Progetto d'esame di Machine Learning

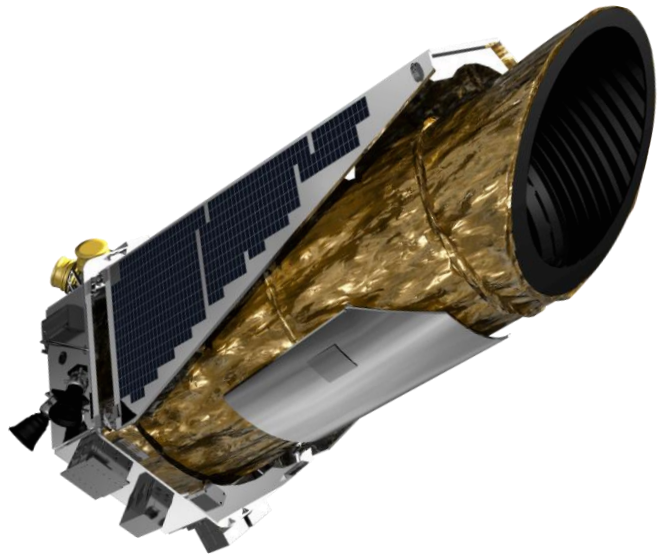
Kepler Exoplanet

Gagianesi Matteo 807316
Talamona Stefano 822452

Dataset

Il dataset originale è composto da 9564 istanze

Ognuna di queste è descritta da 50 variabili, compresa la label target

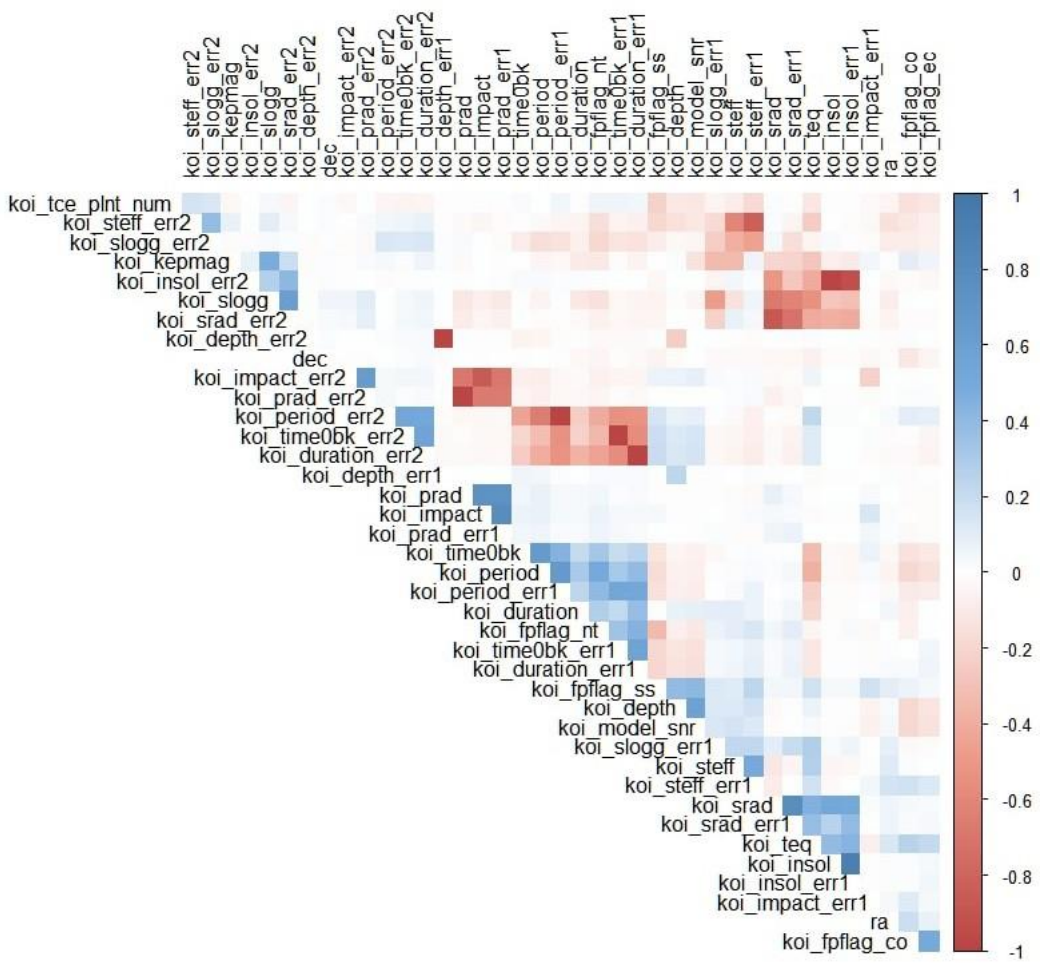


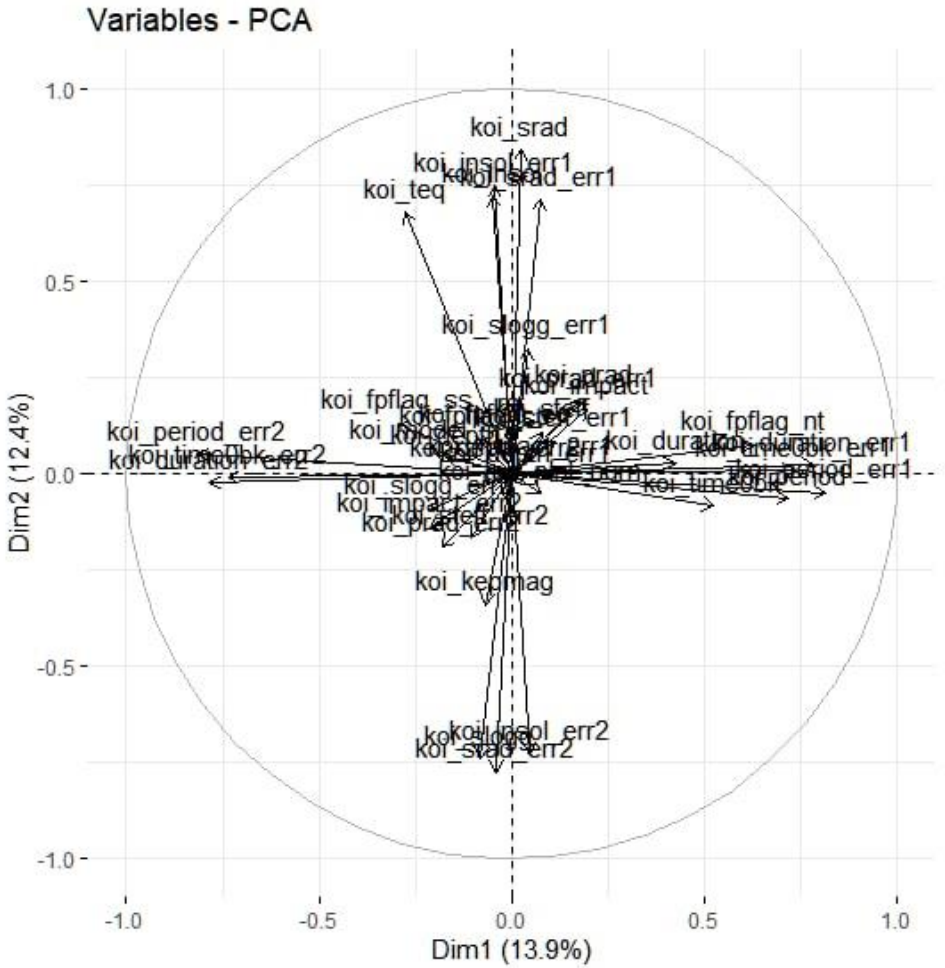
Modifiche ed assunzioni

Vengono rimosse:

- Feature con soli valori nulli
- Istanze con valori nulli
- Istanze con label target "CANDIDATE"

Analisi di Correlazione e PCA





SVM

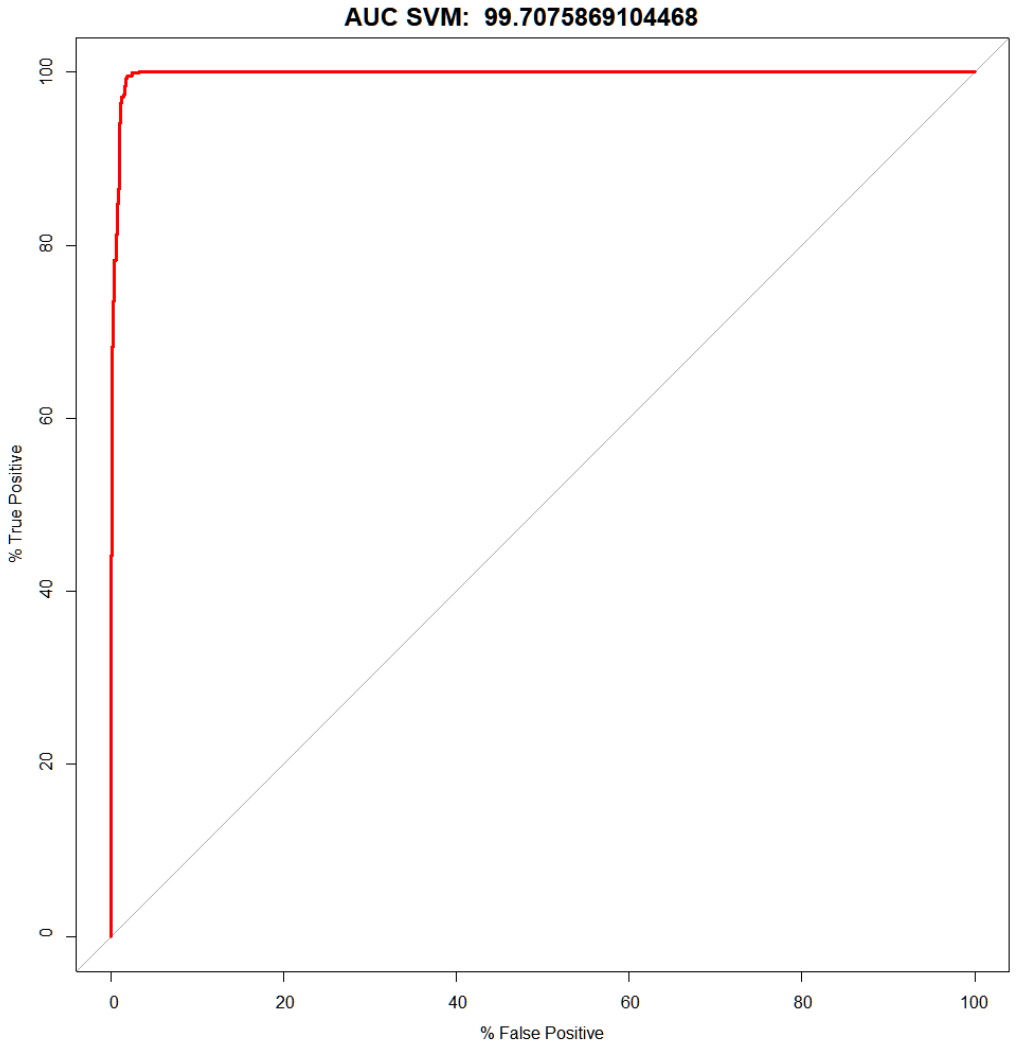
Kernel	Accuracy	Precision	F1 Score	Recall	95% CI
Linear	0.9910	0.9985	0.9869	0.9756	(0.9858, 0.9947)
Polynomial	0.9900	0.9942	0.9855	0.9770	(0.9846, 0.9939)
Radial	0.9925	0.9971	0.9891	0.9813	(0.9877, 0.9958)
Sigmoid	0.9526	0.9426	0.9309	0.9195	(0.9423, 0.9615)

Linear	CONFIRMED	FALSE POSITIVE
CONFIRMED	679	1
FALSE POSITIVE	17	1306

Polynomial	CONFIRMED	FALSE POSITIVE
CONFIRMED	680	4
FALSE POSITIVE	16	1303

Radial	CONFIRMED	FALSE POSITIVE
CONFIRMED	683	2
FALSE POSITIVE	13	1305

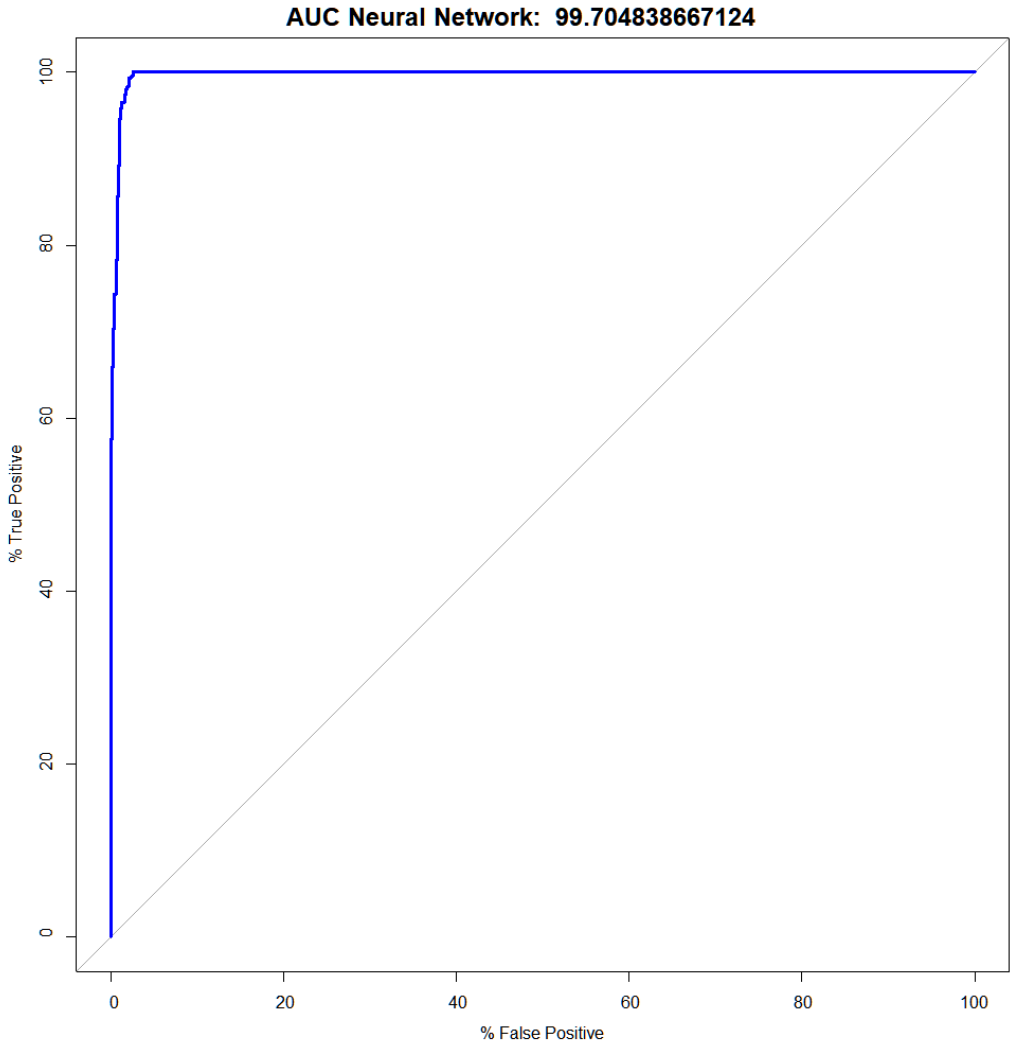
Sigmoid	CONFIRMED	FALSE POSITIVE
CONFIRMED	640	39
FALSE POSITIVE	56	1268



Neural Network

Accuracy	Precision	F1 Score	Recall	95% CI
0.9830	0.9729	0.9756	0.9784	(0.9764, 0.9882)

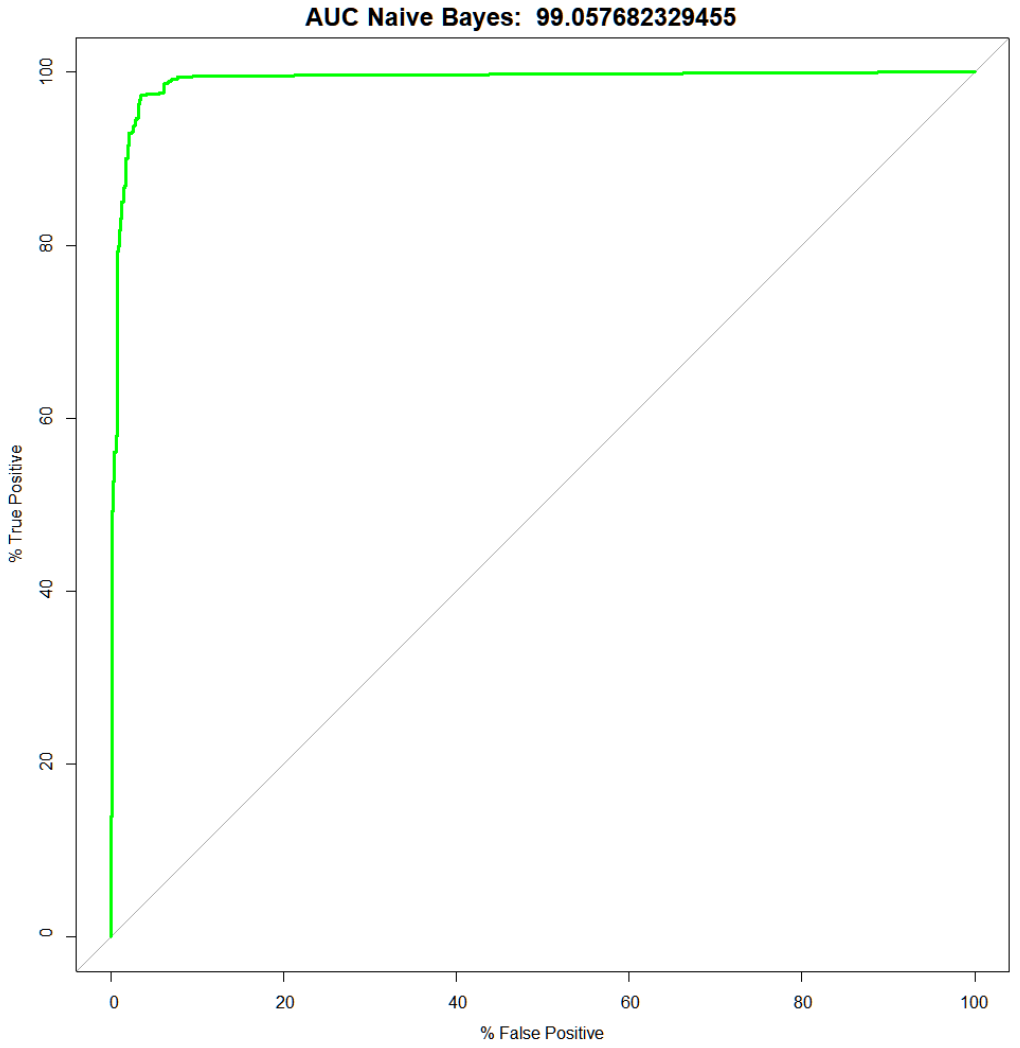
Prediction\Reference	CONFIRMED	FALSE POSITIVE
CONFIRMED	681	19
FALSE POSITIVE	15	1288



Naive Bayes

Accuracy	Precision	F1 Score	Recall	95% CI
0.9651	0.9347	0.9506	0.9670	(0.9561, 0.9727)

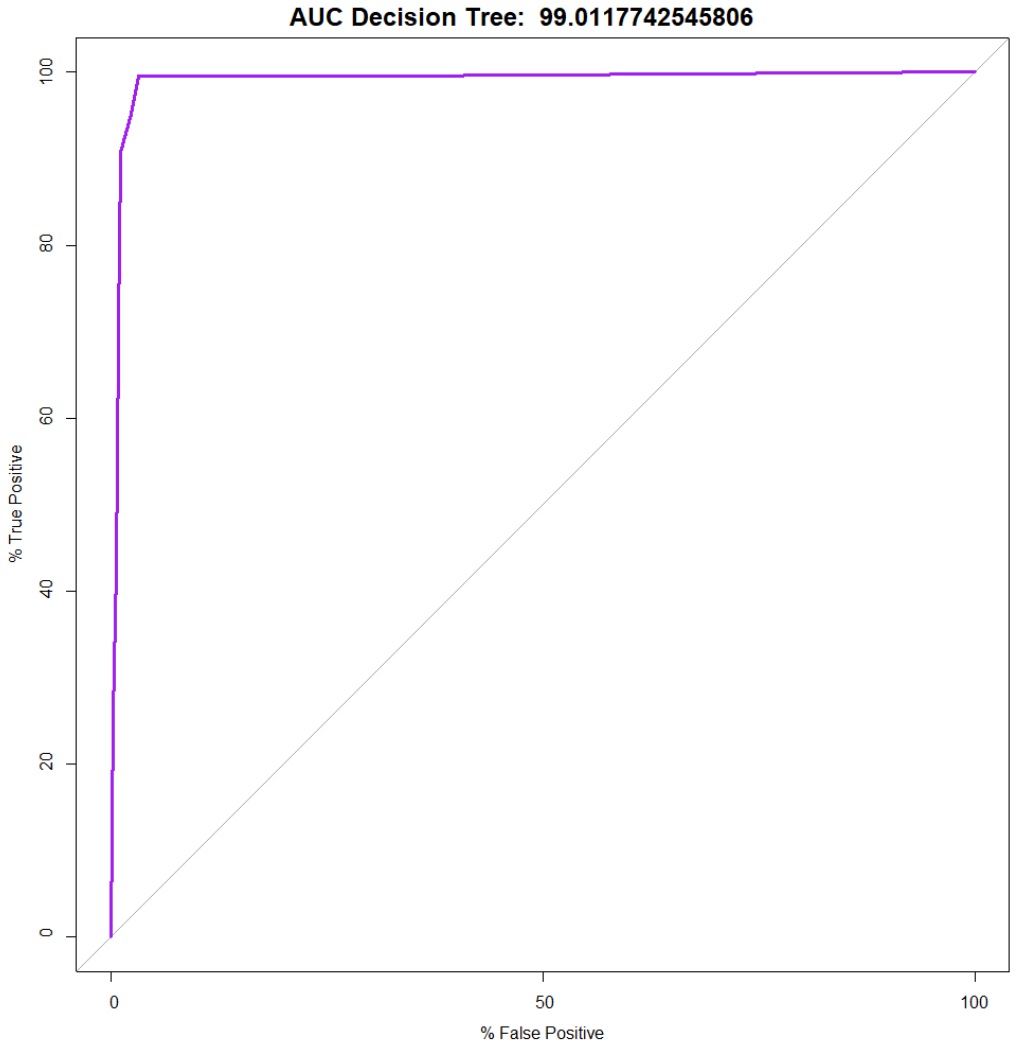
Prediction\Reference	CONFIRMED	FALSE POSITIVE
CONFIRMED	673	47
FALSE POSITIVE	23	1260



Decision Tree

Accuracy	Precision	F1 Score	Recall	95% CI
0.9820	0.9853	0.9738	0.9626	(0.9752, 0.9874)

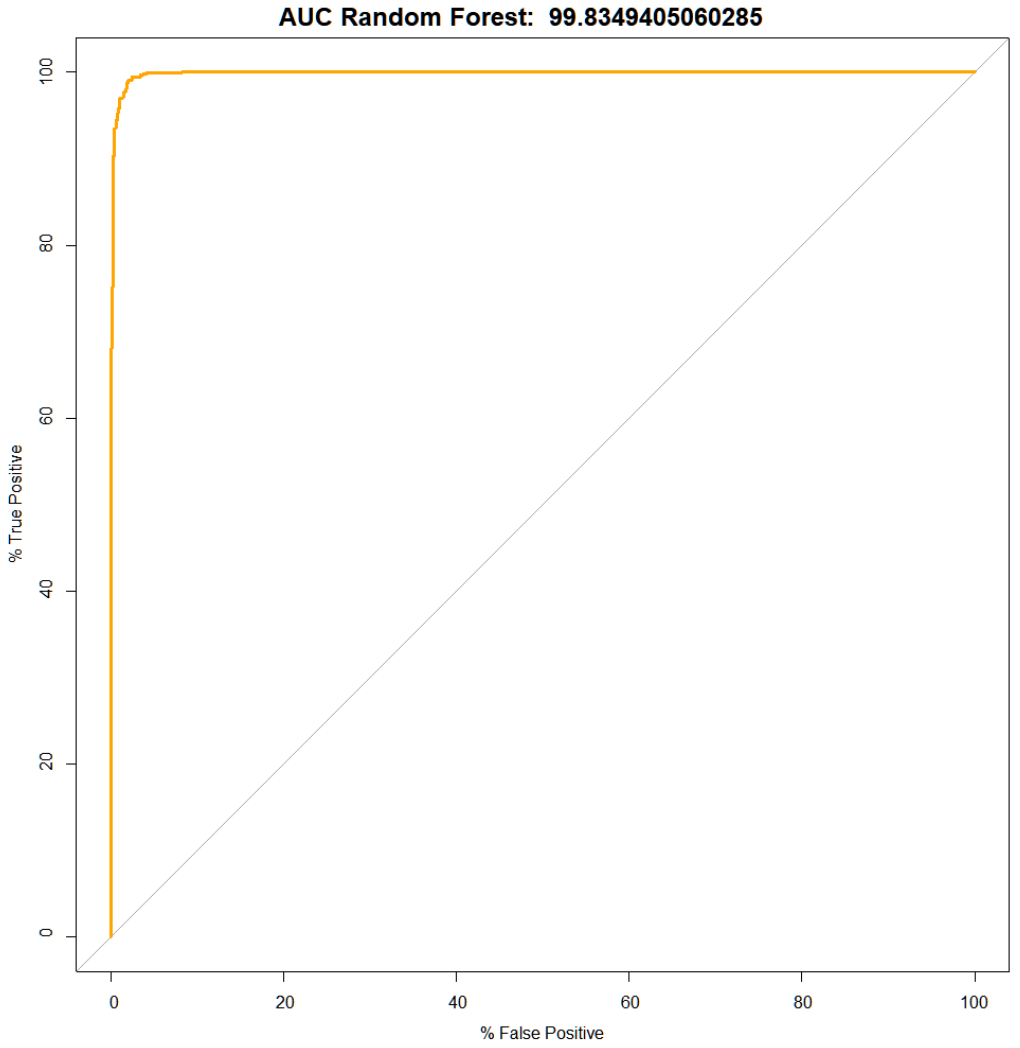
Prediction\Reference	CONFIRMED	FALSE POSITIVE
CONFIRMED	670	10
FALSE POSITIVE	26	1297



Random Forest

Accuracy	Precision	F1 Score	Recall	95% CI
0.9885	0.9985	0.9832	0.9684	(0.9828, 0.9927)

Prediction\Reference	CONFIRMED	FALSE POSITIVE
CONFIRMED	674	1
FALSE POSITIVE	22	1306



Considerazioni sui risultati

Model	Time (s)	Accuracy	Precision	F1 Score	Recall	AUC
SVM	77.25	0.9925	0.9971	0.9891	0.9813	0.9971
Neural Network	132.06	0.9830	0.9729	0.9756	0.9784	0.9970
Naive Bayes	3.75	0.9651	0.9347	0.9506	0.9670	0.9906
Decision Tree	2.41	0.9820	0.9853	0.9738	0.9626	0.9901
Random Forest	318.80	0.9885	0.9985	0.9832	0.9684	0.9983

Considerazioni finali

Performance ottime grazie a:

- Elevata varianza di una buona parte delle variabili del dataset
- Sbilanciamento dei valori delle variabili FLAG
- Sbilanciamento delle label target