

# PROGETTO MACHINE LEARNING

---

BELTRAMELLI FABIO	816912
CAPELLI ALESSANDRO	816302
FINATI DAVIDE	817508

# DOMINIO, OBIETTIVI

---

- ▶ Dominio:

Il dataset preso in esame rappresenta le osservazioni atmosferiche di diverse stazioni meteo in Australia dal 01/11/2007 al 25/06/2017.

- ▶ Obiettivo:

Allenare e valutare diversi modelli per la predizione della possibilità che piova il giorno successivo. Trovare modello con tradeoff migliore tra performance e tempo.

# IPOTESI E ASSUNZIONI

---

- ▶ Ipotesi:

La probabilità che piova il giorno successivo è minore in quanto l'Australia è un territorio caratterizzato da temperature alte e bel tempo. Questo è dimostrato anche dallo sbilanciamento del dataset.

- ▶ Assunzioni:

Abbiamo rimosso dal dataset il valore RISK-MM in quanto variabile target per un task di regressione. Abbiamo anche rimosso le feature con elevata percentuale di valori nulli, quali Evaporation, Sunshine, Cloud9am, Cloud3pm.

# DATASET

---

- ▶ MaxTemp: temperatura massima registrata
- ▶ MinTemp: temperatura minima registrata
- ▶ RainFall: quantità di pioggia caduta
- ▶ WindGustDir: direzione del vento più forte
- ▶ WindGustDir9am: direzione del vento più forte alle nove di mattina
- ▶ WindGustDir3pm: direzione del vento più forte alle tre di pomeriggio
- ▶ WindGustSpeed: velocità del vento più forte
- ▶ WindGustSpeed9am: velocità del vento più forte alle nove di mattina
- ▶ WindGustSpeed3pm: velocità del vento più forte alle tre di pomeriggio
- ▶ Humidity9am: livello di umidità alle nove di mattina
- ▶ Humidity3pm: livello di umidità alle tre di pomeriggio
- ▶ Pressure9am: livello di pressione alle nove di mattina
- ▶ Pressure3pm: livello di pressione alle tre di pomeriggio
- ▶ Temp9am: temperatura registrata alle nove di mattina
- ▶ Temp3pm: temperatura registrata alle tre di pomeriggio
- ▶ RainToday: indica se il giorno precedente alla predizione ha piovuto
- ▶ RainTomorrow: indica il target binario da predire

# CORRELAZIONE FEATURE

# FEATURE SELECTION

---

- ▶ Per quanto riguarda la feature selection abbiamo utilizzato il metodo Correlation-based, in cui un alto valore di correlazione tra due feature indica che avranno lo stesso effetto sulla predizione del target. Per ciò è possibile andare a escluderne una delle due, in modo da ridurre il numero di attributi.

Data la soglia di 0.67, le feature rimosse dal dataset sono:

- ▶ Temp9am
- ▶ MaxTemp
- ▶ Temp3pm
- ▶ Pressure3pm
- ▶ Humidity9am
- ▶ WindGustSpeed

# TRAINING SET & TEST SET

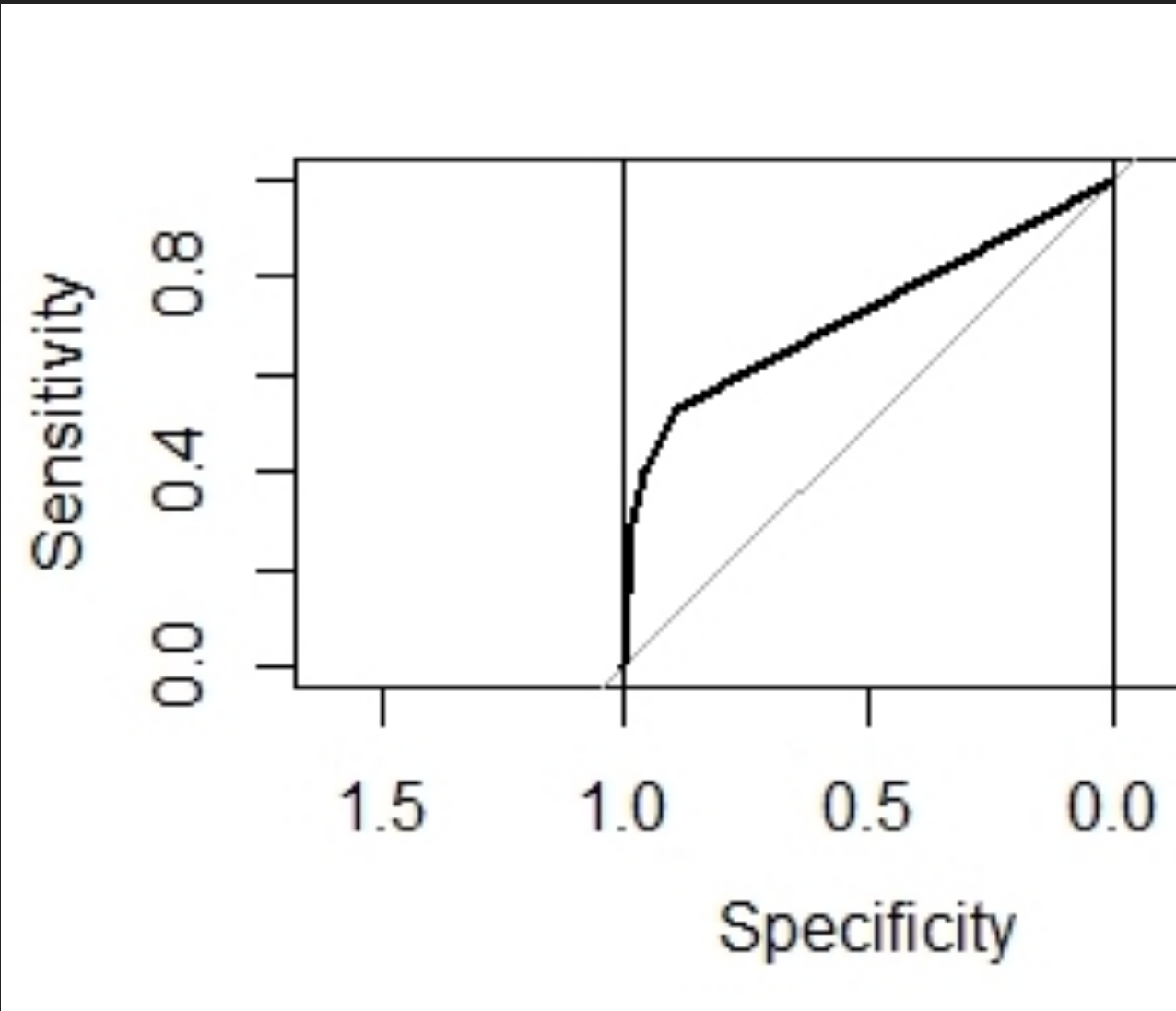
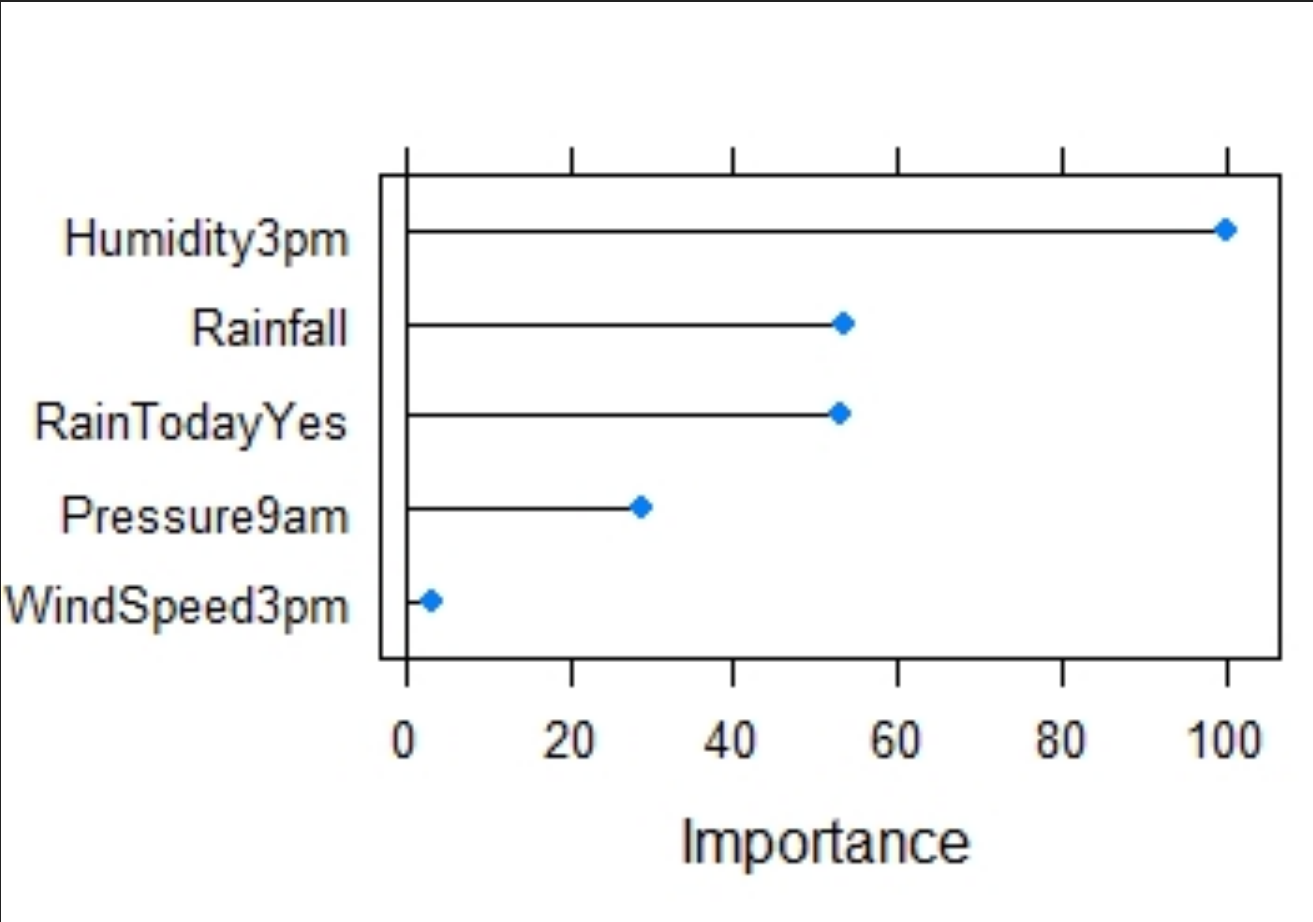
---

- ▶ Abbiamo inizialmente utilizzato la tecnica HoldOut per individuare i modelli più promettenti sui quali eseguire la CrossValidation, in modo da risparmiare tempo successivamente.
- ▶ La partizione HoldOut usata è formata da 70% training e da 30% test, in modo randomico.
- ▶ I modelli scelti dopo la fase di HoldOut sono: DecisionTree, RandomForest, NaiveBayes, NeuralNetwork. SVM è stata esclusa, in quanto è risultata troppo onerosa l'ottimizzazione dei parametri.
- ▶ La CrossValidation applicata è del tipo k-Fold uguale a 10.

# DECISION TREE

	YES	NO
YES	3009	4564
NO	997	25260

Accuracy	0,83
Precision	0,75
Recall	0,39
F1-score	0,51
AUC	0,72



Tempo creazione modello:  
19 sec

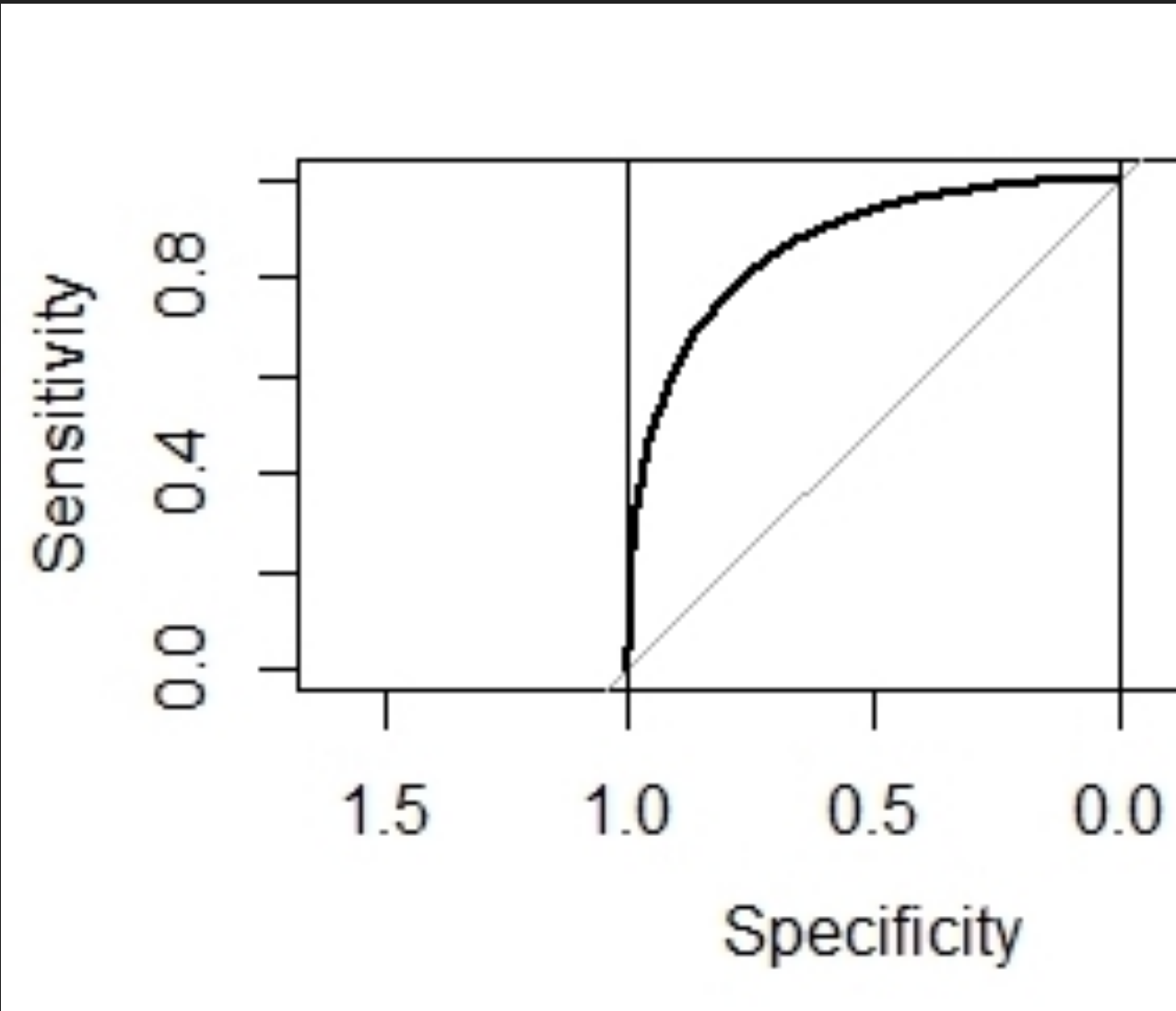
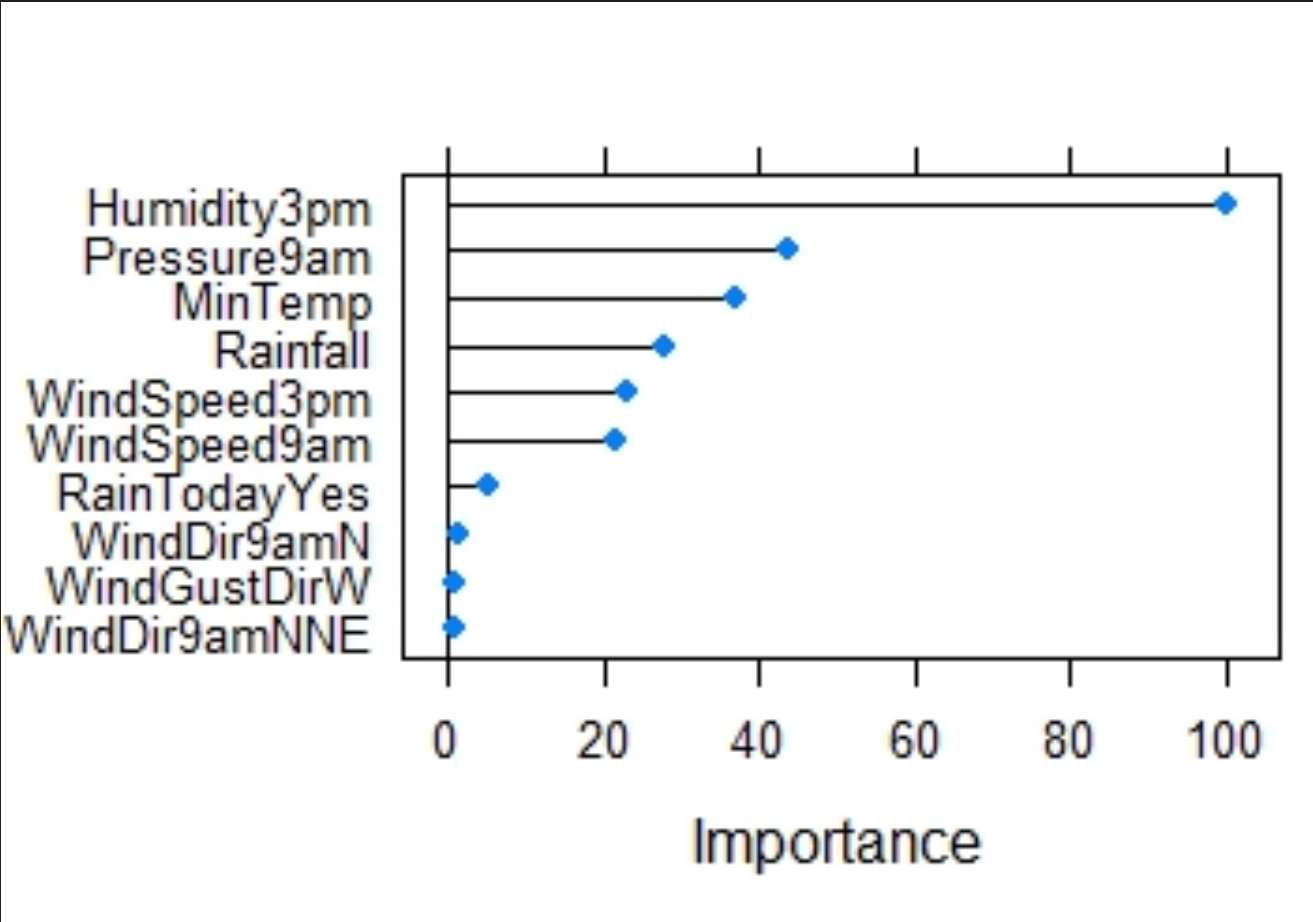
Tempo calcolo predizione:  
1 sec



# RANDOM FOREST

	YES	NO
YES	3714	3859
NO	1338	24919

Accuracy	0,84
Precision	0,73
Recall	0,48
F1-score	0,58
AUC	0,86



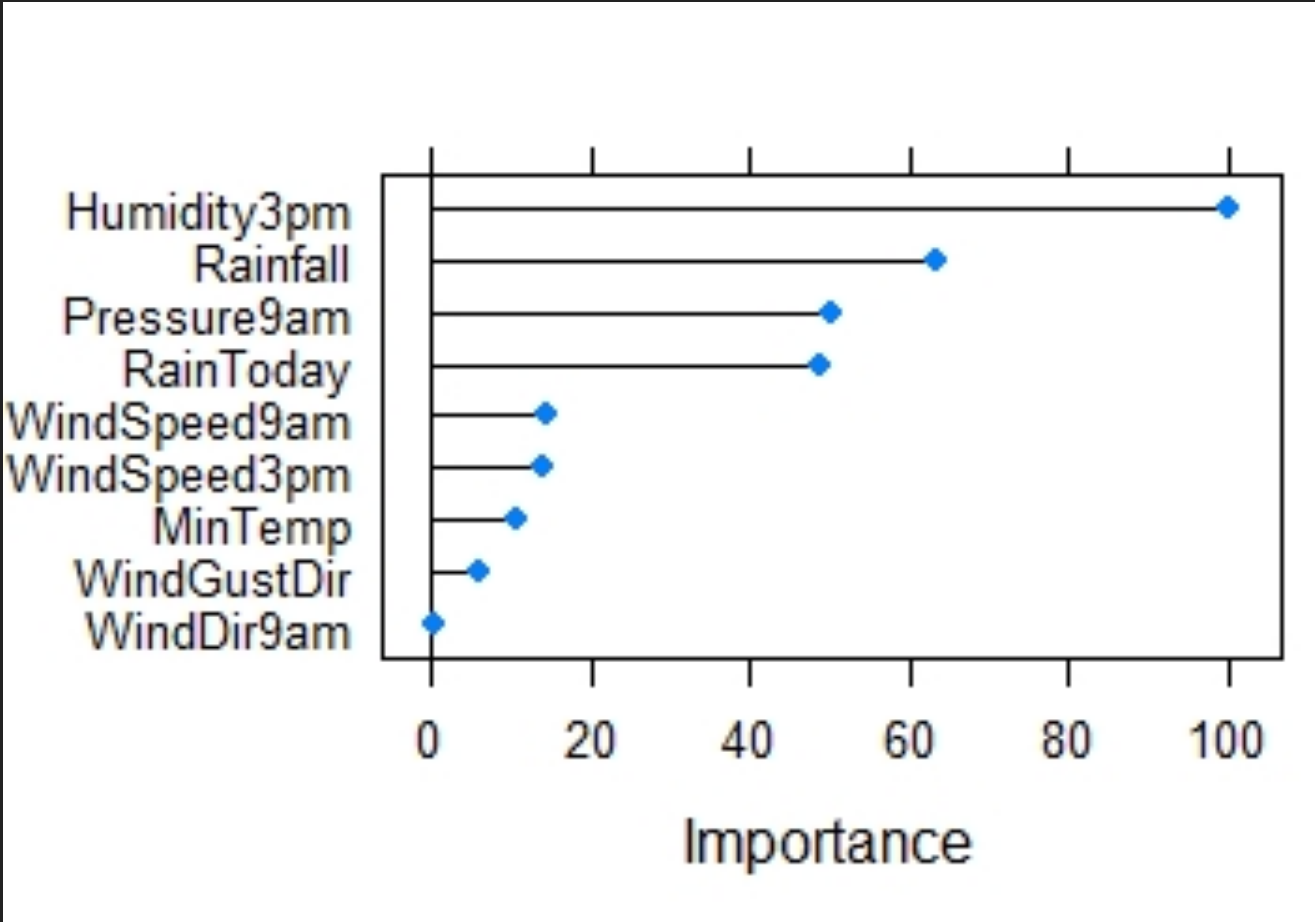
Tempo creazione modello:  
4 ore

Tempo calcolo predizione:  
4 sec

# NAIVE BAYES

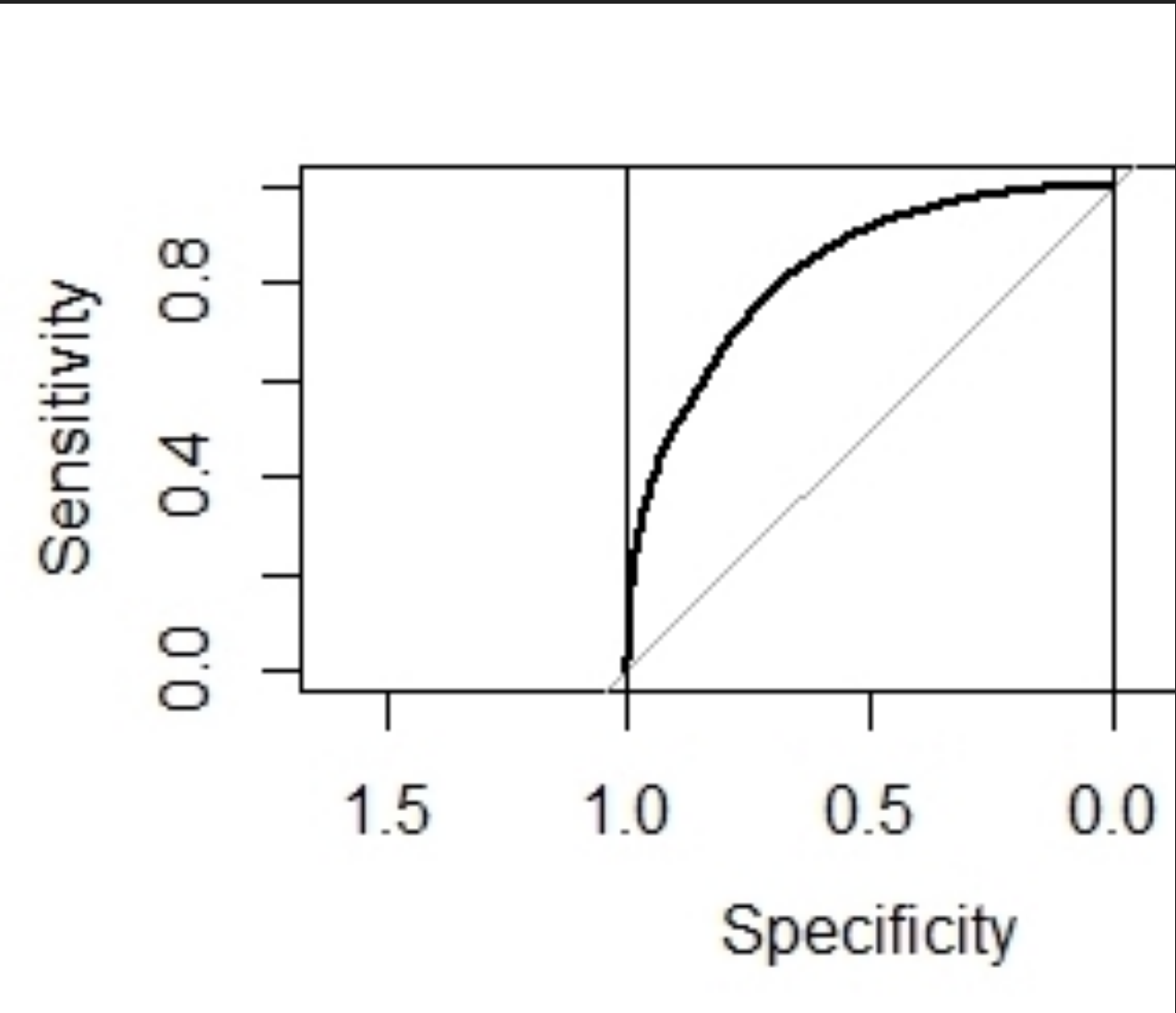
	YES	NO
YES	18	7555
NO	12	26245

Accuracy	0,77
Precision	0,60
Recall	0,002
F1-score	0,004
AUC	0,82



Tempo creazione modello:  
22 sec

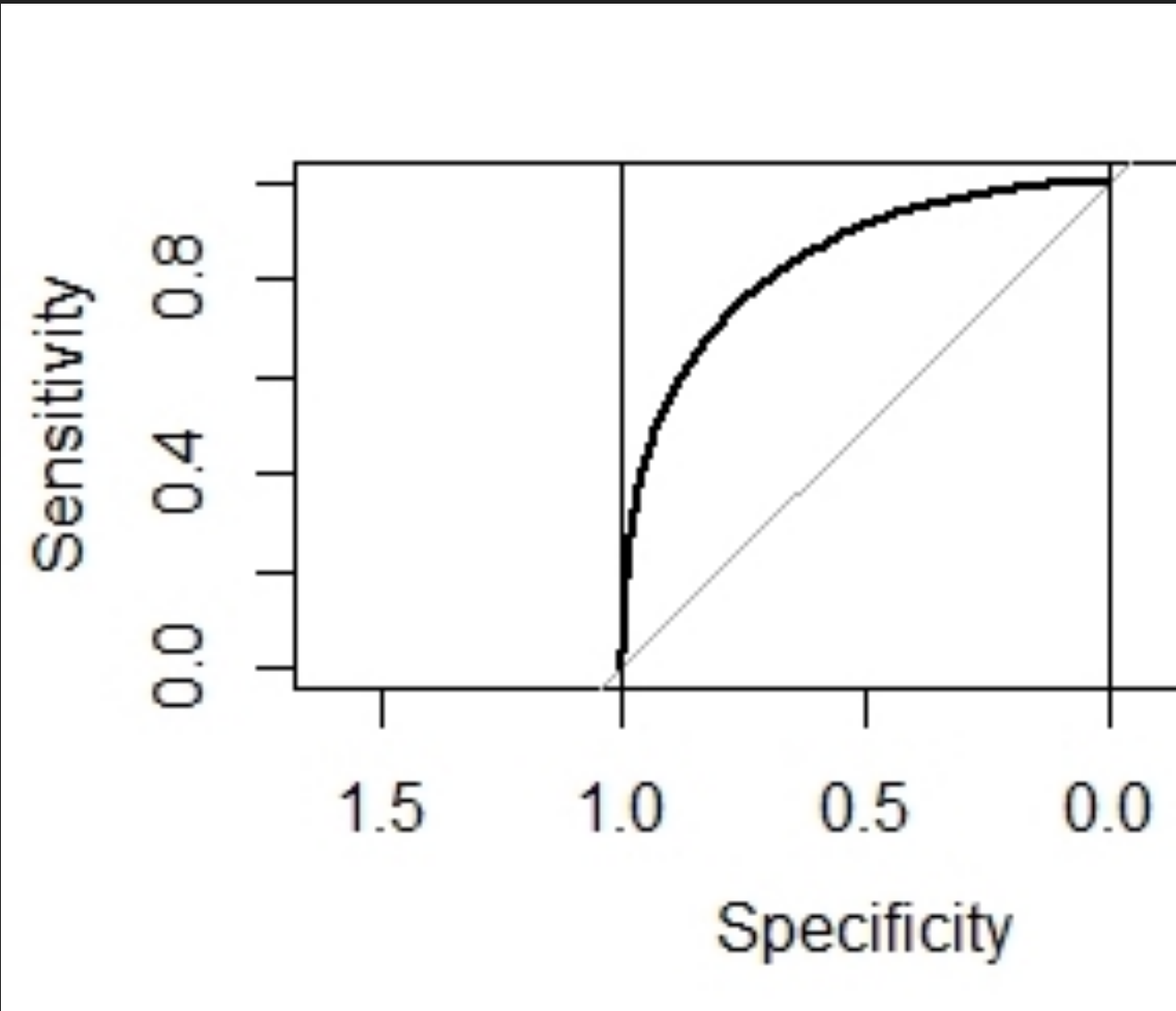
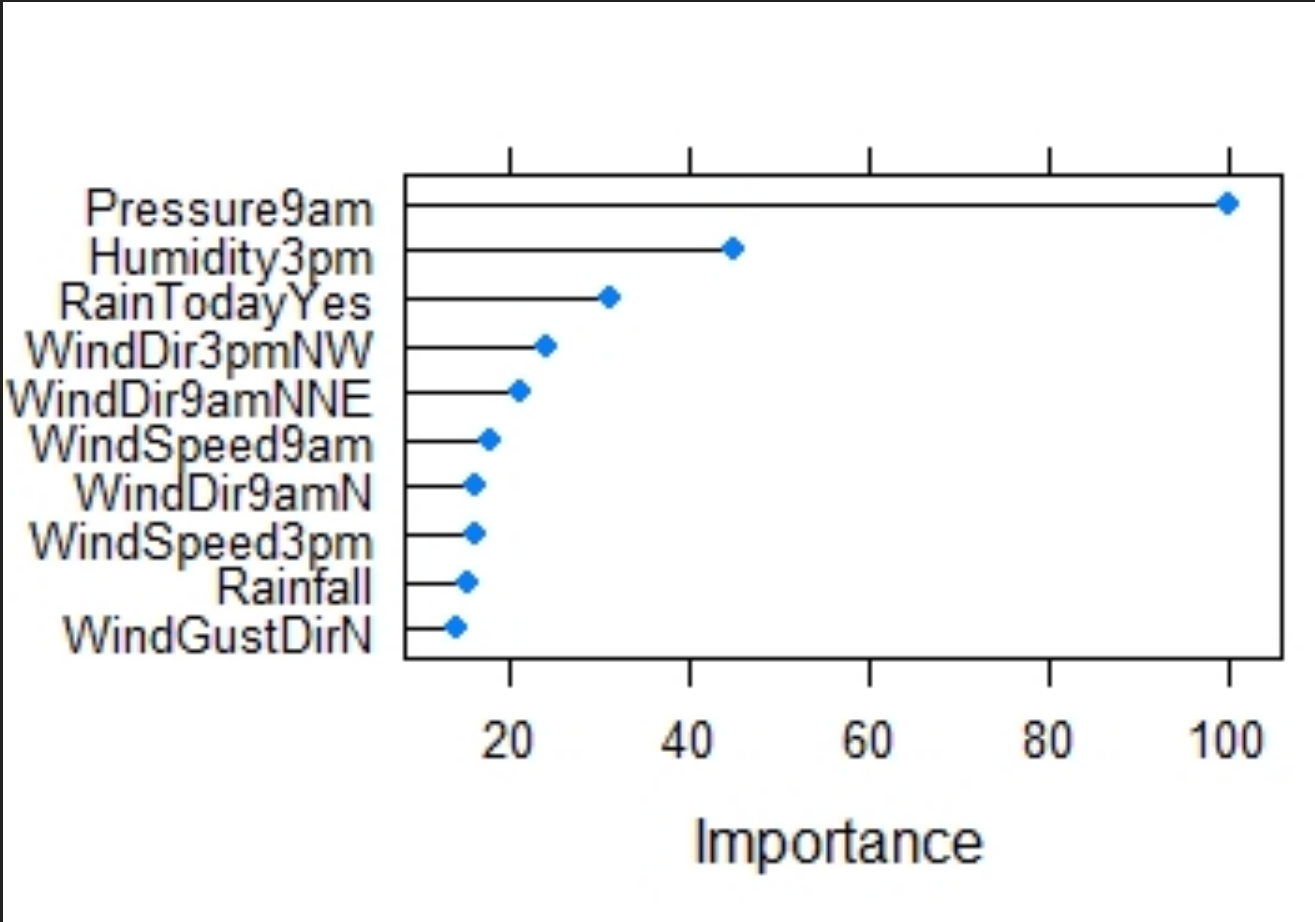
Tempo calcolo predizione:  
1 sec



# NEURAL NETWORK

	YES	NO
YES	3483	4090
NO	1447	24810

Accuracy	0,83
Precision	0,70
Recall	0,46
F1-score	0,55
AUC	0,84



Tempo creazione modello:  
13 min

Tempo calcolo predizione:  
1 sec

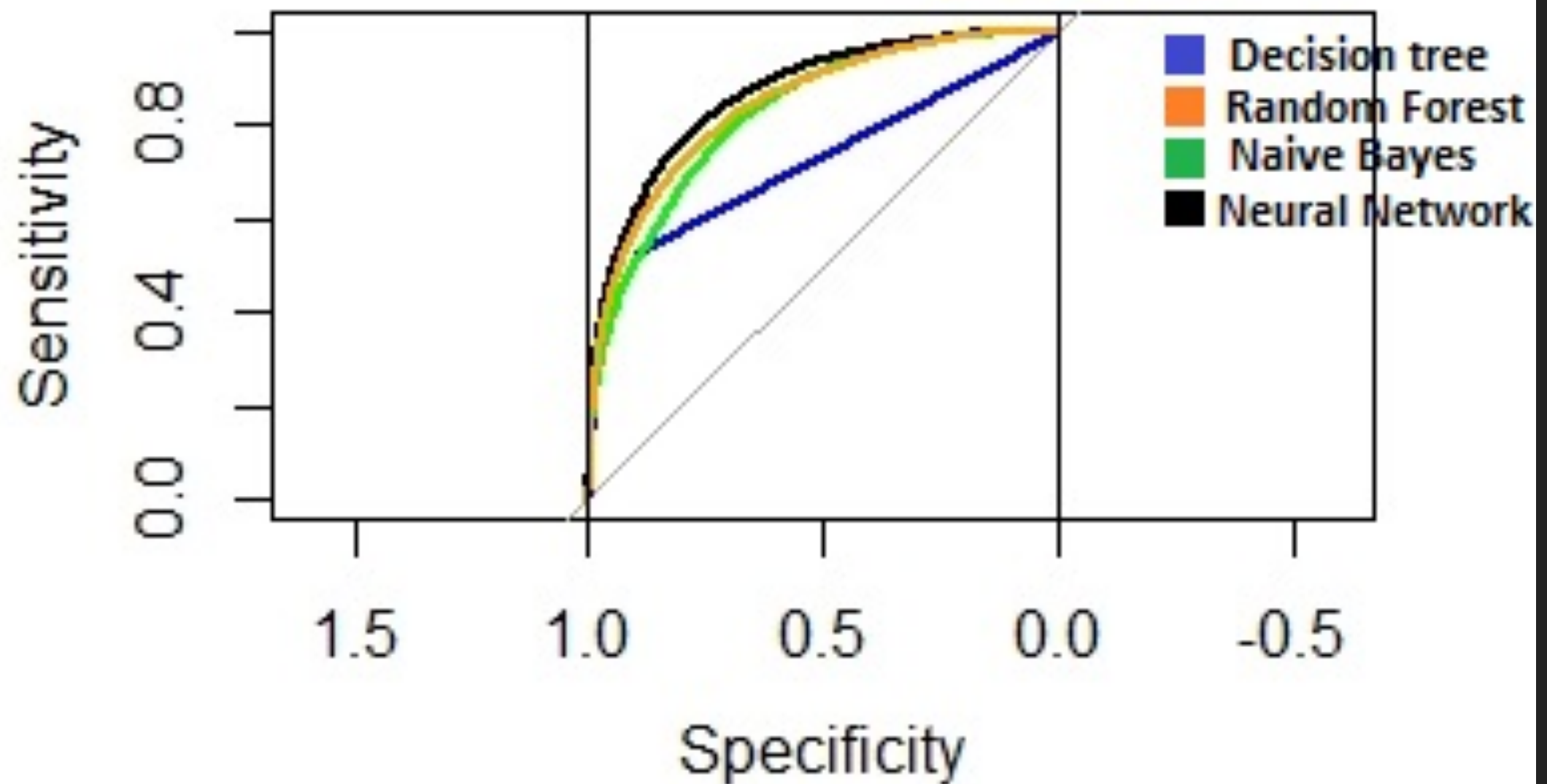
# SVM – RADIAL KERNEL (HOLDOUT)

	YES	NO
YES	2990	787
NO	4583	25470

Accuracy	0,84
Precision	0,79
Recall	0,39
F1-score	0,52

Tempo creazione modello:  
17 min

Tempo calcolo predizione:  
2 min



Dal grafico è facile intuire come i modelli con ROC migliore siano NeuralNetwork e RandomForest, mentre il peggiore è DecisionTree.

# CONCLUSIONI

---

- ▶ La scelta del modello con migliore rapporto performance-tempo è strettamente legata al contesto di utilizzo.
- ▶ Se l'obiettivo è un modello molto veloce, anche sacrificando leggermente le performance, la scelta migliore è DecisionTree, in quanto è risultato veloce, semplice da interpretare e con un buone performance.
- ▶ Al contrario se l'obiettivo è performance elevate, i modelli migliori sono RandomForest e NeuralNetwork.
- ▶ Il modello NaiveBayes è da escludere, infatti ha performance di predizione della classe positiva molto bassa. Mentre il modello SVM è risultato troppo oneroso da allenare.

# SVILUPPI FUTURI

---

- ▶ Migliorare fase di preprocessing, andando a modificare nel modo opportuno i valori nulli.
- ▶ Partire dai risultati di feature importance dei vari modelli, per selezionare le feature più importanti già in fase di preprocessing.
- ▶ Migliorare scelta iperparametri per i vari modelli.
- ▶ Utilizzare modelli più complessi, come ad esempio Deep Learning.