

Classificazione dei Supercomputer

Progetto per il corso di Statistica del Prof. Marco Romito

Rambod Rahmani

Corso di Laurea Magistrale in
Artificial Intelligence and Data Engineering

13 Dicembre 2020

Indice

1	Introduzione	1
2	Dati	1
2.1	Contenuto della tabella	2
2.2	Importazione e pulizia	2
3	Analisi	2
3.1	Preliminare analisi delle Componenti Principali	3
3.2	Classificazione per mezzo di Analisi Discriminante Lineare	3
3.3	Classificazione per mezzo di Analisi Discriminante Quadratica	6
3.4	Classificazione per mezzo di Regressione Logistica	8
4	Conclusioni	8

1 Introduzione

Lo scopo della presente analisi è quello di costruire un modello di classificazione per poter determinare il segmento di mercato di appartenenza di una Supercomputer a partire dalle specifiche delle sue caratteristiche hardware e dalle prestazioni ottenute nei principali benchmarks utilizzati in questo settore. A partire dalla tabella dei dati, tramite l'utilizzo di R, a seguito di una preliminare analisi delle componenti principali, sono stati valutati un modello di classificazione discriminante lineare, uno di classificazione discriminante quadratica e uno di classificazione con la regressione logistica.

Per quanto riguarda il contesto applicativo ipotizzato, possiamo immaginarci che un tale modello di classificazione possa essere utilizzato, al momento dell'installazione di un nuovo Supercomputer, per individuare la fascia di mercato più idonea in base alle sue prestazioni.

2 Dati

La tabella dei dati è stata scaricata dal sito dell'organizzazione **TOP500**. La TOP500 mantiene una graduatoria, ordinata secondo le loro prestazioni, dei Supercomputer attualmente installati e in funzione. Tale graduatoria viene aggiornata con cadenza semestrale.

Link di download diretto: https://www.top500.org/lists/top500/2020/11/download/TOP500_202011.xlsx

Credenziali di accesso:

Login : rambodrahmani@yahoo.it
Password : GCgFH6yuZYFMeCr

2.1 Contenuto della tabella

La tabella dei dati contiene 37 colonne per un totale di 500 osservazioni. Per la presente analisi ho utilizzato le seguenti colonne: **Site, Manufacturer, Country, Year, Segment, TotalCores, Rmax, Rpeak, Nmax, HPCG, Architecture, Processor, ProcessorTechnology, ProcessorSpeed, OperatingSystem, CoProcessor, CoresPerSocket, ProcessorGeneration, SystemModel, SystemFamily, InterconnectFamily, Interconnect, Continent**. A parte le colonne di significato ovvio, penso sia doveroso fornire maggiori informazioni i seguenti fattori:

- **Rmax [TFlop/s]**: massime prestazioni raggiunte nel benchmark LINPACK;
- **Rpeak [TFlop/s]**: massime prestazioni teoriche;
- **Nmax**: dimensione del problema sul quale è stato raggiunto il punteggio Rmax.
- **HPCG [TFlop/s]**: massime prestazioni raggiunte nel benchmark HPCG (High Performance Conjugate Gradient);

2.2 Importazione e pulizia

Sui dati, non è stata effettuata alcuna operazione precedente la loro importazione in R. Il file originale, in formato `.xlsx`, è stato però convertito in `.csv` per facilitare l'importazione. Prima di iniziare l'analisi, ho rimosso le colonne che ritengo che non influenzano la classificazione del segmento di mercato di un Supercomputer (Rank TOP500, "Name", "Computer", "Power.Source", "OS.Family", ecc...), mentre come fattore per la classificazione è stato utilizzato il valore della colonna "Segment".

```
> with(data, table(Segment))
```

Segment	Academic	Government	Industry	Others	Research	Vendor
	67	34	273	14	103	9

Nelle colonne che ho scelto, ho rilevato 351 valori mancanti in "Accelerator/Co-Processor Cores", 426 in "HPCG [TFlop/s]" e 488 in "Nhalf". Dato che il numero di valori mancanti è elevato rispetto al totale delle 500 osservazioni, le suddette colonne sono state eliminate.

3 Analisi

Subito dopo l'importazione e la pulizia dei dati, sono state effettuate due preliminari classificazioni ottenendo i seguenti risultati:

- **Analisi Discriminante Lineare**: una accuratezza non soddisfacente del 73.33%

Confusion Matrix and Statistics

	Reference					
Prediction	Academic	Government	Industry	Others	Research	Vendor
Academic	42	2	8	0	22	2
Government	4	16	1	0	2	0
Industry	5	6	245	11	17	0
Others	0	0	7	2	1	0
Research	15	10	8	1	53	1
Vendor	1	0	4	0	4	5

Overall Statistics

Accuracy : 0.7333

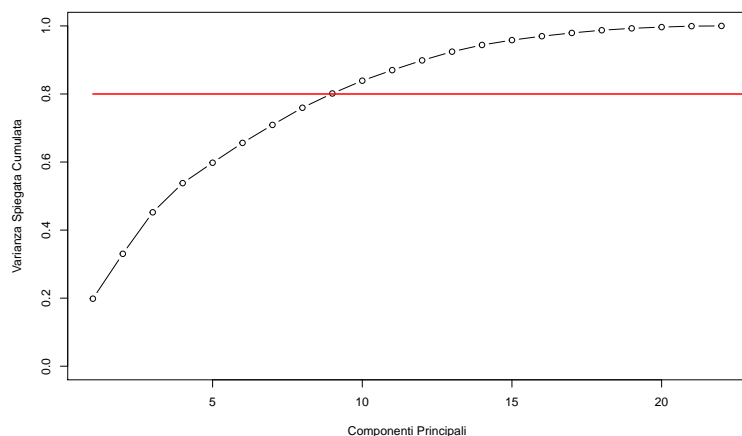
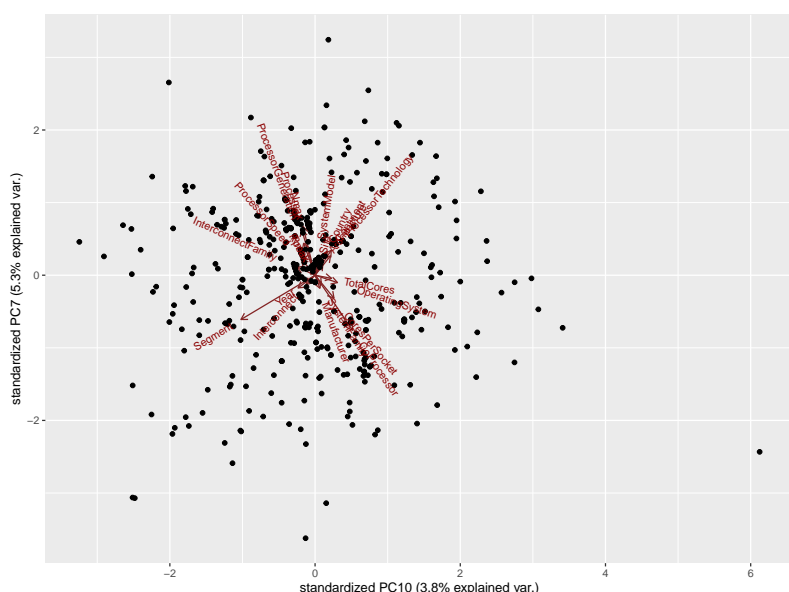
- **Analisi Discriminante Quadratica**: stiamo utilizzando 22 fattori con un numero di osservazioni pari a 495

```
Error in qda.default(x, grouping, ...) :  
  some group is too small for 'qda'
```

Una analisi delle componenti principali per ridurre la dimensione del problema si rivela quindi inevitabile.

3.1 Preliminare analisi delle Componenti Principali

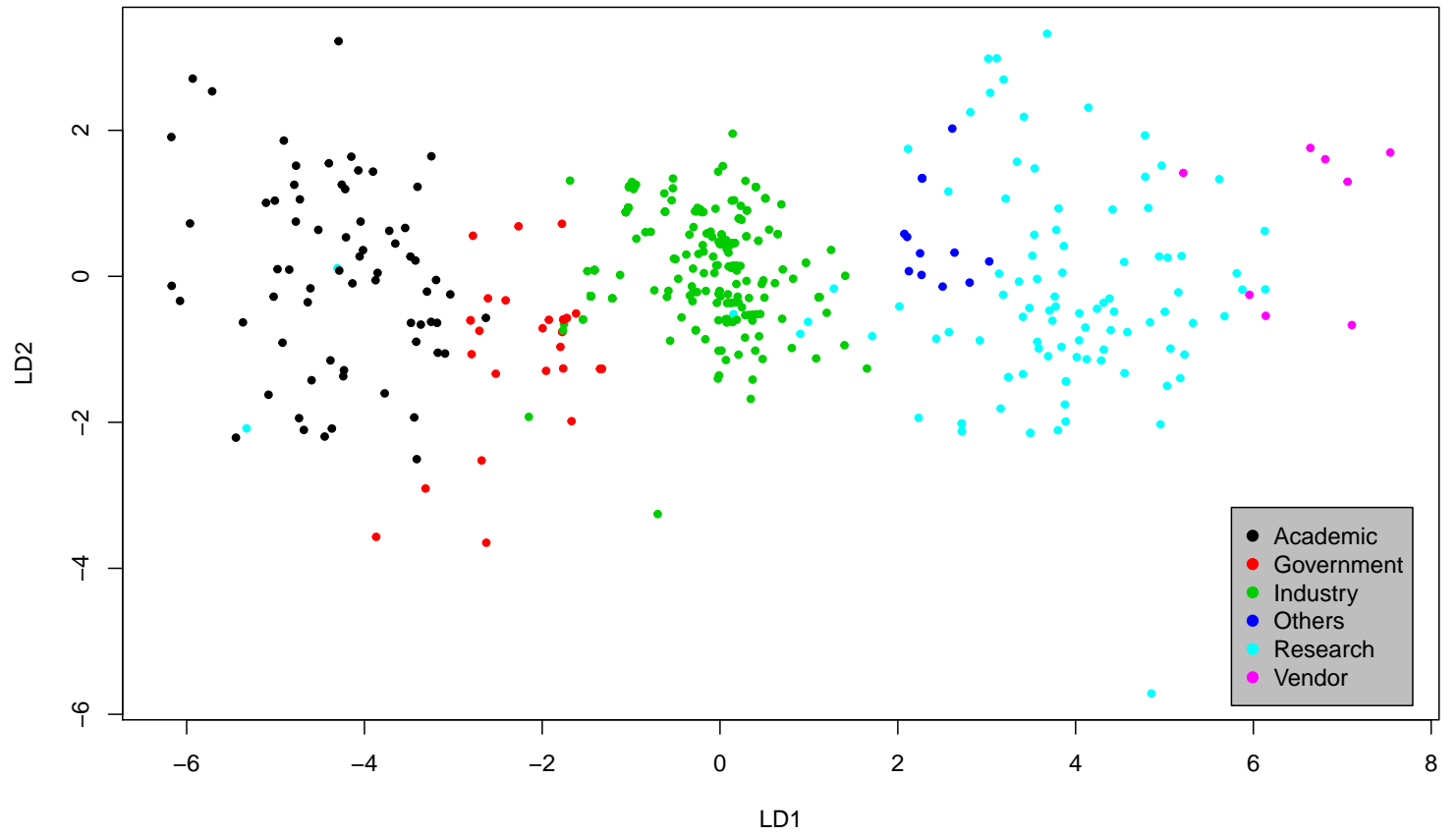
Una prima analisi delle componenti principali, non prendendo in considerazione la classe delle osservazioni, nonostante ci permetta di ridurre il numero di fattori presi in considerazione, porta a risultati persino peggiori: l'accuratezza della classificazione per mezzo di LDA scende al 68%. Ho quindi valutato un secondo modello di PCA dove ho preso in considerazione anche la classe delle osservazioni.



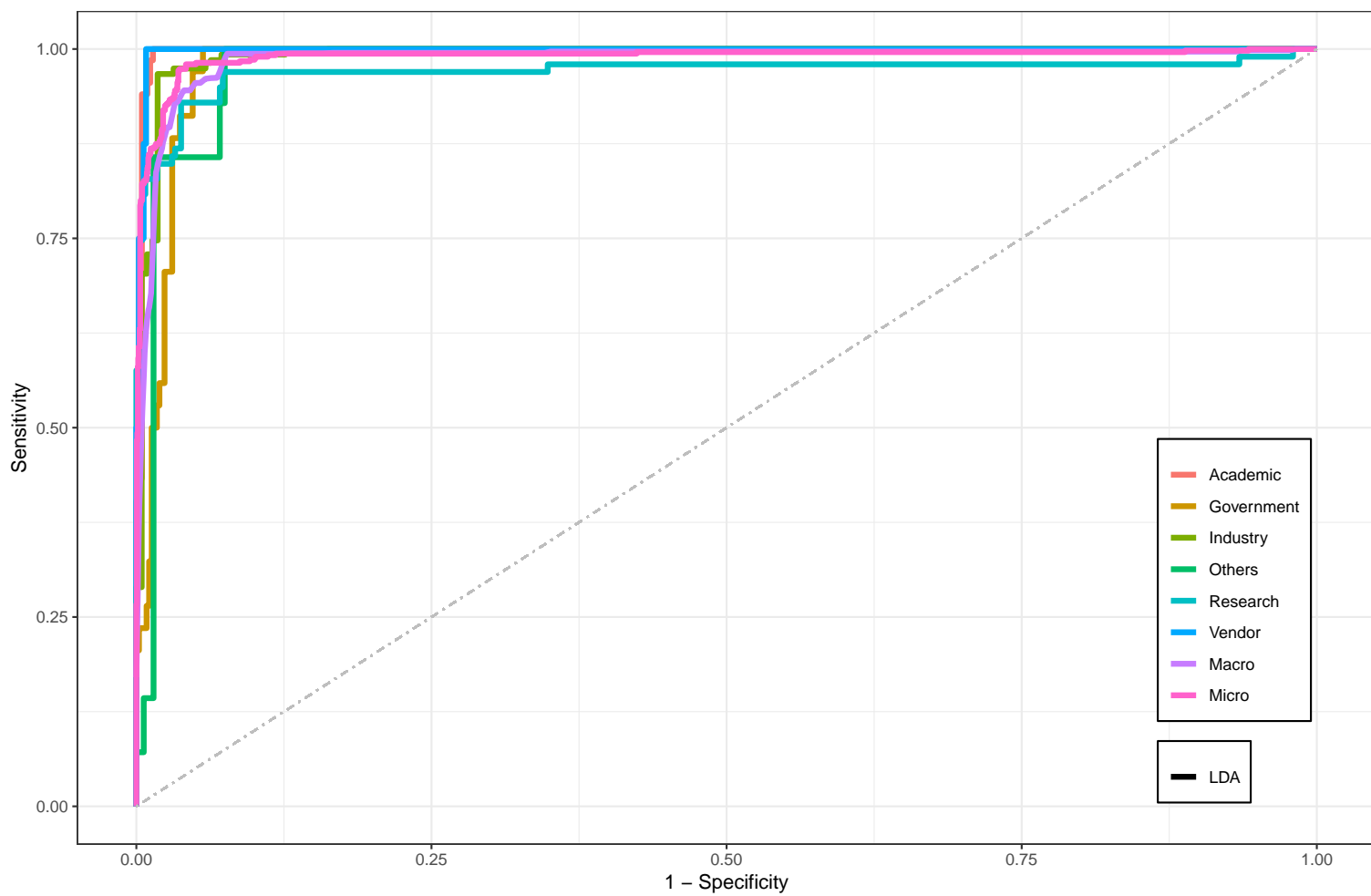
Nonostante non è certamente un problema che definiremmo "passibile di riduzione", possiamo notare che con un numero di fattori pari a 10 catturiamo circa l'85% della struttura. Meno della metà quindi del numero di fattori iniziale. È certamente un miglioramento. Ho identificato le componenti principali che meglio catturavano la varianza del fattore originario "Segment". Con questo nuovo insieme di fattori, la dimensione di analisi del problema è diminuita dall'iniziale 22 a 5. Sono quindi stati valutati a questo punto un modello di classificazione per mezzo di Analisi Discriminante Lineare, uno di classificazione per mezzo di Analisi Discriminante Quadratica e uno di classificazione per mezzo di Regressione Logistica. I risultati ottenuti sono esposti nelle sezioni a seguire.

3.2 Classificazione per mezzo di Analisi Discriminante Lineare

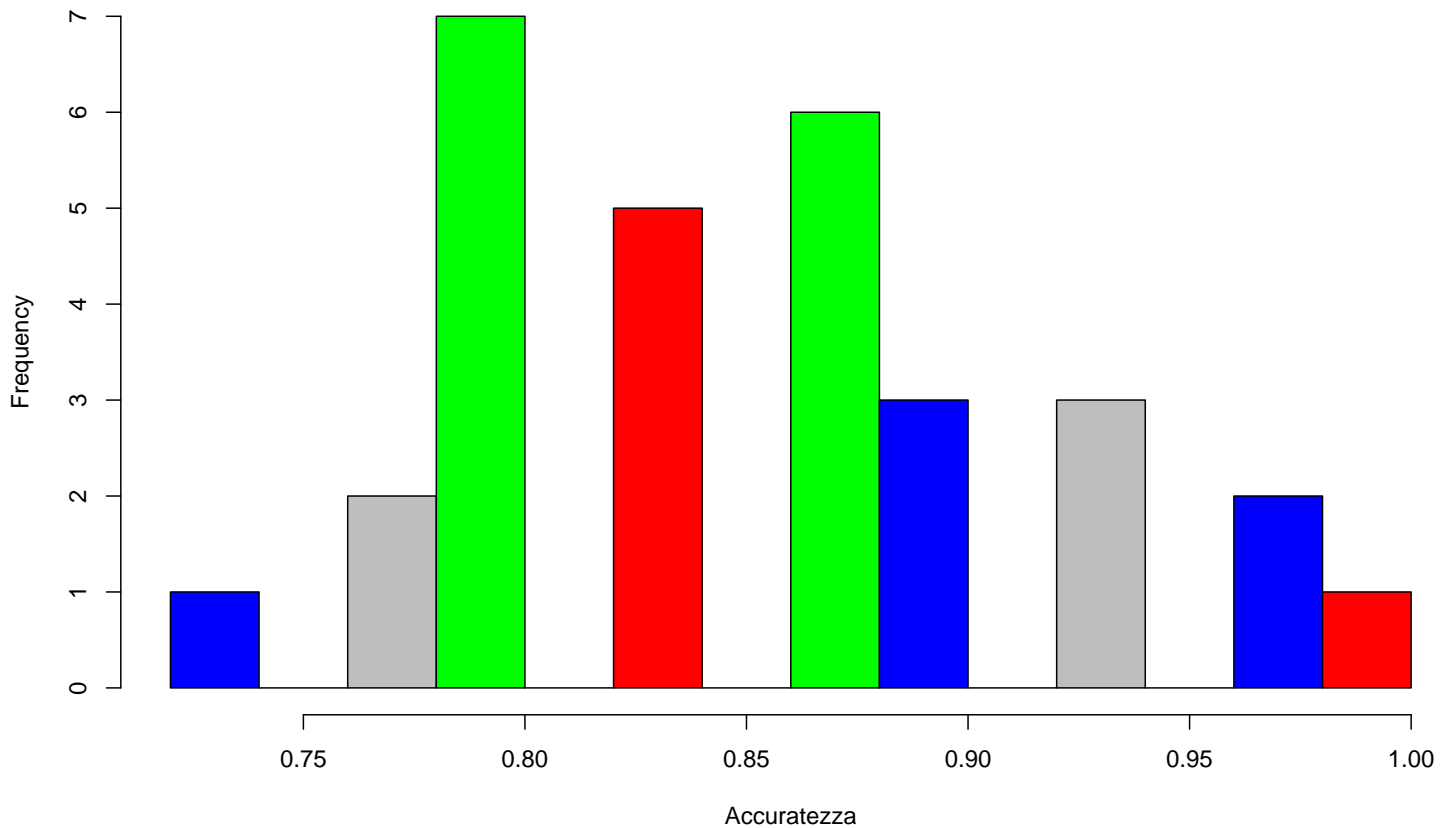
Sono stati utilizzati 5 fattori ottenendo un'accuratezza di classificazione pari a circa 89%. Per ottenere risultati statisticamente significativi l'esperimento è stato ripetuto 30 volte ottenendo una precisione media dell'85.55% e una deviazione standard del 6.57%.



Confusion Matrix and Statistics							
Prediction	Reference						
	Academic	Government	Industry	Others	Research	Vendor	
Academic	65	4	0	0	2	0	
Government	2	9	2	0	0	0	
Industry	0	21	271	0	6	0	
Others	0	0	0	8	6	0	
Research	0	0	0	6	82	1	
Vendor	0	0	0	0	3	7	
Overall Statistics							
Accuracy : 0.8929							



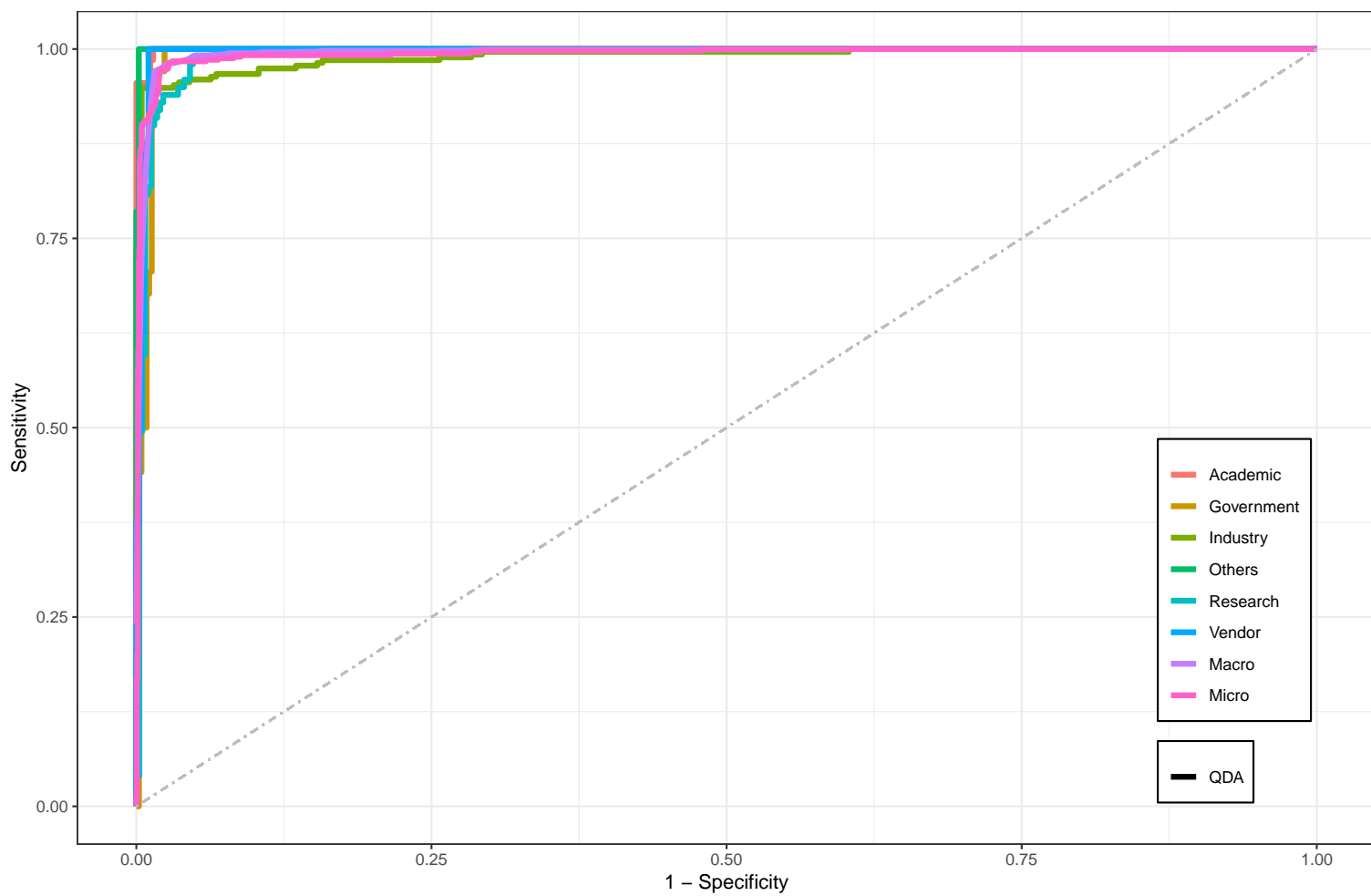
Accuratezza della Classificazione per mezzo di LDA



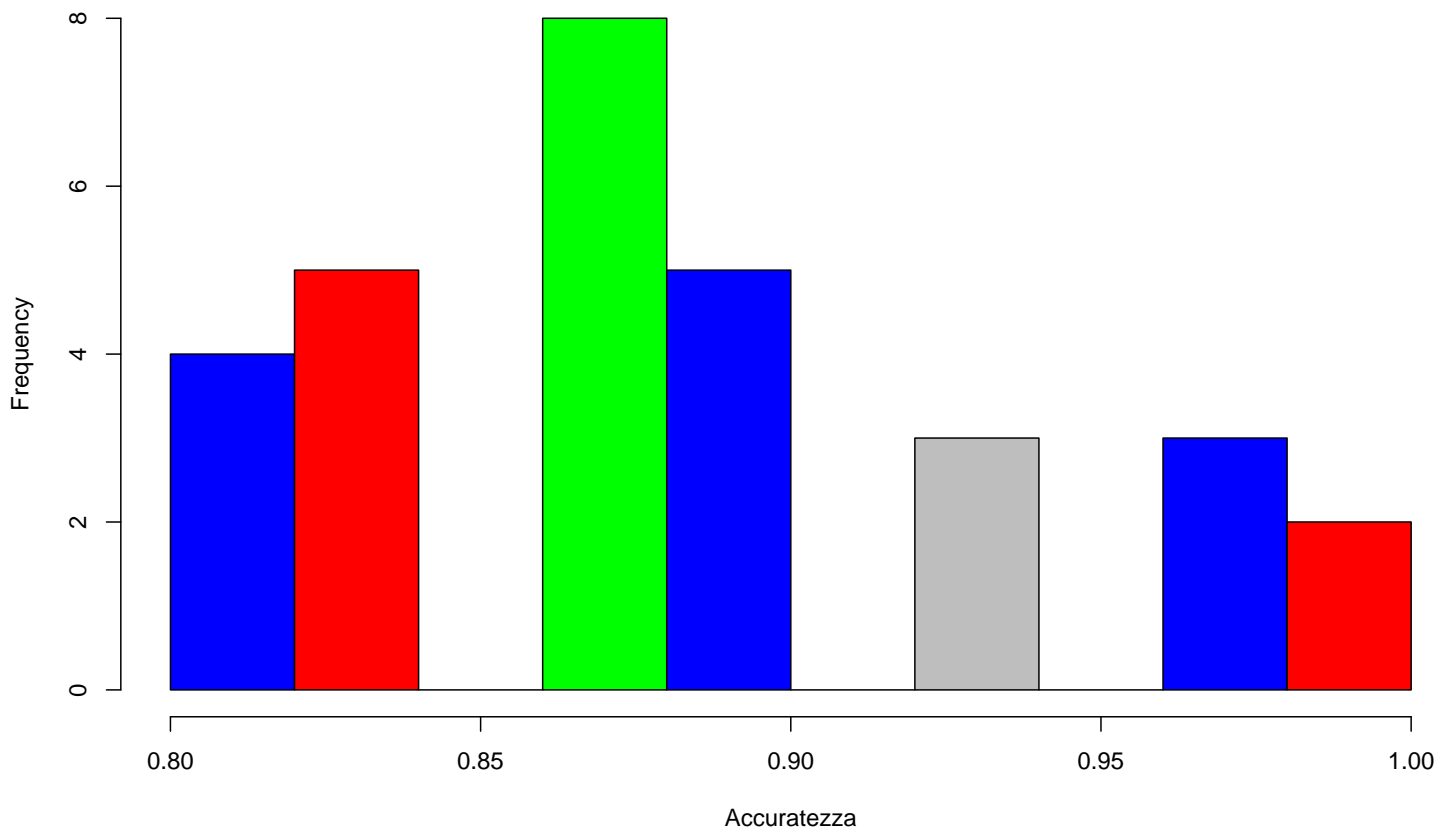
3.3 Classificazione per mezzo di Analisi Discriminante Quadratica

Sono stati utilizzati 5 fattori ottenendo un'accuratezza di classificazione pari a circa 89%. Per ottenere risultati statisticamente significativi l'esperimento è stato ripetuto 30 volte ottenendo una precisione media dell'88.33% e una deviazione standard del 5.85%.

Confusion Matrix and Statistics						
Prediction	Reference					
	Academic	Government	Industry	Others	Research	Vendor
Academic	64	3	0	0	0	0
Government	3	21	1	0	0	0
Industry	0	10	262	0	1	0
Others	0	0	0	12	1	0
Research	0	0	10	2	94	3
Vendor	0	0	0	0	3	5
Overall Statistics						
Accuracy : 0.9253						

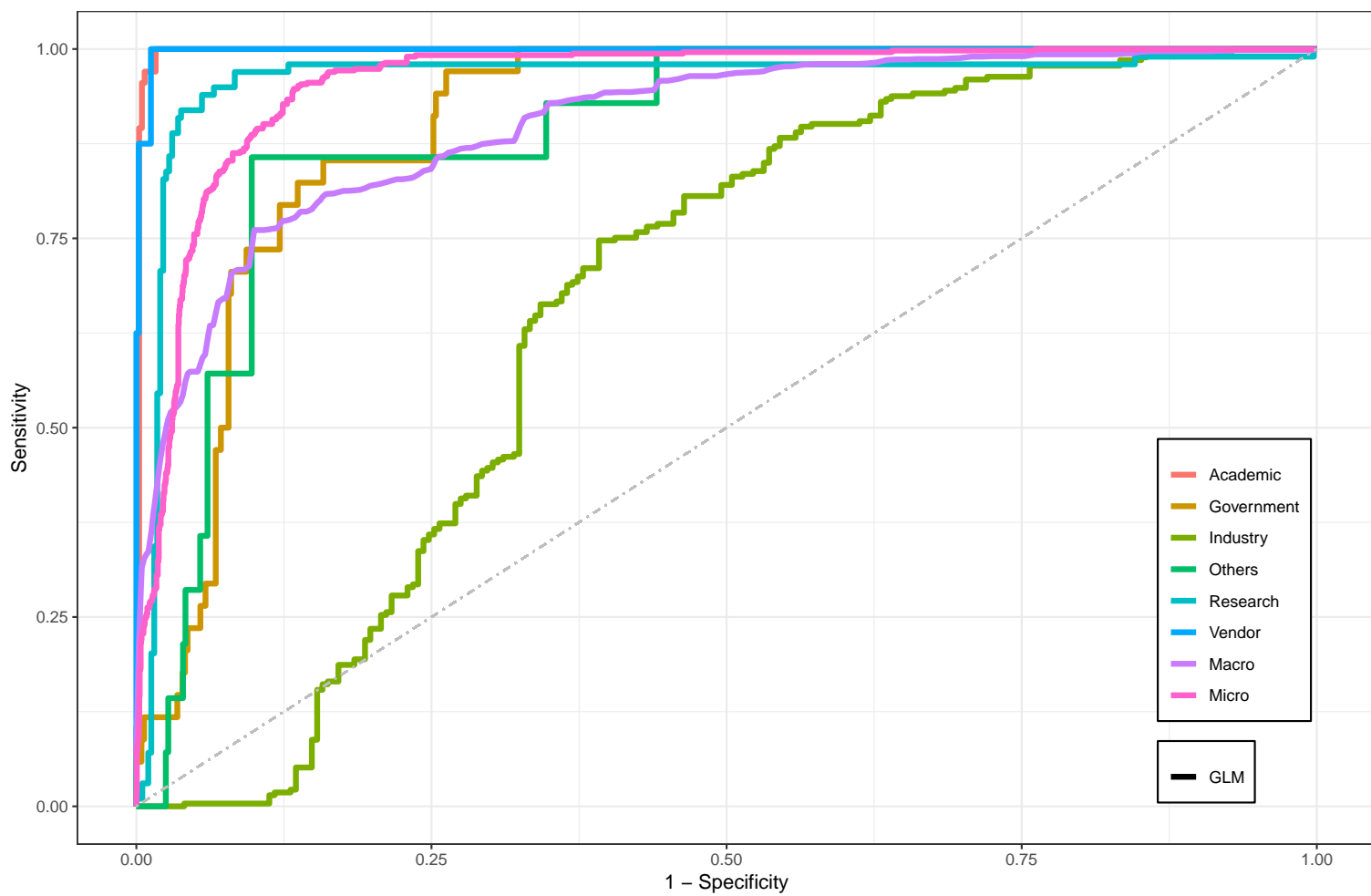


Accuratezza della Classificazione per mezzo di QDA



3.4 Classificazione per mezzo di Regressione Logistica

Sono stati utilizzati 5 fattori ottenendo un'accuratezza di classificazione pari a circa 69%: non è proprio soddisfacente rispetto ai precedenti due modelli analizzati.



4 Conclusioni

In ultima analisi