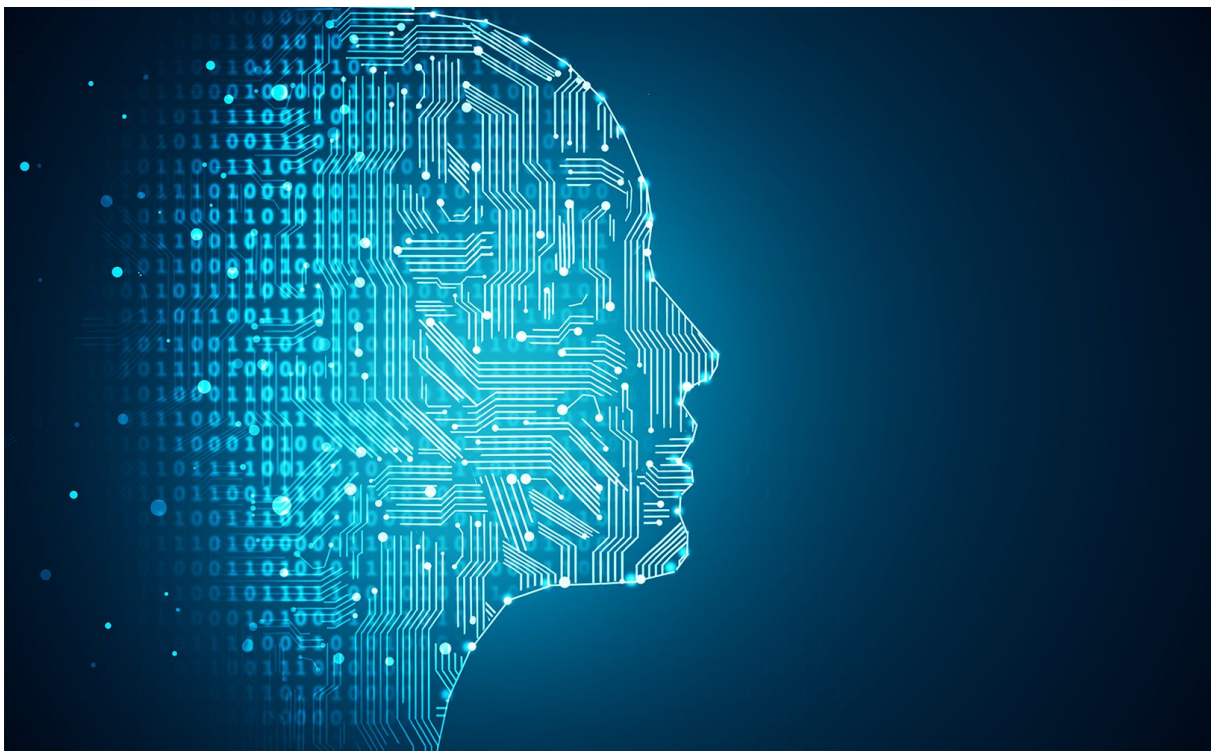


UNIVERSITÀ DEGLI STUDI DI MILANO - BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

PROGETTO DI DATA TECHNOLOGY E MACHINE LEARNING



Anno accademico 2018/2019

Autori

Luca Virgilio - 794866

Samuele Ventura - 793060

Indice

1	Introduzione	3
2	Data Technology	4
2.1	Raccoglimento dati iniziali	4
2.2	Descrizione dei dati iniziali	4
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	7
2.3.1	Tu_Passi_2018 DQ	8
2.3.2	Dati_sensori_meteo DQ	9
2.3.3	Feste DQ	9
2.4	Data cleaning	9
2.5	Data Integration	11
2.6	Analisi Descrittiva	13
3	Machine Learning	15
3.1	Calcolo dei parametri	15
3.2	Feature selection	16
3.3	Analisi descrittiva del training set	18
3.4	Scelta dei modelli	22
3.4.1	Decision tree	22
3.4.2	Random forest	26
4	Analisi dei risultati	29
4.1	Performance	29
4.1.1	Classificazione	29
4.1.2	Regressione	29
4.2	10-fold cross validation	31
5	Conclusioni	33

1 Introduzione

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[1], 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. 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 dell'università Milano-Bicocca 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.

2 Data Technology

2.1 Raccoglimento 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 dati dell'ospedale

In questo modo si è ottenuto un dataset contenente dati fondamentali per predire il tempo di attesa:

- **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 meteorologiche 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 i precedenti o successivi ad una 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 contenute al loro interno.

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

Tabella 1: Dati presenti nel dataset tu_Passi_18_per_andamento_servizio

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 tabella 1 qui sotto.

* 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 meteorologica 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 dovuti 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, <https://www.dati.lombardia.it/Ambiente/stazioni-meteo/pevf-9zqp>, è stato possibile individuare facilmente la stazione più vicina all'ospedale e filtrare i dati direttamente con il numero del sensore individuato, come mostrato in figura 1.

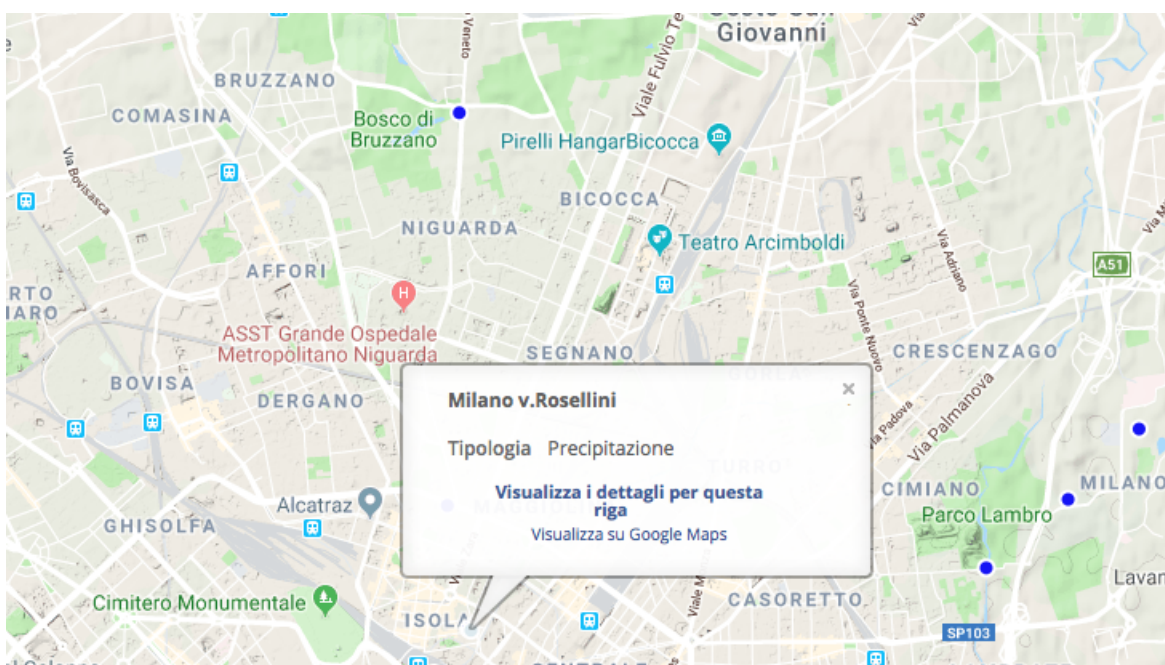


Figura 1: Immagine interfaccia grafica di stazioni meteo lombardia

Il dataset da cui si sono estratte le informazioni sulle precipitazioni è stato scaricato dal sito dell'arpa, al seguente link <http://www.arpalombardia.it/siti/arpalombardia/meteo/richiesta-dati-misurati/Pagine/RichiestaDatiMisurati.aspx>, dal quale è possibile selezionare provincia, tipo di rilevazione, stazione meteorologica e intervallo temporale.

I dati sono inviati via email, come csv o pdf, e sono descritti nella Tabella 2.

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.

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	//

Tabella 2: Dati presenti nel dataset scaricato dal sito arpa

2.2.3 Feste

Il dataset Feste, mostrato nella Tabella 3, è stato creato a mano, non esistendo un dataset con gli attributi richiesti.

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)	//

Tabella 3: Dati presenti nel dataset Festa

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 migliorarne la qualità.

Per verificare la qualità dei dati si è utilizzato il software Talend, nello specifico 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 possibile 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, specificata nella Tabella 4.

Nei prossimi paragrafi verranno mostrati i risultati delle misure di qualità proposte sui singoli dataset.

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

Tabella 4: Dimensioni e metriche di qualità impiegate

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	$(347647 - 3612)/347647 = 0.98$

Tabella 5: Risultati delle completezza degli attributi sul dataset Tu_Passi_2018

Come mostrato nella Tabella 5, ogni attributo ha in media una completezza di 0.97, come viene confermato dalla completezza a livello di tabella, calcolata nella Tabella 6. Per quanto riguarda la consistenza sono stati definiti i seguenti vincoli interni alla tabella:

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

Tabella	Completezza su tabella
Tu_passi_2018	5443856/5562352 = 0.97

Tabella 6: Risultato di completezza a livello di tabella sul dataset Tu_passi_2018

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

Tabella 7: Risultati sui vincoli di consistenza sul dataset Tu_Passi_2018

Nella Tabella 7 vengono mostrati i risultati sulla consistenza delle tuple. Nella Tabella 8 vengono mostrati i risultati riguardo l'unicità delle tuple.

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)

Tabella	Unicità delle tuple
Tu_Passi	1

Tabella 8: Risultati sull'unicità delle tuple nel dataset Tu_Passi_2018

- 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* fossero 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 sfruttabili nella successiva fase di predizione.
 - Sono state eliminate le righe in cui il valore nella colonna *Ultima_operazione_alle* era vuoto, così da poter migliorare la consistenza.

Vincolo	Consistenza
Chiamato_alle_ore >Presente_alle_ore && Ultima_operazione_alle >Chiamato_alle_ore	1

Tabella 9: Risultati sui vincoli di consistenza sul dataset Tu_Passi_2018 dopo la fase di data cleaning

- 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.

2.5 Data Integration

Il dataset finale è il consolidamento dei tre dataset descritti precedentemente. Essi sono stati uniti sulla base della data e dell'ora dei pazienti presenti nel dataset Tu_Passi, come mostrato in Figura 2.

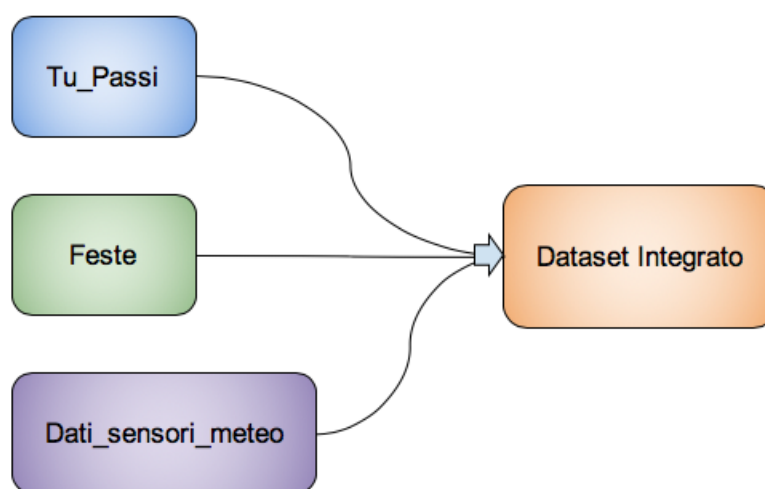


Figura 2: Integrazione dei dataset

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

Tabella 10: Risultati sulla completezza degli attributi sul dataset finale dopo la fase di data cleaning.

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

2.6 Analisi Descrittiva

Di seguito sono mostrati alcuni grafici che descrivono il dataset integrato.

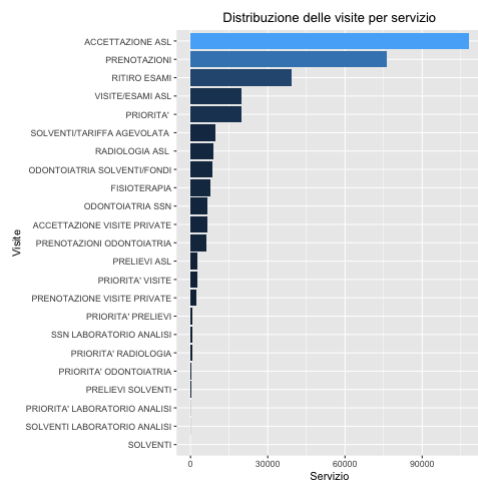


Figura 3: Distribuzione dei servizi

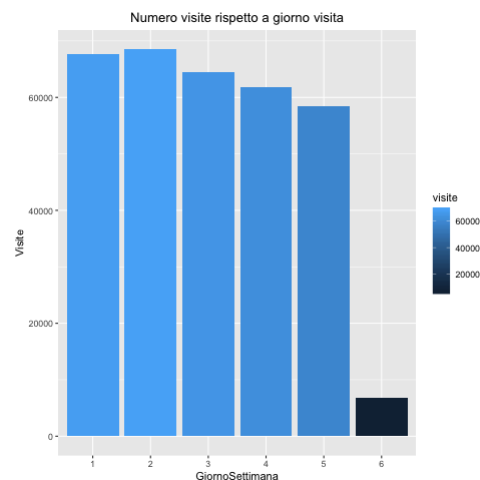


Figura 4: Numero di visite suddivise per giorno della settimana

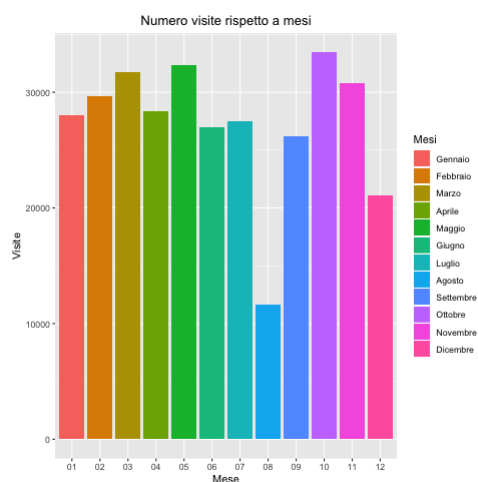


Figura 5: Numero di visite suddivise per mese

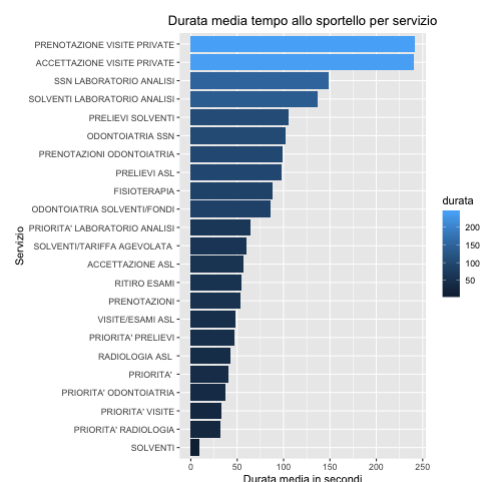


Figura 6: Durata media dei diversi servizi allo sportello

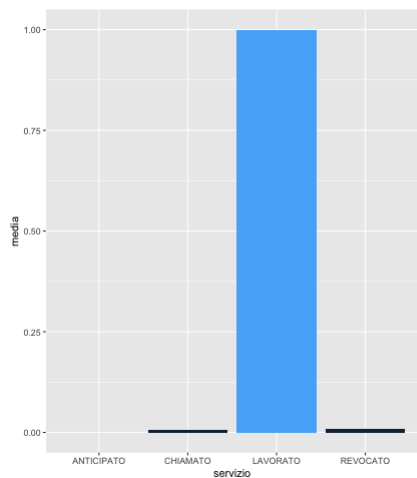


Figura 7: Distribuzione dello stato delle richieste

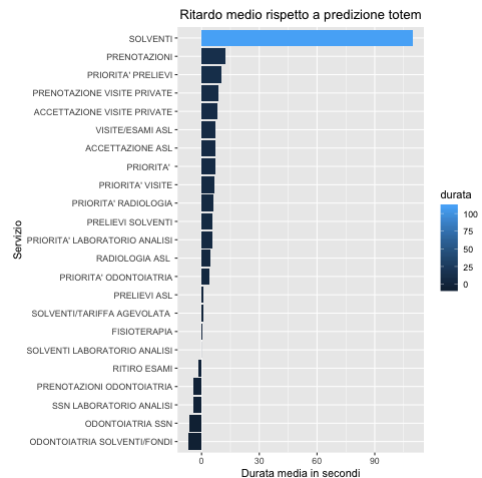


Figura 8: Ritardo medio rispetto alle predizioni del totem

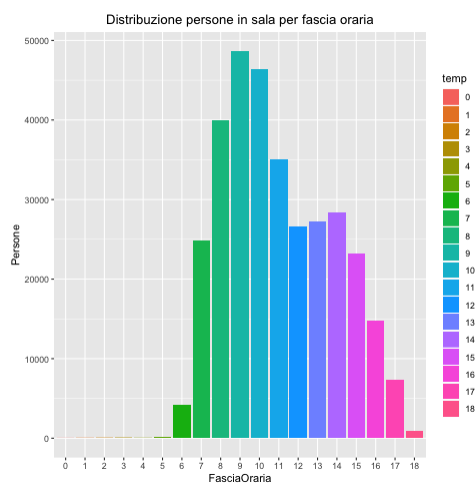


Figura 9: Distribuzione delle richieste per fascia oraria

3 Machine Learning

3.1 Calcolo dei 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, è stato necessario calcolarla. Questa sarà la variabile target per quanto riguarda i modelli predittivi.

Si è deciso di prendere in considerazione solamente gli ultimi 3 mesi dell'anno 2018, dato che molti servizi non sono stati registrati durante l'intero anno. Questa scelta permette di utilizzare un dataset il più simile possibile alla situazione attuale. La distribuzione dei servizi durante l'anno è mostrata nella Figura 10.

	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

Figura 10: Distribuzione dei servizi durante l'anno

Dato che le informazioni contenute nel dataset non risultavano sufficienti per addestrare un modello predittivo efficace, è stato necessario calcolare nuovi attributi basandoci sui dati in nostro possesso, in modo da esplicitare informazioni che prima non lo erano e che potessero essere utili nella fase di training del modello.

Grazie agli attributi che indicavano l'orario della prima e dell'ultima operazione effettuata da un operatore ad uno sportello, si è stimata la durata del servizio allo sportello. Analizzando il dominio di riferimento, si è 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, dopo aver discretizzato in ore l'orario in cui i pazienti vengono chiamati, per ogni ora sono stati contati il numero di sportelli che hanno effettuato almeno una chiamata.

Questo lavoro si basa due articoli pubblicati sul Journal of the American College of Radiology [2] [3]. 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, si è deciso di calcolare dei parametri risultati utili anche negli articoli appena citati:

- 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 Feature selection

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

Utilizzando vari subset di attributi (per aumentare la leggibilità), ogni volta tenendo la colonna target, sono state calcolate le seguenti matrici di correlazione, mostrate in Figura 11, 12, 13.

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, sono state escluse 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. Il dataset finale si è ridotto a 17 caratteristiche, elencate in Tabella 11.

Prima di addestrare il modello di machine learning si è reputato opportuno eseguire un'ulteriore fase di feature selection, per avere un'altra metodologia, oltre alla matrice di correlazione, per valutare l'importanza delle nostre features. La Figura 14 mo-

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

Tabella 11: Dataset finale utilizzato per i modelli di Machine learning

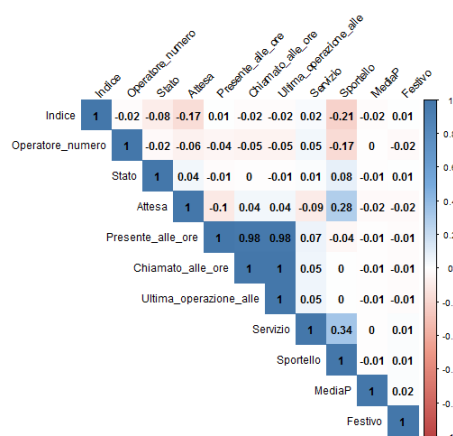


Figura 11: Matrice di correlazione del primo subset

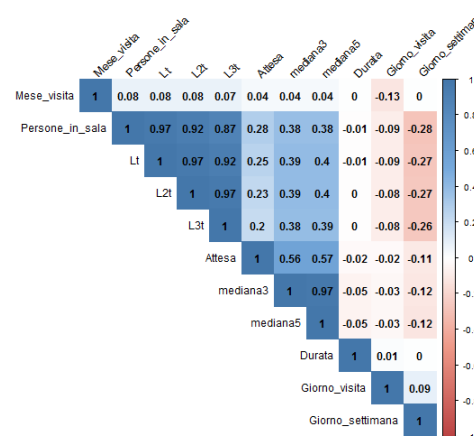


Figura 12: Matrice di correlazione del secondo subset

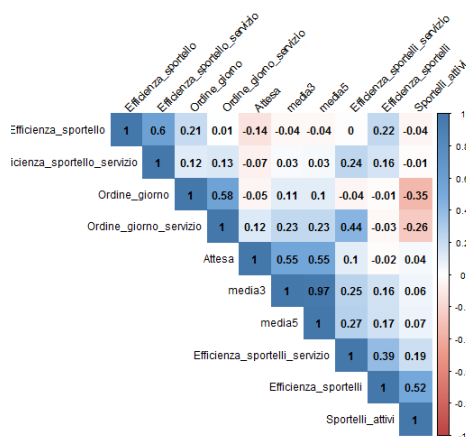


Figura 13: Matrice di correlazione del terzo subset

	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

Figura 14: Feature selection con caret

stra i risultati del metodo `features()` del package `caret` di R. Questo metodo conferma l'importanza delle *features* evidenziate nella matrice di correlazione.

Per rimuovere gli outlier, si è deciso di escludere i record che registravano un valore di attesa superiore al 95 percentile, come mostrato nella Figura 15, così facendo si è migliorata la distribuzione dei dati.

3.3 Analisi descrittiva del 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.

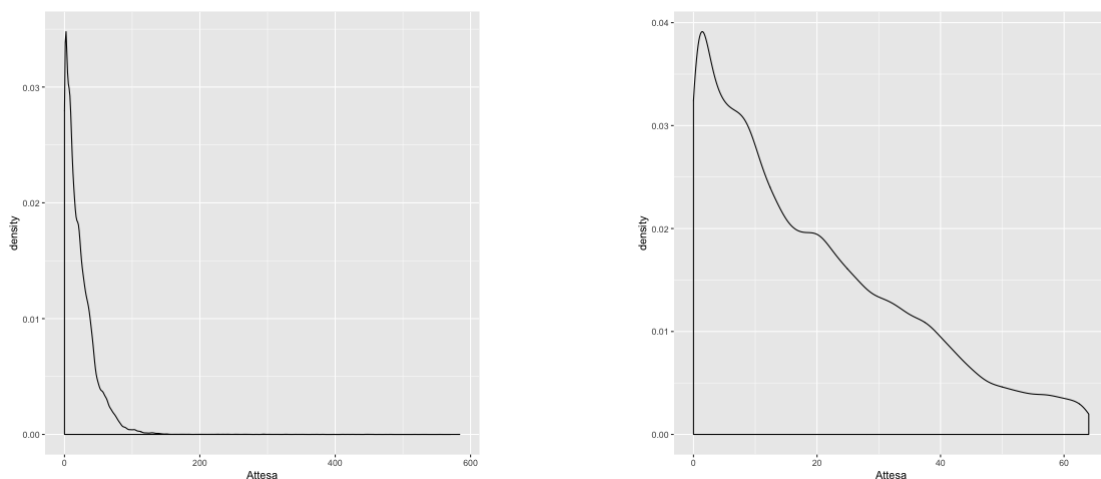
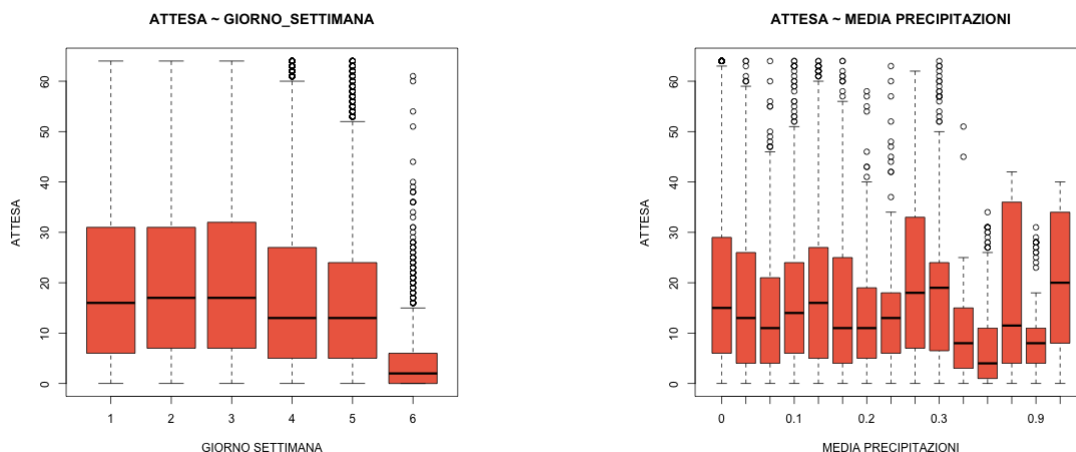
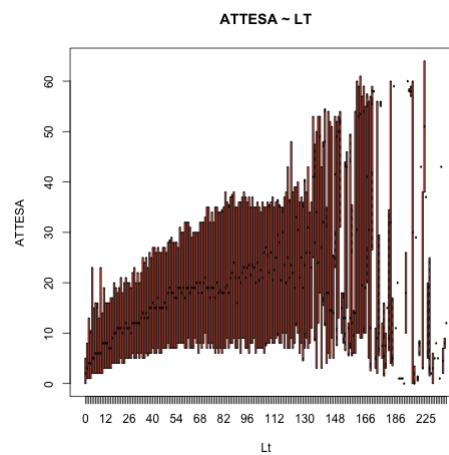
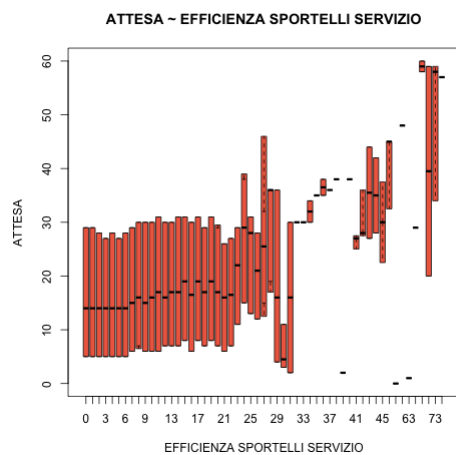
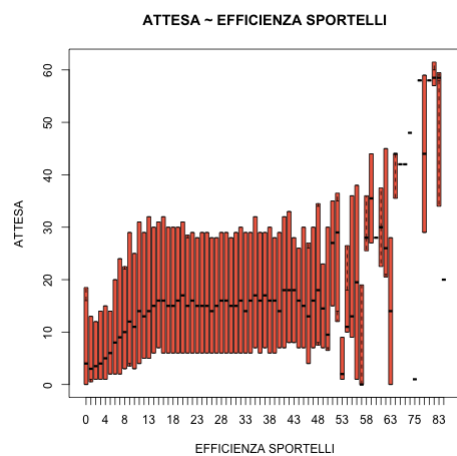
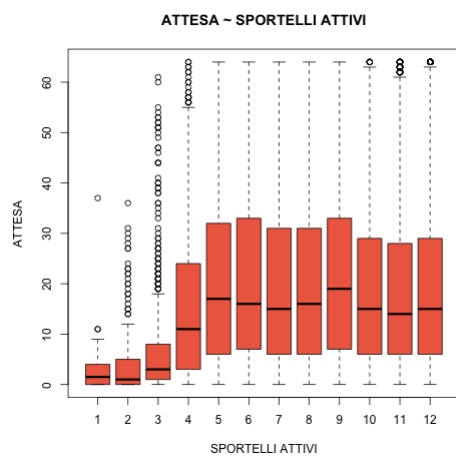
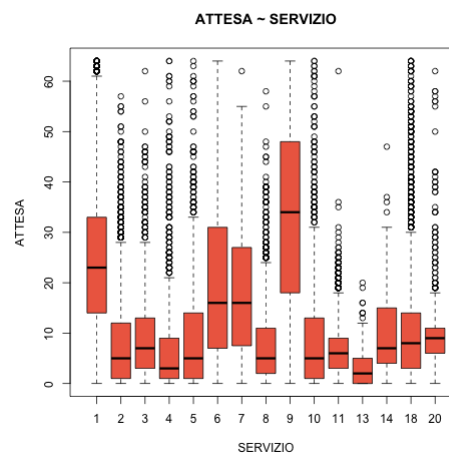
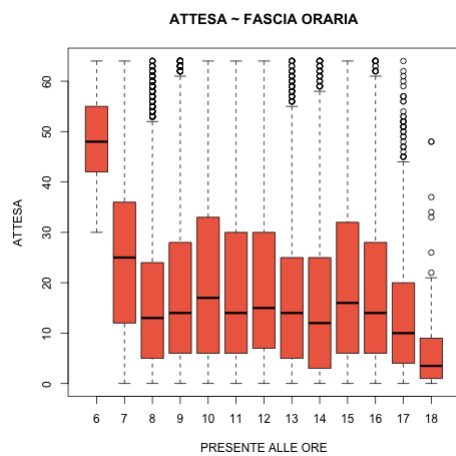


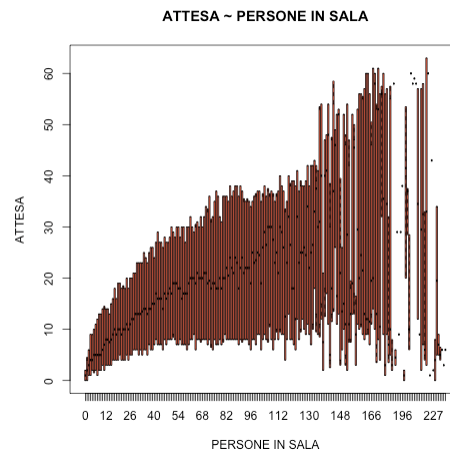
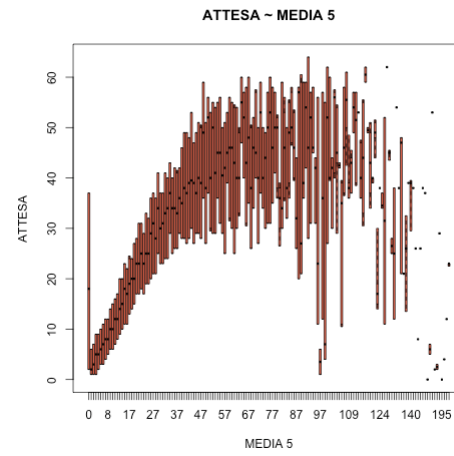
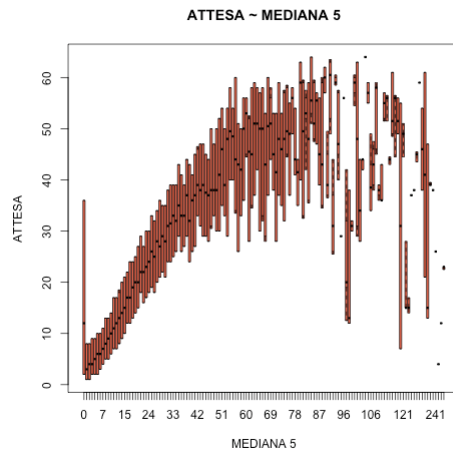
Figura 15: Distribuzione dei dati prima e dopo la rimozione degli outlier

In questa sezione si effettuerà un'analisi esplorativa dati contenuti nel training set, per capire meglio la relazione tra i diversi attributi. Sotto vengono riportati i grafici tra la variabile target attesa e le principali caratteristiche.

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



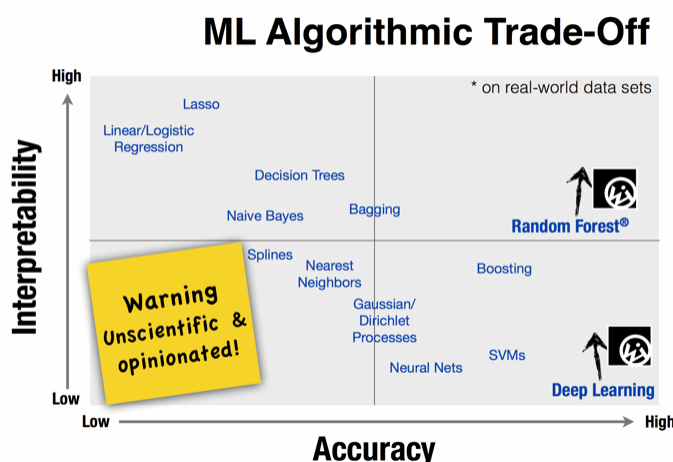




3.4 Scelta dei modelli

Confrontando i risultati ottenuti dalla feature 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 quelle identificate nell'articolo di Curtis, Catherine, et al [2], si è deciso di utilizzare random forest, uno tra i modelli che in questo studio ha ottenuto i migliori risultati.

Oltre a Random Forest si è deciso di utilizzare un Decision tree, un modello più semplice e meno accurato, ma con una maggiore interpretabilità.



Il problema è trattato sia come un problema di classificazione, sia come un problema di regressione. Per la versione di regressione si è utilizzato solo random forest mentre per la versione di classificazione si è utilizzato anche decision tree.

Per trattare il problema nella forma di classificazione è stato necessario suddividere in classi la variabile target attesa, come indicato in Tabella 12.

VALORE ATTESA IN MINUTI	CLASSE
attesa ≤ 10	1
$11 \leq \text{attesa} \leq 20$	2
$21 \leq \text{attesa} \leq 30$	3
attesa > 30	4

Tabella 12: Classificazione della variabile attesa

3.4.1 Decision tree

Inizialmente si è provato un albero di decisione sfruttando tutti gli attributi del training set e con parametri di default. L'albero ottenuto, DT1, viene mostrato in Figura 16.

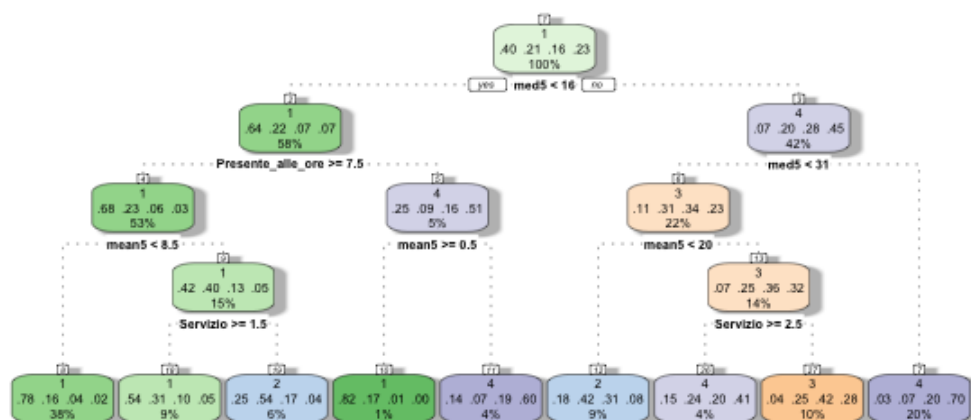


Figura 16: Rappresentazione di DT1

Come mostrato nella Figura 16, si ottiene 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 input. Le misure di performance sono riportate nella Figura 17.

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 17: Metriche DT1

Come si evince dalle misure di performance, i problemi maggiori si riscontrano nel riconoscimento delle classi 2 e 3, le quali già inizialmente avevano una numerosità minore rispetto alle altre classi.

Per cercare di ottenere un albero con accuracy migliore è stato modificato il valore CP, ovvero 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 viene mostrato in Figura 18.

Come mostrato nella Figura 18 si ottiene un albero con troppi split e si perde leggibilità, caratteristica degli alberi di decisione. Come si nota dalle metriche di performance, in Figura 19, avendo aumentato di molto il numero di split, l'accuracy è salita da un valore

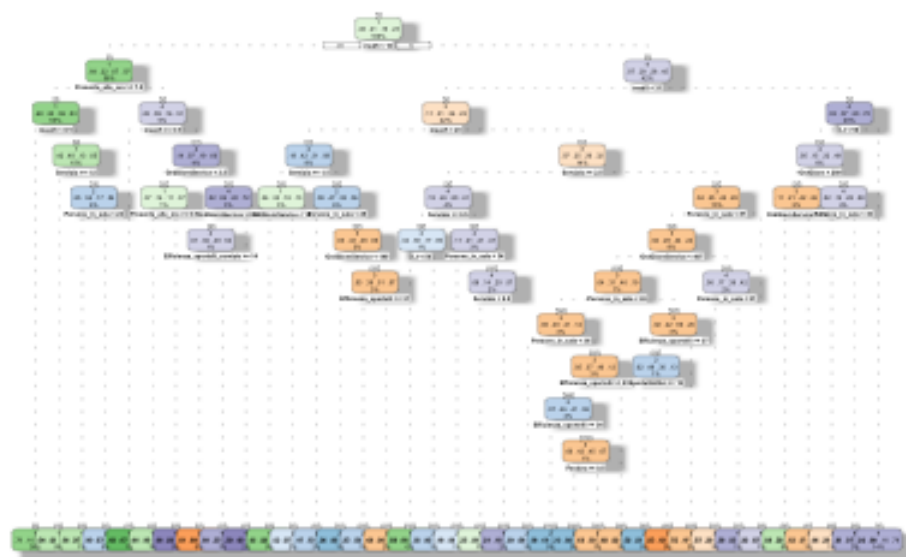


Figura 18: Rappresentazione DT2

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 19: Metriche DT2

iniziale di 0.63 ad uno di 0.66. L'incremento 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 grafico mostrato in Figura 20. 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 DT2 con un valore $CP = 0.012$, ottenendo l'albero mostrato in Figura 21.

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 in Figura 22.

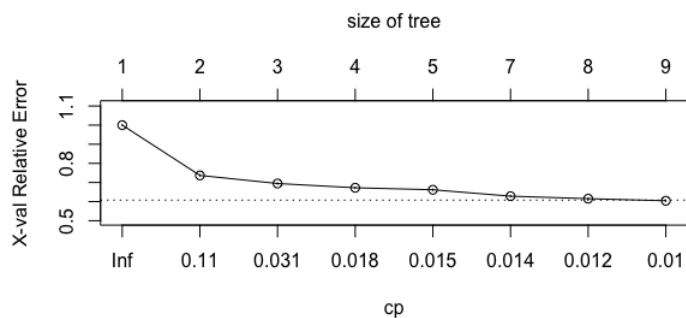


Figura 20: Stima del parametro CP

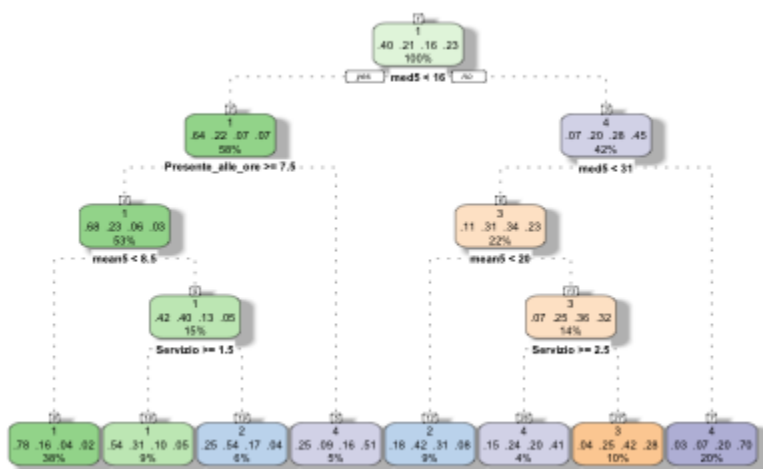


Figura 21: Rappresentazione DT3

```
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 22: Metriche DT3

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 tipo di 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 [4].

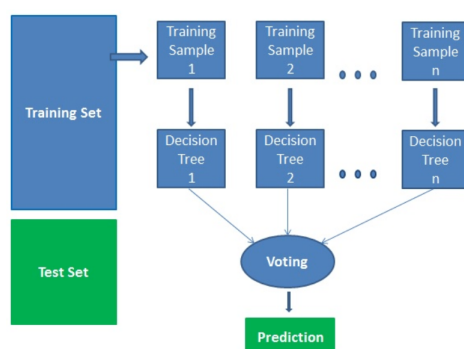


Figura 23: Algoritmo Random forest

Per eseguire random forest vengono impostati due parametri:

- **ntree**, ovvero il numero di alberi generati
- **mtry** il numero di features scelte randomicamente prima di ogni split

Per la scelta di mtry, come suggerito dal libro *"The elements of statistical learning"* [4], considerando come N il numero di features del dataset, nel caso di classificazione si utilizza un valore tra $\sqrt{N} < x < \log(N)$ mentre 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 media dell'errore di predizione di ciascun campionamento x_i dal trainset, usando solo gli alberi che non hanno x_i nel proprio *bootstrap sample* [5].

Si consideri random forest nel problema di classificazione. Per prima cosa si è eseguito random forest su tutto il dataset per stimare il valore dei parametri ottimali. Come si

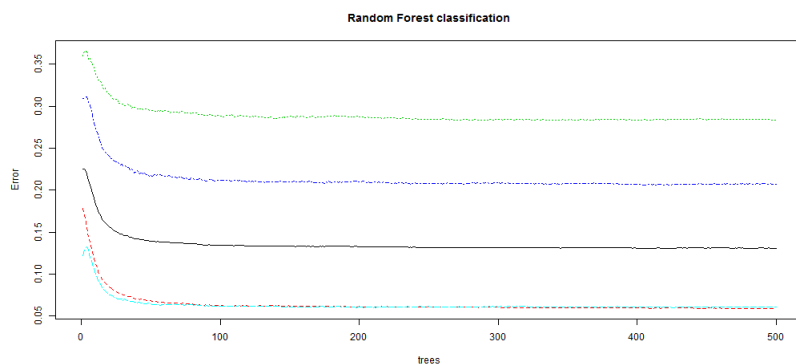


Figura 24: Stima di ntree per random forest

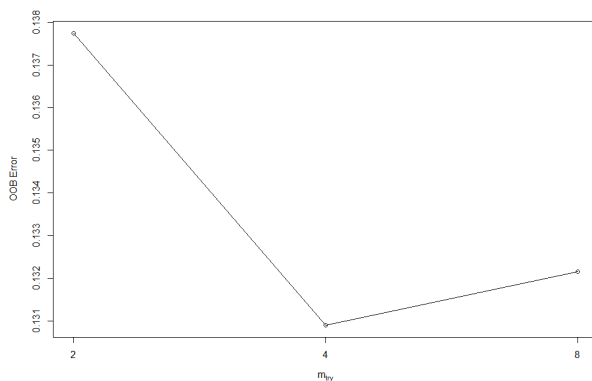


Figura 25: Stima di mtry per random forest

evince dal grafo, Figura 24, sono più che sufficienti 400 alberi per ottenere un OOB error costante per ciascuna classe.

Poi, si è verificato il valore ottimale di mtry, che come si vede nella Figura 25 risulta essere mtry=4.

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, mostrato in Figura 26, corrisponde a quanto già individuato nella feature selection.

Ora il problema viene trattato come un problema di regressione. In questo caso sono stati considerati come outlier i record con valori di attesa superiori al 99 percentile. Come fatto per 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 training set e test set sono stati suddivisi rispettivamente con il 70% e il 30% del dataset. Come mostrato nella Figura 27, anche impostando ntree con un valore superiore a 300 il modello non ottiene risultati migliori.

Per motivi di carico computazionale non è stato possibile testare il parametro mtry su

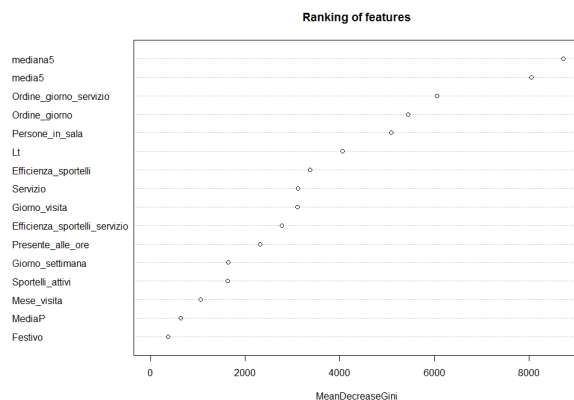


Figura 26: Feature ranking con l'indice di Gini

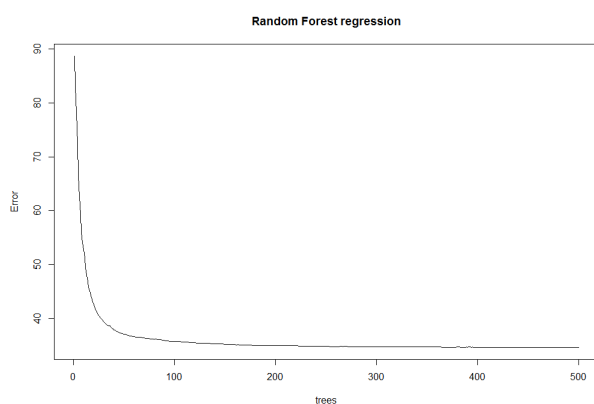


Figura 27: Stima di ntree per random forest regressione

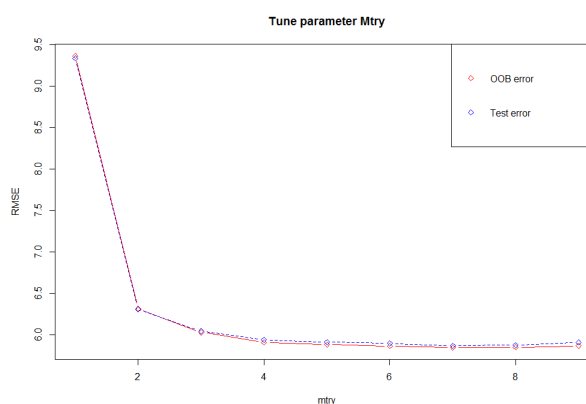


Figura 28: Stima di mtry per random forest regressione

tutti i suoi possibili valori. Quindi è stato testato su un intervallo (1; 9), ottenendo i migliori risultati con $mtry = 7$ (Figura 28).

4 Analisi dei risultati

In questo capitolo verranno analizzati i risultati ottenuti dai diversi modelli.

4.1 Performance

4.1.1 Classificazione

Per la versione di classificazione del problema, si valutano i risultati ottenuti da DT3, ovvero l'albero di decisione con $CP=0.012$ e random forest con parametri $n\text{tree}=400$ e $m\text{try}=4$.

I risultati di DT3 sono riportati nella Figura 29. Malgrado la migliore leggibilità e performance, l'albero di decisione DT3 ottiene risultati insufficienti. Come mostrato nella Roc in Figura 31, questo modello fatica a classificare valori appartenenti classe 2 e 3, cioè l'attesa compresa nella fascia 11-20 e 21-30 minuti. In generale ottiene scarsi risultati.

DT3	<pre> 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 </pre>	<pre> > accuracy_s [1] 0.8634787 > precision_s [1] 0.8427179 > recall_s [1] 0.8552429 > fmeasure_s [1] 0.8489342 > </pre>
-----	--	--

Figura 29: Confronto tra i risultati di DT3 e random forest

Per quanto riguarda il modello di random forest, nella Figura 30 vengono riportate tutte le metriche per ogni classe della variabile target attesa. Confrontando la Figura 31 e la Figura 32, si nota come il modello random forest ottiene risultati migliori rispetto all'albero di decisione. Allo stesso modo dell'albero di decisione si ottengono i risultati peggiori nella classificazione della classe 2 e 3.

	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

Figura 30: Risultati random forest per classe

4.1.2 Regressione

Il diagramma di Taylor è un modello matematico progettato per indicare graficamente quale modello di un sistema è più realistico. Per ciascun modello vengono rappre-

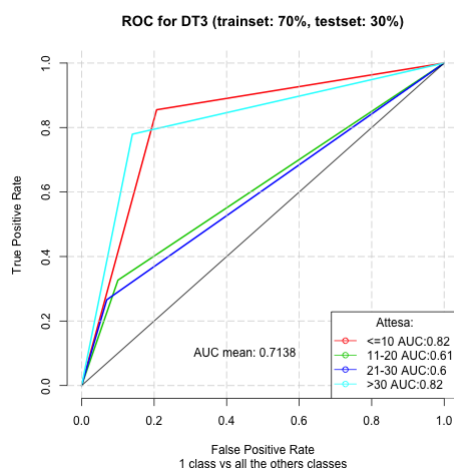


Figura 31: Roc di DT3

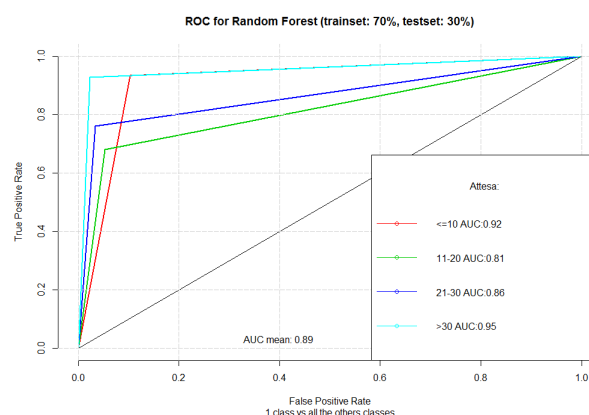


Figura 32: Roc di random forest

sentate tre statistiche, il coefficiente della correlazione di Pearson, indicato dall'angolo azimutale, la deviazione standard, evidenziata dalla distanza radiale ed infine RMSE, proporzionale alla distanza tra l'osservazione e il modello [6].

Per la versione di regressione del problema, secondo quanto si è detto nel paragrafo 3.4.2, 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 $RMSE= 5,8$ minuti con un tempo medio di predizione pari a 21,2 minuti. La Figura 33 visualizza graficamente il risultato ottenuto.

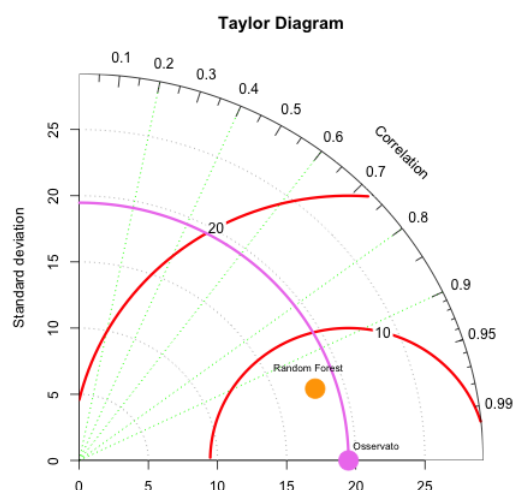


Figura 33: Risultati di random forest regressione

4.2 10-fold cross validation

Sia per il modello DT3 che per random forest si sono validate le misure di performance attraverso una 10-fold 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, ciascuno dei quali a turno è stato preso come test set e gli altri come training set, come mostrato nella Figura 34, e per ogni iterazione del modello sono state calcolate le performance. Infine è stata calcolata la media delle performance ottenute. In Figura 35 sono mostrati i grafici

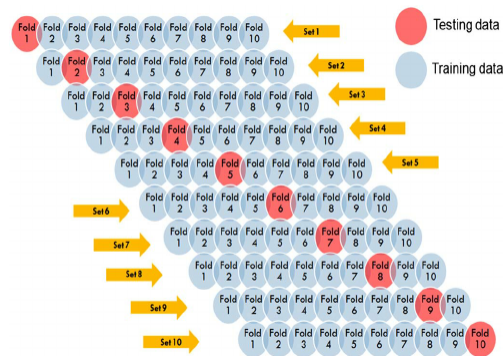


Figura 34: Algoritmo 10-fold cross validation

che rappresentano le 4 misure di performance (accuracy, precision, recall e f-measure) per ciascun modello, per ogni fold preso in considerazione.

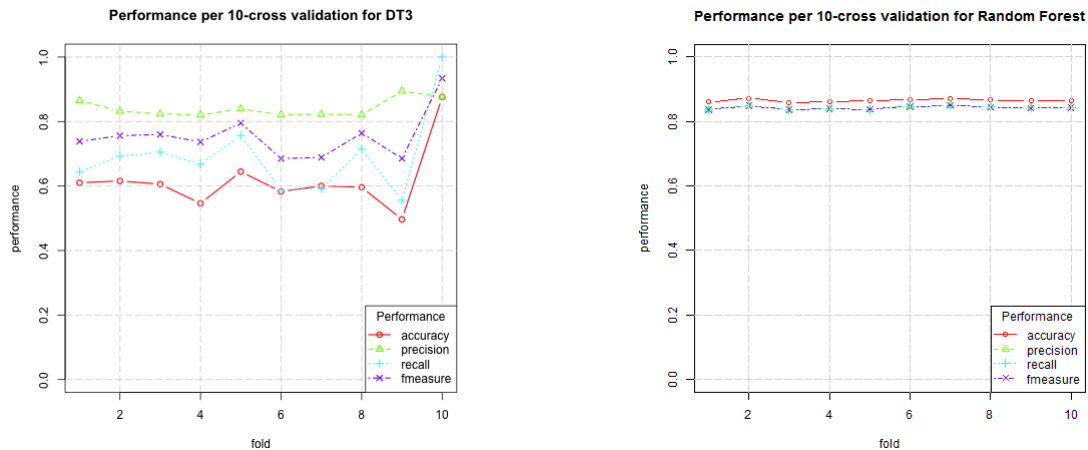


Figura 35: Risultati dei 10-fold cross validation

Nella Tabella 13, vengono mostrate le medie delle performance ottenute nella 10-fold-cross validation.

METRICA	DT3	RANDOM FOREST
accuracy	0.6174974	0.865029
precision	0.8415008	0.8424242
recall	0.6917161	0.8424242
fmeasure	0.7543822	0.8424242

Tabella 13: Performance medie registrate nella 10-fold cross validation

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.

5 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

Riferimenti bibliografici

- [1] David H Maister et al. [The psychology of waiting lines](#). Citeseer, 1984. pages 3
- [2] Catherine Curtis, Chang Liu, Thomas J Bollerman, and Oleg S Pinykh. Machine learning for predicting patient wait times and appointment delays. [Journal of the American College of Radiology](#), 15(9):1310–1316, 2018. pages 16, 22
- [3] Oleg S Pinykh and Daniel I Rosenthal. Can we predict patient wait time? [Journal of the American College of Radiology](#), 12(10):1058–1066, 2015. pages 16
- [4] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. [The elements of statistical learning](#), volume 1. Springer series in statistics New York, 2001. pages 26
- [5] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. [An introduction to statistical learning](#), volume 112. Springer, 2013. pages 26
- [6] Wikipedia contributors. Taylor_diagram — Wikipedia, the free encyclopedia, 2016. URL https://en.wikipedia.org/wiki/Taylor_diagram. [Online; accessed 10-February-2019]. pages 30