Obesity level estimation based on machine learning algorithms

Millone A. (mat. 846588), Rossi S. (mat. 857183), Università degli Studi di Milano-Bicocca,
Laurea Magistrale in Biostatistica, Machine Learning

Introduzione

- Coorte di 2111 pazienti provenienti dallo stato del Messico;
- Dataset con variabili relative a caratteristiche fisiche del paziente, altre storie di obesità in famiglia e abitudine fisiche ed alimentari del soggetto;
- Il 33% dei dati è ottenuto da questionari sottoposti a pazienti reali, mentre il restante 66% è stato generato da una simulazione con lo strumento Weka (learning models) e il filtro SMOTE (che permette operazioni di oversampling)
- **Obiettivo**: trovare il modello che classifica meglio i pazienti obesi in base alle variabili in studio e successivamente valutare le sue performance come modello previsionale per nuovi pazienti

Pillole sull'obesità

- In Italia si stima che il 10-11% della popolazione sperimenti l'obesità
- Vi è una differenza sostanziale tra donne e uomini e tra fasce d'età
- Aumento della probabilità di sviluppare patologie quali diabete, malattie cardiovascolari e tumori

Variabili in esame

- Variabile target: Nobesity, categoriale a 7 livelli: Peso Insufficiente, Peso Normale, Sovrappeso Livello I, II e III, Obesità di tipo I, II e III
- Le variabili sono divise principalmente in 4 macrogruppi:
  - Abitudini alimentari (numero di pasti, consumo di verdura, alcool, livello di idratazione)
  - Abitudini fisiche (attività motoria, fumo, utilizzo dispositivi tecnologici, tipo di trasporto)
  - Caratteristiche fisiche (peso, altezza, età)
  - Suscettibilità famigliare della condizione di salute in studio

Disegno dello studio

PROCEDURE PRELIMINARI

- Missing data
- Ricodifica ed esclusione di alcune variabili

DESCRITTIVE E PREPROCESSING

- Statistiche descrittive
- Correlazione, near zero-variance e model selection

FITTING DEI MODELLI SUL DATASET DI TRAINING

- Stima dei modelli di machine learning
- Valutazione metriche su dataset di training

VALUTAZIONE DEI MODELLI MIGLIORI SUL DATASET DI TEST

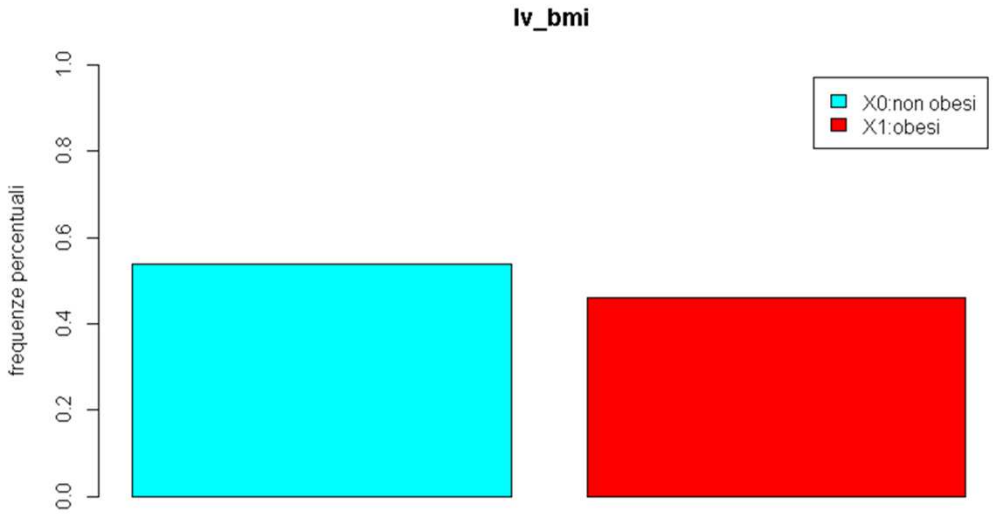
- Analisi delle curve ROC,Lift e overfitting
- Scelta treshold
- Score su nuovi dati

Procedure preliminari

- Controllo dei missing data: nessuna variabile presenta valori mancanti o nulli;
- Ricodifica di alcune variabili da «character» a «factor»
- Eliminazione variabili quali il peso e altezza poiché troppo correlate con la variabile target
- Binarizzazione della variabile target «obesità»:
  - Livello 0: pazienti sottopeso, normopeso e sovrappeso
  - Livello 1: pazienti obesi
- 2 vantaggi:
  - La variabile target è bilanciata (54% «livello 0» e restante 46% «livello 1»)
  - La binarizzazione risponde al quesito clinico che si vuole indagare in questa analisi

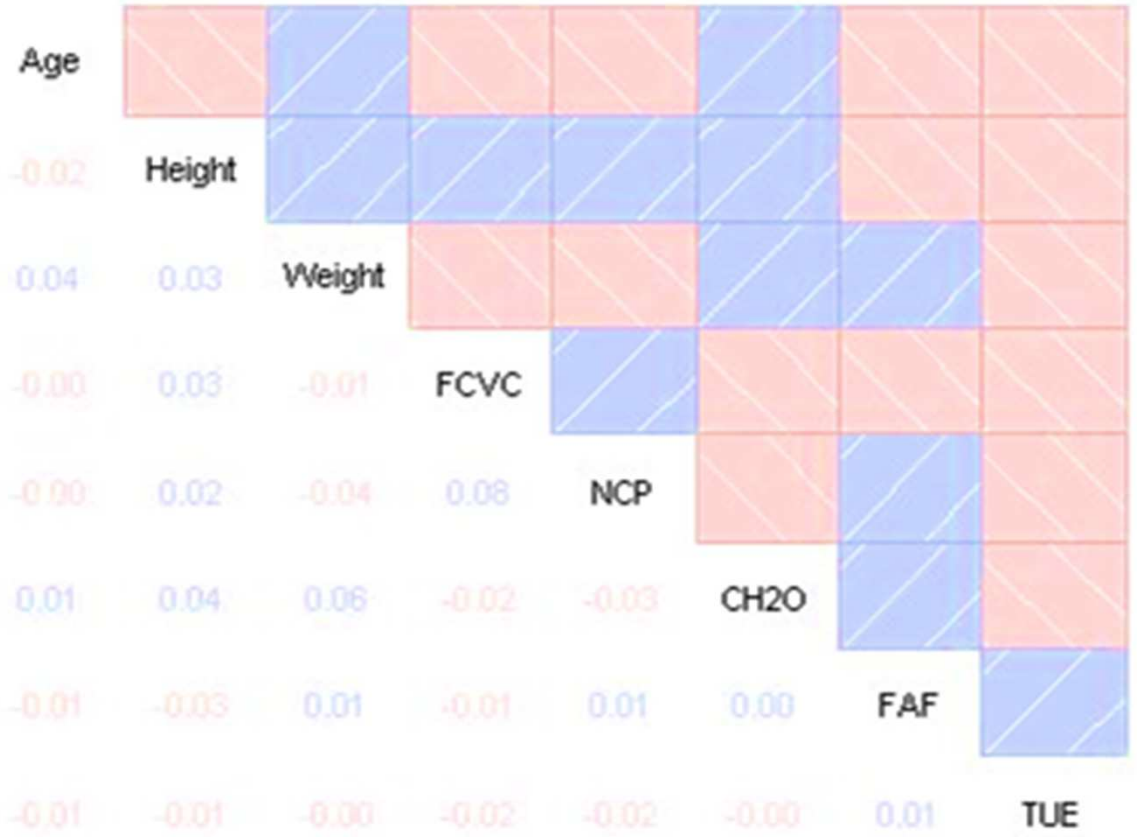
Statistiche descrittive

	Media	1° quart.	Mediana	3° quart.	sd
Age	24.391	19.12	22.185	26.00	19.830
Height	13.598	1.633	1.710	1.786	43.286
Weight	84.21	60.00	80.726	105.03	53.328
FCVC	11.439	2.000	2.497	3.000	45.186
NCP	9.407	2.764	3.000	3.000	42.308
CH20	11.850	1.634	2.000	2.619	44.966
FAF	6.557	0.124	1.000	1.800	30.530
TUE	3.036	0.000	0.625	1.000	19.447



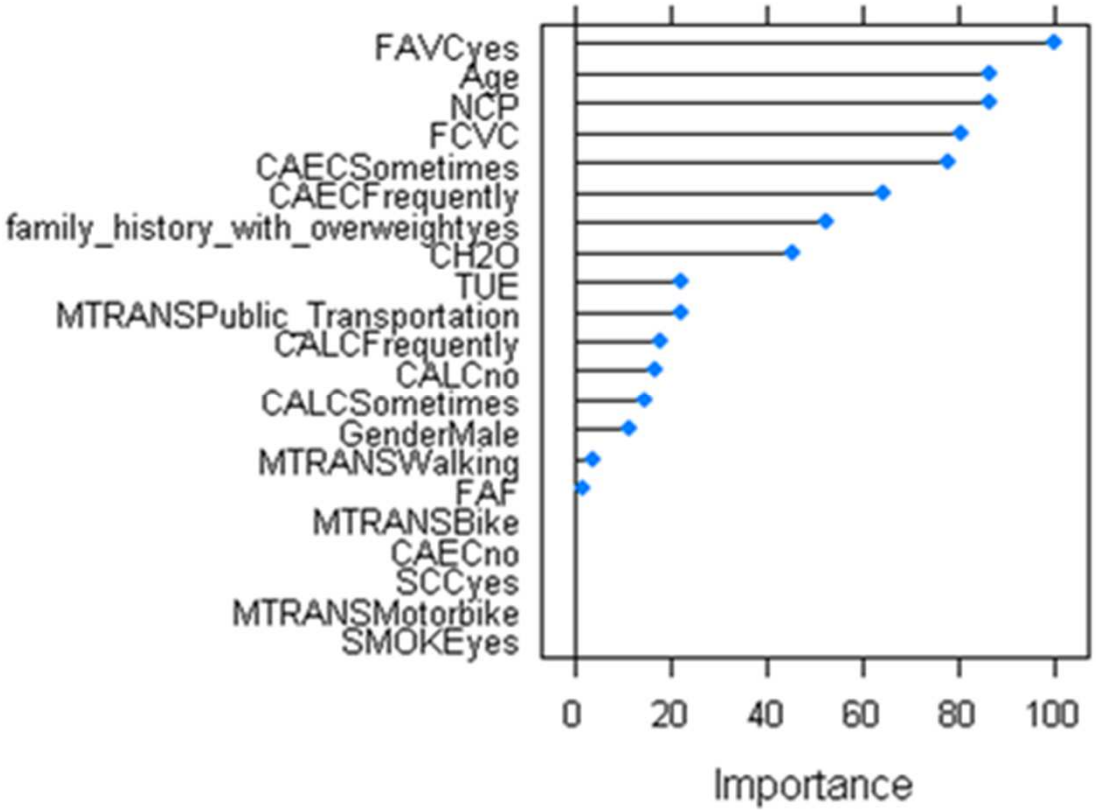
- Frequenza di consumo di verdure (FCVC), Numero di pasti principali (NCP), Frequenza dell'attività fisica (FAF), Consumo di acqua giornaliero (CH20), Tempo di utilizzo di dispositivi tecnologici (TUE)
- Distribuzione del sesso equilibrata
- L'82% dei pazienti ha famigliarità con lo stato di salute «obeso»
- Il 74% usa i trasporti pubblici, il 21% ha l'auto e solo il 5% camminano

Correlazione



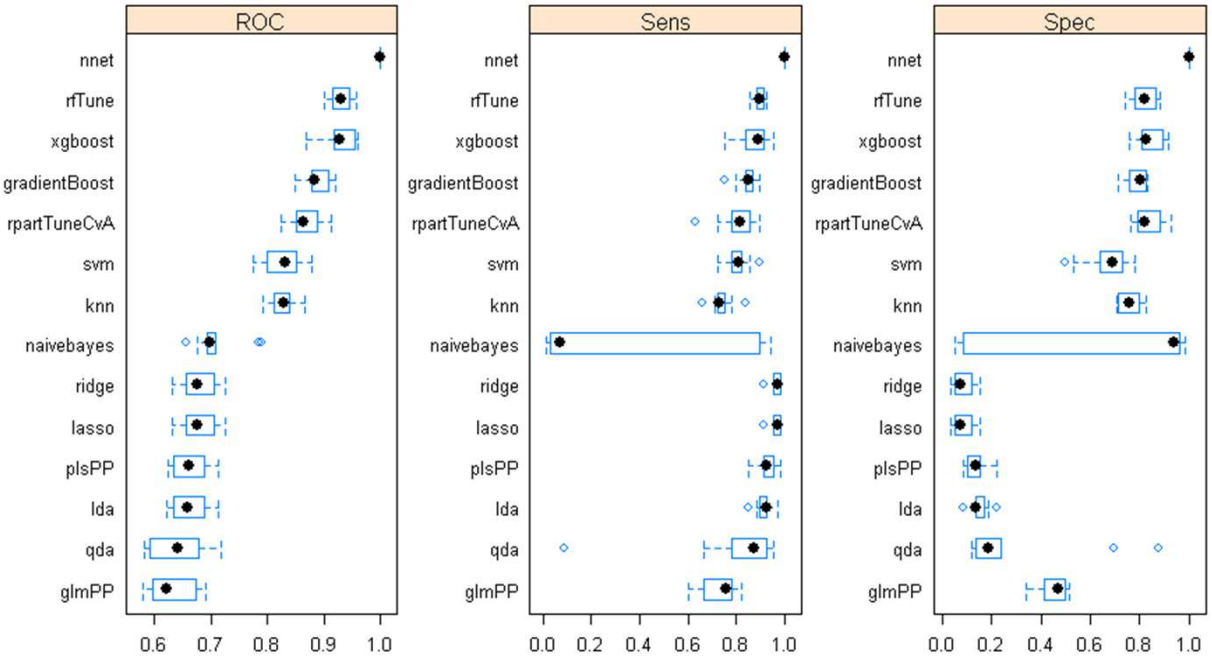
- Assenza di variabili collineari

Model selection



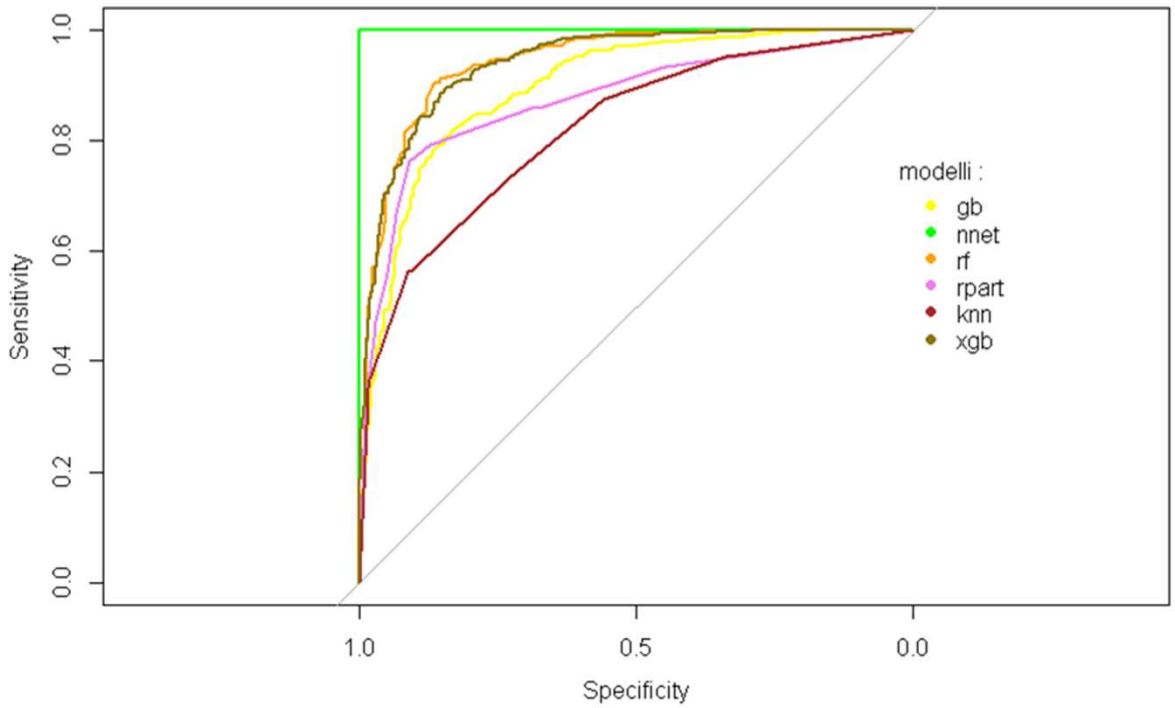
- Selezione delle variabili tramite albero

Fitting dei modelli



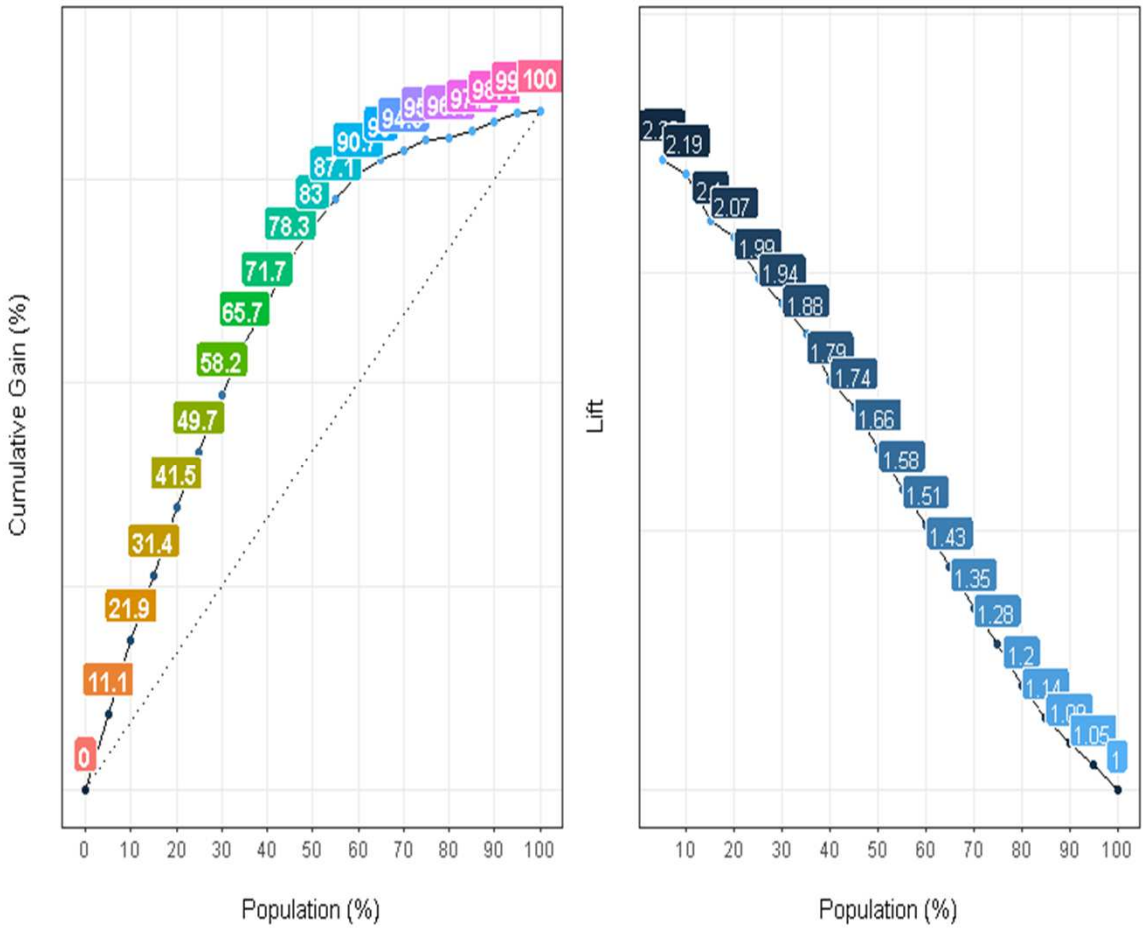
- Ogni algoritmo è stato sviluppato con la tecnica *Cross Validation a 10 Fold*
- Per ogni modello abbiamo svolto il *tuning dei parametri*
- I modelli più performanti (per le metriche: ROC, sensitivity e specificity) sono:
  - *Reti neurali* (potenziale overfitting)
  - Famiglia dei modelli degli alberi: *random forest, gradient boosting* e *xgboost*

Confronto Curve ROC



- Sono calcolate sul dataset di test
- Si conferma il problema dell'overfitting nelle *reti neurali*
- I modelli *random forest, xgboost e gradient boosting* risultano i migliori

Curve Lift dei modelli

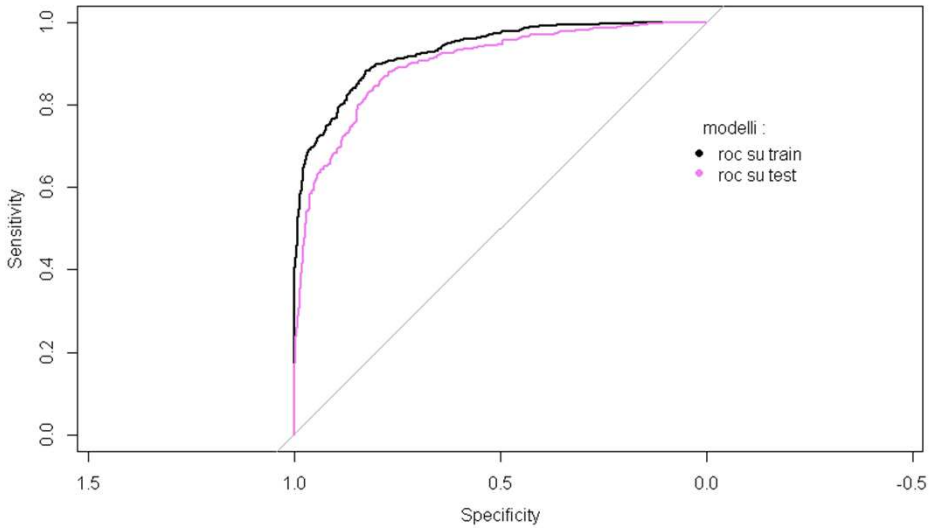


Curve Lift del Gradient boosting

Modello	Popolazione	Gain	Score point
XGboost	20/100	42,53	0,923
Random forest	20/100	43,30	0,850
Gradient boosting	20/100	41,49	0,803

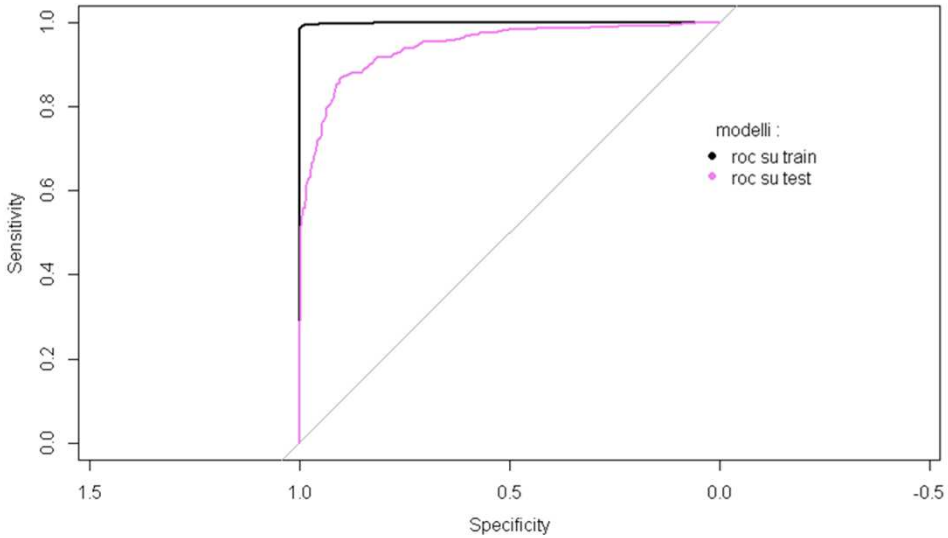
- Le Curve Lift indicano la percentuale di corretti obesi (evento principale dello studio in generale) per ogni porzione di popolazione scelta.
- Tutti i modelli al 20% della popolazione riescono a catturare almeno il 40% della popolazione d'interesse (pazienti obesi)

Differenze curve ROC tra train e test

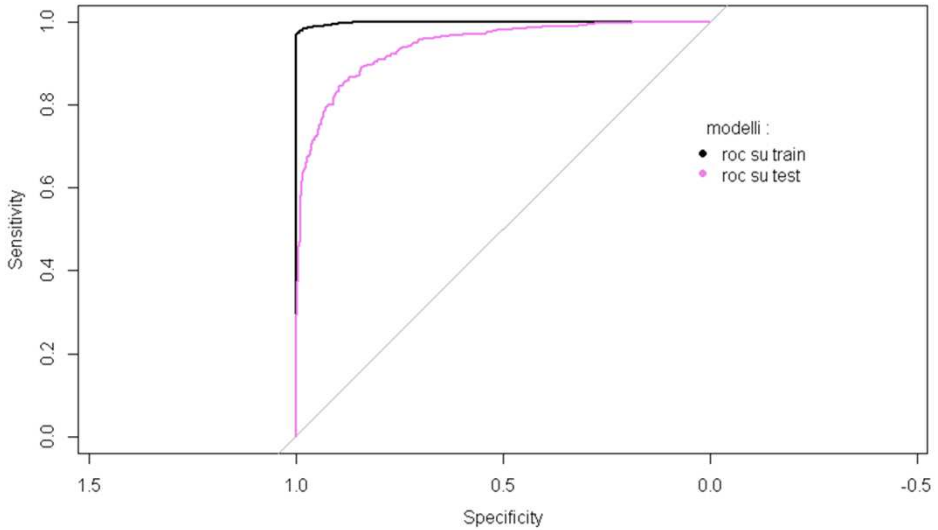


Gradient boosting

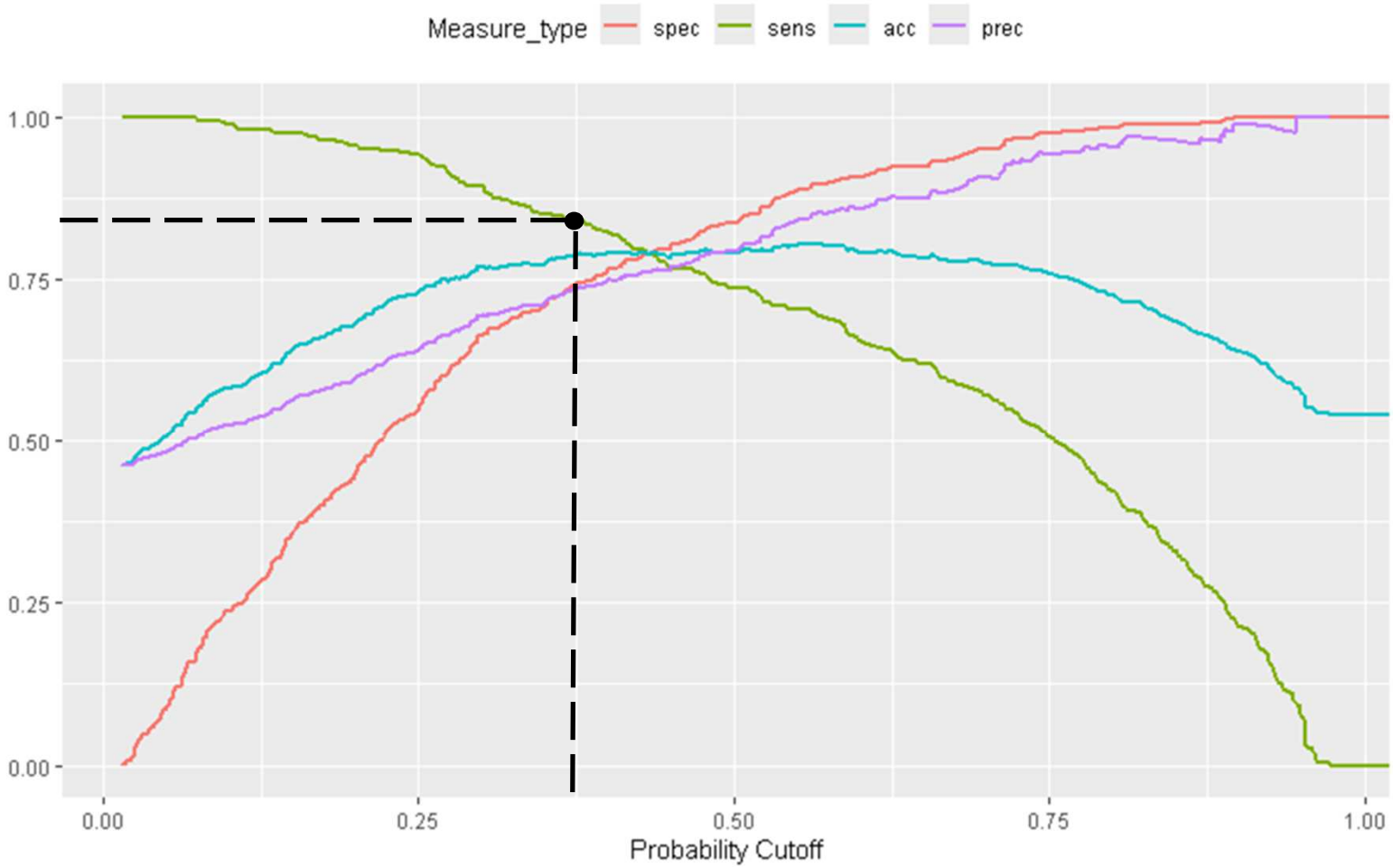
Random forest



Xgboost

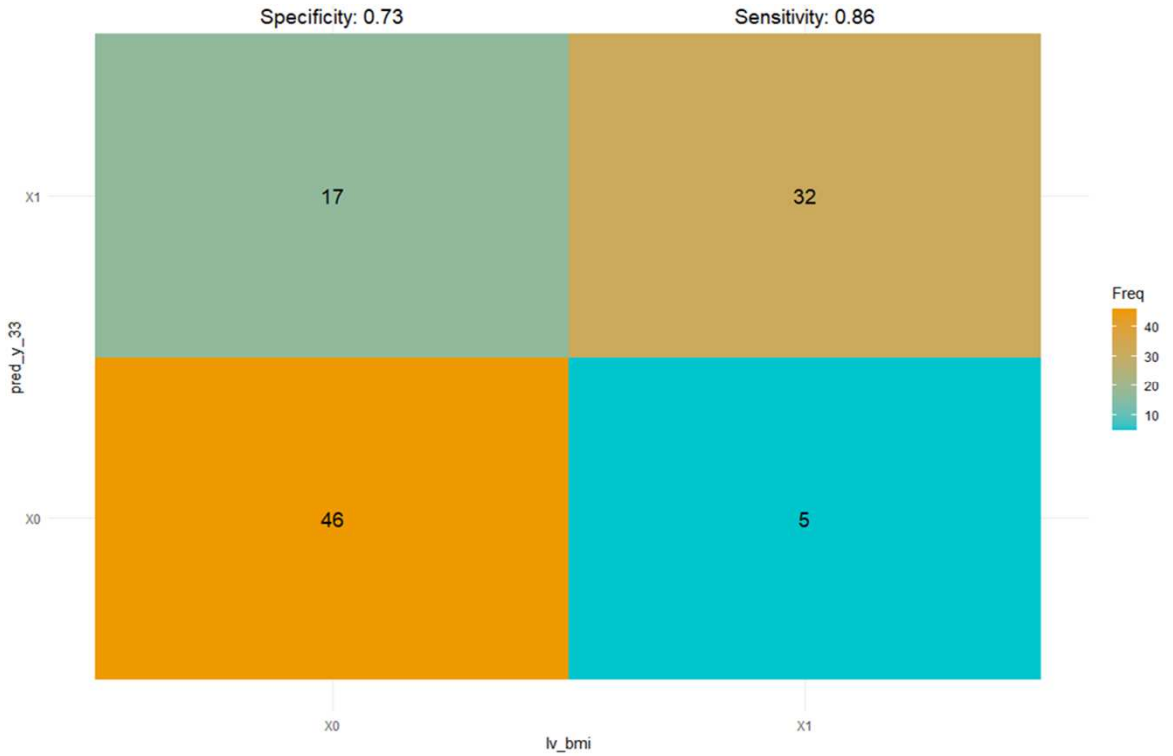


Scelta treshold



- Si sceglie come soglia 0.375, ciò garantisce una sensitivity elevata (all'incirca 0.86) e una buona specificity (all'incirca 0.75)

Score su nuovi dati



- Dataset dei nuovi dati contenente 100 osservazioni
- I valori delle metriche di sensitivity e specificity rimangono simili a quelli mostrati in precedenza (sensitivity:86% e specificty:73%)
- Ulteriore conferma della bontà del modello

Target con 4 e 7 livelli

Modello	Valore AUC (target 7 lv.)	Valore AUC (target 4 lv.)
Reti neurali	100%	100%
Random forest	96%	99,99%
Gradient boosting	94%	99,63%
Knn	81%	66,82%

- Variabile target con tutti e 7 i livelli:
  - *Sottopeso*
  - *Normopeso*
  - *Sovrappeso* I livello, II livello
  - *Obeso* I livello, II livello e III livello
- Variabile target con 4 livelli:
  - *Sottopeso*
  - *Normopeso*
  - *Sovrappeso*
  - *Obeso*
- Impossibilità a calcolare le curve ROC
- I modelli migliori (secondo la metrica AUC) risultano gli stessi di quelli ottenuti con la target binarizzata

Criticità incontrate e potenzialità future

Il modello finale *gradient boosting* garantisce una buona generalizzabilità e una buona replicabilità

PROS

Overfitting nel modello solitamente migliore *(neural network)*, dovuto ai pochi dati presenti

CONS

Valutazione accurata della treshold per garantire risultati buoni nella metrica di maggiore interesse

PROS

66% dati simulati, alcuni valori poco attendibili

CONS

Il modello finale *gradient boosting* ha tempi computazionali ristretti, garantisce risparmio di tempo e risorse

PROS

La target è stata binarizzata: potenziale perdita di informazione

CONS