



# Airline Passenger Satisfaction

What factors lead to customer satisfaction for an Airline?

ANDRONI STEFANO

Mat. 845811

s.androni1@campus.unimib.it

# Indice



## 1 IL DATASET pag.1

*Breve descrizione che fornisce le informazioni principali sul dataset tra cui il contesto, la sorgente e la specifica degli attributi.*

## 2 OBIETTIVO DEL PROBLEMA E STRATEGIA pag.3

*Definizione dell'obiettivo del problema e riepilogo della strategia utilizzata per affrontarlo. Specifica del linguaggio e delle librerie utilizzate.*

## 3 DATA ACQUISITION e DATA PREPROCESSING pag.4

*Analisi dei principali step (anche con riferimento al codice) delle fasi di 'Acquisizione del dato' e 'Pre-processamento dei dati'.*

## 4 EXPLORATORY ANALYSIS pag.9

*Analisi esplorativa del dataset finalizzata al raggiungimento dell'obiettivo prefissato.*

**A - CARATTERISTICHE DEI PASSEGGERI** .....pag.9

**B - SODDISFAZIONE COMPLESSIVA DEI PASSEGGERI** .....pag.10

**C - VALUTAZIONE DEI SERVIZI** .....pag.14

## 5 MODELLI PREDITTIVI (ML) pag.17

*Spiegazione step -by-step del raggiungimento degli obiettivi con l'applicazione di algoritmi di ML tenendo presente i problemi e le scelte affrontate nella loro applicazione.*

**D - DECISION TREE (1)**..... pag.17

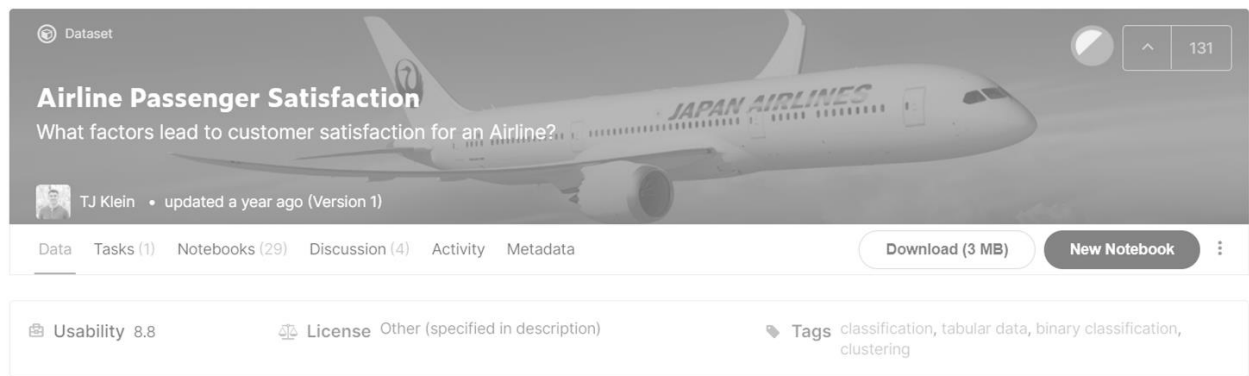
**E – K-NN (2)**..... pag.20

**F – CONFRONTO TRA I MODELLI**..... pag.21

## 6 CONCLUSIONE pag.22

*Breve sintesi dei risultati raggiunti con l'analisi e l'applicazione di algoritmi di ML*

# 1 IL DATASET



**source -** [kaggle](https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction) <https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction>

**textual-description** - il dataset 'Airline Passenger Satisfaction' contiene i dati ricavati da un sondaggio proposto ai passeggeri di una compagnia aerea americana volto ad analizzare la loro soddisfazione differenziata nei servizi offerti. L'autore ha preferito non specificare il nome della compagnia.

**numerical-description** - il dataset conta 129 880 istanze e 25 attributi.

**features** – gli attributi del dataset sono così definiti dall'autore:

## Gender

Sesso del passeggero {Male, Female}

## Customer Type

Tipologia del cliente {Loyal customer, Disloyal customer}

## Age

L'età del passeggero al momento in cui gli è stato sottoposto il sondaggio int>0

## Type of Travel

Scopo del volo del passeggero {Personal Travel, Business Travel}

## Class

La classe di viaggio del volo del passeggero {Business, Eco, Eco Plus}

## Flight Distance

La distanza di volo del viaggio int>0

## Inflight wifi service

Livello di soddisfazione del servizio wifi in volo 0:Not Applicable; 1-5

## Departure/Arrival time convenient

Livello di soddisfazione dell'orario di partenza / arrivo 0:Not Applicable; 1-5

## Ease of Online booking

Livello di soddisfazione della prenotazione online 0:Not Applicable; 1-5

## Gate Location

Livello di soddisfazione dell'ubicazione del gate 0:Not Applicable; 1-5

**Food & Drink**

Livello di soddisfazione di cibo e bevande 0:Not Applicable; 1-5

**Online Boarding**

Livello di soddisfazione dell'imbarco online 0:Not Applicable; 1-5

**Seat comfort**

Livello di soddisfazione del comfort dei sedili 0:Not Applicable; 1-5

**Inflight entertainment**

Livello di soddisfazione dell'intrattenimento in volo 0:Not Applicable; 1-5

**On-board service**

Livello di soddisfazione del servizio a bordo 0:Not Applicable; 1-5

**Leg room service**

Livello di soddisfazione del servizio di spazio aggiuntivo per le gambe 0:Not Applicable; 1-5

**Baggage handling**

Livello di soddisfazione della movimentazione dei bagagli 0:Not Applicable; 1-5

**Check-in service**

Livello di soddisfazione del servizio di check-in 0:Not Applicable; 1-5

**Inflight service**

Livello di soddisfazione del servizio in volo 0:Not Applicable; 1-5

**Cleanliness**

Livello di soddisfazione della pulizia 0:Not Applicable; 1-5

**Departure Delay in Minutes**

Minuti di ritardo rispetto all'orario di partenza schedulato  $int \geq 0$

**Arrival Delay in Minutes**

Minuti di ritardo rispetto all'orario di arrivo schedulato  $int \geq 0$

**Satisfaction**

Livello di soddisfazione complessivo della compagnia aerea {Satisfaction, neutral or dissatisfaction}

\*i VINCOLI ,  e  sono stati dedotti dal contesto reale o espressi esplicitamente dall'autore del dataset (utili per la verifica della coerenza dei dati)

## 2 OBIETTIVO DEL PROBLEMA E STRATEGIA

### OBIETTIVO

La compagnia aerea richiede di poter **predire la soddisfazione** dei suoi passeggeri con l'obiettivo particolare di conoscere quali sono i **fattori che più influiscono su essa**. Il cliente inoltre intende avere un quadro generale sulla **valutazione** che i passeggeri hanno dato **all'esperienza di volo** diversificata nei vari servizi offerti. Obiettivo non secondario quindi è quello di andare a identificare i servizi che hanno una valutazione complessiva peggiore.

### STRATEGIA


Il problema è stato affrontato con la seguente strategia (impostata per steps)\*<sup>1</sup>:

- 1) **Analisi delle caratteristiche principale dei passeggeri** che hanno partecipato al sondaggio in quanto utile a trarre conclusioni più precise negli step successivi [A - CARATTERISTICHE DEI PASSEGGERI]
- 2) **Studio della soddisfazione dei passeggeri** (attributo *\$Satisfaction*) in relazione con gli attributi che esprimono le loro caratteristiche e quelle del volo di cui hanno usufruito [B - SODDISFAZIONE COMPLESSIVA DEI PASSEGGERI]
- 3) **Analisi delle valutazioni dei servizi offerti** nell'esperienza di volo ai passeggeri [C - VALUTAZIONE DEI SERVIZI]
- 4) Impostazione di **modelli per la predizione della soddisfazione** dei passeggeri e valutazione di questi\*<sup>2</sup>  
[D - DECISION TREE] [E - K-NN]

\*<sup>1</sup> – non viene citata nella strategia la fase di pre-processamento dei dati

\*<sup>2</sup> – i modelli scelti per la classificazione sono stati impostati con due tecniche diverse: gli 'alberi decisionali' e i 'K-nearest neighbors'

### LINGUAGGIO E LIBRERIE

Il linguaggio utilizzato per lo studio del dataset e per la risoluzione del problema è  con l'ausilio delle seguenti librerie divise per il tipo di utilizzo:

*Librerie utilizzate per il report*

naniar, ggplot2, tidyverse, plyr, dplyr, GGally

*Librerie utilizzate per l'analisi, ma con risultati (grafici e non) non inseriti nel report*

psych, Amelia, car, visdat, lattice, ggpubr, corrplot, RColorBrewer

*Librerie utilizzate per la parte di ML(1) - Classificazione - DecisionTree*

#ML tree, ctree, rpart #View rattle, rpart.plot, RColorBrewer #ConfusionMatrix e1071, caret #ROC ROC

*Librerie utilizzate per la parte di ML(2) - Classificazione - K-NN*

caret, Class, mlbench, pROC

## 3 DATA ACQUISITION e DATA PREPROCESSING

### PRIMA PARTE

#### Data acquisition

Il dataset viene fornito già diviso in due file entrambi in formato `csv`: `'train.csv'` e `'test.csv'`. Per la prima parte di analisi ed esplorazione del dataset risulta utile unirli utilizzando la funzione `rbind`, che unisce due dataframe per righe.

#### Primo sguardo dei dati

In partenza risulta pratico estrarre alcune istanze del dataset per dare un primo sguardo ai nostri dati: tra le varie opzioni abbiamo `head` #prime 10 righe `tail` #ultime 10 righe `some` #10 righe random. In nostro aiuto in questa fase ci viene incontro la funzione `str`.

```
> str(df)
'data.frame':    129880 obs. of  25 variables:
 $ X              : int  0 1 2 3 4 5 6 7 8 9 ...
 $ id             : int 70172 5047 110028 24026 119299 111157 82113 96462 79485 65725 ...
 $ Gender         : Factor w/ 2 levels "Female","Male": 2 2 1 1 2 1 2 1 1 2 ...
 $ Customer.Type  : Factor w/ 2 levels "disloyal Customer",...: 2 1 2 2 2 2 2 2 2 1 ...
 $ Age           : int 13 25 26 25 61 26 47 52 41 20 ...
 $ Type.of.Travel : Factor w/ 2 levels "Business travel",...: 2 1 1 1 1 2 2 1 1 1 ...
 $ Class         : Factor w/ 3 levels "Business","Eco",...: 3 1 1 1 1 2 2 1 1 2 ...
 $ Flight.Distance : int 460 235 1142 562 214 1180 1276 2035 853 1061 ...
 $ Inflight.wifi.service : int 3 3 2 2 3 3 2 4 1 3 ...
 $ Departure.Arrival.time.convenient : int 4 2 2 5 3 4 4 3 2 3 ...
 $ Ease.of.Online.booking : int 3 3 2 5 3 2 2 4 2 3 ...
 $ Gate.location : int 1 3 2 5 3 1 3 4 2 4 ...
 $ Food.and.drink : int 5 1 5 2 4 1 2 5 4 2 ...
 $ Online.boarding : int 3 3 5 2 5 2 2 5 3 3 ...
 $ Seat.comfort : int 5 1 5 2 5 1 2 5 3 3 ...
 $ Inflight.entertainment : int 5 1 5 2 3 1 2 5 1 2 ...
 $ On.board.service : int 4 1 4 2 3 3 3 5 1 2 ...
 $ Leg.room.service : int 3 5 3 5 4 4 3 5 2 3 ...
 $ Baggage.handling : int 4 3 4 3 4 4 4 5 1 4 ...
 $ Checkin.service : int 4 1 4 1 3 4 3 4 4 4 ...
 $ Inflight.service : int 5 4 4 4 3 4 5 5 1 3 ...
 $ Cleanliness : int 5 1 5 2 3 1 2 4 2 2 ...
 $ Departure.Delay.in.Minutes : int 25 10 11 0 9 4 0 0 ...
 $ Arrival.Delay.in.Minutes : num 18 6 0 9 0 0 23 0 0 0 ...
 $ satisfaction : Factor w/ 2 levels "neutral or dissatisfied",...: 1 1 2 1 2 1 1 2 1 1 ...
```

#### NOTA

- > Il dataset ha **129 880 istanze** e **25 attributi**
- > Gli **attributi 'X' e 'id'** sono **inutili** ai fini di analisi ('id' potrebbe essere utile per verificare la sua unicità nel rispetto della consistenza dei dati)
- > Il **nome degli attributi** potrebbe essere più esplicativo (gli attributi che valutano la soddisfazione da 1 a 5 potrei identificarli facendone iniziare i nomi per 'Sat.\*')
- > Gli attributi che valutano la soddisfazione nei vari campi è utile definirli come **fattori**
- > Potrebbe essere utili aggiungere un **nuovo attributo** che contenga il 'punteggio totale di soddisfazione' per ogni passeggero

### Primo pre-processing dei dati

Possiamo quindi già intervenire sul dataset con:

- la rimozione della prima colonna 'X'
- la ridenominazione di tutte le colonne
- rendendo fattori gli attributi che calcolano la soddisfazione nei vari campi
- aggiungendo una nuova colonna utile ai fini di analisi (in realtà ridondante) →  $\$TotalScore$ : la somma degli attributi che calcolano la soddisfazione degli specifici servizi

Il risultato di questo intervento si può notare da:

```
> str(df) #chek modifiche
'data.frame': 129880 obs. of 25 variables:
 $ Id      : int  70172 5047 110028 24026 119299 111157 82113 96462 79485 65725 ...
 $ Gender  : Factor w/ 2 levels "Female","Male": 2 2 1 1 2 1 2 1 1 2 ...
 $ CustomerType : Factor w/ 2 levels "disloyal Customer",...: 2 1 2 2 2 2 2 2 1 ...
 $ Age     : int  13 25 26 25 61 26 47 52 41 20 ...
 $ TypeOfTravel : Factor w/ 2 levels "Business travel",...: 2 1 1 1 1 2 2 1 1 1 ...
 $ Class    : Factor w/ 3 levels "Business","Eco",...: 3 1 1 1 1 2 2 1 1 2 ...
 $ FlightDistance : int  460 235 1142 562 214 1180 1276 2035 853 1061 ...
 $ Sat.InflightWifiService : Factor w/ 6 levels "0","1","2","3",...: 4 4 3 3 4 4 3 5 2 4 ...
 $ Sat.FlightSchedule : Factor w/ 6 levels "0","1","2","3",...: 5 3 3 6 4 5 5 4 3 4 ...
 $ Sat.EaseOnlineBooking : Factor w/ 6 levels "0","1","2","3",...: 4 4 3 6 4 3 3 5 3 4 ...
 $ Sat.GateLocation : Factor w/ 6 levels "0","1","2","3",...: 2 4 3 6 4 2 4 5 3 5 ...
 $ Sat.FoodDrink : Factor w/ 6 levels "0","1","2","3",...: 6 2 6 3 5 2 3 6 5 3 ...
 $ Sat.OnlineBoarding : Factor w/ 6 levels "0","1","2","3",...: 4 4 6 3 6 3 3 6 4 4 ...
 $ Sat.SeatComfort : Factor w/ 6 levels "0","1","2","3",...: 6 2 6 3 6 2 3 6 4 4 ...
 $ Sat.InflightEntertainment : Factor w/ 6 levels "0","1","2","3",...: 6 2 6 3 4 2 3 6 2 3 ...
 $ Sat.OnBoardService : Factor w/ 6 levels "0","1","2","3",...: 5 2 5 3 4 4 4 6 2 3 ...
 $ Sat.Legroom : Factor w/ 6 levels "0","1","2","3",...: 4 6 4 6 5 5 4 6 3 4 ...
 $ Sat.BaggageHandling : Factor w/ 5 levels "1","2","3","4",...: 4 3 4 3 4 4 4 5 1 4 ...
 $ Sat.CheckinService : Factor w/ 6 levels "0","1","2","3",...: 5 2 5 2 4 5 4 5 5 5 ...
 $ Sat.InflightService : Factor w/ 6 levels "0","1","2","3",...: 6 5 5 5 4 5 6 6 2 4 ...
 $ Sat.Cleanliness : Factor w/ 6 levels "0","1","2","3",...: 6 2 6 3 4 2 3 5 3 3 ...
 $ DepartureDelayMinutes : int  25 10 11 0 9 4 0 0 ...
 $ ArrivalDelayMinutes : num  18 6 0 9 0 0 23 0 0 ...
 $ Satisfaction : Factor w/ 2 levels "neutral or dissatisfied",...: 1 1 2 1 2 1 1 2 1 1 ...
 $ TotalScore : int  67 45 65 55 62 48 52 76 42 54 ...
```

## MISSING VALUES 1

```
> apply(df, function(x) sum(is.na(x)==TRUE))
Id Gender CustomerType Age TypeOfTravel Class FlightDistance Sat.InflightWifiService Sat.FlightSchedule Sat.EaseOnlineBooking
0 0 0 0 0 0 0 0 0 0
Sat.GateLocation Sat.FoodDrink Sat.OnlineBoarding Sat.SeatComfort Sat.InflightEntertainment Sat.OnBoardService Sat.Legroom Sat.BaggageHandling
0 0 0 0 0 0 0 0
Sat.CheckinService Sat.InflightService Sat.Cleanliness DepartureDelayMinutes ArrivalDelayMinutes Satisfaction TotalScore
0 0 0 0 393 0 0
```

➔ **ANALISI** – il dataset presenta missing values per un solo attributo:  $\$ArrivalDelayMinutes$ . I missing values sono 393 e rappresentano lo 0.30% delle istanze. In quanto presenti solo per un singolo attributo e in numero davvero esiguo non risulta così utile analizzarne la distribuzione graficamente.

➔ **STRATEGIA** – sempre per i motivi appena citati si è scelto in questo caso di sostituire i missing values con il corrispettivo valore  $\$DepartureDelayMinutes$  e quindi di non utilizzare nessuno algoritmo di ML per calcolarli. Infatti come evidente nel *Graf.B9* le due features sono altamente correlate (in questo modo si mantiene la correlazione, cosa che non sarebbe successa se li avessimo sostituiti con la media per esempio).

## MISSING VALUES 2

Dalle informazioni fornite non è chiaro se per tutte le variabili che calcolano la soddisfazione da 0 a 5, il valore 0 indichi un NA. (Non avendo la possibilità di interrogare il cliente) Si sceglie questa strada e si va quindi a trasformare tutti gli '0' in NA per queste variabili.

Per ogni variabile che inizia per Sat.\* allora:

```
#Sat.FlightSchedule
levels(df$Sat.FlightSchedule) = c(levels(df$Sat.FlightSchedule),NA) #aggiungo il level "NA"
#levels(df$Sat.FlightSchedule) #check livelli
df$Sat.FlightSchedule[df$Sat.FlightSchedule==0] = NA #sostituisco il livello '0' con il livello 'NA'
df$Sat.FlightSchedule = factor(df$Sat.FlightSchedule) #questo comando elimina i livelli non più utilizzati
```

### NOTA

> Sarà interessante capire se gli NA indicano solamente che il cliente non ha usufruito del servizio (→se sono presenti solo nei servizi opzionali) o siano sintomo di valori mancanti per altri motivi (→se sono presenti anche in campi che avrebbero dovuto essere obbligatori) [nel primo caso sarebbe interessante studiare gli NA per analizzare quante persone hanno usufruito o meno di un certo servizio e se esiste una certa correlazione tra l'utilizzo di questi servizi 'opzionali']

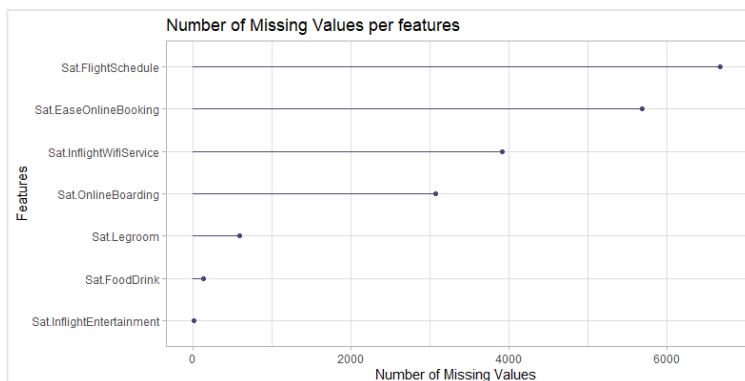
```
> apply(df, function(x) sum(is.na(x)==TRUE))
```

```
Id Gender CustomerType Age TypeOfTravel Class FlightDistance Sat.InflightWifiService Sat.FlightSchedule Sat.EaseOnlineBooking
0 0 0 0 0 0 0 0 3916 6681 5682
Sat.GateLocation Sat.FoodDrink Sat.OnlineBoarding Sat.SeatComfort Sat.InflightEntertainment Sat.OnBoardService Sat.Legroom Sat.BaggageHandling
1 132 3080 1 18 5 598 0
Sat.CheckinService Sat.InflightService Sat.Cleanliness DepartureDelayMinutes ArrivalDelayMinutes Satisfaction TotalScore
1 5 14 0 0 0 0
```

➔ **ANALISI** – il dataset presenta missing values per più attributi. I missing values sono **20 134** e rappresentano lo **0.62%** delle celle totali. Poiché i missing values sono distribuiti su più features risulta utile anche andare a contare le righe incomplete: sono presenti **10 273** istanze incomplete (una istanza può avere più di un missing values) che rappresentano il **7.94%** delle istanze totali.

NA cells	20 134	0.62% del # totale di celle
INCOMPLETE rows	10 273	7.94% del # totale di righe

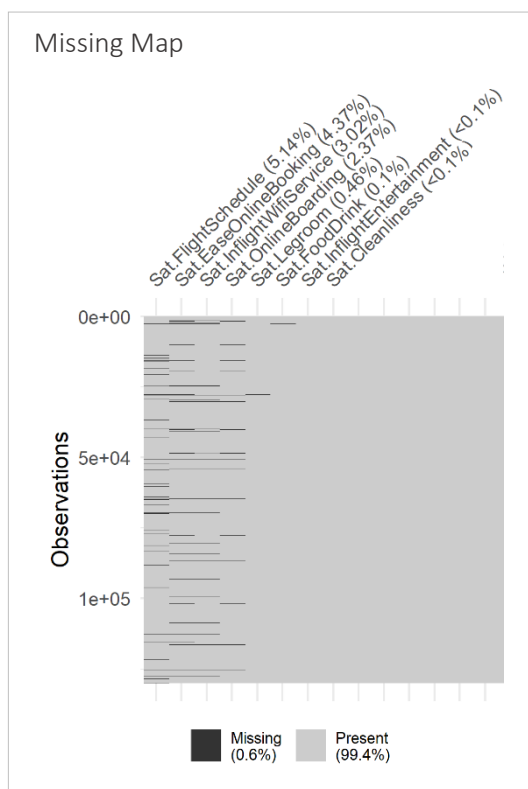
In questo caso un'analisi grafica e più approfondita dei missing values (e sulla loro distribuzione) potrebbe portare a risultati interessanti.



Veloce rappresentazione (ordinata) degli attributi con più missing values



## > ANALISI SULLA DISTRIBUZIONE DEI MISSING VALUES



Il grafico mostra la distribuzione dei missing values per ogni attributo. Gli attributi sono ordinati in ordine decrescente in base al numero di NA. Ogni attributo è etichettato da un numero che va ad indicare la percentuale di missing values sul totale delle istanze per quella colonna. [gli attributi non presenti hanno 0 NA]

### NOTA

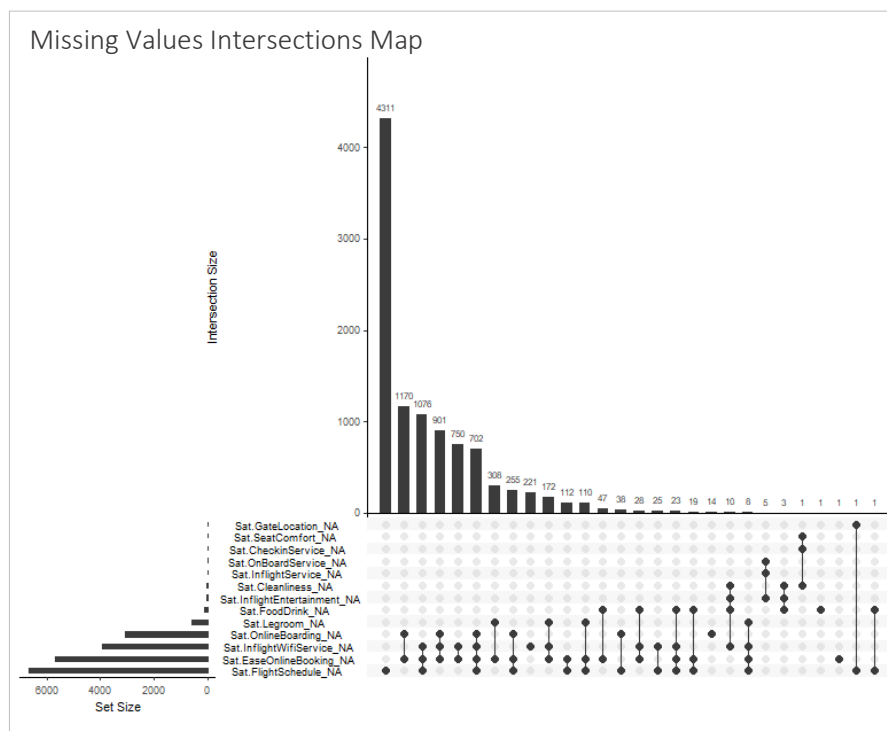
> *\$Sat.FlightSchedule*, per esempio, presenta dei missing values (avrebbe dovuto essere un campo obbligatorio: qualunque passeggero avrebbe potuto dare un voto per la sua soddisfazione sull'orario del volo) quindi potrebbe non essere sempre valida l'ipotesi per cui '0' corrisponde semplicemente ad un servizio di cui il passeggero non abbia usufruito.

> Viene data conferma del fatto che i missing values sono lo 0.6% del totale delle celle del dataset.

> Si possono notare probabili pattern/modelli nella distribuzione dei missing values [in molti casi sono le stesse istanze ad avere più di 1 missing value → questo era già stato ipotizzando quando

abbiamo osservato di avere 20 134 missing values ma 10 273 righe incomplete].

## > ANALISI SULLA COMBINAZIONE DEI MISSING VALUES (PATTERN)



Il grafico mostra la quantità di istanze che hanno come missing values una combinazione dei vari attributi (in ordine decrescente per quantità).

### NOTA

> La maggior parte delle istanze incomplete ha *\$Sat.FlightSchedule* come unico NA.

> È presente una sola istanza che ha *\$Sat.EaseOnlineBooking* come unico NA nonostante sia il secondo attributo per numero di missing values.

> Le istanze che presentano

missing values tra le combinazioni degli attributi *\$Sat.EaseOnlineBooking*, *\$Sat.OnlineBoarding*, *Sat.WifiService* (sommate) sono davvero tante. Tutti e 3 i campi si riferiscono a servizi (opzionali) legati alla tecnologia: si potrebbe dedurre che esista una certa categoria di passeggeri (presumibilmente la parte più anziana) che non utilizza questi servizi (*da verificare*).

➔ **STRATEGIA** – nonostante i missing values non siano molti (0.62%) coprono quasi l’8% delle istanze del dataframe ed inoltre sono presenti solo in attributi del tipo ‘Sat.\*’. Sarebbe stato corretto, per esempio, costruire un modello di classificazione per calcolare gli NA; malgrado ciò è stato deciso di eliminare le istanze incomplete: il campione rimane comunque molto ampio ed inoltre non è questo l’obiettivo principale del progetto.

## CONSISTENCE

### *Vincoli degli attributi*

I vincoli per gli attributi sono stati ricavati in parte dalle informazioni rese disponibili dall’autore del dataset e in parte dedotti dal contesto reale. Questi vincoli sono già stati specificati nel capitolo “1-INFORMAZIONI SUL DATASET”.

### *Ricerca dei vincoli violati*

Per la maggior parte delle features è bastata una verifica con `summary(df)` andando a controllare:

- il valore massimo e minimo per gli attributi numerici
- i livelli (valore che assumono le istanze per quell’attributo) per i fattori

```
> summary(df)
   Id      Gender      CustomerType      Age      TypeOfTravel      Class      FlightDistance      Sat.InflightwifiService
Min.   : 1      Female:60598 OK| disloyal Customer: 19234 OK| Min.   : 7.00 OK| Business travel:82676 OK| Business:58157 OK| Min.   : 31 OK| 1:20771
1st Qu.:33014      Male :58969 OK| Loyal Customer :100333 OK| 1st Qu.:28.00      Personal Travel:36891 OK| Eco   :52631 OK| 1st Qu.: 441 2:30898
Median :65591                                Median :40.00                                Eco Plus: 8779 OK| Median : 868 3:30817
Mean   :65299                                Mean   :39.86                                Mean   :1224 4:23490
3rd Qu.:97640                                3rd Qu.:51.00                                3rd Qu.:1775 5:13591
Max.   :129880                                Max.   :85.00 OK|                                Max.   :4983 OK|

Sat.FlightSchedule Sat.EaseOnlineBooking Sat.GateLocation Sat.FoodDrink Sat.OnlineBoarding Sat.SeatComfort Sat.InflightEntertainment Sat.OnBoardService
1:19049      1:21063      1:20081      1:14584      1:12346      1:13514      1:13957      1:13505
2:21261      2:29069      2:22326      2:25257      2:20784      2:16834      2:19951      2:17014
3:21941 OK|   3:29365 OK|   3:32714 OK|   3:25549 OK|   3:25767 OK|   3:21471 OK|   3:21696 OK|   3:25927 OK|
4:30342      4:23392      4:27963      4:28367      4:36341      4:37085      4:34575      4:36043
5:26774      5:16678      5:16483      5:25810      5:24329      5:30663      5:29388      5:27078

Sat.Legroom Sat.BaggageHandling Sat.checkinService Sat.InflightService Sat.Cleanliness DepartureDelayMinutes ArrivalDelayMinutes
1:11689      1: 7910      1:15007      1: 7803      1:15226      Min.   : 0.00 OK| Min.   : 0.00 OK|
2:22418      2:13138      2:15017      2:13117      2:18360      1st Qu.: 0.00      1st Qu.: 0.00
3:22955 OK|   3:24211 OK|   3:32660 OK|   3:23694 OK|   3:28217 OK|   Median : 0.00      Median : 0.00
4:33733      4:43454      4:33493      4:43895      4:31557      Mean   :14.91      Mean   :15.28
5:28772      5:30834      5:23390      5:31058      5:26207      3rd Qu.:13.00      3rd Qu.:14.00
Max.   :1592.00 OK| Max.   :1584.00 OK|

      Satisfaction      TotalScore
neutral or dissatisfied:68537 OK| Min.   :29.00 OK|
satisfied              :51030 OK| 1st Qu.:52.00
                        Median :59.00
                        Mean   :58.91
                        3rd Qu.:66.00
                        Max.   :83.00 OK|
```

#si poteva tralasciare la colonna ‘Id’ che non ha senso analizzare `summary(df[2:col(df)])`

È stata inoltre effettuata una verifica sulla presenza o meno di dati duplicati. In questo caso l’unico campo di cui aveva senso verificare l’unicità è ‘Id’.

In entrambi i casi non è stata rilevata **nessuna violazione** di un vincolo.

## > ANALISI OUTLIERS

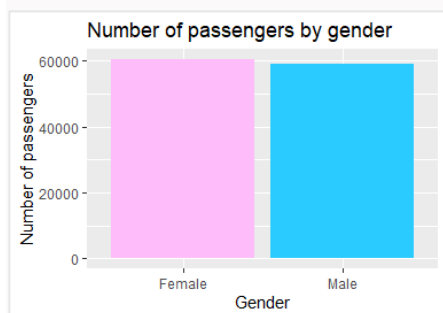
Avendo già verificato il rispetto dei vincoli, in questo caso e in questa fase di pre-processing, si è quasi sicuri che non ci siano outliers da eliminare immediatamente in quanto valori anomali e scorretti dei singoli attributi. Si lascia quindi lo studio più approfondito alla fase successiva.

## A - CARATTERISTICHE DEI PASSEGGERI

## IPOTESI E NOTE SULLE MODALITÀ DEL SONDAGGIO

Quando si parla di *passenger* nella seguente analisi non ci si riferisce per forza a persone fisicamente diverse tra loro: nel caso in cui una persona fisica abbia volato più di una volta per la compagnia e abbia partecipato al sondaggio in più casi, nel data frame sarà presente un'istanza per ogni sua sottoscrizione al sondaggio(?).

## Gender



Graf.A1

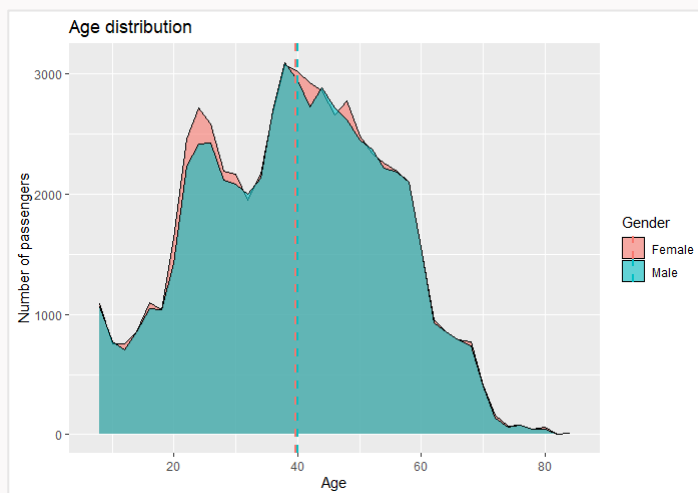
I passeggeri che hanno partecipato al sondaggio si dividono equamente (quasi perfettamente) per genere. Si può notare una leggerissima maggioranza dei passeggeri di sesso femminile.

Female	60 598	50.68%
Male	58 969	49.32%
	119 567	100.00%

## Age

Il passeggero più giovane ha un'età di **7** anni mentre il più anziano di **85**. La media di età per il sesso maschile e femminile è per entrambi intorno ai **40** anni. L'età in cui i passeggeri viaggiano di più con la compagnia è 39 anni (3487 passeggeri).

La distribuzione dell'età dei passeggeri è quasi identica per i due generi. Dopo i 20 anni sempre più passeggeri viaggiano con la compagnia (forte trend positivo). Il trend fortemente negativo inizia dopo i 58 anni.

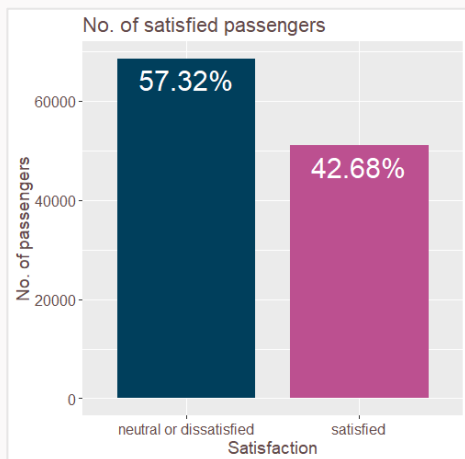


Graf.A2

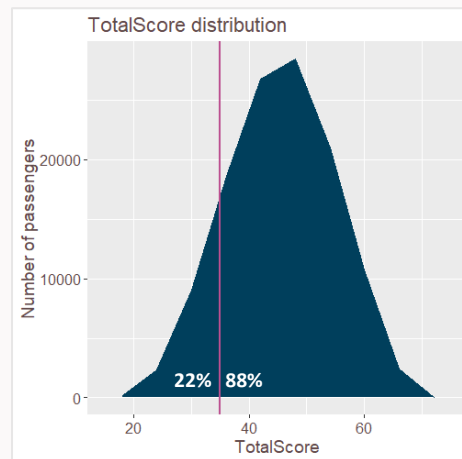
Intorno ai 25 anni si nota una consistente differenza tra i due generi: nel punto di picco viaggiano più di 250 ragazze in più del numero dei ragazzi. Abbiamo una forte inversione di trend che inizia intorno ai 25 anni e termina intorno ai 35.

## B - SODDISFAZIONE COMPLESSIVA DEI PASSEGGERI

### SODDISFAZIONE DEI PASSEGGERI



Graf.B1



Graf.B2

Il *Graf.B1* evidenzia un risultato quasi **bilanciato** con una leggera prevalenza dei passeggeri che si sono dichiarati 'neutrali o non soddisfatti': sono infatti il 57,32% contro il 42,68% che si è invece dichiarato 'soddisfatto'. Nonostante il risultato possa apparentemente sembrare tutt'altro che positivo, è bene sottolineare che tra il 57.32% sono presenti anche i passeggeri che hanno definito la loro soddisfazione come 'neutrale'.

Per evidenziare la positività del risultato è utile andare a rappresentare graficamente la distribuzione della feature \$TotalScore che rappresenta la somma dei punteggi (da 1 a 5) dei 14 attributi che misurano la soddisfazione dei passeggeri nei vari ambiti. Come si nota dal *Graf.B2* la maggior parte dei passeggeri (88%) ha espresso la propria soddisfazione con un punteggio medio  $\geq$  di 2,5 su 5 (l'intercetta è posizionata per TotalScore = 35  $\rightarrow$  punteggio medio = 2,5).

### SODDISFAZIONE DEI PASSEGGERI IN BASE AL GENERE



Graf.B3

La distribuzione della **soddisfazione dei passeggeri per i due sessi 'Female' e 'Male' è pressoché identica** (evidente nel *Graf.B3*). Si potrebbe già osservare quindi che il genere potrebbe non rappresentare un discriminante per l'obiettivo finale. Il fatto che la popolazione si divida quasi perfettamente a metà per genere ci aiuta a giungere più facilmente a tale conclusione.

## SODDISFAZIONE DEI PASSEGGERI IN BASE AL MOTIVO DI VIAGGIO



Graf.B4

Analizzare la soddisfazione dei passeggeri in base al motivo del loro viaggio porta a risultati davvero interessanti. Come evidenziato dal *Graf.B4*, **la maggior parte delle persone che viaggia per motivi di lavoro è rimasta complessivamente soddisfatta** dalla compagnia aerea (il 58%). D'altra parte, **tra le persone che viaggiano per motivi personali, solo l'8% è rimasta soddisfatta**.

*Ipotesi* - Questo risultato potrebbe essere influenzato dal fatto che chi viaggia per motivi personali solitamente viaggia con meno frequenza e potrebbe avere aspettative molto più alte di chi viaggia per lavoro.

## SODDISFAZIONE DEI PASSEGGERI IN BASE AL TIPO DI CLIENTE

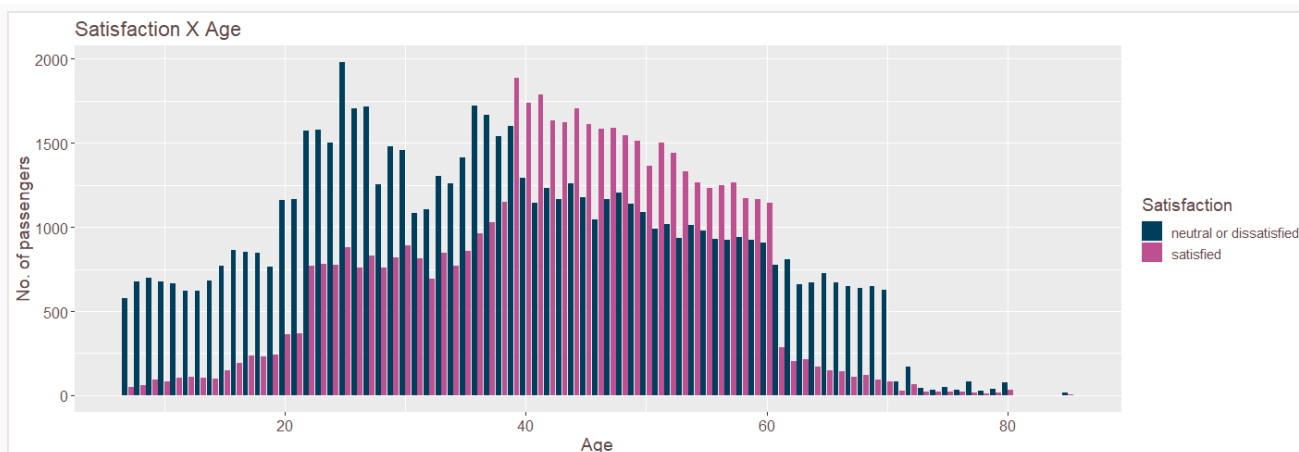


Graf.B5

Come si può notare dal *Graf.B5*, tra i passeggeri fedeli/abituati della compagnia aerea poco meno della metà si ritiene soddisfatto. La situazione è invece ben diversa **per i clienti non abituati dove è il 92% a non ritenersi soddisfatto**.

*Commento* – Risulta abbastanza naturale che la soddisfazione sia maggiore tra i clienti fedeli rispetto ai clienti non abituali.

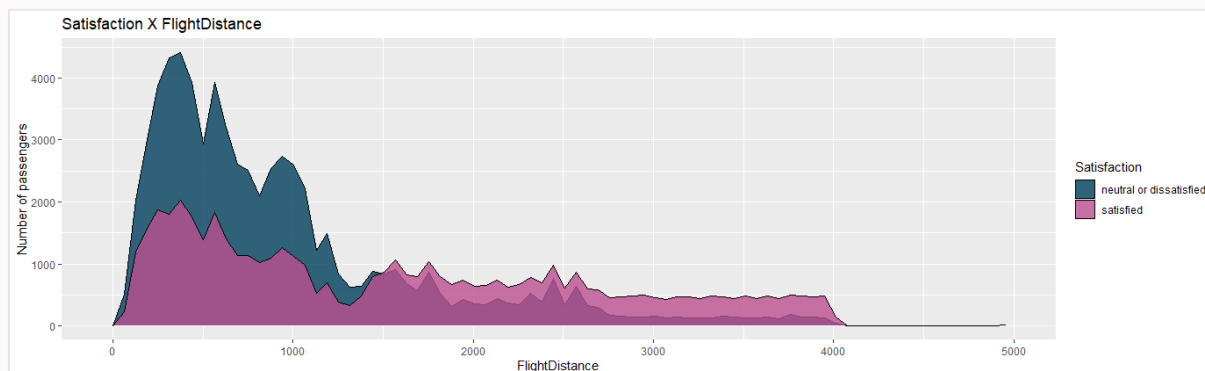
## SODDISFAZIONE DEI PASSEGGERI IN BASE ALL'ETÀ



Graf.B6

La soddisfazione dei passeggeri varia significativamente in base all'età. Come si nota dal *Graf.B6*, **nel range di anni (7-38) e (61-85) il numero di passeggeri 'soddisfatti' è ampiamente inferiore al numero di 'passeggeri neutrali o non soddisfatti'**. **Accade il contrario nel range (39-60)**.

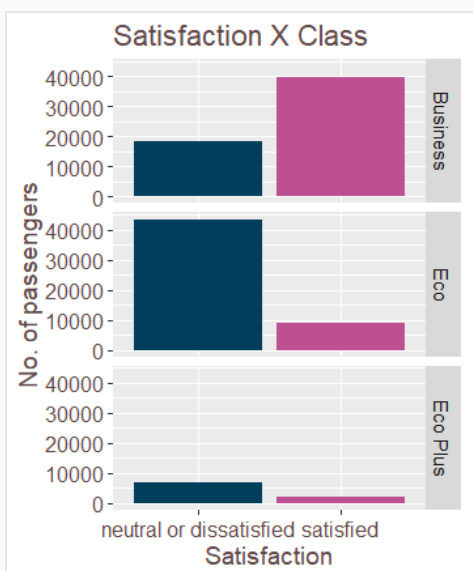
## SODDISFAZIONE DEI PASSEGGERI IN BASE ALLA DISTANZA DEL VOLO



Graf.B7

Si nota dal *Graf.B7* che **per i voli più brevi** (fino a 1500mi) **il numero di passeggeri soddisfatti è molto inferiore** (quasi sempre la metà) rispetto al numero di passeggeri 'neutrali o non soddisfatti'. Per i voli più lunghi invece si presenta la situazione opposta, ma con una differenza più moderata.

## SODDISFAZIONE DEI PASSEGGERI IN BASE ALLA CLASSE

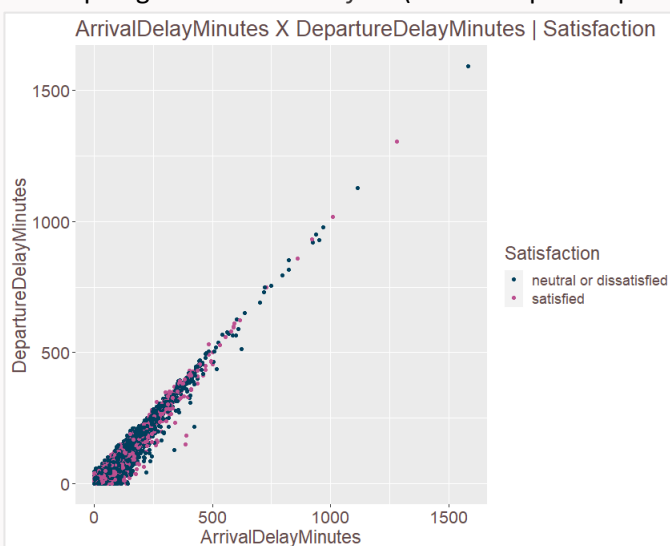


Graf.B8

Lo studio della soddisfazione dei passeggeri in base alla classe di volo porta a delle osservazioni molto interessanti. Come si può notare dal *Graf.B8*, **per le classi 'Eco' e 'Eco Plus' il numero di passeggeri 'soddisfatti' è molto inferiore** al numero dei passeggeri 'neutrali o non soddisfatti'. **Per la classe 'Business' invece addirittura i 2/3 dei passeggeri si ritengono soddisfatti del volo.**

## SODDISFAZIONE DEI PASSEGGERI IN BASE AI MINUTI DI RITARDO ALL'ARRIVO/ALLA PARTENZA

Come si può già notare dal *Graf.B9* (e come si poteva presupporre) i minuti di ritardo all'arrivo sono fortemente correlati con i minuti di ritardo alla partenza quindi ha senso analizzare questi due attributi contemporaneamente.

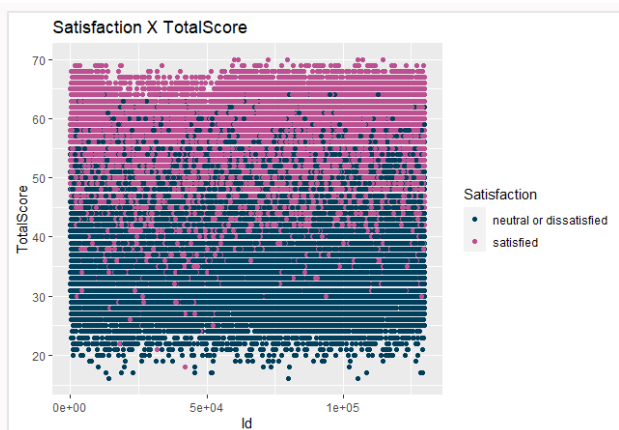


Graf.B9

Dal punto di vista della loro contribuzione alla soddisfazione del cliente sembra non esserci nessun particolare riscontro: ci sono sia **molti passeggeri che hanno viaggiato in ritardo e che si sono ritenuti lo stesso 'soddisfatti'** che passeggeri che hanno viaggiato in orario e si sono ritenuti 'non soddisfatti o neutrali' (si dividono abbastanza equamente). Si può comunque notare che la maggior parte dei passeggeri arrivati con un ampio ritardo non abbiano complessivamente espresso un parere positivo.

## SODDISFAZIONE DEI PASSEGGERI IN BASE AL PUNTEGGIO TOTALE

Esiste un rapporto tra la soddisfazione complessiva del passeggero e il punteggio totale che egli ha dato ai singoli servizi?



Graf.B10

Il risultato che ci si aspetta è che 'più i passeggeri hanno valutato positivamente i singoli servizi, più è alta la probabilità che siano rimasti complessivamente 'soddisfatti'. Come si nota dal *Graf.B10* ciò accade, ma non in modo così netto: **ci sono molti passeggeri che hanno totalizzato un TotalScore molto alto, ma che non si sono ritenuti soddisfatti (e viceversa)**. Esistono anche alcuni casi molto estremi: probabilmente sono degli outliers da trattare in modo opportuno.

Andando ad analizzare molto brevemente gli outliers tramite un boxplot (*Graf.B11*) ciò che risulta più strano è che esiste in particolar modo un grande numero di passeggeri che nonostante abbia valutato non positivamente i servizi si è ritenuto soddisfatto della compagnia aerea. Si tratta di dati errati o di un pattern interessante?

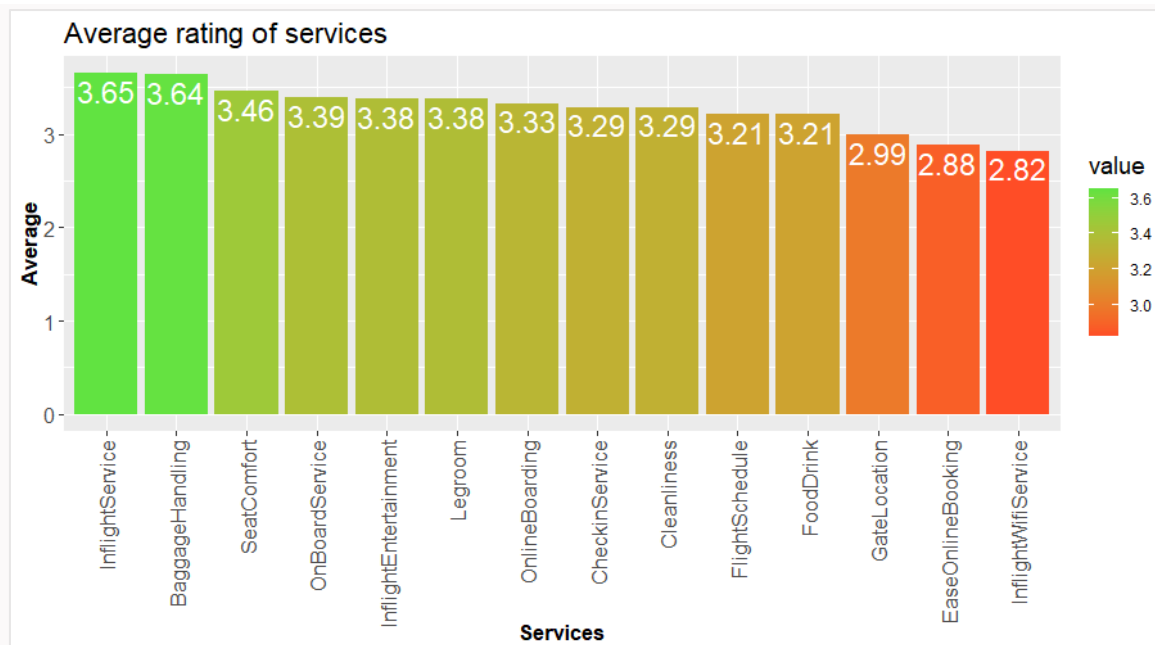
satisfied	52
neutral or dissatisfied	42
TotalScore mean	



Graf.B11

*Commento* - Questo risultato è sicuramente in parte dovuto al fatto che una delle due categorie comprenda sia passeggeri 'neutrali' che 'non soddisfatti'. Potrebbe anche essere che le categorie di valutazione del sondaggio non coprano tutte le sfaccettature dell'esperienza di volo o che contribuiscano alla soddisfazione in modo non equo (con pesi diversi).

## C - VALUTAZIONE DEI SERVIZI



Graf.C1

Come già si era intuito dal *Graf.B2*, complessivamente la valutazione dei servizi è molto buona. Il **voto medio complessivo è 3.28 su 5** risultante dalla media delle medie dei vari servizi (*Graf.C1*). I passeggeri sono rimasti **pienamente soddisfatti** dal **Servizio in volo** e dalla **Gestione dei bagagli** mentre **non completamente soddisfatti** dal **Servizio Wifi** (in volo), dalla **Facilità della prenotazione online** e dalla **Locazione del gate**.

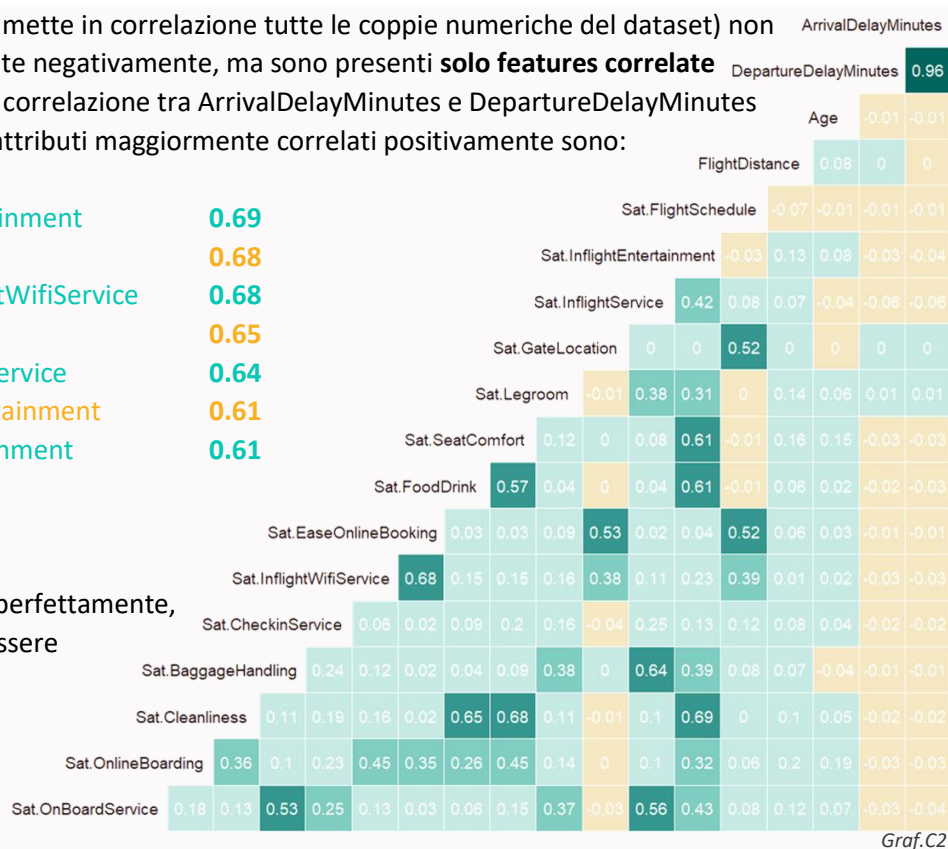
### CORRELAZIONE (lineare) TRA ATTRIBUTI NUMERICI

Come si nota dal *Graf.C2* (che mette in correlazione tutte le coppie numeriche del dataset) non sono presenti features correlate negativamente, ma sono presenti **solo features correlate positivamente**. Escludendo la correlazione tra *ArrivalDelayMinutes* e *DepartureDelayMinutes* che è alta per ovvi motivi, gli attributi maggiormente correlati positivamente sono:

- > Cleanliness – InflightEntertainment **0.69**
- > Cleanliness – SeatComfort **0.68**
- > EaseOnlineBooking – InflightWifiService **0.68**
- > Cleanliness – FoodDrink **0.65**
- > BaggageHandling – InflightService **0.64**
- > SeatComfort – InflightEntertainment **0.61**
- > FoodDrink – InflightEntertainment **0.61**

#### NOTA

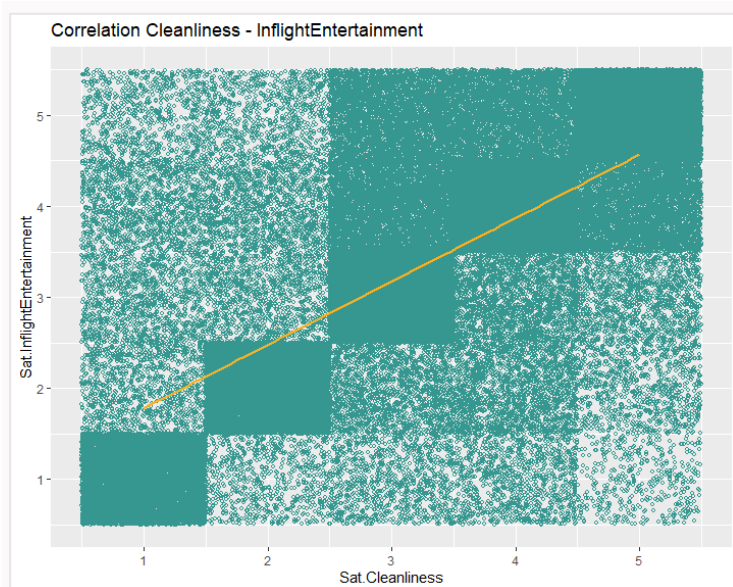
> Nessuna coppia è correlata perfettamente, quindi nessun attributo può essere escluso dall'analisi



Graf.C2



## ANALISI ATTRIBUTI CORRELATI



Graf.C3

### NOTA

> Il Graf.C3 mostra un esempio di come si comportano due variabili correlate positivamente. Poiché si tratta di variabili discrete l'unico modo che abbiamo per ottenere un risultato visivamente utile è grazie alla funzione jitter che aggiunge un piccolo rumore randomico a ogni risposta.

**Cleanliness – InflightEntertainment 0.69**

La soddisfazione sulla pulizia è correlata positivamente con la soddisfazione dell'intrattenimento in volo

**Cleanliness – SeatComfort 0.68**

La soddisfazione sulla pulizia è correlata positivamente con la soddisfazione della comodità dei posti a sedere

**EaseOnlineBooking – InflightWifiService 0.68**

La soddisfazione sulla facilità di prenotazione online è correlata positivamente con la soddisfazione del servizio wifi in volo

**Cleanliness – FoodDrink 0.65**

La soddisfazione sulla pulizia è correlata positivamente con la soddisfazione del servizio 'Cibi e Bevande' in volo

**BaggageHandling – InflightService 0.64**

La soddisfazione sulla movimentazione dei bagagli è correlata positivamente con la soddisfazione del servizio in volo

**SeatComfort – InflightEntertainment 0.61**

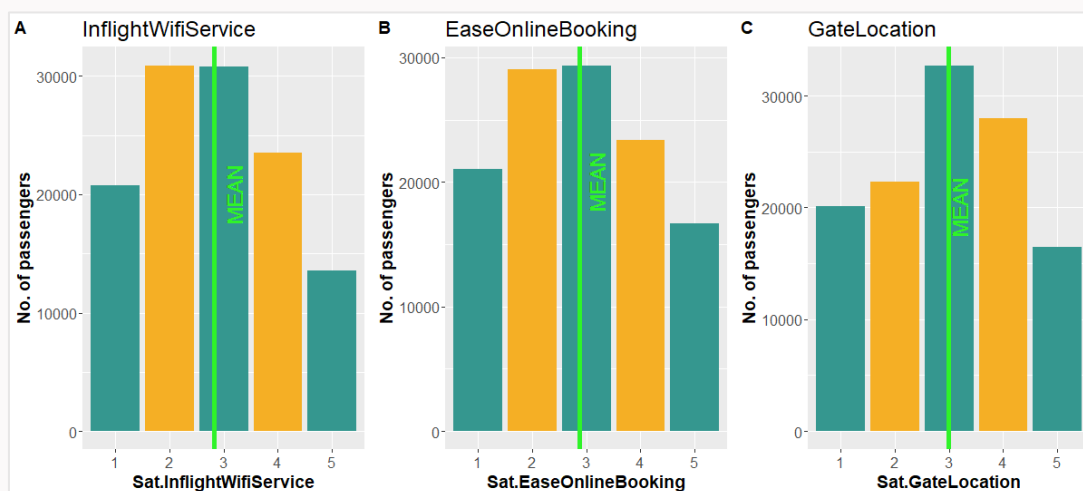
La soddisfazione sulla comodità dei posti a sedere è correlata positivamente con la soddisfazione del servizio in volo

**FoodDrink – InflightEntertainment 0.61**

La soddisfazione del servizio 'Cibi e Bevande' è correlata positivamente con la soddisfazione del servizio in volo

## ANALISI ATTRIBUTI CON VOTO PEGGIORE

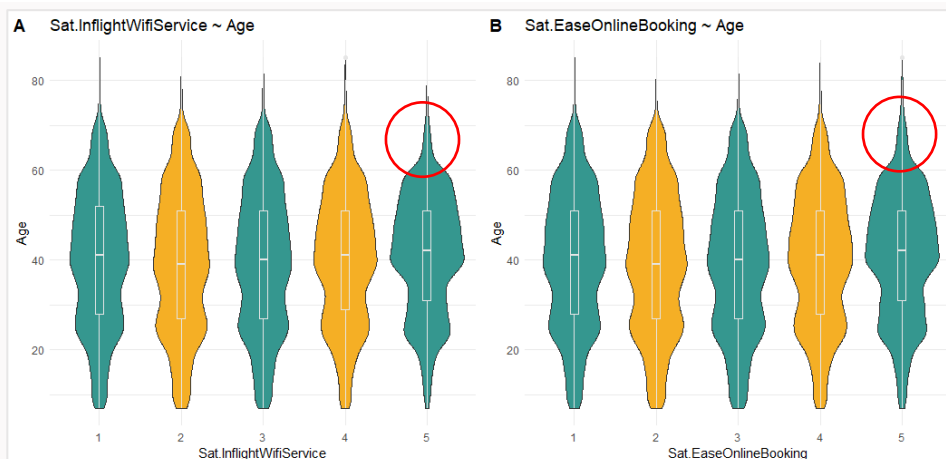
Gli unici 3 servizi con un voto medio minore di 3 sono: Sat.InflightWifiService, Sat.EaseOnlineBooking e Sat.GateLocation. Il Graf.C4 permette di dare un primo sguardo alla distribuzione dei voti per questi attributi.



Graf.C4

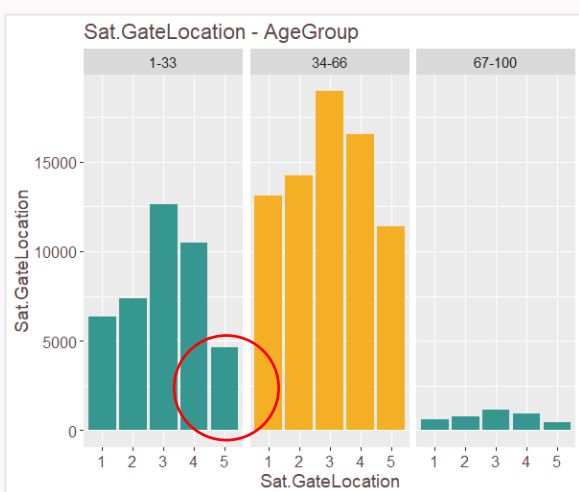
Poiché uno degli obiettivi della compagnia è andare a riconoscere quali servizi contribuiscano negativamente alla soddisfazione dei clienti per poi andare a migliorarli, risulta interessante andare a **capire se esistano alcune categorie di passeggeri che sono rimaste più deluse (rispetto alle altre)** per questi servizi.

Per ognuna delle 3 variabili con voto medio peggiore è stata effettuata un'analisi in base a *TypeOfTravel*, *CustomerType*, *Age*, *Gender*, *Class* e *FlightDistance*. **I voti alle varie categorie sono quasi sempre omogenei: non si notano pattern utili all'analisi.** Seguono solo le osservazioni più interessanti.



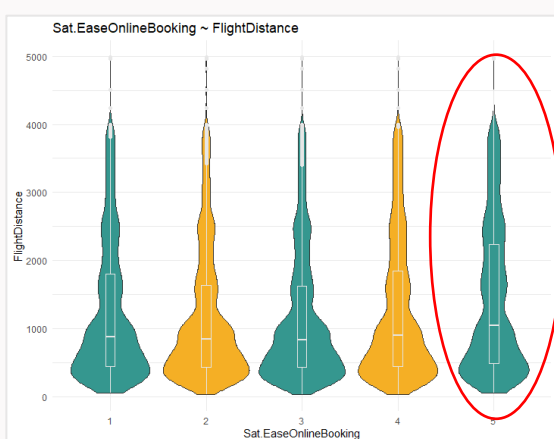
Graf.C5

Nonostante i passeggeri che hanno valutato con il voto massimo **InflightWifiService** e **EaseOnlineBooking** abbiano un'età media più alta, tra questi sono davvero pochi gli **Over60** (hanno valutato il servizio con voto minore). ○ → più fine rispetto agli altri voti



Graf.C6

Nonostante la distribuzione dei voti sia omogenea in tutte e 3 le categorie (di età), è evidente come **nell'intervallo (1-33)**, in proporzione, **molte meno persone abbiano valutato con il massimo dei voti l'attributo GateLocation.**



Graf.C7

Nonostante la distribuzione dei voti per *EaseOnlineBooking* in base a *FlightDistance* sia molto omogenea, è evidente che tra i passeggeri che hanno attribuito il voto massimo ce ne siano meno per viaggi brevi e più per viaggi lunghi (rispetto agli altri voti). → forma più stretta alla base e più larga andando verso l'alto

## 5 MODELLI PREDITTIVI (ML)

Predire la soddisfazione di un passeggero dalle altre features significa trovare una funzione che assegni in modo accurato l'attributo `$Satisfaction` (categoriale) a nuove istanze non classificate: si parla in questo caso di **CLASSIFICAZIONE** (supervised learning). Sono stati costruiti e quindi valutati 2 modelli che utilizzano tecniche di classificazione diverse: **DecisionTree** e **K-NN**.

### D - PREDIRE LA SODDISFAZIONE DEL CLIENTE **DecisionTree**

Tra le varie tecniche di classificazione sono stati scelti gli **alberi di classificazione** (Decision Tree) con costruzione con algoritmo **CART** (Classification And Regression Trees). Uno dei vantaggi degli alberi decisionali è che non hanno ipotesi sulla distribuzione dello spazio.

#### SPLIT DEI DATI

Il dataset è stato splittato in *train* e *test* in proporzioni 70 – 30 (come da approccio standard). L'attributo `$Satisfaction` risulta ben distribuito in entrambi i dataset ottenuti dallo split.

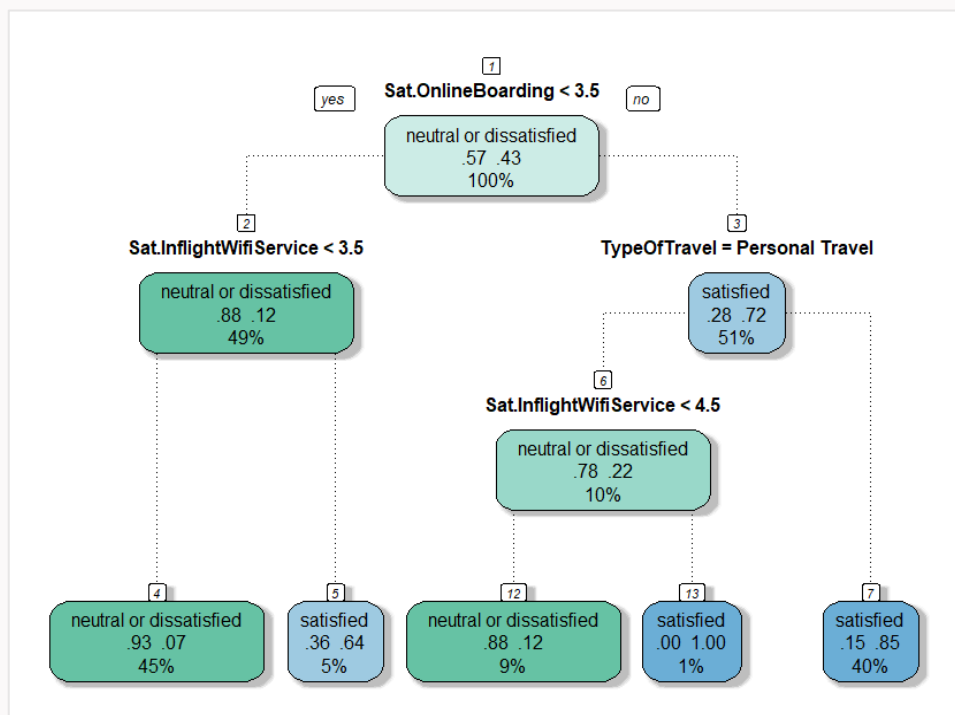
#### PERFORMANCE BASELINE

Viene definito manualmente un modello con la Zero Rule che prevede il valore della classe che ha più osservazioni nel set di dati di addestramento ('neutral or dissatisfied'). L'accuracy per questo modello (ricavato dalla confusion matrix) è pari a 0.57 (in realtà non così male).

#### ALBERO DI DECISIONE - Creazione

Viene creato un DecisionTree con la funzione `rpart` (che di default per calcolare l'impurità del nodo utilizza l'indice di **Gini** per le variabili categoriche). Viene passato come parametro l'intero trainset escludendo manualmente le features `$Id` (fuorviante per la predizione) e `$TotalScore` (ridondante).

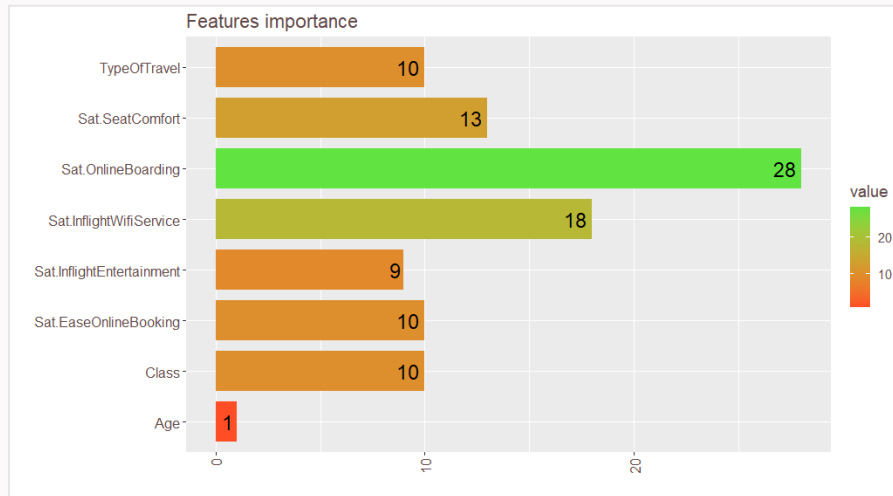
L'albero viene stampato graficamente con `fancyRpartPlot` (libreria rattle).



Graf.D1

Come evidente dal *Graf.D1* l'albero risulta abbastanza semplice: ha 5 nodi foglia (nodi terminali) e 4 nodi di split (con criterio di suddivisione).

## ALBERO DI DECISIONE – Variabili più importanti per il modello predittivo



Graf.D2

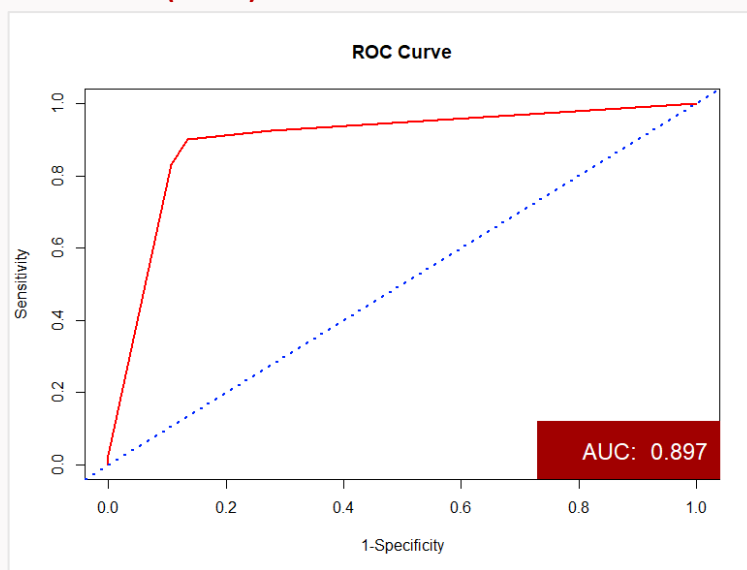
## ALBERO DI DECISIONE - Valutazione

Dopo aver addestrato il nostro albero di decisione possiamo predire le istanze del testset e andare quindi a valutare le prestazioni del modello.

### 1 – MATRICE DI CONFUSIONE

- > **accuracy** = 0.88  $(TP+TN) / (TP+TN+FP+FN) \rightarrow$  error-rate = 0.12
- > **precision** = 0.87  $TP / (TP+FP)$
- > **sensitivity** = 0.92  $TP / (TP+FN)$
- > **specificity** = 0.92  $TN / (TN+FP)$

### 2 – ROC CURVE (e AUC)



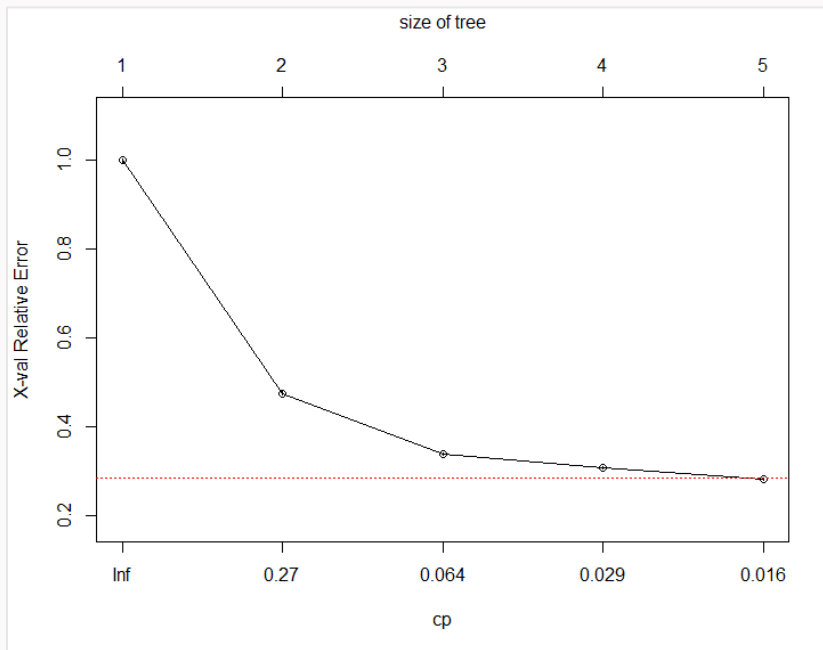
Graf.D3

Come chiaro nel *Graf.D3*, la curva ROC non solo si trova al di sopra della diagonale, ma si avvicina molto alle coordinate ideali (0,1).

L'AUC (Area Under Curve) è pari a 0.897: trovandosi nell'intervallo 0.8-0.9 il modello viene genericamente classificato come 'Good', quasi 'Excellent'.

TruePositiveRate = Sensitivity  
FalsePositiveRate = 1-Specificity

## ALBERO DI DECISIONE - Pruning



Graf. D4

L'errore minimo corrisponde a  $cp=0.016$ . Non dobbiamo effettuare nessun taglio (pruning) in quanto a quel cp (complexity parameter) corrisponde l'attuale dimensione dell'albero.

K-Nearest Neighbor o K-NN è un algoritmo di classificazione non lineare supervisionato. Inoltre è un algoritmo non parametrico, ovvero non fa alcuna supposizione sui dati sottostanti o sulla loro distribuzione.

La funzione utilizzata in questa analisi è `knn` del package 'class' che considera i 'k più vicini' in base alla distanza euclidea (lunghezza del segmento che ha per estremi i due punti).

### NORMALIZZAZIONE DEI DATI

Nella preparazione del dataset si sono convertiti i fattori in numeri (nel caso di variabili binarie in '0' e '1'). Il passaggio successivo è stata la **normalizzazione dei valori del dataset per ogni attributo**, in modo da rendere ininfluente la scala con cui sono stati assegnati.

### SPLIT DEI DATI

Il dataset è stato splittato in *train* e *test* in proporzioni 70 – 30 (come da approccio standard). L'attributo `$Satisfaction` risulta ben distribuito in entrambi i dataset ottenuti dallo split.

### K-NN - Scelta del valore K

Uno dei modi per trovare il valore K ottimale è calcolare la radice quadrata del numero totale di osservazioni nel set di dati. In questo caso nel trainset ci sono 83696 istanze quindi  $k=289.30 \rightarrow \{k=289 \text{ e } k=290\}$

### K-NN - Modello

Sono stati quindi creati i due modelli con i rispettivi valori di  $k=289$  e  $k=290$ . Dall'analisi della matrice di confusione il primo modello risulta molto più prestante.

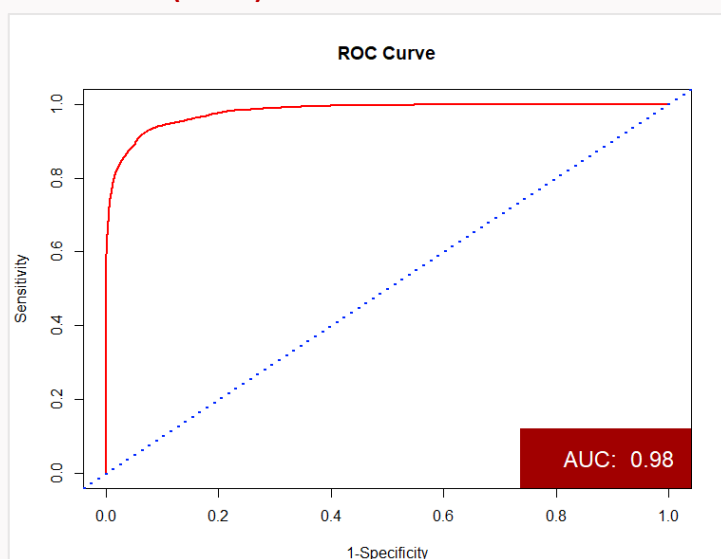
### K-NN - Valutazione

Dopo aver predetto le istanze del testset con K-NN è utile andare a valutare le prestazioni dell'algoritmo.

#### 1 – MATRICE DI CONFUSIONE

- > **accuracy** = 0.91  $(TP+TN) / (TP+TN+FP+FN) \rightarrow \text{error-rate} = 0.09$
- > **precision** = 0.91  $TP / (TP+FP)$
- > **sensitivity** = 0.96  $TP / (TP+FN)$
- > **specificity** = 0.87  $TN / (TN+FP)$

#### 2 – ROC CURVE (e AUC)



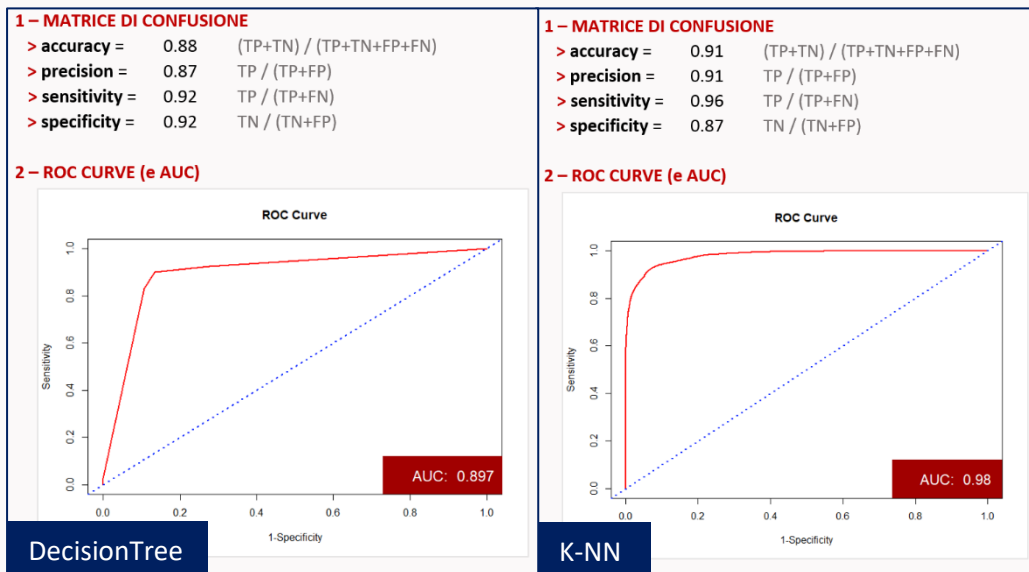
Graf.E1

Come chiaro nel *Graf.E1*, la curva ROC non solo si trova al di sopra della diagonale, ma si avvicina molto alle coordinate ideali (0,1).

L'AUC (Area Under Curve) è pari a 0.98: trovandosi nell'intervallo 0.9-1 il modello viene genericamente classificato come 'Excellent'.

TruePositiveRate = Sensitivity  
FalsePositiveRate = 1-Specificity

## F - CONFRONTO TRA I MODELLI



Sulla base dell'AUC (Area Under Curve nella ROC) e dei principali parametri che misurano le performance dei due modelli (accuracy, precision, sensitivity, specificity) la predizione con K-NN risulta quasi sempre migliore. In generale comunque **entrambi i modelli si possono ritenere 'ECCELLENTI'**: l'accuracy, che rappresenta in questo caso uno dei parametri più importanti, è intorno al 90% per entrambi.

### RAGGIUNGIMENTO OBIETTIVI

#### ✓ **Predizione della soddisfazione dei passeggeri**

Lo studio del dataset ha portato a due eccellenti modelli utili a raggiungere l'obiettivo principale del problema: predire la soddisfazione dei passeggeri.

#### ✓ **Riconoscimento dei fattori che influiscono maggiormente sulla soddisfazione dei passeggeri**

Lo studio del dataset ha portato a riconoscere interessanti categorie di passeggeri (per classe di volo, motivo del volo, età...) che sono rimaste meno soddisfatte delle altre e sulle quali la compagnia può quindi andare a lavorare per migliorare la propria reputazione. Inoltre il modello di classificazione tramite la tecnica degli alberi di decisione ha portato a riconoscere quali sono i fattori più impattanti nel rendere un passeggero più o meno soddisfatto.

#### ✓ **Analisi della valutazione dei servizi offerti**

Dallo studio del dataset è risultato un quadro generale sulla valutazione dei servizi molto positivo. Sono stati inoltre identificati quelli con valutazione peggiore sui quali la compagnia potrebbe andare a lavorare per migliorare la soddisfazione complessiva dei suoi passeggeri. Non è stato riscontrato nessun pattern significativo tra la valutazione dei singoli servizi e le altre features.

### COMMENTO FINALE

Gli obiettivi generati dalle richieste del cliente sono stati discretamente raggiunti e quindi il risultato dello studio del problema può ritenersi soddisfacente.