PRESENTAZIONE DATA TECHNOLOGY & MACHINE LEARNING

Anno accademico 2018/2019

Virgilio Luca 794866

Ventura Samuele 793060

DATA TECHNOLOGY

Obiettivo del progetto

- Predire il tempo di attesa agli sportelli, durante la fase di accettazione, presso l'ospedale Galeazzi di Milano.
- Integrare informazioni contenute in diversi dataset per poter aumentare le informazioni di base da sfruttare nella fase di predizione.
- Utilizzare solo le informazioni disponibili a priori nel momento in cui un paziente si presenta al totem per prenotare il suo turno.

Dataset

- **TuPassi**

  contiene le informazioni registrate dal sistema Tu Passi nel momento in cui un paziente si registra al totem prima di recarsi agli sportelli dell'accettazione

- **Precipitazioni**

  contiene informazioni riguardo le precipitazioni registrate dal sensore posizionato «il più vicino all'ospedale»

- **Festivo**

  per ogni giorno dell'anno 2018 indica la vicinanza ad una data festiva

Data Quality metriche

Dimensione	Tipo	Metrica
Completezza	Completezza per attributo	Numero valori non nulli(per colonna)/ Numero totale di valori
Completezza	Completezza a livello di tabella	Numero valori non nulli/ Numero totale di valori
Consistenza	Vincoli interni alla tabella	Numero valori consistenti / Numero totale di valori
Unicità	Unicità delle tuple	Numero tuple duplicate

Data cleaning

Modifiche effettuate:

- Sostituzione di alcuni valori malformati
- Cancellazione delle tuple nel caso in cui il valore della colonna Presente_alle_ore sia vuoto
- Cancellazione delle tuple nel caso in cui il valore della colonna Ultima_operazione_alle sia vuoto
- Modifica di valori nel caso di errori di data entry

Data Quality

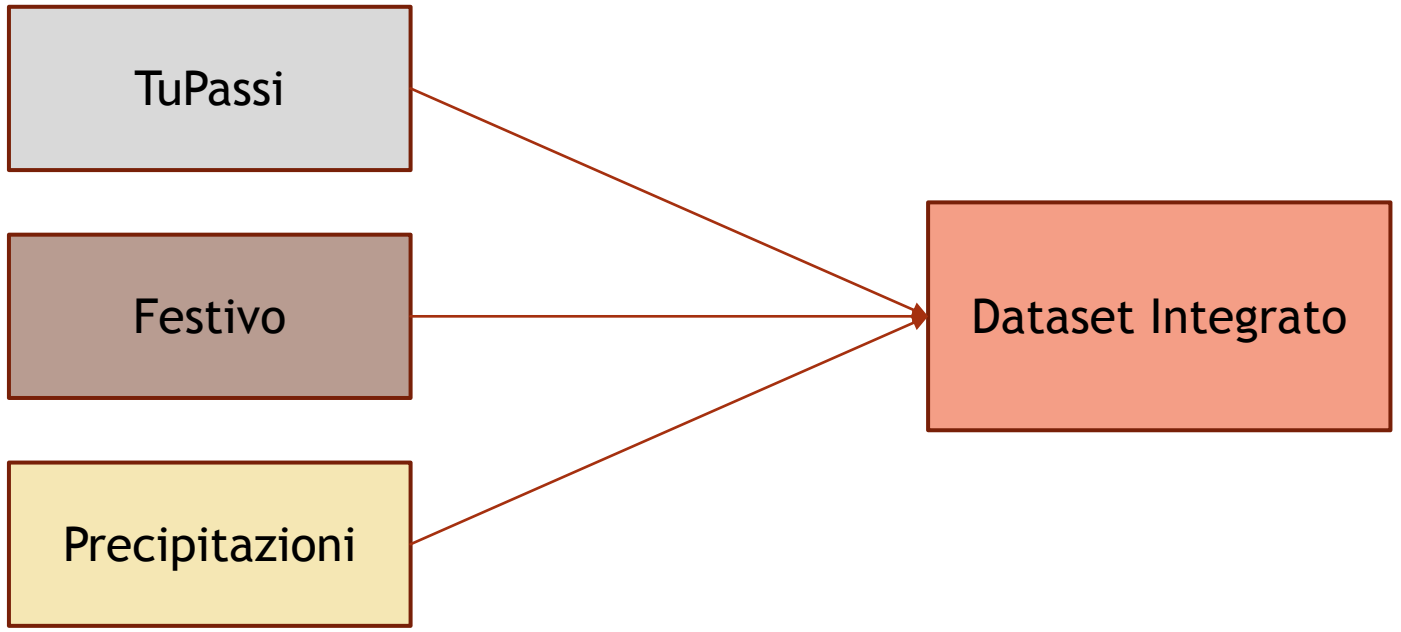
Tu Passi

Metrica	Prima di DQ	Dopo DQ
Completezza della tabella	0.97	0.99
Correttezza dei dati	1	1
Unicità delle tuple	0	0

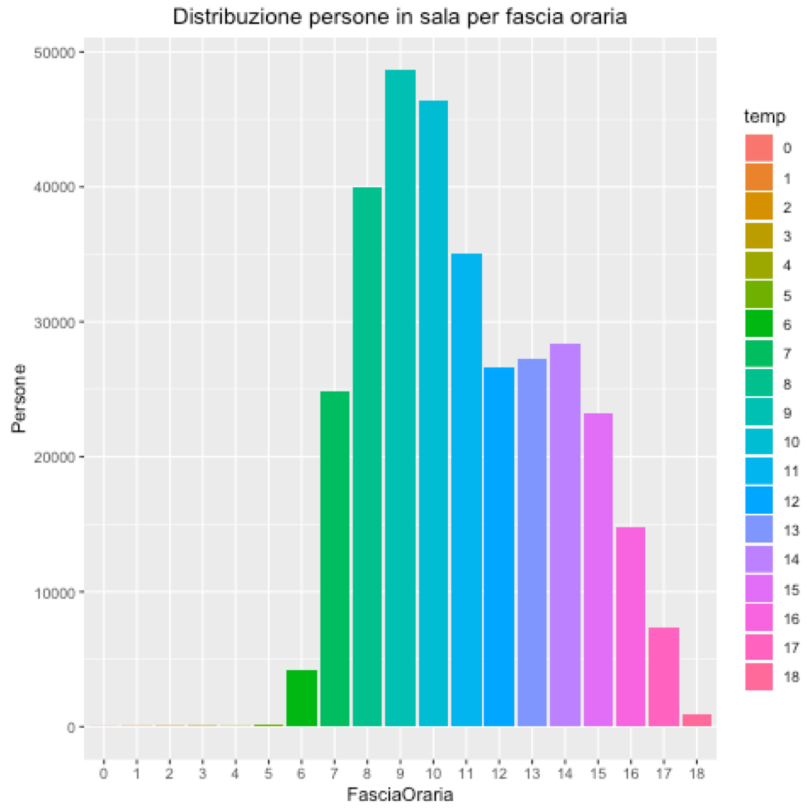
Per quanto riguarda gli altri due dataset non è stata necessaria una fase di data cleaning.

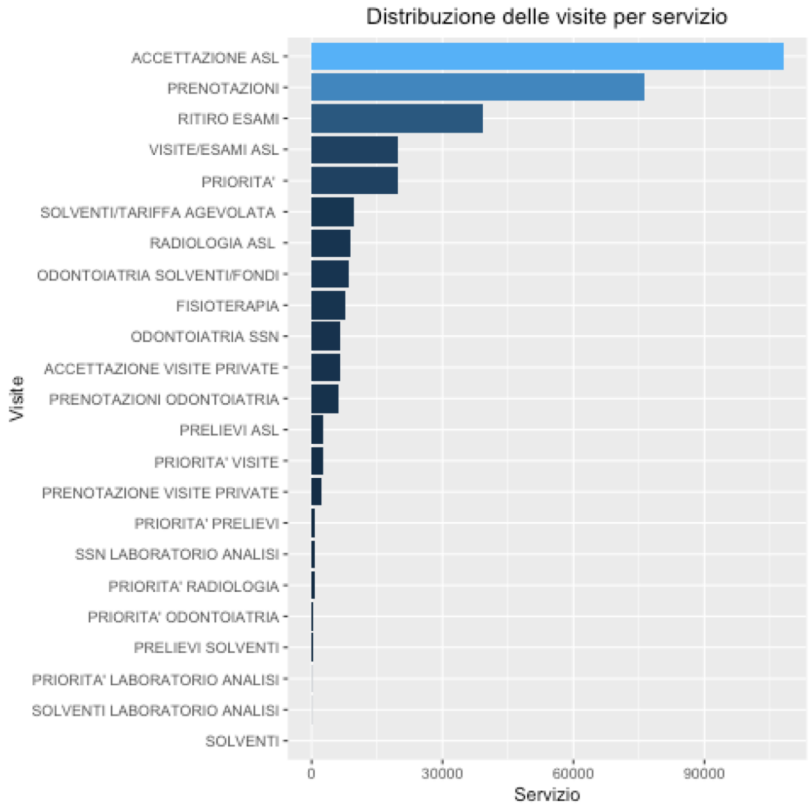
Data integration

- I tre dataset rappresentano domini differenti quindi sono stati integrati con la tecnica del consolidamento, unendo le colonne Festivo e MediaP rispettivamente sulla base del giorno e della fascia oraria nel dataset TuPassi.

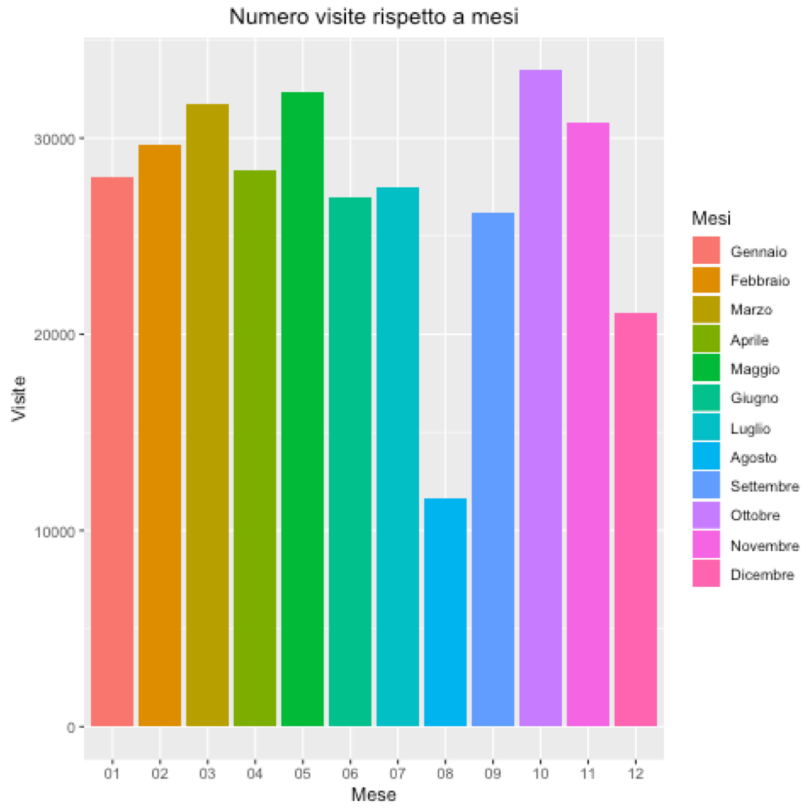


Data exploration 1





Data exploration 2

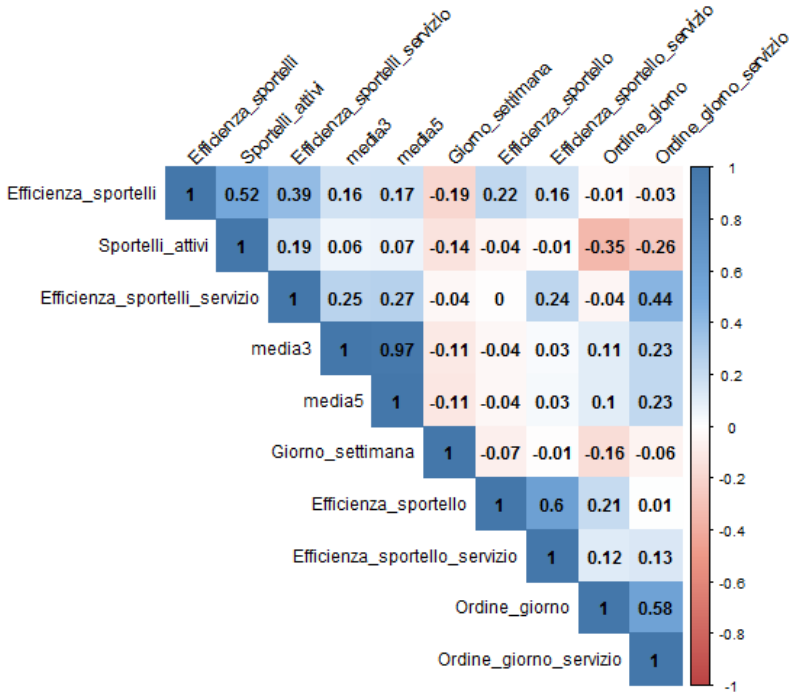


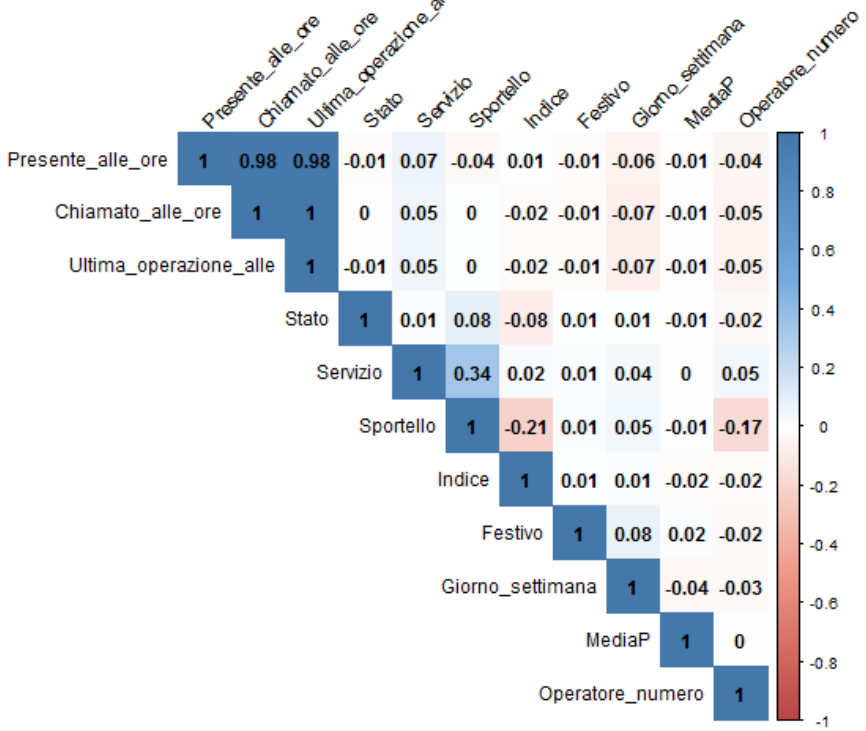
MACHINE LEARNING

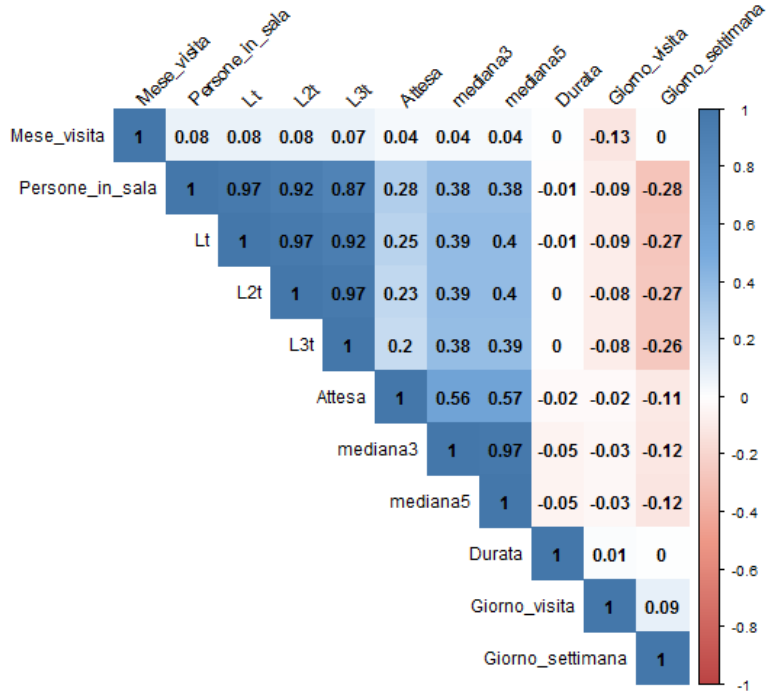
Calcolo parametri

- Calcolo della variabile target Attesa.
- Data la bassa correlazione delle features con la variabile target Attesa è stato necessario calcolare vari parametri, utilizzando solo le informazioni disponibili nel momento in cui un paziente si presenta al totem.
- I parametri principali calcolati sono:
  - Mediana e media per le ultime 3 o 5 persone
  - Persone_in_sala, in istanti differenti di tempo
  - Sportelli_attivi
  - Ordinamento delle persone per giorno e per giorno/servizio
  - Numero di persone servite negli ultimi 10 minuti

Feature selection





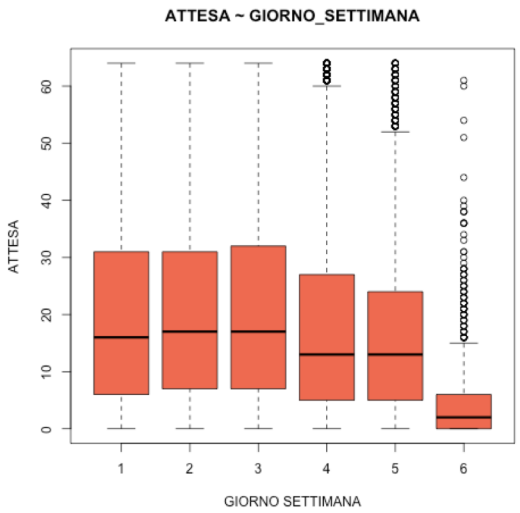


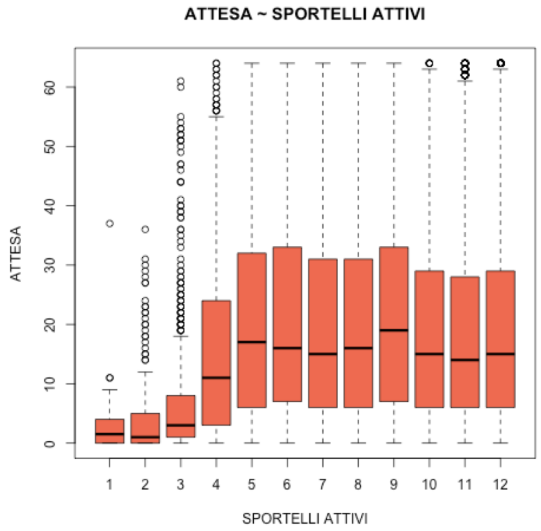
Scelta del modello

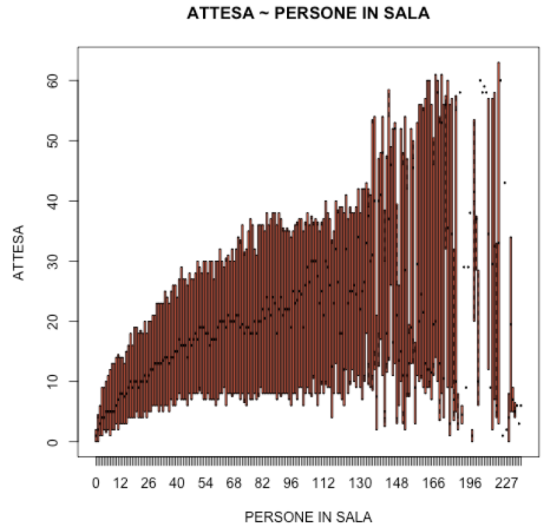
- In entrambi i modelli si sono divisi i dati in due subset (70% e 30%) per il training e la predizione del modello.
- Inizialmente si è scelto si sfruttare un albero di decisione per la sua semplicità, successivamente una random forest per aumentare le performance.
- La variabile target Attesa è stata discretizzata, per potere effettuare una classificazione, secondo le seguenti fasce:

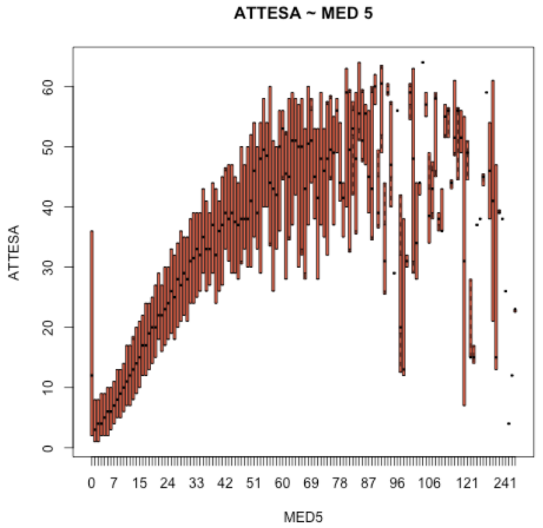
VALORE ATTESA IN MINUTI	CLASSE
attesa >= 10	1
11 <= attesa <=20	2
21 <= attesa <=30	3
attesa > 30	4

Analisi esplorativa Training Set



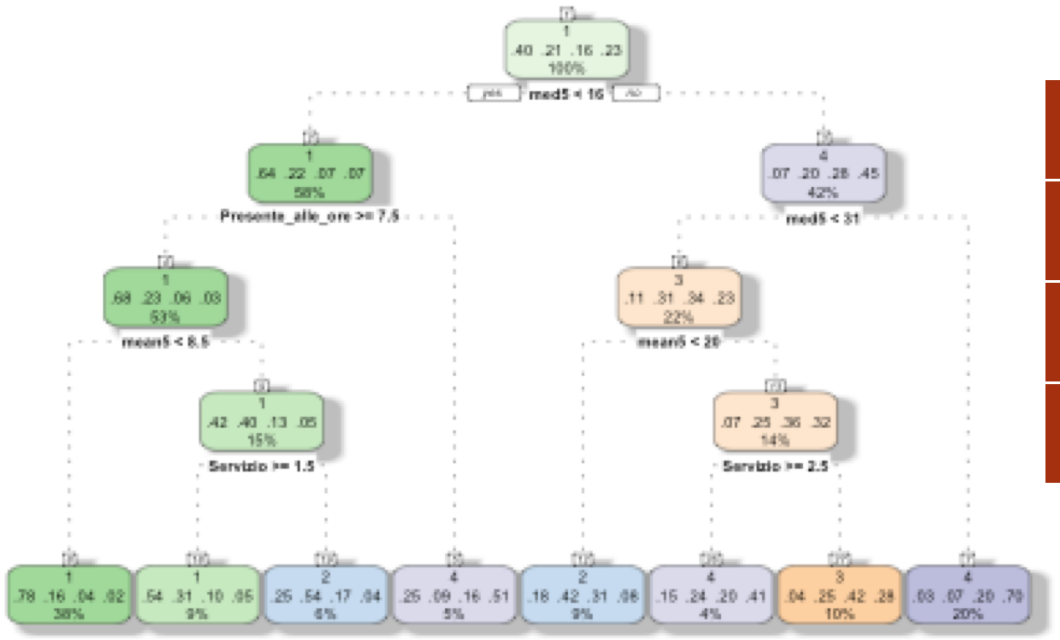






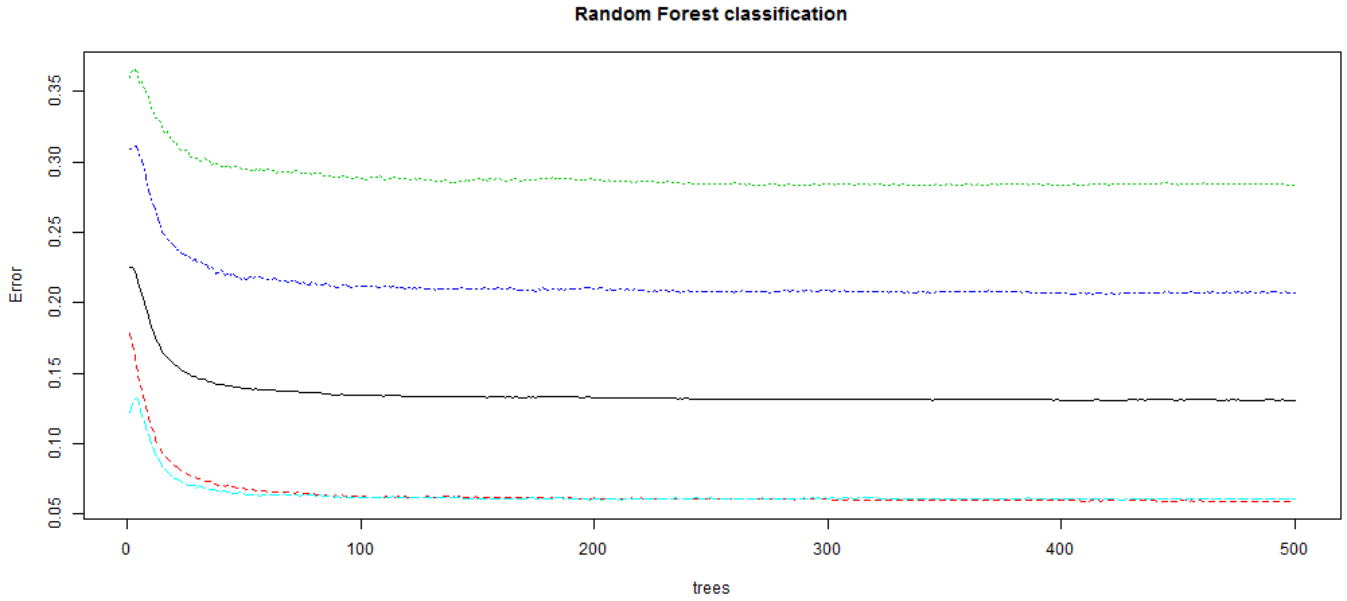
Decision tree

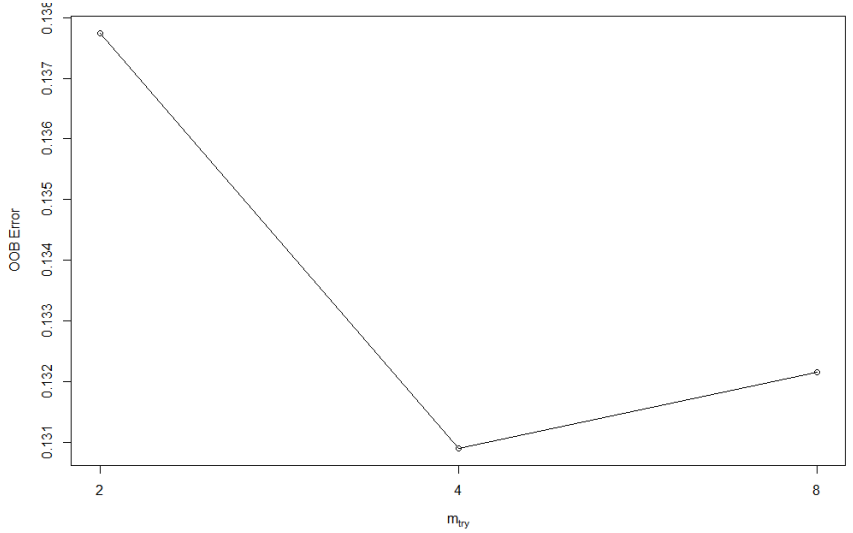
- Dopo aver provato varie configurazioni si è scelta quella con maggiore interpretabilità e con delle performance discrete.



Accuracy	0.6301			
Precision	0.8551	0.3267	0.2657	0.7794
Recall	0.7295	0.4738	0.4187	0.6248
F-measure	0.7873	0.3868	0.3251	0.6936

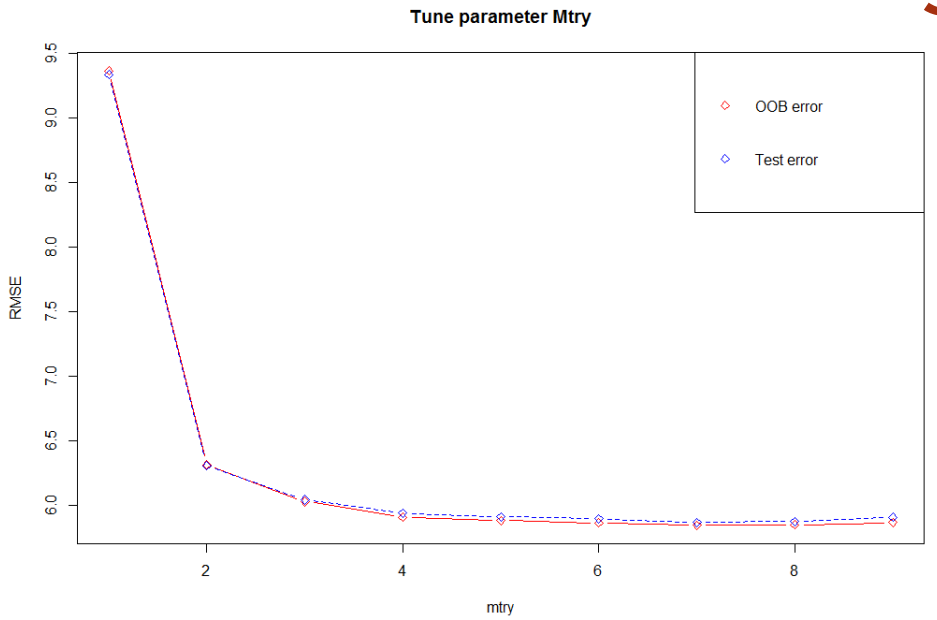
Random forest - classification



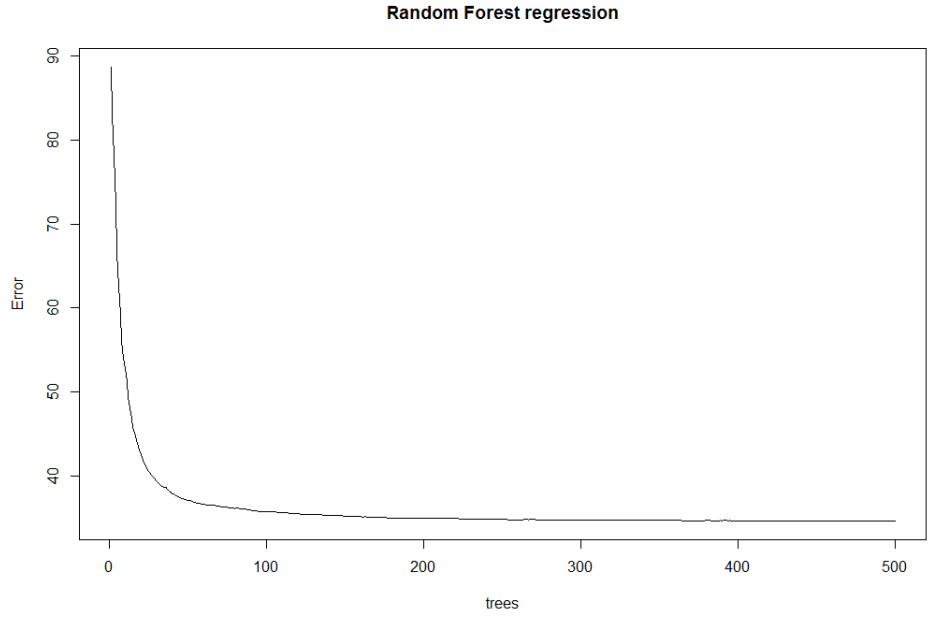


Accuracy	0.0635
Precision	0.8427
Recall	0.8552
F-measure	0.8489

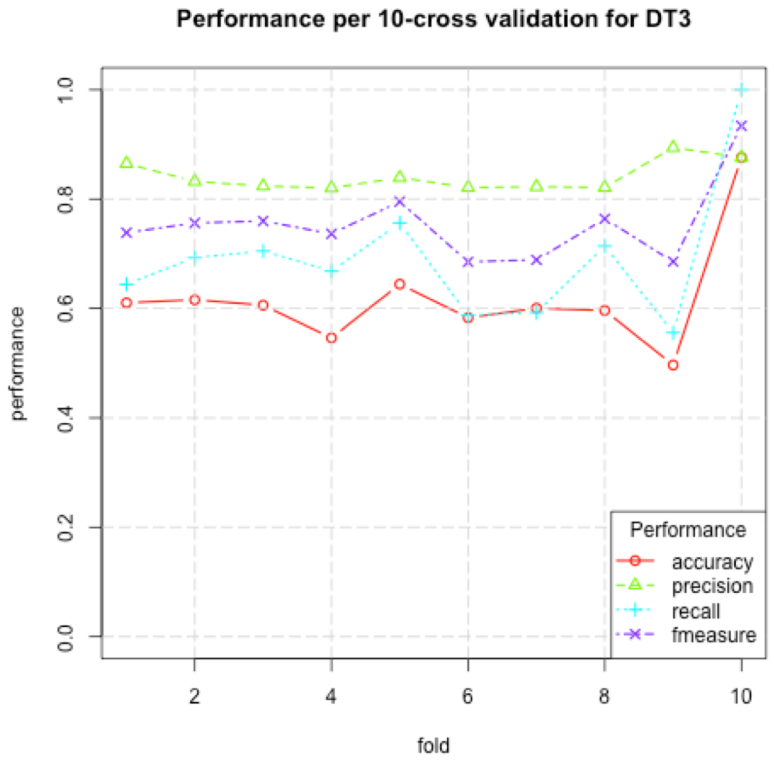
Random forest - regression



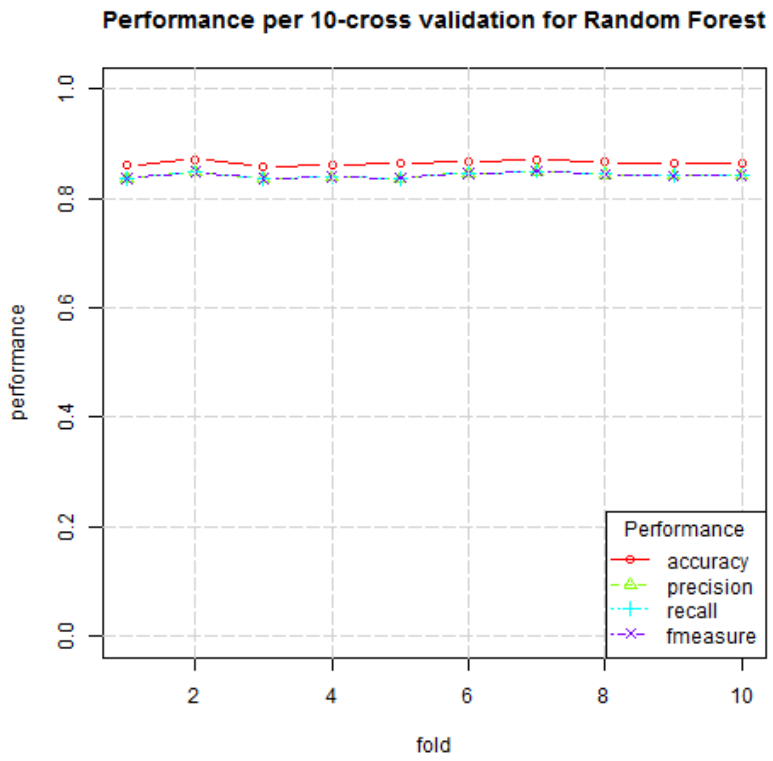
Attesa media	21.2 min
RMSE	5.9 min



10-fold cross validation

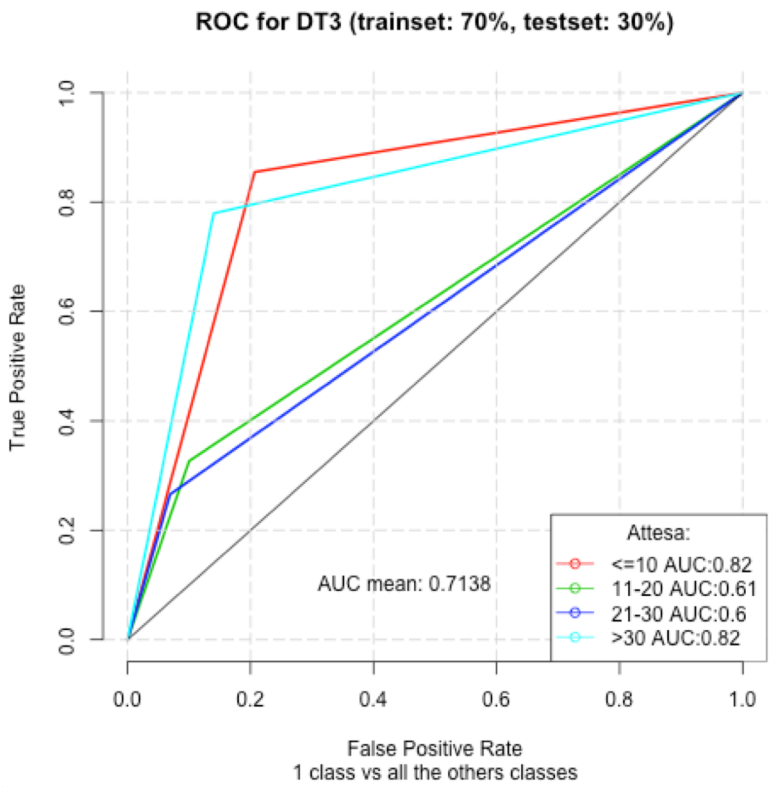


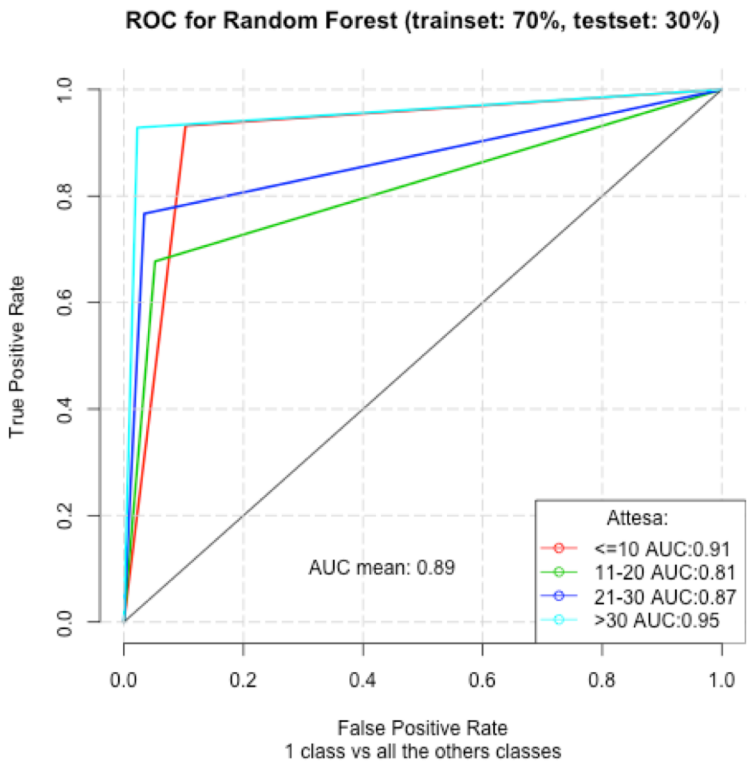

```
> mean_accuracy
[1] 0.6174974
> mean_precision
[1] 0.8415008
> mean_recalls
[1] 0.6917161
> mean_fmeasure
[1] 0.7543822
```




```
> mean(accuracies_res)
[1] 0.865029
> mean(precisions_res)
[1] 0.8424242
> mean(recalls_res)
[1] 0.8424242
> mean(fmeasures_res)
[1] 0.8424242
>
```

Performance





Conclusioni

- Il lavoro di miglioramento della qualità dei dati è stato molto utile per eliminare valori che avrebbero dato problemi nella fase di predizione
- La stima dei parametri è stata fondamentale per ottenere delle buone features da sfruttare dei modelli di machine learning

GRAZIE PER L’ATTEZIONE