Data Science e Tecnologie per le Basi di Dati

Esercitazione #3 – Data mining

BOZZA DI SOLUZIONE

Domanda 1

(a) Come mostrato in Figura 1, l'attributo più selettivo risulta essere "Capital Gain", perché rappresenta il nodo radice dell'albero di decisione.

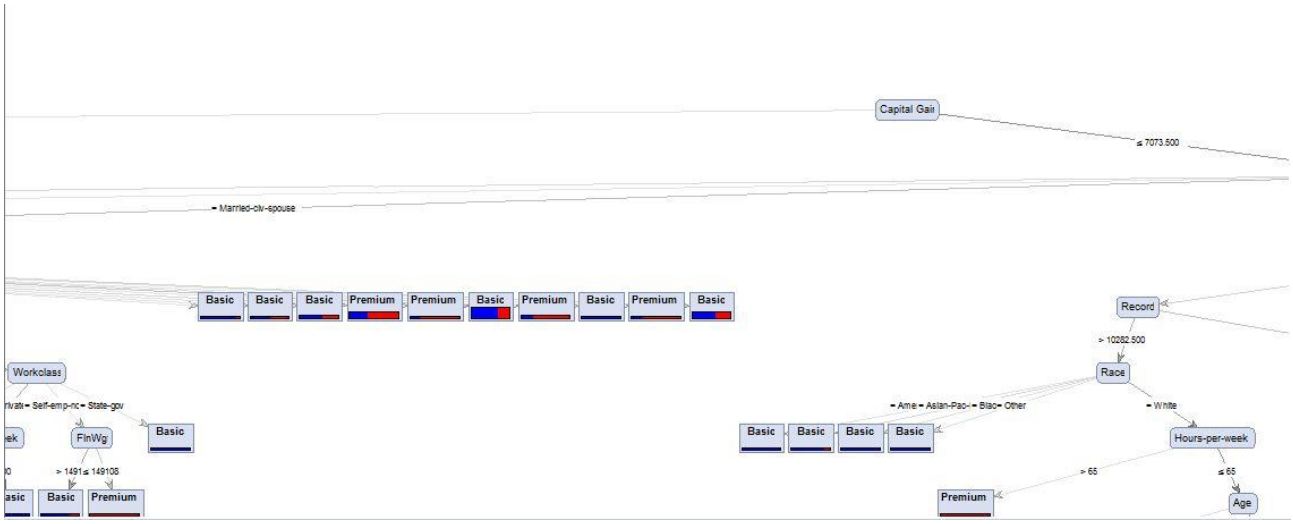


Figura 1

(b) L'altezza dell'albero, ovvero la lunghezza massima di un percorso che collega la radice ad una foglia dell'albero è 15.

(c) Un partizionamento puro è un split sui valori di un attributo tale per cui I record corrispondenti appartengono tutti alla medesima classe. Per esempio, consideriamo la porzione sinistra dell'albero di decisione rappresentato in Figura 2. I valori dell'attributo "Age" sono splittati in due gruppi: > 62.5 and <= 62.5. Mentre la prima partizione è "impura", perché copre record etichettati sia con la classe "Basic" sia con la classe "Premium", la seconda è pura perché tutte le relative istanze appartengono alla classe "Premium".

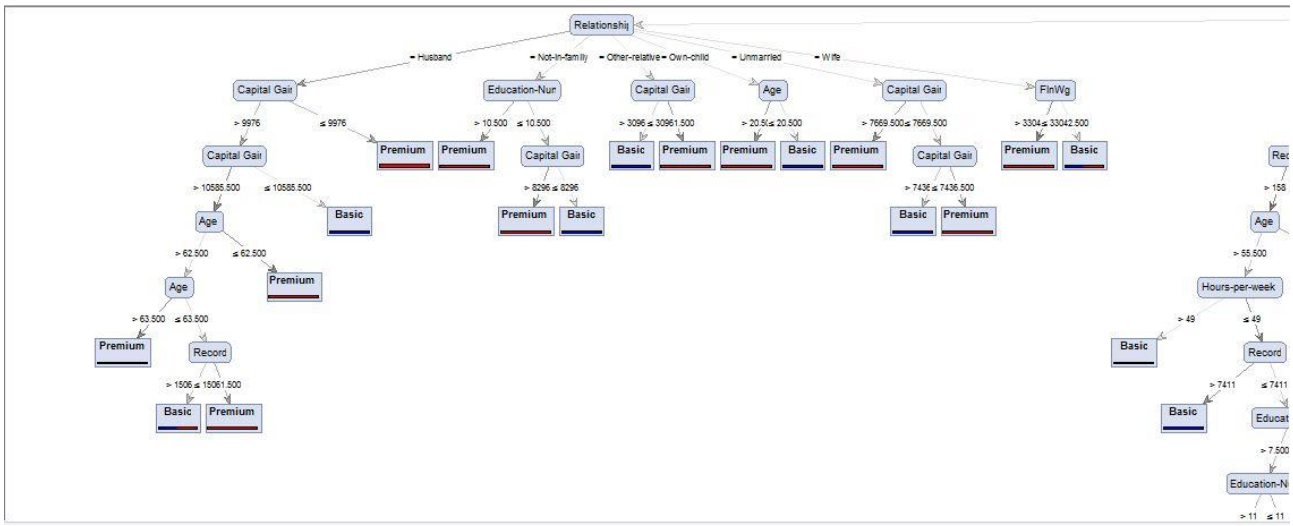


Figura 2

Domanda 2

Il parametro “maximal depth” permette di specificare l’altezza massima dell’albero di decisione generato. Usando la configurazione di default (con maximal depth = 20) l’altezza dell’albero risulta essere 15 e quindi il processo ricorsivo di learning dell’albero viene completato. Al contrario, settando un valore di maximal depth inferior a 15 (ad es. 5) la ricorsione viene interrotta e quindi la qualità del modello generato potenzialmente decresce.

Il parametro “minimal gain” permette di scegliere se splittare ulteriormente un nodo dell’albero oppure no. In particolare, un nodo viene splittato se il suo gain è superiore alla soglia minima (minimal gain). Valori elevati di minimal gain producono un numero limitato di partizionamenti e, di conseguenza, alberi di decisione più piccoli. Valori troppo elevati di minimal gain (ad es., 0.9) impediscono completamente lo split dei valori degli attributi e quindi l’albero risultante conterrà un singolo nodo. Dato che il valore di default di minimal gain è moderatamente basso (ad es., 0.1), esso produce generalmente uno splitting degli attributi abbastanza fitto.

Domanda 3

Impostando come attributo di classe “Native Country” l’analista può predire la nazionalità dell’utente che sottomette una nuova richiesta di servizio sulla base delle richieste passate e delle caratteristiche degli utenti che le hanno sottomesso.

Come mostrato in Figura 3, l’attributo più discriminante in questo caso diventa “Race” in quanto la razza è un informazione saliente per determinare la nazionalità di un individuo.

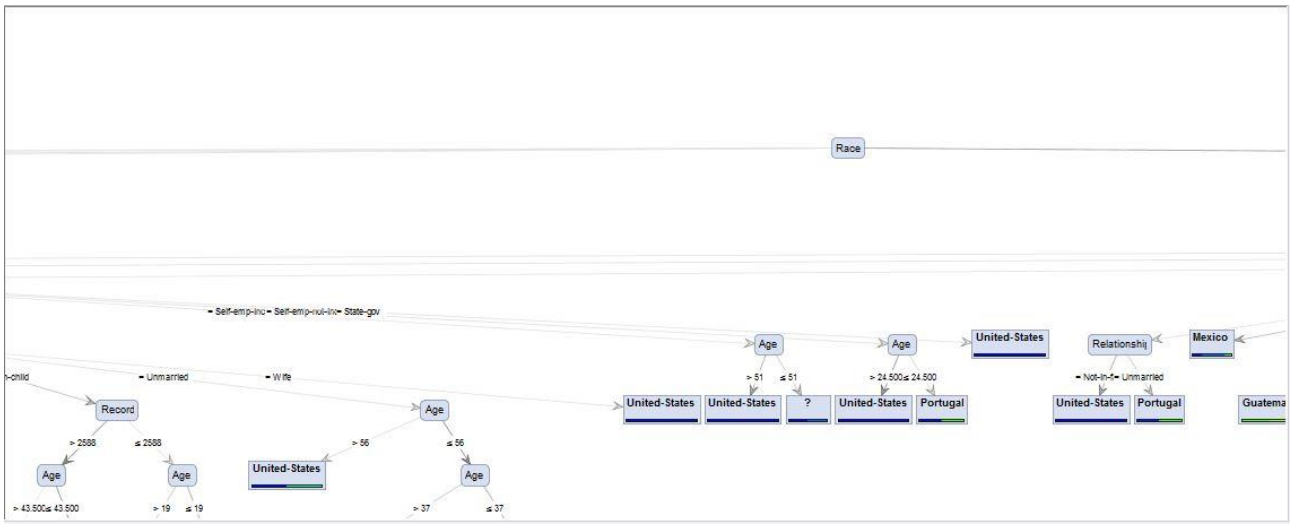


Figura 3

(a) Alcuni esempi di partizionamenti puri sono riportati in Figura 4.

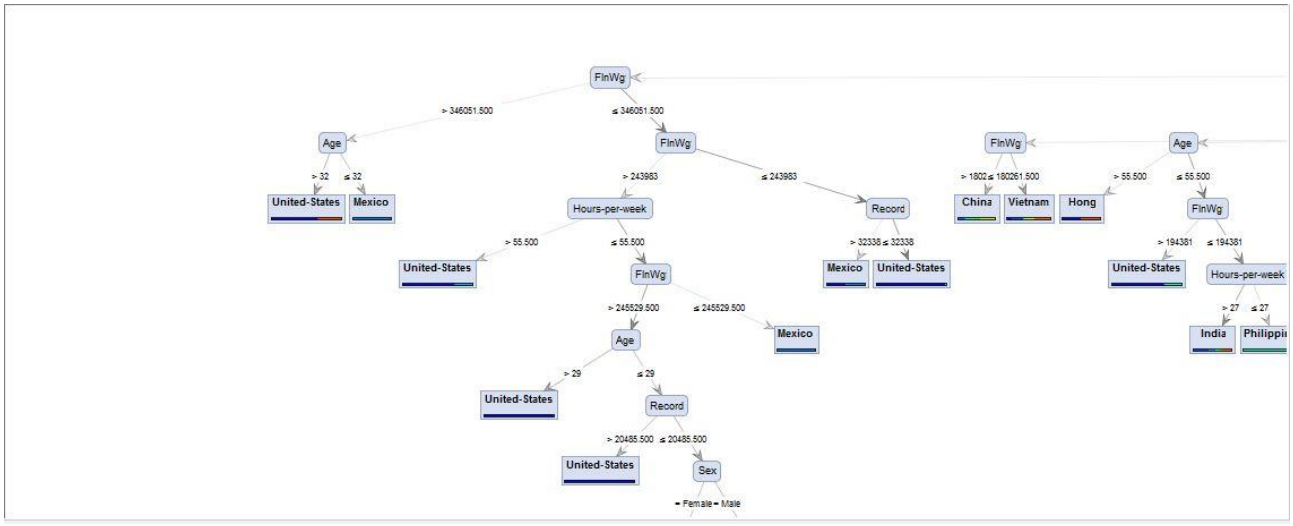
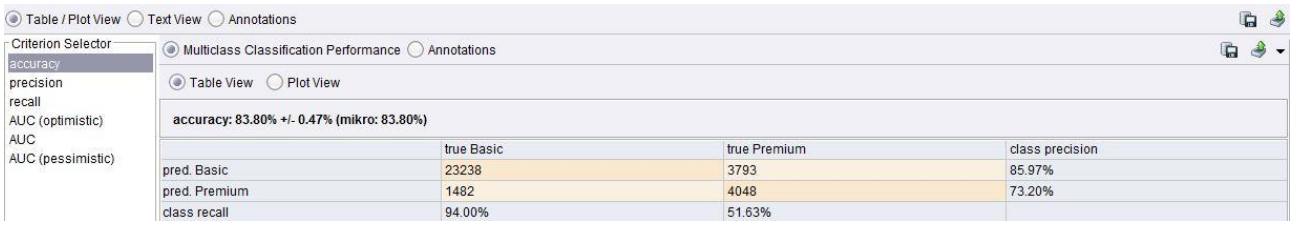


Figura 4

Domanda 4

In generale, riducendo il valore del minimal gain e aumentando la maximal depth si genera un modello di classificazione più dettagliato e quindi più accurato. Tuttavia, sulla base dei risultati riportati nelle Figure 5-13, impostando valori di maximal depth superiori a 5 e minimal gain inferiori a 0.05 si produce l'effetto denominato "overfitting", ovvero il modello risulta troppo "focalizzato" sui dati di train per classificare in modo accurato nuovi dati di test.



	true Basic	true Premium	class precision
pred. Basic	23238	3793	85.97%
pred. Premium	1482	4048	73.20%
class recall	94.00%	51.63%	

Figura 5 – Decision Tree. Minimum gain = 0.1. Maximum depth = 20

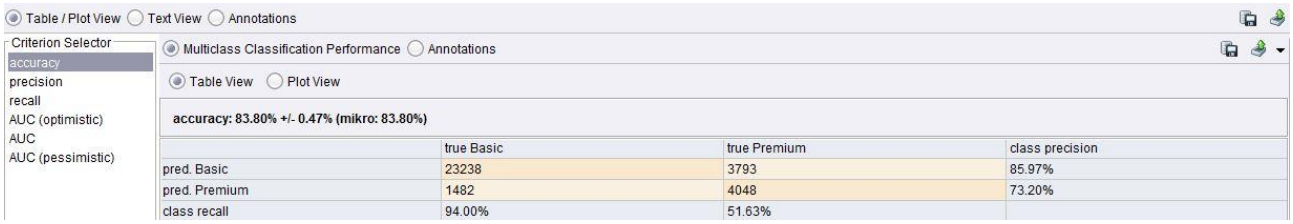


Table / Plot View | Text View | Annotations

Criterion Selector: accuracy, precision, recall, AUC (optimistic), AUC, AUC (pessimistic)

Multiclass Classification Performance | Annotations

Table View | Plot View

accuracy: 83.80% +/- 0.47% (mikro: 83.80%)

	true Basic	true Premium	class precision
pred. Basic	23238	3793	85.97%
pred. Premium	1482	4048	73.20%
class recall	94.00%	51.63%	

Figura 6 – Decision Tree. Minimum gain = 0.1. Maximum depth = 15

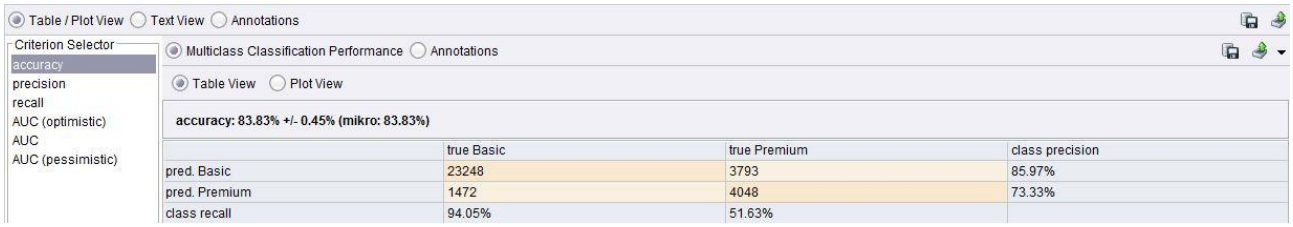


Table / Plot View | Text View | Annotations

Criterion Selector: accuracy, precision, recall, AUC (optimistic), AUC, AUC (pessimistic)

Multiclass Classification Performance | Annotations

Table View | Plot View

accuracy: 83.83% +/- 0.45% (mikro: 83.83%)

	true Basic	true Premium	class precision
pred. Basic	23248	3793	85.97%
pred. Premium	1472	4048	73.33%
class recall	94.05%	51.63%	

Figura 7 – Decision Tree. Minimum gain = 0.1. Maximum depth = 10

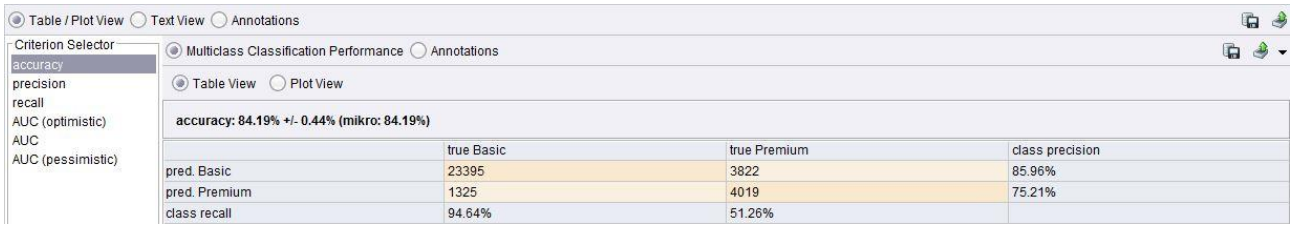


Table / Plot View | Text View | Annotations

Criterion Selector: accuracy, precision, recall, AUC (optimistic), AUC, AUC (pessimistic)

Multiclass Classification Performance | Annotations

Table View | Plot View

accuracy: 84.19% +/- 0.44% (mikro: 84.19%)

	true Basic	true Premium	class precision
pred. Basic	23395	3822	85.96%
pred. Premium	1325	4019	75.21%
class recall	94.64%	51.26%	

Figura 8 – Decision Tree. Minimum gain = 0.1. Maximum depth = 5

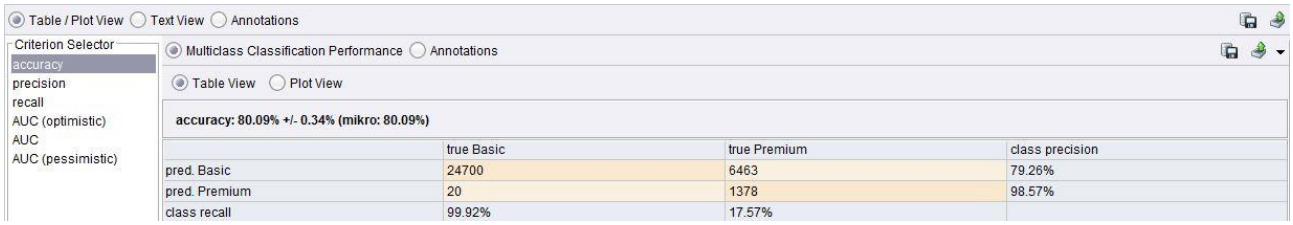


Table / Plot View | Text View | Annotations

Criterion Selector: accuracy, precision, recall, AUC (optimistic), AUC, AUC (pessimistic)

Multiclass Classification Performance | Annotations

Table View | Plot View

accuracy: 80.09% +/- 0.34% (mikro: 80.09%)

	true Basic	true Premium	class precision
pred. Basic	24700	6463	79.26%
pred. Premium	20	1378	98.57%
class recall	99.92%	17.57%	

Figura 9 – Decision Tree. Minimum gain = 0.1. Maximum depth = 3

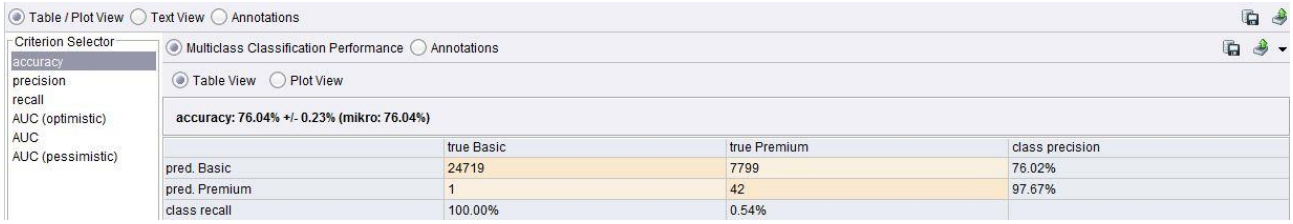


Table / Plot View | Text View | Annotations

Criterion Selector: accuracy, precision, recall, AUC (optimistic), AUC, AUC (pessimistic)

Multiclass Classification Performance | Annotations

Table View | Plot View

accuracy: 76.04% +/- 0.23% (mikro: 76.04%)

	true Basic	true Premium	class precision
pred. Basic	24719	7799	76.02%
pred. Premium	1	42	97.67%
class recall	100.00%	0.54%	

Figura 10 – Decision Tree. Minimum gain = 0.2. Maximum depth = 5

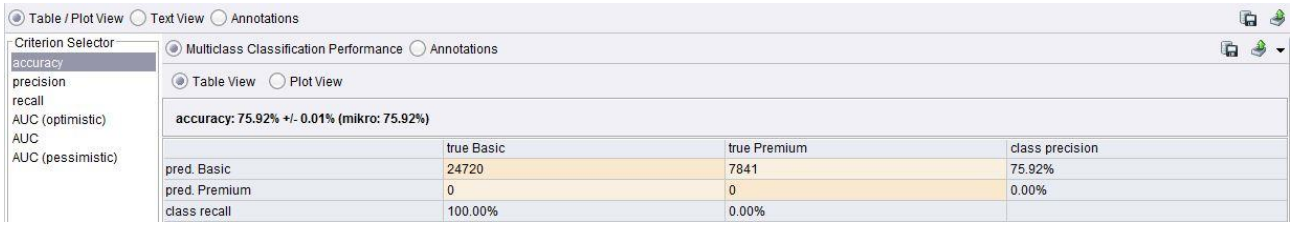


Table / Plot View | Text View | Annotations

Criterion Selector: accuracy, precision, recall, AUC (optimistic), AUC, AUC (pessimistic)

Multiclass Classification Performance | Annotations

Table View | Plot View

accuracy: 75.92% +/- 0.01% (mikro: 75.92%)

	true Basic	true Premium	class precision
pred. Basic	24720	7841	75.92%
pred. Premium	0	0	0.00%
class recall	100.00%	0.00%	

Figura 11 – Decision Tree. Minimum gain = 0.5. Maximum depth = 5

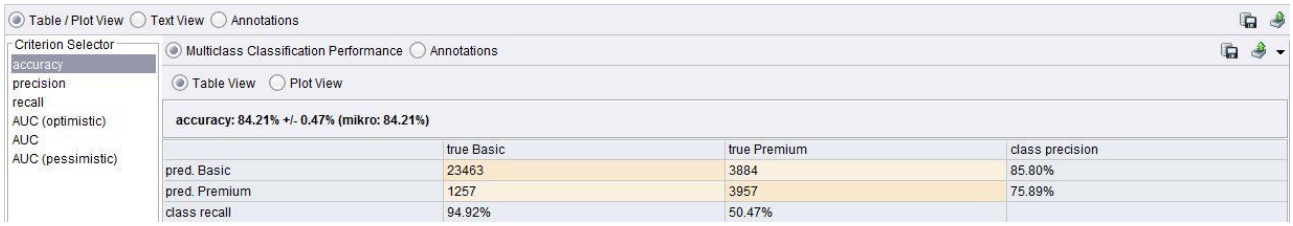


Figura 12 – Decision Tree. Minimum gain = 0.05. Maximum depth = 5

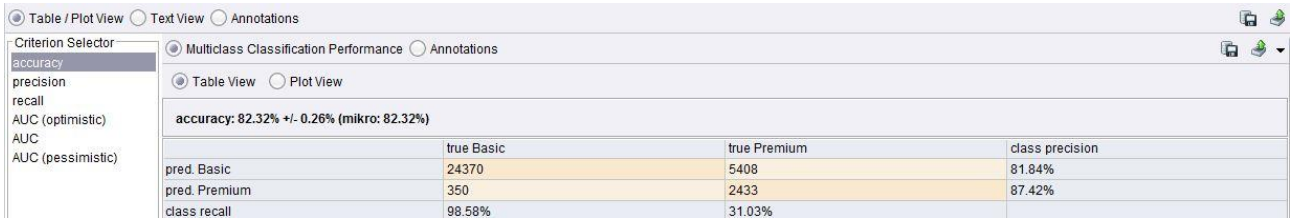


Figura 13 – Decision Tree. Minimum gain = 0.01. Maximum depth = 5

Domanda 5

Incrementando il valore di K, il classificatore considera un numero maggiore di dati di train "vicini" al dato di test e quindi l'accuratezza media cresce: 69.80% con K=1, 74.68% con K=3, 76.80% con K=5, 79.29% con K=15 (Figure 14-19). Considerando un numero molto elevato di record di train "vicini" (ad es., K>15) la presenza di dati rumorosi comincia ad inficiare le performance di classificazione e dunque l'accuratezza media di classificazione diminuisce leggermente (Figura 20).

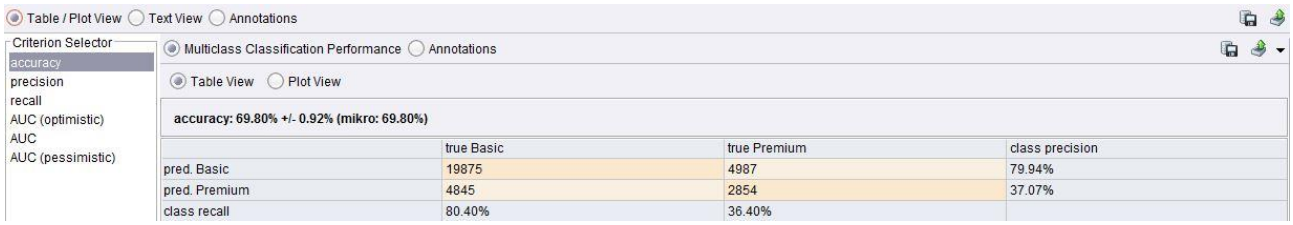


Figura 14 – K-NN. Matrice di confusione. K=1

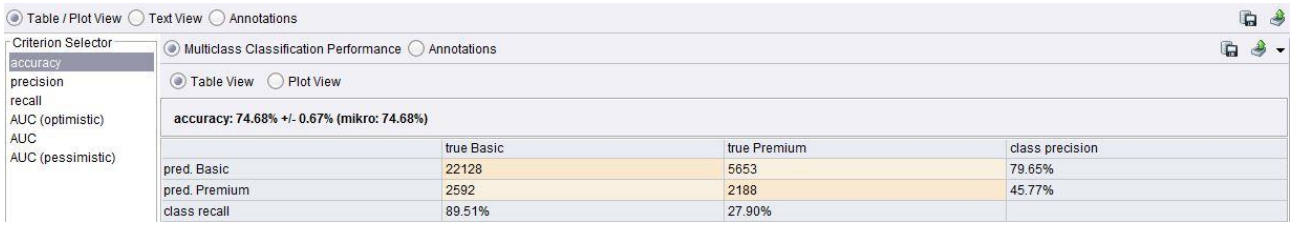


Figura 15 – K-NN. Matrice di confusione. K=3

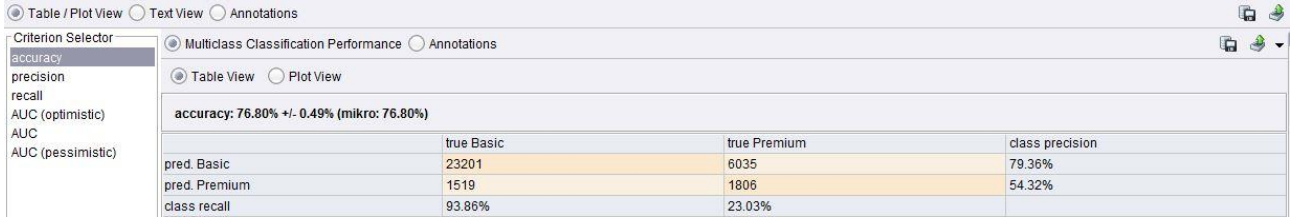


Figura 16 – K-N. Matrice di confusione. K=5

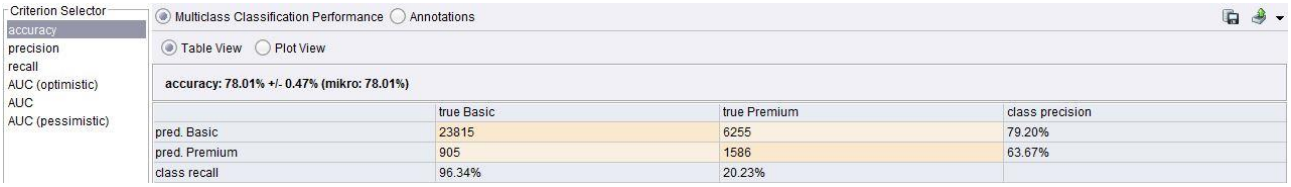


Figura 17 – K-N. Matrice di confusione. K=7

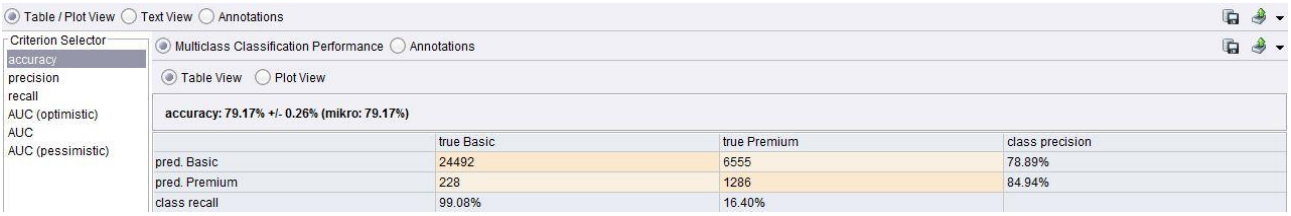


Figura 18 – K-NN. Matrice di confusione. K=10

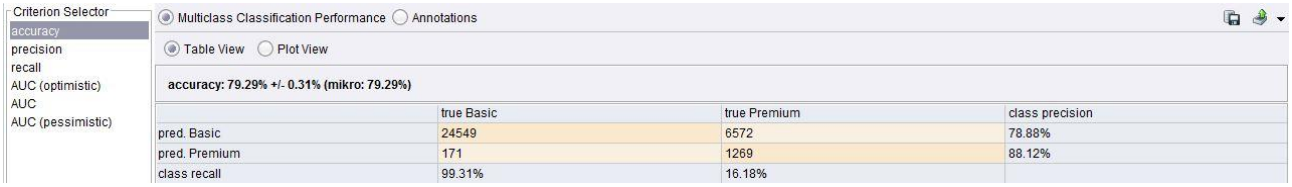


Figura 19 – K-N. Matrice di confusione. K=15

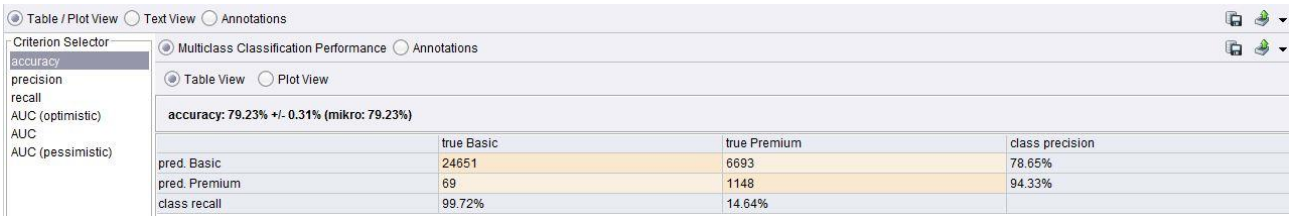


Figura 20 – K-NN. Matrice di confusione. K=20

Come mostrato in Figura 21, Naïve Bayes ottiene un'accuratezza media più elevata di K-NN (83.34% contro79.29%) sul dataset analizzato.

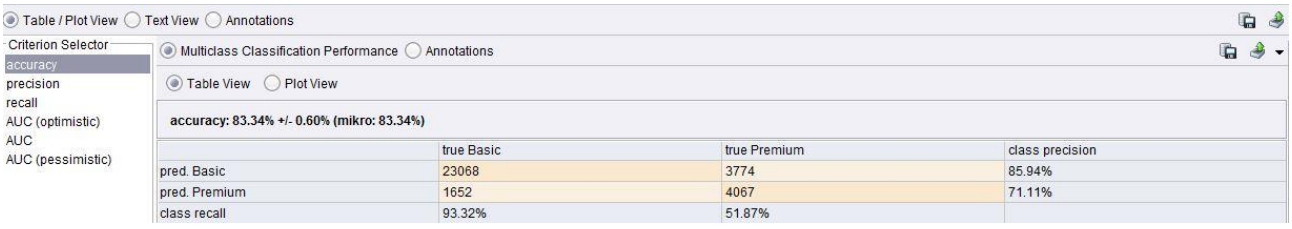


Figura 21 – Naïve Bayes. Matrice di confusione. K=3

Domanda 6

Figura 22 mostra la matrice di correlazione ottenuta dal dataset analizzato. Essa riporta la correlazione mutua (e simmetrica) tra coppie di attributi. Per esempio, l'attributo "Age" risulta essere molto correlato con l'attributo "Marital status" Dato che sussistono correlazioni significative tra attributi, ad es., tra "Age" e "Marital Status" (correlazione = 0.425), tra "Sex" e "Relationship" (correlazione = 0.273), l'ipotesi Naïve

risulta essere irrealistica per il dataset analizzato. Tuttavia, le performance di Naïve Bayes risultano essere mediamente buone (vedi riposta alla domanda precedente).

Table View | Pairwise Table | Plot View | Annotations

Attributes	Age	Workclass	FlnWgt	Education	Record	Education-...	Marital Status	Occupation	Relationship	Race	Sex	Capital Gain	Capital loss	Hours-per-...	Native Cou...
Age	1	0.082	-0.077	0.008	0.001	0.037	0.425	0.017	-0.218	-0.040	-0.089	0.078	0.058	0.069	-0.012
Workclass	0.082	1	-0.006	0.012	-0.001	0.011	0.036	0.217	0.016	0.009	0.019	0.041	0.013	-0.028	-0.010
FlnWgt	-0.077	-0.006	1	0.024	-0.003	-0.043	-0.024	0.008	0.017	0.000	-0.027	0.000	-0.010	-0.019	0.036
Education	0.008	0.012	0.024	1	0.017	-0.280	0.009	0.075	0.044	0.030	0.001	0.024	-0.003	-0.050	0.068
Record	0.001	-0.001	-0.003	0.017	1	-0.001	-0.001	0.000	0.006	0.006	0.002	0.002	-0.001	0.001	0.004
Education-N	0.037	0.011	-0.043	-0.280	-0.001	1	-0.066	-0.243	-0.141	-0.040	-0.012	0.123	0.080	0.148	-0.066
Marital Statu	0.425	0.036	-0.024	0.009	-0.001	-0.066	1	0.007	0.029	0.013	0.182	0.004	0.007	-0.000	0.002
Occupation	0.017	0.217	0.008	0.075	0.000	-0.243	0.007	1	-0.016	0.017	-0.148	-0.045	-0.024	-0.045	0.020
Relationship	-0.218	0.016	0.017	0.044	0.006	-0.141	0.029	-0.016	1	0.097	0.273	-0.044	-0.050	-0.185	0.042
Race	-0.040	0.009	0.000	0.030	0.006	-0.040	0.013	0.017	0.097	1	0.068	-0.008	-0.017	-0.033	0.242
Sex	-0.089	0.019	-0.027	0.001	0.002	-0.012	0.182	-0.148	0.273	0.068	1	-0.048	-0.046	-0.229	0.006
Capital Gain	0.078	0.041	0.000	0.024	0.002	0.123	0.004	-0.045	-0.044	-0.008	-0.048	1	-0.032	0.078	-0.009
Capital loss	0.058	0.013	-0.010	-0.003	-0.001	0.080	0.007	-0.024	-0.050	-0.017	-0.046	-0.032	1	0.054	-0.004
Hours-per-w	0.069	-0.028	-0.019	-0.050	0.001	0.148	-0.000	-0.045	-0.185	-0.033	-0.229	0.078	0.054	1	-0.010
Native Coun	-0.012	-0.010	0.036	0.068	0.004	-0.066	0.002	0.020	0.042	0.242	0.006	-0.009	-0.004	-0.010	1

Figura 22 – Matrice di correlazione