OverseerAI

Identificazione di contenuti Truffa tramite analisi testuale

Antonio Maiorano

Università degli Studi di Salerno

Corso di Fondamenti di Intelligenza Artificiale

Gennaio 2025

https://github.com/BiBooBap/OverseerAI

Introduzione

Problema: La crescente diffusione di contenuti truffaldini su piattaforme come YouTube.
Obiettivo: Creare un modello di Machine Learning per classificare i titoli dei video come "scam" o "legit".
Motivazione:

- Proteggere gli utenti da contenuti fuorvianti.
- Ridurre l'impatto dei contenuti generati automaticamente con scopi dannosi.

Per praticità, il linguaggio usato per il Modello è l'inglese.

Specifica PEAS

P.E.A.S. in questo progetto:

- **Performance:** correttezza di classificazion;
- **Environment:** dataset sintetico;
- **Actuators:** classificatore che determinerà l'appartenenza di un'istanza ad una delle classi;
- **Sensors:** la finestra di input.

Caratteristiche dell'ambiente

Ambiente in cui agisce l'agente:

- **Osservabilità:** parzialmente osservabile;
- **Determinismo:** parzialmente stocastico;
- **Episodicità:** episodico;
- **Dinamismo:** statico;
- **Tipo di agente:** singolo agente.

Approccio alla risoluzione: CRISP-DM

Le fasi principali:

1. **Business Understanding:** Comprendere il problema e definire gli obiettivi.
2. **Data Understanding:** Analisi dei dati e individuazione delle caratteristiche.
3. **Data Preparation:** Pulizia, normalizzazione e trasformazione dei dati.
4. **Modeling:** Addestramento del modello e ottimizzazione.
5. **Evaluation:** Valutazione delle prestazioni.
6. **Deployment:** Implementazione e utilizzo del modello.

Business Understanding

Analisi della problematica:
i contenuti hanno informazioni che risaltano la loro natura di tipo truffaldino, ad esempio determinate parole come "free", o presenza di link nella descrizione (per evitare una possibile moderazione);

Obiettivi di Business:
categorizzare i contenuti in base alle informazioni su di essi;

Criteri di successo:

- un corretto riconoscimento dei contenuti intenzionati a truffare;
- un corretta classificazione dei contenuti normali come leciti;
- evitare di creare bias con chi ha link legittimi nei propri contenuti.

Data Understanding (1)

Origine dei Dati:

- Dati sintetici generati con LLMs (Large Language Models).
- Tre principali caratteristiche:
  - `title` - Titolo del video.
  - `description` - Descrizione del video.
  - `link_desc` - Presenza di link nella descrizione.
- Caratteristiche (con valori) aggiunte per migliorare il dataset:
  - `upload_hour` - Ora di caricamento del contenuto.
  - `likes` - Numero di "mi piace".
  - `dislikes` - Numero di "non mi piace".
  - `comments` - Numero di "commenti".

Data Understanding (2)

```
title,label,description,upload_hour,likes,dislikes,comments
How to Cook Pasta in 10 Minutes,legit,Learn how to cook delicious pasta in just 10 minutes. No experience needed!,20.0,29187.0,2101.0,4003.0
FREE iPhone Giveaway! Just Click Now!!!,scam,Claim your free iPhone now! Visit http://scam-website.com for more details.,11.0,99399.0,9827.0,965.0
Win a Brand New Tesla - Enter Now!,scam,Your chance to win a Tesla! Visit: http://tesla-giveaway.com,11.0,38331.0,3574.0,824.0
Discover the Secrets of Success!,legit,Explore proven strategies for achieving personal and professional success.,10.0,11243.0,801.0,8093.0
WARNING: Your Account is Compromised!,scam,Secure your account now at http://security-alert.com,21.0,47063.0,5698.0,5594.0
Learn Python Programming - Beginner to Advanced,legit,Master Python with our step-by-step guide. Start coding today!,0.0,36959.0,8379.0,7953.0
Get Rich Quick with This Simple Trick!,scam,Discover the secret formula to wealth! More info at http://fakeprofits.net,15.0,4666280.0,727310.0,1901.0
5 Easy Exercises for a Healthier Life,legit,Quick and effective exercises to boost your health and fitness.,,49933.0,3630.0,100.0
Click Here to Claim Your Reward!,scam,Your exclusive prize is waiting. Claim it here: http://reward-link.com,19.0,67550.0,1317.0,8486.0
Shocking Discovery That Banks Don't Want You to Know!,scam,Learn the truth now: http://bank-secrets.com,15.0,53616.0,2958.0,5028.0
Your Computer is at Risk! Fix It Now!,scam,Fix your system vulnerabilities at http://quickfix.com,2.0,12119.0,3144.0,
Top 10 Travel Destinations for 2024,legit,Explore the most beautiful travel spots for 2024. Plan your trip now!,19.0,3045804.0,536687.0,588.0
Your Account Will Be Deleted in 24 Hours!,scam,Immediate action required! Secure your account at http://urgentfix.net,22.0,22941.0,4162.0,2983.0
10 Facts You Didn't Know About Space,legit,Learn amazing facts about space and the universe.,4.0,99983.0,7338.0,4874.0
###ERROR###,scam,Don't lose your chance! Register now at http://exclusive-offer.com,12.0,39817.0,9534.0,1862.0
Become a Python Expert in 30 Days,legit,Intensive Python course designed to make you an expert in one month.,19.0,38801.0,491.0,4827.0
Immediate Action Required to Secure Your Funds!,scam,Protect your funds now! Follow the link: http://safe-bank.com,10.0,87361.0,8806.0,2084.0
New AI Technology That Changes Everything!,legit,Explore groundbreaking AI innovations changing the world.,8.0,6611.0,214.0,5200.0
Congratulations! You've Won a $1000 Gift Card!,scam,Claim your $1000 gift card at http://prizezone.net,16.0,13257.0,3977.0,7572.0
Get a Brand New Laptop for FREE!,scam,Claim your free laptop here: http://free-laptop.org,13.0,9984.0,6628.0,5855.0
Understanding Machine Learning Algorithms,legit,Discover the basics of ML algorithms and how they work.,8.0,43485.0,3385.0,9585.0
```

Figure: Parte del dataset di partenza

Data Preparation (1)

Data Cleaning:

- Rimozione manuale di anomalie non risolvibili e outliers (tramite IQR);
- Imputazione statistica dei valori mancanti (per mediana);
- Normalizzazione dei testi tramite metodologie standard:
  - *Contraction Expansion*;
  - *Tokenizzazione*;
  - *Conversione dei numeri (in cifre) in parole*;
  - *Lemmatizzazione*;
  - *Trasformazione in Minuscolo*;
  - *Stopword Removal*.

Data Preparation (2)

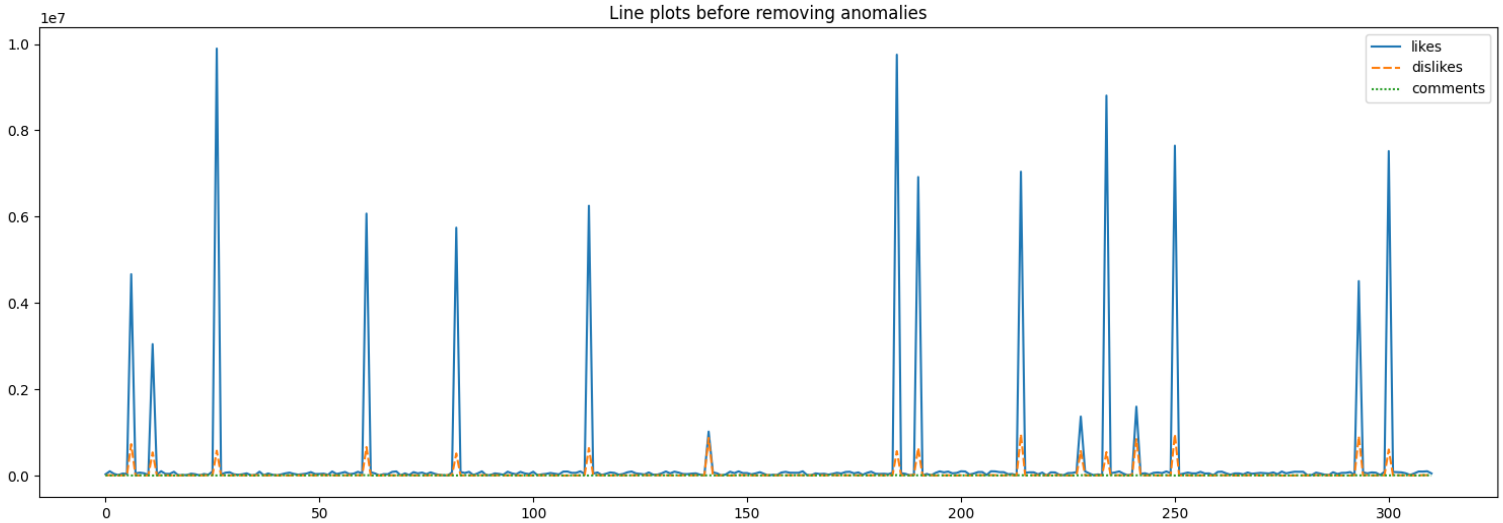


Figure: Prima della rimozione degli Outliers

Data Preparation (3)

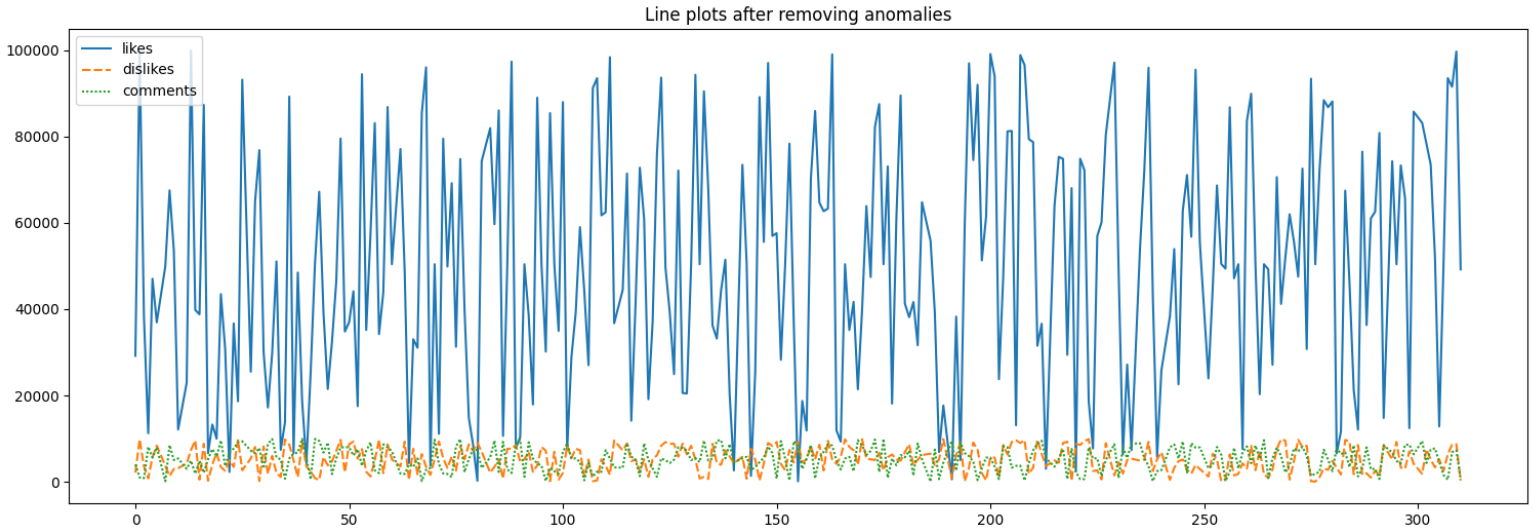


Figure: Dopo la rimozione degli Outliers

Data Preparation (4)

Feature Scaling:
Normalizziamo i valori numerici, con rilevanza matematica e non in un contesto proprio, usando la **Min-Max Normalization**.

$$x' = a + \frac{(x - min(x)) \cdot (b - a)}{max(x) - min(x)}$$

Data Preparation (5)

Feature Selection:
La feature riguardante la descrizione è divisa in due elementi separati al proprio interno:

- il testo della descrizione;
- possibili link.

Dunque andiamo ad estrarre i link, eliminandoli dalla feature di provenienza, inserendoli in una nuova feature "`link_desc`".
Possiamo ora eseguire la normalizzazione del testo anche su essa.
Le feature numeriche possono essere rimosse, per la loro natura indipendente e casuale (non sono collegate al contenuto, ma all'interazione su quest'ultimo).

Data Preparation (6)

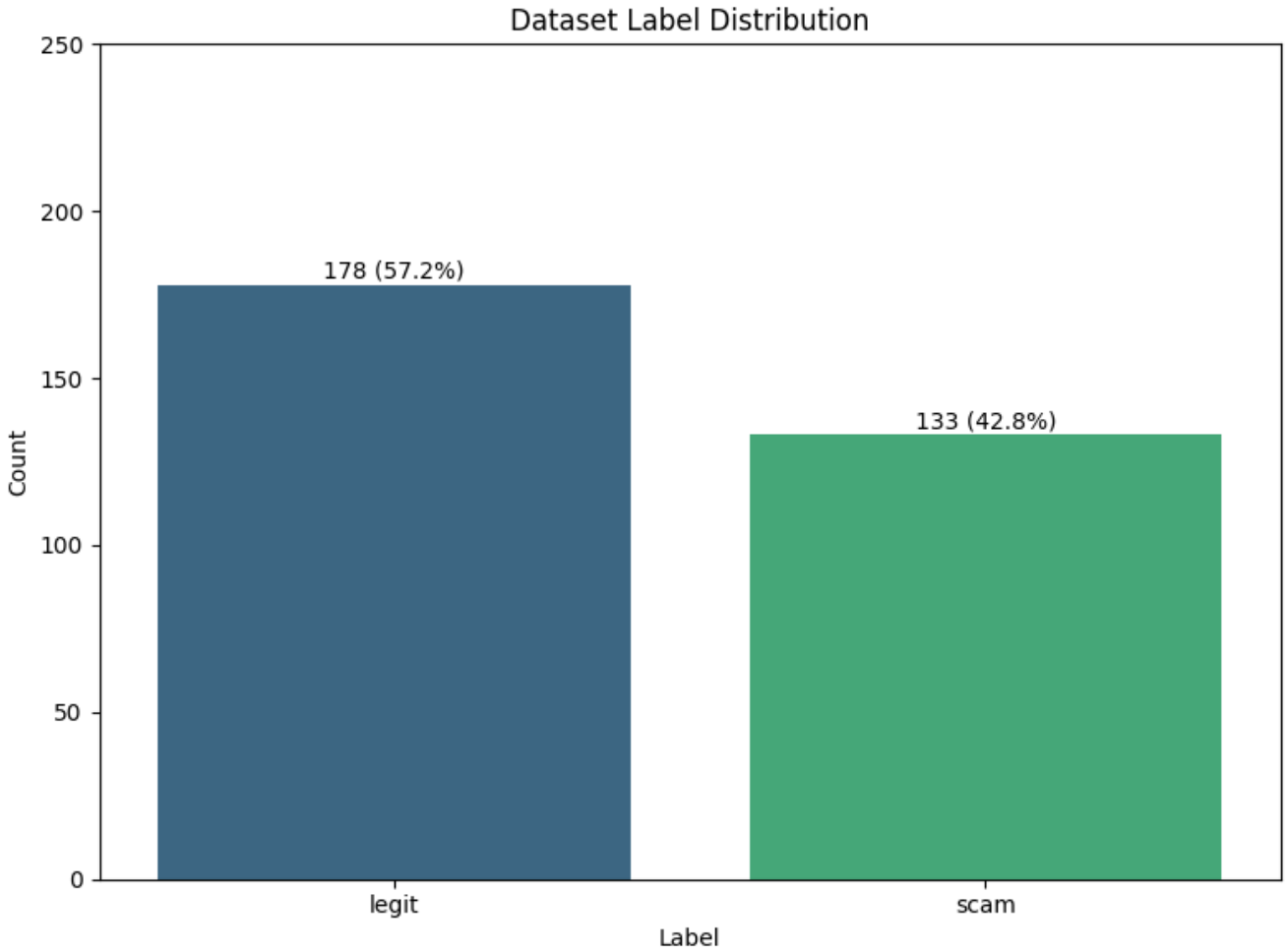


Figure: Dataset equilibrato, non necessario il Data Balancing

Data Preparation (7)

```
title,label,description,link_desc
cook pasta ten minutes,legit,learn cook delicious pasta ten minute experience needed,no link
free iphone giveaway click,scam,claim free iphone visit detail,http scam website com
win brand new tesla enter,scam,chance win tesla visit,http tesla giveaway com
discover secrets success,legit,explore proven strategy achieving personal professional success,no link
warning account compromised,scam,secure account,http security alert com
learn python programming beginner advanced,legit,master python stepbystep guide start coding today,no link
get rich quick simple trick,scam,discover secret formula wealth info,http fakeprofits net
five easy exercises healthier life,legit,quick effective exercise boost health fitness,no link
click claim reward,scam,exclusive prize waiting claim,http reward link com
shocking discovery banks dont want know,scam,learn truth,http bank secrets com
computer risk fix,scam,fix system vulnerability,http quickfix com
top ten travel destinations two thousand and twenty-four,legit,explore beautiful travel spot two thousand and twenty-four plan trip,no link
account deleted twenty-four hours,scam,immediate action required secure account,http urgentfix net
ten facts didnt know space,legit,learn amazing fact space universe,no link
become python expert thirty days,legit,intensive python course designed make expert one month,no link
immediate action required secure funds,scam,protect fund follow link,http safe bank com
new ai technology changes everything,legit,explore groundbreaking ai innovation changing world,no link
congratulations youve one thousand gift card,scam,claim one thousand gift card,http prizezone net
get brand new laptop free,scam,claim free laptop,http free laptop org
understanding machine learning algorithms,legit,discover basic ml algorithm work,no link
win free gift card limited time,scam,dont miss click,http freeswag com
```

Figure: Parte del dataset finale

Data Modeling: Algoritmi Considerati

Naïve Bayes:

- Semplice e veloce;
- Limite principale: assume l'indipendenza tra le feature, non adatto per dati correlati.

Random Forest:

- Robustezza elevata;
- Capacità di gestire feature correlate e pattern nascosti;
- Ensemble di alberi e risultato finale a "voto di maggioranza".

Scelta Finale: Random Forest, per la sua capacità di adattarsi meglio al problema in analisi.

Data Modeling: Training del Modello

Training:

- Preprocessing delle feature con TF-IDF (Term Frequency-Inverse Document Frequency).
- Classificazione tramite Random Forest con 100 alberi paralleli.

Validazione del Training:

- *K-Fold Cross Validation* (10-fold), split del 90% per i dati di training e 10% per i dati di validazione. Ad ogni iterazione lo split di validazione viene re-immesso nello split di training e da quello di training viene presto il prossimo 1/10.
- Metriche calcolate per ogni fold:
  - Accuracy;
  - Precision;
  - Recall;
  - F1-Score.

Evaluation: Metriche di Valutazione (1)

Fold	Accuracy	Precision	Recall	F1
1	0.968	0.981	0.917	0.945
2	1.000	1.000	1.000	1.000
3	0.903	0.900	0.921	0.902
4	0.871	0.900	0.867	0.868
5	0.903	0.932	0.875	0.892
6	0.742	0.818	0.765	0.735
7	0.935	0.950	0.923	0.932
8	0.968	0.976	0.955	0.964
9	0.967	0.972	0.962	0.966
10	0.900	0.925	0.885	0.894

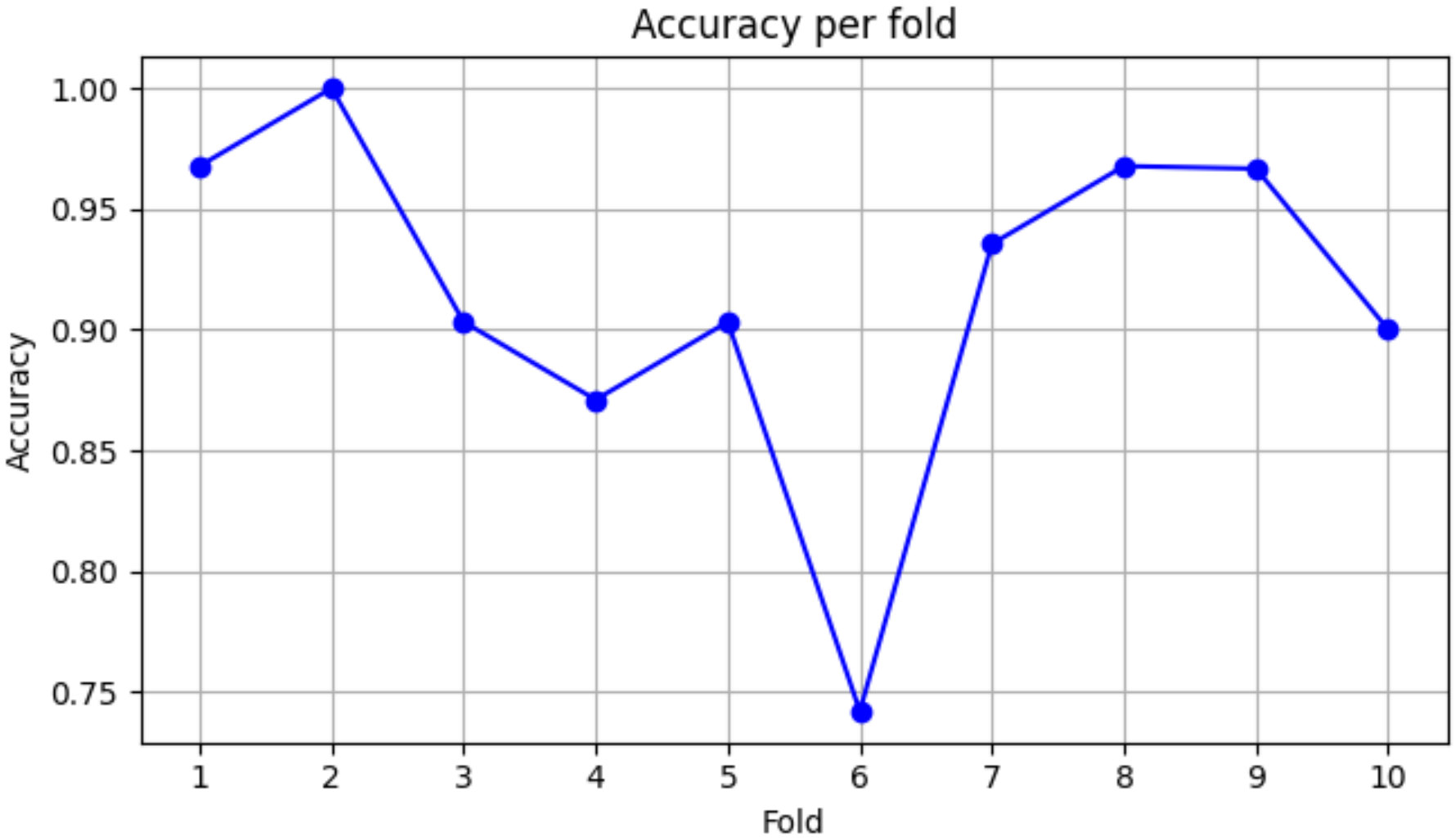
Figure: Tabella delle metriche di valutazione, per-fold

Evaluation: Metriche di Valutazione (2)

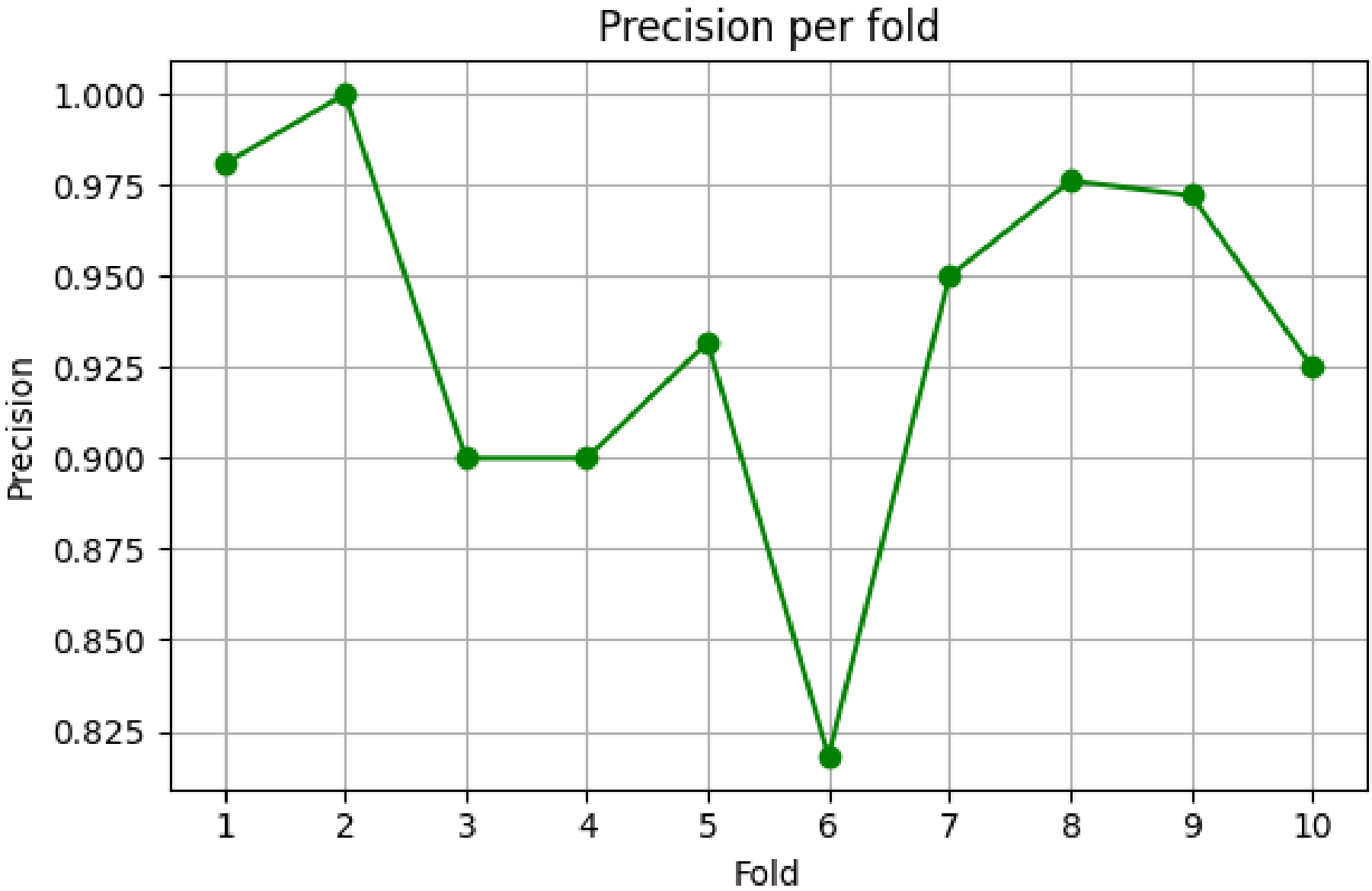
Accuracy	Precision	Recall	F1
0.916	0.935	0.907	0.910

Figure: Tabella delle metriche di valutazione, medie globali

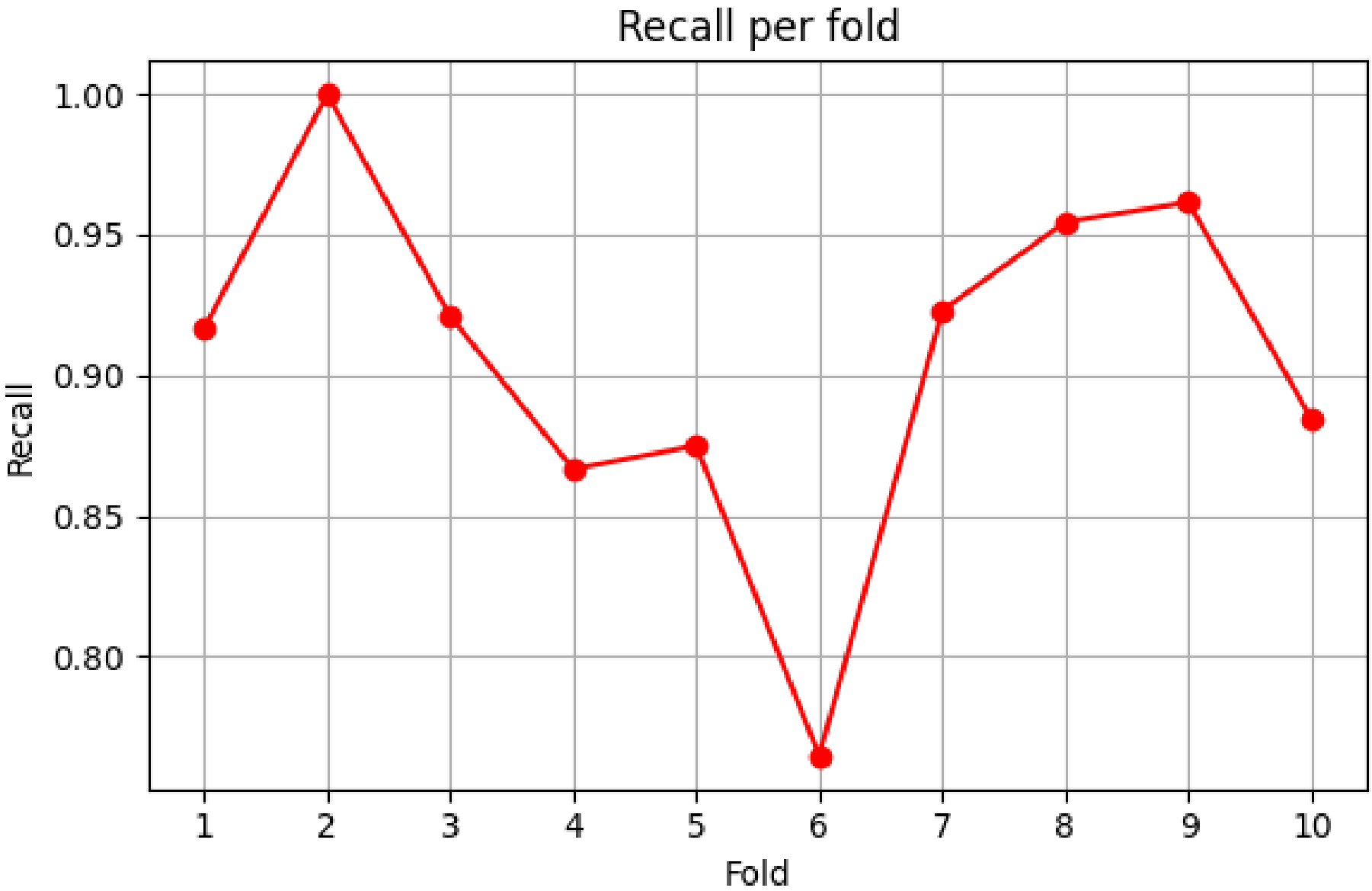
Evaluation: Metriche di Valutazione (3)



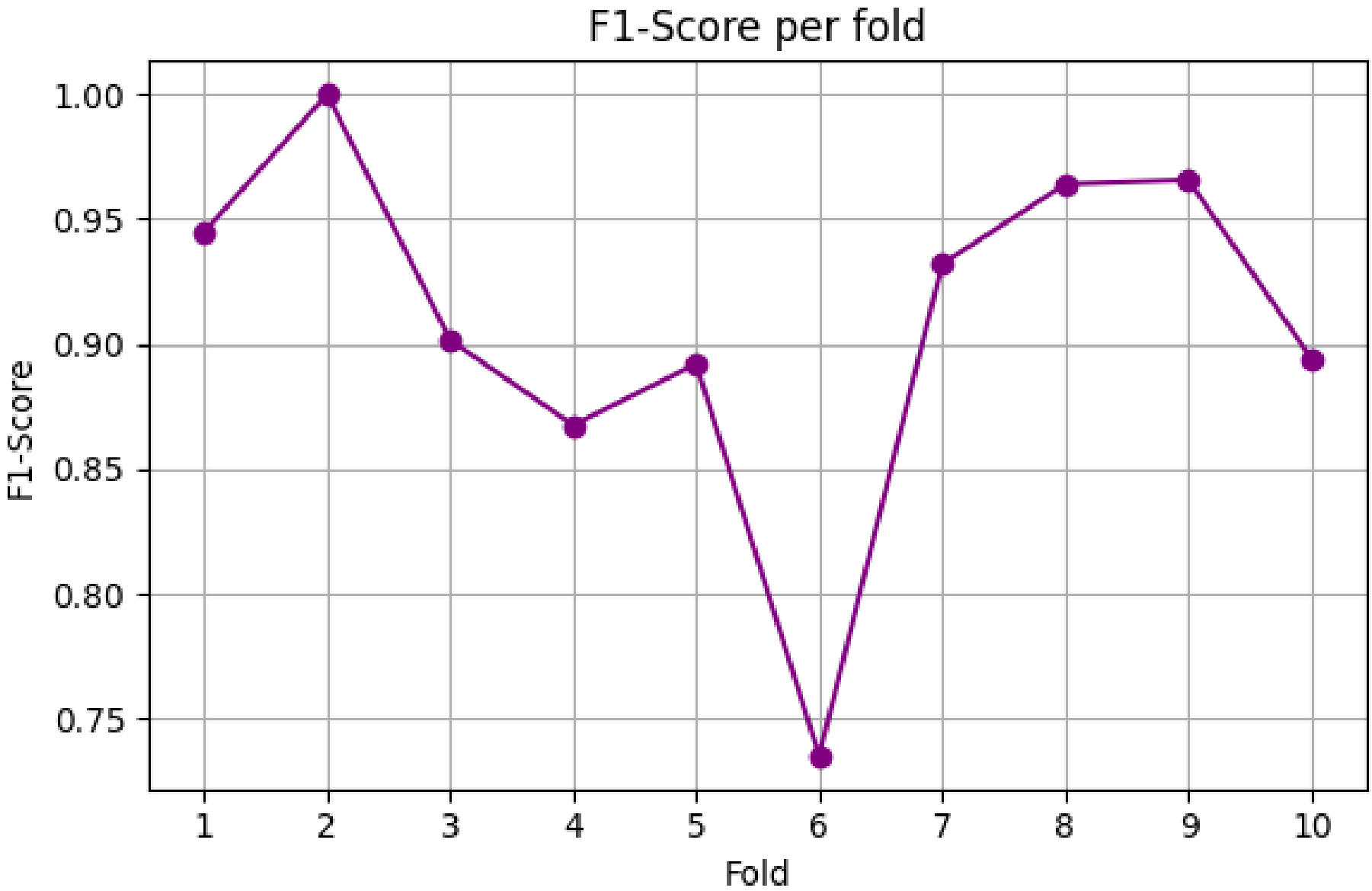
Evaluation: Metriche di Valutazione (4)



Evaluation: Metriche di Valutazione (5)



Evaluation: Metriche di Valutazione (6)



Evaluation: Metriche di Valutazione (7)

Valutazione sulle metriche:
Il traning risulta essere stato ottimo, i valori sono tutti > 0.8, il che significa un ottimo Modello! Addirittura in un fold le metriche assumono tutte valore $= 1$!
...
Giusto?
Non proprio. Vedremo a breve perchè.

Evaluation: Confusion Matrix (1)

Fold	TN	FP	FN	TP
Fold 1	25	0	1	5
Fold 2	18	0	0	13
Fold 3	12	0	3	16
Fold 4	16	0	4	11
Fold 5	19	0	3	9
Fold 6	14	0	8	9
Fold 7	18	0	2	11
Fold 8	20	0	1	10
Fold 9	17	0	1	12
Fold 10	17	0	3	10

Figure: Tabella dei valori della Confusion Matrix, per-fold

Evaluation: Confusion Matrix (2)

TN	FP	FN	TP
176	0	26	106

Figure: Tabella globale e finale della Confusion Matrix

Evaluation: Confusion Matrix (3)

Valutazione sulla Confusion Matrix:
Questo risultato ci potrebbe soddisfare, se non fosse per l'ammontare di **False Negatives (FN)**, che risulta essere rilevante (in proporzione al numero di istanze totali).
Il Modello è dunque soggetto in minima parte ad un problema di erronea classificazione di istanze "legit" come istanze "scam", che potrebbe essere sorvolabile nel caso unico in cui il Modello debba essere usato **solo come tool di supporto**.
Nella fase di Deploy proveremo a capire il perchè di questo problema, e magari dare una possibile soluzione.

Deployment (1)

Utilizzo del Modello:

- Interfaccia terminale per predire il tipo di contenuto.
- Preprocessing degli input (normalizzazione e pulizia).
- Risultato restituito come "scam" o "legit".

Test Realizzati:

- Istanza legit senza link;
- Istanza legit con link;
- Istanza scam senza link;
- Istanza scam con link.

Sono state usate istanze realistiche per questo Test.

Deployment (2)

```
Inserisci il titolo:
How to train your personal AI
Inserisci la descrizione:
In only 8 minutes, you will become the AI master!
Risultato della classificazione: LEGIT
```

Figure: Istanza legit SENZA link = **Corretto**

Deployment (3)

```
Inserisci il titolo:
Top 10 coding languages to learn
Inserisci la descrizione:
Today we are going to cover the top coding languages to learn for yourself. https:// www.instagram.com/big_guy_data
Risultato della classificazione: SCAM
```

Figure: Istanza legit CON link = **Sbagliato**

Deployment (4)

```
Inserisci il titolo:
Watch this video to get money!
Inserisci la descrizione:
If you watch this video you will become rich!
Risultato della classificazione: SCAM
```

Figure: Istanza scam SENZA link = **Corretto**

Deployment (5)

```
Inserisci il titolo:
Learn how to win Jackpots easy and fast!
Inserisci la descrizione:
Follow these easy steps to learn how to win everytime! https://www.growbig.com/
Risultato della classificazione: SCAM
```

Figure: Istanza scam CON link = **Corretto**

Evaluation post-deployment

Problematica:
L'alto caso, visto in precedenza, di False Negatives porta alla erronea classificazione delle istanze legittime, molto probabilmente per un bias del modello, che va a classificare le istanze contenenti "link" più come scam che come legit.
Da dunque molta più importanza di quanto dovrebbe ad essa.

Possibile soluzione:
Una soluzione potrebbe essere proprio quella di assegnare un peso alle istanze, ora come ora equilibrate automaticamente dall'algoritmo, diverso dalla feature "link", e eseguire nuovamente la fase di training.
Per mancanza di tempo questa rimane una ipotesi, implementata nella repository solo in parte.

Conclusioni

Pensieri finali:
Il progetto è stato assai divertente da realizzare, anche fuori dalla seduta di esame continuerò a lavorare sopra ad un modello con scopo simile a quello esposto, nella speranza di portarlo a realizzazione con conoscenze anche avanzate rispetto al corso di Fondamenti.