

Università degli Studi di Milano Bicocca

**Scuola di Scienze**

**Dipartimento di Informatica, Sistemistica e Comunicazione**

**Corso di laurea in Informatica**

Machine learning

Progetto d’esame

Link dataset: <https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction>

Riccardo Moschi 856243 r.moschi@campus.unimib.it Luca Poli 852027 l.poli6@campus.unimib.it

## Indice

[1. Descrizione del dominio di riferimento e obiettivi dell’elaborato 3](#_Toc123813135)

[2. Scelte di design per la creazione del data set, eventuali ipotesi o assunzioni 3](#_Toc123813136)

[3. Descrizione del training set: analisi esplorativa del training set (analisi delle covariate e/o PCA) 3](#_Toc123813137)

[4. Descrizione e motivazione dei modelli di machine learning scelti (almeno due modelli) 5](#_Toc123813138)

[Esecuzione di una 10-fold cross validation e stima delle seguenti misure di performance: 6](#_Toc123813139)

[Matrice di confusione complessiva 6](#_Toc123813140)

[Precision, recall, f-measure, ROC e AUC 6](#_Toc123813141)

[Analisi dei risultati ottenuti 6](#_Toc123813142)

[Conclusioni 6](#_Toc123813143)

## Descrizione del dominio di riferimento e obiettivi dell’elaborato

Questo studio fornisce un'analisi delle caratteristiche dei biglietti aerei di cui hanno usufruito alcuni utenti e delle medesime metriche del volo.

La ricerca mirava a stabilire la relazione tra la soddisfazione del cliente e queste variabili di influenza, utilizzando un metodo di correlazione per verificarne l'effetto, col fine di riuscire a predire la soddisfazione dei futuri clienti. Un modello svm è stato sviluppato dividendo il dataset per il training, testing e validation. I risultati mostrano che, sebbene esista una relazione tra il prezzo del biglietto aereo e i costi operativi, è oscurata dal fatto che il principale fattore determinante dei prezzi è spiegato dal rapporto tra domanda e offerta.

## Scelte di design per la creazione del data set, eventuali ipotesi o assunzioni

Si è deciso di analizzare questo dataset visto il dominio applicativo di nostra conoscenza.

Abbiamo effettuato una ristrutturazione del dataset, cambiando attributi stringhe in numerici, ed abbiamo effettuato uno scaling che ci permette di normalizzare il range di variazione delle caratteristiche (feature).

Per motivi progettuali è stato scelto di utilizzare solo un campione di dati nel dataset, visto l’enorme carico computazionale che avrebbe comportato all’elaboratore.

## Descrizione del training set: analisi esplorativa del training set (analisi delle covariate e/o PCA)

Dato il numero elevato di features, è stata implementata una strategia di feature reduction, la PCA,

Abbiamo deciso di scegliere le prime 7 componenti per raggiungere il nostro obiettivo, in quanto volevamo ottenere circa il 74% di varianza cumulata.

Immagine che contiene tavolo

Descrizione generata automaticamente

Figura 1 Parametri ottenuti dalla PCA

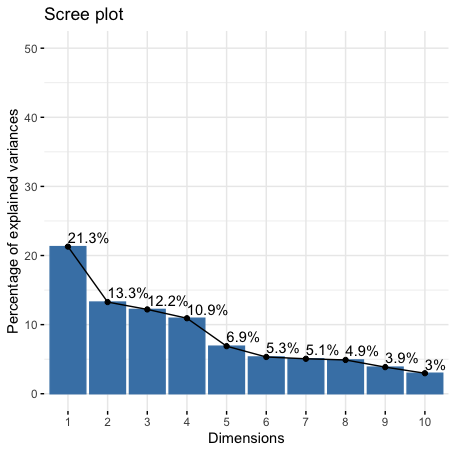


Figura 2 Dimensioni PCA in grafico

Notando il risultato prenderò i primi 7 valori in quanto ottengo circa il 74% di varianza cumulativa e sono soddisfatto con quei dati.

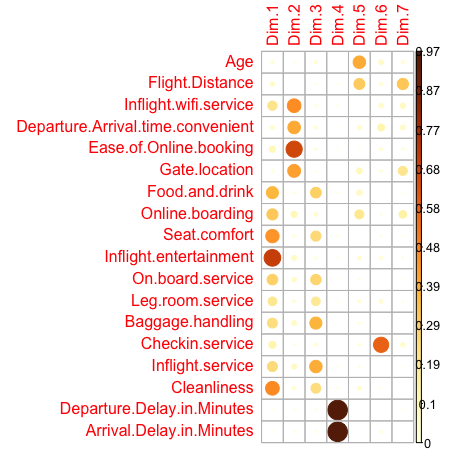


Figura 3 Cos2 di tutte le feature in corrispondenza della dimensione

Grazie a questo grafico è possibile visualizzare il cos2 delle variabili su tutte le dimensioni, in particolare è possibile osservare nella dimensione 1 e 2 gli attributi con valore maggiore, in quanto seguiranno successive analisi.

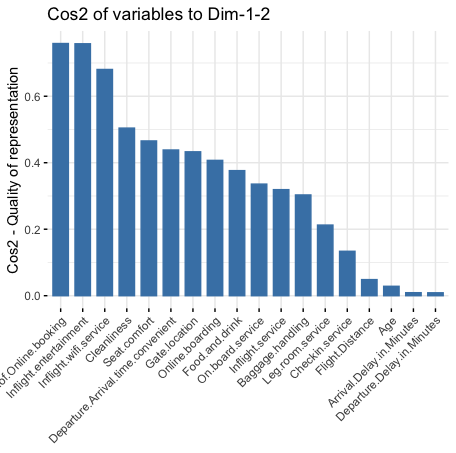


Figura 4 Cos2 delle features rispetto alla dimensione 1-2 della PCA

È anche possibile creare un grafico a barre sul cos2 delle variabili, come è possibile notare “Ease.of.Online.booking” risulta la feature che contribuisce maggiormente con la varianza, rispetto al totale **nella dimensione 1-2**, quindi che si distacca maggiormente dalla media. Infatti, osservando l’immagine sopra, gli attributi in corrispondenza della colonna “Dim.1” e “Dim.2” con più alta varianza (di colore arancione) sono proprio gli stessi che presentano valori più importanti nell’istogramma.

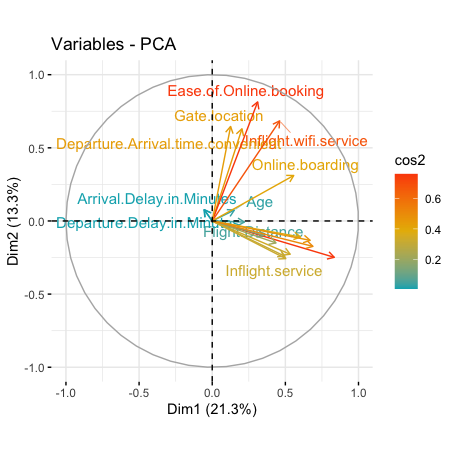


Figura 5 Cos2 rappresentata nel grafo

I medesimi attributi, con varianza più rilevante, è possibile osservarli anche in questo grafo, in particolare “Ease.of.Online.booking”, “Inflight.wifi.service” e “Inflight.entertainment” risultano i più rossastri, ovvero con cos2 maggiore.

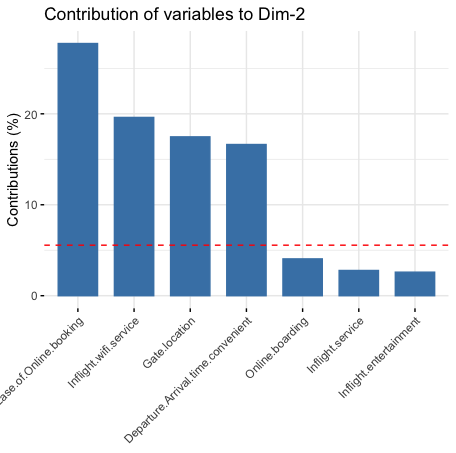
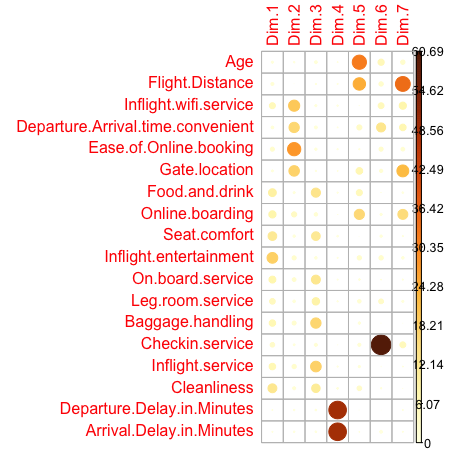


Figura 6 Contributo delle singole variabili nella dimensione 2 della PCA

A partire dal grafo a sinistra si può esaminare il contributo della variabile originale a una determinata componente principale, in questo caso nella dimensione 2, che come è possibile analizzare nell’immagine a destra la feature con contributo maggiore risulta essere “Ease.of.Online.booking”.



Figura 7 Matrice correlazione degli attributi

La **correlation matrix**, ovvero la matrice di correlazione indica se due variabili sono in qualche modo legate tra loro, in particolare è possibile osservare una buona correlazione tra le features “Cleanliness” e “food.and.drink”, “Seat.comfort”, “Inflight.entertainment”, e soprattutto tra “Arrival.Delay.in.Minutes” e “Departure.Delay.in.Minutes” che presentano il massimo legame.

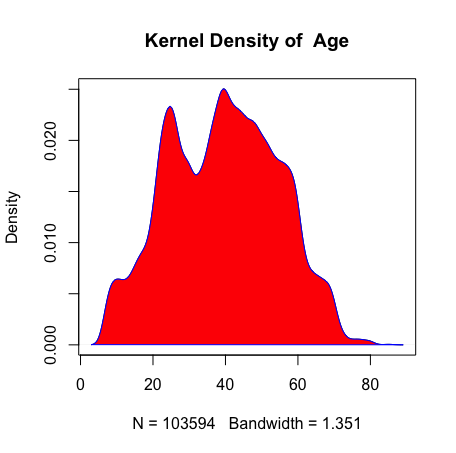


Figura 8 Kernel density dell'età dei passeggeri

È possibile osservare come l’età sia maggiormente addensata al centro, mostrando così una maggiore probabilità sull’età intorno ai 40 anni.

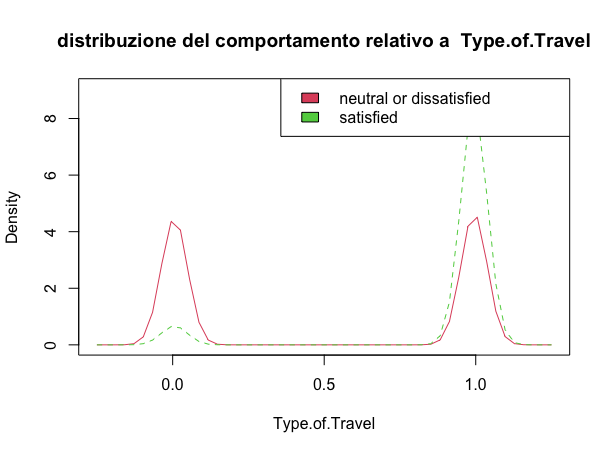


Figura 9 Distribuzione comportamento rispetto al tipo di viaggio di lavoro (se per svago o per lavoro)

E’ possibile osservare la distribuzione del comportamento (soddisfatto o insoddisfatto), in relazione al tipo di viaggio (0= “Personal Travel”,1 =” Business.travel”) e contrariamente a quanto potremmo aspettarci, tendenzialmente i viaggi maggiormente graditi sono quelli per Business.

Immagine che contiene tavolo

Descrizione generata automaticamente

Figura 10 Grafo riassuntivo di gran parte delle variabili in relazione al target

Il grafico presenta il target in relazione agli altri features, quindi la soddisfazione o meno del cliente in base agli altri dati del dataset.

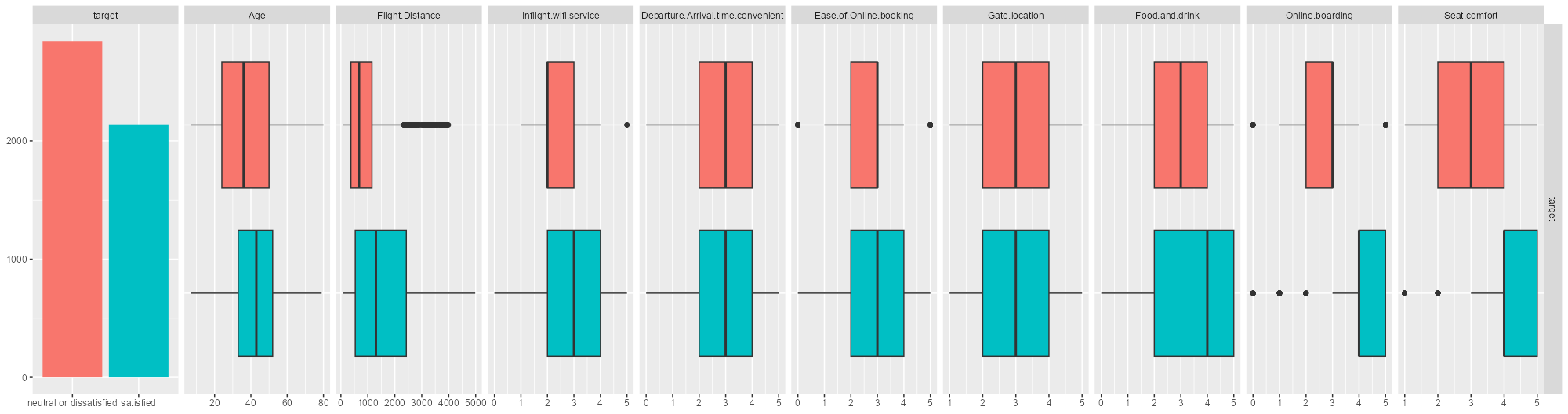


Figura 11 Grafo riassuntivo dei boxplot in relazione al target

Il grafico mostra la distribuzione del target in relazione alle altre features sottoforma di boxplot, quindi la soddisfazione o meno del cliente in base agli altri dati del dataset.

Nei seguenti boxplot è possibile osservare alcuni dati che si discostano significativamente dal resto dei dati (outliers).

Infatti, è possibile osservare come il boxplot dell’attributo “Flight.distance” abbia distribuzione tendente a sinistra (come mostrato anche nella figura sopra sopra) e come ci aspettiamo, presenta diversi outliers in quanto è probabile possa assumere molteplici valori di distanza che si distacchino di molto dalla media.

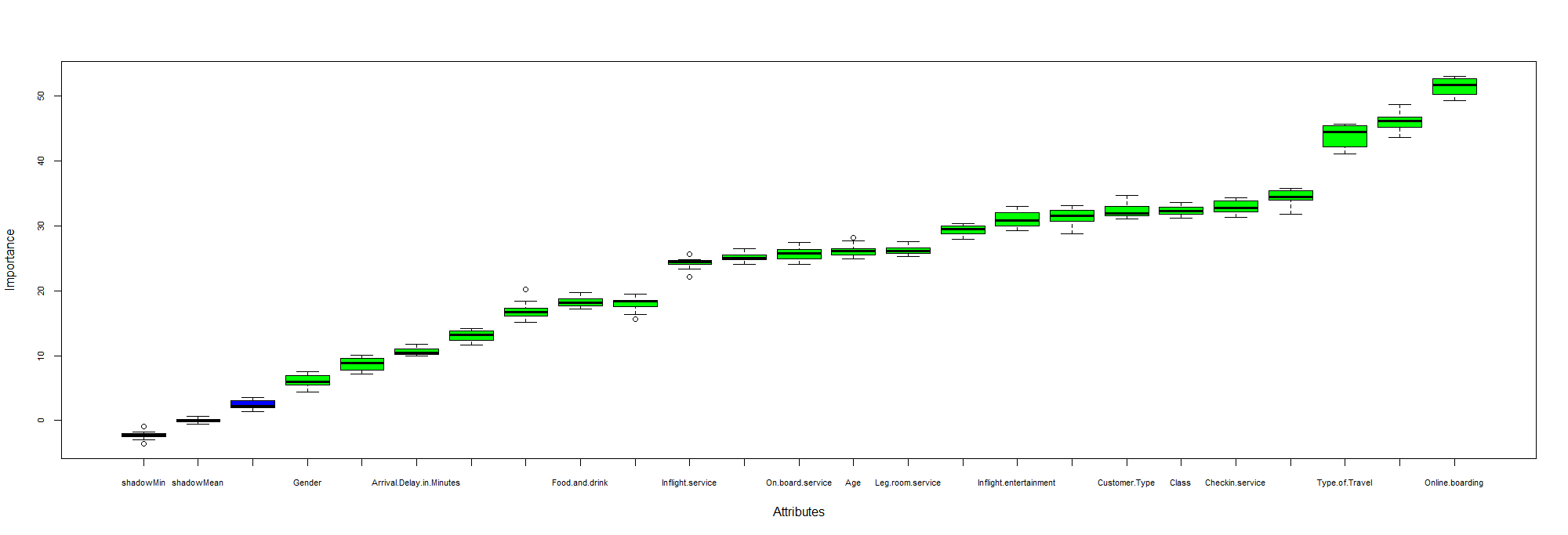


Figura 12 Grafo di raggruppamento dei boxplot

Il seguente grafico consente di confrontare tutti i boxplot per ogni feature, mostrando in particolare i valori anomali in blu e in verde quelli più importanti.

## Descrizione e motivazione dei modelli di machine learning scelti (almeno due modelli)

Si è deciso di utilizzare 2 modelli di Machine Learning: Support Vector Machine e Neural Networks.

Svm è un tipo di modello di classificazione che utilizza i vettori di supporto per costruire una decisione di separazione della superficie tra le diverse classi. La scelta della superficie di separazione viene effettuata in modo da massimizzare la distanza tra le classi, il che significa che il modello cerca di trovare la decisione di separazione che ha il maggior margine possibile. Ciò rende il modello SVM robusto alle variazioni dei dati e generalmente buono per la classificazione. Il modello SVM può anche essere utilizzato per il riconoscimento dei pattern e la regressione

.

Immagine che contiene testo

Descrizione generata automaticamente

## Esecuzione di una 10-fold cross validation e stima delle seguenti misure di performance:

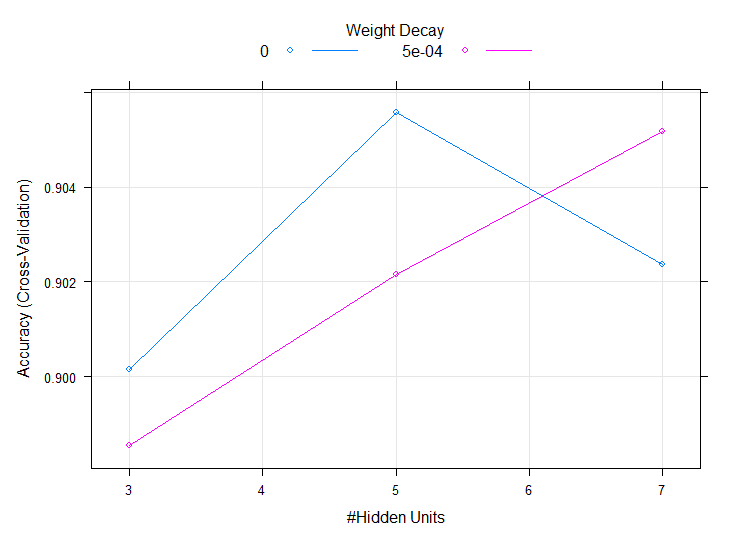


Figura 13 Grafo di Cross-validation di Neural Networks

cross-validation di neural networks (grid search)-> FA RICERCA PARAMETRI

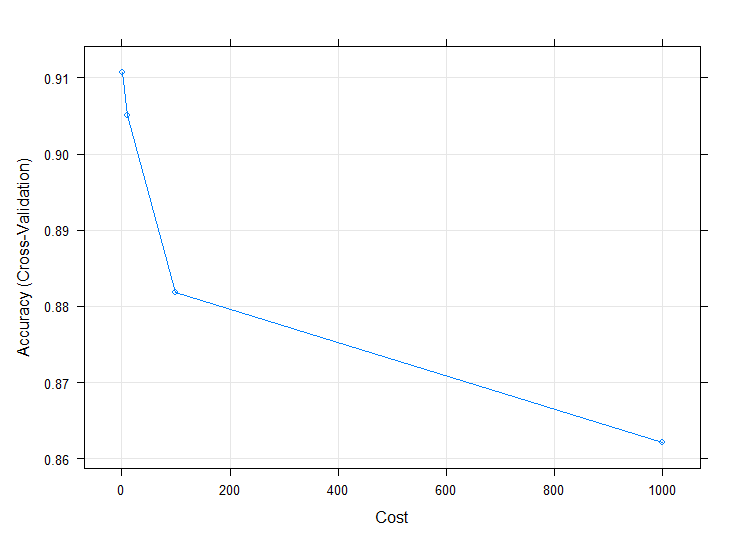


Figura 14 Grafo di Cross-validation di SVM

Cross validation svm che varia sul costo (da 1 a mille)

Immagine che contiene testo, ricevuta, screenshot

Descrizione generata automaticamente

Figura 15 Confusion Matrix dopo la k-fold validation sul dataset di test

## Matrice di confusione complessiva

Immagine che contiene tavolo

Descrizione generata automaticamente

Figura 16 Matrice di confusione di SVM

Immagine che contiene testo

Descrizione generata automaticamente

Figura 17 Matrice di confusione di Neural Networks

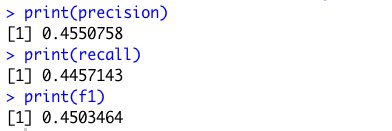
neural networks test

## Precision, recall, f-measure, ROC e AUC

Figura 18 Grafo che mostra la performance di SVM

La curva AUC è uno strumento utile per valutare il rendimento di un modello di classificazione binaria perché è meno sensibile alle variazioni della soglia di classificazione rispetto ad altre metriche come la precisione o il recall. Ciò significa che la curva AUC fornisce una misura più stabile delle prestazioni del modello indipendentemente dal livello di confidenza desiderato.

In questo caso la curva AUC pari a 0.92 indica una buona capacità di classificazione del modello.



## performance di svm

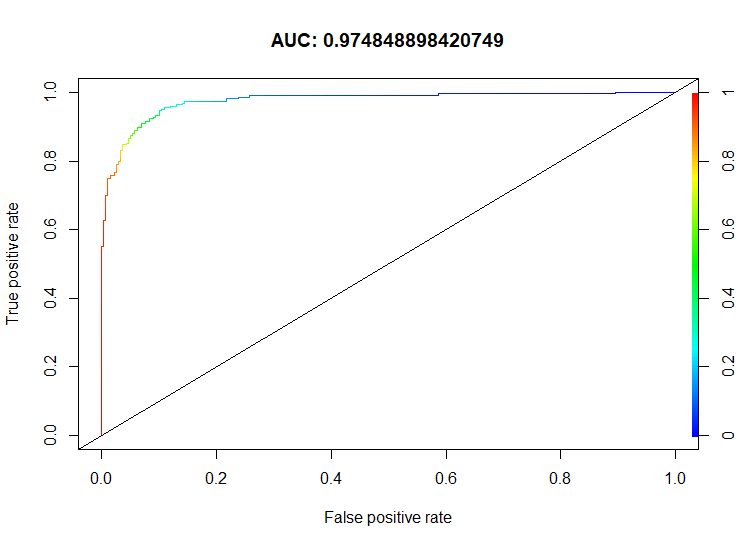


Figura 20 Grafo che mostra la performance di Neural Networks

## Analisi dei risultati ottenuti

## Conclusioni

## Introduzione

## Studio del dominio di riferimento

Questo studio fornisce un'analisi delle caratteristiche dei biglietti aerei di cui hanno usufruito alcuni utenti e delle medesime metriche del volo.

La ricerca mirava a stabilire la relazione tra la soddisfazione del cliente e queste variabili di influenza, utilizzando un metodo di correlazione per verificarne l'effetto, col fine di riuscire a predire la soddisfazione dei futuri clienti. Un modello svm è stato sviluppato dividendo il dataset per il training, testing e validation. I risultati mostrano che, sebbene esista una relazione tra il prezzo del biglietto aereo e i costi operativi, è oscurata dal fatto che il principale fattore determinante dei prezzi è spiegato dal rapporto tra domanda e offerta.

## Dataset

* 1. **Analisi esplorativa**

## Preprocessing

## PCA

## Modelli

## Introduzione (perché uso i modelli ) -> opzionale

## Reti neurale

## SVM

## Risultati

## Confusion matrix

## Roc Auc

## Conclusioni ->opzionale