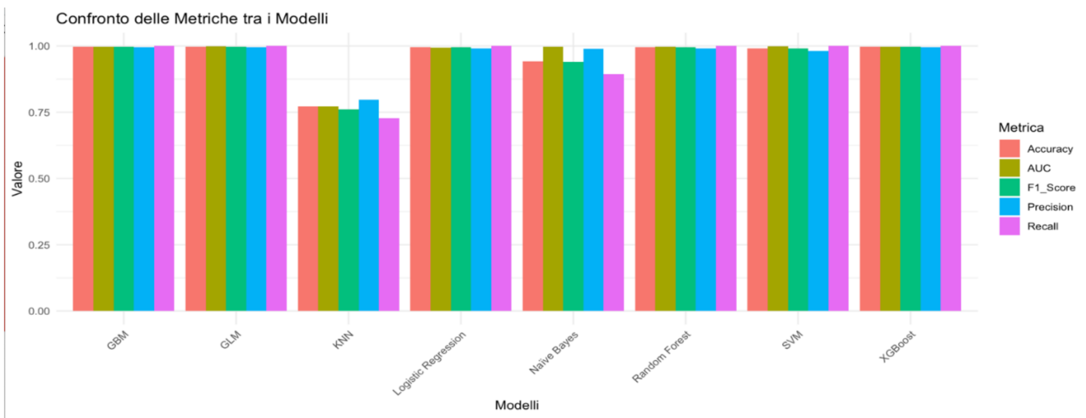
Appendice – Grafici sui risultati ottenuti

Prestazioni sul Dataset Human-generated E-mails

- **Caso 1**: I modelli ML classici (RF, LR, XGBoost) hanno ottenuto ottime prestazioni; KNN è risultato meno efficace; le feature numeriche iniziali sono risultate già altamente predittive;
- **Caso 2**: L'aggiunta di feature semantiche ha migliorato tutte le metriche; XGBoost ha raggiunto performance perfette comportandosi come un modello ideale; CNN, RNN, LSTM mostrano miglioramenti notevoli;
- **Caso 3**: I modelli LLM di tipo Transformer (RoBERTa, DeBERTa, BERT, DistilBERT, XLNet) hanno superato le reti neurali classiche grazie alla capacità di discriminazione semantica avanzata; i modelli RoBERTa, DeBERTa, DistilBERT, BERT, XLNet si distinguono con F1 e AUC prossimi a 1, confermando l'efficacia dell'approccio LLM nella comprensione del contenuto testuale.

Caso 1 – Feature numeriche iniziali – dataset Human Generated E-mail:



	Modello	Accuracy	Precision	Recall	F1_Score	AUC
1	Random Forest	0.9953917	0.9908676	1.0000000	0.9954128	0.9970906
2	XGBoost	0.9976959	0.9954128	1.0000000	0.9977011	0.9976959
3	Logistic Regression	0.9953917	0.9908676	1.0000000	0.9954128	0.9946697
4	GLM	0.9976959	0.9954128	1.0000000	0.9977011	0.9994054
5	SVM	0.9907834	0.9819005	1.0000000	0.9908676	0.9991930
6	KNN	0.7718894	0.7979798	0.7281106	0.7614458	0.7718894
7	GBM	0.9976959	0.9954128	1.0000000	0.9977011	0.9978233
8	Naïve Bayes	0.9423963	0.9897959	0.8940092	0.9394673	0.9976959

Figura 1 - Confronto metriche modelli ML (Accuracy, Precision, Recall, F1, AUC) – Human generated - Caso 1

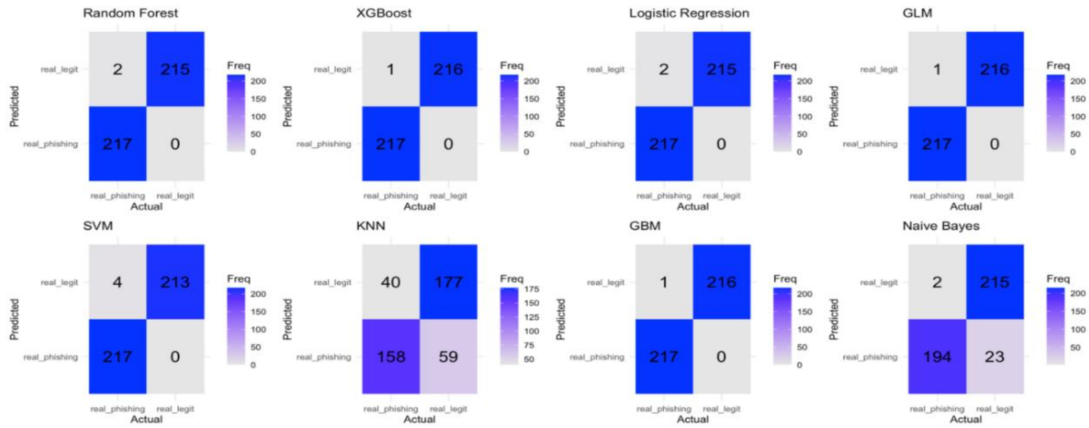


Figura 2 - Matrici di confusione - Human generated - Caso 1

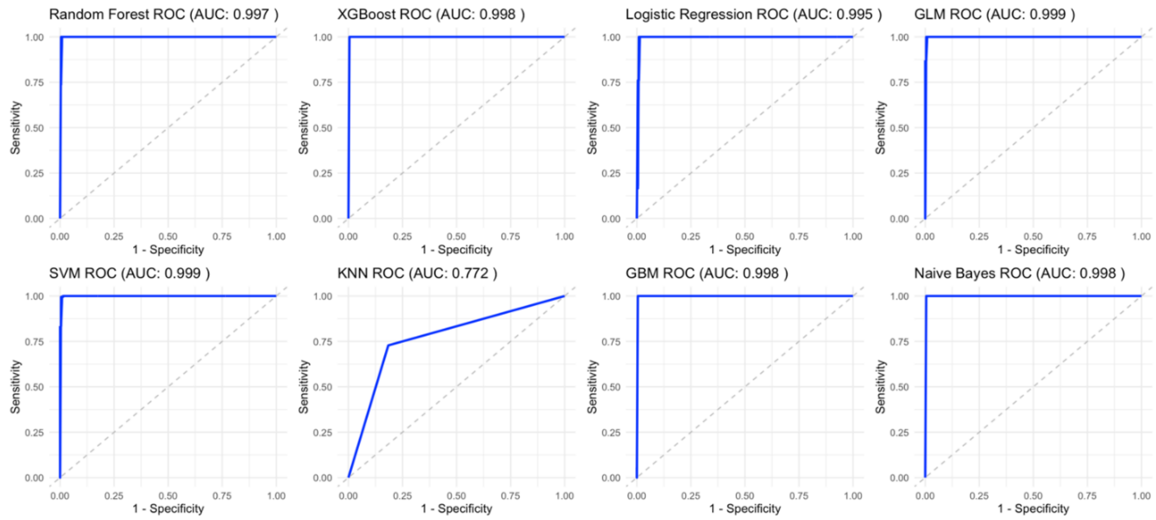


Figura 3 - Curve ROC-AUC - Human generated - Caso 1

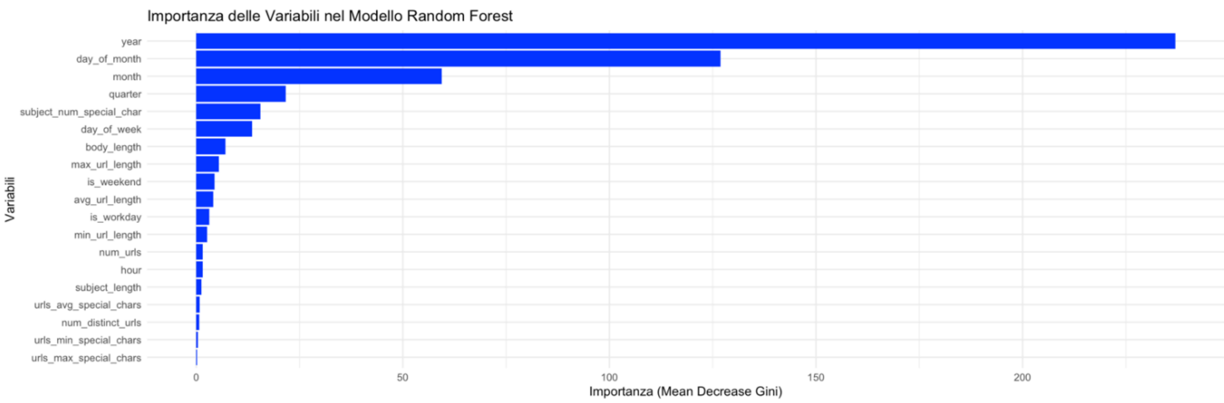
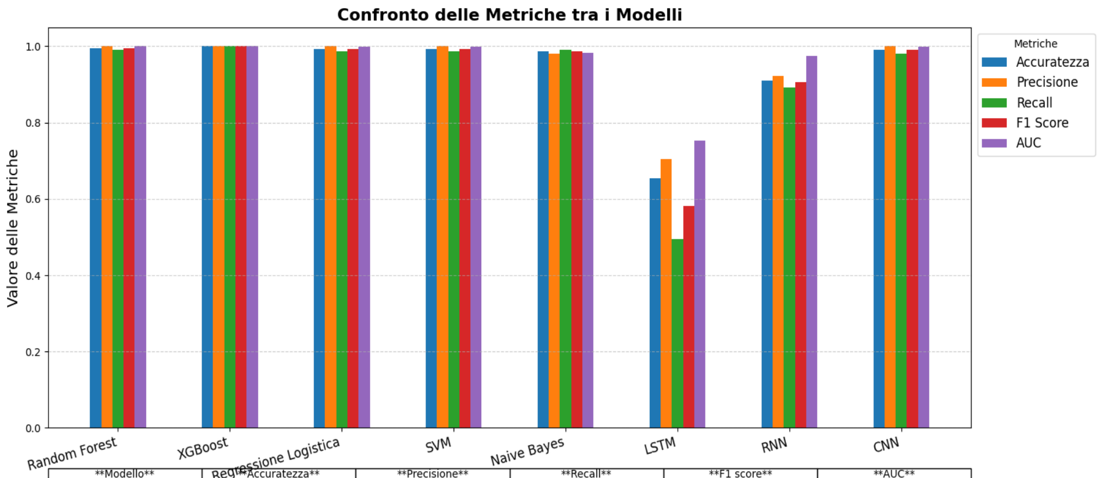


Figura 4 -Importanza variabili Random Forest - Human generated - Caso 1

Caso 2 – Tutte le feature numeriche (con topic, analisi del sentimento e embeddings):



Modello	**Accuratezza**	**Precisione**	**Recall**	**F1 score**	**AUC**
Random Forest	0.9954	1.0	0.9906	0.9953	1.0
XGBoost	1.0	1.0	1.0	1.0	1.0
Regressione Logistica	0.9931	1.0	0.9858	0.9929	0.9992
SVM	0.9931	1.0	0.9858	0.9929	0.9988
Naive Bayes	0.9862	0.9813	0.9906	0.9859	0.9821
LSTM	0.6537	0.7047	0.4953	0.5817	0.7532
RNN	0.9106	0.922	0.8915	0.9065	0.9746
CNN	0.9908	1.0	0.9811	0.9905	0.9977

Figura 5 - Confronto metriche - Human generated - Caso 2

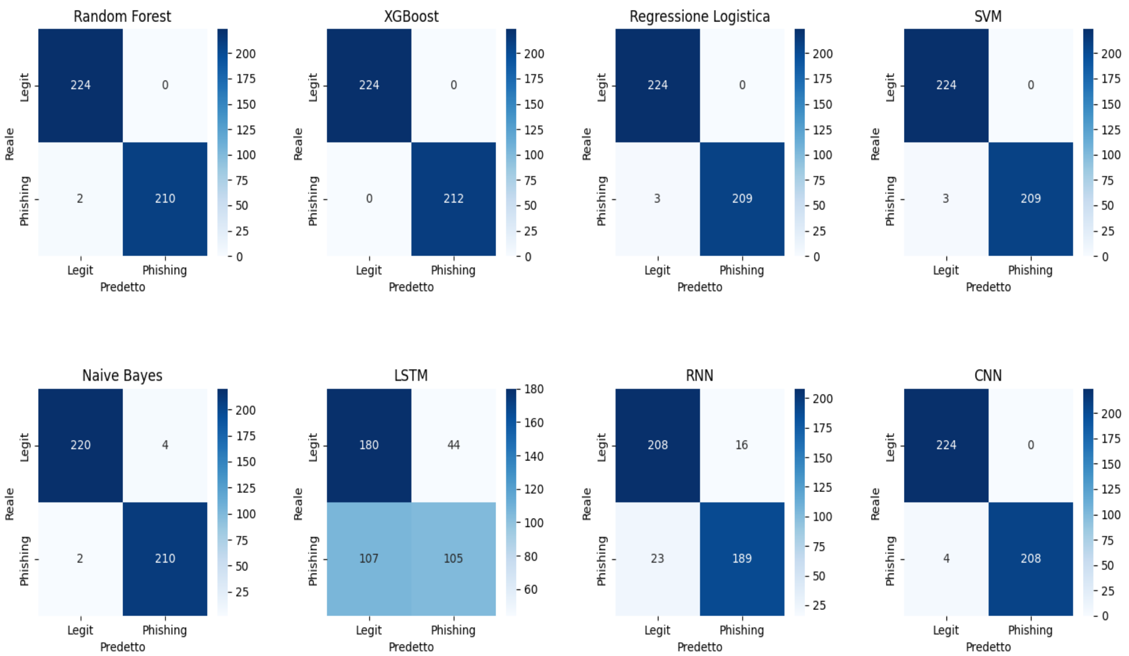


Figura 6 - Matrici di confusione Human generated - Caso 2

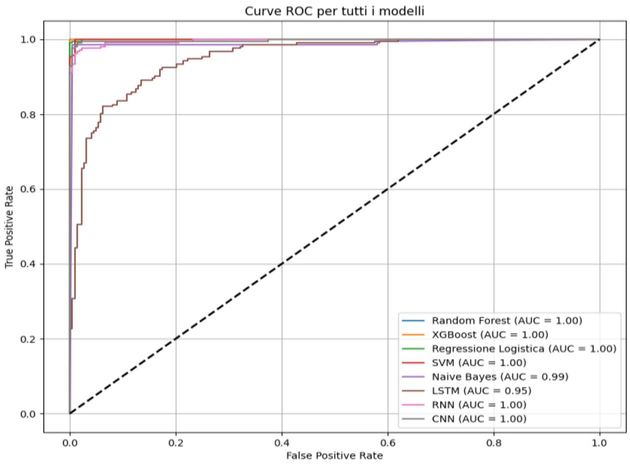


Figura 7 - Curve ROC - Human generated - Caso 2

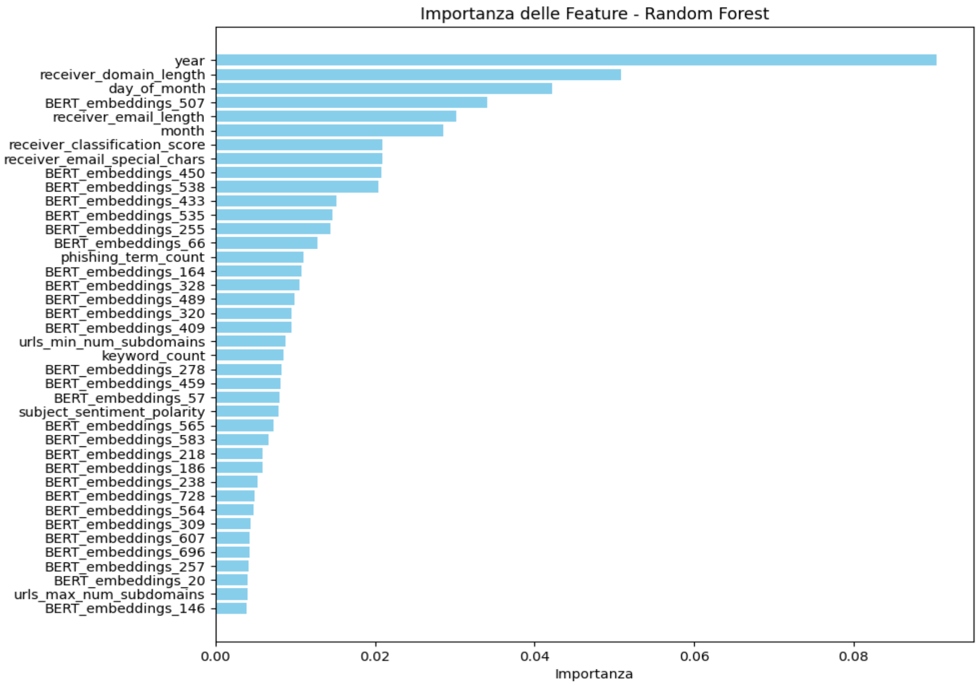


Figura 8 - Importanza feature per Random Forest - Human generated - Caso 2

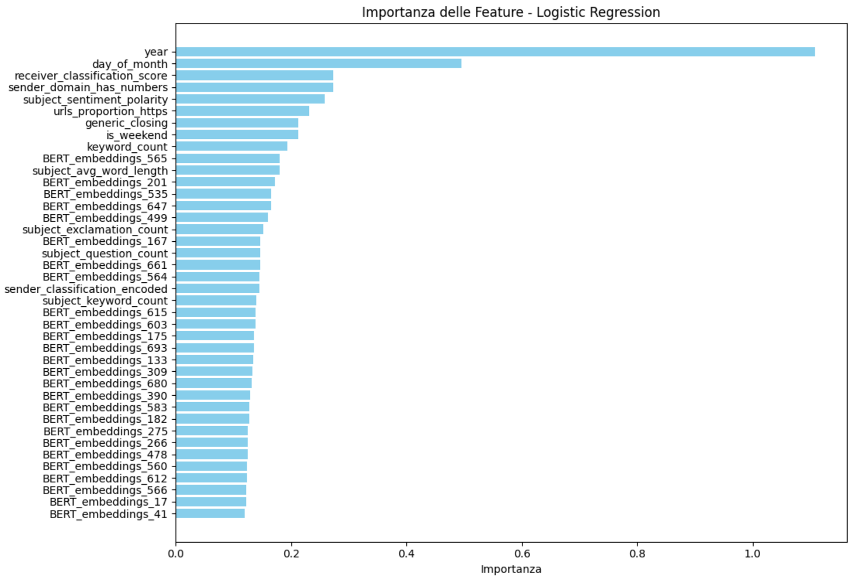
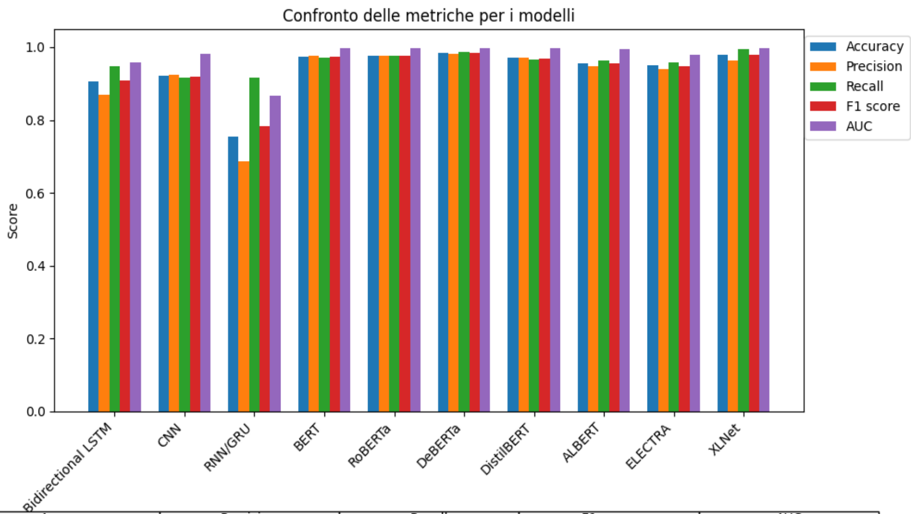


Figura 9 - Importanza feature per Regressione Logistica - Human generated - Caso 2

Caso 3 – Solo cleaned_body testuale:



	Accuracy	Precision	Recall	F1 score	AUC
Bidirectional LSTM	0.91	0.87	0.95	0.91	0.96
CNN	0.92	0.92	0.92	0.92	0.98
RNN/GRU	0.75	0.69	0.92	0.78	0.87
BERT	0.97	0.98	0.97	0.97	1.0
RoBERTa	0.98	0.98	0.98	0.98	1.0
DeBERTa	0.98	0.98	0.99	0.98	1.0
DistilBERT	0.97	0.97	0.97	0.97	1.0
ALBERT	0.96	0.95	0.96	0.96	0.99
ELECTRA	0.95	0.94	0.96	0.95	0.98
XLNet	0.98	0.96	1.0	0.98	1.0

Figure 10 - Confronto metriche - Human generated - caso 3

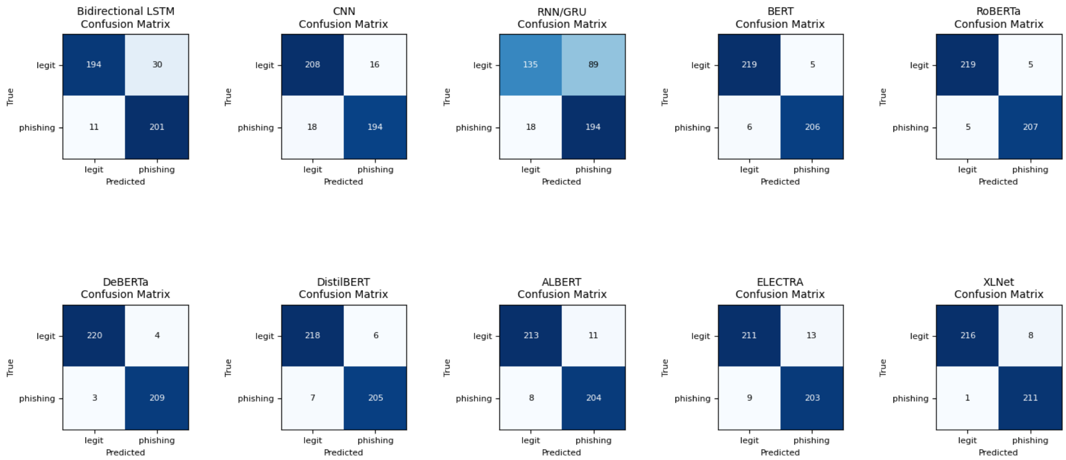


Figura 1 - Matrici di confusione - Human generated - caso 3

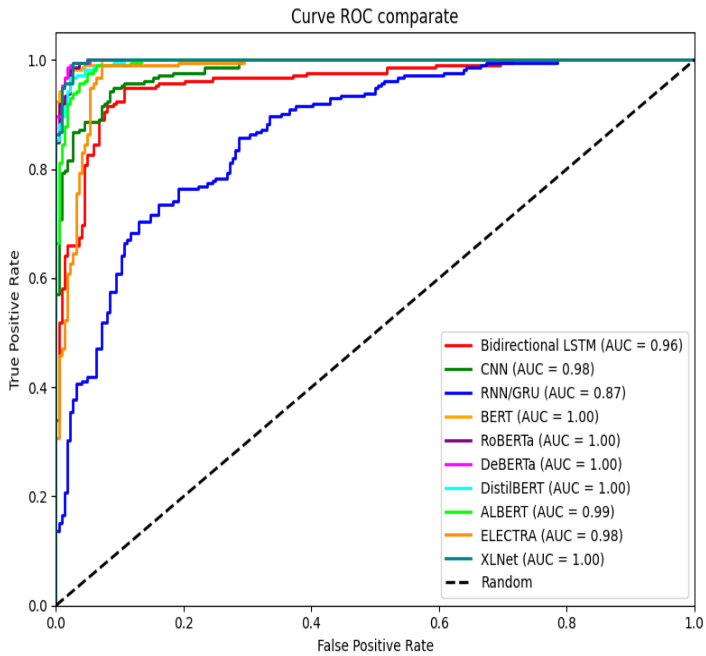


Figura 3 - Curve ROC - Human generated - caso 3

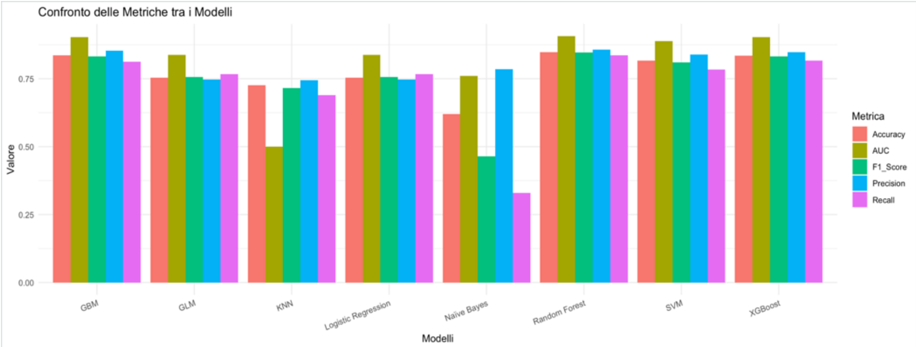
Prestazioni sul Dataset LLM-generated E-mails

I risultati mostrano una diminuzione delle performance complessive, evidenziando come le e-mail generate da LLM siano meno prevedibili e più difficili da rilevare rispetto a quelle generate da esseri umani:

- **Caso 1**: Tutti i modelli hanno registrato un calo significativo delle prestazioni; le feature numeriche iniziali non sono sufficienti a rilevare i testi generati da LLM;
- **Caso 2**: L'integrazione di feature semantiche migliora nettamente le performance; la Regressione Logistica risulta essere il miglior modello; CNN, SVM e XGBoost mostrano performance molto elevate, mentre LSTM risulta debole;

- **Caso 3**: I modelli LLM (DistilBERT, RoBERTa, XLNet) mostrano le migliori performance, confermando la loro efficacia nel riconoscere testo artificiale; DistilBERT e XLNet si confermano come i modelli più performanti.

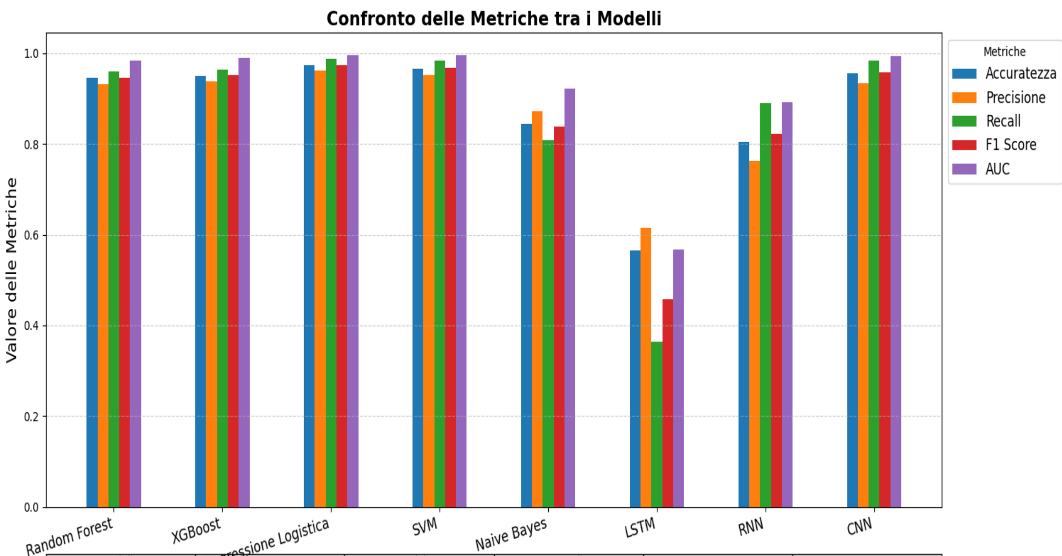
Caso 1 – Feature numeriche iniziali:



	Modello	Accuracy	Precision	Recall	F1_Score	AUC
1	Random Forest	0.8483333	0.8566553	0.8366667	0.8465430	0.9064167
2	XGBoost	0.8350000	0.8477509	0.8166667	0.8319185	0.9036278
3	Logistic Regression	0.7533333	0.7467532	0.7666667	0.7565789	0.8372000
4	GLM	0.7533333	0.7467532	0.7666667	0.7565789	0.8372111
5	SVM	0.8166667	0.8392857	0.7833333	0.8103448	0.8888333
6	KNN	0.7266667	0.7446043	0.6900000	0.7162630	0.5000000
7	GBM	0.8366667	0.8531469	0.8133333	0.8327645	0.9036278
8	Naïve Bayes	0.6200000	0.7857143	0.3300000	0.4647887	0.7605778

Figura 13 - Confronto metriche - LLM generated - caso 1

Caso 2 – Tutte le feature numeriche:



Modello	**Accuratezza**	**Precisione**	**Recall**	**F1 score**	**AUC**
Random Forest	0.945	0.9325	0.9603	0.9462	0.9841
XGBoost	0.95	0.9387	0.9636	0.951	0.9904
Regressione Logistica	0.9733	0.9613	0.9868	0.9739	0.9957
SVM	0.9667	0.9519	0.9834	0.9674	0.9958
Naive Bayes	0.8433	0.8714	0.8079	0.8385	0.9219
LSTM	0.565	0.6145	0.3642	0.4574	0.5673
RNN	0.805	0.762	0.8907	0.8214	0.892
CNN	0.9567	0.934	0.9834	0.9581	0.9939

Figura 14 - Confronto metriche - LLM generated - Caso 2

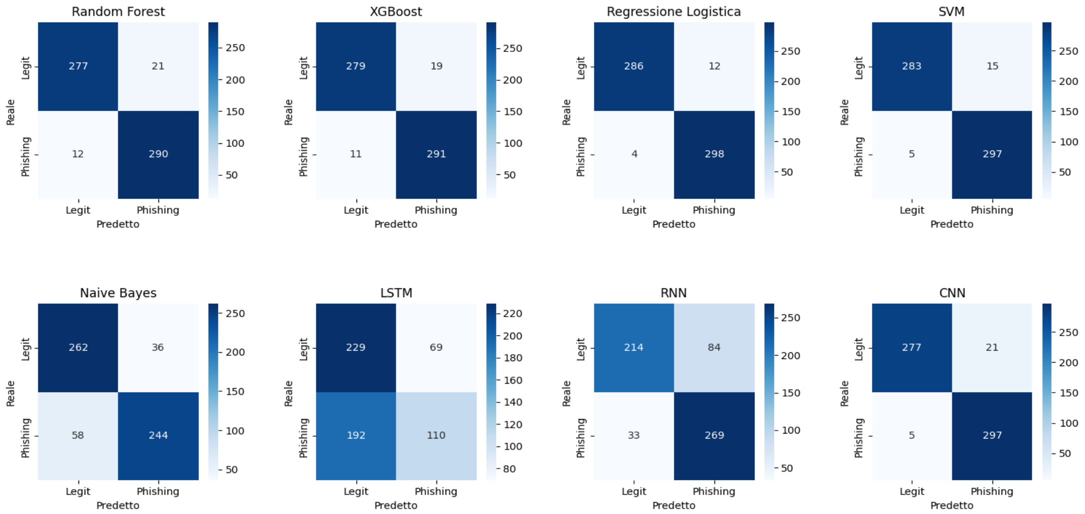


Figura 15 - Matrici di confusione - LLM generated - Caso 2

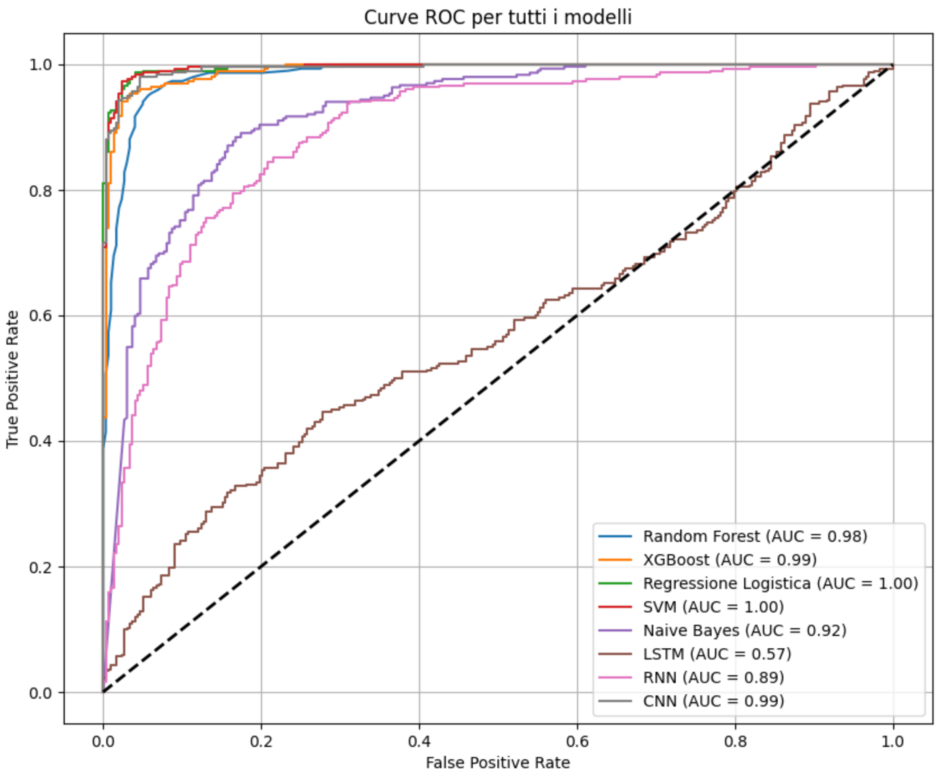


Figura 16 - Curve ROC - LLM generated - Caso 2

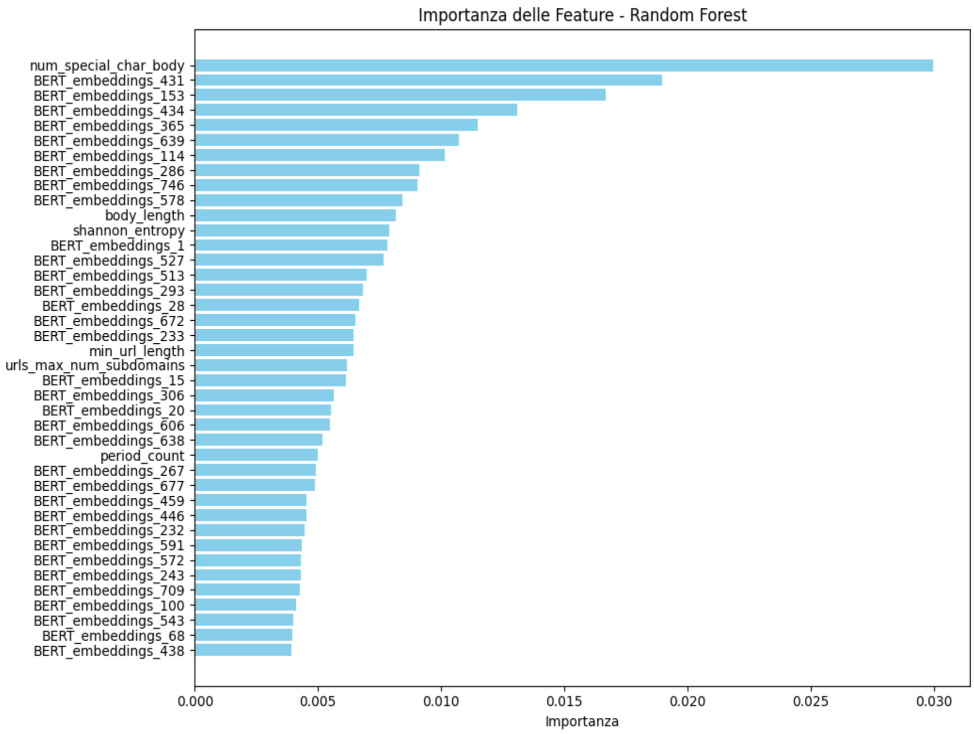


Figura 17 - Importanza delle variabili per Random Forest - LLM generated - Caso 2

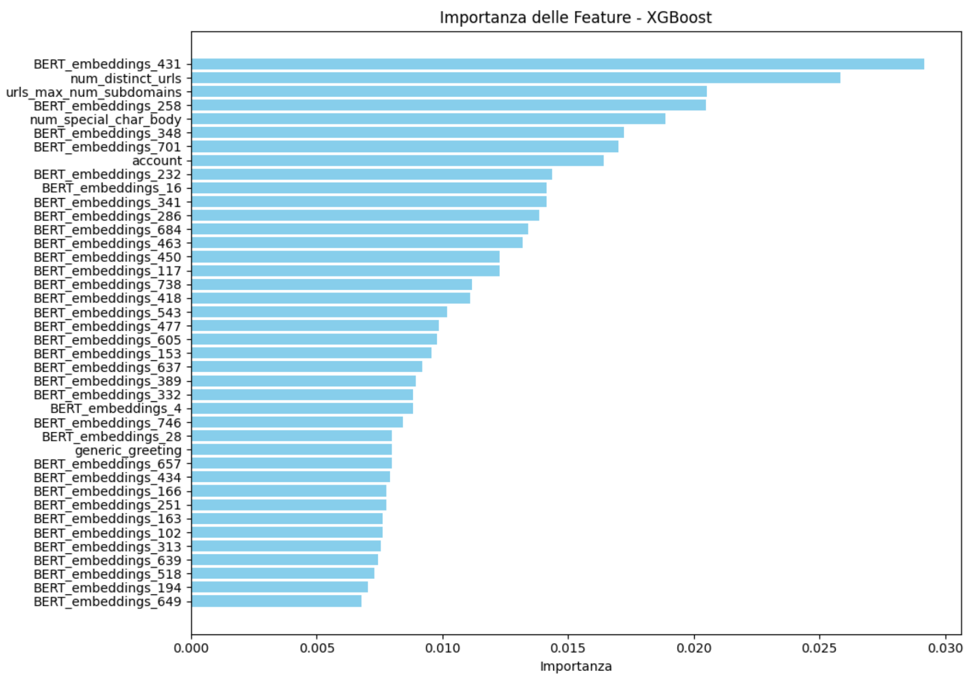


Figura 18 - Importanza variabili per XGBoost - LLM generated - Caso 2

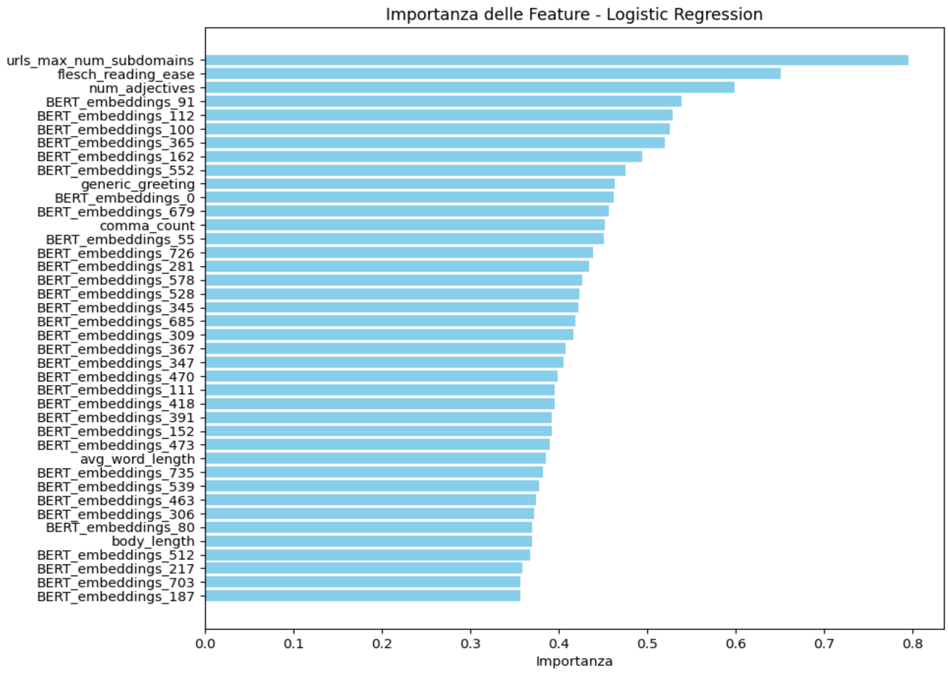
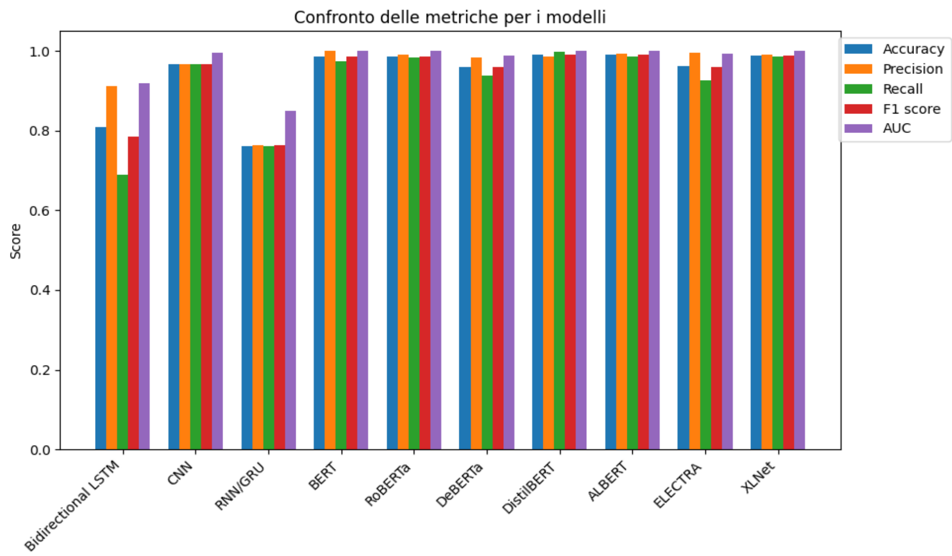


Figura 19 - Importanza variabili per Regressione Logistica - LLM generated - Caso 2

Caso 3 – Solo Body testuale ripulito:



	Accuracy	Precision	Recall	F1 score	AUC
Bidirectional LSTM	0.81	0.91	0.69	0.78	0.92
CNN	0.97	0.97	0.97	0.97	1.0
RNN/GRU	0.76	0.76	0.76	0.76	0.85
BERT	0.99	1.0	0.97	0.99	1.0
RoBERTa	0.99	0.99	0.98	0.99	1.0
DeBERTa	0.96	0.98	0.94	0.96	0.99
DistilBERT	0.99	0.99	1.0	0.99	1.0
ALBERT	0.99	0.99	0.99	0.99	1.0
ELECTRA	0.96	1.0	0.93	0.96	0.99
XLNet	0.99	0.99	0.99	0.99	1.0

Figura 20 - Confronto metriche - LLM generated - Caso 3

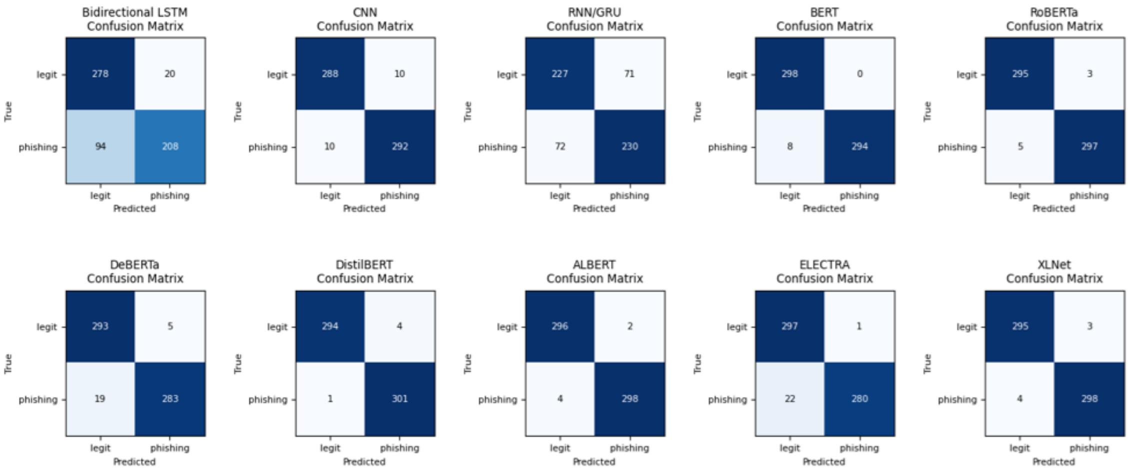


Figura 21 - Matrici di confusione - LLM generated - Caso 3

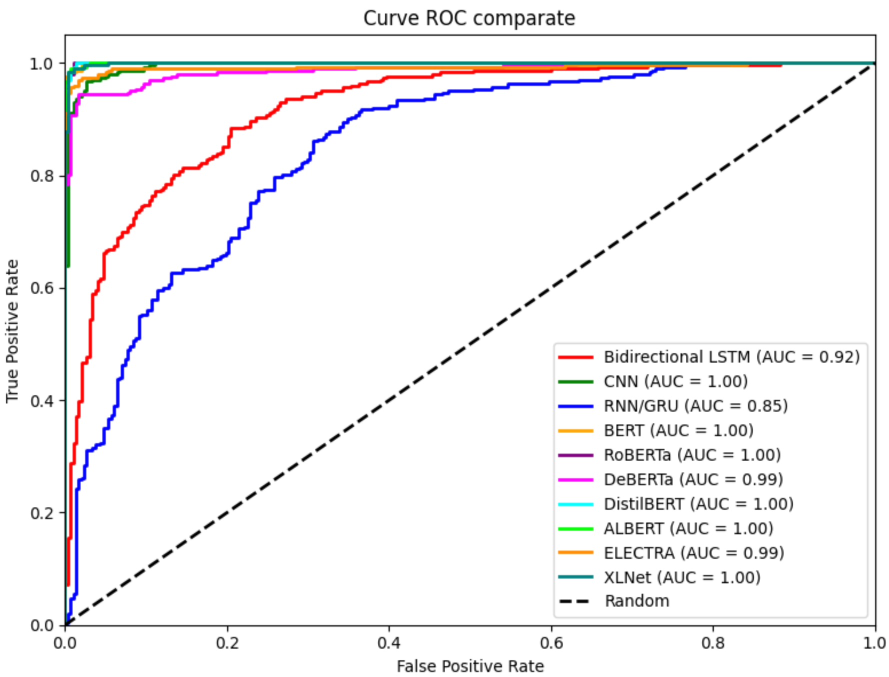


Figura 22 - Curve ROC-AUC - LLM generated - Caso 3