Modelli predittivi per la biodegradabilità di una sostanza chimica

F. Bekollari, A. Giabelli, L. Mandelli, F. Marigo

1 INTRODUZIONE

L'obiettivo principale del Regolamento Europeo *REACH*[1] è quello di migliorare la conoscenza riguardo i pericoli derivanti da prodotti chimici e in quest'ottica prevede la registrazione[2], la validazione e l'autorizzazione delle sostanze chimiche allo scopo di assicurare un maggiore livello di protezione della salute umana e dell'ambiente[1]. *REACH* allo stesso tempo incoraggia l'uso di metodi alternativi alla sperimentazione sugli animali, come ad esempio i modelli *QSAR*[3], modelli matematici che possono essere utilizzati per prevedere determinate proprietà - ad esempio fisico-chimiche o biologiche - delle sostanze, a partire dalla conoscenza della loro struttura chimica[2]. Lo scopo di questo lavoro consiste nel costruire modelli di previsione riguardo la biodegradabilità di una sostanza sulla base della sua struttura chimica. Per fare ciò sono stati utilizzati tre tipi di modelli, ovvero i modelli lineari generalizzati (*GLM*), l'analisi discriminante lineare (*LDA*) e gli alberi di classificazione (*ADC*), utilizzando, per quanto riguarda gli ultimi due, diversi tipi di configurazioni. Il modello migliore tra tutti quelli implementati è risultato essere l'albero di classificazione sviluppato considerando solo le variabili indicate come più significative dal GLM e applicando l'algoritmo *bagging*, ottenendo un valore di *AUC* pari a 0.866.

2 DESCRIZIONE DEL DATASET

Il dataset *QSAR* sulla biodegradazione delle sostanze è stato costruito dal "Milano Chemometrics and QSAR Research Group" (Università degli Studi Milano Bicocca, Milano, Italia). Esso contiene 1055 osservazioni riguardanti sostanze chimiche delle quali sono state analizzate 42 proprietà riportate in Tabella 1. I valori sperimentali di biodegradabilità riguardanti queste sostanze chimiche sono stati raccolti dalla pagina web del "National Institute of Technology and Evaluation of Japan". Per classificare una sostanza chimica come *Ready Biodegradable* (RB), è stata misurata la richiesta biochimica di ossigeno della stessa se immersa in acqua per 28 giorni; nel caso in cui tale richiesta presenti un valore superiore a 60, allora il test risulta positivo. In caso contrario, la sostanza viene classificata come *Not Ready Biodegradable* (NRB)[3].

Per maggiore chiarezza nell'applicazione dei metodi e per mantenere coerenza con la teoria, nei seguenti capitoli i dati vengono suddivisi e denotati come segue. Si definisce la matrice

$$\underset{1055\times 41}{\mathbf{X}} = \begin{pmatrix} x_{1,1} & \dots & x_{1,41} \\ \vdots & \ddots & \vdots \\ x_{1055,1} & \dots & x_{1055,41} \end{pmatrix} = \left(\underline{x}_{\bullet 1}, \dots, \underline{x}_{\bullet 41}\right) = \left(\underline{x}_{1\bullet}, \dots, \underline{x}_{1055\bullet}\right)^T$$

la quale costituisce la matrice del disegno nei vari modelli. Questa matrice $\mathbf{X}$ racchiude 1055 realizzazioni indipendenti del vettore casuale:

$$\underline{X} = (X_1, \dots, X_{41})$$

dove le X_l $(l = 1, \dots, 41)$ rappresentano le variabili esplicative. Per quanto riguarda infine il campione di 1055 determinazioni della variabile dipendente Y, la quale assume nell'esperimento i due valori 0 (NRB) e 1 (RB), si utilizza la notazione $\underline{y} = (y_1, \dots, y_{1055})^T$.

[1] Registration, Evaluation, Authorisation and Restriction of Chemicals.

[2] La registrazione avviene in un database comune a tutti gli Stati membri dell'UE.

[3] Quantitative Structure–Activity Relationship.

Var	Tipo	Breve descrizione	Supporto
X_1	Continua	*SpMax_L*: Autovalori dominanti della matrice di Laplace	[2.00, 6.50]
X_2	Continua	*J_Dz(e)*: Indice di Balaban	[0.80, 9.18]
X_3	Discreta	*nHM*: Numero di atomi pesanti	{0, 1,..., 12}
X_4	Discreta	*F01[N-N]*: Frequenza di azoto-azoto a distanza topologica 1	{0, 1,..., 3}
X_5	Discreta	*F04[C-N]*: Frequenza di carbonio-azoto a distanza topologica 4	{0, 1,..., 36}
X_6	Discreta	*NssssC*: Numero di atomi di tipo ssssC	{0, 1,...,13}
X_7	Discreta	*nCb-*: Numero di carboni di tipo benzenico sostituiti (sp2)	{0, 1,...,18}
X_8	Continua	*C%*: Percentuale di atomi di Carbonio	[0.00, 60.70]
X_9	Discreta	*nCp*: Numero di carboni terminali primari (sp3)	{0, 1,...,24}
X_{10}	Discreta	*nO*: Numero di atomi di ossigeno	{0, 1,...,12}
X_{11}	Discreta	*F03[C-N]*: Frequenza di carbonio-azoto a distanza topologica 3	{0, 1,...,44}
X_{12}	Continua	*SdssC*: Numero di dssC E-states	[-5.26, 4.72]
X_{13}	Continua	*HyWi_B(m)*: Indice Hyper-Wiener dalla matrice di Burden pesata dalla massa	[1.54, 5.70]
X_{14}	Continua	*LOC*: Indice di lopping	[0.00, 4.50]
X_{15}	Continua	*SM6_L*: Momento spettrale di ordine 6 dalla matrice di Laplace	[4.17, 12.61]
X_{16}	Discreta	*F03[C-O]*: Frequenza di Carbonio-Ossigeno a distanza topologica 3	{0, 1,...,40}
X_{17}	Continua	*Me*: Media dell'elettronegatività atomica di Sanderson	[0.96, 1.31]
X_{18}	Continua	*Mi*: Ionizzazione media	[1.02, 1.38]
X_{19}	Discreta	*nN-N*: Numero di idrazine	{0, 1, 2}
X_{20}	Discreta	*nArNO2*: Numero di gruppi nitro aromatici	{0, 1,...,3}
X_{21}	Discreta	*nCRX3*: Numero di CRX3	{0, 1,...,3}
X_{22}	Continua	*SpPosA_B(p)*: Somma autov. della matrice di B. pesata dalla polarizzabilità	[0.86, 1.64]
X_{23}	Discreta	*nCIR*: Numero di circuiti	{0, 1,...,147}
X_{24}	Binaria	*B01[C-Br]*: Presenza/assenza di Carbonio - Bromo alla distanza topologica 1	{0,1}
X_{25}	Binaria	*B03[C-Cl]*: Presenza/assenza di Carbonio - Cloro alla distanza topologica 3	{0,1}
X_{26}	Discreta	*N-073*: Ar2NH / Ar3N / Ar2N-Al / R..N..R	{0, 1,...,3}
X_{27}	Continua	*SpMax_A*: Autovalori dominanti della matrice di adiacenza (indice L-P)	[1.00, 2.86]
X_{28}	Continua	*Psi_i_1d*: Indice di pseudoconnettività intrinseca (tipo 1d)	[-1.09, 1.07]
X_{29}	Binaria	*B04[C-Br]*: Presenza/assenza di carbonio - bromo a distanza topologica 4	{0,1}
X_{30}	Continua	*SdO*: Somma di dO E-states	[0.00, 71.17]
X_{31}	Continua	*TI2_L*: Secondo indice di Mohar dalla matrice di Laplace	[0.44, 17.54]
X_{32}	Discreta	*nCrt*: Numero di anelli terziari C(sp3)	{0, 1,...,8}
X_{33}	Discreta	*C-026*: R–CX–R	{0, 1,...,12}
X_{34}	Discreta	*F02[C-N]*: Frequenza di C-N a distanza topologica 2	{0, 1,...,18}
X_{35}	Discreta	*nHDon*: Numero di atomi donatori per legami idrogeno (N e O)	{0, 1,...,7}
X_{36}	Continua	*SpMax_B(m)*: Autovalori dominanti dalla matrice di B. pesati dalla massa	[2.27, 10.70]
X_{37}	Continua	*Psi_i_A*: Indice di pseudoconnettività intrinseca	[1.47, 5.83]
X_{38}	Discreta	*nN*: Numero di atomi di azoto	{0, 1,...,8}
X_{39}	Continua	*SM6_B(m)*: Momenti sesti spettrali dalla matrice di B. pesata dalla massa	[4.92, 14.70]
X_{40}	Discreta	*nArCOOR*: Numero di esteri aromatici	{0, 1,...,4}
X_{41}	Discreta	*nX*: Numero di atomi alogeni	{0, 1,...,27}
Y	Binaria	*Biodegradabilità*: "Ready Biodegradable" (1) e "Not-Ready-Biodegradable" (0)	{RB, NRB}

Tabella 1: Descrizione delle variabili osservate e misurate nei 1055 esperimenti.

3 ANALISI DESCRITTIVE ED ESPLORATIVE

3.1 Analisi descrittiva

Nel dataset non sono presenti dati mancanti; la variabile risposta presenta 699 sostanze registrate come non biodegradabili e 356 come biodegradabili. La codifica utilizzata, come detto in precedenza, è la seguente: NRB = 0 (rosso) e RB = 1 (verde).

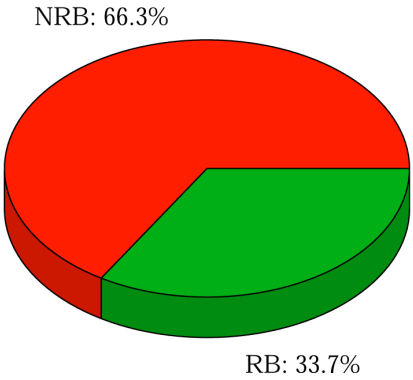

Figura 1: Grafico a torta della variabile risposta.

Si sposta ora l'attenzione sulle variabili esplicative, studiando le correlazioni, solo per quanto riguarda le variabili continue, e le distribuzioni marginali.

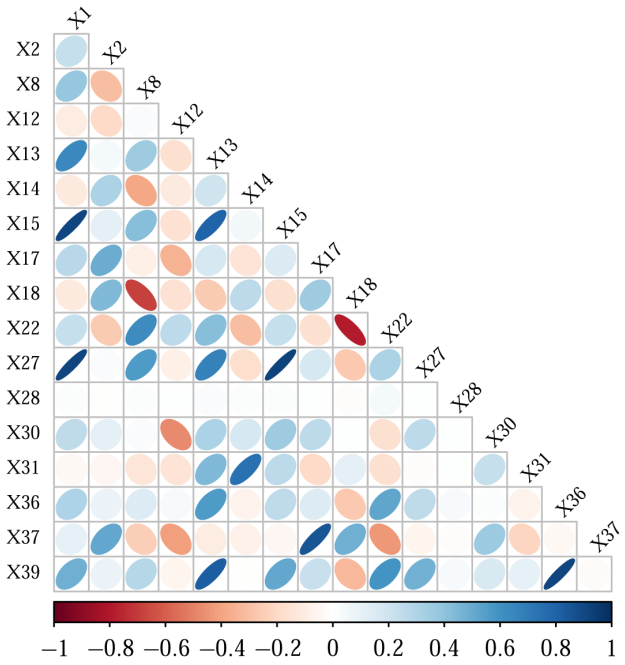

Figura 2: Correlazioni tra le variabili esplicative continue.

Da tale grafico si può osservare come vi siano alcune variabili fortemente correlate tra loro[4] (ad esempio le variabili X_{27} e X_1), mentre altre presentino una correlazione quasi nulla (per esempio la variabile X_{28} è incorrelata con tutte le altre variabili). Di seguito si riportano i boxplot delle stesse variabili continue (Figura 3) e i barplot delle variabili discrete (Figura 4).

[4]questo fatto non sorprende in quanto i valori delle variabili molto correlate spesso corrispondono a misure estratte da una stessa matrice (ad esempio dalla matrice di Burden pesata dalla massa).

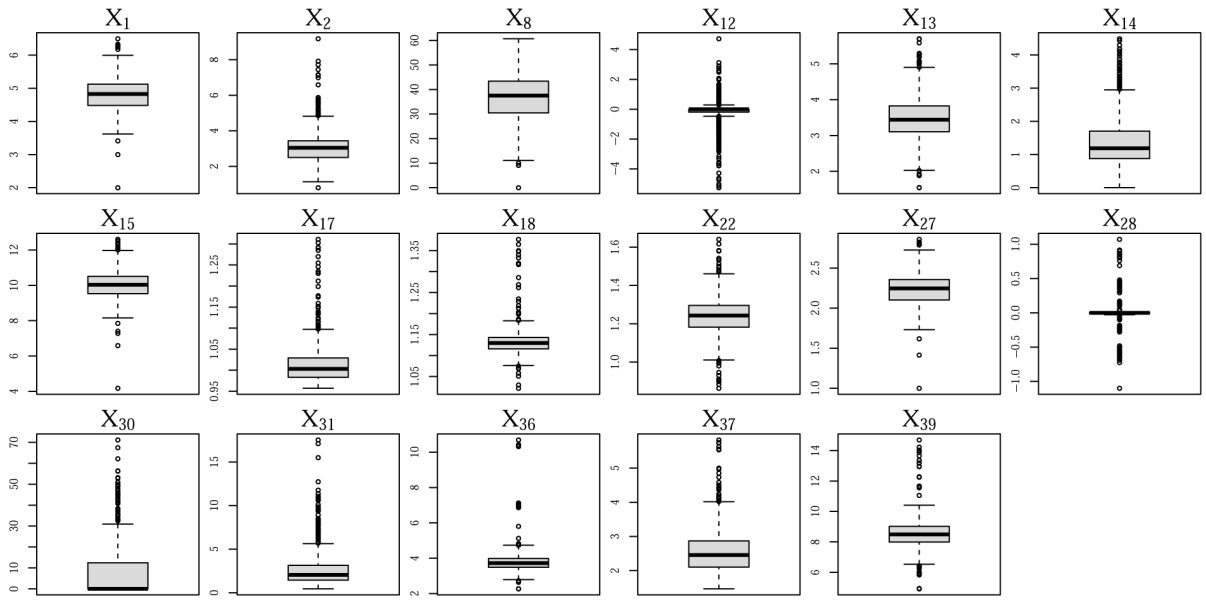

Figura 3: Boxplot delle variabili esplicative continue

Si può notare come le variabili X_{17}, X_{30}, X_{31} e X_{36} presentino una forte asimmetria, mentre le variabili $X_2, X_{14}, X_{15}, X_{18}$ e X_{37} presentano una asimmetria più lieve.

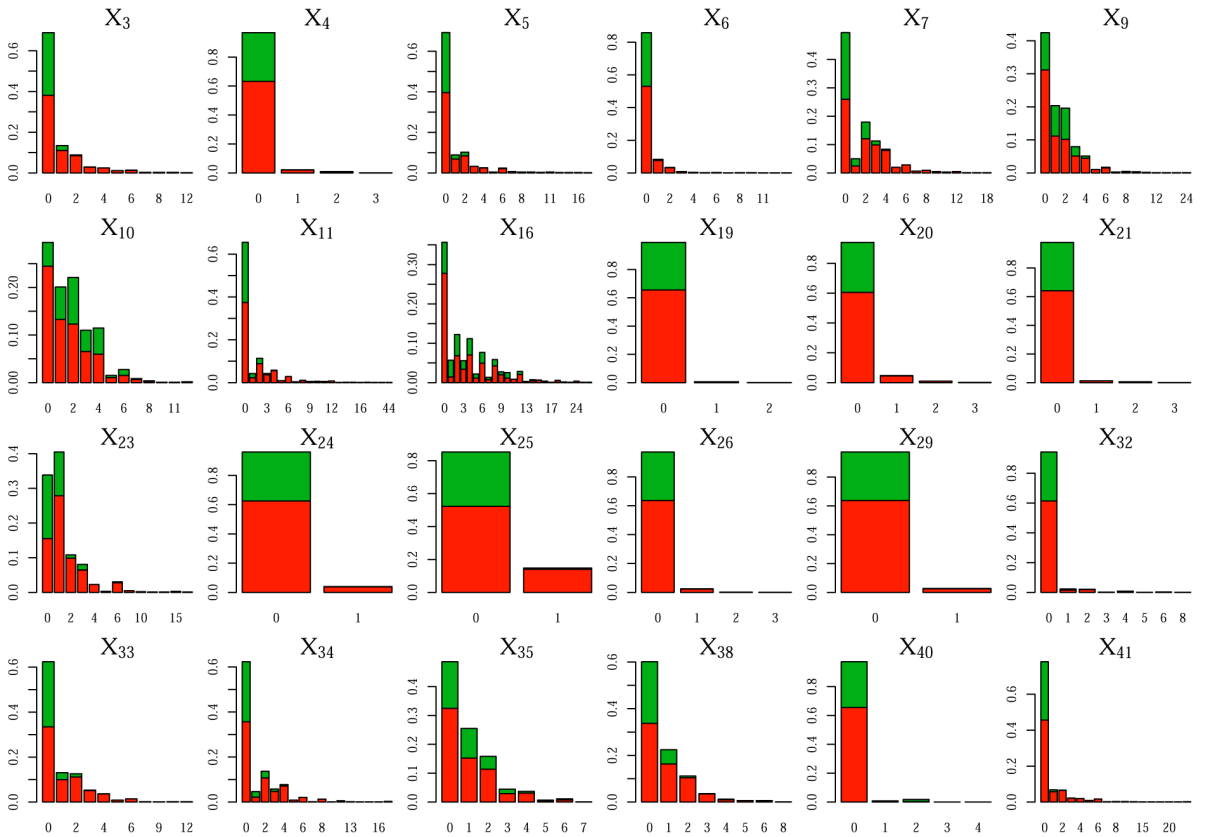

Figura 4: Barplot delle variabili esplicative discrete suddivise in base a Y.

Dai barplot emerge un fatto degno di nota: in molte variabili discrete (come ad esempio la X_4 e la X_{21}), le osservazioni la cui classe è 1 (RB) presentano sempre e solo la modalità dominante (più del 90% delle osservazioni assume tale modalità). Ciò implica che gli "outlier", ovvero tutte le osservazioni con modalità diverse dalla dominante, saranno evidentemente classificate come NRB.

3.2 Analisi esplorativa

Dall'analisi descrittiva è emerso che diverse variabili discrete risultano essere potenzialmente molto influenti in un modello predittivo. Per verificare se anche qualche variabile esplicativa continua possa discriminare in modo evidente la classe di Y, in Figura 5 si riportano i boxplot condizionati a Y.

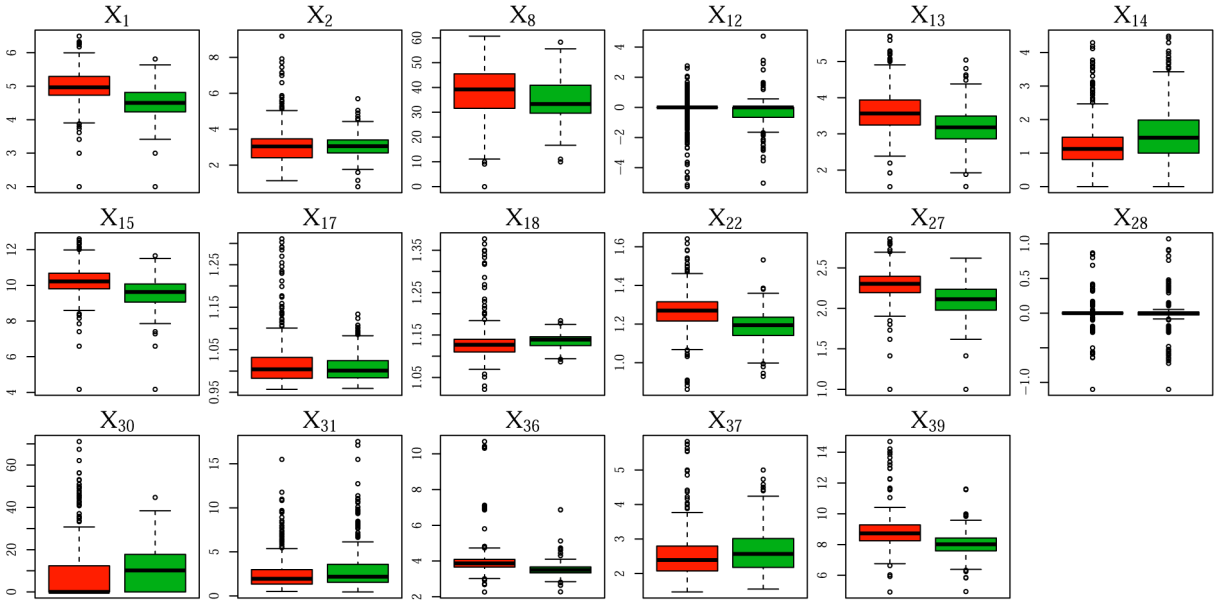

Figura 5: Boxplot delle variabili esplicative continue condizionate alla variabile risposta Y.

I boxplot – delimitati da primo e terzo quartile – sono sempre sovrapposti: ciò fa supporre che non vi sia una singola variabile in grado di effettuare una distinzione netta tra le due classi di Y. Inoltre si osserva la presenza di alcune variabili nelle quali gli outlier sono un forte fattore discriminante, come ad esempio le variabili X_{17} e X_{18}, dove la maggior parte degli outlier appartengono alla classe delle sostanze non biodegradabili, in analogia con quanto detto per alcune variabili discrete. Queste situazioni sono molto importanti e determinanti nella creazione degli alberi di classificazione.

Si effettuano, infine, gli scatterplot delle variabili esplicative, i quali si possono osservare in Figura 8 a pagina 7; sono riportate solo le variabili continue che nel modello di regressione logistica presentano coefficiente significativo. Nei grafici i punti vengono colorati in base alla classe di appartenenza (RB e NRB). Anche questa analisi non si rivela particolarmente utile dal momento che le osservazioni appartenenti alle due diverse classi di Y risultano essere molto sovrapposte.

4 MODELLO DI PREVISIONE GLM

La variabile risposta Y in analisi è di tipo `factor` (binario); il modello di regressione parametrica più indicato per questo tipo di previsione è la regressione logistica. Tale metodo, inoltre, riesce a fornire informazioni sull'importanza delle variabili esplicative, tramite i test di significatività eseguiti sui coefficienti delle stesse. Interpretando quindi i `pvalue` dei coefficienti, è possibile effettuare una selezione delle variabili allo scopo di specificare un modello più sintetico e performante.

In questo capitolo, come nei successivi, si adopera una suddivisione del dataset in training-set e test-set, estraendo casualmente con il software `R` il 70% delle osservazioni dopo aver fissato un seme costante con la funzione `set.seed(1)`. In questo modo sarà possibile confrontare le performance dei diversi metodi applicati.

Il modello si basa sull'assunzione che ogni estrazione y_i $(i = 1, ..., 1055)$ da $Y_i \sim Bernoulli(\mu_i)$, distribuzione che appartiene alla famiglia di dispersione esponenziale, sia indipendente. Il *link* utilizzato

è la funzione $g(\mu_i) = \log \mu_i (1-\mu_i)^{-1}$ (`logit`). La specificazione del modello, dunque, è la seguente:

$$\mathbb{E}(Y_i) = g^{-1}(\beta_0 + \beta_1\, x_{i,1} + ... + \beta_{41}\, x_{i,41}) \qquad i = 1, ..., 1055$$

I coefficienti β_l $(l = 0, ..., 41)$ sono ignoti e vengono stimati inserendo nel modello il vettore campionario $\underline{y} = (y_1, ..., y_{1055})^T$, grazie al quale con metodi numerici si ottengono le stime di massima verosimiglianza $\hat{\beta}_l$. Allo scopo di mantenere nella specificazione soltanto le variabili con coefficienti significativi, viene applicato l'algoritmo `stepAIC` bidirezionale. Il modello finale ottenuto mantiene l'intercetta e 21 variabili esplicative; i coefficienti corrispondenti sono riassunti in Figura 6.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.16800    2.54162  -2.427 0.015233 *
X1          -2.59236    0.86887  -2.984 0.002849 **
X2          -0.93959    0.23995  -3.916 9.01e-05 ***
X3          -0.66079    0.27731  -2.383 0.017180 *
X6          -2.10363    0.45909  -4.582 4.60e-06 ***
X7          -0.98943    0.15676  -6.312 2.76e-10 ***
X8           0.11774    0.02372   4.963 6.95e-07 ***
X10          0.31027    0.14255   2.177 0.029512 *
X12          0.46883    0.19870   2.360 0.018298 *
X13         -4.53286    0.95822  -4.731 2.24e-06 ***
X14          1.19495    0.24239   4.930 8.23e-07 ***
X15          2.75778    0.75126   3.671 0.000242 ***
X20         -2.04644    0.83033  -2.465 0.013717 *
X24          2.52028    0.94588   2.664 0.007711 **
X26          1.89601    0.88654   2.139 0.032463 *
X30          0.04382    0.02142   2.046 0.040783 *
X32         -1.77211    0.37822  -4.685 2.79e-06 ***
X34         -0.36883    0.13574  -2.717 0.006585 **
X35          0.34378    0.14149   2.430 0.015111 *
X37          1.56220    0.36141   4.323 1.54e-05 ***
X38         -0.81952    0.24770  -3.309 0.000938 ***
X40          1.06456    0.41318   2.576 0.009981 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 6: Stima, standard error e `pvalue` dei coefficienti $\hat{\underline{\beta}}$

Per valutare le performance del modello esso viene implementato sul training-set ottenuto a partire dalla matrice $\mathbf{X}_{\texttt{glm}}$ contenente le 21 variabili identificate in precedenza. Applicando poi il classificatore al test-set si sono ottenuti i valori riportati di seguito in Figura 7.

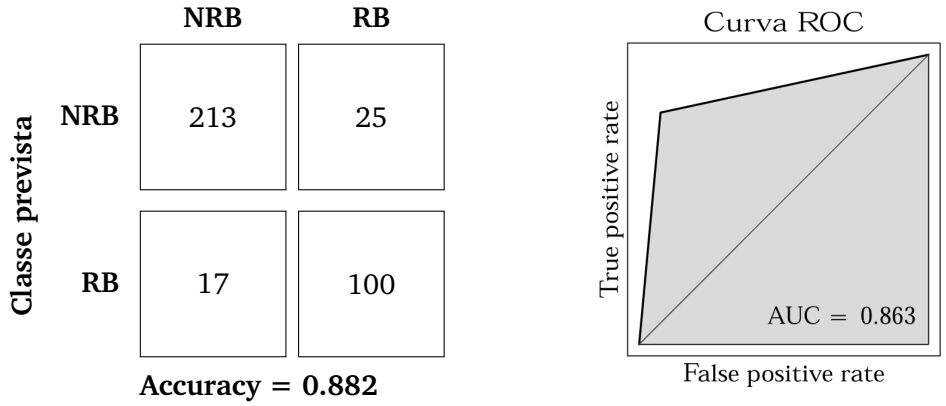

Figura 7: Matrice di confusione e curva ROC del modello predittivo GLM.

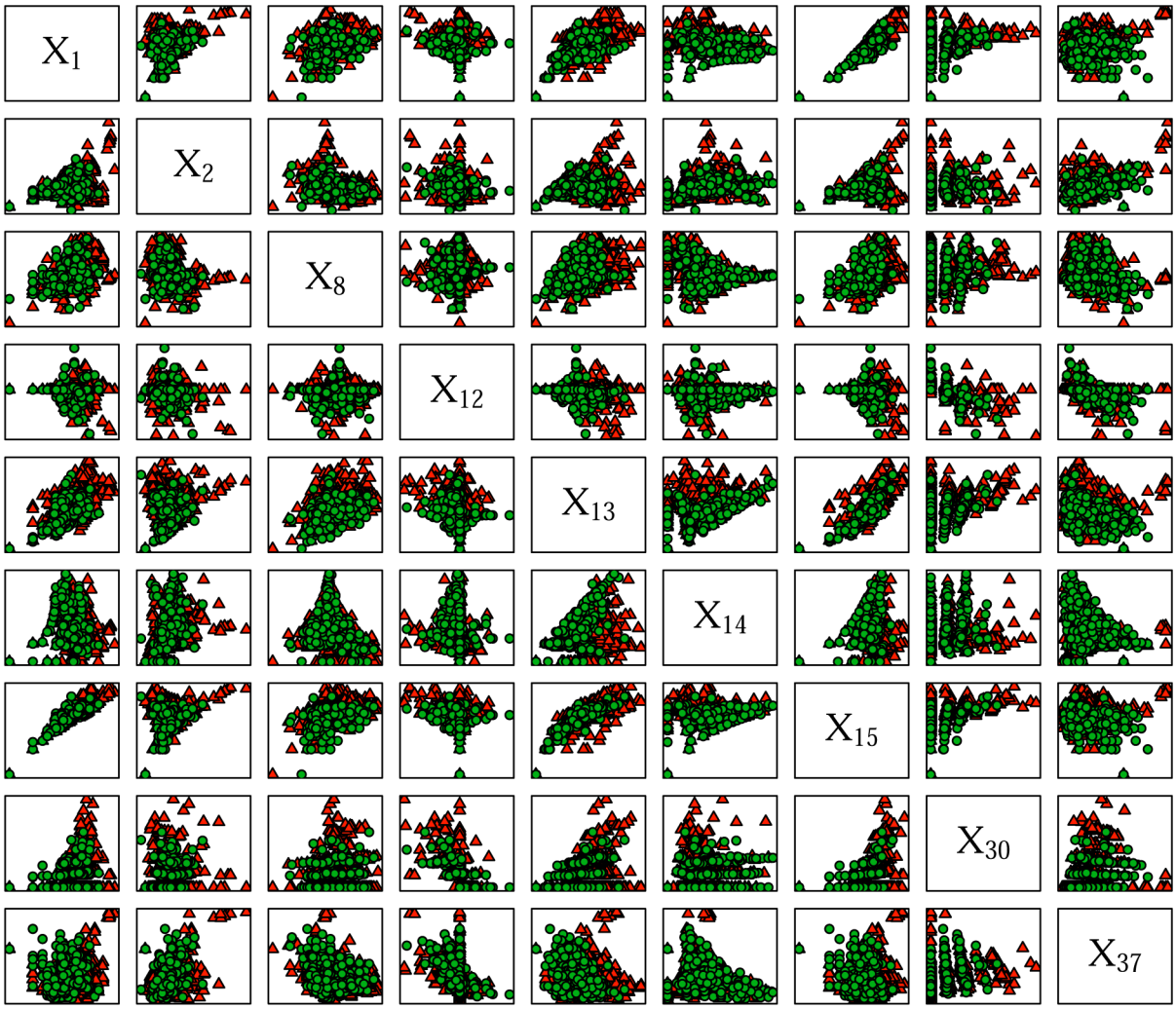

Figura 8: Scatterplot delle variabili esplicative. I punti sono colorati in base alla classe di appartenenza (verde=RB; rosso=NRB).

5 MODELLI DI PREVISIONE CON ANALISI DISCRIMINANTE LINEARE

5.1 Teoria generale

Allo scopo di implementare un'analisi discriminante lineare, si considera il vettore casuale $\underline{X}^C \in \mathbb{R}^{17}$ composto esclusivamente da variabili esplicative continue[5]. Tale modello di previsione si basa su una partizione di $\mathbb{R}^{17}$, ossia lo spazio generato dalle variabili esplicative, in modo tale che, a seconda della regione in cui l'osservazione cade, le venga assegnato un certo valore della variabile risposta Y.

Le ipotesi di tale modello sono:

- Le variabili esplicative devono essere continue, infatti si considera $\underline{X}^C$;
- Le variabili condizionate $X_l|Y = 0$ e $X_l|Y = 1$ ($l = 1, ..., 41$) si distribuiscono come delle $N(\mu_l, \sigma^2)$, con σ^2 costante (ovvero sussiste omoschedasticità).

Per controllare se tali ipotesi siano verificate, di seguito si riportano gli istogrammi delle distribuzioni delle variabili continue condizionate al valore assunto dalla variabile risposta.

[5] $\underline{X}^C = (X_1, X_2, X_8, X_{12}, X_{13}, X_{14}, X_{15}, X_{17}, X_{18}, X_{22}, X_{27}, X_{28}, X_{30}, X_{31}, X_{36}, X_{37}, X_{39})$

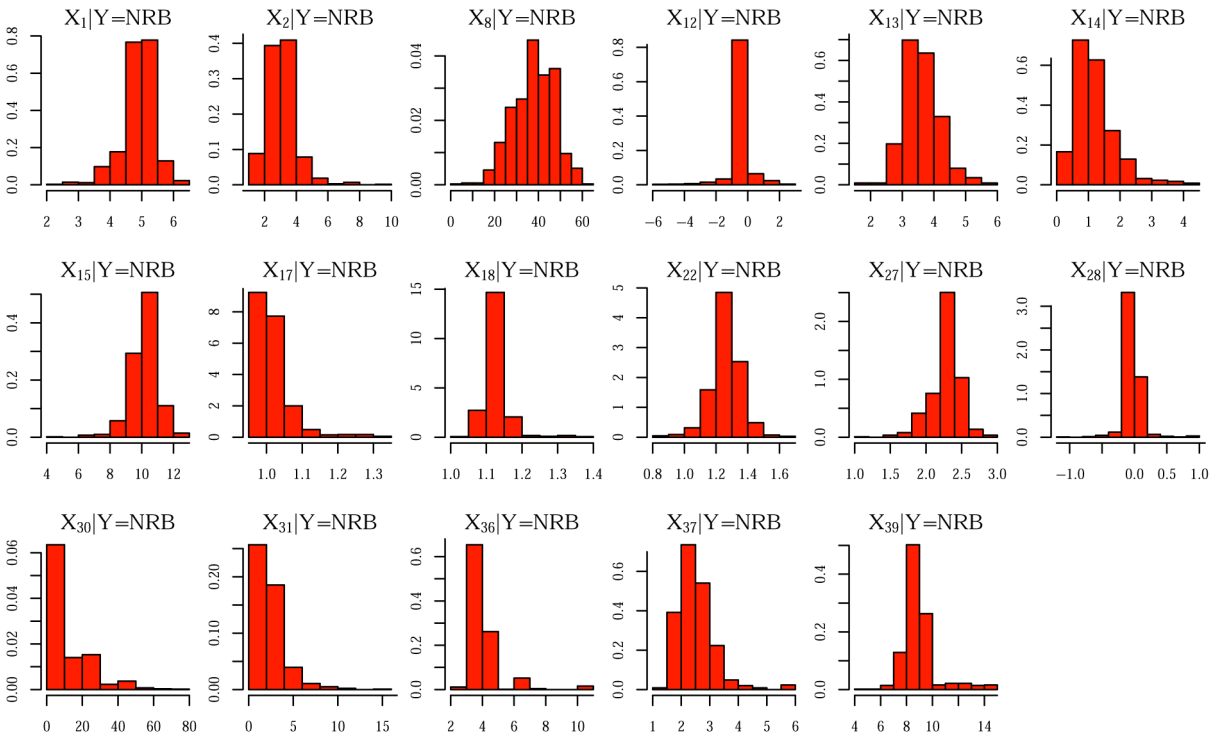

Figura 9: Distribuzioni empiriche delle variabili esplicative, condizionate a $Y = 0$ (NRB).

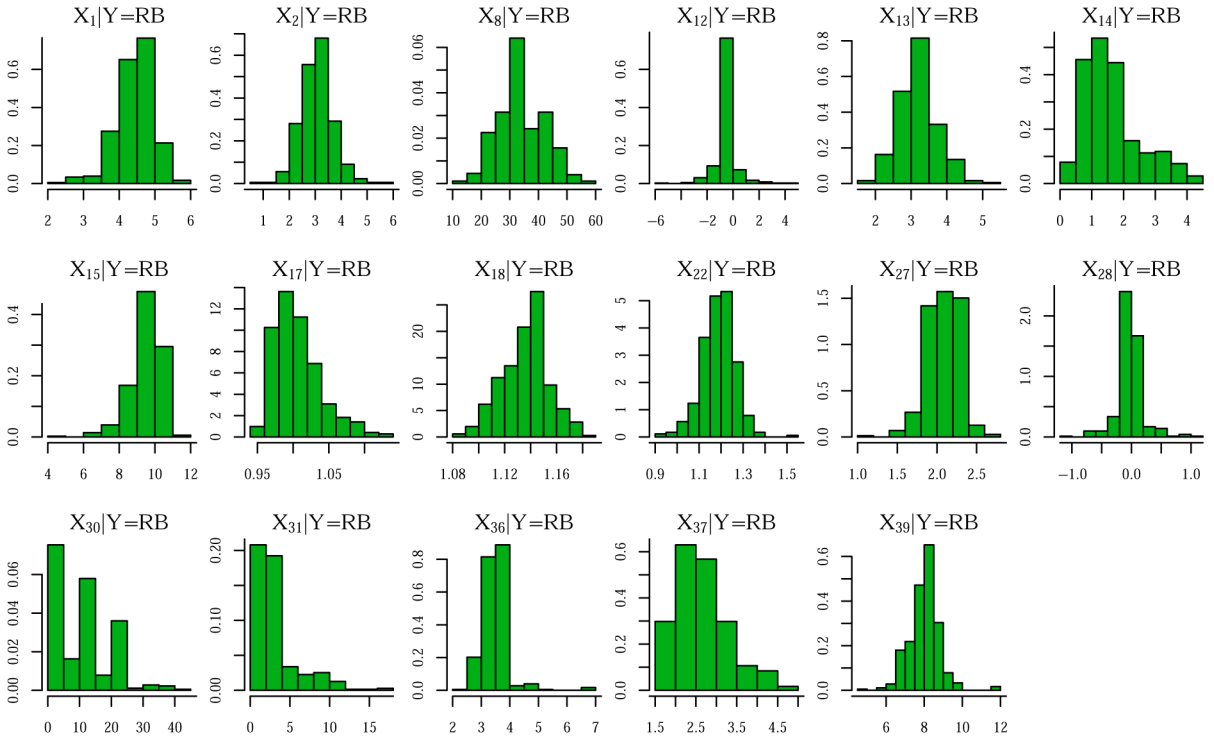

Figura 10: Distribuzioni empiriche delle variabili esplicative, condizionate a $Y = 1$ (RB).

L'ipotesi di normalità è verificata per alcune delle variabili, mentre altre presentano una distribuzione asimmetrica, che sembrerebbe più simile ad una distribuzione Esponenziale o Chi quadro. Mentre le prime variabili che presentano una distribuzione asimmetrica possano comunque essere considerate approssimativamente come delle Normali, la variabile X_{30} ha una distribuzione condizionata troppo irregolare e asimmetrica, quindi si è deciso di non inserirla nei modelli presentati in questa sezione.

L'ipotesi di omoschedasticità non sembra rispettata dalle variabili esplicative condizionate. Per questo motivo, si è effettuata anche un'analisi discriminante quadratica, che non necessita di tale ipotesi. Questo modello predittivo, tuttavia, ha prodotto dei risultati nettamente peggiori rispetto a quelli ottenuti con un'analisi discriminante lineare; quindi nel proseguo delle analisi è stata considerata solo quest'ultima. Il classificatore basato sull'analisi discriminante lineare è costruito in modo tale che l'osservazione $\underline{x}_{i\bullet} = (x_{i,1}, ..., x_{i,41})$ venga assegnata alla classe j se:

$$\mathbb{P}(Y = j|\underline{X}^C = \underline{x}_{i\bullet}) = \max_{h\in\{0,1\}} \mathbb{P}(Y = h|\underline{X}^C = \underline{x}_{i\bullet}) = \max_{h\in\{0,1\}} \frac{\pi_h f_h(\underline{x}_{i\bullet})}{\sum_{u\in\{0,1\}} \pi_u f_u(\underline{x}_{i\bullet})}$$

dove si denota con π_j la probabilità a priori che Y si trovi nella classe j. Quindi si assegna $\underline{x}_{i\bullet}$ alla classe j se

$$\pi_j f_j(\underline{x}_{i\bullet}) = \max_{h\in\{0,1\}} \pi_h f_h(\underline{x}_{i\bullet})$$

Poiché le probabilità a priori π_0 e π_1 sono ignote, esse vengono stimate con la proporzione di osservazioni che assumono ciascuna delle due classi di Y.

5.2 Dataset non bilanciato

5.2.1 Con tutte le variabili continue

Il primo modello di LDA proposto è costruito a partire dalla matrice $\mathbf{X}^C$ dei dati campionari[6] corrispondenti alle realizzazioni del vettore casuale continuo $\underline{X}^C$. Tramite la funzione `lda` su `R` è stato implementato un modello predittivo che prende in input il training-set che considera tutte le variabili contenute in $\mathbf{X}^C$. La *validation* del modello è avvenuta applicando una *Leave One Out Cross Validation*, la quale non ha evidenziato possibili problemi, avendo ottenuto una *Accuracy* di 0.82 e una *AUC* pari a 0.78. Le previsioni effettuate sul test-set hanno ottenuto le seguenti performance: *Accuracy* = 0.831, *AUC* = 0.787.

5.2.2 Con le variabili tenute dal GLM continue

Per semplificare il modello, la matrice di dati $\mathbf{X}^C$ è stata modificata incrociando le informazioni prodotte dal GLM. Infatti, le variabili continue che presentano un coefficiente β non significativo sono state escluse dalla specificazione del modello LDA, che viene costruito dunque sul training-set della nuova matrice $\mathbf{X}^C_{\texttt{glm}}$ che conta in totale 8 variabili[7]. Come nel caso precedente viene applicata una *Leave One Out Cross Validation* che produce una *Accuracy* pari a 0.79 e una *AUC* di 0.743. Questo modello predittivo porta a performance globalmente peggiori del modello precedente: l'*Accuracy* risulta essere pari a 0.825 mentre l'*AUC* è pari a 0.772. La ragione di tale peggioramento nelle prestazioni del classificatore potrebbe essere dovuta all'aver mantenuto un numero troppo basso di variabili.

5.3 Training-set bilanciato

Le modalità della variabile risposta si presentano con frequenze sbilanciate: il 66,3% delle osservazioni sono, infatti, classificate come *Not-Ready-Biodegradable*, mentre il 33,7% sono classificate come *Ready-Biodegradable*; per questo motivo si è ipotizzato che il modello predittivo potesse migliorare se sviluppato sulla base di un training-set bilanciato, in cui la proporzione di osservazioni in ciascuna classe di Y fosse la stessa. È stato dunque costruito un nuovo training-set bilanciato formato dalle 231 osservazioni del training tali che $Y = 1$ (RB) e altrettante tali che $Y = 0$ (NRB), estratte casualmente dal training-set originale.

[6] $\mathbf{X}^C = (\underline{x}_{\bullet 1}, \underline{x}_{\bullet 2}, \underline{x}_{\bullet 8}, \underline{x}_{\bullet 12}, \underline{x}_{\bullet 13}, \underline{x}_{\bullet 14}, \underline{x}_{\bullet 15}, \underline{x}_{\bullet 17}, \underline{x}_{\bullet 18}, \underline{x}_{\bullet 22}, \underline{x}_{\bullet 27}, \underline{x}_{\bullet 28}, \underline{x}_{\bullet 31}, \underline{x}_{\bullet 36}, \underline{x}_{\bullet 37}, \underline{x}_{\bullet 39}) \in \mathbb{R}^{16}$

[7] $\mathbf{X}^C_{\texttt{glm}} = (\underline{x}_{\bullet 1}, \underline{x}_{\bullet 2}, \underline{x}_{\bullet 8}, \underline{x}_{\bullet 12}, \underline{x}_{\bullet 13}, \underline{x}_{\bullet 14}, \underline{x}_{\bullet 15}, \underline{x}_{\bullet 37}) \in \mathbb{R}^8$

5.3.1 Modello semplice

Per quanto riguarda le variabili esplicative, sono state considerate tutte le continue contenute nel dataset originale in seguito ai risultati ottenuti nel paragrafo 5.2.1. Il modello di analisi discriminante lineare applicato sul dataset bilanciato ha prodotto i risultati riportati nella figura 11.

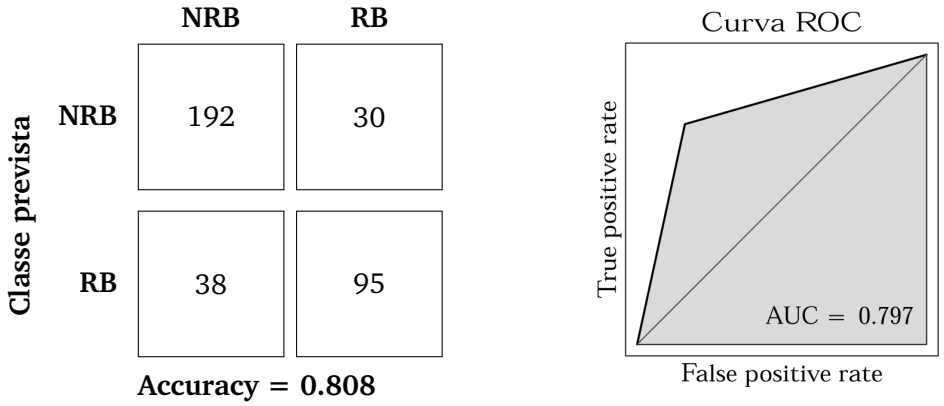

Figura 11: LDA con tutte le variabili continue e dataset bilanciato

L'*AUC* di questo modello è più elevata delle due ottenute precedentemente, il che fa pensare che possa essere stata una buona idea considerare il training-set bilanciato.

5.3.2 Modello migliorato

Per provare a migliorare ulteriormente le performance è stata implementata una funzione che estrae ricorsivamente il training set, lasciando invariati in esso i dati con $Y = 1$ e variando di volta in volta quelli con $Y = 0$. Si effettua per ciascuna estrazione, e quindi per ciascun training-set, una *LDA* e si classifica un'osservazione come biodegradabile o non biodegradabile in base alla classe cui è stata assegnata il maggior numero di volte. Nel caso in cui un'osservazione sia stata assegnata lo stesso numero di volte ad entrambe le classi, essa verrà assegnata in modo casuale. Il test set rimane invariato, così da poterlo confrontare con i risultati precedenti. Ci si aspetta che tale risultato non solo sia migliore, in quanto tale modello è in grado di sfruttare anche i dati che erano stati eliminati nella fase di bilanciamento, ma che sia anche più affidabile in quanto il buon risultato ottenuto avrebbe potuto essere attribuito all'estrazione fortuita di un training-set adatto per prevedere il test-set selezionato. Con questo metodo sono stati ottenuti i risultati migliori basati sulla *LDA* in termini di *AUC* – misura ritenuta più attendibile – come possiamo osservare dalla figura 12.

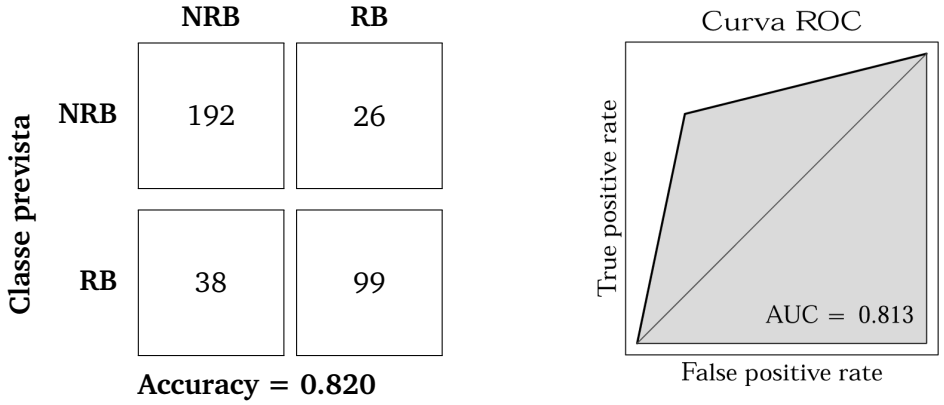

Figura 12: LDA con tutte le variabili continue e dataset bilanciato (modello migliorato)

6 MODELLI DI PREVISIONE CON ALBERI DI CLASSIFICAZIONE

6.1 Teoria generale

In questa sezione si propongono diversi modelli di previsione basati sugli alberi di classificazione. Si consideri la matrice dei dati $\mathbf{X} = (\underline{x}_{\bullet 1}, ..., \underline{x}_{\bullet 41})$; i modelli vengono costruiti sulla base di condizioni binarie poste sulle variabili casuali X_l $(l = 1, ..., 41)$. Per far crescere l'albero ad ogni passo dell'algoritmo si divide lo spazio $\mathbb{R}^{41}$ generato da $\underline{X}$ in due ulteriori regioni sulla base di una X_l e di un punto di suddivisione c tali che $R_1 = \{X|X_l < c\}$ e $R_2 = \{X|X_l \geq c\}$. X_l e c sono scelti in modo che venga massimizzata la quantità:

$$i(R) - p_1 i(R_1) - p_2 i(R_2)$$

dove p_k è la proporzione di osservazioni che appartengono alla regione k, mentre $i(R_k)$ è l'impurità della regione R_k, calcolata a partire dall'indice di Gini:

$$i(R_k) = \sum_{j \in \{0,1\}} \hat{p}_{kj}(1 - \hat{p}_{kj}) \qquad \text{con } \hat{p}_{kj} = \frac{1}{|R_k|} \sum_{i | x_{i\bullet} \in R_k} I(y_i = j)$$

Dopo aver fatto crescere l'albero è necessario che esso venga potato[8]. Infatti un albero con troppe foglie rischia di adattarsi eccessivamente al training-set sul quale viene implementato (*overfitting*).

6.2 Albero di classificazione con tutte le variabili

Il primo albero di classificazione proposto viene costruito con la funzione `tree` sulla matrice del disegno $\mathbf{X}$ nella sua totalità, senza escludere alcuna variabile. La *Leave One Out CV* suggerisce di settare il numero di foglie totali pari a 16; l'albero viene dunque potato con la funzione `prune.tree` settando `best` = 16. Le previsioni sul test-set portano alle seguenti performance: *Accuracy* = 0.808, *AUC* = 0.805, piuttosto buone considerando che si tratta di un modello base e costruito su tutte le variabili.

6.3 Albero di classificazione con le variabili selezionate dal GLM

Questo modello è perfettamente analogo al precedente, salvo per il fatto che il training-set proviene dalla matrice $\mathbf{X}_{\texttt{glm}}$ [9]. Esattamente come prima, la *LOOCV* suggerisce di potare a 16 foglie; questo classificatore applicato al test-set ha un'*Accuracy* di 0.839 e una *AUC* pari a 0.792.

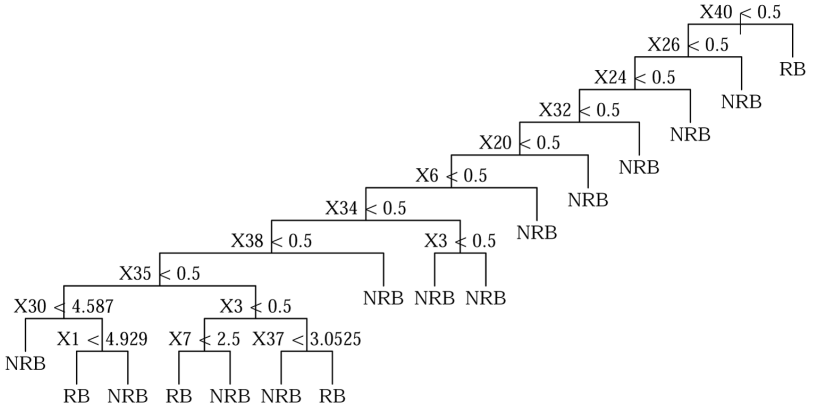

Figura 13: Albero con variabili glm.

Da questo grafico si nota come le variabili che determinano le prime suddivisioni sono quelle che nei *barplot* risultavano avere tutti gli *outlier* classificati come *NRB*.

[8] La potatura è effettuata tramite una *Leave One Out Cross Validation* che permette di individuare il migliore parametro di complessità α; maggiore è α minore sarà la dimensione dell'albero stimato.

[9] $\mathbf{X}_{\texttt{glm}} = (\underline{x}_{\bullet 1}, \underline{x}_{\bullet 2}, \underline{x}_{\bullet 3}, \underline{x}_{\bullet 6}, \underline{x}_{\bullet 7}, \underline{x}_{\bullet 8}, \underline{x}_{\bullet 10}, \underline{x}_{\bullet 12}, \underline{x}_{\bullet 13}, \underline{x}_{\bullet 14}, \underline{x}_{\bullet 15}, \underline{x}_{\bullet 20}, \underline{x}_{\bullet 24}, \underline{x}_{\bullet 26}, \underline{x}_{\bullet 30}, \underline{x}_{\bullet 32}, \underline{x}_{\bullet 34}, \underline{x}_{\bullet 35}, \underline{x}_{\bullet 37}, \underline{x}_{\bullet 38}, \underline{x}_{\bullet 40})$

6.4 Alberi di classificazione con tutte le variabili - Algoritmo *bagging*

Un potenziamento del modello presentato nella sezione 6.2 si ottiene applicando l'algoritmo cosiddetto *bagging* (bootstrap aggregating)[4]. Esso richiede soltanto di fissare due parametri: θ e N e funziona nel seguente modo.

1. Dal training-set viene estratto casualmente un sottoinsieme di osservazioni con reinserimento di cardinalità pari alla numerosità del training di partenza moltiplicata per $\theta = 0.5$;
2. Si stima un albero sulla base dei dati contenuti nel sottoinsieme, senza praticare alcuna potatura;
3. Si esegue la previsione sui dati del test-set;
4. Eseguendo i primi 3 passi per $N = 500$ volte, si ottengono N previsioni dello stesso test-set; si salva dunque per ogni osservazione la classe predetta il maggior numero di volte.

Dunque l'output di questo algoritmo è un vettore di previsioni, con il quale è possibile costruire matrice di confusione e curva ROC. Questa tecnica viene utilizzata per ottenere stime più robuste e per evitare *overfitting*: la varianza della previsione infatti diminuisce. Inoltre, a fronte della riduzione della varianza totale è possibile non potare i singoli alberi di classificazione, scelta che porta solitamente ad aumenti di variabilità, compensati tuttavia da una dimensione sufficientemente elevata di N. Le performance – riassunte in Figura 14 – registrano, come ci si aspettava, un miglioramento.

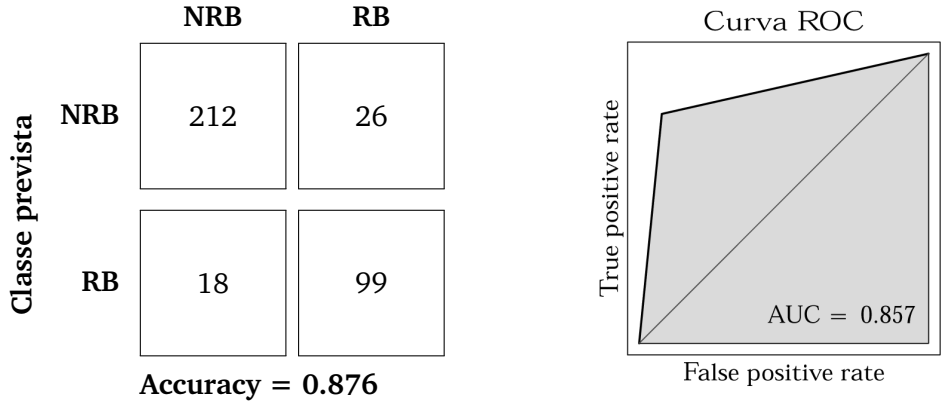

Figura 14: Albero di classificazione con tutte le esplicative + *Bagging*

6.5 Alberi di classificazione con variabili selezionate dal GLM - Algoritmo *bagging*

L'algoritmo *Bagging* presentato nel paragrafo precedente viene poi applicato sulla matrice di dati $\mathbf{X}_{\texttt{glm}}$. I risultati, che sono i migliori finora ottenuti, sono riassunti in Figura 15.

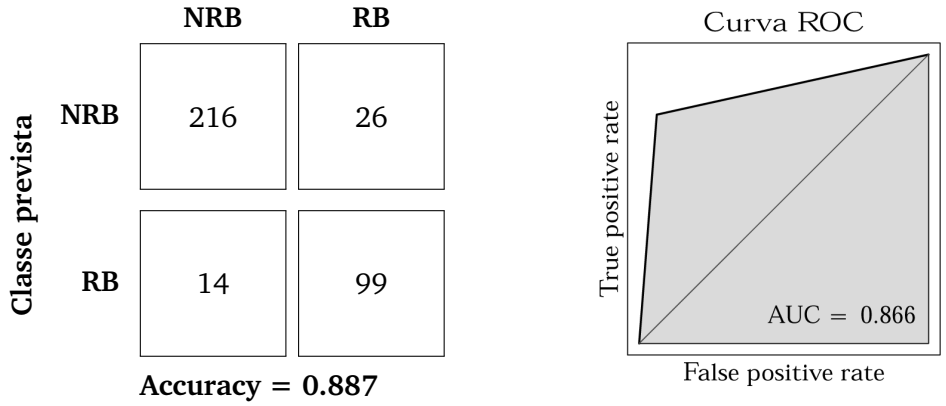

Figura 15: Albero di classificazione con le variabili selezionate dal GLM + *Bagging*

6.5.1 Training-set bilanciato

Modificando leggermente l'algoritmo *bagging* è possibile costruire N sotto-training bilanciati (in cui il numero di osservazioni in ciascuna classe è il medesimo). Le performance tuttavia risultano inferiori sia utilizzando la matrice di dati $\mathbf{X}$ (*Accuracy* = 0.811, *AUC* = 0.830) sia considerando la matrice ridotta $\mathbf{X}_{\texttt{glm}}$ (*Accuracy* = 0.848, *AUC* = 0.853).

7 CONCLUSIONI

In conclusione si riporta una sintesi delle performance dei diversi classificatori.

Classificatore	**Accuracy**	**AUC**
GLM	0.882	0.863
LDA con $\mathbf{X}^C$	0.831	0.787
LDA con $\mathbf{X}^C_{\texttt{glm}}$	0.825	0.772
LDA con $\mathbf{X}^C$ con training-set bilanciato	0.808	0.797
LDA con $\mathbf{X}^C$ con training-set bilanciato e modello migliorato	0.820	0.813
ADC con $\mathbf{X}$	0.808	0.805
ADC con $\mathbf{X}_{\texttt{glm}}$	0.839	0.792
ADC con $\mathbf{X}$ + *bagging*	0.876	0.857
ADC con $\mathbf{X}_{\texttt{glm}}$ + *bagging*	**0.887**	**0.866**
ADC con $\mathbf{X}$ training-set bilanciato + *bagging*	0.811	0.830
ADC con $\mathbf{X}_{\texttt{glm}}$ training-set bilanciato + *bagging*	0.848	0.853

Figura 16: Sintesi delle performance dei classificatori.

Si può quindi concludere che il modello di previsione migliore, sia in termini di *Accuracy* che di *AUC*, sia l'albero di classificazione costruito utilizzando le variabili selezionate dal *GLM*, che a sua volta è un buon classificatore. Era prevedibile il fatto che l'analisi discriminante lineare fosse meno efficace data la presenza di numerose variabili esplicative discrete che in tale modello non possono essere utilizzate.

RIFERIMENTI BIBLIOGRAFICI

[1] *European Chemical Agency (ECHA)*. URL: https://echa.europa.eu/it/regulations/reach/understanding-reach.

[2] ECHA. *Uso di alternative alla sperimentazione sugli animali per adempiere le prescrizioni in materia di informazione relative alla registrazione ai sensi del regolamento REACH*. 2016. URL: http://www.reach.gov.it/sites/default/files/allegati/Guida%20pratica%20alla%20sperimentazione%20sugli%20animali_0.pdf.

[3] K. R Mansouri et al. «Quantitative Structure - Activity Relationship models for ready biodegradability of chemicals». In: *Journal of Chemical Information and Modeling* (2013), pp. 867–878.

[4] Leo Breiman. «Bagging predictors». In: *Machine learning* 24.2 (1996), pp. 123–140.