Università degli Studi di Milano Bicocca

Scuola di Scienze

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di laurea in Informatica

PROGETTO D'ESAME

MACHINE LEARNING

Relazione progetto di:

Nicolae Alexandru Andrei
Matricola 829570

Andrea Premate
Matricola 829777

Anno Accademico 2020-2021

Indice

Capitolo 1

Descrizione del dominio e obiettivo

Questo report ha lo scopo di mostrare l'utilizzo di alcune tecniche di machine learning ideate per la classificazione. Il dominio in cui è stato scelto il data set è quello dell'astronomia.
Le "pulsar" sono un raro tipo di stelle di neutroni che producono emissioni radio rilevabili dalla Terra. Sono considerate di interesse scientifico nel campo dello spazio-tempo, degli stati della materia ed in diversi altri ambiti. Quando le pulsar ruotano, il loro raggio di emissione attraversa il cielo e quando questo incrocia la nostra linea di vista produce un pattern di emissioni radio a banda larga che possiamo rilevare. Dato che le pulsar ruotano rapidamente questo pattern si ripete periodicamente, leggermente modificato e leggermente differente per ogni pulsar.
Gli attributi presenti nel data set HTRU2 vanno quindi ad analizzare alcune caratteristiche dei segnali rilevati considerate importanti. In particolare ogni istanza "osservazione" è descritta da 8 variabili continue. Le prime quattro sono semplici statistiche ottenute dal profilo integrato (vedere [1] per maggiori dettagli) e le rimanenti quattro, in modo simile, sono statistiche ottenute dalla curva DM-SNR (vedere [1] per maggiori dettagli). Data la complessità del dominio ci limitiamo di seguito ad elencare solamente una breve descrizione degli attributi, senza una effettiva spiegazione:

1. Mean of the integrated profile.
2. Standard deviation of the integrated profile.
3. Excess kurtosis of the integrated profile.(grado di appiattimento [2])
4. Skewness of the integrated profile.(grado di simmetria [2])

5. Mean of the DM-SNR curve.
6. Standard deviation of the DM-SNR curve.
7. Excess kurtosis of the DM-SNR curve.
8. Skewness of the DM-SNR curve.

L'obiettivo è quello di poter di distinguere, sulla base degli attributi appena elencati, quando si è effettivamente identificata una pulsar oppure quando si è rilevato solo rumore.
In seguito alcuni esempi provenienti dal dataset. Nell'ultima colonna si può distinguere tra 0 (solo rumore) e 1(pulsar rilevata).

	MeanIP	StdIP	EkIP	SkIP	MeanDM.SNR	StdDM.SNR	EkDM.SNR	SkDM.SNR	PULSAR
37	99.492188	41.25511	0.339316810	0.87442001	4.8461538	29.47106	6.4776624	42.36337684	0
38	44.867188	45.69333	2.888739412	8.81067255	176.1195652	59.73772	-1.7853766	2.94091343	1
39	120.812500	45.16182	0.451277319	0.31494367	2.4598662	19.13626	9.0731149	90.11590515	0
40	131.906250	53.77041	0.219247987	-0.18883459	2.6998328	14.73073	8.3369976	97.53171305	0

Capitolo 2

Scelta di design per la creazione del data set

Il data set iniziale era privo di nomi delle colonne quindi, come prima cosa abbiamo dato i seguenti nomi alle colonne:

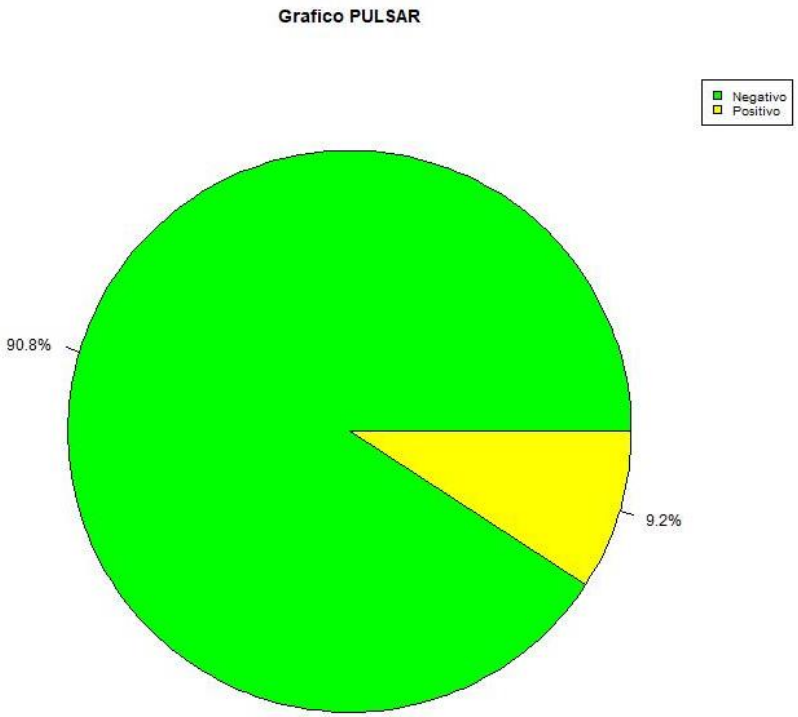
1. MeanIP = Mean of the integrated profile.
2. StdIP = Standard deviation of the integrated profile.
3. EkIP = Excess kurtosis of the integrated profile.
4. SkIP = Skewness of the integrated profile.
5. MeanDM.SNR = Mean of the DM-SNR curve.
6. StdDM.SNR = Standard deviation of the DM-SNR curve.
7. EkDM.SNR = Excess kurtosis of the DM-SNR curve.
8. SkDM.SNR = Skewness of the DM-SNR curve.
9. PULSAR = pulsar detected.

In seguito abbiamo reso "factor" l'attributo rispetto al quale volevamo fare la classificazione: PULSAR, il quale è stato così trasformato dal suo formato originale (numerico) a factor.

Capitolo 3

Descrizione del training set

Il data set contiene un totale di 17 898 istanze di cui 1 639 positive (sì pulsar) e 16 259 negative (no pulsar).



Abbiamo diviso il dataset in train set e test set rispettivamente contenenti il 70% ed il 30% del data set iniziale. In seguito il grafico che mostra le istanze positive e negative successive alla divisione del data set.

Grafico PULSAR Trainset

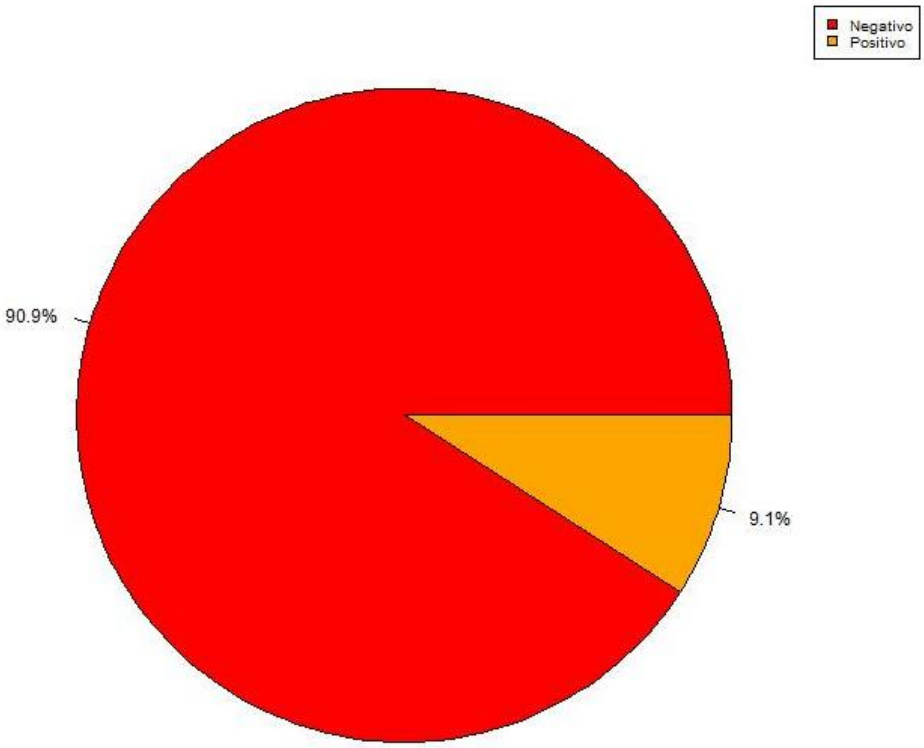
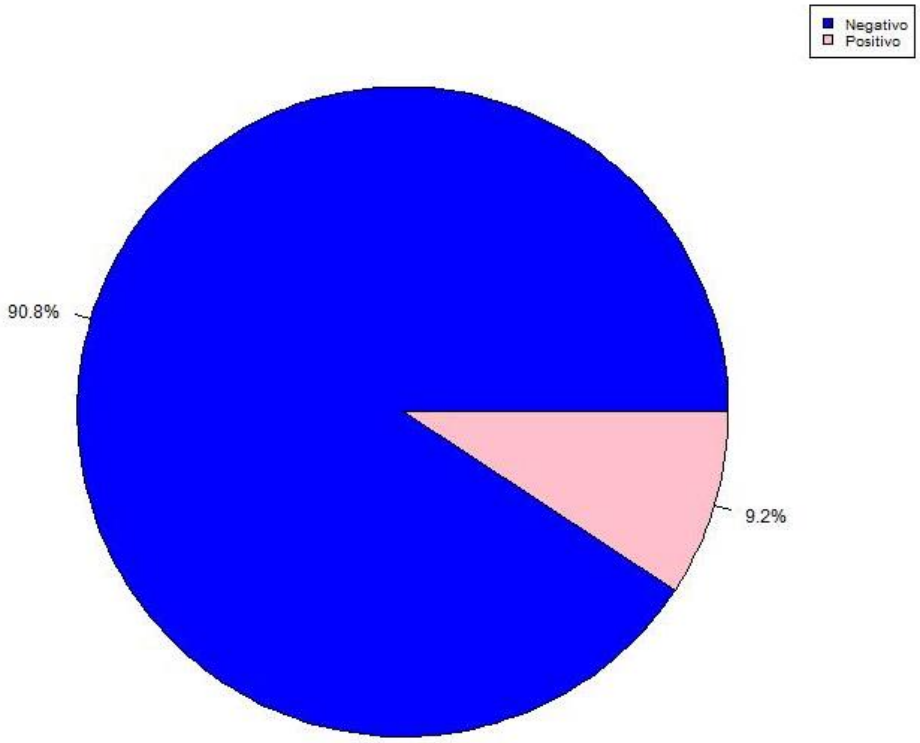
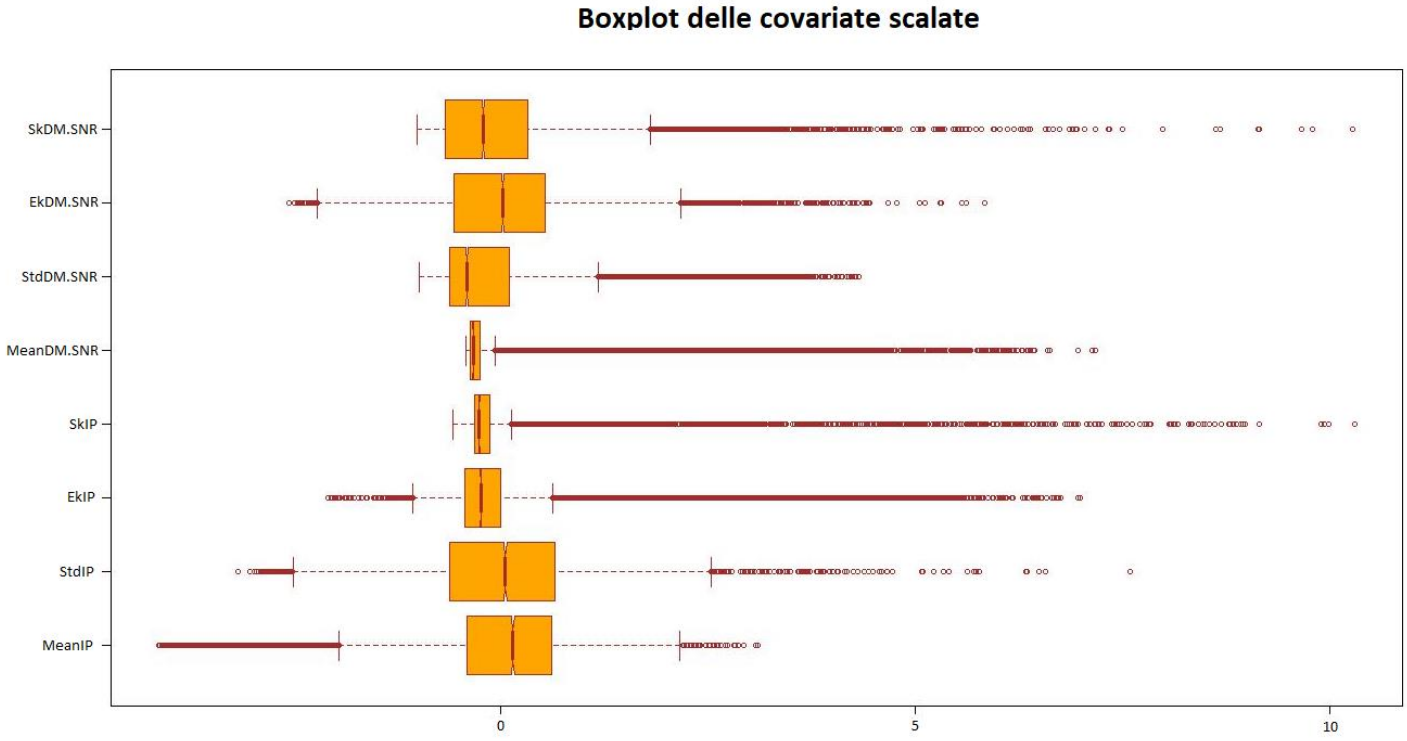


Grafico PULSAR Testset

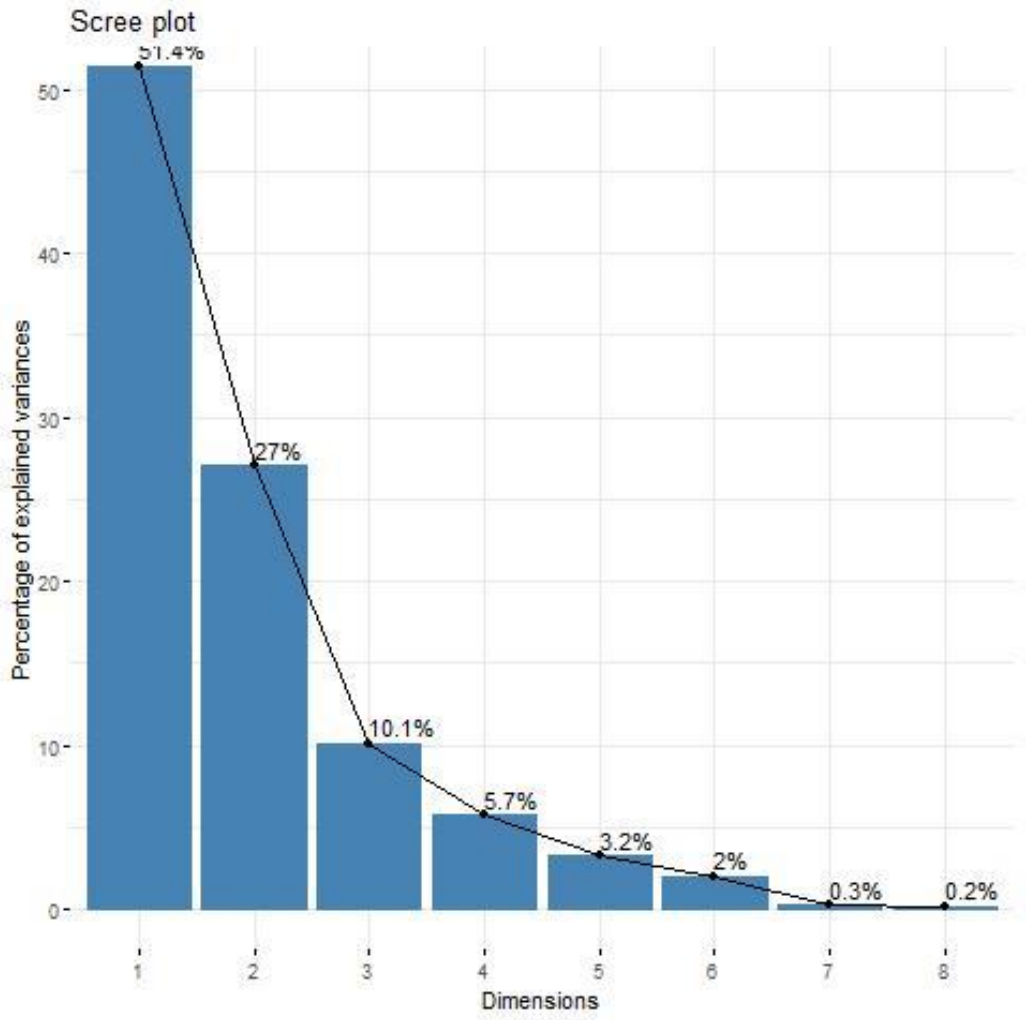


Il data set dunque risulta abbastanza sbilanciato tra istanze positive e negative, così come, di conseguenza, train set e test set. Abbiamo ritenuto che tale sbilanciamento non fosse esageratamente marcato (es. 99% e 1%) e quindi non abbiamo utilizzato tecniche di ricampionamento per renderlo più bilanciato. Inoltre osserviamo che non vi sono valori nulli nel data set, grazie alla funzione anyNA().

In seguito è rappresentato il boxplot delle covariate (del train set) dopo aver applicato la funzione scale(), utile per farsi un'idea di come siano distribuiti i valori degli attributi.

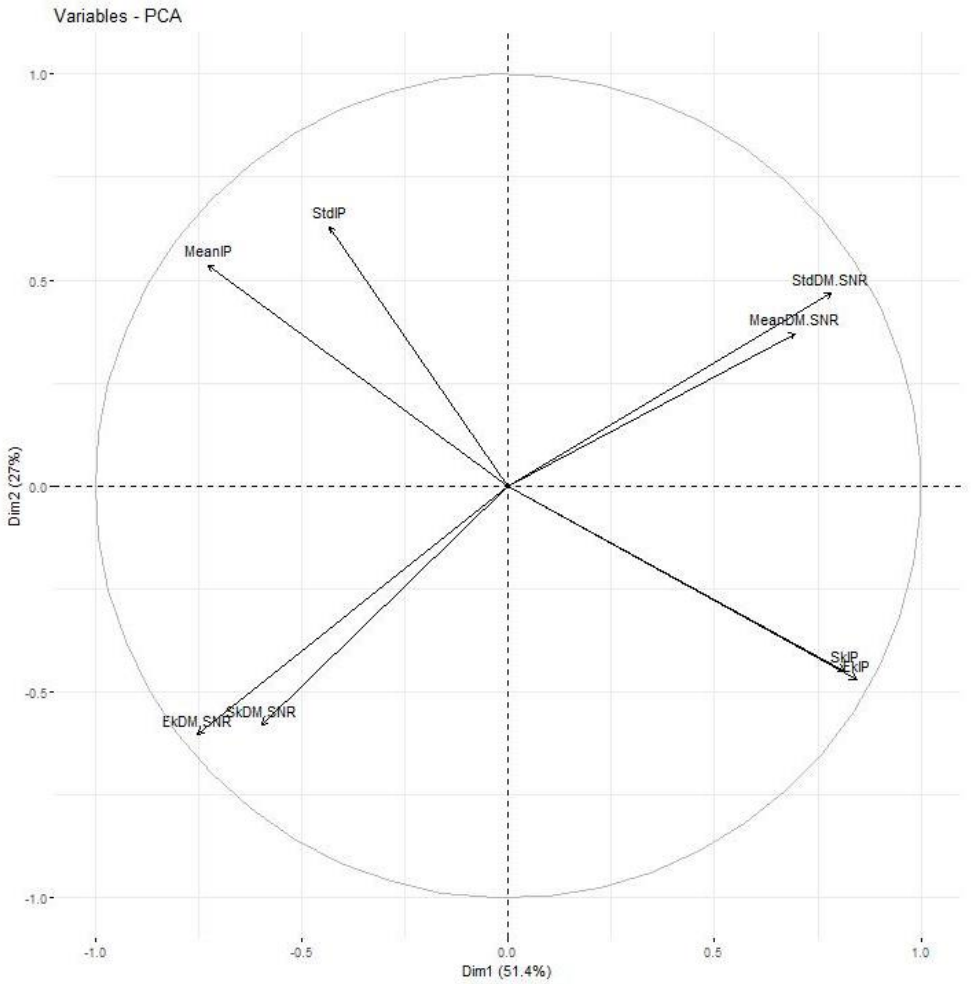


Per ridurre il numero di covariate abbiamo utilizzato la PCA sul train set. Nel grafico sottostante si può visualizzare la varianza percentuale spiegata dalle diverse componenti principali:

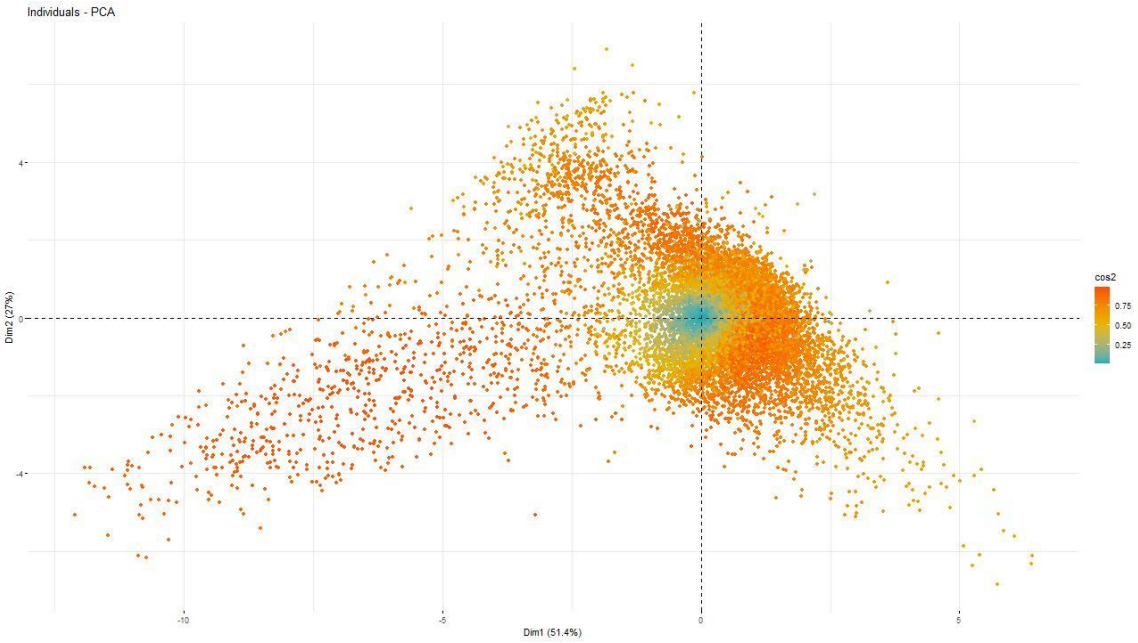


Abbiamo deciso di mantenere tutte le componenti che spiegano almeno il 10% della varianza, trovandoci così a ridurre lo spazio di input da 8 a 3 dimensioni. La percentuale totale di varianza spiegata da queste tre dimensioni è di 88,55154%.

Nel seguente grafico invece si può visualizzare come le diverse covariate originali sono rappresentate tramite l'utilizzo delle prime due componenti principali. Ad una maggiore lunghezza del vettore è associata una maggiore influenza della covariata.



Quest'altro grafico invece rappresenta le varie istanze del train set nello spazio delle due componenti principali. Il colore indica, sfruttando l'indice cos2, la qualità della rappresentazione per ogni istanza.



Capitolo 4

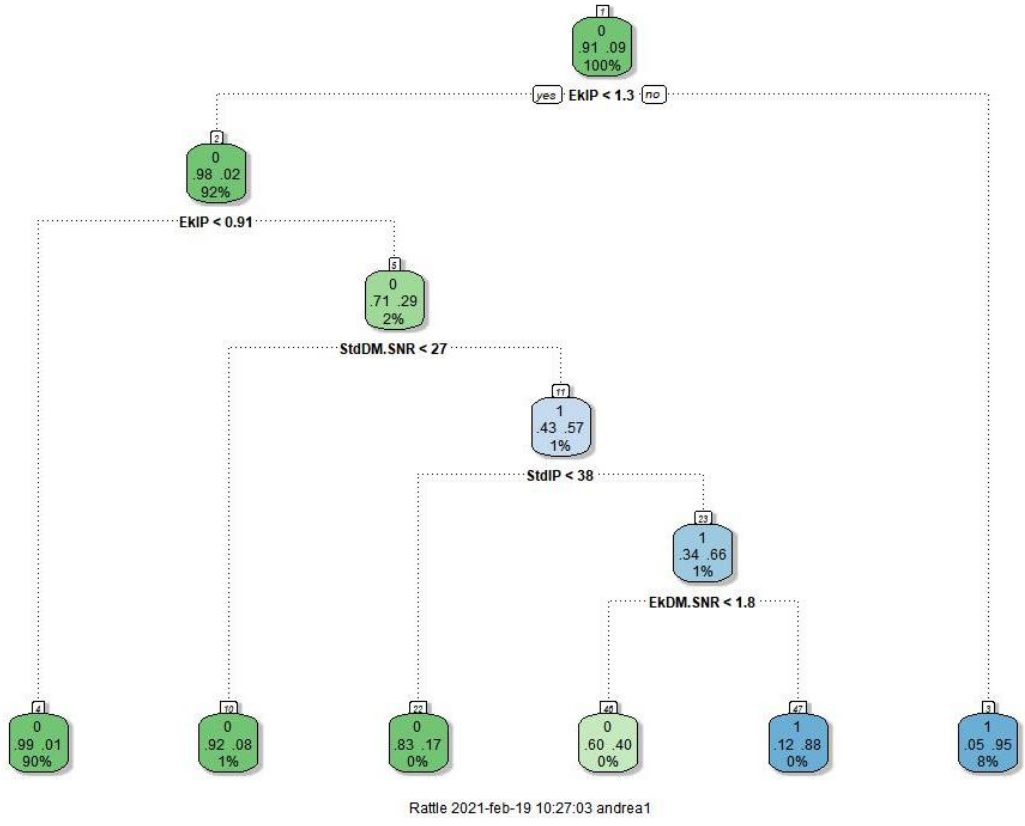
Decision tree

Un albero di decisione è un modello predittivo, dove ogni nodo interno rappresenta una variabile, un arco verso un nodo figlio rappresenta un possibile valore per quella proprietà e una foglia il valore predetto per la variabile obiettivo a partire dai valori delle altre proprietà, che nell'albero è rappresentato dal cammino (path) dal nodo radice (root) al nodo foglia.
Abbiamo scelto di utilizzare questo modello di apprendimento supervisionato perché può essere interpretato molto facilmente: è immediato riuscire a distinguere in base a quali attributi-valori vengono fatte le diverse distinzioni. In particolare, proprio per questo motivo, abbiamo deciso di analizzare l'albero di decisione sia effettuando la PCA che non effettuandola, in modo tale da vedere quali attributi del train set originale facessero maggiormente da spartiacque.

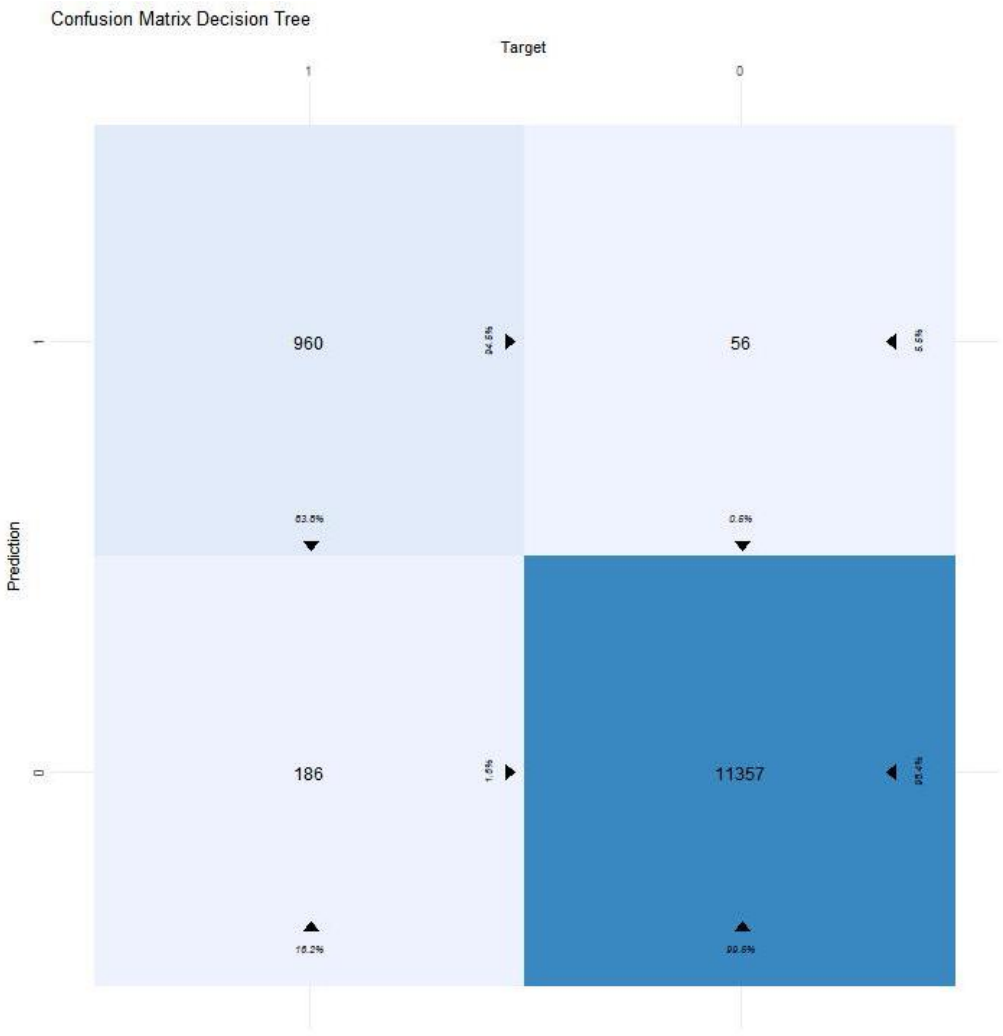
4.1. Versione senza PCA

Durante l'addestramento abbiamo utilizzato una 10-fold cross validation. Il modello che è stato prodotto è il seguente:

Come si può notare l'attributo relativo alla curtosi in eccesso del profilo integrato risulta essere di gran lunga il principale attributo rispetto al quale la classificazione dell'intero albero viene fatta.



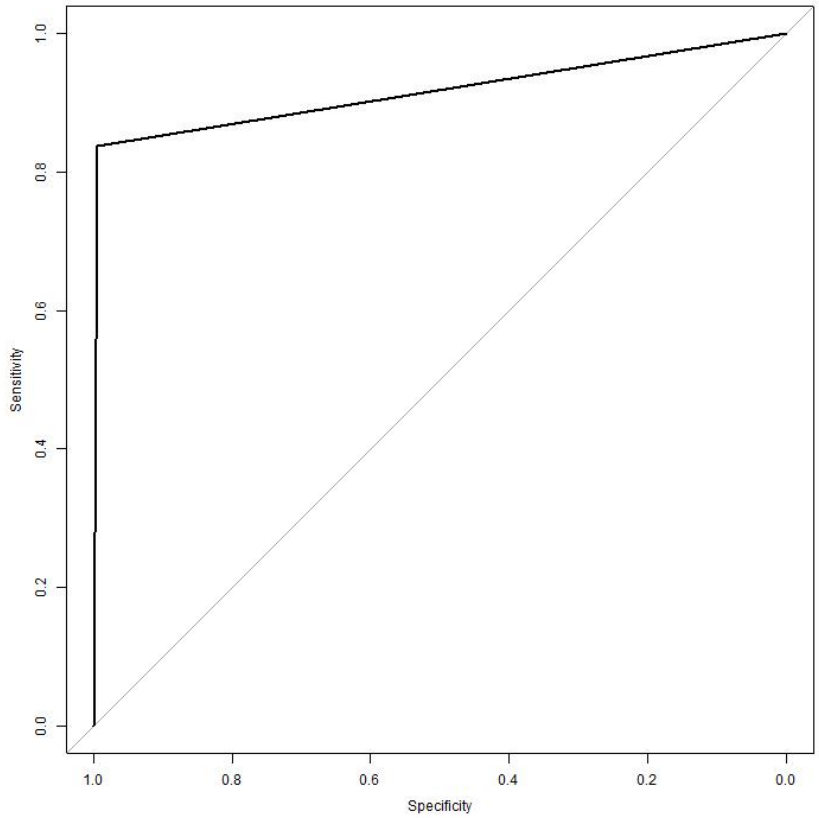
Ecco invece la matrice di confusione:



In base ad essa sono state calcolate le seguenti misure, considerando come valore positivo l'1 (pulsar rilevata):

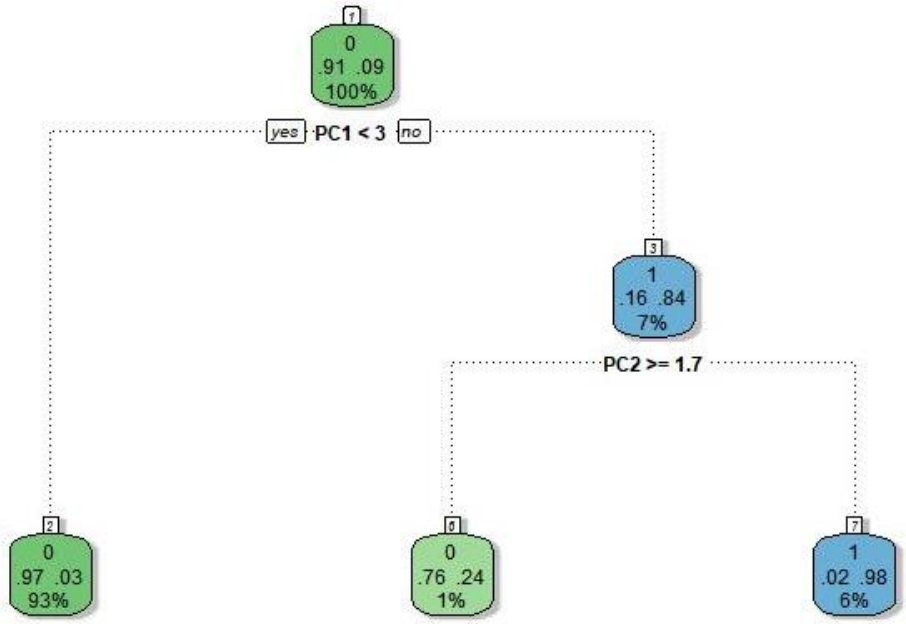
Accuracy (test)	0,979022
Accuracy (10-fold cv)	0,979059
Precision	0,944882
Recall	0,837696
F-measure	0,888067
AUC	0,916395

In seguito la curva ROC dell'albero di decisione



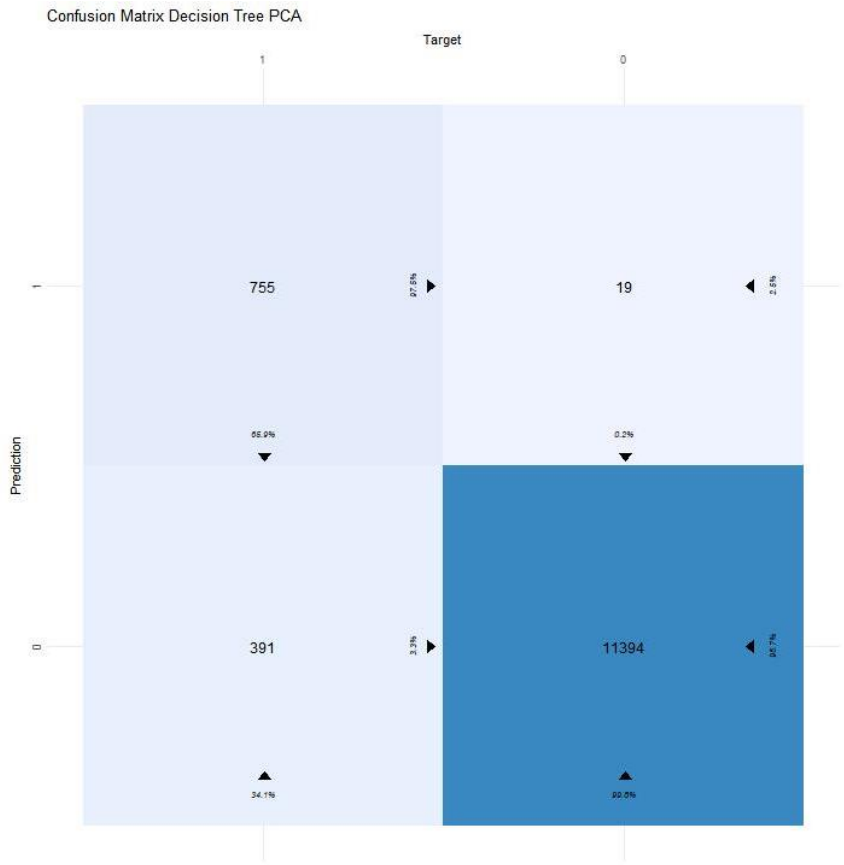
4.2. Versione con PCA

Durante l'addestramento abbiamo utilizzato una 10-fold cross validation. Il modello che è stato prodotto è il seguente:



La prima componente principale è di gran lunga la più discriminante.

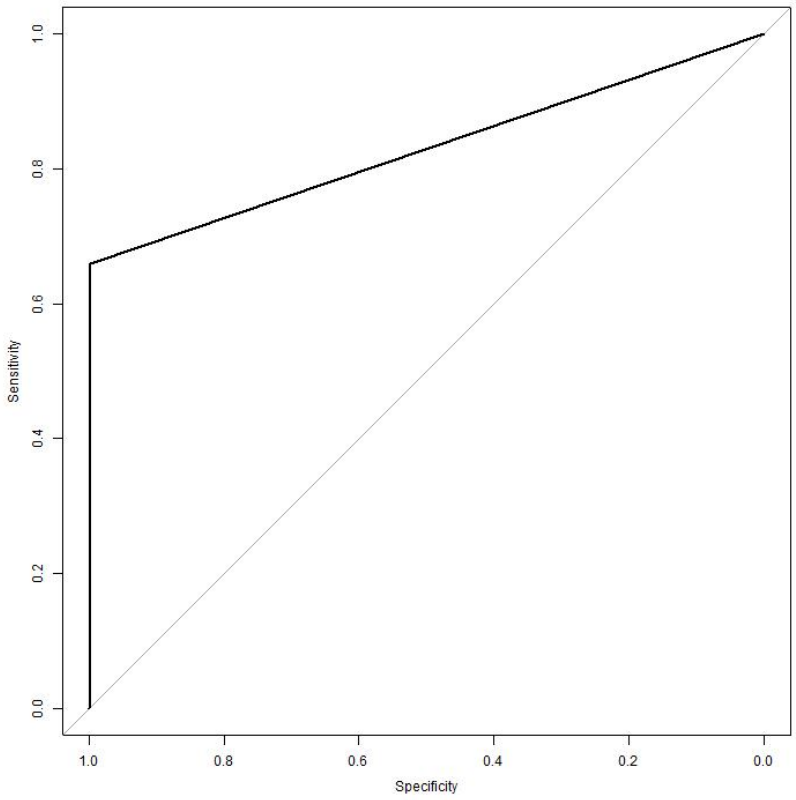
La matrice di confusione risulta invece essere la seguente:
(si noti come in questo caso il valore dei falsi negativi risulti essere rilevante, probabilmente a causa dello sbilanciamento del train set)



In base ad essa sono state calcolate le seguenti misure, considerando come valore positivo l'1 (pulsar rilevata):

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

In seguito la curva ROC dell'albero di decisione

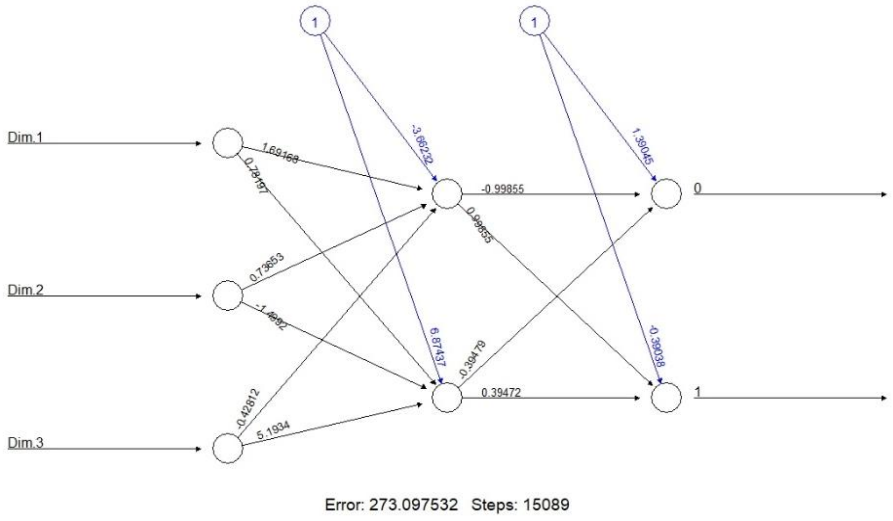


Capitolo 5

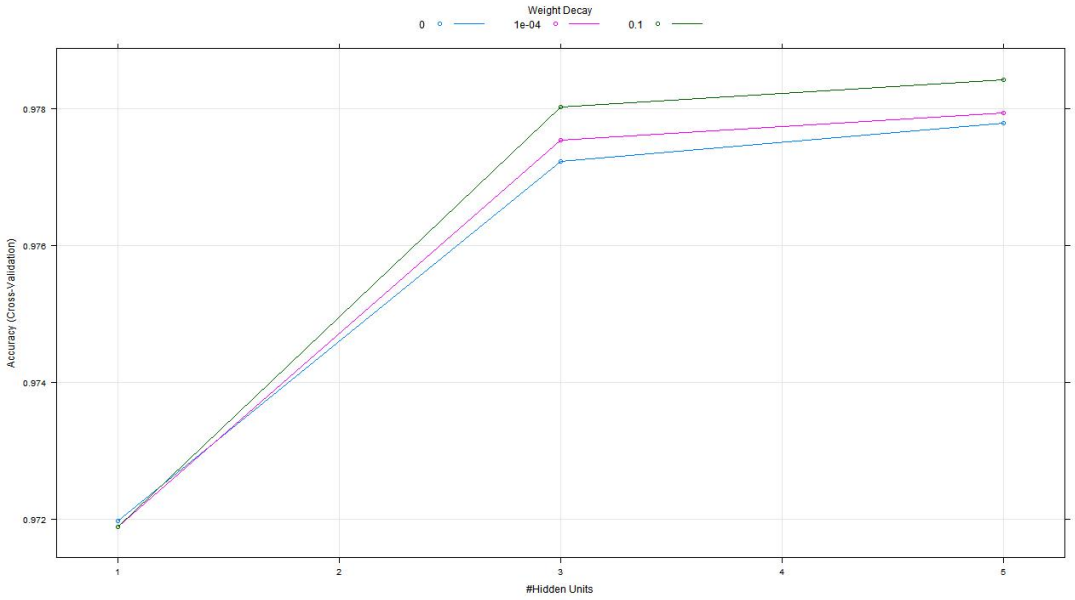
Neural Network

Una rete neurale è un modello di classificazione supervisionata costituita da diversi nodi (neuroni), con il loro peso associato, collegati tra di loro tramite archi pesati. Possiamo distinguere tra neuroni di input, i quali prendono direttamente i dati in base ai quali vogliamo allenare la nostra rete neurale, neuroni di output, dai quali possiamo stabilire quale è la previsione del modello ed infine i neuroni nascosti, utili, se utilizzati correttamente, a migliorare la capacità di previsione della rete neurale.
Abbiamo deciso di utilizzare la rete neurale come secondo modello, sfruttando la PCA, perché volevamo testare se questo modello, più complesso strutturalmente e da allenare, riuscisse ad avere capacità maggiori di classificazione, a dispetto di una minore interpretabilità.

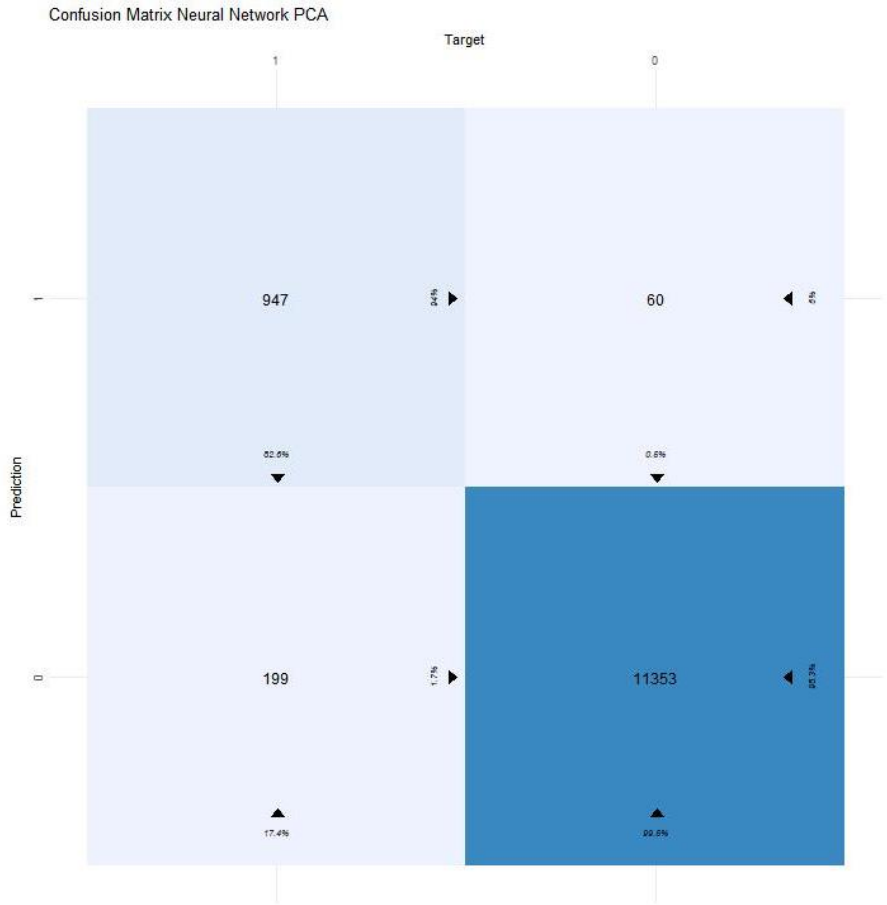
Durante l'addestramento abbiamo utilizzato una 10-fold cross validation. Il modello che è stato prodotto è il seguente:



In seguito il grafico che mostra l'accuracy(Cross-Validation) al variare dei neuroni nascosti.



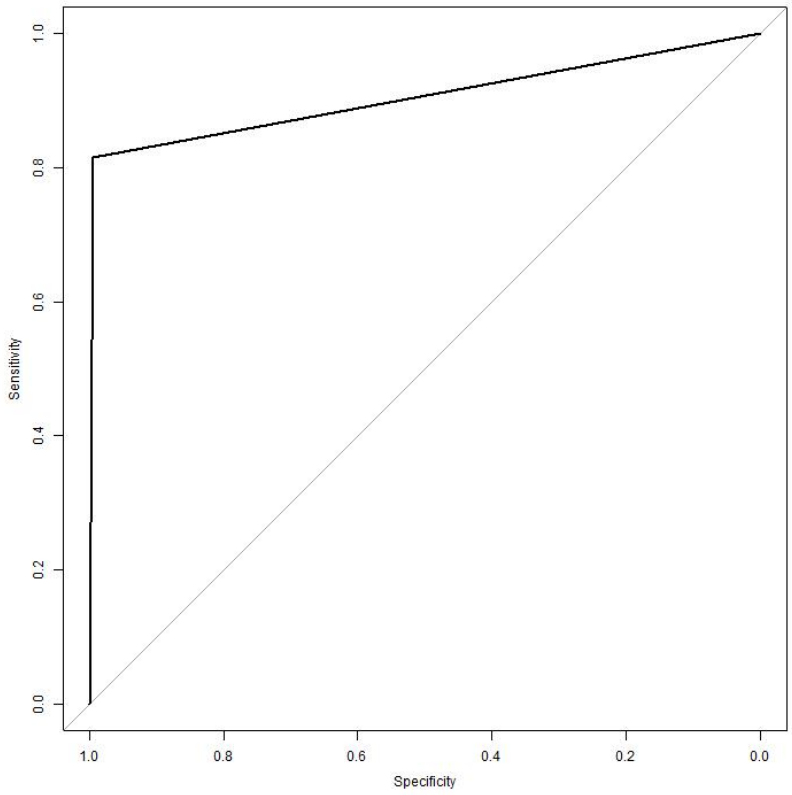
La matrice di confusione risulta invece essere la seguente:



in base ad essa sono state calcolate le seguenti misure, considerando come valore positivo l'1 (pulsar rilevata):

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

In seguito la curva ROC della rete neurale



Capitolo 6

Analisi dei risultati ottenuti

La tabella seguente riepiloga in maniera sintetica i risultati ottenuti:

	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

In primis possiamo affermare che il modello Decision Tree (PCA) risulta essere quello meno performante dei tre. Ciò è conseguenza del fatto che la semplicità del tipo di modello, unito alla riduzione dimensionale, non consente di andare troppo a fondo del processo di apprendimento. Infatti, nonostante la precision risulti più alta rispetto agli altri modelli, si perde troppo in recall: quando il modello dice che è stata rilevata una pulsar la sua probabilità di sbagliare è la minore, ma le pulsar che non vengono rilevate sono troppe.

Andiamo quindi a confrontare il Decision Tree con la Neural Network (PCA): troviamo valori molto simili in tutte le misure, ma quelli del decision tree battono di poco quelli della rete neurale, rendendolo, anche se di pocchissimo, il modello migliore dei tre. Per contro è giusto sottolineare che la rete neurale è riuscita quasi ad eguagliare l'albero di decisione ma lavorando su ben 5 dimensioni in meno.

È stato importante, per la natura del problema, considerare come classe positiva della matrice di confusione l'1(pulsar identificata) perché chiaramente l'oggetto di interesse è il riconoscimento delle pulsar, non il riconoscimento di quando non si ha una pulsar.

Di seguito riportiamo inoltre una tabella riassuntiva con i tempi di addestramento e di predizione da parte dei vari modelli.

I test sono stati eseguiti sulla CPU AMD Ryzen 2600X (6 core, 12 thread, freq. Base 3,6 GHz, freq. Boost 4,2 GHz, chache 16 MB).

	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

Capitolo 7

Conclusioni

In questo progetto abbiamo implementato complessivamente 3 modelli, di cui 2 alberi di decisione, riducendo in un caso lo spazio iniziale tramite la PCA e nell'altro caso utilizzando direttamente le covariate del data set originale, e 1 rete neurale. Il nostro obiettivo era quello di migliorare il più possibile la capacità di riconoscimento di stelle pulsar partendo da un train set contenente statistiche relative ad emissioni radio registrate durante osservazioni astronomiche. Le misure di accuracy, precision, recall, F-measure e AUC ci hanno consentito, grazie ai loro significati, di stabilire che l'albero di decisione (senza PCA), costituisce il modello migliore creato. D'altra parte stupisce come il comportamento della rete neurale sia riuscita pressoché ad eguagliare l'albero di decisione (senza PCA) lavorando su meno della metà degli attributi, grazie alla analisi delle componenti principali.

Bibliografia

[1] D. R. Lorimer and M. Kramer, "Handbook of Pulsar Astronomy", Cambridge University Press, 2005.

[2]https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm#:~:text=Skewness%20is%20a%20measure%20of,relative%20to%20a%20normal%20distribution.

[3]https://archive.ics.uci.edu/ml/datasets/HTRU2