Modelli per la classificazione di soggetti fumatori

Giulia de Innocentiis mat. 864084 Giorgia Faccanoni mat. 871869

23 Novembre 2022

Sommario

Nel presente articolo si è affrontata un'analisi di classificazione con l'obiettivo di individuare, in base a dei valori biologici, quali tra i soggetti fruitori del sistema sanitario coreano NHIS sono fumatori. Sono stati presentati come possibili soluzioni al problema diversi metodi di classificazione di data mining come la regressione logistica, il K-nearest neighbours e il Random Forest, individuando dopo diverse analisi come modello migliore il Random Forest.

1 Introduzione

Il fumo è un'abitudine comune a molte persone, tuttavia esso è una delle principali cause di sviluppo di diverse malattie più o meno gravi. Dopo l'età, il fumo è il fattore di rischio più importante per le malattie cardiovascolari. La speranza di vita di un fumatore è otto anni inferiore a quella di un non fumatore, inoltre chi fuma ha una probabilità doppia di essere colpito da infarto rispetto a chi non fuma e ha una probabilità dieci volte superiore di essere colpito da cancro ai polmoni e alla laringe.

Per le organizzazioni che gestiscono i sistemi sanitari, diventa quindi importante conoscere lo stato di fumatore di una persona poichè questo potrebbe portare a un aumento delle spese sanitarie. Un esempio di questo problema è stato presentato nel 2014 dal National Health Insurance Service, il sistema unico di assicurazione nazionale in Corea del Sud. L'organizzazione ha iniziato una causa legale contro KT&G, il principale produttore di tabacco coreano, chiedendo un risarcimento di 53,3 miliardi di won per le spese sostenute per il trattamento di pazienti affetti da cancro ai polmoni e alla laringe. Di conseguenza, il NHIS ha raccolto diverse informazioni sui loro clienti tra cui il loro stato di fumatore. L'analisi da noi svolta mira a fornire un modello di classificazione in grado di prevedere la condizione di fumatore di nuovi soggetti di cui si conoscono determinati valori biologici, grazie al quale il NHIS sarà in grado di quantificare il numero di fumatori tra i loro clienti. Per costruire questo modello abbiamo utilizzato i principali metodi di classificazione, ovvero la regressione logistica, il K-Nearest Neighbours e il Random Forest, ottenendo un modello più efficiente con il Random Forest.

Nella sezione Metodi e Materiali verrà presentato il dataset e i metodi utilizzati per svolgere l'analisi. Nella sezione Risultati verrano esposti i principali risultati ottenuti dalle nostre analisi mentre nell'ultima sezione Discussioni verrano riportate le considerazioni che hanno portato alla scelta del modello migliore.

2 Materiali e metodi

Questa sezione è dedicata a una descrizione dei materiali e a un'analisi dei metodi utilizzati per la classificazione delle osservazioni del dataset.

2.1 Materiali

Il dataset contiene informazioni riguardanti le persone a carico del NHIS raccolte dal sistema sanitario nazionale coreano nel 2020 tramite dei controlli sanitari generali. Il dataset ha come variabile risposta la variabile smoking che assume valore 0 nel caso di persona non fumatrice e valore 1 nel caso di persona

fumatrice. Esso contiene 55692 osservazioni, di cui il 36,72% appartenenti alla classe dei fumatori. Per ogni osservazione sono state rilevate 27 variabili:

1. ID: codice identificativo del paziente;
2. gender: sesso del paziente; assume valori F (donna) e M (uomo);
3. age: età del paziente; assume valori tra 20 e 80 ad intervalli di 5;
4. height.cm: altezza del paziente in centimetri;
5. weight.kg : peso del paziente in kilogrammi;
6. waist.cm: circonferenza della vita in centimetri;
7. eyesight.left: diottrie mancanti per l'occhio sinistro; assume valori tra 0 e 10;
8. eyesight.right: diottrie mancanti per l'occhio destro; assume valori tra 0 e 10;
9. hearing.left: condizione dell'udito dell'orecchio sinistro; assume valori 1 corrispondente a normalità e 0 in corrispondenza di sospetta malattia;
10. hearing.right: condizione dell'udito dell'orecchio destro; assume valori 1 in corrispondenza di normalità e 0 in corrispondenza di sospetta malattia;
11. systolic: indica il valore di pressione arteriosa massima, ovvero nel momento in cui il cuore è in fase di contrazione;
12. relaxation: indica il valore di pressione arteriosa minima, ovvero nel momento in cui il cuore è in fase di rilassamento;
13. fasting blood sugar: indica la glicemia, ovvero la concentrazione di glucosio nel sangue;
14. Cholesterol: esprime il livello di colesterolo totale, un lipide presente anche nel sangue proveniente sia dal fegato sia dalla dieta alimentare;
15. triglyceride: esprime la quantità di trigliceridi, lipidi presenti nel sangue provenienti per la maggior parte dalla dieta alimentare;
16. HDL: lipoproteine ad alta intensità, che trasportano il colesterolo dalle periferie al fegato riducendone il deposito nelle arterie;
17. LDL: lipoproteine a basse intesità, che trasportano il colesterolo dal fegato alle cellule del corpo;
18. hemoglobin: livello di emoglobina, una proteina contenuta nei globuli rossi che si occupa di trasportare l'ossigeno del sangue;
19. Urine protein: livello delle proteine presenti nelle urine, rilevato tramite il dipstick test. Assume valori tra 1 e 6;
20. serum creatinine: livello di creatinina, una proteina presente in abbondanza nei muscoli che ha il compito di conservare l'energia chimica delle cellule muscolari;
21. AST: livello di transaminasi glutammico-ossalacetica, un'enzima misurato per valutare la salute del fegato, cuore, muscoli e reni;
22. ALT: livello di alanina amino transferasi, un'enzima misurato per valutare la salute del fegato;
23. y-Gtp: livello di gamma glutammiltransferasi, un'enzima estremamente sensibile ad anomalie del fegato e ai dotti biliari;

24. oral: indica se il paziente ha effettuato una visita dentistica generale. Assume valori 0(no) e 1(si);

25. tartar: indica lo stato del tartaro al momento della visita. Assume valori "Yes" o "No";

26. dental caries: indica se il paziente presenta delle carie al momento della visita. Assume valori 0(no) e 1(si);

27. smoking: indica se il paziente è un fumatore(1) o no (0). Lo stato di fumatore non include le sigarette elettroniche.

A causa delle sostanze presenti nel fumo di sigaretta, molte di queste variabili subiscono variazioni in corrispondenza di un soggetto fumatore. Ad esempio la nicotina stimola il corpo a produrre adrenalina, che rende il battito cardiaco più veloce, costringendo il cuore a un maggior lavoro e favorendo la formazione di coaguli nel sangue (trombosi). Per questo motivo, ci aspettiamo quindi un'innalzamento della pressione sanguigna. Inoltre, come viene confermato dalla Figura 2, nei fumatori il valore di emoglobina è più elevato, poichè producono un numero maggiore di globuli rossi. Questo è dovuto al fatto che i globuli rossi si legano al monossido di carbonio (contenuto nel fumo di sigaretta) il quale è più affine al gruppo EME localizzato nell'emoglobina. Per compensare questo fenomeno è necessario una produzione più elevata di globuli rossi in modo tale che parte dell'emoglobina possa comunque legarsi all'ossigeno. In aggiunta, come dimostrato in Figura 1, il colesterolo considerato "buono" HDL è inferiore nei soggetti fumatori quindi il rischio di malattie cardiovarcolari è più alto. Ci attendiamo quindi che i modelli da noi trattati considerino queste variabili come significative.

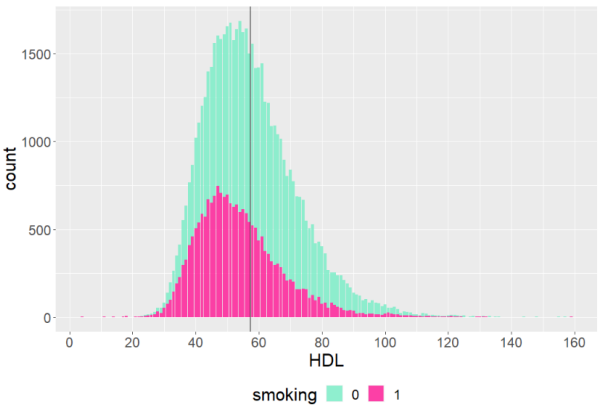

Figura 1: Distribuzione di HDL rispetto alla classe

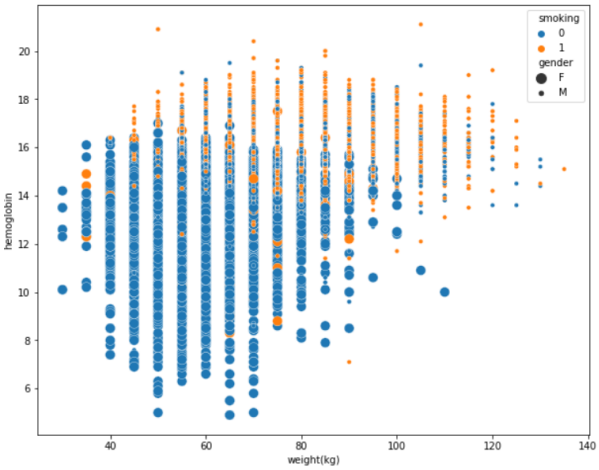

Figura 2: Scatterplot delle variabili hemoglobin e weight classificate per gender e smoking

2.2 Metodi

In quest'analisi sono stati utilizzati dei modelli di data mining supervisionati, ossia algoritmi di apprendimento che, attraverso dati etichettati, sono in grado di classificare e fare previsione. Per poter utilizzare questo tipo di modelli il dataset deve contenere sia la variabile risposta dipendente che le variabili esplicative indipendenti. La tecnica supervisionata cerca di indentificare le relazioni tra variabili indipendenti e dipendenti e di costruire un modello che mostri queste dipendenze. Il modello stimato viene quindi applicato ai dati per i quali il valore target è sconosciuto. I metodi utilizzati sono stati:

1. Regressione logistica

   Il modello di regressione logistica è un modello utilizzato per descrivere la dipendenza tra una variabile dipendente dicotomica che può assumere valori 0 e 1, corrispondenti all'assenza e alla presenza di un attributo, e una o più variabili indipendenti di qualsiasi natura. La regressione

logistica è utilizzata per stimare la probabilità di possesso dell'attributo di un'osservazione e, fissato un valore soglia per tale probabilità, classificare l'osservazione alla classe percui la probabilità stimata è sopra la soglia decisa. La probabilità a posteriori stimata dal modello è data da:

$$\pi(x) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_j}}$$

Il vantaggio di questo metodo è che produce una formula semplice di classificazione e permette di comprendere quali sono le variabili indipendenti significative per la determinazione del possesso dell'attributo o meno. Lo svantaggio è che non riesce ad affrontare adeguatamente problemi in cui le variabili indipendenti hanno effetti non lineari.

Il modello di regressione logistica è un modello molto simile al modello di regressione lineare ma si distingue da esso per la distribuzione degli errori; infatti, nella regressione logistica gli errori non seguono una distribuzione normale ma una distribuzione bernoulliana.

2. K-nearest neighbours

   Si tratta di un algoritmo che si basa sulla distanza tra le osservazioni del dataset, non fa alcuna assunzione sulle variabili indipendenti e non è in grado di dimostrare quali features sono significative o meno per la classificazione. Quando viene fornita una nuova osservazione l'algoritmo calcola le distanze tra l'osservazione e il resto dei dati presenti nel dataset, le ordina, considera le k osservazioni più vicine e classifica la nuova osservazione alla classe maggiormente presente tra le k osservazioni vicine. Il principale vantaggio di questo metodo è che non è necessario stabilire un modello predittivo prima della classificazione. Gli svantaggi sono invece che non produce una semplice formula di probabilità e che la sua accuratezza predittiva è fortemente influenzata dalla misura di distanza utilizzata e dal numero di osservazioni k considerate vicine.

3. Alberi decisionali e Random Forest

   Gli alberi decisionali rappresentano una tecnica di apprendimento per la risoluzione di problemi di classificazione e di previsione. Essi utilizzano tecniche statistiche per individuare la relazione tra una variabile target e tutte le variabili indipendenti. Nei problemi di classificazione, l'algoritmo inizia raggruppando tutti i soggetti del dataset iniziale in un nodo padre. In seguito ripartisce i soggetti in base alle covariate in gruppi finali/foglie mutuamente esclusivi in modo che ogni nodo finale/foglia è più puro possibile in termini di distribuzione della variabile target. Il grado di impurità di un nodo esprime la distribuzione della variabile target in quel nodo e viene misurato con degli indici statistici, come l'indice di eterogeneità di Gini oppure l'Entropia. La decisione per dividere i nodi da padre in figli viene presa calcolando ad ogni iterazione tutte le possibili partizioni per tutte le covariate e per diversi livelli delle stesse, scegliendo tra tutte la partizione che massimizza il decremento del grado di impurità. Oltre a definire il criterio di partizione di un nodo, bisogna anche denifire una regola di arresto che permetta di trovare l'albero che classifica meglio le osservazioni. I problemi di questo metodo sono l'instabilità, ovvero il fatto che è molto sensibile ai dati del training, e l'alta varianza. Come possibili soluzioni sono stati proposti metodi più evoluti, come il metodo Random Forest proposto da Breiman.

   Il Random Forest è una collezione di alberi decisionali. La prima fase dell'algoritmo chiamata bootstrapping consiste nel creare nuovi training di eguale dimensione dal dataset originale, selezionando le osservazioni tramite campionamento con reinserimento. Per ogni nuovo training viene costruito un albero decisionale utilizzando un sottoinsieme casuale delle covariate del dataset completo. Il numero ottimale delle features è dato dal logaritmo oppure dalla radice quadrata del numero di features totali. Una volta costruita la foresta di alberi, si passa alla fase di aggregation che consiste nell'individuare per una nuova osservazione per ogni albero costruito la classe di appartenenza. L'algoritmo sceglie come classe di appartenenza finale della nuova osservazione la classe a cui è stata assegnata più volte. Utilizzando diversi training si risolve il problema dell'instabilità degli alberi decisionali mentre selezionando casualmente le features si riduce la correlazione tra di essi e si ottiene quindi un modello con minore varianza e maggiore capacità di generalizzare.

3 Risultati

Questa sezione è dedicata a mostrare i risultati ottenuti durante l'analisi da noi svolta.

3.1 Risultati principali

In questa sezione illustreremo i diversi passaggi svolti durante la fase di analisi dei dati. I passaggi svolti sono stati i seguenti:

1. Data pre-processing e preparation: analisi e pulizia del dataset;
2. Modeling: costruzione dei modelli di classificazione sul training set e miglioramento dei modelli e dei loro parametri sul validation set;
3. Evaluation: valutazione della performance e della capacità previsiva dei modelli nel test set;

Data pre-processing

Innanzitutto abbiamo svolto alcune analisi descrittive dei dati. Tramite il grafico mostrato in Figura 3 possiamo notare che la distribuzione di soggetti fumatori è maggiormente concentrata nei soggetti più giovani.

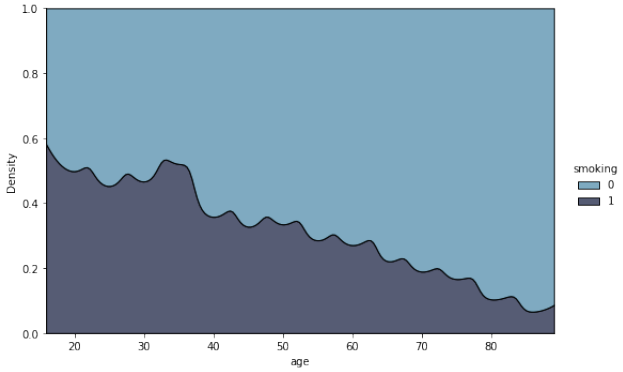

Figura 3: Distribuzione dell'età per classe

Abbiamo poi controllato la presenza di valori mancanti ma non ne abbiamo riscontrati. Dopodichè abbiamo osservato le variabili singolarmente, verificando l'assenza di valori non conformi per le variabili factor e osservando per la variabile "oral" la presenza di un solo livello pari a "Yes" decidendo quindi di rimuoverla. Inoltre abbiamo rimosso anche la variabile "ID" poichè non significativa per le analisi successive. Per continuare le analisi di pulizia dei dati, abbiamo diviso il dataset in training, validation e test assegnando il 60% delle osservazioni al training, il 20% al validation e il 20% al test.

Per verificare la multicollinearità del dataset, abbiamo valutato la correlazione tra le variabili. Dalla matrice di correlazione mostrata in Figura 4, si può notare che le variabili in generale non presentano una situazione di multicollinearità, ad eccezione della variabile AST e ALT che presentano un coefficiente di correlazione pari a 0.75 e delle variabili waist.cm e weight.kg che presentano un coefficiente di correlazione pari a 0.82. Abbiamo quindi deciso di eliminare la variabile ALT, essendo più specifica per le malattie del fegato, e waist.cm. Abbiamo notato inoltre che anche le variabili systolic e relaxation hanno un coefficiente di correlazione alto, ma abbiamo deciso comunque di mantenerle entrambe poichè stimando i nostri modelli risultano entrambe significative.

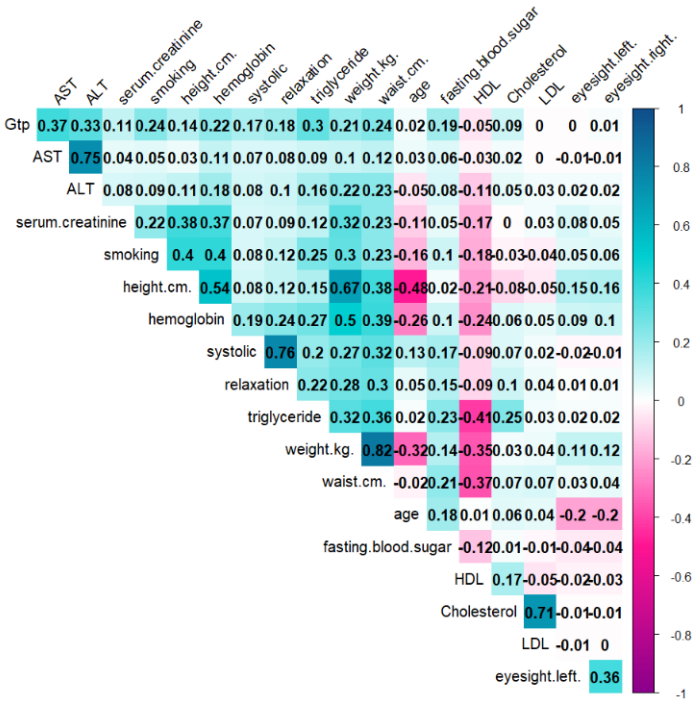

Figura 4: Correlazione delle variabili

Per valutare la presenza di valori anomali abbiamo confrontato il range di variazione delle variabili con quello di una persona sana come mostrato in Figura 5.

Variabili	Range di variazione		Range di variazione di una persona sana
	Minimo	*Massimo*	*Valori*
Systolic	71	240	<140
Relaxation	40	146	<90
Fasting blood sugar	46	505	da 60 a 700 mg/dl
Cholesterol	55	445	<200 mg/dl
Triglyceride	8	999	<400 mg/dl
HDL	4	618	>40 mg/dl
LDL	1	1860	da 100 a 130 mg/dl
Hemoglobin	4,9	21,1	da 12 a 18 g/dl
Serum creatinine	0,1	11,6	da 0,84 a 1,21 mg/dl
AST	6	1311	da 8 a 48 U/l
ALT	1	2914	da 7 a 55 U/l
Gtp	1	999	da 2 a 50 U/l

Figura 5: Range di variazione delle variabili

Abbiamo quindi deciso di eliminare alcune osservazioni per le variabili AST, systolic, fasting blood sugar, HDL, LDL, Cholesterol, triglyceride e serum creatinine, impostando per ognuna di queste variabili un valore massimo che ci ha permesso di eliminare le osservazioni con valori anomali dovuti probabilmente ad errori di misurazione. Il ridotto numero di osservazioni eliminate non impatta sulla nostra analisi.

Analzzando gli outlier per ogni variabile abbiamo individuato una situazione critica per alcune di esse come mostrato in Figura 6.

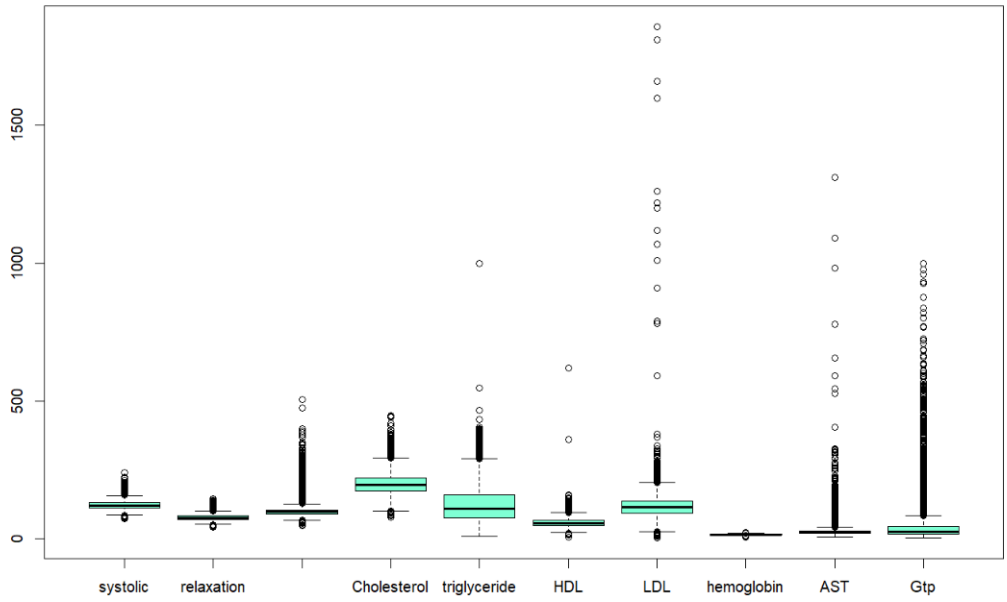

Figura 6: Boxplot di alcune variabili

Abbiamo quindi provato a applicare alcune trasformazioni su di esse riscontrando però miglioramenti tramite la trasformazione logaritmica soltando per 3 variabili. In Figura 7 e Figura 8 riportiamo un esempio di trasformazione logaritmica applicata alla variabile triglyceride.

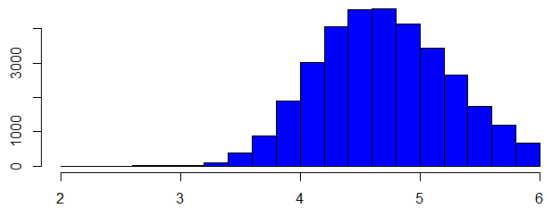

Figura 7: Istogramma della variabile triglyceride

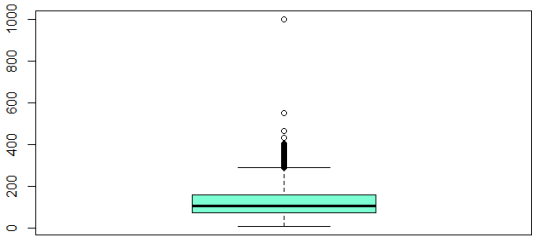

Figura 8: Boxplot della variabile triglyceride

In seguito abbiamo standardizzato i dati data la presenza di osservazioni con unità di misura differenti. Abbiamo poi verificato alcune assunzioni sulle variabili esplicative numeriche. In particolare attraverso l'analisi dei boxplot condizionati alla variabile smoking abbiamo notato che le variabili hanno una varianza simile tra le due classi come mostrato in Figura 9.

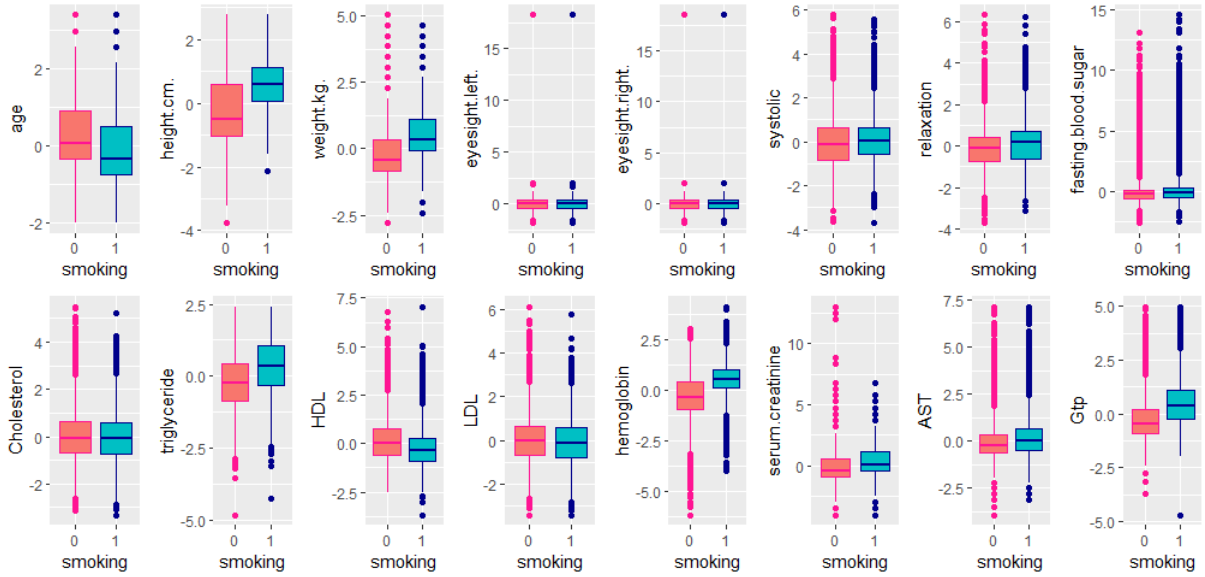

Figura 9: Boxplot condizionati

Abbiamo poi verificato la covarianza delle variabili condizionata alla classe e notato che è simile nella maggioranza delle variabili. In particolare il grafico in Figura 10 evidenzia alcune variabili in cui questa situazione è maggiormente presente.

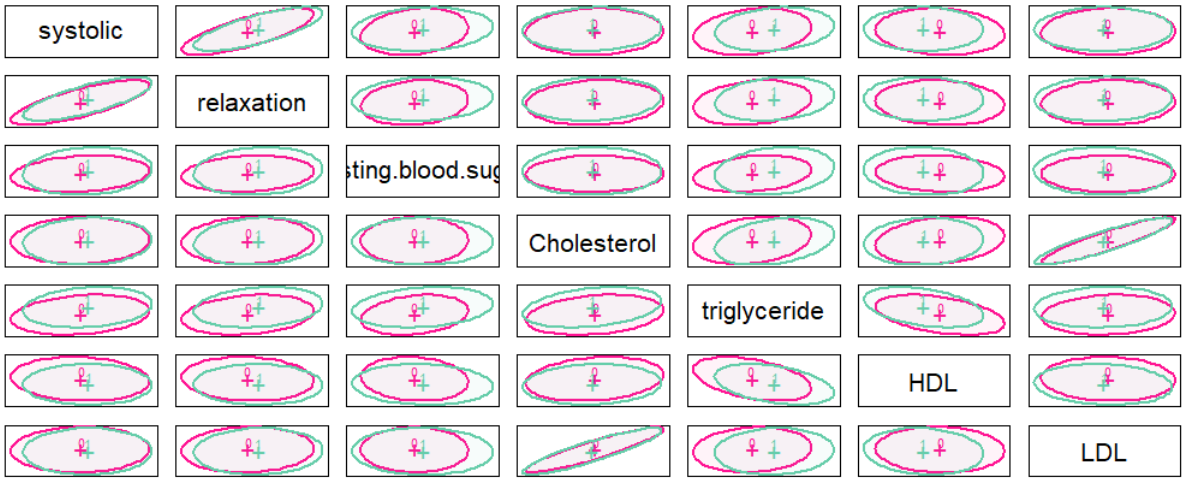

Figura 10: CovEllipse di alcune variabili

Attraverso l'analisi grafica delle curve di densità condizionate alla classe, mostrate in Figura 11, abbiamo notato che solo poche delle variabili esplicative seguono una distribuzione normale, per questo motivo abbiamo deciso che non verranno applicati i metodi di analisi discriminante lineare e quadratica LDA e QDA.

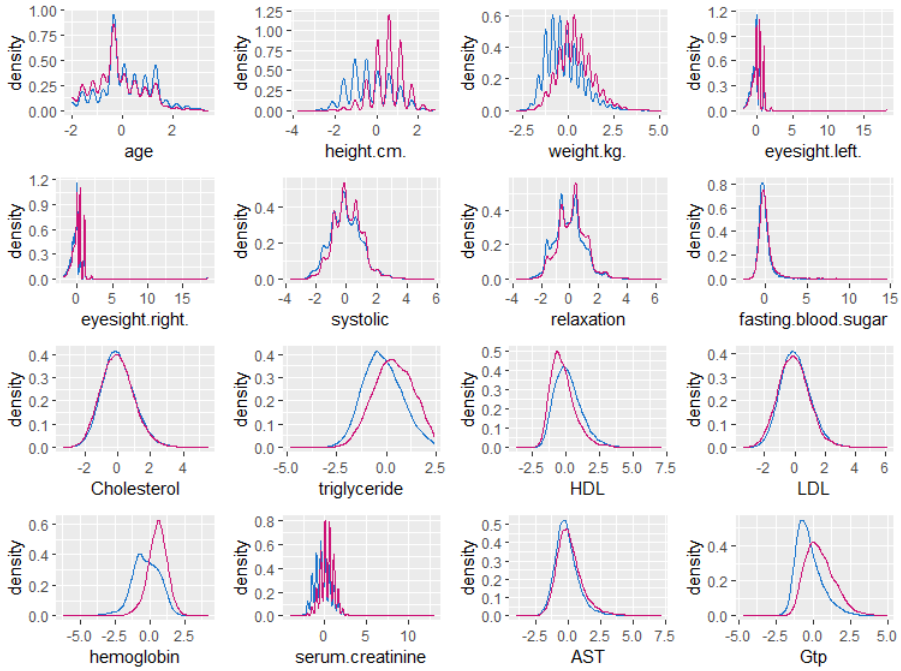

Figura 11: Densità delle variabili condizionata alla classe

Modeling

Dopo aver concluso le procedure di pulizia del dato, abbiamo stimato i modelli ritenuti da noi più adatti per la tipologia di variabile target.

1. Regressione logistica

   Come primo modello abbiamo deciso di sviluppare il modello logistico utilizzando tutte le variabili esplicative e di applicare una selezione stepwise delle variabili basata sull'AIC per rimuovere le variabili non significative. Inoltre, dopo aver individuato i punti influenti e averli rimossi, abbiamo ristimato il modello ottenendone uno con un AIC migliore. Dopo aver applicato le trasformazioni svolte precedentemente sul training anche sul validation, abbiamo quindi utilizzato il modello logistico per stimare le probabiltà di appartenenza alla classe smoking dei dati presenti nel validation, ottenendo un error rate pari a 25,19%, un'accuracy pari a 74,81% e un coefficiente di AUC pari a 0,8301. Come si può notare dalla Tabella 1, il modello classifica in modo più efficace le osservazioni appartenti alla classe non fumatore, cioè quella maggioritaria.

		Valori attuali	
		0	1
Valori previsti	0	5470	1227
	1	1577	2857

Tabella 1: Matrice di confusione della regressione logistica

2. K-nearest neighbours

   Un altro modello da noi sviluppato è stato il K-nearest neighbours. L'unico parametro da scegliere in questo modello è il k, cioè il numero di osservazioni più vicine alla nuova osservazione. Il k ottimale ottenuto sul training è stato 44 e testando direttamente questo metodo sul validation abbiamo ottenuto la matrice di confusione in Tabella 2. Dalla tabella si può notare che l'algoritmo

fallisce maggiormente nel classificare le osservazioni nella classe 1 (fumatore), risultato dovuto allla presenza della classe maggioritaria in 0 (non fumatore). Nonostante quest'errore trovo un error rate di 26,32%, un'accuracy pari 73,67% e un coefficiente di AUC pari a 0,8105.

	Valori attuali	
Valori previsti	0	1
0	5546	1429
1	1577	2857

Tabella 2: Matrice di confusione del KNN

3. Random forest

   Per applicare il modello Random Forest si è deciso di utilizzare il training completo, ottendo dei risultati ottimi. Infatti si è osservata un'accuracy pari al 100% , non commettendo quindi nessun errore di classificazione.

Evaluation

Dato che i risultati ottenuti in termini di accuracy e di curva di ROC sono molto simili per il modello logistico e il KNN, abbiamo deciso di considerare entrambi per la fase di valutazione sul test, insieme al modello Random Forest. Per prima cosa abbiamo riunito il training e il validation formando così un unico training contenente l'80% delle osservazioni, a cui abbiamo applicato le stesse trasformazioni svolte nella fase di pre-processing.

1. Regressione logistica

   Avendo cambiato il training, abbiamo ricalcolato il modello logistico eliminando le variabili non significative tramite la selezione stepwise su AIC e verificando tramite il grafico in Figura 12 la presenza di punti influenti.

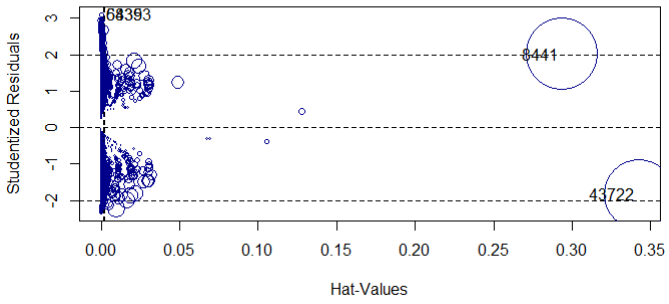

Figura 12: Punti influenti per il modello logistico

Dopo aver rimosso questi punti abbiamo ristimato il modello e ottenuto il modello finale :

$smoking = -3.09915 + 2.8339 gender + 0.25847 height.cm. - 0.2302 weight.kg. - 0.02698 eyesight.left. - 0.13433 hearing.left - 0.19638 systolic + 0.06581 relaxation + 0.03765 fasting.blood.sugar - 0.11707 Cholesterol + 0.3191 triglyceride + 0.04646 HDL + 0.18551 hemoglobin - 0.17821 Urine.protein - 0.1837 serum.creatinine - 0.20123 AST + 0.53988 Gtp + 0.3078 dental.caries + 0.33419 tartar$

Valutando la performance sul test, abbiamo ottenuto un error rate pari a 24,75%, un'accuracy pari a 75,24%, un coefficiente di AUC pari a 0,8355 e la matrice di confusione mostrata in Tabella 3.

		Valori attuali	
		0	1
Valori previsti	0	5494	1244
	1	1509	2875

Tabella 3: Matrice di confusione per il modello logistico

Tuttavia, come possiamo vedere in Figura 13, il modello logistico non è adatto ai nostri dati perchè la relazione tra le variabili e $log\frac{\pi}{1-\pi}$, dove π indica la probabiltà a posteriori della classe fumatori, non è lineare.

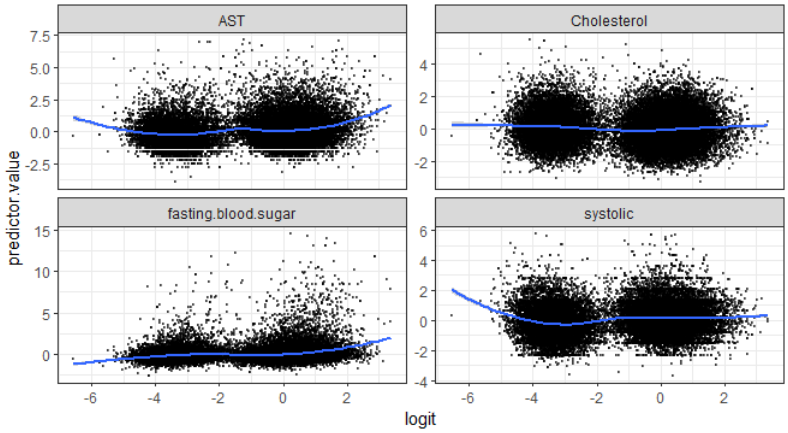

Figura 13: Relazione lineare nel modello logistico

2. K-nearest neighbours

   Applicando invece l'algoritmo KNN sul test, mantenendo lo stesso k ottimale stimato in precedenza, ovvero 44, abbiamo ottenuto un error rate pari a 26,57%, un'accuracy pari a 73,43% e un coefficiente di AUC pari a 0,816. In Tabella 4 si mostrano i valori da noi stimati contro i valori attuali.

		Valori attuali	
		0	1
Valori previsti	0	5563	1515
	1	1440	2604

Tabella 4: Matrice di confusione KNN

3. Random Forest

   Utilizzando il modello Random Forest, stimato sul training completo, abbiamo previsto le classi finali del test set, ottendendo la matrice di confusione indicata in Tabella 5, con un'accuracy del 83,82%.

		Valori attuali	
		0	1
Valori previsti	0	6038	809
	1	991	3288

Tabella 5: Matrice di confusione random Forest

Grazie al modello Random Forest si può individuare anche l'importanza delle covariate, calcolata come l'importanza di Gini che descrive il miglioramento dell'impurità di Gini. Come mostrato nella Figura 14, le variabili più importanti per la classificazione secondo questo modello sono gender, hemoglobin e Gtp.

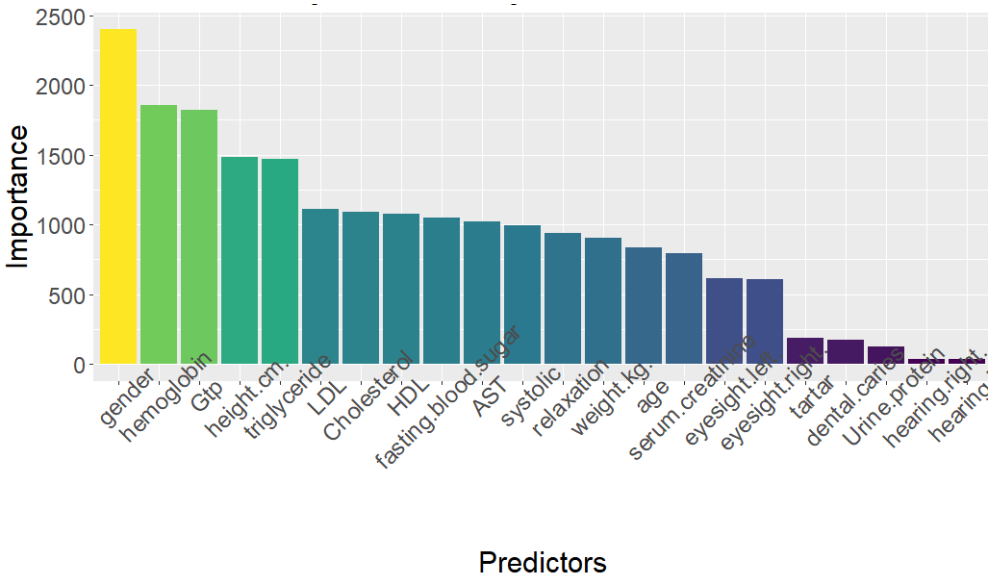

Figura 14: Importanza delle covariate

4. Confronto tra i modelli

   Per confrontare il modello di regressione logistica e il KNN le metriche più adatte sono la sensitivity e la specificity che vengono rappresentate nel grafico della curva di ROC, mostrato in Figura 15. La sensitivity mostra la capacità del modello di classificare correttamente i soggetti appartenti alla classe fumatori mentre la specificity indica la performance del modello nel classificare i soggetti appartenenti alla classe non fumatori. Calcolando l'AUC (area sotto la curva di ROC) si ottengono risultati migliori con la regressione logistica.

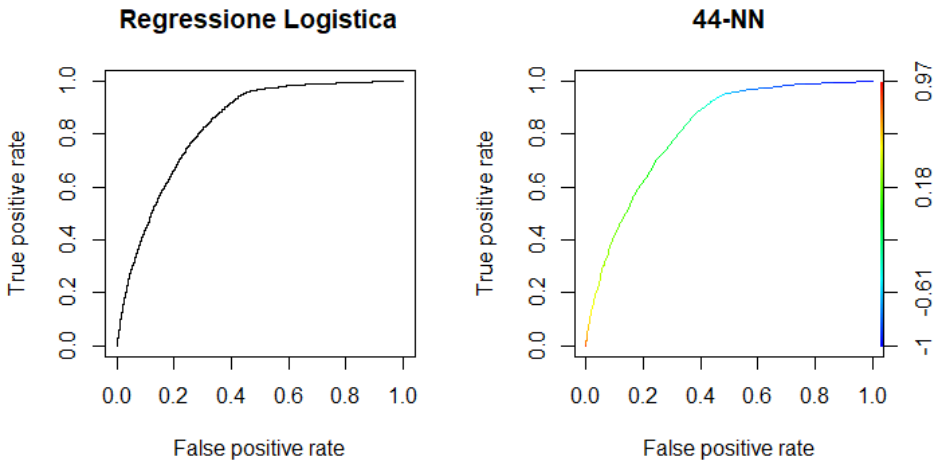

Figura 15: Curva ROC

Nella tabella in Figura 16 vengono mostrati i principali indici per il confronto tra i metodi. Il KNN, rispetto al modello logistico, presenta un valore più elevato di specificity, tuttavia la metrica in questione non è così rilevante per il nostro obiettivo poichè rappresenta la classe dei non fumatori. La regressione logistica ha quindi una performance migliore rispetto al KNN. Ciononostante il random forest presenta valori più elevati in tutte le metriche, in particolare abbiamo notato un aumento significativo nell'accuracy dovuto in gran parte della percentuale dei fumatori classificati correttamente. Quindi il modello migliore è il Random Forest.

Metodo di classificazione	*Accuracy*	*Sensitivity*	*Specificity*	*1-specificity*
Regressione logistica	75,25%	0,698	0,7845	0,2155
44-NN	73,44%	0,6322	0,7944	0,2056
Random Forest	83,82%	0,8025	0,8590	0,141

Figura 16: Tabella di confronto dei due metodi

3.2 Ulteriori risultati

Uno dei principali problemi riscontrati durante la stima del modello logistico e del modello KNN è stato l'elevata dimensionalità del dataset. Abbiamo quindi deciso di provare a stimare diverse volte i modelli scelti su training diversi con un numero inferiore di variabili. Un primo tentativo è stato quello di togliere le variabili HDL, LDL e triglyceride e di lasciare solo la variabile più generica Cholesterol sapendo che quest'ultima è combinazione lineare delle altre tre. Un'altro tentativo è stato quello di togliere la variabile cholestrol e mantenere HDL, LDL e triglyceride, ma in entrambi i casi il modello della regressione logistica presentava un AIC più elevato rispetto al modello di partenza e quindi si è deciso di scartare queste ipotesi. Abbiamo poi calcolato e utilizzato la media delle variabili eyesight.right ed eyesight.left, ma anche in questo caso il modello stimato mostrava un AIC più elevato. Per risolvere il problema abbiamo poi provato ad applicare l'analisi delle componenti principali sulle variabili numeriche pur sapendo che in questo modo avremmo perso l'interpretabilità delle variabili. Abbiamo tenuto in considerazione 11 componenti principali con le quali è stato stimato il modello logistico e KNN, é stato però riscontrato che anche con questo tipo di approccio il modello stimato presentava un'accuracy minore e un AIC più alto. Per questi motivi abbiamo deciso di continuare a considerare tutte le variabili contenute nel dataset.

4 Discussioni

In questo articolo sono state sviluppate delle possibili soluzioni al problema presentato dal sistema sanitario coreano, ovvero l'identificazione di soggetti fumatori in base a valori biologici facilmente reperibili. Sono stati utilizzati la regressione logistica, il K-nearest neighbours e il Random Forest ottenendo migliori risultati attraverso il Random Forest. Esso infatti classifica correttamente circa l'84% delle osservazioni, ottenendo performance ottimali sia nel classificare soggetti appartenti alla classe non fumatori che quelli alla classe fumatori.

Grazie al modello Random Forest si può notare come le variabili più importanti per la classificazione di un soggetto fumatore sono l'emoglobina, il colesterolo in tutte le sue componenti, soprattutto in HDL, e la pressione sanguigna. Questo rispecchia le considerazioni fatte inizialmente sulla conoscenza a priori delle variabili in corrispondenza di un soggetto fumatore, come il fatto che i valori di colesterolo buono HDL sono inferiori e il livello dell'emoglobina è più elevato in un soggetto fumatore. Inoltre si nota che le variabili riguardanti l'igiene e la salute dentale non sono importanti per la classificazione dei soggetti fumatori.

Riferimenti bibliografici

[1] F. Mannocci. and L. Giannarelli , *Il sistema sanitario coreano*, Salute internazionale https://www.saluteinternazionale.info/2011/04/il-sistema-sanitario-sudcoreano/, 2011.

[2] L. Breiman., *Random Forest*, Machine Learning 45.1 (2001): 5-32.

[3] Sconosciuto, *Corea del sud: il giudice da ragione a produttori i tabacco contro il sistema sanitario nazionale*, Intermedia channel, https://www.intermediachannel.it/2020/11/23/corea-del-sud-giudice-da-ragione-a-produttori-di-tabacco-contro-il-servizio-sanitario-nazionale 2020.

[4] Sconosciuto, *I danni del fumo*, My personal trainer , https://www.my-personaltrainer.it/salute/danni-fumo.html.

[5] Sconosciuto, *Fumo e malattieicardiovascolare*, Istituto superiore di sanità , https://www.cuore.iss.it/prevenzione/fumo, 2020.

[6] Sconosciuto, *National Health Insurance Service Health Checkup Information*, e-Government, https://www.data.go.kr/data/15007122/fileData.do, 2021.

[7] Sconosciuto, *Valori alti di emoglobina fumo: carbossiemoglobinemia*, Emoglobina, https://www.emoglobina.info/valori-alti-di-emoglobina-fumo-carbossiemoglobinemia/, 2014.