PROGETTO D'ESAME MACHINE LEARNING

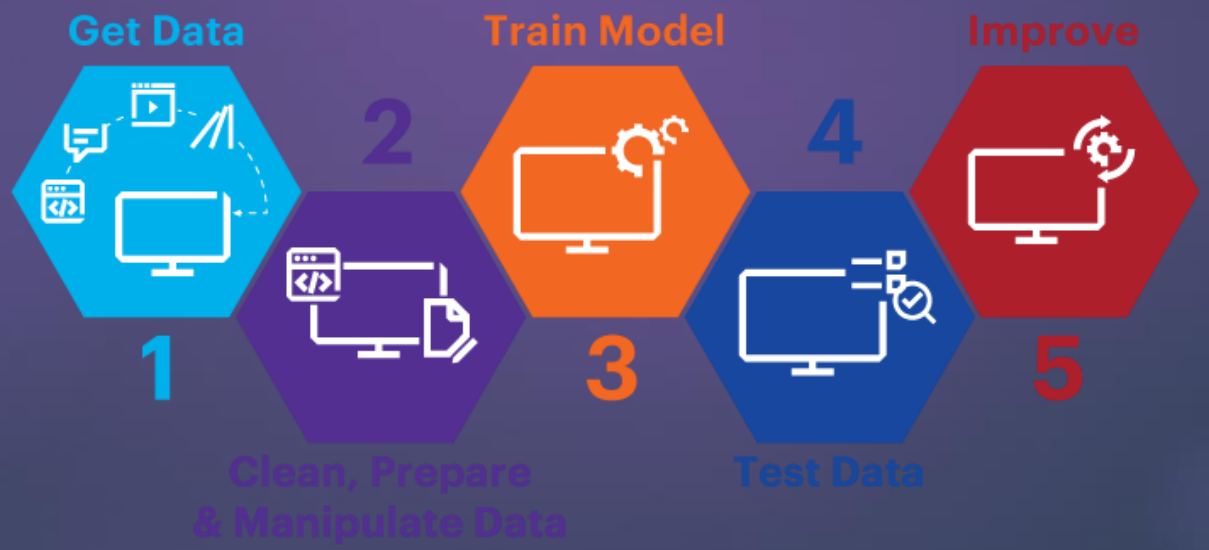
Nicolae Alexandru Andrei
Matricola 829570

Andrea Premate
Matricola 829777

Descrizione del dominio e obiettivi

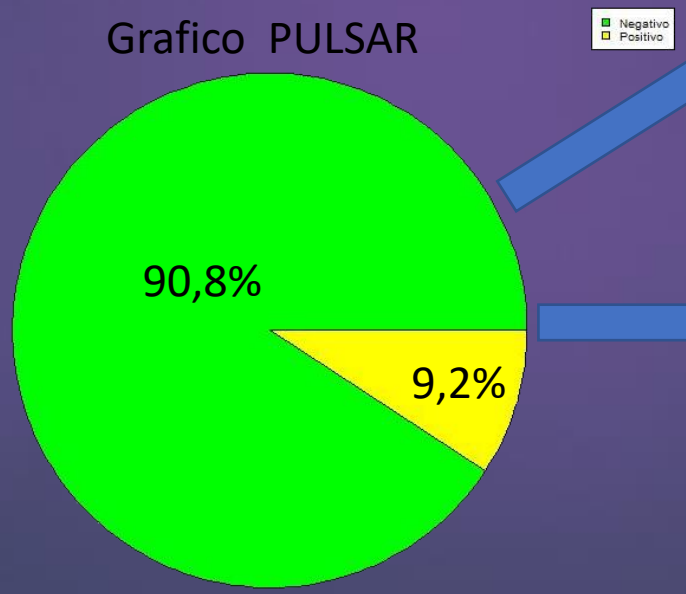
Le “pulsar” sono un raro tipo di stelle di neutroni che producono emissioni radio rilevabili dalla Terra. Alcune statistiche relative a queste emissioni radio sono state raccolte nel dataset HTRU2.

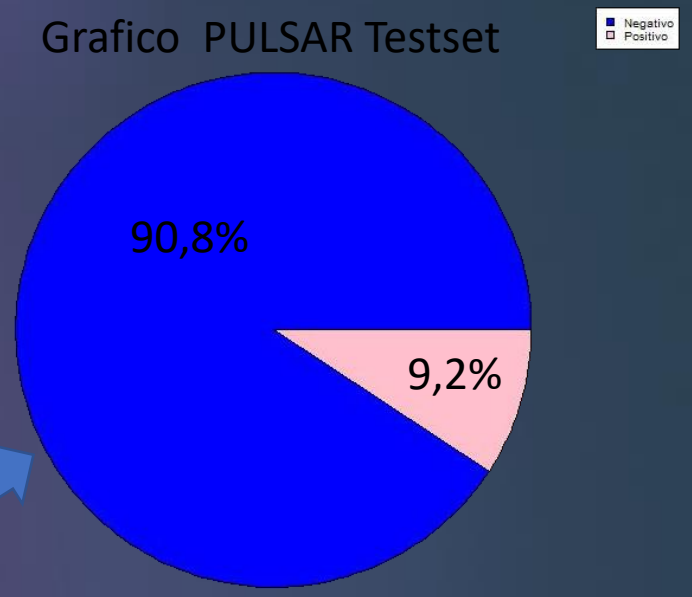
L’obiettivo è quello di poter di distinguere, sulla base di queste statistiche, quando si è effettivamente identificata una pulsar oppure quando si è rilevato solo rumore.

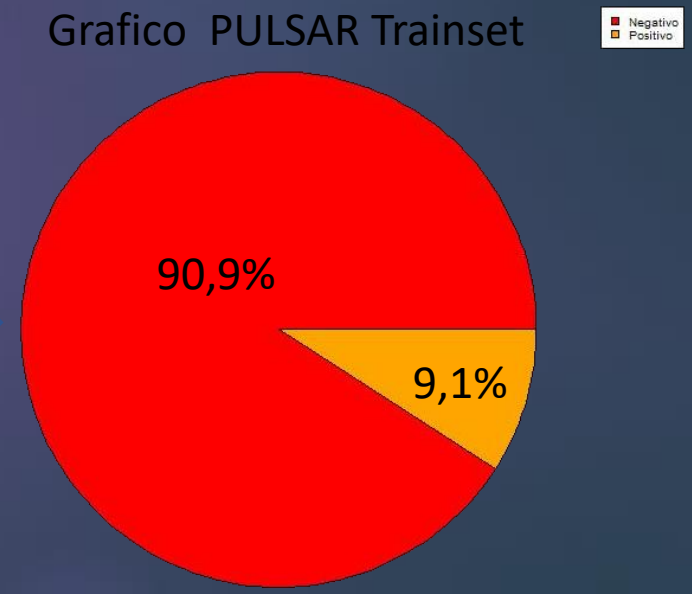


Descrizione del dataset

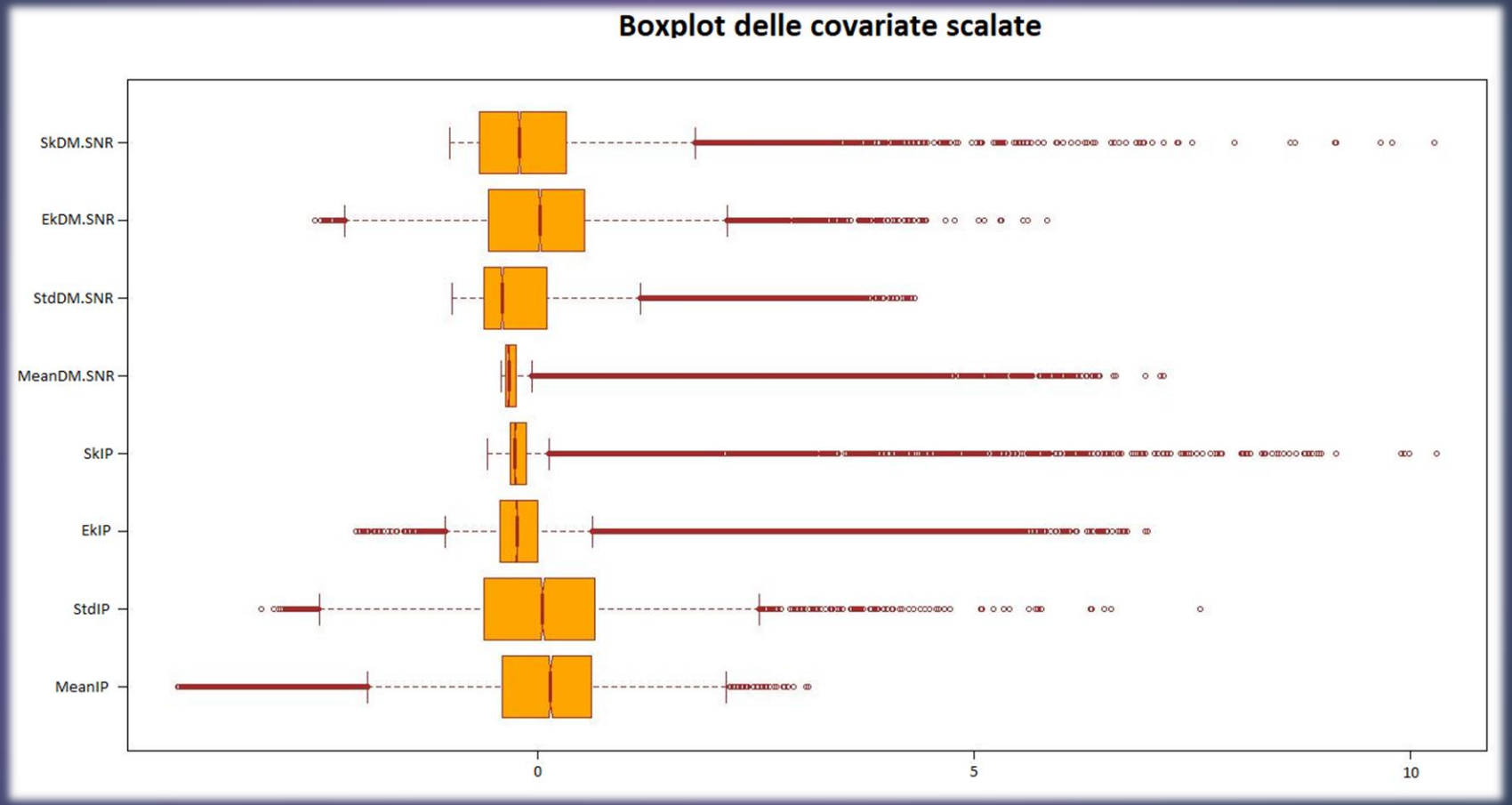
- Dimensione del dataset: 9x17898 (8 attributi numerici e una label)
- Dataset sbilanciato, ma non in modo troppo esagerato





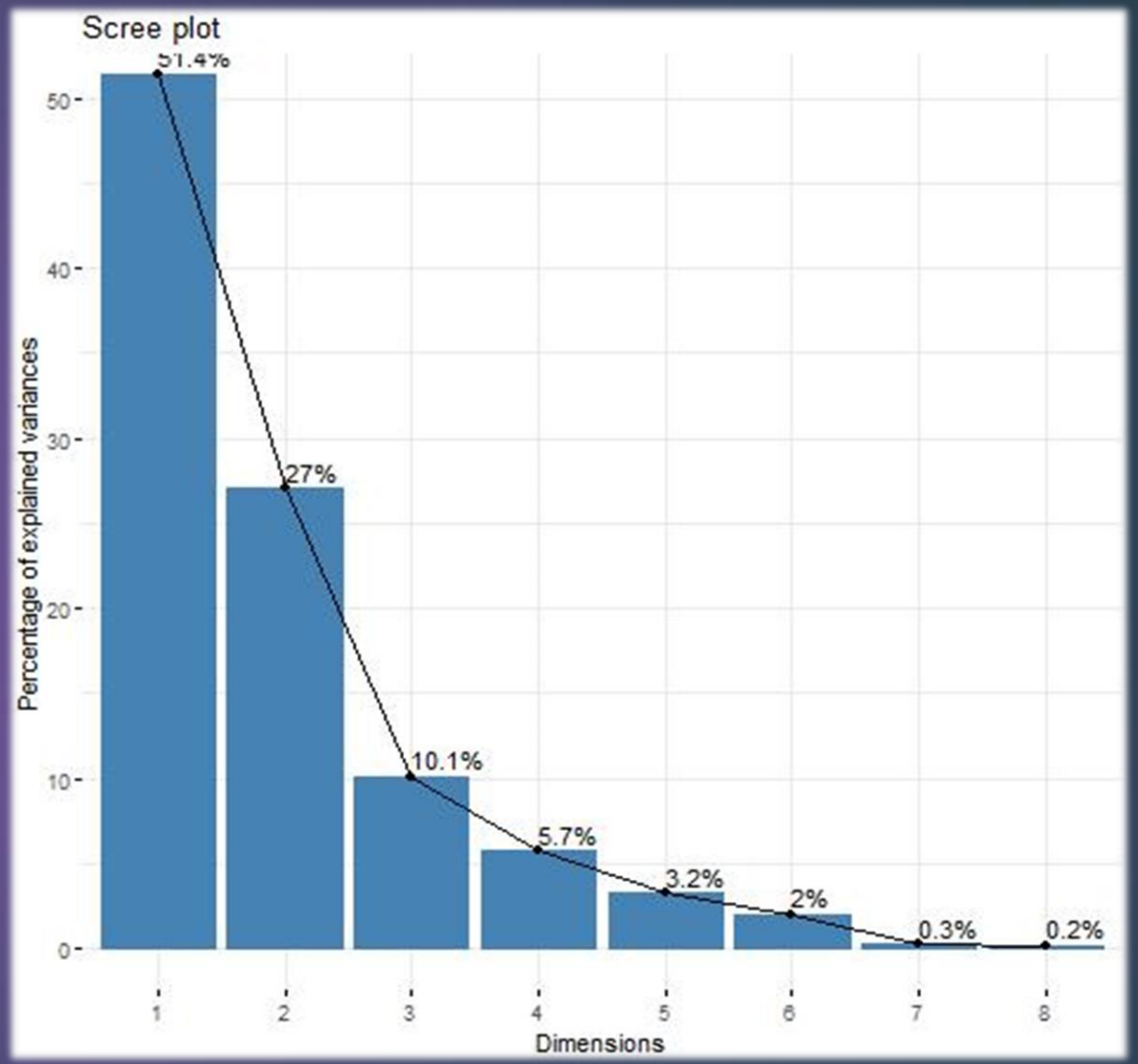


Analisi delle covariate

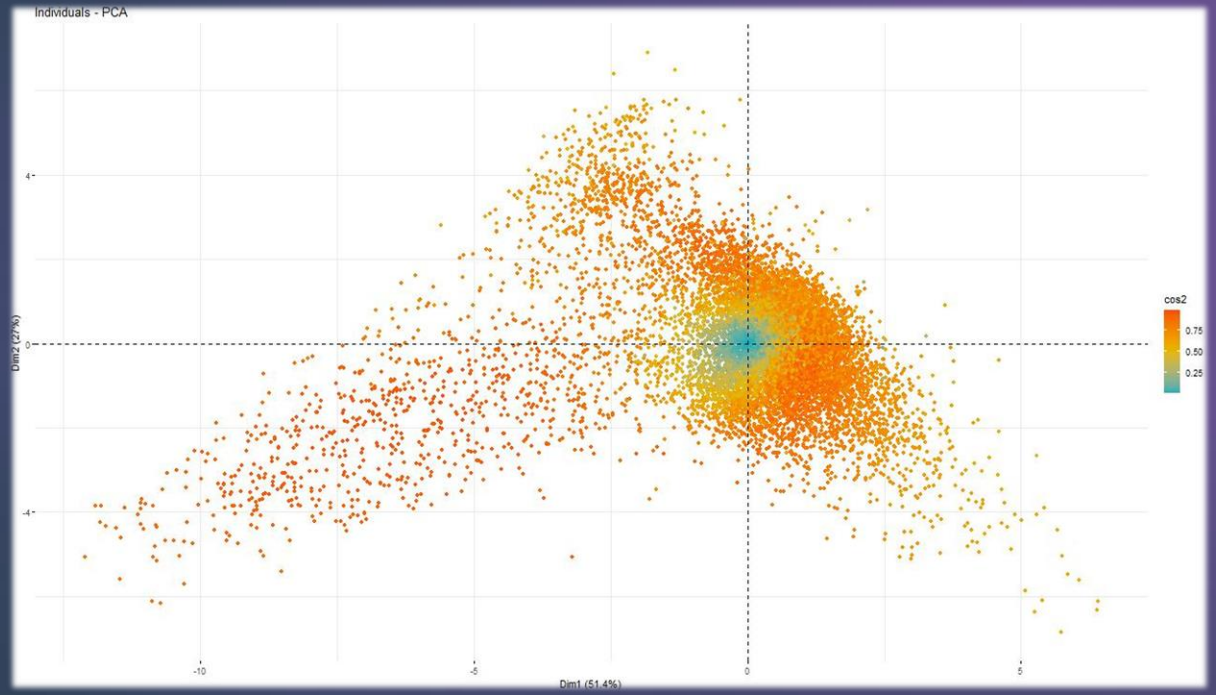


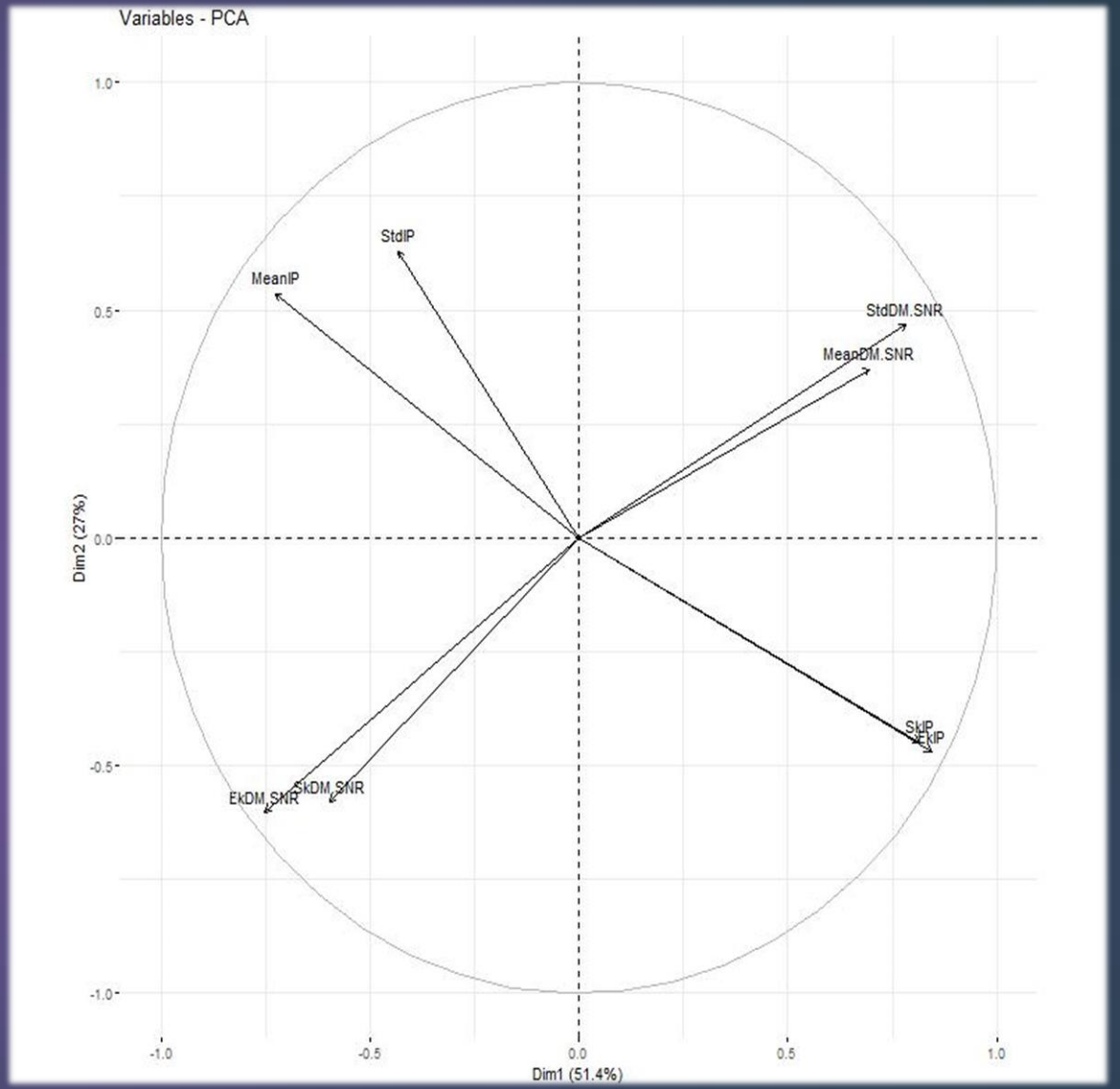
PCA-1

- Per ridurre il numero di covariate abbiamo utilizzato la PCA sul train set. Abbiamo deciso di mantenere tutte le componenti che spiegano almeno il 10% della varianza
- Da 8 a 3 dimensioni.
- Tot varianza spiegata = 88,55%



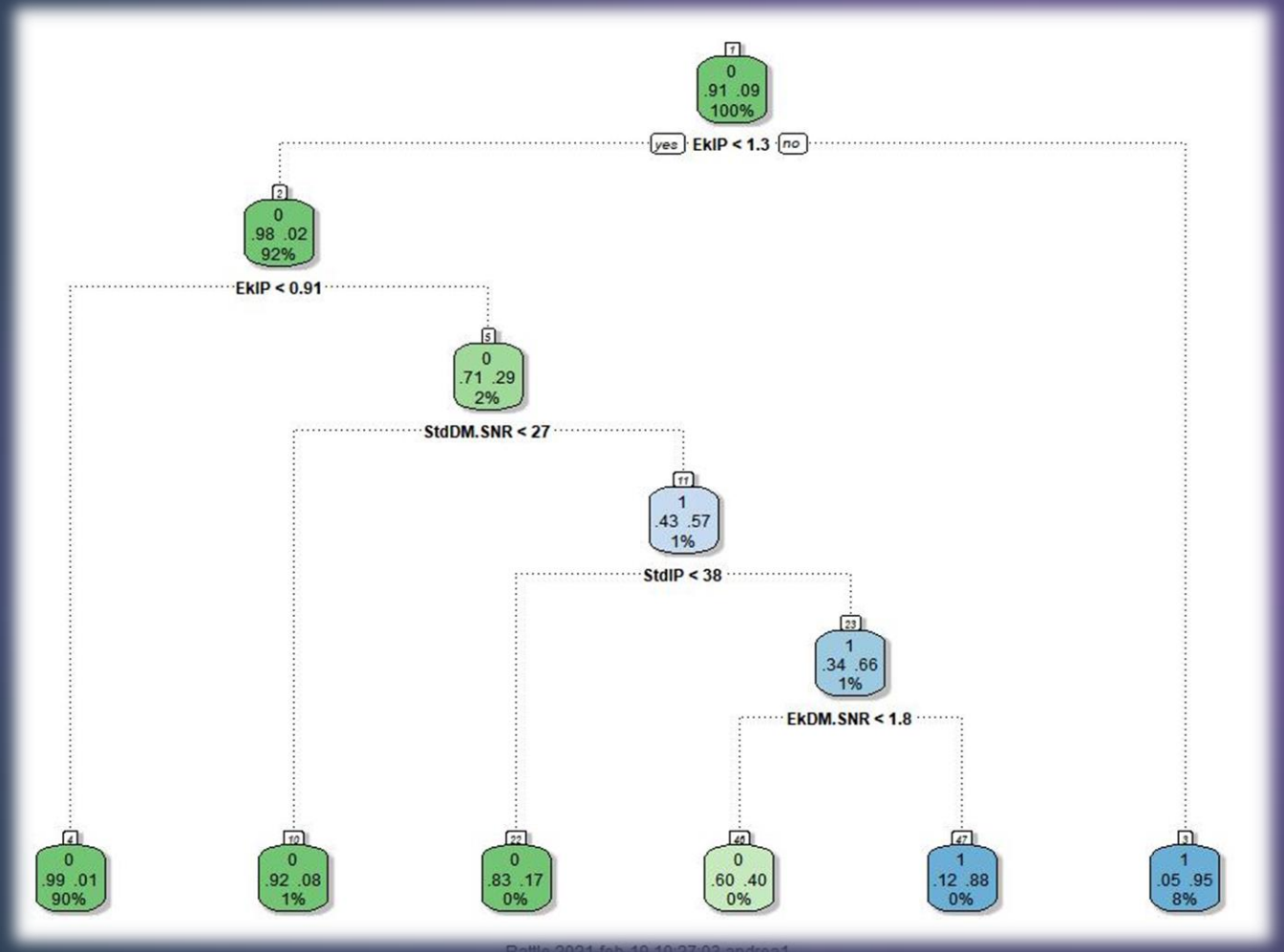
PCA-2

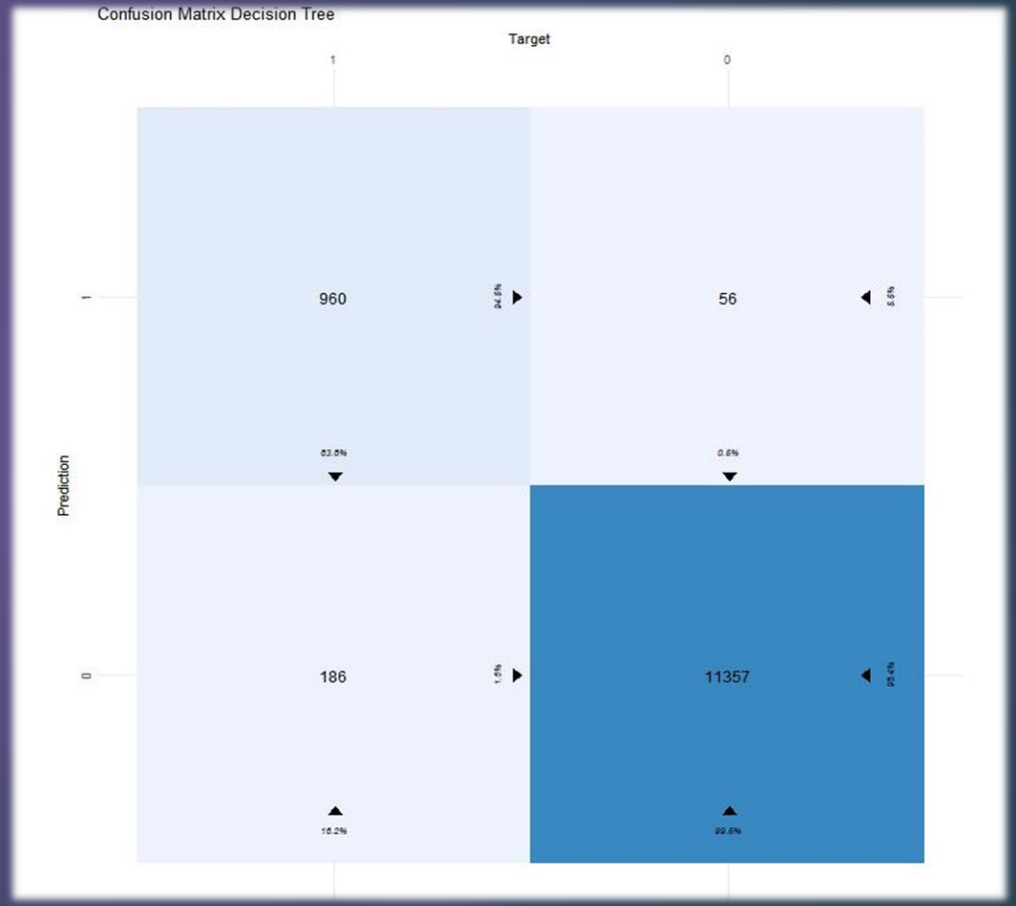

Individuals - PCA
cos2
0.75
0.50
0.25
Dim2 (27%)
Dim1 (51.4%)


Variables - PCA
StdIP
MeanIP
StdDM.SNR
MeanDM.SNR
SkIP
EkIP
EkDM.SNR
SkDM.SNR
Dim2 (27%)
Dim1 (51.4%)

Decision tree (no PCA)

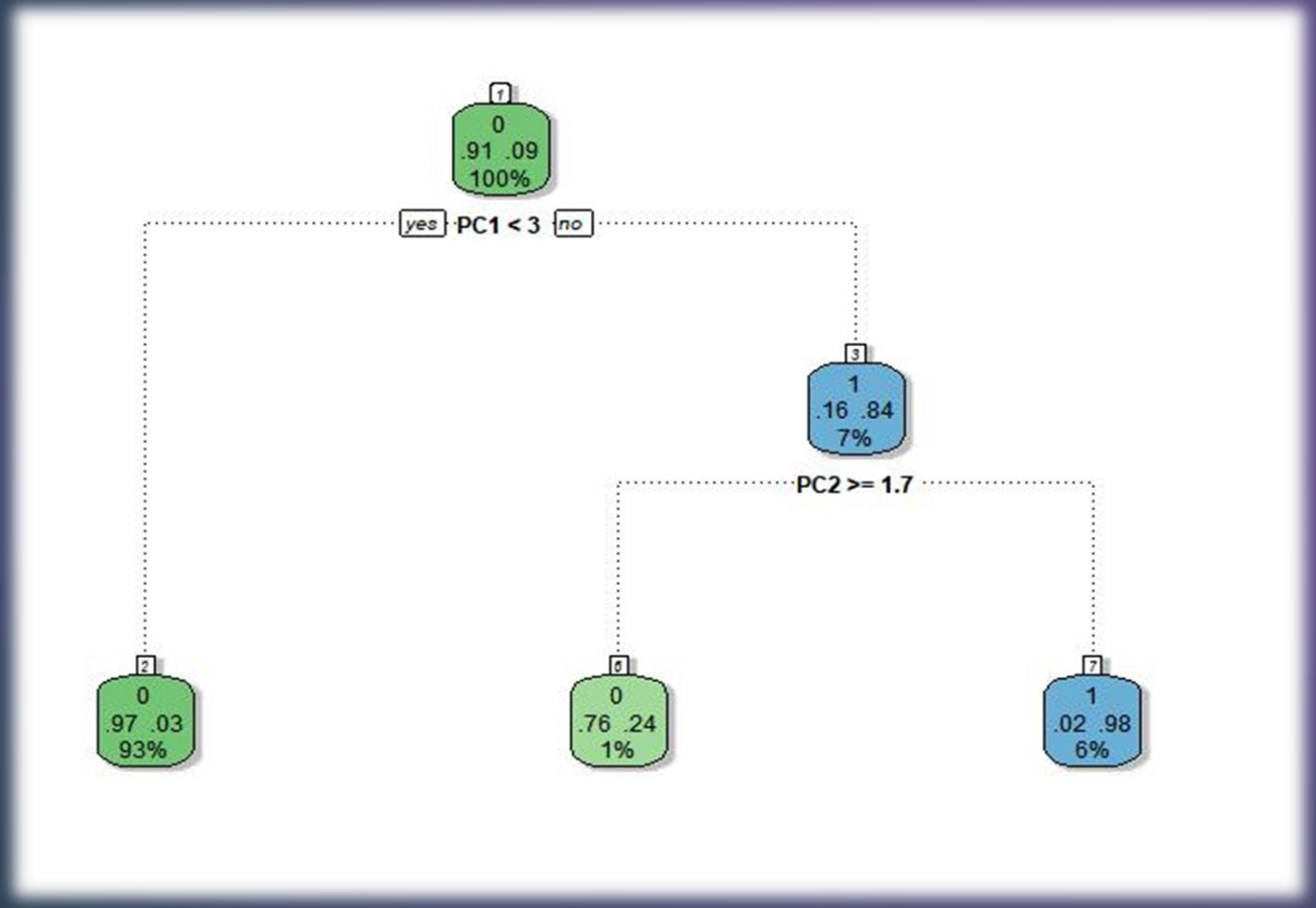
- **Accuracy(test): 0,979022**
- **Accuracy(10-fold cv): 0,979059**
- **Precision: 0,944882**
- **Recall: 0,837696**
- **F-measure: 0,888067**
- **AUC: 0,916395**





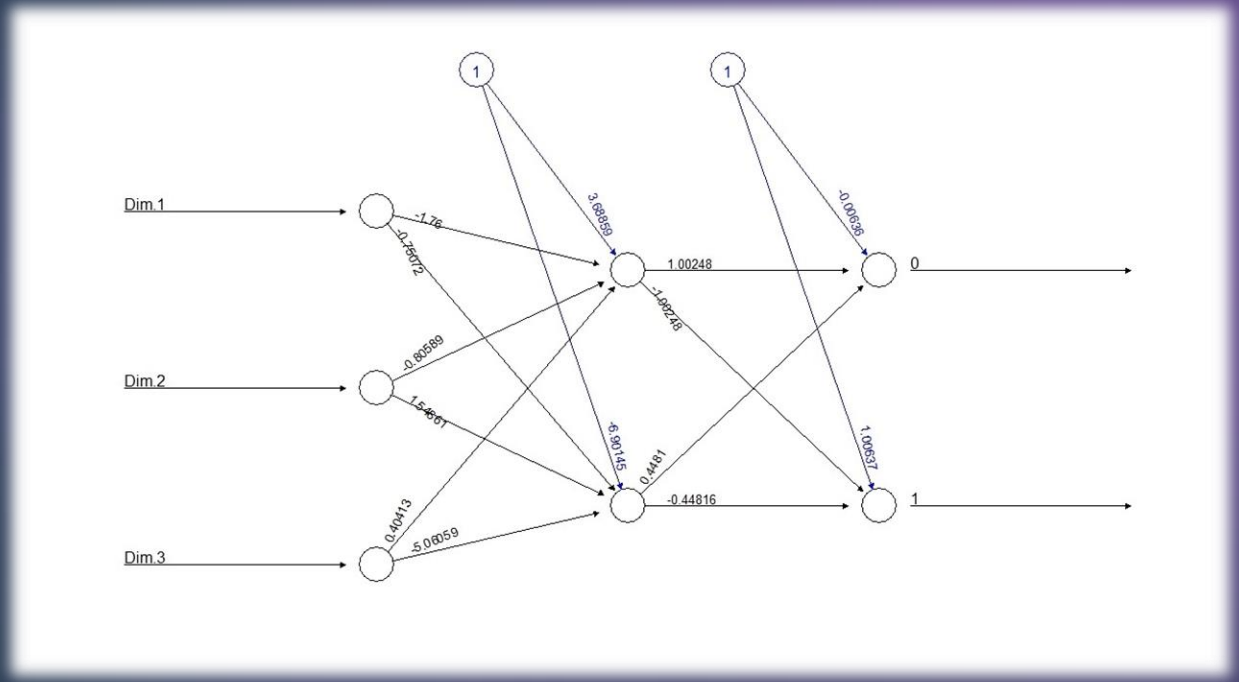
Decision tree (PCA)

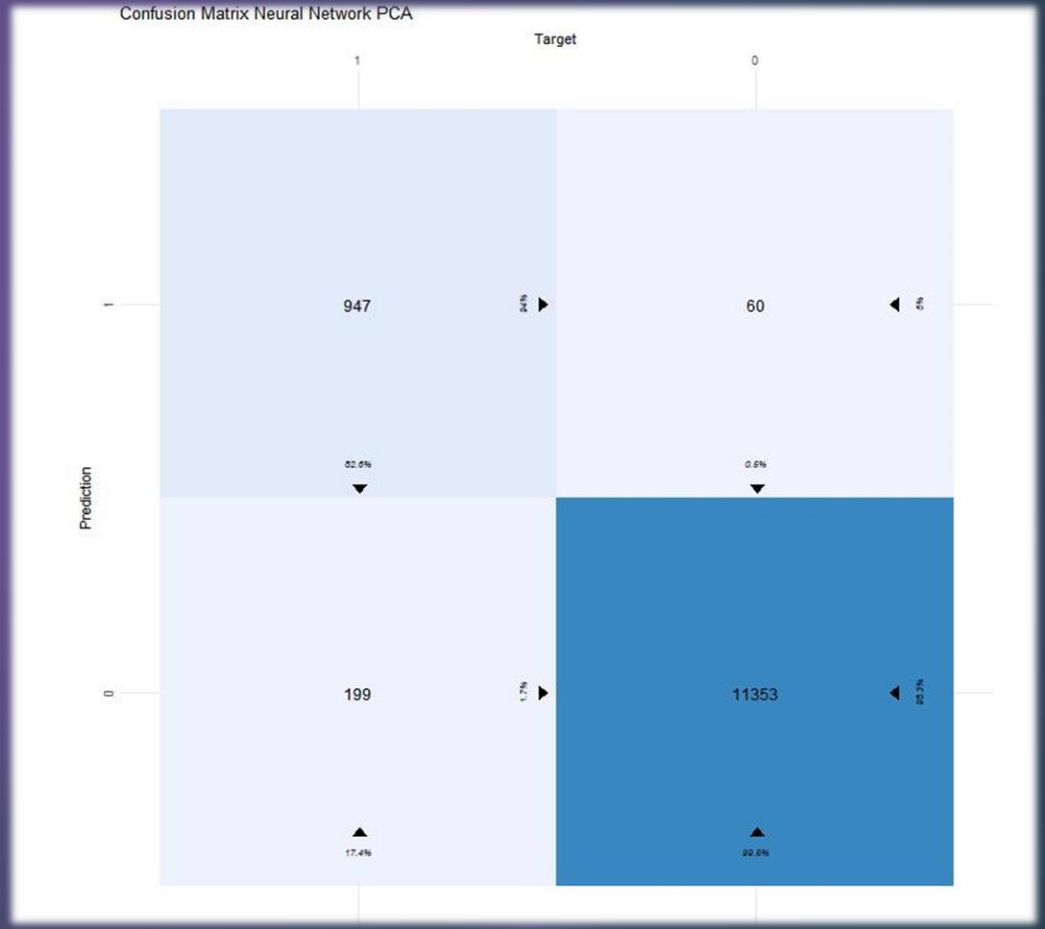
- **Accuracy(test): 0,967972**
- **Accuracy(10-fold cv): 0,969504**
- **Precision: 0,975452**
- **Recall: 0,658813**
- **F-measure: 0,786458**
- **AUC: 0,828574**



Neural network

- **Accuracy(test): 0,977337**
- **Accuracy(10-fold cv): 0,978422**
- **Precision: 0,940417**
- **Recall: 0,826353**
- **F-measure: 0,879703**
- **AUC: 0,910548**





Analisi risultati ottenuti

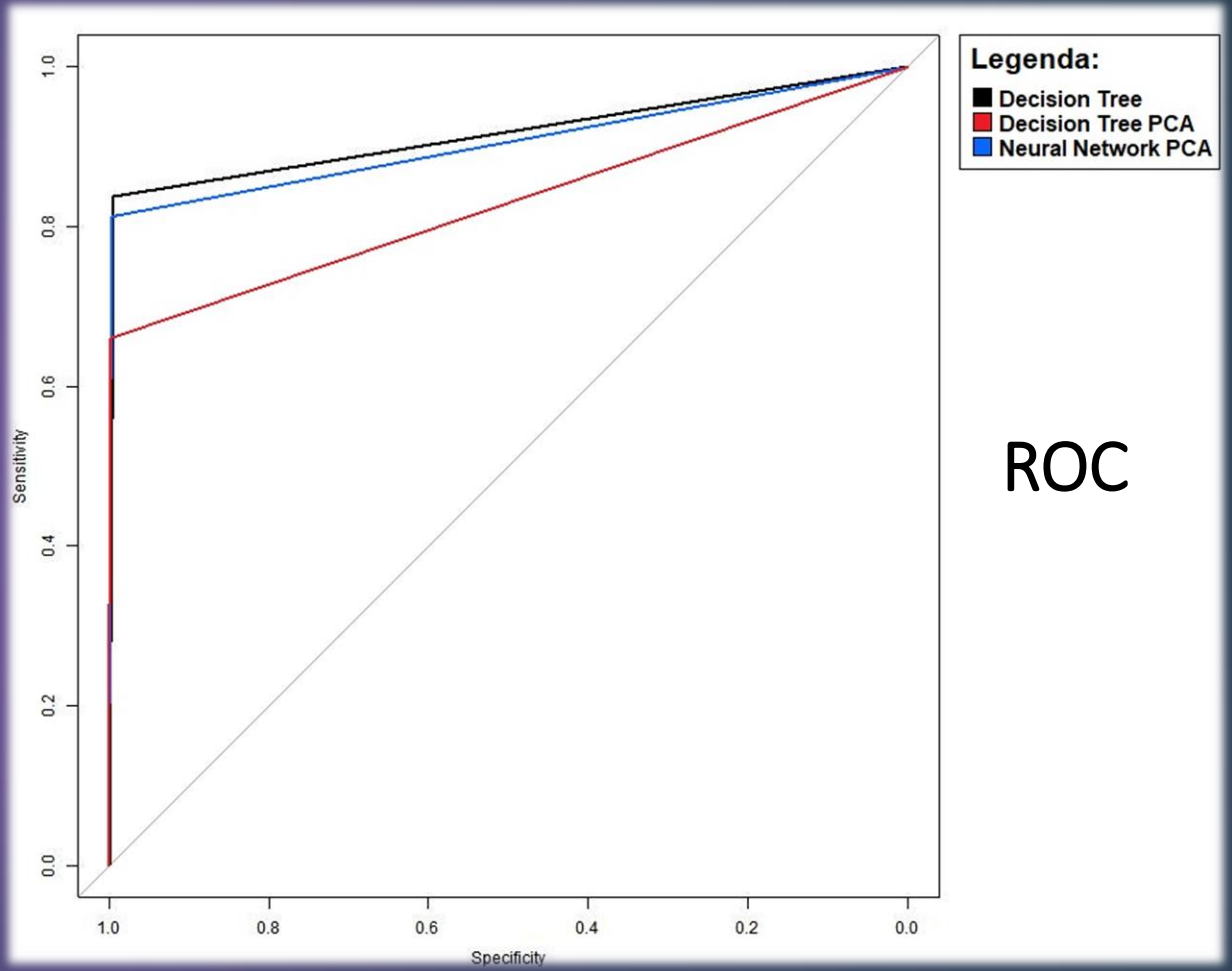
- Modello peggiore: Decision Tree(PCA)
- Altri due modelli pressoché equivalenti

	Train (s)	Predict test (s)	Predict train (s)
Decision Tree	1,479345	0,003003	0,005004
Decision Tree PCA	0,880801	0,003002	0,004004
Neural Network PCA	93,30887	0,006006	0,014012

	Decision Tree	Decision Tree (PCA)	Neural Network (PCA)
Accuracy (test)	0,979022	0,967972	0,977337
Accuracy (10-fold cv)	0,979059	0,969504	0,978422
Precision	0,944882	0,975452	0,940417
Recall	0,837696	0,658813	0,826353
F-measure	0,888067	0,786458	0,879703
AUC	0,916395	0,828574	0,910548

Conclusioni

- Le misure di accuracy, precision, recall, F-measure e AUC ci hanno consentito di stabilire che l'albero di decisione (no PCA), costituisce, con un piccolissimo margine di vantaggio, il modello migliore creato. D'altra parte stupisce come il comportamento della rete neurale sia riuscita quasi ad eguagliare l'albero di decisione(no PCA) lavorando su meno della metà degli attributi, grazie alla analisi delle componenti principali.



ROC