Data Mining - Prova d'esame del 25.6.2020

Analisi dei dati con R

Soluzione

E' stato proposto un dataset reale, il dataset *Pima Indians Diabetes* disponibile nel *repository* di dati di UCI Machine Learning. L'obiettivo è prevedere se un individuo ha il diabete ("Yes") oppure no ("No").

Il modello di benchmark poteva essere facilmente migliorato selezionando il valore di K attraverso il metodo della convalida incrociata:

```
library(kknn)
( K = train.kknn(y ~ ., data=tr, kmax=100)$best.parameters$k )
```

```
[1] 59
```

```
fit = kknn(y ~ ., tr, te, , k = K)
yhat = fit$fitted.values
( ACC = mean( yhat == y.te ) )
```

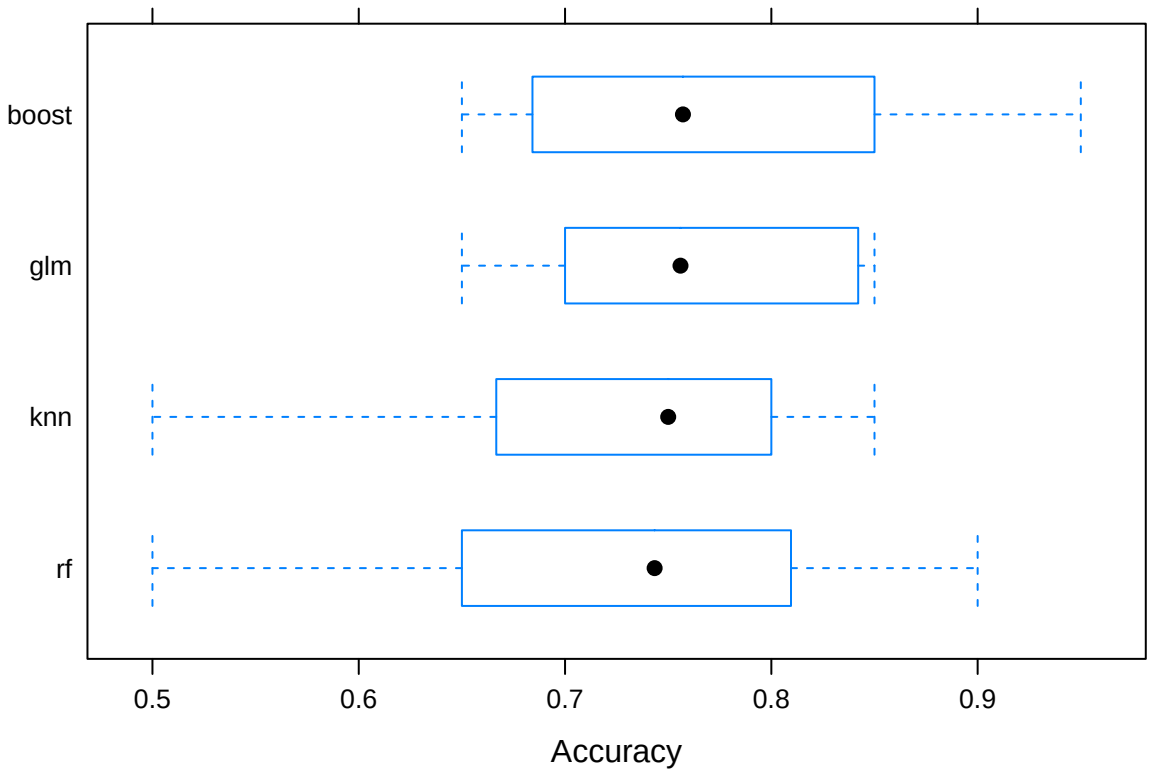
```
[1] 0.7680723
```

oppure considerando un semplice modello logistico

```
fit = glm(y ~ ., tr, family=binomial())
phat = predict(fit,te, type="response")
yhat = ifelse(phat > 0.5, "Yes","No")
( ACC = mean( yhat == y.te ) )
```

```
[1] 0.8012048
```

Infine, adottando l'approccio "forza bruta" si potevano confrontare i diversi modelli con il metodo della convalida incrociata:




```
$knn
[1] 0.753012

$glm
[1] 0.8012048

$rf
[1] 0.7650602

$boost
[1] 0.7650602
```