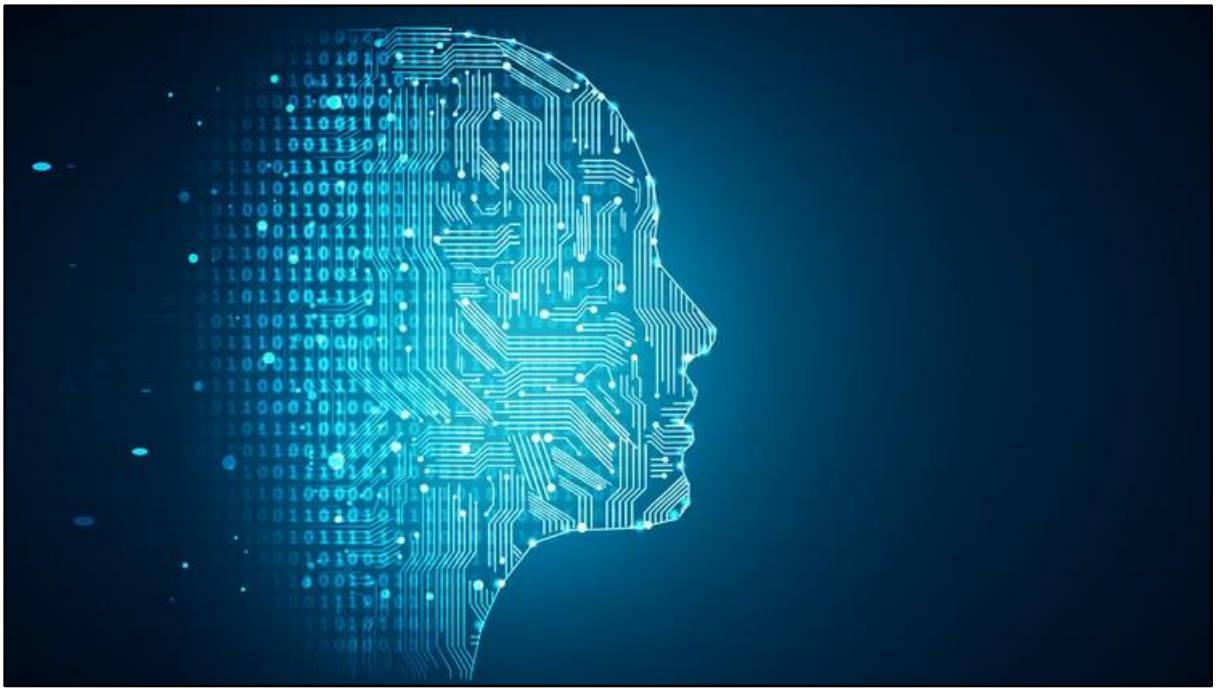
PROGETTO DATA TECHNOLOGY E MACHINE LEARNING



Anno accademico 2018/2019

Virgilio Luca 794866

Ventura Samuele 793060

Sommario

1 OBIETTIVI DEL PROGETTO 3

2 DATA TECHNOLOGY 4

2.1 RACCOLTA DATI INIZIALI 4
2.2 DESCRIZIONE DEI DATI INIZIALI 5
2.2.1 Tu_Passi_2018_per_andamento_servizio 5
2.2.2 Dati_sensori_meteo 6
2.2.3 Feste 7
2.3 DATA QUALITY 8
2.3.1 Tu_Passi_2018 DQ 9
2.3.2 Dati_sensori_meteo DQ 10
2.3.3 Feste DQ 10
2.4 DATA CLEANING 10
2.5 DATA INTEGRATION 13
2.6 ANALISI DESCRITTIVA 15

3 MACHINE LEARNING 22

3.1 CALCOLO PARAMETRI 22
3.2 FEATURES SELECTION 23
3.3 ANALISI DESCRITTIVA TRAINING SET 27
3.4 SCELTA DEI MODELLI 31
3.4.1 Decision Tree 32
3.4.2 Random Forest 36
3.5 10-FOLD CROSS VALIDATION 41
3.6 RISULTATI 44

4 CONCLUSIONI 45

1 OBIETTIVI DEL PROGETTO

L'ambito della sanità è un argomento molto delicato, poiché viene coinvolta la sfera emotiva di una persona. Tra i diversi problemi nell'ambito della sanità, l'attesa dei pazienti è un argomento che riscuote una certa attenzione. Con il passare del tempo gli ospedali hanno iniziato ad occuparsi del problema dell'attesa, introducendo display e codici nel tentativo di migliorare la fase dell'accettazione. Secondo quanto affermato da Maister, la soddisfazione di un servizio dipende dalla percezione e dall'aspettativa dello stesso. Per questo motivo una percezione del servizio è determinante per una migliore soddisfazione da parte dei pazienti. Inoltre un'attesa incerta risulta più lunga rispetto ad un'attesa certa[1]. Per queste ragioni abbiamo deciso di costruire un modello predittivo per stimare il tempo d'attesa nella fase di accettazione.

Questo lavoro è stato reso possibile grazie alla collaborazione con il professor Cabitza e con l'ospedale Galeazzi di Milano. Essendo un dominio molto specifico, l'obiettivo di questo lavoro è quello di predire il tempo di attesa agli sportelli, durante la fase di accettazione, presso l'ospedale Galeazzi di Milano.

[1]Maister, David H. *The psychology of waiting lines*. Boston, MA: Harvard Business School, 1984.

2 DATA TECHNOLOGY

2.1 Raccolta dati iniziali

Il dataset relativo alla fase di accettazione presso gli sportelli dell'ospedale Galeazzi di Milano è stato ottenuto contattando direttamente il reparto IT dell'ospedale stesso per due motivi principali:

1. essendo un dominio molto specifico, riguardo un singolo ospedale, era necessario contattare chi possedeva i dati a noi utili

2. ragioni di privacy sui pazienti e sull'ospedale stesso

In questo modo siamo riusciti ad ottenere un dataset contenente le informazioni a noi necessarie:

- **Tupassi_2018_per_andamento_servizio:** dataset contenente varie informazioni registrate dal sistema ospedaliero

Come detto precedentemente essendo un problema relativo ad un dominio molto specifico è stato difficoltoso cercare altri dataset contenenti informazioni utili da poter sfruttare per risolvere il problema, nonostante questo si è deciso di cercare dei dati relativi alle condizioni climatiche nella città di Milano:

- **Dati_sensori_meteo:** dataset che rappresenta i valori registrati dai vari sensori

Oltre a questi dati si è pensato che potesse essere utile aggiungere informazioni rilevanti la vicinanza di una data ad un giorno festivo, si è quindi creato un dataset per rappresentare giorni in procinto o successivi a festività:

- **Feste:** dataset che rappresenta giorni in procinto di festività

I dataset originali, come sono stati scaricati, sono salvati nella cartella:

ProgettoML&DT/Dataset/Originali

2.2 Descrizione dei dati iniziali

In questa sezione saranno analizzati i dataset originali e descritte le informazioni al loro interno.

2.2.1 Tu_Passi_2018_per_andamento_servizio

Questo è il dataset principale del progetto e rappresenta le informazioni registrate in modo automatico dal sistema TuPassi dell'ospedale Galeazzi di Milano nella fase di accettazione agli sportelli.
Il dataset contiene 347647 righe, una per ogni paziente che ha effettuato l'accettazione nel corso dell'anno 2018. Esistono due modalità per prenotare una richiesta degli sportelli:

- tramite il totem, presente in ospedale
- tramite applicazione TuPassi

Ogni riga è descritta con 16 attributi, elencati nella seguente tabella:

Attributo	Tipo	Descrizione	Formato
Giorno	Data	Indica la data in cui si presenta il paziente	dd/mm/yy
Sede	Stringa	Indica la sede in cui si presta la visita	//
Servizio*	Stringa	Indica il tipo di servizio ospedaliero	//
Appuntamento**	Ora	Indica la stima dell'ora in cui si verrà chiamati	hh:mm:ss
Presente alle ore	Ora	Indica l'ora in cui si ritira il biglietto al totem	hh:mm:ss
Chiamato alle ore	Ora	Indica l'ora in cui si è effettivamente chiamati allo sportello	hh:mm:ss
Chiamato da operatore numero	Intero	Indica l'operatore da cui si è stati chiamati	//
Sportello chiamata	Intero	Indica lo sportello da cui si è stati chiamati	range [1-12]
Prenotato il	Data e Ora	Indica il giorno e l'ora in cui si è prenotato	dd:mm:yy hh:mm:ss
Prenotato	Booleano	Indica se si è prenotato	//
Attesa	Ora	Indica l'attesa rispetto alla predizione del totem	+- mm:ss

Attesa in sec	Intero	Indica l'attesa rispetto alla predizione del totem in secondi	+- ss
Stato	Stringa	Indica se il servizio è stato effettuato oppure altri casi	range[LAVORATO, CHIAMATO, REVOCATO, ANTICIPATO]***
Operatore numero	Intero	Indica l'operatore da cui si è stati chiamati	//
Sportello	Intero	Indica lo sportello da cui si è stati chiamati	range [1-12]
Ultima operazione alle	Ora	Indica l'ultima operazione effettuata dall'operatore e registrata dal sistema	hh:mm:ss

* Servizio può assumere uno tra questi 23 valori:
['ACCETTAZIONE ASL', 'ACCETTAZIONE VISITE PRIVATE', 'FISIOTERAPIA', 'ODONTOIATRIA SOLVENTI/FONDI', 'ODONTOIATRIA SSN', 'PRELIEVI ASL,' 'PRELIEVI SOLVENTI', 'PRENOTAZIONI VISITE PRIVATE', 'PRENOTAZIONI', 'PRENOTAZIONI ODONTOIATRIA', 'PRIORITÀ', 'PRIORITÀ LABORATORIO ANALISI', 'PRIORITÀ ODONTOIATRIA', 'PRIORITÀ PRELIEVI', 'PRIORITÀ RADIOLOGIA', 'PRIORITÀ VISITE', 'RADIOLOGIA ASL', 'RITIRO ESAMI', 'SOLVENTI', 'SOLVENTI LABORATORIO ANALISI', 'SOLVENTI/TARIFFA AGEVOLATA', 'SSN LABORATORIO ANALISI', 'VISITE/ESAMI ASL']

** Appuntamento indica la predizione del totem con il quale ci si registra in coda di attesa agli sportelli

*** Per quanto riguarda l'attributo Stato il significato dei possibili valori è il seguente:

- LAVORATO: è stato chiamato e si è presentato allo sportello
- CHIAMATO: è stato chiamato ma non si è presentato allo sportello
- REVOCATO: è stata annullata l'operazione
- ANTICIPATO: non è stato definito il significato del valore per mancanza di informazioni aggiuntive

2.2.2 Dati_sensori_meteo

Inizialmente si era trovato un dataset contenente informazioni generali su ogni stazione metereologica lombarda, e sui sensori installati. Da questo dataset si sarebbero poi estratti i sensori riguardanti le precipitazioni, confrontando questo con il dataset di tutte le rilevazioni

dei sensori lombardi si sarebbero ottenute solo le rilevazioni delle precipitazioni. Su questo infine si sarebbero dovuto filtrare le righe in modo da tenere solo i sensori nelle vicinanze dell'ospedale Galeazzi.

Questa parte di filtraggio e data management non è stata necessaria perché al seguente indirizzo Stazioni meteo lombarde, è stato facilmente possibili individuare la stazione più vicina all'ospedale e filtrare i dati direttamente con il numero del sensore individuato.

Il dataset da cui si sono estratte le informazioni sulle precipitazioni è stato scaricato dal sito dell'arpa, al seguente link Richiesta Dati ARPA, dal quale è possibile selezionare provincia, tipo di rilevazione, stazione metereologica e intervallo temporale.

I dati sono inviati via email, come csv o pdf, e sono descritti nella seguente tabella:

Attributo	Tipo	Descrizione	Formato
Id Sensore	Intero	ID univoco del sensore	//
Data-Ora	Data e Ora	Indica la data e l'ora della rilevazione	yyyy/mm/dd hh:mm:ss
Valore Cumulato	Double	Indica la misura della rilevazione	//

Il dataset contiene circa 52418 righe, tutte riguardanti un singolo sensore di precipitazioni (il più vicino all'ospedale in esame) e la rilevazione è registrata ogni 10 minuti.

2.2.3 Feste

Il dataset Feste è stato creato a mano (precisare la scelta delle feste)

Attributo	Tipo	Descrizione	Formato
Giorno	Data	Giorno dell'anno	dd/mm/yy
Festivo	Booleano	Festività o vicino ad una festività (prima o dopo 2 giorni)	//

2.3 Data quality

Dopo aver individuato e scaricato/creato i dataset contenenti le informazioni che secondo noi potessero essere utili per il nostro obiettivo è stato necessario verificare la qualità dei dati in nostro possesso, così da poter individuare possibili debolezze e/o incompletezza delle informazioni e definire una strategia per poterne aumentare la qualità.

Per verificare la qualità dei dati si è utilizzato il software Talend, nello specifico il il tool Talend Open Studio for Data Quality. Grazie a questo software si possono impostare delle analisi statistiche più o meno complesse a livello di colonna o tabella del file letto; inoltre è possibili eseguire analisi di frequenza dei dati e di particolari pattern per verificare la distribuzione delle informazioni.
Si è deciso di misurare le seguenti dimensioni di qualità:

- Completezza
- Consistenza
- Unicità

Per ogni dimensione si è decisa una metrica da utilizzare nella misurazione:

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

Nel seguito si mostreranno i risultati delle sui singoli dataset delle misure di qualità proposte.

2.3.1 Tu_Passi_2018 DQ

Attributo	Completezza su attributo
Giorno	1
Sede	1
Servizio	(347647 - 1050)/347647 = 0.99
Appuntamento	(347647 - 31985)/347647 = 0.90
Presente alle ore	(347647 - 1255)/347647 = 0.99
Chiamato alle ore	1
Chiamato da operatore numero	(347647 - 4229)/347647 = 0.98
Sportello chiamata	(347647 - 4229 - 1046)/347647 = 0.98
Prenotato il	(347647 - 34002)/347647 = 0.90
Prenotato	(347647 - 33244)/347647 = 0.90
Attesa	(347647 - 288)/347647 = 0.99
Attesa in sec	1
Stato	1
Operatore numero	1
Sportello	(347647 - 3556)/347647 = 0.98
Ultima operazione alle	(347647 - 3612)/347647 = 0.98

In media quindi abbiamo una *completezza per attributo* di 0.97, come viene confermato dalla *completezza a livello di tabella* mostrato di seguito:

Tabella	Completezza su tabella
Tu_passi_2018	5443856/5562352 = 0.97

Per quanto riguarda la *consistenza* si sono definiti i seguenti vincoli interni alla tabella:

- Tu_passi_2018.Chiamato_alle_ore > Tu_passi_2018.Presente_alle_ore
- Tu_passi_2018.Ultima_operazione_alle > Tu_passi_2018.Chiamato_alle_ore

Vincolo	Consistenza
Chiamato_alle_ore > Presente_alle_ore && Ultima_operazione_alle > Chiamato_alle_ore	(347647 - 15000 - 1255 - 3612) / 347647 = 0.94

Tabella	Unicità delle tuple
Tu_Passi	1

2.3.2 Dati_sensori_meteo DQ

Il dataset non presenta valori null o empty, o celle duplicate.

2.3.3 Feste DQ

Il dataset essendo stato creato a mano, ed avendo una struttura molto semplice, non presenta problemi di data quality.

2.4 Data cleaning

Per migliorare la qualità dei dati è stato utilizzato il tool Talend Opend Studio for Data Integration, con il quale sono state impostate delle pipeline di processing che saranno mostrate in seguito.

L'intero dataset è stato diviso in due subset sulla base dei seguenti vincoli:

- assenza di valori empty o not define, che chiameremo D1
- presenza di valori empty e not define, che chiameremo D2

Per entrambi i subset si sono eseguiti i seguenti controlli:

a. verificare che i valori nella colonna *Giorno* fossero formattati secondo il pattern [dd/mm/yy] e che fossero delle date reali (corrispondenza mese giorno)

b. verificare che i valori nella colonna *Appuntamento* fossero formattati secondo il pattern [hh:mm:ss] e che fossero delle ore reali secondo il pattern [0-24]:[0-59][0-59]

c. per quanto riguarda la colonna *Operatore_Numero*, dopo un'analisi di qualità iniziale sui pattern presenti si sono identificati i seguenti valori anomali [mail, TUPASSITO, TUPASSITO1, TUPASSITO2, IOGTUPASSI3] nonostante la colonna dovesse avere solo valori interi. Sono stati quindi scartati i record con questi valori e gestiti successivamente

d. nella colonna *Sportello_chiamata* sono stati identificati i seguenti valori anomali [120, 7\]

e. verificare che i valori nella colonna *Prenotato_il* fossero formattati secondo il pattern [dd-mm-yyyy hh:mm:ss] e che fossero delle date reali

f. verificare che i valori della colonna *Prenotato* siano in ["Si", "No"]

Per quanto riguarda D1 sono stati identificati 5 record che non rispettano il vincolo d e si è deciso di modificare i due valori in questo modo: 120 -> 12, 7\ -> 7

Per quanto riguarda D2 dopo aver eseguito i controlli elencati in precedenza si sono ottenuti due ulteriori subset, D2.1 e D2.2, i quali sono stati trattati in modo distinto:

- D2.1: questo subset non presenta anomalie rispetto ai controlli precedenti però in alcune righe (567) si è identificata l'assenza di valori per le colonne *Chiamato_da_operatore_numero* e *Sportello_chiamata*, il problema di completezza è stato risolto copiando i valori presenti nella colonne *Operatore_numero* e *Sportello,* sapendo a priori che le coppie di colonne rappresentavano lo stesso attributo. E' stato quindi molto importante non eliminare inizialmente una delle due coppie di colonne perché in tal caso si sarebbero perse informazioni, in quanto si sarebbe dovuto eliminare le righe con valori empty.

- D2.2: in questo subset si sono rilevati deficit maggiori di Data Quality.

  Per prima cosa è stato necessario eliminare 1255 records in cui non era presente il valore nella colonna Presente_alle_ore, in quanto fondamentale per la nostra successiva fase di predizione.

  Analizzando le righe rimanenti si sono identificati 3612 record con valore Stato == REVOCATO ed è stato necessario eliminarli perché non sfruttati nella successiva fase di predizione.

Sono state eliminate le righe in cui il valore nella colonna *Uiltima_operazione_alle* era vuoto, così da poter verificare migliorare la qualità per quanto riguarda la consistenza.

Su alcune colonne si sono identificati valori sfalsati, come per il subset D2.1 è stato quindi necessario riscrivere i valori esatti.

Al termine della fase di data cleaning è stato necessario eliminare 19867 righe dalla tabella, quindi si è perso circa il 6% delle informazioni inizialmente contenute.

Analisi di qualità dopo data cleaning

Vincolo	Consistenza
Chiamato_alle_ore > Presente_alle_ore && Ultima_operazione_alle > Chiamato_alle_ore	1

Tabella	Unicità delle tuple
Tu_Passi	1

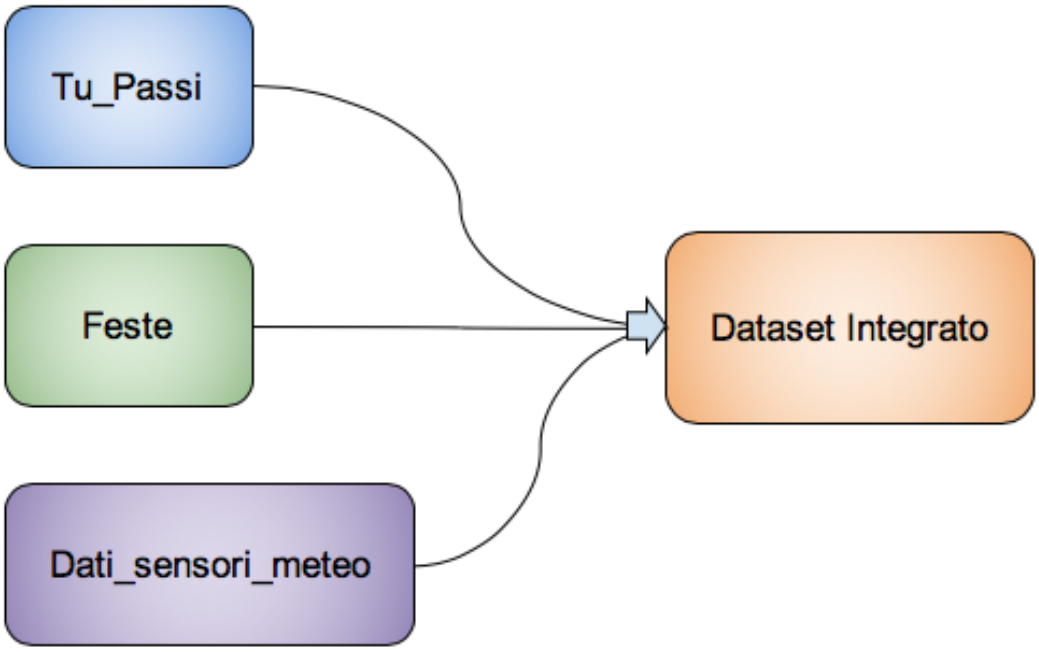
Attributo	Completezza su attributo
Giorno	1
Sede	1
Servizio	1
Appuntamento	1
Presente alle ore	1
Chiamato alle ore	1
Chiamato da operatore numero	1
Sportello chiamata	1
Prenotato il	1

Prenotato	(347647 - 33035) / 347647 = 0.90
Attesa	1
Attesa in sec	1
Stato	1
Operatore numero	1
Sportello	1
Festivo *	1
mediaP *	1

* Nel dataset integrato sono stati aggiunti i dati relativi ai giorni festivi e alla media delle precipitazioni per fascia oraria.

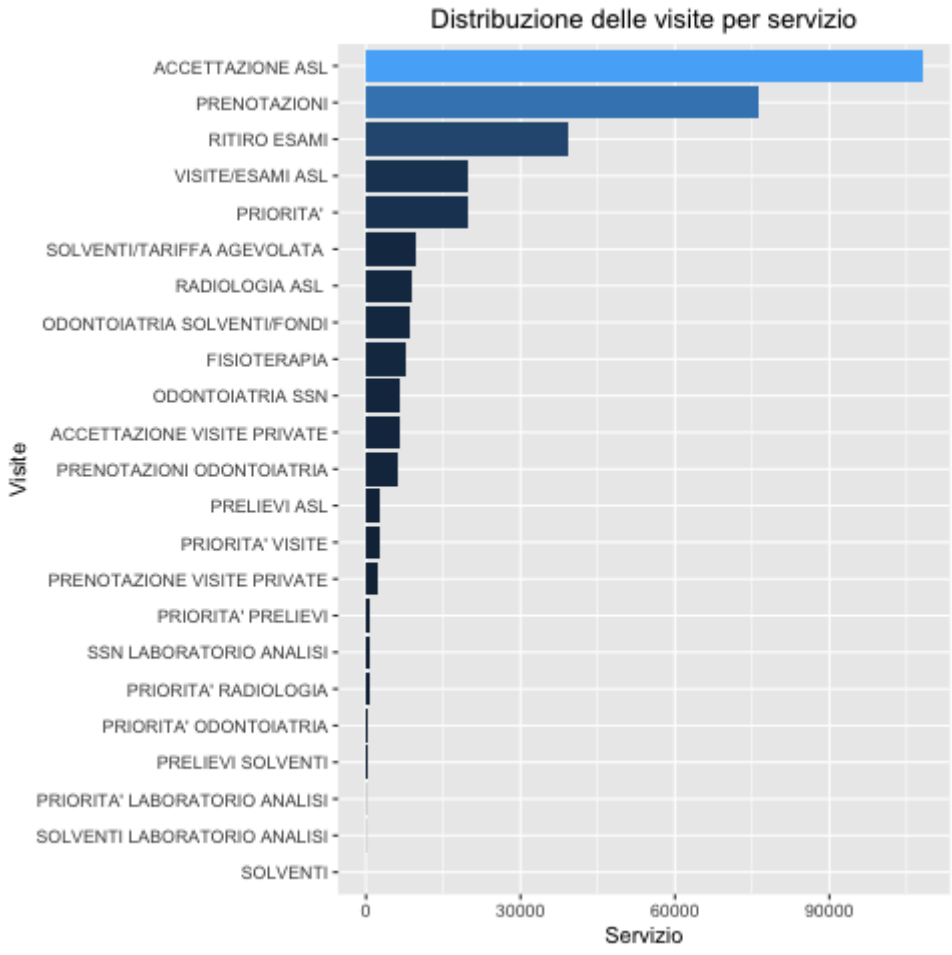
2.5 Data integration

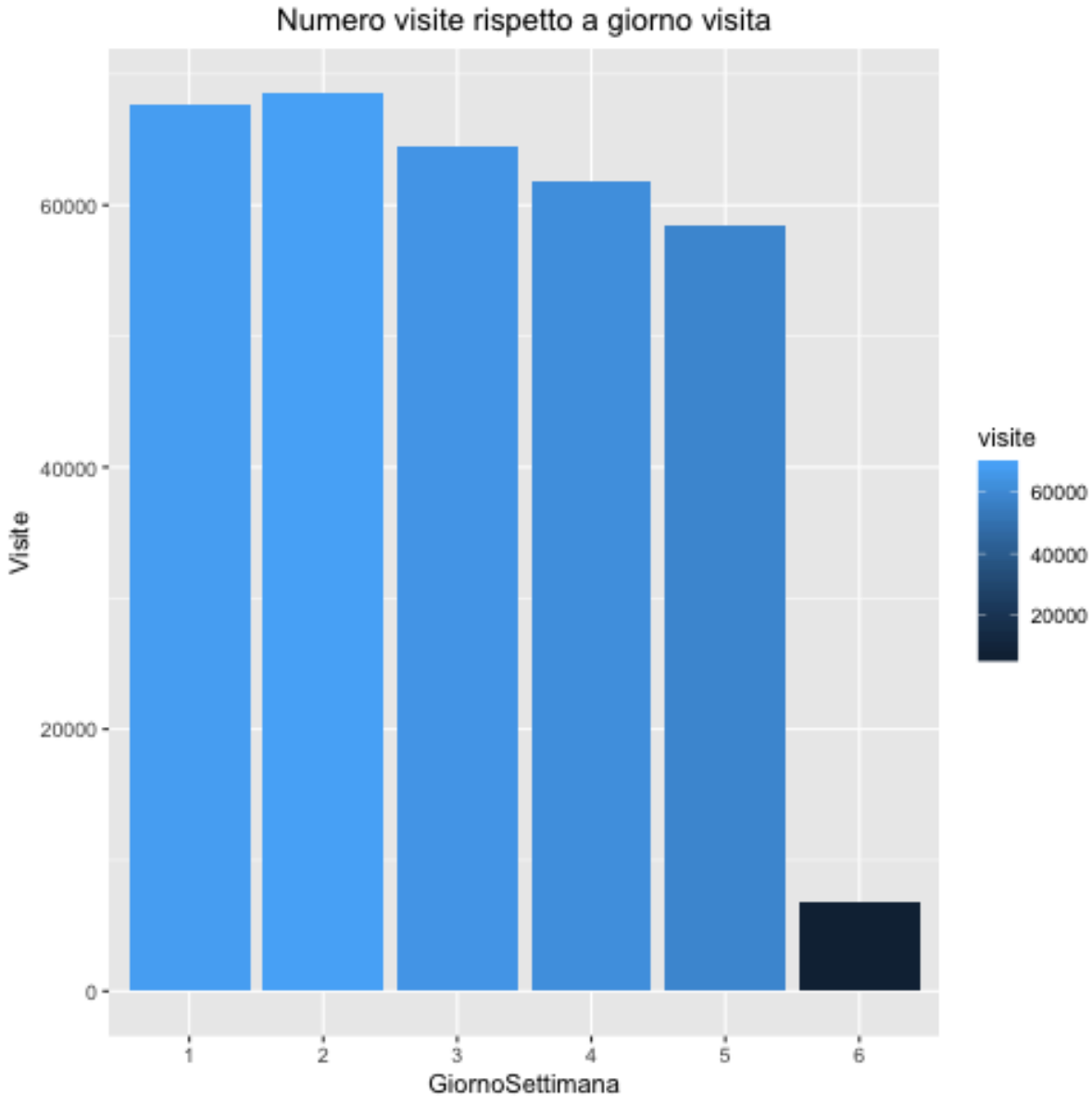
Il dataset finale è il consolidamento dei tre dataset di partenza, in quanto rappresentano diversi realtà, sulla base della data e dell'ora dei pazienti presenti nel dataset Tu_Passi.

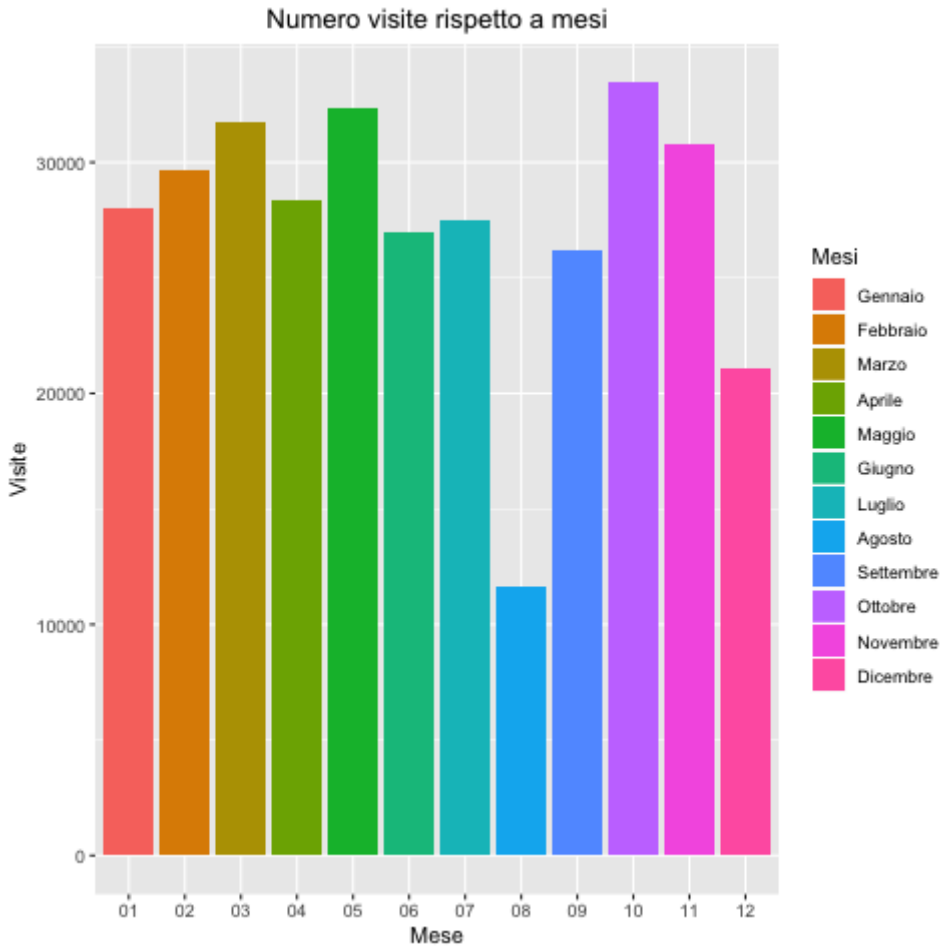

Tu_Passi
Feste
Dati_sensori_meteo
Dataset Integrato

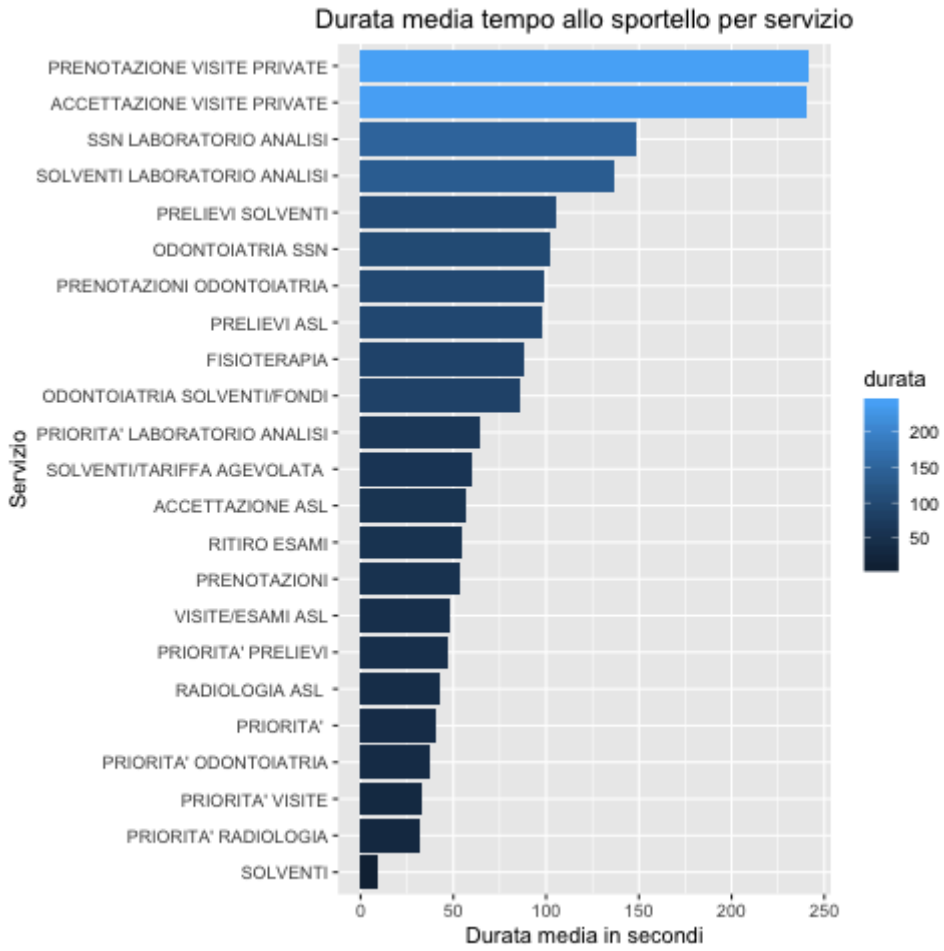
2.6 Analisi Descrittiva

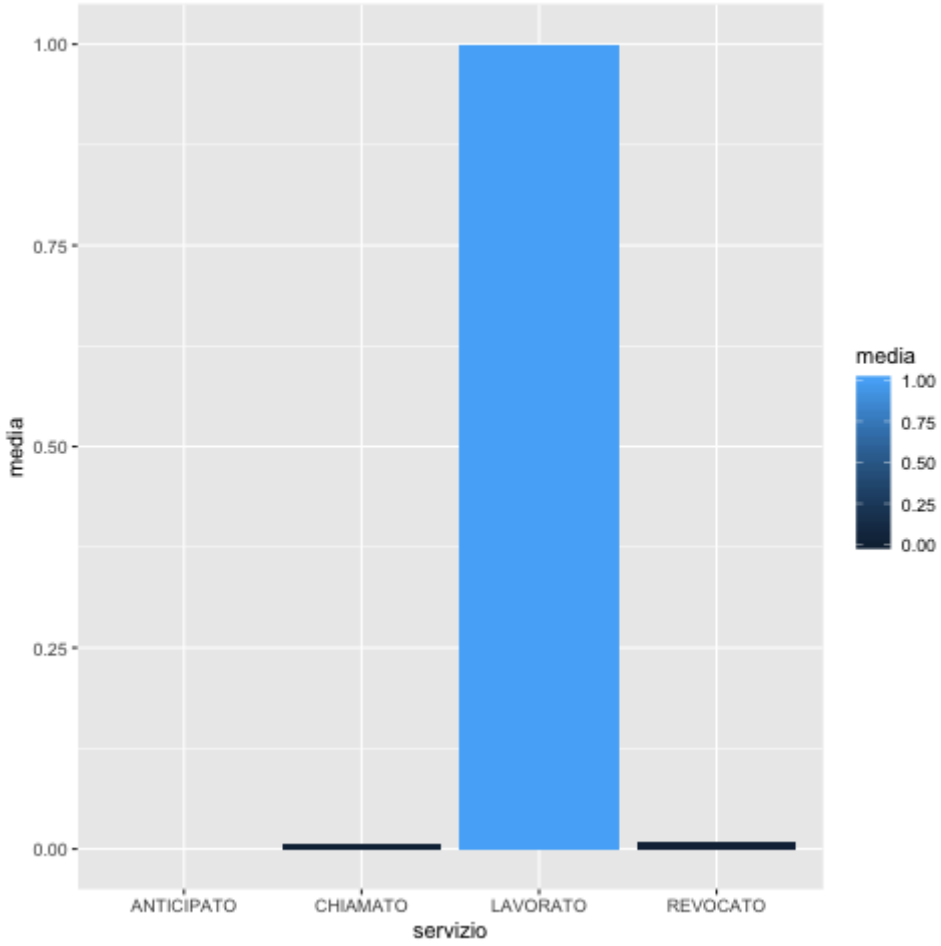
Di seguito sono mostrati alcuni grafici che descrivono il dataset integrato

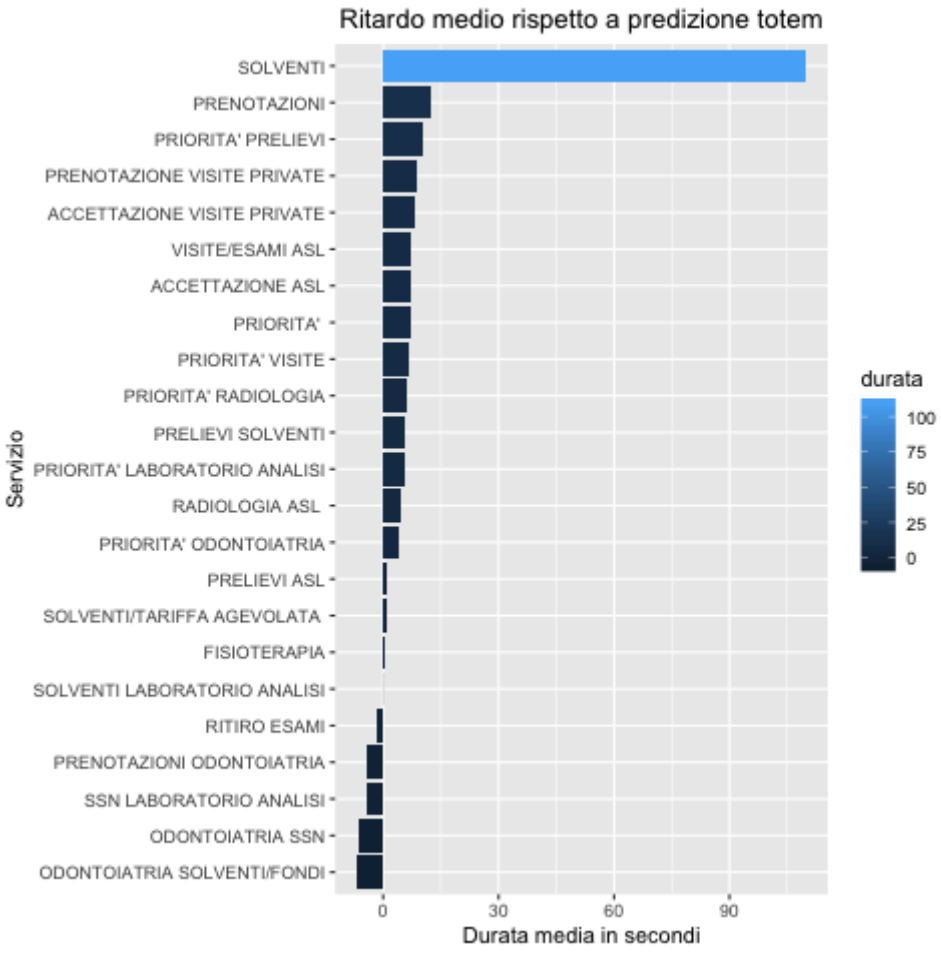


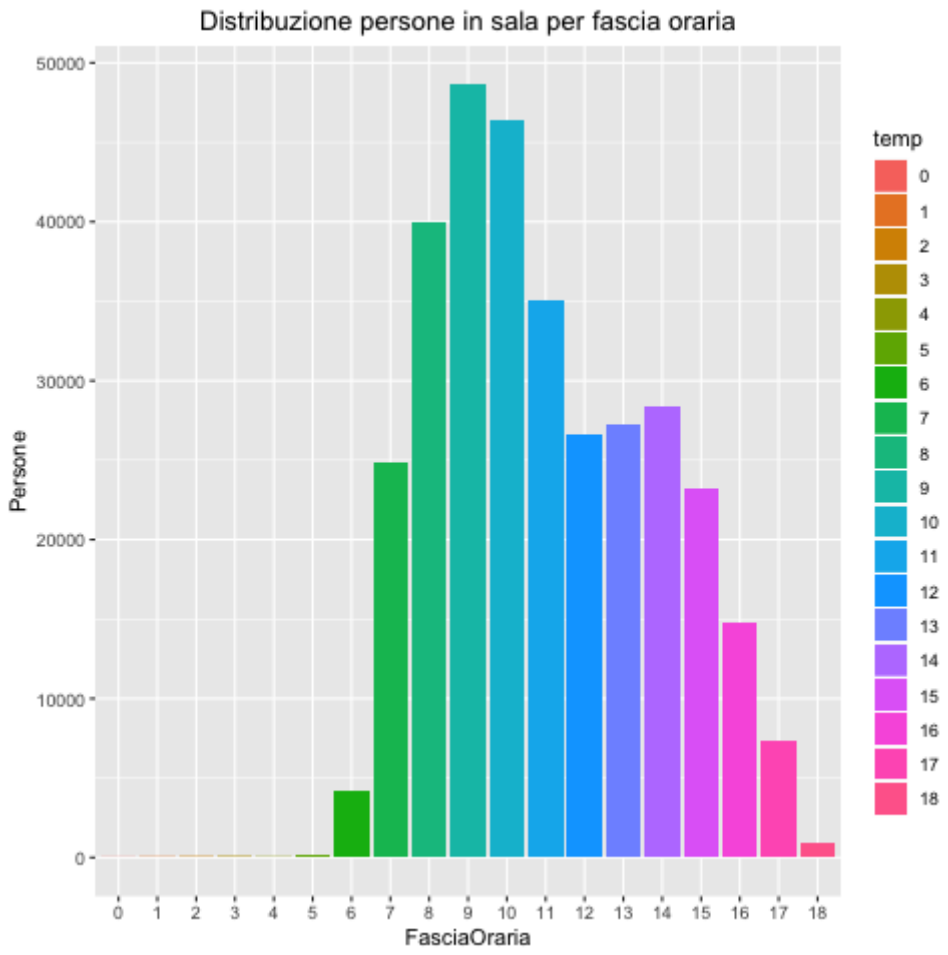

Numero visite rispetto a giorno visita
Visite
60000
40000
20000
0
1
2
3
4
5
6
GiornoSettimana
visite
60000
40000
20000


Numero visite rispetto a mesi
Visite
30000
20000
10000
0
01
02
03
04
05
06
07
08
09
10
11
12
Mese
Mesi
Gennaio
Febbraio
Marzo
Aprile
Maggio
Giugno
Luglio
Agosto
Settembre
Ottobre
Novembre
Dicembre


Durata media tempo allo sportello per servizio
PRENOTAZIONE VISITE PRIVATE
ACCETTAZIONE VISITE PRIVATE
SSN LABORATORIO ANALISI
SOLVENTI LABORATORIO ANALISI
PRELIEVI SOLVENTI
ODONTOIATRIA SSN
PRENOTAZIONI ODONTOIATRIA
PRELIEVI ASL
FISIOTERAPIA
ODONTOIATRIA SOLVENTI/FONDI
PRIORITA' LABORATORIO ANALISI
SOLVENTI/TARIFFA AGEVOLATA
ACCETTAZIONE ASL
RITIRO ESAMI
PRENOTAZIONI
VISITE/ESAMI ASL
PRIORITA' PRELIEVI
RADIOLOGIA ASL
PRIORITA'
PRIORITA' ODONTOIATRIA
PRIORITA' VISITE
PRIORITA' RADIOLOGIA
SOLVENTI
Servizio
0
50
100
150
200
250
Durata media in secondi
durata
200
150
100
50


1.00
0.75
0.50
0.25
0.00
media
ANTICIPATO
CHIAMATO
LAVORATO
REVOCATO
servizio
media
1.00
0.75
0.50
0.25
0.00


Ritardo medio rispetto a predizione totem
SOLVENTI
PRENOTAZIONI
PRIORITA' PRELIEVI
PRENOTAZIONE VISITE PRIVATE
ACCETTAZIONE VISITE PRIVATE
VISITE/ESAMI ASL
ACCETTAZIONE ASL
PRIORITA'
PRIORITA' VISITE
PRIORITA' RADIOLOGIA
PRELIEVI SOLVENTI
PRIORITA' LABORATORIO ANALISI
RADIOLOGIA ASL
PRIORITA' ODONTOIATRIA
PRELIEVI ASL
SOLVENTI/TARIFFA AGEVOLATA
FISIOTERAPIA
SOLVENTI LABORATORIO ANALISI
RITIRO ESAMI
PRENOTAZIONI ODONTOIATRIA
SSN LABORATORIO ANALISI
ODONTOIATRIA SSN
ODONTOIATRIA SOLVENTI/FONDI
Servizio
0
30
60
90
Durata media in secondi
durata
100
75
50
25
0


Distribuzione persone in sala per fascia oraria
Persone
50000
40000
30000
20000
10000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
FasciaOraria
temp
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

3 MACHINE LEARNING

3.1 Calcolo parametri

Dato che il sistema TuPassi non calcola l'attesa effettiva, cioè il tempo intercorso dal momento dell'arrivo del paziente alla chiamata dello stesso da parte di uno sportello, ci siamo adoperati a calcolarla. Questa sarà la nostra variabile target per quanto riguarda i modelli predittivi.

Il dataset in nostro possesso non forniva di base molte informazioni utili per un modello predittivo, è stato necessario calcolare nuovi attributi basandoci sui dati in nostro possesso, in modo da esplicitare informazioni che prima non lo erano e che secondo noi potessero essere utili nella fase di training del modello.

Attraverso l'attributo che indica l'orario dell'ultima operazione effettuata da un operatore ad uno sportello, si è stimata la durata del servizio allo sportello.

Analizzando il dominio di riferimento, abbiamo deciso di stimare il numero di pazienti presenti in sala e il numero di sportelli attivi.

- Per stimare il numero di pazienti presenti in sala è stato sufficiente calcolare, dato l'orario di registrazione di un paziente, il numero di pazienti che si erano già registrati presso il totem ma che non erano ancora stati chiamati da alcuno sportello in quell'orario.

- Per stimare il numero di sportelli attivi abbiamo discretizzato in ore l'orario in cui i diversi pazienti vengono chiamati e ogni giorno, per ogni ora abbiamo contato il numero di sportelli che hanno effettuato almeno una chiamata.

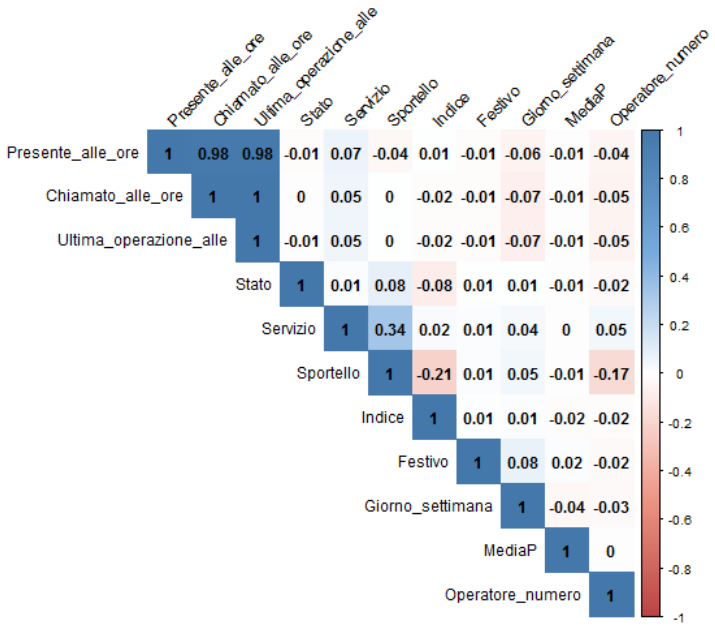
Durante la ricerca esplorativa su eventuali lavori svolti in questo ambito, abbiamo trovato due articoli pubblicati sul Journal of the American College of Radiology. In questi articoli, i ricercatori si occupano di predire il tempo di attesa dei pazienti nel reparto di radiologia del Massachusetts General hospital. Per predire il tempo d'attesa vengono utilizzati modelli statistici (regressione) e di ML.
Grazie ad un dominio applicativo molto simile, abbiamo deciso di calcolare dei parametri utilizzati anche nei loro modelli:

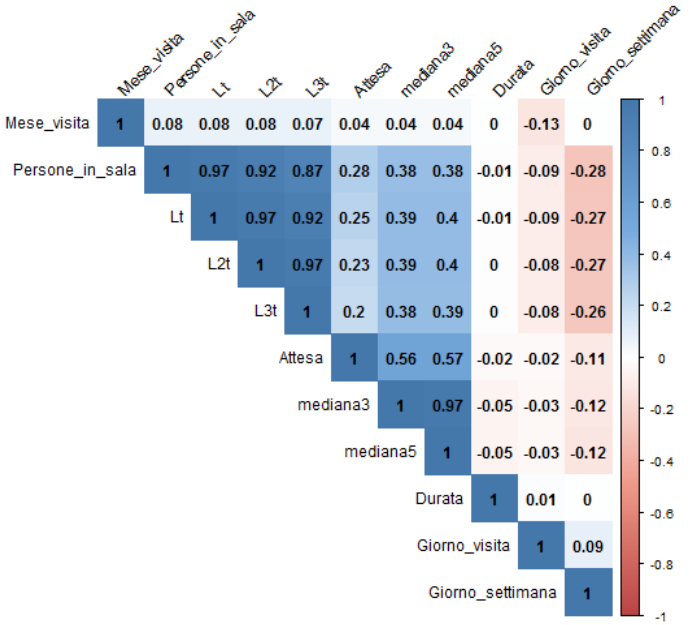
- numero di pazienti presenti in sala nei 10, 20 e 30 minuti precedenti
- media e mediana del tempo di attesa degli ultimi 3 e 5 pazienti, suddivisi per servizio
- numero di pazienti serviti nei 10 minuti precedenti dallo stesso sportello, numero di pazienti serviti nei 10 minuti precedenti da tutti gli sportelli
- numero di pazienti serviti nei 10 minuti precedenti con lo stesso servizio e numero di pazienti serviti nei 10 minuti precedenti con lo stesso servizio da tutti gli sportelli
- ordinamento di pazienti per giorno
- ordinamento di pazienti per servizio

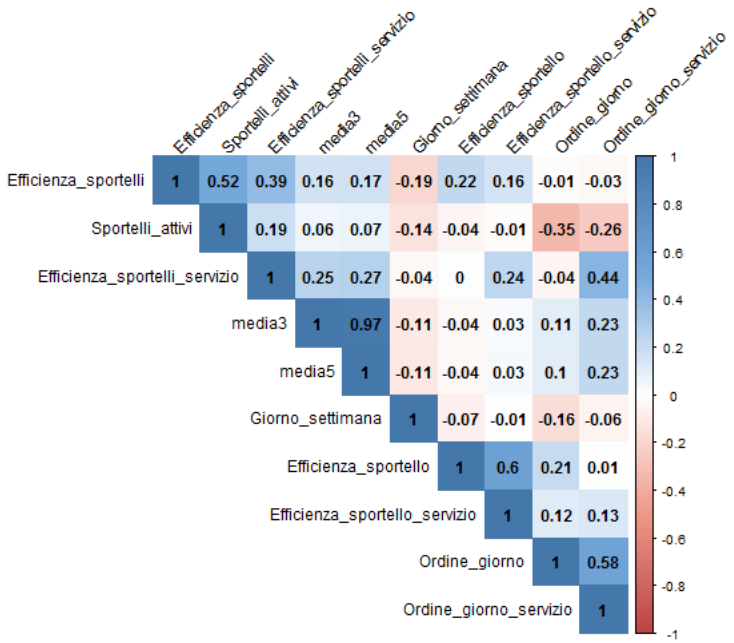
3.2 Features selection

Prima di poter lavorare realmente sui modelli di machine learning è stata necessaria una fase di features selection, così da capire quali attributi potessero essere utili nei modelli di predizione e quali invece si potessero eliminare a priori.

Utilizzando ogni volta vari subset di attributi (per aumentare la leggibilità), ogni volta tenendo la colonna target, sono state calcolate le seguenti matrici di correlazione:







Prima di addestrare un qualsiasi modello, abbiamo escluso tutte quelle features che non potrebbero essere utilizzate per la predizione del tempo di attesa in un sistema in real-time (es: non si può conoscere a priori l'ora a cui si verrà chiamati, oppure lo sportello).

Inoltre, abbiamo escluso alcune features molto correlate tra di loro, perché non aggiungono effettivamente "conoscenza" al modello. Per esempio, media del tempo di attesa degli ultimi 5 pazienti è stata preferita a media degli ultimi 3 pazienti

Quindi il nostro dataset si è ridotto a 17 caratteristiche, elencate qui sotto:

Attributo	Tipo	Descrizione
Giorno_visita	Intero	Indica il giorno in cui si presenta il paziente
Mese_visita	Intero	Indica il mese in cui si presenta il paziente
Giorno_settimana	Intero (1-6)	Indica il giorno della settimana (lun-sab)
Servizio	intero	indica il tipo di servizio ospedaliero
Presente_alle_ore	Intero	Indica la fascia oraria in cui si ritira il biglietto al totem
media5	Double	indica la media del tempo di attesa dei 5 pazienti precedenti con lo stesso servizio
mediana5	Double	indica la mediana del tempo di attesa dei 5 pazienti precedenti con lo stesso servizio

Festivo	Booleano	Festività o vicino ad una festività (prima o dopo 2 giorni)
mediaP	Double	indica il valore medio di precipitazioni accumulate nella fascia oraria in cui è arrivato il paziente
Persone_in_sala	Intero	numero di persone registrate dal sistema e non ancora servite
Lt	Intero	numero di persone in sala registrate nei 10 minuti precedenti
Sportelli_attivi	Intero	stima del numero di sportelli attivi durante il tempo d'attesa del paziente
Efficienza_sportelli	Intero	numero di pazienti accettati da tutti gli sportelli negli ultimi 10 minuti
Efficienza_sportelli_servizio	Intero	numero di pazienti con lo stesso servizio accettati da tutti gli sportelli negli ultimi 10 minuti
Ordine_giorno	Intero	Ordinamento giornaliero dei pazienti per orario d'arrivo
Ordine_giorno_servizio	Intero	Ordinamento giornaliero dei pazienti con lo stesso servizio per orario d'arrivo
Attesa	Double	l'attesa effettiva del paziente

Prima di addestrare il modello di machine learning si è reputato opportuno eseguire un'ulteriore fase di features selection, per avere un'altra metodologia, oltre alla matrice di correlazione, per valutare l'importanza delle nostre features.

Questi sono i risultati ottenuti:

	Overall
MediaP	641.1083
Festivo	369.6652
Servizio	3117.9416
Presente_alle_ore	2313.1333
Giorno_visita	3099.4088
Mese_visita	1053.3145
Giorno_settimana	1641.8172
Persone_in_sala	5085.3480
Lt	4059.0750
mediana5	8722.6055
media5	8047.6431
Efficienza_sportelli	3370.3712
Efficienza_sportelli_servizio	2775.9525
Ordine_giorno	5439.2059
Ordine_giorno_servizio	6058.9711
Sportelli_attivi	1624.8602

Per rimuovere gli outliers, abbiamo deciso di escludere i record che registravano un valore di attesa superiore al 95 percentile, come mostrato nelle immagini seguenti così facendo si è migliorata leggermente la distribuzione dei dati.

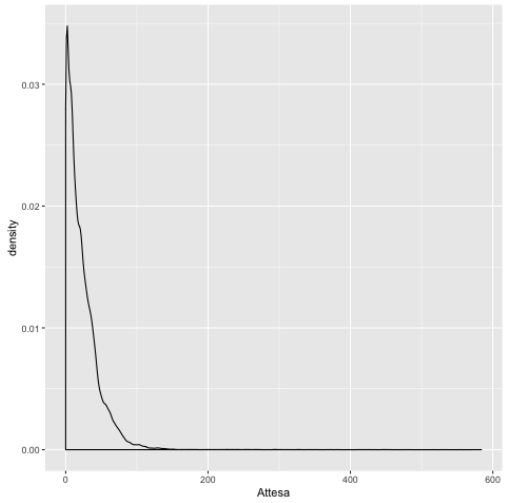


Figura 6 Distribuzione dei dati i base all'attesa

```
> perc = quantile(temp, probs = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.97, 0.99))
> perc
10% 20% 30% 40% 50% 60% 70% 80% 90% 95% 97% 99%
  1   4   8  11  16  22  28  37  51  65  75 101
```

Figura 7 Percentili del dataset sulla base dell'attesa

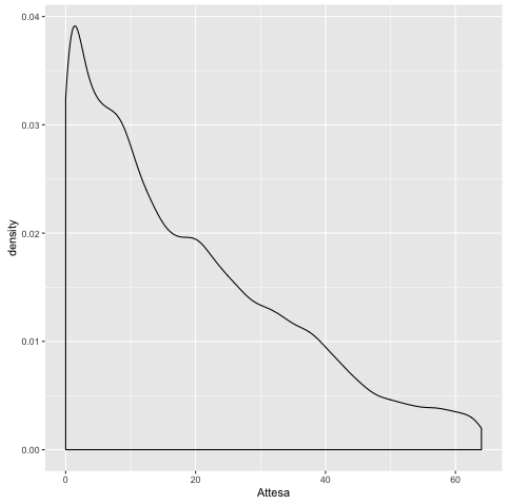


Figura 8 Distribuzione dei dati dal prima del 95 percentile

Oltre alla rimozione dei dati sulla base dell'attesa si è deciso anche di prendere come mesi di analisi solamente gli ultimi 3 dell'anno 2018, questo perché alcuni servizi sono stati attivi

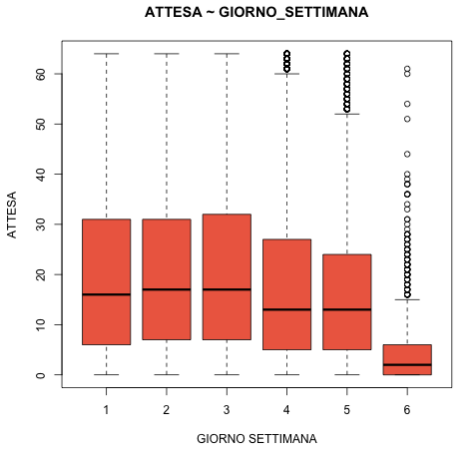
solo per alcuni mesi durante l'anno e inoltre la situazione attuale dell'ospedale Galeazzi risulta essere quella. La distribuzione dei servizi durante l'anno è mostrata nella seguente tabella:

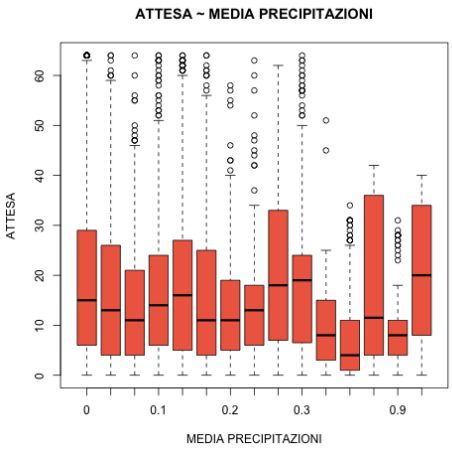
	Gennaio	Febbraio	Marzo	Aprile	Maggio	Giugno	Luglio	Agosto	Settembre	Ottobre	Novembre	Dicembre
ACCETTAZIONE ASL	0	0	10642	11854	13462	10979	11764	4963	11117	13282	11948	8018
ACCETTAZIONE VISITE PRIVATE	0	0	0	0	0	0	0	0	251	2357	2176	1808
FISIOTERAPIA	778	733	686	604	773	633	655	375	714	787	680	399
ODONTOIATRIA SOLVENTI/FONDI	688	757	858	677	876	769	802	24	761	906	863	565
ODONTOIATRIA SSN	558	599	685	600	671	576	624	15	553	757	630	408
PRELIEVI ASL	0	0	273	300	319	284	266	259	283	330	339	198
PRELIEVI SOLVENTI	0	0	15	14	22	17	21	21	23	19	16	8
PRENOTAZIONE VISITE PRIVATE	0	0	0	0	0	0	0	0	66	809	705	536
PRENOTAZIONI	7003	6966	7912	7084	7845	6722	6365	3321	5908	6693	6083	4187
PRENOTAZIONI ODONTOIATRIA	552	527	634	526	613	524	604	28	513	683	590	367
PRIORITA'	0	0	1867	2350	2680	2407	2054	619	1880	2213	2274	1409
PRIORITA' LABORATORIO ANALISI	23	40	10	0	0	0	0	0	0	0	0	0
PRIORITA' ODONTOIATRIA	0	0	0	0	0	0	24	2	51	78	57	37
PRIORITA' PRELIEVI	0	0	71	88	90	82	90	61	89	90	71	39
PRIORITA' RADIOLOGIA	223	422	69	0	0	0	0	0	0	0	0	0
PRIORITA' VISITE	808	1628	323	0	0	0	0	0	0	0	0	0
RADIOLOGIA ASL	3941	3967	1102	3	1	0	0	0	0	0	0	0
RITIRO ESAMI	3317	3717	3911	3350	3941	3202	3385	1614	3185	3530	3507	2442
SOLVENTI	0	0	0	0	0	0	0	0	8	0	0	0
SOLVENTI LABORATORIO ANALISI	26	26	3	0	0	0	0	0	0	0	0	0
SOLVENTI/TARIFFA AGEVOLATA	716	833	924	881	1020	793	859	353	815	918	825	664
SSN LABORATORIO ANALISI	292	355	72	0	0	0	0	0	0	0	0	0
VISITE/ESAMI ASL	9096	9095	1719	0	0	0	0	0	0	0	0	0

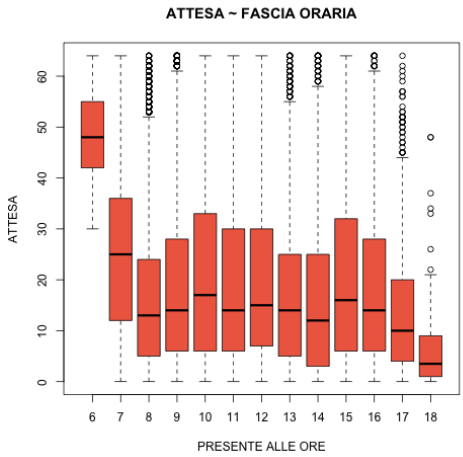
3.3 Analisi descrittiva training set

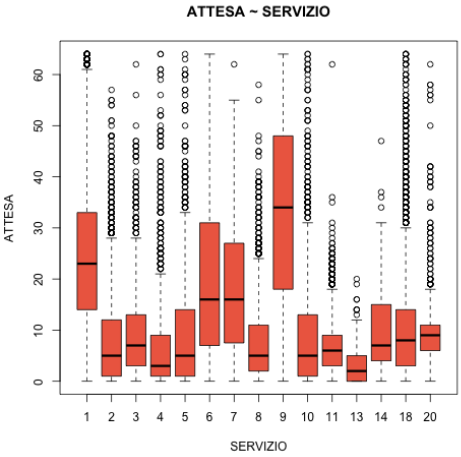
Una volta costruito il dataset da utilizzare, si è passati alla creazione del training set e del test set. Si è deciso di dividere il dataset in due parti: il 70% come training set e il 30% come test set.

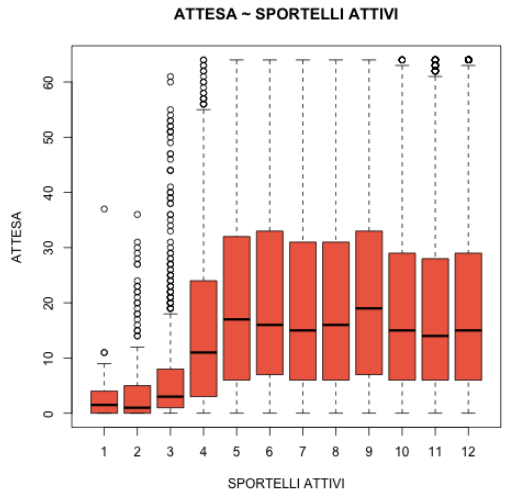
In questa sezione si effettuerà un’analisi esplorativa dati contenuti nel training set, per meglio capire la relazione tra i diversi attributi.

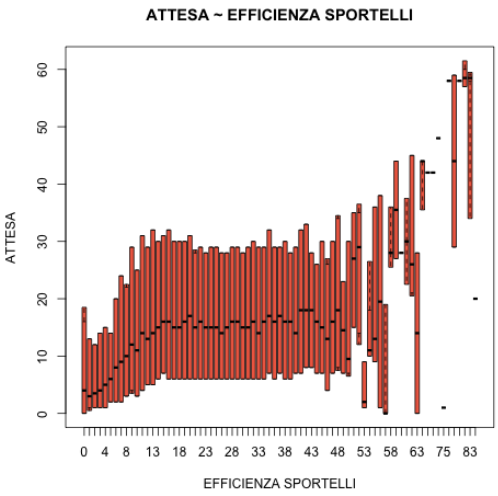


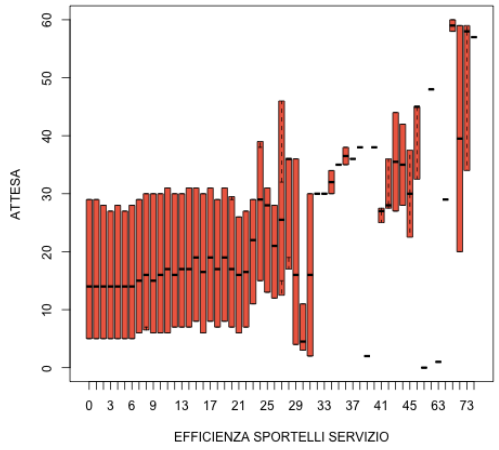


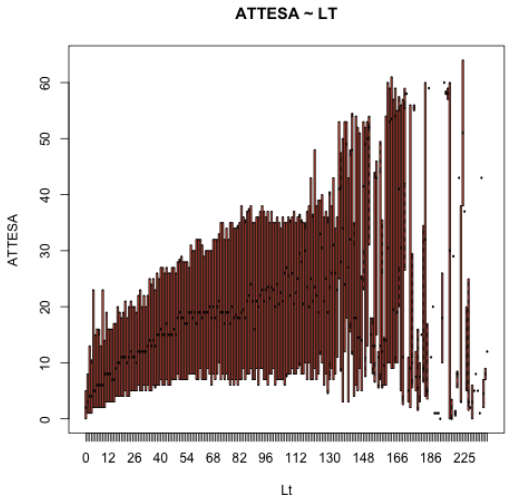


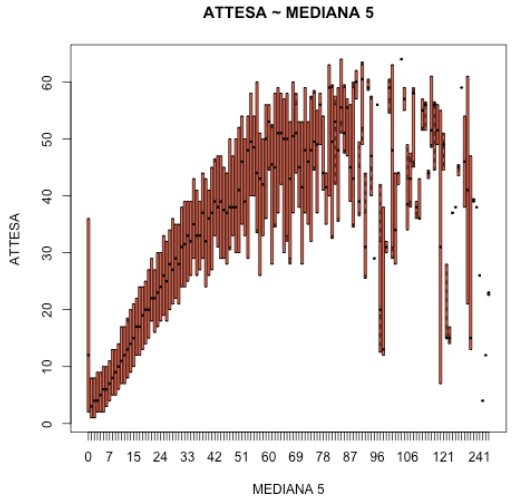


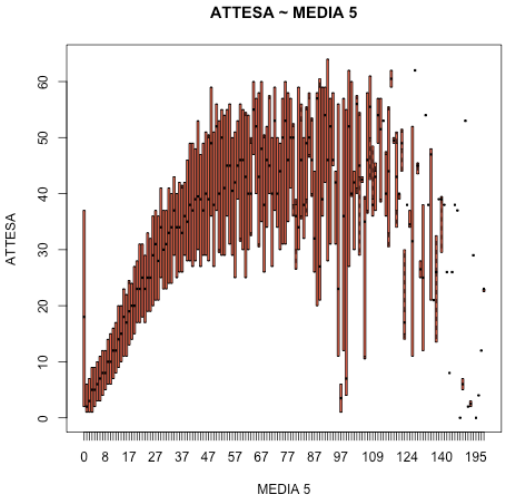

ATTESA ~ SPORTELLI ATTIVI
ATTESA
0 10 20 30 40 50 60
1 2 3 4 5 6 7 8 9 10 11 12
SPORTELLI ATTIVI

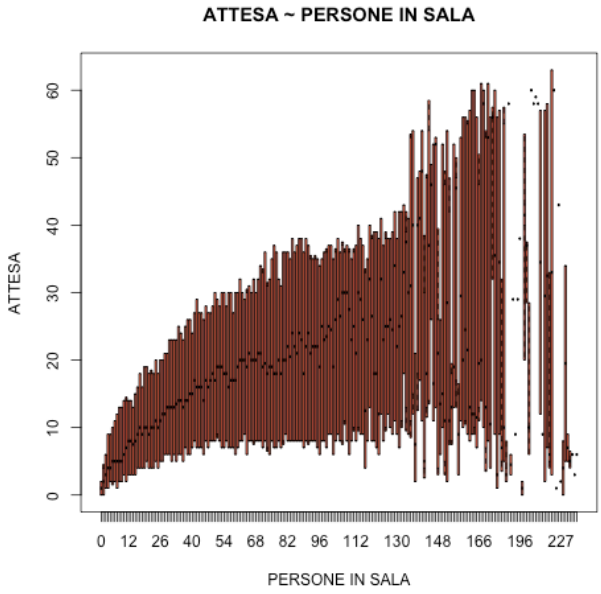

ATTESA ~ EFFICIENZA SPORTELLI
ATTESA
0 10 20 30 40 50 60
0 4 8 13 18 23 28 33 38 43 48 53 58 63 75 83
EFFICIENZA SPORTELLI


ATTESA ~ EFFICIENZA SPORTELLI SERVIZIO
ATTESA
0 10 20 30 40 50 60
0 3 6 9 13 17 21 25 29 33 37 41 45 63 73
EFFICIENZA SPORTELLI SERVIZIO


ATTESA ~ LT
ATTESA
0 10 20 30 40 50 60
0 12 26 40 54 68 82 96 112 130 148 166 186 225
Lt


ATTESA ~ MEDIANA 5
ATTESA
0 10 20 30 40 50 60
0 7 15 24 33 42 51 60 69 78 87 96 106 121 241
MEDIANA 5


ATTESA ~ MEDIA 5
ATTESA
0 10 20 30 40 50 60
0 8 17 27 37 47 57 67 77 87 97 109 124 140 195
MEDIA 5

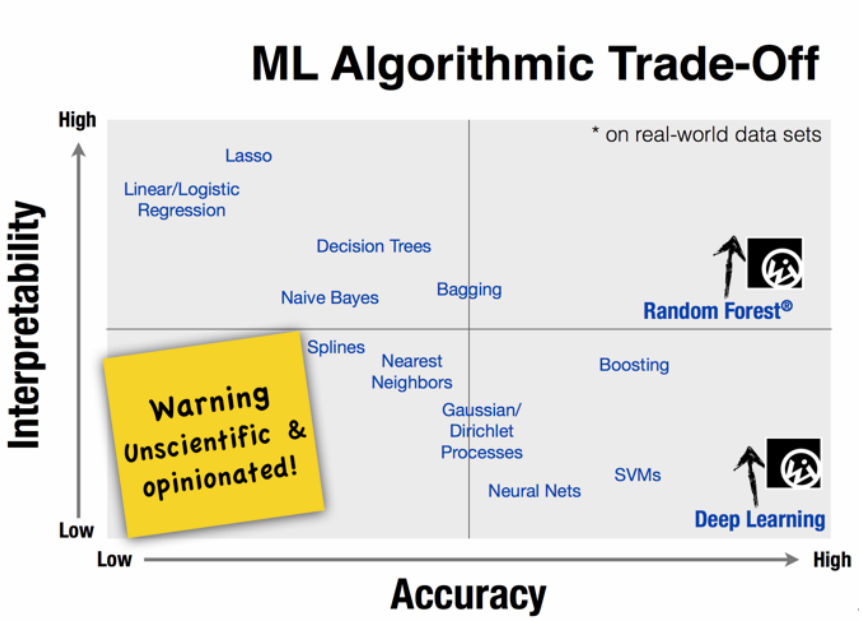


Come si vede dai boxplot precedenti gli attributi che sono maggiormente in relazione con l'attesa sono Persone_in_sala, Media5, Mediana5, Sportelli_attivi, Persone_in_sala, come ulteriore conferma della matrice di correlazione.

3.4 Scelta dei modelli

Confrontando i risultati ottenuti dalla features selection e dalla matrice di correlazione è stato possibile evincere le caratteristiche più significative per predire il tempo d'attesa. Dato che sussisteva una corrispondenza con i migliori modelli identificati nell'articolo di Curtis, Catherine, et al[2]. abbiamo deciso di utilizzare random Forest, uno tra i modelli che in questo studio ha ottenuto i migliori risultati.

Oltre a Random Forest abbiamo deciso di utilizzare un modello più semplice e meno accurato, ma con una maggiore interpretabilità, così da capire effettivamente quali fossero gli attributi maggiormente importanti, nello specifico il modello Decision Tree.



Abbiamo trattato il nostro problema sia come un problema di classificazione e sia come un problema di regressione. In entrambi i casi abbiamo utilizzato random Forest.

In tutti i modelli sfruttati si è discretizzato il tempo di attesa, per poter risolvere il problema sfruttando tecniche di classificazione, nel seguente modo:

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

[2] Curtis, Catherine, et al. "Machine learning for predicting patient wait times and appointment delays." *Journal of the American College of Radiology* 15.9 (2018): 1310-1316.

3.4.1 Decision Tree

Inizialmente si è testato un albero di decisione sfruttando tutti gli attributi del training set e con parametri di default. L'albero ottenuto (DT1) è il seguente:

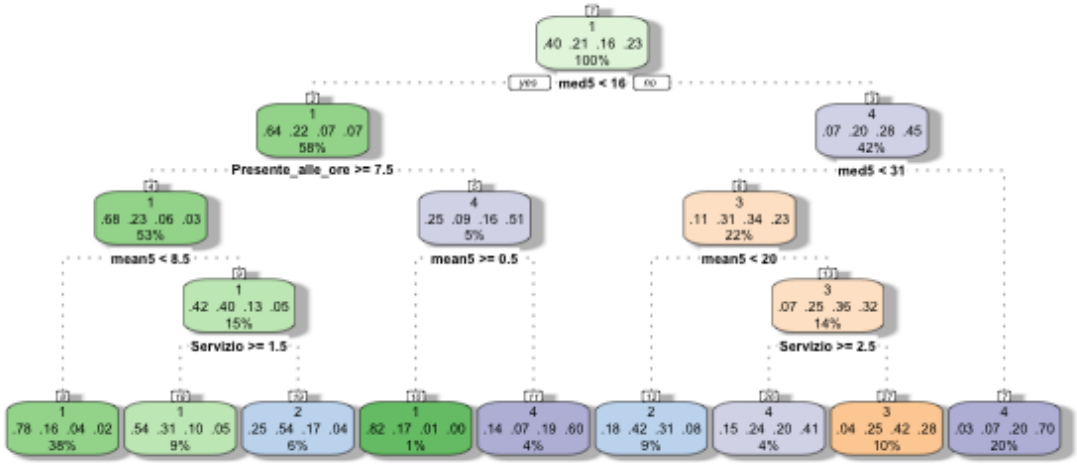


Figura 9 Decision Tree DT1 (default)

Come mostrato nell'immagine precedente otteniamo un albero con 8 split, numero nodi che dividono il training set prima delle foglie, e l'albero sfrutta effettivamente 4 attributi dei 17 passati come parametro.

Le misure di performance sono riportate nella seguente figura:

```
DT1
          accuracy        0.636356
          precision       0.8710149 0.3267837 0.2657963 0.7792720
          recall          0.7312215 0.4738589 0.4187577 0.6415707
          F-measure       0.7950198 0.3868127 0.3251877 0.7037487
```

Figura 10 Performance DT1

Come si può vedere dalle misure di performance i problemi maggiori si riscontrano nella classe 2 e 3, le quali già inizialmente avevano una numerosità minore rispetto alle altre 2 classi.

Per cercare di ottenere un albero di accuracy migliore è stato modificato il valore CP, Il parametro di complessità dell'albero che gestisce il livello di potatura e permette di avere un albero più o meno profondo.
Capire dove potare l'albero può migliorare le performance e soprattutto a parità di accuracy può evitare overfitting sui dati.

Si è deciso quindi di utilizzare un modello con un valore di CP relativamente basso, per capire quanto il modello potesse realmente migliorare le proprie performance.

L'albero DT2 con CP = 0.001 è il seguente:

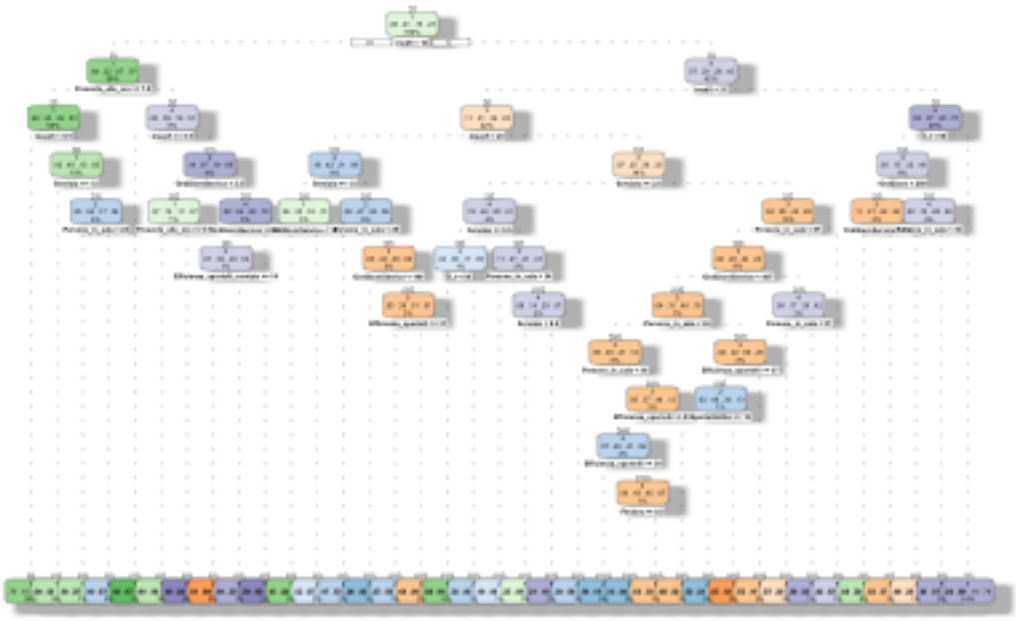


Figura 11 DT2 con cp = 0.001

Come mostrato nell'immagine precedente otteniamo un albero con troppi split e quindi perdiamo leggibilità, punto a favore degli alberi di decisione, a favore di una maggiore accuratezza come si nota dalla seguente immagine delle performance di DT2:

DT2

accuracy	0.6641064			
precision	0.9134525	0.3462419	0.3362924	0.7602654
recall	0.7150806	0.5119887	0.4749263	0.7389334
F-measure	0.8021847	0.4131103	0.3937634	0.7494476

Figura 12 Performance DT2

Come si può notare avendo aumentato di molto il numero di split anche l'accuracy è salita da uno 0.63 iniziale ad uno 0.66, l'incremento minimo di accuracy però non giustifica la perdita di leggibilità dell'albero.

Per cercare di ottenere una buona accuracy e migliorare la leggibilità rispetto a DT2 si è modificato ulteriormente il valore del parametro cp, sulla base del seguente grafico:

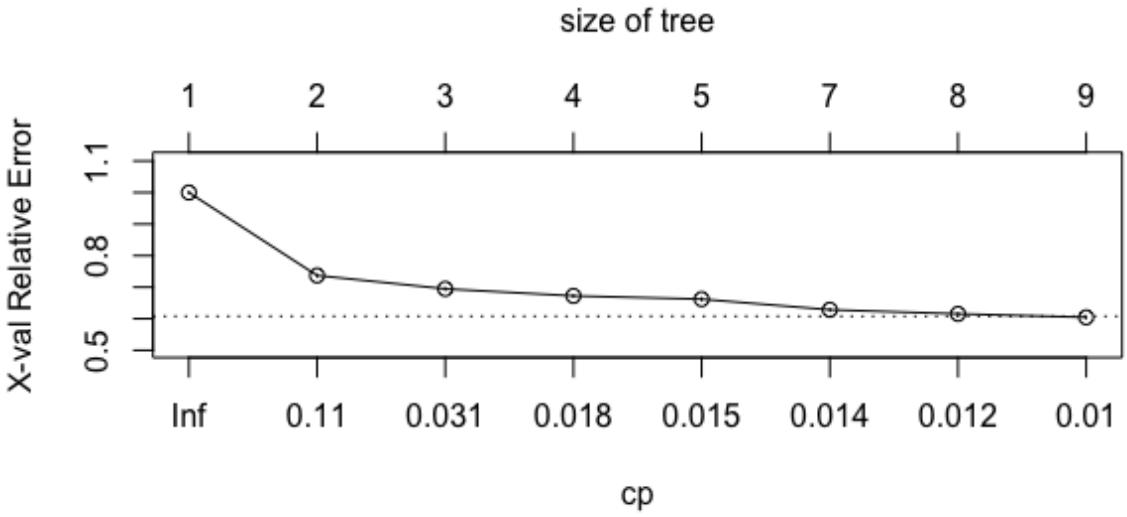


Figura 13 Plot CP

Il grafico mette in relazione l'errore relativo e la profondità dell'albero, e osservandolo si nota come dopo cp = 0.012 l'errore non diminuisca più.

Si è quindi potato il precedente albero con un valore cp = 0.012, ottenendo il seguente albero:

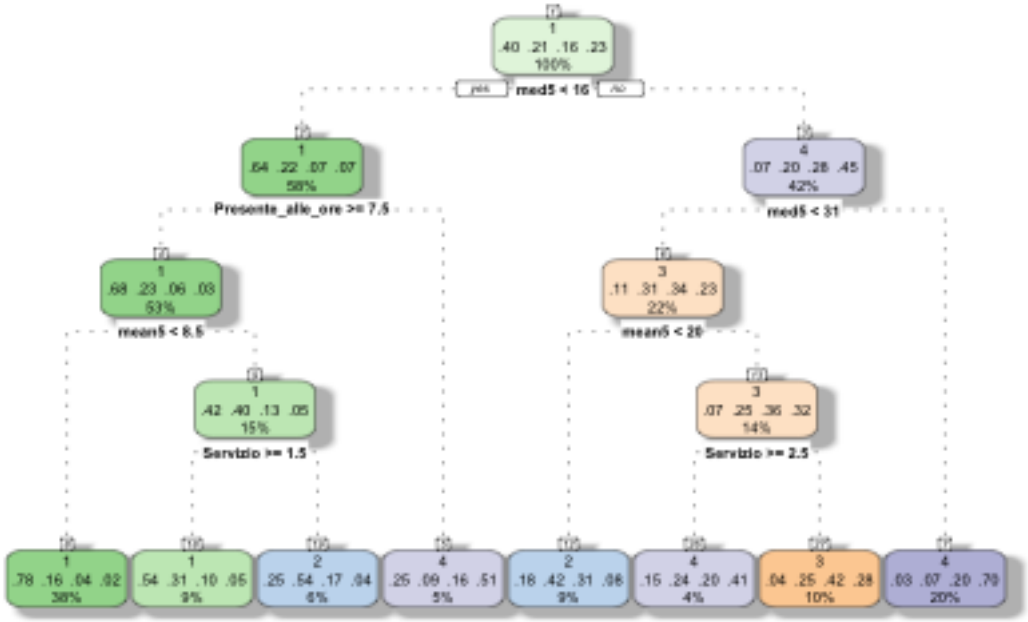


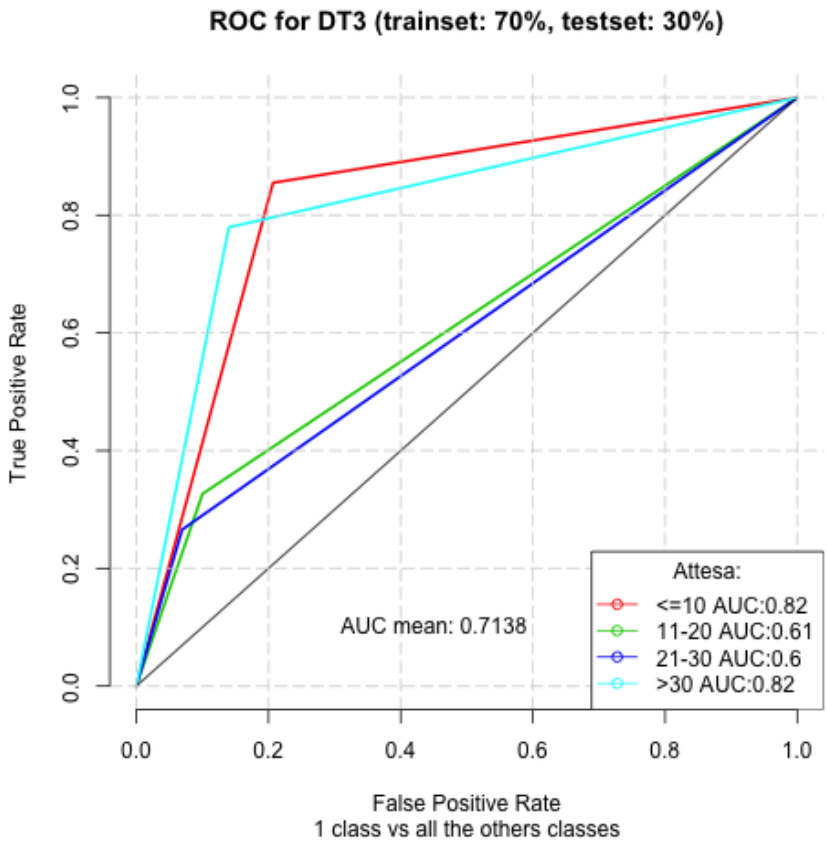
Figura 14 Plot DT3

L’albero ottenuto è molto simile al primo, con valori default, l’unica differenza è la presenza di uno split in meno, 7 a confronto dei precedenti 8. Quindi si è migliorata, anche se di poco, la leggibilità dell’albero, mantenendo le stesse performance, come mostrato di seguito

```
DT3
          accuracy          0.6301206
          precision         0.8551270 0.3267837 0.2657963 0.7794513
          recall            0.7295996 0.4738589 0.4187577 0.6248383
          F-measure         0.7873917 0.3868127 0.3251877 0.6936333
```

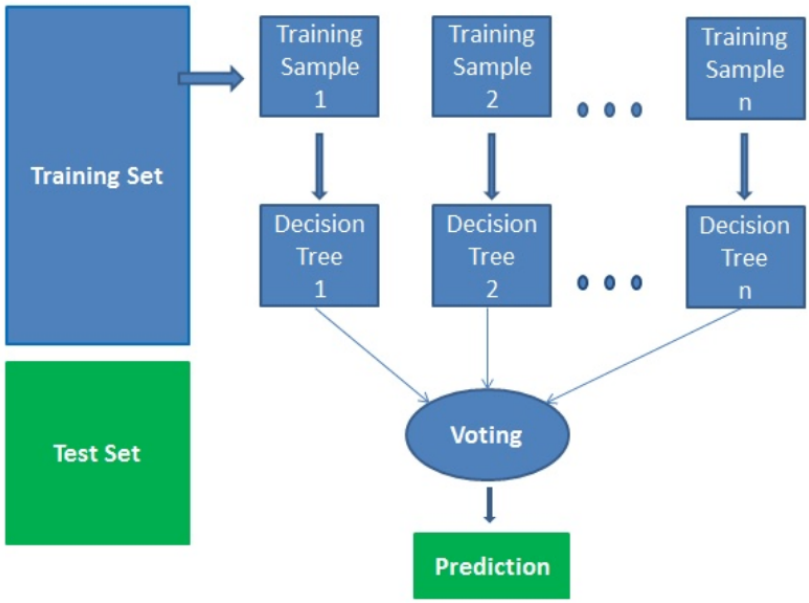
Figura 15 Performance DT3

Di seguito è mostrata la ROC per il solo albero di decisione DT3, migliore per leggibilità e performance



3.4.2 Random Forest

Dopo una prima fase di studio di Random-forest sono state comprese le motivazioni grazie alle quali questo tipo di modello riesce ad ottenere buoni risultati per questo problema. Infatti, random-forest non è altro che un insieme di alberi di decisione generati su differenti parti dello stesso training set, con l'obiettivo di ridurre la varianza. Alberi decisionali profondi hanno la caratteristica di individuare pattern con alta irregolarità, tramite overfitting sul training set, producendo un basso bias ma varianza molto alta. Quindi random forest è un modo per mediare i diversi alberi decisionali profondi, addestrati su differenti parti dello stesso training set, con l'obiettivo di ridurre la varianza[3].



PARAMETRI

Per eseguire Random Forest vengono impostati due parametri:

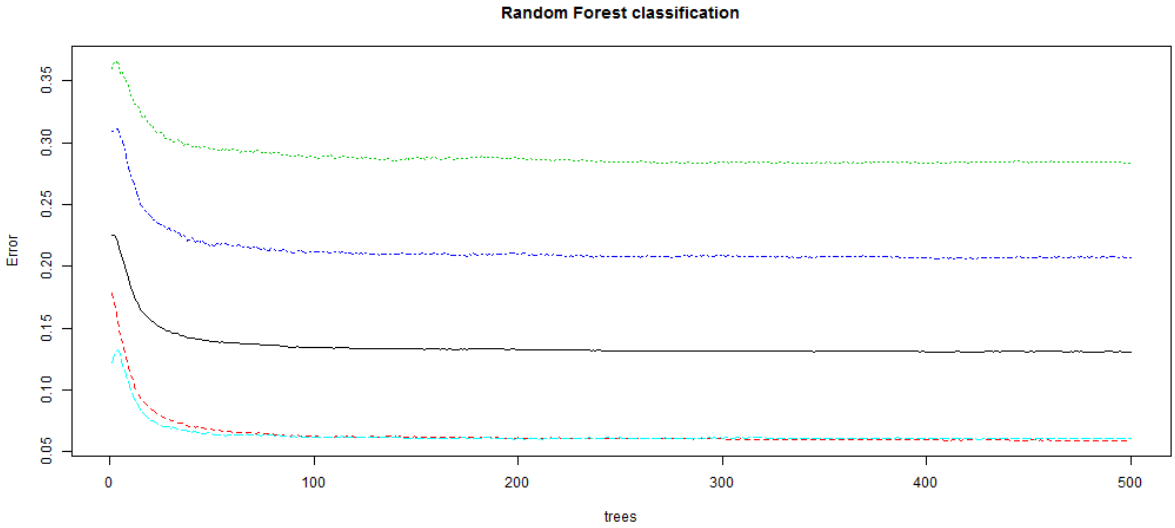
- ntree: ovvero il numero di alberi generati
- mtry: il numero di features scelte randomicamente prima dello split

Per la scelta di mtry, come suggerito dal libro "*The elements of statistical learning*"[3], considerando come Nil numero di features del dataset, nel caso di classificazione si utilizza un valore tra $\sqrt{N}$e $log(N)$ mentre per per la regressione gli si assegna un valore pari a $N/3$. Quindi dato che le nostre features sono 17 dovremmo utilizzare come valore del parametro 4. Questo viene verificato anche dalla stima dell' OOB error al variare del valore di mtry. OOB error, cioè out-of-bag error, è un metodo per calcolare l'errore di predizione su random forest e altri modelli di machine learning che utilizzano la tecnica di bagging. In pratica è la

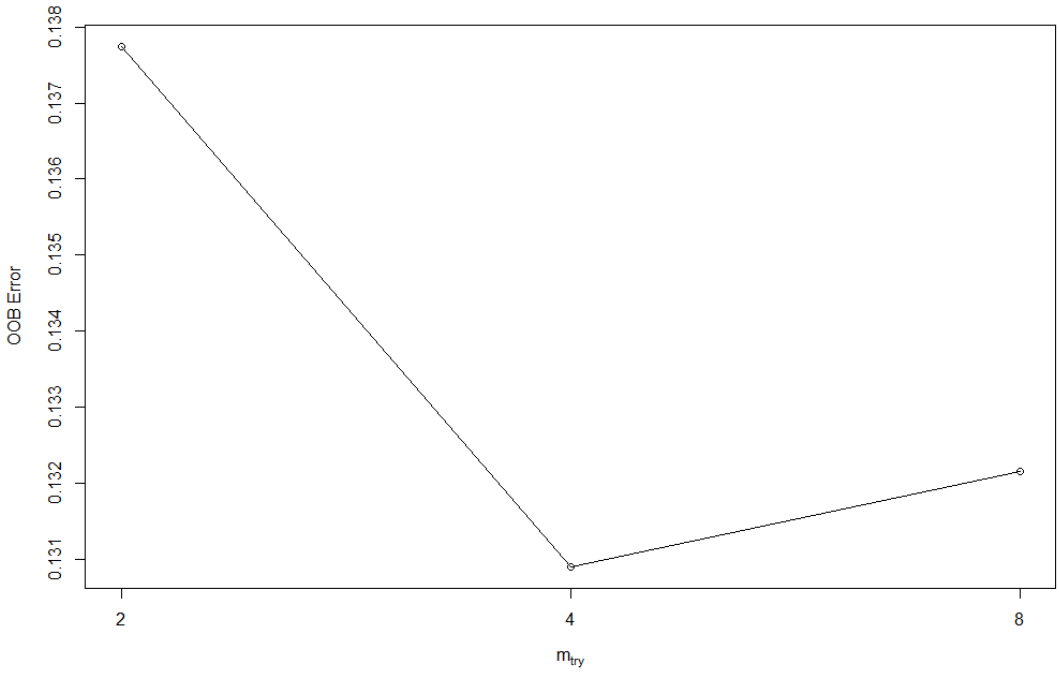
[3]Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. No. 10. New York, NY, USA:: Springer series in statistics, 2001.

media dell'errore di predizione di ciascun campionamento xi dal trainset, usando solo gli alberi che non hanno xi nel proprio bootstrap sample.[4]

Per prima cosa si è eseguito random forest su tutto il dataset per stimare il valore dei parametri ottimali. Come si evince dal grafo, sono sufficienti 400 alberi per ottenere un OOB error costante per ciascuna classe.

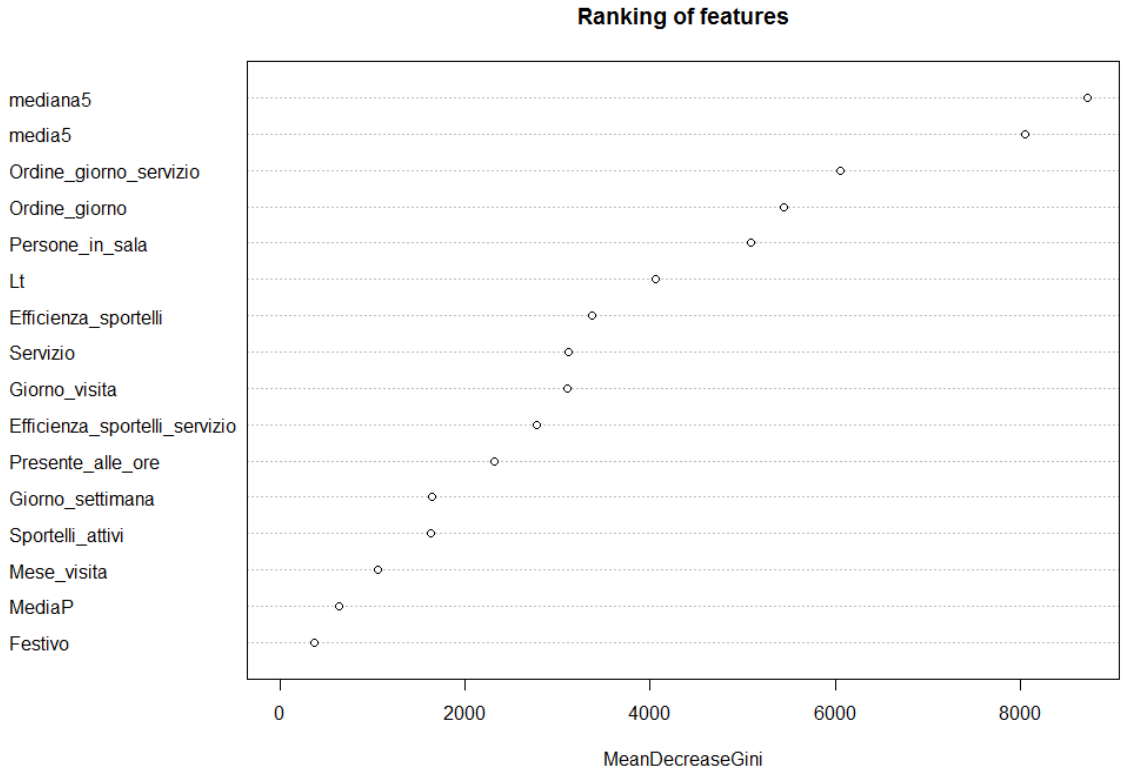


Poi, abbiamo verificato il valore di mtry, che come vediamo nella figura sottostante nel caso di mtry = 4 genera un OOB error inferiore.

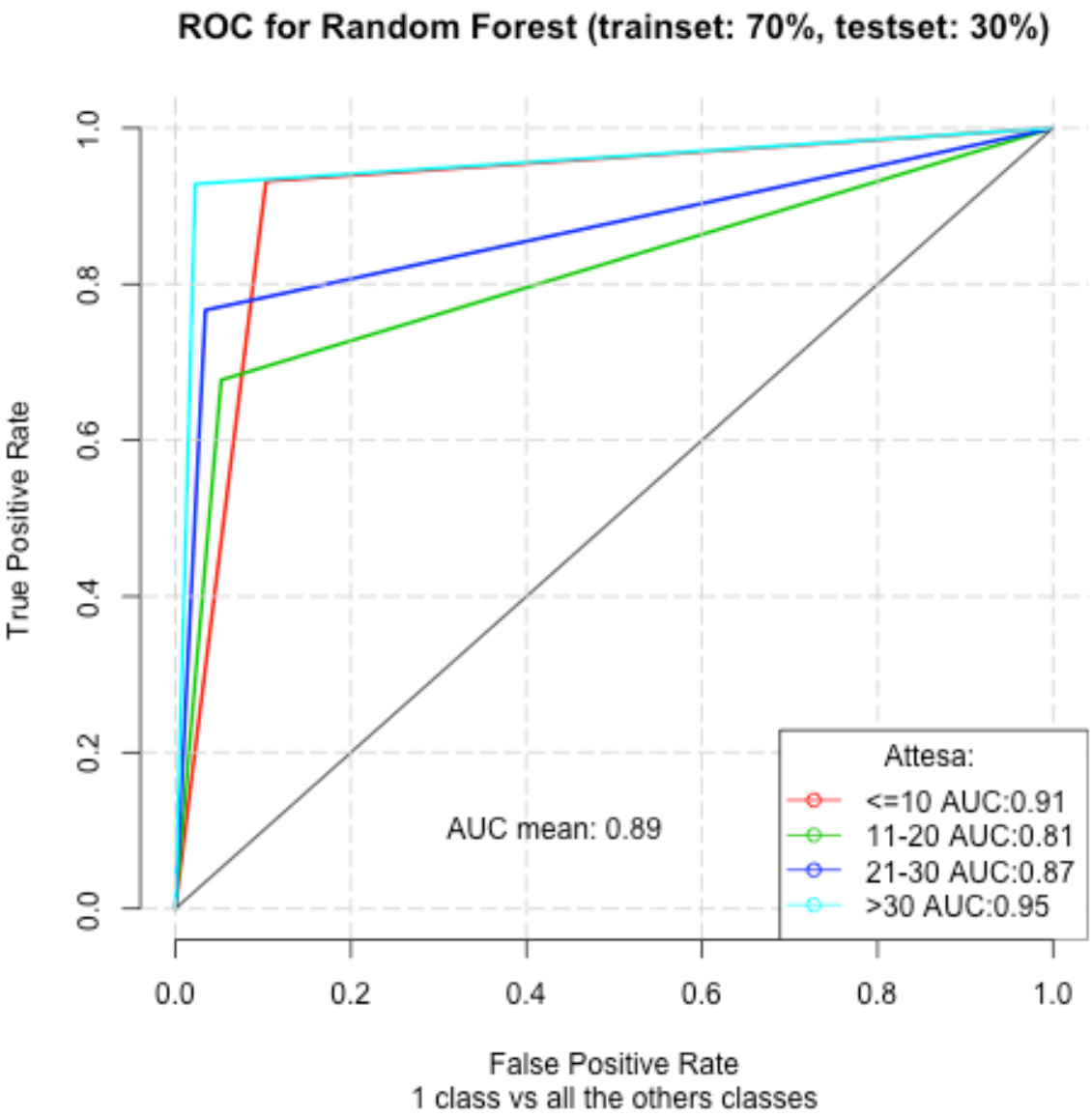


[4] James, Gareth, et al. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.pp. 316–321.

Inoltre, sono state identificate le features di maggior valore. Random Forest ha una sua funzione per assegnare un punteggio alle diverse caratteristiche, basata sull'indice di Gini. Il risultato di questa operazione corrisponde a quanto già individuato nella feature selection.



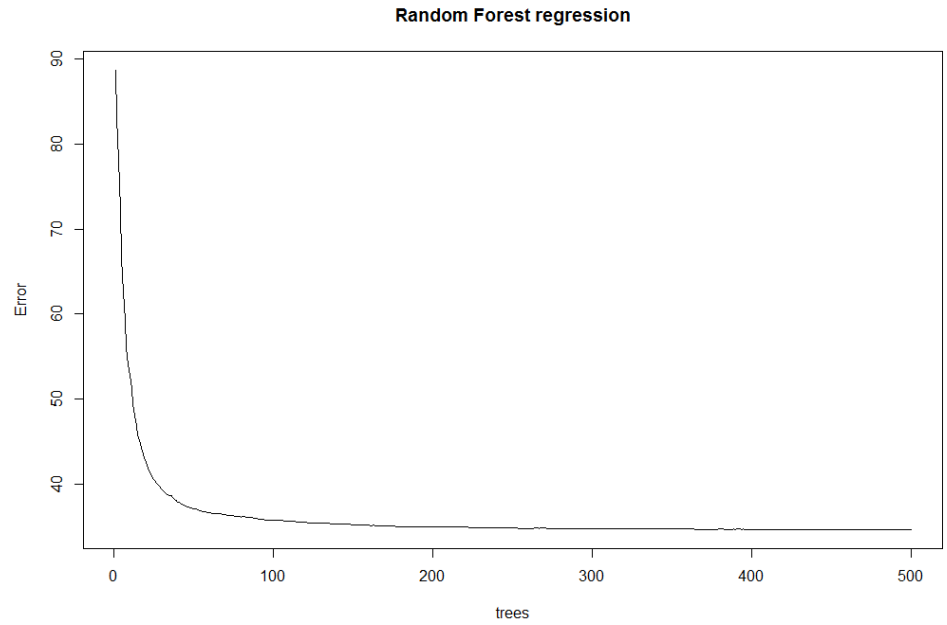
Vengono ora mostrati i risultati del modello:



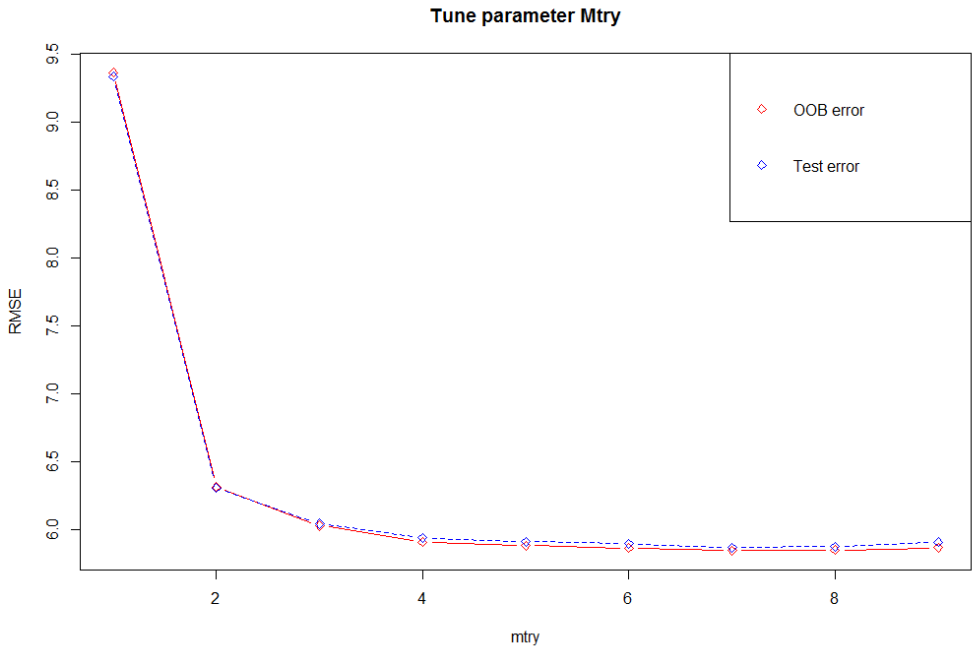
RANDOM FOREST REGRESSION

Ora il problema viene trattato come un problema di regressione. In questo caso sono stati considerati come outliers i record con valori di attesa superiori al 99 percentile.

Come fatto per il Random forest utilizzato nella classificazione, è necessario individuare il miglior valore dei parametri mtree e mtry. Nel primo test sono stati utilizzati i parametri con i valori di default e train set e test set sono stati suddivisi rispettivamente con il 70% e il 30% del dataset. Come mostrato nell'immagine qui sotto, anche impostando ntree con un valore superiore a 300 il modello non ottiene risultati migliori.



Per motivi di computazione non è stato possibile testare il parametro mtry su tutti i suoi possibili valori. Quindi è stato testato con un valore da 1 a 9, ottenendo i migliori risultati con mtry = 7

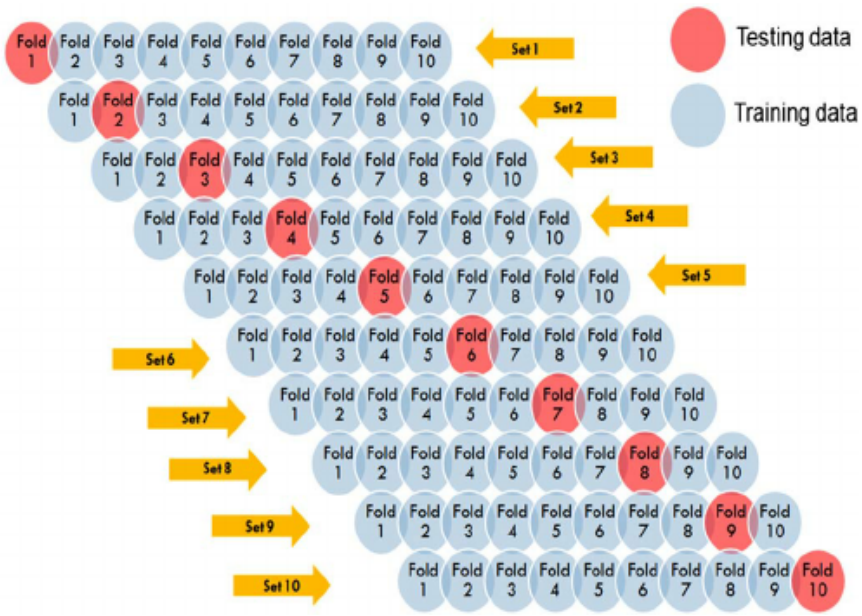


Quindi si è deciso di addestrare il modello random forest con mtry=7 e mtree=300. Il risultato viene valutato utilizzando la metrica RMSE (root mean square error). Il risultato ottenuto è RMSE= 5,9 con un tempo medio di predizione pari a 21.2

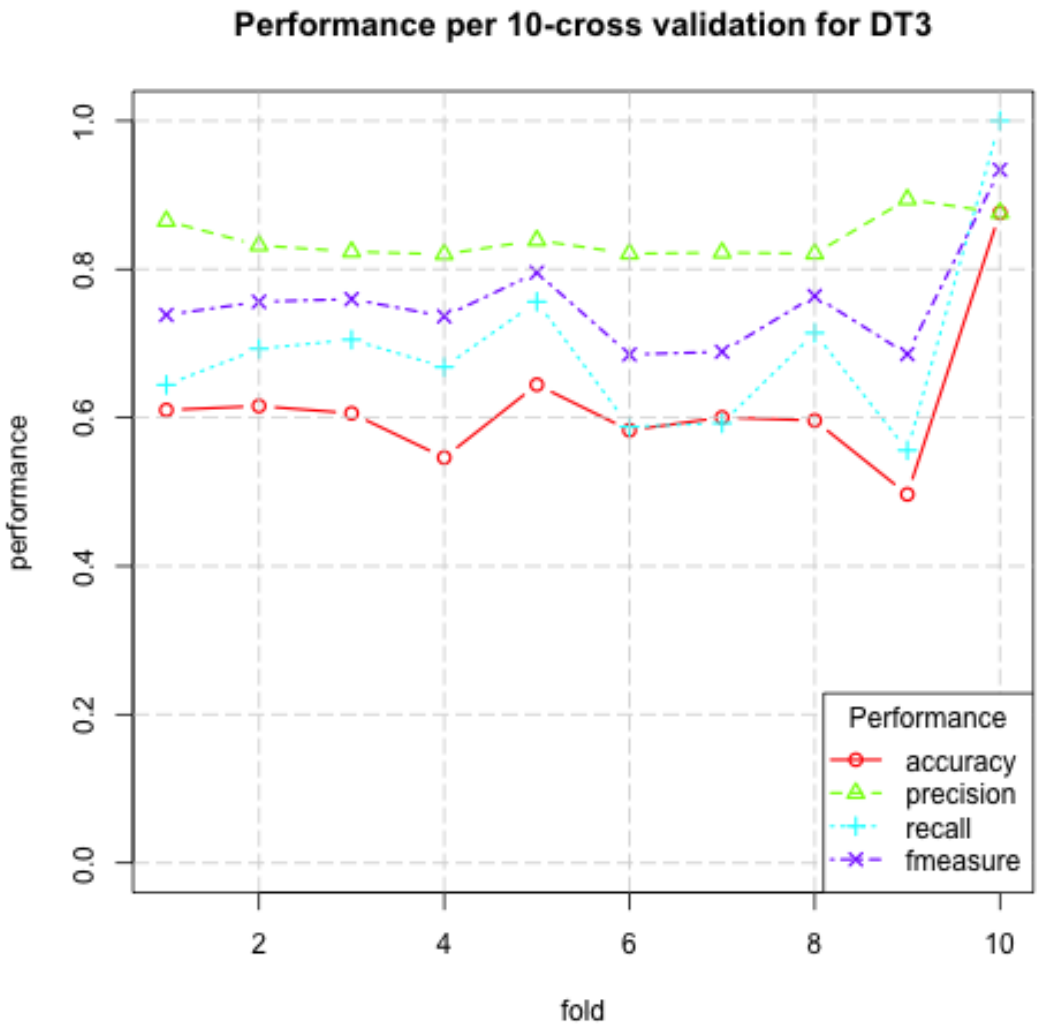
3.5 10-fold cross validation

Sia per il modello DT3 che per random forest si sono validate le misure di performance attraverso una 10-cross validation. Questa tecnica è molto utile per verificare che le performance del modello siano reali e non causali sulla base della scelta del training e test set.

L'intero dataset è stato diviso in 10 fold casuali, i quali a turno erano presi come test set e il resto come training set, come mostrato nella figura successiva, e per ogni modello si sono calcolate le performance. Infine ne è stata calcolata la media per ognuna.



Di seguito è mostrato un grafico che rappresenta le 4 misure di performance (accuracy, precision, recall e f-measure) per ogni fold preso in considerazione:

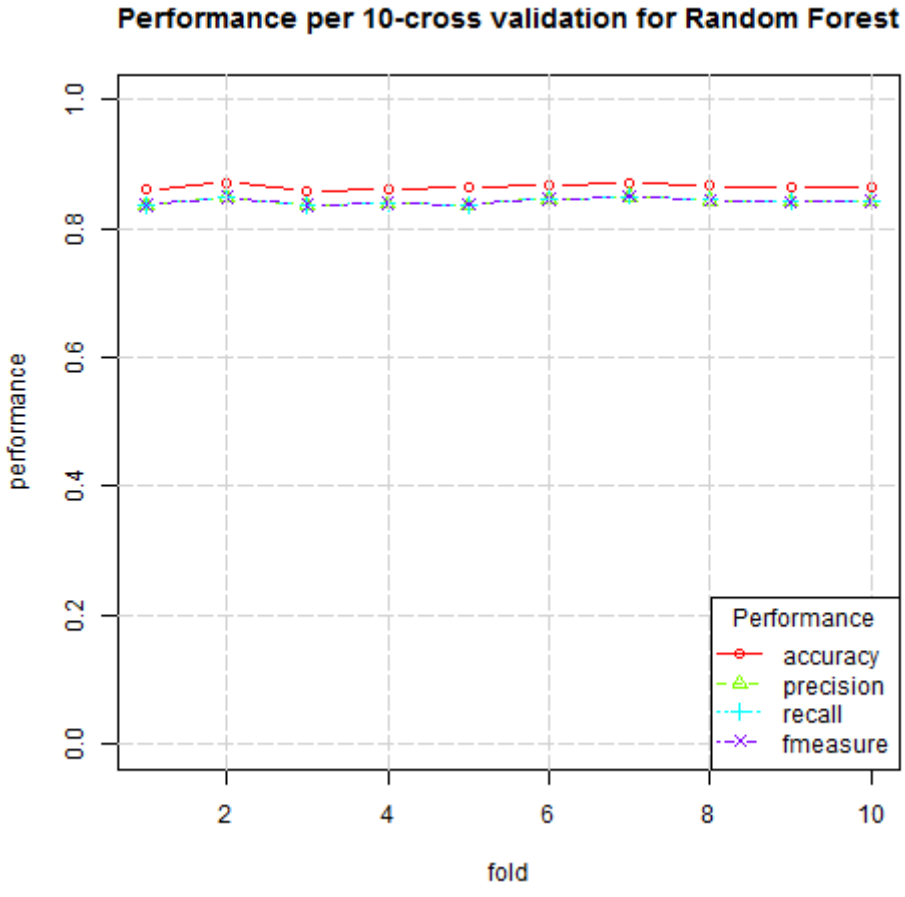


Le misure di performance medie dopo l'esecuzione della 10-cross validation sono le seguenti:

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

Le performance medie mostrano come i valori di accuracy inferiore a 0.6 e maggiore di 0.8 mostrati nel plot precedente siano probabilmente causati dalla scelta nella divisione dei dati in training set e test set, inoltre confermano le misure calcolate sul singolo modello iniziale.

Per quanto riguarda random forest, viene mostrato come le misure di perfomance dopo la 10-fold cross validation sono leggermente inferiori a quelle ottenute dal modello nella singola computazione, ma con un'accuratezza corrispondente a quella stimata dall'OOB error.



Le misure di performance medie dopo l'esecuzione della 10-cross validation sono le seguenti:

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

Per motivi computazionali non è stata effettuata la 10-fold cross validation sul random forest nel problema di regressione, ma è possibile valutare l'efficacia del modello grazie al RMSE stimato a partire dall' OOB error.

3.6 Risultati

Di seguito è mostrato un riassunto delle misure di performance sui due modelli di classificazione utilizzati

```
DT3
          accuracy          0.6301206
          precision         0.5567895
          recall            0.5617636
          F-measure         0.5482563
```

```
DT3
          accuracy          0.6301206
          precision         0.8551270 0.3267837 0.2657963 0.7794513
          recall            0.7295996 0.4738589 0.4187577 0.6248383
          F-measure         0.7873917 0.3868127 0.3251877 0.6936333
```

Random forest:

```
> accuracy_s
[1] 0.8634787
> precision_s
[1] 0.8427179
> recall_s
[1] 0.8552429
> fmeasure_s
[1] 0.8489342
>
         Sensitivity Specificity Pos Pred Value Neg Pred Value Precision    Recall        F1
Class: 1   0.9348575   0.9108424      0.8728070      0.9552880 0.8728070 0.9348575 0.9027673
Class: 2   0.7172414   0.9512003      0.8015414      0.9244820 0.8015414 0.7172414 0.7570519
Class: 3   0.7876448   0.9660617      0.8160000      0.9596892 0.8160000 0.7876448 0.8015717
Class: 4   0.9311280   0.9794477      0.9306233      0.9796049 0.9306233 0.9311280 0.9308756
         Prevalence Detection Rate Detection Prevalence Balanced Accuracy
Class: 1  0.3955649      0.3697968            0.4236868         0.9228499
Class: 2  0.2155600      0.1546085            0.1928890         0.8342208
Class: 3  0.1604311      0.1263627            0.1548563         0.8768532
Class: 4  0.2284440      0.2127106            0.2285679         0.9552878
```

4 CONCLUSIONI

DATA TECHNOLOGY

La parte di Data Technology è stata fondamentale per permetterci di lavorare con dati puliti e significativi.

MACHINE LEARNING

Per quanto riguarda la parte di machine learning, l'utilizzo dell'albero di decisione non riesce a gestire l'alta variabilità dei dati. Al contrario il modello random forest ha permesso di ottenere risultati soddisfacenti sia nel problema di classificazione che in quello di regressione.
Il principale limite è stato di natura hardware e software. Infatti, non è stato possibile stimare i parametri sull'intero dataset, ed effettuare la 10-fold cross validation sul random forest nel problema di regressione. Questi stessi limiti non hanno permesso l'utilizzo di modelli più avanzati ed onerosi computazionalmente.

sviluppi futuri:

- utilizzare un dataset più ampio (un anno)
- utilizzare altri modelli di ML come GBM