DATASET MAGIC TELESCOPE

SIGNORI ELENA - MATR. 843017

MAGIC (acronimo per Major Atmospheric Gamma-ray Imaging Cherenkov Telescopes) è un sistema di due telescopi a raggi gamma che si trova nelle isole Canarie. Questo sistema di telescopi è in grado di rilevare raggi gamma di origine extraterrestre per studiare l'origine dei raggi cosmici.
Quando il raggio gamma dallo spazio arriva nell'alta atmosfera provoca una serie di decadimenti con una specie di effetto doccia. Un prodotto di queste reazioni vicino alla superficie terrestre è un fascio di luce blu, detta luce Cherenkov, che viene captata dai telescopi a terra. A seconda dell'energia della gamma primaria, vengono raccolti da un totale di poche centinaia a circa 10.000 fotoni Cherenkov, in schemi (chiamati immagine della doccia), consentendo di discriminare statisticamente quelli causati dai gamma primari (segnale) dalle immagini delle docce adroniche avviate dai raggi cosmici nell'atmosfera superiore (fondo).

Il dataset analizzato contiene una serie di dati, generati per simulare la registrazione di particelle gamma ad alta energia da parte di un telescopio MAGIC usando la tecnica di imaging.

Il dataset è composto da 19.020 osservazioni e 11 variabili che sono le seguenti:

- fLength: lunghezza dell'asse maggiore di un'ellisse (variabile numerica in mm)
- fWidth: lunghezza dell'asse minore di un'ellisse (variabile numerica in mm)
- fSize: log in base 10 della somma del contenuto di tutti i pixel, numero di fotoni (variabile numerica)
- fConc: rapporto tra la somma dei due pixel più alti e la variabile fSize (variabile numerica)
- fConc1: rapporto tra il più alto pixel e la variabile fSize (variabile numerica)
- fAsym: distanza del pixel più alto dal centro, proiettata sull'asse maggiore (variabile numerica in mm)
- fM3Long: radice cubica del terzo momento lungo l'asse maggiore (variabile numerica in mm)
- fM3Trans: radice terza del terzo momento lungo l'asse minore (variabile numerica in mm)
- fAlpha: angolo tra l'asse maggiore dell'ellisse e il vettore che collega l'origine al punto (variabile numerica in gradi)
- fDist: distanza dall'origine al centro dell'ellisse (variabile numerica in mm)
- Class: variabile risposta che classifica i raggi come gamma (segnale) e adroni (fondo)

Per ragioni tecniche il numero degli eventi «adroni» è stato sottostimato. Nei dati reali la classe «adroni» rappresenta la maggioranza degli eventi.

Nelle informazioni sul dataset viene anche specificato che classificare un evento di fondo come un segnale è peggio che commettere l'errore opposto. Viene inoltre richiesto che tale errore si mantenga al di sotto di una delle seguenti soglie: 0.01, 0.02, 0.05, 0.1, 0.2.

Dopo aver suddiviso il dataset in training, validation e test set (75% per il training+validation set e 25% per il test set), osservo che le proporzioni delle classi nei diversi set sono le stesse. Inoltre, si tratta di un problema caratterizzato da classi di risposta non bilanciate ma non verranno attuate tecniche di bilanciamento.
Il dataset non presenta missing quindi si può procedere con l'analisi del training set.

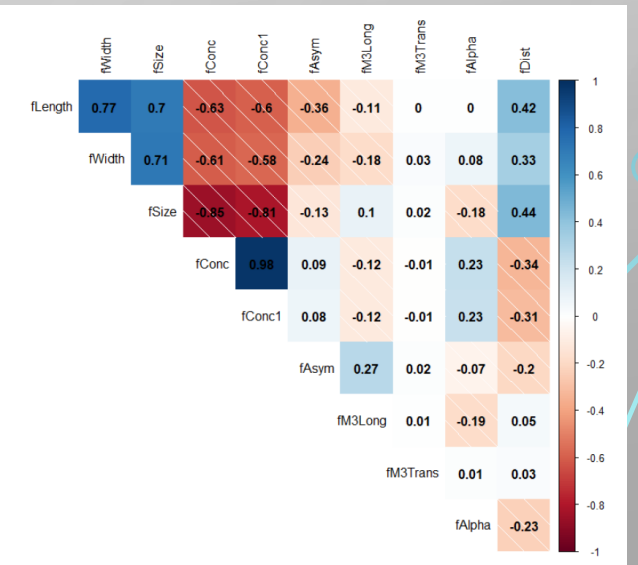
TRAINING SET

Da una prima analisi generale del training set, si nota che la variabile fWidth presenta alcuni valori anomali. In diverse osservazioni l'esplicativa in questione ha valore 0 che, per il significato della variabile, non ha molto senso quindi si procede con l'imputazione della mediana.

Proseguendo con lo studio delle distribuzioni delle variabili, si rileva la presenza di un numero elevato di outliers per le variabili fWidth, fAsym e fM3Long (superiori al 5% delle osservazioni). Si valuta l'applicazione di trasformazioni che, però, non risolvono il problema. Nel caso invece della variabile fLength, caratterizzata da distribuzione fortemente asimmetrica e 5% di outliers, la trasformazione logaritmica risolve entrambe le problematiche.

Analizzando la correlazione tra le variabili è evidente una forte correlazione tra le prime 5 esplicative, come si osserva dal grafico, quindi si decide di utilizzare la PCA per ridurre la dimensionalità. Prima di procedere con l'analisi delle componenti principali si standardizzano le variabili.

Dopo aver analizzato le variabili si decide di applicare la PCA alle prime tre variabili utilizzando la prima componente principale; si ottiene quindi una variabile che viene chiamata fLenWidSiz. Si applica poi l'analisi delle componenti principali anche alle variabili fConc e fConc1 considerando la prima componente principale. Si riduce così il numero di variabili che passa da 11 a 8 (inclusa la variabile Class).



Dopo aver ridotto la dimensionalità del training set, si verificano le assunzioni per l'applicazione dei modelli di classificazione.

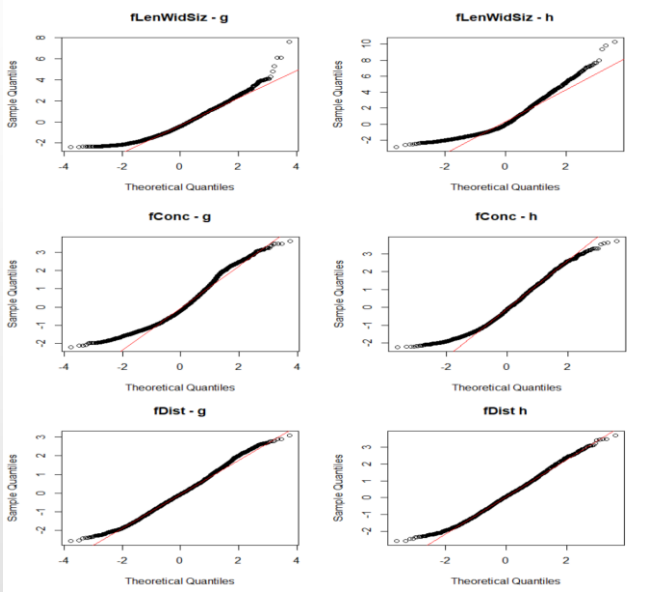
OSSERVAZIONI:

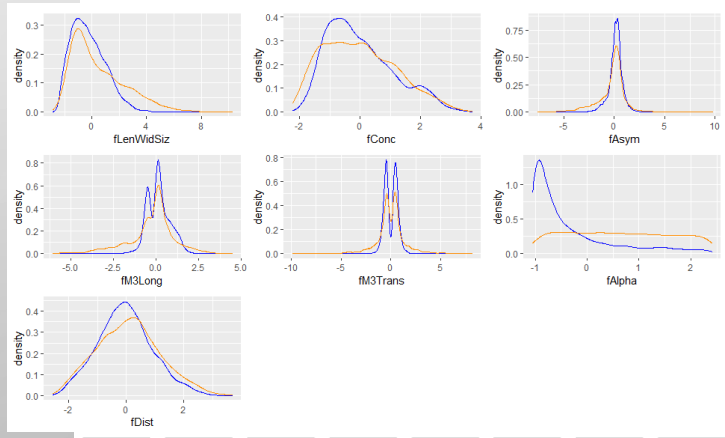
- Normalità: le variabili che sembrano avere una distribuzione che più si avvicina a una normale (almeno per una delle due classi) sono fLenWidSiz, fConc e fDist, seppur con code pesanti probabilmente dovute alla presenza di outliers.
- Covarianza: dal grafico ad ellissi si osserva che le variabili non sembrano avere covarianza in comune.

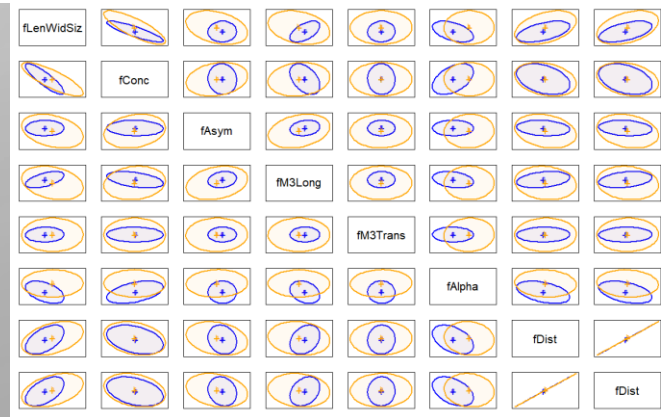
In seguito a queste osservazioni si decide di procedere considerando come possibili metodi di classificazione:

- Regressione logistica
- QDA (forzando l'ipotesi di normalità e quindi consapevole di commettere un errore)
- KNN

La LDA viene esclusa perché non vengono rispettate le ipotesi alla base del metodo e, inoltre, il numero di osservazioni è molto elevato.







1. Regressione logistica

Per il metodo della regressione logistica viene inizialmente stimato un modello che tiene conto di tutte le esplicative disponibili. Si osserva che alcune variabili non sono significative e viene eseguita una selezione stepwise basata sull'AIC. Successivamente, si verifica con un InfluencePlot la presenza di outliers, i quali vengono eliminati. A questo punto si ripete la stima del modello sul training set escludendo gli outliers ed è possibile notare che l'AIC migliora. Ripetendo la selezione stepwise, l'AIC migliora ulteriormente e le covariate risultano tutte significative per qualunque livello di α. Il modello che ne risulta presenta le seguenti covariate: fLenWidSiz ,fConc, fAsym, fM3Long e fAlpha.

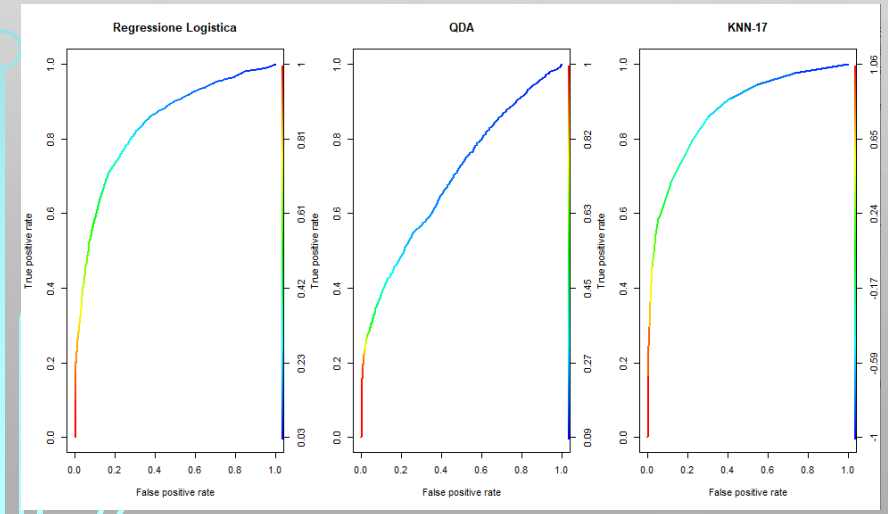
2. Analisi discriminante quadratica
Per la stima del modello della QDA vengono scelte solo le variabili che presentano una distribuzione simile a quella di una normale, quindi: fLenWidSiz, fConc e fDist.

3. K-Nearest Neighbours
Il KNN viene rimandato al validation set perché si tratta di un algoritmo che non prevede ipotesi sulle variabili di input.

VALIDATION SET
Nel validation set vengono applicate le scelte fatte nel training set in merito alla trasformazione della variabile fLength, all'imputazione della mediana per la variabile fWidth, alla PCA e alla standardizzazione delle esplicative.
In questa fase, i metodi vengono confrontati sulla base delle informazioni deducibili dalla loro matrice di confusione per valutarne la capacità di classificazione (accuracy, specificity e sensitivity). Vengono poi rappresentate le curve di ROC dei metodi e viene calcolata l'area sotto la curva (AUC).

Per il metodo KNN, dopo aver osservato che per valori molto alti di k l'accuratezza si riduce, sono stati valutati valori da 1 a 50. Il valore di k che ha restituito il maggior livello di accuratezza è stato 17.

Indicatori	Reg. logistica	QDA	KNN – 17
Accuracy	0,7885	0,7208	0,8185
Sensitivity	0,7979	0,7156	0,8058
Specificity	0,7635	0,7493	0,8583
AUC	0.8378	0.6965	0.8739



Visti i valori nella tabella e osservate le curve di ROC si decide di procedere solo con la Regressione logistica e il KNN.
In particolare, dato che è peggio classificare un evento di fondo come segnale, siamo interessati a minimizzare il «false positive rate» e quindi vorremmo un livello di specificità che sia il più alto possibile.

TRAINING + VALIDATION SET

In questa fase vengono stimati di nuovo i due modelli con cui si è scelto di proseguire. Vengono applicate tutte le scelte fatte sulle trasformazioni delle variabili, imputazione della mediana, PCA e standardizzazione. Per il KNN non è necessaria nessuna operazione, mentre per la Regressione logistica viene rieseguita una selezione stepwise basata sull'AIC che porta a selezionare le stesse variabili che erano state individuate nel training set.

TEST SET

Giunti alla fase finale del confronto tra i metodi di classificazione, si applicano nuovamente le scelte fatte in precedenza e si procede a testare la capacità di classificazione dei modelli confrontando le informazioni ricavabili dalla matrice di confusione.

Dalla tabella è possibile notare che il KNN risulta essere migliore della regressione logistica per tutti e tre gli indicatori; si decide quindi di considerare il KNN con k=17 come modello finale da valutare.

Indicatori	Reg. logistica	KNN - 17
Accuracy	0.7876	0.8223
Sensitivity	0.8001	0.8108
Specificity	0.7558	0.8565

Calcolando l'area sottesa alla curva di ROC per il modello KNN si ottiene AUC=0,88

Infine, si calcolano il training error e il test error ottenendo:

- Training error = 0.1815
- Test error = 0.1777

L'errore nel test è inferiore quindi le performance migliorano tra training e test.
Per completezza, poiché si desidera minimizzare il «false positive rate», viene calcolato anche il suo valore che risulta:
FPR = 0.1435

CONCLUSIONI

Il metodo che risulta più adatto alla classificazione del fenomeno è il KNN considerando un k pari a 17.
Si osserva che tale metodo permette di ottenere un livello di «false positive rate» inferiore ad una delle soglie richieste (0.01, 0.02, 0.05, 0.1, 0.2) e un buon livello di accuratezza.