Dipartimento di Informatica, Sistemistica e Comunicazione

Mosquito Disease Spread Rate Prediction

Data Technology & Machine Learning

Giorgia Adorni
Matricola 806787

Elia Cereda
Matricola 807539

Eric Nisoli
Matricola 807147

Anno Accademico 2018-2019

Indice

I Introduzione **3**

1 Descrizione del progetto **4**
1.1 Dominio di riferimento 4
1.2 Obiettivi 4
1.3 Design del dataset 5

II Data Technology **6**

1 Dataset **7**
1.1 Acquisizione dei dataset 7
1.2 Descrizione dei tre dataset 8

2 Data Quality **14**
2.1 Completezza 14
2.1.1 Acquisizione di nuovi dati 18
2.2 Leggibilità 19
2.3 Acquisizione di nuovi dati 20

3 Data Integration **22**
3.1 Deduplication 22
3.2 Record Linkage 23
3.3 Data Fusion 24
3.3.1 Features 24

4 Analisi di almeno 2 dimensioni di qualità e relative metriche delle features successivamente utilizzate **30**
4.1 Completezza 30
4.2 Leggibilità 30

5 Analisi descrittive dei dati integrati **32**

III Machine Learning **33**

1 Creazione dei training set **34**

2 Analisi esplorativa del training set **35**

3 Descrizione e motivazione dei modelli di machine learning utilizzati **40**
3.1 Random Forest 40
3.2 Support Vector Machine 40
3.3 Neural Network 41

4 Esperimenti **42**
4.1 Random Forest 42
4.1.1 Holdout 42
4.1.2 10-fold cross validation 43
4.2 Support Vector Machine 45

5 Analisi dei risultati ottenuti **49**

6 Conclusioni **50**
6.1 Sviluppi futuri 50

Riferimenti bibliografici **51**

Parte I

Introduzione

Capitolo 1

Descrizione del progetto

1.1 Dominio di riferimento

Il virus del Nilo occidentale (noto anche come West Nile Virus, WNV) infetta ogni anno migliaia di persone, provocando nel 20% circa dei casi sintomi che variano da forti febbri a gravi complicazioni neurologiche fino ad arrivare anche alla morte.

Oltre agli esseri umani, il virus colpisce anche animali quali uccelli e cavalli, causando in questi ultimi tassi di mortalità che raggiungono il 40%. Isolato per la prima volta nel 1937 nel distretto di West Nile in Uganda, da cui prende il nome, si è oggi diffuso in tutto il mondo.

Per questi motivi, il controllo e la prevenzione delle infezioni da WNV risultano essere argomenti di grande interesse.

Le zanzare infette costituiscono il principale vettore di trasmissione del virus agli esseri umani. Quando nel 2002 furono riportati i primi casi umani a Chicago, il Dipartimento di Sanità Pubblica della città ha avviato un esteso programma di sorveglianza e controllo che resta tutt'oggi in vigore.

Dalla fine della primavera all'inizio dell'autunno, numerose trappole per zanzare vengono distribuite sul territorio, andando ogni settimana a testare la presenza del virus negli esemplari catturati. Sulla base di queste informazioni, la città programma l'utilizzo di insetticidi per controllare la popolazione di zanzare adulte.

1.2 Obiettivi

Con una media di 91 trappole attive per 14 settimane ogni anno, questo programma presenta significativi costi di raccolta dei campioni ed esecuzione dei test clinici per determinare la presenza del virus.

L'obiettivo di questo elaborato è progettare un modello di classificazione supervisionata capace di prevedere se ogni settimana verrà rilevato o meno il virus in

una certa trappola. Le predizioni di questo modello permetteranno di indirizzare gli sforzi di raccolta verso quelle trappole in cui ci si aspetta di trovare esemplari infetti con probabilità maggiore.

Dato che i test effettuati negli anni sono risultati positivi al virus solo nell'8,6% dei casi, un modello sufficientemente accurato ha le potenzialità di ridurre fortemente il numero di trappole da analizzare ogni settimana e di conseguenza i costi del programma.

Si ritene che un clima caldo e secco sia favorevole alla diffusione del WNV. Per questo motivo le predizioni si basano principalmente su dati relativi alle condizioni meteorologiche e all'ubicazione delle trappole.

1.3 Design del dataset

Il problema descritto è stato proposto dalla città di Chicago sulla piattaforma Kaggle[1] nel 2015. Rispetto a quanto richiesto in quella competizione, per questo progetto abbiamo però adottato una strategia leggermente diversa.

Abbiamo potuto constatare che i dati forniti su Kaggle erano già stati oggetto di un'elaborazione preliminare e le attività svolte si sovrapponevano parzialmente a quanto richiesto per il modulo di Data Technology. Abbiamo quindi deciso di non utilizzare questi dataset, ma piuttosto ricostruirli a partire dalle loro fonti originali.

Siamo riusciti ad individuare i dati necessari sui portali Open Data del Dipartimento di Sanità Pubblica di Chicago (CDPH) e della National Oceanic and Atmospheric Administration (NOAA). Come descritto nei Capitoli 2 e 3, abbiamo poi effettuato un'analisi di qualità dei dati grezzi, la correzione delle anomalie rilevate e l'integrazione delle due fonti.

Ricostruendo i dataset abbiamo inoltre potuto ottenere dati più recenti di quelli pubblicati su Kaggle, permettendoci di analizzare il periodo 2007–2017 invece che limitarci agli anni 2007–2013.

Uno svantaggio dell'approccio adottato è che, basandosi su dati differenti, i nostri risultati non potranno essere confrontati direttamente con lo stato dell'arte visibile sulla classifica della competizione. **FIXME completare questa parte quando abbiamo effettivamente dei risultati**

[1] `https://www.kaggle.com/c/predict-west-nile-virus`

Parte II

Data Technology

Capitolo 1

Dataset

1.1 Acquisizione dei dataset

Il dataset WNV Mosquito contiene la posizione di ogni trappola installata nella città attraverso il programma di Salute Ambientale del Dipartimento di Sanità Pubblica di Chicago negli anni tra il 2007 e il 2018. Contiene inoltre la specie e il numero di zanzare trovate in ciascuna trappola ogni settimana, nonchè i risultati dei test per il virus del Nilo occidentale. È stato reperito dal sito Chicago Data Portal[1].

I dataset Weather e Stations provengono invece dal database NOAA Quality Controlled Local Climatological Data (QCLCD) disponibile sul sito dalla National Oceanic and Atmospheric Administration[2]. I dati grezzi sono composti da un file CSV per ciascun mese nel periodo 2007–2017, al cui interno sono presenti le rilevazioni di tutte le stazioni meteorologiche degli Stati Uniti.

Come prima cosa abbiamo scritto una serie di script Bash con i quali estrarre esclusivamente le stazioni meteorologiche nei pressi della città di Chicago e aggregarle in un unico file contenente l'intero arco temporale di interesse. Attraverso un ulteriore script Bash abbiamo analizzato le colonne presenti nei file CSV, per accertarci che la loro struttura fosse rimasta costante negli anni.

Si può notare che vi è una discrepanza nei periodi temporali coperti dalle due fonti, in quanto i dati QCLCD 2018 non sono ad oggi disponibili online. Come verrà descritto più in dettaglio nella Sezione 3.3, questo ci ha costretto a limitarci al periodo 2007–2017, escludendo quindi l'anno 2018 dal dataset WNV Mosquito.

[1] `https://data.cityofchicago.org/Health-Human-Services/West-Nile-Virus-WNV-Mosquito-Test-Results/jqe8-8r6s/data`

[2] `https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/quality-controlled-local-climatological-data-qclcd`

1.2 Descrizione dei tre dataset

Vengono di seguito descritti gli attributi di ciascun dataset.

Attributo	**Descrizione**	**Valori**
season_year	anno del test	categorico 2007 – 2017
week	settimana dell'anno del test	numerico 1 – 52 (20 – 40)
test_id	id del record	numerico
block	indirizzo della trappola	xxXX = numero civico STREET = nome della via xxXX STREET
trap	id della trappola	identificatore T001X – T925X
trap_type	tipologia di trappola	OVI CVC GRAVID SENTINEL
test_date	data del test	mm/dd/yyyy hh:mm:ss P
number_of_mosquitoes	numero di zanzare catturate	numerico 1 – 50
result	risultato del test	positive = WNV presente negative = WNV non presente
species	specie di zanzare	CULEX PIPIENS/RESTAUANS CULEX PIPIENS CULEX RESTAUANS CULEX ERRATICUS CULEX SALINARIUS CULEX TARSALIS CULEX TERRITANS UNSPECIFIED CULEX
latitude	latitudine dell'indirizzo	decimale 40.00 – 42.00
longitude	longitudine dell'indirizzo	decimale -87.00 – -88.00

Tabella 1.2: Attributi del dataset WNV Mosquito

Attributo	**Descrizione**	**Valori**
wban	id della stazione	numerico
year_month_day	data della rilevazione	yyyymmdd
t_max	temperatura massima giornaliera (°F)	numerico (*) estremo del mese
t_max_flag	qualità t_max	blank = valore non sospetto s = valore sospetto
t_min	temperatura minima giornaliera (°F)	numerico (*) estremo del mese
t_min_flag	qualità t_min	blank = valore non sospetto s = valore sospetto
t_avg	temperatura media giornaliera (°F)	numerico
t_avg_flag	qualità t_avg	blank = valore non sospetto s = valore sospetto
depart	scostamento dalla temperatura normale (°F)	numerico
depart_flag	qualità depart	blank = valore non sospetto s = valore sospetto
dew_point	temperatura, a pressione costante, in cui l'aria diventa satura (°F)	numerico
dew_point_flag	qualità dew_point	blank = valore non sospetto s = valore sospetto
wet_bulb	temperatura di equilibrio di scambio convettivo e di massa d'aria in moto turbolento dell'acqua (°F)	numerico
wet_bulb_flag	qualità wet_bulb	blank = valore non sospetto s = valore sospetto
heat	numero di gradi in cui la temperatura media giornaliera è inferiore a 65 °F (temperatura al di sotto della quale gli edifici devono essere riscaldati)	numerico
heat_flag	qualità heat	blank = valore non sospetto s = valore sospetto

Continua nella pagina seguente

Continua dalla pagina precedente

Attributo	**Descrizione**	**Valori**
cool	numero di gradi in cui la temperatura media giornaliera è superiore a 65 °F (temperatura al di sopra della quale gli edifici devono essere raffreddati)	numerico
cool_flag	qualità cool	blank = valore non sospetto s = valore sospetto
sunrise	ora dell'alba (LST)	hhmm
sunrise_flag	qualità sunrise	blank = valore non sospetto s = valore sospetto
sunset	ora del tramonto (LST)	hhmm
sunset_flag	qualità sunset	blank = valore non sospetto s = valore sospetto
code_sum	lista di condizioni meteorologiche giornaliere	BR = foschia DU = polvere DZ = pioggerella FC = nubi a imbuto FG = nebbia FU = fumo GR = grandine GS = piccola grandine HZ = foschia PL = pioggia gelata RA = pioggia SN = neve SQ = raffiche di vento TS = temporale UP = sconosciuto
code_sum_flag	qualità code_sum	blank = valore non sospetto s = valore sospetto
depth	quantità di neve depositata a terra (millimetri)	numerico
depth_flag	qualità depth	blank = valore non sospetto s = valore sospetto
water1	quantità di acqua ottenuta sciogliendo la neve a terra (millimetri)	numerico

Continua nella pagina seguente

Continua dalla pagina precedente

Attributo	**Descrizione**	**Valori**
water1_flag	qualità water1	blank = valore non sospetto s = valore sospetto
snow_fall	quantità giornaliera di neve caduta (inch)	decimale
snow_fall_flag	qualità snow_fall	blank = valore non sospetto s = valore sospetto
precip_total	quantità di precipitazioni giornaliere (inch)	decimale
precip_total_flag	qualità precip_total	blank = valore non sospetto s = valore sospetto
stn_pressure	pressione media giornaliera (pollici di mercurio)	decimale
stn_pressure_flag	qualità stn_pressure	blank = valore non sospetto s = valore sospetto
sea_level	pressione media giornaliera a livello del mare (pollici di mercurio)	decimale
sea_level_flag	qualità sea_level	blank = valore non sospetto s = valore sospetto
result_speed	velocità di punta giornaliera del vento (miglia all'ora)	decimale
result_speed_flag	qualità result_speed	blank = valore non sospetto s = valore sospetto
result_dir	direzione giornaliera del vento durante la velocità di punta	numerico
result_dir_flag	qualità result_dir	blank = valore non sospetto s = valore sospetto
avg_speed	velocità media giornaliera del vento (miglia all'ora)	decimale
avg_speed_flag	qualità avg_speed	blank = valore non sospetto s = valore sospetto
max5_speed	velocità massima giornaliera raggiunta dal vento per 5 minuti (miglia all'ora)	numerico
max5_speed_flag	qualità max5_speed	blank = valore non sospetto s = valore sospetto

Continua nella pagina seguente

Continua dalla pagina precedente

Attributo	Descrizione	Valori
max5_dir	direzione giornaliera del vento durante la massima velocità raggiunta per 5 minuti	numerico
max5_dir_flag	qualità max5_dir	blank = valore non sospetto s = valore sospetto
max2_speed	velocità massima giornaliera raggiunta dal vento per 2 minuti (miglia all'ora)	numerico
max2_speed_flag	qualità max2_speed	blank = valore non sospetto s = valore sospetto
max2_dir	direzione giornaliera del vento durante la massima velocità raggiunta per 2 minuti	numerico
max2_dir_flag	qualità max2_dir	blank = valore non sospetto s = valore sospetto

Tabella 1.4: Attributi del dataset WEATHER

Attributo	Descrizione	Valori
wban	codice WBAN della stazione	numerico
wmo	codice WMO della stazione	numerico
callsign	codice IATA dell'aeroporto in cui si trova la stazione	categorico
climate_division_code climate_division_state_code climate_division_station_code	codici che identificano la regione climatica a cui la stazione appartiene	numerico numerico numerico
name	nome della stazione	categorico
state	sigla dello stato	categorico
location	posizione della stazione	categorico
latitude	latitudine della stazione	decimale 40.00 – 42.00
longitude	longitudine della stazione	decimale -87.00 – -88.00
ground_height	altitudine del terreno sul	decimale

Continua nella pagina seguente

Continua dalla pagina precedente

Attributo	**Descrizione**	**Valori**
	livello del mare (piedi)	
station_height	altitudine della stazione sul livello del mare (piedi)	decimale
barometer	altitudine del barometro sul livello del mare (piedi)	decimale
time_zone	fuso orario della stazione	numerico

Tabella 1.6: Attributi del dataset STATIONS

Capitolo 2

Data Quality

Abbiamo sottoposto i dataset ad un processo di analisi finalizzato all'esplorazione dei dati e all'incremento della loro qualità. In particolare, data la natura dei dati e il nostro scopo, siamo ricorsi alle dimensioni di completezza, consistenza e leggibilità.

2.1 Completezza

La prima dimensione di qualità analizzata è la completezza, ovvero la copertura con cui vengono rappresentati gli insiemi di dati. Le metriche misurate sono rispettivamente:

- Completezza di attributo: il rapporto tra il numero di valori non nulli nella colonna e la cardinalità della colonna (ovvero il numero di tuple);
- Completezza di tupla: il rapporto tra il numero di tuple che non contengono nessun valore nullo e il numero totale di tuple;
- Completezza di tabella: il rapporto tra il numero di valori non nulli nella tabella e la cardinalità della tabella (ovvero il prodotto tra numero di tuple e numero di colonne).

La completezza complessiva dei quattro dataset è del 93,72%, mentre la completezza delle tuple è solo del 56,86%.

Dataset	Completezza di tupla	Completezza dataset
WNV Mosquito	84,89%	97,48%
Weather	0,00%	86,09%
Stations	20,00%	82,67%
Totale	56,86%	93,72%

Tabella 2.1: Analisi di completezza sui tre dataset

Vengono in seguito mostrate le analisi relative ai dati di ciascun dataset.

WNV Mosquito

Il dataset WNV Mosquito presenta una completezza del 97,48%. Esso risulta essere il più completo tra i dataset presentati. In particolare solo due attributi su dodici, `latitude` e `longitude`, presentano dei valori mancanti.
In tabella 2.2 sono presentate le analisi di completezza relative a ciascun attributo e la percentuale della completezza di tupla e di schema.

Attributo	Istanze	Valori nulli	Completezza
season_year	27196	0	100%
week	27196	0	100%
test_id	27196	0	100%
block	27196	0	100%
trap	27196	0	100%
trap_type	27196	0	100%
test_date	27196	0	100%
number_of_mosquitoes	27196	0	100%
result	27196	0	100%
species	27196	0	100%
latitude	27196	4108	84,89%
longitude	27196	4108	84,89%
Tuple	27196	4108	84,89%
Dataset	326352	8216	97,48%

Tabella 2.2: Analisi di completezza sul dataset WNV Mosquito

Weather

Il dataset WEATHER presenta una completezza del 86,09%. In particolare 26 attributi su 50 risultano completi, mentre tutti gli altri presentano dei valori mancanti. In tabella 2.4 sono presentate le analisi di completezza relative a ciascun attributo e la percentuale della completezza di tupla e di schema.

Attributo	**Istanze**	**Valori nulli**	**Completezza**
wban	13400	0	100,00%
year_month_day	13400	0	100,00%
t_max	13400	1723	87,14%
t_max_flag	13400	0	100,00%
t_min	13400	1723	87,14%
t_min_flag	13400	0	100,00%
t_avg	13400	1807	86,51%
t_avg_flag	13400	0	100,00%
depart	13400	9561	28,65%
depart_flag	13400	0	100,00%
dew_point	13400	61	99,54%
dew_point_flag	13400	0	100,00%
wet_bulb	13400	172	98,72%
wet_bulb_flag	13400	0	100,00%
heat	13400	1807	86,51%
heat_flag	13400	0	100,00%
cool	13400	1807	86,51%
cool_flag	13400	0	100,00%
sunrise	13400	9533	28,86%
sunrise_flag	13400	0	100,00%
sunset	13400	9533	28,86%
sunset_flag	13400	0	100,00%
code_sum	13400	6250	53,36%
code_sum_flag	13400	0	100,00%
depth	13400	10922	18,49%
depth_flag	13400	0	100,00%
water1	13400	13400	0,00%
water1_flag	13400	0	100,00%
snow_fall	13400	10912	18,57%
snow_fall_flag	13400	0	100,00%
precip_total	13400	1745	86,98%
precip_total_flag	13400	0	100,00%

Continua nella pagina seguente

Continua dalla pagina precedente

Attributo	**Istanze**	**Valori nulli**	**Completezza**
stn_pressure	13400	1749	86,95%
stn_pressure_flag	13400	0	100,00%
sea_level	13400	1771	86,78%
sea_level_flag	13400	0	100,00%
result_speed	13400	41	99,69%
result_speed_flag	13400	0	100,00%
result_dir	13400	41	99,69%
result_dir_flag	13400	0	100,00%
avg_speed	13400	1736	87,04%
avg_speed_flag	13400	0	100,00%
max5_speed	13400	1728	87,10%
max5_speed_flag	13400	0	100,00%
max5_dir	13400	1728	87,10%
max5_dir_flag	13400	0	100,00%
max2_speed	13400	1719	87,17%
max2_speed_flag	13400	0	100,00%
max2_dir	13400	1719	87,17%
max2_dir_flag	13400	0	100,00%
Tuple	13400	13400	0,00%
Dataset	670000	93188	86,09%

Tabella 2.4: Analisi di completezza sul dataset WEATHER

Stations

Il dataset STATIONS presenta una completezza del 82,67%. In percentuale esso risulta essere il meno completo tra i dataset presentati, in valori assoluti è però molto più piccolo degli altri, dato che contiene solo 5 tuple. Inoltre, dei quattro attributi su 15 che presentano dei valori mancanti, possiamo trascurare `wmo`, `climate_division_code` e `climate_division_station_code`, trattandosi di codici identificativi delle stazioni meteorologiche che non sono importanti per i nostri scopi.

In tabella 2.5 sono presentate le analisi di completezza relative a ciascun attributo, alle tuple e all'intero schema.

Attributo	Istanze	Valori nulli	Completezza
wban	5	0	100,00%
wmo	5	3	40,00%
callsign	5	0	100,00%
climate_division_code	5	4	20,00%
climate_division_state_code	5	0	100,00%
climate_division_station_code	5	4	20,00%
name	5	0	100,00%
state	5	0	100,00%
location	5	0	100,00%
latitude	5	0	100,00%
longitude	5	0	100,00%
ground_height	5	0	100,00%
station_height	5	0	100,00%
barometer	5	2	60,00%
time_zone	5	0	100,00%
Tuple	5	4	20,00%
Dataset	75	13	82,67%

Tabella 2.5: Analisi di completezza sul dataset STATIONS

2.1.1 Acquisizione di nuovi dati

Per migliorare la completezza abbiamo effettuato un completamento dei dati mancanti con tecniche che sfruttano la conoscenza del dominio specifico.
Nel dataset WNV MOSQUITO è stato possibile popolare i valori `latitude` e `longitude` mancanti, utilizzando il servizio di **Geocoding**[1] offerto da Google Maps. In particolare, abbiamo richiesto le coppie di latitudine e longitudine corrispondenti agli indirizzi presenti nella colonna `block`. Per verificare l'accuratezza delle posizioni ottenute è stata calcolata la distanza geografica tra le coordinate presenti e quelle ottenute dal servizio. In questo modo abbiamo potuto verificare che il 95% delle istanze si trova ad una distanza inferiore a 925 metri.
Essendo queste posizioni utilizzate principalmente per individuare le stazioni meteorologiche più vicine ad una data trappola, abbiamo ritenuto sufficiente l'accuratezza dei risultati del servizio di Geocoding. Quindi, sfruttando le coordinate ottenute, sono stati popolati i campi mancanti nel dataset originale, raggiungendo una completezza di schema, tupla e attributi pari al 100,00%.

[1] `https://developers.google.com/maps/documentation/geocoding/start`

2.2 Leggibilità

La leggibilità è una dimensione di qualità che misura la facilità e l'immediatezza di comprensione di una determinata rappresentazione dei dati. Una rappresentazione dalla buona leggibilità si presta ad analisi (almeno sommarie) da parte degli utilizzatori dei dati, che possono in questo modo trarne informazioni utili ai loro processi decisionali.
I dataset grezzi scelti per il progetto presentano diverse problematiche sotto questo punto di vista. In primo luogo, le unità di misura utilizzate sono quelle del sistema anglosassone, che risultano di difficile lettura al di fuori degli Stati Uniti. Inoltre, vi sono attributi che impiegano codici alfanumerici impossibili da interpretare senza consultare la documentazione.
Nel seguito discutiamo le modifiche che abbiamo apportato ai vari dataset per migliorarne la leggibilità.

- WEATHER:
  - Questo dataset utilizza il carattere `M` per indicare i valori mancanti. È stato scelto di sostituirlo con il valore `null` poiché più esplicito e poiché riconosciuto automaticamente anche dai software di analisi dei dati e dai DBMS.
  - `date`: inizialmente l'attributo rappresentava una data attraverso il pattern *yyyyMMdd*, per migliorarne la leggibilità è stato trasformato nella forma *dd/MM/yyyy*. Ad esempio la data *20070501* è stata modificata in *01/05/2007*.
  - `t_max`, `t_min`, `t_avg`, `depart`, `dew_point`, `wet_bulb`, `heat`, `cool`: i valori di temperatura sono stati convertiti da gradi Fahrenheit (°F) a gradi Celsius (°C).
  - `snow_depth`, `snow_water`, `snow_fall`, `precip_total`: i valori sono stati convertiti da pollici (inch) a millimetri (mm).
  - `stn_pressure`, `sea_level`: i valori di pressione sono stati convertiti da pollici di mercurio (inHg) a millimetri di mercurio (mmHg).
  - `result_speed`, `avg_speed`, `max5_speed`, `max2_speed`: i valori di velocità sono stati convertiti da miglia orarie (mph) a chilometri orari (km/h).
  - `code_sum`: questo attributo corrisponde ad una lista di codici che possono essere assegnati a ciascuna rilevazione, indicanti i diversi tipi di precipitazioni e altri fenomeni meteorologici. I codici alfanumerici originali sono stati trasformati nei corrispondenti nomi in linguaggio naturale (ad es. pioggia, neve, nebbia, ...). Inoltre, dato che a ciascuna rilevazione può

essere assegnato più di un codice, abbiamo trasformato la singola colonna del dataset originale in una colonna per ciascun codice possibile, in cui un valore booleano indica la presenza o l'assenza della rispettiva condizione meteo.

- STATION:
  - `ground_height`, `station_height`, `barometer`: i valori di altitudine sono stati convertiti da piedi (feet) a metri (m).
- WNV MOSQUITO:
  - `trap`: nel dataset le trappole vengono contrassegnate dalla lettera T seguita da tre cifre. Trappole "satellite" sono spesso installate vicino ad una trappola principale per potenziare la sorveglianza. Per questo motivo sono indicate dallo stesso codice della trappola principale seguito da una lettera. Ad esempio T220A è la prima trappola satellite per T220. Essendo questo di difficile lettura sono state aggiunte due colonne: `main_trap` che riporta il codice della trappola principale e `sub_trap` che contiene nel caso delle trappole satellite la lettera ad essa associata.
  - `result`: inizialmente la colonna conteneva dei valori testuali, *positive* e *negative*, che sono stati trasformati in valori booleani.

2.3 Acquisizione di nuovi dati

Nel dataset WNV MOSQUITO abbiamo voluto aggiungere l'attributo `day_of_week` che corrisponde al giorno della settimana in cui è stato effettuato il test. questo è stato utile anche per derivare delle statistiche, come per esempio il giorno della settimana in cui è stato effettuato il maggior numero di test.

Un analisi del dataset WEATHER ha mostrato come alcuni codici della colonna `code_sum` comparissero molto raramente. Per rendere questi valori più adatti all'utilizzo come feature, abbiamo deciso di aggregarli manualmente in categorie più ampie, in base alla somiglianza dei fenomeni meteorologici rappresentati:

- I codici `GR` (grandine, 1 occorrenza), `GS` (piccola grandine, 3 occorrenze) e `PL` (pioggia gelata, 33 occorrenze) sono stati trasformati nel codice aggregato `HAIL` (grandine, 37 occorrenze).

- I codici `SQ` (raffiche di vento, 8 occorrenze), `FC` (nubi a imbuto, 1 occorrenza) sono stati trasformati nel codice aggregato `WIND` (forte vento, 9 occorrenze).
- I codici `FU` (fumo, 19 occorrenze), `DU` (polvere, 1 occorrenza) e sono stati trasformati nel codice aggregato `SMOKE` (fumo e polvere, 20 occorrenze).
- Tutti i rimanenti codici compaiono più di 110 volte nell'arco temporale analizzato e sono stati mantenuti invariati.

Nello stesso dataset, a seguito dell'aggregazione delle osservazioni giornaliere in settimanali, è stata aggiunta la colonna `days_per_week`, rappresentante il numero di giorni settimanali aggregati. Nella maggior parte dei casi questo valore corrisponde a 7, ovvero l'aggregazione di tutti i giorni della settimana.

Capitolo 3

Data Integration

3.1 Deduplication

In seguito alla valutazione dei dataset tramite le dimensioni di qualità è stata eseguita l'operazione di deduplicazione sui singoli dataset, ovvero l'identificazione di quelle coppie o gruppi di tuple corrispondenti ad uno stesso oggetto del mondo reale.

È stato utilizzato il software Power BI di Microsoft, in particolare ricorrendo allo strumento Power Query, per l'inserimento, la trasformazione, l'integrazione e l'arricchimento dei dati.

È stata effettuata la deduplicazione del dataset WNV Mosquito, analizzando l'attributo `block`. In particolare è stata riscontrata la presenza di alcune tuple duplicate, che differivano per i campi `latitude` e `longitude`. Calcolando la distanza tra le diverse coordinate abbiamo verificato che le differenze erano solo di alcuni metri. Dunque è stato deciso di tenere **FIXME randomicamente** una delle tuple duplicate data la scarsa differenza tra le due e la poca importanza della precisione, poiché nel nostro modello predittivo le coordinate di ogni indirizzo vengono utilizzate solo per cercare una corrispondenza con la stazione di rilevamento meteo più vicina.

Sempre sul dataset WNV Mosquito è stata analizzata la duplicazione delle colonne `latitude`, `longitude` e `block` rispetto a `trap`. In questo caso abbiamo rilevato la presenza di una stessa trappola situata in luoghi diversi. Aggiungendo l'attributo `season_year` non abbiamo più riscontrato tuple duplicate. Abbiamo quindi ipotizzato che la stessa trappola fosse stata posizionata diversamente negli anni.

Una situazione analoga si verifica analizzando il campo `trap_type` rispetto a `trap`. Vengono rilevate 17 tuple duplicate, ma ancora una volta è possibile discriminarle attraverso l'attributo `season_year`.

3.2 Record Linkage

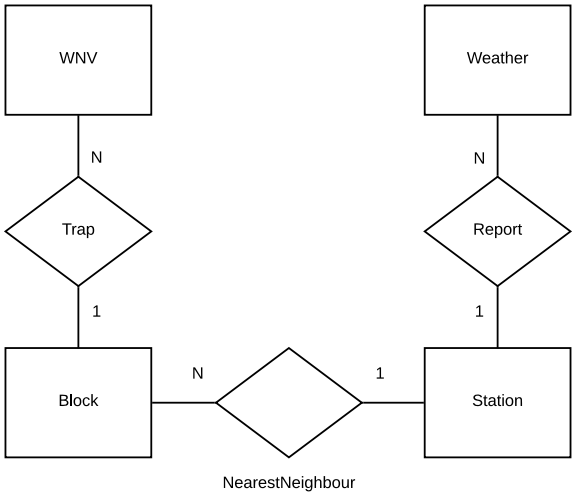


Figura 3.1: Schema concettuale per il record linkage

L'analisi dei dataset ha portato alla stesura dello schema concettuale riportato in Figura 3.1 per le operazioni di record linkage, in particolare sono state definite le seguenti chiavi primarie per le entità:

- WNV: (`test_id`, `season_year`, `week`).
- BLOCK: (`block`).
- STATION: (`wban`).
- WEATHER: (`wban`, `year`, `week_of_year`).

Le relazioni tra entità invece si basano sulle seguenti chiavi e in particolare per la relazione **NearestNeighbour** sul concetto di vicinanza fra elementi distribuiti nello spazio.

- **Trap**: WNV(`block`), BLOCK(`block`).
- **Observation**: WEATHER(`wban`), STATION(`wban`).

– **NearestNeighbour**: il concetto di vicinanza spaziale è ottenuto effettuando il prodotto cartesiano per tutte le coppie BLOCK×STATION e calcolando la distanza tra le coordinate geografiche ottenuti dagli attributi `latitude` e `longitude` sia di BLOCK che di STATION, coinvolgendo nella relazione solamente la STATION più vicina ad un BLOCK.

È stato scelto di utilizzare la tecnica di record linkage empirico, dato che le chiavi scelte per le relazioni non presentano errori, dunque la distanza di edit pari a zero tra le possibili corrispondenze porta sempre ad un matching esatto delle chiavi.

3.3 Data Fusion

Il processo di data fusion consiste nell'esecuzione di una query di join fra le entità definite nella Sezione 3.2 attraverso le corrispondenti relazioni individuate e descritte.

Non è stato necessario applicare nessuna tecnica di risoluzione dei conflitti poiché nei dataset utilizzati non vi sono ambiguità.

FIXME I record di weather e wnv descrivono proprietà dell'ambiente misurate in un certo momento temporale: in un caso le condizioni meteorologiche, nell'altro quante zanzare sono state catturate.

FIXME bo magari dovremmo scrivere qualche altra cosa

FIXME: dobbiamo spiegare che delle stazioni meteo ne abbiamo potute tenere solo 2 e che di WNV abbiamo dovuto escludere il 2018 perché non avevamo il meteo

3.3.1 Features

Dopo le operazioni di integrazione dei quattro dataset, tre iniziali e uno generato, sono state selezionate 66 features.

In particolare sono stati eliminati dal dataset STATIONS i record relativi alle stazioni con identificativo `wban` 4807 e 4879, poiché essendo attive dal 2015 non forniscono alcuna rilevazione dei dati meteo per i primi 8 anni in cui sono stati effettuati i test del WNV.
Inoltre di questa tabella non sono state considerate come features le colonne `wban`, `wmo`, `call_sign`, `climate_division_code`, `climate_division_state_code`,

`climate_division_station_code`, `name` e `state`, poiché abbiamo ritenuto che non fornissero dati rilevanti per il nostro obiettivo.

Per quanto riguarda il dataset WEATHER, sono stati eliminati gli attributi `snow_water`, `snow_depth`, `depart`, `sunrise`, `sunset`, `snow_fall` e le relative colonne `flag`, poiché per le stazioni in esame non sono stati misurati questi parametri e inoltre anche `year_month_day` poiché non influisce sulla predizione.

Del dataset BLOCK sono state utilizzate come feature solamente le colonne `block_latitude`, `block_longitude` e `distance`, mentre le altre sono state rimosse poiché duplicate a seguito dell'integrazione.

Dal dataset WNV MOSQUITO sono state escluse solamente le colonne `latitude` e `longitude` poiché sostituite da `block_latitude` e `block_longitude` derivanti dal dataset BLOCK.

Dopo l'integrazione, il database risultante contiene 25482 righe e 66 colonne, in particolare, analizzando dataset per dataset si hanno:

- WNV MOSQUITO: contenente 27196 righe e 10 attributi;
- WEATHER: contenente 1701 righe e 47 attributi;
- STATION: contenente 3 righe e 6 attributi.
- BLOCK: contenente 3 righe e 3 attributi.

Vengono di seguito descritte le features scelte:

Attributo	**Descrizione**	**Valori**
result	risultato del test	booleano
season_year	anno del test	categorico 2007 – 2017
week	settimana dell'anno	numerico 1 – 52 (20 – 40)
trap_type	tipologia di trappola	OVI CVC GRAVID SENTINEL
test_date	data del test	mm/dd/yyyy hh:mm:ss
number_of_mosquitoes	num. zanzare catturate	numerico 1 – 50

Continua nella pagina seguente

Continua dalla pagina precedente

Attributo	**Descrizione**	**Valori**
species	specie di zanzare	categorico
main_trap	id trappola principale	identificatore T001 – T925
sub_trap	trappola secondaria	categorico X = carattere
day_of_week	giorno della settimana	numerico
days_per_week	osservazioni giornaliere	numerico
t_max	temperatura massima settimanale (°C)	numerico
t_max_is_suspicious	occorrenze settimanali di t_max sospette	numerico
t_min	temperatura minima settimanale (°C)	numerico
t_min_is_suspicious	occorrenze settimanali di t_min sospette	numerico
t_avg	temperatura media settimanale (°C)	numerico
t_avg_is_suspicious	occorrenze settimanali di t_avg sospette	numerico
dew_point	media settimanale di dew_point (°C)	numerico
dew_point_is_suspicious	occorrenze settimanali di dew_point sospette	numerico
wet_bulb	media settimanale di wet_bulb (°C)	numerico
wet_bulb_is_suspicious	occorrenze settimanali di wet_bulb sospette	numerico
heat	media settimanale di heat	numerico
heat_is_suspicious	occorrenze settimanali di heat sospette	numerico
cool	media settimanale di cool	numerico
cool_is_suspicious	occorrenze settimanali di cool sospette	numerico
code_sum_is_suspicious	occorrenze settimanali di code_sum sospetti	numerico
code_sum_ra	numero di giorni della settimana in cui occorre	numerico

Continua nella pagina seguente

Continua dalla pagina precedente

Attributo	**Descrizione**	**Valori**
	la condizione RA	
code_sum_br	numero di giorni della settimana in cui occorre la condizione BR	numerico
code_sum_hz	numero di giorni della settimana in cui occorre la condizione HZ	numerico
code_sum_ts	numero di giorni della settimana in cui occorre la condizione TS	numerico
code_sum_smoke	numero di giorni della settimana in cui occorrono le condizioni FU e DU	numerico
code_sum_dz	numero di giorni della settimana in cui occorre la condizione DZ	numerico
code_sum_wind	numero di giorni della settimana in cui occorrono le condizioni FC e SQ	numerico
code_sum_fg	numero di giorni della settimana in cui occorrono le condizioni FG e HZ	numerico
code_sum_sn	numero di giorni della settimana in cui occorre la condizione SN	numerico
code_sum_hail	numero di giorni della settimana in cui occorrono le condizioni GS, GR e PL	numerico
code_sum_up	numero di giorni della settimana in cui occorrono condizioni sconosciute	numerico
precip_total	media di precipitazioni settimanali (mm)	decimale
precip_total_is_suspicious	occorrenze settimanali di precip_total sospette	numerico
stn_pressure	media settimanale di	decimale

Continua nella pagina seguente

Continua dalla pagina precedente

Attributo	**Descrizione**	**Valori**
	stn_pressure (mmHg)	
stn_pressure_is_suspicious	occorrenze settimanali si stn_pressure sospette	numerico
sea_level	media settimanale di sea_level (mmHg)	decimale
sea_level_is_suspicious	occorrenze settimanali di sea_level sospette	numerico
result_speed	media settimanale di result_speed (km/h)	decimale
result_speed_is_suspicious	occorrenze settimanali di result_speed sospette	numerico
result_dir	media settimanale di result_dir	numerico
result_dir_is_suspicious	occorrenze settimanali di result_dir sospette	numerico
avg_speed	media settimanale di avg_speed (km/h)	decimale
avg_speed_is_suspicious	occorrenze settimanali di avg_speed sospette	numerico
max5_speed	media settimanale di max5_speed (km/h)	numerico
max5_speed_is_suspicious	occorrenze settimanali di max5_speed sospette	numerico
max5_dir	media settimanale di max5_dir	numerico
max5_dir_is_suspicious	occorrenze settimanali di max5_dir sospette	numerico
max2_speed	media settimanale di max2_speed (km/h)	numerico
max2_speed_is_suspicious	occorrenze settimanali di max2_speed sospette	numerico
max2_dir	media settimanale di max2_dir	numerico
max2_dir_is_suspicious	occorrenze settimanali di max2_dir sospette	numerico
station_location	posizione della stazione	categorico
station_latitude	latitudine della stazione	decimale

Continua nella pagina seguente

Continua dalla pagina precedente

Attributo	**Descrizione**	**Valori**
		40.00 – 42.00
station_longitude	longitudine della stazione	decimale -87.00 – -88.00
station_ground_height	altitudine del terreno sul livello del mare (metri)	decimale
station_station_height	altitudine della stazione sul livello del mare (metri)	decimale
station_barometer	altitudine del barometro sul livello del mare (metri)	decimale
block_latitude	latitudine del blocco	decimale 40.00 – 42.00
block_longitude	longitudine del blocco	decimale -87.00 – -88.00
block_station_distance	distanza in tra la stazione e il blocco (metri)	decimale

Tabella 3.2: Features

Capitolo 4

Analisi di almeno 2 dimensioni di qualità e relative metriche delle features successivamente utilizzate

4.1 Completezza

In seguito alla selezione delle 66 features, la completezza complessiva del dataset risulta essere del 100%, così come quella di attributo e di tupla.

4.2 Leggibilità

Dopo le modifiche apportate al dataset, in particolare quelle per convertire le unità di misura dal sistema anglosassone a quello internazionale e ... la leggibilità risulta notevolmente migliorata.

- **Dataset finale**: il dataset finale contiene per lo più attributi di tipo numerico che esprimono valori di grandezze fisiche (distanze, temperature, velocità), che sono autoesplicativi. Potrebbe essere utile invece rendere immediata all'utente la visualizzazione del risultato delle analisi, del giorno in cui sono state effettuate e soprattutto le condizioni meteo poiché ora vengono utilizzati dei codici numerici. A causa di ciò è stato pensato di utilizzare una vista SQL per evidenziare questi dati. Una query sugli attributi discussi in precedenza produce i risultati visibile in Tabella 4.1, mentre applicando la query SQL mostrata nel Listing 4.1 vengono restituiti i record mostrati nella Tabella 4.2.

test_id	result	day_of_week	cod_sum_ra	cod_sum_ts
23505	0	2	3	0
23109	0	2	0	0
45618	1	3	1	2
40026	1	4	1	0

Tabella 4.1: Risultato query dataset senza view

```
SELECT test_id,
result,
day_of_week,
weather_conditions
FROM "ReadableFusedDataset";
```

Listing 4.1: Query SQL per la leggibilità degli attributi

In particolare è possibile notare come gli attributi `code_sum_ra` e `code_sum_ts` sono stati collassati in un unico attributo `weather_conditions` visibile nella Tabella 4.2 che racchiude la rappresentazione testuale di tutti i fenomeni meteorologici registrati durante il test.

test_id	result	day_of_week	weather_conditions
23505	Negativo	Mercoledì	Pioggia
23109	Negativo	Mercoledì	
45618	Positivo	Giovedì	Pioggia,Temporale
40026	Positivo	Venerdì	Pioggia

Tabella 4.2: Risultato query dataset attraverso la view

Capitolo 5

Analisi descrittive dei dati integrati

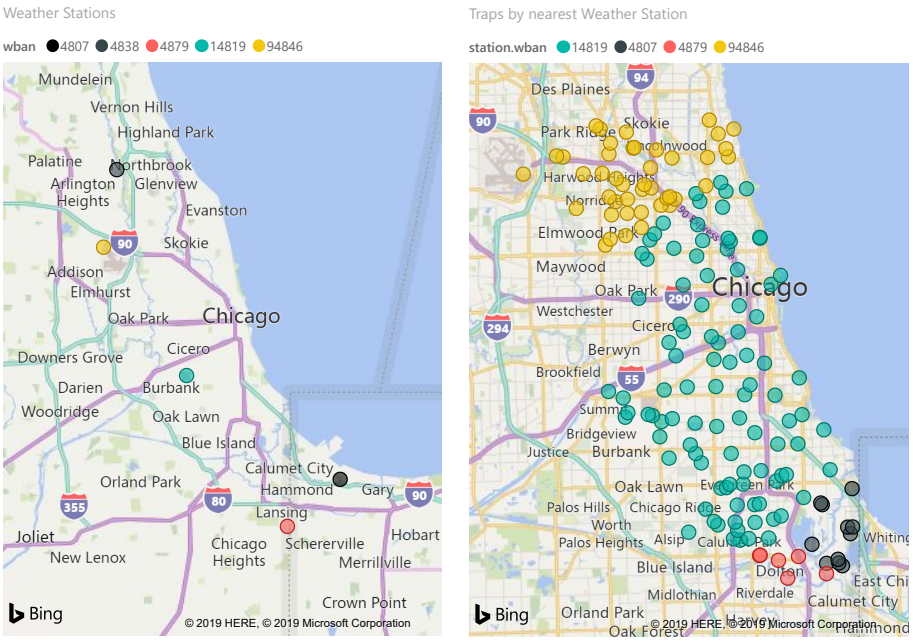


Figura 5.1: FIXME Visualizzazione dei 64 filtri convoluzionali appresi dal layer **conv1** della rete ResNet-34. Il blu scuro rappresenta i valori negativi, il verde quelli prossimi allo zero e il giallo quelli positivi

Parte III

Machine Learning

Capitolo 1

Creazione dei training set

Capitolo 2

Analisi esplorativa del training set

L'analisi esplorativa è stata effettuata sull'intero dataset ripulito. L'esplorazione dei dati è stata eseguita utilizzando il package R **DataExplorer** e tramite il software **Power BI**.
Il primo passo è stato visualizzare la distribuzione di colonne con valori continui, discreti e mancanti, utilizzando la funzione `plot_intro`.

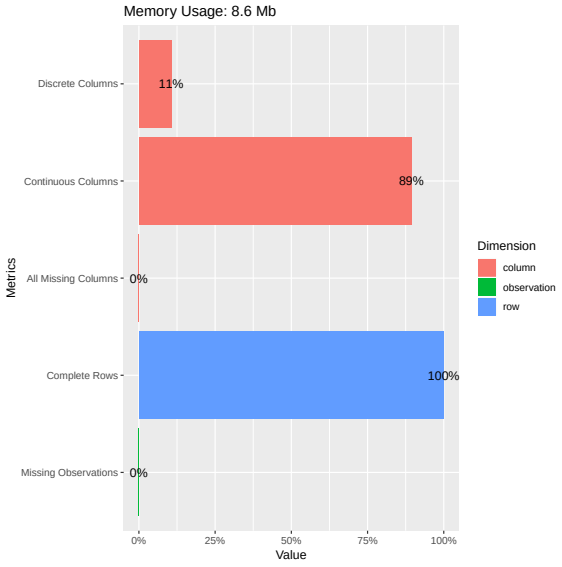


Figura 2.1

La maggior parte degli attributi del dataset, precisamente 59, sono di tipo continuo, mentre solamente 7 sono di tipo discreto. Di tutto il dataset, risulta che, in seguito alla selezione degli attributi alla fine del procedimento di integrazione, tutte le righe sono complete al 100%.

Per andare nel dettaglio di queste variabili, sono stati generati dei diagrammi a barre, con la funzione `plot_bar`, per quanto riguarda le variabili discrete, e gli istogrammi, tramite la funzione `plot_histogram`, delle variabili continue.

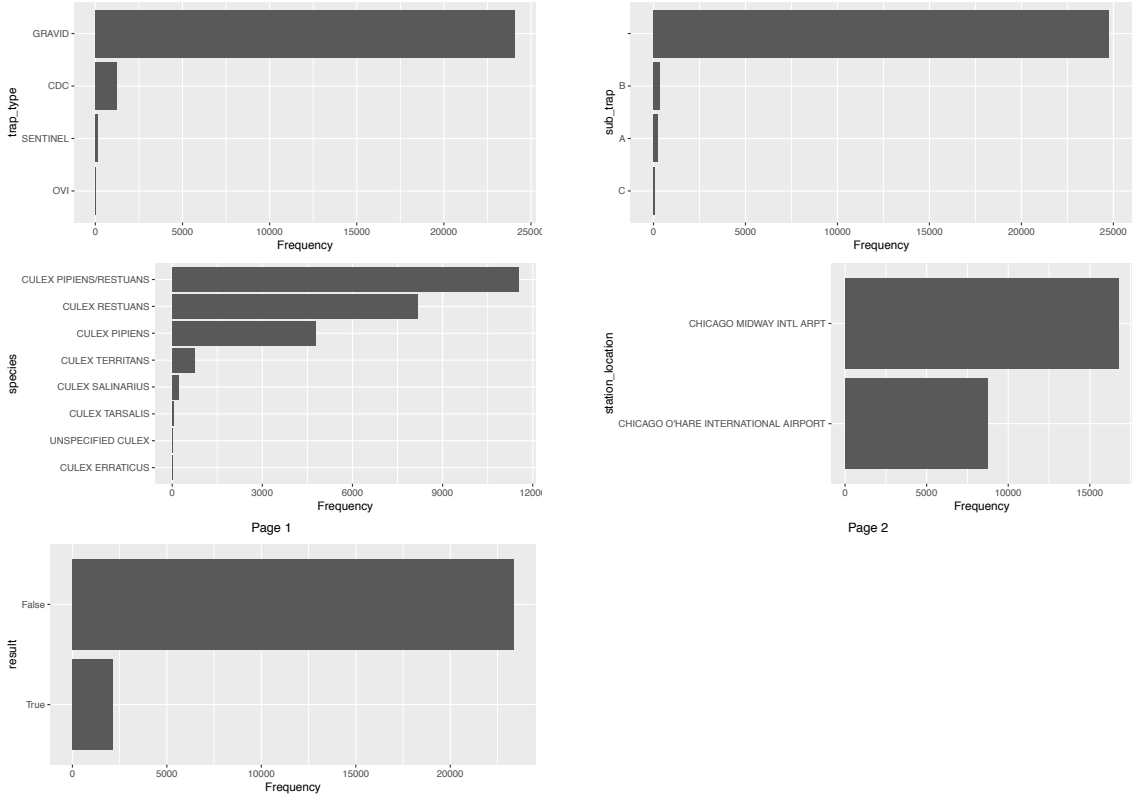
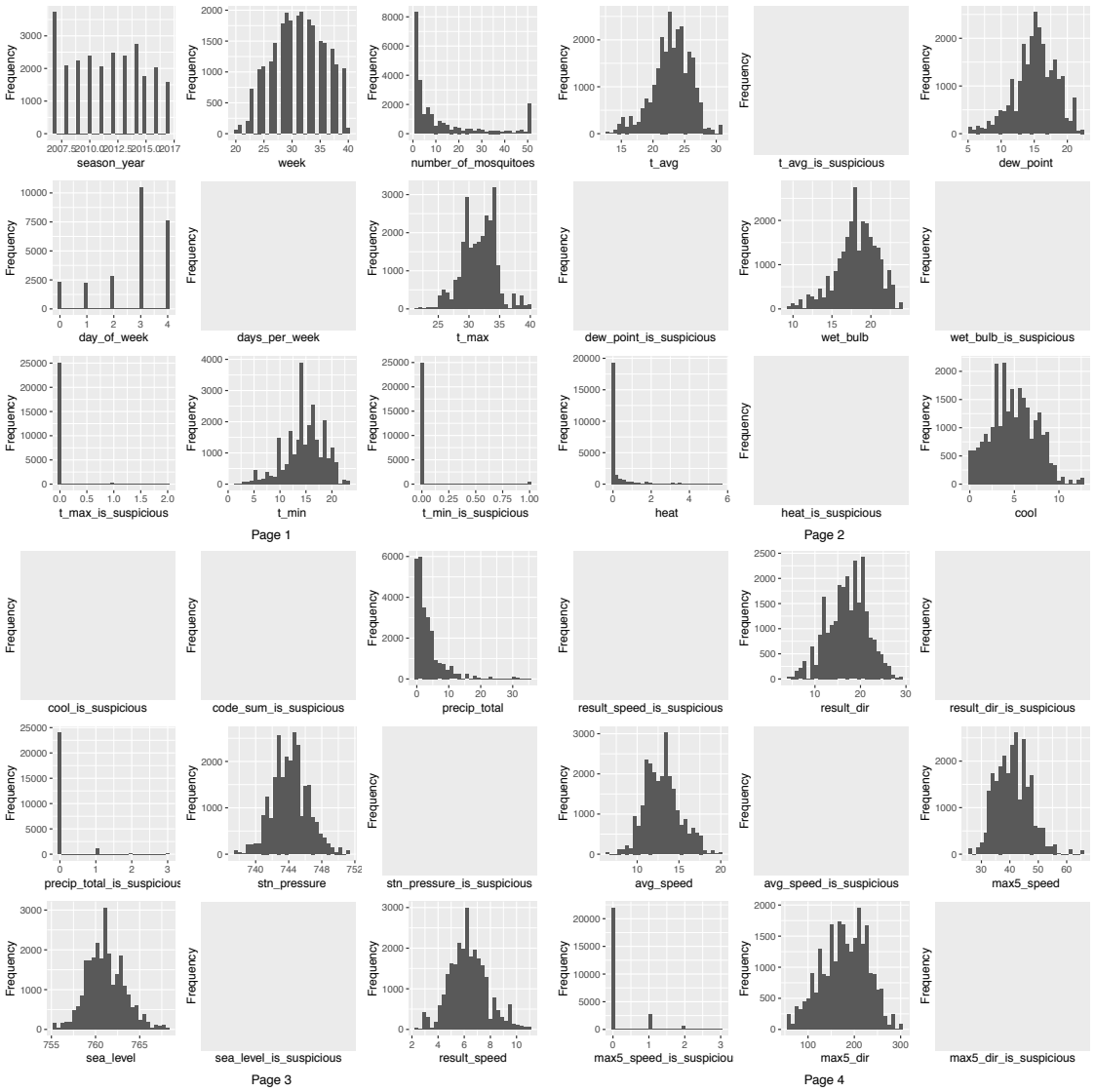


Figura 2.2: GRAFICI bar

Da questa analisi si evince che ...
In seguito, per evidenziale eventuali correlazioni tra le variabili del dataset, è stata visualizzata una matrice di correlazione, grazie la funzione `plot_correlation`.
In figura 2.4 viene mostrato il risultato.
Nell'analisi di questa matrice poniamo particolare attenzione alla variabile `result`, ovvero il target del nostro modello.


Frequency
season_year
week
number_of_mosquitoes
t_avg
t_avg_is_suspicious
dew_point
day_of_week
days_per_week
t_max
dew_point_is_suspicious
wet_bulb
wet_bulb_is_suspicious
t_max_is_suspicious
t_min
t_min_is_suspicious
heat
heat_is_suspicious
cool
Page 1
Page 2
cool_is_suspicious
code_sum_is_suspicious
precip_total
result_speed_is_suspicious
result_dir
result_dir_is_suspicious
precip_total_is_suspicious
stn_pressure
stn_pressure_is_suspicious
avg_speed
avg_speed_is_suspicious
max5_speed
sea_level
sea_level_is_suspicious
result_speed
max5_speed_is_suspiciou
max5_dir
max5_dir_is_suspicious
Page 3
Page 4

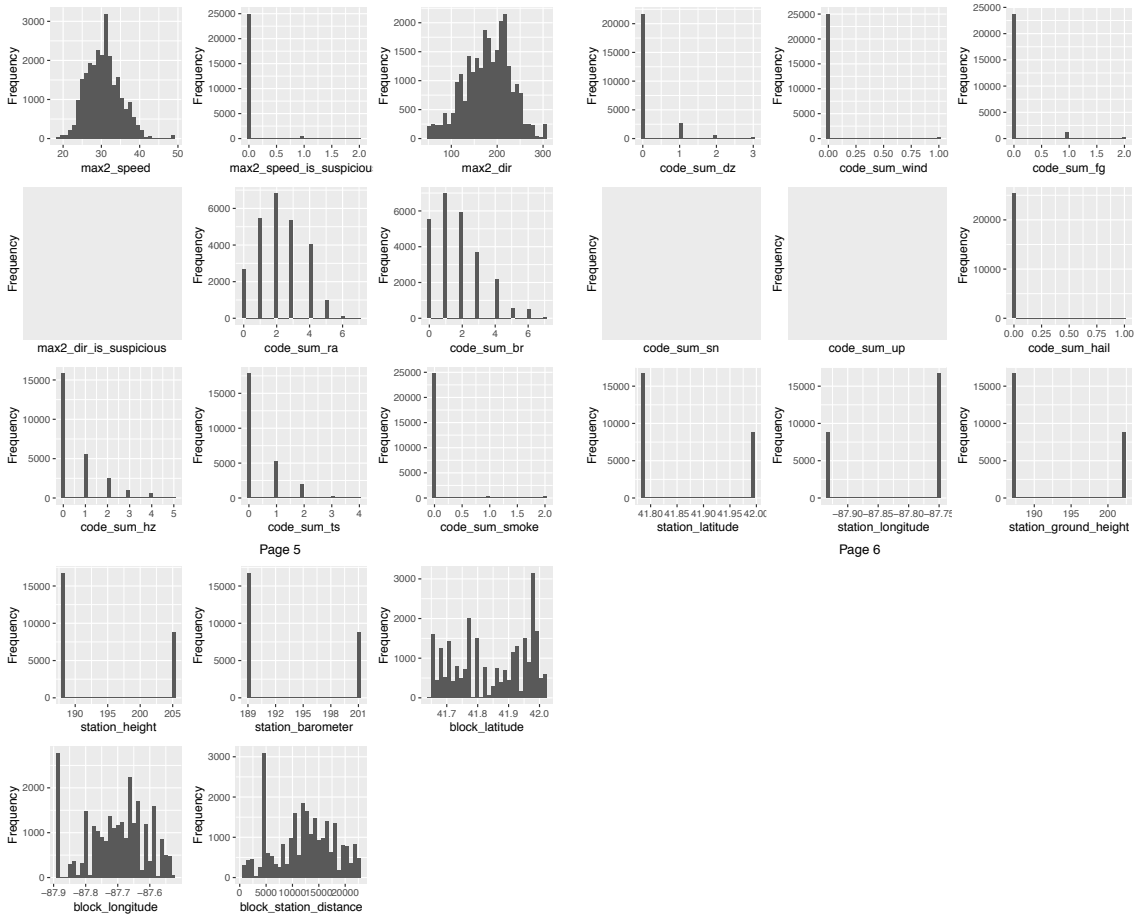


Figura 2.3: GRAFICI istogramma

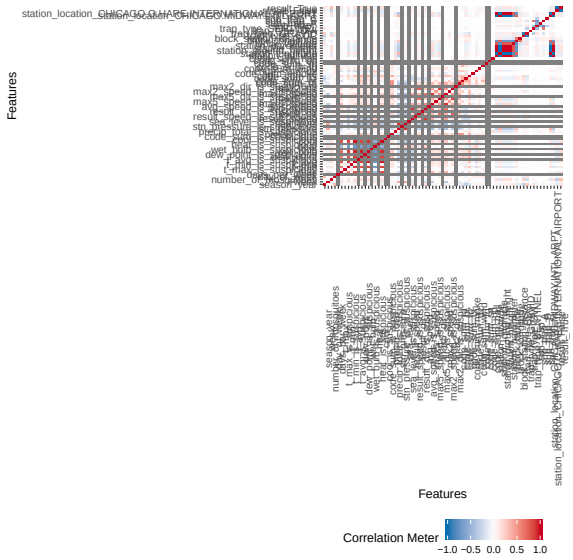


Figura 2.4

Capitolo 3

Descrizione e motivazione dei modelli di machine learning utilizzati

3.1 Random Forest

Il primo modello implementato è Random Forest.
Questo è stato scelto poiché in grado di maneggiare diversi tipi di variabili, perché è robusto rispetto agli outlier, da una stima dell'errore e dell'importanza delle variabili. I modelli di tipo Random Forest convergono sempre e non presentano problemi di overfitting. Per questo motivo Random Forest non necessita della messa in atto di tecniche di validazione incrociata o della verifica su un set separato di variabili per avere una valutazione imparziale dell'errore, in quanto ciò è garantito intrinsecamente dal metodo.
Il modello è di semplice utilizzo, in quanto prevede l'inserimento di soltanto due parametri (il numero di variabili nel sottoinsieme di variabili casuali usate in ogni nodo ed il numero di alberi nella foresta) e non è molto sensibile ai loro valori. Per questo motivo è stata utilizzata la configurazione di default.

3.2 Support Vector Machine

È stato scelto di provare ad utilizzare il modello di classificazione Support Vector Machine (SVM) poichè la classificazione di un dataset contenente dati reali e comprensivo di circa 70 features potrebbe risultare difficile, a causa della bassa possibilità che i dati siano linearmente separabili. Con SVM è possibile utilizzare il metodo Kernel tramite delle funzioni che permettono di aumentare la dimensionalità dei dati senza dover calcolare nuovamente la loro posizione all'interno del feature space; in questo modo è possibile trovare un iperpiano di separazione in grado di classificare i dati.

La funzione kernel (K) utilizzata è di tipo Radial Basis Function (RBF) ed è definita nell'Equazione 3.1.

$$K(x, x') = e^{-\gamma\|x-x'\|^2} \tag{3.1}$$

Il classificatore può essere addestrato utilizzando due parametri per controllare il grado di linearità dell'iperpiano (γ) e il margine (C) tra i vettori di supporto:

- γ: è il parametro della funzione kernel 3.1, minore è il valore e maggiormente l'iperpiano assumerà una forma lineare. Aumentando troppo questo valore si rischia l'overfitting del modello.
- C: è responsabile di controllare l'ampiezza del margine tra i vettori di supporto, maggiore è il suo valore minore sarà l'ampiezza del margine portando a una diminuzione della proprietà di generalizzazione.

Attraverso il processo di tuning del modello è stato scelto di utilizzare il valore 0.01 per il parametro γ e 10 per C.

3.3 Neural Network

Capitolo 4

Esperimenti

Per ogni modello, la creazione di training e validation set è stata realizzata utilizzando due tecniche diverse: è stato effettuato un esperimento in cui viene utilizzato un *holdout* 80–20 ed uno che ricorre ad una *10-fold cross validation* con dataset tutti della stessa dimensione.

4.1 Random Forest

Per l'addestrare è stato utilizzato il package R `RandomForest`.

4.1.1 Holdout

La stima degli errori di classificazione OOB è del 7,06%.

Le misure di accuracy, precision, recall e f-measure sono le seguenti:

	Accuracy	**Precision**	**Recall**	**F-Measure**
Training	94,41%	87,63%	38,65%	53,64%
Validation	93,11%	70,41%	28,33%	40,41%

Tabella 4.1: Analisi performance dell'holdout

Di seguito vengono mostrate le curve ROC per il traning e il validation set:

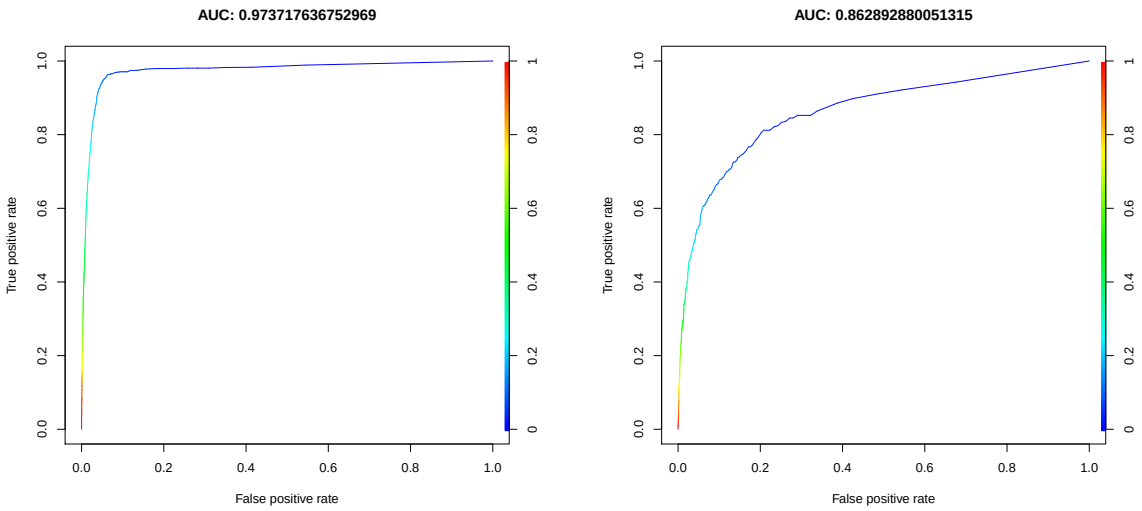


Figura 4.1: ROC sul training e validation set

4.1.2 10-fold cross validation

La stima degli errori di classificazione OOB è del 7,06%, ed stata calcolata come la media della misura su ogni fold.
Vengono mostrate nella tabella 4.2 la stima degli errori OOB per ogni fold.

Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
7,22%	7,32%	7,25%	7,36%	7,01%	7,29%	6,78%	6,84%	6,82%	6,75%

Tabella 4.2: Stima degli errori di classificazione OOB

Per quanto riguarda le misure finali di accuracy, precision, recall e f-measure, come per l'errore di classificazione OOB sono state calcolare come la media delle misure su ogni fold. Vengono mostrate nella tabella 4.3 i dati relativi ad ogni fold e in tabella 4.4 le misure complessive.

		Accuracy	**Precision**	**Recall**	**F-Measure**
Training	Fold1	94,23%	84,42%	40,57%	54,80%
	Fold2	94,43%	84,30%	43,81%	57,65%
	Fold3	94,17%	85,68%	37,93%	52,59%
	Fold4	94,09%	85,98%	39,02%	53,68%
	Fold5	94,38%	85,22%	39,86%	54,32%
	Fold6	93,99%	85,52%	37,46%	52,10%
	Fold7	94,53%	85,00%	36,83%	51,40%
	Fold8	94,72%	87,09%	40,01%	54,83%
	Fold9	94,37%	84,93%	35,51%	50,08%
	Fold10	94,50%	86,60%	35,77%	50,63%
Validation	Fold1	94,51%	78,57%	7,43%	13,58%
	Fold2	94,78%	75,00%	6,47%	11,92%
	Fold3	94,62%	92,11%	20,71%	33,82%
	Fold4	95,68%	57,14%	10,62%	17,91%
	Fold5	92,35%	63,64%	6,97%	12,55%
	Fold6	95,45%	76,92%	8,13%	14,71%
	Fold7	88,85%	66,14%	25,85%	37,17%
	Fold8	90,93%	73,55%	30,90%	43,52%
	Fold9	89,29%	58,71%	30,33%	40,00%
	Fold10	89,76%	66,67%	36,36%	47,06%

Tabella 4.3: Analisi di performance per ogni fold

	Accuracy	**Precision**	**Recall**	**F-Measure**
Training	94,34%	85,47%	38,68%	53,21%
Validation	92,62%	70,85%	18,38%	161,67%

Tabella 4.4: Analisi performance cross validation

Vengono infine mostrate le curve ROC per ognuno dei fold, sia sul traning che sul validation set:
Possiamo vedere che in nessun fold il modello ha un valore di Area Under Curve (AUC) inferiore al valore 0,80, che viene considerato come un indice di buone prestazioni. Mediamente si ha un valore di 0,85.

4.2 Support Vector Machine

Le misure finali di accuracy, precision, recall e f-measure sono state calcolare come la media delle misure su ogni fold. Vengono mostrate nella tabella 4.5 i dati relativi ad ogni fold e in tabella 4.6 le misure complessive.

		Accuracy	**Precision**	**Recall**	**F-Measure**
Training	Fold1	93,28%	76,90%	31,66%	44,85%
	Fold2	93,24%	78,12%	30,56%	43,93%
	Fold3	93,28%	77,73%	29,80%	43,08%
	Fold4	93,11%	78,03%	30,01%	43,34%
	Fold5	93,46%	77,92%	30,80%	44,14%
	Fold6	93,17%	77,45%	30,71%	43,98%
	Fold7	93,63%	76,31%	27,38%	40,30%
	Fold8	93,66%	78,97%	28,41%	41,78%
	Fold9	93,66%	78,61%	28,00%	41,29%
	Fold10	93,63%	77,58%	27,02%	40,08%
Validation	Fold1	94,15%	78,57%	6,08%	11,28%
	Fold2	94,70%	60,00%	8,63%	15,08%
	Fold3	94,30%	80,00%	18,93%	30,61%
	Fold4	95,83%	62,06%	15,92%	25,33%
	Fold5	92,58%	64,28%	13,43%	22,21%
	Fold6	95,44%	66,66%	11,38%	19,44%
	Fold7	89,16%	66,66%	30,15%	41,52%
	Fold8	90,58%	75,43%	29,86%	42,78%
	Fold9	89,44%	59,62%	32,00%	41,64%
	Fold10	90,07%	72,60%	33,22%	45,58%

Tabella 4.5: Analisi di performance per ogni fold

	Accuracy	**Precision**	**Recall**	**F-Measure**
Training	94,412%	77,34%	29,43%	42,67%
Validation	92,62%	68,50%	19,96%	29,54%

Tabella 4.6: Analisi performance cross validation

Vengono infine mostrate le curve ROC per ognuno dei fold, sia sul traning che sul validation set:

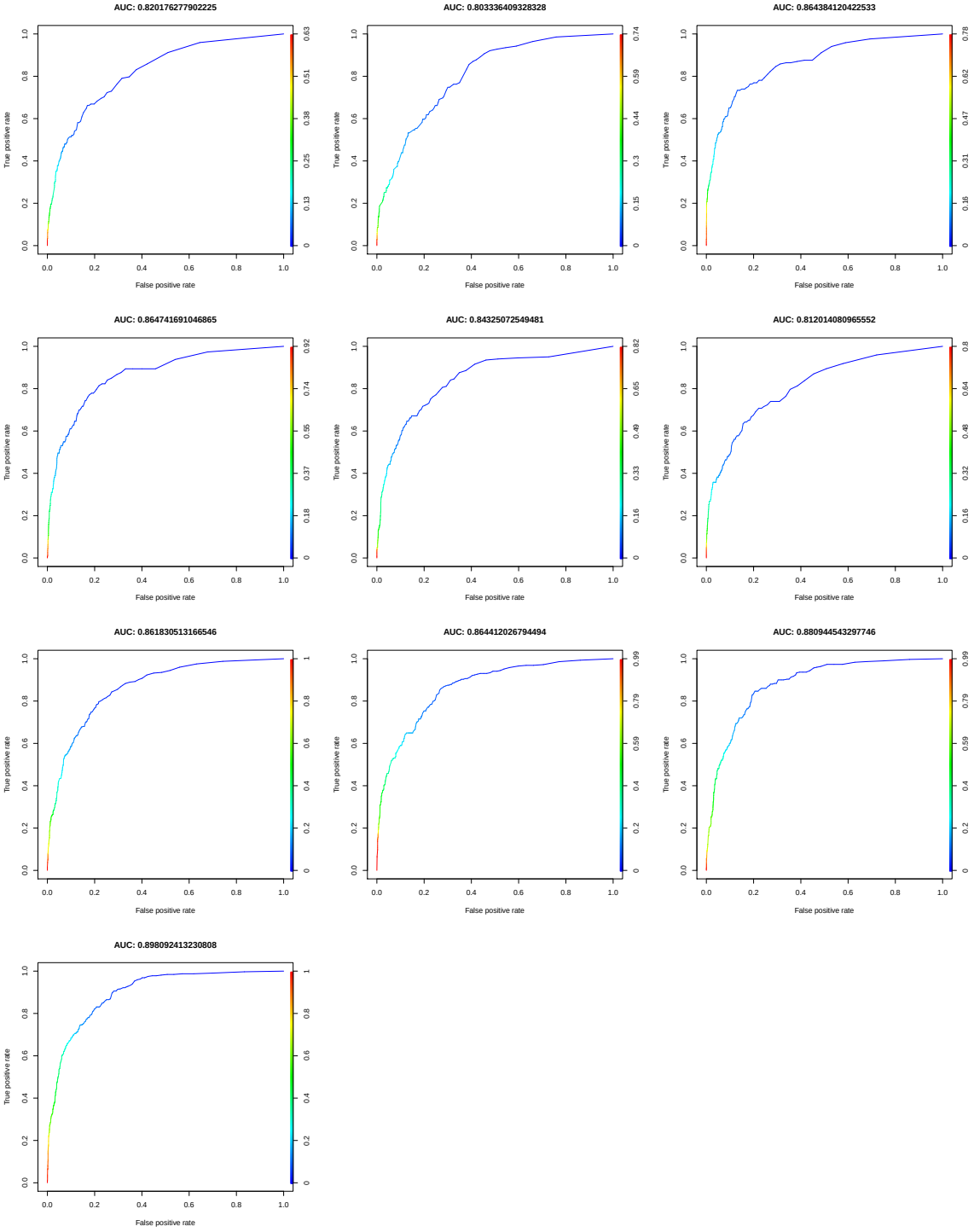


Figura 4.2: ROC sul validation set

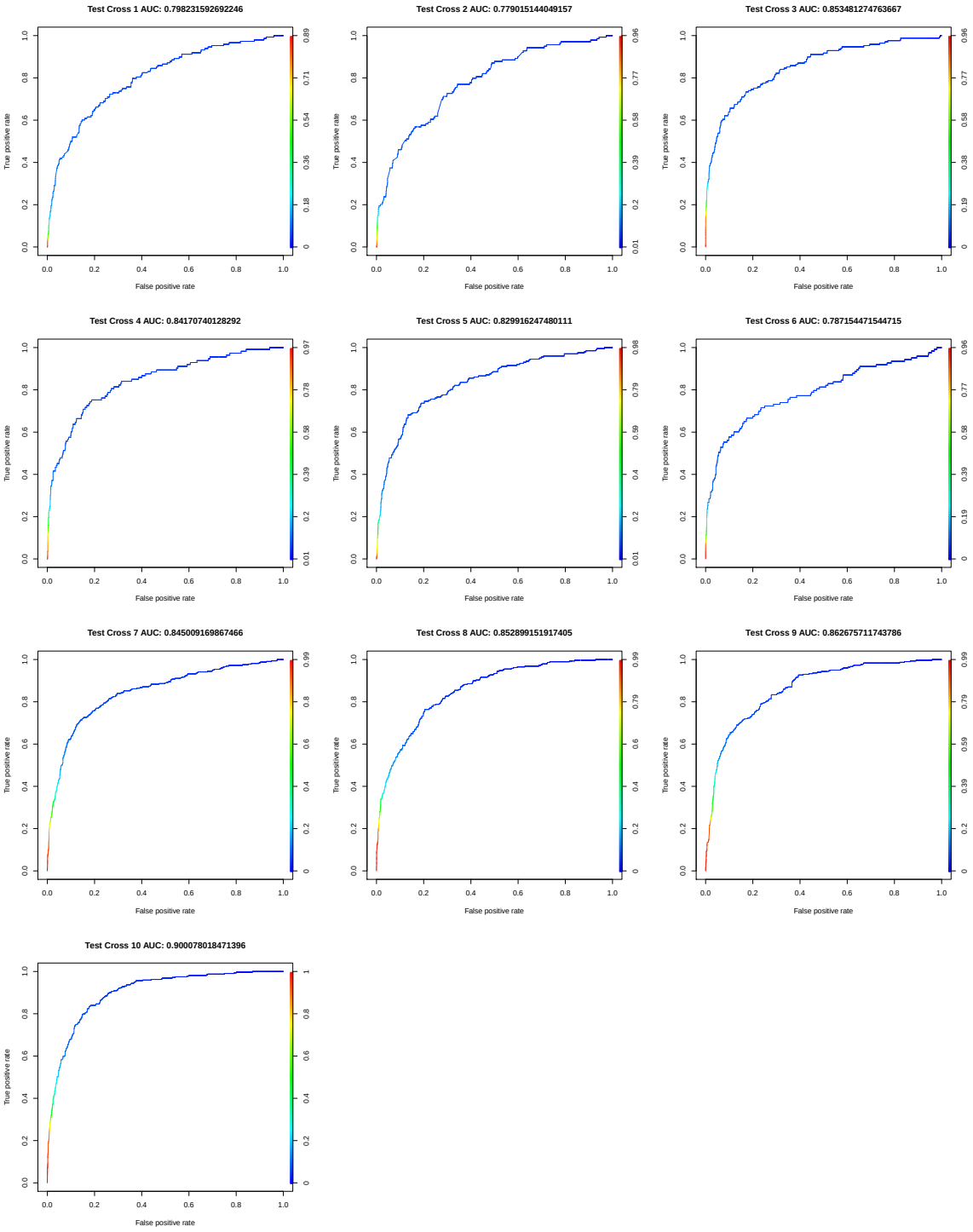


Figura 4.3: ROC sul validation set

Capitolo 5

Analisi dei risultati ottenuti

Capitolo 6

Conclusioni

6.1 Sviluppi futuri

Per quanto riguarda Random Forest, poiché l'algoritmo offre la possibilità di misurare l'importanza della variabile predittore, uno sviluppo futuro potrebbe essere utilizzare l'importanza delle variabili, mostrata nella figura 6.1, per classificare l'utilità delle variabili, ed utilizzare solo le più importanti come features del modello (ristimare il modello usando solo le variabili d'interesse).

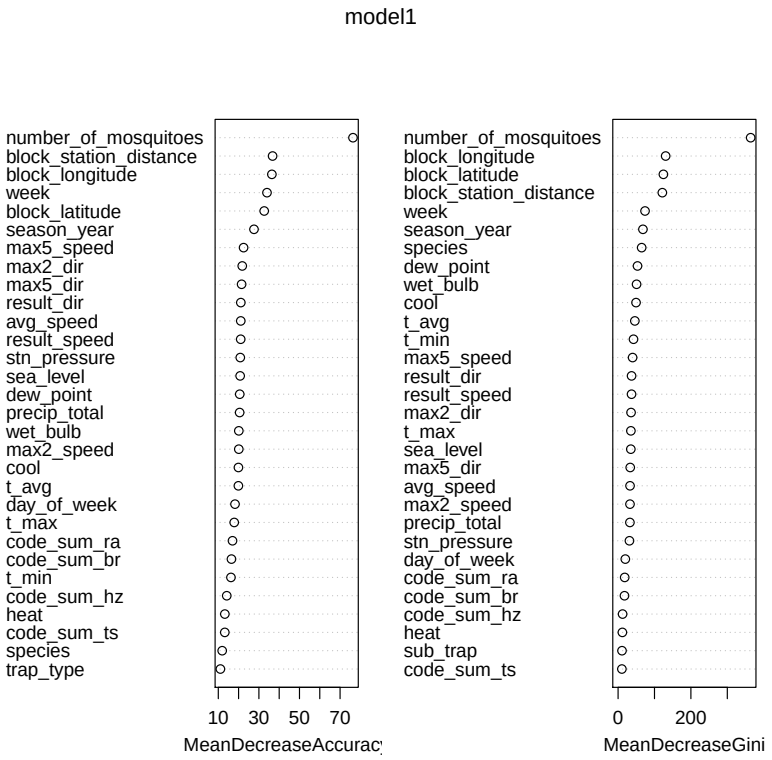


Figura 6.1: Random Forest Importance

L'importanza variabile "globale" è la diminuzione media della precisione su tutte le previsioni out-of-bag con convalida incrociata. L'importanza variabile locale è la diminuzione media della precisione di ogni singola previsione out-of-bag con convalida incrociata.
L'importanza di GINI misura il guadagno medio di purezza mediante la divisione di una determinata variabile. Se la variabile è utile, tende a dividere i nodi con etichetta mista in nodi di una singola classe pura. La divisione con variabili permutate non tende né a incrementare né a diminuire la purezza dei nodi. Permutando una variabile utile, si tende a dare una diminuzione relativamente grande del guadagno gini medio. L'importanza di GINI è strettamente correlata alla funzione di decisione locale, utilizzata dalla foresta casuale per selezionare la migliore suddivisione disponibile. Pertanto, non ci vuole molto tempo extra per calcolare. D'altra parte, il guadagno gini medio nelle divisioni locali, non è necessariamente ciò che è più utile misurare, al contrario del cambiamento delle prestazioni generali del modello. L'importanza di Gini è di importanza generale inferiore a (basata sulla permutazione) in quanto è relativamente più distorta, più instabile e tende a rispondere a una domanda più indiretta.