LAVORO IN USA: GRANDI OPPORTUNITA` E PREMIO AL MERITO

Eva Pelagalli[1], Dario Piovesana[2]

Abstract

State cercando la facoltà a cui iscrivervi in università o state cercando lavoro?
Se avete in mente di andare a studiare o a lavorare negli USA, ci sono buone possibilità.
L'economia statunitense infatti, si caratterizza per un forte dinamismo occupazionale e per l'offerta di un ventaglio molto ampio di professioni, sebbene non tutti i mestieri siano richiesti in egual misura.
Prima di richiedere il visto, è necessario porsi qualche domanda: quali sono le professioni più ambite negli USA? Quanto guadagnano gli stranieri? Ci sono discriminazioni razziali e di genere?
Il presente lavoro vuole analizzare la situazione lavorativa degli USA, l'origine della ricchezza, la distribuzione dei redditi, il contesto socio-economico e sviluppare un modello previsivo al fine di indirizzare chi fosse interessato a trasferirsi in America, nella scelta di un lavoro redditizio.

[1]Università degli Studi di Milano Bicocca, CdLM CLAMSES
[2]Università degli Studi di Milano Bicocca, CdLM CLAMSES

Indice

Introduzione

Per un'analisi approfondita sul mercato del lavoro in USA è stato analizzato il dataset *Adult* (Tabella 1) reso disponibile dall'ufficio censimenti degli USA e contente informazioni sul lavoro. E`una raccolta di 32561 osservazioni e quindici variabili.
La variabile *Income* misura il reddito lordo prima del pagamento delle imposte ed è stata trattata come variabile di riposta ovvero, sulla base delle informazioni disponibili dalle restanti variabili, si è cercato di capire, con opportune tecniche di *Machine Learning*, se una persona riuscirà a guadagnare più o meno di $50000 l'anno. La presente analisi è strutturata nel seguente modo:

Tabella1: dataset Adult

Variabili	Descrizione
age	età
workclass	categoria della professione
fnlwgt	peso in base alle caratteristiche demografiche
education e education num	livello di istruzione
marital status e relationship	relazione familiare
occupation	occupazione
race	razza
sex	sesso
capital gain	plusvalenza degli investimenti
capital loss	minusvalenza degli investimenti
hours per week	ore di lavoro settimanali
native Country	Pease di origine
income	<$50000 >$50000

- ✓ nel primo paragrafo sono riportati risultati di una prima analisi esplorativa e descrittiva;

- ✓ nel secondo, il dataset è stato sottoposto a tecniche di *data processing*;
- ✓ nel terzo sono stati implementati e migliorati i modelli di classificazione;
- ✓ nel quarto sono riportate le conclusioni e gli ulteriori sviluppi.

1 Esplorazione e visualizzazione

Da prime analisi descrittive ed esplorative, è emerso che:

- ✓ il 76% dei lavoratori guadagna meno di $50000 l'anno;
- ✓ il 74% dei lavoratori ha un impiego nel settore privato;
- ✓ prevale la famiglia tradizionale con più del 46% di lavoratori sposati.

Si è cercato di capire graficamente quali fossero le variabili con un maggiore impatto sulla variabile di risposta.
Nel *bar plot* della Figura 1 il colore rosso indica un reddito inferiore a $50000, il celeste superiore a $50000: è evidente che nelle famiglie con un coniuge assente o con coniugi separati si registrano redditi più bassi.

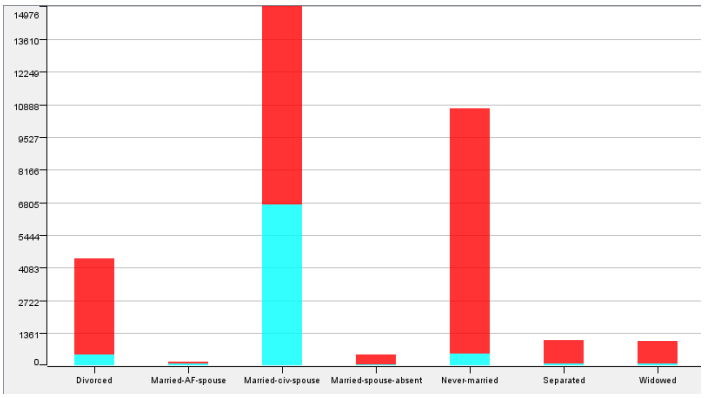

Figura1 Bar plot marital status

Anche il numero di anni di studio sembra influenzare la variabile di risposta. Infatti, nel *box plot* della Figura 2 è dimostrato che il reddito tende ad aumentare con il grado d'istruzione (soprattutto dopo un *master* o un dottorato o in seguito ad una laurea conseguita in una *professional school*[1]).
Dai dati della Figura 3 è inoltre emerso che, le professioni più redditizie, sono manageriali e professionali, cui seguono quelle di supporto tecnico. Le professioni meno redditizie sono quelle manuali e presso le Forze Armate; quelle mediamente redditizie sono i lavori di ufficio, tecnici e presso l'*American Federal Protective Service*.

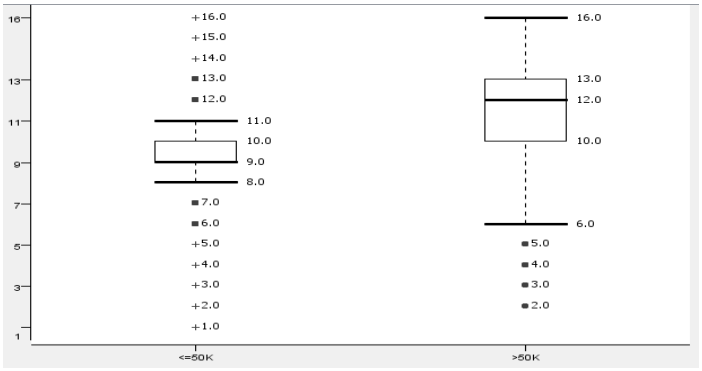

Figura2 Box plot education num

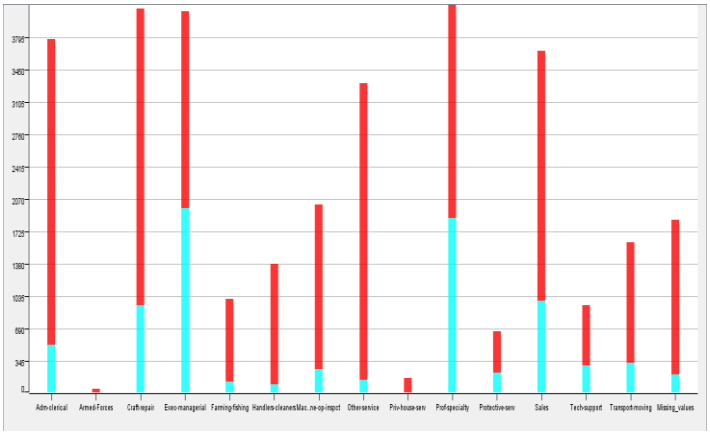

Figura3 Bar plot occupation

I valori più elevati per la variabile *hours per week*, tendono a registrarsi in corrispondenza della classe più elevata di reddito, tuttavia la mediana pari a quaranta ore settimanali, è la stessa per le due classi. Si noti che un cospicuo numero di lavoratori ha un reddito inferiore a $50000 pur in corrispondenza delle modalità più alte (Figura4). Ciò fa pensare che il tipo di professione abbia un impatto più forte sulla variabile di risposta rispetto al numero di ore di lavoro.

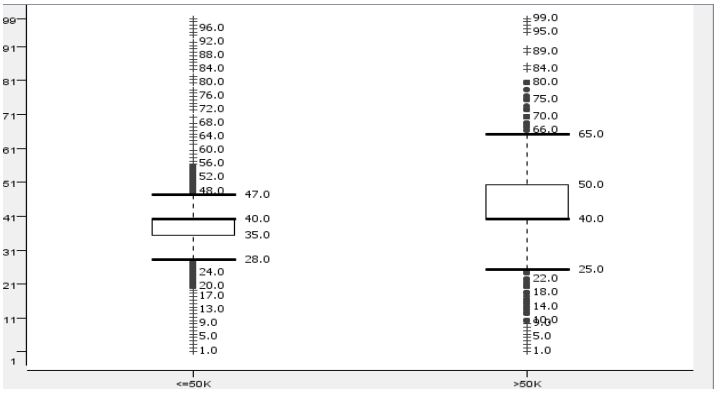

Figura4 Box plot hours per week

[1] Studi necessari per accedere alle professioni mediche e legali

La variabile *age* conferma che il reddito cresce con l'età: le due mediane sono trentaquattro e quarantaquattro per le due classi di reddito (Figura5).

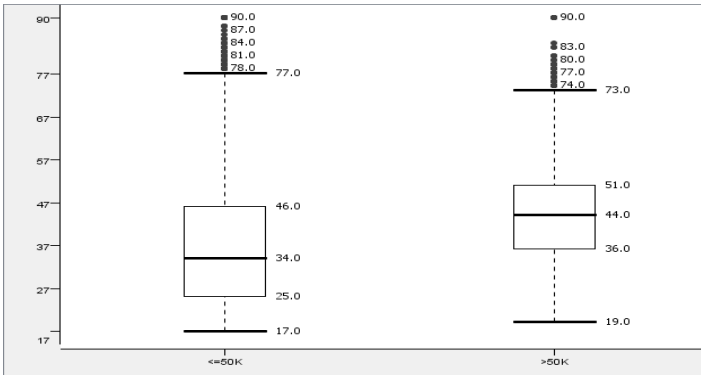

Figura5 Box plot age

Le donne e gli uomini che guadagno più di $50000 sono rispettivamente l'11% e il 31% (Figura6)

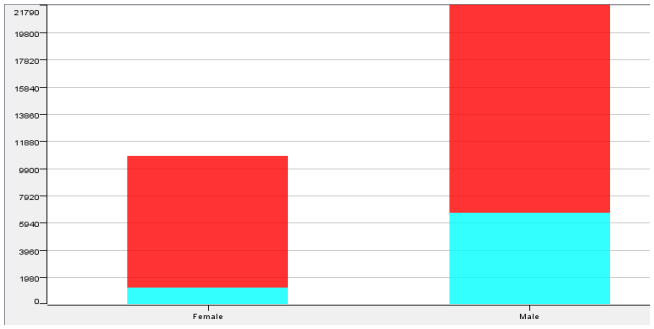

Figura6 Bar plot sex

I nativi americani, gli asiatici, i neri, i bianchi e la restante categoria *other* con più di $50000 annui registrano le seguenti percentuali: 11.58%, 26,56%, 12.39%, 25.59% e 9.22% (Figura 7).

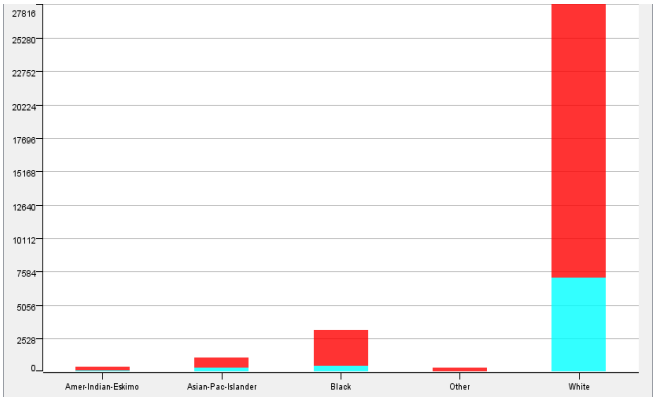

Figura7 Bar plot race

Le Figure 8 e 9 descrivono le variabili *capital gain* e *capital loss*. Le due distribuzioni sono entrambe asimmetriche verso destra a indicare che la gran parte dei lavoratori ha una propensione bassa ad investire o non investe affatto. In particolare, chi ha avuto plusvalenze più alte, appartiene alla classe di reddito superiore, chi invece ha perso grandi capitali, appartiene alla classe inferiore.

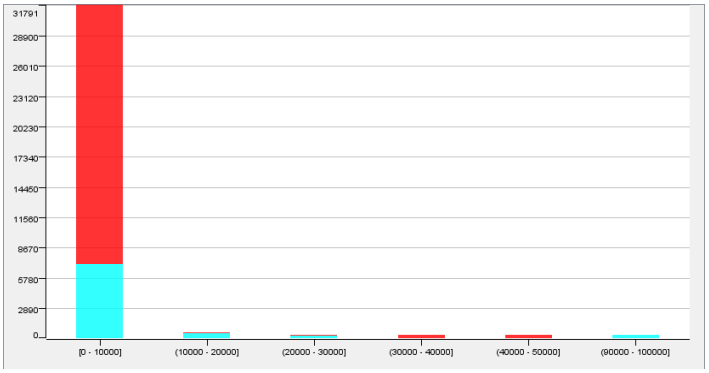

Figura8 Bar plot capital gain

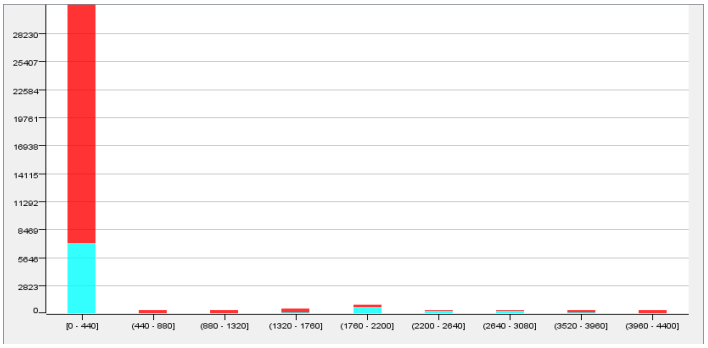

Figura9 Bar plot capital loss

2 *Preprocessing*

2.1 *Missing replacement*

Prima di implementare i modelli di classificazione, sono stati sostituiti 4262 dati mancanti con valori plausibili.

In particolare, la variabile *workclass* ne ha 1836, *occupation* 1843 e *native country* 583. La variabile *occupation* ha quindi sette valori mancanti in più rispetto a *workclass*. E`stato notato che, in corrispondenza di queste sette osservazioni mancanti, la variabile *workclass* assume la modalità *never worked*[2]. Allora è ragionevole supporre che costoro, non lavorando, non sappiano rispondere alla domanda "Qual è la tua occupazione?", pertanto è stata creata una modalità chiamata *none* per indicare "nessuna occupazione".

Poiché i restanti valori mancanti delle variabili *workclass* e *occupation* (1836 ciascuna) sono riferiti agli stessi individui, sono stati sostituiti con una nuova modalità chiamata *ND* per indicare rispettivamente una categoria e un'occupazione non

[2] La modalità *never worked* è assunta solo da queste sette unità

dichiarata. Si è ritenuto utile non cancellare queste unità giacché il loro 90% appartiene alla classe inferiore di reddito; quindi i valori mancanti sono stati trattati come una modalità in grado di discriminare le due classi. Invero, un motivo plausibile per cui una persona non dichiara la sua occupazione è che abbia una situazione economica non molto stabile (magari cambia spesso lavoro), è quindi facile che abbia un reddito inferiore a $50000.
Per quanto riguarda la variabile *native country*, i valori mancanti sono stati sostituiti con la moda *United-States*, poiché osservando le frequenze percentuali dei lavoratori statunitensi rispetto alle due modalità della variabile Income, e quelle dei lavoratori di cui non si conosce la nazionalità, sempre rispetto alle due modalità della variabile di risposta, è emerso che la percentuale di coloro che hanno un reddito >$50000 è in entrambi i casi pari al 25%. Con questa sostituzione, non sono quindi alterate le percentuali della modalità *United-States* rispetto alla variabile Income.

2.2 *Feature creation*

Una volta completata questa fase, è stato deciso di eliminare le variabili non utili alla discriminazione fra le due classi di reddito, di semplificare quelle utili creandone nuove *ad hoc.*
Nella Tabella 2 sono riportate nuove variabili e nuove modalità ottenute sintetizzando le variabili originali. Le variabili con un significato simile sono state fuse in un'unica variabile, le cui nuove modalità sono state definite in base alla somiglianza che le precedenti modalità hanno rispetto alla variabile di risposta.
Per esempio, per la variabile *native country* sono state definite tre modalità: *rich*, *poor* e *United-States*. Il criterio di scelta tra *rich* e *poor* è stato classificare in *rich* le nazioni che avessero più del 20% di lavoratori con reddito superiore a $50000[3], *poor* in caso contrario.
Le variabili *relationship* e *marital status* sono state fuse in *family type*, mentre *capital gain* e *capital loss* nella nuova variabile *dummy* chiamata *investement*, dove la modalità uno è definito in corrispondenza di valori non nulli di *capital gain* e *capital loss*. Questa variabile individua le persone che hanno una parte di reddito generata da investimenti e non da lavoro.

Tabella2 *Feature creation*

NUOVA VARIABILE	NUOVE MODALITA`
native country	poor, rich, United-States
workclass	gov+self-not, ND, federal, private, self-emp-inc, no worker
education	no graduated, assoc, doc+ prof, bachelors, HS-grad, masters, some-college
family type	family with chidren, married, one parent with child, single
investements	1 ,0
occupation	low paid, hand+other, middle paid, trans+rep, sup+serv, manager+spec,

La Figura 10 mostra che tra i lavoratori che non investono, il 19% guadagna più di $50000, mentre tra i lavoratori che investono, questa percentuale arriva a 58%. Ciò vuol dire che i più ricchi tendono a investire di più o che gli investimenti sono una fonte di ricchezza.

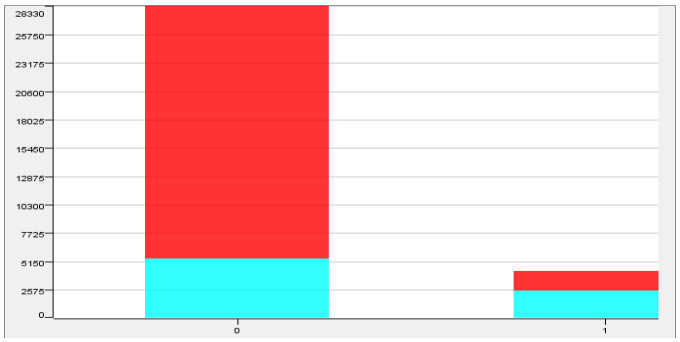

Figura10 Bar plot Investement

La variabile *fnlwgt* è stata eliminata poiché dal *box plot* è stato visto che non è in grado di discriminare in modo opportuno tra le due modalità della variabile di risposta.
Le variabili *education* e *education num* sono state messe insieme poiché simili. Come nuove modalità, sono stati aggregati tutti i livelli inferiori al diploma e distinti con maggiore precisione quelli superiori.
Sebbene *workclass* e *occupation*, abbiano un significato simile, è stato deciso di non aggregarle e di comporne separatamente le modalità.
Le restanti variabili non sono state modificate poiché già strutturate correttamente.

[3] Una sola eccezione è stata fatta l'Olanda, classificata in *rich*, benché l'unica lavoratrice olandese avesse un reddito appartenente alla classe inferiore di reddito.

2.3 *Feature selection* e misure di valutazione

Una volta giunti a questo punto, al fine di migliorare l'interpretabilità della classificazione, è opportuno individuare le variabili più utili, eliminando le ridondanti e le irrilevanti.
A tale scopo, è stato diviso il dataset in due parti chiamate *train* e *test*, di cui la prima contiene il 70% delle osservazioni e la seconda il rimanente 30%. Dal momento che vi è un problema di sbilanciamento nella variabile di risposta, ovvero i lavoratori con un reddito superiore a $50000 occupano una percentuale piuttosto ridotta, è stato utilizzato un campionamento stratificato per selezionare le unità sia del *train* sia del *test*. In tal modo, vi è garanzia che le unità rare siano presenti in fase di *training* e si migliori così l'accuratezza previsiva della classe minoritaria. In assenza di una correzione al problema del bilanciamento, le previsioni sarebbero non soddisfacenti.
È stato eseguito allora, un confronto tra il filtro univariato, il filtro multivariato (entrambi con modello *J48)*, il filtro multivariato con *random forest* e il *Wrapper* con modello *J48*. Come misura di associazione tra gli attributi candidati, è stato scelto l'*InfoGain* per il filtro univariato, *CfsSubset* per i due filtri multivariati e il *WrapperSubset* per il *Wrapper*. Per un confronto oggettivo, è stato usato lo stesso seme prima di implementare i tre filtri e il Wrapper. Per selezionare le variabili più importanti, sono state usate l'*accuracy*, l'*AUC* e l'*F-measure* come misure di valutazione.
Quest'ultima, è particolarmente utile nei casi di classificazione in presenza di una classe rara. Infatti, l'*accuracy* misurando la percentuale di classificazioni corrette, può essere alta anche quando il classificatore non è in grado di classificare correttamente le unità rare, se riesce però a essere molto performante nella classificazione delle unità della classe maggioritaria. L'*AUC* e l'*F-measure* invece, pongono maggiore enfasi sul problema della classe rara. L'*AUC* è l'area sotto la curva roc, la quale si costruisce dalla sensibilità e da 1- specificità, dove la sensibilità è la frazione delle unità della classe minoritaria predette correttamente, la specificità è la frazione delle unità della classe maggioritaria predette senza errori dal classificatore; l'*AUC* sarà tanto più elevata quanto più queste misure saranno congiuntamente elevate.
L'*F-mesure* si ricava dalla media armonica di *precision* e *recall*, dove la prima, è la frazione delle unità rare predette correttamente tra queste ultime e *recall* coincide con la sensibilità.

Dal confronto delle misure di valutazione (Tabella 3), è risultato che il filtro multivariato *J48* ha i migliori risultati di *accuracy* e *F-measure*. Generalmente un filtro multivariato è migliore di quello univariato perché seleziona le migliori variabili in seguito a un'analisi congiunta che permette di ottenere variabili non correlate fra loro e fortemente associate con la variabile di risposta.
Invece, il *Wrapper J48* è risultato migliore tra le performance avute usando come misura di valutazione l'*AUC*.

Tabella3 Misure di valutazione per la *future selection*

	Accuracy	AUC	F-measure
Filtro univariato J48	0.841	0.856	0.625
Filtro multivariato J48	0842	0.868	0.636
Filtro multivariate RF	0.824	0.849	0.611
Wrapper J48	0.836	0.875	0.616

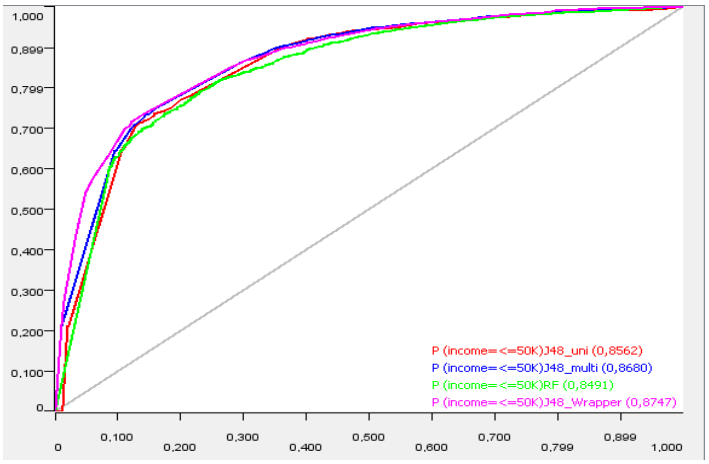

Figura 11 Curva ROC per la Future Selection

Le variabili selezionati da questi due metodi sono riportate in Tabella 4.

Tabella4 Risultati della future selection

Modelli	Variabili selezionate
Filtro multivariato J48	age, education, occupation, family_type, investments
Wrapper J48	hours per week, education, occupation, family type, investments

I due modelli selezionano le stesse variabili con la differenza di *age* per il filtro multivariato *J48* e *hours per week* per il *Wrapper J48*. La somiglianza tra questi due risultati conferma la robustezza dell'importanza delle variabili scelte. In entrambi i casi, né la variabile *sex*, né la variabile *race* è stata selezionata tra le più importanti per prevedere la variabile di risposta, il ché vuol dire che dai dati risulta che, negli USA, non ci sono discriminazioni razziali o di genere, piuttosto l'avere un reddito superiore o inferiore a $50000 sembra dipendere dal grado d'istruzione, dal tipo di lavoro, dall'età, dal numero di ore di lavoro, dalla famiglia e dalla decisione o meno di investire la propria ricchezza.

4 Modelli di classificazione

4.1 Classi e implementazione dei modelli

Si è ritenuto opportuno applicare i modelli di classificazione la prima volta utilizzando le variabili selezionate dal *filtro multivariato J48* e la seconda volta le variabili selezionate con il *Wrapper J48* per poi confrontare i risultati.
Allora, il dataset è stato suddiviso in *train* e *test* di dimensioni pari a 70% e 30% applicando sempre un campionamento stratificato per le ragioni illustrate in precedenza. Al fine di eseguire un confronto oggettivo, è stato usato lo stesso seme per tutti i modelli di classificazione che sono stati raggruppati in quattro classi, come riportato di seguito:

- ✓ **HEURISTIC** (*Decision Trees*, *Random Forest*)
- ✓ **REGRESSION BASED** (*Logistic*)
- ✓ **SEPARATION** (*Support Vector Machine*, *Artificial Neural Networks*)
- ✓ **PROBABILISTIC** (*Naive Bayes Classifiers*, *Bayesian Net Classifiers*)

Nella prima classe, è stato usato il modello *Decision Tree* implementato da *Knime* e come misura è stata scelta l'indice di Gini. E`stato implementato anche il modello *J48* del software *Weka* e sono stati scelti tre *folds*. Per la *Random Forest*, sempre implementata da *Weka*, il numero di alberi è stato scelto pari a dieci.
Nella seconda classe, sono stati definiti due modelli implementati dal software *Weka*: *Logistic* e *SimpleLogistic* per eseguire una regressione logistica.
Nella terza classe sono stati implementati i modelli *SPegasos* e *SMO* del software *Weka* (con *kernel=PolyKernel*) per la *Support Vector Machine* e il *Multilayer Perceptron* di *Weka* come modello di *Artificial Neural Network*.
Nella quarta classe, sono stati usati i modelli *Naive Bayes Learner* e *Naive Bayes*, il primo di *Knime* e l'altro di *Weka*, per definire un *Naive Bayes Classifier*. In seguito, sono stati sviluppati i modelli *BayesNet* (con algoritmo *TAN*), *BayesNet* (con algoritmo *K2*) e *NBTree* per definire invece un *Bayes Net Classifier*.
Nelle Tabelle5 e 6, sono riportati i valori di *accuracy*, di *AUC* e di *F-measure* ottenuti con le variabili selezionate con il filtro multivariato *J48* e il *Wrapper J48.*

Tabella5 Misure di valutazione dei modelli di classificazione con filtro multivariato J48

Modelli	Accuracy	AUC	F-measure
Decision Tree	0.823	0.853	0.606
J48	0.84	0.87	0.633
Random Forest	0.819	0.844	0.599
Logistic	0.838	0.887	0.636
Simple Logistic	0.839	0.887	0.638
SPegasos	0.84	0.887	0.636
SMO	0.84	0.751	0.635
MLP	0.839	0.889	0.617
Naive Bayes Learner	0.84	0.889	0.641
Naive Bayes	0.837	0.889	0.652
Bayes Net TAN	0.834	0.89	0.64
NBTree	0.831	0.888	0.65
Bayes Net K2	0.831	0.888	0.65

Tabella6 Misure di valutazione dei modelli di classificazione con Wrapper J48

Modelli	Accuracy	AUC	F-measure
Decision Tree	0.831	0.868	0.614
J48	0.838	0.873	0.636
Random Forest	0.827	0.863	0.615
Logistic	0.836	0.888	0.627
Simple Logistic	0.836	0.887	0.627
SPegasos	0.838	0.887	0.631
SMO	0.836	0.746	0.627
MLP	0.837	0.885	0.608
Naive Bayes Learner	0.837	0.884	0.642
Naive Bayes	0.834	0.883	0.647
Bayes Net TAN	0.835	0.889	0.625
NBTree	0.83	0.883	0.641
Bayes Net K2	0.83	0.883	0.641

Dal confronto dei valori delle tre misure di valutazione dei modelli sviluppati, prima con le variabili selezionate dal filtro multivariato e poi con quelle selezionate dal *Wrapper*, sebbene i risultati siano piuttosto simili, è emerso che gli scores più alti si sono avuti utilizzando la variabile *age* piuttosto che *hours per week*.

In Tabella 5 sono evidenziati in giallo, verde e celeste i valori più elevati ottenuti, in corrispondenza delle tre misure di valutazione.

4.2 Modelli migliori con *cross validation* e *iterated holdout*

Giacché i risultati ottenuti con i modelli sviluppati in precedenza, dipendono fortemente dalla scelta del *test set*, è stato deciso di implementare una *cross validation* e un *iterated holdout* al fine di capire quali modelli fossero più performanti, variando ad ogni passo dell'algoritmo il *test set*.
Dato che, sia la *cross validation* sia l'*iterated holdout*, sono piuttosto onerosi computazionalmente, sono stati scelti i modelli con una migliore *AUC* per ognuna delle quattro classi utilizzando le variabili selezionate dal filtro multivarato *J48*. Sono quindi stati scelti i modelli: *J48*, *Simple Logistic*[4], *MLP* e *Bayes Net* con algoritmo *TAN*. Il numero di iterazioni è stato fissato a trenta per tutti i modelli ed è stato scelto lo stesso seme per un confronto oggettivo.
Con la *cross validation*, il dataset viene diviso in k parti di cui una è il *test* e le restanti formano il *train*. Si allena il modello sul *train* e si testa sul *test*. Tale procedimento è ripetuto k volte, in modo tale che ognuno dei k insiemi abbia avuto il ruolo di insieme di verifica. Per ognuno dei quattro modelli, la *cross validation* permette di ottenere la previsione per ogni unità, poiché questa solo una volta entra nel *test*, dunque ad ogni unità corrisponde una previsione. I risultati della *cross validation* sono riportati in Tabella 7.
Dai risultati, il modello *Bayes Net* raggiunge gli scores più alti sia per l'*AUC* sia per l'*F-measure* e ha risultati alquanto analoghi con i modelli *J48* e *Simple Logstic* per quanto concerne l'*accuracy*.
Considerando congiuntamente tutte e tre le misure di valutazione, la *cross validation* conferma il *Bayes Net*, come modello ottimale per risolvere il problema di classificazione binaria. I modelli *Simple Logistic* e *MLP* hanno risultati piuttosto simili, mentre il modello *J48* si distacca maggiormente con il minimo valore di *AUC*. In Figura 12, la curva *ROC* illustra graficamente i risultati dell'*AUC*.
Nel caso dell'*iterated holdout* invece, ad ogni passo dell'algoritmo, il *train* e il *test* cambiano senza alcuna garanzia che le unità finiscano una sola volta nel *test*. Allora, sono stati sviluppati trenta modelli separatamente per *J48*, *SimpleLogistic*, *MLP* e *Bayes*

[4] Sia *Logistic*, sia *Simple Logistic* hanno le medesime prestazione in termini di *AUC*, dalle altre due misure di valutazione, sembra più performante *Simple Logistic*

Tabella7 Misure di valutazione dei modelli di classificazione con CV

	J48	SL	MLP	BN
Accuracy	0.839	0.839	0.835	0.838
AUC	0.864	0.887	0.888	0.891
F-measure	0.632	0.628	0.631	0.639

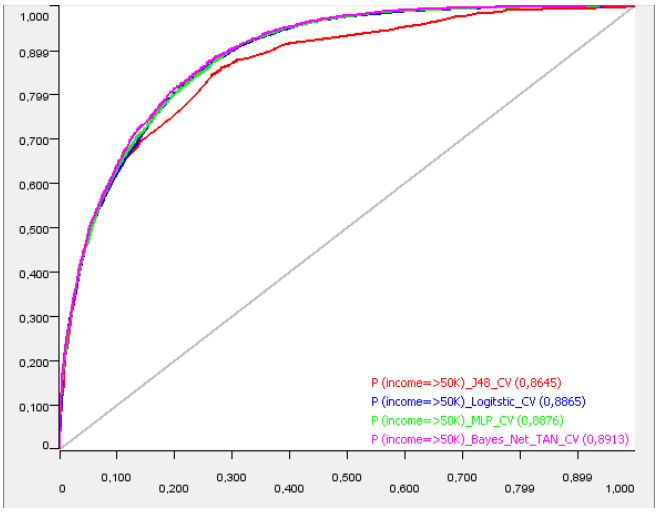

Figura12 Curva ROC modelli con CV

Net. Ad ogni passo dell'algoritmo, sono state selezionate, con campionamento stratificato, le unità del *train* e del *test*; per avere un *train* e un *test* diversi ogni volta, non è stato fissato il seme. Pertanto, i trenta modelli ottenuti alla fine di una prima simulazione sono leggermente diversi da altri trenta generati da una seconda simulazione. Invero, questi risultati sono però piuttosto simili perché i modelli sono stati testati ben trenta volte, con la conseguenza che la differenza nei risultati tra una simulazione e l'altra è minima. Come nel caso della *cross validation*, anche l'*iterated hold out* ha confermato il modello *Naive Bayes* come il migliore a risolvere un problema di classificazione binaria secondo le misure di *AUC* e *F-measure*. Di seguito, nelle Figure 13, 14 e 15 sono riportati i *box plot* delle tre misure di valutazione di una delle simulazioni implementate.
I risultati sono conformi a quelli della cross validation, ovvero:

- ✓ secondo l'*accuracy*, i modelli *Bayes Net*, *J48* e *Simple Logistic* sono similmente performanti, mentre il *MLP* ha una varianza molto più grande in confronto;
- ✓ secondo l'*AUC*, il modello *J48* è il peggiore, il migliore è il *Bayes Net* con i minimi e i

massimi più alti, mentre i restanti modelli tendono ad assomigliarsi;

- ✓ secondo l'*F-measure*, il *Bayes Net* è il migliore, il MLP il più variabile e il *J48* poco più performante del *Simple Logistic*.

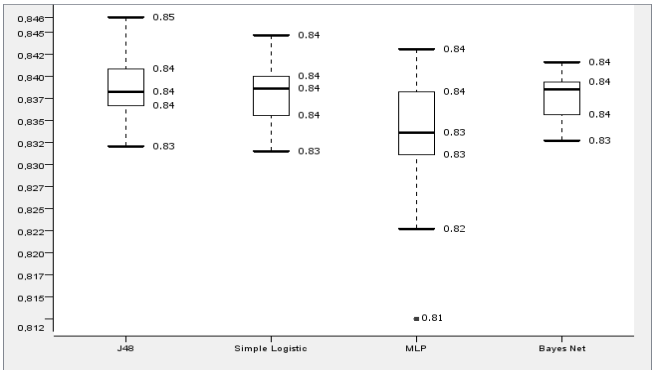

Figura13 Box plot dell'Accuracy con metodo iterated holdout

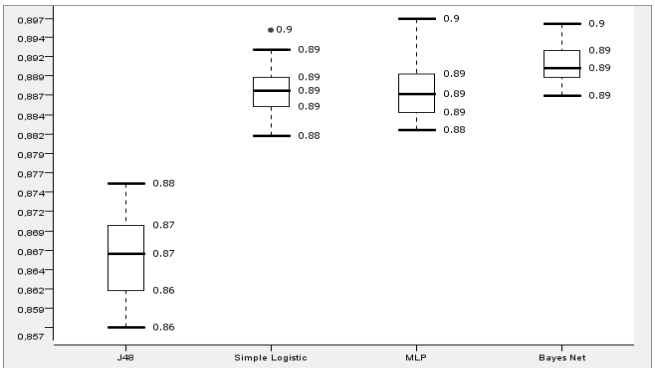

Figura14 Box plot dell'AUC con metodo iterated holdout

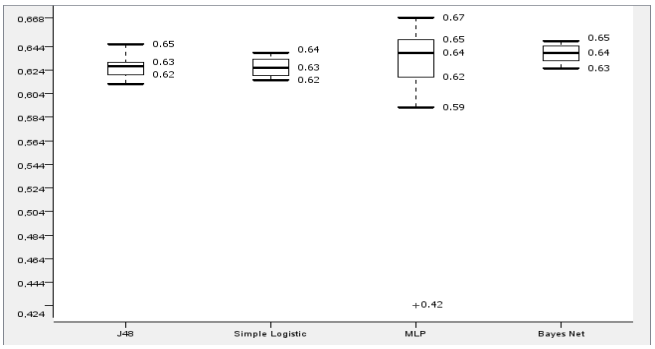

Figura15 Box plot dell'F-measure con metodo iterated holdout

Le potenzialità della *cross validation* e dell'*iterated hold out*, hanno consolidato i valori delle tre misure di valutazione. Dovendo però scegliere un modello, il *Naive Bayes* con algoritmo *TAN* è il più performante in accordo con tutte e tre.

5 Conclusioni

In conclusione, mediante opportune tecniche di *Machine Learning*, il presente lavoro ha cercato di individuare l'origine della ricchezza negli Stati Uniti con una particolare attenzione al reddito generato da lavoro. Al fine di migliorare i risultati sono stati implementati più modelli e, dopo aver selezionato i migliori, sono stati sviluppati metodi per alzare i valori delle tre misure di valutazione. I modelli implementati e valutati, non sono però fini a sè stessi: possono essere riutilizzati per nuovi dati di cui non ci conosce la variabile di risposta, ma si dispone del resto delle altre variabili. Dall'esito della *feature selection*, è emerso che nel mercato del lavoro in Usa non ci sono significative discriminazioni di genere, di razza e di provenienza: il reddito tende invece ad essere più elevato in corrispondenza di professioni che richiedono un livello più alto d'istruzione e al crescere del numero di ore di lavoro. Inoltre, il reddito tende ad aumentare considerevolmente con l'esperienza, ciò è conseguenza del maggior impiego nel settore privato rispetto al settore pubblico. Infatti, in quello privato, l'avanzamento della carriera avviene generalmente in tempi più veloci. La selezione della variabile *family type* tra le variabili più significative, spiega che dietro ai redditi c'è un determinato tipo di famiglia e in particolare, nelle famiglie con coniugi sposati si registrano redditi più elevati.
Nell'ottica di un miglioramento dell'analisi, sarebbe interessante avere maggiori informazioni circa il tipo di investimenti che i lavoratori delle due classi intraprendono. E`noto infatti, che gli investimenti più rischiosi sono anche più redditizi. Nel dataset però, non si hanno informazioni sulla natura degli investimenti, quindi non si può sapere quali investimenti sono stati la causa di un aumento o una diminuzione della ricchezza in corrispondenza delle due classi di reddito. Un successivo sviluppo del presente lavoro potrebbe riguardare l'associazione che c'è tra i diversi Stati o le città rispetto alla variabile di risposta per poi esaminare quanto è intensa la mobilità della forza lavoro. Infine, la disposizione di informazioni circa la tassazione (la variabile *Income* misura il reddito lordo) e il sistema redistributivo della ricchezza in USA, possono completare l'analisi in esame.

5 Riferimenti

https://www.kaggle.com/uciml/adult-census-income

https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/