Università degli Studi di Napoli Federico II

Polo delle Polo delle Scienze Umane e Sociali

Dottorato di ricerca in Statistica

XXVI ciclo: Esame finale nell'Indirizzo "Statistica computazionale"

Candidato

Ing. Filomena Mauriello

Tecniche di ricampionamento per dataset con classi di risposta sbilanciate. Una proposta metodologica per dataset con predittori di natura numerica e categorica.

Coordinatore di dottorato:

Prof. Carlo Lauro

Relatori:

Prof. Massimo Aria

Prof. Marina Marino

Nessuno sa abbastanza, ed abbastanza presto.

E. Pound

Ringraziamenti

Giunta alla fine di questo lavoro penso sia più che giusto spendere due parole anche verso coloro che mi hanno aiutato e sostenuto in questo lungo e faticoso percorso.

Un fortissimo ringraziamento per la costante fiducia e il grande affetto va alla mia famiglia. In particolar modo ai miei genitori senza i quali non sarei mai potuta arrivare sin qui, pronti a soccorrermi nei momenti di difficoltà. A mia madre che è stata la prima a credere in me e a dirmi di inseguire i miei sogni. A mio padre che è sempre stato pronto a correre in mio aiuto.

Un grazie urlato alla mia sorellona Giovanna, anche se fisicamente lontana, mi ha particolarmente aiutato, mi è sempre stata vicina e in tanti momenti mi ha spronato ad andare avanti e non arrendermi; non è cambiato nulla anche se sei lontana!... e per l'avermi regalato una bellissima nipotina, Maria Vittoria.

Grazie a Leopoldo per la positività che mi ha regalato in ogni occasione, per l'aiuto che mi ha offerto nei modi più disparati e per aver cercato sempre di farmi sorridere.

Un grande grazie al professore Massimo Aria, che non solo è stato il mio tutor e un punto fondamentale nella attività di ricerca, ma anche perché è soprattutto un amico.

Un grande grazie anche a alla professoressa Marina Marino, per la disponibilità, i consigli, la comprensione e la cordialità che ha sempre mostrato nei miei confronti durante tutto il periodo.

Tanta stima e riconoscenza vanno ai Professor Roberta Siciliano e Antonio D'Ambrosio, per avermi seguito con passione i miei studi e dedicandomi parte del

loro prezioso tempo per formulare proposte capaci di migliorare il lavoro di ricerca, per il proficuo scambio di idee e conoscenze, e per tutte le iniziative che mi ha suggerito per migliorare scientificamente il lavoro.

Desidero ringraziare, il professor Alfonso Montella, per la disponibilità, i consigli, la comprensione e la cordialità che ha sempre mostrato nei miei confronti durante tutto il periodo di tesi.

Desidero inoltre ringraziare una serie di persone che in circostanze spesso diverse mi hanno dato serenità, amicizia e supporto morale nell'arco di questo variegato percorso.

In primo luogo Lorenza, con cui ho condiviso gioie, dolori, frustrazioni e soddisfazioni non soltanto in questi tre anni di dottorato ma di tutta la mia lunga avventura universitaria.

Roberta la mia amica di mille pazzie, che ha ascoltato ogni mia parola e sopportato ogni mio sfogo…mi ha donato tanta amicizia.

Laura, che ha sempre ascoltato e analizzato i miei frequenti monologhi, sia quelli seri che quelli più deliranti.

Andrea, il mio amico vecchietto, che è sempre pronto a correre ovunque io sia.

Carmela, che mi ha aiutato a comprendere non solo la statistica, ma anche l'amicizia.

Salvatore il mio consulente informatico, con cui ho condiviso tanti pomeriggi.

Ringrazio Francesco, che tutti i giorni mi ha sopportato in ufficio e che pazientemente ha subito i miei nervosismi pre-cosegna, e non solo.

Infine un grazie a tutti gli amici che sono andati e che sono tornati, a quelli che non ci sono e a quelli che arriveranno e alle mie colleghe Lella, Manuela e Elisabetta.

*Grazie per davvero….**Mena***

Sommario

Notazione

S rappresenta il dataset formato da *n* osservazioni. S=[$Y; X$] dove:

- Xè la matrice dei predittori di dimensione $n \times p$;
- x_i la generica osservazione della matrice X tale che $i = 1, \dots, n$
- Y è un vettore delle risposte di dimensione $n \times 1$, di natura categorica attribuite ad ogni osservazione della matrice X dei predittori;
- $y_i = c$ per $c = 0, \dots C$, è la classe di risposta assunta dalla *i*-esima osservazione;

Considerando $C = \{0,1\}$, si definiscono:

- S_{min} è il sottoinsieme di S avente per classi di risposta le osservazioni appartenetti alla classe minoritaria, tale che $S_{min} \subset S$;
- S_{max} è il sottoinsieme di S avente per classi di risposta le osservazioni appartenetti alla classe maggioritaria, tale che $S_{max} \subset S$.

così che:

- $S_{min} \cap S_{max} = \{\Phi\}$ e
- $S_{min} \cup S_{max} = \{S\}$.

Con E si indicano i campioni generati da procedure di campionamento su S, sono definiti:

- E_{min} campione generato per la classe di risposta minoritaria;
- E_{max} campione generato per la classe di risposta maggioritaria.

X_i^{Knn} è la sottomatrice di X formata dalle k osservazioni più vicine a x_i, identificate secondo l'algoritmo K-nearest neighbors. $X_i^{Knn} \subset X$

- $x_{i,k}^{Knn}$ è la generica osservazione della matrice X_i^{Knn} tale che la distanza $d(x_{i,k}^{Knn}, x_i) \leq d(x_i, x_{i*})$ con x_{i*} una generica osservazione della matrice X tale che $x_{i*} \notin X_i^{Knn}$.

Introduzione

Lo studio dei dati con classi di risposta sbilanciate è un argomento di notevole importanza, soprattutto nella medicina, nella finanze, nella sicurezza stradale ed altri campi. In presenza di una distribuzione della variabile di risposta estremamente sbilanciata il processo di apprendimento può essere distorto, perché il modello tende a focalizzarsi sulla classe prevalente e ignorare gli eventi rari, che possono essere pazienti aventi un cancro, incidenti stradali mortali, oppure cattivi creditori. La regressione logistica, per esempio, nota come uno dei metodi parametrici tradizionali più utilizzati per la classificazione binaria, non è consigliabile quando le classi sono sbilanciate, perché la probabilità condizionale della classe rara è sottostimata. Nemmeno i metodi più flessibili non parametrici come gli alberi di classificazione e le regole associative sono immuni alle conseguenze di una distribuzione asimmetrica delle classi. Gli alberi di classificazione, per esempio, sono costruiti trovando divisioni successive tale che sia massimo il decremento di impurità. Questo è tipicamente tradotto in modelli comuni aventi una accuratezza elevata nella classe prevalente e una precisione molto bassa della classificazione dell'evento raro. È importante sottolineare che la classe minoritaria di solito rappresenta il concetto di interesse, ad esempio la diagnosi medica di pazienti con malattie rare, come il cancro.

Diverse sono le soluzioni che sono state proposte nel tempo per affrontare il problema dei dati estremamente squilibrati, e si possono distinguere due approcci comuni, Tecniche di Cost-Sensitive Learning e Tecniche di campionamento. A differenza dei modelli tradizionali di apprendimento, le tecniche Cost-Sensitive utilizzano una funzione di costo di errata classificazione per pesare le diverse classi di risposta e così limitare gli effetti dovuti allo sbilanciamento della distribuzione delle classi stesse. L'obiettivo dell'apprendimento Cost-Sensitive è

minimizzare i costi di errata classificazione pesati sulla base di una funzione di penalità.

Le tecniche di campionamento effettuano un lavoro di pre-processing sui dati, in modo da fornire una distribuzione bilanciata tra le classi. L'uso di metodi di campionamento consiste nella modifica di un set di dati sbilanciati attraverso alcuni meccanismi in modo da fornire una distribuzione equilibrata. Le tecniche più comuni sono il *random oversampling* che attua un campionamento con ripetizione delle osservazioni appartenenti alla classe rara e il *random undersampling* che, al contrario, effettua un campionamento senza ripetizione tra le osservazioni appartenenti alla classe maggioritaria. In altre parole, Il *random oversampling* è un metodo che mira a bilanciare la distribuzione di classe attraverso la replicazione casuale di esempi appratenti alla classe minoritaria. Diversi autori concordano sul fatto che il *random oversampling* può aumentare la probabilità che si verifichino problemi di overfitting. Ciò implica che un classificatore potrebbe pervenire alla definizione di regole apparentemente accurate, ma che in realtà lo sono solo per il dataset replicato e non per la popolazione di riferimento. Inoltre, il *random oversampling* aumenta il costo computazionale del processo di apprendimento accrescendo in maniera massiva la dimensione della matrice dei dati (in merito al numero di osservazioni da trattare).

Il *random undersampling*, estraendo solo una parte delle osservazioni che compongono la distribuzione all'interno della classe maggioritaria, potrebbe portare a risultati insoddisfacenti in quanto parte dell'informazione contenuta nella matrice iniziale verrà scartata riducendo la dimensione del campione.

Nonostante gli svantaggi delle tecniche di ricampionamento, queste sono molto più popolari delle tecniche di Cost-Sensitive Learning. La ragione più ovvia è che esistono implementazioni Cost-Sensitive solo per alcuni algoritmi di

apprendimento e quindi spesso l'unica via per trattare la problematica dei dati con classi sbilanciate è quella di ricorrere a tecniche di ricampionamento.

Infatti le tecniche di ricampionamento agiscono come una fase di pre-elaborazione, consentendo al sistema di apprendimento di ricevere le osservazioni, come se appartenessero a un insieme di dati ben equilibrato.

Nel corso degli anni, molte tecniche sono state sviluppate con l'obiettivo di superare i limiti del *random sampling*. Molti studi hanno focalizzato l'attenzione su metodi di ricampionamento dei dati in modo da avere delle classi non sbilanciate mantenendo al contempo una struttura informativa coerente con il dataset originario, si ricordano in particolare il *Synthetic Minority Over-sampling TEchnique* (SMOTE), *ADAptive SYnthetic sampling* (ADASYN) e *Random OverSampling Examples* (ROSE).

Tutte queste tecniche generano osservazioni "sintetiche" dalla classe di minoranza e le aggiungono al set di dati esistenti. I record artificiali della classe di minoranza sono generati basandosi sulla similarità nello spazio dei predittori. In particolare, la similarità tra le osservazioni è misurata attraverso l'impiego di alcune misure di distanza. Analizzando i risultati ottenuti in tali studi si può osservare come l'impiego di tecniche di generazione di dati artificiali consenta di migliorare sensibilmente le misure di performance dei modelli di classificazione rispetto agli approcci classici di ricampionamento.

Un limite, che accomuna tutte le tecniche appena citate, è rappresentato dalla impossibilità di trattare dataset con classi sbilanciate in cui vi sia la presenza sia di predittori numerici sia predittori qualitativi. Solitamente, per superare questo problema, ci si limita ad ignorare le variabili categoriche nel processo di generazione di dati artificiali.

Nel presente lavoro si propone una nuova metodologia di *synthetic sampling*, chiamato *"Synthetic Over-sampling for Numerical and Categorical variables*

(SONCA)" che possa essere utilizzato con dataset caratterizzati dalla presenza di predittori di natura eterogenea.

L'idea chiave di SONCA consiste nella generazione di osservazioni artificiali attraverso la definizione di una funzione di probabilità inversamente proporzionale alla distanza tra le osservazioni. Distanza misurata dopo una codifica ad hoc della matrice originaria così che possano essere considerate contemporaneamente sia variabili numeriche sia categoriche.

Al fine di valutare l'efficacia di SONCA, l'algoritmo è testato usando differenti dataset, che appartengono alla banca dati dell'UCI Machine Learning Repository. In particolare, si è scelto di valutare le prestazioni dell'algoritmo SONCA rispetto a due aspetti principali: la sensibilità delle performance del metodo rispetto ai diversi parametri che lo caratterizzano; la comparazione delle performance rispetto alle principali proposte metodologiche già presenti in letteratura.

Per valutare le prestazioni dell'algoritmo SONCA, sono state considerate diverse misure di performance. Le consuete misure di accuratezza, come ad esempio il tasso di errata classificazione, possono condurre a risultati fuorvianti perché dipendono fortemente dalla distribuzione di classe. Dallo studio della letteratura, è stato identificato un set di misure di performance che si caratterizzano per un funzionamento che sia indipendente dalla distribuzione della classe di risposta.

Inoltre il processo di valutazione è completato con l'utilizzo di SONCA su 2 dataset reali che riguardano il problema, di grande interesse nella letteratura sulla sicurezza stradale, dello studio delle determinanti degli incidenti stradali mortali.

Il presente lavoro di tesi è stato suddiviso in cinque capitoli:

- Nel primo capitolo è affrontato il problema delle classi di risposta sbilanciate e le diverse soluzioni presenti in letteratura;
- Nel secondo capitolo è presentato l'algoritmo di ricampionamento Synthetic Over-sampling for Numerical and Categorical variables

(SONCA), che può essere utilizzato sia per predittori di natura numerica, sia per dati di natura categorica;

- Nel terzo capitolo sono riportate le diverse misure di performance utilizzate per la valutazione dell'accuratezza dei classificatori;
- Nel quarto capitolo sono utilizzati numerosi dataset per valutare le prestazioni dell'algoritmo SONCA rispetto a due aspetti principali: La sensibilità delle performance del metodo rispetto ai diversi parametri che lo caratterizzano; La comparazione delle performance rispetto alle principali proposte metodologiche già presenti in letteratura;
- Nel quinto capitolo il processo di valutazione è completato con l'utilizzo di SONCA su 2 dataset reali che riguardano il problema dello studio delle determinanti degli incidenti stradali mortali.

Capitolo 1 - Analisi della letteratura

1.1. Il problema delle classi di risposta sbilanciate

Il problema delle classi di risposta sbilanciate è stato riconosciuto in molti campi applicativi, come nelle telecomunicazioni, nel rilevamento di sversamenti di petrolio nelle immagini radar satellitari, nella gestione del rischio, attività di filtraggio, diagnosi medica (ad esempio le malattie rare e rare mutazioni genetiche), l'identificazione di frodi in conti bancari, la text classification, l'intrusione in reti informatiche (Chawla, 2003; Guo et al., 2008), severità degli incidenti stradale (Tesema et al., 2005, Emerson et al., 2011; Nayak et al., 2011, Torrão et al., 2014).

In ciascuno di questi ambiti una classe ha una frequenza estremamente elevata rispetto a un'altra con rapporti dell'ordine del tipo 100:1, 1000:1 e 10000:1 (Chawala et al., 2002; Chawala et al., 2004). In Figura 1 è illustrato un esempio di confronto tra popolazione con classe di risposta bilanciata e popolazione con classe di risposta sbilanciata (Cao et al., 2011).

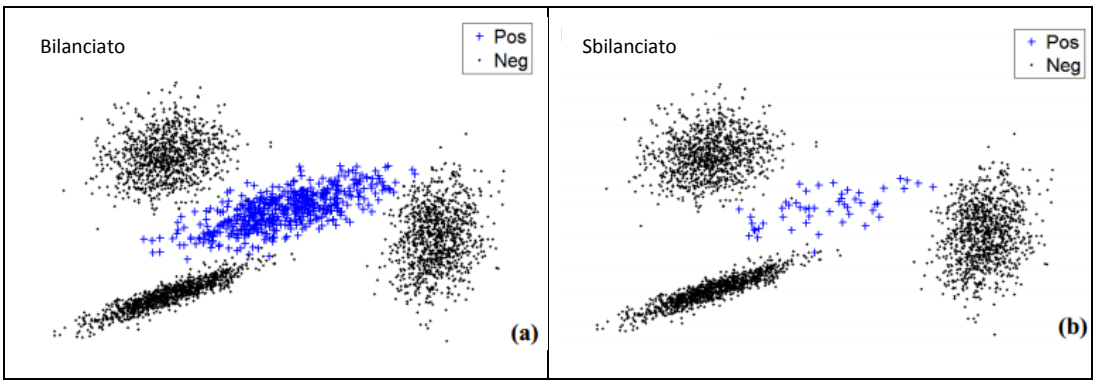

Figura 1 – Esempi di data set con classe di risposta bilanciata in (a) e di banche dati con classe di risposta sbilanciata in (b) in funzione dello spazio 2D (Cao et al., 2011).

In presenza di una distribuzione della variabile di risposta estremamente sbilanciata il processo di apprendimento può portare a risultati distorti (He e

Garcia, 2009; Ndour e Dossou-Gbété, 2012). La maggior parte dei metodi di apprendimento sono concepiti per identificare la regola di classificazione che meglio si adatta ai dati secondo qualche criterio di accuratezza globale, il loro obiettivo è di minimizzare l'errore globale al quale la classe minoritaria contribuisce poco. Inoltre, essi assumono che ci sia un'equa distribuzione dei dati per tutte le classi, e assumono che gli errori provenienti dalle diverse classi, ovvero dalla classe maggioritaria e dalla classe minoritaria, hanno lo stesso peso. Per queste diverse ragioni si hanno scarse prestazioni degli algoritmi di classificazione esistenti in presenza classi di risposta sbilanciate (Ganganwar, 2012). Quando i dati sono sbilanciati, il modello tende a concentrarsi sulla classe prevalente e ignorare gli eventi rari. Di conseguenza, le osservazioni che appartengono alla classe minoritaria sono classificate erroneamente con maggiore frequenza di quelle appartenenti alla classe maggioritaria.

L'inefficacia di tali algoritmi nel predire la classa rara in presenza di dati squilibrati è dimostrata da diversi studi presenti in letteratura. È stato osservato che sia metodi parametrici sia metodi non parametrici sono sensibili alle conseguenze della distribuzione squilibrata delle classi di risposta. La regressione logistica, per esempio, nota come uno dei metodi parametrici tradizionali più utilizzati per la classificazione binaria, non è consigliabile quando le classi sono sbilanciate, perché la probabilità condizionale della classe rara è sottostimata (King 2001; Menardi e Torelli, 2009). Nemmeno i metodi più flessibili, non parametrici, come gli alberi di classificazione e le regole associative sono immuni alle conseguenze di una distribuzione asimmetrica delle classi. Gli alberi di classificazione, per esempio, sono costruiti trovando divisioni successive tale che sia massimo il decremento di impurità. Tuttavia, tale metodologia potrebbe non essere appropriata quando le classi di risposta sono squilibrate. In tal caso, accade che c'è un'alta accuratezza per la classe prevalente e una bassa accuratezza per la classe rara (Menardi e Torelli, 200; Chawla, 2003). Anche le regole associative

sono influenzate dalle distribuzioni della classe di risposta. In presenza di classi di risposte sbilanciate, molte regole non sono selezionate perché tendono ad avere supporto molto basso (Gue et al., 2003, Ndour e Dossou-Gbété, 2012).

È importante rilevare che la classe minoritaria di solito rappresenta il concetto di interesse, ad esempio la diagnosi medica di pazienti con malattie rare, come il cancro (Cao et al., 2011). Se un malato di cancro, che in genere appartiene alla classe rara, è classificato come sano, non farà nulla per curare la sua malattia e ciò porterà all'aggravarsi delle sue condizioni di salute. Data la grande importanza, il problema dei dati con classi di risposta sbilanciate negli anni ha ricevuto l'attenzione dalle comunità di machine learning e di data mining in forma di workshop (Japkowicz, 2000, Chawla et al., 2003, Dietterich et al., 2003, Ferri et al., 2004) e numeri speciali (Chawla et al., 2004), suggerendo diverse soluzioni che saranno trattate in seguito.

1.2. Cosa sono le classi di risposta sbilanciate

Un dataset che presenta una distribuzione non equa tra le classi di risposta può essere definito "imbalanced" o sbilanciato. La maggior parte di tali squilibri sono indicati come intrinseci, cioè, lo squilibrio è una diretta conseguenza della natura del dataset. Tuttavia, ci sono casi in cui la causa dei dati squilibrati non è legata alla variabilità intrinseca. Per esempio, supponiamo che un insieme di dati sia generato da un flusso di dati continuo, bilanciati in un intervallo di tempo specifico, e se durante questo intervallo, la trasmissione ha interruzioni sporadiche in cui i dati non sono trasmessi, è possibile che i dati acquisiti possano essere sbilanciati, nel qual caso, il set di dati sarebbe squilibrato. Gli squilibri di questo tipo sono considerati estrinseci, cioè quando lo squilibrio non è direttamente legato alla natura del database, ma è legato a fattori esterni come il tempo o la raccolta dei dati. Squilibri estrinseci, sono altrettanto interessanti quanto le loro controparti intrinseche, poiché può benissimo accadere che lo spazio dati, da cui si

ottiene un insieme di dati squilibrato estrinseco non può essere sbilanciato a tutti (He e Garcia, 2009).

Oltre a squilibrio intrinseco ed estrinseco, è importante comprendere la differenza tra squilibrio relativo e squilibrio dovuto a rari casi (o "rarità assoluta"). La differenza tra squilibrio relativo e squilibrio dovuto a rari casi è legato alla dimensione del dataset. Per dataset di dimensioni elevate (configurabili come big data, milioni di osservazioni), anche in presenza di elevato squilibrio, la classe minoritaria sarà caratterizzata da una numerosità molto elevata. Se n=100000 osservazioni e il rapporto tra le classi 1:100 la classe minoritaria avrà una numerosità di circa 1000 osservazioni. In tal caso si parla di squilibrio relativo, infatti, pur essendo la classe di risposta sbilanciata, la sua frequenza è tale da consentire l'uso di metodi statistici classici senza nessuna fase di pre-processing. Al contrario si parla di squilibrio assoluto quando la classe rara assume frequenze basse anche quando il dataset è di tipo big data (He e Garcia, 2009).

1.3. Trattamento dei Dataset Sbilanciati

Diverse sono le soluzioni che sono state proposte nel tempo per affrontare il problema dei dati estremamente squilibrati, e si possono distinguere in due approcci comuni (Chen et al., 2004; Guo et al., 2008; He e Garcia, 2009):

1. Tecniche di Cost-Sensitive Learning lavorano a livello di algoritmo e si basano sull'assegnazione di un costo elevato per errata classificazione della classe di minoranza, e cerca di minimizzare il costo complessivo.
2. Tecniche di campionamento eseguono un lavoro di pre-processing sui dati, in modo da fornire una distribuzione bilanciata tra le classi. Sono distinte due metodologie, nel caso in cui si aggiungono record si parla di oversampling, se invece si eliminano dei record per bilanciare la distribuzione tra classi si tratta di undersampling.

1.3.1. Tecniche di Cost-Sensitive Learning

A differenza dei modelli tradizionali di apprendimento, le tecniche Cost-Sensitive utilizzano i costi di errata classificazione per bilanciare la differenza tra le classi di dati (Sun et al., 2007). In generale, i costi di errata classificazione utilizzano delle penalità per le classificazioni errate attraverso una matrice di costo U, dove U(i, j) è il costo di previsione di un'osservazione appartenente alla classe i, quando in realtà appartiene alla classe j. Con questa notazione, U(+,-) è il costo dovuto all'errata classificazione di un'osservazione positiva (classe rara) come negativa, diversamente U(-,+) è il costo dovuto all'errata classificazione di un'osservazione negativa (classe prevalente) come positiva. Nel trattare il problema delle classi di risposta sbilanciate, l'importanza del riconoscimento delle osservazioni positive è superiore a quello delle osservazioni negative. Il costo di errata classificazione di un'osservazione positiva è superiore al costo di errata classificazione di un'osservazione negativa (cioè, U(+,-)>U(-,+)), mentre il costo di una corretta classificazione ha penalità pari a 0 (cioè, U(+,+) = U(-,-) = 0). L'obiettivo dell'apprendimento Cost-Sensitive è minimizzare i costi di errata classificazione (Liu et al. 2006; Liu et al., 2010).

1.3.2. Tecniche di ricampionamento

Le tecniche di ricampionamento hanno ricevuto una notevole attenzione per superare i problemi legati a dataset con variabile di risposta sbilanciate (Chen et al., 2004; Japkowicz, 2000; Chawla et al., 2002; Weiss e Provost, 2003).

In genere, l'uso di metodi di campionamento consiste nella modifica di un set di dati sbilanciati attraverso alcuni meccanismi in modo da fornire una distribuzione equilibrata. Alcuni studi hanno dimostrato che per diversi classificatori, l'utilizzo di queste tecniche fornisce migliori prestazioni complessive rispetto a un insieme di dati squilibrato (He e Garcia, 2009 ; Weiss e Provost , 2003).

Questi risultati giustificano l'uso di metodi di campionamento in presenza di dataset con classi di risposta sbilanciate. Essi sono un facile metodo per il bilanciamento dei dati e possono essere applicati a qualsiasi sistema di apprendimento, poiché agiscono come una fase di pre-elaborazione, consentendo al sistema di apprendimento di ricevere le osservazioni, come se appartenessero a un insieme di dati ben equilibrato. Qualsiasi errore del sistema verso la classe di maggioranza causato dalla diversa percentuale delle classe, dovrebbe essere eliminato (Ganganwar, 2013).

Diverse metodi di ricampionamento sono stati proposti, i più utilizzati sono l'oversampling e l'undersampling.

Definito un dataset S con *n* osservazioni ($|S| = n$), dove $x_i \in X$ è un'osservazione nello spazio n-dimensionale, e $y_i \in Y$ è la classe di risposta associata con x_i. Considerando $C = 2$, si definiscono:

- $S_{min} \subset S$ il sottoinsieme di S avente per classi di risposta le osservazioni appartenetti alla classe minoritaria;
- $S_{max} \subset S$ il sottoinsieme di S avente per classi di risposta le osservazioni appartenetti alla classe maggioritaria.

così che $S_{min} \cap S_{max} = \{\Phi\}$ e $S_{min} \cup S_{maj} = \{S\}$.

Infine, i campioni generati da procedure di campionamento su S sono indicati con E, con sottoinsiemi disgiunti E_{min} e E_{max} che rappresentano, rispettivamente, le classi minoritarie e maggioritarie della classe di risposta (He e Garcia, 2009).

L'oversampling è un metodo non euristico che mira a bilanciare la distribuzione di classe attraverso la replicazione casuale di esempi di classe minoritari. I meccanismi di *random oversampling* prevedono l'aggiunta di un insieme E ottenuto dalle osservazioni appartenenti alla classe di minoranza. In questo modo, il numero di osservazioni totali nel S_{min}è aumentato di $|E|$ e la distribuzione della

classe di risposta di S viene equilibrata. Questo fornisce un meccanismo per variare il grado di equilibrio di una distribuzione di classi di risposta sbilanciate a qualsiasi livello desiderato (He e Garcia, 2009).

Il *random undersampling* è un metodo non euristico che mira a bilanciare la distribuzione di classe attraverso l'eliminazione casuale di osservazioni appartenenti alla classe di maggioranza. In particolare, sono scelti una serie di osservazioni appartenenti alla di classe di maggioranza (S_{max}) e rimossi in modo tale da avere $|S| = |S_{min}| + |S_{max}| - |E|$ (He e Garcia, 2009). Poiché molti esempi di classe maggioranza sono eliminati, il training set diventa più equilibrato e il processo di apprendimento diventa più veloce (Liu et al., 2009).

A prima vista, i metodi di *random oversampling* e di *random undersampling* sembrano essere funzionalmente equivalenti poiché entrambi alterano la dimensione dei dati originali e possono effettivamente fornire la stessa proporzione di equilibrio. Tuttavia, questa comunanza è solo superficiale, ogni metodo presenta una serie problematiche che possono potenzialmente ostacolare l'apprendimento (He e Garcia, 2009). Diversi autori concordano che il *random oversampling* può aumentare la probabilità che si verifichi un problema di overfitting, dal momento che sono replicate copie esatte delle osservazioni appartenenti alla classe di minoritaria. In questo modo, un qualsiasi processo di apprendimento, potrebbe costruire regole che sono apparentemente accurate, ma in realtà copre un esempio replicato. Inoltre, il *random oversampling* può introdurre un costo computazionale aggiuntivo se il set di dati è già abbastanza grande ma squilibrato (Kotsiantis et al., 2006). Il principale inconveniente del *random undersampling* è che questo metodo può scartare dati potenzialmente utili che potrebbero essere importanti per il processo di apprendimento (Kotsiantis et al., 2006).

Nonostante gli svantaggi delle tecniche di ricampionamento, queste sono molto più popolari delle tecniche di Cost-Sensitive Learning. La ragione più ovvia è che non ci sono implementazioni per tutti gli algoritmi di apprendimento. Diversamente, le tecniche di ricampionamento sono un facile strumento per il bilanciamento dei dati e possono essere applicate a qualsiasi sistema di apprendimento, in quanto agiscono come una fase di pre-elaborazione, consentendo al sistema di apprendimento di ricevere le osservazioni, come se appartenessero a un insieme di dati ben equilibrato (Ganganwar, 2013).

1.4. Database artificiali (o sintetici)

Nel corso degli anni, molte tecniche sono state sviluppate con l'obiettivo di superare i limiti del random sampling. La creazione di database sintetici risolve questi limiti cercando di generalizzare la regione di decisione della classe di minoranza.

In Tabella 1 Tabella 2 sono riportati alcuni studi presenti in letteratura riguardante le tecniche per la creazione di database sintetici.

In particolare dall'analisi della letteratura, si può osservare che le tecniche di oversampling più famose e più utilizzate sono: Synthetic Minority Over-sampling Technique (SMOTE), ADAptive SYnthetic sampling (ADASYN) e Random OverSampling Examples (ROSE).

Tabella 1: Metodi di ricampionamento con over-sampilng (Kell algoritm, 2013)

Full Name	Short Name	Reference
Synthetic Minority Over-sampling TEchnique	SMOTE-I	N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling TEchnique. Journal of Artificial Intelligence Research 16 (2002) 321-357.
Synthetic Minority Over-sampling TEchnique + Edited Nearest Neighbor	SMOTE_ENN-I	G.E.A.P.A. Batista, R.C. Prati, M.C. Monard. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations 6:1 (2004) 20-29.
Synthetic Minority Over-sampling TEchnique + Tomek's modification of Condensed Nearest Neighbor	SMOTE_TL-I	G.E.A.P.A. Batista, R.C. Prati, M.C. Monard. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations 6:1 (2004) 20-29.
ADAptive SYNthetic Sampling	ADASYN-I	H. He, Y. Bai, E.A. Garcia, S. Li. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. 2008 International Joint Conference on Neural Networks (IJCNN08). Hong Kong (Hong Kong Special Administrative Region of the Peo, 2008) 1322-1328.
Borderline-Synthetic Minority Over-sampling TEchnique	Borderline SMOTE-I	H. Han, W.Y. Wang, B.H. Mao. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. 2005 International Conference on Intelligent Computing (ICIC05). LNCS 3644, Springer 2005, Hefei (China, 2005) 878-887.
Safe Level Synthetic Minority Over-sampling TEchnique	Safe Level SMOTE-I	C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap. Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD09). LNCS 5476, Springer 2009, Bangkok (Thailand, 2009) 475-482.
Random over-sampling	ROS-I	G.E.A.P.A. Batista, R.C. Prati, M.C. Monard. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations 6:1 (2004) 20-29.
Adjusting the Direction Of the synthetic Minority clasS examples	ADOMS-I	S. Tang, S. Chen. The Generation Mechanism of Synthetic Minority Class Examples. 5th Int. Conference on Information Technology and Applications in Biomedicine (ITAB 2008). Shenzhen (China, 2008) 444-447.
Selective Preprocessing of Imbalanced Data	SPIDER-I	J. Stefanowski, S. Wilk. Selective pre-processing of imbalanced data for improving classification performance. 10th International Conference in Data Warehousing and Knowledge Discovery (DaWaK2008). LNCS 5182, Springer 2008, Turin (Italy, 2008) 283-292.
Aglomerative Hierarchical Clustering	AHC-I	G. Cohen, M. Hilario, H. Sax, S. Hugonnet, A. Geissbuhler. Learning from imbalanced data in surveillance of nosocomial infection. Artificial Intelligence in Medicine 37 (2006) 7-18.
Selective Preprocessing of Imbalanced Data 2	SPIDER2-I	K. Napierala, J. Stefanowski, S. Wilk. Learning from Imbalanced Data in Presence of Noisy and Borderline Examples. 7th International Conference on Rough Sets and Current Trends in Computing (RSCTC2010). Warsaw (Poland, 2010) 158-167.
Hybrid Preprocessing using SMOTE and Rough Sets Theory	SMOTE RSB-I	E. Ramentol, Y. Caballero, R. Bello, F. Herrera. SMOTE-RSB*: A Hybrid Preprocessing Approach based on Oversampling and Undersampling for High Imbalanced Data-Sets using SMOTE and Rough Sets Theory. Knowledge and Information Systems (2011) In press.

Tabella 2: Metodi di ricampionamento con under-sampilng (Kell algoritm, 2013)

Full Name	Short Name	Reference
Tomek's modification of Condensed Nearest Neighbor	TL-I	I. Tomek. Two modifications of CNN. IEEE Transactions on Systems, Man and Cybernetics 6 (1976) 769-772.
Condensed Nearest Neighbor	CNN-I	P.E. Hart. The Condensed Nearest Neighbour Rule. IEEE Transactions on Information Theory 14:5 (1968) 515-516.
Random under-sampling	RUS-I	G.E.A.P.A. Batista, R.C. Prati, M.C. Monard. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations 6:1 (2004) 20-29.
One Sided Selection	OSS-I	M. Kubat, S. Matwin. Addressing the curse of imbalanced training sets: one-sided selection. 14th International Conference on Machine Learning (ICML97). Tennessee (USA, 1997) 179-186.
Condensed Nearest Neighbor + Tomek's modification of Condensed Nearest Neighbor	CNNTL-I	G.E.A.P.A. Batista, R.C. Prati, M.C. Monard. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations 6:1 (2004) 20-29.
Neighborhood Cleaning Rule	NCL-I	J. Laurikkala. Improving Identification of Difficult Small Classes by Balancing Class Distribution . 8th Conference on AI in Medicine in Europe (AIME01). LNCS 2001, Springer 2001, Cascais (Portugal, 2001) 63-66.
Undersampling Based on Clustering	SBC-I	S. Yen, Y. Lee. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. International Conference on Intelligent Computing (ICIC06). Kunming (China, 2006) 731-740.
Class Purity Maximization	CPM-I	K. Yoon, S. Kwek. An unsupervised learning approach to resolving the data imbalanced issue in supervised learning problems in functional genomics. 5th International Conference on Hybrid Intelligent Systems (HIS05). Rio de Janeiro (Brazil, 2005) 303-308.

1.4.1. Synthetic Minority Over-sampling Technique (SMOTE)

Synthetic Minority Over-sampling Technique (SMOTE) è uno dei metodi di oversampling più popolare. Questo approccio si ispira a una tecnica che ha avuto successo nel riconoscimento dei caratteri scritti a mano. SMOTE genera osservazioni "sintetiche" a partire dalla classe di minoranza e li aggiunge al set di dati esistenti. I record artificiali della classe di minoranza vengono generati basandosi sulla similarità nello spazio dei predittori. Per ciascun record x_i appartenente alla classe di minoranza vengono creati k osservazioni e solo quelle più vicine sono prese (Chawala et al., 2002).

Per ciascuna osservazione x_i appartenente alla classe di minoritaria vengono presi in considerazione i K-nearest neighbors. I K-nearest neighbors sono definiti come le k osservazioni delle classi minoritaria la cui distanza euclidea con x_i presenta il

minimo valore lungo le n dimensioni dello spazio dei predittori. Una nuova osservazione artificiatale si otterrà attraverso la seguente formula:

$$x_j^{SMOTE} = x_i + (\tilde{x}_{i,k}^{Knn} - x_i)\delta_j$$

dove $\delta_j \in [0, 1]$ rappresenta un numero casuale.

In altre parole, per creare un'osservazione artificiale si sceglie in maniera casuale uno dei K vicini, dopo di che si perturba l'osservazione x_i con una porzione casuale della differenza tra la stessa osservazione e un'estratta a caso tra i vicini più vicini (Figura 2).

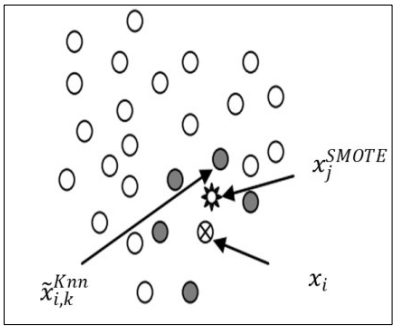

Figura 2 – Esempio di una sintetica osservazione (Chi, 2010)

Operando in questo modo l'osservazione creata si trova sulla linea che collega, in uno spazio euclideo, x_i con $\tilde{x}_{i,k}^{Knn}$ nello spazio dei predittori.

In Figura 3 è illustrato il funzionamento dell'algoritmo.

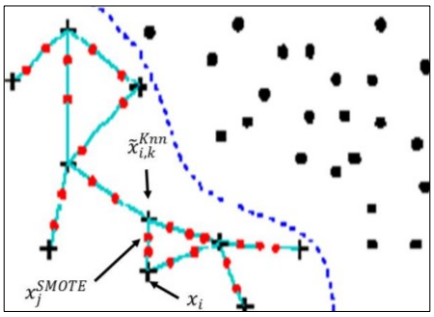

Figura 3 - Applicazione SMOTE (Bakka e Ali touché, 2013)

Le osservazioni sintetiche vengono generate seguendo le fasi di seguito elencate:

1) Si trovano i K vicini per ciascuna osservazione appartenente alla classe di minoranza;
2) Si selezionano in maniera casuale il vicino $\tilde{x}_{i,k}^{Knn}$
3) Si calcola la differenza tra le osservazioni x_i e $\tilde{x}_{i,k}^{Knn}$: $diff = \tilde{x}_{i,k}^{Knn} - x_i$;
4) Si genera un numero casuale δ (compreso tra 0 e 1)
5) Si crea l'osservazione sintetica: $x_j^{SMOTE} = x_i + diff\delta_j$

Inoltre nella classe di maggioranza sono eliminate alcune osservazioni in maniera casuale, fino a che la percentuale della popolazione della classe di minoranza non è pari a una determinata percentuale della classe di maggioranza (Chawla et al., 2002).

I risultati ottenuti utilizzando tale algoritmo hanno evidenziato che le prestazioni degli stimatori sono migliorate. L'analisi di tali risultati ha, inoltre, evidenziato le prestazioni degli stimatori migliorano all'aumentare delle percentuali di ricampionamento fino ad arrivare a un bilanciamento del 200%. Da lì in avanti non sono stati notati ulteriori miglioramenti, piuttosto con percentuali elevate, ad esempio del 500%, le prestazioni degli stimatori sono addirittura peggiori rispetto a quelle che si ottengono lavorando su data set originali.

L'algoritmo di SMOTE non permette di utilizzare dati con predittori categorici, così Chawala nel 2003 propose l'utilizzo della Value Distance Metric (VDM) per calcolare i vicini più prossimi per dataset aventi sia predittori nominali che continui. L'algoritmo con tale variazione della distanza metrica è stato definito SMOTE-NC. I risultati ottenuti hanno evidenziato, che l'utilizzo dell'algoritmo di SMOTE-NC, per tale dataset, non porta nessun miglioramento nell'implementazione del processo di stima.

1.4.2. SMOTE- borderline

Han et al. (2005) hanno presentato due metodi di campionamento, borderline-SMOTE1 e borderline-SMOTE2. Questi generano nuove osservazioni appartenenti, rispettivamente, alla classe di minoranza e maggioranza utilizzando solo i record vicino alla decisione di confine

Al fine di ottenere una migliore previsione, la maggior parte degli algoritmi di classificazione tentano di apprendere dalle regione di confine di ogni classe. Le osservazioni di confine e quelle vicine sono osservazioni borderline e, essendo più suscettibili a essere erroneamente classificati rispetto a quelle più lontane dal confine, hanno un maggior peso nel processo di classificazione. Su tali osservazioni si basano i due nuovi metodi di oversampling, in cui solo le osservazioni nella regione di confine della classe di minoranza sono sovra-campionate.

Tali metodi sono basati sull'Algoritmo di SMOTE. In primo luogo, sono individuate le osservazioni appartenenti alla classe minoritaria lungo la linea di confine, successivamente vengono individuate le k osservazioni più vicine a esse. Dopo di che, le nuove osservazioni sintetiche vengono generate lungo la linea tra le osservazioni appartenenti alla classe minoritaria e i suoi vicini più prossimi selezionati.

In Figura 4 è illustrato un esempio della procedura di Borderline – SMOTE.

In primo luogo, per ogni osservazione appartenente alla classe minoritaria, $x_i \epsilon\, S_{min}$, bisogna determinare l'insieme di vicini più vicini, $S_{i:min-NN}$, con m il numero di vicini più vicini a x_i.

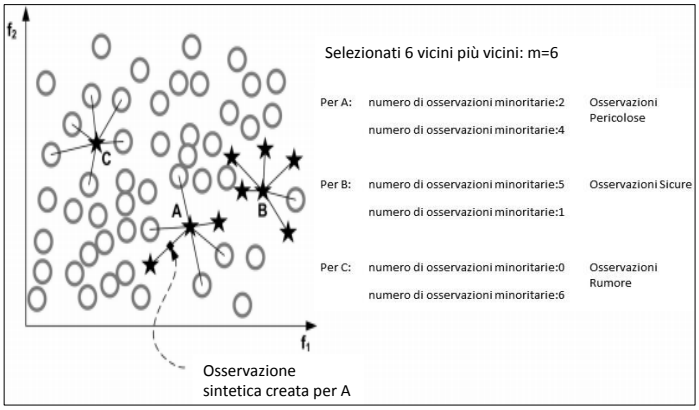

Figura 4 – Creazione di nuovi dati da osservazioni borderline (He et al., 2009).

Successivamente, per ogni x_i, bisogna individuare il numero di vicini più vicini che appartiene alla classe di maggioranza, vale a dire $|S_{i:min-NN} \cap S_{max}|$. Infine, sono selezionati tutti gli x_i che soddisfano la seguente condizione $\frac{m}{2} < |S_{i:min-NN} \cap S_{max}| < m$. Tale condizione suggerisce che solo le osservazioni x_i che hanno più vicini appartenenti alla classe di maggioranza che di minoranza sono selezionati per formare il gruppo Danger. Le osservazioni appartenenti al gruppo Danger, sono le osservazioni appartenenti alla classe di minoranza lungo la linea di confine che hanno più probabilità di essere erroneamente classificati. Le osservazioni x_i appartenenti al gruppo Danger sono poi utilizzate per generare nuove osservazioni sintetiche attraverso l'algoritmo SMOTE. Si deve notare che, se $|S_{i:min-NN} \cap S_{max}| = m$, cioè se tutti i vicini più prossimi di x_i sono esempi di maggioranza, come ad esempio l'istanza di C in Figura 4, allora questa x_i è considerata come rumore e nessuna osservazione sintetica è generata da esso (He et al., 2009).

In Figura 5 è illustrato un esempio di applicazione con SMOTE Borderline: (a) è la rappresentazione del set di dati originali, i punti neri sono le osservazioni della classe di maggioranza, mentre il rosso rappresenta la classe di minoranza; (b) sono individuate le osservazioni che si trovano al confine con la classe di maggioranza

e sono circondate in blu; (c) le osservazioni cerchiate in blu sono le nuove osservazioni generate con l'algoritmo di SMOTE Borderline.

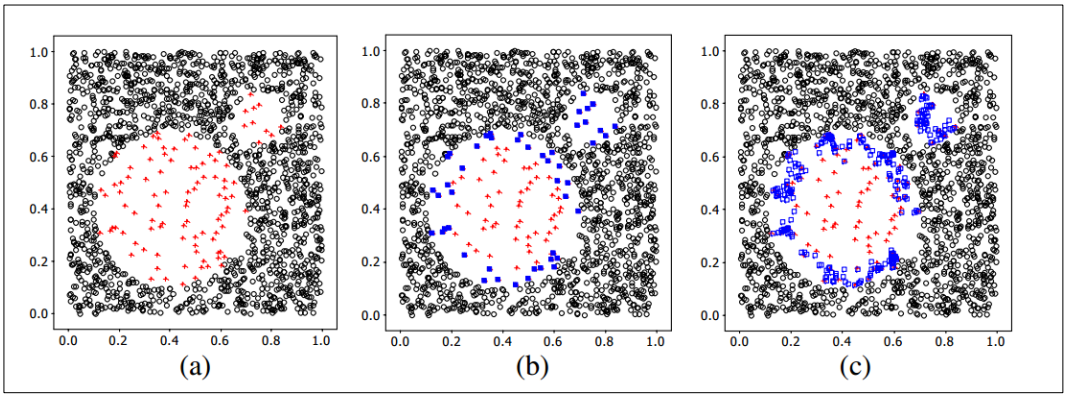

Figura 5 - SMOTE Borderline applicazione (Han et al., 2005)

L'algoritmo di SMOTE Borderline produce risultati migliori rispetto al SMOTE originale, dal momento che le osservazioni situate ai confini sono quelle che più probabilmente possono essere erroneamente classificate.

1.4.3. ADAptive SYNthetic Sampling Approach for Imbalanced (ADASYN)

He et al. (2009) proposero un miglioramento dell'algoritmo di SMOTE. L'idea di ADASYN è di decidere in maniera automatica il numero di osservazioni sintetiche da generare per ogni osservazione appartenente alla classe rara utilizzando una funzione di distribuzione di densità. Il numero di osservazioni sintetiche, generate per ogni osservazione appartenete alla classe minoritaria, è determinato dalla percentuale di osservazioni appartenenti alla classe maggioritaria nelle sue vicinanze.

Prima, calcola il numero di osservazioni sintetiche che devono essere generate per l'intera classe di minoranza:

$$G = (|S_{max}| - |S_{min}|) \times \beta$$

dove $\beta\epsilon[0,1]$ ed è un parametro utilizzato per specificare il livello di bilanciamento desiderato dopo il processo di generazione dei dati sintetici. Successivamente, per ciascun'osservazione $x_i\epsilon S_{min}$, bisogna trovare i vicini K-nearest secondo la distanza euclidea e calcolare il rapporto Γ_{xi} definito come (Ding, 2011):

$$\Gamma_{x_i} = \frac{{\Delta_i}/{K}}{Z} \; con \; i = 1, \dots \; S_{min}$$

Dove Δ_i è il numero di osservazioni vicine a x_i appartenenti alla classe maggioritaria, e Z è il termine normalizzazione in modo che Γ_{x_i} sia funzione di distribuzione ($\sum \Gamma_{x_i} = 1$) .

Il numero di osservazioni da generare per ogni x_i è :

$$N'_{x_i} = \Gamma_{x_i} \times (N^- - N^+) \times \beta$$

dove N^- è il numero di osservazioni della classe maggioritaria e N^+ è il numero di osservazioni della classe minoritaria. Ovviamente, ogni x_i non avrà uguale numero di osservazioni artificiali, in quanto dipende dalla loro distribuzione. Più alto è il numero di osservazioni appartenenti alla classe maggioritaria più saranno le osservazioni sintetiche che dovranno essere generate.

I risultati delle diverse analisi effettuate dagli autori, hanno evidenziato che l'algoritmo ADASYN raggiunge risultati competitivi, fornisce una maggior accuratezza sia per la classe di minoranza sia per la classe di maggioranza, e non sacrifica una classe per preferirne un'altra (He et al., 2009).

1.4.4. Random Over Sampling Examples (ROSE)

Menardi e Torelli (2012) proposero un nuovo algoritmo di ricampionamento, Random Over Sampling Examples (ROSE). Lo scopo di ROSE è di generare nuove osservazioni sintetiche attraverso la stima di una funzione di densità kernel.

Le osservazioni sintetiche vengono generate seguendo le fasi di seguito elencate:

1. Si estrae in maniera casuale una delle due classi, assegnando a ognuna delle stesse probabilità ½ .
2. Si estrae un'osservazione x_i dal sottoinsieme della matrice X composta dalle unità che hanno come risposta la classe estratta, con probabilità pari a $p = 1/n_c$, dove n_c è il numero di osservazioni appartenente alla classe di risposta.
3. Si stima una distribuzione dell'intorno di x_i attraverso una funzione di smooting kernel K_{H_i}:

   $\hat{f}(x|y = Y_j) = \sum_i^{n_j} p_i Pr(x|x_i) = \sum_i^{n_j} \frac{1}{n_j} Pr(x|x_i) = \sum_i^{n_j} \frac{1}{n_j} K_{H_j}(x|x_i)$
4. La nuova osservazione sintetica, x_j^{SMOTE}, si otterrà da un'estrazione casuale della stessa da una funzione di probabilità proporzionale a K_{H_i}.

I risultati delle diverse analisi effettuate dagli autori, hanno evidenziato che utilizzando l'algoritmo ROSE si hanno modelli più accurati con buone misure di performance.

1.5. Sintesi dell'analisi della letteratura

Dall'analisi della letteratura si è osservato che in presenza di una distribuzione della variabile di risposta estremamente sbilanciata il processo di apprendimento può essere distorto. Quando i dati sono sbilanciati, il modello tende a concentrarsi sulla classe prevalente e ignorare gli eventi rari. Di conseguenza, le osservazioni,

che appartengono alla classe minoritaria, sono classificate erroneamente più spesso di quelle appartenenti alla classe di maggioranza. È importante sottolineare che la classe minoritaria di solito rappresenta il concetto di interesse, ad esempio la diagnosi medica di pazienti con malattie rare, come il cancro. Se un malato di cancro, che in genere appartiene alla classe rara, è classificato come sano, non farà nulla per curare la sua malattia e ciò porterà all'aggravarsi delle sue condizioni di salute.

In letteratura, numerose soluzioni sono state presentate per affrontare tale problematica. In particolare si possono distinguere due approcci differenti, Tecniche di Cost-Sensitive Learning e Tecniche di sampling. Le Tecniche di Cost-Sensitive Learning utilizzano i costi di errata classificazione per bilanciare la distribuzione asimmetrica delle classi, mentre le Tecniche di sampling consistono nella modifica di un set di dati sbilanciati attraverso alcuni meccanismi in modo da fornire una distribuzione equilibrata. Questi ultimi rappresentano la soluzione maggiormente adottata nelle applicazioni su problemi reali. Ciò è dovuto a diverse motivazioni:

- In letteratura non esistono proposte metodologiche di tipo cost-sensistive per molti degli algoritmi di classificazione quotidianamente utilizzati.
- Le tecniche di ricampionamento rappresentano uno strumento di facile applicazione nei settori più svariati (medicina, ingegneria, economia, ecc.) in quanto intervengono in una fase di pre-elaborazione dei dati senza necessità di intervenire sul sistema di apprendimento.

Nel corso degli anni, molte tecniche sono state sviluppate con l'obiettivo di superare i limiti del random sampling. La creazione di database sintetici risolve questi limiti cercando di generalizzare la regione di decisione della classe di minoranza.

Dall’analisi della letteratura è stato osservato che le tecniche di ricampionamento più utilizzate sono: Synthetic Minority Over-sampling Technique (SMOTE), ADAptive SYNthetic sampling (ADASYN) e Random OverSampling Examples (ROSE).

La maggior parte di queste tecniche generano osservazioni "sintetiche" dalla classe di minoranza e le aggiungono al set di dati esistenti. I record artificiali della classe di minoranza sono generati basandosi sulla similarità nello spazio dei predittori. In particolare, la similarità tra le osservazioni è misurata attraverso delle misure di distanza come la distanza euclidea.

Analizzando i risultati ottenuti in tali studi si può osservare che in presenza di dataset con variabile di risposta sbilanciate si ha una scarsa accuratezza nei modelli di classificazione, mentre utilizzando algoritmi di ricampionamento si osservano dei valori migliori nelle misure di performance.

Tali approcci proposti non consentono l'elaborazione di insiemi di dati in cui sono contemporaneamente presenti predittori di natura numerica e categorica, e le proposte sino ad ora presenti in letteratura per ovviare a questo limite non sono apparse soddisfacenti.

Capitolo 2 – SONCA: Synthetic Over-sampling for Numerical and Categorical variables

Dalla analisi della letteratura si è osservato che in presenza di una distribuzione della variabile di risposta estremamente sbilanciata il processo di apprendimento può essere distorto.

Nel corso degli anni, molte tecniche sono state sviluppate con l'obiettivo di affrontare tale problematica. Le tecniche di ricampionamento hanno ricevuto una notevole attenzione per contrastare l'effetto delle banche dati con variabile risposta sbilanciate. In particolare, molti studi si sono concentrati sulla creazione di database sintetici.

Le tecniche di ricampionamento sono un facile strumento per il bilanciamento dei dati e possono essere applicate a qualsiasi sistema di apprendimento, agiscono come una fase di pre-elaborazione, consentendo al sistema di ricevere le osservazioni, come se appartenessero a un insieme di dati ben equilibrato.

Come illustrato nel capitolo precedente, le tecniche di ricampionamento più utilizzate sono: Synthetic Minority Over-sampling Technique (SMOTE), ADaptive SYNThetic (ADASYN) sampling e Random OverSampling Examples (ROSE). Nessuno degli approcci proposti consente l'elaborazione di insiemi di dati in cui sono contemporaneamente presenti predittori di natura numerica e categorica.

Nel presente lavoro si propone una nuova metodologia *"Synthetic Over-sampling for Numerical and Categorical variables (SONCA)"* che può essere utilizzato per

dataset caratterizzato sia per predittori di natura numerica, sia per dati di natura categorica.

2.1. Synthetic Over-sampling for Numerical and Categorical variables (SONCA)

Si consideri un dataset costituito dal vettore risposta di dimensione n, $Y_{n\times1}$, e dalla matrice dei predittori di dimensione $n \times q$, $X_{n\times q}$, dove $Y_{n\times1}$ si manifesta con k=C+1 classi, mentre X è costituita da due sotto matrici X_{Num}, matrice dei p predittori numeri, e X_{Cat}, matrice dei q-p predittori categorici.

Si supponga che una delle classi di Y sia sottorappresentata, attraverso SONCA è possibile ottenere un nuovo dataset sintetico affinché la variabile di risposta sia bilanciata per ogni classe.

Ogni osservazione, estratta casualmente dal set di dati originale, è sostituita da un'altra osservazione selezionata all'interno dell'intera matrice dei predittori. La scelta casuale è effettuata ipotizzando una funzione di probabilità inversamente proporzionale alla distanza, tra l'osservazione estratta casualmente dal set di dati originale e tutte le osservazioni della matrice dei predittori. In questo modo le osservazioni con minore distanza avranno una maggiore probabilità di essere scelte.

Le osservazioni sintetiche sono generate seguendo le fasi di seguito elencate:

1. Si estrae in maniera casuale la classe di risposta assunta dalla *i*-esima osservazione $y_i = c$ per $c = 0, \dots C$, attribuendo a ogni modalità una probabilità uniforme $1/(C+1)$, ricordando che $(C+1)$ è numero di classi della variabile di risposta.
2. Si estrae un'osservazione x_i dal sottoinsieme della matrice dei predittori composta dalle unità che hanno come risposta la classe estratta, $X|(Y =$

c), con probabilità pari a $p = 1/n_c$, dove n_c è il numero di osservazioni appartenente alla classe di risposta.

3. Si calcolano le distanze $d(x_i; x_{i*})$ tra x_i e tutte le osservazioni $x_{i*} \in X_{-i}$, con $i* = 1,2,\ldots,i-1,i+1,\ldots n$.
4. Si stima una distribuzione di probabilità di x_i attraverso una funzione monotona decrescente delle distanze calcolate al punto precedente $P(x_{i*}|x_i) \propto f(d(x_i; x_{i*}))$;
5. Si estrae casualmente dalla distribuzione di probabilità stimata la nuova osservazione sintetica, x_j^{SONCA};
6. Si ripetono i passi 1-5, da 1 a [(C+1)×m] volte, dove m è il numero di osservazioni sintetiche da ottenere per ogni classe della variabile di risposta.

2.1.1. Codifica della sotto-matrice dei predittori categorici e scelta della distanza metrica e della funzione di probabilità

La matrice dei predittori, $X_{n\times q}$ può essere costituita da due sotto matrici X_{Num}, matrice dei p predittori numeri, e X_{Cat}, matrice dei q-p predittori categorici.

Affinché si possa calcolare la distanza $d(x_i; x_{i*})$ e quindi usare l'algoritmo SONCA è necessario pretrattare la matrice dei q-p predittori categorici, X_{Cat}, attraverso una codifica disgiuntiva completa, che consiste nello scomporre il carattere in tante variabili dicotomiche che misurano la presenza/assenza di ognuna delle modalità che compongono la variabile originaria. In altre parole la codifica disgiuntiva consiste nel creare, per ogni variabile, tante colonne quante sono le proprie modalità, dove le colonne rappresentano le indicatrici di ogni modalità. In tal modo la sotto matrice X_{Cat}, con q-p predittori categorici dove ogni variabile categorica si esplica rispettivamente in n_{p+1}, $n_{p+2},\ldots, n_{q-p}$ modalità,

attraverso la codifica disgiuntiva completa si trasforma in una sotto matrice X_{Binary} con n_{binary}variabili dicotomiche dove:

$$n_{binary} = \sum_{j=p+1}^{q-p} (n_{p+j})$$

Dopo aver effettuato una codifica disgiuntiva completa della matrice dei predittori categorici, è necessario calcolare le distanze $d(x_i; x_{i*})$ tra x_i e tutte le osservazioni $x_{i*} \in X_{-i}$, con $i* = 1,2,..,i-1,i+1,...n$. In letteratura sono presenti differenti misure come la distanza Euclidea, la matrice di Minkowski, la distanza di Manhattan, la distanza di Mahalanobis e altre (Esposito et al., 2002), ma tali misure sono inappropriate in presenza di variabili misti. La letteratura propone diverse soluzioni, come normalizzare i dati oppure usare distanze metriche normalizzate. Suarez-Alvarez et al. (2012) confrontarono le diverse soluzioni per diversi set di dati e videro che l'accuratezza degli algoritmi è aumentata utilizzando metriche normalizzate.

Quando ci troviamo in presenza di dataset con predittori di natura sia numerica sia categorica anche la distanza euclidea normalizzata è inefficace, in tal caso deve essere utilizzata la distanza euclidea ponderata (Greenacre, 2008; Schultz, e Joachims, T., 2003):

$$d(x_i; x_{i*}) = \sqrt{\left(\sum_j w_j (x_{i,j} - x_{i*,j})^2 \right)}$$

dove w_j è il coefficiente di ponderazione, ovvero indica il peso attribuito alla variabile j. Il coefficiente w_j è pari $1/\sigma_j^2$ se la $j-esima$ variabile è di tipo numerica, se invece la $j-esima$ variabile è di tipo categorica allora w_j è pari $1/c_j$ dove $c_j = \frac{n_{p+j}}{n_{binary}}$.

Infine, l'ultimo punto da definire nell'algoritmo è la scelta della funzione di probabilità monotona decrescente $P(x_{i*}|x_i) \propto f(d(x_i; x_{i*}))$. In letteratura esistono diverse funzioni di probabilità, utilizzate anche per le funzioni kernel come (Figura 6): Triangular; Gaussian, Epanechnikov; Quartic; etc.

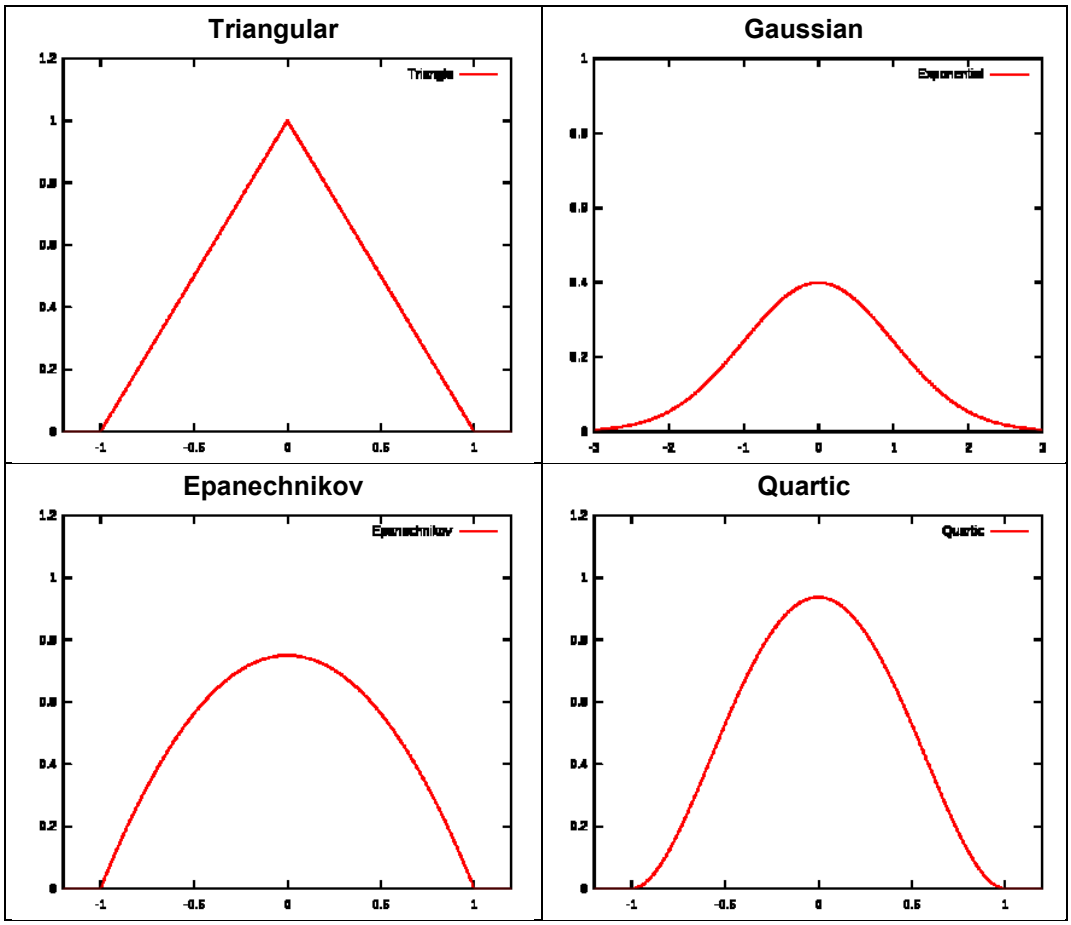

Figura 6 – Funzioni di distribuzioni

Dato che uno degli scopi del presente lavoro è la fruibilità della metodologia si è deciso di utilizzare e di confrontare le distribuzioni di probabilità più note come la funzione di distribuzione triangolare e gaussiana. La distribuzione più importante è senz'altro la gaussiana, è considerata il caso base delle distribuzioni di probabilità continue a causa del suo ruolo nel teorema del limite centrale. Diversi

fenomeni continui sembrano seguire, almeno approssimativamente, una distribuzione normale e può essere utilizzata per approssimare numerose distribuzioni di probabilità discrete. Le distribuzioni triangolari sono interessanti sia perché permettono di applicare concetti generali su funzioni matematiche elementari, sia perché rappresentano schemi abbastanza realistici per modellizzare l'incertezza su grandezze fisiche.

2.2. Considerazioni sull'algoritmo SONCA

SONCA è un facile strumento per il bilanciamento dei dati, agisce come una fase di pre-elaborazione, consentendo al sistema di apprendimento di ricevere le osservazioni come se appartenessero a un insieme di dati ben equilibrato, pertanto può essere applicato a qualsiasi sistema di apprendimento. SONCA è un algoritmo di bilanciamento, attraverso il quale è possibile ottenere un nuovo dataset sintetico affinché la variabile di risposta sia bilanciata per ogni classe. Il dataset sintetico è ottenuto in modo tale che ogni osservazione, estratta casualmente dal set di dati originale, è sostituita da un'altra osservazione selezionata all'interno dell'intera matrice dei predittori. La scelta casuale è effettuata ipotizzando una funzione di probabilità, triangolare o gaussiana, inversamente proporzionale alla distanza, tra l'osservazione estratta casualmente dal set di dati originale e tutte le osservazioni della matrice dei predittori.

Diversamente dagli altri algoritmi presenti in letteratura, attraverso SONCA è possibile trattare dataset sia dati di natura numerica sia dati di natura categorica, attraverso una codifica disgiuntiva completa della matrice dei predittori categorici. Inoltre trovandoci in presenza di dati misti, la distanza, tra l'osservazione estratta casualmente dal set di dati originale e tutte le osservazioni della matrice dei predittori, è calcolata attraverso la distanza euclidea normalizzata.

Capitolo 3 - Misure di performance

Un problema legato alle classi di risposta sbilanciate riguarda la valutazione dell'accuratezza del classificatore ed emerge sia nella scelta della misura dell'errore sia nella stima di esso. Le consuete misure di accuratezza, come ad esempio il tasso di errata classificazione, possono condurre a risultati fuorvianti perché dipendono fortemente dalla distribuzione di classe. Per esempio, in un problema in cui la classe rara è rappresentata in solo l'1% dei dati, seguendo una strategia di assegnazione della risposta basata sul criterio della classe più frequente (criterio della moda), il modello presenterebbe un errore complessivo pari all'1%. Apparentemente, il modello sembrerebbe avere un'ottima performance, ma in realtà se ciò è vero per la classe più frequente, al contrario, per la classe rara il modello sarà completamente inefficace (Menardi, 2009).

La scelta della misura di prestazione deve orientarsi verso la scelta di grandezze che siano indipendenti dalla distribuzione della classe di risposta. In letteratura, esistono numerose misure di performance che godono di questa proprietà e nel capitolo ne vengono descritte alcune.

3.1. Matrice di confusione e misure di performance

Il modo più semplice per valutare le prestazioni dei classificatori si basa sull'analisi della matrice di confusione. La matrice di confusione è la matrice contenente le informazioni circa lo stato della realtà e la classificazione ottenuta. Una formalizzazione generale di una matrice di questo tipo è proposta nella Tabella 3, le righe della matrice sono classi reali, e le colonne sono le classi previste. L'elemento sulla riga i e sulla colonna j è il numero di casi in cui il

classificatore ha classificato la classe "vera" j come classe i. Attraverso questa matrice è osservabile se vi è "confusione" nella classificazione di diverse classi.

Gli elementi della diagonale principale rappresentano il numero di osservazioni correttamente interpretati per ogni classe, gli altri sono il numero di errori ottenuti per ogni classe.

Tabella 3: Matrice di confusione k × k

		Classificazione						
		1	**2**	**…**	**j**	**…**	**k**	**Totale**
Eventi reali	**1**	n_{11}	n_{12}		n_{1j}		n_{1k}	$n_{1.}$
	2	n_{21}	n_{22}		n_{2j}		n_{2k}	$n_{2.}$
	…							
	i	n_{i1}	n_{i2}		n_{ij}		n_{ik}	$n_{i.}$
	k	n_{k1}	n_{k2}		n_{kj}		n_{kk}	$n_{k.}$
	Totale	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.k}$	n

Per semplificare la comprensione della matrice di confusione, in Tabella 4 è riportata una matrice di confusione per un problema di due classi, aventi valori di classe positiva e negativa. TP e TN indicano il numero di osservazioni positive e negative che sono classificate correttamente, mentre FN e FP indicano rispettivamente il numero di osservazioni positive e negative erroneamente classificati.

Tabella 4: Matrice di confusione 2 × 2

	Positive Prediction	Negative Prediction
Actual Positive Class	True Positive (TP)	False Negative (FN)
Actual Negative Class	False Positive (FP)	True Negative (TN)

Dalla matrice di confusione è possibile ricavare quattro metriche di performance che misurano direttamente le prestazioni di classificazione in classi positive e negative in modo indipendente (Weiss e Provost, 2003; Prati et al 2004; Chen et al., 2004; Guo et al 2008):

- False negative rate: $FN_{rate} = \frac{FN}{\mathrm{TP+FN}}$ è la porzione di casi positivi erroneamente classificati come appartenenti alla classe negativa;

- False positive rate: $FP_{rate} = \frac{FP}{FP+TN}$ è la porzione di casi negativi erroneamente classificati come appartenenti alla classe positiva;
- True negative rate: $\mathrm{TN_{rate}} = \mathrm{Acc^{-}} = \frac{\mathrm{TN}}{\mathrm{FP+TN}}$ è la porzione di casi negativi correttamente classificati come appartenenti alla classe negativa;
- True positive rate: $\mathrm{TP_{rate}} = \mathrm{Recall} = \mathrm{Acc^{+}} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$ è la porzione di casi positivi correttamente classificati come appartenenti alla classe positiva;
- $Precision = \frac{TP}{TP+FP}$ è la porzione di casi predetti positivi;
- $Err = \frac{FP+FN}{TP+FN+FP+TN}$ è la porzione di casi erroneamente classificati;
- $Acc = \frac{TP+TN}{TP+FN+FP+TN} = 1 - Err$ è la porzione di casi correttamente classificati.

Queste misure di performance hanno il vantaggio di essere indipendenti dalle probabilità a priori. Lo scopo di un classificatore è di minimizzare le percentuali di falsi positivi e negativi, ovvero massimizzare le percentuali dei veri negativi e positivi. Purtroppo, per la maggior parte delle applicazioni reali, vi è un compromesso tra FN_{rate} e FP_{rate} e, analogamente, tra TN_{rate} e TP_{rate}. È desiderabile avere un classificatore che dà elevata accuratezza della stima sulla classe di minoranza (Acc+), mantenendo un'accuratezza ragionevole per la classe di maggioranza (Acc-) (Chen et al., 2004).

Inoltre, l'accuratezza della stima sulla classe di minoranza (Acc+), definita anche Recall, è una misura di completezza mentre la Precision è una misura dell'esattezza (He e Garcia, 2008). Come stato visto l'obiettivo principale degli algoritmi di ricampionamento è di migliorare la Recall senza danneggiare la Precision. Tuttavia, gli obiettivi della Recall e della Precision possono essere spesso in conflitto, poiché quando si aumentano i veri positivi per la classe di minoranza, il numero di falsi positivi potrebbe aumentare riducendo la Precision (Ding, 2011).

Altre metriche di valutazione sono spesso adottate per incorporare con un'unica misura due o più parametri, in modo da avere valutazioni più complete dei problemi di squilibrio e di apprendimento, come (Weiss e Provost, 2003; Prati et al 2004; Chen et al., 2004; Kotsiantis, et al., 2006; Guo et al 2008; Weng, e Poon, 2008; Cao et al., 2011; Ding, 2011):

- $F-measure = \frac{(1+\beta)^2 \times Precision \times Recall}{\beta^2 \times Precision + Recall} = \frac{(1+\beta)^2 \times Precision \times Acc^+}{\beta^2 \times Precision + Acc^+}$ è una media armonica tra la Recall e la Precision. Dove β è un coefficiente di aggiustamento relativo all'importanza della Precision verso la Recall. Esso è generalmente pari a 1.
- $G-mean = (Acc^- \times Acc^+)^{\frac{1}{2}}$ è la media geometrica tra le classi

F-measure è una misura che combina la Recall e la Precision, può essere definita come una media armonica tra la Recall e la Precision ed emette un numero unico che riflette la "bontà" di un classificatore, alla presenza di classi rare (Chawala, 2010; Ding, 2011). Un'altra metrica fornita dalla letteratura è G - mean, considera le prestazioni di entrambe le classi, positiva e negativa, e utilizza la media geometrica per combinarle. Un alto valore di G-Mean può essere raggiunto solo con elevata accuratezza della classe positiva e negativa (Ding, 2011).

Sebbene F-measure e G - mean sono misure che apportano grandi miglioramenti rispetto la Precision e la Recall, sono ancora inefficaci nel definire le prestazioni degli stimatori (Lui e Garcia, 2009; Chen et al., 2004).

3.2. Curva ROC (Receiver Operating Characteristic) e AUC (Area sottesa alla curva ROC)

Uno strumento appropriato per valutare le prestazioni dei classificatori è fornito dalla curva ROC (Receiver Operating Characteristic) (Bradley, 1997; Chawla, et

al., 2002; Chawla, 2010; Chen et al., 2004; Kotsiantis et al., 2006; Guo et al., 2008 Ding, 2011).

La curva ROC è stata introdotta per la prima volta da alcuni ingegneri, durante la seconda guerra mondiale, per l'analisi delle immagini radar e lo studio del rapporto segnale/disturbo. Il problema era quello di riconoscere il segnale causato dalla presenza di oggetti nemici su campi di battaglia, distinguendolo dal rumore di fondo presenti nei segnali radar. Più tardi, tra il 1970 e il 1980, diventò evidente l'importanza dell'utilizzo di questa tecnica nella valutazione dei test diagnostici, in campi quali la radiologia, cardiologia, chimica clinica ed epidemiologia. Recentemente è entrata anche nell'ambito del data mining (Azzalini e Scarpa, 2009).

A partire dalla matrice di confusione definiamo:

- La sensitivity che esprime la proporzione di Veri Positivi rispetto al numero totale di osservazioni positive ed è definita come:

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

- La specificity che esprime la proporzione di Veri Negativi rispetto al numero totale di osservazioni negative ed è definita come:

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

La curva ROC (Figura 7) è un diagramma in cui sono riportate in ordinata la sensitivity e in ascissa il complemento a 1 della specificity (1- specificity, ossia il tasso dei falsi positivi) (Rocchi, 2001). La Sensitivity è condizionata negativamente dal numero di falsi negativi: pertanto un test molto sensibile dovrà associarsi a una quota molto bassa di falsi negativi. La Specificity è influenzata

dal numero di falsi positivi; ovvero un modello sarà tanto più accurato, quanto più bassa sarà la quota dei falsi positivi.

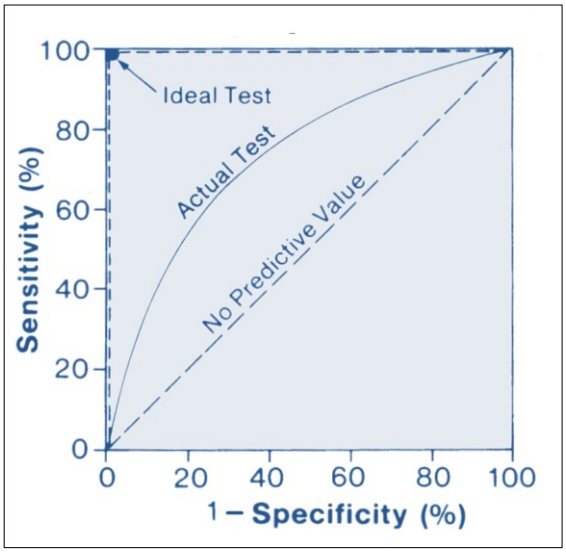

Figura 7 - Esempio di Curva ROC

La curva ROC per la classificazione casuale, (1- specificity, sensitivity), è pari alla diagonale del quadrato. Una regola di classificazione è tanto migliore quanto più la curva ROC si discosta dalla diagonale verso il punto (0; 1).

In particolare, accade che per ogni classificatore, il numero di veri positivi non può aumentare senza che aumentino anche i falsi positivi, e poiché la curva visualizza contemporaneamente i TP_{rate} e i FP_{rate}, la curva ROC rappresenta un valido metodo di analisi.

Infine, l'area compresa tra la curva ROC e la diagonale del quadrato è definita AUC (Area Under Curve). AUC equivale alla probabilità che il classificatore identifichi correttamente un'osservazione estratta a caso dal gruppo dei positivi. Pertanto, AUC è comunemente usata come misura globale per la valutazione delle prestazioni dei classificatori (Huang e Ling, 2005; Ramentol et al., 2012).

3.3. Alcune applicazioni delle misure di performance con gli algoritmi di ricampionamento

Di seguito sono state riportate le misure di performance calcolate da diversi autori per testare i seguenti algoritmi di ricampionamento:

- Synthetic Minority Over-sampling Technique (SMOTE);
- Synthetic Minority Over-sampling Technique per dati nominali e continui (SMOTE-NC);
- SMOTE- borderline;
- Adaptive Synthetic Sampling Approach for Imbalanced (ADASYN);
- Learning Random Over Sampling Examples (ROSE).

3.3.1. Synthetic Minority Over-sampling Technique (SMOTE)

Chawala et al., (2002) per i loro esperimenti hanno usato 3 processi differenti di stima (C4.5, Ripper e Naive Bayes Classifier) e nove dataset con differenti distribuzioni (Tabella 5).

Tabella 5: Distribuzione dei dataset utilizzati da Chawala et al. (2002)

	Majority Class		Minority Class		
Dataset	N	%	N	%	Total
Pima	500	0.65	268	0.35	768
Phoneme	3818	0.71	1586	0.29	5404
Adult	37155	0.76	11687	0.24	48842
E-state	46869	0.88	6351	0.12	53220
Satimage	5809	0.90	626	0.10	6435
Forest Cover	35754	0.93	2747	0.07	38501
Oil	896	0.96	41	0.04	937
Mammography	10923	0.98	260	0.02	11183
Can	435512	0.98	8360	0.02	443872

Sono stati provati valori crescenti di ricampionamento: 50%, 100%, 200%, etc. fino a 500%. Da Figura 8 a Figura 11 sono riportate alcune delle curve ROC ottenute per tale studio.

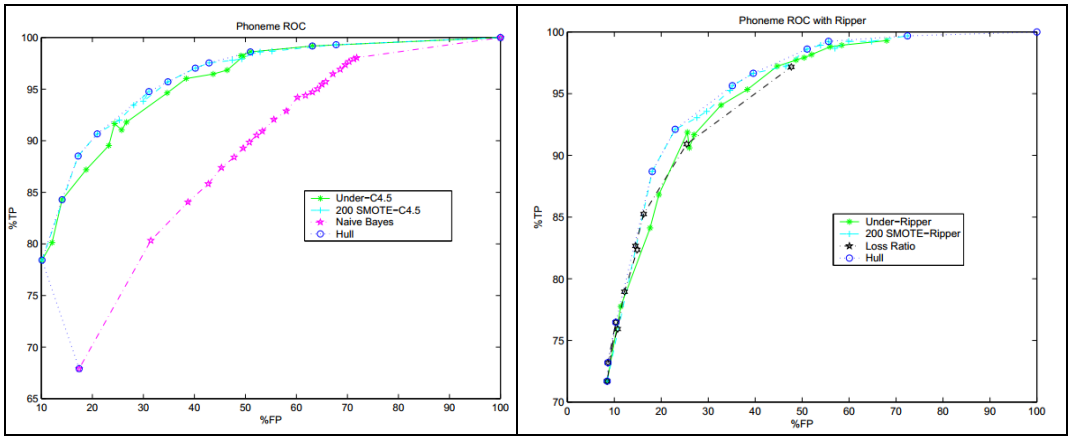

Figura 8 – Curve ROC con e senza ricampionamento con SMOTE per il dataset Phoneme (Chawala et al., 2002)

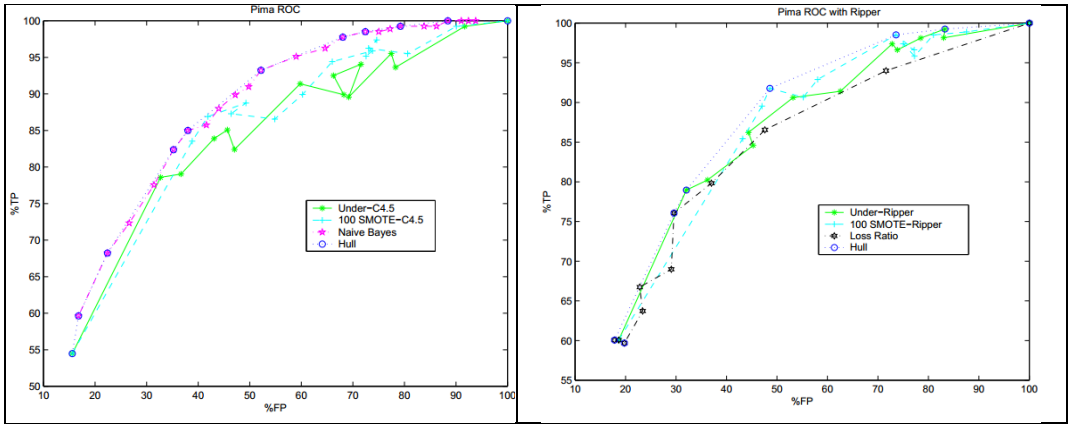

Figura 9 – Curve ROC con e senza ricampionamento con SMOTE per il dataset Pima (Chawala et al., 2002)

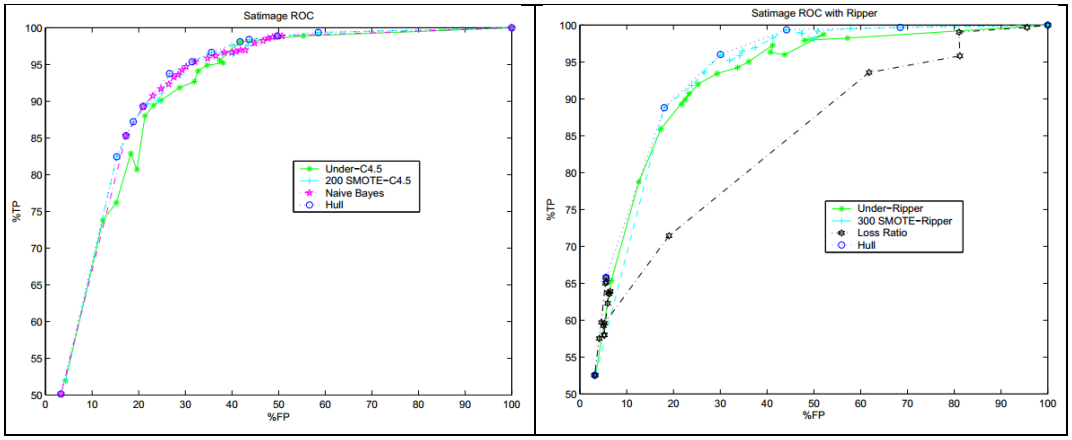

Figura 10 – Curve ROC con e senza ricampionamento con SMOTE per il dataset Satimage (Chawala et al., 2002)

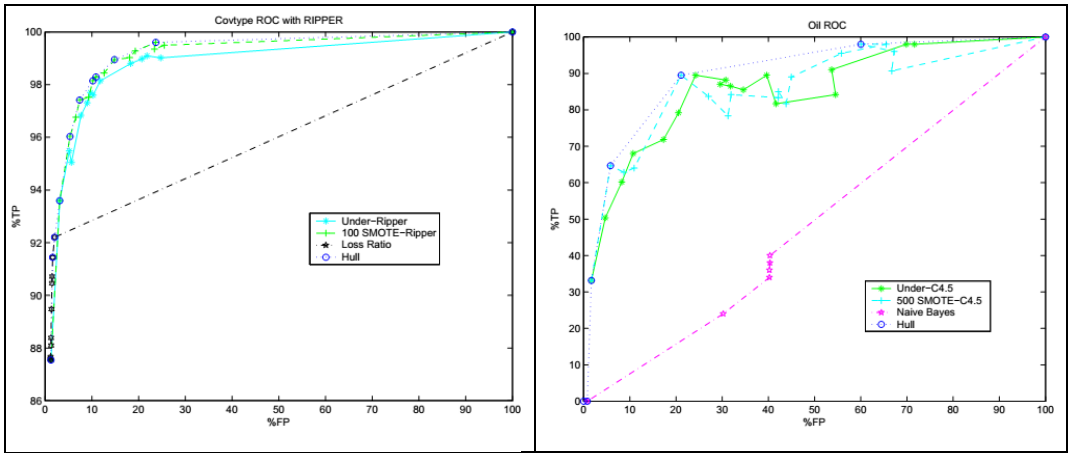

Figura 11 – Curve ROC con e senza ricampionamento con SMOTE per i dataset Forest Cover e Oil (Chawala et al., 2002)

L'analisi di tali curve, ha evidenziato che le curve migliorano all'aumentare della percentuale fino ad arrivare a un ricampionamento del 200%. Da lì in avanti non si notano altri miglioramenti e con percentuali elevate, come del 500%, le curve ROC sono addirittura peggiori rispetto a quelle che si ottengono lavorando su dataset originali. Questo è indice del fatto che, con tali percentuali, si creano troppe osservazioni artificiali che vanno a sovraffollare lo spazio delle osservazioni rendendo difficile la distinzione tra classi.

3.3.2. Synthetic Minority Over-sampling Technique per dati nominali e continui (SMOTE-NC)

Chawala (2003) per testare l'algoritmo utilizzò un solo dataset, adult dataset, avente anche i predittori categorici.

In Figura 12, sono riportate le curve ROC ottenute dai diversi modelli stimati.

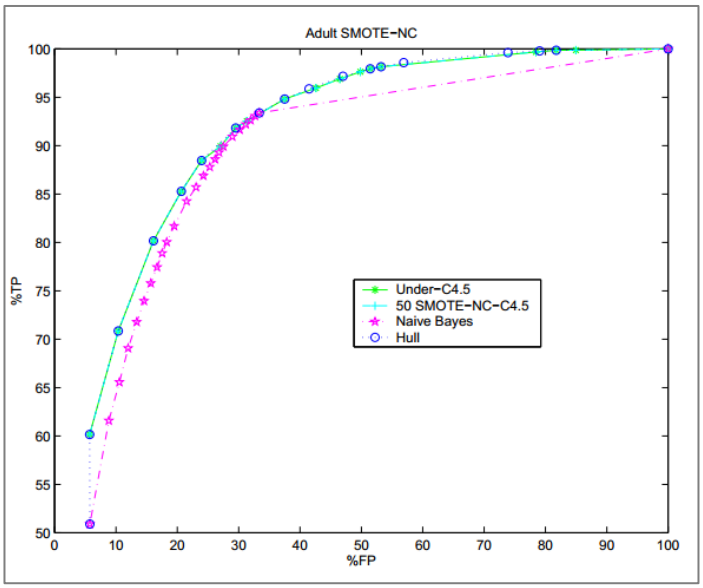

Figura 12- Risultati di SMOTE-NC (Chawala, 2003)

I risultati ottenuti hanno evidenziato che l'utilizzo dell'algoritmo di SMOTE-NC, per tale dataset, non porta nessun miglioramento nell'implementazione del processo di stima utilizzando la regressione ad albero C.4.5.

3.3.3. SMOTE- borderline

Gli autori per i loro esperimenti hanno usato l'albero di regressione C.4.5, e 5 database con differenti distribuzioni. In Tabella 6 sono riportati i dataset utilizzati con le relative distribuzione di classe.

Tabella 6: Distribuzione dei dataset utilizzati (Han et al., 2005)

Data set	Majority Class	%	Minority Class	%	Total
Circle(Simulation)	1500	0.94	100	0.06	1600
Pima(UCI)	501	0.65	267	0.35	768
Satimage(UCI)	5809	0.90	626	0.10	6435
Haberman(UCI)	225	0.74	81	0.26	306

In Figura 13 sono illustrate le misure di performance, F-value e di TP_{rate}, per la classe di minoranza usando dataset originali e dataset sintetici ottenuti con gli

algoritmi di ricampionamento: SMOTE, borderline-SMOTE1, borderline-SMOTE2, e random oversampling.

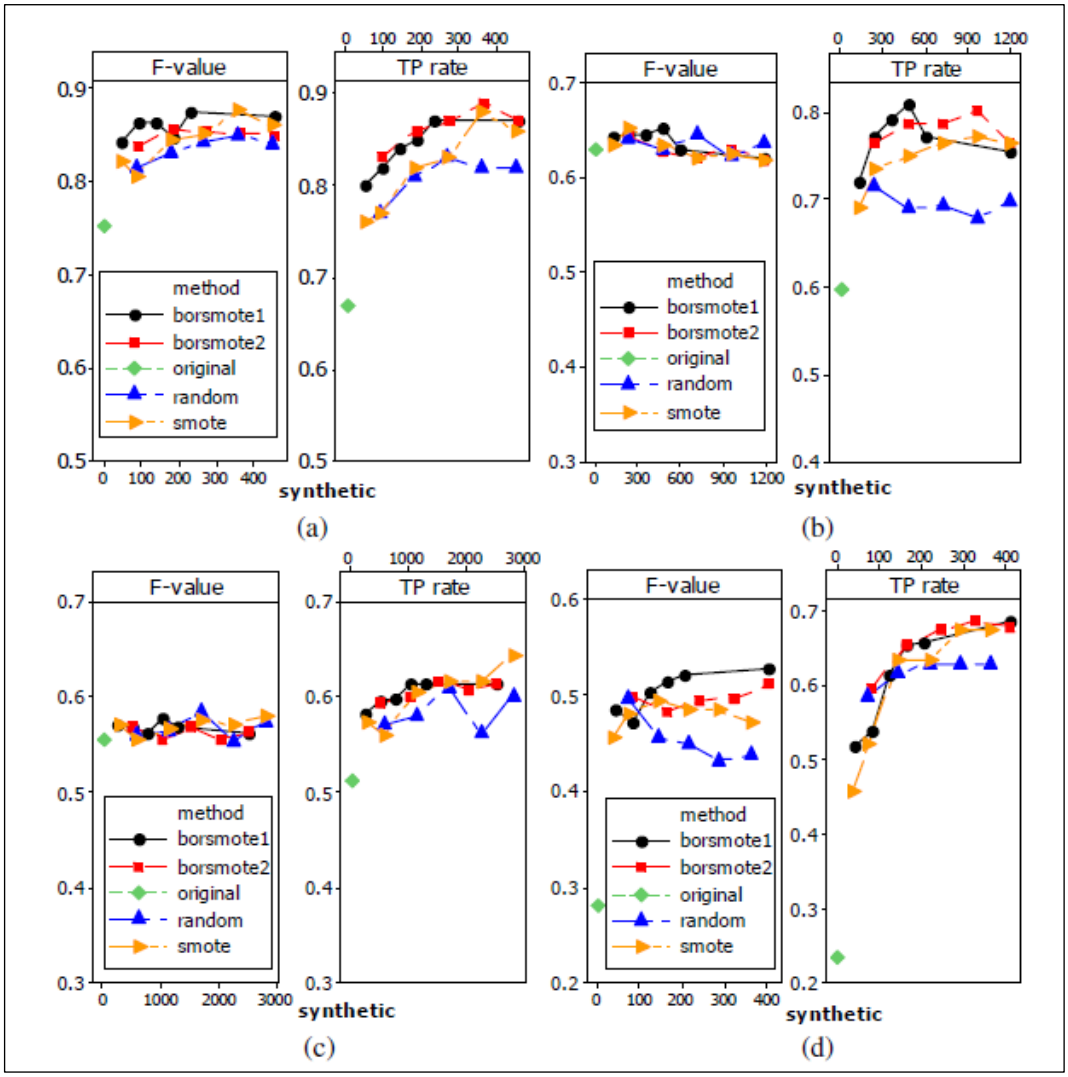

Figura 13 –Risultati di borderline-SMOTE1 e borderline-SMOTE2 (Han et al., 2005)

I risultati hanno dimostrato che utilizzando borderline-SMOTE1 si hanno migliori valori di F-value e TP_{rate}. Borderline-SMOTE2 raggiunge valore di TP_{rate} ancora migliori ma F-value diminuisce a causa della sovrapposizione tra le due classi (Han et al., 2005).

3.3.4. Adaptive Synthetic Sampling Approach for Imbalanced (ADASYN)

He et al. (2008), per testare l'efficacia dell'algoritmo di ricampionamento ADASYN, hanno utilizzato 5 dataset appartenenti alla banca dati dell'UCI Machine Learning Repository, con diverse distribuzioni (Tabella 7).

Tabella 7: Distribuzione dei dataset utilizzati (He et al., 2009)

Data set	Majority Class	%	Minority Class	%	Total
Vehicle	647	0.76	199	0.24	846
Diabetes	500	0.65	268	0.35	768
Vowel	900	0.91	90	0.09	990
Ionosphere	225	0.64	126	0.36	351
Abalone	689	0.94	42	0.06	731

Tali dati sono stati usati per stimare alberi di classificazione, senza nessun processo di ricampionamento e con un processo di ricampionamento, utilizzando come algoritmi SMOTE e ADASYN. In Tabella 8, sono riportate le diverse misure di performance calcolate: Accuracy (OA); Precision; F-measure; G-mean.

Tabella 8: Comparazione delle metriche di valutazione (He et al., 2009)

Dataset	Methods	OA	Precision	Recall	F_measure	G_mean
Vehicle	Decision tree	0.9220	**0.8454**	0.8199	0.8308	0.8834
	SMOTE	0.9239	0.8236	0.8638	0.8418	0.9018
	ADASYN	**0.9257**	0.8067	**0.9015**	**0.8505**	**0.9168**
Pima Indian Diabetes	Decision tree	0.6831	**0.5460**	0.5500	0.5469	0.6430
	SMOTE	0.6557	0.5049	**0.6201**	0.5556	0.6454
	ADASYN	**0.6837**	0.5412	0.6097	**0.5726**	**0.6625**
Vowel recognition	Decision tree	**0.9760**	**0.8710**	0.8700	0.8681	0.9256
	SMOTE	0.9753	0.8365	0.9147	**0.8717**	0.9470
	ADASYN	0.9678	0.7603	**0.9560**	0.8453	**0.9622**
Ionosphere	Decision tree	0.8617	**0.8403**	0.7698	0.8003	0.8371
	SMOTE	0.8646	0.8211	0.8032	0.8101	0.8489
	ADASYN	**0.8686**	0.8298	**0.8095**	**0.8162**	**0.8530**
Abalone	Decision tree	**0.9307**	**0.3877**	0.2929	**0.3249**	0.5227
	SMOTE	0.9121	0.2876	0.3414	0.3060	0.5588
	ADASYN	0.8659	0.2073	**0.4538**	0.2805	**0.6291**
Winning times	Decision tree	2	**5**	0	1	0
	SMOTE	0	0	1	1	0
	ADASYN	**3**	0	**4**	**3**	**5**

I risultati hanno evidenziato che l'algoritmo ADASYN raggiunge risultati competitivi. ADASYN fornisce le migliori prestazioni in termini di G-mean per tutti i set di dati utilizzati. Questo significa che tale algoritmo permette di ottenere una maggior accuratezza per entrambe le classi, sia di minoranza sia di maggioranza, e non sacrifica una classe per preferirne a un'altra (He et al., 2009).

3.3.5. Learning Random Over Sampling Examples (ROSE)

Menardi e Torelli (2013), per testare l'efficacia dell'algoritmo di ricampionamento ROSE, hanno utilizzato 20 dataset, con diverse distribuzioni (Tabella 7).

Tabella 9: Distribuzione dei dataset utilizzati da Menardi e Torelli (2013)

Data set	Majority Class	%	Minority Class	%	Total
Wine quality	4878	99.6	20	0.4	4898
Forest cover	38308	99.5	193	0.5	38501
Infocamere	11121	99.3	78	0.7	11199
Abalone	4144	99.2	33	0.8	4177
Hypothyroid	2374	99	24	1	2398
Adult	47865	98	977	2	48842
Phoneme ELENA Oral/nasal sounds	5269	97.5	135	2.5	5404
Breast cancer	552	97	17	3	569
Cardiotocography	1882	96.5	68	3.5	1950
Transfusion	718	96	30	4	748
Glass	204	95.5	10	4.5	214
Pima indians	730	95	38	5	768
Cylinder bands	343	94	22	6	365
Vehicle silhouettes	399	93	30	7	429
Image segmentation	2125	92	185	8	2310
Spectf	243	91	24	9	267
Vertebral column	279	90	31	10	310
Parkinsons	171	87.5	24	12.5	195
Concrete compressive strength	876	85	155	15	1030
Credit screening	522	80	131	20	653

In Tabella 10, è riportata l'AUC per i differenti dataset per diversi algoritmi di ricampionamento.

Osservando i valori di AUC per i dataset sbilanciati si evidenzia che in presenza di squilibri assoluti (con poche osservazioni appartenenti alla classe minoritaria) si ha una scarsa accuratezza della classificazione, come per i dataset Wine quality, Transfusion, Glass e Vehicle silhouettes. Quando invece lo squilibrio è relativo, come per i dataset Forest cover, Adult, Phonemeand, Concrete Compressive, Strength, gli effetti dello squilibrio di classe sono meno critici e si ha una più alta accuratezza della classificazione.

Tabella 10: Comparazione della AUC (Menardi e Torelli, 2013)

Data set	Imbalanced data	Undersampling	SMOTE	ROSE
Wine quality	0.557	0.604	0.624	0.782
Forest cover	0.785	0.855	0.889	0.903
Infocamere	0.536	0.567	0.781	0.768
Abalone	0.571	0.714	0.663	0.717
Hypothyroid	0.658	0.927	0.959	0.974
Adult	0.624	0.796	0.752	0.794
Phoneme ELENA Oral/nasal sounds	0.758	0.726	0.751	0.828
Breast cancer	0.822	0.886	0.879	0.898
Cardiotocography	0.673	0.795	0.848	0.852
Transfusion	0.558	0.584	0.561	0.66
Glass	0.533	0.533	0.712	0.76
Pima indians	0.571	0.583	0.665	0.714
Cylinder bands	0.571	0.539	0.534	0.644
Vehicle silhouettes	0.551	0.536	0.521	0.698
Image segmentation	0.789	0.877	0.906	0.904
Spectf	0.574	0.668	0.565	0.685
Vertebral column	0.686	0.816	0.816	0.851
Parkinsons	0.629	0.725	0.719	0.752
Concrete compressive strength	0.799	0.826	0.804	0.819
Credit screening	0.682	0.757	0.741	0.759
pvalues	$<10^{-3}$	0.001	$<10^{-3}$	-

Inoltre, utilizzando algoritmi di ricampionamento si osservano dei migliori valori di AUC. In particolare, sono stati ottenuti valori di AUC più alti per 18 dataset su 20, eccetto che per Forest cover e Image segmentation.

Capitolo 4 - Analisi dei Risultati

Nel presente capitolo, utilizzando numerosi dataset ben noti in letteratura (dal repository UCI Machine Learning), si sono volute valutare le prestazioni dell'algoritmo SONCA rispetto a due aspetti principali:

- La sensibilità delle performance del metodo rispetto ai diversi parametri che lo caratterizzano;
- La comparazione delle performance rispetto alle principali proposte metodologiche già presenti in letteratura;

Inoltre il processo di valutazione è completato con l'utilizzo di SONCA su 2 dataset reali che riguardano il problema dello studio delle determinanti degli incidenti stradali mortali. I risultati di quest'ultima analisi sono riportati nel capitolo successivo.

In Tabella 11, sono riportati i quattro dataset utilizzati con le relative statistiche descrittive e il relativo numero di predittori.

Tabella 11: Machine Learning Repository: UCI database

		Un-success		Success		
Dataset	**# Var.**	**N**	**%**	**N**	**%**	**Tot**
Cover type	12	35754	92.87	2747	7.13	38501
Adult	14	24720	75.92	7841	24.08	32561
Pima Indian Diabetes	8	500	65.10	268	34.90	768
Glass	9	197	92.06	17	7.94	214

In Figura 14, sono riportate le fasi eseguite nel processo di elaborazione di ogni dataset.

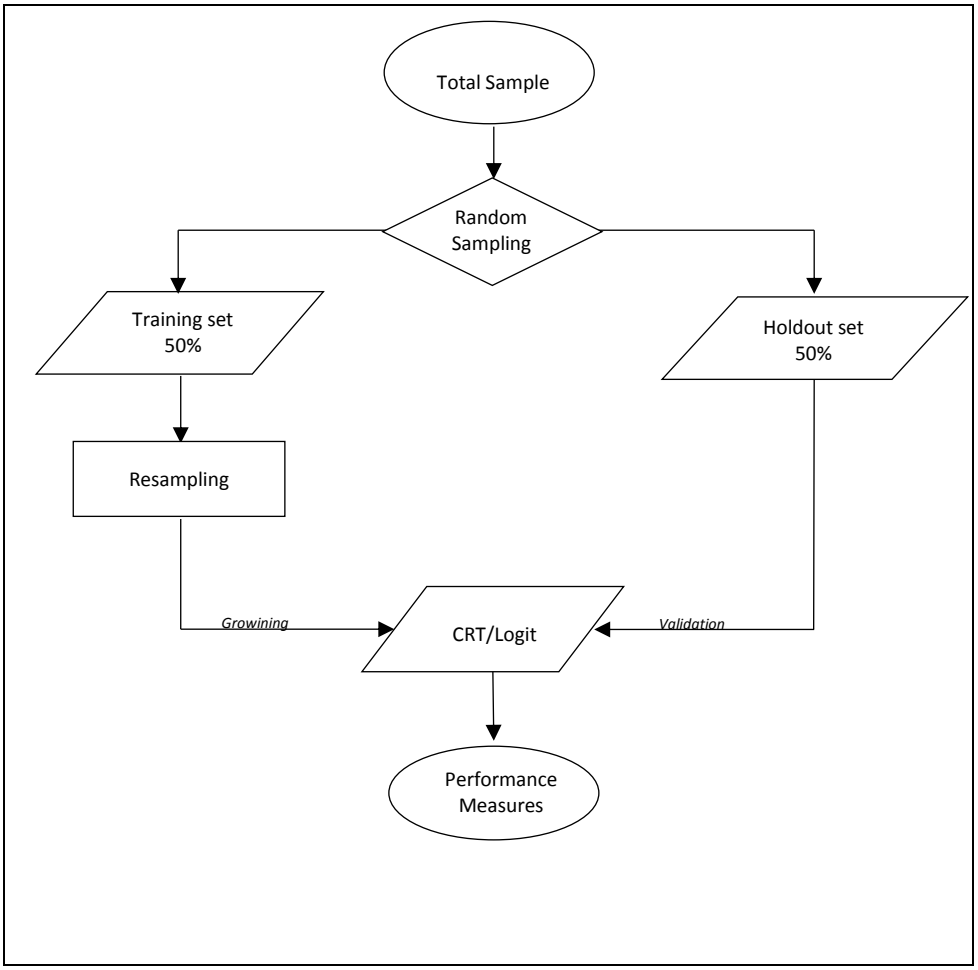

Figura 14 – Fasi del processo di apprendimento e di validazione

Per ogni dataset, volendo individuare modelli che siano più vicini ai dati cercando di evitare problemi di overfitting, è stato necessario dividere la popolazione in due set di dati:

- Training set (con il 50% delle osservazioni) utilizzato per la stima dei modelli;
- Holdout set (con il restante 50% delle osservazioni) utilizzato per la successiva verifica dei modelli.

Le due partizioni sono state estratte casualmente; inoltre le osservazioni del holdout set sono state estratte in modo da essere indipendenti dai casi di training,

l'unica relazione fra i due sets di dati è l'appartenenza alla stessa popolazione (Dulli et al., 2009).

Dopo aver suddiviso il dataset in due parti, il training set è stato oggetto di una procedura di ricampionamento per bilanciare la classe di risposta.

I diversi modelli sono stati stimati utilizzando il nuovo dataset bilanciato.

In particolare, si è voluto verificare l'efficacia di SONCA, non solo per modelli di natura non parametrica ma anche per modelli di natura parametrica. Data la natura binomiale della variabile risposta, sono stati scelti come modelli di stima:

- Il logistic regression models (Logit, Berkson 1944) è un modello di regressione per variabili di risposta dicotomiche che ha l'obiettivo di stimare la probabilità condizionata dell'evento successo;
- L'Albero di classificazione CART (Classification And Regression Trees, Brieman et al., 1984) è un metodo gerarchico per descrivere una partizione dell'insieme delle unità statistiche basata sulla relazione tra una variabile risposta e un set di predittori di natura mista (quantitativi e categorici)

Dopo aver stimato i modelli per i dataset bilanciati, le misure di performance sono state calcolate attraverso le matrici di confusione, ottenute dal processo di verifica dei modelli stimati utilizzando l'holdout set.

Pertanto, il presente capitolo, è stato suddiviso in due fasi:

1) Studio della sensibilità del parametro m e delle finzioni di distribuzione di probabilità: sono state confrontate le misure di performance ottenute con SONCA, utilizzando diverse curve di distribuzione e diversi valori di m.
2) Confronto dell'efficacia di SONCA rispetto agli algoritmi SMOTE e ROSE.

4.1. Dataset utilizzati

I dataset utilizzati appartengono alla banca dati dell'UCI Machine Learning Repository, le cui distribuzioni di classe sono riportate in Tabella 12.

La scelta dei dataset è stata fatta in modo che ci siano differenti distribuzioni, offrendo così diversi domini per le analisi.

Tabella 12: Machine Learning Repository: UCI database

		Insuccesso		Successo		
Dataset	**# Var.**	**N**	**%**	**N**	**%**	**Tot**
Cover type	12	35754	92.87	2747	7.13	38501
Adult	14	24720	75.92	7841	24.08	32561
Pima Indian Diabetes	8	500	65.10	268	34.90	768
Glass	9	197	92.06	17	7.94	214

I dataset scelti in tale lavoro, sono stati utilizzati anche in altri studi simili, come:

- Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002);
- SMOTE- borderline (Han et al., 2005);
- SMOTEBoost algorithm (Chawla et al., 2003);
- Adaptive Synthetic Sampling Approach for Imbalanced (ADASYN) (He et al., 2009);
- Learning Random Over Sampling Examples (ROSE) (Menardi et al., 2012).

4.1.1. Cover type

Obiettivo del dataset è di determinare il tipo copertura forestale sulla base di variabili cartografiche. Il dataset è costituito solo da predittori numerici (Tabella 13).

Originariamente il dataset era composto di 581012 osservazioni e la variabile di risposta era costituita da sette classi, per confrontare i risultati con quelli della letteratura sono state considerate solo due classi, Ponderosa Pine con 35754 osservazioni e Cottonwood / Willow con 2747 osservazioni.

Tabella 13: Descrizione del database Cover Type

Variabile	Tipo di variabile
Elevation	continua
Aspect	continua
Slope	continua
Horizontal_Distance_To_Hydrology	continua
Vertical_Distance_To_Hydrology	continua
Horizontal_Distance_To_Roadways	continua
Hillshade_9am	continua
Hillshade_Noon	continua
Hillshade_3pm	continua
Horizontal_Distance_To_Fire_Points	continua
Wilderness_Area (4 binary columns)	continua
Soil_Type (40 binary columns)	continua
Cover_Type (7 types)	intero

Come si osserva in Tabella 14, la classe di risposta è molto sbilanciata, la classe 0 ha una frequenza del 92.9%, mentre la classe 1 ha una frequenza del 7.1%. In presenza di una distribuzione della variabile di risposta estremamente sbilanciata il processo di apprendimento può essere distorto, perché il modello tende a focalizzarsi sulla classe prevalente e ignorare gli eventi rari.

Tabella 14: Distribuzioni di classe per la variabile Cover Type

Cover_Type	N	$F=N/N_{tot}$ [%]
0	35754	92.9
1	2747	7.1
Totale	**38501**	100.0

Attraverso un processo di estrazione casuale il dataset è stato suddiviso in due dataset:

- Training set (Tabella 15) costituito da 19191 osservazioni pari al 49.8% delle osservazioni totali;

Tabella 15: Distribuzioni di classe del training set per la variabile Cover Type

Cover_Type	N	F=N/N_{tot} [%]
0	17817	92.8
1	1374	7.2
Totale	**19191**	100.0

- Holdout test (Tabella 16) costituito da 19310 pari 50.2% delle osservazioni totali.

Tabella 16: Distribuzioni di classe del dataset test per la variabile Cover Type

Cover_Type	N	F=N/N_{tot} [%]
0	17937	92.9
1	1373	7.1
Totale	**19310**	100.0

4.1.2. Adult dataset

L'obiettivo del dataset è di determinare se una persona avente determinate caratteristiche, guadagna più di 50 $K/anno.

Il dataset è costituito da 6 predittori continui e 8 predittori nominali (Tabella 17).

Tabella 17: Predittori per Adult dataset

Variabile	Tipo di variabile
Age	Continua
Workclass	Categorica – 8 classi (Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked).
Fnlwgt	Continua
Education	Categorica – 16 classi (Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool).
Education-Num	Continua
Marital-Status	Categorica – 7 classi (Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse).
Occupation	Categorica – 13 classi (Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces).
Relationship	Categorica – 6 classi (Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried).

Variabile	Tipo di variabile
Race	Categorica – 5 classi (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black).
Sex	Categorica –2 classi (Female, Male).
Capital-Gain	Continua
Capital-Loss	Continua
Hours-Per-Week	Continua
Native-Country	Categorica – 41 classi (United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands).

Il dataset è composto da 32561 osservazioni, la variabile di risposta è costituita da 7841 osservazioni appartenente alla classe minoritaria, pari al 24% del totale, e 27720 osservazioni appartenenti alla classe maggioritaria, pari al 76% del totale (Tabella 18).

Tabella 18: Distribuzione della variabile risposta Adult dataset

	N	f=N/Ntot
0	24720	75.9
1	7841	24.1
Totale	**32561**	100.0

Attraverso un processo di estrazione casuale il dataset è stato suddiviso in due dataset:

- Training test (Tabella 19) costituito da 16256 osservazioni pari al 49.9% delle osservazioni totali;

Tabella 19: Distribuzioni di classe del training set per il dataset Adult

Adult	N	%
1	3914	24.1
0	12342	75.9
Totale	**16256**	**100.0**

- Holdout test (Tabella 20) è costituito da 16305 pari 50.1% delle osservazioni totali.

Tabella 20: Distribuzioni di classe dell'holdout set per il dataset Adult

Adult	N	%
1	3927	24.1
0	12378	75.9
Totale	**16305**	**100.0**

4.1.3. Glass

Obiettivo del dataset è quello di determinare il tipo vetro sulla base della composizione del vetro. Il dataset è costituito solo da predittori numerici (Tabella 21).

Tabella 21: Descrizione del database Glass

Codice	Variabile	Tipo di dato
RI	refractive index	continua
Na	Sodium	continua
Mg	Magnesium	continua
Al	Aluminum	continua
Si	Silicon	continua
K	Potassium	continua
Ca	Calcium	continua
Ba	Barium	continua
Fe	Iron	continua
Glass	Type of glass	Intero(1 a 7)

Originariamente la classe di risposta era costituita da 7 tipi di vetro differenti, per confrontare i risultati con i risultati della letteratura, la variabile di risposta è stata trasformata in una variabile binaria, con modalità pari a 1 per la tipologia di vetro "vehicle windows float processed" e modalità pari a 0 per tutte le altre tipologie di vetro. La variabile di risposta è costituita da 17 osservazioni appartenente alla classe minoritaria, pari all'8% del totale, e 197 osservazioni appartenenti alla classe maggioritaria, pari al 92% del totale (Tabella 22).

Tabella 22: Distribuzione di classe della variabile risposta Type of Glass

Glass	N	f=N/Ntot
0	197	92.1
1	17	7.9
Totale	**214**	100.0

Attraverso un processo di estrazione casuale il dataset è stato suddiviso in 2 dataset:

- Training test (Tabella 23) costituito da 110 osservazioni pari al 49.8% delle osservazioni totali;

Tabella 23: Distribuzioni di classe del training set per il dataset Glass

Glass	N	$F=N/N_{tot}$ [%]
0	101	91.82
1	9	8.18
Totale	**110**	100.0

- Holdout test (Tabella 24) costituito da 19'310 pari 50.2% delle osservazioni totali.

Tabella 24: Distribuzioni di classe del holdout set per il dataset Glass

Glass	N	$F=N/N_{tot}$ [%]
0	96	92.31
1	8	7.69
Totale	**104**	100.0

4.1.4. Pima Indian Diabetes

Obiettivo del dataset è di determinare i casi di diabete positivi in una popolazione vicino a Phoenix, Arizona. Il dataset è costituito da solo 8 predittori di natura continua (Tabella 25).

Tabella 25: Descrizione del database Pima Indian Diabetes

Variabile	Tipo di dato
Number of times pregnant	intero
Plasma glucose	continua
Diastolic blood pressure	continua
Triceps skin fold thickness	continua
2-Hour serum insulin	continua
Body mass index	continua
Diabetes pedigree function	continua
Age	intero

La varabile di risposta ha due classi e 768 osservazioni (Tabella 26). Il numero di osservazioni positive della classe sono 268.

Tabella 26: distribuzioni di classe per la variabile Pima Indian Diabetes

Pima Indian Diabetes	N	**F=N/N_{tot}**
0	500	65.1
1	268	34.9
Totale	**768**	100.0

Attraverso un processo di estrazione casuale il dataset è stato suddiviso in due dataset:

- Training set (Tabella 27) costituito da 382 osservazioni pari al 49.7% delle osservazioni totali;

Tabella 27: Distribuzioni di classe per il training test per il dataset Pima Indian Diabetes

Pima Indian Diabetes	N	**F=N/N_{tot}**
0	239	62.6
1	143	37.4
Totale	**382**	100.0

- Holdout test (Tabella 28) costituito da 386 osservazioni pari al 50.3% delle osservazioni totali.

Tabella 28: Distribuzioni di classe dell'holdout set per il dataset Pima Indian Diabetes

Pima Indian Diabetes	N	**F=N/N_{tot}**
0	261	67.6
1	125	32.4
Totale	**386**	100.0

4.2. Sensibilità al parametro m

In tale paragrafo è valutata la sensibilità al parametro m, in altre parole quante osservazioni sintetiche devono essere generate affinché si minimizzino le percentuali di falsi positivi e negativi, e si massimizzino le percentuali dei veri negativi e positivi, senza accrescere inutilmente la complessità del dataset e quindi il costo computazionale dell'analisi.

A tal fine sono stati analizzati due dataset, Cover type e Adult dataset. Per entrambi i dataset sono stati stimati il CART e il Logit sia con il training set originale sia con il training set ricampionato con SONCA, utilizzando una distribuzione di probabilità sia triangolare sia gaussiana, per diversi valori di m.

Nel seguito sono riportati i risultati di sintesi delle diverse analisi, mentre in Appendice 1 sono riportate tutti i risultati delle analisi

4.2.1. Cover type

Il dataset Cover type è composto da 38501 osservazioni, con tutti predittori di natura numerica. La variabile di risposta cover type è estremamente sbilanciata, la modalità 0 (la classe di maggioranza) ha una frequenza del 92.8%, mentre la classe 1 (la classe di minoranza) ha una frequenza del 7.1%.

Per tale dataset sono stati stimati l'albero di classificazione CART e il logit nei seguenti casi:

- Dataset originale senza ricampionamento;
- Dataset bilanciato con SONCA usando la distribuzione di probabilità triangolare e con:
    - m=3500;
    - m=5500;
    - m=7500;

- m=9500.

- Dataset bilanciato con SONCA usando la distribuzione di probabilità gaussiana e con:
  - m=3500;
  - m=5500;
  - m=7500;
  - m=9500.

In Tabella 29 e Tabella 30 sono riportate le distribuzioni di frequenza per il training e l'holdout set, sia per il dataset originale sia per i dataset bilanciati con SONCA.

Tabella 29: Distribuzioni di classe per il training e l'holdout set, per il dataset originale e per i dataset bilanciati con SONCA per i diversi valori di m con distribuzione di probabilità triangolare

		Training set		Holdout set	
m	$m/\#1_{originale}$	#1	#0	#1	#0
-	-	1374	17817	1373	17937
3500	2.5	3491	3453	1373	17937
5500	4.0	5457	5453	1373	17937
7500	5.5	7428	7599	1373	17937
9500	6.9	9445	9451	1373	17937

Tabella 30: Distribuzioni di classe per il training e l'holdout set, per il dataset originale e per i dataset bilanciati con SONCA per i diversi valori di m con distribuzione di probabilità gaussiana

		Training set		Holdout set	
m	$m/\#1_{originale}$	#1	#0	#1	#0
-	-	1374	17817	1373	17937
3500	2.5	3482	3592	1373	17937
5500	4.0	5412	5478	1373	17937
7500	5.5	7477	7593	1373	17937
9500	6.9	9406	9511	1373	17937

Dalla matrice di confusione è stato possibile ricavare le diverse metriche di performance che misurano l'accuratezza dei classificatori con e senza il ricampionamento.

In Tabella 31 sono riportate le misure di performance ottenute dalla stima del CART, ricampionando il set training con SONCA e con distribuzione di probabilità triangolare. In Figura 15 e Figura 16 sono raffigurate alcune delle principali misure di performance.

I risultati evidenziano che per il CART senza il ricampionamento, FN_{rate} è pari a 0.543, ciò significa che all'incirca il 54% dei casi positivi sono classificati come casi negativi. Si può però osservare come FN_{rate} si riduce bilanciando il data set con l'utilizzo di SONCA e il valore FN_{rate} è all'incirca pari al 7-10%, il che equivale a dire a una riduzione dei falsi negativi di circa il 70-85%. Contemporaneamente si ha anche un aumento dei TP_{rate} che dal 46% passa all'incirca al 90%. Tali miglioramenti comportano, però, un peggioramento del TN_{rate} che tende a diminuire e di FP_{rate} che tende ad aumentare.

Tabella 31: Confronto delle misure di performance per il CART, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità triangolare per i diversi valori di m

m	$m/\#1_{originale}$	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
-	0	0.543	0.008	0.992	0.457	0.820	0.587	0.673	0.046	0.954	0.927
3500	2.5	0.143	0.114	0.886	0.857	0.366	0.513	0.871	0.116	0.884	0.931
5500	4.0	0.079	0.147	0.853	0.921	0.324	0.480	0.886	0.142	0.858	0.943
7500	5.5	0.103	0.108	0.892	0.897	0.388	0.541	0.894	0.108	0.892	0.933
9500	6.9	0.074	0.099	0.901	0.926	0.417	0.575	0.913	0.098	0.902	0.943

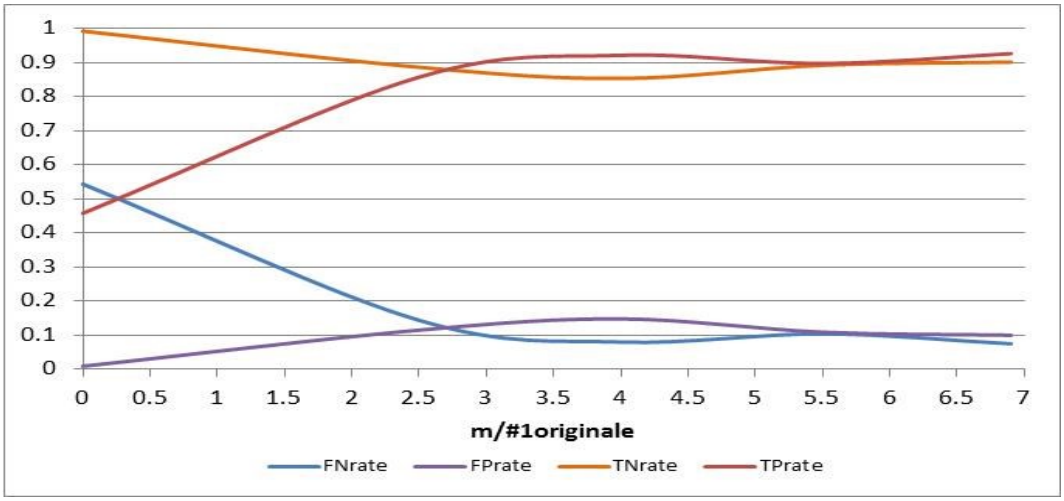

Figura 15 - FN rate, FPrate, TNrate, TPrate per il CART, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità triangolare per i diversi valori di m

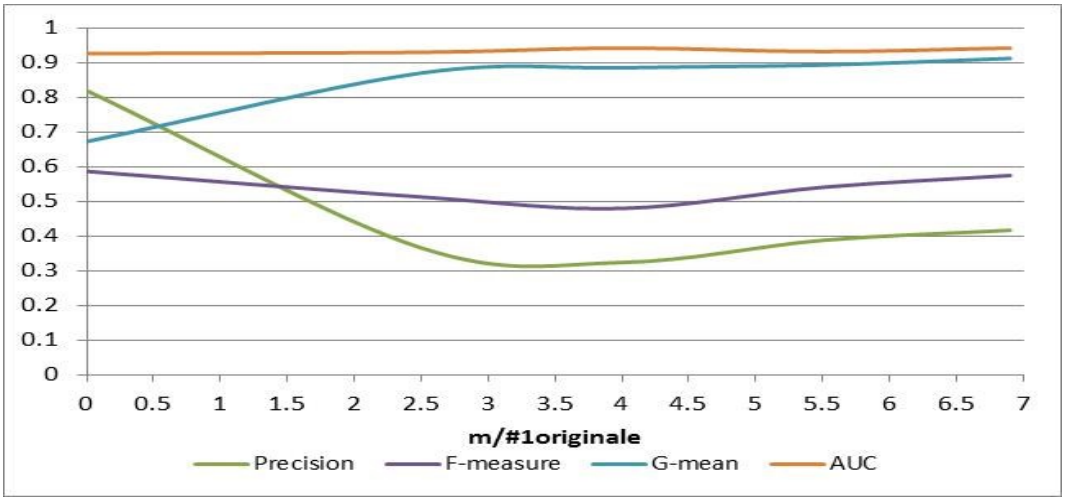

Figura 16 – Precision, F-measure, G-mean, AUC per il CART, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità triangolare per i diversi valori di m

Per avere una visione più completa si possono usare misure come la Precision e la Recall (spesso indicata anche come ACC+ oppure TP_{rate}) (Figura 16). In tale caso, si osserva che la Recall aumenta mentre la Precision diminuisce. Questo risultato è concorde con la letteratura, come discusso da Chawala (2010), aumentando i veri positivi per la classe di minoranza, aumentano anche il numero di falsi positivi, riducendo così la Precision. Pertanto, per ovviare a tali problemi è possibile osservare i valori di F-measure e G-mean. I valori di F-measure, che combinano la Recall e la Precision, sono più o meno costanti nella stima del CART, mentre i valori di G-mean, aumentano. Un altro strumento per valutare l'accuratezza di uno stimatore è fornito dalla curva ROC e dall'area sottesa alla curva ROC (AUC), anche tale parametro resta più o meno costante nei diversi scenari di studio.

I risultati evidenziano che per il dataset sbilanciato la stima del CART ha prodotto un modello poco accurato con valori di FN_{rate}=0.54 e TP_{rate}=0.46, F-measure=0.59 e G-mean=0.67, ma bilanciando i dataset con SONCA prima della stima dei modelli si ha una riduzione di casi positivi erroneamente classificati come appartenenti alla classe negativa e un aumento di casi positivi correttamente classificati come appartenenti alla classe positiva, anche con un miglioramento

delle misure di performance. Al variare di m non è stata riscontrata una forte variazione delle misure di performance, comunque il modello più accurato è stato ottenuto per m=5500, in altre parole con un rapporto di ricampionamento pari a quattro. La percentuale di casi positivi erroneamente classificati come appartenenti alla classe negativa (FN_{rate}) è passata dallo 0.54 allo 0.08, quindi si è avuta una riduzione dell'85%, mentre i valori di G-mean sono passati da 0.67 a 0.89.

In Tabella 32, Figura 17 e Figura 18 sono riportate le misure di performance ottenute dalla stima del CART con distribuzione di gaussiana.

Tabella 32: Confronto delle misure di performance per il CART, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità gaussiana per i diversi valori di m

m	$m/\#1_{originale}$	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
-	0	0.543	0.008	0.992	0.457	0.820	0.587	0.673	0.046	0.954	0.927
3500	2.5	0.092	0.114	0.886	0.908	0.379	0.535	0.897	0.112	0.888	0.956
5500	4.0	0.100	0.133	0.867	0.900	0.341	0.495	0.883	0.131	0.869	0.936
7500	5.5	0.108	0.107	0.893	0.892	0.390	0.543	0.892	0.107	0.893	0.932
9500	6.9	0.114	0.106	0.894	0.886	0.389	0.541	0.890	0.107	0.893	0.935

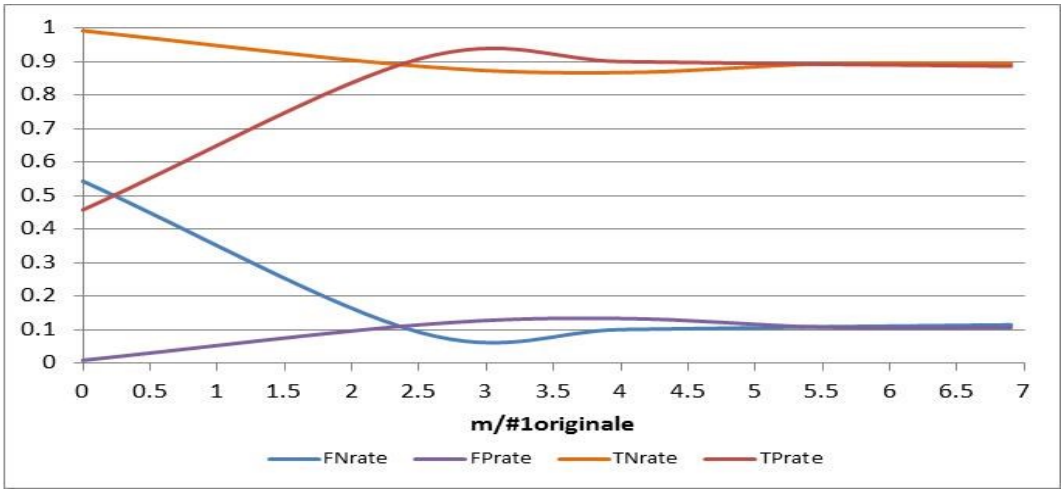

Figura 17 - FN rate, FPrate, TNrate, TPrate per il CART, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità gaussiana per i diversi valori di m

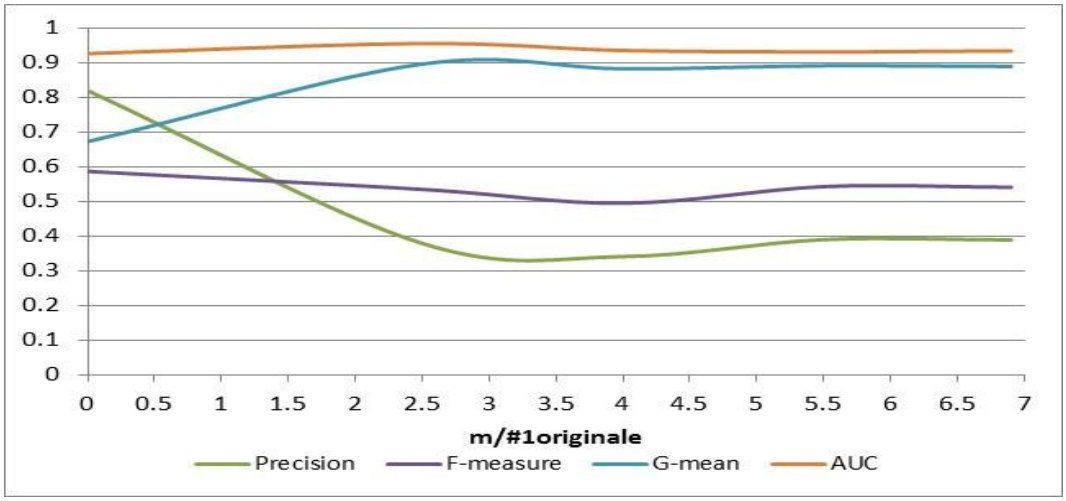

Figura 18 – Precision, F-measure, G-mean, AUC per il CART, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità gaussiana per i diversi valori di m

I risultati sono simili a quelli ottenuti per la distribuzione di tipo triangolare. I modelli stimati con il dataset bilanciato con SONCA utilizzando una distribuzione di probabilità gaussiana presentano dei valori delle misure di performance migliori. All'aumentare di m le diverse misure di performance restano pressoché costanti. I valori migliori sono ottenuti per m=3500, ovvero con un rapporto di ricampionamento pari a 2.5. La percentuale di casi positivi erroneamente classificati come appartenenti alla classe negativa (FN_{rate}) è passata da 0.54 allo 0.09, quindi si è avuta una riduzione dell'83%, mentre i valori di G-mean sono passati da 0.67 a 0.90.

In Tabella 33, Figura 19 e Figura 20 sono riportate le misure di performance ottenute dalla stima del Logit con distribuzione di probabilità triangolare.

I risultati confermano quelli visti con la stima dei modelli del CART. I risultati evidenziano che per il Logit senza il ricampionamento FN_{rate} è pari a 0.99, ciò significa che all'incirca il 99% dei casi positivi sono classificati come casi negativi. Si può però osservare come FN_{rate} si riduce ricampionando il datset con SONCA, il valore FN_{rate} per i diversi valori di m sono all'incirca pari al 20%.

Contemporaneamente si ha anche un aumento dei TP_{rate} che dal 46% passa all'incirca al 70-80%.

Tabella 33: Confronto delle misure di performance per il Logit, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità triangolare per i diversi valori di m

m	$m/\#1_{originale}$	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
-	0	0.985	0.286	0.714	0.015	0.401	0.028	0.103	0.936	0.064	0.895
3500	2.5	0.192	0.185	0.815	0.808	0.251	0.383	0.811	0.185	0.815	0.894
5500	4.0	0.290	0.130	0.870	0.710	0.296	0.417	0.786	0.141	0.859	0.890
7500	5.5	0.193	0.166	0.834	0.807	0.272	0.406	0.820	0.168	0.832	0.898
9500	6.9	0.183	0.179	0.821	0.817	0.259	0.394	0.819	0.179	0.821	0.897

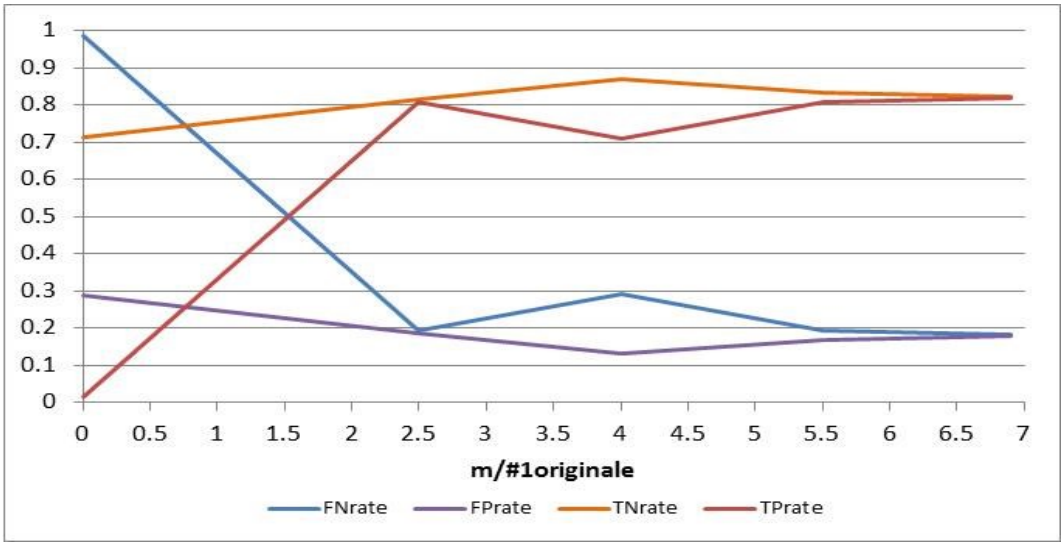

Figura 19 - FNrate, FPrate, TNrate, TPrate per il Logit, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità triangolare per i diversi valori di m

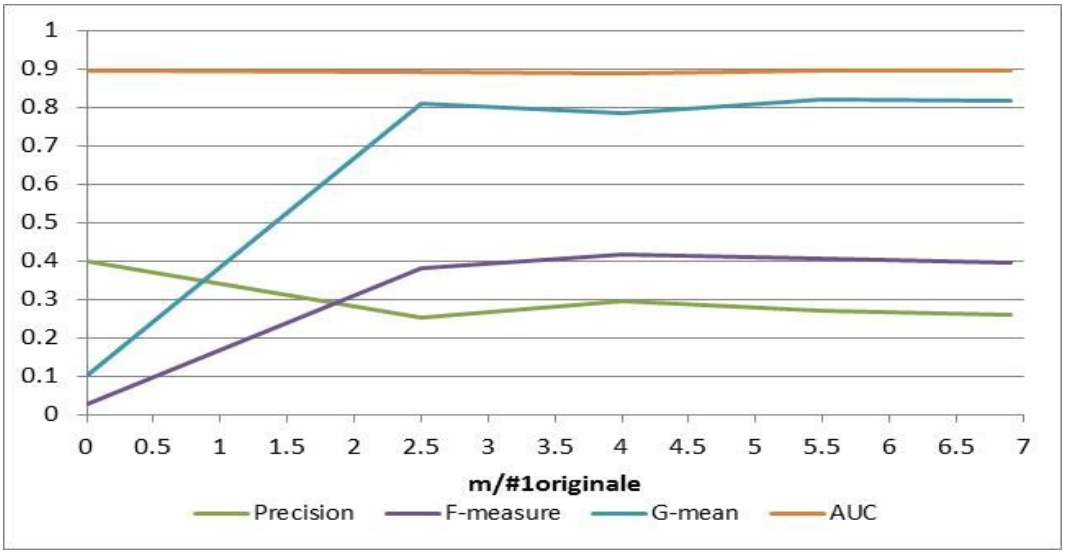

Figura 20 – Precision, F-measure, G-mean, AUC per il Logit, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità triangolare per i diversi valori di m

I valori di F-measure e G-mean aumentano, mentre la Precision diminuisce.

In Tabella 33, Figura 21 e Figura 22 sono riportate le misure di performance ottenute dalla stima del Logit con distribuzione di probabilità gaussiana.

Tabella 34: Confronto delle misure di performance per il Logit, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità gaussiana per i diversi valori di m

m	$m/\#1_{originale}$	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
-	0	0.985	0.286	0.714	0.015	0.401	0.028	0.103	0.936	0.064	0.895
3500	2.5	0.209	0.189	0.811	0.791	0.243	0.371	0.801	0.190	0.810	0.890
5500	4	0.100	0.133	0.867	0.900	0.341	0.495	0.883	0.131	0.869	0.894
7500	5.5	0.179	0.170	0.830	0.821	0.270	0.406	0.825	0.171	0.829	0.899
9500	6.9	0.183	0.170	0.830	0.817	0.269	0.405	0.823	0.171	0.829	0.898

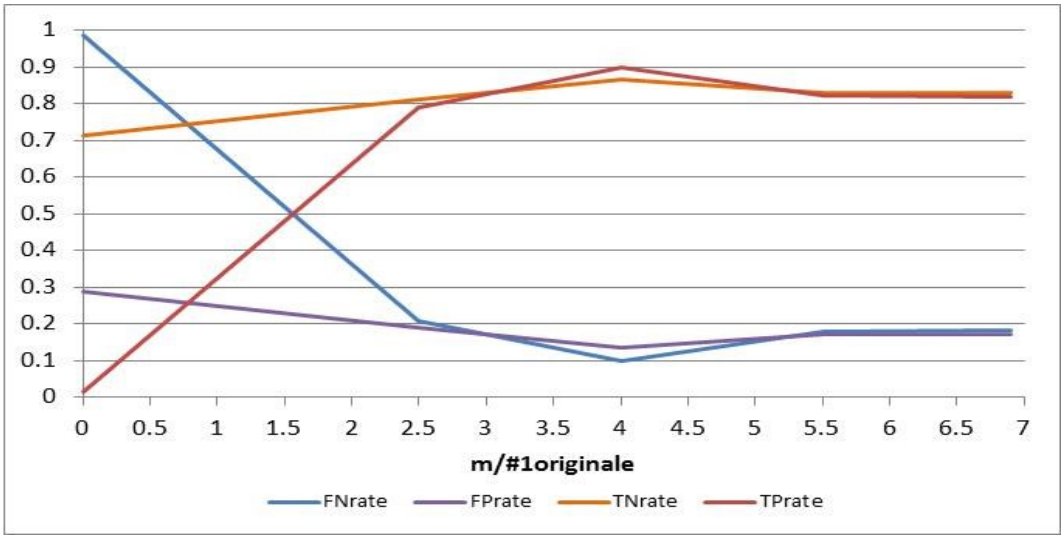

Figura 21 - FNrate, FPrate, TNrate, TPrate per il Logit, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità gaussiana per i diversi valori di m

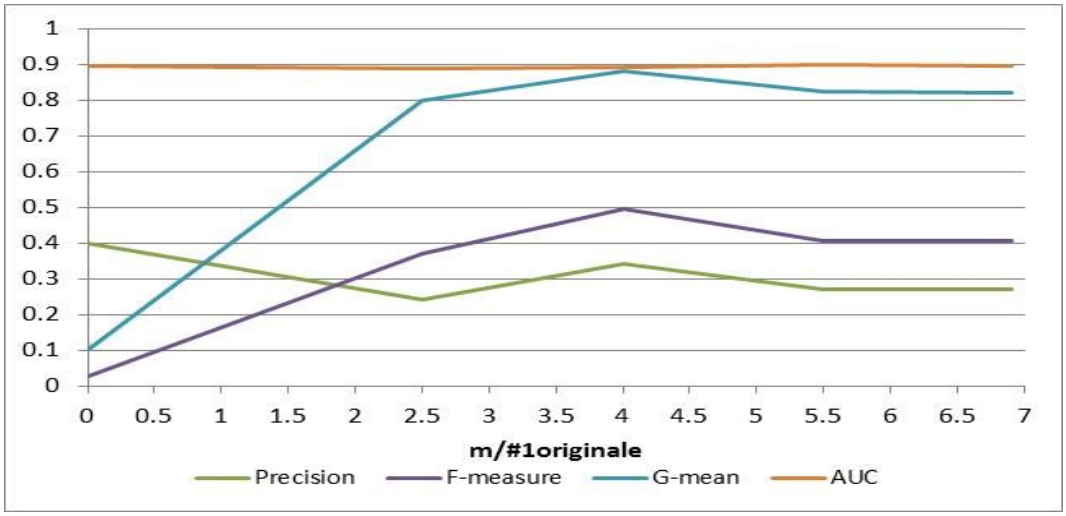

Figura 22 – Precision, F-measure, G-mean, AUC per il Logit, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità gaussiana per i diversi valori di m

I risultati confermano quelli visti finora. I modelli stimati con il dataset bilanciato, con SONCA e con una distribuzione di probabilità gaussiana, sono molto più accurati dei modelli dei modelli stimati con il dataset sbilanciato. La percentuale di casi positivi erroneamente classificati come appartenenti alla classe negativa (FN_{rate}) è passata da 0.99 allo 0.10, mentre i valori di G-mean sono passati da 0.10 a 0.88.

4.2.2. Adult dataset

Il dataset Adult dataset è composto da 32561 osservazioni, con predittori di natura sia numerica sia categorica. La variabile di risposta è costituita da 7841 osservazioni appartenente alla classe minoritaria, pari al 24% del totale, e 27720 osservazioni appartenenti alla classe maggioritaria, pari al 76% del totale.

Per tale dataset sono stati stimati l'albero di regressione e il logit nei seguenti casi:

- Dataset originale senza ricampionamento;
- Dataset bilanciato con SONCA usando la distribuzione di probabilità triangolare e con:
  - m=4000;
  - m=8000;
  - m=12000.
- Dataset bilanciato con SONCA usando la distribuzione di probabilità gaussiana e con:
  - m=4000;
  - m=8000;
  - m=12000.

In Tabella 35 e Tabella 36 sono riportate le distribuzioni di frequenza per il training e l'holdout set, sia per il dataset originale sia per i dataset bilanciati con SONCA.

Tabella 35: Distribuzioni di classe per il training e l'holdout set, per il dataset originale e per i dataset bilanciati con SONCA per i diversi valori di m con distribuzione di probabilità triangolare

		Training set		Holdout set	
m	$m/\#1_{originale}$	#1	#0	#1	#0
-	-	3914	12342	3927	12378
4000	1.0	4003	4025	3927	12378
8000	2.0	8005	8078	3927	12378
12000	3.1	12087	11980	3927	12378

Tabella 36: Distribuzioni di classe per il training e l'holdout set, per il dataset originale e per i dataset bilanciati con SONCA per i diversi valori di m con distribuzione di probabilità gaussiana

		Training set		Holdout set	
m	$m/\#1_{originale}$	#1	#0	#1	#0
-	-	3914	12342	3927	12378
4000	1.0	4017	3963	3927	12378
8000	2.0	7957	8044	3927	12378
12000	3.1	12078	11989	3927	12378

Dalla matrice di confusione è stato possibile ricavare le diverse metriche di performance che misurano le prestazioni degli stimatori con e senza il ricampionamento.

In Tabella 37, Figura 23 e Figura 53 sono riportate le misure di performance ottenute dalla stima del CART, ricampionando il set training con SONCA e distribuzione di probabilità triangolare.

I risultati evidenziano che per il CART senza il ricampionamento FN_{rate} è pari a 0.519, ciò significa che all'incirca il 52% dei casi positivi sono classificati come casi negativi. Si può però osservare come FN_{rate} si riduce con l'utilizzo di SONCA, il valore FN_{rate} è all'incirca pari al 15%. Contemporaneamente si ha anche un aumento dei TP_{rate} che dal 48% passa all'incirca al 80%.

Pertanto, tali miglioramenti comportano, però, un peggioramento del TN_{rate} che tende a diminuire e di FP_{rate} che tende ad aumentare. I valori di F-measure hanno un piccolo incremento, mentre i valori di G-mean aumentano, passando da 0.60 all'incirca a 0.80. I valori di AUC restano costanti.

Tabella 37: Confronto delle misure di performance per il CART, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità triangolare per i diversi valori di m

m	$m/\#1_{originale}$	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
-	-	0.519	0.035	0.965	0.481	0.814	0.605	0.681	0.151	0.849	0.893
4000	1.0	0.152	0.231	0.769	0.848	0.537	0.658	0.807	0.212	0.788	0.895
8000	2.0	0.214	0.177	0.823	0.786	0.585	0.671	0.804	0.186	0.814	0.887
12000	3.1	0.160	0.217	0.783	0.840	0.551	0.665	0.811	0.204	0.796	0.893

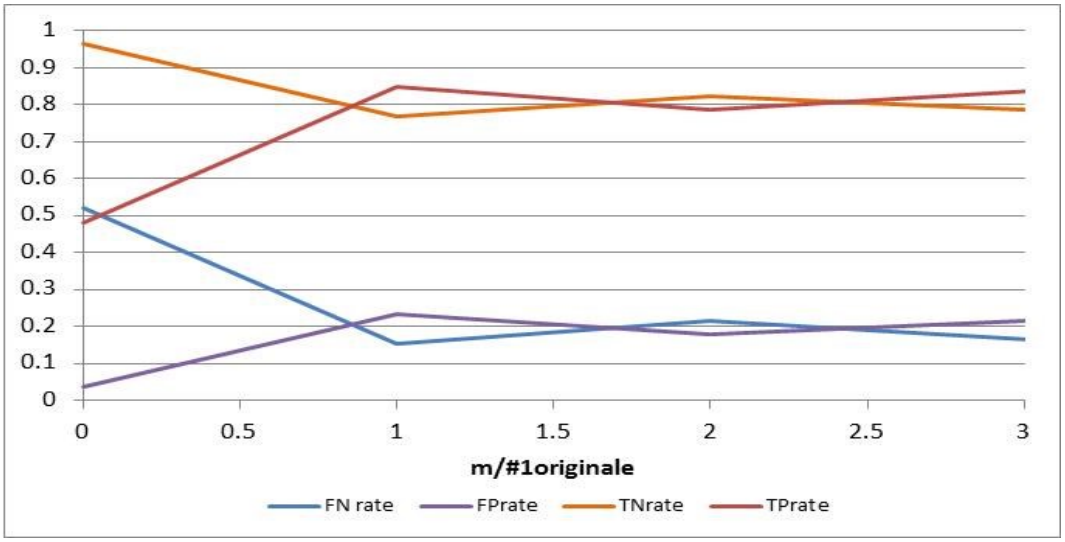

Figura 23 - FN rate, FPrate, TNrate, TPrate per il CART, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità triangolare per i diversi valori di m

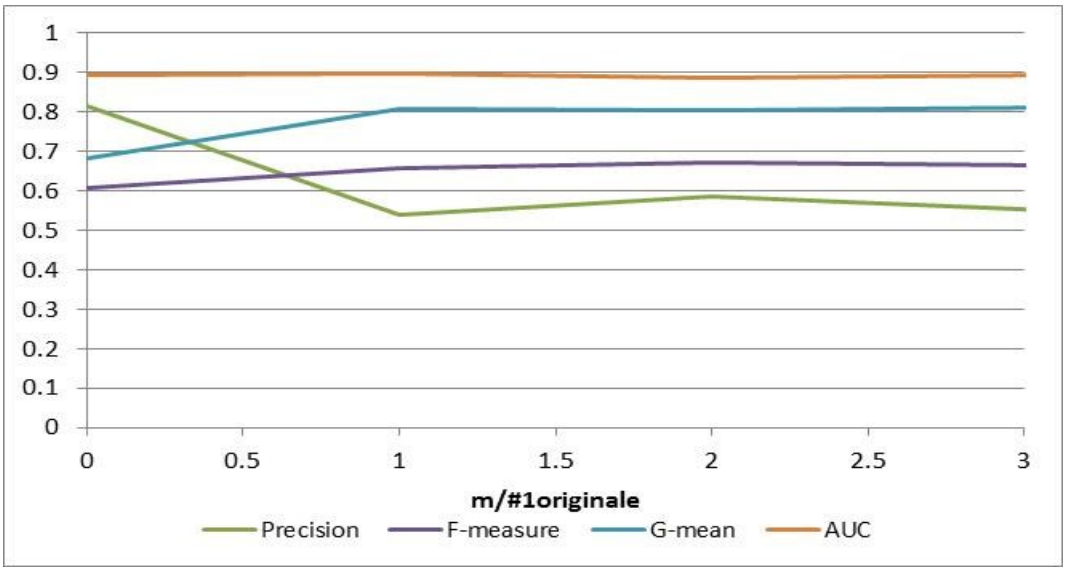

Figura 24 – Precision, F-measure, G-mean, AUC per il CART, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità triangolare per i diversi valori di m

Inoltre, i risultati evidenziano che, anche, con un rapporto di bilanciamento pari a 1 si ottengono dei buoni risultati. In altre parole anche senza incrementare il numero di osservazioni, ma semplicemente ottenendo un nuovo dataset sintetico che abbia le caratteristiche del precedente dataset è possibile stimare dei modelli aventi buone misure di performance.

In Tabella 38, Figura 25 e Figura 26 sono riportate le misure di performance ottenute dalla stima del CART con distribuzione di probabilità gaussiana.

I risultati evidenziano che per il CART senza il ricampionamento FN_{rate} è pari a 0.52, e si riduce con l'utilizzo di SONCA passando a FN_{rate}= 0.14. Contemporaneamente si ha anche un aumento dei TP_{rate} che dal 48% passa all'incirca al 86%. Pertanto, tali miglioramenti comportano, un peggioramento del TN_{rate} che tende a diminuire e di FP_{rate} che tende ad aumentare.

Tabella 38: Confronto delle misure di performance per il CART, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità gaussiana per i diversi valori di m

m	m/#1$_{originale}$	FN $_{rate}$	FP$_{rate}$	TN$_{rate}$	TP$_{rate}$	Precision	F-measure	G-mean	Err	Acc	AUC
-	-	0.519	0.035	0.965	0.481	0.814	0.605	0.681	0.151	0.849	0.893
4000	1.0	0.143	0.237	0.763	0.857	0.535	0.659	0.809	0.214	0.786	0.891
8000	2.0	0.142	0.232	0.768	0.858	0.540	0.663	0.812	0.211	0.789	0.889
12000	3.1	0.162	0.220	0.780	0.838	0.547	0.662	0.808	0.206	0.794	0.893

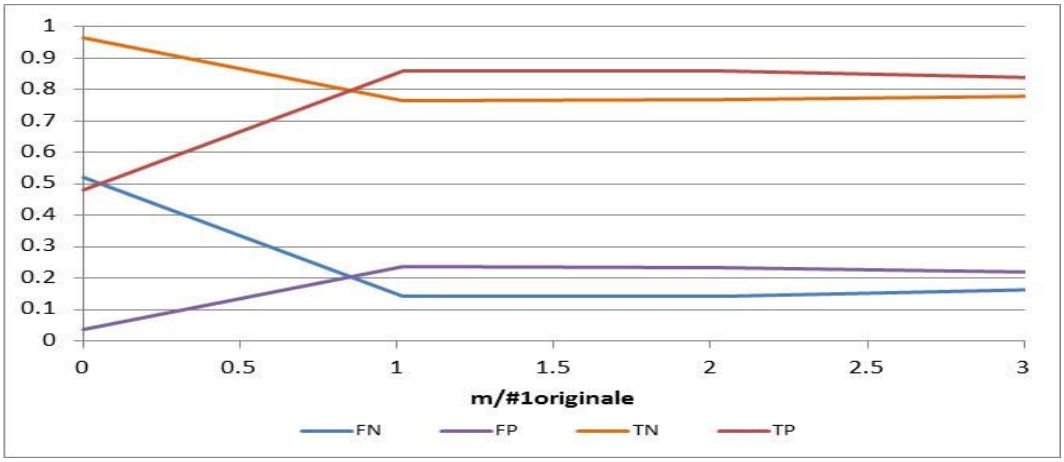

Figura 25 - FN rate, FPrate, TNrate, TPrate per il CART, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità gaussiana per i diversi valori di m

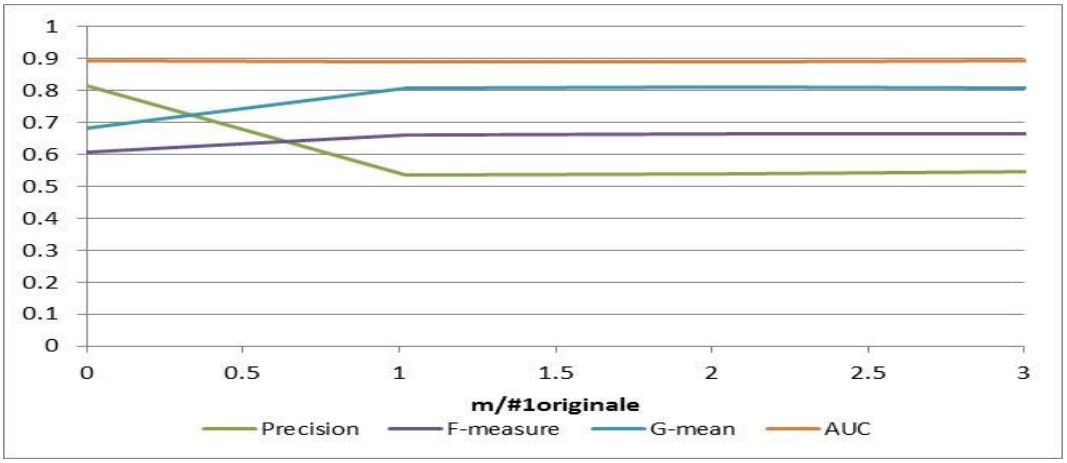

Figura 26 – Precision, F-measure, G-mean, AUC per il CART, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità gaussiana per i diversi valori di m

I valori di F-measure hanno un piccolo incremento, lo stesso accade per i valori di G-mean che passano da 0.60 all'incirca a 0.81. I valori di AUC restano costanti all'incirca pari a 0.89.

In Tabella 39, Figura 27 e Figura 28 sono riportate le misure di performance ottenute dalla stima del Logit con distribuzione di probabilità triangolare.

I risultati evidenziano che per il dataset sbilanciato la stima del Logit ha prodotto un modello poco accurato con valori di FN_{rate}=0.61 e TP_{rate}=0.39, F-measure=0.50 e G-mean=0.60, ma bilanciando il dataset con SONCA prima della stima dei modelli si ha un miglioramento dei modelli. In particolare si osserva una riduzione di casi positivi erroneamente classificati come appartenenti alla classe negativa e un aumento di casi positivi correttamente classificati come appartenenti alla classe positiva, FN_{rate} passa all'incirca a valori del 10-20% e un aumento di TPrate che passa a valori del 70-80%. Inoltre, i valori di F-measure, G-mean e AUC aumentano in maniera sostanziale.

Tabella 39: Confronto delle misure di performance per il Logit, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità triangolare per i diversi valori di m

m	$m/\#1_{originale}$	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
-	-	0.612	0.050	0.950	0.388	0.712	0.503	0.607	0.185	0.815	0.833
4000	1.0	0.231	0.226	0.774	0.769	0.519	0.620	0.772	0.227	0.773	0.856
8000	2.0	0.269	0.240	0.760	0.731	0.491	0.587	0.745	0.247	0.753	0.830
12000	3.1	0.162	0.220	0.780	0.838	0.547	0.662	0.808	0.206	0.794	0.893

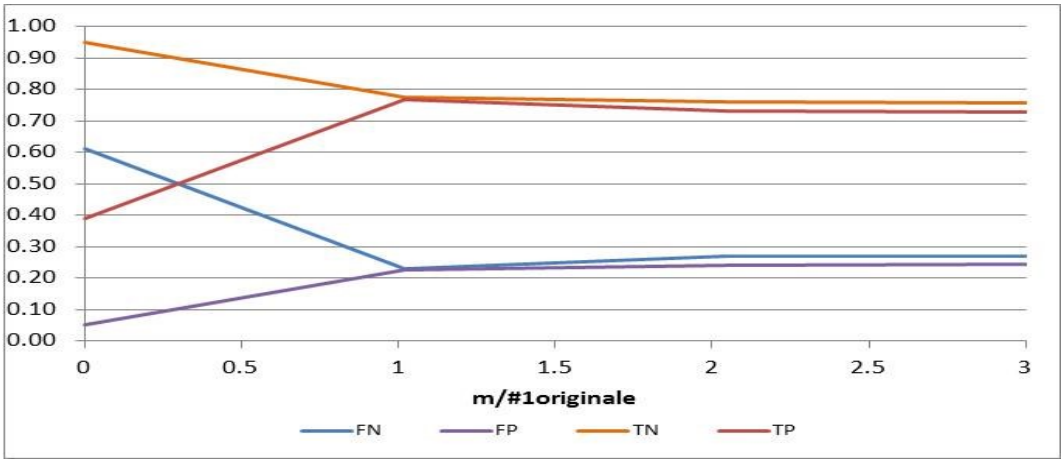

Figura 27 - FNrate, FPrate, TNrate, TPrate per il Logit, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità triangolare per i diversi valori di m

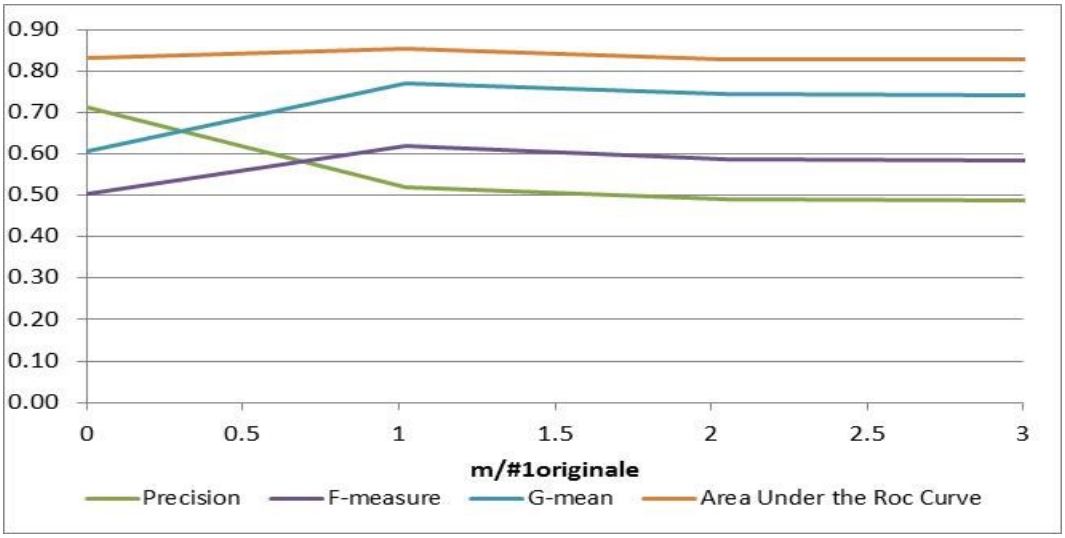

Figura 28 – Precision, F-measure, G-mean, AUC per il Logit, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità triangolare per i diversi valori di m

In Tabella 40, Figura 29 e Figura 30 sono riportate le misure di performance ottenute dalla stima del Logit con distribuzione di probabilità gaussiana.

Tabella 40: Confronto delle misure di performance per il Logit, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità gaussiana per i diversi valori di m

m	$m/\#1_{originale}$	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
-	-	0.612	0.050	0.950	0.388	0.712	0.503	0.607	0.185	0.815	0.833
4000	1.0	0.264	0.240	0.760	0.736	0.493	0.590	0.748	0.246	0.754	0.841
8000	2.0	0.261	0.242	0.758	0.739	0.493	0.591	0.748	0.246	0.754	0.832
12000	3.1										

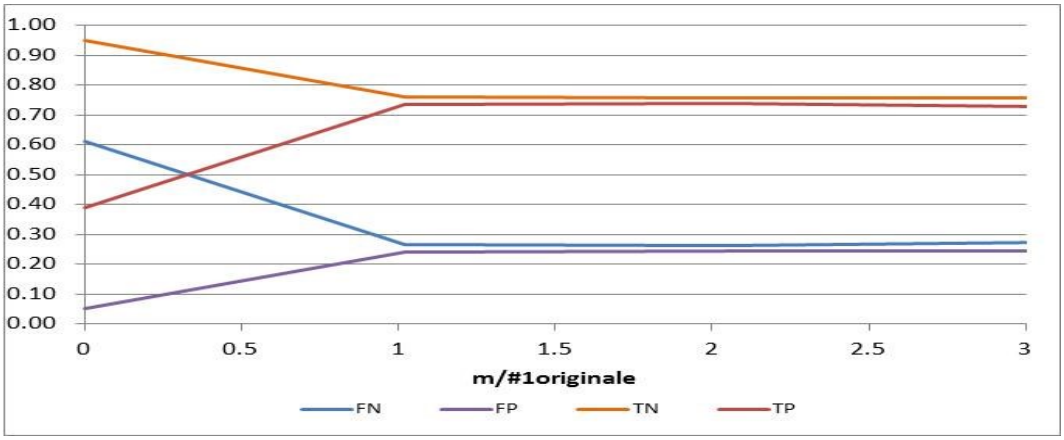

Figura 29 - FNrate, FPrate, TNrate, TPrate per il Logit, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità gaussiana per i diversi valori di m

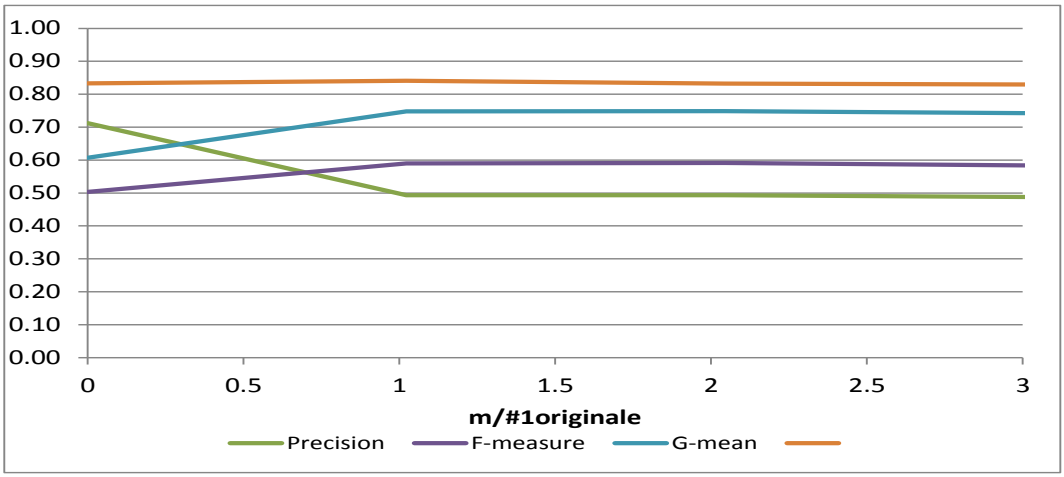

Figura 30 – Precision, F-measure, G-mean, AUC per il Logit, tra il dataset originale e il dataset ricampionato con distribuzione di probabilità gaussiana per i diversi valori di m

I risultati evidenziano che per il dataset sbilanciato la stima del Logit ha prodotto un modello poco accurato con valori di FN_{rate}=0.612 e TP_{rate}=0.388, F-measure=0.503 e G-mean=0.607, ma bilanciando il dataset con SONCA prima della stima dei modelli si ha un miglioramento dei modelli. In particolare si osserva una riduzione di casi positivi erroneamente classificati come appartenenti alla classe negativa e un aumento di casi positivi correttamente classificati come appartenenti alla classe positiva, FN_{rate} passa all'incirca a valori del 20% e un aumento di TP_{rate} che passa a valori del 75-80%. Inoltre, i valori di F-measure, G-mean e AUC aumentano in maniera sostanziale.

4.3. Comparazione di SONCA con altri studi

Nel presente paragrafo sono state confrontate le misure di performance ottenute dalla stima del CART e del Logit bilanciando i dataset con SONCA (utilizzando entrambe le distribuzioni di probabilità) con altri studi.

Per ogni dataset sono stati stimati il Cart e il Logit nei seguenti casi:

- Dataset originale;
- Ricampionamento con SONCA;
- Ricampionamento con SMOTE;
- Ricampionamento con ROSE.

In tale analisi, per confrontare SONCA con SMOTE e ROSE, è stato necessario usare dataset (Tabella 41) aventi predittori solo di natura numerica.

Tabella 41: Dataset utilizzati per il confronto con altri studi

		Un-success		Success		
Dataset	**# Var.**	**N**	**%**	**N**	**%**	**Tot**
Cover type	12	35754	92.87	2747	7.13	38501
Pima Indian Diabetes	8	500	65.10	268	34.90	768
Glass	9	197	92.06	17	7.94	214

Nel seguito sono riportati i risultati di sintesi delle diverse analisi, mentre in Appendice 2 sono riportate tutti i risultati delle analisi

4.3.1. Cover type

In Tabella 42 e Tabella 43 sono riportate le misure di performance ottenute dalla stima del CART e del Logit nei diversi casi studio.

Tabella 42: Confronto delle misure di performance (CART)

Algoritmi	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
Originale	0.543	0.008	0.992	0.457	0.820	0.587	0.673	0.046	0.954	0.927
SONCA (Triangolare)	0.074	0.099	0.901	0.926	0.417	0.575	0.913	0.098	0.902	0.943
SONCA (Gaussiana)	0.114	0.106	0.894	0.886	0.389	0.541	0.890	0.107	0.893	0.935
SMOTE	0.214	0.177	0.823	0.786	0.585	0.671	0.804	0.186	0.814	0.944
ROSE	0.078	0.184	0.816	0.922	0.277	0.426	0.867	0.177	0.823	0.922

Tabella 43: Confronto delle misure di performance (Logit)

Algoritmi	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
Originale	0.985	0.286	0.714	0.015	0.401	0.028	0.103	0.936	0.064	0.895
SONCA (Triangolare)	0.183	0.179	0.821	0.817	0.259	0.394	0.819	0.179	0.821	0.897
SONCA (Gaussiana)	0.183	0.170	0.830	0.817	0.269	0.405	0.823	0.171	0.829	0.898
SMOTE	0.185	0.213	0.787	0.815	0.788	0.801	0.801	0.199	0.801	0.901
ROSE	0.349	0.065	0.935	0.651	0.758	0.700	0.780	0.133	0.867	0.900

I risultati evidenziano che usando il dataset sbilanciato, si hanno dei modelli poco accurati: per il CART, FN_{rate} è pari a 0.54, e per il logit FN_{rate} è pari a 0.99. Utilizzando uno qualunque degli algoritmi si osserva una riduzione di FN_{rate}, anche le altre misure di performance come G-mean e AUC aumentano.

L'agoritmo SONCA produce migliori risultati per il CART utilizzando SONCA con distribuzione di probabilità triangolare FN_{rate} =0.074, G-mean=0.913 e AUC=0.94. L'agoritmo SONCA per il Logit produce risultati piuttosto simili confrontando le due distribuzioni di probabilità.

4.3.2. Glass

In Tabella 44 e Tabella 45 sono riportate le misure di performance ottenute dalla stima del CART e del Logit nei diversi casi studio.

Tabella 44: Confronto delle misure di performance (CART)

Algoritmo	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
Originale	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.077	0.923	0.000
SONCA (Triangolare)	0.375	0.125	0.875	0.625	0.294	0.400	0.740	0.144	0.856	0.851
SONCA (Gaussiana)	0.500	0.219	0.781	0.500	0.160	0.242	0.625	0.240	0.760	0.654
SMOTE	0.625	0.146	0.854	0.375	0.176	0.240	0.566	0.183	0.817	0.639
ROSE	0.125	0.510	0.490	0.875	0.125	0.219	0.655	0.481	0.519	0.708

Tabella 45: Confronto delle misure di performance (Logit)

Algoritmo	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
Originale	1.000	0.010	0.990	0.000	0.000	0.000	0.000	0.087	0.913	0.000
SONCA (Triangolare)	0.250	0.281	0.719	0.750	0.182	0.293	0.734	0.279	0.721	0.802
SONCA (Gaussiana)	0.250	0.250	0.750	0.750	0.200	0.316	0.750	0.250	0.750	0.750
SMOTE	0.500	0.292	0.708	0.500	0.125	0.200	0.595	0.308	0.692	0.656
ROSE	0.250	0.563	0.438	0.750	0.100	0.176	0.573	0.538	0.462	0.660

La variabile di risposta di glass dataset presenta uno sbilanciamento assoluto e i modelli di stima, CART e il Logit, sono inefficaci nel predire la classa rara. In tal caso, accade che c'è un'alta accuratezza per la classe prevalente e una bassa accuratezza per la classe rara: per il CART TN_{rate}=1, FN_{rate}=1, e TP_{rate}=0, mentre per il Logit TN_{rate}=0.990, FN_{rate}=1, e TP_{rate}=0. Utilizzando degli algoritmi di ricampionamento si ha un miglioramento delle misure di performance:

- Per il CART le migliori prestazioni sono ottenute dall'algoritmo SONCA con distribuzione triangolare (FN_{rate}=0.375, TP_{rate}=0.625, G-mean=0.740 e AUC=0.851) e per ROSE (FN_{rate}=0.125, TP_{rate}=0.875, G-mean=0.655 e AUC=0.708);
- Per il Logit le migliori prestazioni sono ottenute dall'algoritmo SONCA con distribuzione gaussiana (FN_{rate}=0.250, TP_{rate}=0.750, G-mean=0.750 e

AUC=0.750) e per ROSE (FN_{rate}=0.250, TP_{rate}=0.750, G-mean=0.573 e AUC=0.660).

4.3.3. Pima Indian Dataset

In Tabella 46 e Tabella 47 sono riportate le misure di performance ottenute dalla stima del CART e del Logit nei diversi casi studio.

Tabella 46: Confronto delle misure di performance (CART)

Algoritmo	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
Originale	0.336	0.203	0.797	0.664	0.610	0.636	0.727	0.246	0.754	0.771
SONCA (Triangolare)	0.216	0.330	0.670	0.784	0.533	0.634	0.725	0.293	0.707	0.727
SONCA (Gaussiana)	0.384	0.318	0.682	0.616	0.481	0.540	0.648	0.339	0.661	0.703
SMOTE	0.136	0.448	0.552	0.864	0.480	0.617	0.690	0.347	0.653	0.703
ROSE	0.696	0.107	0.893	0.304	0.576	0.398	0.521	0.298	0.702	0.754

Tabella 47: Confronto delle misure di performance (Logit)

Algoritmo	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
Originale	0.424	0.123	0.877	0.576	0.692	0.629	0.711	0.220	0.780	0.854
SONCA (Triangolare)	0.184	0.372	0.628	0.816	0.513	0.630	0.716	0.311	0.689	0.772
SONCA (Gaussiana)	0.216	0.249	0.751	0.784	0.601	0.681	0.767	0.238	0.762	0.855
SMOTE	0.032	0.579	0.421	0.968	0.445	0.610	0.638	0.402	0.598	0.695
ROSE	0.336	0.169	0.831	0.664	0.654	0.659	0.743	0.223	0.777	0.854

I risultati ottenuti sono differenti da quelli visti finora. Sia per il Logit che per il CART. Osservando le misure di performance ottenute usando il dataset originale con il dataset ricampionato, sia per la stima del CART e del Logit, si notano già dei buoni valori delle diverse misure di performance. I valori di FN_{rate} e FP_{rate} diminuiscono utilizzando gli algoritmi di ricampionamento. I valori TP_{rate} aumentano mentre i valori di TN_{rate} diminuiscono. Gli altri valori come F-measure, G-mean e AUC aumentano lievemente. Il motivo di ciò, è dovuto al fatto che il dataset non è molto sbilanciato, pertanto il dataset non ha bisogno di essere ricampionato.

4.4. Sintesi dei risultati

I risultati evidenziano che usando un dataset con variabile di risposta molto sbilanciato, si hanno dei modelli di stima poco accurati, con un'alta percentuale di casi negativi erroneamente classificati come appartenenti alla classe positiva, FN_{rate}, e una bassa percentuale di casi positivi correttamente classificati come appartenenti alla classe positiva, TP_{rate}, insieme a bassi valori delle misure di performance come F-measure, G-mean e AUC.

Utilizzando gli algoritmi di bilanciamento, alla presenza di dataset con variabile di risposta estremamente sbilanciata, in particolare se lo squilibrio è assoluto, i modelli di stima sono più accurati. Le diverse misure di performance migliorano: FN_{rate} si riduce, mentre TP_{rate} aumenta insieme a F-measure, G-mean e AUC. Contemporaneamente al aumento dei TP_{rate} si ha anche un aumento dei falsi positivi, riducendo così la Precision.

Dal confronto dell'algoritmo di SONCA con SMOTE e ROSE, è stato visto che i migliori valori delle misure di performance sono ottenuti utilizzando SONCA.

SONCA è efficace sia per dataset con predittori numerici, sia con predittori di natura sia numerica sia categorica.

Confrontando le misure di performance ottenute dai modelli di stima usando SONCA con le due differenti distribuzioni di probabilità non si riscontrano grandi variazioni.

Al variare di m non è stata riscontrata una forte variazione delle misure di performance, così da poter utilizzare anche rapporti di bilanciamento bassi senza aggiungere un ulteriore costo computazionale.

Capitolo 5 - SONCA e gli incidenti stradali

Gli incidenti stradali rappresentano un problema di grave rilevanza per l'alto numero di morti e di feriti. Ogni anno si contano almeno 1,3 milioni di morti e 50 milioni di feriti a causa degli incidenti stradali. L'Organizzazione Mondiale della Sanità prevede che entro il 2020 le vittime della strada aumenteranno, dal livello attuale di 1.3 milioni, a oltre 1.9 milioni. In Europa gli incidenti stradali sono una delle prime cause di morte, con più di 120000 morti e più di 2.4 milioni feriti ogni anno. In Italia, la mortalità da incidente stradale è fra le più elevate rispetto al resto dell'Europa. Ogni anno in Italia circa 5000 persone perdono la vita a causa di incidenti stradali, 120000 sono feriti gravi di cui 15000 restano invalidi gravi. Tali numeri danno un quadro di come la questione della sicurezza stradale sia un argomento di enorme importanza per i dipartimenti di Prevenzione e i sistemi sanitari di tutti i Paesi.

La politica per la sicurezza stradale della Commissione Europea si "*prefigge di aumentare il livello della sicurezza stradale e garantire una mobilità sicura e rispettosa dell'ambiente per i cittadini di tutta Europa*". Per riuscire in tale intento, la Commissione Europea ha proposto di ridurre entro il 2020 del 50% il numero delle vittime sulle strade europee rispetto al 2010 (WHO, 2004; COM, 2010a; COM, 2010b; ANCI-UPI, 2010).

Il successo dei programmi di miglioramento della sicurezza stradale dipende strettamente dalla disponibilità di metodi che forniscono stime affidabili del livello di sicurezza associato con le strade esistenti. A tal fine bisogna analizzare i fattori che determinano gli incidenti stradali, con particolare attenzione agli incidenti mortali.

Gli studi connessi con la sicurezza stradale hanno messo in luce che l'incidentalità stradale è un fenomeno molto complesso da analizzare. Il rischio di incidente è legato a diversi fattori, quali le caratteristiche strutturali della rete e della strada, i comportamenti dei conducenti e i fattori ambientali.

In tali studi, di grande importanza riveste lo studio delle determinanti che determinano gli incidenti stradali, con particolare attenzione agli incidenti mortali. Un problema legato allo studio degli incidenti stradali mortali è dovuto allo squilibrio della variabile di risposta. Diversi studi hanno evidenziato che la percentuale degli incidenti mortali è inferiore al 3% degli incidenti totali (Montella et al., 2011; Montella et al., 2012). Come è stato visto in precedenza, molti algoritmi sono inefficaci nel predire la classa rara, in questo caso nel predire gli incidenti mortali (Tesema et al., 2005, Emerson et al., 2011; Nayak et al., 2011, Torrão et al., 2014).

In tale studio, si vuole verificare come l'algoritmo di ricampionamento SONCA possa risolvere il problema delle classi sbilanciate nei dataset degli incidenti stradali. A tal fine sono stati presi due dataset con differenti distribuzioni e predittori (Tabella 48).

Tabella 48: Database degli incidenti stradali

		Non Mortali		Mortali		
Dataset	**# Var.**	**N**	**%**	**N**	**%**	**Tot**
PTW - ISTAT	26	225017	98.26	3980	1.74	228997
A56 Crashes	31	1688	93.11	125	6.89	1813

5.1. PTW crashes

I dati analizzati sono i microdati ISTAT (Istituto nazionale di STATistica) dell'intero territorio nazionale nel triennio 2008-2010. Tali dati riportano per ciascun incidente, in formato ASCII, 159 campi in cui sono presenti tutte le

informazioni contenute nel rapporto statistico di incidente stradale ISTAT CTT.INC: data, localizzazione, luogo, natura, circostanze accertate o presunte, tipo di veicoli coinvolti, conseguenze alle persone e conseguenze ai veicoli (Montella et al., 2008; Montella et al., 2012).

Dall'analisi del dataset e da uno studio della letteratura, dai 159 campi forniti sono state scelte solo 11 variabili categoriche (Tabella 49).

Tabella 49: Descrizione del database PTW-ISTAT

Variabile	Modalità Variabile
Severity	Fatal; Injury
Week Day	Week end; Week day
Season	Autumn;Springer; Winter; Summer
Lighting	Day; Night
Weather	Clear; Rainy; Other
Area	Rural; Urban
Road type	Urban municipal; Urban provincial; Urban national; Rural municipal; Rural provincial; Rural national; Motorway
Alignment	Tangent; Curve; Unsignalized intersection; Signalized intersection; Other
Pavement	Dry; Wet ; Slippery; Frozen; Snowy
Involved Vehicle	Single vehicle; Multi-vehicles
Collision type	Angle; Falling from the vehicle; Head-on; Hit fixed obstacle in carriageway; Hit temporary obstacle in carriageway; Hit parked vehicle; Hit pedestrian; Hit stopped vehicle; Hit train; Rear-end; Run-off-the-road; Sideswipe; Sudden braking
Speed related	Speed related; No Speed related

Il dataset originale consisteva di 645772 incidenti stradali, con 12184 incidenti mortali e 633588 incidenti con feriti. Poiché i motociclisti sono più vulnerabili degli altri conducenti, a causa della mancanza di protezioni in caso di incidente, in tale studio si è voluto porre l'attenzione sugli incidenti in cui è stato coinvolto almeno un veicolo a due ruote (Powered two-wheeler –PTW). Il dataset PTW-ISTAT consta di 228997 osservazioni. Come si osserva in Tabella 50, la variabile di risposta è molto sbilanciata, la classe 0 (incidenti con feriti) ha una frequenza del 98.3%, mentre la classe 1 (incidenti mortali) ha una frequenza dell'1.7%.

Tabella 50: Distribuzioni di classe per la variabile Severity

Severity	N	$F=N/N_{tot}$ [%]
0 (Con feriti)	225017	98.26

1 (Mortali)	3980	1.74
Totale	**228997**	100.00

Attraverso un processo di estrazione casuale il dataset è stato suddiviso in due dataset:

- Training set (Tabella 51) costituito da 114237 osservazioni pari al 49.9% delle osservazioni totali;

Tabella 51: Distribuzioni di classe del training set per la variabile Severity

Severity	N	$F=N/N_{tot}$ [%]
0	112250	98.26
1	1987	1.74
Totale	**114237**	100.00

- Holdout test (Tabella 52) costituito da 114760 pari 50.1% delle osservazioni totali.

Tabella 52: Distribuzioni di classe del dataset test per la variabile Severity

Severity	N	$F=N/N_{tot}$ [%]
0	112767	98.26
1	1993	1.74
Totale	**114760**	100.00

Per tale dataset sono stati stimati l'albero di classificazione CART e il Logit nei seguenti casi:

- Dataset originale senza ricampionamento;
- Dataset bilanciato con SONCA usando la distribuzione di probabilità triangolare;
- Dataset bilanciato con SONCA usando la distribuzione di probabilità gaussiana.

5.1.1. Dataset originale

Albero di regressione senza il ricampionamento

In Figura 31 e Figura 32 sono riportati gli alberi di regressione senza bilanciare la variabile di risposta.

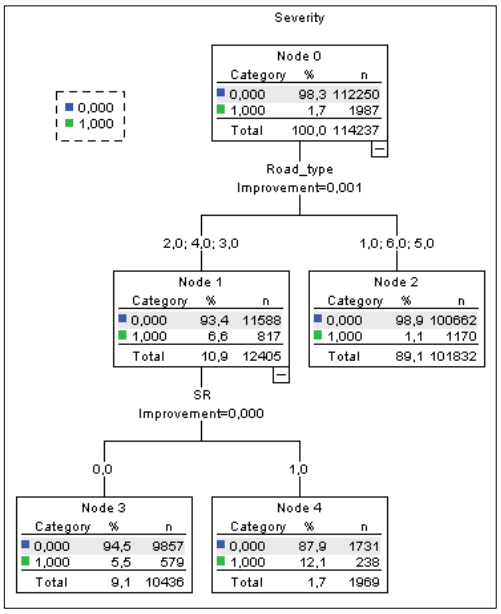

Figura 31 – Albero di regressione per il training set senza ricampionamento

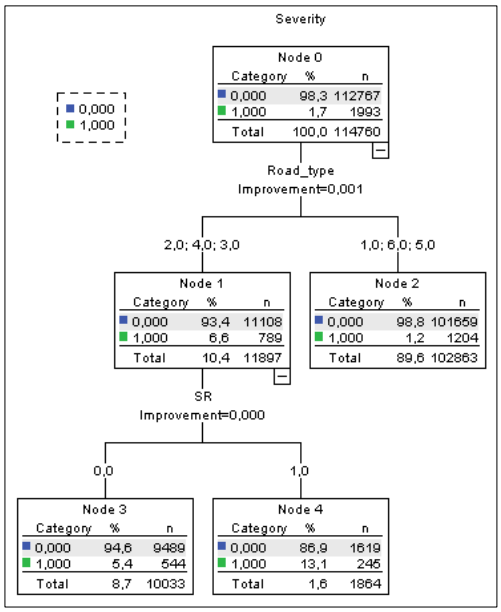

Figura 32 – Albero di regressione per l'holdout set senza ricampionamento

In Tabella 53 e Tabella 54 sono riportate sia la matrice di confusione sia le misure di prestazioni per entrambi i set dati, training e test.

Tabella 53: Matrice di confusione senza ricampionamento

Sample		Predicted		
		1	0	Percent Correct
Training	1	0	1987	0.0%
	0	0	112250	100.0%
	Overall Percentage	0.0%	100.0%	98.3%
Holdout	1	0	1993	0.0%
	0	0	112767	100.0%
	Overall Percentage	0.0%	100.0%	98.3%

Tabella 54: Misure di performance senza ricampionamento

Dataset	FNrate	FPrate	TNrate	TPrate	Precision	F-measure	G-mean	Err	Acc	AUC
Training	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.017	0.983	
Holdout	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.017	0.983	0

Dai risultati ottenuti si osserva che il modello tende a concentrarsi sulla classe prevalente e ignorare gli eventi rari. Tutte le osservazioni che appartengono alla classe minoritaria sono classificate erroneamente.

Logit senza il ricampionamento

In Tabella 55 sono riportate le variabili nell'equazione.

Tabella 55: Variabili nell'equazione senza il ricampionamento

	B	S.E.	Wald	df	Sig.	Exp(B)
SR(1)	-0.784	0.057	188.859	1.000	0.000	0.457
zWeek_Dy(1)	0.344	0.049	48.848	1.000	0.000	1.411
zSeason			23.668	3.000	0.000	
zSeason(1)	-0.258	0.069	14.174	1.000	0.000	0.772
zSeason(2)	0.057	0.056	1.039	1.000	0.308	1.059
zSeason(3)	-0.163	0.076	4.552	1.000	0.033	0.849
Lighting(1)	0.503	0.052	93.606	1.000	0.000	1.654
Road_type			1.070.199	5.000	0.000	
Road_type(1)	0.954	0.070	183.626	1.000	0.000	2.597
Road_type(2)	1.278	0.116	120.914	1.000	0.000	3.588
Road_type(3)	1.808	0.076	559.275	1.000	0.000	6.097
Road_type(4)	1.917	0.066	855.035	1.000	0.000	6.800
Road_type(5)	1.124	0.129	75.396	1.000	0.000	3.078
Alignment			43.598	4.000	0.000	
Alignment(1)	0.189	0.069	7.465	1.000	0.006	1.209
Alignment(2)	-0.260	0.057	20.595	1.000	0.000	0.771
Alignment(3)	-0.352	0.124	8.023	1.000	0.005	0.703
Alignment(4)	0.109	0.186	0.346	1.000	0.556	1.115
Pavment			30.228	4.000	0.000	
Pavment(1)	-0.332	0.089	13.952	1.000	0.000	0.718
Pavment(2)	-1.115	0.273	16.692	1.000	0.000	0.328
Pavment(3)	0.227	0.465	0.238	1.000	0.626	1.254
Pavment(4)	-0.387	1.011	0.146	1.000	0.702	0.679
Crash_type			377.246	12.000	0.000	
Crash_type(1)	1.581	0.157	101.673	1.000	0.000	4.862
Crash_type(2)	-0.436	0.169	6.636	1.000	0.010	0.647
Crash_type(3)	1.514	0.235	41.662	1.000	0.000	4.547
Crash_type(4)	-16.940	22.639.765	0.000	1.000	0.999	0.000
Crash_type(5)	0.950	0.123	59.823	1.000	0.000	2.585
Crash_type(6)	-0.418	0.087	23.246	1.000	0.000	0.658
Crash_type(7)	1.623	0.185	76.709	1.000	0.000	5.068
Crash_type(8)	0.701	0.074	88.578	1.000	0.000	2.016
Crash_type(9)	-0.581	0.087	44.765	1.000	0.000	0.559
Crash_type(10)	0.460	0.425	1.169	1.000	0.280	1.584
Crash_type(11)	0.791	0.167	22.325	1.000	0.000	2.205
Crash_type(12)	1.340	0.145	85.306	1.000	0.000	3.819
Vehicle_Number	0.777	0.081	91.703	1.000	0.000	2.174
Constant	-5.631	0.193	848.504	1.000	0.000	0.004

In

Tabella 56 e Tabella 57 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, senza ricampionamento.

Tabella 56: Matrice di confusione – Logit senza il ricampionamento

Sample		Predicted		
		1	0	Percent Correct
Training	1	0	1987	0.0%
	0	0	112250	100.0%
	Overall Percentage	0.0%	100.0%	98.3%
Holdout	1	1	1992	0.1%
	0	0	112767	100.0%
	Overall Percentage	0.0%	100.0%	98.3%

Tabella 57: Misure di performance – Logit senza il ricampionamento

Dataset	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
Training	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.017	0.983	
Holdout	0.999	0.000	1.000	0.001	0.000	0.000	0.022	0.017	0.983	0.792

In Figura 34 e Tabella 58sono riportate la curva ROC e l'area sottesa alla curva ROC.

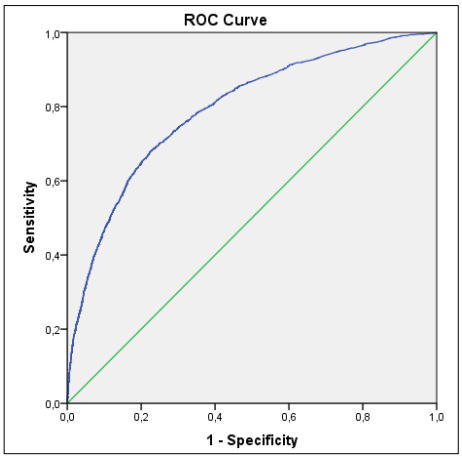

Figura 33 - Curva ROC per l'holdout set

Tabella 58: Area sotto la curva ROC per l'holdout set

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.792	0.005	0	0.782	0.802

Il modello stimato nonostante abbia un AUC pari a 0.792, il modello presenta un tasso di falsi negativi piuttosto alto, FN_{rate}=0.999, con valori di G-mean=0.022.

5.1.2. SONCA con distribuzione di probabilità triangolare

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo triangolare. In Tabella 59 sono riportate le distribuzioni di frequenza per il set di training.

Tabella 59: Distribuzioni di classe per il training set: SONCA e distribuzione di probabilità triangolare

Variabile	N	%
0	2025	50.76
1	1964	49.23
Totale	**3989**	100.00

Albero di regressione con SONCA e distribuzione di probabilità triangolare

In Figura 34 e Figura 35 sono riportati gli alberi di regressione training holdout set avendo implementato SONCA con distribuzione triangolare.

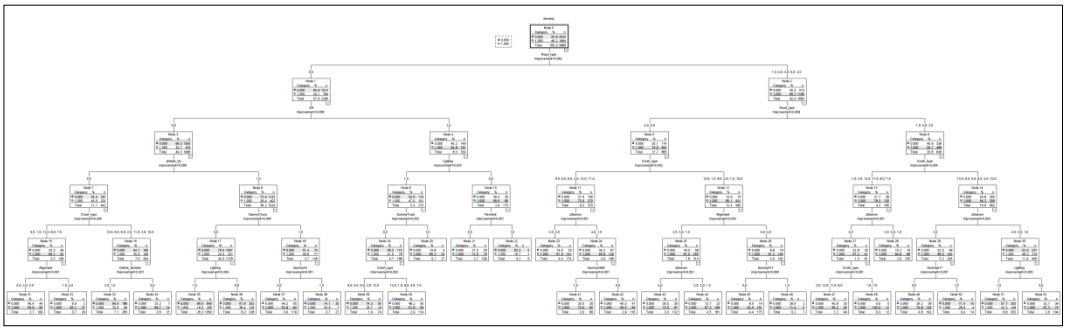

Figura 34 – Albero di regressione per il training set: SONCA e distribuzione triangolare

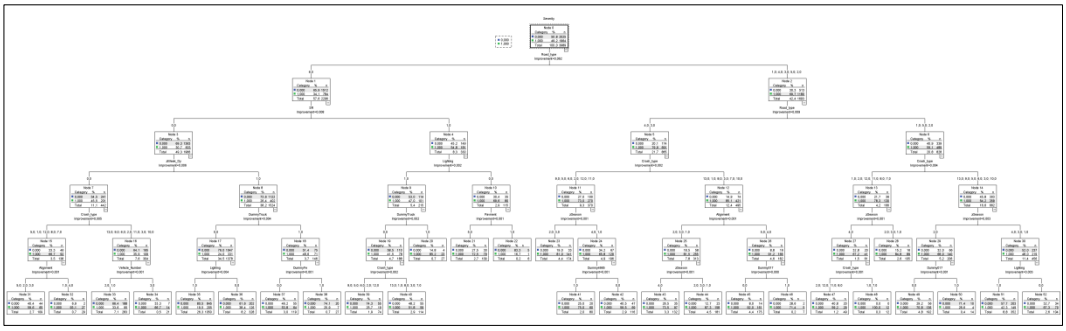

Figura 35 – Albero di regressione per l'holdout set: SONCA e distribuzione triangolare

In Tabella 60 e Tabella 61 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA.

Tabella 60: Matrice di confusione - CART: SONCA

Sample		Predicted		
		1	0	Percent Correct
Training	1	1358	606	69.1%
	0	497	1528	75.5%
	Overall Percentage	46.5%	53.5%	72.3%
Holdout	1	1303	690	65.4%
	0	28670	84097	74.6%
	Overall Percentage	26.1%	73.9%	74.4%

Tabella 61: Misure di performance - CART: SONCA

Dataset	FNrate	FPrate	TNrate	TPrate	Precision	F-measure	G-mean	Err	Acc	AUC
Training	0.309	0.245	0.755	0.691	0.732	0.711	0.722	0.277	0.723	
Holdout	0.346	0.254	0.746	0.654	0.043	0.082	0.698	0.256	0.744	0.751

In Figura 36 e Tabella 62 sono riportate la curva ROC e l'area sottesa alla curva ROC.

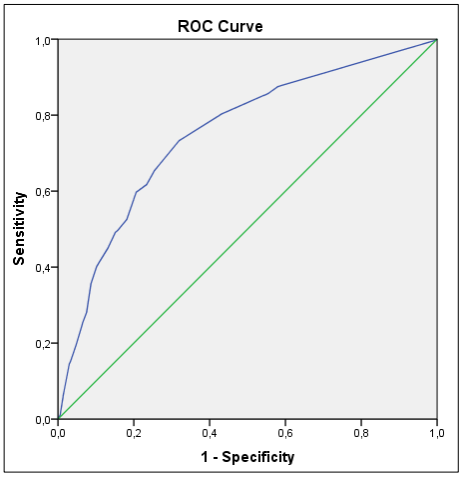

Figura 36 - Curva ROC per l'holdout set

Tabella 62: Area sotto la curva ROC per l'holdout set

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.751	0.006	0	0.740	0.762

In tale modello si osserva che il tasso di casi positivi erroneamente classificati come appartenenti alla classe negativa diminuisce FN_{rate} passa infatti da 1 nel modello senza ricampionamento a 0.346, si ha così una riduzione del 65%. Aumentano anche le altre misure di performance, in particolare G-mean=0.698 e AUC=0.751.

Logit con SONCA e distribuzione di probabilità triangolare

In Tabella 63 sono riportate le variabili nell'equazione.

Tabella 63: Variabili nell'equazione: SONCA con distribuzione di probabilità triangolare

	B	S.E.	Wald	df	Sig.	Exp(B)
SR(1)	-0.698	0.103	45.594	1.000	0	0.498
zWeek_Dy(1)	0.591	0.083	50.857	1.000	0	1.807
zSeason			56.812	3.000	0	
zSeason(1)	-0.112	0.109	1.056	1.000	0	0.894
zSeason(2)	0.575	0.089	41.499	1.000	0	1.776
zSeason(3)	0.039	0.120	0.104	1.000	1	1.039
Lighting(1)	0.786	0.088	79.461	1.000	0	2.194
Road_type			396.367	5.000	0	
Road_type(1)	1.026	0.100	104.614	1.000	0	2.789
Road_type(2)	1.287	0.215	35.824	1.000	0	3.623
Road_type(3)	1.929	0.148	170.186	1.000	0	6.879
Road_type(4)	2.015	0.131	235.438	1.000	0	7.500
Road_type(5)	1.424	0.232	37.542	1.000	0	4.154
Alignment			37.647	4.000	0	
Alignment(1)	-0.226	0.126	3.194	1.000	0	0.798
Alignment(2)	-0.471	0.087	29.685	1.000	0	0.624
Alignment(3)	-0.662	0.177	14.031	1.000	0	0.516
Alignment(4)	0.268	0.344	0.603	1.000	0	1.307
Pavment			9.960	4.000	0	
Pavment(1)	-0.143	0.147	0.945	1.000	0	0.867
Pavment(2)	-1.249	0.412	9.178	1.000	0	0.287
Pavment(3)	0.127	0.668	0.036	1.000	1	1.136
Pavment(4)	-0.313	1.520	0.042	1.000	1	0.731
Crash_type			130.174	11.000	0	
Crash_type(1)	1.436	0.288	24.880	1.000	0	4.204
Crash_type(2)	-0.573	0.241	5.632	1.000	0	0.564
Crash_type(3)	0.806	0.434	3.450	1.000	0	2.239
Crash_type(4)	0.660	0.217	9.221	1.000	0	1.935
Crash_type(5)	-0.674	0.141	22.800	1.000	0	0.510
Crash_type(6)	1.968	0.416	22.438	1.000	0	7.160
Crash_type(7)	0.654	0.135	23.540	1.000	0	1.922
Crash_type(8)	-0.589	0.117	25.456	1.000	0	0.555
Crash_type(9)	-0.514	0.692	0.553	1.000	0	0.598
Crash_type(10)	0.823	0.275	8.982	1.000	0	2.277
Crash_type(11)	1.021	0.273	13.965	1.000	0	2.776
Vehicle_Number	1.017	0.168	36.886	1.000	0	2.766
Constant	-2.317	0.374	38.297	1.000	0	0.099

In Tabella 64 e Tabella 65 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA.

Tabella 64: Matrice di confusione - Logit: SONCA

Sample		Predicted		
		1	0	Percent Correct
Training	1	1131	833	57.6%
	0	509	1516	74.9%
	Overall Percentage	41.1%	58.9%	66.4%
Holdout	1	1167	826	58.6%
	0	31528	81239	72.0%
	Overall Percentage	28.5%	71.5%	71.8%

Tabella 65: Misure di performance - Logit: SONCA

Dataset	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
Training	0.424	0.251	0.749	0.576	0.690	0.628	0.657	0.336	0.664	
Holdout	0.414	0.280	0.720	0.586	0.036	0.067	0.649	0.282	0.718	0.689

In Figura 37 e Tabella 66 sono riportate la curva ROC e Area sotto la curva ROC.

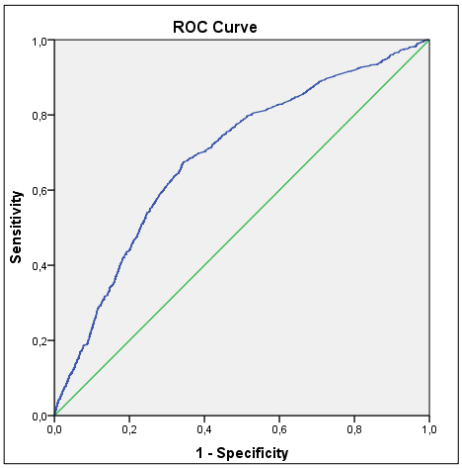

Figura 37 - Curva ROC per l'holdout set - Logit: SONCA

Tabella 66: Area sotto la curva ROC per l'holdout set - CART: SONCA con distribuzione di probabilità triangolare e m=4000

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.689	0.006	0.000	0.678	0.701

Analogamente a quanto visto per il CART, il Logit stimato, dopo aver ricampionato il training set con SONCA usando una distribuzione di probabilità triangolare, mostra dei miglioramenti. In tale modello, si osserva che FN_{rate} diminuisce, passa da 1 nel modello senza ricampionamento a 0.414, si ha così una riduzione del 58%. Aumentano anche le altre misure di performance, in particolare G-mean=0.649 e AUC=0.689.

5.1.3. SONCA con distribuzione di probabilità gaussiana

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo gaussiana. In Tabella 67 sono riportate le distribuzioni di frequenza per il set di training.

Tabella 67: Distribuzioni di classe per il training set: SONCA e distribuzione di probabilità gaussiana

Variabile	N	%
0	1006	51.54
1	946	48.46
Totale	**1952**	100.00

Albero di regressione con SONCA e distribuzione di probabilità gaussiana

In Figura 38 e Figura 39 sono riportati gli alberi di regressione training holdout set avendo implementato SONCA con distribuzione gaussiana.

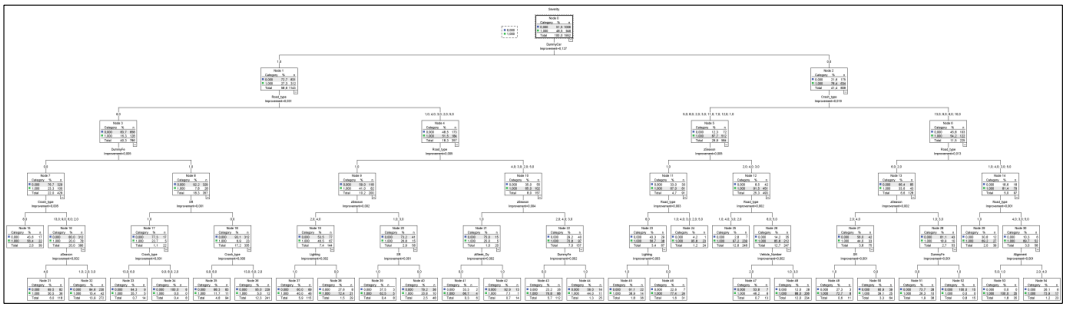

Figura 38 – Albero di regressione per il training set: SONCA e distribuzione gaussiana

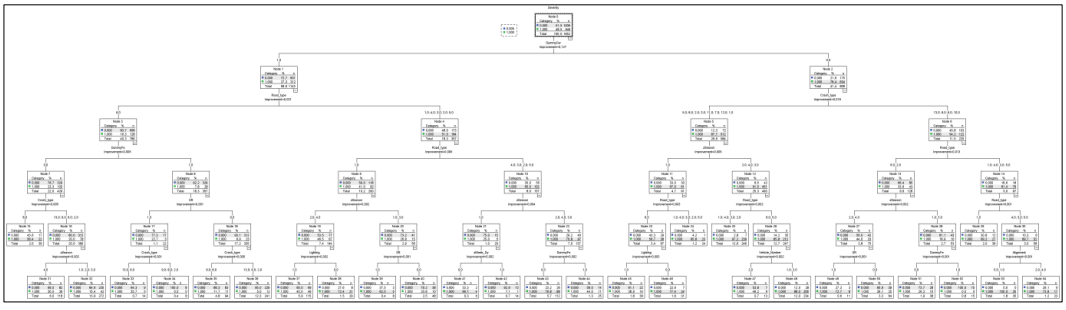

Figura 39 – Albero di regressione per l'holdout set: SONCA e distribuzione gaussiana

In

Tabella 68 e Tabella 69 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA.

Tabella 68: Matrice di confusione - CART: SONCA

Sample		Predicted		
		1	0	Percent Correct
Training	1	717	229	75.8%
	0	120	886	88.1%
	Overall Percentage	42.9%	57.1%	82.1%
Holdout	1	1231	762	61.8%
	0	31111	81656	72.4%
	Overall Percentage	28.2%	71.8%	72.2%

Tabella 69: Misure di performance - CART: SONCA

Dataset	FNrate	FPrate	TNrate	TPrate	Precision	F-measure	G-mean	Err	Acc	AUC
Training	0.242	0.119	0.881	0.758	0.857	0.804	0.817	0.179	0.821	
Holdout	0.382	0.276	0.724	0.618	0.038	0.072	0.669	0.278	0.722	0.712

In Figura 40 e Tabella 70 sono riportate la curva ROC e l'area sottesa alla curva ROC.

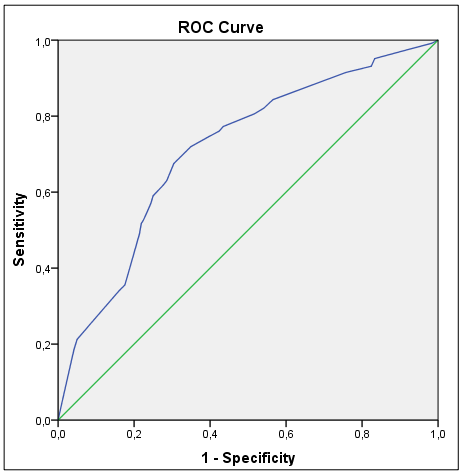

Figura 40 - Curva ROC per l'holdout set

Tabella 70: Area sotto la curva ROC per l'holdout set

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.712	0.006	0.000	0.701	0.723

In tale modello si osserva che il tasso di casi positivi erroneamente classificati come appartenenti alla classe negativa diminuisce, FN_{rate} passa da 1 nel modello senza ricampionamento a 0.382, si ha così una riduzione del 61%. Aumentano anche le altre misure di performance, in particolare G-mean=0.669 e AUC=0.712.

Logit con SONCA e distribuzione di probabilità gaussiana

In Tabella 71 sono riportate le variabili nell'equazione.

Tabella 71: Variabili nell'equazione: SONCA con distribuzione di probabilità gaussiana

Variabili	B	S.E.	Wald	df	Sig.	Exp(B)
SR(0)	- 1.180	0.194	36.854	1.000	0.000	0.307
Week_Dy(0)	0.313	0.145	4.675	1.000	0.031	1.367
Season			67.639	3.000	0.000	
Season(1)	- 1.394	0.174	64.204	1.000	0.000	0.248
Season(2)	- 0.412	0.151	7.436	1.000	0.006	0.662
Season(3)	- 0.807	0.202	15.878	1.000	0.000	0.446
Lighting(1)	0.555	0.154	12.929	1.000	0.000	1.742
Area(1)	- 20.286	12984.444	0.000	1.000	0.999	0.000
Road_type			72.532	5.000	0.000	
Road_type(1)	1.268	0.154	67.571	1.000	0.000	3.555
Road_type(2)	22.058	12984.444	0.000	1.000	0.999	3800591820.604
Road_type(3)	23.009	12984.444	0.000	1.000	0.999	9832886063.187
Road_type(4)	22.401	12984.444	0.000	1.000	0.999	5352099385.928
Road_type(5)	22.055	12984.444	0.000	1.000	0.999	3786885478.986
Crash_type			46.050	10.000	0.000	
Crash_type(1)	- 0.685	0.464	2.174	1.000	0.140	0.504
Crash_type(2)	0.782	0.333	5.512	1.000	0.019	2.187
Crash_type(3)	- 1.126	0.734	2.356	1.000	0.125	0.324
Crash_type(4)	- 0.627	0.364	2.975	1.000	0.085	0.534
Crash_type(5)	0.008	0.213	0.001	1.000	0.971	1.008
Crash_type(6)	0.987	1.064	0.861	1.000	0.353	2.683
Crash_type(7)	1.280	0.231	30.792	1.000	0.000	3.595
Crash_type(8)	0.107	0.184	0.338	1.000	0.561	1.113
Crash_type(9)	- 1.973	0.770	6.560	1.000	0.010	0.139
Crash_type(10)	- 0.647	0.498	1.685	1.000	0.194	0.524
Involved_Vehicle(0)	2.666	0.460	33.577	1.000	0.000	14.382
Vehicle_Number	- 0.716	0.345	4.302	1.000	0.038	0.489
Constant	1.223	0.732	2.793	1.000	0.095	3.396

In Tabella 72 e

Tabella 73 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA.

Tabella 72: Matrice di confusione - Logit: SONCA

Sample		Predicted		
		1	0	Percent Correct
Training	1	701	245	74.1%
	0	287	719	71.5%
	Overall Percentage	50.6%	49.4%	72.7%
Holdout	1	1373	620	68.9%
	0	41557	71210	63.1%
	Overall Percentage	37.4%	62.6%	63.2%

Tabella 73: Misure di performance - Logit: SONCA

Dataset	FNrate	FPrate	TNrate	TPrate	Precision	F-measure	G-mean	Err	Acc	AUC
Training	0.259	0.285	0.715	0.741	0.710	0.725	0.728	0.273	0.727	
Holdout	0.311	0.369	0.631	0.689	0.032	0.061	0.660	0.368	0.632	0.722

In Figura 41 e Tabella 74 sono riportate la curva ROC e Area sotto la curva ROC.

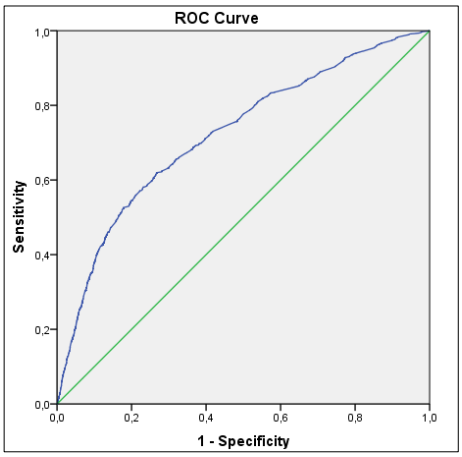

Figura 41 - Curva ROC per l'holdout set - Logit: SONCA

Tabella 74: Area sotto la curva ROC per l'holdout set - CART: SONCA con distribuzione di probabilità gaussiana

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.722	0.006	0.000	0.710	0.734

Analogamente a quanto visto per il CART, il Logit stimato dopo aver ricampionato il training set con SONCA usando una distribuzione di probabilità gaussiana mostra dei miglioramenti. In tale modello, si osserva che FN_{rate} diminuisce, passa da 1 nel modello senza ricampionamento a 0.694, si ha così una riduzione del 31%. Aumentano anche le altre misure di performance, in particolare G-mean=0.497 e AUC=0.581.

5.1.4. Confronto dei risultati PTW

Confrontando i risultati del CART tra le due distribuzioni di frequenza (Tabella 75) si osservano risultati non molto differenti. Con la distribuzione di

probabilità triangolare, FN_{rate} =0.346 per la distribuzione triangolare, mentre FN_{rate} =0.382 per la distribuzione gaussiana. Per la distribuzione triangolare, il tasso di casi positivi erroneamente classificati come appartenenti alla classe negativa si riduce del 65%, mentre per la distribuzione gaussiana si riduce del 61%. Contemporaneamente si osserva che il tasso di corretti positivi sono più alti usando una distribuzione di probabilità triangolare piuttosto che quella gaussiana (TP_{rate}=0.654 vs TPrate=0.618). I valori di Precision e F-measure non sono molto differenti tra di loro, ma si osserva che i valori di G-mean e AUC sono più alti per la distribuzione triangolare.

Tabella 75: Confronto dei risultati PTW ISTAT - CART

Dataset	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
Originale	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.017	0.983	0.000
SONCA (Triangolare)	0.346	0.254	0.746	0.654	0.043	0.082	0.698	0.256	0.744	0.751
SONCA (Gaussiana)	0.382	0.276	0.724	0.618	0.038	0.072	0.669	0.278	0.722	0.712

Diversamente da quanto visto per il CART, confrontando i risultati del Logit tra le due distribuzioni di frequenza (Tabella 76), si osservano migliori risultati con la distribuzione di probabilità triangolare, FN_{rate} =0.414 per la distribuzione triangolare, mentre FN_{rate} =0.311 per la distribuzione gaussiana, anche i tassi di corretti positivi sono più alti usando una distribuzione di probabilità gaussiana (TP_{rate}=0.586 vs TPrate=0.689). I valori di Precision e F-measure non sono molto differenti tra di loro, ma si osserva che i valori di G-mean e AUC sono più alti per la distribuzione gaussiana.

Tabella 76: Confronto dei risultati PTW ISTAT - Logit

Dataset	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
Originale	0.999	0.000	1.000	0.001	0.000	0.000	0.022	0.017	0.983	0.792
SONCA (Triangolare)	0.414	0.280	0.720	0.586	0.036	0.067	0.649	0.282	0.718	0.689
SONCA (Gaussiana)	0.311	0.369	0.631	0.689	0.032	0.061	0.660	0.368	0.632	0.722

Confrontando i risultati tra il CART e il Logit, si osserva che il CART offre dei risultati migliori. Il CART non richiede alcuna assunzione sulla variabile obiettivo e nemmeno con riguardo alle relazioni fra le variabili in esame si impongono ipotesi specifiche. Tutto questo rende la tecnica più flessibile e potenzialmente più efficiente delle metodologie tradizionali (Zani e Cerioli, 2007).

5.2. A56 crashes - UniNa

I dati riguardanti gli incidenti stradali utilizzati in questo lavoro, sono stati acquisiti mediante la consultazione dei rapporti degli incidenti stradali relativi al periodo 2006 - 2011, presso la Sottosezione della Polizia Stradale di Fuorigrotta Napoli, avvenuti sull'A56 Tangenziale di Napoli. La raccolta delle informazioni è stata possibile grazie al previo rilascio di un'autorizzazione del Ministero degli Interni e la successiva richiesta di autorizzazione inviata al dirigente della sezione Polizia Stradale di Napoli e al Comandante della Sottosezione Polizia Stradale di Napoli, la quale specificava che tutto sarebbe stato eseguito nel pieno rispetto della normativa sulla privacy.

Tali dati riportano per ciascun incidente, 135 campi in cui sono presenti tutte le informazioni contenute nel "Prontuario per le annotazioni e gli accertamenti urgenti relativi agli incidenti stradali" fornito dal Ministero degli Interni (Montella et al., 2012). Da tali informazioni sono state estratte 16 predittori di natura mista (Tabella 77). La Tangenziale Est-Ovest di Napoli è un'autostrada urbana (tipo A secondo il D.M. 5.11.01 "Norme funzionali e geometriche per la costruzione delle strade") costituita da due carreggiate separate da spartitraffico, ciascuna con tre corsie di marcia.

Il dataset originale consisteva di 2357 incidenti stradali, poiché lo studio è stato focalizzato sugli incidenti avvenuti in autostrada, non negli svincoli e nelle aree limitrofe, sono stati analizzati 1831 incidenti.

Tabella 77: Descrizione del database A56 crashes - UniNa

Variabile	Attributi
workday	festivo;feriale
Fasce orario	
Sezione trasversale	Rilevato ; A raso; Trincea; Viadotto; Galleria
Geometria orizzontale	Rettifilo; Curva
Dir. Curva	dx; sx;
Raggio curva	dimensioni curva [m]
Angolo di deviazione	ampiezza della curva [gon]
Fondo stradale	Asciutto; Bagnato; Sdrucciolevole; Ghiacciato; Innevato
Meteo	Sereno; Coperto; Pioggia; Grandine; Neve; Nebbia; Vento forte; Altro
Illuminazione	Giorno; Notte; Notte, con illuminazione artificiale
Tipo di incidente	Frontale; Fronto-laterale; Laterale; Tamponamento; Investimento pedone; Investimento animale; Urto con veicolo fermo; Urto con veicolo in sosta; Urto con ostacolo accidentale in carreggiata; Fuoriuscita; Frenata improvvisa; Caduta da veicolo; Ribaltamento in carreggiata
Incidente secondario	1 se l'incidente è causato da incidente precedente; altrimenti 0
Speed related	1 se almeno uno dei conducenti coinvolti nell'incidente è stato multato per eccesso di velocità; altrimenti 0
Tipo di veicolo	Auto; Pesante; Moto; Altro
Genere	Maschio; Femmina; Nd;
Età	[18-25[; [25-45[; [45-60[; [60-+∞[

Come si osserva in Tabella 78, la variabile di risposta è sbilanciata, la classe 0 (incidenti con feriti e con solo danni materiali) ha una frequenza del 93.1%, mentre la classe 1 (incidenti mortali) ha una frequenza del 6.9%.

Tabella 78: Distribuzioni di classe per la variabile Severity

Severity	N	$F=N/N_{tot}$ [%]
0 (Incidenti con feriti e con solo Danni materiali)	1688	93.11
1 (Incidenti Mortali)	125	6.89
Totale complessivo	**1813**	100.00

Attraverso un processo di estrazione casuale il dataset è stato suddiviso in due dataset:

- Training set (Tabella 79) costituito da 932 osservazioni pari al 51.4% delle osservazioni totali;

Tabella 79: Distribuzioni di classe del training set per la variabile Severity

Severity	N	$F=N/N_{tot}$ [%]
0	861	92.38
1	71	7.62
Totale	**932**	100.00

- Holdout test (Tabella 80) costituito da 881 pari 49.6% delle osservazioni totali.

Tabella 80: Distribuzioni di classe del dataset test per la variabile Severity

Severity	N	$F=N/N_{tot}$ [%]
0	827	93.87
1	54	6.13
Totale	**881**	100.00

Per tale dataset sono stati stimati l'albero di classificazione CART e il logit nei seguenti casi:

- Dataset originale senza ricampionamento;
- Dataset bilanciato con SONCA usando la distribuzione di probabilità triangolare;
- Dataset bilanciato con SONCA usando la distribuzione di probabilità gaussiana.

5.2.1. Dataset originale

Albero di regressione senza il ricampionamento

In Figura 42 e Figura 43sono riportati gli alberi di regressione senza bilanciare la variabile di risposta.

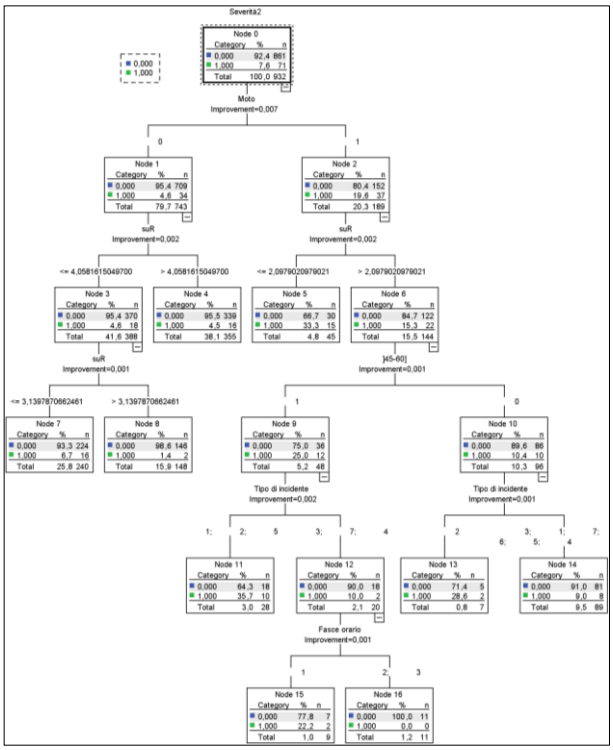

Figura 42 – Albero di regressione per il training set senza ricampionamento

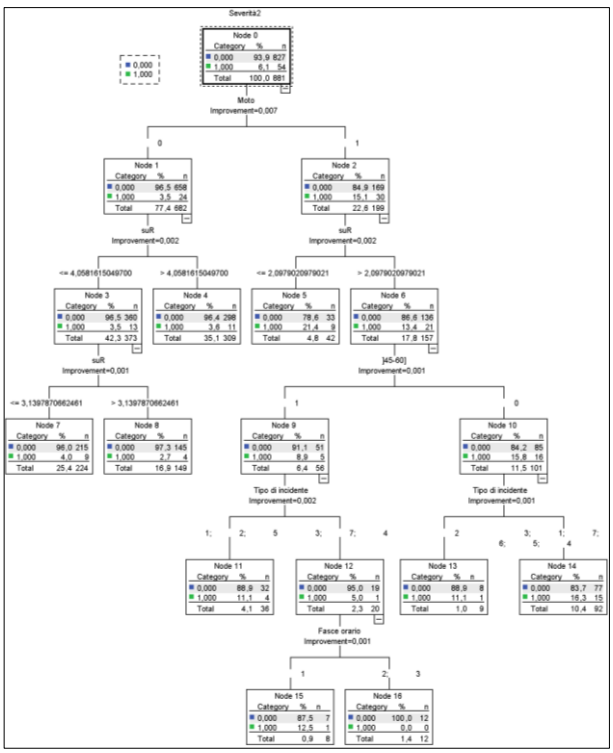

Figura 43 – Albero di regressione per l'holdout set senza ricampionamento

Sample		Predicted		
		1	0	Percent Correct
Training	1	0	71	0.0%
	0	0	861	100.0%
	Overall Percentage	0.0%	100.0%	92.4%
Holdout	1	0	54	0.0%
	0	0	827	100.0%
	Overall Percentage	0.0%	100.0%	93.9%

Tabella 81: Misure di performance senza ricampionamento

Dataset	FNrate	FPrate	TNrate	TPrate	Precision	F-measure	G-mean	Err	Acc	AUC
Training	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.076	0.924	
Holdout	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.061	0.939	0

Dai risultati ottenuti si osserva che il modello tende a concentrarsi sulla classe prevalente e ignorare gli eventi rari. Tutte le osservazioni che appartengono alla classe minoritaria sono classificate erroneamente.

Logit senza il ricampionamento

In Tabella 82 sono riportate le variabili nell'equazione. Come si osserva non è presente nessuna variabile, solo la costante.

Tabella 82: Variabili nell'equazione senza il ricampionamento

Variabile	B	S.E.	Wald	df	Sig.	Exp(B)
Geometriaorizzontale(0)	-0.686	0.274	6.261	1	0.012	0.504
Tipodiincidente			14.291	6	0.027	
Tipodiincidente(1)	0.552	1.043	0.281	1	0.596	1.738
Tipodiincidente(2)	0.520	1.059	0.241	1	0.624	1.681
Tipodiincidente(3)	0.389	1.091	0.127	1	0.722	1.475
Tipodiincidente(4)	0.853	1.157	0.544	1	0.461	2.347
Tipodiincidente(5)	1.945	1.078	3.258	1	0.071	6.996
Tipodiincidente(6)	-17.901	4777.868	0.000	1	0.997	0.000
Incidentesecondario(0)	-0.768	0.269	8.134	1	0.004	0.464
Constant	-2.388	1.046	5.211	1	0.022	0.092

In Tabella 83 e Tabella 84 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, senza ricampionamento.

Tabella 83: Matrice di confusione – Logit senza il ricampionamento

Sample		Predicted		
		1	0	Percent Correct
Training	1	7	64	9.9%
	0	240	621	72.1%
	Overall Percentage	26.5%	73.5%	67.4%
Holdout	1	9	45	16.7%
	0	211	616	74.5%
	Overall Percentage	23.6%	70.9%	70.9%

Tabella 84: Misure di performance – Logit senza il ricampionamento

Dataset	FNrate	FPrate	TNrate	TPrate	Precision	F-measure	G-mean	Err	Acc	AUC
Training	0.901	0.279	0.721	0.099	0.028	0.044	0.267	0.326	0.674	
Holdout	0.833	0.255	0.745	0.167	0.000	0.000	0.352	0.291	0.709	0.439

In

Figura 44 e Tabella 85 sono riportate la curva ROC e l'area sottesa alla curva ROC.

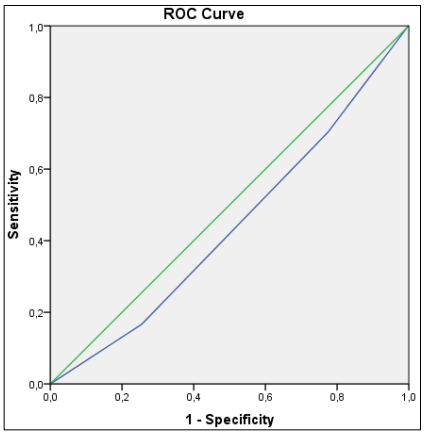

Figura 44 - Curva ROC per l'holdout set

Tabella 85: Area sotto la curva ROC per l'holdout set

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.439	0.039	0.135	0.362	0.517

In tale modello, l'83% dei casi positivi sono erroneamente classificati come appartenenti alla classe negativa, tale fenomeno è evidenziato dalla curva ROC, la quale si trova al di sotto della diagonale principale e l'AUC=0.439, ciò significa che la probabilità che il classificatore identifichi correttamente un'osservazione

estratta a caso dal gruppo dei positivi, quindi dagli incidenti mortali, è inferiore del 50%.

5.2.2. SONCA con distribuzione di probabilità triangolare

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo triangolare. In Tabella 86 sono riportate le distribuzioni di frequenza per il set di training.

Tabella 86: Distribuzioni di classe per il training set: SONCA e distribuzione di probabilità triangolare

Variabile	N	%
0	1956	49.31
1	2011	50.69
Totale	3967	100.00

Albero di regressione con SONCA e distribuzione di probabilità triangolare

In Figura 45 e Figura 46 sono riportati gli alberi di regressione training e holdout set avendo implementato SONCA con distribuzione triangolare.

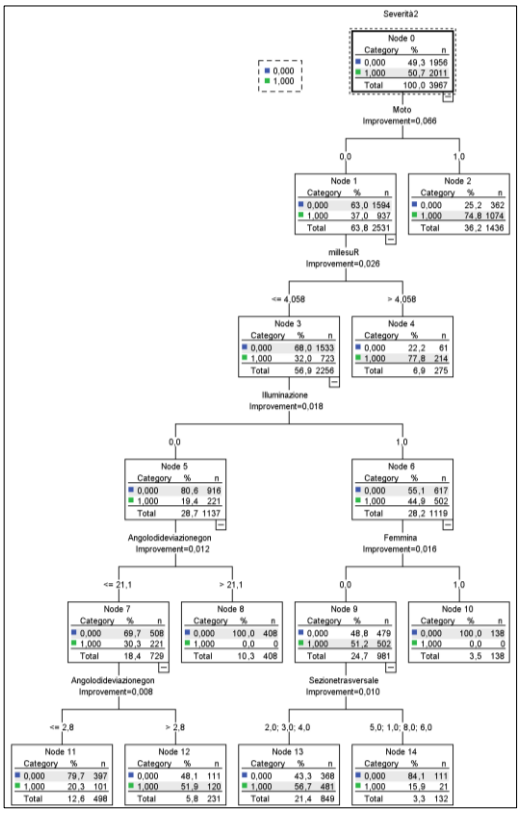

Figura 45 – Albero di regressione per il training set: SONCA e distribuzione triangolare

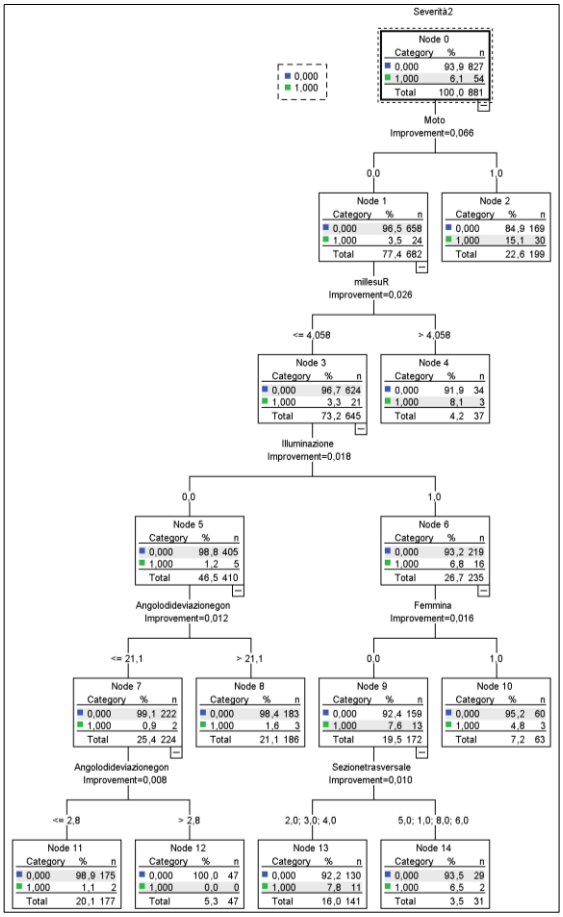

Figura 46 – Albero di regressione per l'holdout set: SONCA e distribuzione triangolare

In Tabella 87 e Tabella 88 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 87: Matrice di confusione - CART: SONCA e distribuzione triangolare

Sample		Predicted		
		1	0	Percent Correct
Training	1	1889	122	93.9%
	0	902	1054	53.9%
	Overall Percentage	70.4%	29.6%	74.2%
Holdout	1	44	10	81.5%
	0	380	447	54.1%
	Overall Percentage	48.1%	51.9%	55.7%

Tabella 88: Misure di performance - CART: SONCA

Dataset	FNrate	FPrate	TNrate	TPrate	Precision	F-measure	G-mean	Err	Acc	AUC
Training	0.061	0.461	0.539	0.939	0.677	0.787	0.711	0.258	0.742	
Holdout	0.185	0.459	0.541	0.815	0.104	0.184	0.664	0.443	0.557	0.712

In Figura 47 e Tabella 89 sono riportate la curva ROC e l'area sottesa alla curva ROC.

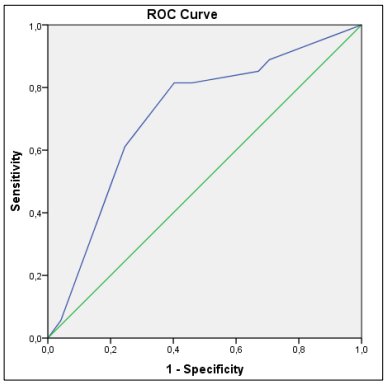

Figura 47 - Curva ROC per l'holdout set

Tabella 89: Area sotto la curva ROC per l'holdout set

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.712	0.036	0.000	0.642	0.782

In tale modello si osserva che il tasso di casi positivi erroneamente classificati come appartenenti alla classe negativa diminuisce FN_{rate} passa da 1 nel modello senza ricampionamento a 0.185, si ha così una riduzione dell'81%, contemporaneamnet TP_{rate} passa da 0 nel modello senza ricampionamento a 0.685. Aumentano anche le altre misure di performance, in particolare G-mean=0.664 e AUC=0.712.

Logit con SONCA e distribuzione di probabilità triangolare

In Tabella 90 sono riportate le variabili nell'equazione.

Tabella 90: Variabili nell'equazione: SONCA e distribuzione di probabilità triangolare

Variabile	B	S.E.	Wald	df	Sig.	Exp(B)
workday(1)	0.323	0.077	17.718	1	0.000	1.381
Sezionetrasversale			32.450	6	0.000	
Sezionetrasversale(1)	20.653	14115.854	0.000	1	0.999	932580848.840
Sezionetrasversale(2)	20.682	14115.854	0.000	1	0.999	959396884.853
Sezionetrasversale(3)	20.793	14115.854	0.000	1	0.999	1072471282.904
Sezionetrasversale(4)	21.125	14115.854	0.000	1	0.999	1494007752.966
Sezionetrasversale(5)	20.686	14115.854	0.000	1	0.999	963006117.712
Sezionetrasversale(6)	21.339	14115.854	0.000	1	0.999	1850942939.648
Geometriaorizzontale(1)	-0.939	00.127	54.334	1	0.000	0.391
millesuR	0.138	0.058	05.619	1	0.018	1.148
Angolodideviazionegon	-0.016	.003	20.476	1	0.000	.985
Fondostradale(1)	-0.590	.174	11.509	1	0.001	.554

Variabile	B	S.E.	Wald	df	Sig.	Exp(B)
Meteo			26.492	2	0.000	
Meteo(1)	-0.061	0.107	0.331	1	0.565	.940
Meteo(2)	-1.108	0.216	26.439	1	0.000	.330
Illuminazione(1)	-0.335	0.078	18.200	1	0.000	.716
Tipodiincidente			168.197	6	0.000	
Tipodiincidente(1)	0.725	0.252	8.253	1	0.004	2.064
Tipodiincidente(2)	0.424	0.261	2.635	1	0.105	1.529
Tipodiincidente(3)	0.378	0.267	2.002	1	0.157	1.459
Tipodiincidente(4)	0.694	0.287	5.859	1	0.015	2.001
Tipodiincidente(5)	2.108	0.276	58.393	1	0.000	8.228
Tipodiincidente(6)	-20.337	3371.989	0.000	1	0.995	.000
Incidentesecondario(1)	-0.792	0.074	112.989	1	0.000	.453
Speedrelated(1)	0.263	0.075	12.327	1	0.000	1.300
Constant	-19.950	14115.854	0.000	1	0.999	.000

In Tabella 91 e Tabella 92 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, con SONCA e distribuzione di probabilità triangolare.

Tabella 91: Matrice di confusione – Logit: SONCA e distribuzione triangolare

Sample		Predicted		
		1	0	Percent Correct
Training	1	1303	708	64.8%
	0	733	1223	62.5%
	Overall Percentage	51.3%	48.7%	63.7%
Holdout	1	32	22	59.3%
	0	305	522	63.1%
	Overall Percentage	8.5%	13.7%	62.9%

Tabella 92: Misure di performance – Logit: SONCA e distribuzione triangolare

Dataset	FNrate	FPrate	TNrate	TPrate	Precision	F-measure	G-mean	Err	Acc	AUC
Training	0.352	0.375	0.625	0.648	0.640	0.644	0.363	0.637		0.352
Holdout	0.407	0.369	0.631	0.593	0.095	0.164	0.371	0.629	0.652	0.407

In Figura 48 e Tabella 93 sono riportate la curva ROC e l'area sottesa alla curva ROC.

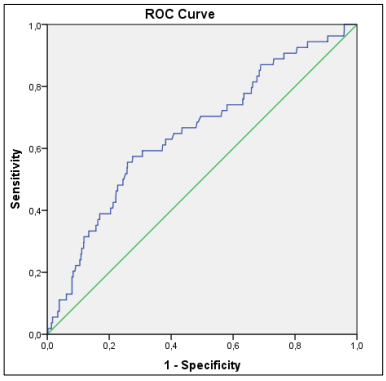

Figura 48 - Curva ROC per l'holdout set

Tabella 93: Area sotto la curva ROC per l'holdout set

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.652	0.039	0.000	0.575	0.729

5.2.3. SONCA con distribuzione di probabilità gaussiana

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo gaussiana. In Tabella 94 sono riportate le distribuzioni di frequenza per il set di training.

Tabella 94: Distribuzioni di classe per il training set: SONCA e distribuzione di probabilità gaussiana

Variabile	N	%
0	800	50.70
1	778	49.30
Totale	1578	100.00

Albero di regressione con SONCA e distribuzione di probabilità gaussiana

In Figura 49 e Figura 50 sono riportati gli alberi di regressione training e holdout set avendo implementato SONCA con distribuzione gaussiana.

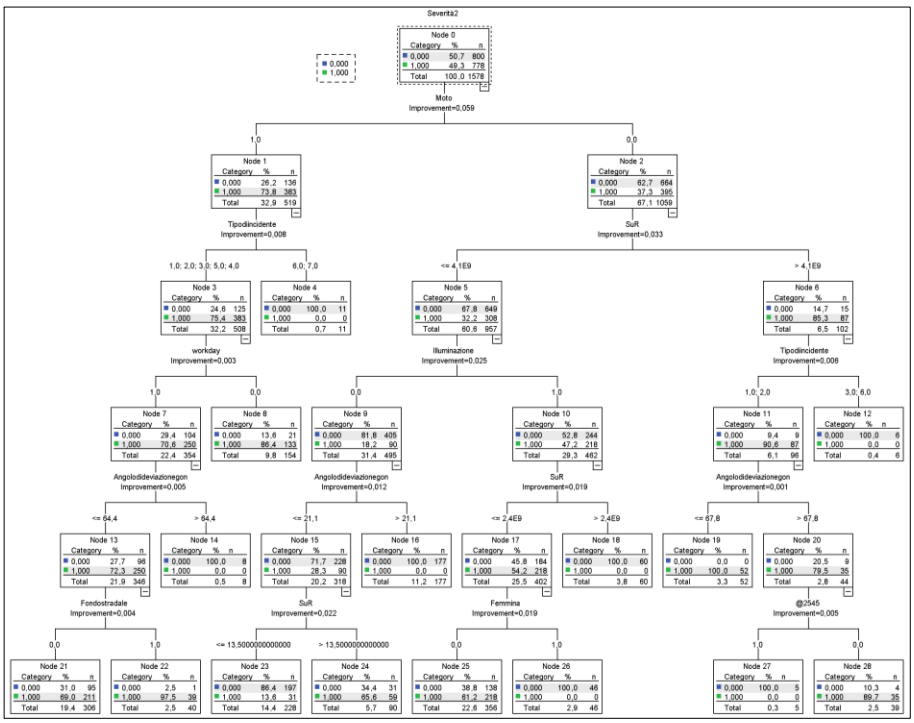

Figura 49 – Albero di regressione per il training set: SONCA e distribuzione gaussiana

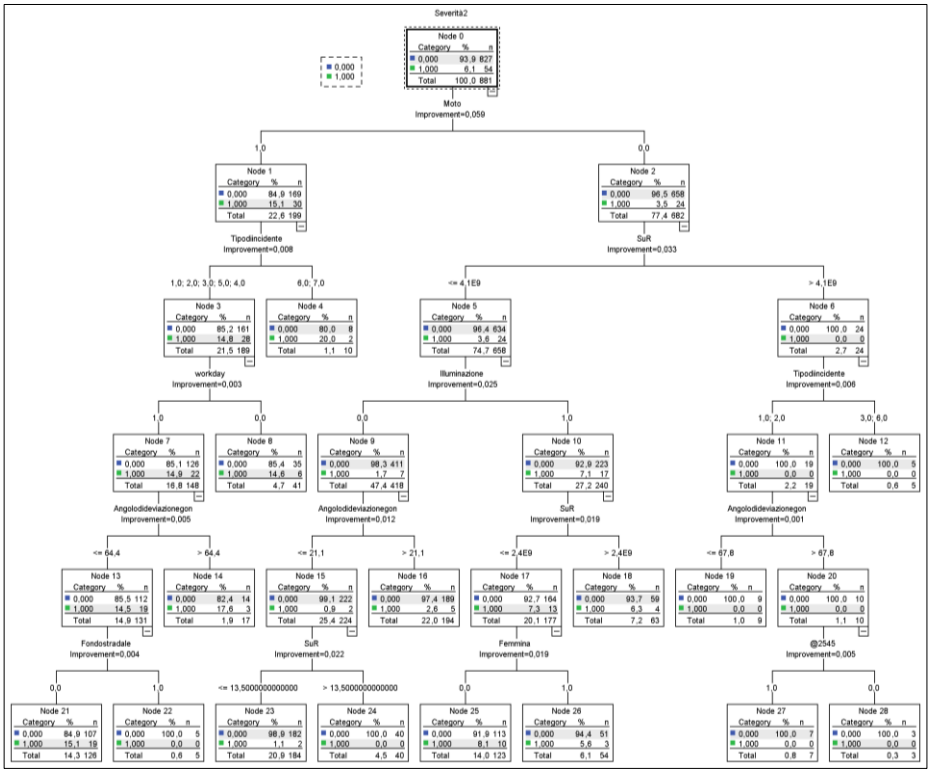

Figura 50 – Albero di regressione per l'holdout set: SONCA e distribuzione gaussiana

In Tabella 95 e Tabella 96 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 95: Matrice di confusione - CART: SONCA

Sample		Predicted		
		1	0	Percent Correct
Training	1	747	31	96.0%
	0	290	510	63.8%
	Overall Percentage	65.7%	34.3%	79.7%
Holdout	1	35	19	64.8%
	0	312	515	62.3%
	Overall Percentage	39.4%	60.6%	62.4%

Tabella 96: Misure di performance - CART: SONCA

Dataset	FNrate	FPrate	TNrate	TPrate	Precision	F-measure	G-mean	Err	Acc	AUC
Training	0.040	0.363	0.638	0.960	0.720	0.823	0.782	0.203	0.797	
Holdout	0.352	0.377	0.623	0.648	0.101	0.175	0.635	0.376	0.624	0.862

In Figura 51 e Tabella 97 sono riportate la curva ROC e l'area sottesa alla curva ROC.

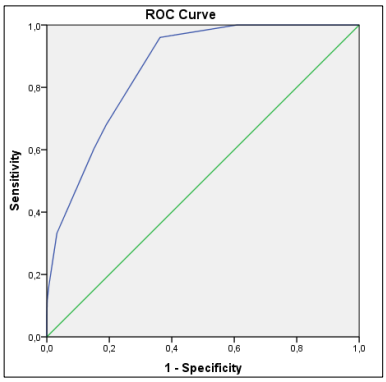

Figura 51 - Curva ROC per l'holdout set

Tabella 97: Area sotto la curva ROC per l'holdout set

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.862	0.009	0.000	0.844	0.880

Logit con SONCA e distribuzione di probabilità gaussiana

In Tabella 98 sono riportate le variabili nell'equazione.

Tabella 98: Variabili nell'equazione: SONCA con distribuzione di probabilità gaussiana

Variabili	B	S.E.	Wald	df	Sig.	Exp(B)
workday(1)	0.401	0.125	10.336	1	0.001	1.493
Sezionetrasversale			30.485	6	0.000	
Sezionetrasversale(1)	22.462	16,386.126	0.000	1	0.999	5,687,837,399.603
Sezionetrasversale(2)	20.416	16,386.126	0.000	1	0.999	735,525,274.990
Sezionetrasversale(3)	20.928	16,386.126	0.000	1	0.999	1,226,647,539.941
Sezionetrasversale(4)	21.004	16,386.126	0.000	1	0.999	1,323,777,980.978
Sezionetrasversale(5)	20.439	16,386.126	0.000	1	0.999	752,392,526.243
Sezionetrasversale(6)	21.247	16,386.126	0.000	1	0.999	1,688,830,288.231
Geometriaorizzontale(1)	- 0.693	0.123	31.604	1	0.000	0.500
Illuminazione(1)	- 0.475	0.125	14.364	1	0.000	0.622
Tipodiincidente			80.600	6	0.000	
Tipodiincidente(1)	0.765	0.457	2.806	1	0.094	2.149
Tipodiincidente(2)	0.489	0.462	1.121	1	0.290	1.631
Tipodiincidente(3)	0.060	0.485	0.015	1	0.902	1.062
Tipodiincidente(4)	0.517	0.506	1.042	1	0.307	1.676
Tipodiincidente(5)	2.238	0.493	20.649	1	0.000	9.377
Tipodiincidente(6)	- 20.329	4,638.130	0.000	1	0.997	0.000
Incidentesecondario(1)	- 0.940	0.120	61.220	1	0.000	0.391
Speedrelated(1)	0.424	0.118	12.903	1	0.000	1.529
Constant	- 20.710	16,386.126	0.000	1	0.999	0.000

In Tabella 99 e Tabella 100 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 99: Matrice di confusione - Logit: SONCA con distribuzione di probabilità gaussiana

Sample		Predicted		
		1	0	Percent Correct
Training	1	515	263	66.2%
	0	263	537	67.1%
	Overall Percentage	49.3%	50.7%	66.7%
Holdout	1	32	22	59.3%
	0	309	518	62.6%
	Overall Percentage	21.6%	34.2%	62.4%

Tabella 100: Misure di performance – Logit: SONCA con distribuzione di probabilità gaussiana

Dataset	FNrate	FPrate	TNrate	TPrate	Precision	F-measure	G-mean	Err	Acc	AUC
Training	0.338	0.329	0.671	0.662	0.662	0.662	0.667	0.333	0.667	
Holdout	0.407	0.374	0.626	0.593	0.094	0.162	0.609	0.376	0.624	0.661

Analogamente a quanto visto per il CART, il Logit stimato dopo aver ricampionato il training set con SONCA usando una distribuzione di probabilità gaussiana mostra dei miglioramenti. In tale modello, si osserva che FN_{rate} diminuisce, passa da 0.833 nel modello senza ricampionamento a 0.407, si ha così una riduzione del 51%. Aumentano anche le altre misure di performance, in particolare G-mean=0.609 e AUC=0.661.

5.2.4. Confronto dei risultati A56 crashes

Confrontando i risultati del CART tra le due distribuzioni di frequenza (Tabella 101) si osservano migliori risultati con la distribuzione di probabilità triangolare, FN_{rate} =0.185 per la distribuzione triangolare, mentre FN_{rate} =0.352 per la distribuzione gaussiana. Per la distribuzione triangolare, il tasso di casi positivi erroneamente classificati come appartenenti alla classe negativa si riduce del 81%, mentre per la distribuzione gaussiana si riduce del 65%. Contemporaneamente si osserva che il tasso di corretti positivi sono più alti usando una distribuzione di probabilità triangolare piuttosto che quella gaussiana (TP_{rate}=0.815 vs TPrate=0.648). I valori di Precision e F-measure non sono molto

differenti tra di loro, ma si osserva che i valori di G-mean e AUC sono più alti per la distribuzione triangolare.

Tabella 101: Confronto dei risultati PTW ISTAT - CART

Dataset	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
Originale	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.061	0.939	0.000
SONCA (Triangolare)	0.185	0.459	0.541	0.815	0.104	0.184	0.664	0.443	0.557	0.712
SONCA (Gaussiana)	0.352	0.377	0.623	0.648	0.101	0.175	0.635	0.376	0.624	0.862

Diversamente da quanto visto sino ad ora per l'A56, confrontando i risultati del Logit tra le due distribuzioni di frequenza (Tabella 102), si osserva che non ci sono forti differenze. Il tasso di casi positivi erroneamente classificati come appartenenti alla classe negativa sono uguali (FN_{rate} =0.407), anche i valori di Precision, F-measure, G-mean e AUC non sono molto differenti tra loro.

Tabella 102: Confronto dei risultati PTW ISTAT - Logit

Dataset	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	G-mean	Err	Acc	AUC
Originale	0.833	0.255	0.745	0.167	0.000	0.000	0.352	0.291	0.709	0.439
SONCA (Triangolare)	0.407	0.369	0.631	0.593	0.095	0.164	0.612	0.371	0.629	0.652
SONCA (Gaussiana)	0.407	0.374	0.626	0.593	0.094	0.162	0.609	0.376	0.624	0.661

Analogamente a quanto visto per i dati PTW crashes, confrontando i risultati tra il CART e il Logit, si osserva che il CART offre dei risultati migliori

Conclusioni

Dall'analisi della letteratura è stato osservato che in presenza di una distribuzione della variabile di risposta estremamente sbilanciata il processo di apprendimento può essere distorto.

Le tecniche di ricampionamento hanno ricevuto una notevole attenzione per superare i problemi legati a dataset con variabile di risposta sbilanciate. Queste agiscono come una fase di pre-elaborazione, consentendo al sistema di apprendimento di ricevere le osservazioni, come se appartenessero a un insieme di dati ben equilibrato. Le tecniche più comuni sono il *random oversampling* e il *random undersampling*. Diversi autori concordano sul fatto che il *random oversampling* può aumentare la probabilità che si verifichino problemi di overfitting, inoltre aumenta il costo computazionale del processo di apprendimento accrescendo in maniera massiccia la dimensione della matrice dei dati (in merito al numero di osservazioni da trattare). Il principale inconveniente del *random undersampling* è che questo metodo può scartare dati potenzialmente utili che potrebbero essere importanti per il processo di apprendimento. Nel corso degli anni, molte tecniche sono state sviluppate con l'obiettivo di superare i limiti del random sampling, è stato osservato che la creazione di database sintetici risolve questi limiti cercando di generalizzare la regione di decisione della classe di minoranza.

Tali approcci proposti non consentono l'elaborazione d'insiemi di dati in cui sono contemporaneamente presenti predittori di natura numerica e categorica, e le proposte sino ad ora presenti in letteratura per ovviare a questo limite non sono apparse soddisfacenti.

Nel presente lavoro è stata presentata una nuova metodologia di *synthetic sampling*, chiamato *"Synthetic Over-sampling for Numerical and Categorical variables (SONCA) "* che possa essere utilizzato con dataset caratterizzati dalla presenza di predittori di natura eterogenea. Dato un dataset costituito dal vettore risposta, $Y_{n\times1}$, e dalla matrice dei predittori, $X_{n\times q}$. Dove $Y_{n\times1}$ si manifesta con k=C+1 classi, mentre X è costituita da due sotto matrici X_{Num}, matrice dei p predittori numeri, e X_{Cat}, matrice dei q-p predittori categorici. Supponendo che una delle classi di Y sia sottorappresentata, attraverso SONCA è possibile ottenere un nuovo dataset sintetico affinché la variabile di risposta sia bilanciata per ogni classe.

Dalla comparazione dei risultati ottenuti per i diversi dataset appartenenti alla banca dati dell'UCI Machine Learning Repository, è stato evidenziato che:

- Usando un dataset con variabile di risposta molto sbilanciato, si hanno dei modelli di stima poco accurati, con un alto tasso di casi negativi erroneamente classificati come appartenenti alla classe positiva, FN_{rate}, e un basso tasso di casi positivi correttamente classificati come appartenenti alla classe positiva, TP_{rate}, insieme a bassi valori delle misure di performance come F-measure, G-mean e AUC.
- Utilizzando gli algoritmi di bilanciamento, alla presenza di dataset con variabile di risposta estremamente sbilanciata, in particolare se lo squilibrio è assoluto, i modelli di stima sono più accurati. Le diverse misure di performance migliorano: FN_{rate} si riduce, mentre TP_{rate} aumenta insieme a F-measure, G-mean e AUC. Contemporaneamente al aumento dei TPrate si ha anche un aumento dei falsi positivi, riducendo così la Precision.
- Dal confronto dell'algoritmo di SONCA con SMOTE e ROSE, è stato visto che i migliori valori delle misure di performance sono ottenuti utilizzando SONCA.

- SONCA è efficace sia per dataset con predittori numerici, sia con predittori di natura sia numerica sia categorica.
- Confrontando le misure di performance ottenute dai modelli di stima usando SONCA con le due differenti distribuzioni di probabilità non si riscontrano grandi variazioni.
- Al variare di m non è stata riscontrata una forte variazione delle misure di performance, così da poter utilizzare anche rapporti di bilanciamento bassi senza aggiungere un ulteriore costo computazionale.

Inoltre analizzando i risultati ottenuti utilizzando SONCA su 2 dataset reali che riguardano il problema dello studio delle determinanti degli incidenti stradali mortali, è stato osservato che:

- Sia per il CART sia per il Logit il tasso di casi positivi erroneamente classificati come appartenenti alla classe negativa si riduce sia per la distribuzione triangolare sia per la distribuzione gaussiana. Contemporaneamente si osserva che il tasso di corretti positivi sono più alti, TPrate sia per la distribuzione di probabilità triangolare sia per la distribuzione di probabilità gaussiana. I valori di G-mean e AUC aumentano notevolmente.
- Confrontando i risultati tra il CART e il Logit, si osserva che il CART offre dei risultati migliori. Il CART non richiede alcuna assunzione sul variabile obiettivo e nemmeno con riguardo alle relazioni fra le variabili in esame si impongono ipotesi specifiche. Tutto questo rende la tecnica più flessibile e potenzialmente più efficiente delle metodologie tradizionali.

BIBLIOGRAFIA

Ahmad, A., Dey, L., 2007. A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. Pattern Recognition Letters, Vol 28(1), pp. 110-118.

Azzalini, A., Scarpa, B., 2009. Analisi dei dati e data mining. Springer

ANCI-UPI; 2010. La Sicurezza Stradale: Dal Quadro Europeo E Nazionale Alle Buone Pratiche Locali. Osservatorio Nazionale delle Autonomie Locali sulla Sicurezza Stradale ANCI-UPI.

Azimi, M., Zhan, Y.; 2010. *Categorizing Freeway Flow Conditions Using Clustering Methods*. Transportation Research Record: Journal of the Transportation Research Board, Vol. 2173, pp. 105-114.

Bekkar, M., Alitouche, T. A., 2013. Imbalanced data learning approaches review. International Journal of Data Mining & Knowledge Management Process, Vol 3 (4), pp. 15-33.Berkson, J.,1944. Application of the logistic function to bio-assay. Journal of the American Statistical Association, Vol. 39 (227), pp. 357-365.

Bottarelli, E., Parodi, S., 2003. Un approccio per la valutazione della validità dei test diagnostici: le curve R.O.C. (Receiver Operating Characteristic). Ann. Fac. Medic. Vet. di Parma, Vol. 23, pp. 49-68.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J.,1984. Classification and regression trees. Wadsworth & Brooks. Monterey, CA.

Cao, H., Li, X. L., Woon, Y. K., & Ng, S. K., 2011. SPO: Structure preserving oversampling for imbalanced time series classification. In Data Mining (ICDM), 2011 IEEE 11th International Conference, pp. 1008-1013.

Chawla, N. V., Bowyer, Hall, L. O., Kegelmeyer W. P.. 2002. Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, Vol. 16, pp. 321–357.

Chawla, N. V., Japkowicz, N., and Kołcz, A., 2004. SIGKDD Special Issue on Learning from Imbalanced Dataset.

Chawla, N.V., 2003. C4.5 and Imbalanced Data sets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure Workshop on Learning from Imbalanced Dataset II, ICML, Washington DC, 2003.

Chawla, N.V., Lazarevic, A., O. Hall L., Bowyer, K., 2003. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. In Proceedings of the Principles of Knowledge Discovery in Databases, PKDD-2003.

Chen, C., A., Liaw, L., Breiman, 2004. Using random forests to learn unbalanced data. Technical Report 666, Statistics Department, University of California at Berkeley.

Choi, J. M., 2010. A selective sampling method for imbalanced data learning on support vector machines. Tesi di dottorato presso Iowa State University.

COM, Commission for Global Road Safety; 2010a. *Un decennio di iniziative per la sicurezza stradale*.

COM, Commissione delle Comunità Europee; 2010b. *Verso uno spazio europeo della sicurezza stradale: orientamenti 2011-2020 per la sicurezza stradale*. COM(2010) 389/3

Cost, S., Salzberg, S., 1993. A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. *Machine Learning,* vol. 10, no. 1, pp. 57-78.

Das, G., Mannila, H., 2000. Context-based similarity measures for categorical databases. In *Principles of Data Mining and Knowledge Discovery*, pp. 201-210. Springer Berlin Heidelberg.

Dulli, S., Furini, S., Data mining, Peron, E., 2009. Metodi e strategie. Springer.

Ding, Z., 2011. Diversified Ensemble Classifiers for Highly Imbalanced Data Learning and their Application in Bioinformatics.

Dietterich, T., Margineantu, D., Provost, F., and Turney, P., editors (2003).Proceedings of the ICML'2000 Workshop on COST-SENSITIVE LEARNING.

Esposito, F., Malebra, D., Tamma, V., Bock, H.H., 2002. Classical resemblance measures. Analysis of Symbolic Data. Springer, pp. 139-152.

Emerson, D., Nayak, R., Weligamage, J., 2011. Using data mining to predict road crash count with a focus on skid resistance values.

Ferri, C., Flach, P., Orallo, J., and Lachice, N., editors (2004).ECAI' 2004 First Workshop on ROC Analysis in AI. ECAI.

Ganganwar, V., 2012. An overview of classification algorithms for imbalanced. International Journal of Emerging Technology and Advanced Engineering, Vol. 2 (4), pp. 42-47.

Greenacre, M., 2008. Measures of distance between samples: Euclidean. http://www.econ.upf.edu/~michael/stanford/

Guo X., Yin, Y., Dong, C., Yang, G., e G. Zhou., 2008. On the Class Imbalance Problem. Fourth International Conference on Natural Computation.

Han, H., Wang, W. Y., Mao, B.H, 2005 Borderline-smote: A new over-sampling method in imbalanced data sets learning. In Advances in Intel-ligent Computing, pp. 878-887.

He, H., Bai, Y., Garcia, E., & Li, S. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of international joint conference on neural networks pp. 1322–1328.

He, H., Garcia, E. A., 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, VOL. 21 (9), pp. 1263-1284.

Huang, J., Ling, CX., 2005. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng, Vol. 17(3), pp. 299–310

Japkowicz, N. (2000b). Learning from Imbalanced Data sets: A Comparison of Various Strategies. In Proceedings of the AAAI'2000 Workshop on Learning from Imbalanced Data Sets, Austin, TX.

Keel, Knowledge Extraction based on Evolutionary Learning, 2013. http://sci2s.ugr.es/keel/algorithms.php#sub4.

Kotsiantis, S., Kanellopoulos, D., Pintelas, P. (2006). Handling imbalanced dataset: A review. GESTS International Transactions on Computer Science and Engineering, Vol. 30.

Liu, W., Chawla, S., Cieslak, D. A., Chala, N. V. (2010). A Robust Decision Tree Algorithm for Imbalanced Data Sets. SDM. Vol. 10, pp. 766-777.

Liu, X.Y., Wu, J., Zhou, Z.H., (2009). Exploratory Under Sampling for Class Imbalance Learning. IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, VOL. 39(2).

Menardi, G. (2009). Some Issues Emerging In Evaluating the Risk of Default for SMEs. S. Co. 2009. Sixth Conference. Complex Data Modeling and Prediction.

Menardi, G., Torelli, N., 2012. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discover.*

Montella A., Lista A., Mauriello F., 2008. L'incidentalità nei tratti urbani delle strade provinciali e statali. XVII Convegno. Nazionale della Società Italiana di Infrastrutture Viarie - Le Reti di Tras porto Urbano. Progettazione, Costruzione, Gestione, Enna

Montella A., Andreassen D., Tarko A., Turner S., Mauriello F., Imbriani L. L., Singh R., 2012. Critical Review of the International Crash Databases and Proposals for Improvement of the Italian National Databases. Procedia - Social and Behavioral Sciences, Vol. 53, pp. 49-61.

Nayak, R., Emerson, D., Weligamage, J., Piyatrapoomi, N. 2011. Road crash proneness prediction using data mining. In Proceedings of the 14th International Conference on Extending Database Technology, pp. 521-526, ACM.

Ndour, C., Dossou-Gbété, S., 2012. Classification approach based on association rules mining for unbalanced data. arXiv preprint arXiv:1202.5514.

Ramentol, E., Caballero, Y., Bello, R., Herrera, F., 2012. SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowledge and Information Systems, 33(2), 245-265.

Rocchi, M.B.L., 2001. La valutazione di un esame diagnostico mediante curve ROC: Alcune osservazioni. Biochimica clinica, vol., 25, pp. 382-389.

Schultz, M., Joachims, T., 2003. Learning a Distance Metric from Relative Comparisons. In Neural Information Processing Systems.

Suarez-Alvarez, M. M., Pham, D. T., Prostov, M. Y., Prostov, Y. I., 2012. Statistical approach to normalization of feature vectors and clustering of mixed datasets. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, Vol. *468*(2145), pp. 2630-2651.

Sun, Y., Kamel, M. S., Wong, A. K., & Wang, Y.,2007. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition, Vol. 40(12), pp. 3358-3378.

Tesema, T. B., Abraham, A., Grosan, C., 2005. Rule mining and classification of road traffic accidents using adaptive regression trees. International Journal of Simulation, Vol., 6(10), 80-94.

Torrão, G. A., Coelho, M. C., Rouphail, N. M., 2014. Modeling the impact of subject and opponent vehicle on crash severity in two-vehicle collisions. In Proceedings of the 93rd Annual Meeting of the Transportation Research Board.

Zani, S. e Cerioli, A. (2007): Analisi dei dati e data mining per le decisioni aziendali. Giu®rµe editore, Milano.

Wang, H., 2006. Nearest neighbors by neighborhood counting. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol *28* (6), pp. 942-953.

Weller, G. , Schlag, B. , Gatti, G. , Jorna, R. ; van de Leur, M. , 2006. *Human Factors in Road Design. State of the art and empirical evidence.* RIPCORD–ISEREST

WHO, World Health Organization, 2004. World report on road traffic injury prevention. Ginevra

Weiss, G. M. , e Provost F., (2003). Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. Journal of Artificial Intelligence Research, Vol. 19, pp. 315-354.

Weng, C. G., Poon, J. 2008. A new evaluation measure for imbalanced datasets. In Proceedings of the 7th Australasian Data Mining Conference, Australian Computer Society, Inc., Vol. 87, pp. 27-32.

Appendice 1 – Analisi della sensibilità del parametro m

A.1.1. Cover type

Come è stato osservato, la variabile di risposta cover type è estremamente sbilanciata, la modalità 0 (la classe di maggioranza) ha una frequenza del 92.8%, mentre la classe 1 (la classe di minoranza) ha una frequenza del 7.1% inferiore del 10%. Per tale dataset sono stati stimati l'albero di regressione e il logit nei seguenti casi:

- Dataset originale;
- SONCA con distribuzione di probabilità triangolare e con m=3500;
- SONCA con distribuzione di probabilità triangolare e con m=5500;
- SONCA con distribuzione di probabilità triangolare e con m=7500;
- SONCA con distribuzione di probabilità triangolare e con m=9500;
- SONCA con distribuzione di probabilità gaussiana e con m=3500;
- SONCA con distribuzione di probabilità gaussiana e con m=5500;
- SONCA con distribuzione di probabilità gaussiana e con m=7500;
- SONCA con distribuzione di probabilità gaussiana e con m=9500.

Dataset originale

Albero di regressione senza il ricampionamento

In Figura 52 e Figura 53 sono riportati gli alberi di regressione senza bilanciare la variabile di risposta.

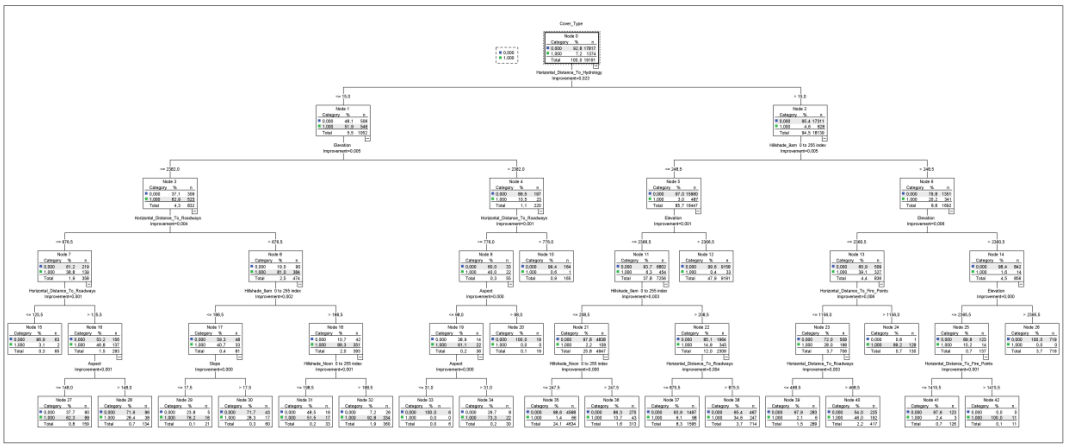

Figura 52 – Albero di regressione per il training set senza ricampionamento

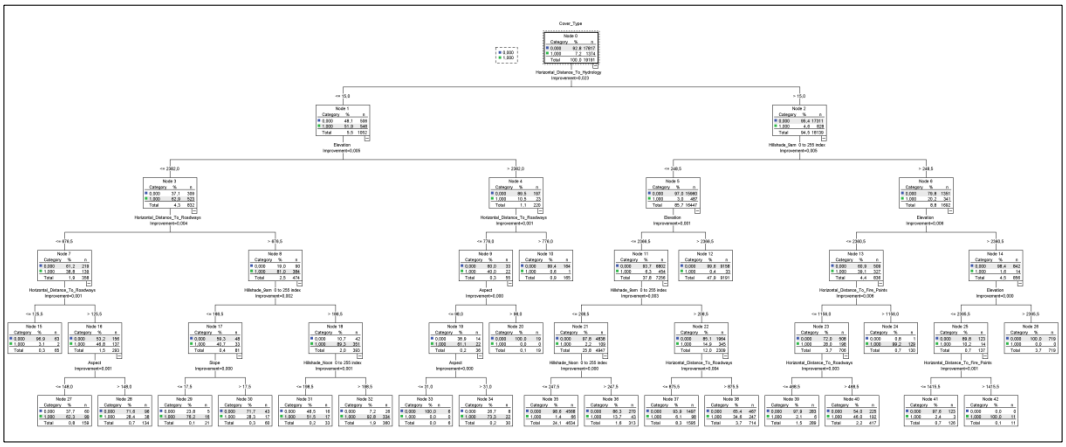

Figura 53 – Albero di regressione per l'holdout set senza ricampionamento

In Tabella 104 e in Tabella 105 sono riportate sia la matrice di confusione sia le misure di prestazioni per entrambi i set dati, training e holdout.

Tabella 103: Matrice di confusione senza ricampionamento

Sample	Observed	Predicted		
		0	1	Percent Correct
Training	0	17701	116	99,3%
	1	746	628	45,7%
	Overall Percentage	96,1%	3,9%	95,5%
Holdout	0	17799	138	99,2%
	1	746	627	45,7%
	Overall Percentage	96,0%	4,0%	95,4%

Tabella 104: Misure di performance senza ricampionamento

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.543	0.007	0.993	0.457	0.844	0.593	0.045	0.955
Holdout	0.543	0.008	0.992	0.457	0.820	0.587	0.046	0.954

In Figura 54 e in Tabella 105 sono riportate la curva ROC e l'area sottesa alla curva ROC.

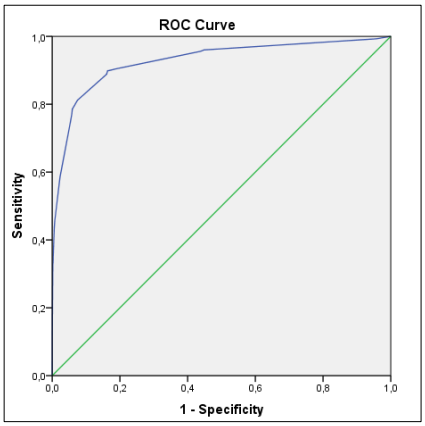

Figura 54 - Curva ROC per l'holdout set senza ricampionamento

Tabella 105 - Area sotto la curva ROC per l'holdout set senza ricampionamento

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.927	0.004	0.000	0.919	0.936

Logit senza il ricampionamento

In Tabella 106 sono riportate le variabili nell'equazione.

Tabella 106: Variabili nell'equazione per il training set senza ricampionamento

	B	S.E.	Wald	df	Sig.	Exp(B)
Elevation	-0.011	0.000	1210.839	1	0.000	0.989
Slope	-0.052	0.005	102.605	1	0.000	0.949
Horizontal_Distance_To_Hydrology	-0.008	0.001	190.233	1	0.000	0.992
Vertical_Distance_To_Hydrology	0.016	0.001	156.222	1	0.000	1.017
Horizontal_Distance_To_Roadways	0.002	0.000	419.905	1	0.000	1.002
Hillshade_9am0to255index	0.036	0.001	726.456	1	0.000	1.037
Hillshade_Noon0to255index	0.014	0.002	61.461	1	0.000	1.014
Horizontal_Distance_To_Fire_Points	0.001	0.000	181.593	1	0.000	1.001
Constant	11.095	.749	219.302	1	0.000	65849.961

In Tabella 107 e in Tabella 108 sono riportate la matrice di confusione che le misure di prestazioni sia per il dataset di training che di holdout.

Tabella 107: Matrice di confusione senza ricampionamento

Sample		**Predicted**		
		1	0	Percent Correct
Training	1	238	17579	1.3%
	0	398	976	71.0%
	Overall Percentage	3.3%	96.7%	6.3%
Holdout	1	262	17675	1.5%
	0	392	981	71.4%
	Overall Percentage	3.4%	97.2%	6.4%

Tabella 108: Misure di performance senza ricampionamento

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.987	0.290	0.710	0.013	0.374	0.026	0.937	0.063
Holdout	0.985	0.286	0.714	0.015	0.401	0.028	0.936	0.064

In Figura 55 e in Tabella 109 sono riportate la curva ROC e l'area sottesa alla curva ROC.

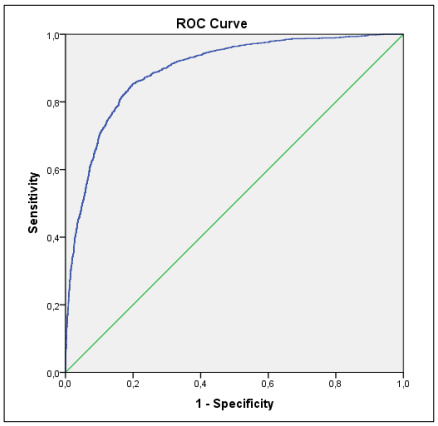

Figura 55 - Curva ROC per l'holdout set senza ricampionamento

Tabella 109 - Area sotto la curva ROC per l'holdout set senza ricampionamento

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.895	0.004	0.000	0.886	0.903

SONCA con distribuzione di probabilità triangolare e m=3500

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo triangolare e m=3500. In Tabella 110 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA con distribuzione di probabilità triangolare e m=3500

Tabella 110: Distribuzioni di classe per il training set: SONCA con distribuzione di probabilità triangolare e m=3500

Variabile	N	%
1	3'491	50.3
0	3'453	49.7
Totale	6'944	100.0

Albero di regressione con il ricampionamento: SONCA con distribuzione di probabilità triangolare e m=3500

In Figura 56 e Figura 57 sono riportati gli alberi di regressione training avendo implementato SONCA con distribuzione di probabilità triangolare e m=3500.

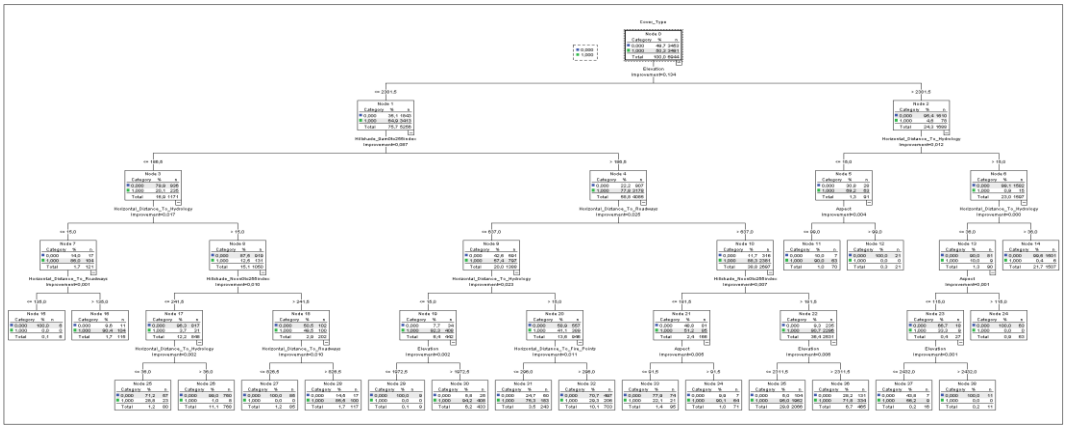

Figura 56 – Albero di regressione per il trainig set: SONCA con distribuzione di probabilità triangolare e m=3500

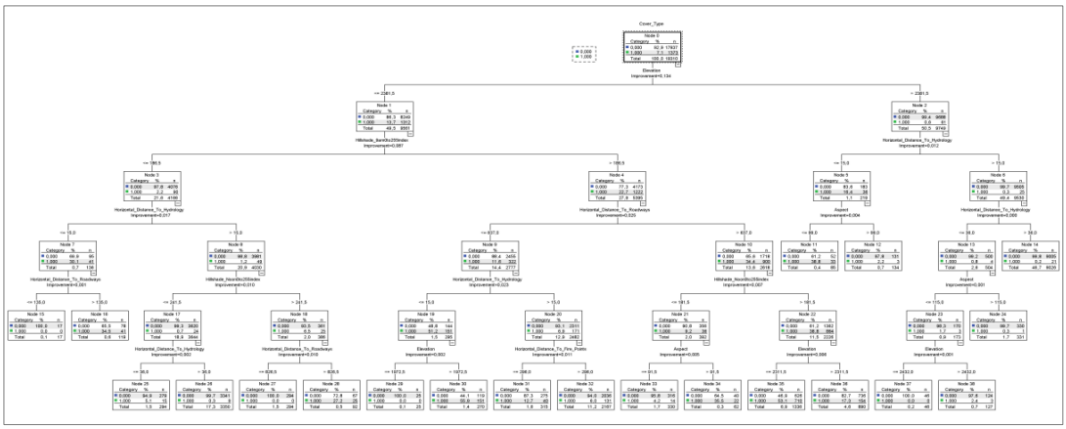

Figura 57 – Albero di regressione per l'holdout set: SONCA con distribuzione di probabilità triangolare e m=3500

In Tabella 111 e Tabella 112 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 111: Matrice di confusione-CART: SONCA con distribuzione di probabilità triangolare e m=3500

Sample		Predicted		
		1	0	Percent Correct
Training	1	3'227	264	92.4%
	0	369	3'084	89.3%
	Overall Percentage	51.8%	48.2%	90.9%
Holdout	1	1'176	197	85.7%
	0	2039	15'898	88.6%
	Overall Percentage	16.6%	83.4%	88.4%

Tabella 112: Misure di performance - CART: SONCA con distribuzione di probabilità triangolare e m=3500

	FN	FP	TN	TP	Precision	F-measure	Err	Acc
Training	0.076	0.107	0.893	0.924	0.897	0.911	0.091	0.909
Holdout	0.143	0.114	0.886	0.857	0.366	0.513	0.116	0.884

In Figura 59 e Tabella 113 sono riportate la curva ROC e Area sotto la curva ROC.

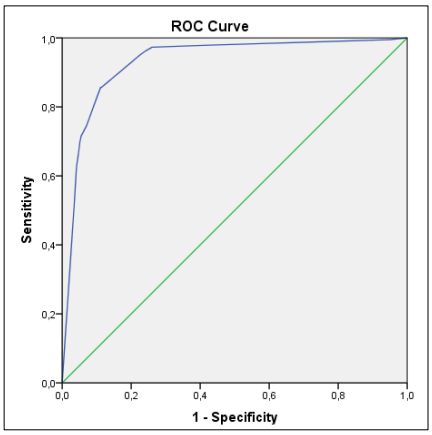

Figura 58 - Curva ROC per l'holdout set - CART: SONCA con distribuzione di probabilità triangolare e m=3500

Tabella 113: Area sotto la curva ROC per l'holdout set - CART: SONCA con distribuzione di probabilità triangolare e m=3500

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.931	0.003	0.000	0.925	0.938

Logit con il ricampionamento: SONCA con distribuzione di probabilità triangolare e m=3500

In Tabella 120 sono riportate le variabili nell'equazione.

Tabella 114: Variabili nell'equazione: SONCA con distribuzione di probabilità triangolare e m=3500

Variables in the Equation	B	S.E.	Wald	df	Sig0.	Exp(B)
Elevation	-0.011	0.000	1151.882	1	0.000	0.989
Slope	-0.110	0.015	57.661	1	0.000	0.896
Horizontal_Distance_To_Hydrology	-0.005	0.000	173.137	1	0.000	0.995
Vertical_Distance_To_Hydrology	0.009	0.001	79.143	1	0.000	1.009
Horizontal_Distance_To_Roadways	0.002	0.000	410.811	1	0.000	1.002
Hillshade_9am0to255index	-0.037	0.015	6.546	1	0.011	0.963
Hillshade_Noon0to255index	0.081	0.012	46.761	1	0.000	1.084
Hillshade_3pm0to255index	-0.062	0.012	27.711	1	0.000	0.939
Horizontal_Distance_To_Fire_Points	0.001	0.000	45.319	1	0.000	1.001
Constant	24.647	2.558	92.805	1	0.000	0.5056E+11

In Tabella 115 e Tabella 116 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 115: Matrice di confusione - Logit: SONCA con distribuzione di probabilità triangolare e m=3500

Sample		Predicted		
		1	0	Percent Correct
Training	1	2984	507	85.5%
	0	589	2864	82.9%
	Overall Percentage	51.5%	48.5%	84.2%
Holdout	1	1109	264	80.8%
	0	3314	14623	81.5%
	Overall Percentage	22.9%	77.1%	81.5%

Tabella 116: Misure di performance - Logit: SONCA con distribuzione di probabilità triangolare e m=3500

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.145	0.171	0.829	0.855	0.835	0.845	0.158	0.842
Holdout	0.192	0.185	0.815	0.808	0.251	0.383	0.185	0.815

In Figura 59 e Tabella 117 sono riportate la curva ROC e Area sotto la curva ROC.

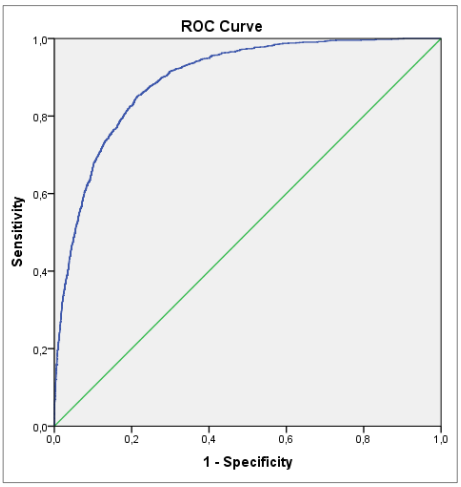

Figura 59 - Curva ROC - Logit: SONCA con distribuzione di probabilità triangolare e m=3500

Tabella 117 - Area sotto la curva ROC - Logit: SONCA con distribuzione di probabilità triangolare e m=3500

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.894	0.004	0.000	0.887	0.902

SONCA con distribuzione di probabilità triangolare e m=5500

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo triangolare e m=5500. In Tabella 118 sono riportate le

distribuzioni di frequenza per il set di training avendo implementato SONCA con distribuzione di probabilità triangolare e m=5500

Tabella 118: Distribuzioni di classe per il training set: SONCA con distribuzione di probabilità triangolare e m=5500

Variabile	N	%
1	5’457	50.0
0	5’453	50.0
Totale	10’910	100.0

Albero di regressione con il ricampionamento: SONCA con distribuzione di probabilità triangolare e m=5500

In Figura 60 e Figura 61 sono riportati gli alberi di regressione training avendo implementato SONCA con distribuzione di probabilità triangolare e m=5500.

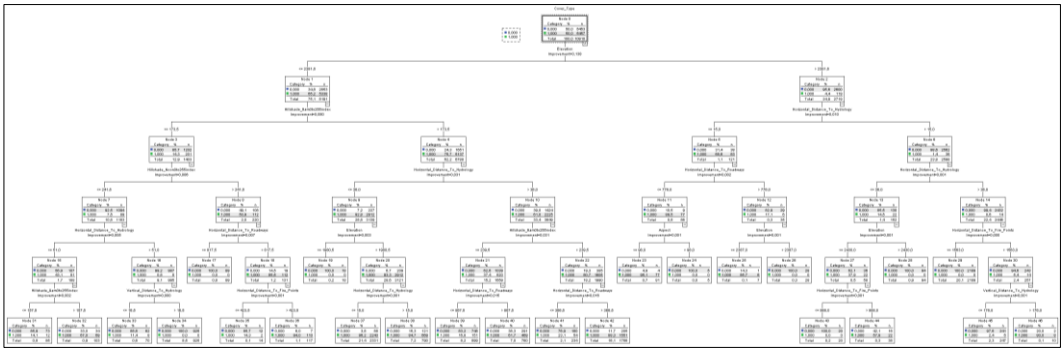

Figura 60 – Albero di regressione per il training set: SONCA con distribuzione di probabilità triangolare e m=5500

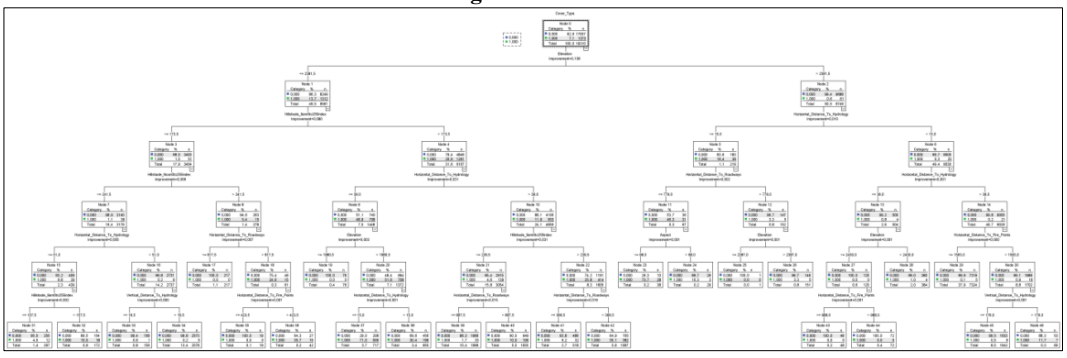

Figura 61 – Albero di regressione per l’holdout set: SONCA con distribuzione di probabilità triangolare e m=5500

In Tabella 119 e Tabella 120 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l’holdout set.

Tabella 119: Matrice di confusione - CART: SONCA con distribuzione di probabilità triangolare e m=5500

Sample		Predicted		
		1	0	Percent Correct
Training	1	5224	233	95.7%
	0	769	4684	85.9%
	Overall Percentage	54.9%	45.1%	90.8%
Holdout	1	1264	109	92.1%
	0	2635	15302	85.3%
	Overall Percentage	20.2%	79.8%	85.8%

Tabella 120: Misure di performance - CART: SONCA con distribuzione di probabilità triangolare con m=5500

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.043	0.141	0.859	0.957	0.872	0.912	0.092	0.908
Holdout	0.079	0.147	0.853	0.921	0.324	0.480	0.142	0.858

In Figura 62 e Tabella 121 sono riportate la curva ROC e Area sotto la curva ROC.

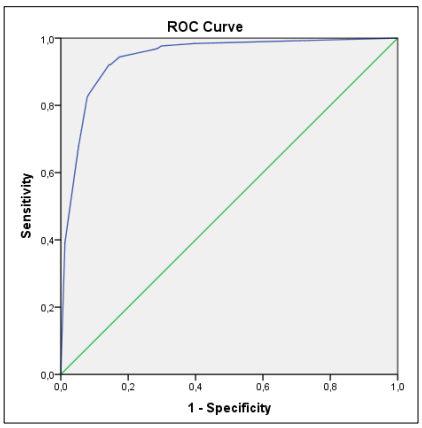

Figura 62 - Curva ROC - CART: SONCA con distribuzione di probabilità triangolare e m=5500

Tabella 121: Area sotto la curva ROC - CART: SONCA con distribuzione di probabilità triangolare e m=5500

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.943	0.003	0	0.937	0.948

Logit con il ricampionamento: SONCA con distribuzione di probabilità triangolare e m=5500

In Tabella 122 sono riportate le variabili nell'equazione.

Tabella 122: Variabili nell'equazione: SONCA con distribuzione di probabilità triangolare e m=5500

Variables in the Equation (9 step)	B	S.E.	Wald	df	Sig.	Exp(B)
Elevation	-0.011	0.000	1774.887	1	0.000	0.989
Slope	-0.080	0.011	53.950	1	0.000	0.923
Horizontal_Distance_To_Hydrology	-0.005	0.000	263.916	1	0.000	0.995
Vertical_Distance_To_Hydrology	0.006	0.001	72.325	1	0.000	1.006
Horizontal_Distance_To_Roadways	0.002	0.000	651.540	1	0.000	1.002
Hillshade_9am0to255index	-0.026	0.011	5.794	1	0.016	0.974
Hillshade_Noon0to255index	0.073	0.009	67.275	1	0.000	1.075
Hillshade_3pm0to255index	-0.052	0.009	34.406	1	0.000	0.949
Horizontal_Distance_To_Fire_Points	0.000	0.000	49.787	1	0.000	1.000
Constant	21.520	1.903	127.857	1	0.000	2.21886E+9

In Tabella 123 e Tabella 124 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 123: Matrice di confusione - Logit: SONCA con distribuzione di probabilità triangolare con m=5500

Sample		Predicted		
		1	0	Percent Correct
Training	1	4221	1236	77.4%
	0	782	4671	85.7%
	Overall Percentage	45.9%	54.1%	81.5%
Holdout	1	975	398	71.0%
	0	2323	15614	87.0%
	Overall Percentage	17.1%	82.9%	85.9%

Tabella 124: Misure di performance - Logit: SONCA con distribuzione di probabilità triangolare con m=5500

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.226	0.143	0.857	0.774	0.844	0.807	0.185	0.815
Holdout	0.290	0.130	0.870	0.710	0.296	0.417	0.141	0.859

In Figura 74 e Tabella 125 sono riportate la curva ROC e Area sotto la curva ROC.

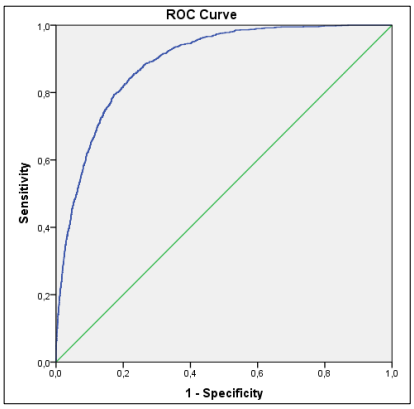

Figura 63 - Curva ROC - Logit: SONCA con distribuzione di probabilità triangolare con m=5500

Tabella 125: Area sotto la curva ROC - Logit: SONCA con distribuzione di probabilità triangolare con m=5500

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.890	0.004	0.000	0.882	0.898

SONCA con distribuzione di probabilità triangolare e m=7500

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo triangolare e m=7500. In Tabella 126 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA con distribuzione di probabilità triangolare e m=7500

Tabella 126: Distribuzioni di classe per il training set: SONCA con distribuzione di probabilità triangolare e m=7500

Variabile	N	%
1	7'428	49.4
0	7'599	50.6
Totale	15'027	100.0

Albero di regressione con il ricampionamento: SONCA con distribuzione di probabilità triangolare e m=7500

In Figura 64 e Figura 65 sono riportati gli alberi di regressione training avendo implementato SONCA con distribuzione di probabilità triangolare e m=7500.

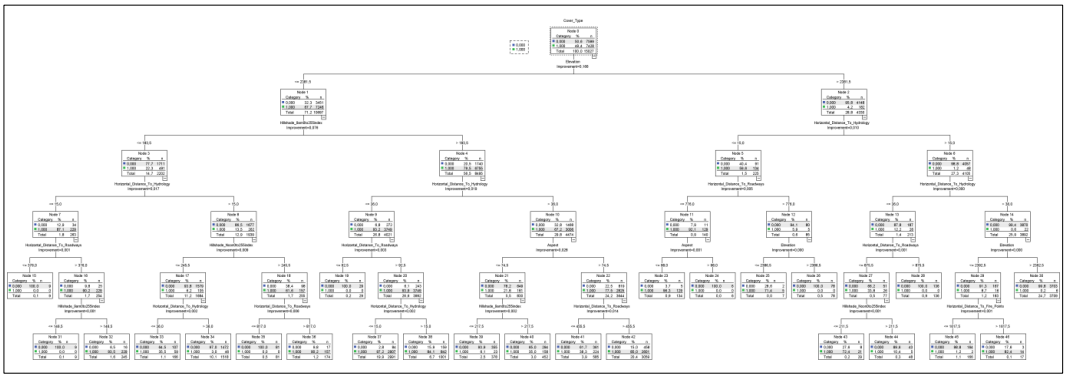

Figura 64 – Albero di regressione per il trainnig set: SONCA con distribuzione di probabilità triangolare e m=7500

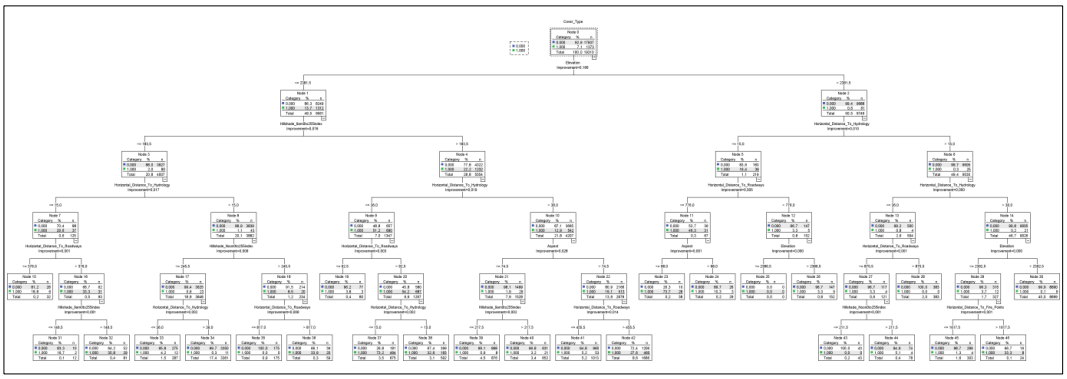

Figura 65 – Albero di regressione per l'holdout set: SONCA con distribuzione di probabilità triangolare con m=7500

In Tabella 127 e Tabella 128 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 127: Matrice di confusione - CART: SONCA con distribuzione di probabilità triangolare e m=7500

Sample		Predicted		
		1	0	Percent Correct
Training	1	6905	523	93.0%
	0	752	6847	90.1%
	Overall Percentage	51.0%	49.0%	91.5%
Holdout	1	1232	141	89.7%
	0	1946	15991	89.2%
	Overall Percentage	16.5%	83.5%	89.2%

Tabella 128: Misure di performance - CART: SONCA con distribuzione di probabilità triangolare con m=7500

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.070	0.099	0.901	0.930	0.902	0.915	0.085	0.915
Holdout	0.103	0.108	0.892	0.897	0.388	0.541	0.108	0.892

In Figura 66e Tabella 129 sono riportate la curva ROC e Area sotto la curva ROC.

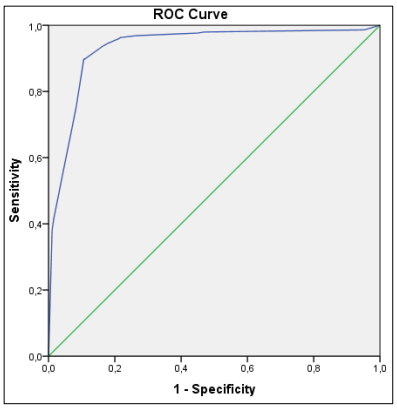

Figura 66 - Curva ROC - CART: SONCA con distribuzione di probabilità triangolare e m=7500

Tabella 129: Area sotto la curva ROC - CART: SONCA con distribuzione di probabilità triangolare e m=7500

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.933	0.004	0.000	0.926	0.940

Logit con il ricampionamento: SONCA con distribuzione di probabilità triangolare con m=5500

In Tabella 130 sono riportate le variabili nell'equazione.

Tabella 130: Variabili nell'equazione: SONCA con distribuzione di probabilità triangolare con m=7500

Variables in the Equation	B	S.E.	Wald	df	Sig.	Exp(B)
Elevation	-0.01	0.00	2680.25	1.00	0.00	0.99
Slope	-0.08	0.01	72.56	1.00	0.00	0.92
Horizontal_Distance_To_Hydrology	-0.01	0.00	339.70	1.00	0.00	0.99
Vertical_Distance_To_Hydrology	0.01	0.00	122.54	1.00	0.00	1.01
Horizontal_Distance_To_Roadways	0.00	0.00	628.74	1.00	0.00	1.00
Hillshade_9am0to255index	-0.02	0.01	5.29	1.00	0.02	0.98
Hillshade_Noon0to255index	0.07	0.01	84.57	1.00	0.00	1.08
Hillshade_3pm0to255index	-0.05	0.01	41.56	1.00	0.00	0.95
Horizontal_Distance_To_Fire_Points	0.00	0.00	20.64	1.00	0.00	1.00
Constant	24.72	1.75	199.95	1.00	0.00	54248709246.52

In Tabella 131 e Tabella 132 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 131: Matrice di confusione - Logit: SONCA con distribuzione di probabilità triangolare e m=7500

Sample		Predicted		
		1	0	Percent Correct
Training	1	6455	973	86.9%
	0	1119	6480	85.3%
	Overall Percentage	50.4%	49.6%	86.1%
Holdout	1	1108	265	80.7%
	0	2972	14965	83.4%
	Overall Percentage	21.1%	78.9%	83.2%

Tabella 132: Misure di performance per- Logit: SONCA con distribuzione di probabilità triangolare con m=7500

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.131	0.147	0.853	0.869	0.852	0.861	0.139	0.861
Holdout	0.193	0.166	0.834	0.807	0.272	0.406	0.168	0.832

In Figura 67 e Tabella 133 sono riportate la curva ROC e Area sotto la curva ROC.

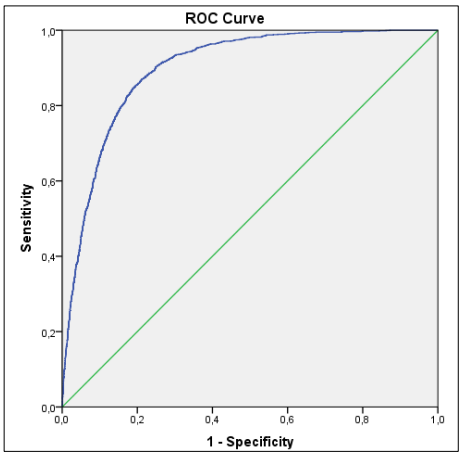

Figura 67 - Curva ROC - Logit: SONCA con distribuzione di probabilità triangolare con m=5500

Tabella 133: Area sotto la curva ROC - Logit: SONCA con distribuzione di probabilità triangolare con m=5500

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.898	0.004	0.000	0.891	0.905

SONCA con distribuzione di probabilità triangolare e m=9500

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo triangolare e m=9500. In Tabella 134 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA con distribuzione di probabilità triangolare e m=9500

Tabella 134: Distribuzioni di classe per il training set: SONCA con distribuzione di probabilità triangolare e m=9500

Variabile	N	%
1	9'445	49.98
0	9'451	50.02
Totale	18896	100.00

Albero di regressione con il ricampionamento: SONCA con distribuzione di probabilità triangolare con m=9500

In Figura 69 e Figura 72 sono riportati gli alberi di regressione training avendo implementato SONCA con distribuzione di probabilità triangolare e m=9500.

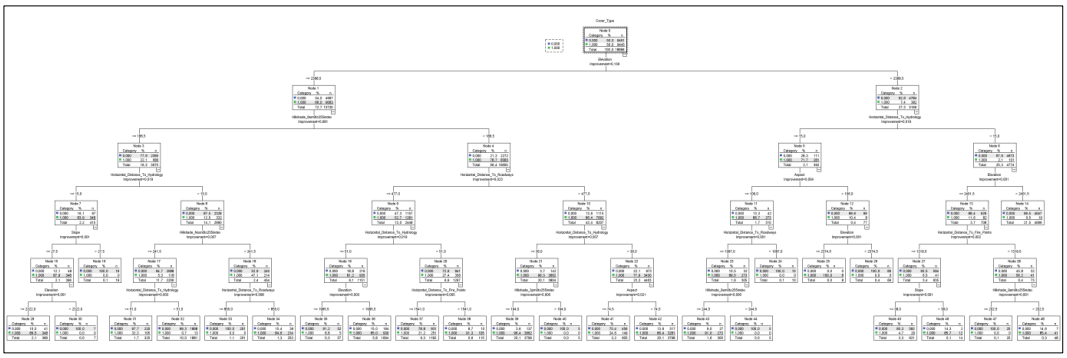

Figura 68 – Albero di regressione per il training set: SONCA con distribuzione di probabilità triangolare e m=9500

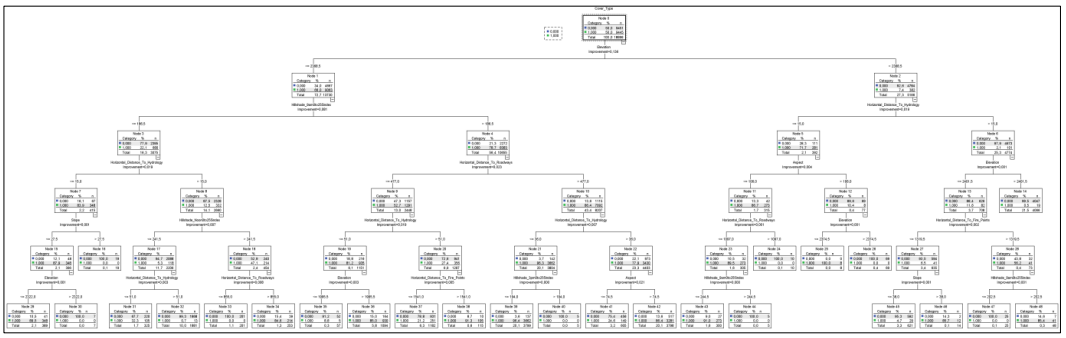

Figura 69 – Albero di regressione per l'holdout set: SONCA con distribuzione di probabilità triangolare e m=9500

In Tabella 135 e Tabella 136 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 135: Matrice di confusione - CART: SONCA con distribuzione di probabilità triangolare e m=9500

Sample		Predicted		
		1	0	Percent Correct
Training	1	8874	571	94.0%
	0	944	8507	90.0%
	Overall Percentage	52.0%	48.0%	92.0%
Holdout	1	1272	101	92.6%
	0	1782	16155	90.1%
	Overall Percentage	15.8%	84.2%	90.2%

Tabella 136: Misure di performance - CART: SONCA con distribuzione di probabilità triangolare e m=9500

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.060	0.100	0.900	0.940	0.904	0.921	0.080	0.920
Holdout	0.074	0.099	0.901	0.926	0.417	0.575	0.098	0.902

In Figura 70 e Tabella 137 sono riportate la curva ROC e Area sotto la curva.

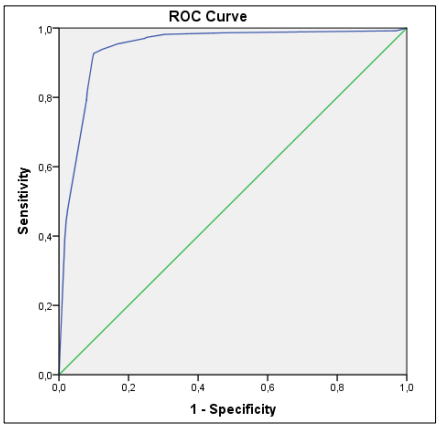

Figura 70 - Curva ROC - CART: SONCA con distribuzione di probabilità triangolare e m=9500

Tabella 137: Area sotto la curva ROC - CART: SONCA con distribuzione di probabilità triangolare con m=9500

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.943	0.003	0.000	0.937	0.949

Logit con il ricampionamento: SONCA con distribuzione di probabilità triangolare e m=9500

In Tabella 138 sono riportate le variabili nell'equazione.

Tabella 138: Variabili nell'equazione: SONCA con distribuzione di probabilità triangolare e m=9500

Variables in the Equation	B	S.E.	Wald	df	Sig.	Exp(B)
Elevation	-0.01	0.00	3135.06	1.00	0.00	0.99
Slope	-0.12	0.01	164.81	1.00	0.00	0.89
Horizontal_Distance_To_Hydrology	-0.01	0.00	544.85	1.00	0.00	0.99
Vertical_Distance_To_Hydrology	0.01	0.00	203.26	1.00	0.00	1.01
Horizontal_Distance_To_Roadways	0.00	0.00	1030.38	1.00	0.00	1.00
Hillshade_9am0to255index	-0.05	0.01	34.02	1.00	0.00	0.95
Hillshade_Noon0to255index	0.10	0.01	164.40	1.00	0.00	1.10
Hillshade_3pm0to255index	-0.08	0.01	102.92	1.00	0.00	0.93
Horizontal_Distance_To_Fire_Points	0.00	0.00	179.84	1.00	0.00	1.00
Constant	27.10	1.61	282.69	1.00	0.00	590119900590.10

In Tabella 139 e Tabella 140 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 139: Matrice di confusione - Logit: SONCA con distribuzione di probabilità triangolare con m=9500

Sample		Predicted		
		1	0	Percent Correct
Training	1	8035	1410	85.1%
	0	1630	7821	82.8%
	Overall Percentage	51.1%	48.9%	83.9%
Holdout	1	1122	251	81.7%
	0	3203	14734	82.1%
	Overall Percentage	22.4%	77.6%	82.1%

Tabella 140: Misure di performance - Logit: SONCA con distribuzione di probabilità triangolare con m=7500

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.149	0.172	0.828	0.851	0.831	0.841	0.161	0.839
Holdout	0.183	0.179	0.821	0.817	0.259	0.394	0.179	0.821

In Figura 71 e Tabella 141 sono riportate la curva ROC e Area sotto la curva ROC.

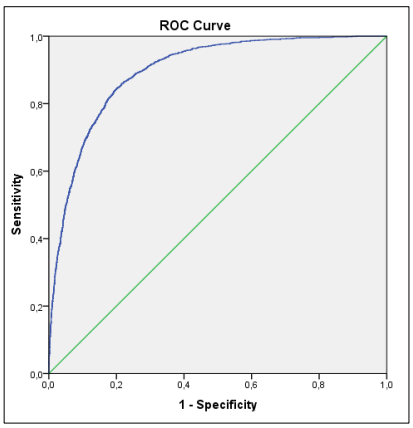

Figura 71 - Curva ROC - Logit: SONCA con distribuzione di probabilità triangolare e m=5500

Tabella 141: Area sotto la curva ROC - Logit: SONCA con distribuzione di probabilità triangolare e m=5500

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.897	0.004	0.000	0.889	0.904

SONCA con distribuzione di probabilità gaussiana e m=3500

Il training set è stato ricampionato utilizzando SONCA, con una distribuzione di probabilità di tipo gaussiano e m=3500. In Tabella 142 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA con distribuzione di probabilità gaussiana e m=3500

Tabella 142: Distribuzioni di classe per il training set: SONCA con distribuzione di probabilità gaussiana e m=3500

Variabile	N	%
1	3’482	49.22
0	3’592	50.78
Totale	7’074	100.00

Albero di regressione con il ricampionamento: SONCA con distribuzione di probabilità gaussiana e m=3500

In Figura 72 e Figura 73 sono riportati gli alberi di regressione training e holdout avendo implementato SONCA con distribuzione di probabilità gaussiana e m=3500.

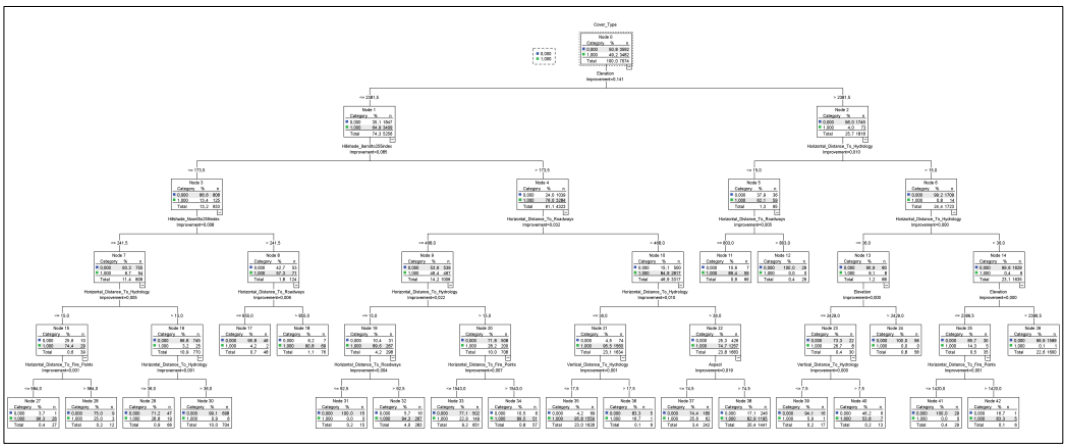

Figura 72 – Albero di regressione per il trainig set: SONCA con distribuzione di probabilità gaussiana e m=3500

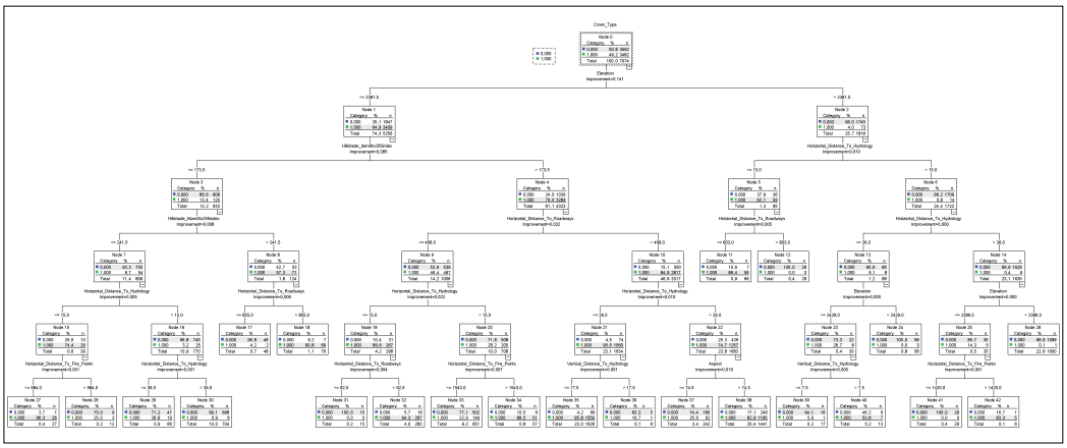

Figura 73 – Albero di regressione per l’holdout set: SONCA con distribuzione di probabilità gaussiana e m=3500

In Tabella 143 e Tabella 144 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 143: Matrice di confusione-CART: SONCA con distribuzione di probabilità gaussiana e m=3500

Sample		Predicted		
		1	0	Percent Correct
Training	1	3238	244	93.0%
	0	359	3233	90.0%
	Overall Percentage	50.8%	49.2%	91.5%
Holdout	1	1247	126	90.8%
	0	2043	15894	88.6%
	Overall Percentage	17.0%	83.0%	88.8%

Tabella 144: Misure di performance - CART: SONCA con distribuzione di probabilità gaussiana e m=3500

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.070	0.100	0.900	0.930	0.900	0.915	0.085	0.915
Holdout	0.092	0.114	0.886	0.908	0.379	0.535	0.112	0.888

In Figura 74 e Tabella 145 sono riportate la curva ROC e Area sotto la curva ROC.

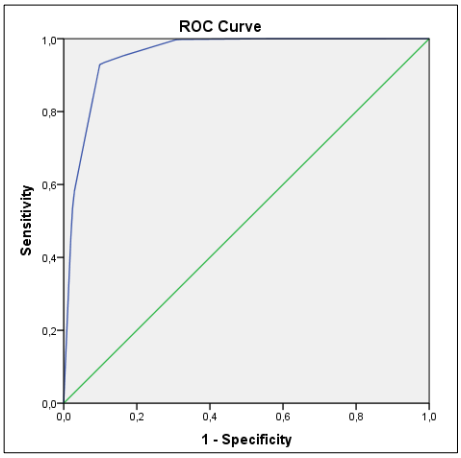

Figura 74 - Curva ROC - CART: SONCA con distribuzione di probabilità gaussiana e m=3500

Tabella 145: Area sotto la curva ROC - CART: SONCA con distribuzione di probabilità gaussiana e m=3500

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.956	0.002	0.000	0.951	0.960

Logit con il ricampionamento: SONCA con distribuzione di probabilità gaussiana e m=3500

In Tabella 146 sono riportate le variabili nell'equazione.

Tabella 146: Variabili nell'equazione: SONCA con distribuzione di probabilità gaussiana e m=3500

Variabili nell'equazione	B	S0.E0.	Wald	df	Sig0.	Exp(B)
Elevation	-0.011	0.000	1144.119	1	0.000	0.989
Slope	-0.122	0.014	71.171	1	0.000	0.885
Horizontal_Distance_To_Hydrology	-0.006	0.000	195.629	1	0.000	0.994
Vertical_Distance_To_Hydrology	0.008	0.001	65.341	1	0.000	1.008
Horizontal_Distance_To_Roadways	0.002	0.000	485.144	1	0.000	1.002
Hillshade_9am0to255index	-0.059	0.015	16.494	1	0.000	0.943
Hillshade_Noon0to255index	0.098	0.012	67.857	1	0.000	1.102
Hillshade_3pm0to255index	-0.079	0.012	44.108	1	0.000	0.924
Horizontal_Distance_To_Fire_Points	0.001	0.000	37.295	1	0.000	1.001
Constant	270.564	20.556	116.305	1	0.000	934726970705

In Tabella 147 e Tabella 148 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 147: Matrice di confusione - Logit: SONCA con distribuzione di probabilità gaussiana e m=3500

Sample		Predicted		
		1	0	Percent Correct
Training	1	2916	566	83.7%
	0	593	2999	83.5%
	Overall Percentage	49.6%	50.4%	83.6%
Holdout	1	1086	287	79.1%
	0	3388	14549	81.1%
	Overall Percentage	23.2%	76.8%	81.0%

Tabella 148: Misure di performance - Logit: SONCA con distribuzione di probabilità gaussiana e m=3500

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.163	0.165	0.835	0.837	0.831	0.834	0.164	0.836
Holdout	0.209	0.189	0.811	0.791	0.243	0.371	0.190	0.810

In Figura 75 e Tabella 149 sono riportate la curva ROC e Area sotto la curva ROC.

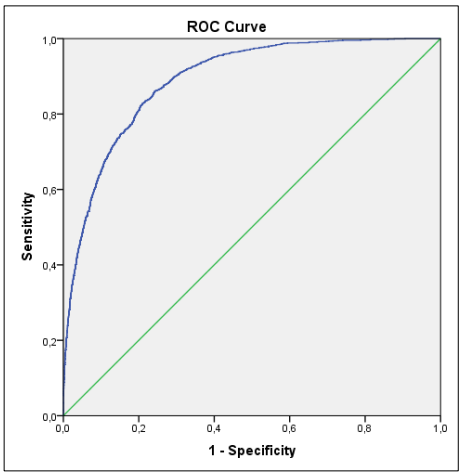

Figura 75 - Curva ROC - Logit: SONCA con distribuzione di probabilità gaussiana e m=3500

Tabella 149 - Area sotto la curva ROC - Logit: SONCA con distribuzione di probabilità gaussiana e m=3500

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.890	0.004	0.000	0.882	0.898

SONCA con distribuzione di probabilità gaussiana e m=5500

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo gaussiana e m=5500. In Tabella 150 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA con distribuzione di probabilità gaussiana e m=5500

Tabella 150: Distribuzioni di classe per il training set: SONCA con distribuzione di probabilità gaussiana e m=5500

Variabile	N	%
1	5’412	49.7
0	5’478	50.3
Totale	10’890	100.0

Albero di regressione con il ricampionamento: SONCA con distribuzione di probabilità gaussiana e m=5500

In Figura 76 e Figura 77 sono riportati gli alberi di regressione training avendo implementato SONCA con distribuzione di probabilità gaussiana e m=5500.

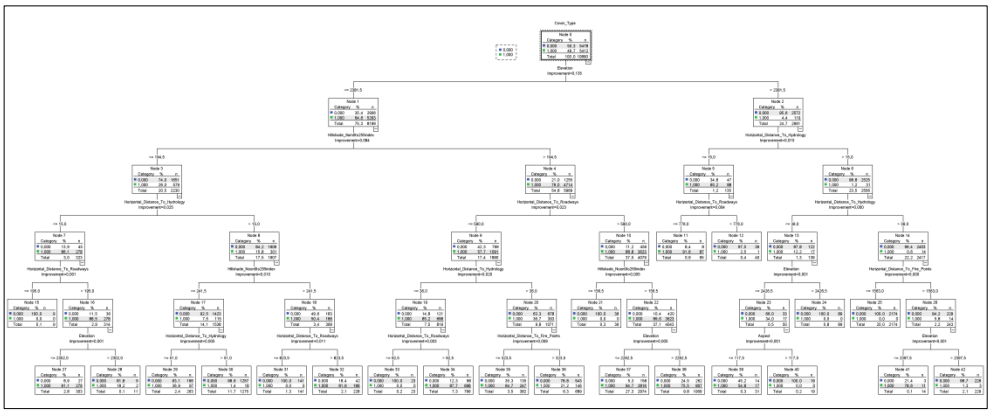

Figura 76 – Albero di regressione per il training set: SONCA con distribuzione di probabilità gaussiana e m=5500

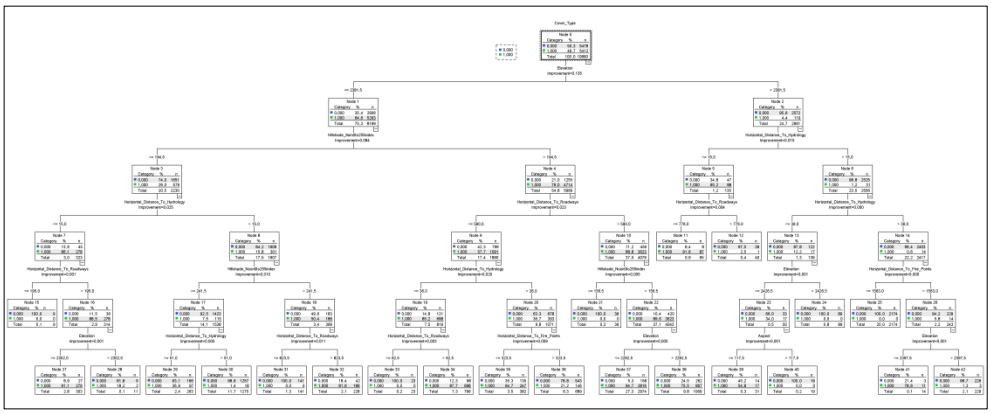

Figura 77 – Albero di regressione per l'holdout set: SONCA con distribuzione di probabilità gaussiana e m=5500

In Tabella 151 e Tabella 152 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set avendo implementato SONCA con distribuzione di probabilità gaussiana e m=5500.

Tabella 151: Matrice di confusione - CART: SONCA con distribuzione di probabilità gaussiana e m=5500

Sample		Predicted		
		1	0	Percent Correct
Training	1	5145	267	95.1%
	0	747	4731	86.4%
	Overall Percentage	54.1%	45.9%	90.7%
Holdout	1	1236	137	90.0%
	0	2386	15551	86.7%
	Overall Percentage	18.8%	81.2%	86.9%

Tabella 152: Misure di performance - CART: SONCA con distribuzione di probabilità gaussiana con m=5500

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.049	0.136	0.864	0.951	0.873	0.910	0.093	0.907
Holdout	0.100	0.133	0.867	0.900	0.341	0.495	0.131	0.869

In Figura 78 e Tabella 153 sono riportate la curva ROC e Area sotto la curva ROC.

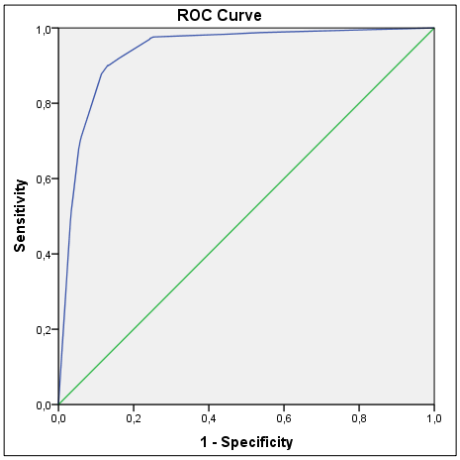

Figura 78 - Curva ROC - CART: SONCA con distribuzione di probabilità gaussiana e m=5500

Tabella 153: Area sotto la curva ROC - CART: SONCA con distribuzione di probabilità gaussiana e m=5500

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.936	0.003	0.000	0.930	0.942

Logit con il ricampionamento: SONCA con distribuzione di probabilità gaussiana e m=5500

In Tabella 154 sono riportate le variabili nell'equazione.

Tabella 154: Variabili nell'equazione: SONCA con distribuzione di probabilità gaussiana e m=5500

Variable (11 step)	B	S.E.	Wald	df	Sig.	Exp(B)
Elevation	-0.012	0.000	1791.945	1	0.000	0.988
Slope	-0.144	0.012	136.492	1	0.000	0.865
Horizontal_Distance_To_Hydrology	-0.005	0.000	246.308	1	0.000	0.995
Vertical_Distance_To_Hydrology	0.007	0.001	75.374	1	0.000	1.007
Horizontal_Distance_To_Roadways	0.002	0.000	665.747	1	0.000	1.002
Hillshade_9am0to255index	-0.090	0.012	52.694	1	0.000	0.914
Hillshade_Noon0to255index	0.124	0.010	152.795	1	0.000	1.132
Hillshade_3pm0to255index	-0.105	0.010	107.841	1	0.000	0.900
Horizontal_Distance_To_Fire_Points	0.001	0.000	102.146	1	0.000	1.001
Constant	33.612	2.200	233.330	1	0.000	0.395928*1E+15

In Tabella 155 e Tabella 156 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 155: Matrice di confusione - Logit: SONCA con distribuzione di probabilità gaussiana con m=5500

Sample		Predicted		
		1	0	Percent Correct
Training	1	4582	830	84.7%
	0	897	4581	83.6%
	Overall Percentage	50.3%	49.7%	84.1%
Holdout	1	1098	275	80.0%
	0	3231	14706	82.0%
	Overall Percentage	22.4%	77.6%	81.8%

Tabella 156: Misure di performance - Logit: SONCA con distribuzione di probabilità gaussiana con m=5500

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.153	0.164	0.836	0.847	0.836	0.841	0.159	0.841
Holdout	0.200	0.180	0.820	0.800	0.254	0.385	0.182	0.818

In Figura 79 e Tabella 157 sono riportate la curva ROC e Area sotto la curva ROC.

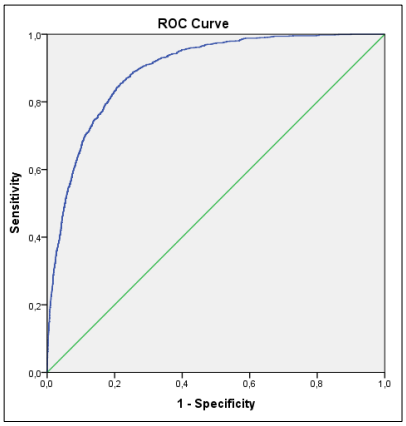

Figura 79 - Curva ROC - Logit: SONCA con distribuzione di probabilità gaussiana con m=5500

Tabella 157: Area sotto la curva ROC - Logit: SONCA con distribuzione di probabilità gaussiana con m=5500

Area	Std. Error.	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.894	0.004	0.000	0.887	0.902

SONCA con distribuzione di probabilità gaussiana e m=7500

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo gaussiana e m=7500. In Tabella 126 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA con distribuzione di probabilità gaussiana e m=7500

Tabella 158: Distribuzioni di classe per il training set: SONCA con distribuzione di probabilità gaussiana e m=7500

Variabile	N	%
1	7’477	49.62
0	7’593	50.38
Totale	15’070	100.00

Albero di regressione con il ricampionamento: SONCA con distribuzione di probabilità gaussiana e m=7500

In Figura 80 e Figura 81 sono riportati gli alberi di regressione training avendo implementato SONCA con distribuzione di probabilità gaussiana e m=7500.

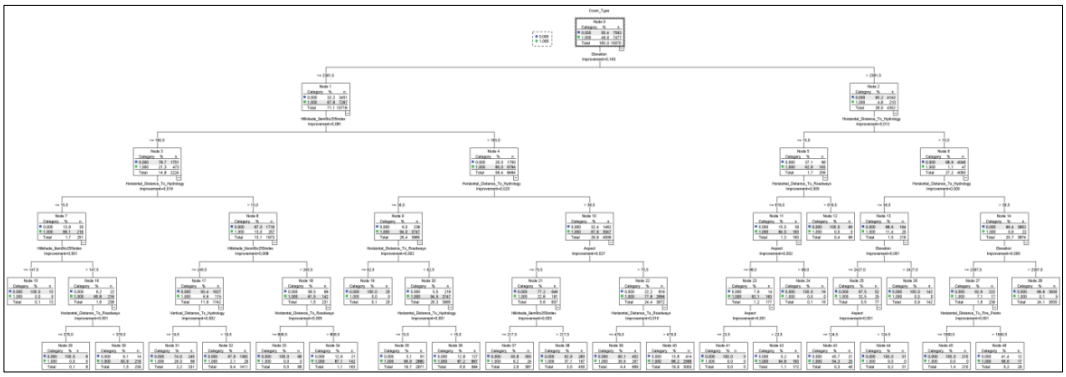

Figura 80 – Albero di regressione per il trainnig set: SONCA con distribuzione di probabilità gaussiana e m=7500

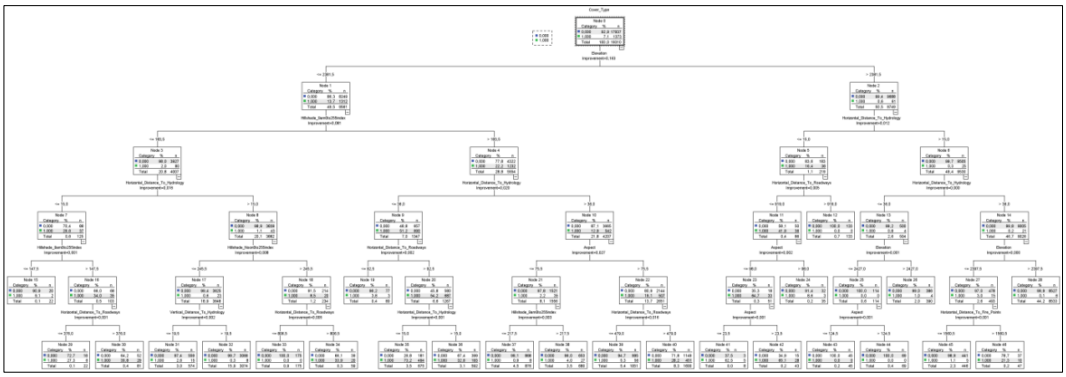

Figura 81 – Albero di regressione per l'holdout set: SONCA con distribuzione di probabilità gaussiana con m=7500

In Tabella 159 e Tabella 160 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 159: Matrice di confusione - CART: SONCA con distribuzione di probabilità gaussiana e m=7500

Sample		Predicted		
		1	0	Percent Correct
Training	1	6899	578	92.3%
	0	709	6884	90.7%
	Overall Percentage	50.5%	49.5%	91.5%
Holdout	1	1225	148	89.2%
	0	1917	16020	89.3%
	Overall Percentage	16.3%	83.7%	89.3%

Tabella 160: Misure di performance - CART: SONCA con distribuzione di probabilità gaussiana con m=7500

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.077	0.093	0.907	0.923	0.907	0.915	0.085	0.915
Holdout	0.108	0.107	0.893	0.892	0.390	0.543	0.107	0.893

In Figura 82 e Tabella 161 sono riportate la curva ROC e Area sotto la curva ROC.

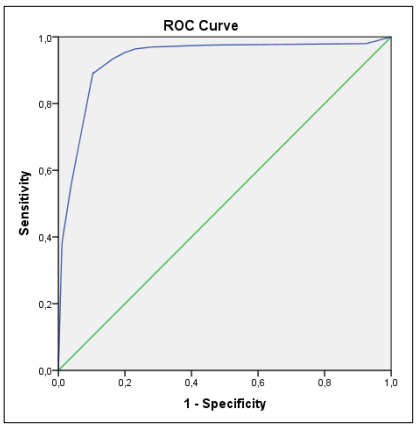

Figura 82 - Curva ROC - CART: SONCA con distribuzione di probabilità gaussiana e m=7500

Tabella 161: Area sotto la curva ROC - CART: SONCA con distribuzione di probabilità gaussiana e m=7500

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.932	0.004	0.000	0.924	0.940

Logit con il ricampionamento: SONCA con distribuzione di probabilità gaussiana e m=7500

In Tabella 162 sono riportate le variabili nell'equazione.

Tabella 162: Variabili nell'equazione: SONCA con distribuzione di probabilità gaussiana con m=7500

Variables in the Equation	B	S.E.	Wald	df	Sig.	Exp(B)
Elevation	-0.013	0.000	2686.212	1	0.000	0.987
Slope	-0.093	0.011	76.397	1	0.000	0.911
Horizontal_Distance_To_Hydrology	-0.006	0.000	378.781	1	0.000	0.994
Vertical_Distance_To_Hydrology	0.008	0.001	148.880	1	0.000	1.008
Horizontal_Distance_To_Roadways	0.002	0.000	616.847	1	0.000	1.002
Hillshade_9am0to255index	-0.030	0.011	7.552	1	0.006	0.970
Hillshade_Noon0to255index	0.082	0.009	85.116	1	0.000	1.086
Hillshade_3pm0to255index	-0.058	0.009	42.706	1	0.000	0.943
Horizontal_Distance_To_Fire_Points	0.000	0.000	22.787	1	0.000	1.000
Constant	25.697	1.909	181.263	1	0.000	144570109209.921

In Tabella 163 e Tabella 164 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA con distribuzione di probabilità gaussiana e m=7500.

Tabella 163: Matrice di confusione - Logit: SONCA con distribuzione di probabilità gaussiana e m=7500

Sample		Predicted		
		1	0	Percent Correct
Training	1	6523	954	87.2%
	0	1104	6489	85.5%
	Overall Percentage	50.6%	49.4%	86.3%
Holdout	1	1127	246	82.1%
	0	3050	14887	83.0%
	Overall Percentage	21.6%	78.4%	82.9%

Tabella 164: Misure di performance per- Logit: SONCA con distribuzione di probabilità gaussiana con m=7500

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.128	0.145	0.855	0.872	0.855	0.864	0.137	0.863
Holdout	0.179	0.170	0.830	0.821	0.270	0.406	0.171	0.829

In Figura 83 e Tabella 165 sono riportate la curva ROC e Area sotto la curva ROC.

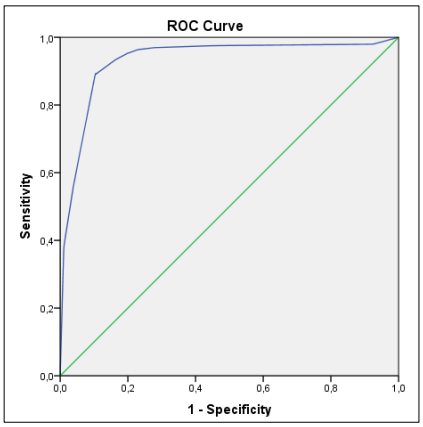

Figura 83 - Curva ROC - Logit: SONCA con distribuzione di probabilità gaussiana con m=5500

Tabella 165: Area sotto la curva ROC - Logit: SONCA con distribuzione di probabilità gaussiana con m=5500

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.899	0.004	0.000	0.891	0.906

SONCA con distribuzione di probabilità gaussiana e m=9500

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo gaussiana e m=9500. In Tabella 166 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA con distribuzione di probabilità gaussiana e m=9500

Tabella 166: Distribuzioni di classe per il training set: SONCA con distribuzione di probabilità gaussiana e m=9500

Variabile	N	%
1	9'406	49.72
0	9'511	50.28
Totale	18'917	100.00

Albero di regressione con il ricampionamento: SONCA con distribuzione di probabilità gaussiana con m=9500

In Figura 84 e Figura 85 sono riportati gli alberi di regressione training avendo implementato SONCA con distribuzione di probabilità gaussiana e m=9500.

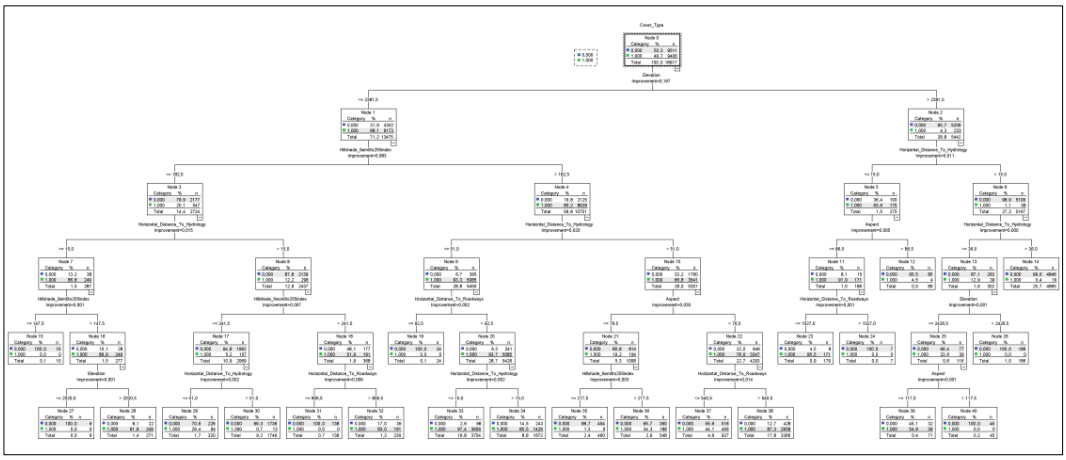

Figura 84 – Albero di regressione per il training set: SONCA con distribuzione di probabilità gaussiana e m=9500

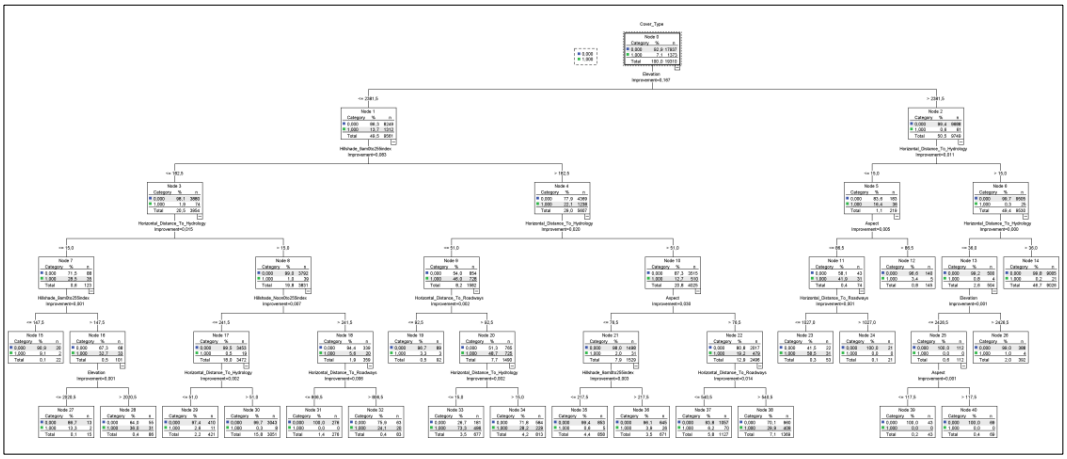

Figura 85 – Albero di regressione per l'holdout set: SONCA con distribuzione di probabilità gaussiana e m=9500

In Tabella 167 e Tabella 168 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA con distribuzione di probabilità gaussiana e m=9500.

Tabella 167: Matrice di confusione - CART: SONCA con distribuzione di probabilità gaussiana e m=9500

Sample		Predicted		
		1	0	Percent Correct
Training	1	8673	733	92.2%
	0	870	8641	90.9%
	Overall Percentage	50.4%	49.6%	91.5%
Holdout	1	1216	157	88.6%
	0	1908	16029	89.4%
	Overall Percentage	16.2%	83.8%	89.3%

Tabella 168: Misure di performance - CART: SONCA con distribuzione di probabilità gaussiana e m=9500

	FN	FP	TN	TP	Precision	F-measure	Err	Acc
Training	0.078	0.091	0.909	0.922	0.909	0.915	0.085	0.915
Holdout	0.114	0.106	0.894	0.886	0.389	0.541	0.107	0.893

In Figura 86 e Tabella 169 sono riportate la curva ROC e Area sotto la curva.

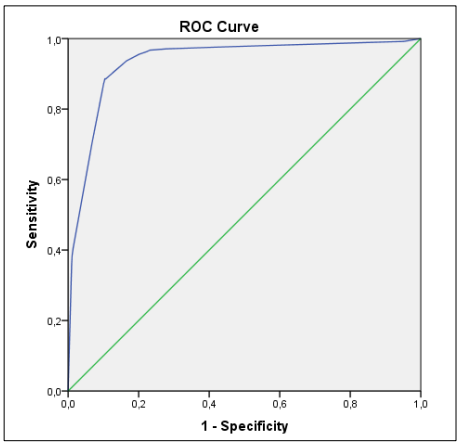

Figura 86 - Curva ROC - CART: SONCA con distribuzione di probabilità gaussiana e m=9500

Tabella 169: Area sotto la curva ROC - CART: SONCA con distribuzione di probabilità gaussiana con m=9500

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.935	0.004	0.000	0.928	0.942

Logit con il ricampionamento: SONCA con distribuzione di probabilità gaussiana e m=9500

In Tabella 170 sono riportate le variabili nell'equazione.

Tabella 170: Variabili nell'equazione: SONCA con distribuzione di probabilità gaussiana e m=9500

Variables in the Equation	B	S.E.	Wald	df	Sig.	Exp(B)
Elevation	-0.01	0.00	3284.76	1.00	0.00	0.99
Slope	-0.13	0.01	179.61	1.00	0.00	0.88
Horizontal_Distance_To_Hydrology	-0.01	0.00	470.00	1.00	0.00	0.99
Vertical_Distance_To_Hydrology	0.01	0.00	178.20	1.00	0.00	1.01
Horizontal_Distance_To_Roadways	0.00	0.00	759.23	1.00	0.00	1.00
Hillshade_9am0to255index	-0.07	0.01	47.67	1.00	0.00	0.93
Hillshade_Noon0to255index	0.11	0.01	198.07	1.00	0.00	1.12
Hillshade_3pm0to255index	-0.09	0.01	123.22	1.00	0.00	0.91
Horizontal_Distance_To_Fire_Points	0.00	0.00	21.56	1.00	0.00	1.00
Constant	32.71	1.77	343.17	1.00	0.00	1608.02E+11

In Tabella 171 e Tabella 172 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA con distribuzione di probabilità gaussiana e m=9500.

Tabella 171: Matrice di confusione - Logit: SONCA con distribuzione di probabilità gaussiana con m=9500

Sample		Predicted		
		1	0	Percent Correct
Training	1	8240	1166	87.6%
	0	1332	8179	86.0%
	Overall Percentage	50.6%	49.4%	86.8%
Holdout	1	1122	251	81.7%
	0	3051	14886	83.0%
	Overall Percentage	21.6%	78.4%	82.9%

Tabella 172: Misure di performance - Logit: SONCA con distribuzione di probabilità gaussiana con m=7500

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.124	0.140	0.860	0.876	0.861	0.868	0.132	0.868
Holdout	0.183	0.170	0.830	0.817	0.269	0.405	0.171	0.829

In Figura 87 e Tabella 173 sono riportate la curva ROC e Area sotto la curva ROC.

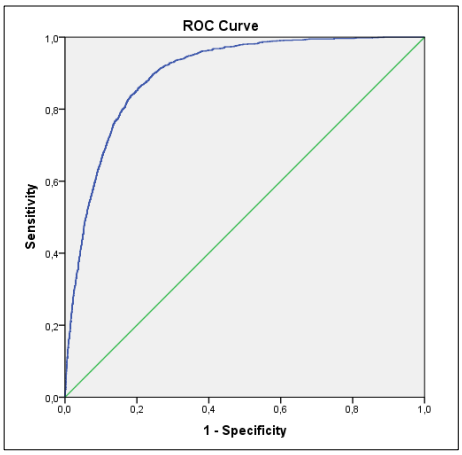

Figura 87 - Curva ROC - Logit: SONCA con distribuzione di probabilità gaussiana e m=5500

Tabella 173: Area sotto la curva ROC - Logit: SONCA con distribuzione di probabilità gaussiana e m=5500

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.898	0.004	0.000	0.891	0.905

A.1.2. Adult dataset

Il dataset è composto da 32’561 osservazioni, il compito di previsione relative a questo insieme di dati è quello di determinare se una persona caratterizzata rende più di 50 $K/anno.

La variabile di risposta è costituita da 7’841 osservazioni appartenente alla classe minoritaria, pari al 24% del totale, e 27'720 osservazioni appartenenti alla classe maggioritaria, pari al 76% del totale

- Dataset originale;
- SONCA con distribuzione di probabilità triangolare e con m=4000;
- SONCA con distribuzione di probabilità triangolare e con m=8000;
- SONCA con distribuzione di probabilità triangolare e con m=12000;
- SONCA con distribuzione di probabilità gaussiana e con m=4000;
- SONCA con distribuzione di probabilità gaussiana e con m=8000;
- SONCA con distribuzione di probabilità gaussiana e con m=12000.

Dataset originale

Albero di regressione senza il ricampionamento

In Figura 88 e Figura 89 sono riportati gli alberi di regressione senza bilanciare la variabile di risposta.

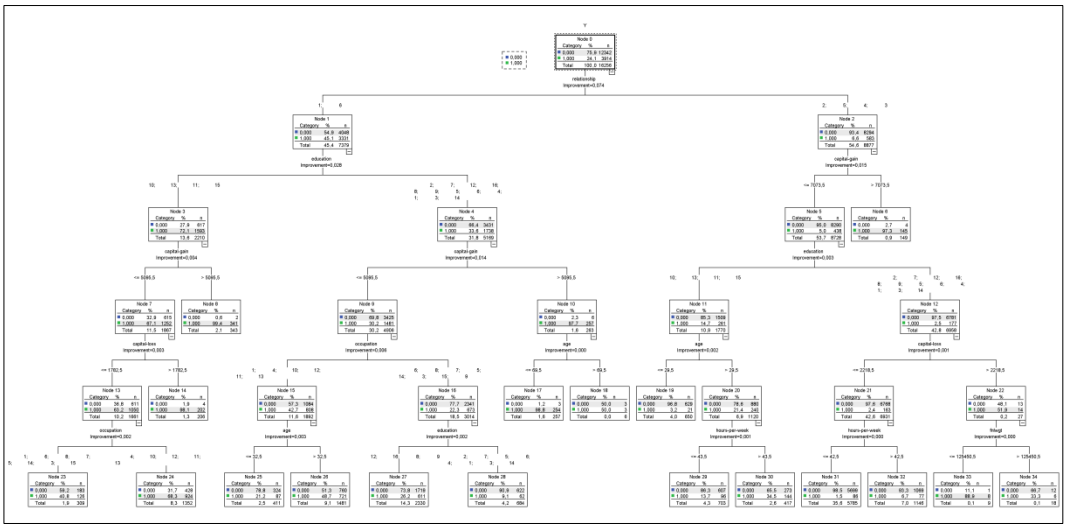

Figura 88 – Albero di regressione per il training set senza ricampionamento

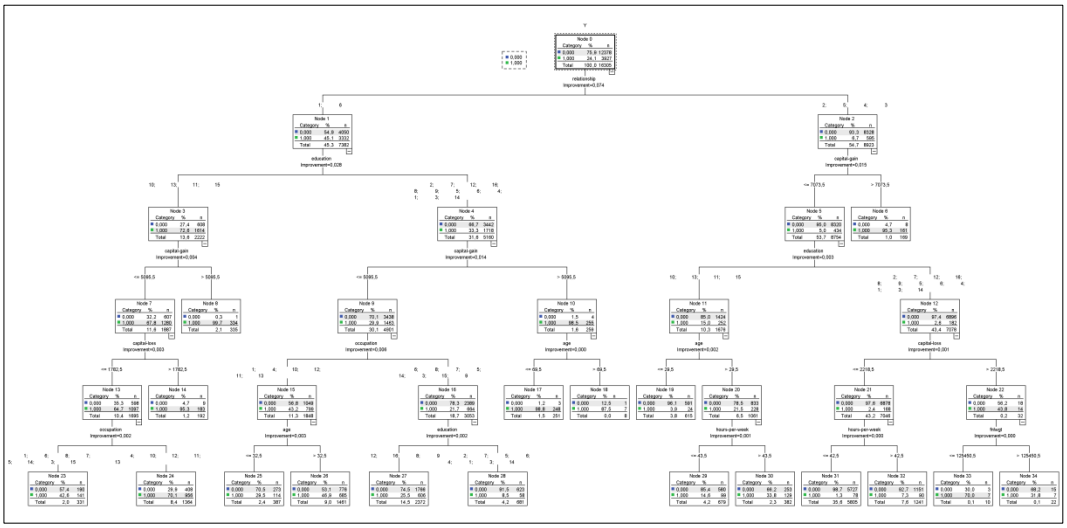

Figura 89 – Albero di regressione per l'holdout set senza ricampionamento

In Tabella 174 e Tabella 175 sono riportate sia la matrice di confusione sia le misure di prestazioni per entrambi i set dati, training e holdout.

Tabella 174: Matrice di confusione senza ricampionamento

Sample		Predicted		
		1	0	Percent Correct
Training	1	1874	2040	47.9%
	0	442	11900	96.4%
	Overall Percentage	14.2%	85.8%	84.7%
Holdout	1	1889	2038	48.1%
	0	432	11946	96.5%
	Overall Percentage	14.2%	85.8%	84.9%

Tabella 175: Misure di performance senza ricampionamento

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.521	0.036	0.964	0.479	0.809	0.602	0.153	0.847
Holdout	0.519	0.035	0.965	0.481	0.814	0.605	0.151	0.849

In Figura 90 e in Tabella 176 sono riportate la curva ROC e l'area sottesa alla curva ROC.

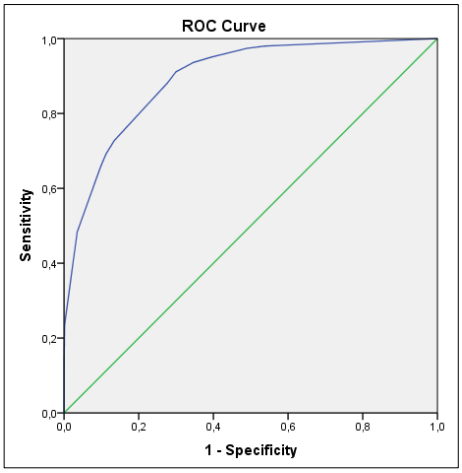

Figura 90 - Curva ROC per l'holdout set senza ricampionamento

Tabella 176: Area sotto la curva ROC per l'holdout set senza ricampionamento

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.893	0.003	0.000	0.888	0.899

Logit senza il ricampionamento

In Tabella 177 sono riportate le variabili nell'equazione.

Tabella 177: Variabili nell'equazione per il training set senza ricampionamento

	B	S.E.	Wald	df	Sig.	Exp(B)
age	0.04	0.00	666.19	1.00	0.00	1.05
educationnum	0.31	0.01	1046.40	1.00	0.00	1.36
capitalgain	0.00	0.00	555.58	1.00	0.00	1.00
capitalloss	0.00	0.00	255.84	1.00	0.00	1.00
hoursperweek	0.04	0.00	451.21	1.00	0.00	1.04
Constant	-8.22	0.16	2581.73	1.00	0.00	0.00

In Tabella 178 e Tabella 179 sono riportate la matrice di confusione che le misure di prestazioni sia per il dataset di training che di holdout.

Tabella 178: Matrice di confusione senza ricampionamento

Sample		Predicted		
		1	0	Percent Correct
Training	1	1524	2390	38.9%
	0	626	11716	94.9%
	Overall Percentage	13.2%	86.8%	81.4%
Holdout	1	1525	2402	38.8%
	0	617	11761	95.0%
	Overall Percentage	13.1%	86.9%	81.5%

Tabella 179: Misure di performance senza ricampionamento

	FN	FP	TN	TP	Precision	F-measure	Err	Acc
Training	0.611	0.051	0.949	0.389	0.709	0.503	0.186	0.814
Holdout	0.612	0.050	0.950	0.388	0.712	0.503	0.185	0.815

In Figura 91 e in Tabella 180 sono riportate la curva ROC e l'area sottesa alla curva ROC.

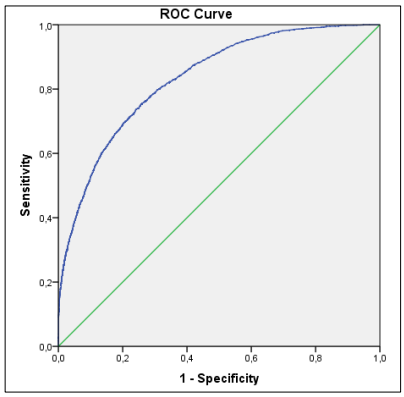

Figura 91 - Curva ROC per l'holdout set senza ricampionamento

Tabella 180: Area sotto la curva ROC per l'holdout set senza ricampionamento

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.833	0.004	0.000	0.826	0.840

SONCA con distribuzione di probabilità triangolare e m=4000

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo triangolare e m=4000. In Tabella 181 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA con distribuzione di probabilità triangolare e m=4000

Tabella 181: Distribuzioni di classe per il training set: SONCA con distribuzione di probabilità triangolare e m=4000

Variabile	N	%
1	4003	49.9
0	4025	50.1
Totale	8028	100.00

Albero di regressione con il ricampionamento: SONCA con distribuzione di probabilità triangolare e m=4000

In Figura 92 e Figura 93 sono riportati gli alberi di regressione training avendo implementato SONCA con distribuzione di probabilità triangolare e m=4000.

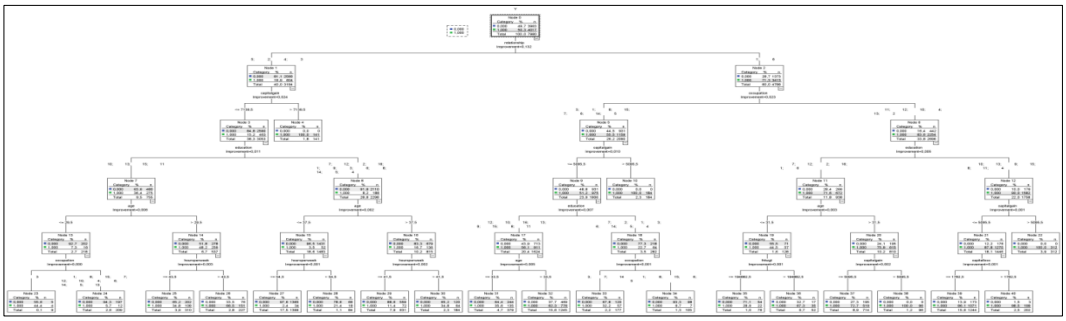

Figura 92 – Albero di regressione per il trainig set: SONCA con distribuzione di probabilità triangolare e m=4000

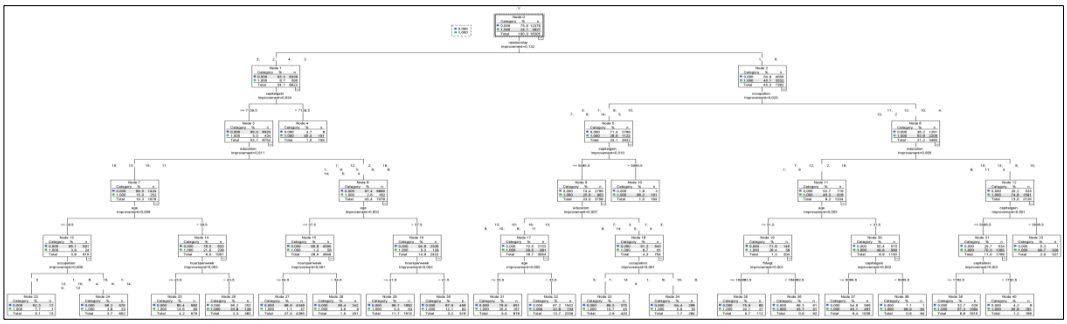

Figura 93 – Albero di regressione per l'holdout set: SONCA con distribuzione di probabilità triangolare e m=4000

In Tabella 182 e Tabella 183 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 182: Matrice di confusione-CART: SONCA con distribuzione di probabilità triangolare e m=4000

Sample		Predicted		
		1	0	Percent Correct
Training	1	3484	533	86.7%
	0	933	3030	76.5%
	Overall Percentage	55.4%	44.6%	81.6%
Holdout	1	3367	560	85.7%
	0	2932	9446	76.3%
	Overall Percentage	38.6%	61.4%	78.6%

Tabella 183: Misure di performance - CART: SONCA con distribuzione di probabilità triangolare e m=4000

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.141	0.227	0.773	0.859	0.790	0.823	0.184	0.816
Holdout	0.152	0.231	0.769	0.848	0.537	0.658	0.212	0.788

In Figura 94 e Tabella 184 sono riportate la curva e l'area sotto la curva ROC.

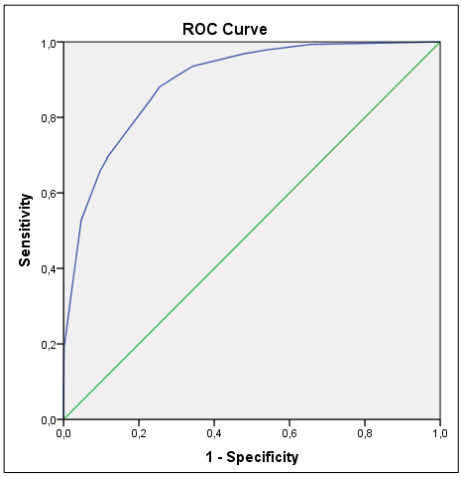

Figura 94 - Curva ROC per l'holdout set - CART: SONCA con distribuzione di probabilità triangolare e m=4000

Tabella 184: Area sotto la curva ROC per l'holdout set - CART: SONCA con distribuzione di probabilità triangolare e m=4000

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.895	0.003	0.000	0.889	0.900

Logit con il ricampionamento: SONCA con distribuzione di probabilità triangolare e m=4000

In Tabella 185 sono riportate le variabili nell'equazione.

Tabella 185: Variabili nell'equazione: SONCA con distribuzione di probabilità triangolare e m=4000

	B	S.E.	Wald	df	Sig.	Exp(B)
age	0.0496	0.0023	476.8742	1.0000	0.0000	1.0508
educationnum	0.3191	0.0120	706.9254	1.0000	0.0000	1.3760
capitalgain	0.0003	0.0000	236.0045	1.0000	0.0000	1.0003
capitalloss	0.0006	0.0001	95.9043	1.0000	0.0000	1.0006
hoursperweek	0.0415	0.0025	272.8698	1.0000	0.0000	1.0423
Constant	-7.4126	0.2062	1292.5592	1.0000	0.0000	0.0006

In Tabella 186 e Tabella 187 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 186: Matrice di confusione - Logit: SONCA con distribuzione di probabilità triangolare e m=4000

Sample		Predicted		
		1	0	Percent Correct
Training	1	3080	923	76.9%
	0	890	3135	77.9%
	Overall Percentage	49.5%	50.5%	77.4%
Holdout	1	3021	906	76.9%
	0	2797	9581	77.4%
	Overall Percentage	72.5%	130.6%	77.3%

Tabella 187: Misure di performance - Logit: SONCA con distribuzione di probabilità triangolare e m=4000

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.231	0.221	0.779	0.769	0.776	0.773	0.226	0.774
Holdout	0.231	0.226	0.774	0.769	0.519	0.620	0.227	0.773

In Figura 95 e Tabella 188sono riportate la curva e l'area sotto la curva ROC.

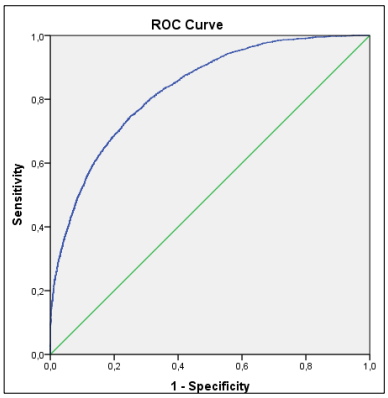

Figura 95 - Curva ROC - Logit: SONCA con distribuzione di probabilità triangolare e m=4000

Tabella 188: Area sotto la curva ROC - Logit: SONCA con distribuzione di probabilità triangolare e m=4000

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.856	0.003	0.000	0.850	0.862

SONCA con distribuzione di probabilità triangolare e m=8000

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo triangolare e m=8000. In Tabella 189 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA con distribuzione di probabilità triangolare e m=8000

Tabella 189: Distribuzioni di classe per il training set: SONCA con distribuzione di probabilità triangolare e m=8000

Variabile	N	%
0	8'078	50.2
1	8'005	49.8
Totale	16'083	100.0

Albero di regressione con il ricampionamento: SONCA con distribuzione di probabilità triangolare e m=8000

In Figura 96 e Figura 97sono riportati gli alberi di regressione training avendo implementato SONCA con distribuzione di probabilità triangolare e m=8000.

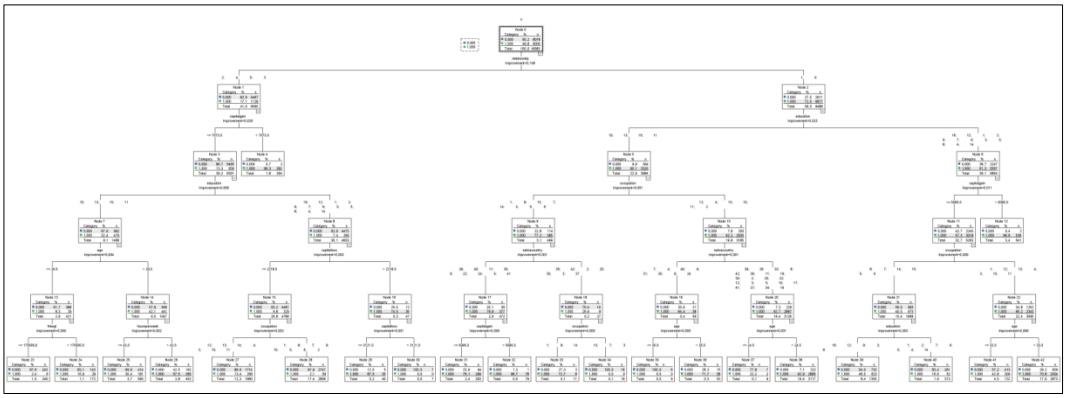

Figura 96 – Albero di regressione per il training set: SONCA con distribuzione di probabilità triangolare e m=8000

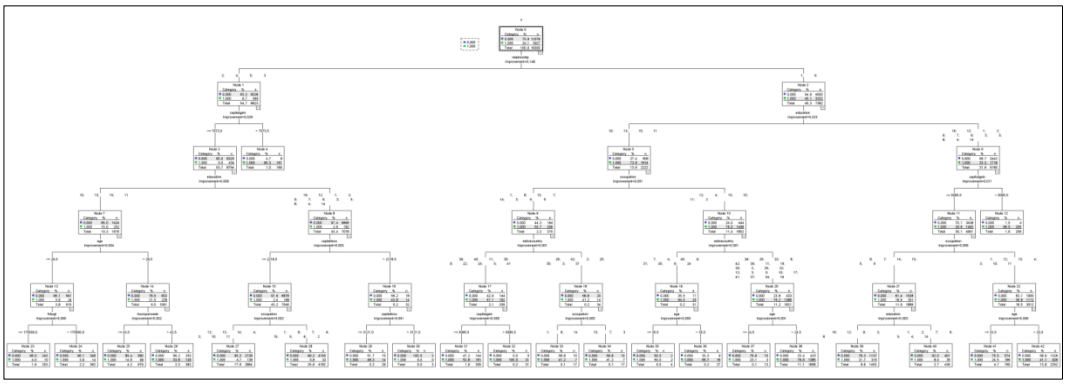

Figura 97 – Albero di regressione per l'holdout set: SONCA con distribuzione di probabilità triangolare e m=8000

In Tabella 190 e Tabella 191 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 190: Matrice di confusione - CART: SONCA con distribuzione di probabilità triangolare e m=8000

Sample		Predicted		
		1	0	Percent Correct
Training	1	6478	1527	80.9%
	0	1375	6703	83.0%
	Overall Percentage	48.8%	51.2%	82.0%
Holdout	1	3087	840	78.6%
	0	2192	10186	82.3%
	Overall Percentage	32.4%	67.6%	81.4%

Tabella 191: Misure di performance - CART: SONCA con distribuzione di probabilità triangolare con m=8000

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.191	0.170	0.830	0.809	0.825	0.817	0.180	0.820
Holdout	0.214	0.177	0.823	0.786	0.585	0.671	0.186	0.814

In Figura 98 e Tabella 192 sono riportate la curva e l'area sotto la curva ROC.

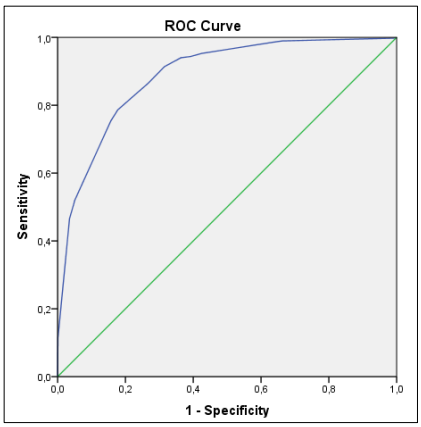

Figura 98 - Curva ROC - CART: SONCA con distribuzione di probabilità triangolare e m=8000

Tabella 192: Area sotto la curva ROC - CART: SONCA con distribuzione di probabilità triangolare e m=8000

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.887	0.003	0.000	0.881	0.892

Logit con il ricampionamento: SONCA con distribuzione di probabilità triangolare e m=8000

In Tabella 193 sono riportate le variabili nell'equazione.

Tabella 193: Variabili nell'equazione: SONCA con distribuzione di probabilità triangolare e m=8000

	B	S.E.	Wald	df	Sig.	Exp(B)
age	0.051	0.002	967.649	1	0.000	1.052
fnlwgt	0.000	0.000	132.404	1	0.000	1.000
educationnum	0.322	0.009	1392.473	1	0.000	1.380
capitalgain	0.000	0.000	483.760	1	0.000	1.000
capitalloss	0.001	0.000	249.672	1	0.000	1.001
hoursperweek	0.043	0.002	559.981	1	0.000	1.044
Constant	-8.064	0.156	2662.344	1	0.000	0.000

In Tabella 194 e Tabella 195sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 194: Matrice di confusione - Logit: SONCA con distribuzione di probabilità triangolare con m=8000

Sample		Predicted		
		1	0	Percent Correct
Training	1	5703	2302	71.2%
	0	1863	6215	76.9%
	Overall Percentage	47.0%	53.0%	74.1%
Holdout	1	2870	1057	73.1%
	0	2975	9403	76.0%
	Overall Percentage	35.8%	64.2%	75.3%

Tabella 195: Misure di performance - Logit: SONCA con distribuzione di probabilità triangolare con m=8000

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.288	0.231	0.769	0.712	0.754	0.733	0.259	0.741
Holdout	0.269	0.240	0.760	0.731	0.491	0.587	0.247	0.753

In Figura 99 e Tabella 196 sono riportate la curva e l'area sotto la curva ROC.

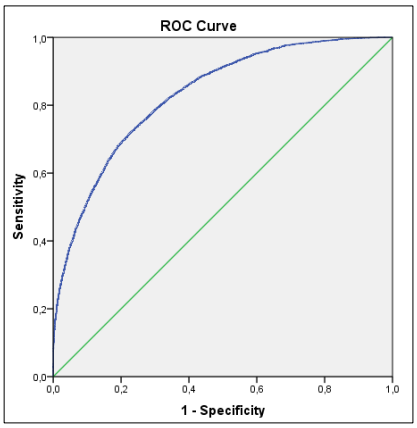

Figura 99 - Curva ROC - Logit: SONCA con distribuzione di probabilità triangolare con m=8000

Tabella 196: Area sotto la curva ROC - Logit: SONCA con distribuzione di probabilità triangolare con m=8000

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.830	0.004	0.000	0.823	0.837

SONCA con distribuzione di probabilità triangolare e m=12000

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo triangolare e m=12000. In Tabella 197 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA con distribuzione di probabilità triangolare e m=12000

Tabella 197: Distribuzioni di classe per il training set: SONCA con distribuzione di probabilità triangolare e m=12000

Variabile	N	%
1	12’087	50.2
0	11’980	49.8
Totale	24’067	100.0

Albero di regressione con il ricampionamento: SONCA con distribuzione di probabilità triangolare e m=12000

In Figura 100 e Figura 101sono riportati gli alberi di regressione training avendo implementato SONCA con distribuzione di probabilità triangolare e m=12000.

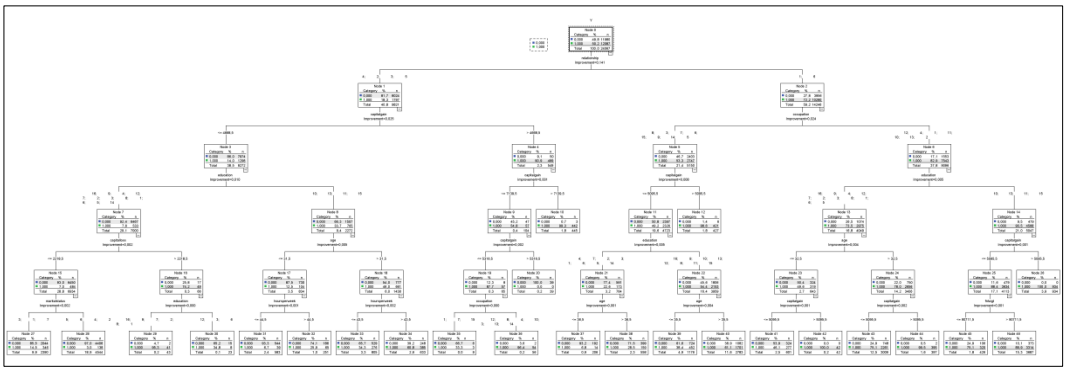

Figura 100 – Albero di regressione per il trainnig set: SONCA con distribuzione di probabilità triangolare e m=12000

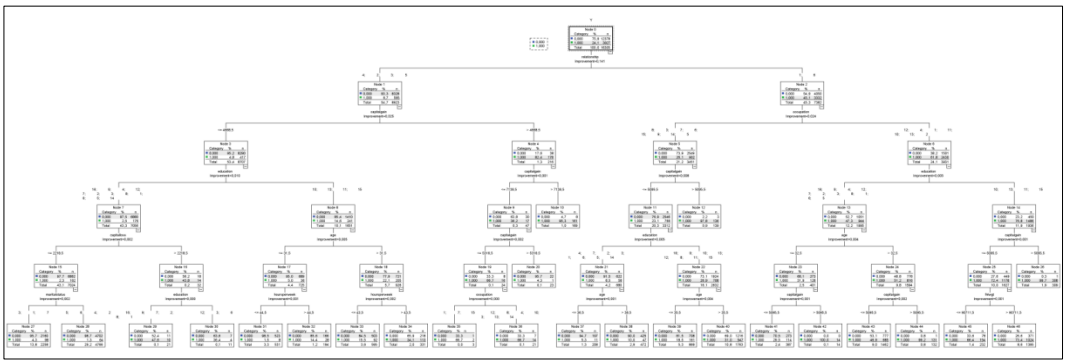

Figura 101 – Albero di regressione per l'holdout set: SONCA con distribuzione di probabilità triangolare con m=12000

In

Tabella 198 e Tabella 199 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 198: Matrice di confusione - CART: SONCA con distribuzione di probabilità triangolare e m=12000

Sample		Predicted		
		1	0	Percent Correct
Training	1	10310	1777	85.3%
	0	2572	9408	78.5%
	Overall Percentage	53.5%	46.5%	81.9%
Holdout	1	3297	630	84.0%
	0	2691	9687	78.3%
	Overall Percentage	36.7%	63.3%	79.6%

Tabella 199: Misure di performance - CART: SONCA con distribuzione di probabilità triangolare con m=12000

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.147	0.215	0.785	0.853	0.800	0.826	0.181	0.819
Holdout	0.160	0.217	0.783	0.840	0.551	0.665	0.204	0.796

In Figura 102 e Tabella 200 sono riportate la curva e l'area sotto la curva ROC.

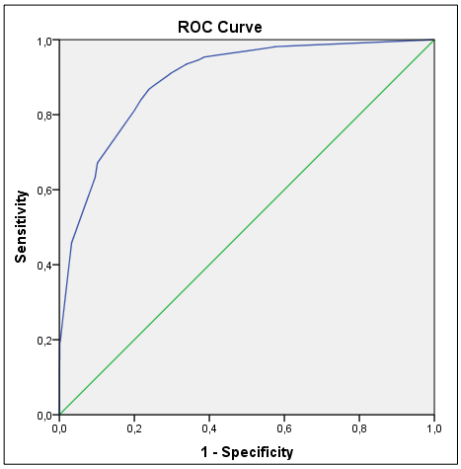

Figura 102 - Curva ROC - CART: SONCA con distribuzione di probabilità triangolare e m=12000

Tabella 200: Area sotto la curva ROC - CART: SONCA con distribuzione di probabilità triangolare e m=12000

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.893	0.003	0.000	0.887	0.898

Logit con il ricampionamento: SONCA con distribuzione di probabilità triangolare con m=5500

In Tabella 201 sono riportate le variabili nell'equazione.

Tabella 201: Variabili nell'equazione: SONCA con distribuzione di probabilità triangolare con m=12000

	B	S.E.	Wald	df	Sig.	Exp(B)
age	0.052	0.001	1560.512	1	0.000	1.053
fnlwgt	0.000	0.000	226.081	1	0.000	1.000
educationnum	0.309	0.007	1971.200	1	0.000	1.362
capitalgain	0.000	0.000	726.756	1	0.000	1.000
capitalloss	0.001	0.000	376.325	1	0.000	1.001
hoursperweek	0.042	0.001	815.919	1	0.000	1.043
Constant	-7.980	0.127	3951.219	1	0.000	0.000

In Tabella 202 e Tabella 203 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA con distribuzione di probabilità triangolare e m=12000.

Tabella 202: Matrice di confusione - Logit: SONCA con distribuzione di probabilità triangolare e m=12000

Sample		Predicted		
		1	0	Percent Correct
Training	1	8614	3473	71.3%
	0	2897	9083	75.8%
	Overall Percentage	47.8%	52.2%	73.5%
Holdout	1	2861	1066	72.9%
	0	3015	9363	75.6%
	Overall Percentage	36.0%	64.0%	75.0%

Tabella 203: Misure di performance per- Logit: SONCA con distribuzione di probabilità triangolare con m=12000

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.287	0.242	0.758	0.713	0.748	0.730	0.265	0.735
Holdout	0.271	0.244	0.756	0.729	0.487	0.584	0.250	0.750

In

Figura 103 e Tabella 204 sono riportate la curva e l'area sotto la curva ROC.

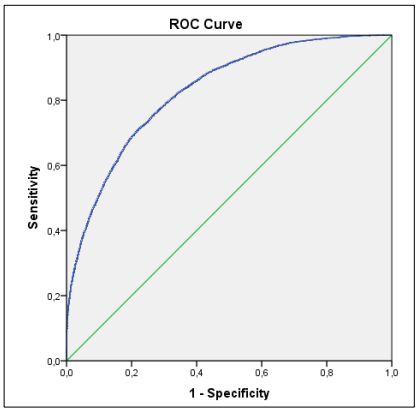

Figura 103 - Curva ROC - Logit: SONCA con distribuzione di probabilità triangolare con m=5500

Tabella 204: Area sotto la curva ROC - Logit: SONCA con distribuzione di probabilità triangolare con m=5500

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.829	0.004	0.000	0.822	0.836

SONCA con distribuzione di probabilità gaussiana e m=4000

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo gaussiana e m=4000. In Tabella 205 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA con distribuzione di probabilità gaussiana e m=4000

Tabella 205: Distribuzioni di classe per il training set: SONCA con distribuzione di probabilità gaussiana e m=4000

Variabile	N	%
1	4'017	50.34
0	3'963	49.66
Totale	7'980	100.00

Albero di regressione con il ricampionamento: SONCA con distribuzione di probabilità gaussiana e m=4000

In Figura 92 e Figura 93 sono riportati gli alberi di regressione training avendo implementato SONCA con distribuzione di probabilità gaussiana e m=4000.

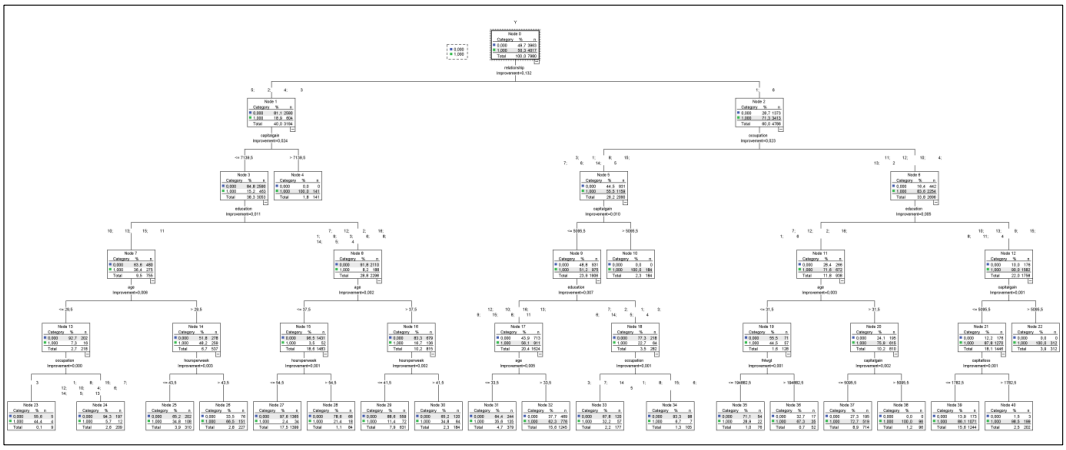

Figura 104 – Albero di regressione per il trainig set: SONCA con distribuzione di probabilità gaussiana e m=4000

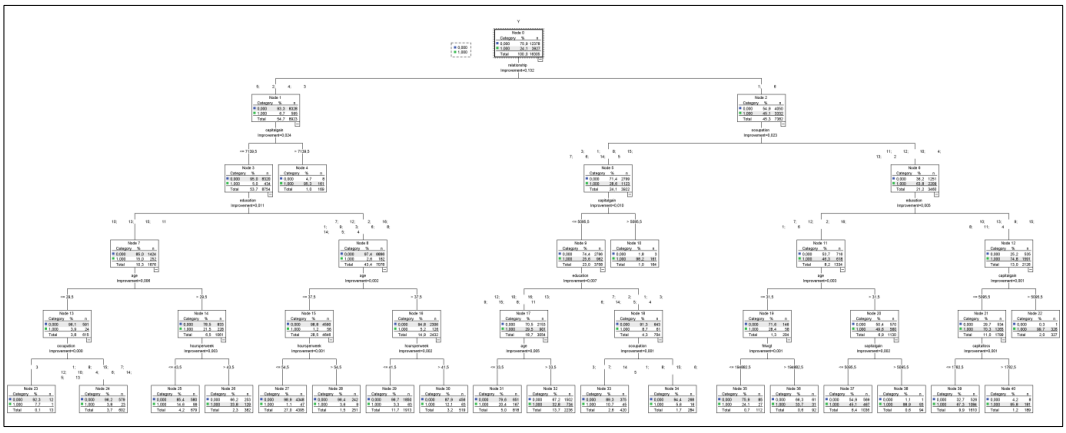

Figura 105 – Albero di regressione per l'holdout set: SONCA con distribuzione di probabilità gaussiana e m=4000

In Tabella 206 e Tabella 207sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 206: Matrice di confusione-CART: SONCA con distribuzione di probabilità gaussiana e m=4000

Sample		Predicted		
		1	0	Percent Correct
Training	1	3484	533	86.7%
	0	933	3030	76.5%
	Overall Percentage	55.4%	44.6%	81.6%
Holdout	1	3367	560	85.7%
	0	2932	9446	76.3%
	Overall Percentage	38.6%	61.4%	78.6%

Tabella 207: Misure di performance - CART: SONCA con distribuzione di probabilità gaussiana e m=4000

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.133	0.235	0.765	0.867	0.789	0.826	0.184	0.816
Holdout	0.143	0.237	0.763	0.857	0.535	0.659	0.214	0.786

In Figura 106 e Tabella 208 sono riportate la curva e l'area sotto la curva ROC.

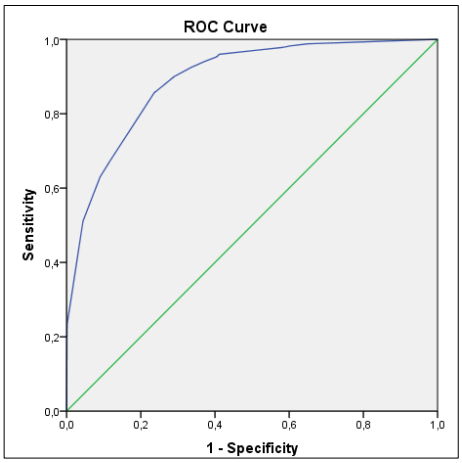

Figura 106 - Curva ROC per l'holdout set - CART: SONCA con distribuzione di probabilità gaussiana e m=4000

Tabella 208: Area sotto la curva ROC per l'holdout set - CART: SONCA con distribuzione di probabilità gaussiana e m=4000

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.891	0.003	0.000	0.886	0.897

Logit con il ricampionamento: SONCA con distribuzione di probabilità gaussiana e m=4000

In Tabella 209 sono riportate le variabili nell'equazione.

Tabella 209: Variabili nell'equazione: SONCA con distribuzione di probabilità gaussiana e m=4000

Variables	B	S.E.	Wald	df	Sig.	Exp(B)
age	0.044	0.002	351.467	1	0.000	1.045
educationnum	0.325	0.012	704.322	1	0.000	1.384
maritalstatus	-0.300	0.022	192.630	1	0.000	0.741
capitalgain	0.000	0.000	231.017	1	0.000	1.000
capitalloss	0.001	0.000	94.411	1	0.000	1.001
hoursperweek	0.038	0.003	219.906	1	0.000	1.038
Constant	-6.090	0.225	733.184	1	0.000	0.002

In

Tabella 210 e Tabella 211sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA con distribuzione di probabilità gaussiana e m=4000.

Tabella 210: Matrice di confusione - Logit: SONCA con distribuzione di probabilità gaussiana e m=4000

Sample		Predicted		
		1	0	Percent Correct
Training	1	2918	1099	72.6%
	0	1002	2961	74.7%
	Overall Percentage	49.1%	50.9%	73.7%
Holdout	1	2904	1023	73.9%
	0	2990	9388	75.8%
	Overall Percentage	36.1%	63.9%	75.4%

Tabella 211: Misure di performance - Logit: SONCA con distribuzione di probabilità gaussiana e m=4000

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.274	0.253	0.747	0.726	0.744	0.735	0.263	0.737
Holdout	0.261	0.242	0.758	0.739	0.493	0.591	0.246	0.754

In Figura 107 e Tabella 212 sono riportate la curva e l'area sotto la curva ROC.

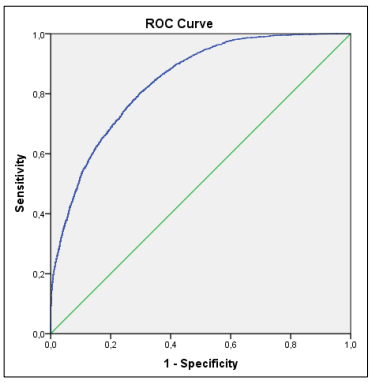

Figura 107 - Curva ROC - Logit: SONCA con distribuzione di probabilità gaussiana e m=4000

Tabella 212: Area sotto la curva ROC - Logit: SONCA con distribuzione di probabilità gaussiana e m=4000

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.841	0.003	0.000	0.835	0.848

SONCA con distribuzione di probabilità gaussiana e m=8000

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo gaussiana e m=8000. In Tabella 213 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA con distribuzione di probabilità gaussiana e m=8000

Tabella 213: Distribuzioni di classe per il training set: SONCA con distribuzione di probabilità gaussiana e m=8000

Variabile	N	%
0	8'044	50.3
1	7'957	49.7
Totale	16'001	100.00

Albero di regressione con il ricampionamento: SONCA con distribuzione di probabilità gaussiana e m=8000

In Figura 108 e Figura 109sono riportati gli alberi di regressione training avendo implementato SONCA con distribuzione di probabilità gaussiana e m=8000.

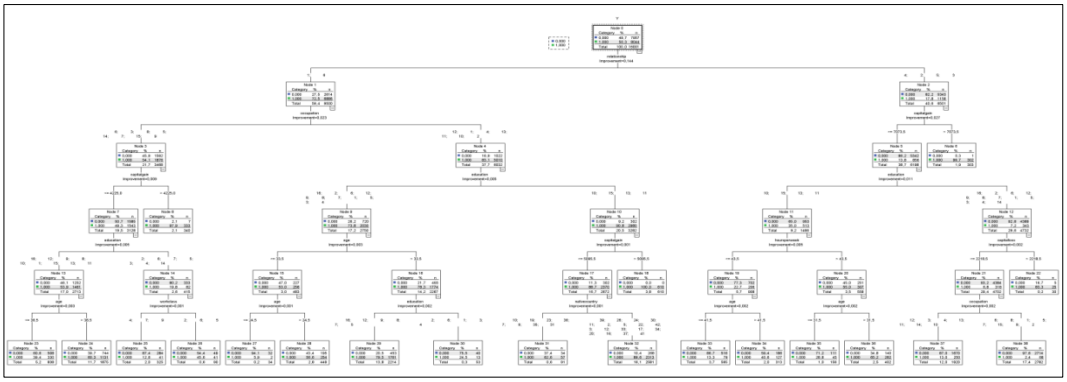

Figura 108 – Albero di regressione per il training set: SONCA con distribuzione di probabilità gaussiana e m=8000

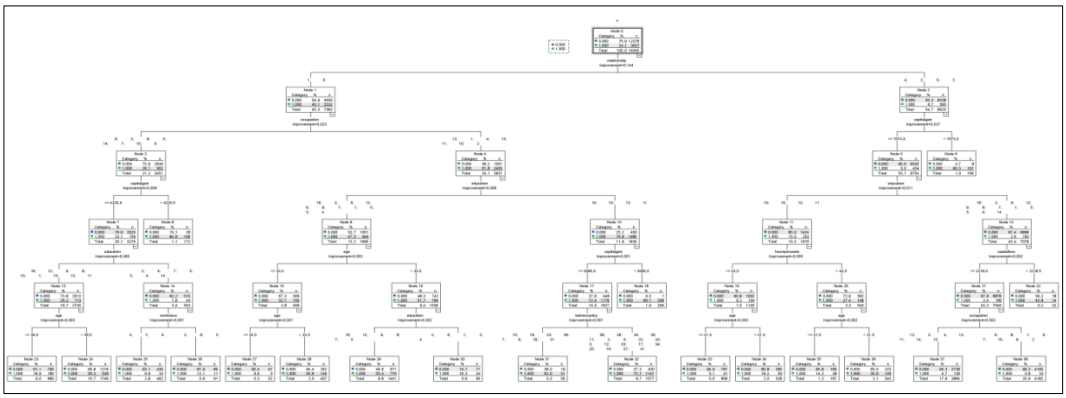

Figura 109 – Albero di regressione per l'holdout set: SONCA con distribuzione di probabilità gaussiana e m=8000

In Tabella 214 e Tabella 215 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 214: Matrice di confusione - CART: SONCA con distribuzione di probabilità gaussiana e m=8000

Sample		Predicted		
		1	0	Percent Correct
Training	1	7048	996	87.6%
	0	1847	6110	76.8%
	Overall Percentage	55.6%	44.4%	82.2%
Holdout	1	3370	557	85.8%
	0	2876	9502	76.8%
	Overall Percentage	38.3%	61.7%	78.9%

Tabella 215: Misure di performance - CART: SONCA con distribuzione di probabilità gaussiana con m=8000

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.124	0.232	0.768	0.876	0.792	0.832	0.178	0.822
Holdout	0.142	0.232	0.768	0.858	0.540	0.663	0.211	0.789

In Figura 110 e Tabella 216 sono riportate la curva e l'area sotto la curva ROC.

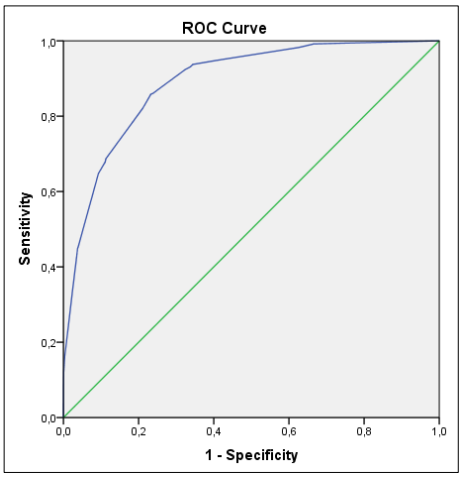

Figura 110 - Curva ROC - CART: SONCA con distribuzione di probabilità gaussiana e m=8000

Tabella 216: Area sotto la curva ROC - CART: SONCA con distribuzione di probabilità gaussiana e m=8000

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.889	0.003	0.000	0.884	0.895

Logit con il ricampionamento: SONCA con distribuzione di probabilità gaussiana e m=8000

In Tabella 217 sono riportate le variabili nell'equazione.

Tabella 217: Variabili nell'equazione: SONCA con distribuzione di probabilità gaussiana e m=8000

	B	S.E.	Wald	df	Sig.	Exp(B)
age	0.050	0.002	959.066	1	0.000	1.051
fnlwgt	2.365E-06	0.000	94.546	1	0.000	1.000
educationnum	0.312	0.009	1327.902	1	0.000	1.366
capitalgain	3.093 E-04	0.000	464.652	1	0.000	1.000
capitalloss	0.001	0.000	226.612	1	0.000	1.001
hoursperweek	0.049	0.002	669.604	1	0.000	1.050
Constant	-8.113	0.158	2624.052	1	0.000	0.000

In Tabella 218 e Tabella 219 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 218: Matrice di confusione - Logit: SONCA con distribuzione di probabilità gaussiana con m=8000

Sample		Predicted		
		1	0	Percent Correct
Training	1	5781	2263	71.9%
	0	1892	6065	76.2%
	Overall Percentage	48.0%	52.0%	74.0%
Holdout	1	2889	1038	73.6%
	0	2976	9402	76.0%
	Overall Percentage	36.0%	64.0%	75.4%

Tabella 219: Misure di performance - Logit: SONCA con distribuzione di probabilità gaussiana con m=8000

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.281	0.238	0.762	0.719	0.753	0.736	0.260	0.740
Holdout	0.264	0.240	0.760	0.736	0.493	0.590	0.246	0.754

In Figura 111 e Tabella 220 sono riportate la curva e l'area sotto la curva ROC.

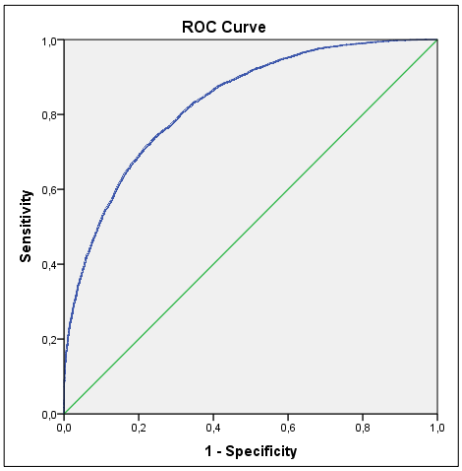

Figura 111 - Curva ROC - Logit: SONCA con distribuzione di probabilità gaussiana con m=8000

Tabella 220: Area sotto la curva ROC - Logit: SONCA con distribuzione di probabilità gaussiana con m=8000

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.832	0.004	0.000	0.825	0.839

SONCA con distribuzione di probabilità gaussiana e m=12000

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo gaussiana e m=12000. In Tabella 221 sono riportate le

distribuzioni di frequenza per il set di training avendo implementato SONCA con distribuzione di probabilità gaussiana e m=12000

Tabella 221: Distribuzioni di classe per il training set: SONCA con distribuzione di probabilità gaussiana e m=12000

Variabile	N	%
0	11989	49.82
1	12078	50.18
Totale	24067	100.00

Albero di regressione con il ricampionamento: SONCA con distribuzione di probabilità gaussiana e m=12000

In Figura 112 e Figura 113 sono riportati gli alberi di regressione training avendo implementato SONCA con distribuzione di probabilità gaussiana e m=12000.

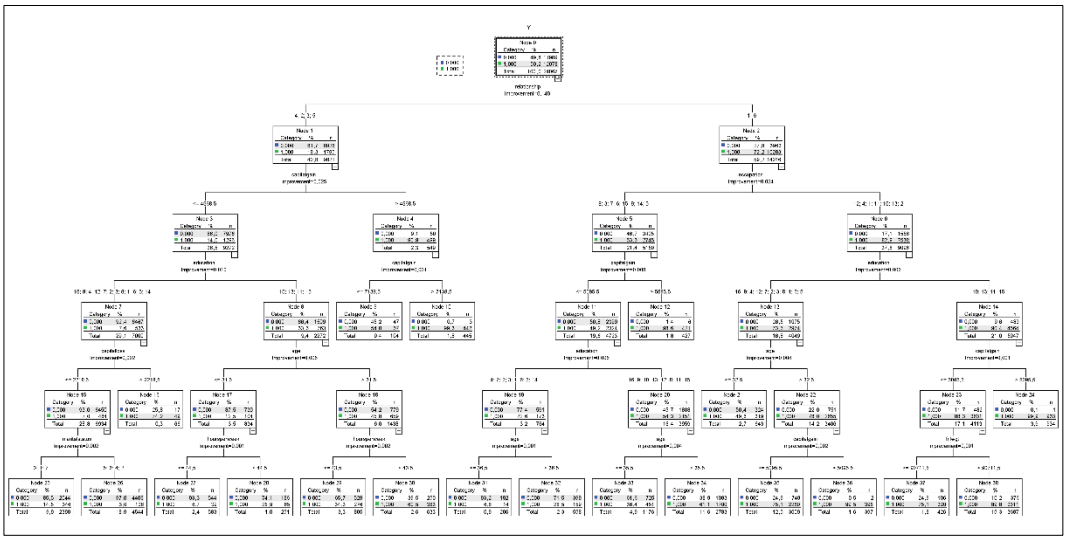

Figura 112 – Albero di regressione per il training set: SONCA con distribuzione di probabilità gaussiana e m=12000

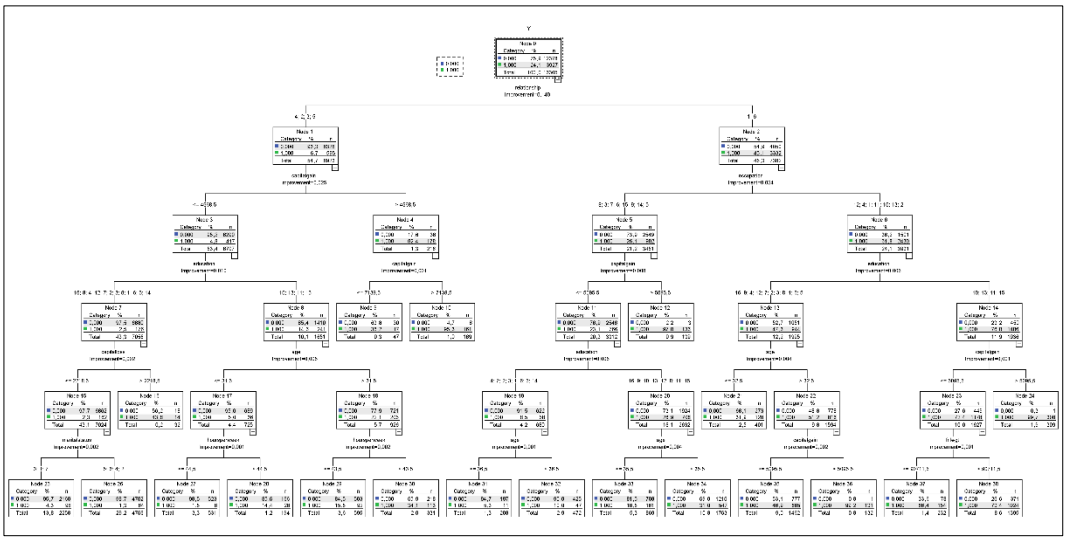

Figura 113 – Albero di regressione per l'holdout set: SONCA con distribuzione di probabilità gaussiana e m=12000

In Tabella 214 e Tabella 215 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 222: Matrice di confusione - CART: SONCA con distribuzione di probabilità gaussiana e m=12000

Sample		Predicted		
		1	0	Percent Correct
Training	1	10271	1807	85.0%
	0	2640	9349	78.0%
	Overall Percentage	53.6%	46.4%	81.5%
Holdout	1	3290	637	83.8%
	0	2721	9657	78.0%
	Overall Percentage	36.9%	63.1%	79.4%

Tabella 223: Misure di performance - CART: SONCA con distribuzione di probabilità gaussiana con m=12000

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.150	0.220	0.780	0.850	0.796	0.822	0.185	0.815
Holdout	0.162	0.220	0.780	0.838	0.547	0.662	0.206	0.794

In Figura 114 e Tabella 216 sono riportate la curva e l'area sotto la curva ROC.

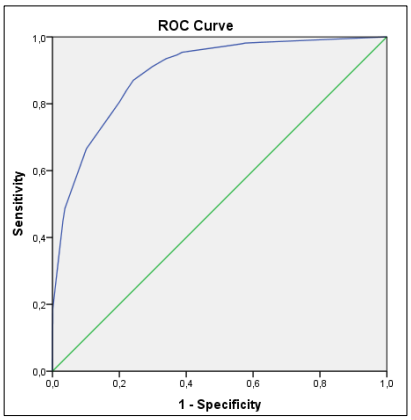

Figura 114 - Curva ROC - CART: SONCA con distribuzione di probabilità gaussiana e m=12000

Tabella 224: Area sotto la curva ROC - CART: SONCA con distribuzione di probabilità gaussiana e m=12000

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.893	0.003	0.000	0.888	0.899

Logit con il ricampionamento: SONCA con distribuzione di probabilità gaussiana e m=8000

In Tabella 225 sono riportate le variabili nell'equazione.

Tabella 225: Variabili nell'equazione: SONCA con distribuzione di probabilità gaussiana e m=12000

	B	S.E.	Wald	df	Sig.	Exp(B)
age	0.052	0.001	1558.34	1	0.000	1.053
fnlwgt	0.000	0.000	224.76	1	0.000	1.000
educationnum	0.308	0.007	1965.17	1	0.000	1.361
capitalgain	0.000	0.000	726.41	1	0.000	1.000
capitalloss	0.001	0.000	375.57	1	0.000	1.001
hoursperweek	0.042	0.001	810.87	1	0.000	1.043
Constant	-7.959	0.127	3943.63	1	0.000	0.000

In Tabella 226 e Tabella 227 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set.

Tabella 226: Matrice di confusione - Logit: SONCA con distribuzione di probabilità triangolare e m=12000

Sample		Predicted		
		1	0	Percent Correct
Training	1	8605	3473	71.2%
	0	2896	9093	75.8%
	Overall Percentage	47.8%	52.2%	73.5%
Holdout	1	2857	1070	72.8%
	0	3006	9372	75.7%
	Overall Percentage	24.4%	43.4%	75.0%

Tabella 227: Misure di performance per- Logit: SONCA con distribuzione di probabilità triangolare con m=12000

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.288	0.242	0.758	0.712	0.748	0.730	0.265	0.735
Holdout	0.272	0.243	0.757	0.728	0.487	0.584	0.250	0.750

In Figura 115 e Tabella 228 sono riportate la curva e l'area sotto la curva ROC.

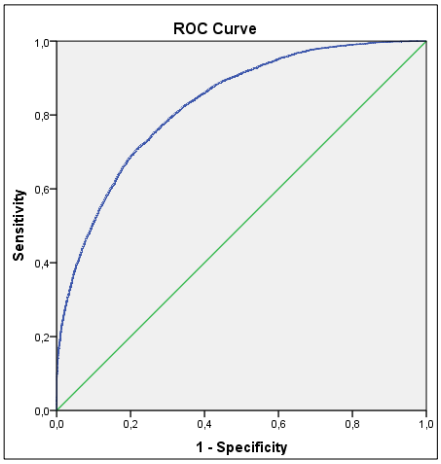

Figura 115 - Curva ROC - Logit: SONCA con distribuzione di probabilità triangolare con m=12000

Tabella 228: Area sotto la curva ROC - Logit: SONCA con distribuzione di probabilità triangolare con m=12000

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.829	0.004	0.000	0.822	0.836

Appendice 2 – Comparazione di SONCA con altri studi

A.2.1. Cover type

Come è stato osservato la variabile di risposta cover type è estremamente sbilanciata, la modalità 0 (la classe di maggioranza) ha una frequenza del 92.8%, mentre la classe 1 (la classe di minoranza) ha una frequenza del 7.1% inferiore del 10%. Per tale dataset sono stati stimati l'albero di regressione e il logit nei seguenti casi:

- Dataset originale;
- Ricampionamento con SONCA;
- Ricampionamento con SMOTE;
- Ricampionamento con ROSE.

Dataset originale

Albero di regressione senza il ricampionamento

In Figura 116 e Figura 117 sono riportati gli alberi di regressione senza bilanciare la variabile di risposta.

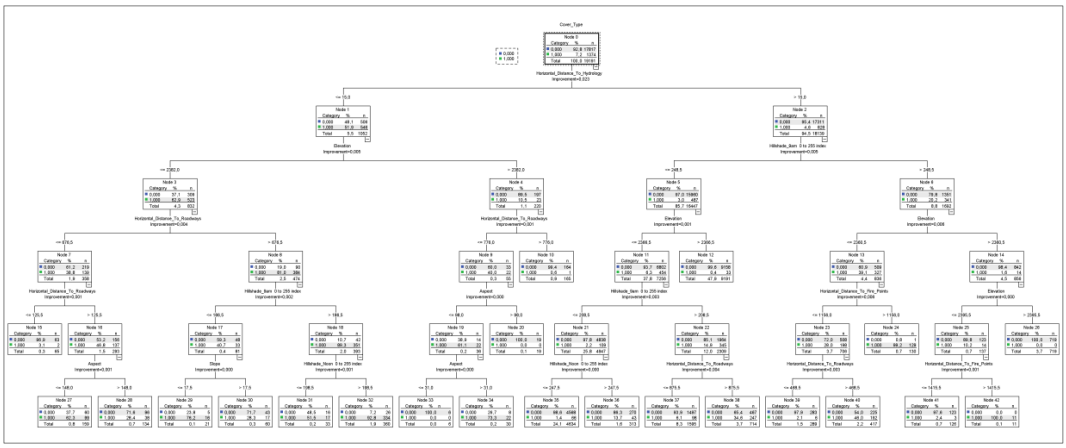

Figura 116 – Albero di regressione per il training set senza ricampionamento

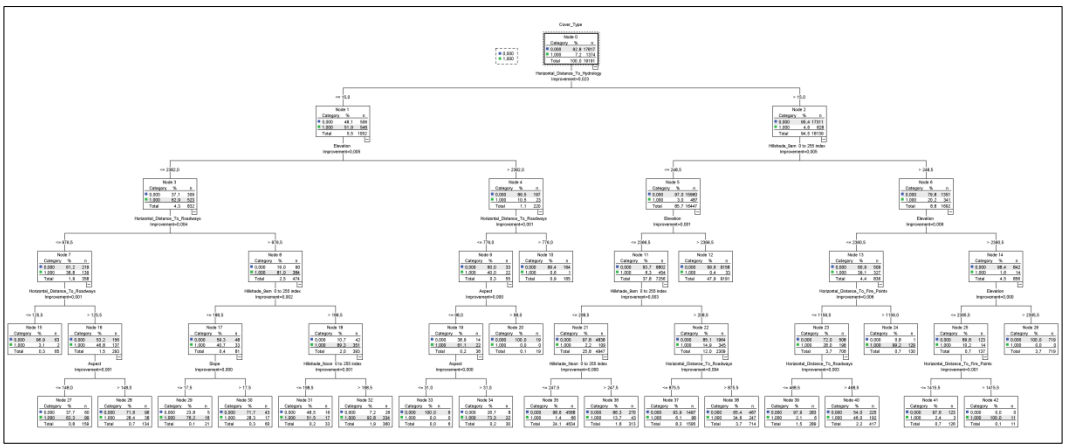

Figura 117 – Albero di regressione per l'holdout set senza ricampionamento

In Tabella 229 e Tabella 230 sono riportate sia la matrice di confusione sia le misure di prestazioni per entrambi i set dati, training e holdout.

Tabella 229: Matrice di confusione senza ricampionamento

Sample	Observed	Predicted		
		0	1	Percent Correct
Training	0	17701	116	99,3%
	1	746	628	45,7%
	Overall Percentage	96,1%	3,9%	95,5%
Holdout	0	17799	138	99,2%
	1	746	627	45,7%
	Overall Percentage	96,0%	4,0%	95,4%

Tabella 230: Misure di performance senza ricampionamento

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.543	0.007	0.993	0.457	0.844	0.593	0.045	0.955
Holdout	0.543	0.008	0.992	0.457	0.820	0.587	0.046	0.954

In Figura 118 e in Tabella 231 sono riportate la curva ROC e l'area sottesa alla curva ROC.

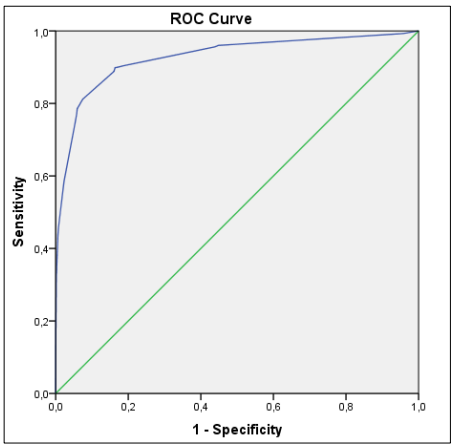

Figura 118 - Curva ROC per l'holdout set senza ricampionamento

Tabella 231: Area sotto la curva ROC per l'holdout set senza ricampionamento

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.927	0.004	0.000	0.919	0.936

Logit senza il ricampionamento

In Tabella 232 sono riportate le variabili nell'equazione.

Tabella 232: Variabili nell'equazione per il training set senza ricampionamento

	B	S.E.	Wald	df	Sig.	Exp(B)
Elevation	-0.011	0.000	1210.839	1	0.000	0.989
Slope	-0.052	0.005	102.605	1	0.000	0.949
Horizontal_Distance_To_Hydrology	-0.008	0.001	190.233	1	0.000	0.992
Vertical_Distance_To_Hydrology	0.016	0.001	156.222	1	0.000	1.017
Horizontal_Distance_To_Roadways	0.002	0.000	419.905	1	0.000	1.002
Hillshade_9am0to255index	0.036	0.001	726.456	1	0.000	1.037
Hillshade_Noon0to255index	0.014	0.002	61.461	1	0.000	1.014
Horizontal_Distance_To_Fire_Points	0.001	0.000	181.593	1	0.000	1.001
Constant	11.095	.749	219.302	1	0.000	65849.961

In Tabella 233 e Tabella 234 sono riportate la matrice di confusione che le misure di prestazioni sia per il dataset di training che di holdout.

Tabella 233: Matrice di confusione senza ricampionamento

Sample		Predicted		
		1	0	Percent Correct
Training	1	238	17579	1.3%
	0	398	976	71.0%
	Overall Percentage	3.3%	96.7%	6.3%
Holdout	1	262	17675	1.5%
	0	392	981	71.4%
	Overall Percentage	3.4%	97.2%	6.4%

Tabella 234: Misure di performance senza ricampionamento

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.987	0.290	0.710	0.013	0.374	0.026	0.937	0.063
Holdout	0.985	0.286	0.714	0.015	0.401	0.028	0.936	0.064

In Figura 119 e in Tabella 235 sono riportate la curva ROC e l'area sottesa alla curva ROC.

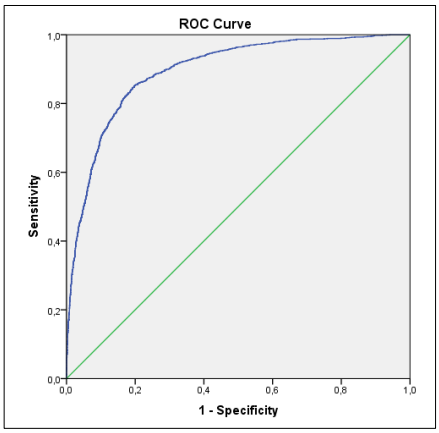

Figura 119 - Curva ROC per l'holdout set senza ricampionamento

Tabella 235: Area sotto la curva ROC per l'holdout set senza ricampionamento

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.895	0.004	0.000	0.886	0.903

SONCA con distribuzione di probabilità triangolare

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo triangolare. In

Tabella 236 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA.

Tabella 236: Distribuzioni di classe per il training set: SONCA

Variabile	N	%
1	9’445	49.98
0	9’451	50.02
Totale	18896	100.00

Albero di regressione con il ricampionamento: SONCA

In Figura 120 e Figura 121 sono riportati gli alberi di regressione training avendo implementato SONCA.

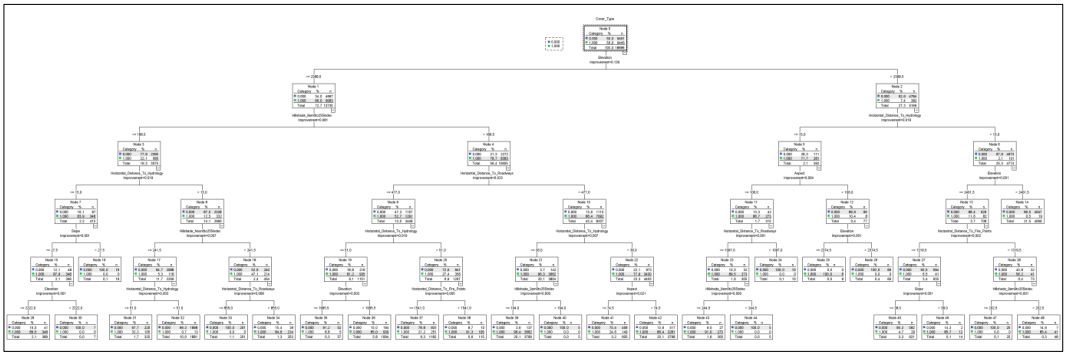

Figura 120 – Albero di regressione per il training set: SONCA

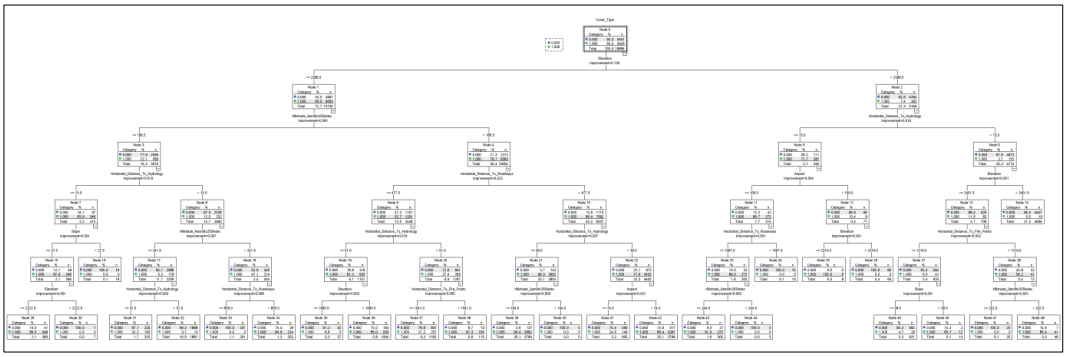

Figura 121 – Albero di regressione per l’holdout set: SONCA

In Tabella 237 e Tabella 238 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l’holdout set, avendo implementato SONCA.

Tabella 237: Matrice di confusione - CART: SONCA

Sample		Predicted		
		1	0	Percent Correct
Training	1	8874	571	94.0%
	0	944	8507	90.0%
	Overall Percentage	52.0%	48.0%	92.0%
Holdout	1	1272	101	92.6%
	0	1782	16155	90.1%
	Overall Percentage	15.8%	84.2%	90.2%

Tabella 238: Misure di performance - CART: SONCA

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.060	0.100	0.900	0.940	0.904	0.921	0.080	0.920
Holdout	0.074	0.099	0.901	0.926	0.417	0.575	0.098	0.902

In Figura 122 e Tabella 239 sono riportate la curva ROC e Area sotto la curva.

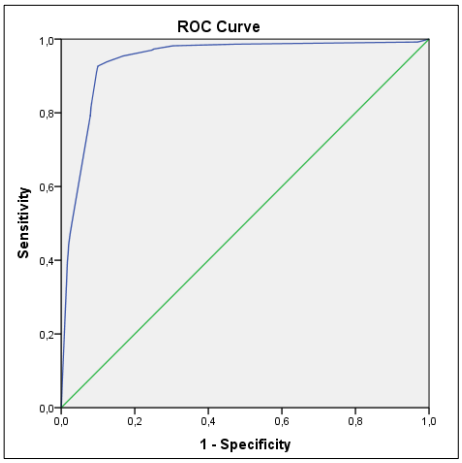

Figura 122 - Curva ROC - CART: SONCA

Tabella 239: Area sotto la curva ROC - CART: SONCA

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.943	0.003	0.000	0.937	0.949

Logit con il ricampionamento: SONCA

In Tabella 240 sono riportate le variabili nell'equazione.

Tabella 240: Variabili nell'equazione: SONCA con distribuzione di probabilità triangolare

Variables in the Equation	B	S.E.	Wald	df	Sig.	Exp(B)
Elevation	-0.01	0.00	3135.06	1.00	0.00	0.99
Slope	-0.12	0.01	164.81	1.00	0.00	0.89
Horizontal_Distance_To_Hydrology	-0.01	0.00	544.85	1.00	0.00	0.99
Vertical_Distance_To_Hydrology	0.01	0.00	203.26	1.00	0.00	1.01
Horizontal_Distance_To_Roadways	0.00	0.00	1030.38	1.00	0.00	1.00
Hillshade_9am0to255index	-0.05	0.01	34.02	1.00	0.00	0.95
Hillshade_Noon0to255index	0.10	0.01	164.40	1.00	0.00	1.10
Hillshade_3pm0to255index	-0.08	0.01	102.92	1.00	0.00	0.93
Horizontal_Distance_To_Fire_Points	0.00	0.00	179.84	1.00	0.00	1.00
Constant	27.10	1.61	282.69	1.00	0.00	590119900590.10

In Tabella 241 e Tabella 242 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA.

Tabella 241: Matrice di confusione - Logit: SONCA

Sample		Predicted		
		1	0	Percent Correct
Training	1	8035	1410	85.1%
	0	1630	7821	82.8%
	Overall Percentage	51.1%	48.9%	83.9%
Holdout	1	1122	251	81.7%
	0	3203	14734	82.1%
	Overall Percentage	22.4%	77.6%	82.1%

Tabella 242: Misure di performance - Logit: SONCA

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.149	0.172	0.828	0.851	0.831	0.841	0.161	0.839
Holdout	0.183	0.179	0.821	0.817	0.259	0.394	0.179	0.821

In Figura 123 e Tabella 243 sono riportate la curva e l'area sotto la curva ROC.

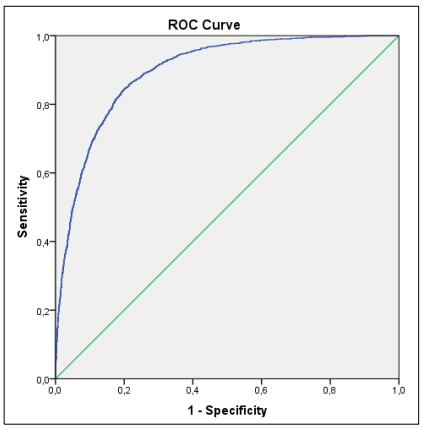

Figura 123 - Curva ROC - Logit: SONCA

Tabella 243: Area sotto la curva ROC - Logit: SONCA

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.897	0.004	0.000	0.889	0.904

SONCA con distribuzione di probabilità gaussiana

Il set di training è stato ricampionato utilizzando SONCA. In Tabella 244 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA.

Tabella 244: Distribuzioni di classe per il training set: SONCA

Variabile	N	%
1	9'406	49.72
0	9'511	50.28
Totale	18'917	100.00

Albero di regressione con il ricampionamento: SONCA

In Figura 124 e Figura 125 sono riportati gli alberi di regressione training avendo implementato SONCA.

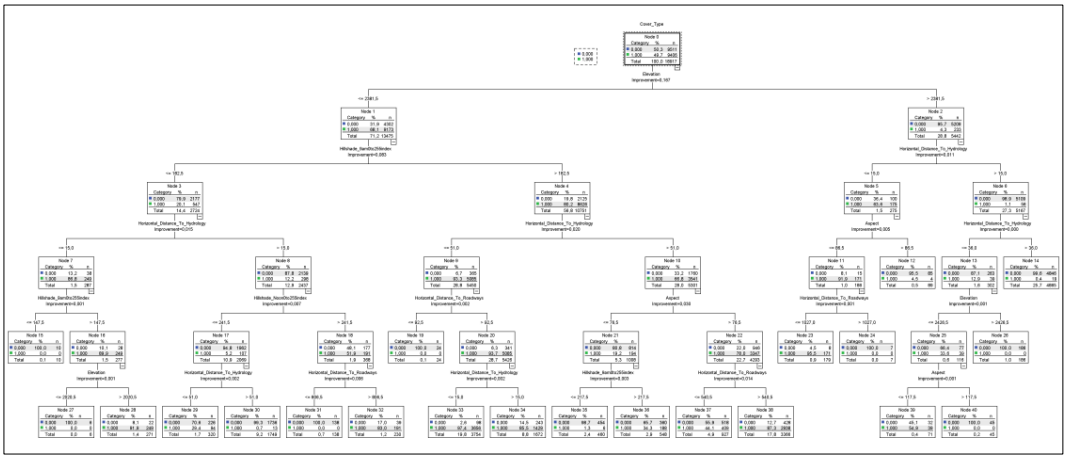

Figura 124 – Albero di regressione per il training set: SONCA

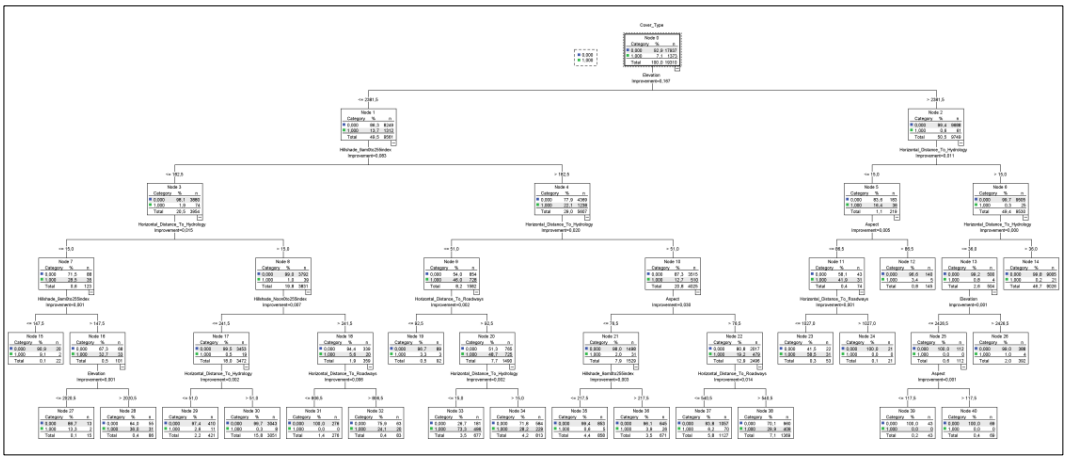

Figura 125 – Albero di regressione per l'holdout set: SONCA

In Tabella 245 e Tabella 246 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA.

Tabella 245: Matrice di confusione - CART: SONCA

Sample		Predicted		
		1	0	Percent Correct
Training	1	8673	733	92.2%
	0	870	8641	90.9%
	Overall Percentage	50.4%	49.6%	91.5%
Holdout	1	1216	157	88.6%
	0	1908	16029	89.4%
	Overall Percentage	16.2%	83.8%	89.3%

Tabella 246: Misure di performance - CART: SONCA

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.078	0.091	0.909	0.922	0.909	0.915	0.085	0.915
Holdout	0.114	0.106	0.894	0.886	0.389	0.541	0.107	0.893

In Figura 126e Tabella 247 sono riportate la curva ROC e Area sotto la curva.

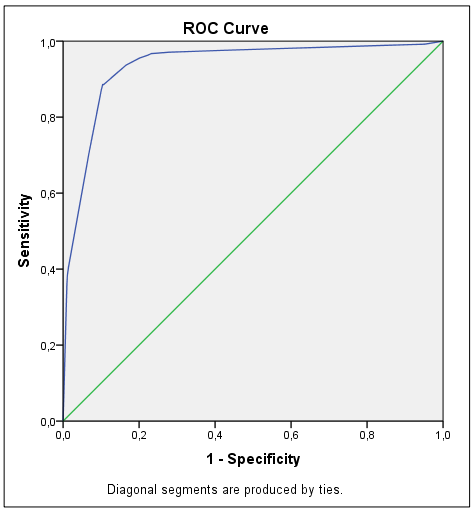

Figura 126 - Curva ROC - CART: SONCA

Tabella 247: Area sotto la curva ROC - CART: SONCA

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.935	0.004	0.000	0.928	0.942

Logit con il ricampionamento: SONCA

In Tabella 248 sono riportate le variabili nell'equazione.

Tabella 248: Variabili nell'equazione: SONCA

Elevation	-0.01	0.00	3284.76	1.00	0.00	0.99
Slope	-0.13	0.01	179.61	1.00	0.00	0.88
Horizontal_Distance_To_Hydrology	-0.01	0.00	470.00	1.00	0.00	0.99
Vertical_Distance_To_Hydrology	0.01	0.00	178.20	1.00	0.00	1.01
Horizontal_Distance_To_Roadways	0.00	0.00	759.23	1.00	0.00	1.00
Hillshade_9am0to255index	-0.07	0.01	47.67	1.00	0.00	0.93
Hillshade_Noon0to255index	0.11	0.01	198.07	1.00	0.00	1.12
Hillshade_3pm0to255index	-0.09	0.01	123.22	1.00	0.00	0.91
Horizontal_Distance_To_Fire_Points	0.00	0.00	21.56	1.00	0.00	1.00
Constant	32.71	1.77	343.17	1.00	0.00	1608.02E+11

In Tabella 249 e Tabella 250 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA.

Tabella 249: Matrice di confusione - Logit: SONCA

Sample		Predicted		
		1	0	Percent Correct
Training	1	8240	1166	87.6%
	0	1332	8179	86.0%
	Overall Percentage	50.6%	49.4%	86.8%
Holdout	1	1122	251	81.7%
	0	3051	14886	83.0%
	Overall Percentage	21.6%	78.4%	82.9%

Tabella 250: Misure di performance - Logit: SONCA

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.124	0.140	0.860	0.876	0.861	0.868	0.132	0.868
Holdout	0.183	0.170	0.830	0.817	0.269	0.405	0.171	0.829

In Figura 127 e Tabella 251 sono riportate la curva e l'area sotto la curva ROC.

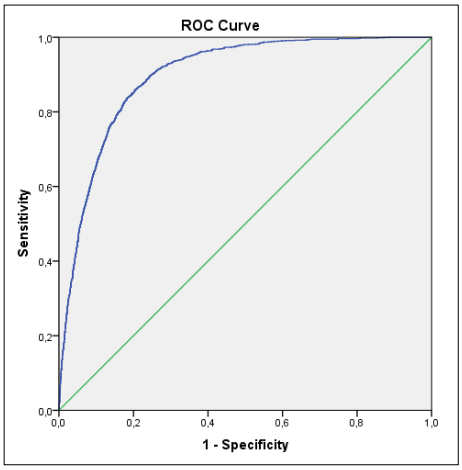

Figura 127 - Curva ROC - Logit: SONCA

Tabella 251: Area sotto la curva ROC - Logit: SONCA

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.898	0.004	0.000	0.891	0.905

SMOTE

Il set di training è stato ricampionato utilizzando SMOTE. In Tabella 252 sono riportate le distribuzioni di frequenza per il set di training.

Tabella 252: Distribuzioni di classe per il training set: SMOTE

Cover_Type	N	F=N/N$_{tot}$[%]
1	5'496	23.8
0	17'564	76.2
Totale	23’060	100.0

Albero di regressione con il ricampionamento: SMOTE

In Figura 128 e Figura 129 sono riportati gli alberi di regressione training avendo implementato SMOTE.

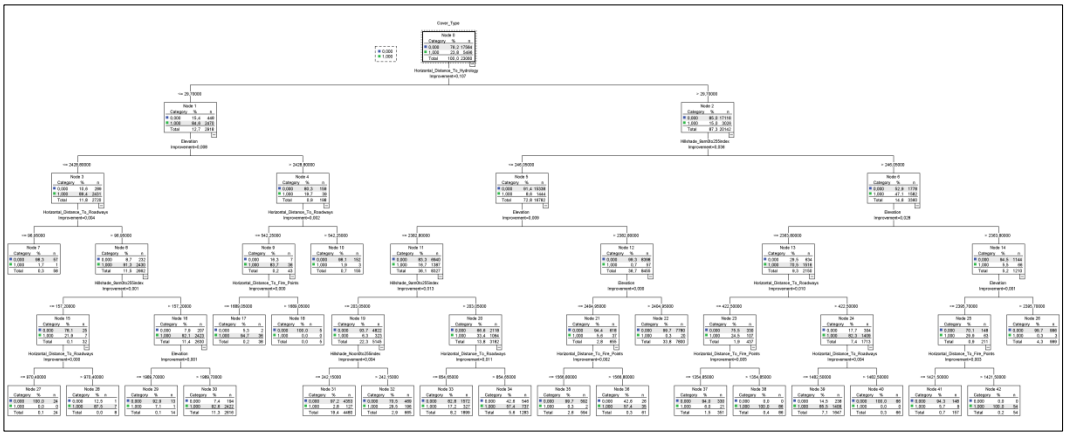

Figura 128 – Albero di regressione per il training set: SMOTE

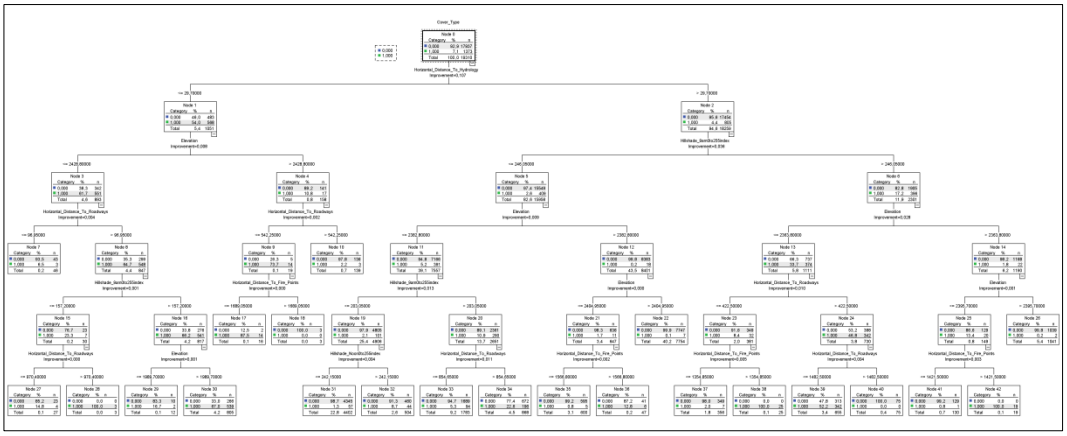

Figura 129 – Albero di regressione per l’holdout set: SMOTE

In Tabella 253 e Tabella 254 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SMOTE.

Tabella 253: Matrice di confusione - CART: SMOTE

Sample		Predicted		
		1	0	Percent Correct
Training	1	4786	710	87.1%
	0	1007	16557	94.3%
	Overall Percentage	25.1%	74.9%	92.6%
Holdout	1	1144	229	83.3%
	0	1294	16643	92.8%
	Overall Percentage	12.6%	87.4%	92.1%

Tabella 254: Misure di performance - CART: SONCA

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.129	0.057	0.943	0.871	0.826	0.848	0.074	0.926
Holdout	0.167	0.072	0.928	0.833	0.469	0.600	0.079	0.921

In Figura 130 e Tabella 255 sono riportate la curva ROC e Area sotto la curva.

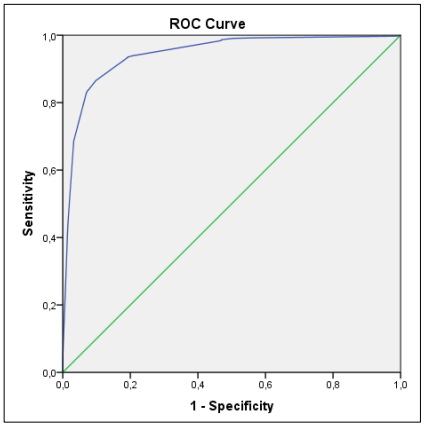

Figura 130 - Curva ROC - CART: SMOTE

Tabella 255: Area sotto la curva ROC - CART: SMOTE

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.944	0.003	0.000	0.938	0.950

Logit con il ricampionamento: SMOTE

In Tabella 256 sono riportate le variabili nell'equazione.

Tabella 256: Variabili nell'equazione: SMOTE

Variables in the Equation	B	S.E.	Wald	df	Sig.	Exp(B)
Elevation	-0.013	0.000	3390.665	1	0.000	0.987
Slope	-0.108	0.009	146.328	1	0.000	0.898
Horizontal_Distance_To_Hydrology	-0.006	0.000	423.891	1	0.000	0.994
Vertical_Distance_To_Hydrology	0.013	0.001	317.432	1	0.000	1.013
Horizontal_Distance_To_Roadways	0.002	0.000	911.840	1	0.000	1.002
Hillshade_9am0to255index	-0.026	0.009	8.304	1	0.004	0.974
Hillshade_Noon0to255index	0.073	0.007	97.263	1	0.000	1.076
Hillshade_3pm0to255index	-0.054	0.007	53.922	1	0.000	0.947
Horizontal_Distance_To_Fire_Points	0.001	0.000	549.629	1	0.000	1.001
Constant	24.561	1.575	243.131	1	0.000	46439560915.8295

In Tabella 257 e Tabella 258 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SMOTE.

Tabella 257: Matrice di confusione - Logit: SMOTE

Sample		Predicted		
		1	0	Percent Correct
Training	1	3577	1919	65.1%
	0	1144	16420	93.5%
	Overall Percentage	20.5%	79.5%	86.7%
Holdout	1	870	503	63.4%
	0	1377	16560	92.3%
	Overall Percentage	11.6%	88.4%	90.3%

Tabella 258: Misure di performance - Logit: SMOTE

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.349	0.065	0.935	0.651	0.758	0.700	0.133	0.867
Holdout	0.366	0.077	0.923	0.634	0.387	0.481	0.097	0.903

In Figura 131 e Tabella 259 sono riportate la curva e l'area sotto la curva ROC.

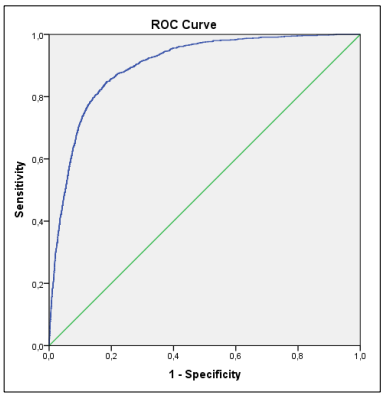

Figura 131 - Curva ROC - Logit: SMOTE

Tabella 259: Area sotto la curva ROC - Logit: SMOTE

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.900	0.004	00.000	0.893	0.908

ROSE

Il set di training è stato ricampionato utilizzando ROSE. In Tabella 260 sono riportate le distribuzioni di frequenza per il set di training.

Tabella 260: Distribuzioni di classe per il training set: ROSE

Variabile	N	%
1	9'464	49.3
0	9'727	50.7
Totale	19'191	100.0

Albero di regressione con il ricampionamento: ROSE

In Figura 132 e Figura 133 sono riportati gli alberi di regressione training avendo implementato ROSE.

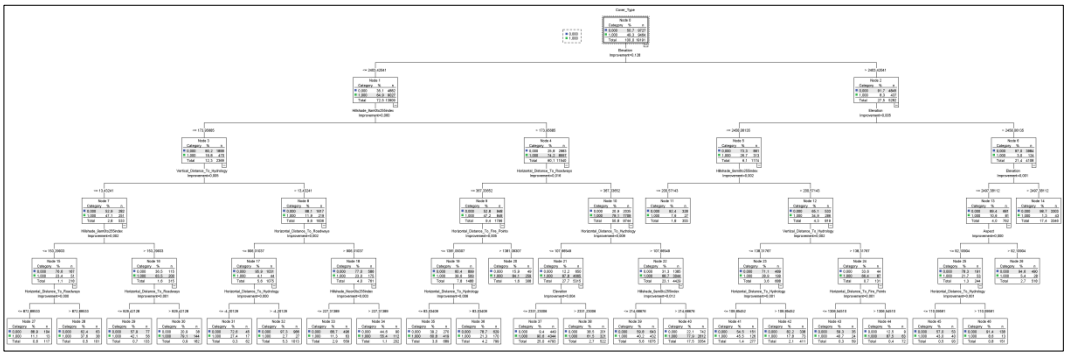

Figura 132 – Albero di regressione per il training set: ROSE

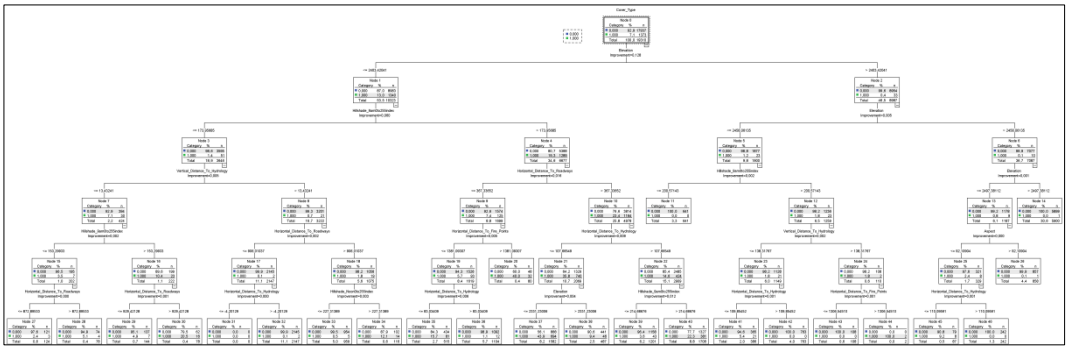

Figura 133 – Albero di regressione per l'holdout set: ROSE

In Tabella 261 e Tabella 262 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato ROSE.

Tabella 261: Matrice di confusione - CART: ROSE

Sample		Predicted		
		1	0	Percent Correct
Training	1	8274	1190	87.4%
	0	1848	7879	81.0%
	Overall Percentage	52.7%	47.3%	84.2%
Holdout	1	1266	107	92.2%
	0	3302	14635	81.6%
	Overall Percentage	23.7%	76.3%	82.3%

Tabella 262: Misure di performance - CART: ROSE

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.126	0.190	0.810	0.874	0.817	0.845	0.158	0.842
Holdout	0.078	0.184	0.816	0.922	0.277	0.426	0.177	0.823

In Figura 134 e Tabella 263 sono riportate la curva ROC e Area sotto la curva.

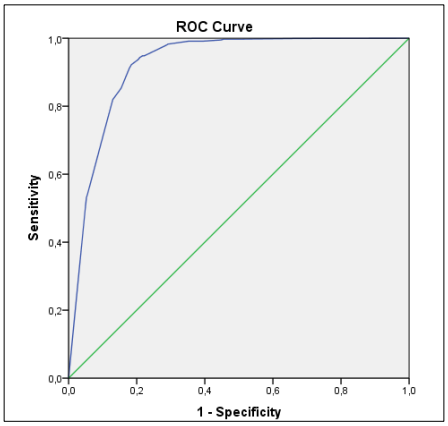

Figura 134 - Curva ROC - CART: ROSE

Tabella 263: Area sotto la curva ROC - CART: ROSE

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.922	0.003	0.000	0.917	0.927

Logit con il ricampionamento: ROSE

In Tabella 264 sono riportate le variabili nell'equazione.

Tabella 264: Variabili nell'equazione: ROSE

Variables in the Equation	B	S.E.	Wald	df	Sig.	Exp(B)
Elevation	-0.009	0.000	3264.95	1	0.000	0.991
Slope	-0.030	0.002	171.83	1	0.000	0.971
Horizontal_Distance_To_Hydrology	-0.003	0.000	391.63	1	0.000	0.997
Vertical_Distance_To_Hydrology	0.003	0.000	49.11	1	0.000	1.003
Horizontal_Distance_To_Roadways	0.001	0.000	369.34	1	0.000	1.001
Hillshade_9am0to255index	0.021	0.001	698.96	1	0.000	1.021
Hillshade_Noon0to255index	0.019	0.001	391.49	1	0.000	1.019
Hillshade_3pm0to255index	-0.007	0.001	159.50	1	0.000	0.993
Horizontal_Distance_To_Fire_Points	0.000	0.000	154.33	1	0.000	1.000
Constant	12.172	0.366	1107.826	1	0.000	193352.309

In Tabella 265 e Tabella 266 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato ROSE.

Tabella 265: Matrice di confusione - Logit: ROSE

Sample		Predicted		
		1	0	Percent Correct
Training	1	1146	227	83.5%
	0	2954	14983	83.5%
	Overall Percentage	21.2%	78.8%	83.5%
Holdout	1	7712	1752	81.5%
	0	2070	7657	78.7%
	Overall Percentage	51.0%	49.0%	80.1%

Tabella 266: Misure di performance - Logit: ROSE

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.165	0.165	0.835	0.835	0.280	0.419	0.165	0.835
Holdout	0.185	0.213	0.787	0.815	0.788	0.801	0.199	0.801

In Figura 135 e Tabella 173 sono riportate la curva e l'area sotto la curva ROC.

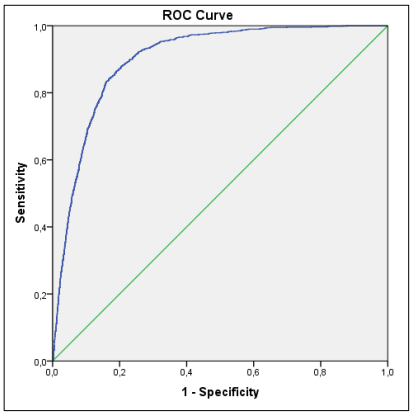

Figura 135 - Curva ROC - Logit: ROSE

Tabella 267: Area sotto la curva ROC - Logit: ROSE

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.901	0.003	0.000	0.894	0.908

A.2.2. Glass

Come è stato osservato la variabile di risposta type of glass è estremamente sbilanciata, la modalità 0 (la classe di maggioranza) ha una frequenza del 92.1%, mentre la classe 1 (la classe di minoranza) ha una frequenza del 7.9% inferiore del 10%. Per tale dataset sono stati stimati l'albero di regressione e il logit nei seguenti casi:

- Dataset originale;
- Ricampionamento con SONCA;
- Ricampionamento con SMOTE;
- Ricampionamento con ROSE.

Dataset originale

Albero di regressione senza il ricampionamento

In Figura 136 e Figura 137 sono riportati gli alberi di regressione senza bilanciare la variabile di risposta.

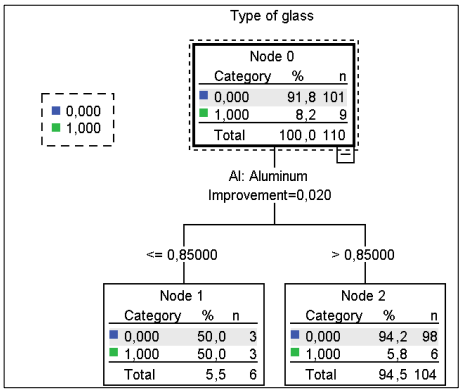

Figura 136 – Albero di regressione per il training set senza ricampionamento

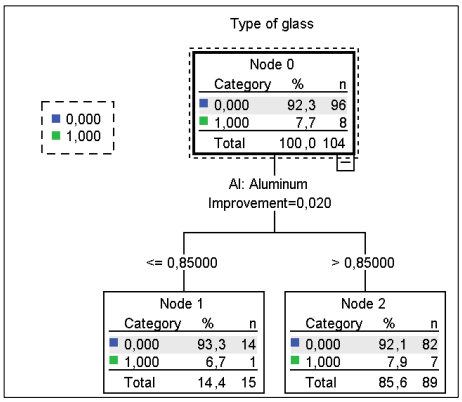

Figura 137 – Albero di regressione per l'holdout set senza ricampionamento

In Tabella 268 e Tabella 269 sono riportate sia la matrice di confusione sia le misure di prestazioni per entrambi i set dati, training e holdout.

Tabella 268: Matrice di confusione senza ricampionamento

Sample		Predicted		
		1	0	Percent Correct
Training	1	0	9	0.0%
	0	0	101	100.0%
	Overall Percentage	0.0%	100.0%	91.8%
Holdout	1	0	8	0.0%
	0	0	96	100.0%
	Overall Percentage	0.0%	100.0%	92.3%

Tabella 269: Misure di performance senza ricampionamento

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	1.000	0.000	1.000	0.000	0.000	0.000	0.082	0.918
Holdout	1.000	0.000	1.000	0.000	0.000	0.000	0.077	0.923

A causa dei risultati ottenuti non è stato possibile calcolare la curva ROC.

Logit senza il ricampionamento

In Tabella 270 sono riportate le variabili nell'equazione.

Tabella 270: Variabili nell'equazione per il training set senza ricampionamento

	B	S.E.	Wald	df	Sig.	Exp(B)
Aluminum	-2.453	1.131	4.700	1	.030	.086
Constant	.800	1.395	.329	1	.566	2.226

In

Tabella 271 e Tabella 272 sono riportate la matrice di confusione che le misure di prestazioni sia per il dataset di training che di holdout.

Tabella 271: Matrice di confusione senza ricampionamento

Sample		Predicted		
		1	0	Percent Correct
Training	1	0	9	0.0%
	0	0	101	100.0%
	Overall Percentage	0.0%	100.0%	91.8%
Holdout	1	0	8	0.0%
	0	1	95	99.0%
	Overall Percentage	0.9%	93.6%	91.3%

Tabella 272: Misure di performance senza ricampionamento

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	1.000	0.000	1.000	0.000	0.000	0.000	0.082	0.918
Holdout	1.000	0.010	0.990	0.000	0.000	0.000	0.087	0.913

A causa dei risultati ottenuti non è stato possibile calcolare la curva ROC.

SONCA con distribuzione di probabilità triangolare

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo triangolare. In Tabella 273 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA.

Tabella 273: Distribuzioni di classe per il training set: SONCA

Variabile	N	%
1	97	48.7
0	102	51.3
Totale	199	100.00

Albero di regressione con il ricampionamento: SONCA

In Figura 138 e Figura 139 sono riportati gli alberi di regressione training avendo implementato SONCA.

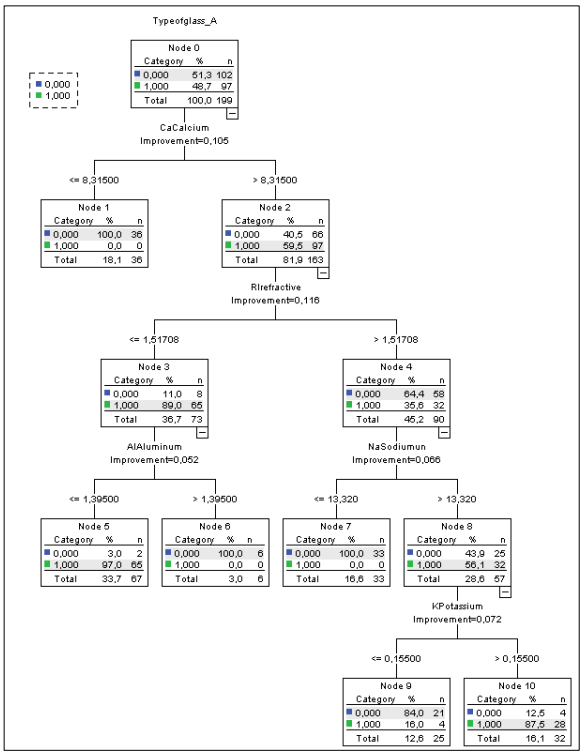

Figura 138 – Albero di regressione per il training set: SONCA

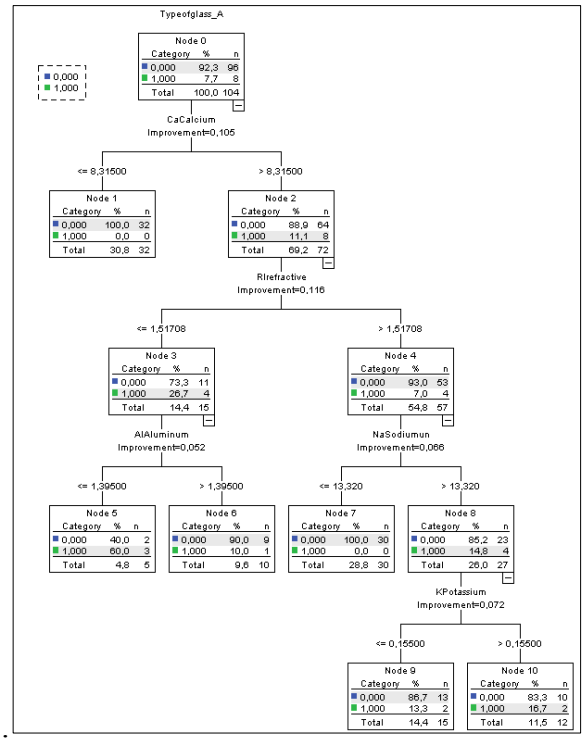

Figura 139 – Albero di regressione per l'holdout set: SONCA

In Tabella 274 e Tabella 275 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA.

Tabella 274: Matrice di confusione - CART: SONCA

Sample		Predicted		
		1	0	Percent Correct
Training	1	93	4	95.9%
	0	6	96	94.1%
	Overall Percentage	49.7%	50.3%	95.0%
Holdout	1	5	3	62.5%
	0	12	84	87.5%
	Overall Percentage	16.3%	83.7%	85.6%

Tabella 275: Misure di performance - CART: SONCA

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.041	0.059	0.941	0.959	0.939	0.949	0.050	0.950
Holdout	0.375	0.125	0.875	0.625	0.294	0.400	0.144	0.856

InFigura 140 e Tabella 276 sono riportate la curva ROC e Area sotto la curva.

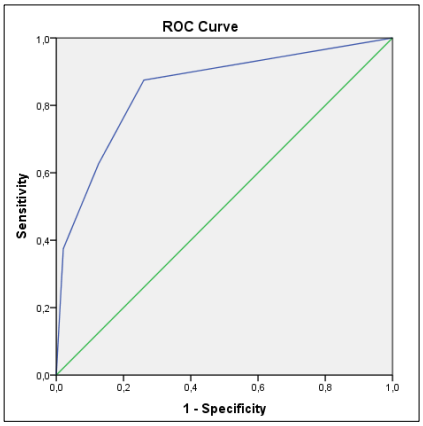

Figura 140 - Curva ROC - CART: SONCA

Tabella 276: Area sotto la curva ROC - CART: SONCA

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.851	0.077	0.001	0.700	1.000

Logit con il ricampionamento: SONCA

In

Tabella 277 sono riportate le variabili nell’equazione.

Tabella 277: Variabili nell'equazione: SONCA con distribuzione di probabilità triangolare

	B	S.E.	Wald	df	Sig.	Exp(B)
RIrefractive	-4989.778	719.287	48.124	1	0.000	0.000
MgMagnesium	8.262	1.903	18.840	1	0.000	3873.524
AlAluminum	-6.574	1.838	12.793	1	0.000	0.001
SiSilicon	-7.435	1.511	24.219	1	0.000	0.001
CaCalcium	10.335	2.025	26.042	1	0.000	30803.172
BaBarium	20.808	4.522	21.174	1	0.000	1088424086.873
Constant	8002.954	1157.579	47.797	1	0.000	-

In Tabella 278 e Tabella 279 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA.

Tabella 278: Matrice di confusione - Logit: SONCA

Sample		**Predicted**		
		1	0	Percent Correct
Training	1	97	0	100.0%
	0	5	97	95.1%
	Overall Percentage	51.3%	48.7%	97.5%
Holdout	1	6	2	75.0%
	0	27	69	71.9%
	Overall Percentage	16.6%	35.7%	72.1%

Tabella 279: Misure di performance - Logit: SONCA

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.000	0.049	0.951	1.000	0.951	0.975	0.025	0.975
Holdout	0.250	0.281	0.719	0.750	0.182	0.293	0.279	0.721

In Figura 141 e Tabella 281 sono riportate la curva e l'area sotto la curva ROC.

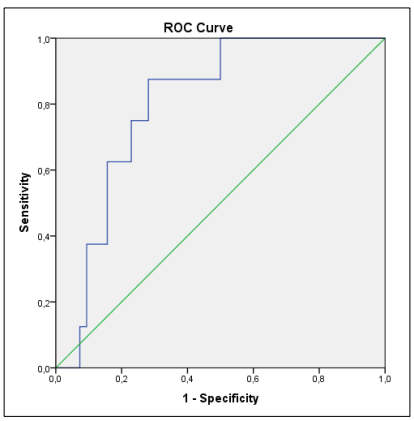

Figura 141 - Curva ROC - Logit: SONCA

Tabella 280: Area sotto la curva ROC - Logit: SONCA

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.802	0.057	0.005	0.691	0.913

SONCA con distribuzione di probabilità gaussiana

Il set di training è stato ricampionato utilizzando SONCA. In Tabella 281 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA.

Tabella 281: Distribuzioni di classe per il training set: SONCA

Variabile	N	%
1	96	45.6
0	106	54.4
Totale	195	100.0

Albero di regressione con il ricampionamento: SONCA

In Figura 142 e Figura 143 sono riportati gli alberi di regressione training avendo implementato SONCA.

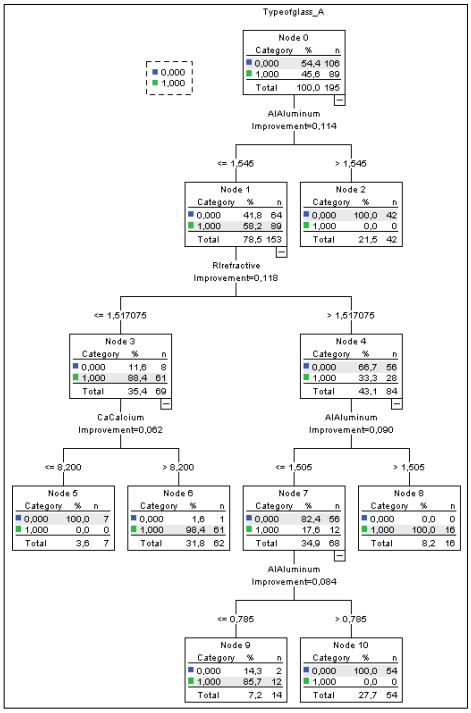

Figura 142 – Albero di regressione per il training set: SONCA

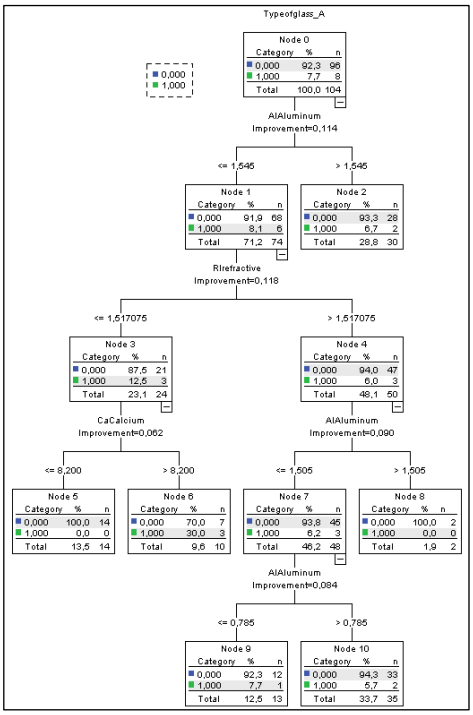

Figura 143 – Albero di regressione per l'holdout set: SONCA

In Tabella 282 e Tabella 283 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA.

Tabella 282: Matrice di confusione - CART: SONCA

Sample		Predicted		
		1	0	Percent Correct
Training	1	89	0	100.0%
	0	3	103	97.2%
	Overall Percentage	47.2%	52.8%	98.5%
Holdout	1	4	4	50.0%
	0	21	75	78.1%
	Overall Percentage	24.0%	76.0%	76.0%

Tabella 283: Misure di performance - CART: SONCA

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.000	0.028	0.972	1.000	0.967	0.983	0.015	0.985
Holdout	0.500	0.219	0.781	0.500	0.160	0.242	0.240	0.760

In Figura 144 e Tabella 284 sono riportate la curva ROC e Area sotto la curva

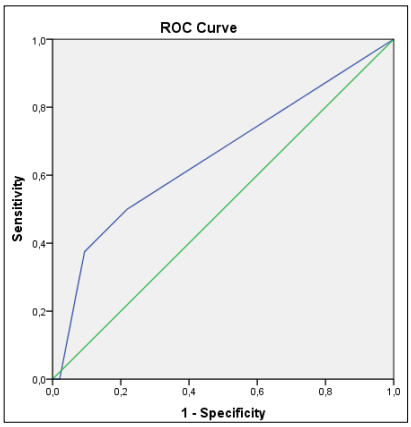

Figura 144 - Curva ROC - CART: SONCA

Tabella 284: Area sotto la curva ROC - CART: SONCA

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.654	0.113	0.148	0.433	0.875

Logit con il ricampionamento: SONCA

In Tabella 285 sono riportate le variabili nell'equazione.

Tabella 285: Variabili nell'equazione: SONCA

	B	S.E.	Wald	df	Sig.	Exp(B)
RIrefractive	-8282.793	1506.178	30.241	1	0.000	0.000
MgMagnesium	13.262	3.068	18.683	1	0.000	574706.201
AlAluminum	-9.504	2.733	12.091	1	0.001	.000
SiSilicon	-11.940	2.566	21.657	1	0.000	.000
CaCalcium	16.736	3.428	23.841	1	0.000	18554568.558
BaBarium	34.114	7.627	20.006	1	0.000	653736675809203.000
Constant	13257.974	2418.950	30.040	1	0.000	

In Tabella 286 e Tabella 287 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA.

Tabella 286: Matrice di confusione - Logit: SONCA

Sample		Predicted		
		1	0	Percent Correct
Training	1	84	5	94.4%
	0	4	102	96.2%
	Overall Percentage	45.1%	54.9%	95.4%
Holdout	1	6	2	75.0%
	0	24	72	75.0%
	Overall Percentage	15.4%	37.9%	75.0%

Tabella 287: Misure di performance - Logit: SONCA

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.056	0.038	0.962	0.944	0.955	0.949	0.046	0.954
Holdout	0.250	0.250	0.750	0.750	0.200	0.316	0.250	0.750

In Figura 145 e Tabella 288 sono riportate la curva e l'area sotto la curva ROC.

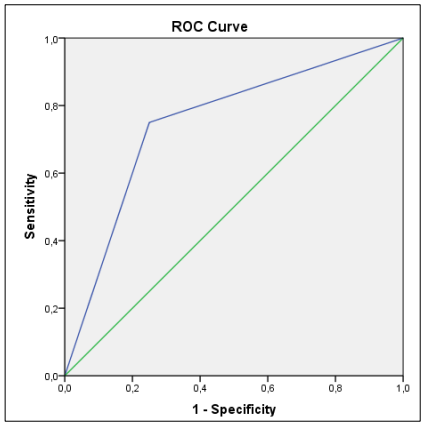

Figura 145 - Curva ROC - Logit: SONCA

Tabella 288: Area sotto la curva ROC - Logit: SONCA

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.750	0.092	0.019	0.569	0.931

SMOTE

Il set di training è stato ricampionato utilizzando SMOTE. In Tabella 289 sono riportate le distribuzioni di frequenza per il set di training.

Tabella 289: Distribuzioni di classe per il training set: SMOTE

Cover_Type	N	$F=N/N_{tot}[\%]$
1	36	29.8
0	85	70.2
Totale	121	100.0

Albero di regressione con il ricampionamento: SMOTE

In Figura 146 e Figura 147 sono riportati gli alberi di regressione training avendo implementato SMOTE.

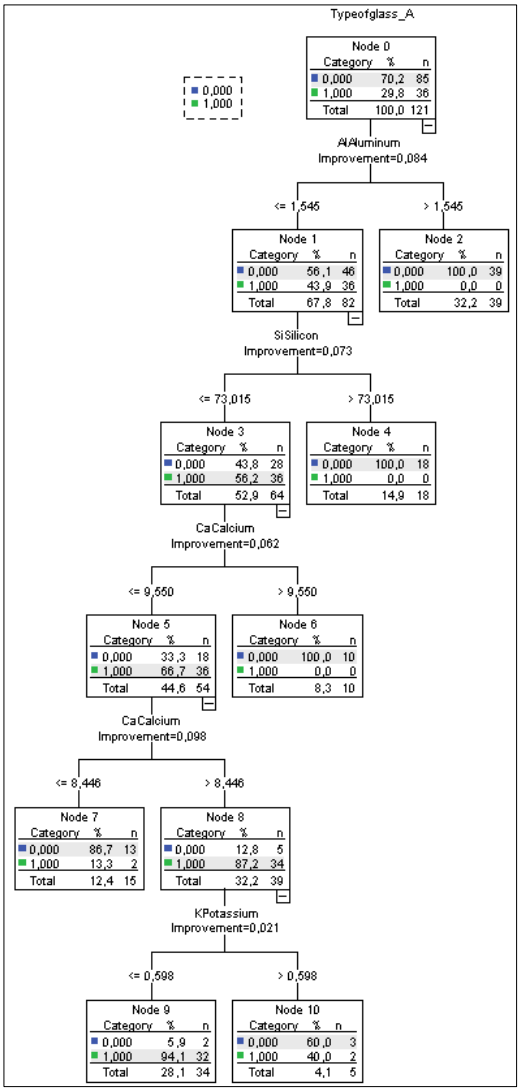

Figura 146 – Albero di regressione per il training set: SMOTE

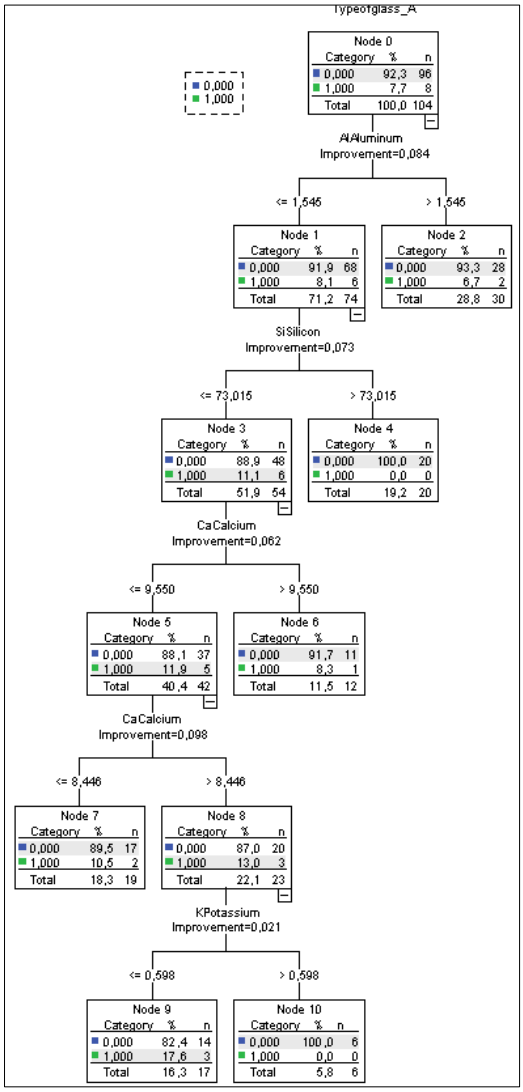

Figura 147 – Albero di regressione per l'holdout set: SMOTE

In Tabella 290 e Tabella 291 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SMOTE.

Tabella 290: Matrice di confusione - CART: SMOTE

Sample		Predicted		
		1	0	Percent Correct
Training	1	32	4	88.9%
	0	2	83	97.6%
	Overall Percentage	28.1%	71.9%	95.0%
Holdout	1	3	5	37.5%
	0	14	82	85.4%
	Overall Percentage	16.3%	83.7%	81.7%

Tabella 291: Misure di performance - CART: SONCA

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.111	0.024	0.976	0.889	0.941	0.914	0.050	0.950
Holdout	0.625	0.146	0.854	0.375	0.176	0.240	0.183	0.817

In Figura 148 e Tabella 292 sono riportate la curva ROC e Area sotto la curva

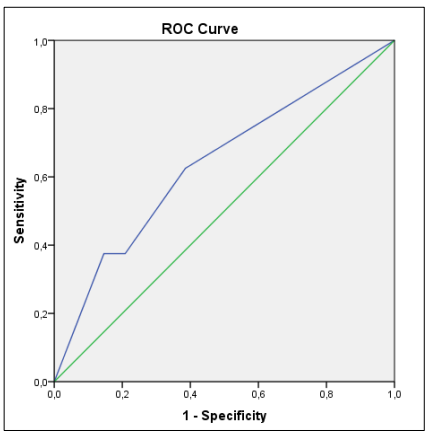

Figura 148 - Curva ROC - CART: SMOTE

Tabella 292: Area sotto la curva ROC - CART: SMOTE

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.639	0.108	0.194	0.428	0.850

Logit con il ricampionamento: SMOTE

In Tabella 293 sono riportate le variabili nell'equazione.

Tabella 293: Variabili nell'equazione: SMOTE

Variables in the Equation	B	S.E.	Wald	df	Sig.	Exp(B)
RIrefractive	-2742.691	691.103	15.750	1	.000	0.000
AlAluminum	-18.674	4.584	16.596	1	.000	0.000
SiSilicon	-9.414	2.443	14.846	1	.000	0.000
BaBarium	6.111	2.541	5.785	1	.016	450.571
FeIron	-34.173	10.085	11.482	1	.001	0.000
Constant	4870.843	1223.743	15.843	1	.000	

In Tabella 294 e Tabella 295 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SMOTE.

Tabella 294: Matrice di confusione - Logit: SMOTE

Sample		Predicted		
		1	0	Percent Correct
Training	1	31	5	86.1%
	0	4	81	95.3%
	Overall Percentage	28.9%	71.1%	92.6%
Holdout	1	4	4	50.0%
	0	28	68	70.8%
	Overall Percentage	26.4%	59.5%	69.2%

Tabella 295: Misure di performance - Logit: SMOTE

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.139	0.047	0.953	0.861	0.886	0.873	0.074	0.926
Holdout	0.500	0.292	0.708	0.500	0.125	0.200	0.308	0.692

In Figura 149 e Tabella 296sono riportate la curva e l'area sotto la curva ROC.

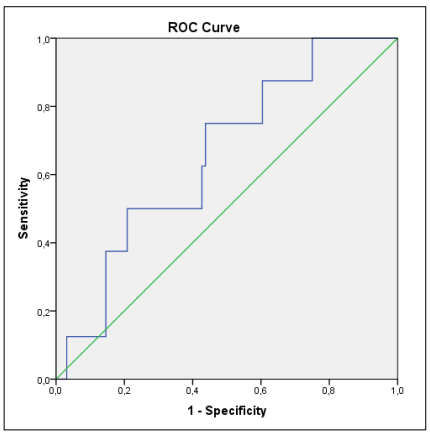

Figura 149 - Curva ROC - Logit: SMOTE

Tabella 296: Area sotto la curva ROC - Logit: SMOTE

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.656	0.089	0.143	0.481	0.832

ROSE

Il set di training è stato ricampionato utilizzando ROSE. In Tabella 297 sono riportate le distribuzioni di frequenza per il set di training.

Tabella 297: Distribuzioni di classe per il training set: ROSE

Variabile	N	%
1	52	47.3
0	58	52.7
Totale	110	100.0

Albero di regressione con il ricampionamento: ROSE

In Figura 150 e Figura 151 sono riportati gli alberi di regressione training avendo implementato ROSE.

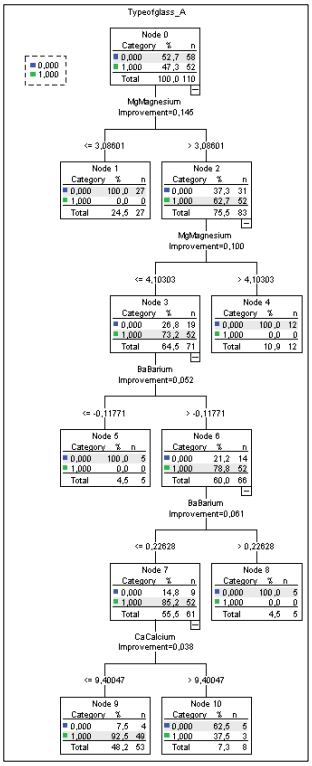

Figura 150 – Albero di regressione per il training set: ROSE

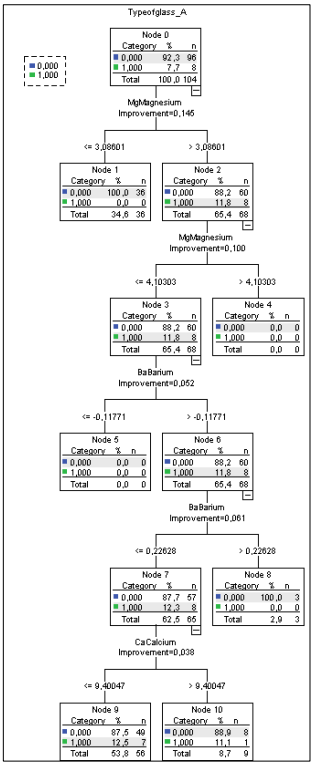

Figura 151 – Albero di regressione per l'holdout set: ROSE

In Tabella 298 e Tabella 299 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato ROSE.

Tabella 298: Matrice di confusione - CART: ROSE

Sample		Predicted		
		1	0	Percent Correct
Training	1	49	3	94.2%
	0	4	54	93.1%
	Overall Percentage	48.2%	51.8%	93.6%
Holdout	1	7	1	87.5%
	0	49	47	49.0%
	Overall Percentage	50.9%	43.6%	51.9%

Tabella 299: Misure di performance - CART: ROSE

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.058	0.069	0.931	0.942	0.925	0.933	0.064	0.936
Holdout	0.125	0.510	0.490	0.875	0.125	0.219	0.481	0.519

In Figura 152e Tabella 300 sono riportate la curva ROC e Area sotto la curva.

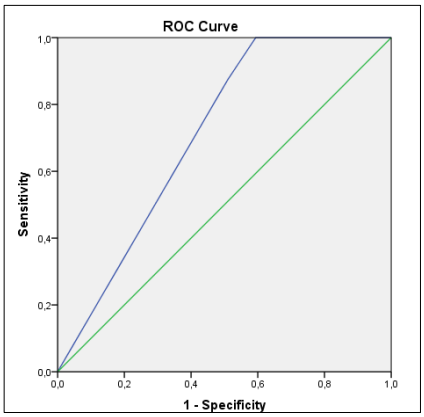

Figura 152 - Curva ROC - CART: ROSE

Tabella 300: Area sotto la curva ROC - CART: ROSE

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.708	0.069	0.052	0.572	0.844

Logit con il ricampionamento: ROSE

In Tabella 301 sono riportate le variabili nell'equazione.

Tabella 301: Variabili nell'equazione: ROSE

Variables in the Equation	B	S.E.	Wald	df	Sig.	Exp(B)
MgMagnesium	0.573	0.198	8.373	1	0.004	1.774
Constant	-1.945	0.689	7.969	1	0.005	0.143

In Tabella 302 e Tabella 303 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato ROSE.

Tabella 302: Matrice di confusione - Logit: ROSE

Sample		Predicted		
		1	0	Percent Correct
Training	1	49	3	94.2%
	0	4	54	93.1%
	Overall Percentage	48.2%	51.8%	93.6%
Holdout	1	7	1	87.5%
	0	49	47	49.0%
	Overall Percentage	50.9%	43.6%	51.9%

Tabella 303: Misure di performance - Logit: ROSE

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.058	0.069	0.931	0.942	0.925	0.933	0.064	0.936
Holdout	0.125	0.510	0.490	0.875	0.125	0.219	0.481	0.519

In Figura 153 e Tabella 304 sono riportate la curva e l'area sotto la curva ROC.

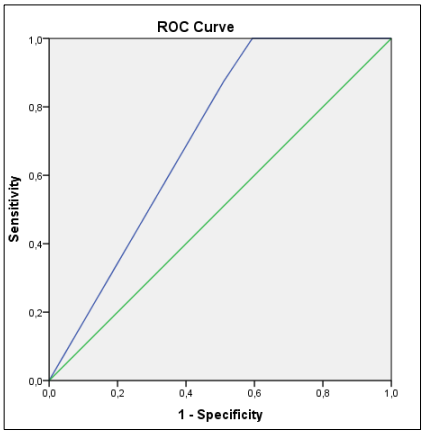

Figura 153 - Curva ROC - Logit: ROSE

Tabella 304: Area sotto la curva ROC - Logit: ROSE

Area	Std. Error[a]	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.708	0.069	0.052	0.572	0.844

A.2.3. Pima Indian Diabetes

Come è stato osservato la variabile di risposta Pima Indian Diabetes non è sbilanciata, la modalità 0 (la classe di maggioranza) ha una frequenza del 65.1%,

mentre la classe 1 (la classe di minoranza) ha una frequenza del 34.9%. Per tale dataset sono stati stimati l'albero di regressione e il logit nei seguenti casi:

- Dataset originale;
- Ricampionamento con SONCA;
- Ricampionamento con SMOTE;
- Ricampionamento con ROSE.

Dataset originale

Albero di regressione senza il ricampionamento

In Figura 154 e Figura 155 sono riportati gli alberi di regressione senza bilanciare la variabile di risposta.

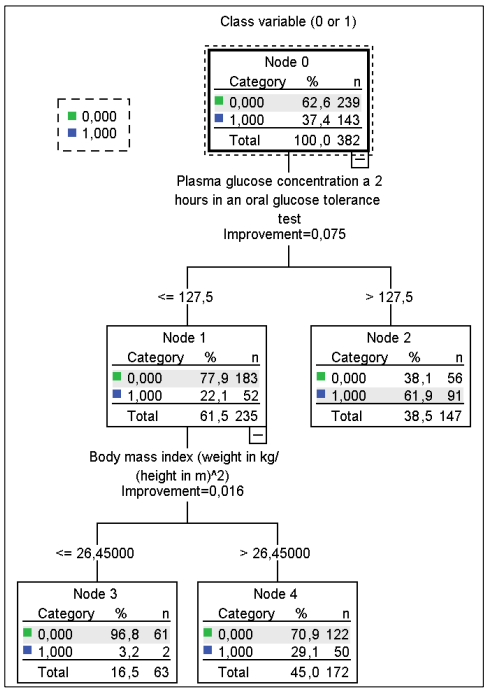

Figura 154 – Albero di regressione per il training set senza ricampionamento

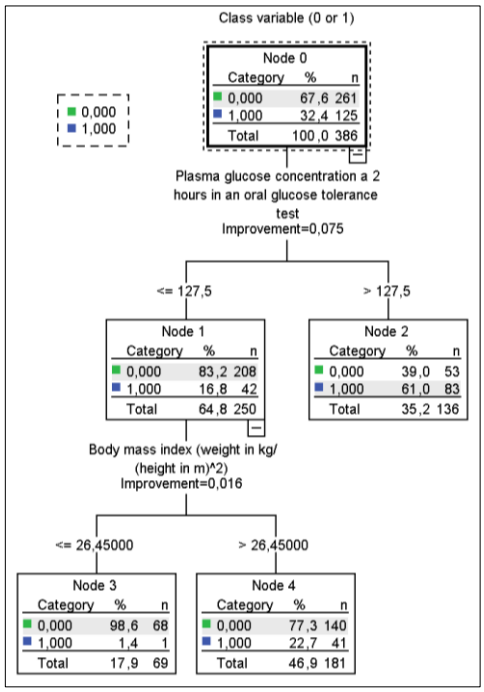

Figura 155 – Albero di regressione per l'holdout set senza ricampionamento

In Tabella 305 e Tabella 306 sono riportate sia la matrice di confusione sia le misure di prestazioni per entrambi i set dati, training e holdout.

Tabella 305: Matrice di confusione senza ricampionamento

Sample		Predicted		
		1	0	Percent Correct
Training	1	91	52	63.6%
	0	56	183	76.6%
	Overall Percentage	38.5%	61.5%	71.7%
Test	1	83	42	66.4%
	0	53	208	79.7%
	Overall Percentage	35.2%	64.8%	75.4%

Tabella 306: Misure di performance senza ricampionamento

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.364	0.234	0.766	0.636	0.619	0.628	0.283	0.717
Test	0.336	0.203	0.797	0.664	0.610	0.636	0.246	0.754

In Figura 156 e in Tabella 307 sono riportate la curva ROC e l'area sottesa alla curva ROC.

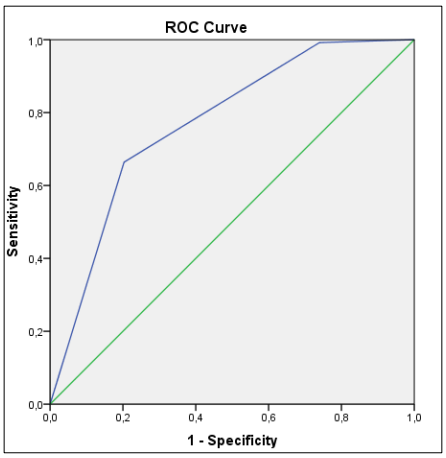

Figura 156 - Curva ROC per l'holdout set senza ricampionamento

Tabella 307: Area sotto la curva ROC per l'holdout set senza ricampionamento

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.771	0.025	0.000	0.723	0.819

Logit senza il ricampionamento

In Tabella 308 sono riportate le variabili nell'equazione.

Tabella 308: Variabili nell'equazione per il training set senza ricampionamento

	B	S.E.	Wald	df	Sig.	Exp(B)
Numberoftimespregnant	0.147	0.039	14.425	1	0.000	1.158
Plasmaglucoseconcentration	0.031	0.005	47.987	1	0.000	1.032
Bodymassindex	0.069	0.018	14.726	1	0.000	1.072
Diabetespedigreefunction	1.187	0.398	8.881	1	0.003	3.278
Constant	-7.873	0.875	80.956	1	0.000	0.000

In Tabella 309 e Tabella 310 sono riportate la matrice di confusione che le misure di prestazioni sia per il dataset di training che di holdout.

Tabella 309: Matrice di confusione senza ricampionamento

Sample		Predicted		
		1	0	Percent Correct
Training	1	72	53	57.6%
	0	32	229	87.7%
	Overall Percentage	26.9%	73.1%	78.0%
Test	1	82	61	57.3%
	0	35	204	85.4%
	Overall Percentage	30.6%	69.4%	74.9%

Tabella 310: Misure di performance senza ricampionamento

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.424	0.123	0.877	0.576	0.692	0.629	0.220	0.780
Test	0.427	0.146	0.854	0.573	0.701	0.631	0.251	0.749

In Figura 157 e in Tabella 311 sono riportate la curva ROC e l'area sottesa alla curva ROC.

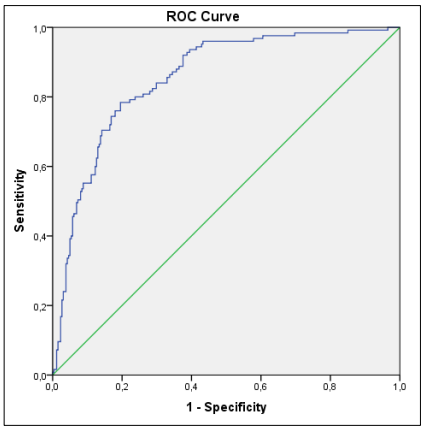

Figura 157 - Curva ROC per l'holdout set senza ricampionamento

Tabella 311: Area sotto la curva ROC per l'holdout set senza ricampionamento

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.854	0.020	0.000	0.815	0.894

SONCA con distribuzione di probabilità triangolare

Il set di training è stato ricampionato utilizzando SONCA, con una distribuzione delle probabilità di tipo triangolare. In Tabella 312 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA.

Tabella 312: Distribuzioni di classe per il training set: SONCA

Variabile	N	%
0	588	49.58
1	598	50.42
Totale	1186	100.00

Albero di regressione con il ricampionamento: SONCA

In Figura 158 e Figura 159 sono riportati gli alberi di regressione training avendo implementato SONCA.

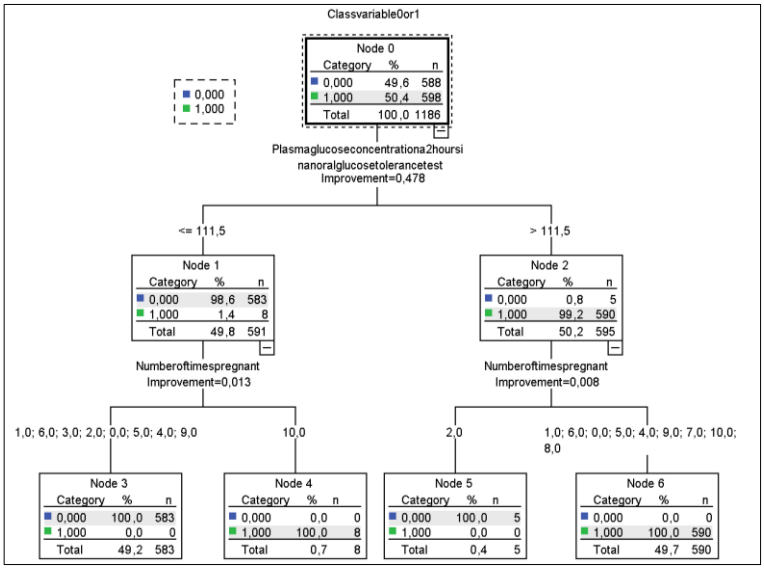

Figura 158 – Albero di regressione per il training set: SONCA

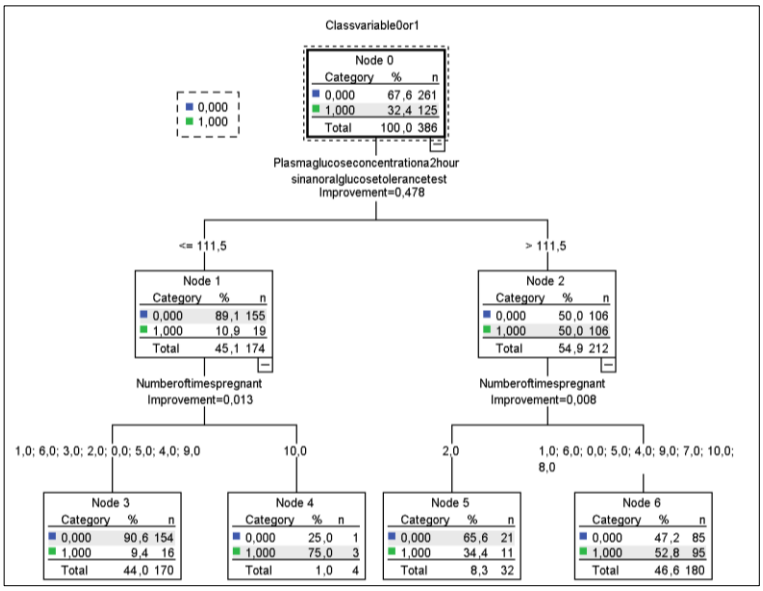

Figura 159 – Albero di regressione per l'holdout set: SONCA

In Tabella 313 e Tabella 314 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA.

Tabella 313: Matrice di confusione - CART: SONCA

Sample		Predicted		
		1	0	Percent Correct
Training	1	598	0	100.0%
	0	0	588	100.0%
	Overall Percentage	50.4%	49.6%	100.0%
Holdout	1	98	27	78.4%
	0	86	175	67.0%
	Overall Percentage	47.7%	52.3%	70.7%

Tabella 314: Misure di performance - CART: SONCA

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.000	0.000	1.000	1.000	1.000	1.000	0.000	1.000
Holdout	0.216	0.330	0.670	0.784	0.533	0.634	0.293	0.707

In Figura 160 e Tabella 315 sono riportate la curva ROC e Area sotto la curva.

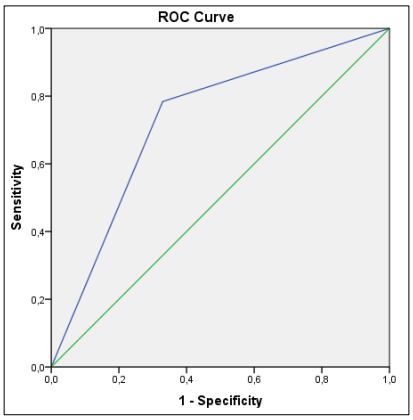

Figura 160 - Curva ROC - CART: SONCA

Tabella 315: Area sotto la curva ROC - CART: SONCA

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.727	0.027	0.000	0.674	0.781

Logit con il ricampionamento: SONCA

In Tabella 316 sono riportate le variabili nell'equazione.

Tabella 316: Variabili nell'equazione: SONCA

	B	S.E.	Wald	df	Sig.	Exp(B)
Numberoftimespregnant	6.067	168.963	0.001	1	0.971	431.473
Plasmaglucoseconcentration	5.724	75.915	0.006	1	0.940	306.026
Constant	-679.065	8782.052	0.006	1	0.938	0.000

In Tabella 317 e Tabella 318 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA.

Tabella 317: Matrice di confusione - Logit: SONCA

Sample		Predicted		
		1	0	Percent Correct
Training	1	598	0	100.0%
	0	0	588	100.0%
	Overall Percentage	50.4%	49.6%	100.0%
Holdout	1	102	23	81.6%
	0	97	164	62.8%
	Overall Percentage	51.6%	48.4%	68.9%

Tabella 318: Misure di performance - Logit: SONCA

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.000	0.000	1.000	1.000	1.000	1.000	0.000	1.000
Holdout	0.184	0.372	0.628	0.816	0.513	0.630	0.311	0.689

In Figura 161 e Tabella 319 sono riportate la curva e l'area sotto la curva ROC.

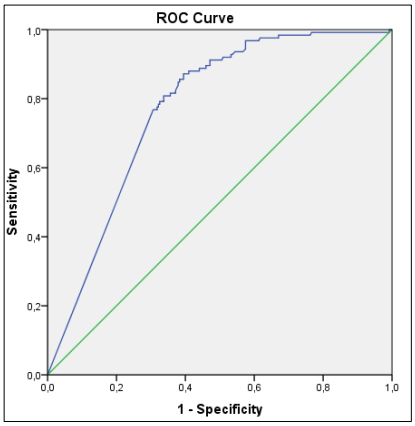

Figura 161 - Curva ROC - Logit: SONCA

Tabella 319: Area sotto la curva ROC - Logit: SONCA

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.772	0.024	0.000	0.725	0.818

SONCA con distribuzione di probabilità gaussiana

Il set di training è stato ricampionato utilizzando SONCA. In Tabella 320 sono riportate le distribuzioni di frequenza per il set di training avendo implementato SONCA.

Tabella 320: Distribuzioni di classe per il training set: SONCA

Variabile	N	%
0	615	49.71706
1	622	50.28294
Totale	1237	100

Albero di regressione con il ricampionamento: SONCA

In Figura 162 e Figura 163 sono riportati gli alberi di regressione training avendo implementato SONCA.

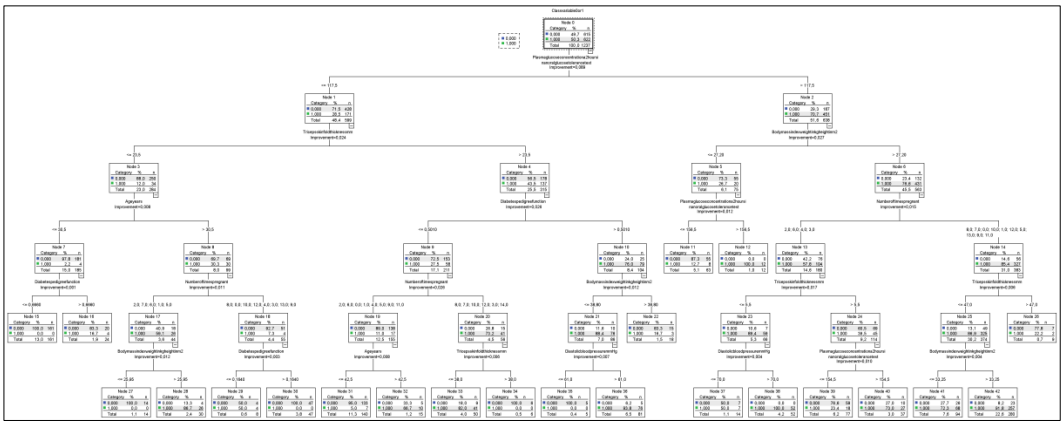

Figura 162 – Albero di regressione per il training set: SONCA

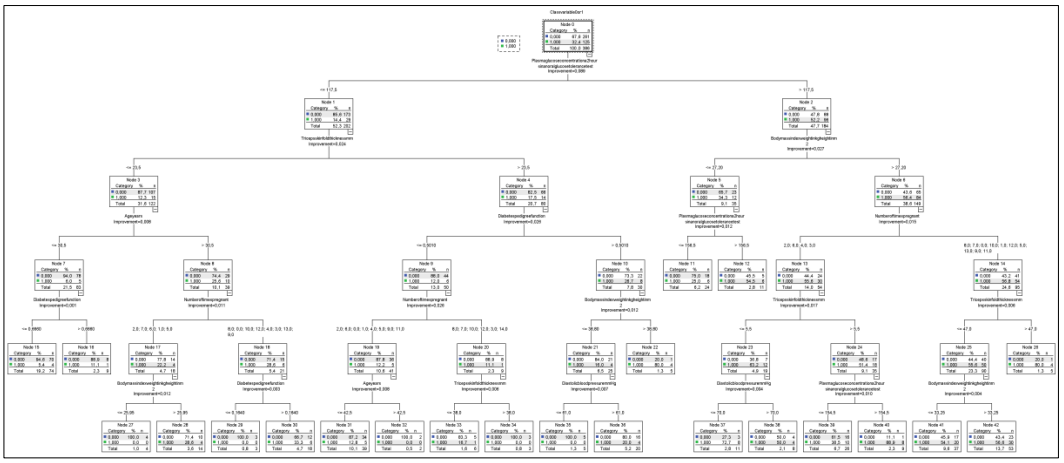

Figura 163 – Albero di regressione per l'holdout set: SONCA

In Tabella 321 e

Tabella 322 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA.

Tabella 321: Matrice di confusione - Logit: SONCA

Sample		Predicted		
		1	0	Percent Correct
Training	1	569	53	91.5%
	0	82	533	86.7%
	Overall Percentage	52.6%	47.4%	89.1%
Test	1	77	48	61.6%
	0	83	178	68.2%
	Overall Percentage	41.5%	58.5%	66.1%

Tabella 322: Misure di performance - CART: SONCA

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.085	0.133	0.867	0.915	0.874	0.894	0.109	0.891
Test	0.384	0.318	0.682	0.616	0.481	0.540	0.339	0.661

In Figura 164 e Tabella 323 sono riportate la curva ROC e Area sotto la curva.

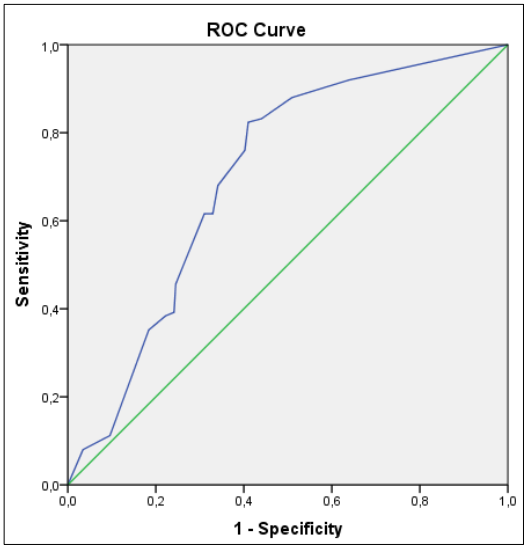

Figura 164 - Curva ROC - CART: SONCA

Tabella 323: Area sotto la curva ROC - CART: SONCA

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.703	0.027	0.000	0.650	0.755

Logit con il ricampionamento: SONCA

In Tabella 324 sono riportate le variabili nell'equazione.

Tabella 324: Variabili nell'equazione: SONCA

	B	S.E.	Wald	df	Sig.	Exp(B)
Numberoftimespregnant	0.130	0.021	38.389	1	0.000	1.139
Plasmaglucoseconcentration	0.038	0.003	170.652	1	0.000	1.039
2HourseruminsulinmuUml	-0.002	.001	5.232	1	0.022	0.998
Bodymassindexweightinkgheightinm2	0.069	0.010	49.987	1	0.000	1.071
Diabetespedigreefunction	1.555	0.226	47.373	1	0.000	4.737
Constant	-7.974	0.502	252.386	1	0.000	0.000

In Tabella 325 e Tabella 326 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SONCA.

Tabella 325: Matrice di confusione - Logit: SONCA

Sample		Predicted		
		1	0	Percent Correct
Training	1	445	177	71.5%
	0	145	470	76.4%
	Overall Percentage	47.7%	52.3%	74.0%
Holdout	1	98	27	78.4%
	0	65	196	75.1%
	Overall Percentage	42.2%	57.8%	76.2%

Tabella 326: Misure di performance - Logit: SONCA

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.285	0.236	0.764	0.715	0.754	0.734	0.260	0.740
Test	0.216	0.249	0.751	0.784	0.601	0.681	0.238	0.762

In Figura 165 e Tabella 327 sono riportate la curva e l'area sotto la curva ROC.

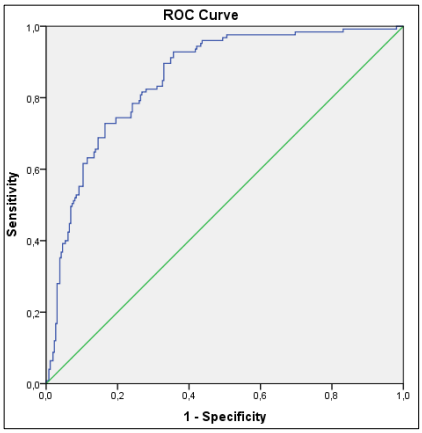

Figura 165 - Curva ROC - Logit: SONCA

Tabella 327: Area sotto la curva ROC - Logit: SONCA

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.855	0.020	0.000	0.816	0.894

SMOTE

Il set di training è stato ricampionato utilizzando SMOTE. In Tabella 328 sono riportate le distribuzioni di frequenza per il set di training.

Tabella 328: Distribuzioni di classe per il training set: SMOTE

Cover_Type	N	F=N/N$_{tot}$[%]
0	124	17.8
1	571	82.2
Totale	**695**	100.0

Albero di regressione con il ricampionamento: SMOTE

In Figura 166 e Figura 167 sono riportati gli alberi di regressione training avendo implementato SMOTE.

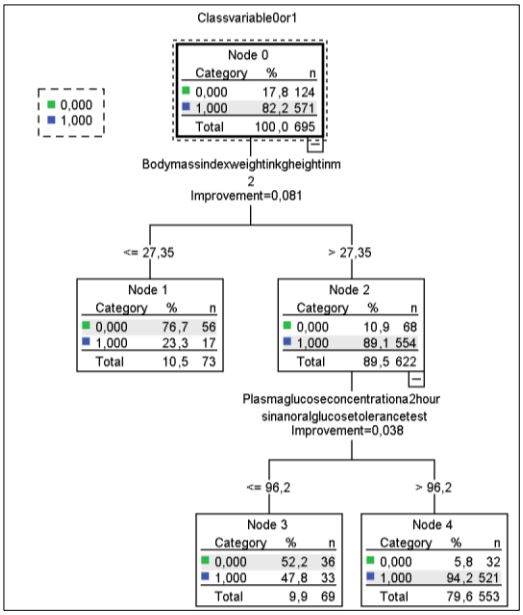

Figura 166 – Albero di regressione per il training set: SMOTE

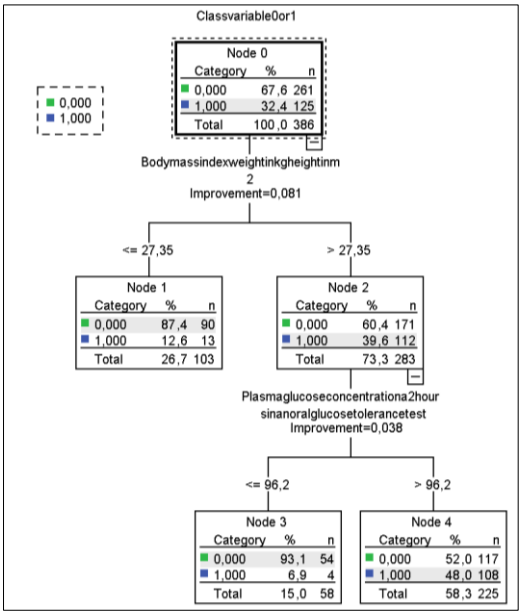

Figura 167 – Albero di regressione per l'holdout set: SMOTE

In Tabella 329 e Tabella 330 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SMOTE.

Tabella 329: Matrice di confusione - CART: SMOTE

Sample		Predicted		
		1	0	Percent Correct
Training	1	521	50	91.2%
	0	32	92	74.2%
	Overall Percentage	79.6%	20.4%	88.2%
Holdout	1	108	17	86.4%
	0	117	144	55.2%
	Overall Percentage	58.3%	41.7%	65.3%

Tabella 330: Misure di performance - CART: SONCA

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.088	0.258	0.742	0.912	0.942	0.927	0.118	0.882
Holdout	0.136	0.448	0.552	0.864	0.480	0.617	0.347	0.653

In Figura 168 e Tabella 331 sono riportate la curva ROC e Area sotto la curva.

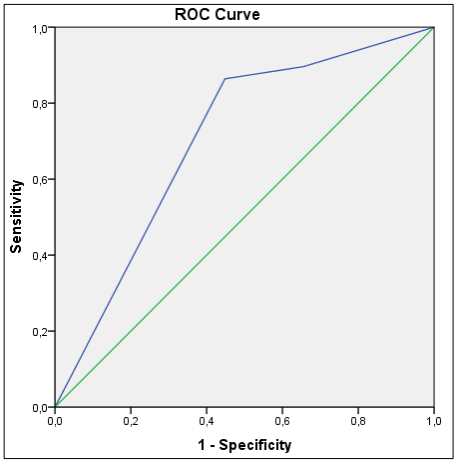

Figura 168 - Curva ROC - CART: SMOTE

Tabella 331: Area sotto la curva ROC - CART: SMOTE

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.703	0.027	0.000	0.649	0.756

Logit con il ricampionamento: SMOTE

In Tabella 332 sono riportate le variabili nell'equazione.

Tabella 332: Variabili nell'equazione: SMOTE

	B	S.E.	Wald	df	Sig.	Exp(B)
Numberoftimespregnant	0.259	0.049	27.705	1	0.000	1.296
Plasmaglucoseconcentration	0.052	0.006	66.256	1	0.000	1.053
Bodymassindexweightinkgheightinm2	0.132	0.022	35.900	1	0.000	1.141
Diabetespedigreefunction	2.190	0.510	18.404	1	0.000	8.931
Constant	-10.851	1.123	93.359	1	0.000	0.000

In Tabella 333 e Tabella 334 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato SMOTE.

Tabella 333: Matrice di confusione - Logit: SMOTE

Sample		Predicted		
		1	0	Percent Correct
Training	1	546	25	95.6%
	0	58	66	53.2%
	Overall Percentage	86.9%	13.1%	88.1%
Holdout	1	121	4	96.8%
	0	151	110	42.1%
	Overall Percentage	70.5%	29.5%	59.8%

Tabella 334: Misure di performance - Logit: SMOTE

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.044	0.468	0.532	0.956	0.904	0.929	0.119	0.881
Holdout	0.032	0.579	0.421	0.968	0.445	0.610	0.402	0.598

In Figura 169 e Tabella 335 sono riportate la curva e l'area sotto la curva ROC.

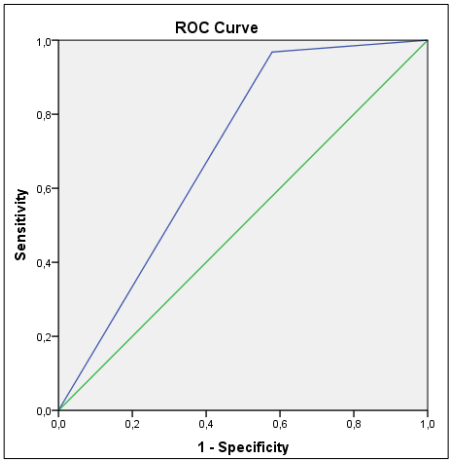

Figura 169 - Curva ROC - Logit: SMOTE

Tabella 335: Area sotto la curva ROC - Logit: SMOTE

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.695	0.026	0.000	0.643	0.746

ROSE

Il set di training è stato ricampionato utilizzando ROSE. In Tabella 336 sono riportate le distribuzioni di frequenza per il set di training.

Tabella 336: Distribuzioni di classe per il training set: ROSE

Variabile	N	%
0	200	52.3
1	182	48.7
Totale	**382**	100.0

Albero di regressione con il ricampionamento: ROSE

In Figura 170 e Figura 171 sono riportati gli alberi di regressione training avendo implementato ROSE.

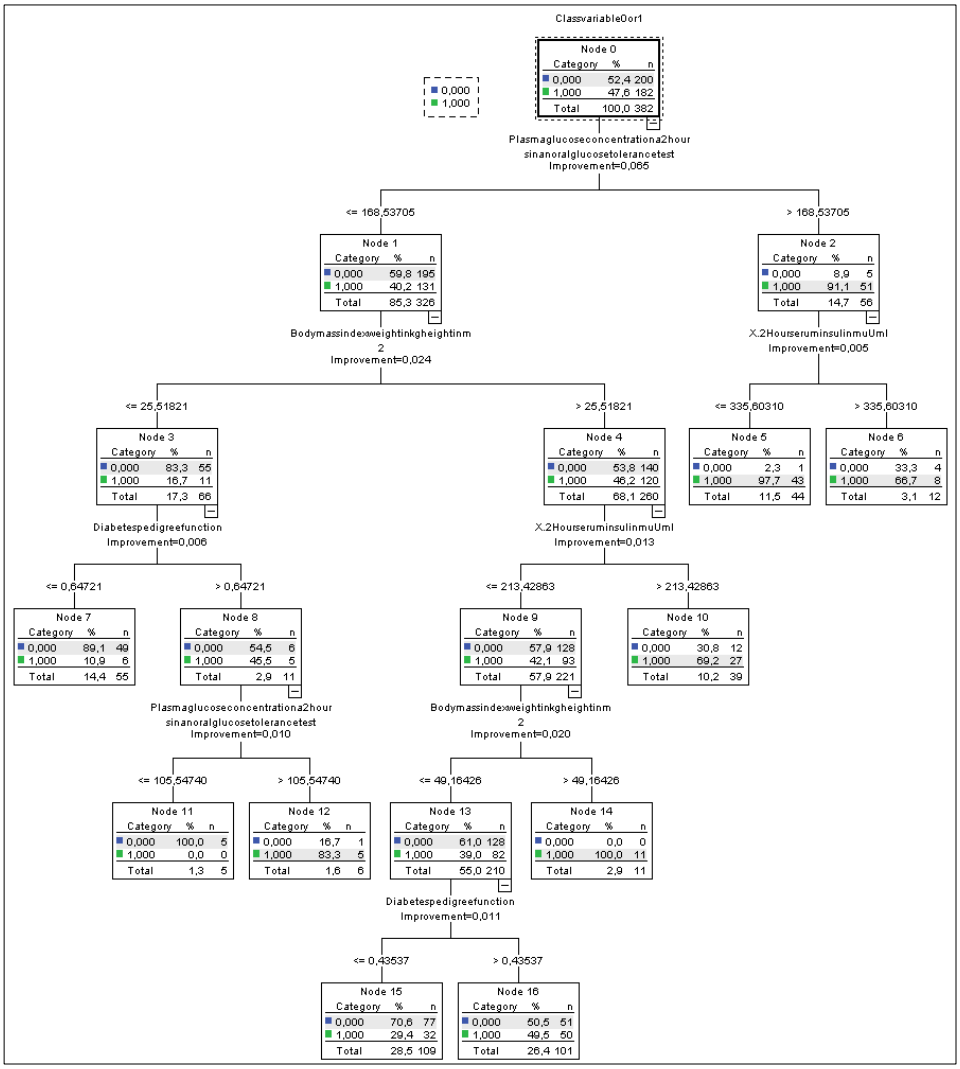

Figura 170 – Albero di regressione per il training set: ROSE

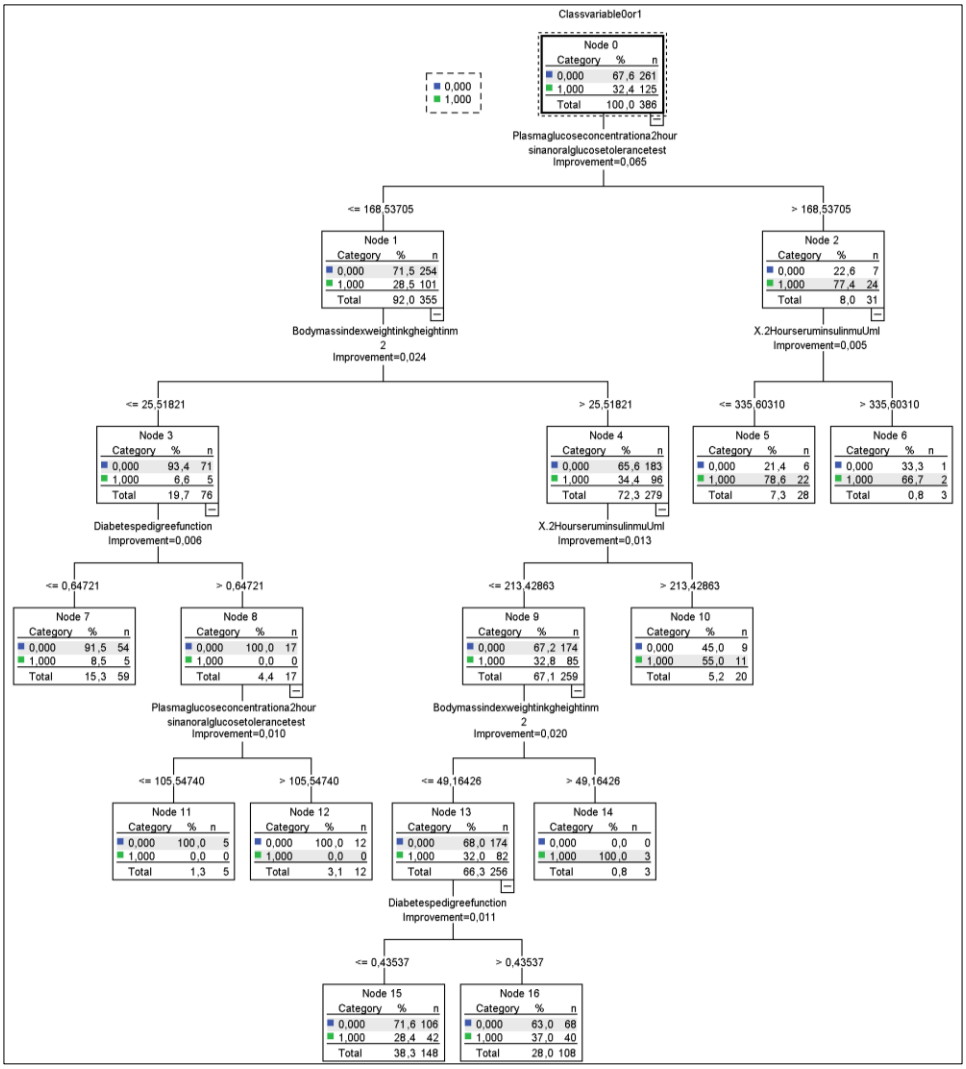

Figura 171 – Albero di regressione per l'holdout set: ROSE

In Tabella 337 e

Tabella 338 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato ROSE.

Tabella 337: Matrice di confusione - CART: ROSE

Sample		Predicted		
		1	0	Percent Correct
Training	1	94	88	51.6%
	0	18	182	91.0%
	Overall Percentage	29.3%	70.7%	72.3%
Holdout	1	38	87	30.4%
	0	28	233	89.3%
	Overall Percentage	17.1%	82.9%	70.2%

Tabella 338: Misure di performance - CART: ROSE

	FN	FP	TN	TP	Precision	F-measure	Err	Acc
Training	0.484	0.090	0.910	0.516	0.839	0.639	0.277	0.723
Test	0.696	0.107	0.893	0.304	0.576	0.398	0.298	0.702

In Figura 172 e Tabella 339 sono riportate la curva ROC e Area sotto la curva.

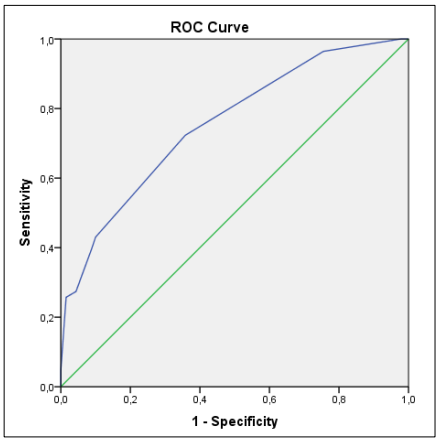

Figura 172 - Curva ROC - CART: SONCA

Tabella 339: Area sotto la curva ROC - CART: SONCA

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.754	0.018	0.000	0.720	0.789

Logit con il ricampionamento: ROSE

In Tabella 340 sono riportate le variabili nell'equazione.

Tabella 340: Variabili nell'equazione: ROSE

	B	S.E.	Wald	df	Sig.	Exp(B)
Plasmaglucoseconcentration	0.019	0.003	31.433	1	0.000	1.019
Bodymassindexweightinkgheightinm2	0.043	0.013	10.667	1	0.001	1.044
Diabetespedigreefunction	0.710	0.316	5.046	1	0.025	2.033
Constant	-4.159	0.591	49.558	1	0.000	0.016

In Tabella 341 e Tabella 342 sono riportate la matrice di confusione e le misure di prestazioni sia per il training set sia per l'holdout set, avendo implementato ROSE.

Tabella 341: Matrice di confusione - Logit: ROSE

Sample		Predicted		
		1	0	Percent Correct
Training	1	106	76	58.2%
	0	50	150	75.0%
	Overall Percentage	40.8%	59.2%	67.0%
Holdout	1	83	42	66.4%
	0	44	217	83.1%
	Overall Percentage	32.9%	67.1%	77.7%

Tabella 342: Misure di performance - LOGIT: ROSE

	FN_{rate}	FP_{rate}	TN_{rate}	TP_{rate}	Precision	F-measure	Err	Acc
Training	0.418	0.250	0.750	0.582	0.679	0.627	0.330	0.670
Holdout	0.336	0.169	0.831	0.664	0.654	0.659	0.223	0.777

In Figura 173 e Tabella 343 sono riportate la curva ROC e Area sotto la curva.

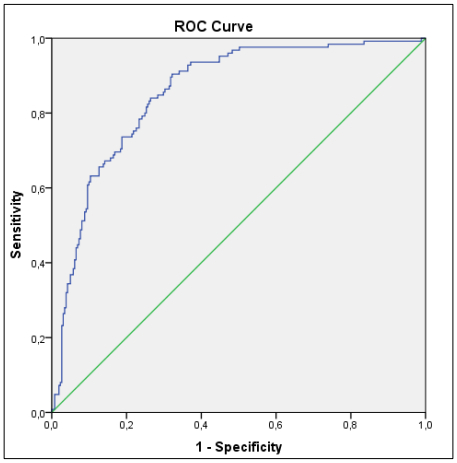

Figura 173 - Curva ROC - LOGIT: ROSE

Tabella 343: Area sotto la curva ROC - LOGIT: ROSE

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.856	0.020	0.000	0.817	0.895