Esplorando l'Universo: cluster e metodi di classificazione dei corpi celesti

Francesca Verna, Edoardo Olivieri, Paolo Portanova

La vastità dell'universo ha da sempre affascinato l'uomo, la cui insaziabile curiosità lo ha portato a spingersi nelle sue più oscure profondità. Il primo passo per comprendere la complessità e la diversità del cosmo è classificare i corpi celesti che vi si trovano come stelle, galassie e quasar. E' proprio questo l'obbiettivo del seguente report che, applicando tre diversi modelli (il Model Based Clustering , la Model Based Classification e il Mixture of Experts Model), vuole studiare le caratteristiche di tali raggruppamenti.

Il dataset preso in considerazione contiene 100.000 osservazioni, provenienti dalla 17' release del SDSS (Sloan Digital Sky Survey), un'indagine astronomica su larga scala che raccoglie e utilizza informazioni fotometriche e spettrografiche per identificare i diversi corpi celesti.
Le variabili presenti sono 18:

- **obj_ID**: valore univoco per catalogare l'oggetto
- **alpha**: indica l'angolo di ascensione (analoga alla longitudine ma proiettata sulla sfera celeste anziché sulla superficie terrestre)
- **delta**: indica l'angolo di declinazione (analoga alla latitudine ma proiettata sulla sfera celeste anziché sulla superficie terrestre)
- **u**: filtro degli ultravioletti
- **g**: filtro verde
- **r**: filtro rosso
- **i**: filtro near-infrared
- **z**: filtro infrarosso
- **run_ID**: numero usato per catalogare la scan eseguita
- **rerun_ID**: numero usato per specificare il modo in cui è stata processata l'immagine
- **cam_col**: numero usato per identificare la scanline
- **field_ID**: numero usato per identificare il field della scan
- **spec_obj_ID**: valore univoco per identificare
- **class**: classe dell'oggetto (STAR, GALAXY, QSO)
- **redshift**: valore del redshift
- **plate**: numero che identifica il filtro
- **MJD**: Modified Julian Date, indica la data in cui è stata eseguita la scan
- **fiber_ID**: numero che identifica la fibra utilizzata

I filtri fotometrici servono a misurare la luce proveniente dai corpi celesti su differenti lunghezze d'onda.

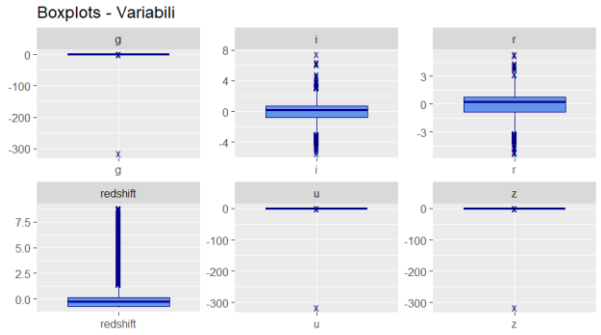
Il redshift è un parametro che consente di determinare la velocità e la distanza dei corpi celesti, nonché l'espansione dell'universo. Esso misura lo spostamento verso il rosso delle lunghezze d'onda della luce provenienti da un oggetto che si sta allontanando dalla Terra. Per l'effetto Doppler, quando un oggetto celeste si allontana da noi, le lunghezze d'onda della luce emessa si allungano, rientrando nello spettro del rosso. Al contrario, rientrano nello spettro del blu qualora un corpo celeste si stia avvicinando. In astronomia uno spostamento verso il blu viene rappresentato da un redshift negativo. Come risulterà evidente nelle fasi successive di questa analisi, il redshift risulta di fondamentale importanza ai fini dell'applicazione dei tre modelli.

Le stelle sono masse gassose che emettono grandi quantità di luce e calore a causa del processo di fusione nucleare che avviene nel loro nucleo.
Le galassie sono vaste aggregazioni di stelle, gas, polvere e materia oscura, legate insieme dalla forza gravitazionale.
I quasar (quasi-stellar radio source, QSO), fonti di energia estremamente luminose e distanti, sono oggetti di aspetto simile alle stelle, solitamente situati al centro delle galassie.

FASE DI ANALISI ESPLORATIVA

Prima di procedere con l'applicazione dei modelli, è necessario analizzare e lavorare sui dati.
Innanzitutto, sono state selezionate solo le variabili rivelanti ai fini della ricerca, ovvero: **u**, **g**, **r**, **i**, **z**, **redshift** che sono state codificate come number e **class** che è stata codificata come factor. Le restanti variabili invece sono state escluse in quanto: **obj_ID**, **run_ID**, **rerun_ID**, **cam_col**, **field_ID**, **spec_obj_ID**, **plate**, **fiber_ID** sono codici identificativi, **alpha** e **delta** rappresentano le coordinate di dove è stato identificato l'oggetto e **MJD** è una data.
Quindi sono stati fatti gli opportuni controlli e si è verificato che nel dataset non sono presenti valori mancanti.

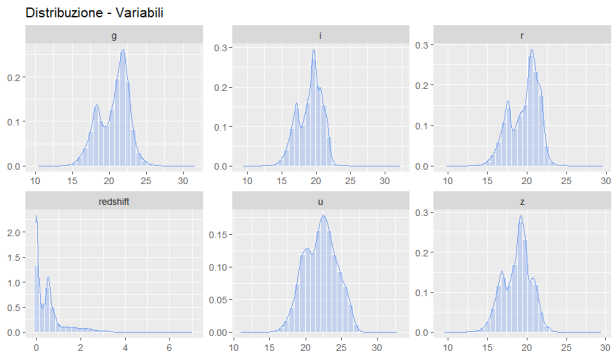
Outlier:



Dai boxplot delle sei variabili si può notare la presenza di diversi outlier. Per coerenza della divisione in gruppi e della classificazione però, è stato valutato di non rimuoverli, ad eccezione di un' unica osservazione, che ha un **g** = -9999, **u** = -9999 e **z** = -9999. Tale osservazione, estremamente anomala sia di per sé che rispetto alle altre, potrebbe essere dovuta ad un errore di inserimento dei dati all'interno del dataset. Dal momento che un valore così basso di **g**, **u** e **z** non risulta affatto plausibile, è stato valutato di rimuoverla dalle analisi successive.

Distribuzione delle variabili:

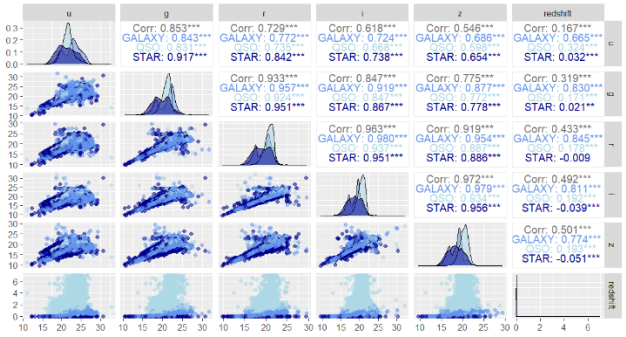
Per ciascuna delle variabili selezionate, esclusa **class**, si è costruito un istogramma con sovraimposta una stima della densità non parametrica.



Dall'analisi di tali grafici emerge che **g**, **i**, **u**, **z** e **r** si muovono nello stesso range di valori e la loro distribuzione risulta essere abbastanza simmetrica e centrata intorno al valore 20. Diversa è la forma della distribuzione della variabile **redshift**, il cui range di valori sembra muoversi tra 0 e 7 e che presenta una notevole asimmetria positiva. Ciò può essere dovuto al fatto che, essendo un indicatore per la distanza tra la Terra e un corpo celeste, a mano a mano che aumenta, e che quindi ci si allontana dalla Terra, il numero di corpi celesti osservati diminuisce.

Le distribuzioni delle variabili sopra rappresentate presentano più picchi, suggerendo una distribuzione multimodale. Ciò è coerente con l'ipotesi di trovarsi di fronte a un modello mistura, con diverse sottopopolazioni che presentano caratteristiche differenti le une dalle altre

Analisi delle correlazioni:



Da tale grafico risulta evidente che le variabili **u**, **g**, **i**, **r** e **z** sono altamente correlate tra di loro, mentre è inferiore la correlazione tra **redshift** e le altre.

Ciò suggerisce che gran parte della variabilità dei dati, e quindi delle informazioni in essi contenute, può essere spiegata da un numero inferiore di variabili e verosimilmente una di queste può essere **redshift.**

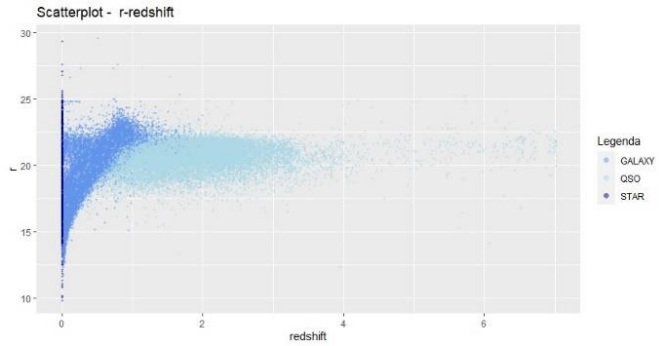
Analisi componenti principali:

Per tale ragione si è deciso di svolgere un'analisi delle componenti principali, al fine di selezionare le variabili che sintetizzano al meglio le informazioni contenute nel dataset, senza perdere troppe informazioni.

In particolar modo si può notare come le prime due componenti principali riescano a spiegare da sole il 90.58996% della varianza totale. Dunque, per ciascuna di esse si è scelto di selezionare la variabile che presenta loading più elevato in valore assoluto e che quindi ha il peso maggiore nel determinare quella specifica componente principale.

Tali variabili, rispettivamente per la prima e la seconda componente principale, sono **r** e **redshift**.

Focus su r e redshift:



Attraverso l'analisi dello scatterplot tra **r** e **redshift** e distinguendo i punti per la variabile **class** è possibile confermare l'accuratezza delle decisioni.

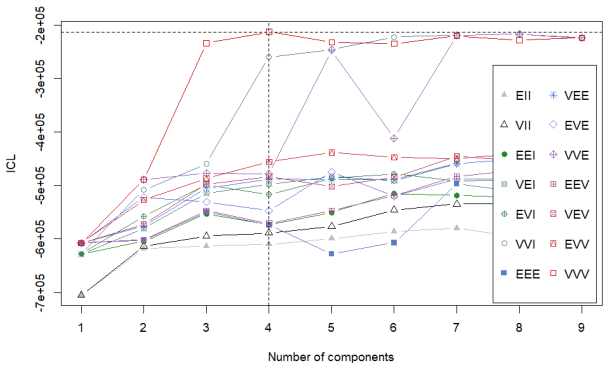
In particolar modo si osserva che per le stelle, il valore del **redshift** è prevalentemente attorno allo zero, mentre **r** varia considerevolmente, in un intervallo che va da 9 a 30.

Per i quasar, il range di valori del **redshift** è molto più vasto rispetto alle stelle, va da 0 a 7. Questo è dovuto al fatto che i quasar si trovano generalmente a grandi distanze dalla Terra, ma grazie alla loro elevata

luminosità, possono essere rilevati, a differenza delle stelle. Inoltre, la variabile **r** per i quasar varia in un range di valori più limitato rispetto alle stelle.
Infine, le galassie presentano dei valori di **redshift** che vanno da 0 a circa 2. Per valori bassi di **redshift**, **r** assume un intervallo simile a quello delle stelle (da 9 a 30), ma con l'aumentare del **redshift**, i valori di **r** tendono a concentrarsi in un range più stretto (da 20 a 25)

MODEL BASED CLUSTERING

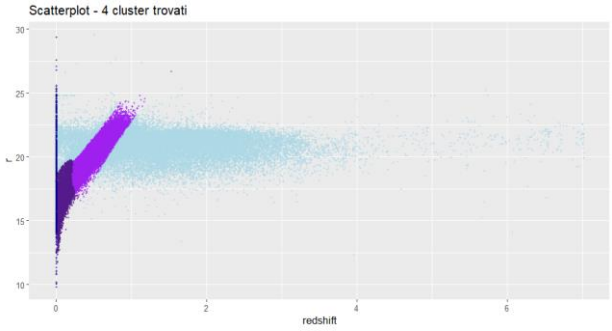
L'obbiettivo è valutare se si riesca a dividere le osservazioni in gruppi corrispondenti alle diverse categorie di corpi celesti.
Si passa all'applicazione del Model Based Clustering, tecnica di unsupervised learning che permette di individuare cluster di osservazioni in cui si ipotizza che i dati provengano da distribuzioni probabilistiche diverse, ognuna con i propri parametri. Nel Model Based Clustering si è sottoposti ad una duplice scelta: una relativa al numero di componenti della mistura (e quindi di cluster) e una relativa ai vincoli da imporre sulla struttura di varianza e covarianza. Il BIC e l'ICL sono i criteri utilizzati per scegliere il numero di componenti. Nella analisi si è scelto inizialmente di non imporre vincoli né sul numero di cluster né sul tipo di modello, lasciando che sia il software a stabilire tra tutti i modelli quale sia il migliore e quale sia il numero ottimale di cluster. Si è preferito usare come criterio l'ICL che rispetto al BIC introduce una penalizzazione per un numero eccessivo di cluster.



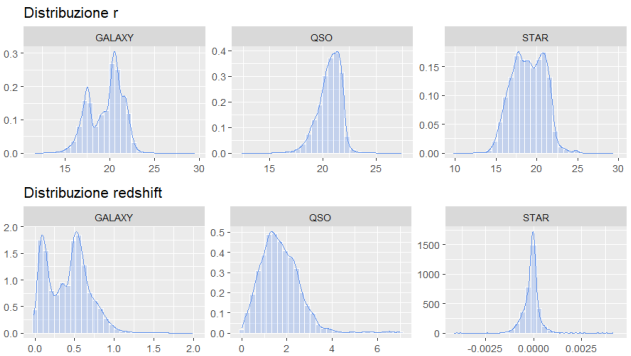
Il plot dell'ICL restituisce come modello migliore il modello VVV con 4 componenti, seguito da un modello VVI con 8 componenti e da un modello VVE con 8 componenti (il simbolo è nascosto dietro a quello del modello VVI). Tali conclusioni risultano essere confermate anche dal comando mclustICL, di cui si riporta il risultato di seguito.

```
> summary(ICL)
Best ICL values:
              VVV,4        VVI,8        VVE,8
ICL      -212349.5 -216555.017 -216606.766
ICL diff      0.0   -4205.519   -4257.267
```

A fronte di tali risultati, è stato eseguito il Model Based Clustering imponendo modelNames = 'VVV' e G = 4. Il modello VVV implica che volume, forma (quindi la lunghezza degli assi) e orientamento (quindi l'inclinazione degli assi) dei cluster siano variabili.



Confrontando questo scatterplot con quello reale, si vede che il modello riesce a cogliere molto bene i gruppi rappresentanti stelle e quasar. Per quanto riguarda le galassie, invece, esse risultano essere divise in due gruppi, uno con valori di **redshift** e **r** più bassi e uno con valori di **redshift** e **r** più elevati.
Si è quindi calcolata l'entropia di tale clusterizzazione, che è risultata pari a 11069.85. Essendo l'entropia una misura non normalizzata, è necessario confrontarla col suo valore soglia, in questo caso pari a 138628. Sulla base di tali risultati si è quindi potuto concludere che la clusterizzazione risultante è buona in quanto il valore dell'entropia è molto inferiore al suo valore di riferimento. Nonostante ciò, tale valore rimane comunque molto lontano da zero, cioè distante dalla situazione in cui si ha incertezza nulla e una suddivisione in cluster delle osservazioni perfetta.
A fronte del fatto che in questo caso le etichette vere sono note e sono in numero pari a tre, ci si è chiesti quale fosse il motivo per cui il numero di cluster ottimale riportato secondo l'ICL fosse quattro.
Si è pensato di fare un istogramma delle variabili **r** e **redshift** per ogni modalità della variabile **class**:

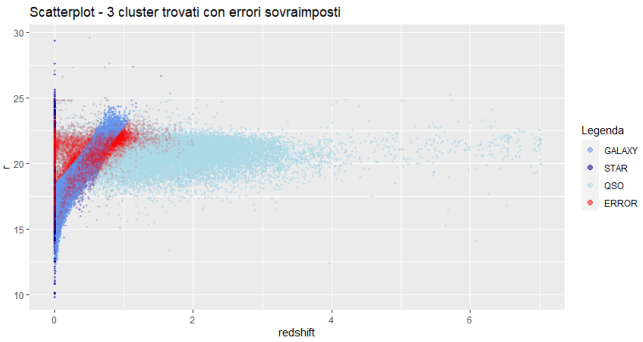


Sulla base di tali istogrammi, si può notare che per la modalità *GALAXY* si ha un andamento bimodale sia per la variabile **r** che per la variabile **redshift**. Ciò fa presupporre che l'algoritmo abbia trovato due sottoinsiemi diversi di galassie.

Mclust con 3 componenti:

Successivamente si è riproposta la stessa tecnica di analisi, imponendo però un numero di componenti pari al numero di etichette reali. In questo modo è possibile calcolare una serie di indici, che non sarebbe possibile

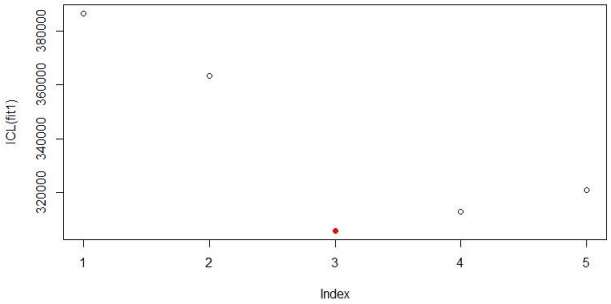
ricavare nel caso in cui le etichette vere non siano note e che permettono di valutare la bontà della clusterizzazione.
Il Classification Error Rate (CER) fornisce la proporzione di osservazioni misclassificate: in questo caso particolare, risulta essere pari a 0.07213072, valore molto basso.
Di seguito viene rappresentato lo scatterplot delle osservazioni suddivise nei diversi gruppi, con sovraimposto in rosso i punti corrispondenti alle osservazioni misclassificate. Esse risultano concentrarsi nella zona del grafico caratterizzata da valori di **r** compresi tra 15 e 25 e valori di **redshift** compresi tra 0 e 2.



L'Adjusted Rand Index (ARI), che misura la similarità tra le etichette reali e le etichette previste, risulta essere pari a 0.7898323, ed essendo vicino a 1, ci indica che le due partizioni sono simili.
Infine è stata calcolata la Confusion Matrix, che confronta le predicted labels con le true labels e permette di ottenere alcune misure che offrono la possibilità di trarre conclusioni circa la bontà della clusterizzazione. L'accuracy (1-CER), che misura sul totale delle osservazioni quante ne sono state assegnate correttamente ai gruppi, è pari al 92.79%. La sensitivity per le classi *GALAXY*, *QSO* e *STAR* è rispettivamente pari al 91.70%, all'89.25% e al 98.88%, ad indicare che il modello presenta un'elevata capacità di identificare correttamente le vere classi. La specificity, invece, è pari al 94.88% per le galassie, al 94.25% per i quasar e al 99.40% per le stelle. Ciò significa che il modello presenta un'elevata capacità di evitare di assegnare osservazioni ad una delle classi, quando realmente non vi appartengono.
Tutti gli indici calcolati permettono di concludere che la suddivisione dei dati in tre cluster è molto buona.

MIXTURE OF EXPERTS MODEL

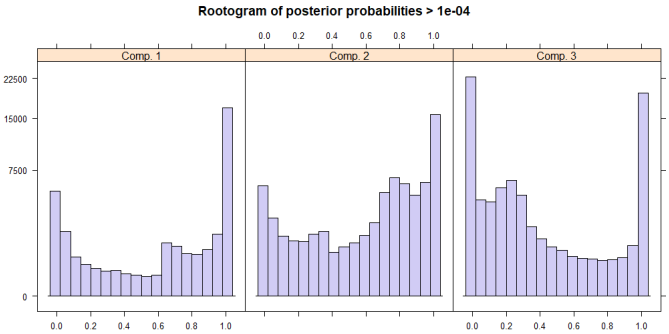
Successivamente ci si è chiesti se sia possibile raggruppare i corpi celesti sulla base di valori simili del filtro rosso (cioè di **r**), e se tali valori dipendano dalla variabile **redshift**, andando anche a verificare se i gruppi di osservazioni che emergono coincidono con le etichette vere che si hanno a disposizione. Si usa **r** come variabile risposta, mentre **redshift** come covariata, ipotizzando che tale covariata vada ad influenzare non solo le componenti della mistura, ma anche i pesi, arrivando così a costruire un Full Mixture of Experts Model. Sull'expert network viene implementata una regressione lineare gaussiana, mentre sul gating network una regressione multinomiale con link logit. Viene quindi implementata la funzione *stepFlexmix* che permette individuare il numero di gruppi ottimale da cui far partire successivamente l'algoritmo sulla base dell'ICL : dai risultati emerge che tale numero è tre.



In seguito, imponendo k=3, viene eseguita la funzione *flexmix* (e quindi il processo di stima) diverse volte al fine di selezionare il modello con ICL più basso. Il modello individuato divide i dati in tre gruppi contenenti rispettivamente 24234, 49986 e 25779 osservazioni.
Andando a stimare e fare inferenza sui parametri dell'expert network, si vede che il coefficiente stimato di **redshift** è significativamente diverso da zero in tutti e tre i gruppi individuati, andando così a confermare l'ipotesi che avesse effettivamente influenza sui valori di **r**.
I p-value del test di significatività sono infatti rispettivamente per i tre gruppi pari a 2.2e-16, 2.2e-16, 8.093e-12 e i coefficienti di **redshift** sono pari a -45.016172 per il gruppo 1, 7.352676 per il gruppo 2 e -0.0459050 per il gruppo 3.
Le stime dei coefficienti della regressione multinomiale sui gating network sono pari a :

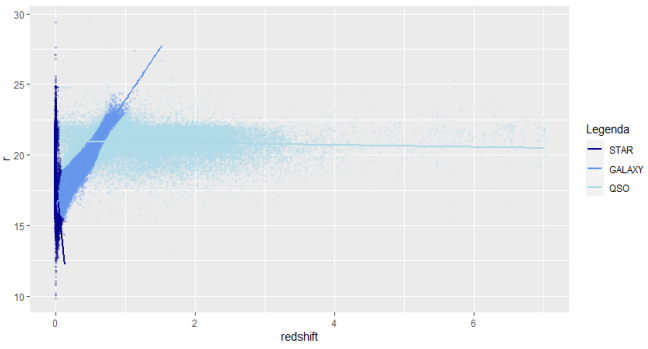
```
> parameters(bestfinal ,which="concomitant")
                1        2         3
(Intercept) 0 -2.91597 -6.719655
redshift    0 65.61614 70.931098
```

Si passa quindi alla rappresentazione del rootogram al fine di vedere dal punto di vista grafico quanto ben sono separati i gruppi:



Tale grafico mostra dei risultati abbastanza soddisfacenti, in quanto per tutti tre i gruppi si hanno dei picchi in corrispondenza dell'uno e dello zero.

Questo suggerisce una buona separazione dei cluster, separazione che non è però perfetta perché ci sono delle aree abbastanza concentrate al centro dell'intervallo, che indicano la presenza di una leggera sovrapposizione tra i gruppi.
Viene successivamente rappresentato il Mixture of Regression Model:



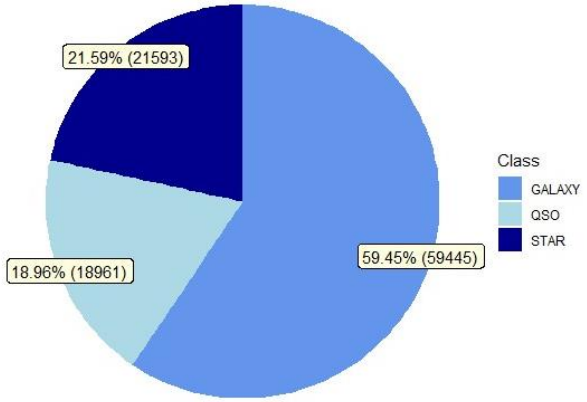
Dal grafico emerge che le rette riescono ad interpolare abbastanza bene i dati, andando a cogliere in maniera adeguata anche l'orientamento dei gruppi.
Grazie alla disponibilità delle reali etichette si riesce a calcolare il *classification error rate*, che risulta pari a 0.1237212.
Si può quindi concludere che anche il Mixture of Experts Model riesce a suddividere in maniera soddisfacente i dati nei diversi gruppi: c'è da tenere presente un *error rate* che non risulta essere particolarmente basso e che ci indica che circa il 12.4% delle osservazioni vengono misclassificate.

MODEL BASED CLASSIFICATION

La Model Based Classification è una tecnica di supervised learning grazie alla quale, tramite l'allenamento dei modelli di mistura di normali sul trainig set, la selezione del modello migliore sulla base del CV e l'applicazione di tale modello sul test set, si riesce a classificare nuove unità statistiche nei diversi gruppi. Dal momento che si è notato nelle analisi precedenti un andamento bimodale nel sottogruppo costituito dalle galassie, nel caso in questione si è scelto di sfruttare il metodo MDA (Mixture Discriminant Analysis). Tale metodo presuppone che la popolazione sia costituita da una serie di component densities che a loro volta sono delle misture di normali d-variate, permettendo così di identificare anche curve di livello di forme non regolari.

Analisi dati Imbalanced:

Prima di iniziare con la Model Based Classification, con la quale si vuole individuare una regola di classificazione in grado di predire il gruppo di appartenenza di nuovi corpi celesti, è stato ritenuto opportuno valutare la presenza o meno di imbalanced data, situazione nella quale le classi di un insieme di osservazioni non sono rappresentate in modo equo. Lo squilibrio dovuto alla presenza di classi sovra e sottorappresentate potrebbe influenzare in maniera negativa i metodi di classificazione, che potrebbero assegnare troppe osservazioni alla classe più grande, impendendo di capire cosa succede in quelle più piccole.



Tale grafico mostra le frequenze relative percentuali e la numerosità delle tre diverse classi. Risulta evidente che la classe *GALAXY* è sovrarappresentata, mentre le classi *QSO* e *STAR* risultano essere sottorappresentate. Anche in questo caso si è deciso di lavorare con le variabili **r** e **redshift**, in quanto, come si è già visto in precedenza con l'analisi delle componenti principali, esse sono le più significative ai fini delle analisi. Per cercare di risolvere la problematica degli imbalanced data, si è deciso di usare la funzione *downSample*, disponibile nella library *caret* di R, che permette di applicare ai dati la tecnica di undersampling. Questa tecnica prevede l'estrazione casuale di unità statistiche dalla classe sovrarappresentata (in questo caso *GALAXY*) e l'eliminazione di tali unità dal dataset, al fine di trovare un equilibrio nelle diverse classi di osservazioni. Solo in seguito a tale passaggio, è stata applicata la Model Based Classification.

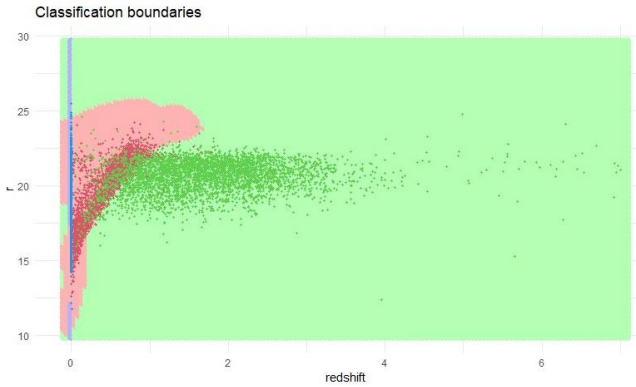
MclustDA:

Il dataset bilanciato viene suddiviso, tramite la funzione di R *CreateDataPartition*, in due sottoinsiemi: il training set, contente l'80% delle osservazioni e il test set, contenente il restante 20%. Successivamente viene implementato un ciclo for in cui viene applicata al training set la funzione *MclustDA*, che permette di ottenere una serie di classificatori, e, sfruttando la funzione *cvMclustDA*, viene eseguita la *V-fold cross validation* con l'obiettivo di selezionare il classificatore migliore, cioè il classificatore con il *cross validation error rate* più basso.
A seguito di cinque iterazioni, si ricava che l'*MclustDA* prevede per la classe *GALAXY* un modello VVV con 5 componenti, per la classe *QSO* un modello VVV con 5 componenti e per la classe *STAR* un modello VVE con 4 componenti. Il *cross validation classification error rate* stimato è pari a 0.0504318, mentre il CER risulta pari a

0.0524789. E' stata calcolata anche la Confusion Matrix che restituisce valori molto elevati di specificity e sensitivity per ciascuna delle classi.

```
> confusionMatrix(predizione$classification, as.factor(testing$class))
Confusion Matrix and Statistics

          Reference
Prediction GALAXY  QSO STAR
    GALAXY   3640  444    0
    QSO       122 3347    0
    STAR       30    1 3792

Overall Statistics

               Accuracy : 0.9475
                 95% CI : (0.9433, 0.9515)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9213

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: GALAXY Class: QSO Class: STAR
Sensitivity                 0.9599     0.8826      1.0000
Specificity                 0.9415     0.9839      0.9959
Pos Pred Value              0.8913     0.9648      0.9919
Neg Pred Value              0.9792     0.9437      1.0000
Prevalence                  0.3333     0.3333      0.3333
Detection Rate              0.3200     0.2942      0.3333
Detection Prevalence        0.3590     0.3049      0.3361
Balanced Accuracy           0.9507     0.9333      0.9980
```

Successivamente si è ritenuto opportuno rappresentare i classification boundaries, al fine di visualizzare anche dal punto di vista grafico dove si trovano le zone di classificazione che sono emerse con l'MDA e dove si trovano i punti veri. In blu vengono visualizzati i classification boundaries e i punti rappresentanti le stelle, in rosso quelli rappresentanti le galassie e in verde i quasar.



Il grafico conferma che il modello permette di cogliere molto bene la vera appartenenza dei punti, che sembrano per la maggior parte rientrare nelle zone di classificazione emerse. Si nota solo un po' di overlapping nella zona intermedia tra quasar e galassie, similmente a quanto era emerso anche dalla rappresentazione dei punti misclassificati nel Model Based Clustering con tre componenti.

In conclusione, la Model Based Classification fornisce riscontri molto positivi circa la precisione con la quale riesce ad assegnare le unità statistiche alle classi di appartenenza. Questa evidenza fornisce un grande supporto per classificazioni successive che potrebbero rendersi necessarie qualora altri corpi celesti venissero osservati e ci fosse il bisogno di determinarne la classe di appartenenza.

Bibliografia:

- https://www.kaggle.com/datasets/fedesoriano/stellar-classification-dataset-sdss17
- https://www.sdss4.org/dr17/
- https://skyserver.sdss.org/dr17/en/home.aspx
- https://www.sdss4.org/instruments/camera/