# Relazione di Statistica Multivariata: Analisi predittiva sullo stato operativo delle imprese polacche

*Gruppo composto da: De Vecchi Federica, Pelagalli Eva, Piovesana Dario, Pisaniello Sara*

## 1. Introduzione

La previsione riguardante lo stato operativo delle imprese è argomento di grande impatto e attualità, soprattutto quando interessa i **mercati emergenti**, detti anche NIC (Nazioni di recente industrializzazione).
Sono comprese, con tale dicotomia, tutte le economie non ancora pienamente sviluppate in possesso, però, di un grande potenziale di crescita a fronte di investimenti il cui rischio è molto elevato.
Lo studio compiuto qui a seguito è finalizzato allo sviluppo di modelli predittivi che consentano di classificare accuratamente se delle aziende polacche incorreranno o meno in uno stato di bancarotta, utilizzando informazioni di tipo economico-finanziario.
Innanzitutto sono state effettuate verifiche preliminari, da cui sono emerse problematiche di vario genere: valori mancanti, collinearità tra le variabili esplicative, presenza di outliers, variabili esplicative non in grado di discriminare condizionatamente alla variabile risposta.
Per cui sono state applicate tecniche di "pulizia del dataset", frequenti in analisi multivariata, inducendo una notevole riduzione delle dimensioni della matrice di dati.
Gli strumenti utilizzati per la previsione del fallimento/non fallimento delle aziende sono: analisi di discriminante lineare (LDA), e alberi di classificazione (ADC).
Per ciascuna previsione, inoltre, sono state sviluppate tecniche di bilanciamento del campione ed è stato sfruttato l'algoritmo *bagging*, al fine di sviluppare un modello più adeguato.
Il modello migliore a cui si è pervenuti è l'approccio *bagging* per creare alberi di classificazione, utilizzando un campione bilanciato ottenuto da una combinazione di over e under-sampling randomizzati.

## 2. Descrizione del dataset e presentazione delle variabili

I dati a disposizione sono stati raccolti da EMIS (*Emerging Markets Information Service)*, un database contenente informazioni relative ai mercati emergenti nel mondo.
Il nostro dataset, in particolare, presenta dati di aziende polacche trattati dal Dottor Sebastian Tomczak, (del Dipartimento di Ricerche Operative dell'Università delle Scienze e della Tecnologia di Wroclaw, Polonia) indagati nel periodo 2000-2012 per aziende in bancarotta, e dal 2007 al 2013 per le aziende operative.
Verrà considerata solo una porzione del dataset originario, che era di natura intertemporale, relativa ad aziende tre anni prima di un possibile fallimento/non fallimento.
L'accuratezza delle nostre indagini, rispetto allo studio di Tomczak[1], potrebbe risultare intaccata dalla mancanza di dimensione intertemporale. Di fatto, è buona pratica economica osservare uno spettro evolutivo aziendale superiore all'anno per valutare il rischio di fallimento effettivo[2].
Il dataset "bankruptcy.txt" contiene 10503 osservazioni, relative al terzo anno di previsione, e 65 variabili. Le prime 64 variabili rappresentano le esplicative, la 65-esima variabile, invece, rappresenta la risposta, con valori 0 e 1 che riflettono se l'azienda sarà in bancarotta o meno dopo tre anni.
Le variabili incluse sono le seguenti:

| Variabili | Tipo | Descrizione della variabile | Supporto |
|---|---|---|---|
| X1 | *Continua* | Profitti Netti/Totale Attivo | ***[-17.69,52.65]*** |
| X2 | *Continua* | Totale Passivo/Totale Attivo | ***[0, 480.73]*** |
| X3 | *Continua* | Capitale Circolante/Totale Attivo | ***[-479.73,17.08]*** |
| X4 | *Continua* | Attivi Correnti/Debiti a Breve Termine | ***[0, 53433]*** |
| X5 | *Continua* | [(Cassa +Titoli a Breve Termine +Crediti-Debiti a Breve Termine) / (Spese Operative-Ammortamenti)]*365 | ***[-11903000,685440]*** |
| X6 | *Continua* | Utili Non Distribuiti/Totale Attivo | ***[-508.12, 45.533]*** |
| X7 | *Continua* | EBIT (Profitti Lordi) /Totale Attivo | ***[-17.69,52.65]*** |

[1] Tomczak, S., 2014b. Comparative analysis of the bankrupt companies of the sector of animal slaughtering and processing. Equilibrium. Quarterly Journal of Economics and Economic Policy 9, 59–86

[2] Claudio Teodori, Analisi di bilancio lettura ed interpretazione.G. Giampichelli Editore, terza edizione, 2017

| | | | |
|---|---|---|---|
| X8 | *Continua* | Patrimonio Netto/Totale Passivo | ***[-2.08,53432]*** |
| X9 | *Continua* | Ricavi da Vendite e Prestazioni/Totale attivo | ***[-1.21, 740.44]*** |
| X10 | *Continua* | Capitale Proprio/Totale Attivo | ***[-479.73, 11.83]*** |
| X11 | *Continua* | (Marginale Operativo Lordo + Sopravvivenze + Oneri Finanziari) /Totale Attivo | ***[-17.69, 52.65]*** |
| X12 | *Continua* | Marginale Operativo Lordo/Debiti a Breve Termine | ***[-1543.8, 8259.4]*** |
| X13 | *Continua* | (Marginale Operativo Lordo + Ammortamenti) /Ricavi da Vendite e Prestazioni | ***[-631.71, 4972]*** |
| X14 | *Continua* | (Marginale Operativo Lordo+ Interessi) / Totale Attivo | ***[-17.69, 52.65]*** |
| X15 | *Continua* | (Totale Passivo*365) /(Marginale Operativo Lordo + Ammortamenti) | ***[-2321800, 10236000]*** |
| X16 | *Continua* | (Marginale Operativo Lordo + Ammortamenti) /Totale Passivo | ***[-204.3, 8259.4]*** |
| X17 | *Continua* | Totale Attivo/Totale Passivo | ***[-0.04, 53433]*** |
| X18 | *Continua* | Marginale Operativo Lordo/Totale Attivo | ***[-17.69,53.68]*** |
| X19 | *Continua* | Marginale Operativo Lordo/Ricavi da Vendite e Prestazioni | ***[-771.65, 123.94]*** |
| X20 | *Continua* | Rimanenze*365)/Ricavi da Vendite e Prestazioni | ***[0, 91600]*** |
| X21 | *Continua* | Ricavi da Vendite e Prestazioni(n)/Ricavi da Vendite e Prestazioni (n-1) | ***[-1.10, 29907]*** |
| X22 | *Continua* | Valore della Produzione/Totale Attivo | ***[-17.69, 47.59]*** |
| X23 | *Continua* | Profitto Netto(EBITDA)/ Ricavi da Vendite e Prestazioni | ***[-771.65, 123.94]*** |
| X24 | *Continua* | Profitto Lordo(in 3 anni)/Totale attivo | ***[ -9.3392 179.9200]*** |
| X25 | *Continua* | (Capitale Proprio-Capitale Sociale)/Totale Attivo | ***[-500.75, 8.83]*** |
| X26 | *Continua* | (Profitto Netto(EBITDA)+ Ammortamenti)/Totale Passivo | ***[-204.3, 8262.3]*** |
| X27 | *Continua* | Valore della Produzione/Oneri Finanziari | ***[-190130, 2723000]*** |
| X28 | *Continua* | Capitale Circolante/Immobilizzazioni | ***[-690.4, 6233.3]*** |
| X29 | *Continua* | Log(Totale Attivo) | ***[-0.35, 9.61]*** |
| X30 | *Continua* | (Totale Passivo-Cassa)/Ricavi delle Vendite e Prestazioni | ***[-6351.7, 2940.5]*** |
| X31 | *Continua* | (Marginale Operativo Lordo +interessi)/Ricavi delle Vendite e Prestazioni | ***[-771.39, 60.43]*** |
| X32 | *Continua* | Debiti Correnti*365/Costi per Materie Prime, Sussidiarie di Consumo e Merci | ***[-9295.6 ,6674200]*** |
| X33 | *Continua* | Costo della Produzione/Debiti a Breve Termine | ***[-1.92,2787.9]*** |
| X34 | *Continua* | Costo della Produzione/Totale Passivo | ***[-1696,6348.5]*** |
| X35 | *Continua* | Profitti Da Vendite/Totale Attivo | ***[-17.07, 47.59]*** |
| X36 | *Continua* | Totale Vendite/Totale Attivo | ***[-8.4465e-05, 169.5]*** |
| X37 | *Continua* | (Attività Correnti- Rimanenze)/Debiti di Lungo Periodo | ***[-2.20, 136090]*** |
| X38 | *Continua* | Capitale Costante/Totale Attivo | ***[-479.73,13.65]*** |
| X39 | *Continua* | (Profitti da Vendite) /Ricavi da Vendite e Prestazioni | ***[-551.11, 293.15]*** |
| X40 | *Continua* | (Attività Correnti-Rimanenze-Crediti) /(Debiti a Breve Termine) | ***[-7.08, 2883]*** |
| X41 | *Continua* | Totale Passivo/((Profitti in attività operative + Ammortamenti)*(12/365)) | ***[-667.73, 288770]*** |
| X42 | *Continua* | Profitti in Attività Operative/Ricavi delle Vendite e Prestazioni | ***[-765.8, 165.95]*** |
| X43 | *Continua* | (Crediti di rotazione + Rimanenze Espresse in Giorni) | ***[-25113, 54030]*** |
| X44 | *Continua* | (Crediti*365) /Ricavi dalle Vendite e Prestazioni | ***[-25113, 254030]*** |
| X45 | *Continua* | Profitto Netto/Rimanenze | ***[-74385, 113280]*** |
| X46 | *Continua* | (Attivo Corrente – Rimanenze )/Debiti a Breve Termine | ***[-6.46, 53433]*** |
| X47 | *Continua* | (Rimanenze*365) /Costo di Materie Prime, sussidiarie, di consumo e merci | ***[-17.30, 2591100]*** |
| X48 | *Continua* | (Profitto Netto(EBITDA) - Ammortamenti) /Totale Attivo | ***[-17.69, 47.59]*** |
| X49 | *Continua* | (Profitto Netto(EBITDA)- Ammortamenti) /Ricavi da Vendite e Prestazioni | ***[-905.75, 178.89]*** |
| X50 | *Continua* | Attivo Corrente/Totale Debiti | ***[0, 53433]*** |
| X51 | *Continua* | Debiti a Breve Termine/Totale Attivo | ***[0,480.73,]*** |
| X52 | *Continua* | (Debiti a Breve Termine*365)/Costo di Materie Prime, sussidiarie, di consumo e Merci | ***[-25.46, 84827]*** |
| X53 | *Continua* | Capitale Proprio/Immobilizzazioni | ***[-869.04, 6234.3]*** |
| X54 | *Continua* | Capitale Costante/Immobilizzazioni | ***[-706.49, 6234.3]*** |
| X55 | *Continua* | Capitale Circolante | ***[-751380, 3380500]*** |
| X56 | *Continua* | (Ricavi di Vendite e Prestazioni - Costo Materie Prime, sussidiarie, di Consumo e Merci) /Ricavi di Vendite e Prestazioni | ***[-5691.7, 293.15]*** |
| X57 | *Continua* | (Attivo Corrente- Rimanenze - Debiti a Breve Termine)/ (Ricavi di Vendite e Prestazioni-Marginale Operativo Lordo-Immobilizzazioni) | ***[-1667.3, 552.64]*** |
| X58 | *Continua* | Costi della Produzione/Valore della Produzione | ***[-198.69, 18118]*** |
| X59 | *Continua* | Debiti a Lungo Termine/Capitale Proprio | ***[-172.07, 7617.3]*** |
| X60 | *Continua* | Ricavi da Vendite e Prestazioni/ Rimanenze | ***[0, 3660200]*** |
| X61 | *Continua* | Ricavi da Vendite e Prestazioni/Crediti | ***[-6.59, 4470.4]*** |
| X62 | *Continua* | Debiti a Breve Termine*365/Ricavi di Vendite e Prestazioni | ***[-2336500, 1073500]*** |
| X63 | *Continua* | Ricavi da Vendite e Prestazioni/Debiti a Breve Termine | ***[0, 1974.5]*** |
| X64 | *Continua* | Ricavi da Vendite e Prestazioni/Immobilizzazioni | ***[0, 21499]*** |
| X65 | *Binaria* | Variabile Risposta (Bancarotta/Non bancarotta) | ***{0,1}*** |

**Tabella 1**: Descrizione delle variabili

# 3. Analisi descrittive ed esplorative

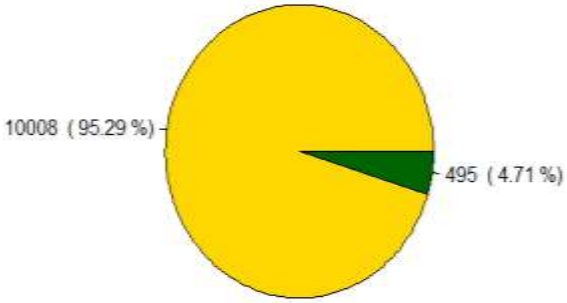
## 3.1 Bilanciamento della variabile risposta

Osservando la variabile risposta binaria X65 lo squilibrio tra classi è evidente: 495 delle 10503 aziende polacche sono state dichiarate in bancarotta (il 4.71% dell'intero campione), mentre 10008 su 10503 sono state dichiarate operative (il 95.29%).
Questo aspetto giustifica le tecniche di bilanciamento operate nei modelli predittivi in seguito.

**Figura 1**: Grafico a torta della variabile risposta X65

## 3.2 Presenza di valori missing

Analizzando il dataset, si è notato un consistente numero di valori mancanti nei predittori, codificati come "?" e convertiti in NA in fase di lettura del dataset. Nello specifico, essi rappresentano 9888 valori in totale.
Per risolvere il problema della presenza dei missing values, è stata eseguita in diversi passaggi una tecnica di imputazione ragionevole.




Come primo passo, sono stati contati tutti gli NA presenti in ciascuna variabile esplicativa. Ordinando le variabili in base al conteggio, sono stati ottenuti i seguenti risultati:

| X1 | X2 | X3 | X6 | X7 | X10 | X11 | X14 | X18 | X22 | X25 | X29 | X35 | X36 | X38 | X48 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| X51 | X55 | X57 | X59 | X9 | X15 | X8 | X16 | X17 | X26 | X34 | X50 | X61 | X4 | X12 | X33 |
| **0** | **0** | **0** | **0** | **3** | **8** | **14** | **14** | **14** | **14** | **14** | **14** | **17** | **18** | **18** | **18** |
| X40 | X46 | X63 | X5 | X58 | X13 | X19 | X20 | X23 | X30 | X31 | X39 | X42 | X43 | X44 | X49 |
| **18** | **18** | **18** | **25** | **29** | **43** | **43** | **43** | **43** | **43** | **43** | **43** | **43** | **43** | **43** | **43** |
| X56 | X62 | X47 | X52 | X32 | X41 | X24 | X28 | X53 | X54 | X64 | X45 | X60 | X27 | X21 | X37 |
| **43** | **43** | **86** | **86** | **86** | **101** | **202** | **227** | **228** | **228** | **228** | **591** | **592** | **715** | **807** | **4736** |

**Tabella 2**: rappresentazione dei valori mancanti in ordine crescente per variabile

Dalla tabella 2 si evince che la variabile X37 presenta un numero eccessivamente alto di missing values (all'incirca il 45% della composizione totale della variabile). Dal momento che norme euristiche consigliano l'eliminazione di una variabile esplicativa nel caso in cui il numero di valori mancanti raggiunga almeno il 20%, si è ritenuto opportuno escludere X37 dall'analisi.

Il passo successivo è stato quello di contare il numero di NA rispetto alle osservazioni del dataset "bank" e, successivamente, di eliminare tutte le righe con almeno 5 valori mancanti. In questo modo il numero di osservazioni si è ridotto a 10295. I restanti valori mancanti sono stati sostituiti con la mediana della relativa variabile di appartenenza. In alternativa, gli stessi potevano essere sostituiti con la media, ma si è preferito lavorare con la mediana, poiché quest'ultima è un indice di posizione più robusto in presenza di outliers.

## 3.3 Problemi di collinearità

Spostando l'attenzione sulla matrice di correlazione, si è voluto investigare la presenza di eventuali correlazioni tra le variabili.
Dalla Figura 2 emergono gruppi di variabili altamente correlate.
Tale evidenza statistica potrebbe essere ricondotta a ragioni di tipo economico. Le variabili X48, X1, X1, X7, X14, X22, per esempio, possono essere sintetizzate dal rapporto *Profitti/Totale Attivo.*
Un altro esempio è fornito dal gruppo di variabili X39, X42, X49, X31, X19, che riassumono il rapporto *Profitti/Ricavi da Vendite e Prestazioni*.
Si spiega economicamente anche la correlazione tra le variabili X12, X16, X26, riassumibili in *Profitti/ Debiti*.
Si noti che queste ultime variabili sono correlate positivamente anche con X9 (*Ricavi da Vendite e Prestazioni/Totale attivo*) e X34 (*Costo della Produzione/Totale Passivo*). A tal proposito, è evidente che un'azienda tenda a sostenere costi proporzionalmente

maggiori in macchinari e/o nuovo personale addetto (*Costo della Produzione, Totale Passivo, Debiti*) per coprire una maggior domanda di beni/servizi (*Ricavi da Vendite e Profitti*.).
Un ulteriore esempio è dato dalle variabili X51 e X2 correlate all'85% e sintetizzabili dal rapporto *Passività/Attività*. Esse risultano correlate negativamente con X3, X25, X10, X38, esprimibili, invece, come rapporto *Capitale/Totale Attivo*.
Intuitivamente, un'azienda inadempiente (*Passività*) risulterà facilmente in situazione sottocapitalizzazione (*Capitale*) per il risarcimento dei debitori.

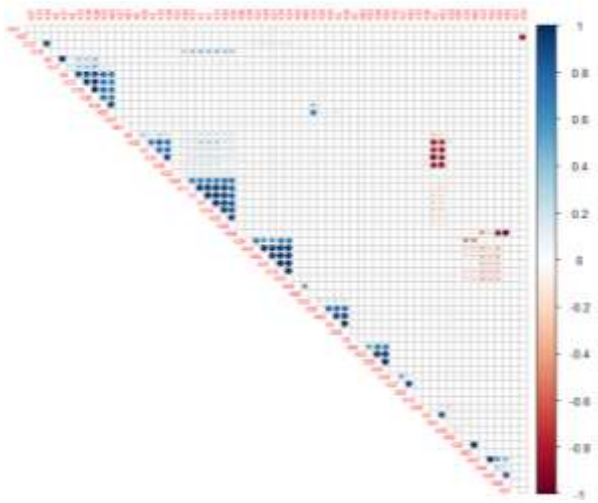


**Figura 2**: Rappresentazione grafica della matrice di correlazione

Dal momento che variabili fortemente correlate ad altre non forniscono alcuna informazione aggiuntiva, ma rischiano persino di peggiorare la previsione, si è deciso di utilizzare il comando "findCorrelation" per stabilire quali variabili eliminare. Questa funzione confronta coppie di variabili che, in valore assoluto, sono correlate almeno al 70% (opzione *cutoff* impostata dall'utente) e restituisce come output un vettore contenente le variabili che dovrebbero essere scartate per evitare collinearità. Nello specifico, ad ogni confronto viene scartata la variabile mediamente più correlata con le restanti.
Eseguendo la funzione, le variabili da scartare risultano essere:

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X1 | X2 | X3 | X4 | X7 | X10 | X11 | X12 | X14 | X16 | X17 | X19 | X20 | X22 | X23 | X25 | X26 |
| X28 | X30 | X34 | X35 | X37 | X40 | X42 | X43 | X47 | X49 | X50 | X51 | X53 | X54 | X56 | X57 | X63 |

In seguito alle eliminazioni, il dataset finale è composto da 29 variabili esplicative e una variabile risposta. Ricalcolando la matrice di correlazione, la rappresentazione grafica di quest'ultima ha confermato l'assenza di collinearità tra le variabili rimanenti.

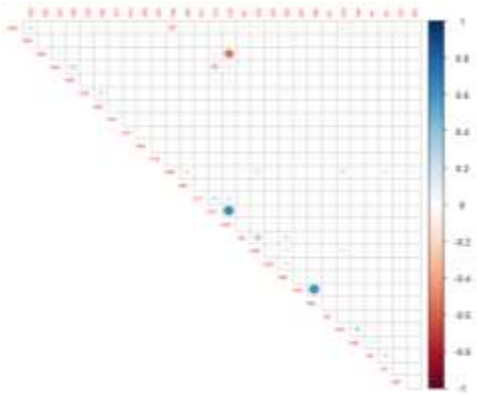


**Figura 3**: Matrice di correlazione dopo l'eliminazione tramite findCorrelation

## 3.4 Analisi dei boxplot

Infine, sono stati considerati i boxplot delle variabili esplicative condizionate alla variabile di risposta.
Osservando un esempio di questi Figura 4a, è stata supposta l'assenza di variabili estremamente significative a fini previsivi.
Tuttavia, studiando gli stessi boxplot senza outliers, risulta visibile che alcune di esse sono in grado di discriminare in maniera abbastanza accurata (come ad esempio le variabili X8, X38, X46, vedi Figura 4b)

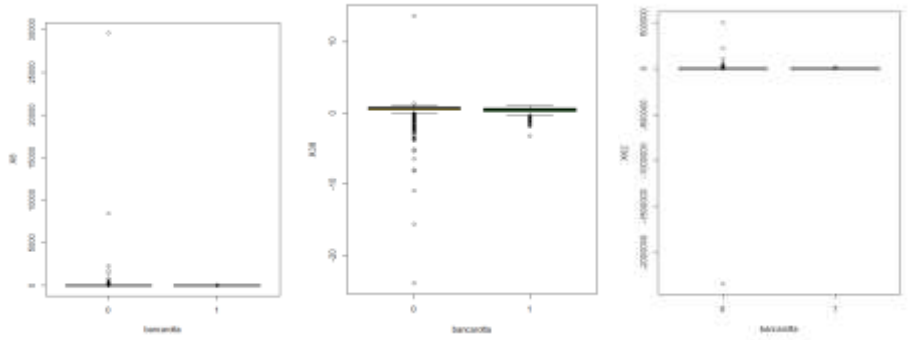


**Figura 4a:** Boxplot di variabili maggiormente significative (X8, X38 e X46), condizionati alla variabile risposta (con outliers)

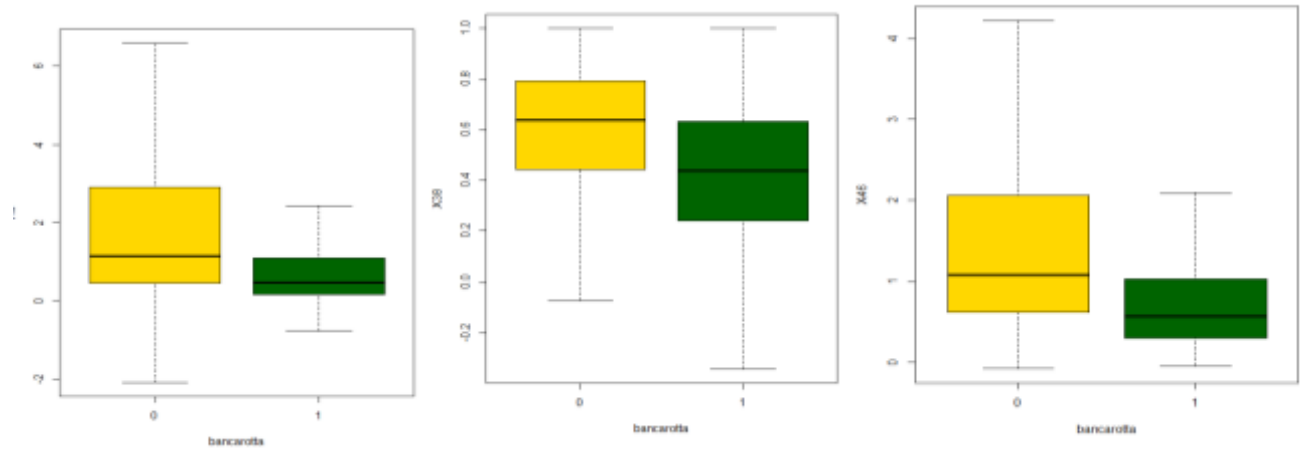



**Figura 4b:** Boxplot di variabili maggiormente significative (X8, X38 e X46), condizionati alla variabile risposta (senza outliers)

Al contrario, altre variabili, quali X9 e X59 (Figura 5), non discriminano affatto se un'azienda possa finire o meno in bancarotta. Dal momento che neppure le variabili X2, X29, X36, X44, X60, X61, X64 sembrano essere particolarmente significative, si è deciso di escluderle.

Il dataset finale presenta quindi 20 variabili esplicative e una variabile risposta.

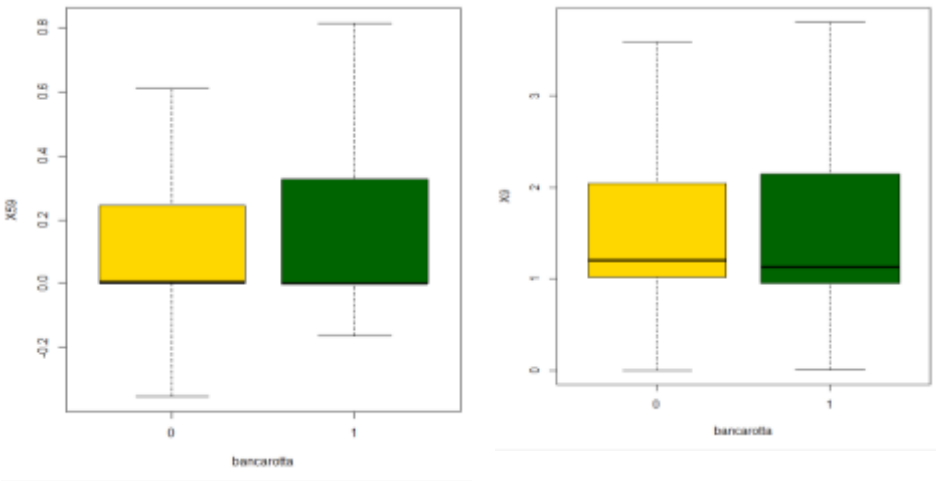



**Figura 5:** Boxplot di variabili poco significative (X9 e X59), condizionati alla variabile risposta (senza outliers)

Il dataset finale presenta quindi 20 variabili esplicative e una variabile risposta.

## 3.5 Presenza di valori outliers

Ulteriore informazione ricavabile dai boxplot è la quantità di valori al di fuori dei cosiddetti "baffi", essi sono definiti outliers deboli, quando non si discostano di molto dal boxplot, e outliers forti quando lo scostamento è maggiore. Notiamo che la percentuale di questi valori è alta in queste variabili, e influisce fortemente sull'asimmetria delle distribuzioni.

Tuttavia, proprio in ragione di tale numerosità, scegliamo di lasciare intaccati questi valori per non stravolgere troppo la reale distribuzione delle variabili.
Riportiamo i valori in seguito:

| Variabile | N° outiliers | % Outliers | Media degli outliers | Media della variabile con outliers | Media della variabile senza outliers |
|---|---|---|---|---|---|
| X5 | **1481** | **17%** | **-9426.8** | **-1370.07** | **1.26** |
| X6 | **2926** | **40.3%** | **-0.05** | **0** | **0.01** |
| X8 | **1097** | **12.1%** | **87.19** | **10.65** | **1.4** |
| X13 | **1040** | **11.4%** | **3.01** | **0.38** | **0.08** |
| X15 | **1726** | **20.4%** | **7537.67** | **2140.78** | **1039.19** |
| X24 | **759** | **8.1%** | **0.74** | **0.23** | **0.19** |
| X27 | **2204** | **27.6%** | **5186.95** | **1123.82** | **1.34** |
| X31 | **1303** | **14.7%** | **-1.83** | **-0.19** | **0.05** |
| X32 | **773** | **8.1%** | **13728.67** | **1108.43** | **83.91** |
| X33 | **789** | **8.4 %** | **38.11** | **7.72** | **5.17** |
| X38 | **284** | **2.9%** | **-0.84** | **0.56** | **0.6** |
| X41 | **1739** | **20.6%** | **170.15** | **29.14** | **0.1** |
| X45 | **1809** | **21.6%** | **92.91** | **16.79** | **0.34** |
| X46 | **1077** | **11.8 %** | **18.6** | **18.6** | **1.18** |
| X48 | **1091** | **12%** | **-0.13** | **0.01** | **0.03** |
| X52 | **761** | **8.1%** | **149.59** | **11.39** | **0.23** |
| X55 | **1724** | **20.4%** | **31327.18** | **6616.56** | **1579.77** |
| X58 | **1201** | **13.4%** | **6.39** | **1.58** | **0.94** |
| X62 | **761** | **8.1%** | **-1050.32** | **-7.29** | **76.96** |

**Tabella 3:** Analisi dei valori outliers

# 4. Modelli di previsione con alberi di classificazione

## 4.1 Definizione della tecnica

Dal momento che la variabile di risposta è una *dummy* Y={0,1}, è possibile definire un modello di previsione mediante alberi di classificazione in grado di partizionare lo spazio delle variabili esplicative in regioni.
Ad ogni passo, la divisione ricorsiva binaria permette di scegliere la variabile $X_j$ con j =1,...,20 e il punto *c* in cui effettuare la divisione, tali per cui le regioni

$$R_1 = \{X|X_j < c\} \quad \text{e} \quad R_2 = \{X|X_j \geq c\}$$

Massimizzano il guadagno dell'informazione

$$i(R) - p_1 i(R_1(j,c)) - p_2 i(R_2(j,c))$$

Dove:

- $p_k$ è la proporzione di osservazioni della regione $R_k(j,c)$
- $i(R_k(j,c))$ è l'impurità della regione $R_k(j,c)$ calcolata mediante l'indice di Gini:

$$\sum_{l \in Y} \hat{p}_{kl}(1 - \hat{p}_{kl}) \quad \text{con} \quad \hat{p}_{kl} = \frac{1}{\text{card}(R_k)} \sum_{t: x_t \in R_k} I(y_t = l)$$

## 4.2 Albero di classificazione con training set non bilanciato

Al fine di poter confrontare i risultati ottenuti dalle varie tecniche di seguito illustrate, è stato diviso il dataset totale in un training set, sul quale costruire il classificatore, e un test set, sul quale testare il potere previsivo, di proporzioni 70% e 30% rispettivamente.
L'albero migliore, ottenuto sul training set attraverso *K-fold cross validation* con k=100, possiede 204 foglie.

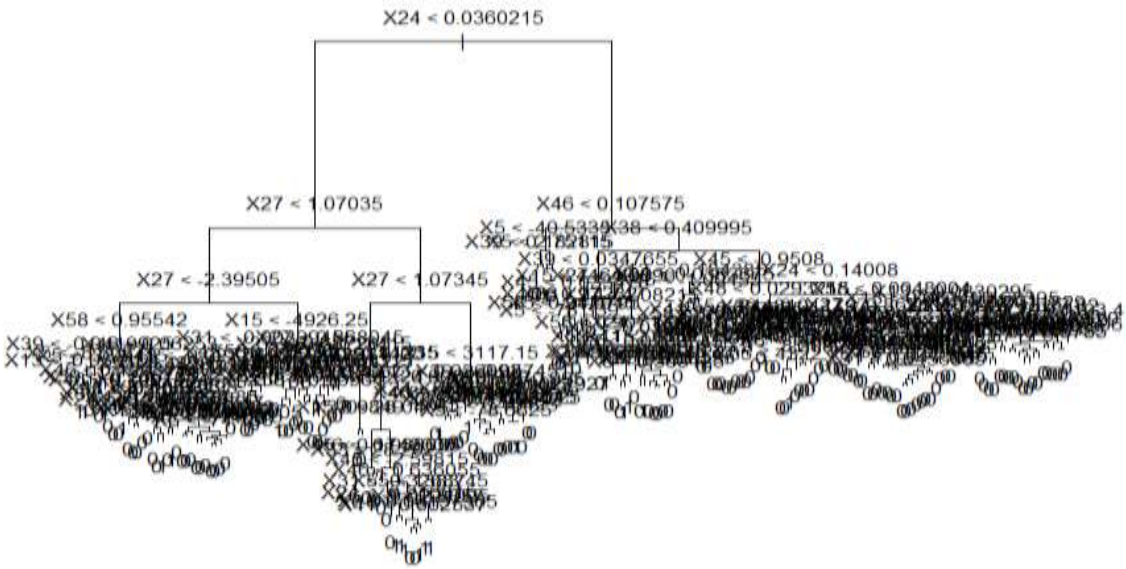



**Figura 6:** Albero migliore ottenuto via CV sul training set non bilanciato

Applicando quest'albero di classificazione alle osservazioni contenute nel test set, sono stati ottenuti i seguenti risultati:

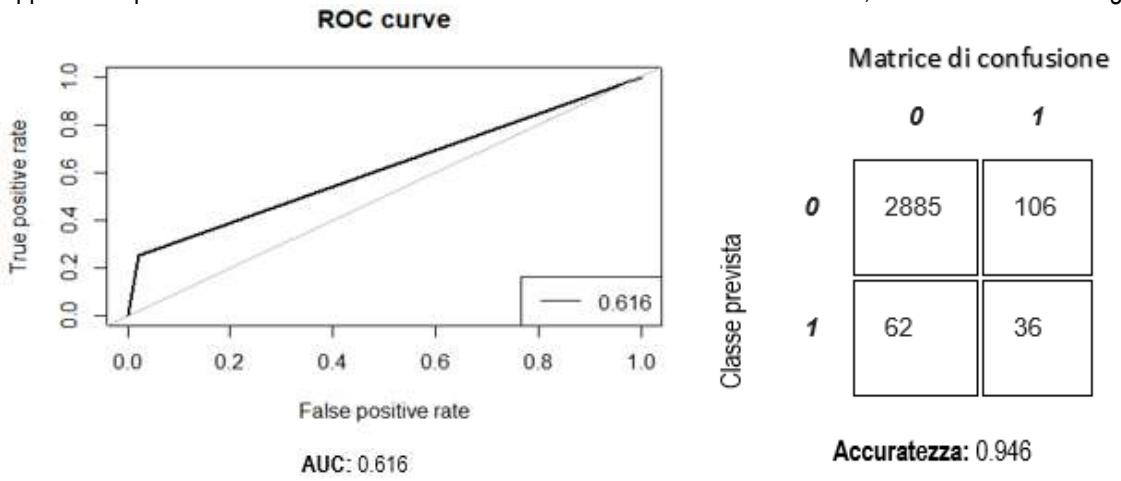



**Figura 7**: Curva ROC e matrice di confusione con training set non bilanciato

Un'accuratezza così elevata è dovuta allo sbilanciamento della variabile risposta: si ottiene un ottimo 98% di classificazioni corrette per la classe 0 ed un misero 25% di classificazioni corrette per la classe 1.
Date le premesse fatte, un valore di AUC (*area under the curve*) pari a 0.611 può ritenersi un discreto successo.

## 4.3 Albero di classificazione con training set bilanciato

È possibile replicare nuovamente il procedimento precedentemente descritto bilanciando il training set, in modo da avere lo stesso numero, pari a 240, di risposte di valore 0 e di valore 1.
Poiché il numero di osservazioni del training set è diminuito considerevolmente da 7206 a 480, è stata implementata una *leave one out CV,* da cui è emerso che l'albero di dimensioni migliori presenta 20 foglie.

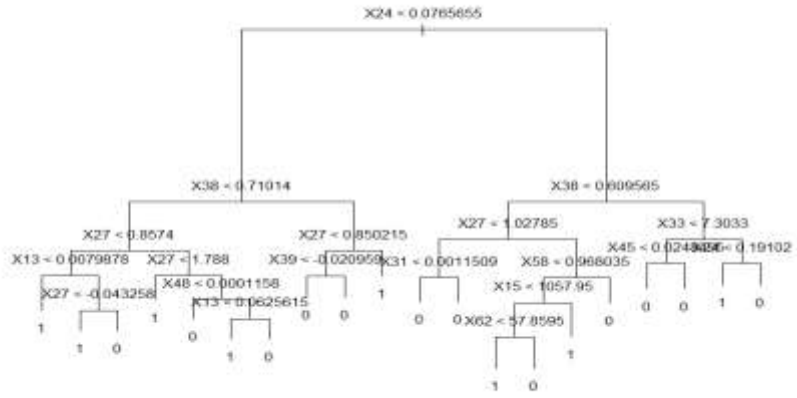


**Figura 8**: albero di classificazione con training set bilanciato

Valutando la performance dell'albero ottenuto sul test set, abbiamo ottenuto i seguenti risultati:

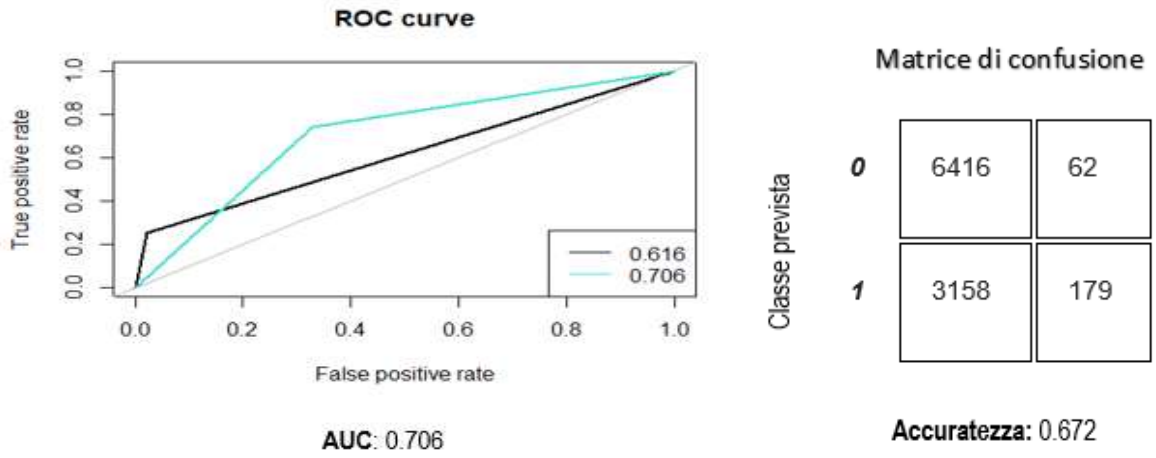



**Figura 9**: curva ROC e matrice di confusione con *training set* bilanciato

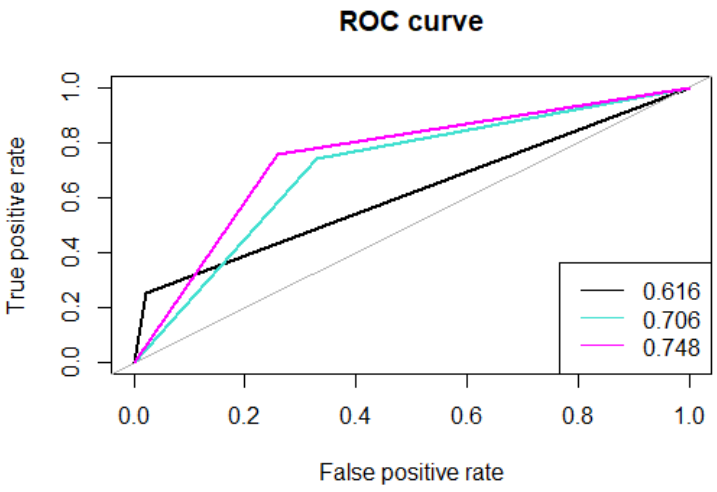
In Figura 9 è possibile osservare la curva ROC del training set bilanciato, in azzurro, sovrapposta a quella in Figura 7.
Come ci si poteva aspettare, l'AUC è aumentato, infatti la maggiore quantità di risposte di valore 1 consente all'algoritmo implementato in R di fare una migliore previsione per queste risposte.
Al miglioramento dell'AUC segue un peggioramento dell'accuratezza (aumentano gli errori per la risposta di valore 0); tuttavia si noti che le corrette classificazioni di entrambe le risposte di valore 0 e di valore 1, rispettivamente pari a 67% e 74%, prevalgono su quelle scorrette.
Globalmente, quindi, si può ritenere che il bilanciamento migliori la previsione.

## 4.4 Albero di classificazione con training set bilanciato + *bagging*

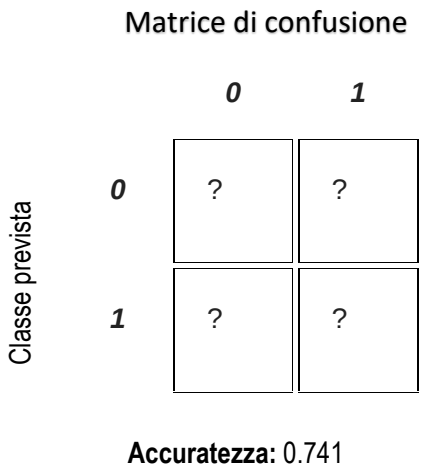
I risultati ottenuti nel paragrafo 4.3 sono migliorabili attraverso l'algoritmo *bagging*, il quale permette di disporre di un certo numero di insiemi di training, mentre il test set rimane invariato.
Più precisamente i passi dell'algoritmo sono:

1. Si estrae con ripetizione dal training set bilanciato, un numero pari alla metà di questo (240 in questo caso).
2. Si stima l'albero senza effettuare la potatura
3. Si fa la previsione sul *test set*
4. Si ripetono i punti 1,2,3 per *n* volte (si è scelto *N*=500) in modo da avere *n* previsioni
5. Si classifica ogni unità con 0 o 1, in base al numero maggiore di volte con cui tale unità è predetta (in caso di parità la classificazione è casuale)

Aggiungendo alla Figura 9 la curva ROC, in rosa, (ottenuta dall'algoritmo sopracitato) otteniamo la Figura 10.
Essendo il *bagging* un multiclassificatore (come *boosting, Adaboosting ecc.*), in quanto tale è dimostrato essere teoricamente e praticamente più performante rispetto a classificatori semplici, quando i singoli classificatori sono indipendenti o parzialmente indipendenti tra di loro (condizione verificata nel *bagging* poiché si classificano porzioni di training differenti). Per cui non ci sorprende che in termini di AUC che in termini di accuratezza le previsioni siano migliori di quelle del paragrafo 4.3.




**AUC**: 0.748




**Accuratezza:** 0.741

**Figura 10**: curva ROC e matrice di confusione con l'aggiunta del *training set* bilanciato + *bagging*

## 4.5 Albero di classificazione con l'algoritmo "*ovun.sample*" + *bagging*

Nella libreria *Rose* di R, è contenuta la funzione *ovun.sample*, che permette di effettuare un bilanciamento più efficiente sovra-campionando la risposta di valore 1 (*over*), sotto-campionando quella di valore 0 (*under*) o applicando entrambi i campionamenti (*both*).
Si è scelto di creare un campione di 1000 aziende con *method=both,* di cui 503 hanno risposta 0 e 497 risposta 1.
L'albero migliore restituito dalla *leave one out CV* ha 39 foglie,

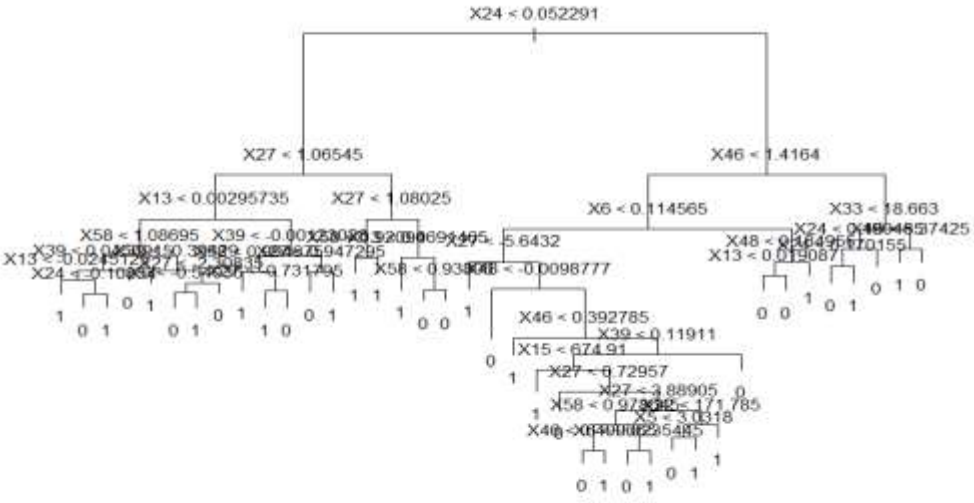
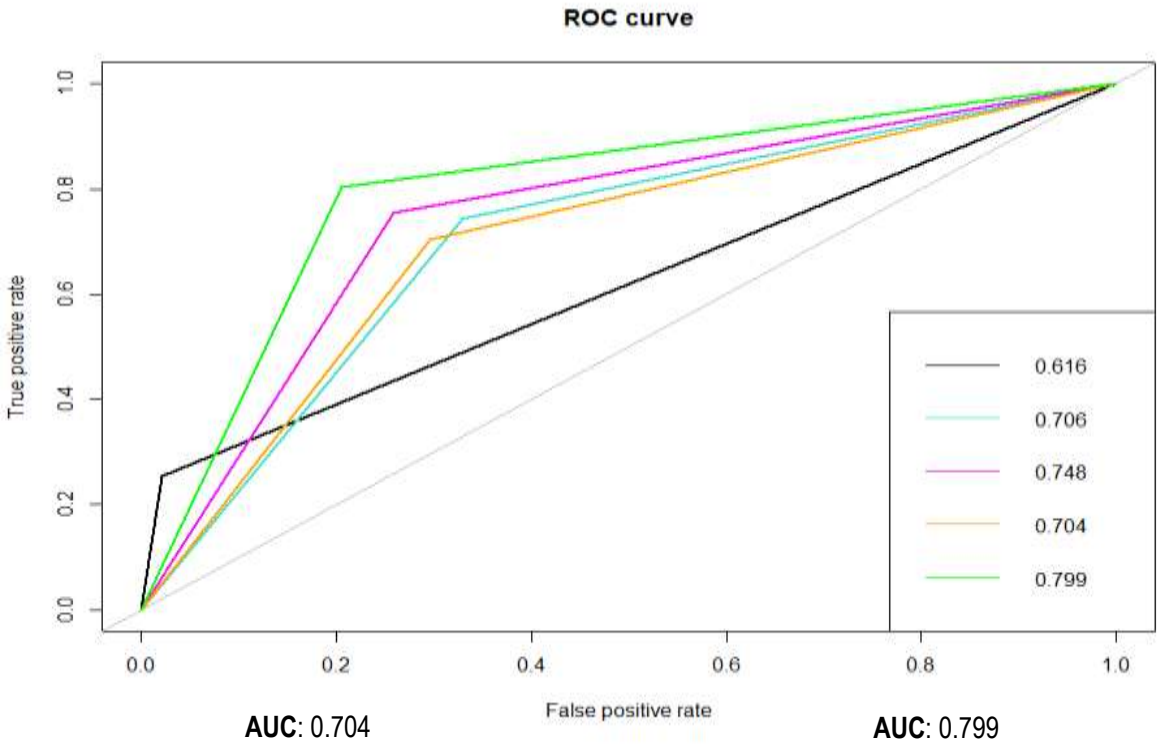



**Figura 11**: albero con *ovun.sample*

mentre le previsioni sul *test set* sono :
Accuratezza= 0.704
AUC= 0.704
Si noti che in tutti gli alberi la variabile che definisce il primo *split* è sempre X24, ovvero il Profitto lordo (in 3 anni) / Totale attivo.
Si vuole precisare che un solo anno di osservazione non è sufficiente per prevedere il fallimento o meno di un'azienda, o meglio, una forte perdita in un anno non pregiudica necessariamente il fallimento dell'impresa. (Le contromisure dettate dalla normativa fallimentare, si noti che Polonia segue la legislazione europea, hanno inizio dopo tre anni di perdite consecutive.)
Dal momento che X24 è l'unica variabile che considera un periodo di tempo superiore all'anno, non è sorprendente che sia scelta dall'algoritmo come miglior discrimine della classe fallimento o non fallimento.

Applicando l'algoritmo *bagging* per 500 volte, otteniamo il miglior risultato in termini di AUC nel corso di tutta l'analisi. In Figura 12 sono riportati in arancione e in verde i valori dell'AUC ottenuti con l'algoritmo ovun.sampling senza e con bagging. L'accuratezza ottenuta con questo algoritmo è pari a 79.5%.




**AUC**: 0.704

**AUC**: 0.799

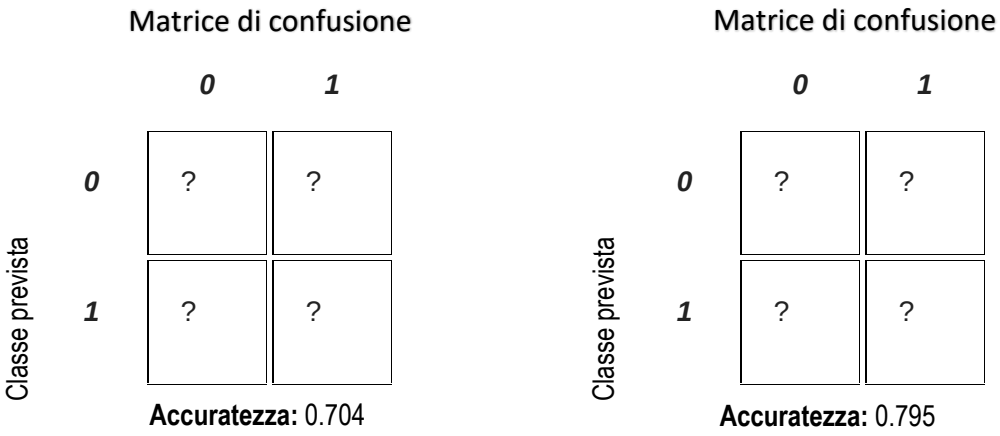



**Figura 12**: curva ROC e matrici di confusione con l'aggiunta di *ovun.sample + bagging*

# 5.Modelli di previsione con analisi discriminante

## 5.1 Definizione della tecnica

L'analisi discriminante è una tecnica supervisionata, che permette di classificare nuove unità di cui non si conosce la classe di appartenenza.
In questa sezione si è deciso di implementare un'analisi discriminante lineare, facendo la seguente assunzione:

- $Xl|Y = j \sim N(\mu_l, \sigma^2)$ con $l = 1, \dots 20$ $e$ $j = 0{,}1$

Quindi j dipende dalla classe, mentre $\sigma^2 > 0$ è costante, ovvero stiamo assumendo la condizione di omoschedasticità
Notiamo che a priori l'omoschedasticità e la normalità risultano essere non particolarmente ragionevoli in virtù della massiccia presenza di valori anomali, anche 15% (vedi paragrafo 3.5).
Per verificare tali assunzioni; sono riportati in Figura 12 gli istogrammi delle distribuzioni delle variabili continue condizionate al valore assunto dalla variabile risposta.

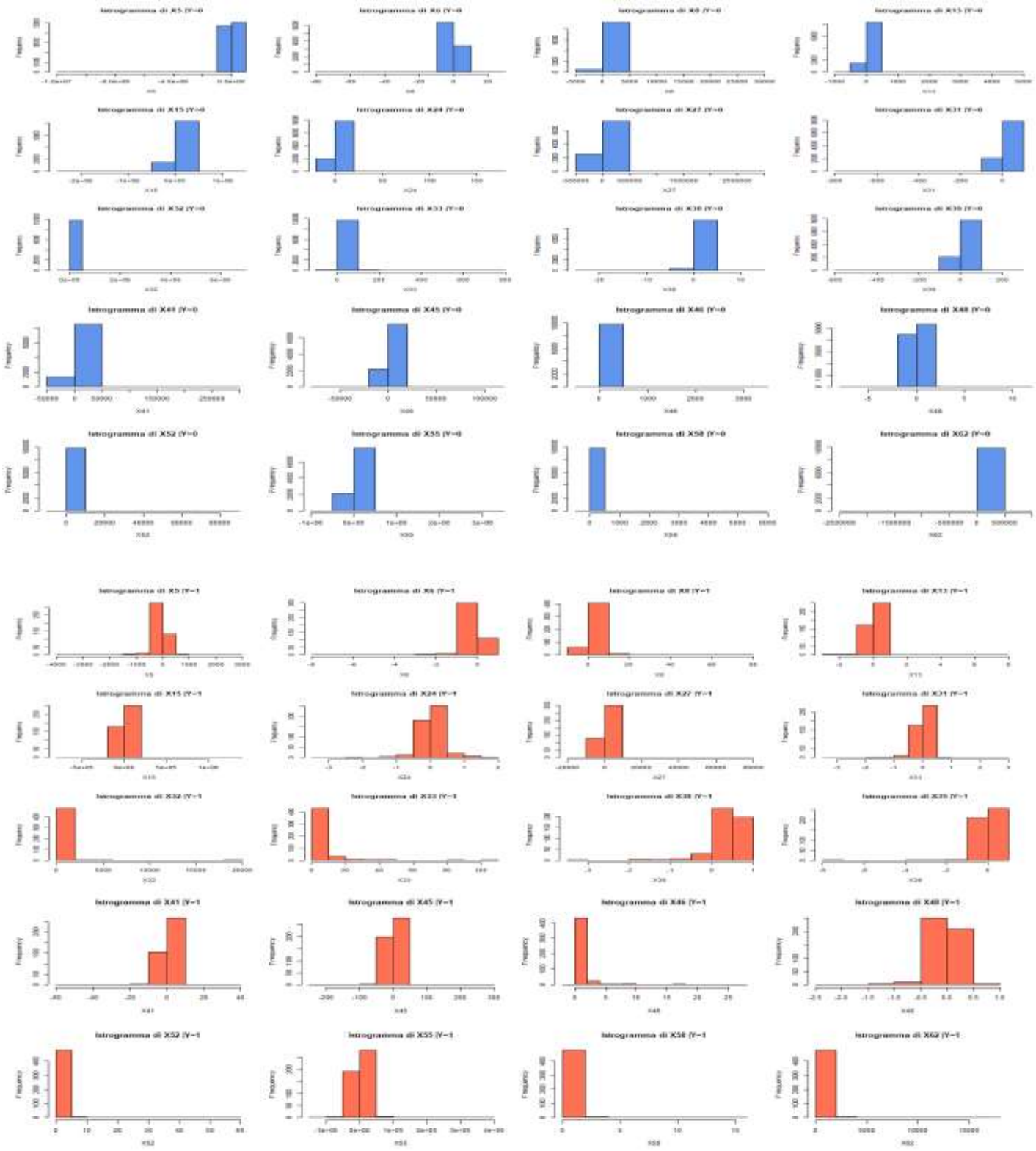
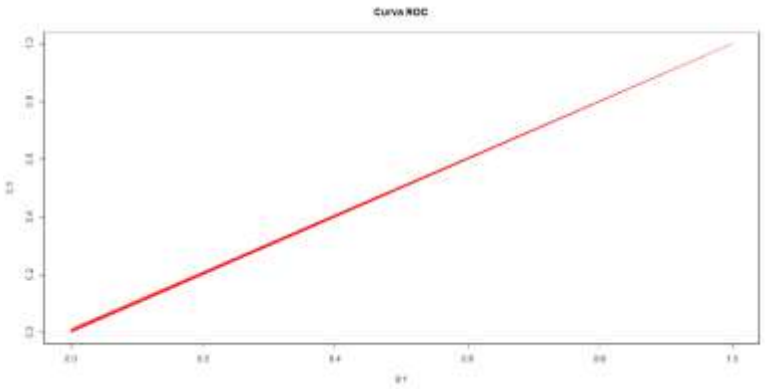


**Figura 13**: Grafici distribuzione empirica delle variabili esplicative, condizionate alla classe Y=1 bankruptcy

Si può osservare dagli istogrammi fortemente asimmetrici che le assunzioni descritte precedentemente non sembrano rispettate. Provando ad effettuare un'analisi quadratica volta evitare il problema dell'omoschedasticità, sono stati osservati risultati peggiori rispetto all'analisi discriminante lineare, quindi si è deciso di soffermarsi sul caso lineare.

## 5.2 Analisi discriminante lineare con training set non bilanciato

Utilizzando un ciclo while, il dataset originario è stato diviso in training set e test set 100 volte, in proporzione 70% e 30% rispettivamente. Ad ogni passo è stata applicata l'analisi discriminante lineare ai dati di training e sono stati calcolati accuratezza e AUC, confrontando i valori previsti dal modello con i dati del test set.
I risultati in media sono i seguenti:




**AUC**: 0.500

Matrice di confusione

| Classe prevista | 0 | 1 |
|---|---|---|
| 0 | ? | ? |
| 1 | ? | ? |

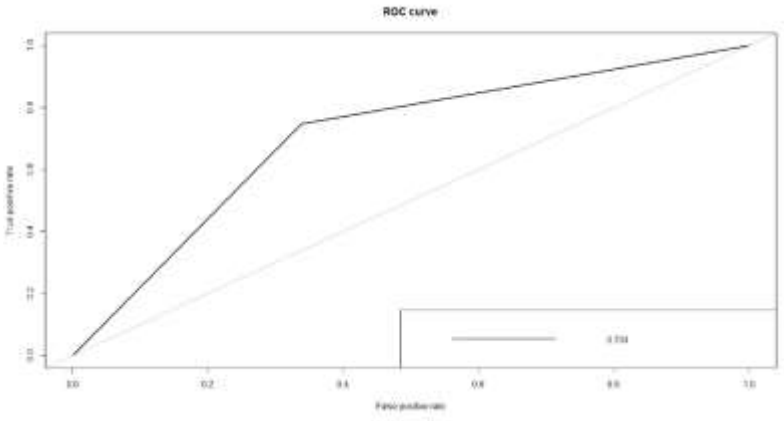
**Accuratezza:** 0.950

**Figura 14**: Curva ROC e matrice di confusione con training set non bilanciato

Si evince che in media l'analisi discriminante lineare non è un buon modello per la previsione. Infatti, l'area sotto la curva ROC afferma che globalmente il classificatore non ha un buon andamento. L'elevata accuratezza è dovuta al fatto che il dataset non è bilanciato e, quindi, vi è un'alta proporzionalità di corrette classificazioni nel caso delle aziende non in bancarotta.

## 5.2 Analisi discriminante lineare con training set bilanciato + bagging

E' stato bilanciato il training set prendendo un campione composto da 240 osservazioni di banche in fallimento e altrettante non in bancarotta, poichè il numero di aziende con classe pari a 1 è 480. Il campione restante è stato usato come test set.
Sfruttando nuovamente l'algoritmo *bagging* con *n*=500, si sono ottenuti i seguenti risultati:




**AUC**: 0.704

Matrice di confusione

| Classe prevista | 0 | 1 |
|---|---|---|
| 0 | 6334 | 3240 |
| 1 | 61 | 180 |

**Accuratezza:** 0.664

**Figura 15**: Curva ROC e matrice di confusione training set bilanciato + bagging

L'accuratezza non è particolarmente elevata, dal momento che il modello classifica correttamente solo il 5% delle banche in fallimento (ovvero con risposta pari a 1). L'AUC, invece, è più elevato rispetto al caso precedente, ovvero con training set sbilanciato, quindi si può concludere che il bilanciamento abbia portato a dei risultati migliori.

# 6.Conclusioni

Per consolidare i risultati ottenuti in precedenza si è voluto ripetere con un ciclo *while*, ciascun metodo predittivo trattato in precedenza *n*=100 volte. Questa operazione è volta ad evitare che i nostri risultati delle analisi (misurati con metriche di AUC e accuracy) siano troppo influenzati da una scelta casuale del campione (particolarmente fortunata o sfortunata). L'implementazione con *while* non sembra contraddire quanto dimostrato in precedenza, ma, anzi, ne è la conferma.

A seguito riportiamo i boxplot che rappresentano la distribuzione dei 100 valori per i metodi ripetuti, opportunamente separata in AUC e Accuracy.

A conclusione si è proposta una tabella che mette a confronto tutti i modelli con e senza ciclo while (come valore unico si è inserita la mediana delle 100 ripetizioni).

Il miglior risultato, considerando AUC ed ACCURACY, è sempre dato dal modello ADC *+ovun.sample + bagging.*

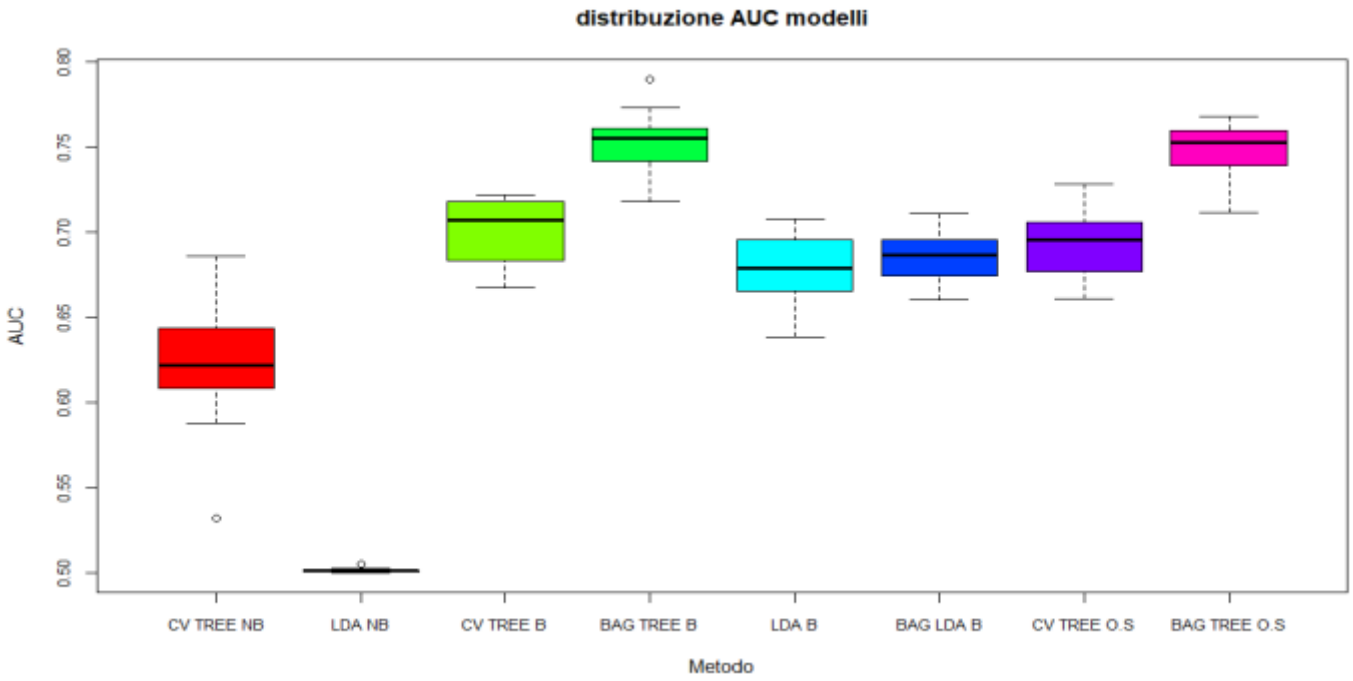



**Figura 16**: AUC ripetuti (N=100) per ogni modello sotto forma di boxplot

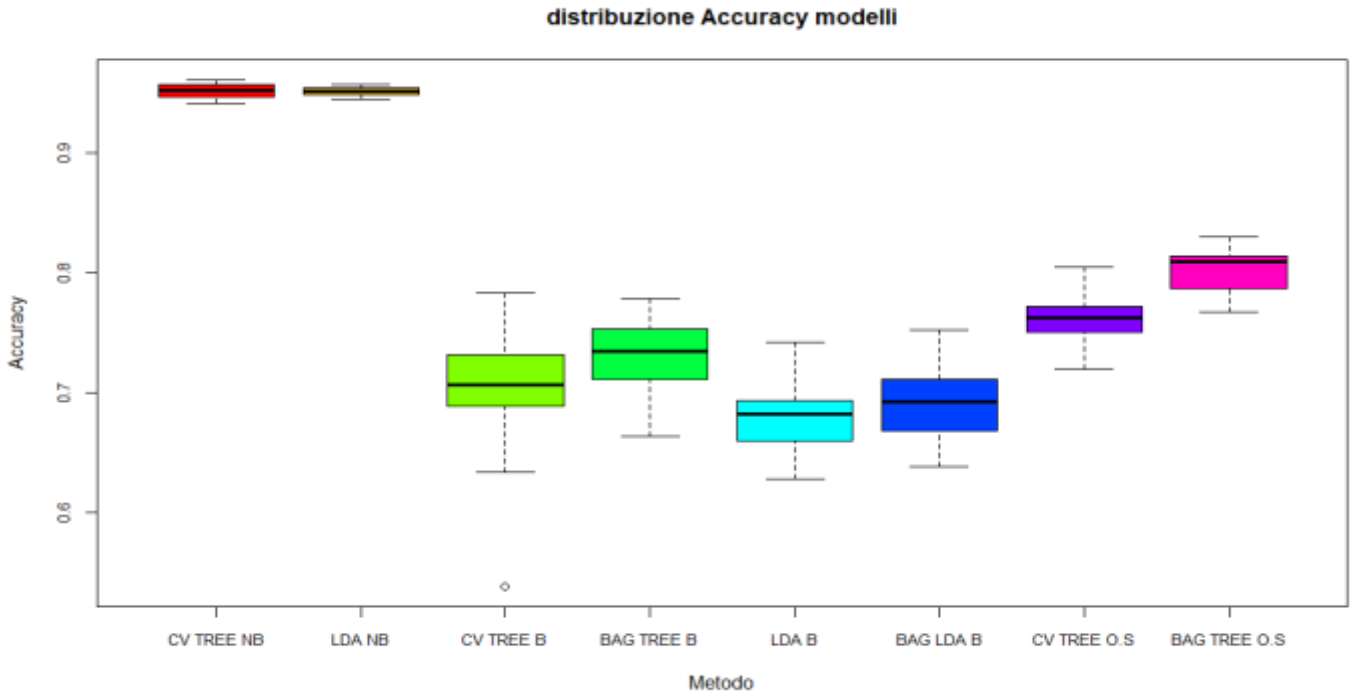



**Figura 17**: Accuracy ripetuta (N=100) per ogni modello sotto forma di boxplot

| Modello | Accuracy Ciclo while | Accuracy | AUC Ciclo while | AUC |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| ADC non bilanciato + *cv* | 0.952 | 0.946 | 0.622 | 0.616 |
| ADC bilanciato + *cv* | 0.706 | 0.706 | 0.707 | 0.672 |
| ADC bilanciato + *bagging* | 0.734 | 0.748 | 0.755 | 0.741 |
| ADC *ovun.sample* | 0.762 | 0.704 | 0.695 | 0.704 |
| ADC *ovun.sample* + *bagging* | 0.809 | 0.795 | 0.752 | 0.799 |
| LDA non bilanciata | 0.951 | 0.950 | 0.500 | 0.500 |
| LDA bilanciata + *bagging* | 0.692 | 0.664 | 0.686 | 0.704 |

**Tabella 4:** Performance dei modelli predittivi

# 7.Bibliografia


[1] Emerging Markets Information Service, *URL:* https://www.emis.com/
[2] Machine Learning Repository *URL:https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data*
[3] Mercati Emergenti URL:http://www.wallstreetitalia.com/trend/mercati-emergenti/
[4] Zieba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction. Expert Systems with Applications
*URL: https://www.ii.pwr.edu.pl/~tomczak/PDF/[MZSTJT].pdf*
[5] L. Breiman, Machine learning, Springer 1996
[6] Rete giudiziaria Europea in materia civile e commerciale URL: http://ec.europa.eu/civiljustice/bankruptcy/bankruptcy_ger_it.htm