Elaborato finale del corso Machine Learning
Giugno 2018

Predire la classe di reddito attraverso la tecnica di classificazione

Bassanese Luca, Vargas Ivan, Viganò Elena

Quali sono le caratteristiche degli individui che contribuiscono alla determinazione del loro reddito? Come possono fare i governi a stabilire quanto tassare i singoli individui? Queste ed altre domande se le sono certamente poste i dipendenti statali addetti alla raccolta dei dati durante i periodi di censimento. Anche noi ci siamo interessati all'argomento e in particolare abbiamo cercato di rispondere al quesito: come si fa a capire se un individuo appartiene o meno alla fascia più abbiente della popolazione? Per comprenderlo abbiamo preso in considerazione un *dataset* fornito dal *Bureau of the Census* americano e in seguito a delle prime analisi descrittive ed esplorative, sono state individuate le variabili più importanti per portare a termine il nostro fine attraverso l'uso di metodi di *feature selection*. Successivamente, tramite l'applicazione di tecniche di classificazione supervisionata abbiamo cercato di individuare i classificatori più performanti. Dei vari modelli analizzati, due in particolare sono stati ritenuti superiori agli altri per capacità discriminante e affidabilità.

Indice

Introduzione

Negli Stati Uniti d'America ogni dieci anni il *Bureau of the Census* effettua il censimento di tutta la popolazione americana. Questa operazione ha due scopi fondamentali il primo dei quali è sicuramente quello di contare la totalità dei cittadini della nazione con la finalità di riuscire a determinare il numero corretto di Deputati che, ogni stato, a seguito di elezioni interne, dovrà inviare come propri rappresentanti al Congresso. Questo procedimento, quindi, risulta essere di estrema importanza, in quanto garantisce la corretta determinazione della composizione dell'organo fondamentale a cui compete il potere legislativo.

Il secondo fine del censimento, invece, è quello della raccolta di grandi quantità di dati relativi alla popolazione, i quali, in un secondo momento, permettono agli esperti di sviluppare delle misure statistiche volte a sintetizzare le informazioni più importanti riguardanti non solo gli abitanti dei 50 stati facenti parte della Repubblica Federale americana, ma anche inerenti all'andamento dell'economia di quest'ultima. I principali dati raccolti, infatti, sono quelli riguardanti il livello d'istruzione, l'occupazione e la fascia di reddito a cui i singoli cittadini appartengono. Questo elaborato nasce, quindi, con lo scopo di estrarre valore, attraverso l'applicazione di tecniche di *machine learning*, dai dati facenti parte del dataset da noi analizzato. L'obiettivo è quello di riuscire a prevedere, attraverso l'uso di classificatori, la fascia di reddito a cui ciascuna unità statistica considerata, ovvero ciascun cittadino, appartiene. Comprendere ciò è, infatti, di estrema importanza per tutti i governi al fine di stabilire una tassazione equa e adeguata alle possibilità monetarie dei cittadini.

Non siamo, tuttavia, i primi ad affrontare questo problema. Solo sulla piattaforma *Kaggle*, infatti, si possono consultare molteplici analisi condotte da diversi statistici su questo tema. Questo report, tuttavia, si vuole leggermente discostare dai risultati precedentemente ottenuti, in quanto si focalizza sul cercare di classificare con maggiore accuratezza la categoria di individui più abbiente che si ritiene essere di maggiore interesse per un qualsiasi stato in quanto dotata di maggiori possibilità di esborso e quindi sottoponibile, eventualmente, a una maggiore pressione fiscale.

1. Dataset & preprocessing

1.1 Descrizione del dataset

Il *dataset* che abbiamo utilizzato per portare a termine il *task* che ci siamo prefissati è denominato *Adult Census Data*. Esso contiene solo una parte dei dati raccolti dal *Bureau of the Census* americano con il censimento del 1994 ed è stato ottenuto grazie all'opera di estrazione portata avanti dagli studiosi Ron Kohavi e Barry Becker. Inizialmente questi studiosi lo utilizzarono per analizzare le implicazioni della *sample selection bias* sulle varie tecniche di classificazione. Nel 1996, lo stesso Ron Kohavi lo riutilizzò per costruire un modello ibrido tra il *Naive Bayes* e gli alberi di classificazione con l'intenzione di ottenere una maggiore *accuracy* nelle previsioni.
I risultati vennero esposti in un *paper* dal titolo "*Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid*".
Nel *dataset* sono presenti 15 attributi di cui solamente 6 sono di tipo quantitativo, mentre i restanti sono tutti di tipo qualitativo (sia nominale che ordinale). Il numero di osservazioni è pari a 32561. Ogni singolo *record* rappresenta i dati che sono stati rilevati su ogni cittadino. In particolare, le variabili osservate sono l'età (*age*), il settore di impiego (*workclass*), i pesi di ponderazione per il ricampionamento (*fnlwgt*), il più alto livello di istruzione raggiunto (*education*, *education.num*), lo stato coniugale (*marital.status*), la professione svolta (*occupation*), il tipo di relazione (*relationship*), l'etnia (*race*), il genere (*sex*), le ore lavorative settimanali (*hours.per.week*), l'ammontare dei profitti e delle perdite derivanti da investimenti (*capital.gain*, *capital.loss*), la nazione di nascita (*native.country*) e la classe di reddito di appartenenza (*income*).

1.2 Trattamento dei *missing values*

La presenza di *missing values* nel *dataset* è circoscritta alle sole variabili *native.country*, *workclass* e *occupation.*

Per prima cosa sono stati trattati i valori mancanti della variabile *native.country*. Per eliminare questi ultimi è stata adottata la tecnica del *mode replacement*, che consiste nel sostituire i valori assenti con la modalità dell'attributo preso in considerazione che si manifesta con maggiore frequenza. Nel trattare i valori mancanti relativi alle variabili *workclass* e *occupation*, invece, è stato seguito un procedimento diverso. Innanzitutto, è emerso dalle analisi che solo in due casistiche è possibile riscontrare la presenza di *missing*. La prima è quella in cui il record esaminato presenta la dicitura "?" in corrispondenza delle celle relative ad entrambe le variabili in questione, la seconda, invece, è quella in cui le osservazioni, pur assumendo la modalità *Never-worked* in corrispondenza dell'attributo *workclass*, manifestano comunque la presenza di un valore mancante in corrispondenza della variabile *occupation*. In tutti questi casi si è provveduto a rimuovere interamente le osservazioni in quanto anche applicando una delle altre possibili metodologie utilizzate per il trattamento dei *missing values* si sarebbero ottenuti dei dati che non sarebbero stati sufficientemente affidabili o particolarmente significativi. Questo procedimento è stato eseguito su 1843 righe del *dataset.*

1.3 Riduzione delle modalità delle variabili qualitative

Una volta eseguita una prima pulizia del *dataset*, è emerso che ognuna delle variabili qualitative (con l'unica eccezione dell'attributo *sex*) presenta un elevato numero di modalità.

E' sembrato necessario, quindi, ricercare dei metodi per ridurre in modo efficiente il numero di livelli di ogni singola variabile. Per poter capire come aggregare al meglio le modalità degli attributi, tramite l'uso del software R, si è eseguita una prima serie di analisi esplorative.

Attraverso l'uso di istogrammi, infatti, è stata rappresentata la relazione esistente tra ognuna delle variabili qualitative e la variabile target *income*.

Successivamente, grazie ai grafici ottenuti e in parte anche sulla base delle conoscenze di dominio, sono state accorpate le modalità delle variabili che o contenevano una percentuale simile di soggetti con la medesima tipologia di reddito o che sono state ritenute appartenenti ad una macrocategoria comune.
Quest'ultimo metodo, ad esempio, è stato applicato all'attributo *native.country* dove i paesi sono stati raggruppati in relazione all'area geografica di appartenenza. Nel caso, invece, dell'attributo *education* per effettuare la riduzione delle classi della variabile si è cercato di tenere conto della struttura del sistema d'istruzione americano e di unire sulla base del grafico sotto riportato le classi che presentavano un livello di reddito pressoché simile.

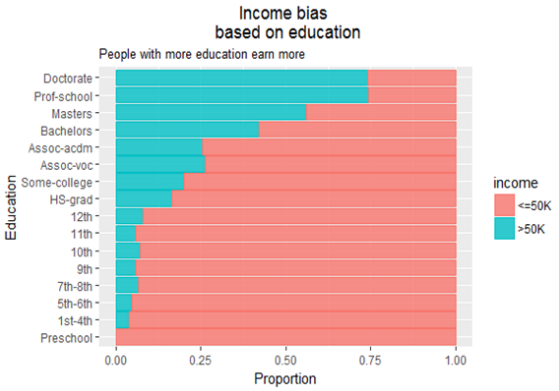

Figura 1. Istogramma condizionato rappresentante la relazione tra la variabile qualitativa *education* e la variabile *target income*.

2. Analisi esplorative

Si passa ora ad un'analisi della dipendenza tra variabili al fine di comprendere se nel *dataset* sono presenti attributi che apportano la stessa informazione.
Procederemo studiando separatamente il grado di correlazione e di associazione esistente tra le variabili quantitative e quelle qualitative. Analizzando la matrice di correlazione delle variabili quantitative si comprende che alcuni attributi non risultano essere significativamente correlati tra di loro. Ciò implica che ogni variabile porta informazioni differenti e da questo consegue che stando unicamente all'osservazione di questo grafico non è possibile effettuare alcun tipo di selezione tra le esplicative.

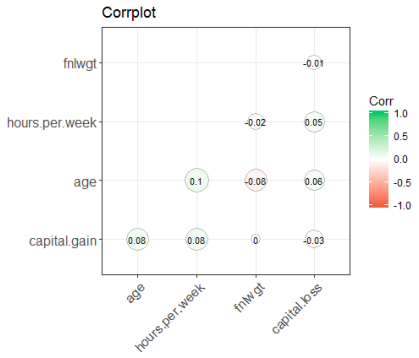

Figura 2. *Corrplot* variabili quantitative.

La dipendenza tra le variabili qualitative è stata misurata, invece, attraverso il calcolo del grado di associazione tra gli attributi. Per determinare tale quantità è stata utilizzata come misura la *V di Cramer*, la quale è una normalizzazione della statistica χ^2 calcolata sulle tabelle di contingenza. Essa assume valori compresi tra 0 e 1. Valori molto vicini allo 0 indicano un'associazione molto debole tra gli attributi, al contrario valori vicini all'1 indicano una forte dipendenza.

La *V di Cramer* è calcolata sulla base della seguente formula:

$$V = \sqrt{\frac{\frac{\chi^2}{n}}{\min(c-1, r-1)}}$$

Dove *n* è il numero totale delle osservazioni, *r* è il numero delle righe della matrice di contingenza e *c* è il numero di colonne di essa.

Tale misura è stata implementata attraverso l'utilizzo di R e dal risultato emerso si è notato che esiste una forte associazione tra alcune coppie di variabili. Un'analisi più approfondita, tuttavia, ci ha portato ad individuare delle motivazioni valide per non escludere tali variabili dall'insieme di dati da utilizzare per implementare la classificazione. I valori di associazione più elevati sono infatti stati riscontrati per le coppie *relationship-sex* (65%) e *marital.status-relationship* (61%). L'elevata associazione tra la prima coppia di attributi può essere giustificata dal fatto che tra le modalità di *relationship* vi sono *wife* e *husband* che ovviamente sono legate al genere dell'individuo considerato. Le variabili *marital.status* e *relationship*, invece, pur risultando essere associate, portano informazioni diverse e per questo si è deciso di non eliminarle.

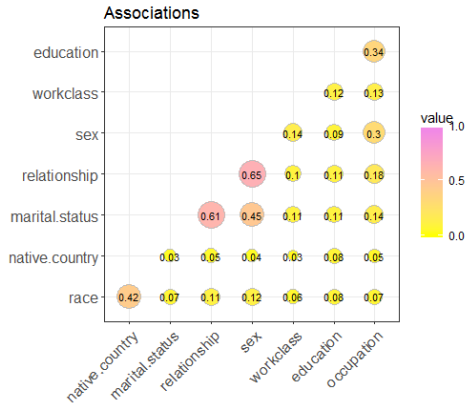

Figura 3. Grafico di associazione tra le variabili qualitative.

2.1 Manipolazione dei dati

Dopo una prima analisi del contenuto del dataset è emerso che quattro delle 15 variabili originarie ovvero *education.num*, *fnlwgt*, *capital.gain* e *capital.loss* non risultano essere utili al fine di implementare con successo gli algoritmi di classificazione.

Education.num, infatti, è una variabile che non apporta nuove informazioni in quanto non fa altro che esprimere in modo diverso il contenuto della variabile qualitativa *education.*

Per verificarlo sono stati osservati dei *box-plot* tra le variabili *Education* e *Education.num*. Tali hanno evidenziato l'esistenza di perfetta correlazione tra i due attributi.

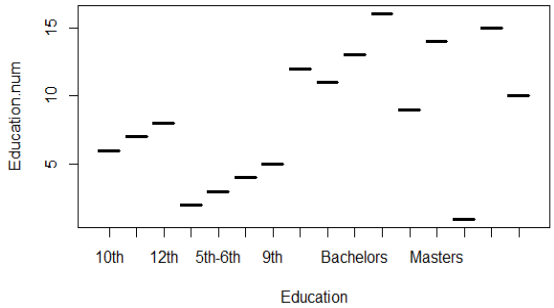

Figura 4. Box-plot tra *Education* e *Education.num*.

La colonna contrassegnata dal nome *fnlwgt*, invece, contiene dei pesi che vengono calcolati mensilmente dalla *Population Division* operante per il *Bureau of the Census*.

Essi fungono da ottimi indicatori delle caratteristiche demografiche della popolazione. In corrispondenza di tutti i soggetti che condividono le medesime condizioni di vita, quindi, dovremo trovare nel *dataset* il medesimo valore della variabile *fnlwgt*.

I motivi principali che, tuttavia, ci hanno condotto a voler scartare questa variabile sono due. In primo luogo, è molto difficile poterli usare poiché il loro calcolo dipende dall'esito di un processo di campionamento. In secondo luogo, *fnlwgt* risulta essere indipendente da *income*. Questo può essere visto osservando la distribuzione di *fnlwgt* condizionata al *target*. Essa, infatti, non varia al variare della classe della variabile risposta presa in considerazione.

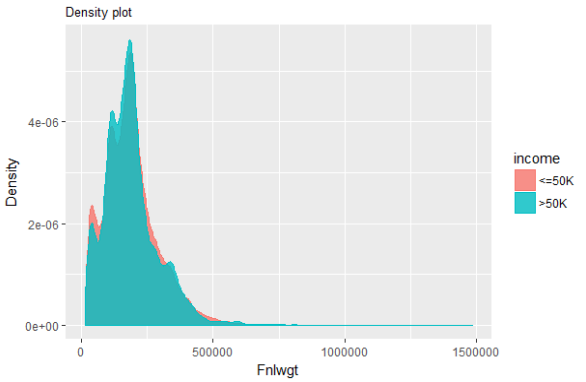

Figura 5. Distribuzione condizionata della variabile *fnlwgt* rispetto alla variabile *income.*

Le variabili capital.*gain* e *capital.loss*, invece, risultano avere una distribuzione schiacciata sulla loro mediana (in entrambi i casi pari a zero) e solamente gli *outlier* assumono valori diversi da essa. In aggiunta, l'osservazione della distribuzione dei *box-plot* condizionati al *target* permette affermare che nessuno dei due attributi considerati risulta avere significative capacità discriminanti che possono essere d'aiuto nell'istruzione del classificatore a differenza, invece, delle altre due variabili quantitative – *age* e *hours.per.week* - facenti parte del *dataset*.

Per i motivi sopra menzionati, si è deciso quindi di eliminare le suddette variabili dall'insieme di dati in analisi.

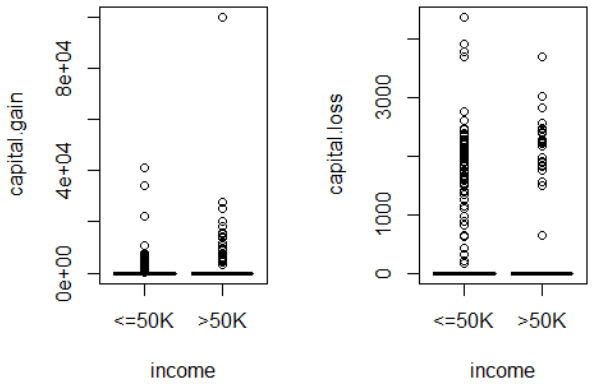

Figura 6. Box-plot condizionati alla variabile *target* per gli attributi *capital.gain* e *capital.loss.*

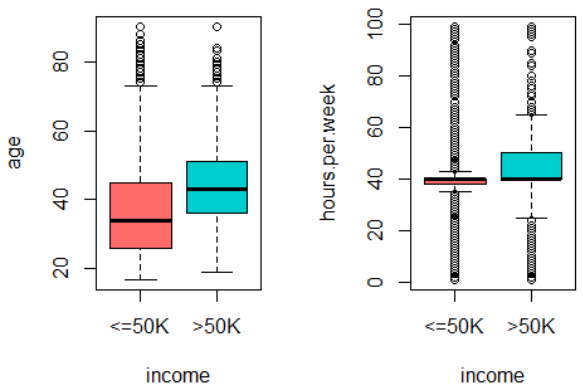

Figura 7. Box-plot condizionati alla variabile *target* per gli attributi *age* e *hours.per.week*.

3. Classification

Completata la fase di *preprocessing*, al fine di riuscire a prevedere in modo corretto la categoria di reddito a cui ogni cittadino appartiene, sono stati implementati diversi algoritmi di classificazione attraverso l'utilizzo del programma KNIME.

3.1 Partizione del dataset

Per procedere nello sviluppo dell'analisi, si è reso necessario partizionare il dataset in 2 parti distinte: nel *training set* è stato inserito il 67% delle osservazioni a nostra disposizione, mentre il restante 33% è andato a comporre il *test set*.

Inoltre, per garantire che con la nostra partizione venga mantenuta la distribuzione di *income* sulla popolazione, è stato utilizzato il metodo dello *stratified sampling*, fissando 123 come seme delle estrazioni in modo tale da garantire la riproducibilità dei nostri risultati.

3.2 I classificatori utilizzati

Per lo studio sono stati utilizzati i seguenti modelli:

- **Modelli euristici:** J48, *Random Forest* (*Number of Trees* = 30), *Decision Tree* (*Gini Index*, *number of threads* = 4, *minimum number of records per node* = 5, *pruning method* basato sul principio del *minimum description length*).
- **Modelli di regressione**: regressione logistica e regressione logistica semplice.
- **Modelli di separazione:** *Support Vector Machine* (SMO poly) con *kernel* polinomiale di grado 1, *Multylayer Perceptron* (*hidden layers* = a, *learning rate* = 0.3, *momentum* = 0.2), *SPegasos*.
- **Modelli probabilistici**: *Naïve Bayes* (*weka*), *Naïve Bayes Learner*, *Bayes Net* (*searchAlgorithm* = K2 / TAN), *Naïve Bayes Tree*.

4. Risultati della classification

Una volta implementati i classificatori è stata comparata la bontà della loro performance attraverso l'uso di diverse misure come l'*accuracy*, la *precision*, la *recall*, la *F-measure* e l'AUC.

4.1 Misure di valutazione

L'*accuracy* permette di stabilire il numero di corrette previsioni di ciascun classificatore e si calcola sulla base dell'utilizzo della *confusion matrix*, cioè la matrice che consente il confronto tra i valori predetti e quelli reali del *test set*.

		Predicted class: *P*	Predicted class: *N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Figura 8. Matrice di confusione

Dalla *confusion matrix* si ottiene il valore dell'*accuracy*:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Questa misura di performance è particolarmente rilevante nel caso in cui si dia la medesima importanza al prevedere correttamente entrambe le modalità della variabile *target*.

Nel caso in cui, invece, si volesse prestare maggiore attenzione alla previsione di solo una delle modalità di quest'ultima ovvero quella che, in genere, è considerata essere la più rara e significativa, si dovrebbe ricorrere ad altri indicatori. In particolare, nel caso del *dataset adult*, si sceglie come classe di riferimento quella relativa al reddito "*>50K*" poiché è numericamente inferiore ed è più difficile da prevedere. Inoltre, si ritiene che un qualsiasi governo, nel momento in cui effettua un censimento, abbia maggiore interesse nell'individuare la parte della popolazione che ha un reddito più elevato in modo tale da applicare ai soggetti in questione una tassazione più

cospicua. Anche queste misure sono ottenute a partire dai valori contenuti nella matrice di confusione. Esse sono:

- la *precision*: $p = \frac{TP}{TP+FP}$
- la *recall*: $r = \frac{TP}{TP+FN}$
- la *F-measure*: $F_1 = \frac{2rp}{r+p} = \frac{2 \times TP}{2 \times TP+FP+FN}$

La *precision* individua la frazione delle osservazioni previste correttamente dal classificatore come facenti parte della classe più rara (*positive class*) rapportandola al totale delle previsioni positive.
Tanto più alto è il valore assunto dalla *precision*, tanto più basso è il numero di falsi positivi previsti dal classificatore.

La *recall*, invece, indica il numero delle osservazioni facenti parte della *positive class* previste correttamente dal classificatore. Tanto più alto è il valore assunto da questo indicatore, tanto più accurata risulta essere la previsione di questa classe.

La *F-measure* rappresenta una media armonica tra la *recall* e la *precision*. Un valore molto alto della *F-measure* è sinonimo di valori elevati di *recall* e *precision*.

Da ultimo, l'AUC (*area under curve*) è una misura che rappresenta l'area sottostante alla curva ROC, la quale è costruita mettendo sull'asse delle x il *true positive ratio*:

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

Sull'asse delle y, invece, viene rappresentato il *false positive rate*:

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

L' AUC della ROC è una misura della bontà dell'andamento globale del classificatore. Essa misura la capacità discriminante, ovvero la capacità che il classificatore ha di distinguere le osservazioni facenti parte del *test set* sulla base della classe di appartenenza.

Nel caso in cui l'AUC è pari a 0.5 il classificatore viene a coincidere con lo *ZeroR*, il quale è un classificatore molto semplice basato sull' utilizzo del *random sampling*.

Dal punto di vista grafico, inoltre, il modello può essere giudicato tanto migliore quanto più la curva tende all'angolo in alto a sinistra del grafico fino ad arrivare al caso in cui il classificatore è ritenuto perfetto ovvero quando AUC = 1.

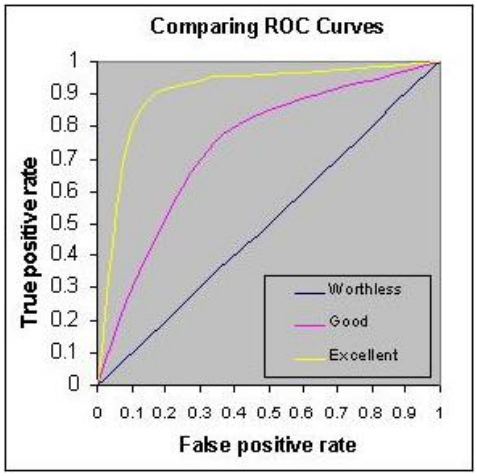

Figura 9. Esempio di andamento di ROC *curve*.

4.2 Risultati e commenti

Analizzando i risultati si nota che tutti i classificatori hanno un'*accuracy* molto elevata (compresa tra il 79% e l'84%). La stessa cosa si può dire per il livello di AUC. Infatti, quasi tutti i modelli hanno un AUC pari o superiore all'80%. Ciò indica che i diversi algoritmi utilizzati hanno una buona capacità discriminante. Tuttavia, se si studia anche il valore assunto dalla *recall*, si può notare, che i classificatori che registrano i valori più alti di *accuracy*, (sopra l'82%), assumono dei valori di *recall* che sono estremamente bassi (tra il 50% e il 60%). Tale fenomeno prende il nome di "*Accuracy paradox*" e sembrerebbe suggerire che la gran parte dei classificatori da noi utilizzati non riesce a prevedere correttamente la classe d'interesse ovvero la classe "*>50K*".

Le uniche eccezioni a questo fenomeno sono il *Naïve Bayes* e il *Bayes Net* con *searchAlgorithm K2*. Essi risultano, in questo caso, essere i classificatori più performanti in quanto mantengono un buon livello di *accuracy* (79.8% e 79.9%) e contemporaneamente sono in grado molto bene di prevedere la classe minoritaria (*recall* 74.4% e 75.3%). I classificatori mostrano, inoltre, di avere anche un'ottima capacità di discriminare tra le classi del *target* (AUC = 87.4% e AUC = 87.6%).

	Accuracy	Recall	Precision	F-measure	AUC
J48	0,829	0,541	0,704	0,612	0,833
Random Forest	0,807	0,566	0,625	0,594	0,846
Decision Tree	0,827	0,569	0,683	0,621	0,859
Logistic	0,831	0,556	0,705	0,622	0,88
Simple Logistic	0,831	0,558	0,703	0,622	0,879
SMO poly	0,812	0,564	0,639	0,599	0,729
Multilayer Perceptron	0,811	0,7	0,605	0,65	0,871
SPegasos	0,828	0,482	0,737	0,583	0,88
Naive Bayes	0,798	0,744	0,573	0,647	0,874
Bayes Net BNC	0,799	0,753	0,573	0,651	0,876
Bayes Net TANB	0,824	0,613	0,658	0,635	0,876
NBTree	0,831	0,574	0,693	0,628	0,875

Figura 10. Risultati dei modelli di classificazione.

5. Bilanciamento

Avendo osservato nella gran parte dei casi valori molto bassi per la *recall*, si è voluto analizzare il valore delle frequenze relative della variabile *income*. Ciò ha permesso di evidenziare la presenza di uno sbilanciamento nei dati a disposizione. Infatti, è stato possibile riscontrare che solo il 24.9% delle unità statistiche appartiene alla classe "*>50K*".

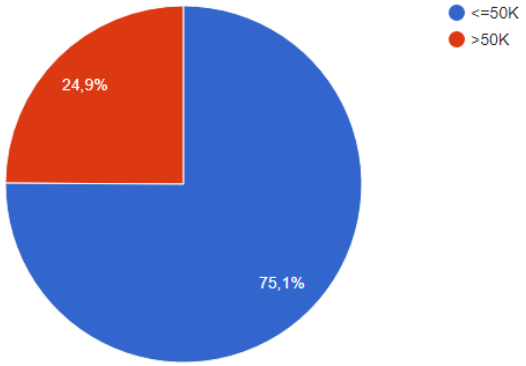

Figura 11. *Pie Chart* delle modalità della variabile *income*.

Questo fatto potrebbe causare una possibile distorsione nel processo di apprendimento dei classificatori, che rischierebbero, quindi, di focalizzarsi prevalentemente sulla classe predominante ("*<= 50K*"), ignorando la classe rara ("*> 50K*").

Al fine di rendere equilibrate le classi del *target* si è deciso di bilanciare il dataset utilizzando la tecnica dell'*oversampling*, il cui scopo consiste nell' aumentare la dimensione della classe minoritaria eseguendo un ricampionamento con ripetizione delle osservazioni appartenenti a quest'ultima. Questo procedimento è stato eseguito fino a che le due classi sono giunte ad avere la medesima numerosità.

L'algoritmo che è stato utilizzato per perseguire questo scopo è: "*ovun.sample*", ovvero una funzione facente parte del pacchetto "ROSE" di R, che esegue questa tecnica attraverso l'uso dell'opzione *method* = "*over*". Tale bilanciamento è stato eseguito unicamente sui dati appartenenti al *training set* originario. Il *test set*, invece, è stato mantenuto inalterato.

Dall'applicazione di questa tecnica si è ottenuto un *training set* di dimensione pari a 30936 osservazioni, il quale è stato utilizzato per istruire gli stessi modelli di classificazione precedentemente implementati.

I risultati ottenuti sono stati i seguenti:

	Accuracy	Recall	Precision	F-measure	AUC
J48	0,784	0,782	0,55	0,646	0,804
Random Forest	0,79	0,644	0,574	0,607	0,843
Decision Tree	0,786	0,808	0,551	0,655	0,86
Logistic	0,779	0,826	0,539	0,653	0,882
Simple Logistic	0,778	0,83	0,539	0,653	0,88
SMO poly	0,751	0,839	0,503	0,629	0,78
Multilayer Perceptron	0,778	0,813	0,539	0,648	0,869
SPegasos	0,437	0,994	0,308	0,471	0,872
Naive Bayes	0,757	0,843	0,511	0,636	0,874
Bayes Net BNC	0,759	0,853	0,512	0,64	0,876
Bayes Net TANB	0,769	0,826	0,526	0,643	0,874
NBTree	0,779	0,837	0,539	0,656	0,871

Figura 12. Risultati dei modelli di classificazione sul *dataset* bilanciato.

Osservando i valori registrati per le diverse misure di valutazione della bontà dei classificatori, si può notare un aumento generalizzato del valore della *recall*. Il valore dell'*accuracy*, invece, sebbene abbia subito solo una lieve diminuzione rispetto al caso di assenza di bilanciamento, in questa parte dell'analisi non verrà più preso in considerazione in quanto ci si vuole focalizzare sul prevedere correttamente la classe positiva. Il livello di *AUC*, infine, è rimasto più o meno costante attorno all'80% indicando che comunque tutti i classificatori presi in considerazione discriminano bene le due classi anche nel caso di bilanciamento.

Si è deciso, inoltre, di commentare unicamente i risultati degli algoritmi migliori: *Naïve Bayes*, *Simple Logistic*, *Bayesian Network (*con come *searchAlgorithm K2* e *TAN)* e *Naive Bayes Tree*

Innanzitutto, è necessario osservare che il *Naïve Bayes* e la *Bayesian Network (*con come *searchAlgorithm K2)*, ovvero i classificatori scelti come più performanti in caso di assenza di bilanciamento, continuano ad essere efficienti. Essi assumono addirittura un valore più elevato di *recall*,

infatti essa passa per il *Naïve Bayes* dal 74.4% all'84.3%, mentre per la *Bayesian Network* essa va dall'75.3% all'85.3%.

Entrambi i classificatori mantengono il medesimo livello di AUC e sebbene assumano valori pressoché simili di entrambe le misure bisogna, tuttavia, sottolineare che il modello più performante, in questo caso, risulta essere la *Bayesian Network*.

L'altra rete bayesiana, inoltre, ottiene risultati molto buoni e di poco inferiori a quelli di queste ultime.

Anche la regressione logistica semplice si dimostra essere un ottimo modello. Essa, infatti, nonostante manifesti una *recall* leggermente al di sotto del valore registrato per gli altri modelli individuati come più performanti, presenta la maggiore area sotto la curva ROC (AUC dell'88%).

Da ultimo, il *Naïve Bayes Tree* giunge a risultati che sono in linea con quelli degli altri classificatori.

E' bene specificare, inoltre, che l'*SPegasos* non può essere considerato come uno dei modelli migliori in quanto, pur presentando valori molto alti di AUC e di *recall*, esso si concentra esclusivamente sul prevedere la classe minoritaria, fallendo nella previsione della categoria "<=50K".

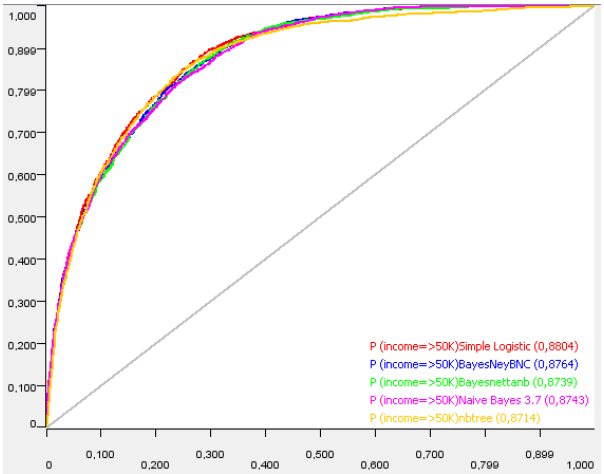

Figura 11. ROC dei migliori classificatori.

6. Feature selection

I risultati precedentemente riportati sono stati ottenuti sulla base di una valutazione effettuata mediante le analisi esplorative.

Si vuole ora verificare se sarebbe stato possibile raggiungere performance migliori applicando altre metodologie per la selezione degli attributi.

Per fare ciò sono state utilizzate delle tecniche di *feature selection* per eliminare le variabili del dataset che risultano essere ridondanti o irrilevanti al fine della classificazione.

Le possibili tecniche utilizzabili sono:

Principal component analysis (PCA): tecnica che permette di individuare nuovi attributi ottenuti come combinazione lineare di quelli originali. Essi risultano essere ortogonali tra di loro e hanno la capacità di catturare gran parte della variabilità presente nei dati. Tale metodo, tuttavia, è particolarmente adatto per i dati di tipo quantitativo.

Filter: l'approccio basato sul *filter* fa una selezione degli attributi prima di istruire il classificatore. Esso, per differenti *subset* del dataset, ottimizza una funzione obbiettivo la quale misura il grado di associazione tra le variabili esplicative e la variabile risposta. Gli attributi selezionati dal *filter* saranno quelli del *subset* ottimale. In base alla funzione obbiettivo ed alla tipologia di *subsets* impiegati si possono distinguere due tipologie di *filter*:

- **Filter univariato**: calcola per ogni attributo l'associazione con la variabile risposta, permettendo di eliminare gli attributi irrilevanti (***Infogain***, ***Gainratio***)
- **Filter multivariato**: analizza congiuntamente più variabili in modo tale da eliminare non solo gli attributi irrilevanti ma anche quelli ridondanti (***CfsSubsetEval***, ***Relief***)

Wrapper: tecnica che si basa sulla selezione degli attributi che permettono di ottenere la migliore performance per uno specifico classificatore. Ogni volta che viene individuato un differente sottogruppo delle variabili del dataset originario viene stimato il corrispondente modello e se ne valuta la bontà. Gli attributi selezionati dal *wrapper* saranno quelli appartenenti al *subset* che consegue la migliore performance.

6.1 Implementazione

Delle differenti tecniche di *feature selection* esposte, si è deciso di utilizzare solamente il *filter*. Non si è ritenuto opportuno in questo caso applicare la PCA in quanto il *dataset* a nostra disposizione contiene attributi per lo più di natura qualitativa. Ciò potrebbe causare dei problemi nel calcolo della matrice delle covarianze su cui si basa questo approccio. Per quanto riguarda il *wrapper*, invece, si è deciso di non

usufruire di questa tecnica in quanto essa, pur essendo potenzialmente molto utile, risulta essere computazionalmente troppo onerosa.

Analisi con InfoGain, Gain ratio e Relief

La *feature selection* eseguita con queste tre tecniche si basa sull'utilizzo del cosiddetto *Ranker* come metodo di selezione. Esso ha lo scopo di ordinare le esplicative. In tutti e tre i casi si giunge al medesimo risultato: le variabili presenti nel *dataset* vengono tutte individuate come significative e non viene effettuata nessun tipo di selezione delle variabili. Essendo l'obiettivo quello di riuscire a spiegare la variabilità del *dataset* usando un ridotto numero di attributi, tale risultato non è stato ritenuto rilevante ai fini della nostra analisi.

Analisi con CfsSubsetEval

Viene poi usato il *filter CfsSubsetEval* per effettuare un ulteriore tentativo di selezione delle variabili. Tale algoritmo utilizza il metodo *GreedyStepwise*.

Come precedentemente sottolineato, la metodologia del *CfsSubsetEval* appartiene alla categoria dei *filter* multivariati e quindi ci permette di studiare come le variabili apportano informazione in modo congiunto.
Dall'implementazione otteniamo che il *subset* più rilevante contiene solo sei variabili ovvero: *education.num*, *marital.status*, *occupation*, *relationship*, *capital.gain* e *capital.loss*.
Quest'ultimo viene utilizzato per istruire i classificatori precedentemente impiegati nell'analisi sul *dataset* equilibrato attraverso il metodo dell'*oversampling* in modo da verificare se si ottiene una migliore performance. I risultati ottenuti sono:

	Accuracy	Recall	Precision	F-measure	AUC
J48	0,799	0,835	0,565	0,674	0,891
Random Forest	0,798	0,824	0,565	0,67	0,897
Decision Tree	0,798	0,829	0,564	0,671	0,893
Logistic	0,784	0,827	0,544	0,657	0,89
Simple Logistic	0,784	0,827	0,544	0,656	0,89
SMO poly	0,766	0,848	0,519	0,643	0,793
Multilayer Perceptron	0,777	0,853	0,533	0,656	0,894
SPegasos	0,563	0,986	0,362	0,529	0,875
Naive Bayes	0,819	0,501	0,686	0,579	0,874
Bayes Net BNC	0,8	0,845	0,566	0,678	0,908
Bayes Net TANB	0,805	0,847	0,574	0,684	0,911
NBTree	0,796	0,853	0,56	0,676	0,908

Figura 12. Risultati dei modelli di classificazione sul *dataset* costituito unicamente da 6 attributi.

Come nell'analisi precedente, per valutare la bontà dei modelli si presterà maggiore attenzione a due misure: la *recall* e l'AUC.

Osservando queste misure, notiamo che alcuni dei classificatori precedentemente individuati come più performanti, ovvero *Bayesian Network (*con come *searchAlgorithm K2* e *TAN)* e *Naive Bayes Tree*, continuano ad esserlo. Il *Simple Logistic*, invece, pur rimanendo complessivamente un buon modello, risulta essere meno preferibile rispetto ad altri.

Un'altra cosa che si può osservare è che la performance del Multilayer Perceptron migliora molto sia dal punto di vista della *recall* che da quello dell'AUC rispetto ai casi di *dataset* bilanciato e sbilanciato.

Una nota di particolare rilievo, infine, deve essere fatta per il *Naïve Bayes* che in precedenza era stato individuato come uno dei modelli più stabili in quanto dava risultati ottimi sia nel caso di assenza che di presenza di bilanciamento. Infatti, esso è l'unico classificatore che peggiora dal punto di vista della *recall*.

Infine, si sottolinea che l'*SPegasos* continua a prevedere quasi esclusivamente la classe minoritaria.

7. Conclusioni

In questo elaborato, attraverso l'uso di tecniche di apprendimento supervisionato si è cercato di classificare al meglio gli individui sottoposti al censimento nel 1994 negli Stati Uniti d'America, al fine di poter comprendere quale fosse la loro classe di reddito.

Dopo aver eseguito le prime analisi descrittive ed esplorative, si è effettuata la pulizia del dataset e la riduzione della sua dimensionalità. Grazie a queste prime analisi è emerso che la gran parte della popolazione studiata apparteneva alla classe meno agiata o comunque alla classe media (reddito "*<= 50K*"), mentre solo pochi individui facevano parte del ceto più abbiente.

Ritenendo più importante per un governo conoscere quali fossero le persone appartenenti a quest'ultima classe al fine di aumentare la pressione fiscale gravante su di esse e di garantire anche un'ipotetica diminuzione delle tasse per gli individui con entrate annue più ridotte, ci si è focalizzati sul cercare di prevedere correttamente la parte più ricca della popolazione.

Si sono, quindi, implementati i medesimi modelli dapprima sul dataset sbilanciato e poi su quello bilanciato attraverso l'*oversampling*, verificando effettivamente un aumento della performance dei classificatori nell'individuare la classe rara. Sono stati commentati e comparati i risultati più soddisfacenti e da queste analisi è emerso come classificatore più performante la rete bayesiana con *SearchAlgorithm K2* in quanto esso è l'algoritmo che riesce a classificare meglio la classe *target* e che ha una capacità discriminante estremamente elevata.

Successivamente è stata implementata la *feature selection* in modo tale da verificare se eventualmente vi fosse un altro metodo per fare selezione degli attributi che garantisse migliori prestazioni da parte dei classificatori. L'implementazione degli algoritmi sul *dataset* costituito unicamente dalle sei variabili rilevanti ha consentito un indubbio aumento nella capacità di individuazione dei soggetti più abbienti, ma di fatto i risultati ottenuti avvallano la nostra tesi secondo cui tra i modelli più performanti per effettuare questo tipo di analisi sono da annoverare senza dubbio le reti bayesiane. Tuttavia, bisogna anche riconoscere che l'algoritmo in assoluto più performante non fa parte di questa categoria, ma è un *Naïve Bayes Tree* come dimostrato da Ron Kohavi nel 1996.

L'analisi qui condotta potrebbe senza dubbio essere ampliata e migliorata.

Sotto la guida di esperti in materia di censimento e di tassazione, potrebbe essere interessante costruire una matrice di costo in modo tale da poter andare ad individuare i modelli più performanti in questa casistica.

Da ultimo, nonostante nel nostro tentativo di applicazione non si siano raggiunti risultati degni di nota, potrebbe essere utile utilizzare tecniche più sofisticate di *clustering* in modo da individuare particolari legami tra i dati.

8. Riferimenti bibliografici

1. Tan, P. e **Steinbach**, M e **Kumar**, V. (2006) Capitolo 5: Classification: Alternative Techniques. *Introduction to Data Mining*. 316-323.

2. http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names, sito visitato il 10 maggio 2018.

3. Figura 8 - source: https://rasbt.github.io/mlxtend/user_guide/evaluate/confusion_matrix/, sito visitato il 15 maggio 2018.

4. Figura 9 - source: http://gim.unmc.edu/dxtests/roc3.htm, sito visitato il 18 maggio 2018.

5. Kohavi, R (1996) Scaling Up the Accuracy of Naïve-Bayes Classifiers: a Decision-Tree Hybrid.

6. https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/, sito visitato il 25 maggio 2018.

7. Gingrich, P http://uregina.ca/~gingrich/ch11a.pdf, sito visitato il 25 maggio 2018.