infine esposti alcuni esempi mirati ad una maggiore comprensione del teorema stesso. Successivamente dopo aver introdotto i concetti di inferenza Bayesiana, viene esaminato l'algoritmo Naive Bayes ed inseguito ne viene presentato un tipico esempio basato sul calcolo del rischio.

Cenni storici sull'autore

Thomas Bayes nacque a Londra nel 1702 e morì il 17 aprile 1761 a Tunbridge Wells. È stato un matematico nonché pastore presbiteriano. È noto in statistica per il suo Teorema di Bayes sulla probabilità condizionata, pubblicato postumo nel 1763.
L'importanza del teorema che porta il suo nome è tale da aver dato vita a un intero approccio alla statistica, la statistica bayesiana. Pur senza aver mai ricoperto cariche accademiche e pubblicato lavori a suo nome, Bayes fu eletto Membro della Royal Society nel 1742.
È sepolto nel cimitero Bunhill Fields di Londra.

Introduzione al teorema di Bayes

Il teorema di Bayes venne presentato nel 1763 nell'articolo *Essay Towards Solving a Problem in the Doctrine of Chances* di Thomas Bayes, pubblicato postumo in *Philosophical Transactions of the Royal Society of London*.
Il teorema di Bayes deriva da tre teoremi fondamentali delle probabilità: il teorema della probabilità condizionata, il teorema della probabilità composta ed il teorema della probabilità assoluta.
Questi tre teoremi dicono rispettivamente che:

- Teorema della probabilità condizionata:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

la probabilità di A condizionata da B è definita come la probabilità che si verifichi l'evento A, a condizione che si verifichi pure l'evento B, entrambi eventi dello spazio S di cui B tale che $P(B) > 0$.

- Teorema della probabilità composta:

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$$

per cui la probabilità che due eventi si verifichino contemporaneamente è pari alla probabilità di uno dei due eventi moltiplicato con la probabilità dell'altro evento condizionato dal verificarsi del primo.

- Teorema della probabilità assoluta:

afferma che se A_1 ……,A_n formano una partizione dello spazio di tutti gli eventi possibili Ω (ossia $A_i \cap A_j = \emptyset \ \forall i,j$ e $U_{i=1}^{n} A_i = \Omega$) e B è un qualsiasi evento, allora:

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$

Date le premesse enunciate, il Teorema di Bayes viene spesso utilizzato per calcolare la probabilità a posteriori conseguentemente a delle osservazioni. La sua formulazione più nota è:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Nel caso però non sia disponibile direttamente $P(B)$ si ricorre alla regola delle alternative, nella forma

$$P(B) = P(B \cap A) + P(B \cap \bar{A})$$

o nella forma

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$$

dove:

- $P(A)$: è la probabilità a priori (o probabilità marginale) di A, ovvero calcola la probabilità di A prima di aver raccolto informazioni addizionali senza tenere conto dell'evento B. Per brevità è detta *prior*.
- $P(A|B)$: è la probabilità a posteriori (o probabilità condizionata) di A dato B, poiché dipende dal verificarsi di B. Questa probabilità è anche chiamata probabilità a posteriori, poiché calcola la probabilità di A dopo che ho preso atto che l'evento B è accaduto. Per brevità è detta *posterior*.
- $P(B|A)$: è la probabilità condizionata di B dato A, che è nota come la verosimiglianza dei dati, o *likelihood*, e rappresenta la probabilità che si verifichino i dati sotto l'ipotesi che A sia vera.
- $P(B)$: è la probabilità a priori (o probabilità marginale) di B ed è chiamata costante di normalizzazione

Esempi di applicazione al teorema di Bayes

In questa sezione viene presentato un semplice esempio che serve per far intuire immediatamente l'importanza dei concetti appena descritti:

Esempio: Individuare la probabilità di una meningite

Un dottore sa che la meningite causa rigidità del collo per il 50% dei casi, quindi

$$P(rigidità\ del\ collo|meningite) = \frac{1}{2} = 0.5$$

La probabilità incondizionata che un paziente possa avere la meningite è

$$P(rigidità\ del\ collo) = \frac{1}{50000} = 0.00002$$

La probabilità incondizionata che un paziente possa avere rigidità del collo è

$$P(rigidità\ del\ collo) = \frac{1}{20} = 0.05$$

Per calcolare la probabilità che un paziente con rigidità del collo abbia la meningite, si applica il teorema di Bayes in questa maniera:

$$P(M|R) = \frac{P(R|M)P(M)}{P(R)} = \frac{0.5\ x\ 0.00002}{0.05} = 0.0002$$

$$dove\ P(M) = P(meningite)\ \ e\ \ P(R) = P(rigidità\ del\ collo)$$

La probabilità che quindi un paziente che ha rigidità del collo abbia la meningite è uguale a 0,002.

Statistica bayesiana

Tradizionalmente la statistica si occupa di raccogliere informazioni, elaborarle ed eventualmente discutere la validità dei modelli. Una volta raccolti i dati, il problema principale della statistica è di usarli per validare o meno delle ipotesi. Il vantaggio di ciò è che non si entra troppo nel dettaglio di quali dati si tratta né di quali ipotesi stiamo considerando. Ovviamente il prezzo è che le affermazioni saranno sempre di tipo probabilistico, ma è possibile quantificare il grado di fiducia (o incertezza) su quanto affermato.

L'approccio Bayesiano ha le sue fondamenta in un concetto di probabilità che è diverso dall'approccio classico della statistica, che secondo la definizione tradizionale, la probabilità di un evento risulta 'il rapporto fra il numero di esperimenti in cui esso si è verificato ed il numero totale di esperimenti eseguiti nelle stesse condizioni, essendo tale numero opportunamente grande'.

La probabilità Bayesiana di un evento assume una concezione del tutto soggettivista, in quanto viene ad esprimere il grado di fiducia che lo sperimentatore ripone nel verificarsi di un dato evento e dipende quindi dallo stato di conoscenza, o di ignoranza di tale evento, che fa parte dell'esperienza acquisita in precedenza da ciascuno.

La teoria bayesiana, quindi, fa ricorso ad una concezione non legata strettamente e unicamente all'evento ma anche e soprattutto al soggetto assegnante la funzione di probabilità. Quest'ultima, non è caratteristica dell'evento ma si viene a trovare tra individuo e mondo esterno, seguendo criteri razionali nell'assegnazione, i quali però non imporranno che la probabilità dell'evento sia uguale per ogni decisore ma potrà invece variare in base a tutto il bagaglio culturale acquisito e alle informazioni a disposizione.

I classificatori

Un sistema di classificazione o di riconoscimento, considerato in senso ampio, ha il compito di fornire ad un utente, che sia uomo o calcolatore, una valutazione della realtà fisica osservata e tale valutazione si avvale di una suddivisione della realtà in insiemi, aventi caratteristiche omogenee, detti classi.

L'obiettivo dell'analisi di classificazione è la verifica dell'esistenza di differenze tra le classi in funzione delle variabili considerate e la formulazione di un modello che sia in grado di assegnare ciascun campione alla classe cui esso appartiene.
Esistono vari tipi di modelli di classificazione e differiscono per il formalismo utilizzato per rappresentare la funzione di classificazione.
Elenco di alcuni modelli di classificazione:

- Logici: (es. Alberi e Regole di Decisione) la funzione di classificazione è espressa mediante condizioni logiche sui valori degli attributi
- Statistici: (es. Naive Bayes) memorizzano i parametri delle varie distribuzioni di probabilità relative alle classi ed agli attributi. Per classificare un generico oggetto si stimano le probabilità di appartenenza alle varie classi
- Basati sugli esempi: (es. Nearest neighbor) memorizzano tutti gli esempi del training set ed assegnano la classe ad un oggetto valutando la "somiglianza" con gli esempi memorizzati.
- Matematici: (es. SVM) la funzione di classificazione è una funzione matematica, di cui si memorizzano i vari parametri

Alberi e regole di decisione

Si tratta di un classificatore con struttura ad albero, in cui ogni nodo può essere o foglia o nodo interno: se foglia, indica il valore della classe assegnata all'istanza; se nodo interno, specifica il test effettuato su un attributo. Per ciascun valore assunto da un attributo in un test, l'algoritmo crea un ramo e il relativo sottoalbero.
Il focus principale dell'algoritmo di crescita dell'albero di decisione è come scegliere gli attributi da testare in ciascun nodo interno dell'albero. L'obiettivo è selezionare gli attributi più utili per classificare le istanze di training attraverso una strategia top down, che consiste in una ricerca avida degli attributi senza tornare a riconsiderare le precedenti scelte. Il criterio di suddivisione con cui crea un nuovo nodo si basa sul massimo guadagno di informazione. In pratica sceglie l'attributo che riesce a dividere "meglio" le istanze appartenenti a classi diverse. Quando tutti gli elementi in un nodo hanno la medesima classe, l'algoritmo non procede con ulteriore suddivisione.

Per evitare overfitting, l'algoritmo inizia una eventuale fase di riduzione: individua gli attributi che non hanno contribuito ad una consistente suddivisione delle istanze ed elimina i rispettivi nodi riunendo le istanze al livello superiore. Quando l'algoritmo termina, è possibile percorrere l'albero dalla radice e, seguendo il percorso risultante dai singoli test presenti su ogni nodo interno, si ottiene la classificazione dell'istanza.
L'albero di decisione non funziona bene quando la classificazione prevede numerose classi e un numero relativamente piccolo di esempi. Inoltre la fase di training può essere computazionalmente costosa perché deve confrontare tutte le possibili suddivisioni ed eventualmente effettuare la riduzione, anch'essa molto costosa.

Naive Bayes

The Naive Bayes è un classificatore lineare particolarmente semplice, basato sul teorema di Bayes e su una forte assunzione di indipendenza, tanto che a volte è noto come "*modello a feature indipendenti*". In pratica si assume che la presenza o l'assenza di una particolare feature di una classe non sia correlata alla presenza o assenza di altre features. Dunque il contributo di ogni feature è considerato indipendente dagli altri.
Sebbene questa assunzione sia spesso falsa, il classificatore si dimostra paradossalmente efficiente nel supervised learning, senza contare l'enorme beneficio in termini di complessità computazionale.
I vantaggi di questo algoritmo comprendono la semplicità del modello stesso facile da implementare, la velocità sia in fase di training che di classificazione, lo spazio di memoria richiesto abbastanza contenuto, la preparazione di un set training non troppo vasto. Sebbene tale modello non sia del tutto "corretto", in quanto assume gli attributi indipendenti tra loro, lavora bene in un sorprendente gran numero di casi perché spesso si è interessati all'accuratezza della classificazione più che all'accuratezza delle stime di probabilità.

K-Nearest-Neighbors

Anche questo classificatore è considerato tra i più semplici del machine learning: esso memorizza le istanze del training, poi, basandosi su un criterio di vicinanza, mette in

relazione l'istanza da classificare con alcune istanze del training set presenti nello spazio delle feature. In pratica, l'istanza è classificata “a maggioranza” in base alla classe più comune tra le k istanze più vicine del training.
Per classificare una nuova istanza, l'algoritmo cerca nel training set l'istanza “più simile” a quella data. Poi consulta le k istanze più vicine e sceglie la classe maggioritaria assegnandola alla nuova istanza. L'accuratezza dell'algoritmo può subire un forte calo per la presenza di feature irrilevanti.
È un algoritmo robusto, computazionalmente costoso sebbene concettualmente semplice, che ottiene spesso buoni risultati. Tuttavia la performance è molto condizionata dalla misura di similarità usata e trovare una buona misura di similarità può essere difficile. In generale esiste un compromesso tra la complessità di un modello e l'accuratezza della classificazione.
Il KNN rappresenta l'estremo della complessità: pur essendo un algoritmo con buone prestazioni, cattura anche il rumore. Il Naive Bayes si pone invece all'estremo opposto: la sua assunzione di indipendenza lo rende un modello computazionale semplice ma il confine tra due regioni di decisione è una linea retta che, come tale, ottiene un adattamento minore.

SVM Support Vector Machine

Senza voler entrare in dettagli formali, l'idea principale di questo classificatore consiste nel rappresentare gli esempi del training come punti nello spazio mappati in modo tale che punti appartenenti a classi differenti siano separati dal più ampio gap possibile. I punti che mappano il test set saranno assegnati ad una categoria o all'altra in base al lato del gap su cui cadono. Più specificatamente, SVM costruisce un iperpiano ed esegue una buona separazione quando l'iperpiano ha la più ampia distanza dai punti di training più vicini di ciascuna classe. Ci sono molti iperpiani che potrebbero classificare il dato.
La miglior scelta è quella di selezionare l'iperpiano che rappresenta la più ampia separazione, o margine, tra due classi, ossia l'iperpiano tale che la distanza tra esso e il punto più vicino su ciascun lato sia massima. Se esiste, tale iperpiano è noto come il massimo margine di iperpiano e il classificatore lineare è definito come classificatore a margine massimo.

Valutazione dei classificatori

In questo paragrafo viene trattata la metodologia con cui viene valutato un classificatore, ossia come riuscire a capire se il classificatore in questione ha prodotto dei buoni risultati oppure no.

Al termine del processo di classificazione è importante valutare l'affidabilità del modello per fini predittivi, che si esamina attraverso l'analisi di una tabella, detta "matrice di confusione", nella quale sono visibili gli oggetti realmente appartenenti a ciascuna classe (classe reale) e gli oggetti assegnati a ciascuna classe dal modello (classe predetta).

Sotto è rappresentata una Matrice di Confusione di un problema di classificazione binaria, è però possibile estendere questo concetto anche a problemi multi-class.

		CLASSE PREDETTA	
		-	**+**
CLASSE REALE	**-**	*True Negative*	*False Negative*
	+	*False Positive*	*True Positive*

- True Positive (TP): numero di record positivi classificati correttamente come positivi.
- False Positive (FP): numero di record negativi classificati erroneamente come positivi.
- True Negative (TN): numero di record negativi classificati correttamente come negativi.
- False Negative (FN): numero di record positivi classificati erroneamente come negativi.

A partire dai valori della matrice è possibile estrarre informazioni e metriche più significative rispetto al modello sviluppato.

Sensibilità: Nota anche come TPR (True Positive Rate) o Recall. Indica la frazione dei record positivi classificati correttamente positivi. Considerando un classificatore come un test diagnostico, la sensibilità rappresenta la proporzione di persone malate che effettivamente vengono etichettate come tali. Può essere vista come la capacità del classificatore di identificare record positivi.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Specificità: Nota anche come TNR (True Negative Rate). Indica la frazione dei record negativi classificati correttamente come negativi. Considerando un problema medico rappresenta la proporzione di persone che il classificatore identifica come tali. Esprime la capacità del classificatore di identificare risultati negativi.

$$TNR = \frac{TN}{TN + FP}$$

False Positive Rate: Indica la frazione dei record negativi classificati come positivi.

$$FPR = \frac{FP}{FP + TN}$$

False Negative Rate: Indica la frazione dei record positivi classificati come negativi.

$$FNR = \frac{FN}{FN + TP}$$

Precision: Essa indica la frazione dei *True Positive* rispetto a tutti i risultati positivi. Maggiore è la *precision* meno sono i falsi positivi commessi dal modello.

$$Precision = \frac{TP}{TP + FP}$$

Tra *sensibilità* e *precision* vi è una sottile differenza: la prima indica la percentuale di pazienti identificati come malati tra tutti quelli che lo sono nella realtà, mentre la seconda

indica la percentuale di pazienti realmente malati tra tutti quelli etichettati come malati dal classificatore.

Misura dell'accuratezza

In generale l'accuratezza da un'idea della qualità di uno strumento in esame, nel caso specifico dei classificatori è definita come il numero di campioni correttamente classificati rispetto al numero totale di campioni classificati. Esistono diverse tecniche per valutare l'accuratezza che un algoritmo di costruzione di un modello può garantire. Solitamente, per i classificatori, si utilizza un set di addestramento, chiamato "*training-set*", costituito da campioni di cui si conosce a priori la classe di appartenenza, curandosi del fatto che tale insieme sia significativo e completo, cioè con un numero sufficiente di campioni rappresentativi di tutte le classi. Per la verifica del metodo di riconoscimento ci si avvale di un set definito "*test-set* o validation-set", anch'esso costituito da campioni la cui classe è nota; esso però è costituito da un insieme di campioni diverso rispetto a quelli del training-set.
In seguito a questa operazione possiamo calcolare l'errore sia di training che di validazione:

$$E_{TRAINING} = \frac{Campioni\ sbagliati\ in\ training}{N.campioni\ training}$$

$$E_{VALIDATION} = \frac{Campioni\ sbagliati\ in\ validation}{N.campioni\ validation}$$

L'errore di validazione ci dà informazioni su quanto bene ha imparato il classificatore.

La curva di ROC

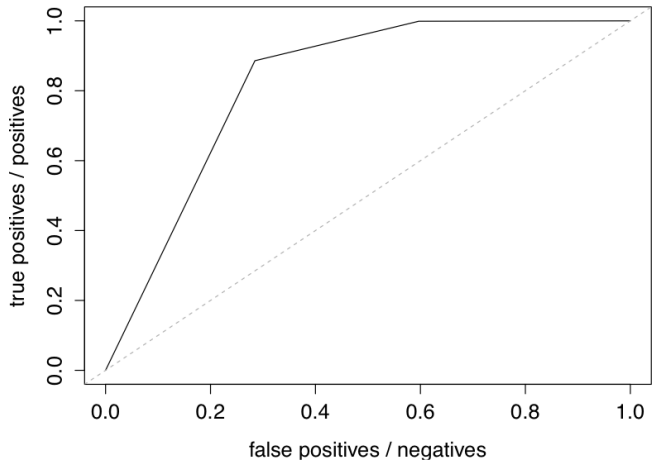
I grafici delle curve ROC (Receiving Operating Characteristics) consentono di valutare in modo visivo l'accuratezza di un classificatore e di confrontare tra loro diversi modelli di

classificazione. Essi corrispondono al contenuto informativo di una sequenza di matrici di confusione e permettono di determinare il compromesso ideale tra il numero di osservazioni positive classificate correttamente e il numero di osservazioni negative classificate in modo errato.

Un grafico ROC è un diagramma bidimensionale che riproduce sull'asse delle ascisse la percentuale di *False Positive* e sull'asse delle ordinate la percentuale di *True Positive.* Il punto di coordinate (0,1) rappresenta il classificatore ideale, che non commette alcun errore di predizione, in quanto la sua percentuale di *False Positive* è nulla e quella di *True Positive* è 1.

Un'utile proprietà delle curve ROC è quella di essere insensibili alla variazione della grandezza dell'insieme dei record utilizzati. Se la proporzione tra record positivi e record negativi cambia in un test set allora la curva non varia. Questo è dovuto al fatto che la ROC è basata su TPR e FPR, ovvero su dei valori che dipendono dal rapporto tra gli elementi di una colonna della Matrice di Confusione e non dalla distribuzione delle classi.

Per comparare due classificatori può essere utile ridurre la ROC ad un singolo valore che rappresenta la performance, ossia l'AUC (Area Under the roc Curve), l'area al di sotto della curva. Dato che si tratta di una porzione di un'area di un quadrato di lato unitario essa ha sempre un valore compreso tra 0 e 1.



Il classificatore Naive Bayes

Questo capitolo si apre presentando il classificatore Naive Bayes, uno tra i metodi più usati in certi problemi di apprendimento. Nonostante la sua estrema semplicità, in molte situazioni riesce a raggiungere delle prestazioni che sono addirittura superiori rispetto ai più noti algoritmi di apprendimento.

Una tipologia di classificatori è basata sulla applicazione del Teorema di Bayes: questi tipi di classificatori sono noti come classificatori Bayesiani. Il principio che regola questa tipologia di classificatori si basa sul fatto che alcuni individui appartengono ad una classe di interesse con una data probabilità sulla base di certe osservazioni.

Tale probabilità è calcolata assumendo che le caratteristiche osservate possano essere tra loro dipendenti o indipendenti; in questo secondo caso il classificatore Bayesiano è detto naive in quanto assume ingenuamente che la presenza o l'assenza di una particolare caratteristica in una data classe di interesse non è correlata alla presenza o assenza di altre caratteristiche semplificando notevolmente il calcolo. Descriviamo di seguito il modello probabilistico che sta alla base dei classificatori Bayesiani naive. Data una classe di interesse C ed un insieme di caratteristiche F1, ..., Fn di un individuo si vuole conoscere con quale probabilità questo appartiene a C. Ovvero si vuole conoscere:

$$P(C|F_1, \dots, F_n) = \frac{P(C)\ x\ P(F_1, \dots, F_n|C)}{P(F_1, \dots, F_n)}$$

Tralasciando il denominatore che rappresenta una costante di normalizzazione è necessario valutare la probabilità $P(F_1, \dots, F_n|C)$, il numeratore per la teoria della probabilità composta equivale a $P(C, F_1, \dots, F_n)$ ed applicando più volte la definizione di probabilità condizionata si ottiene:

$$P(C, F_1, \dots, F_n) = P(C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2) \dots P(F_n|C, F_1, \dots, F_{n-1})$$

Con l'assunzione naive della indipendenza condizionale si assume che ogni caratteristica F_i sia condizionalmente indipendente da ogni altra caratteristica F_j con $i \neq$j.

Per cui

$$P(F_i|C, F_j) = P(F_i|C) \qquad \text{con } i \neq \text{j}$$

per la probabilità composta possiamo quindi

$$P(C, F_1, \dots, F_n) = P(C)P(F_1|C)P(F_2|C) \dots P(F_n|C) = P(C)\prod_{i=1}^{n} P(F_i|C)$$

L'apprendimento Bayesiano Naive richiede il calcolo della probabilità $P(C|F_n)$ per tutte le F_i che nel caso di un problema reale risulta intrattabile. Per semplificare si considerano solo le F_i che massimizzano la probabilità $P(C|F_n)$. Con questo tipologia di decisore, chiamato Maximum A Posteriori (MAP), il classificatore Bayesiano Naive assume la seguente:

$$C_{NB} = argmax_C\, P(C) \prod_{i=1}^{n} P(F_i|C).$$

Esempio di applicazione del classificatore Naive Bayes

Per illustrare l'impiego di un classificatore Bayesiano Naive si considera la seguente tabella, dove sono presenti una serie di dati bancari relativi ai propri clienti e alle loro transazioni:

CLIENTE	A_1 =" **Importo dare bonifici estero a fiscalità agevolata"**	A_2 =" **Importo dare bonifici estero verso paesi a rischio terrorismo"**	A_3 =" **importo operazioni per cassa"**	A_4 =" **Numero assegni liberi richiesti"**	A_5 =" **Numero di deleghe per operare"**	A_6 =" **Numero gare vinte totali"**	A_7 =" **Numero cassette di sicurezza"**	**RISCHIO**
C1	0	0	0	1	0	1	1	Basso
C2	0	0	0	1	0	2	1	Basso
C3	1	1	1	3	1	0	2	Medio
C4	1	1	2	3	2	1	0	Medio
C5	2	0	2	1	2	2	2	Medio
C6	0	0	1	0	0	0	0	Basso
C7	3	3	4	4	2	3	2	Alto
C8	4	4	3	5	3	1	3	Alto
C9	2	1	1	2	0	2	1	Medio
C10	0	0	0	0	0	0	1	Basso
C11	1	1	2	3	2	1	1	Medio
C12	0	0	0	1	0	2	0	Basso
C13	1	1	1	1	1	0	0	Basso
C14	3	2	2	3	1	2	2	Alto

Ove nelle tre colonne dove viene indicato l'importo, sono stati convertiti i valori da continui in discreti, attraverso delle soglie fisse. Infatti 0 indica realmente un importo pari a zero euro, mentre 1 rappresenta una soglia di valore fino ai 50 mila euro, e così via.

Stesso ragionamento è stato applicato negli attributi dove viene indicato numero, in quanto non viene esplicitato il numero stesso ma il range di valori, rendendo più semplice il lavoro da svolgere.

L'obbiettivo di questo esempio è quello di determinare il livello di rischio di corruzione/finanziamento al terrorismo per ogni singolo cliente della banca in base ai dati forniti.
Utilizzeremo quindi il classificatore Naive Bayes e l'insieme di dati in ingesso di questa tabella per classificare la seguente istanza:

CLIENTE	A_1 =" **Importo dare bonifici estero a fiscalità agevolata"**	A_2 =" **Importo dare bonifici estero verso paesi a rischio terrorismo"**	A_3 =" **importo operazioni per cassa"**	A_4 =" **Numero assegni liberi richiesti"**	A_5 =" **Numero di deleghe per operare"**	A_6 =" **Numero gare vinte totali"**	A_7 =" **Numero cassette di sicurezza"**	**RISCHIO**
C15	1	1	1	1	0	2	0	???

L'obbiettivo è quello di predire il valore finale (basso, medio o alto) ossia determinare quale sia il livello di rischio di corruzione/finanziamento al terrorismo per la suddetta istanza.
Sappiamo dalla teoria che:

$$C_{NB} = argmax_C\, P(C_i) \prod_{i=1}^{n} P(F_i\,|C_i) \qquad dove\ C_i \in \{basso, medio\ e\ alto\}$$

E quindi:

$$C_{NB} = argmax_C\, P(C_i)P(A_1 = 1|C_i)P(A_2 = 1|C_i)$$
$$P(A_3 = 1|C_i)P(A_4 = 1|C_i)P(A_5 = 0|C_i)$$
$$P(A_6 = 2|C_i)P(A_7 = 0|C_i)$$

Per calcolare C_{NB} ora devono essere calcolate le 24 probabilità che possono essere stimate partendo dai dati in ingesso; per prima cosa si devono trovare le probabilità dei differenti risultati finali basandoci sulla loro frequenza negli esempi:

$$P(RISCHIO = Basso) = \frac{6}{14} = 0.43$$

$$P(RISCHIO = Medio) = \frac{5}{14} = 0.36$$

$$P(RISCHIO = Alto) = \frac{3}{14} = 0.21$$

Nello stesso modo si possono calcolare le probabilità condizionali, ad esempio per ($A_7 =$ "$numero\ cassette\ di\ sicurezza$") $= 0$ sono:

$$P((A_7 = "numero\ cassette\ di\ sicurezza" = 0 \mid RISCIO = Basso) = \frac{3}{4} = 0.75$$
$$P((A_7 = "numero\ cassette\ di\ sicurezza" = 0 \mid RISCIO = Medio) = \frac{1}{4} = 0.25$$
$$P((A_7 = "numero\ cassette\ di\ sicurezza" = 0 \mid RISCIO = Alto) = \frac{0}{4} = 0$$

Utilizzando queste probabilità appena calcolate, e tutte le rimanenti che si calcolano esattamente nel medesimo modo, riusciamo a calcolare C_{NB}:

$$P(RISCHIO = Basso)P(A_1 = 1|Basso)P(A_2 = 1|Basso)$$
$$P(A_3 = 1|Basso)P(A_4 = 1|Basso)P(A_5 = 0|Basso)$$
$$P(A_6 = 2|Basso)P(A_7 = 0|Basso) = 0.0021414$$

$$P(RISCHIO = Medio)P(A_1 = 1|Medio)P(A_2 = 1|Medio)$$
$$P(A_3 = 1|Medio)P(A_4 = 1|Medio)P(A_5 = 0|Medio)$$
$$P(A_6 = 2|Medio)P(A_7 = 0|Medio) = 0.0003672$$

$$P(RISCHIO = Alto)P(A_1 = 1|Alto)P(A_2 = 1|Alto)$$
$$P(A_3 = 1|Alto)P(A_4 = 1|Alto)P(A_5 = 0|Alto)$$
$$P(A_6 = 2|Alto)P(A_7 = 0|Alto) = 0$$

Normalizzando le sopraindicate quantità sommate una per una possiamo calcolare la probabilità condizionale che il risultato finale sia: "*RISCHIO=Basso*" dati i risultati dei dati osservati.
Quindi per l'esempio corrente la probabilità è:

$$\text{Basso: } \frac{0.0021414}{0.0021414 + 0.0003672} = 0.85$$

$$\text{Medio: } \frac{0.0003672}{0.0021414 + 0.0003672} = 0.15$$

$$Alto: \frac{0}{0.0021414 + 0.0003672} = 0$$

Filtri di Features Selection

Quando si ha a che fare con dataset che riguardano molti campi del mondo reale bisogna porre l'attenzione sulla dimensionalità dei dati, ovvero sul numero delle feature disponibili. Nell'ambito del machine learning vi è un concetto noto come *Curse of Dimensionality* che afferma:
In assenza di ipotesi semplificative la dimensione del training set necessario per stimare una funzione di alcune variabili ad un certo livello di accuratezza cresce esponenzialmente con il numero di variabili.
Cercare di sviluppare un modello a partire da un dataset ad alta dimensionalità può essere molto costoso in termini computazionali e può portare a risultati non accurati.

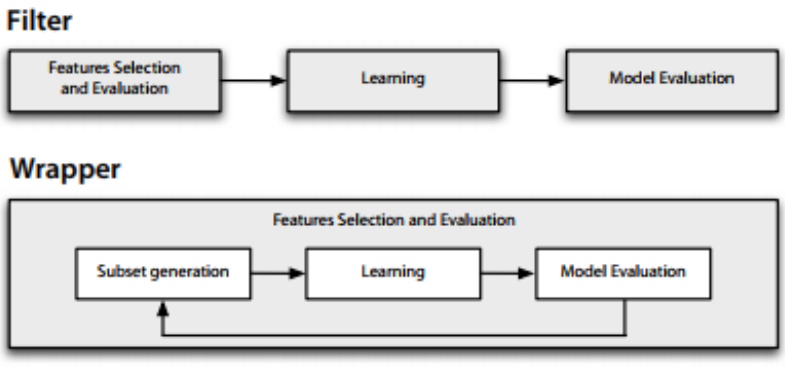
Spesso invece risulta interessante identificare un sottoinsieme di feature che siano più predittive rispetto all'insieme di partenza, inoltre alcune variabili inutili possono comportarsi in maniera simile al rumore ed impattare in modo negativo nella classificazione.

Per risolvere questi problemi esistono due tecniche: Feature Trasformation e Feature Selection. La prima cerca di trasformare le feature originali in un nuovo spazio mantenendo più informazione possibile.

Gli algoritmi di Features Selection prevedono un approccio diverso cercando di localizzare un sottoinsieme minimo ottimale di variabili originali piuttosto che trasformarle in un nuovo spazio. Le tecniche di selezione tipicamente prevedono: una strategia di ricerca per esplorare lo spazio dei sottoinsiemi di feature ed un criterio di valutazione per valutare e analizzare i candidati. Queste tecniche possono essere divise in due grandi categorie:

- Filtri: tentano di rimuovere gli attributi irrilevanti dall'insieme delle feature prima dell'esecuzione dell'algoritmo di learning. I dati sono analizzati per identificare le dimensioni che sono più rilevanti per discriminare la struttura, il sottoinsieme scelto è poi utilizzato per allenare il modello.

- Wrapper: approccio in cui le performance di un classificatore guidano la ricerca nella selezione delle feature.



Capitolo 3 - SVILUPPO DEL PROGETTO

Soluzione di contrasto al finanziamento al terrorismo ed all'individuazione dei comportamenti anomali riferiti alla corruzione

Il ripetersi di terribili attacchi terroristici, in forme e modalità imprevedibili e diversificate, ha determinato l'intensificarsi delle attività di contrasto e di prevenzione da parte delle istituzioni internazionali, comunitarie e nazionali.

Gli intermediari finanziari, i professionisti e gli operatori economici ricompresi tra i destinatari degli obblighi previsti dal d.lgs. 231/2007 sono chiamati a dare il loro contributo per contrastare questa minaccia, individuando le operazioni che possono avere lo scopo di finanziare il terrorismo ed agevolare la corruzione, chiamandole nella cosiddetta "collaborazione attiva".

Per agevolare l'individuazione delle operazioni sospette, il decreto prevede alcuni strumenti operativi: gli indicatori di anomalia, emanati da autorità diverse su proposta della UIF; i modelli e gli schemi rappresentativi di comportamenti anomali, elaborati e diffusi dalla UIF.

A tale scopo, il provvedimento ha dato enfasi al tema della tempestività dell'adeguata verifica della clientela: il professionista che prende in carico un nuovo mandato, infatti, deve essere in grado di garantire nell'immediato l'identificazione del cliente medesimo e del titolare effettivo.

Il controllo, che si sostanzia nell'acquisizione e valutazione di informazioni sul cliente e sullo scopo e natura della prestazione professionale, deve essere effettuato prima dell'instaurazione del rapporto continuativo o del conferimento dell'incarico.

Dall'altro canto, su un tema rilevante e attuale come la prevenzione della corruzione gli obiettivi fondamentali figurano il rafforzamento dell'accountability, della trasparenza e dell'integrità dei poteri pubblici e la prevenzione dei rischi di corruzione e di infiltrazione criminale al loro interno.

Secondo quanto affermato dal dott. Claudio Clemente, direttore dell'Unità di Informazione Finanziaria per l'Italia, costruire e consolidare una diffusa cultura della prevenzione sono progetti ambiziosi e complessi che, oltre a richiedere la convinta adesione a principi etici,

vanno demandati all'azione integrata di più fattori: un sistema di obblighi e divieti, un apparato coordinato di autorità di regolazione, indirizzo e controllo, strumenti e modelli volti ad accrescere la capacità di identificazione e mitigazione dei rischi, presidi organizzativi e procedurali interni, impianti per la gestione e l'analisi delle informazioni.[1]
È essenziale che tali fattori si inscrivano in una cornice di norme generali di qualità elevata, idonee a creare un contesto istituzionale di diffusa reazione, in particolare nel settore pubblico.
L'Unità di Informazione Finanziaria ha da tempo posto in evidenza i legami tra corruzione e riciclaggio, la particolare pericolosità sociale di tali fenomeni, la loro capacità di determinare gravissime distorsioni nell'economia legale. Nella quotidiana attività di analisi e approfondimento delle segnalazioni di operazioni sospette di riciclaggio, l'Unità è attenta a individuare schemi operativi, tipologie, caratteristiche soggettive e fattori di rischio tipicamente riconducibili alla corruzione. Un quadro normativo anticorruzione adeguato, non ridondante, focalizzato su misure volte a "innalzare" il costo dell'attività illecita può favorire anche l'azione di prevenzione e contrasto del riciclaggio.

Sviluppo del modello di calcolo del rischio corruttivo e di finanziamento al terrorismo

Per risolvere i due problemi elencati precedentemente, si vuole creare due modelli probabilistici distinti in grado di individuare possibili clienti con alto rischio di operazioni sospette tramite l'utilizzo di un algoritmo semplice ma al tempo stesso robusto, e che lavoro in maniera ottimale con una grande quantità di dati.
È stato scelto per l'appunto il classificatore Naive Bayes nell'individuazione di possibili clienti che effettuano operazioni sospette, e sono stati eseguiti due procedimenti di creazione di modelli distinti, questo per via delle due finalità diverse su cui è indirizzato lo scopo dell'analisi, ossia la corruzione e il finanziamento al terrorismo.

---

[1] Camera dei deputati, Roma, *Prevenzione del riciclaggio e della corruzione: strategie convergenti, modelli comuni, possibili integrazioni,* 2015. *Intervento del dott. Claudio Clemente Direttore dell'Unità di Informazione Finanziaria per l'Italia*
https://uif.bancaditalia.it/pubblicazioni/interventi/documenti/Clemente-Seminario-Italiadecide-AEqua.pdf9

Uno dei primi problemi che bisogna risolvere per analizzare i dati è il trattamento dei valori mancanti, in quanto una gestione corretta di tali valori è fondamentale per ottenere una modellazione efficace.

Un valore mancante può avere diversi significati, è possibile che il campo non fosse applicabile, che l'evento non si sia verificato o che i dati non fossero disponibili, inoltre potrebbe essere accaduto che la persona che ha immesso i dati non conoscesse il valore corretto o non abbia verificato l'effettiva compilazione di un campo.

Dunque qualora i dati forniti dalle fonti esterne non fossero disponibili, e non vi è alcuna modalità nel reperirli, tale valore mancante viene rimpiazzato con il valore zero, che indica una quantità nulla.

Una volta ripulito il dataset e rimpiazzato i valori mancanti, si procede con la fase di classificazione a due passi, la quale permette di estrarre modelli che descrivono le classi dei dati, dunque di predire le etichette di categorie che verranno applicate ai dati.

Nel primo passo, un modello è costruito per descrivere un set di dati in classi analizzando i record del database descritte dagli attributi. Ogni record assunto è appartenente a una classe predefinita, determinata da un attributo specifico, che è chiamato attributo di etichetta di classe.

Nel contesto della classificazione, i record di dati sono anche chiamati campioni, esempi od oggetti, e l'insieme di questi vengono utilizzati per costruire il modello, formando il training set.

I record singoli che costituiscono il training set sono definiti come training sample e sono selezionati in modo casuale dalla popolazione dei campioni. Fino a che l'etichetta di classe di ogni training sample è fornita, questo passo è anche chiamato supervised learning. Ciò si pone in contrasto con l'unsupervised learning, nel quale, a priori, non si conoscono le etichette di classe di ogni training sample e il numero di classi da apprendere.

I modelli probabilistici appresi saranno due, che si formeranno in base ai diversi parametri presenti nei vari attributi, qualora un attributo come "bonifici esteri verso paesi ad alto rischio di terrorismo" presentasse un parametro elevato allora tale record andrebbe a formare il training set del modello riguardante il finanziamento al terrorismo, mentre valori alti negli attributi come "Numero deleghe per operare" o "Numero gare vinte" andrebbero a costruire il modello sulla corruzione.

Dunque dal primo passo è stato appreso un modello probabilistico, mentre nel passo successivo il modello viene usato per la classificazione, e viene stimata l'accuratezza predittiva di tale modello.
L'accuratezza di un modello su un dato test set è la percentuale di campioni del test set che sono correttamente classificati dal modello stesso.
Per ogni campione di test, l'etichetta di classe conosciuta è comparata con la predizione della classe per quel campione, fornita dal modello.
Si noti che, se l'accuratezza del modello è stimata basandosi sul training data set, questa stima potrebbe essere ottimistica, fino a che il modello costruito tende a fare overfit dei dati, ciò significa che potrebbe avere incorporato alcune particolari anomalie del training set che non sono presenti in tutta la popolazione.
È consigliabile quindi un test set, ossia un dataset con record diversi da quelli usati nel training set ma sempre con l'attributo di classe etichettato. Se l'accuratezza del modello è considerata accettabile, il modello può essere utilizzato per classificare futuri record di dati o oggetti per i quali la classe non è conosciuta.

Casi reali utilizzati e processo induttivo

Secondo la definizione di Aristotele, l'induzione è il procedimento che dai particolari porta all'"universale" e tale concetto è stato sostanzialmente ribadito da filosofi e scienziati nei secoli successivi.
Ma nella seconda metà del ′600, dopo che le scienze hanno cominciato ad usare ampiamente il procedimento induttivo, tale problema è stato affrontato con rinnovato interesse, tant'è che a porlo chiaramente fu il dubbio scettico di Hume*:*

" Tutte le inferenze tratte dall'esperienza suppongono, come loro fondamento, che il futuro rassomiglierà al passato e che poteri simili saranno uniti a simili qualità sensibili. Se ci fosse qualche sospetto che il corso della natura potesse cambiare e che il passato non servisse di regola per il futuro, ogni esperienza diverrebbe inutile e non potrebbe dare origine ad alcuna inferenza o conclusione".

Giustificando tale pensiero, si è giunti a un'interpretazione probabilistica dell'induzione che porta alle due conclusioni seguenti:

1. L'induzione è il solo mezzo per ottenere previsioni
2. Essa è il solo metodo di autocorrelazione

Sull'induzione vi è quasi l'unanimità nel considerarla come un procedimento logico, opposto a quello di deduzione, per cui dall'osservazione di casi particolari si sale ad affermazioni generali o, talvolta, dalla conoscenza di fatti si risale alla conoscenza delle leggi che li regolano.

Pertanto, nella sua più ampia accezione, *l'inferenza deve essere intesa come una qualsivoglia forma di argomentazione che riguardi un processo di induzione oppure un procedimento di deduzione*; nel primo caso si può parlare di *inferenza induttiva*, nel secondo di *inferenza deduttiva*.

Per lo più, tuttavia, nella ricerca scientifica ed operativa, quando si usa il solo termine inferenza, si vuole intendere l'inferenza induttiva, la quale, non di rado, viene specificata come *inferenza statistica* o *inferenza probabilistica.*

Dunque, proprio tramite un processo di induzione, è stato possibile individuare manualmente due aziende[2], le quali mostrano con forte intensità un'attività di corruzione.

Da una prima analisi, è stato notato che l'azienda A riceve appalti da un ente pubblica, e ciò a prima vista non risulta essere inconsueto, ma può apparire già un primo campanello di allarme, in quanto negli ultimi anni sono uscite molte notizie relative alla corruzione rispetto alle vincite di questi appalti, per presunti giri di mazzette coinvolte in queste gare.

In seguito, approfondendo lo studio a riguardo, l'azienda presa in considerazione effettua bonifici di elevata somma verso una seconda azienda, azienda B, che si presenta come una società di consulenza.

La cosa più interessante è apparsa nel momento in cui sono stati individuati i nomi degli intestatari delle due aziende che si scambiavano importi di elevata quantità.

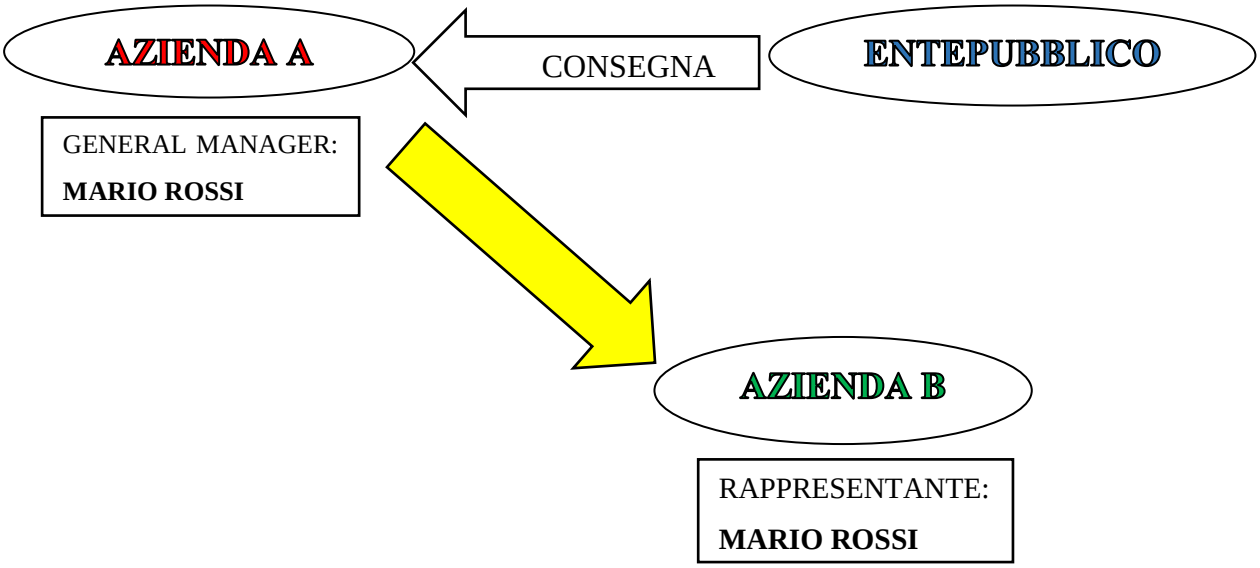
[2] Per diritti di riservatezza delle informazioni personali non sarà possibile utilizzare i nomi reali di enti e persone, dunque il caso posto in seguito presenterà nomi fittizi ma il contesto creatosi è puramente reale.

Infatti il general manager della prima azienda e il rappresentante della seconda risultano essere la stessa persona, ciò è stato possibile, tramite un software, ATOKA, il quale permette di venire a conoscenza degli intestatari dell'azienda, possedendo solo il codice fiscale di questa.

Approfondendo l'analisi sull'azienda B, la quale si occupa di consulenza, questa risulta sospetta in quanto riesce a fatturare circa un milione di euro avendo solo 4 dipendenti.

Inoltre, da riscontri effettuati manualmente sul web, si è evinto che un personaggio rappresentativo di una delle due società era in contatto a personaggi pubblici legati alla politica.

In conclusione, la circostanza analizzata mostra in primo piano un vertice apicale che fattura a sé stesso ed una società con fatturato sproporzionato rispetto a dipendenti e settore.



Da tale analisi, si denota che tali soggetti presentano un rischio alto di operazione sospetta alla corruzione, e quindi in maniera induttiva, altri record che presenteranno le stesse caratteristiche verranno predetti, tramite il classificatore che ha studiato e memorizzato il caso precedente, con il medesimo grado di rischio, tra quelli proposti dalla UIF.

Il caso appena mostrato evidenzia in maniera chiara e lineare un giro di ingenti somme di denaro, senza aver mostrato alcun sospetto agli intermediari finanziari. Ma questo è solo una tipologia di corruzione "banale" tra le tante presenti nel sistema bancario.

Capitolo 4 - RISULTATI SPERIMENTALI

Come anticipato, lo scopo ultimo di questo lavoro di tesi consiste nella sperimentazione di tecniche di machine learning per la classificazione del calcolo di rischio sulla corruzione e sul finanziamento al terrorismo. A tale proposito è stato utilizzato il classificatore "naiveBayes" presente nel software R all'interno del pacchetto "e1071".
Sono stati realizzati, quindi, vari esperimenti atti a studiare la capacità di classificazione dell'algoritmo e la performance migliore nell'attribuire correttamente l'etichetta di classe.

Data set utilizzati

I data set di input forniti per applicare la classificazione Bayesiana mostrano lo stesso numero di attributi, pari a 59, mentre per quanto riguarda la numerosità delle righe, si presentano 45040 record nell'ambito del finanziamento al terrorismo mentre relativa alla corruzione sono presenti 844174 record.

È importante sottolineare che la numerosità degli attributi presi in considerazione in partenza non è elevata in quanto è fondamentale attuare passi graduali per la realizzazione di un modello probabilistico partendo da zero; qualora quest'ultimo presentasse delle performance eccezionali sarà necessario aumentare le variabili esplicative permettendo così al modello di fare previsioni facendo riferimento a più caratteristiche per ogni pattern.
È importante sottolineare che i set di dati utilizzati non presentano ancora l'attributo di etichetta "RISCHIO" in quanto la procedura eseguita è quella di supervised learning, ossia si segue un approccio "top-down" che è applicabile quando è noto il dominio del problema. Tale dominio viene aggiunto manualmente dall'essere umano, in base ai valori dei parametri che caratterizzano ogni attributo.

Nell'apprendimento supervisionato l'algoritmo utilizzato genera una funzione che lega i valori di input ad un output desiderato, il "RISCHIO", attraverso l'osservazione di un set di esempi nei quali ogni dato di input ha il suo relativo dato di output.

Prima di procedere con il primo step, è necessario analizzare la presenza di valori mancanti, ma come detto nei capitoli precedenti, qualora fossero presenti, sarà sufficiente sostituire tali mancanze con il valore nullo, ossia con 0.

Nel software R utilizzato per implementare la classificazione, è stata usata la seguente funzione, per eliminare la presenza di valori mancanti, dunque sostituendo a quelle celle il valore zero.

```
file[is.na(file)]<-0
```

A questo punto si è in grado di procedere con la classificazione, ma ciò è possibile solo attraverso l'osservazione di esempi già classificati.

A tal proposito sono stati etichettati due data set distinti per i due ambiti, per la corruzione è stato prelevato un set di dati da 13000 record ai quali è stata definita un'etichetta per la variabile di target, mentre per quanto riguarda il set relativo al terrorismo sono stati riclassificati 5000 record. L'attribuzione delle etichette segue due criteri distinti, ed in base a relativi attributi è stato possibile distinguere due rischi distinti per i due ambiti.

La realizzazione di un modello di input etichettato è stato effettuato in maniera manuale dall'uomo, anche facendo riferimento a casi reali di cui si conosce già il grado di rischio.

Si può osservare nella seguente tabelle con quali criteri sono stati attribuiti i cinque diversi livelli di rischio (1=basso, 2=medio/basso, 3=medio, 4=medio/alto e 5=alto) per la corruzione.

NDG	Cluster_ID	ClusterAmbito_ID	Vertice	RISCHIO	IMPORTO DARE BONIFICI ESTERI	FREQUENZA DARE BONIFICI ESTERI	IMPORTO AVERE BONIFICI ESTERI	FREQUENZA AVERE BONIFICI ESTERI	IMPORTO DARE BONIFICI ITALIA	FREQUENZA DARE BONIFICI ITALIA	IMPORTO AVERE BONIFICI ITALIA
A	614	18281	0	1	0	0	0	0	0	0	0
B	7936	18378	0	1	0	0	0	0	0	0	0
C	2754	18289	0	2	0	0	1	5	0	0	0
D	2648	18287	0	2	0	0	0	0	0	0	0
E	2189	2184	0	3	0	0	1	9	6	9	0
F	3939	3926	0	3	0	0	0	0	6	9	8
G	11819	11591	0	4	0	0	0	0	7	9	0
H	5458	5427	0	4	0	0	0	0	6	9	8
I	80	80	1	5	7	9	9	9	8	9	8
L	16970	16670	1	5	9	9	0	0	8	9	8

FREQUENZA AVERE BONIFICI ITALIA	…	…	…	IMPORTO PRELIEVI DA CONTO CORRENTE	FREQUENZA PRELIEVI DA CONTO CORRENTE	IMPORTO VERSAMENTI SU RAPPORTI CONTINUATIVI	FREQUENZA VERSAMENTI SU RAPPORTI CONTINUATIVI	NUMERO DI CONTI	NUMERO ALERT GENERATI IN AML	NUMERO CASSETTE SICUREZZA	UTILIZZO CONTANTE
3	…	…	…	0	0	0	0	2	0	0	0
1	…	…	…	0	0	4	9	5	0	0	1
2	…	…	…	0	0	1	3	3	0	0	1
1	…	…	…	0	0	5	9	2	0	0	1
0	…	…	…	0	0	8	9	6	0	0	1
9	…	…	…	0	0	9	9	6	0	1	4
0	…	…	…	9	9	9	9	3	0	0	9
9	…	…	…	9	9	1	9	6	0	0	4
9	…	…	…	0	0	0	0	6	0	0	0
9	…	…	…	0	0	0	0	5	0	0	0

È stato seguito lo stesso procedimento per quanto riguarda il finanziamento al terrorismo, nel quale però a variabili riferenti le onlus islamiche è stata data più importanza per la determinazione del rischio.

Una volta etichettati i due diversi set di dati, si può procedere con la classificazione, ma prima di eseguire tale procedimento è stato necessario applicare una tecnica di riduzione delle features, per far sì che il classificatore fosse più efficiente nel fare previsioni.

Tecnica di Features Selection

Per la realizzazione degli esperimenti, ai data set originari di 59 variabili esplicative è stata diminuita la dimensione dello spazio delle variabili utilizzando la tecnica della Features Selection. Non è stata considerata invece la Features Trasformation dato che è fondamentale ottenere un modello finale i cui attributi abbiano dei valori interpretabili fisicamente.

Adottando l'approccio "Filtro" della Features Selection sono stati rimossi gli attributi irrilevanti, ossia che presentavano un grado di significatività ridotto, prima dell'esecuzione dell'algoritmo di classificazione.

L’output prodotto dalla funzione di riduzione delle variabili esplicative è il seguente:

```
Boruta performed 13 iterations in 4.398076 mins.
 30 attributes confirmed important: `COMUNE DI NASCITA`, `ESPOSIZIONE
POLITICA`, `FAMIGLIA CALABRESE`, `FREQUENZA AVERE BO EST FISCALITA
AGEVOLATA`, `FREQUENZA AVERE BONIFICI ESTERI` and 25 more;
 29 attributes confirmed unimportant: `Cluster che effettua bonifici
esteri per un importo X > Y (stipendio medio del vertice del cluster)`,
`Cluster che riceve bonifici esteri da paesi a fiscalità agevolata per
un importo X > Y (stipendio medio del vertice del cluster)`, `Cluster
che riceve bonifici esteri per un importo X > Y (stipendio medio del
vertice del cluster)`, `Cluster in cui un componente è un pep o pil o
famigliare e vince almeno due gare`, `DELEGHE IN AZIENDE VINCENTI NON
RISCHIOSE` and 24 more;
```

Dal seguente approccio le variabili selezionate sono quelle che maggiormente riescono a fornire informazioni per l’identificazione della classe target.

Una diminuzione delle variabili di circa il 50% può essere considerato un bene, in quanto ad un problema di minore dimensionalità, in particolare, corrisponde una complessità minore in termini di tempi di risoluzione.

È anche possibile che eliminando dell’informazione non rilevante nell’input si possa ottenere una misura di prestazione migliore nello svolgimento dei compiti del programma.

Classificazione

La scelta di un modello di apprendimento automatico dovrebbe essere sempre accompagnata dalla definizione di un dominio funzionale, nel presente lavoro avendo impiegato algoritmi di apprendimento automatico specializzati in classificazione, focalizziamo l’attenzione su questa funzionalità.

La classificazione è una procedura nella quale singoli oggetti vengono assegnati a gruppi definiti, grazie a dataset contenenti informazioni inerenti agli oggetti stessi e sulla base di un training set di voci precedentemente etichettate, Si intende con "classificatore" un sistema che esegue una mappatura tra uno spazio di caratteristiche X e un set di etichette Y.

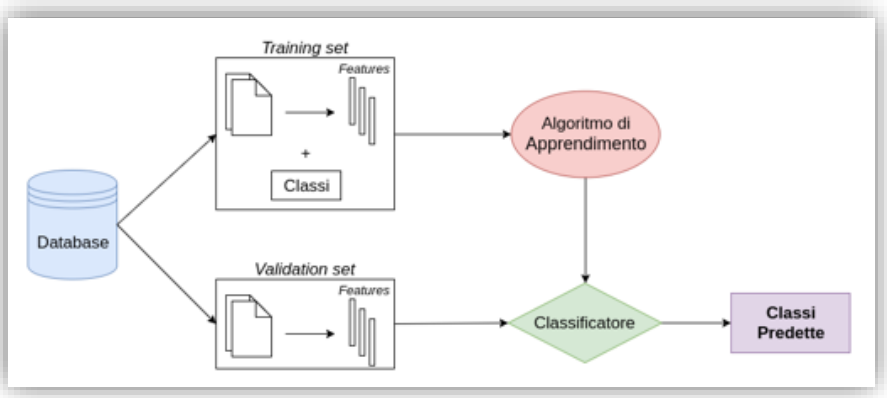
Fondamentalmente un classificatore assegna un’etichetta predefinita per ogni campione.

Durante la classificazione possono verificarsi le seguenti situazioni:

- "**Classificazione esatta**": dato un pattern X in input, il classificatore ritorna la sua etichetta Y (scalare);
- **"Classificazione probabilistica"**: Dato un pattern X in input, il classificatore restituisce un vettore Y che contiene la probabilità che Y_i sia l'etichetta giusta per X. In altre parole cerchiamo, per ogni vettore, la probabilità che esso sia un membro della classe Y_i.

Entrambi i casi possono essere applicati in esperimenti che prevedono due o più classi di classificazione, in questo caso le classi prese in considerazione sono ben cinque.
Le fasi di classificazione sono sostanzialmente tre:

- **Fase di Training**: durante questa fase l'algoritmo viene addestrato a restituire delle valutazioni di qualche tipo, a partire da un insieme di pattern di input, dotati di relativa etichetta della classe di appartenenza;
- **Fase di Testing**: prevede una serie di test durante i quali l'algoritmo "addestrato" prende in ingresso pattern e vettori di "target" mai usati durante il training e restituisce dati statistici, matrici di confusione, errore globale, nonché le etichette di classificazione ricavate per ogni pattern;
- **Fase di valutazione**: In questa fase viene dato all'algoritmo un insieme di dati di cui non si conosce il valore di output. L'algoritmo restituisce le etichette di classificazione per ogni pattern di input in base alle sue capacità ottenute durante le due fasi precedenti.



L’immagine soprastante mostra in maniera lineare lo svolgimento dell’apprendimento automatico.

Durante la fase di training, è stato preso in considerazione il Training set, ossia 2/3 del set di dati etichettato in fase iniziale, sul quale è stato applicato l’algoritmo di apprendimento, ossia il classificatore “naiveBayes”, il quale memorizza i parametri delle varie distribuzioni di probabilità relativa alle classi e agli attributi.
Poi nel Validation set, il rimanente 1/3 del set etichettato, è stato applicato il modello probabilistico il quale ha già imparato le regole per predire la classe di target tramite il set di “allenamento”.

E solo a questo punto il modello creato è in grado di predire la classe target del set di test. Ma quest’ultimo set di dati presentava già l’attributo di etichetta, attribuito inizialmente all’intero set, dunque è possibile analizzare il livello di efficienza del classificatore creato tramite una matrice di confusione.

Infatti per una corretta analisi delle prestazioni di classificazione viene calcolata una matrice di confusione, nella quale ogni riga della matrice rappresenta le istanze di una classe prevista, mentre ciascuna colonna rappresenta le istanze di una classe reale. Uno dei principali benefici della matrice di confusione è quello di verificare in modo semplice se il sistema stia mischiando due classi.
Si può concludere, riassumendo, che un comune esperimento di classificazione necessita di:

- Un dataset di training per "addestrare" il modello;

- Un dataset di test, che viene utilizzato per ottenere un giudizio finale sulla qualità del classificatore. I dati di questo dataset devono essere dati "nuovi", cioè non precedentemente utilizzati nell’addestramento;

- Un dataset di validazione che può essere fornito dall'uomo.

Le funzioni utilizzate in R sono le seguenti:

```
inTrain<-createDataPartition(file$RISCHIO,p=0.67,list=F)
datTrain<-file[inTrain,]
datTest<-file[-inTrain,]
```

Tale funzione permette di dividere il set di dati iniziale, che era stato già etichettato, in due set distinti i quali presentano dimensionalità diverse, ossia il Training set è composto dai 2/3 del set iniziale, mentre il Test set dal rimanente 1/3.
Successivamente, si applica il classificatore "naiveBayes" al Training set:

```
modello<-naiveBayes(RISCHIO ~., data=datTrain)
```

In questo modo viene applicato l'algoritmo Bayesiano al "*datTrain*" in base all'attributo "RISCHIO", che rappresenta la variabile risposta.
Utilizzando la funzione di R, "*predict*" è possibile ricevere in output i valori dell'attributo RISCHIO relativo al "*datTest*", ossia al set di test, ottenendo così la classificazione esatta.

```
previsione<-predict(modello,datTest)
```

Mentre aggiungendo alla funzione precedente *type=" raw"* ottengo la classificazione probabilistica, ossia non il valore della variabile etichetta ma la probabilità di quel valore.

```
previsione.prob<-predict(modello,datTest,type="raw")
```

A questo punto, per studiare la bontà del classificatore utilizzato, eseguo una matrice di confusione al set di test, studiando la corrispondenza tra i valori di etichetta reali e quelli predetti.

```
confusionMatrix(previsione,datTest$RISCHIO,dnn=c("Prediction","True"))
```

L’output prodotto da tale funzione è il seguente, per lo studio del rischio sul tema del finanziamento al terrorismo, dove il numero di record presenti nel set di dati è pari a 5000:

```
Confusion Matrix and Statistics

          True
Prediction   1   2   3   4   5
         1 330  73  29  15  13
         2  84 625  13  16   7
         3  22  32  41  32   9
         4  13   8  16 204   4
         5   3   7   3   3  59

Overall Statistics

               Accuracy : 0.758
                 95% CI : (0.7366, 0.7784)
    No Information Rate : 0.4485
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6502
 Mcnemar's Test P-Value : 0.002343

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            0.7301   0.8389  0.40196   0.7556  0.64130
Specificity            0.8925   0.8690  0.93906   0.9705  0.98980
Pos Pred Value         0.7174   0.8389  0.30147   0.8327  0.78667
Neg Pred Value         0.8984   0.8690  0.96000   0.9534  0.97919
Prevalence             0.2721   0.4485  0.06141   0.1626  0.05539
Detection Rate         0.1987   0.3763  0.02468   0.1228  0.03552
Detection Prevalence   0.2769   0.4485  0.08188   0.1475  0.04515
Balanced Accuracy      0.8113   0.8540  0.67051   0.8630  0.81555
```

Questa è la matrice di confusione per quanto riguarda il modello relativo al terrorismo, che presenta un grado di accuratezza “buono”, pari a 0,758, ciò indica che il 75,8% dei record è stato classificato correttamente.

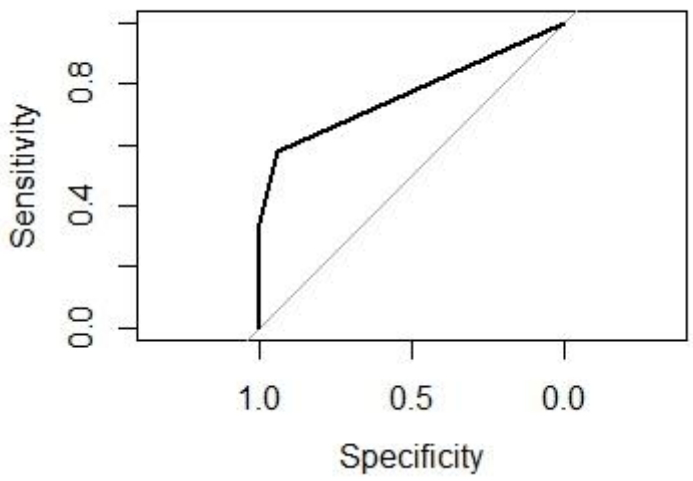
Mentre la capacità del modello di prevedere correttamente è sufficientemente alta, l’errore totale, ossia l’indice complementare dell’accuratezza, indica che l’errore di previsione commesso dal modello è pari a:

$$1 - Accuratezza = 1 - 0{,}785 = 0{,}215$$

Quindi il modello preso in considerazione, riguardante il tema del finanziamento al terrorismo, prevede correttamente per il 75,8% dei casi, contro il 21,5% di errore.
Mentre per quanto riguarda le singole classi, si può notare che la classe 2 presenta un indice di sensitività notevolmente buono, in quanto per l'83,89% dei casi la classe 2 è stata stimata correttamente sul totale delle osservazioni con la medesima classe.
Invece la classe che presenta un tasso di precisione peggiore rispetto alle altre, con 40,19%, è la classe 3, fattore noto nella matrice di confusione, poiché solo 41 pattern su 102 viene predetto correttamente, mentre i rimanenti 61 record sono predetti erroneamente.
L'output della funzione "confusionMatrix" produce inoltre come output il coefficiente statistico che rappresenta il grado di accuratezza e affidabilità in una classificazione statistica, ossia la Kappa di Cohen. Esistono diversi "gradi di concordanza", in base ai quali possiamo definire se Kappa di Cohen è scarso o ottimo:

- $k<0.2$ = concordanza scarsa;
- $0.2<k< 0.4$ = concordanza modesta;
- $0.41<k< 0.61$ = moderata;
- $0.61<k< 0.80$ = buona;
- >0.80 = eccellente.

Il valore della Kappa di Cohen ottenuto dal modello probabilistico sul terrorismo è pari a 0.6502, dunque il grado di accuratezza e affidabilità è ritenuto buono.
Infine si può ritenere il modello accettabile per fare previsione dal momento che più della metà delle volte prevede correttamente, tuttavia lo si deve utilizzare come strumento per prendere decisioni, non ci si può affidare totalmente ai suoi andamenti.
È anche ovvio che in una procedura di apprendimento automatico supervisionata il modello tende a migliorare ogni qualvolta che si è in possesso di casi reali, così da rendere più probabilistico tale classificatore.




```
Multi-class area under the curve: 0.9107
```

Infine per concludere la valutazione finale sulla performance del modello statistico, è utile osservare l'andamento della curva di ROC, e quindi anche dell'area sottostante ad essa.
La curva di ROC consente di esprimere, rappresentare e valutare l'accuratezza di una procedura diagnostica, lungo i due assi si possono rappresentare la sensibilità e la specificità.

Le curve ROC passano per i punti (1,0) e (0,1), avendo inoltre due condizioni che rappresentano due curve limite:

- una che taglia il grafico a 45°, passando per l'origine. Questa retta rappresenta il caso del classificatore casuale, e l'area sottesa AUC è pari a 0,5.
- la seconda curva è rappresentata dal segmento che dal punto (1,0) sale al punto (1,1) e da quello che congiunge il punto (1,1) a (0,0), avendo un'area sottesa di valore pari a 1, ovvero rappresenta il classificatore perfetto.

Il classificatore Bayesiano sul terrorismo presenta un'area sottostante alla curva di ROC pari a 0.9107, ciò è indice di un test altamente accurato.

Per quanto riguarda il modello costruito sul tema della corruzione, la quale presentava un dataset di 13000 record, la matrice di confusione ottenuta è la seguente:

```
Confusion Matrix and Statistics

          True
Prediction    1    2    3    4    5
         1  156  136    3    0    0
         2    8  104   34    3   39
         3    0   71 1642    6    9
         4    0   14  252  515  161
         5    0   13   83   10  992

Overall Statistics

               Accuracy : 0.8019
                 95% CI : (0.7896, 0.8138)
    No Information Rate : 0.4738
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7172
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity           0.95122  0.30769   0.8153   0.9644   0.8260
Specificity           0.96599  0.97853   0.9616   0.8851   0.9652
Pos Pred Value        0.52881  0.55319   0.9502   0.5467   0.9035
Neg Pred Value        0.99798  0.94241   0.8526   0.9943   0.9337
Prevalence            0.03858  0.07951   0.4738   0.1256   0.2825
Detection Rate        0.03670  0.02446   0.3863   0.1211   0.2334
Detection Prevalence  0.06940  0.04422   0.4065   0.2216   0.2583
Balanced Accuracy     0.95860  0.64311   0.8884   0.9248   0.8956
```

La matrice di confusione per quanto riguarda il modello relativo alla corruzione, analizza un set di dati più elevato rispetto alla precedente, 13000 pattern, i quali sono stati partizionati in due set distinti, e sul set di test è stata applicata la funzione di matrice di confusione per analizzare il grado di bontà che produce tale classificazione.

Presenta un grado di accuratezza "molto buono", infatti si denota che ben l'80.19% dei record vengono classificati adeguatamente.

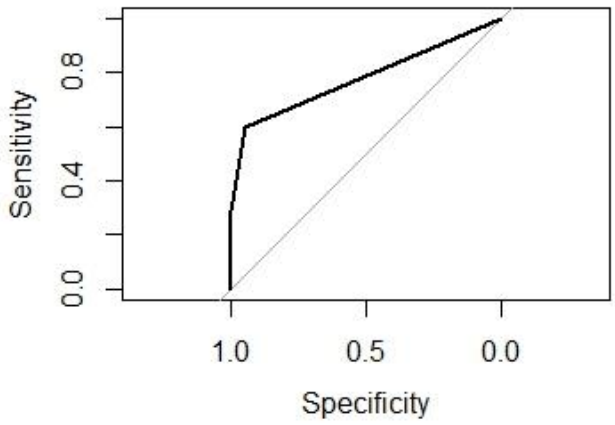
In merito all'errore totale, il livello di errore di previsione commesso dal modello è pari a:

$$1 - Accuratezza = 1 - 0{,}8019 = 0{,}1981$$

Quindi il modello costruito, riesce a produrre delle previsioni di rischio corrette per l'80.19% dei casi, mentre per l'19.81% dei casi produce previsioni errate.

Analizzando la proporzione di record corretti positivamente per le singole classi, si può notare che le classi 1 e 4 presentano un indice di sensitività ottimo, maggiore del 95%.

Invece la classe che presenta un tasso di precisione minore rispetto alle altre, con 30,76%, è la classe 2, circostanza evidente nella matrice, poiché vengono predette erroneamente più record di quanti ne vengono predetti correttamente.
In questo caso la Kappa di Cohen, viene riportata a 0.7172, ciò comporta che anche in questo caso il livello di accuratezza e affidabilità è sostenuto buono.
Alla luce dei risultati dell'analisi si può valutare in modo positivo il modello stimato, considerandolo un buon previsore per i rendimenti futuri.
I risultati del secondo classificatore hanno riportato dei valori maggiori rispetto al caso sull'analisi del terrorismo, questo fattore è anche causato dall'utilizzo di più pattern che comportano al classificatore un miglioramento durante la fase di Training set.
Per concludere l'analisi del secondo modello è interessante osservare che anche la curva di ROC e l'area sottostante sono "leggermente" maggiori rispetto alla precedente. E ciò è fattore di una maggiore bontà del classificatore.




```
Multi-class area under the curve: 0.9216
```

Valutazione

Come ultimo step, la fase di valutazione, è stato applicato il modello creato ad un nuovo dataset il quale non presenta l'attributo etichetta, dunque sarà proprio l'algoritmo creato ad attribuire la previsione del grado di rischio per ciascun pattern.

Sono stati scelti appositamente record che presentano grado di rischio pari a 5, che corrisponde a rischio "ALTO" di corruzione.

NDG	Cluster_ID	ClusterAmbito_ID	Vertice	previsione	PAESE DI NASCITA	FAMIGLIA CALABRESE	COMUNE DI NASCITA
XXX	190664	171676	0	5	-	0	-
YYY	187553	168565	1	5	-	0	-
ZZZ	187530	168542	0	5	-	0	-
KKK	202727	183739	0	5	-	0	-

PROVINCIA DI RESIDENZA	PROVINCIA RISCHIOSA	CITTADINANZA	ESPOSIZIONE POLITICA	IMPORTO DARE BONIFICI ESTERI	FREQUENZA DARE BONIFICI ESTERI	IMPORTO AVERE BONIFICI ESTERI
Roma	0	-	0	0	0	0
Pisa	0	-	0	9	9	9
Siena	0	-	0	0	0	0
Verona	0	-	0	0	0	0

FREQUENZA AVERE BONIFICI ESTERI	IMPORTO DARE BONIFICI ITALIA	FREQUENZA DARE BONIFICI ITALIA	IMPORTO AVERE BONIFICI ITALIA	FREQUENZA AVERE BONIFICI ITALIA	IMPORTO AVERE BO EST FISCALITA AGEVOLATA	FREQUENZA AVERE BO EST FISCALITA AGEVOLATA	IMPORTO PRELIEVI DA CONTO CORRENTE
0	0	0	0	0	0	0	0
9	8	9	8	9	9	9	7
0	8	9	8	9	0	0	9
0	0	9	0	0	0	0	0

FREQUENZA PRELIEVI DA CONTO CORRENTE	IMPORTO VERSAMENTI SU RAPPORTI CONTINUATIVI	FREQUENZA VERSAMENTI SU RAPPORTI CONTINUATIVI	NUMERO DI DELEGHE PER OPERARE	NUMERO SEGNALAZIONI A UIF	NUMERO DI CONTI	NUMERO CASSETTE SICUREZZA	UTILIZZO CONTANTE
0	9	9	0	1	6	0	9
9	0	0	0	0	7	0	0
9	3	9	0	1	8	1	9
0	4	9	0	1	6	0	6

È importante osservare che i record “XXX”, “ZZZ” e “KKK” presentano all’attributo “Numero segnalazioni UIF” il flag 1, che indica che tali soggetti sono già stati segnalati all’Ufficio di Informazione Finanziaria per l’Italia, dunque risulta opportuno che tali vengono riportati con un livello di rischio alto.

Prima di andare ad analizzare il record “YYY”, è opportuno analizzare i valori discreti riportati nei pattern, in quanto rappresentano soglie di valori.

Generalmente gli importi presentano la seguente mappatura, dove l’importo a sinistra è una quantità in euro:

1 = 1:25000
2 = 25001:50000
3 = 50001:75000
4 = 75001:100000
5 = 100000:150000
6 = 150001:200000
7 = 200001:300000
8 = 300001:500000
9 = >500000

Mentre le frequenze hanno un raggruppamento diverso, dove vengono rappresentati soglie di numeri discreti:

1 = 1:2
2 = 3:5
3 = 6:9
4 = 10:15
5= 16:20
6 = 21:30
7 = 31:40
8 = 41:50
9 = >50

Dunque si può osservare che il secondo record, presenta importi e frequenze elevate per quanto riguarda tutti i tipi di bonifici considerati dal modello, anche in merito al prelievo da conto corrente e al numero di conti il valore presente è considerato molto elevato. Dunque anche in questo caso, il classificatore ha riportato correttamente un grado di rischio alto, per quanto riguarda tale soggetto.

In modo analogo, è possibile analizzare i seguenti pattern relativi ai restanti gradi di rischio, ossia 1=basso, 2=medio/basso, 3=medio e 4=medio/alto. Ed è possibile dire che il classificatore ha attribuiti i relativi livelli di rischio in maniera consona ai parametri presenti negli attributi.

NDG	Cluster_ID	ClusterAmbito_ID	Vertice	previsione	PAESE DI NASCITA	FAMIGLIA CALABRESE	COMUNE DI NASCITA
AAA	189753	170765	0	**1**	-	0	-
BBB	188062	169074	0	**1**	-	0	-
CCC	194489	175501	0	**2**	-	0	-
DDD	201600	182612	0	**2**	-	0	GENOVA
EEE	204139	185151	0	**3**	-	0	-
FFF	201679	182691	0	**3**	-	0	-
GGG	188085	169097	0	**4**	-	0	-
HHH	187483	168495	0	**4**	-	0	-

PROVINCIA DI RESIDENZA	PROVINCIA RISCHIOSA	CITTADINANZA	ESPOSIZIONE POLITICA	IMPORTO DARE BONIFICI ESTERI	FREQUENZA DARE BONIFICI ESTERI	IMPORTO AVERE BONIFICI ESTERI
Torino	0	-	0	0	0	0
Trieste	0	-	0	1	9	0
Genova	0	-	0	0	0	0
Genova	0	ITALIA	0	0	0	0
Modena	0	-	0	7	9	0
Piacenza	0	-	0	4	9	0
Siena	0	-	0	0	0	0
Pisa	0	-	0	0	0	0

FREQUENZA AVERE BONIFICI ESTERI	IMPORTO DARE BONIFICI ITALIA	FREQUENZA DARE BONIFICI ITALIA	IMPORTO AVERE BONIFICI ITALIA	FREQUENZA AVERE BONIFICI ITALIA	IMPORTO AVERE BO EST FISCALITA AGEVOLATA	FREQUENZA AVERE BO EST FISCALITA AGEVOLATA	IMPORTO PRELIEVI DA CONTO CORRENTE
0	4	9	7	9	0	0	0
0	1	9	0	0	0	0	0
0	2	9	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	6	9	0	0	0	0	0
0	1	9	0	0	0	0	1
0	4	9	1	9	0	0	7

FREQUENZA PRELIEVI DA CONTO CORRENTE	IMPORTO VERSAMENTI SU RAPPORTI CONTINUATIVI	FREQUENZA VERSAMENTI SU RAPPORTI CONTINUATIVI	NUMERO DI DELEGHE PER OPERARE	NUMERO SEGNALAZIONI A UIF	NUMERO DI CONTI	NUMERO CASSETTE SICUREZZA	UTILIZZO CONTANTE
0	0	0	0	0	2	0	0
0	0	0	0	1	5	0	0
0	0	0	0	0	3	0	0
0	0	0	0	0	4	0	0
0	9	9	0	0	5	0	7
0	9	9	0	0	5	0	4
9	9	9	0	0	6	0	4
9	0	0	0	0	7	0	0

Si deduce da tali risultati che il modello creato, in questo caso è stato applicato il classificatore sulla corruzione, riesce a produrre delle previsioni sufficientemente coerenti con i parametri presenti nei vari attributi.

Prima di passare alle conclusioni, risulta interessante, osservare quali sono i valori che il classificatore Bayesiano fornisce attraverso la sua applicazione.
Come già detto il classificatore utilizzato è il Naive Bayes, il quale fa riferimento alla formula di Bayes:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Dove $P(A|B)$ rappresenta la probabilità dell'attributo RISCHIO condizionata ai 59 attributi, che dopo la features selection ne sono rimasti 30.
Mentre $P(A)$ è la probabilità a priori del rischio, quindi in questo contesto il rischio, avendo cinque classi, è il seguente:

```
CLASSE RISCHIO :            1     2     3     4     5
PROBABILITA' A PRIORI :   0.036 0.082 0.474 0.125 0.283
```

Estratto tramite la seguente funzione di R:

```
modello$apriori
```

Mentre per quanto riguarda $P(B|A)$, ossia le probabilità delle variabili esplicative condizionate rispetto ai cinque livelli del rischio, si ottiene ad esempio:

```
   NUMERO.CASSETTE.SICUREZZA
Y             0           1           2           3           4
  1 0.986928105 0.013071895 0.000000000 0.000000000 0.000000000
  2 0.879943503 0.110169492 0.009887006 0.000000000 0.000000000
  3 0.927001953 0.062744141 0.010253906 0.000000000 0.000000000
  4 0.943518519 0.051851852 0.002777778 0.001851852 0.000000000
  5 0.918609407 0.035582822 0.015541922 0.019222904 0.011042945
```

Dove, nelle righe sono riportati i cinque differenti gradi di rischio, mentre sulle colonne si possono osservare i parametri discreti presenti nell'attributo relativo al numero delle cassette di sicurezze. Considerando il primo valore proposto, 0.986928105, questo rappresenta la probabilità che un pattern che presenta zero cassette di sicurezza ha un rischio di corruzione basso per il 98.69% dei casi. Queste tabelle sono disponibili per ogni singolo attributo utilizzato per costruire il classificatore.

Dunque se un record, viene classificato come 4=medio/alto, e nella variabile esplicativa "NUMERO.CASSETTE.SICUREZZA" presenta il valore discreto 1, ciò vuol dire che questa variabile per tale record ha un "peso" pari allo 0.0518 nella contribuzione del calcolo del rischio finale.

Nel progetto intrapreso infatti, oltre a ottenere delle previsioni sul rischio di finanziamento al terrorismo e corruzione, era interessante riuscire a ottenere delle probabilità condizionate al rischio, le quali rappresentano dei valori, oppure "pesi", che contribuisce al calcolo finale del rischio.

È stato possibile implementare un algoritmo in R affinché si estraesse per ogni pattern le varie $P(Attributo|Rischio)$ attraverso l'uso di quest'ultime tabelle e attraverso il grado di rischio predetto dal classificatore.

Purtroppo R è un software eccezionale per eseguire soltanto il suo lavoro, ossia fare previsioni a record di elevata numerosità; infatti riesce a predire la classe di etichetta "RISCHIO" a 800000 record in soli 14 minuti.

Mentre per quanto riguarda l'algoritmo creato per estrarre le $P(Attributo|Rischio)$, questo impiega una quantità di minuti per cui non vale la pena eseguire tale nel software R, poiché un procedimento del genere in un software come SQL, riporta un'esecuzione del processo minore, circa 10 minuti.

Per estrarre le diverse tabelle $P(Attributo|Rischio)$ è stata eseguita tale procedura in R:

```
modello$tables$
```

Nella finestra che si apre affianco al secondo "$" è possibile, scorrendo in basso, selezionare l'attributo del quale si vuole mostrare la probabilità condizionata rispetto alla classe "RISCHIO".

CONCLUSIONI

In questo elaborato di tesi è stato mostrato come sia possibile utilizzare tecniche di apprendimento automatico supervisionato nel contesto della classificazione del rischio a soggetti delle banche.
Sono stati analizzate due tematiche distinte sulle quali calcolare il rischio, il finanziamento al terrorismo e la corruzione, argomenti che tutto sommato concernano in un unico grande tema, il riciclaggio di proventi illeciti.

È stato mostrato in che modalità viene sviluppato un classificatore che permette l'identificazione di pattern che presentano i cinque diversi livelli di rischio, ricavati utilizzando la suddivisione di rischio di inoltro delle segnalazioni sospette proposte dall'Unità di Informazione Finanziaria.
L'insieme di dati utilizzato per fare la classificazione, conteneva non più di 59 colonne, le quali presentavano al loro interno valori discreti, permettendo così all'algoritmo di eseguire in maniera più lineare le complesse operazioni.
Il modello di classificazione scelto è il Naive Bayes, il quale basa la sua teoria sul teorema di Bayes, permettendo di ottenere le probabilità di rischio condizionate agli attributi, avendo come ipotesi fondamentale l'indipendenza delle variabili esplicative.

Nell'implementare l'algoritmo Bayesiano è stata effettuata la suddivisione del set di dati in due partizioni, 70% e 30%, rispettivamente training set e test set. Ad entrambi i set di dati è stata aggiunta la variabile risposta con relativa etichetta per ogni pattern.
Ma prima di proseguire con l'apprendimento della macchina, è stato ritenuto necessario eseguire una riduzione dello spazio delle variabili, utilizzando la struttura "filtro" della Features Selection, diminuendo il numero delle variabili da 59 a 30, ritenendo quelle da eliminare meno importanti in quanto presentavano poca significatività rispetto alla etichetta di classe.
Con il primo set di dati è stato costruito il modello probabilistico, affinché questo imparasse a modellare i dati avendo a disposizione la variabile risposta "RISCHIO" e le 30 variabili esplicative. Al set minore di dati è stato applicato il modello appena costruito, e tramite questo è stato possibile predire il grado di rischio per ciascun record.

Utilizzando le previsioni e i valori reali del rischio del test set, si è costruito una matrice di confusione, tramite la quale si è analizzata la bontà di adattamento del modello, attraverso alcuni indici particolari come l'accuratezza, la kappa di Cohen e la curva di ROC.
Dopo aver ottenuto un modello probabilistico Bayesiano "buono" è stata effettuata la fase di valutazione, ossia l'applicazione di tale modello su un dataset nuovo senza alcuna variabile di rischio, affinché il rischio venga predetto in automatico basandosi su cosa il modello ha imparato durante la fase di training.

Le matrici di confusione analizzate nei due ambiti distinti hanno riportato indici di accuratezza buoni per il terrorismo mentre per la corruzione i valori di bontà erano migliori. Questo fattore è dovuto al fatto che il modello sul terrorismo ha utilizzato come test di training un set da, 3500, il 70% di 5000, mentre la corruzione su 9100 record, il 70% di 13000. Quindi il modello sulla corruzione ha avuto a disposizione più pattern per capire con quali caratteristiche delle features veniva applicato un livello di rischio rispetto all'altro.

Sviluppi futuri di questa tematica sarà sicuramente l'aumento delle features utilizzate, per garantire ai classificatori di lavorare meglio durante le loro fasi di apprendimento, permettendo una classificazione più raffinata dei dati.
Inoltre la vastità dell'argomento dell'apprendimento automatico propone innumerevoli spunti da poter approfondire nei prossimi anni.

BIBLIOGRAFIA

Davigo Piercamillo, *Il sistema della corruzione*, Editori Laterza.

Razzante Ranieri, *CORRUZIONE RICICLAGGIO E MAFIA la prevenzione e la repressione nel nostro ordinamento giuridico*, Aracne, 2015

Peter M. Lee, *Bayesian statistics: an introduction*, Chichester: Wiley, 4th edition, 2012

Alfonso Maria Stile, *Riciclaggio e reimpiego di proventi illeciti*, 2009

Carlo De Stefano, Luciano Piacentini, Italo Saverio Trento, *I nuovi scenari del terrorismo internazionale di matrice jihadista*, Rubettino.

Carlo Vercellis, *Business Intelligence: modelli matematici e sistemi per le decisioni*, McGraw-Hill.

SITOGRAFIA

Badano Federica Rachele, *La corruzione e la legge 190/2012. Priorità all'azione di prevenzione*, Il fisco oggi, rivista telematica, 2015
http://www.fiscooggi.it/analisi-e-commenti/articolo/corruzione-e-legge-1902012priorita-all-azione-prevenzione-1

Transparency International Italia, CPI 2016: l'Italia guadagna una posizione, ma non basta, 2016
https://www.transparency.it/cpi-2016-l-italia-guadagna-una-posizione-ma-non-basta/

Guidoni Caterina, *La leggenda dei 60 miliardi, stima falsa che tutti citano*, Il Sole 24 ore, 2016

http://www.ilsole24ore.com/art/notizie/2016-08-20/la-leggenda-60-miliardi-stima-falsa-che-tutti-citano-185856.shtml?uuid=ADP3Kxn&refresh_ce=1

Frasca Antonello, *La disciplina amministrativistica della prevenzione della corruzione*, 2013
http://tesi.eprints.luiss.it/12376/1/frasca-antonello-tesi-2014.pdf

Nardi Sara, *"Guerra" al terrorismo e diritto internazionale umanitario*, 2015
https://etd.adm.unipi.it/theses/available/etd-11232015-104630/unrestricted/TESI_TERRORISMO_INTERNAZIONALE.pdf

Palumbo Giovanbattista, *Money Transfer, boom del mercato nero: in Italia spopola la rete parallela*, Il Messaggero.it, 2015
http://www.ilmessaggero.it/pay/edicola/money_transfer_mercato_nero-1199304.html

Boezi Francesco, *La guerra santa della droga*, Gli occhi della guerra, 2017
http://www.occhidellaguerra.it/droga-e-jihad-un-legame-sempre-piu-forte/

Wikipedia, *Thomas Bayes*
https://it.wikipedia.org/wiki/Thomas_Bayes

Tortora Stefano, *Tesina di intelligenza artificiale, Il classificatore Naïve Bayes*
http://www.stefanotortora.altervista.org/progetti/IntelligenzaArtificialeII.pdf

Giangreco Manuela, *Approccio Frequentista e Bayesiano, Due Modi Diversi di Vedere La Stessa Realtà: Applicazioni alla Modulazione del Dolore a Misure Ripetute*, 2013
https://air.unimi.it/retrieve/handle/2434/252855/346387/phd_unimi_R09138.pdf

Tono Raffaella, *Natural Language processing e tecniche semantiche per il supporto alla diagnosi: un esperimento*, 2010
http://tesi.cab.unipd.it/25036/1/Tesi_per_pdf.pdf

Lazzarini Nicola, *Tecniche di approfondimento automatico per l'identificazione dell'iperaldosteronismo primario*, 2012
http://tesi.cab.unipd.it/40187/1/Tesi_NicolaLazzarini.pdf

Russo Diego, Vella Fabio, *Bayesian Filters*, 2011
http://www.dmi.unipg.it/bista/didattica/sicurezza-pg/seminari2010-11/bayes/russo_vella.pdf

Unità di informazione finanziaria per l'Italia, *Segnalazioni di operazioni sospette*.
http://uif.bancaditalia.it/adempimenti-operatori/segnalazioni-sos/

Vianelli Silvio, *Induzione e inferenza statistica.*
http://www.trapaninostra.it/libri/Biblioteca_Fardelliana/La_Fardelliana_1984_n_2-3/La_Fardelliana_1984_n_2-3-03.pdf