# **PROGETTO DI STATISTICA NUMERICA**

# **Gruppo L: Gimigliano Marianna, Popa Luis, Ricci Nicola**

**JOHNS HOPKINS UNIVERSITY**

**IONOSPHERE   
DATABASE**

FONTI:

Link database from Kaggle: <https://www.kaggle.com/prashant111/ionosphere>

Link database from UCI: <http://archive.ics.uci.edu/ml/datasets/Ionosphere>

1. **DATASET INFORMATION**

CLASSIFICAZIONE DEI RITORNI RADAR DALLA IONOSFERA UTILIZZANDO LE RETI NEURALI

Il Dataset *“Ionoshpere”* è costituito da un set di 16 antenne ad alta frequenza con una potenza totale trasmessa dell'ordine di 6,4 kilowatt. I valori radar restituiti sono usati per studiare la fisica della ionosfera. Tale sistema è situato a Goose Bay, Labrador, Canada.

Il radar funziona trasmettendo un modello multi-pulse alla ionosfera. Il ricevitore si attiva tra i diversi impulsi e determina la loro velocità misurando lo sfasamento dei ritorni. I segnali ricevuti dalle antenne sono elaborati tramite una funzione di autocorrelazione *R(t, k)* i cui argomenti sono il tempo di un impulso *t* e il numero di impulso *k*. Sono stati rilevati complessivamente 17 numeri di impulsi.

Le istanze in questo Dataset sono descritte da 2 attributi per ogni numero di impulso, corrispondenti ai valori complessi restituiti dalla funzione *R* che studia tali segnali elettromagnetici.

Target: i rendimenti radar "good" mostrano prove di qualche tipo di struttura nella ionosfera mentre, i rendimenti "bad" non lo fanno.

1. **PULIZIA E ORGANIZZAZIONE DEL DATASET**

Come primo step il Dataset è stato caricato, salvato nella sua versione originale in una seconda variabile e ripulito da eventuali dati mancanti.

Per una maggiore comodità nell’utilizzo e lettura del Dataset sono state rinominate tutte le sue colonne, definendo ogni coppia di colonne numeriche che rappresentano un’antenna come una lettera dell’alfabeto e la colonna finale come “target”, essendo in essa contenuti i risultati della nostra analisi.

Quest’ultima, contenente variabili categoriali, è stata poi trasformata in factor per sfruttare il concetto delle etichette.

Viene successivamente controllato, per ogni colonna, il minimo e il massimo delle variabili in essa contenute, così da determinare se ci sono valori esterni ai range previsti.   
Ad eccezione delle prime due colonne, tutte le altre si sviluppano nell’intervallo [-1; 1], senza necessità di rimuovere alcun elemento identificabile come outlier.

Infine, sono state rimosse le colonne “a” e “b”, essendo non significative e con valori differenti rispetto all’andamento generale del Dataset.

1. **SPLITTING**

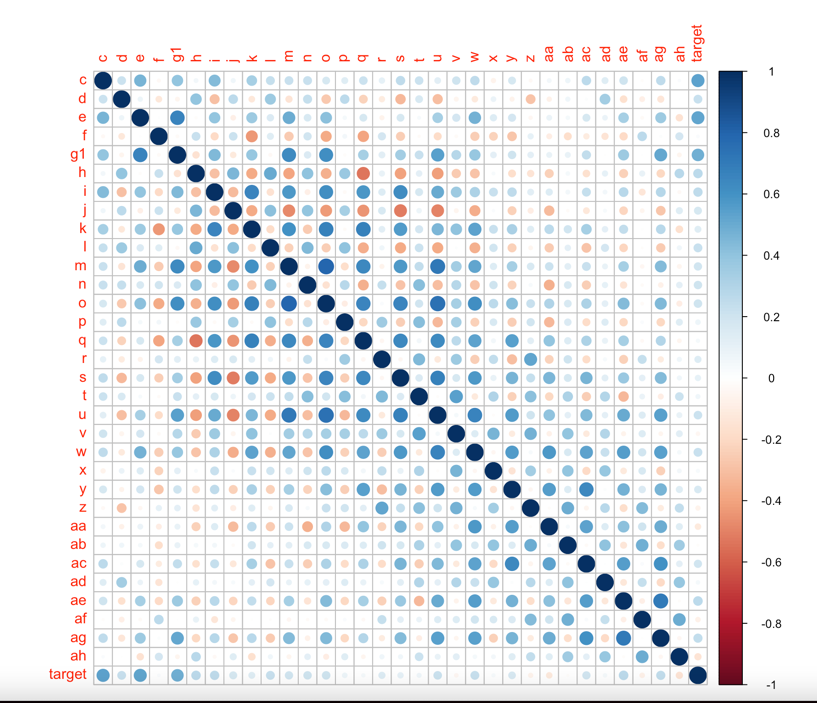
In questa fase il Dataset viene diviso in Training Set, Validation Set e Test Set contenenti rispettivamente 225, 66 e 60 variabili pronte per essere utilizzate per l’analisi.

Viene poi rinominato il Dataset “df” assegnandogli i dati contenuti nel Training Set per comodità di utilizzo nelle fasi successive.

1. **VISUALIZZAZIONE**

Nella fase di visualizzazione sono stati creati più grafici allo scopo di determinare quali colonne, quindi quali antenne, fossero più significative per l’analisi del Dataset.

Inizialmente è stata generata una Matrice di Correlazione:

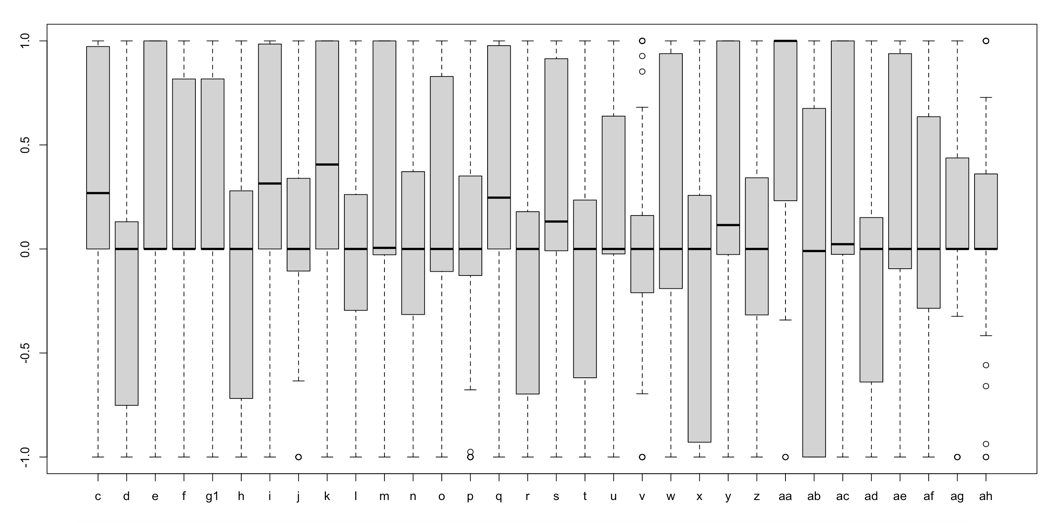


Nella presente matrice si può osservare, nella sezione “target”, quali colonne correlano maggiormente con essa. Si evidenziano:

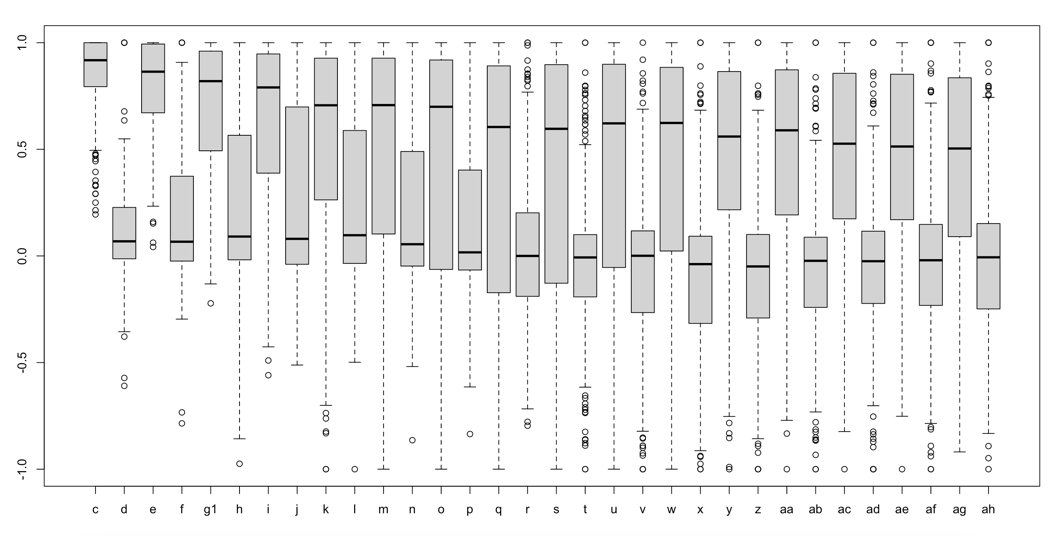
* le prime colonne dispari, caratterizzate da un punto con correlazione compresa tra 0,5 e 1.
* tutte le colonne pari generano punti con correlazione intorno allo 0, quindi risultano essere meno significative.

Per confermare questa ipotesi sono stati creati due grafici Boxplot a partire da due subset contenenti rispettivamente solo le osservazioni con target “bad” e con target “good”:

**BAD**



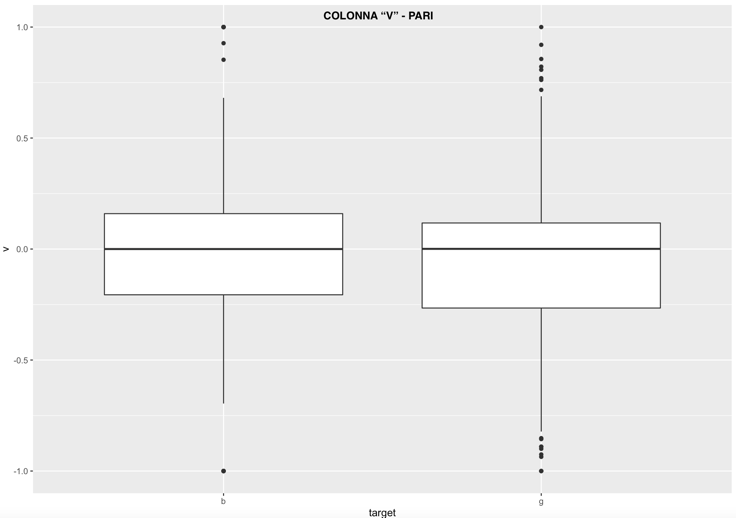
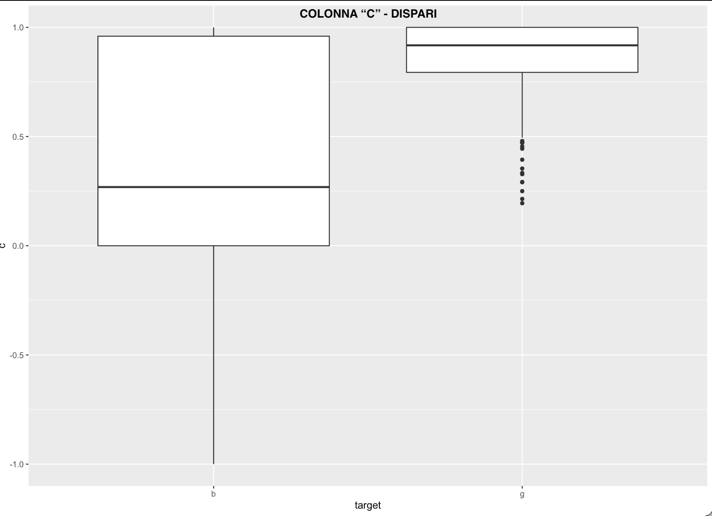
**GOOD**



Si può notare come nel Boxplot “bad” le medie siano prevalentemente intorno allo 0. Nel Boxplot “good” invece, al variare del target, solo la media delle colonne dispari si modifica spostandosi verso valori positivi maggiori.

Questo risultato ci indica che in esse sono contenuti valori che sono significativi per l’analisi.

Osservando più da vicino una colonna dispari e una pari si riconfermano queste ipotesi:



Nella colonna “C” si nota un grande dislivello tra la media dei valori target. I valori “bad” presentano una grande deviazione standard sopra la media, mentre i valori “good” si concentrano in un intervallo totalmente positivo.

Tale difformità tra i dati rende interessante l’analisi delle colonne così strutturate.

Al contrario nella colonna “V” le medie e le deviazioni standard sono all’incirca identiche, quindi questo risultato è da considerare poco rilevante.

A scopo informativo e visivo è stato implementato anche un grafico a torta contenente le percentuali dei due valori target.

1. **SVILUPPO DEL MODELLO DI MACHINE LEARNING**

Durante la fase di addestramento sono stati valutati più modelli SVM con kernel differenti allo scopo di trovare il più efficiente per la struttura del nostro Dataset.

In prima battuta è stato analizzato un modello con kernel polinomiale, sul quale sono state fatte una serie di iterazioni al fine di determinare i parametri ottimali di “cost” e “degree”.

Un altro tentativo viene poi eseguito con un modello avente kernel radiale, usando il medesimo procedimento.

Questi test sono stati sviluppati contestualmente sul Training Set e sul Validation Set.

Tramite le funzioni istanziate in precedenza di Missclassification Rate e Accuracy è stato rilevato sul Test Set che il modello radiale, più efficiente, genera previsioni future molto accurate.

Eseguendo il file “Ionoshpere.SRS”, nel quale è stato implementato un ciclo for che itera su parte del codice del file “Inosphere”, si genera una tabella contenente le metriche di errore per i due modelli:

* Missclassification Rate polinomiale
* Accuracy polinomiale
* Missclassification Rate radiale
* Accuracy radiale

Immagine che contiene tavolo

Descrizione generata automaticamente

Tramite questa tabella viene evidenziato che nella maggior parte delle iterazioni il modello con kernel radiale è più preciso e più efficace in termini di minimizzazione dell’errore e massimizzazione dell’accuratezza.

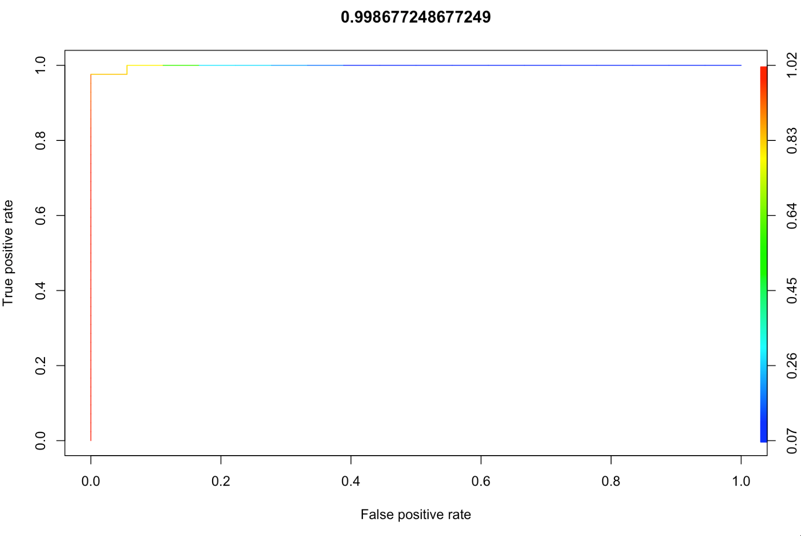
In conclusione, è stato addestrato il modello definitivo usando un kernel radiale con “cost=1” e “gamma=1/9”, il cui errore rilevato viene riportato nella colonna finale della tabella.

1. **INTERPRETAZIONE PROBABILISTICA**

Al fine di sviluppare questo punto sono stati modificati i dati contenuti nella colonna “target” in valori numerici 0 e 1. Viene quindi creato un vettore di probabilità e viene settato un Treshold ottimale per migliorare la performance del modello.

Per provare ciò viene definita una Matrice di Confusione, nella quale si visualizzano valori Falsi Positivi e Falsi Negativi, che sono minimi grazie ad un valore di Treshold uguale a 0.7.

Successivamente è stato implementato il grafico della Curva di ROC per rappresentare visivamente la sensibilità del modello.



In questo grafico ad ogni valore del Treshold nell’intervallo [0; 1] corrisponde una coppia di False Positive Rate e False Negative Rate. Il modello risulta essere estremamente preciso, minimizzando il rapporto tra le coppie.

In più si visualizza il valore AUC, indice della performance del modello, tendente ad 1, ovvero il valore ottimale.

1. **STUDIO STATISTICO**

Tramite strumenti di statistica descrittiva e inferenziale è stata eseguita un’analisi su un campione preso in considerazione.

Sono state studiate le misure del centro (media), di diffusione dei dati (varianza, deviazione standard) e i rispettivi intervalli di confidenza per la media, ovvero gli intervalli per cui i valori trovati sono attendibili.

Questa analisi è stata effettuata su due colonne dispari significative (C ed E) e su due pari non significative (X e F) al fine di evidenziarne nuovamente le differenze.

Immagine che contiene tavolo

Descrizione generata automaticamente

Calcolati gli intervalli di confidenza si nota che le medie trovate rientrano nel range con una precisione del 95%.

Inoltre, come dimostrato nel punto 4, le medie delle colonne dispari si discostano nettamente dallo 0 rispetto a quelle pari.

1. **FEATURE SELECTION**

Grazie all’analisi svolta in precedenza sono state determinate le colonne significative per questa fase. Sono quindi state prese in considerazione, in modo esemplificativo, le prime due colonne dispari del nostro Dataset (C ed E), data la loro forte influenza sulla performance. È stato quindi eseguito il modello finale trovato solo su di esse.

Su quest’ultimo sono state applicate le funzioni di Miscalssification Rate e Accuracy, da cui si nota che se tali colonne sono presenti nel modello, la sua prestazione aumenta notevolmente.