



Loan Approval Classification Model

Corso: Data &
AI – Develhope

- Alessandro Caponigro
- Emanuela Torre
- Francesca Cola

“Un modello predittivo per stimare la corretta approvazione dei prestiti”

Indice:

- Fase EDA & Pre-processing
- Modelli
- Confronto Tra Modelli
- Prossimi passi & Conclusioni

Exploratory Data Analysis

Fonte: Kaggle - loan_data.csv

Osservazioni: 45.000 righe & 14 colonne

Tipi di variabili

- **Categorical:**

- loan_intent,
 - person_home_ownership,
 - person_gender,
 - person_education

- **Numerical:**

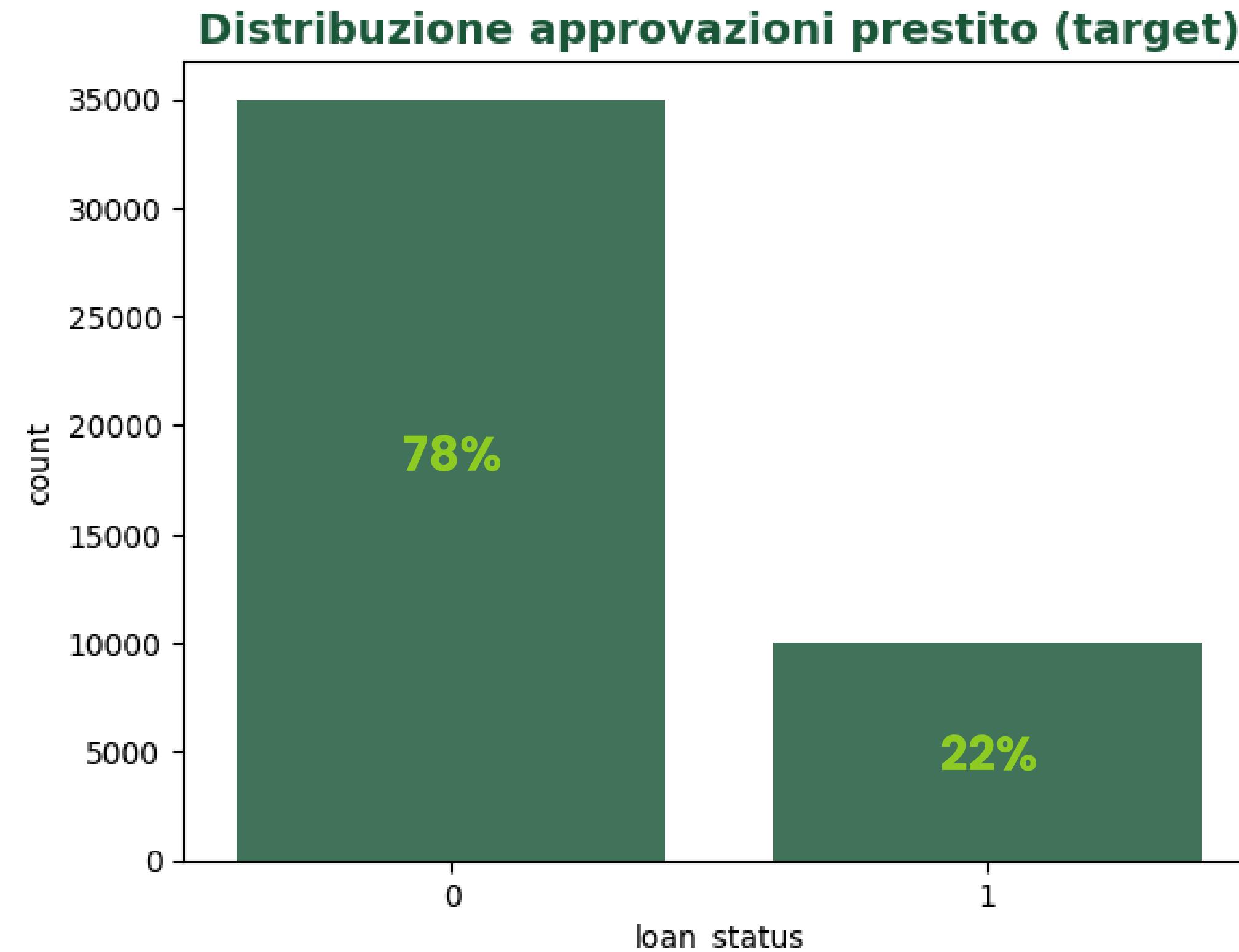
- loan_amnt,
 - loan_int_rate,
 - loan_percent_income,
 - person_age,
 - person_income,
 - person_emp_exp,
 - credit_score

- **Binary:**

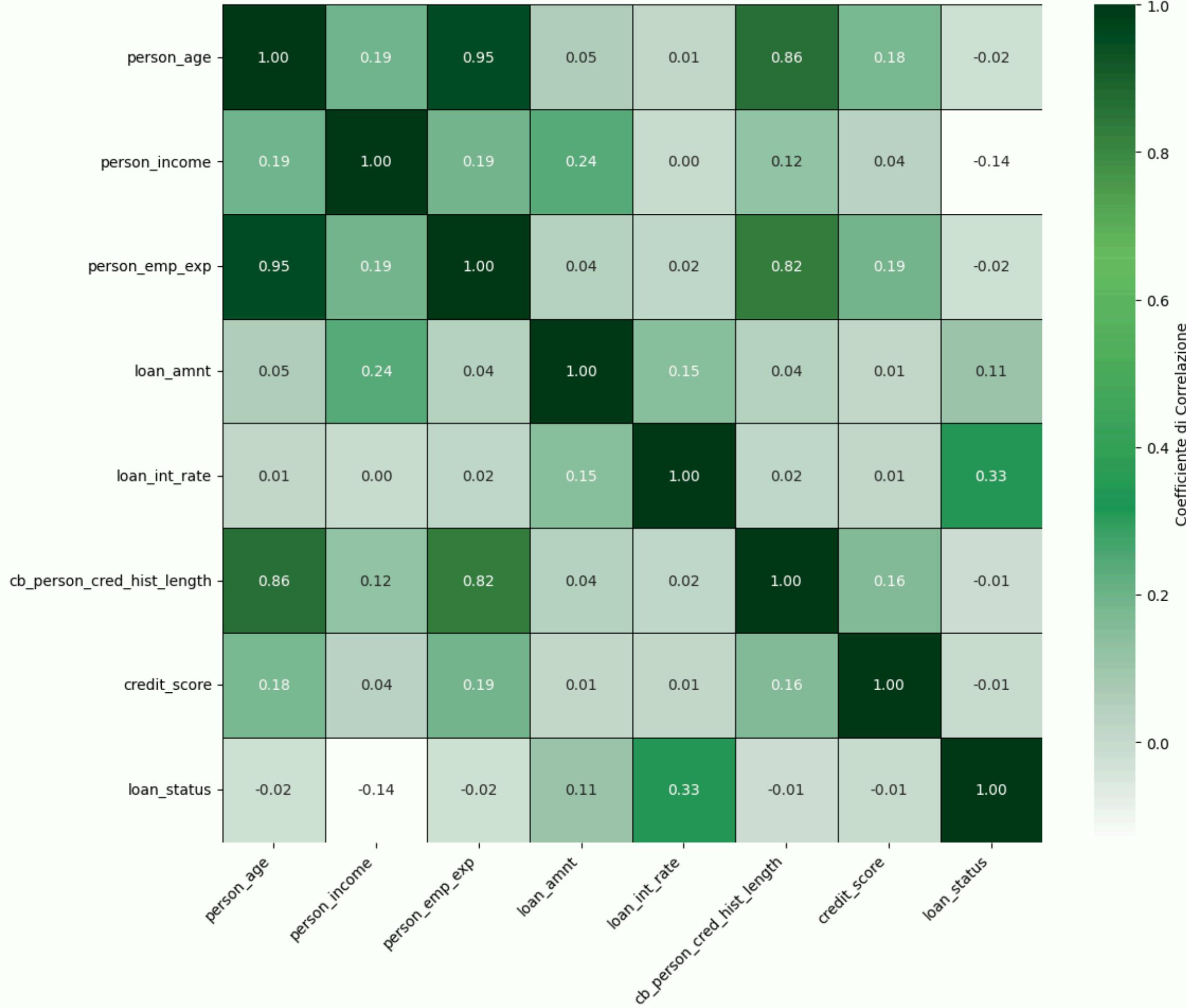
- loan_status,
 - previous_loan_defaults_on_file

Variabile Target (Y)

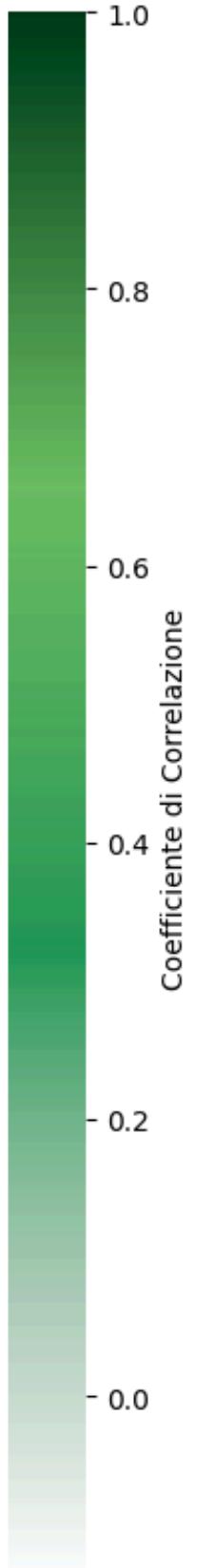
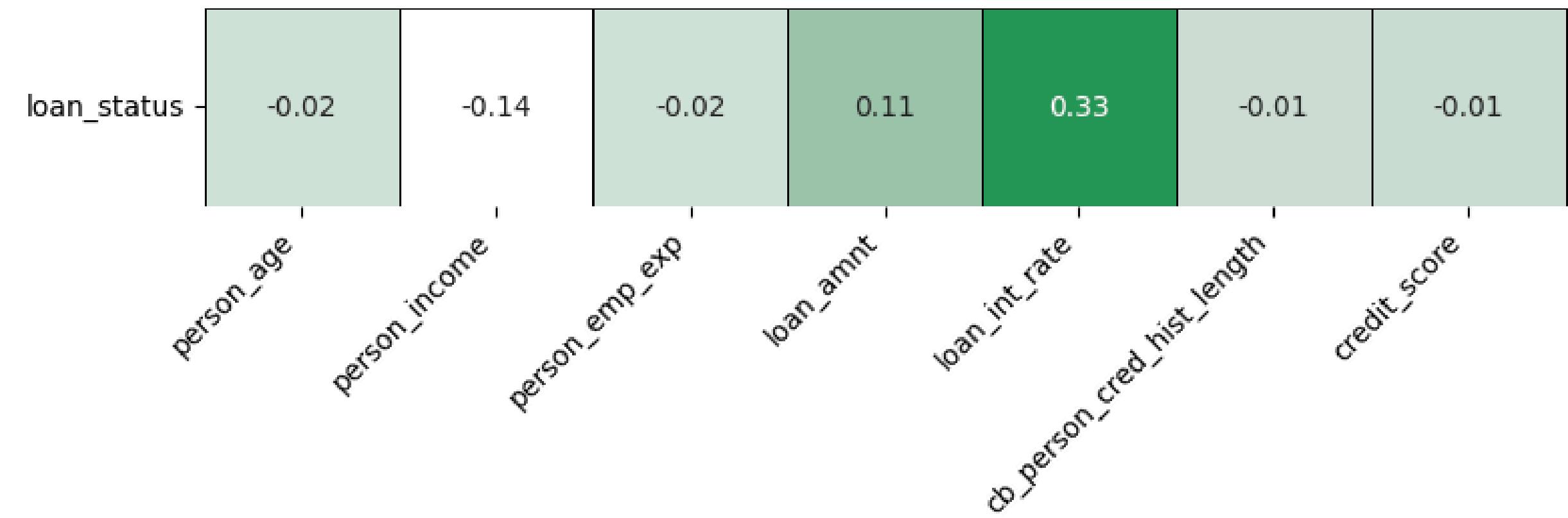
Target da sistemare



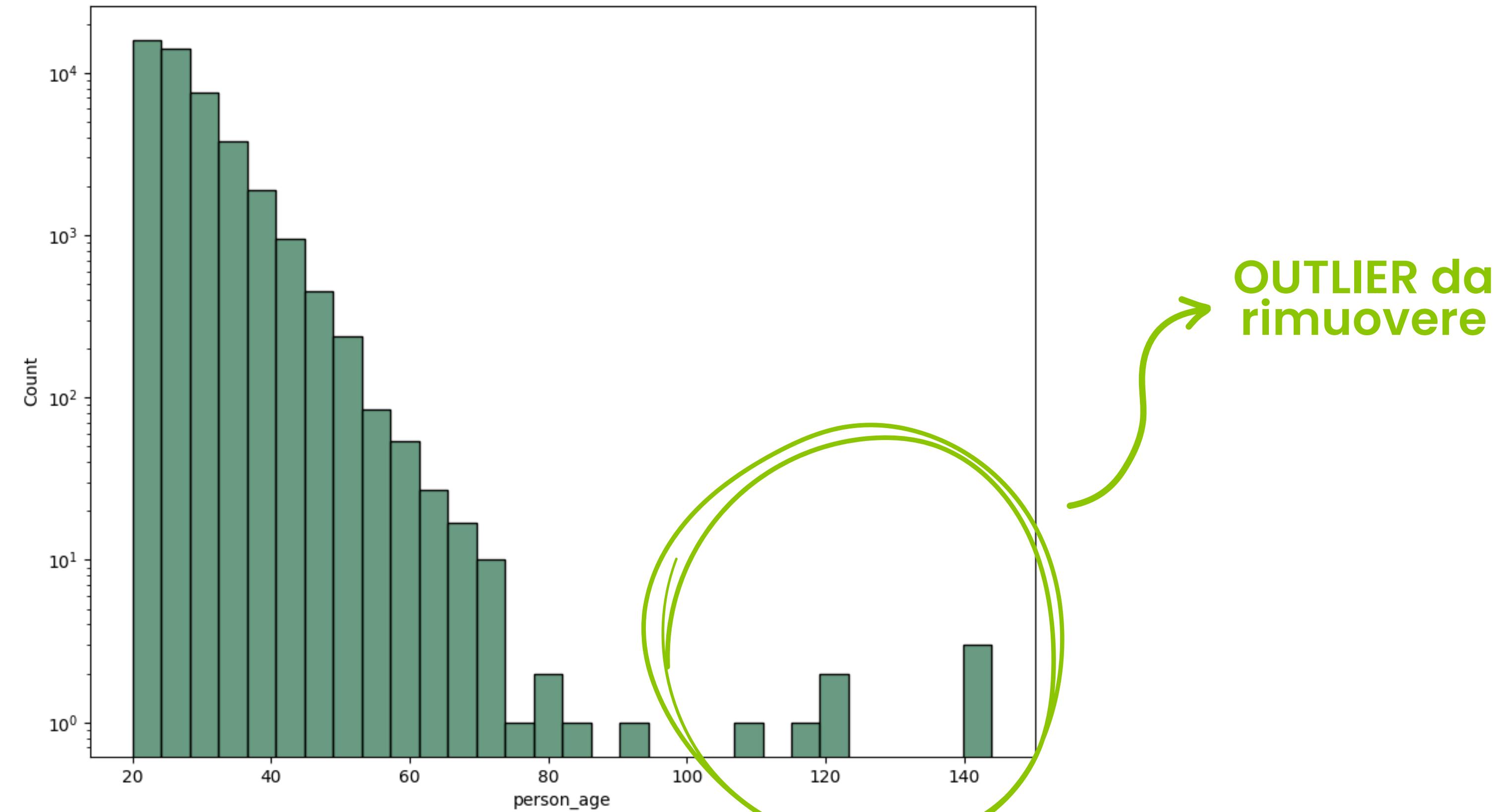
Matrice di Correlazione delle Variabili Numeriche



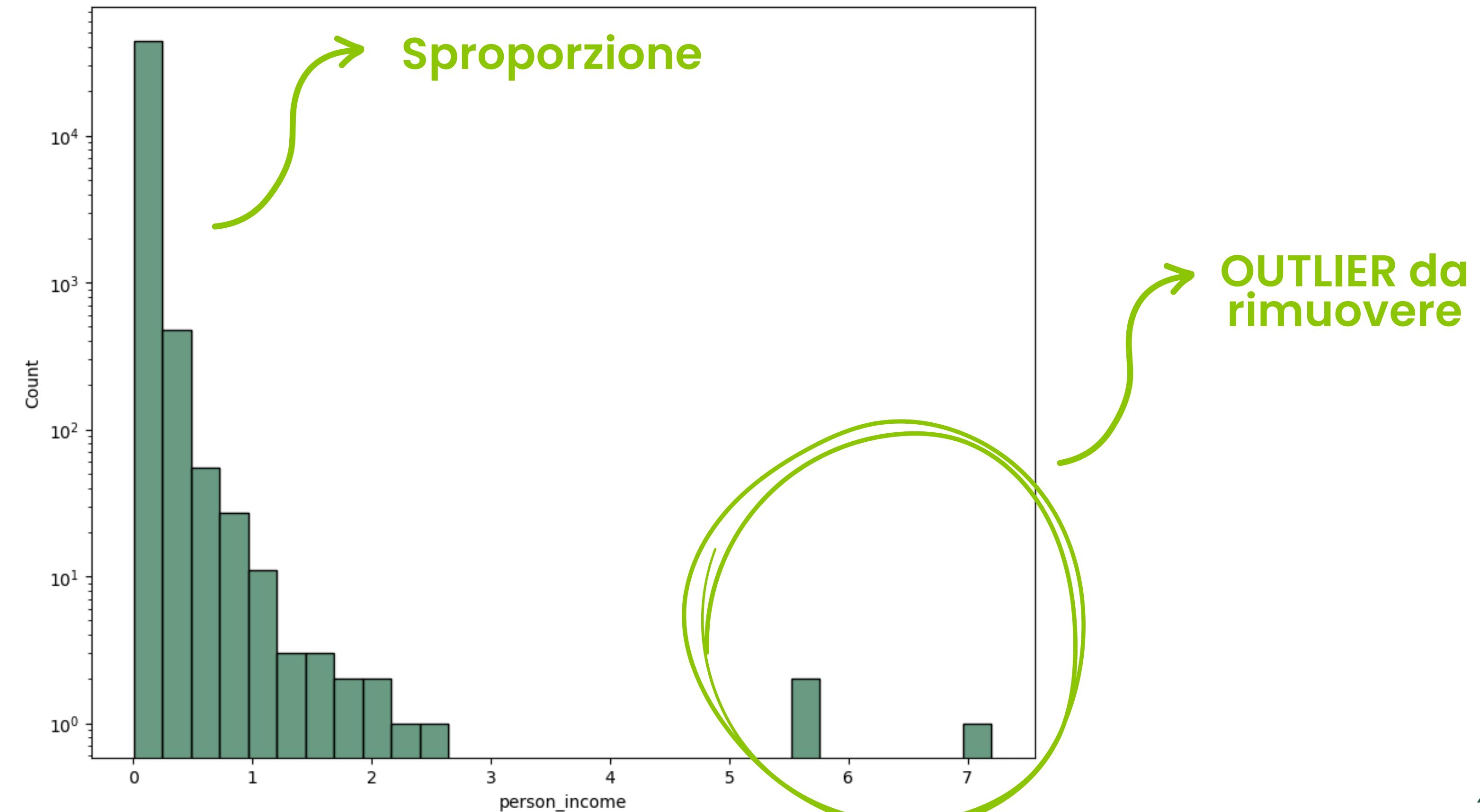
Correlazioni numeriche con la variabile target



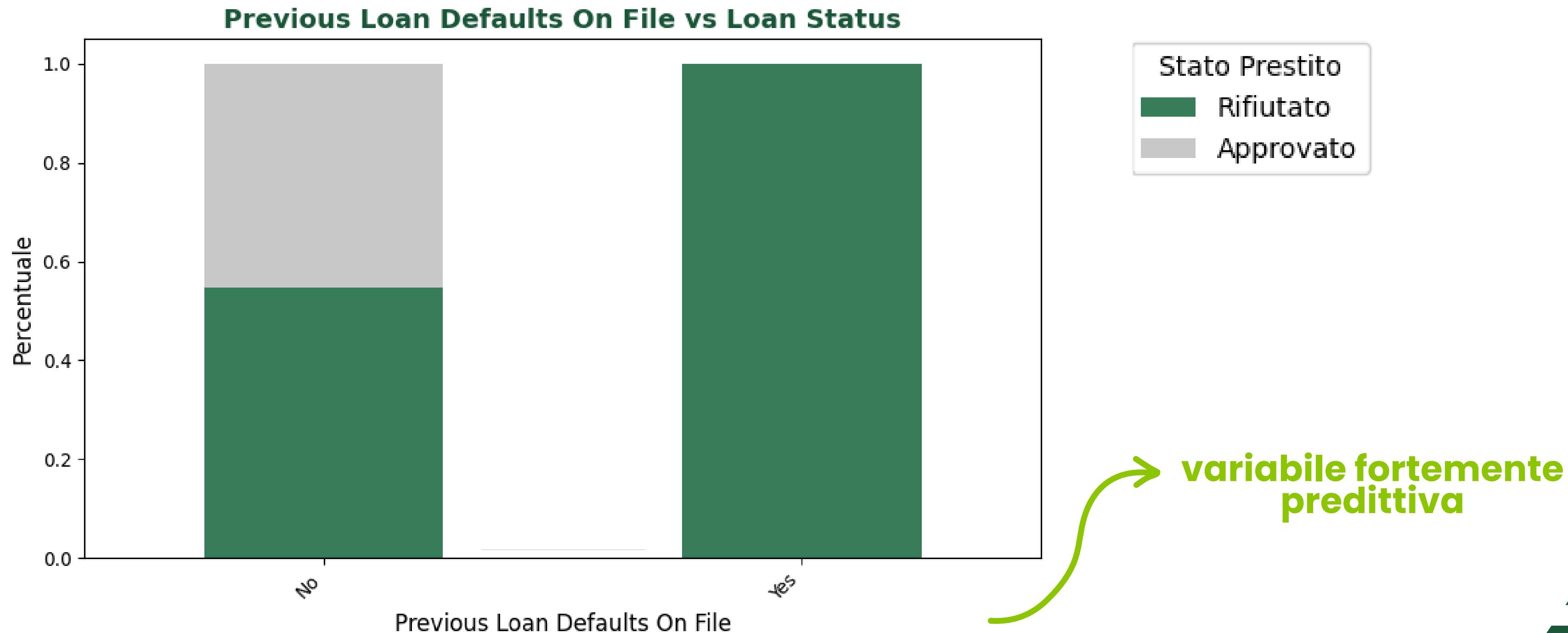
Distribuzione delle età



Distribuzione del reddito



Analisi incrociata tra variabili binarie e target in percentuale





Data Cleaning & Pre-processing



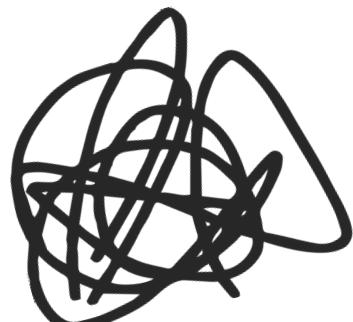
“Un buon modello parte sempre da dati puliti”

~~person_gender~~

~~person_income > 5mln~~

~~person_home_ownership:
OTHER~~

~~person_omp_exp~~



Encoding Delle Variabili



I Modelli

Decision Tree

- E' un modello supervisionato che crea un albero di decisioni basato su regole if–then.
- Suddivide i dati in modo gerarchico attraverso domande binarie su feature.



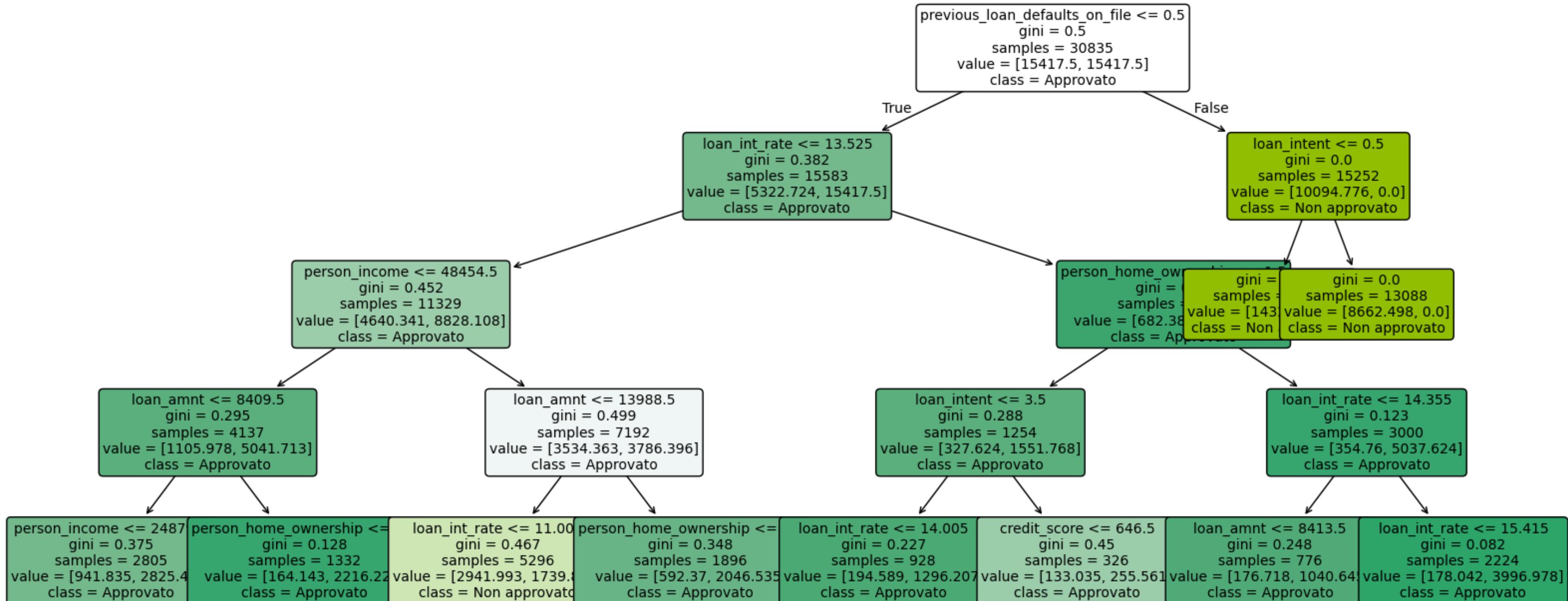
Decision Tree (No-Pruning)

Viene lasciato crescere completamente

Decision Tree (Post-Pruning)

Viene costruito completamente e poi
semplificato dopo l'allenamento

Decision Tree (No-Pruning)

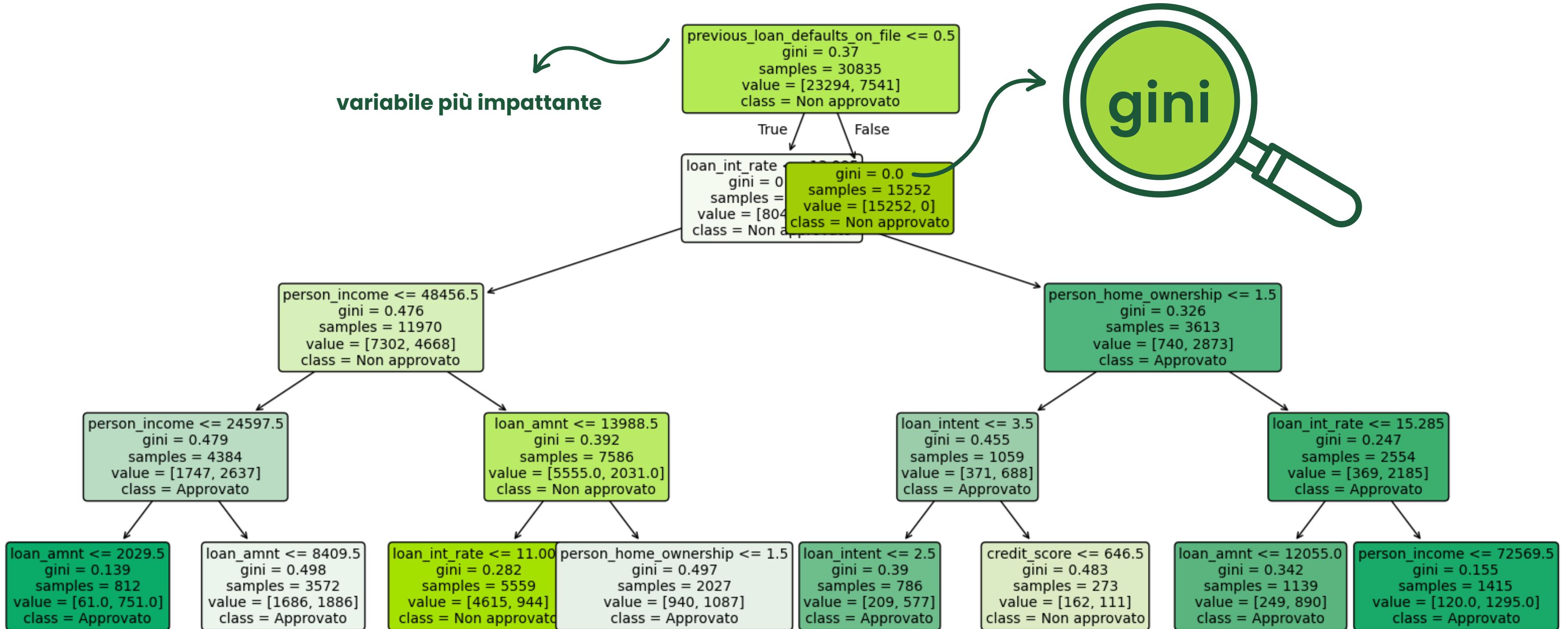


= il nodo tende alla classe Approvato (0)

= il nodo tende alla classe Non approvato (1)

Più scuro è il colore, più il nodo è "sicuro" della sua decisione.

Decision Tree (Post-Pruning)



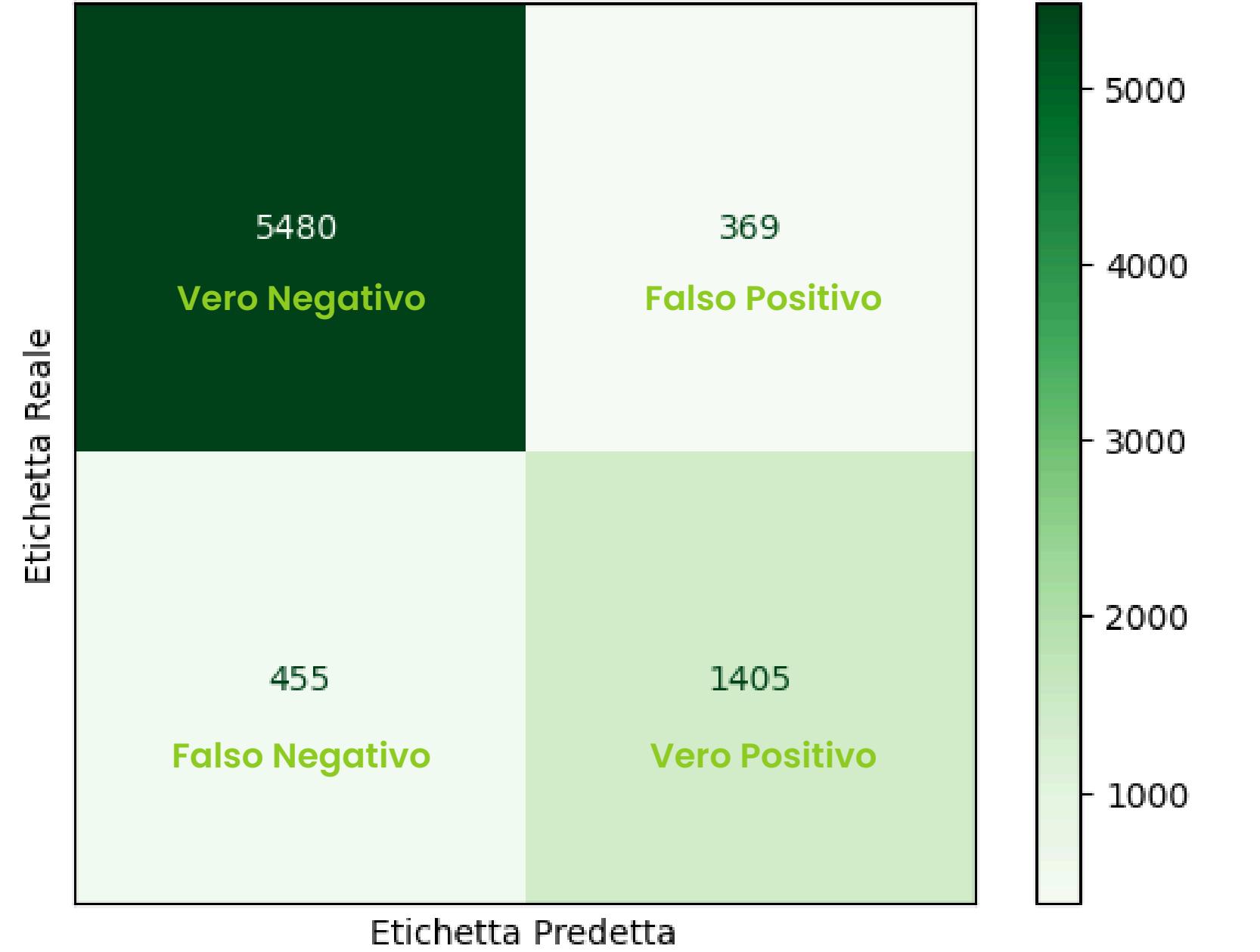
= il nodo tende alla classe Approvato (0)

= il nodo tende alla classe Non approvato (1)

Più scuro è il colore, più il nodo è "sicuro" della sua decisione.

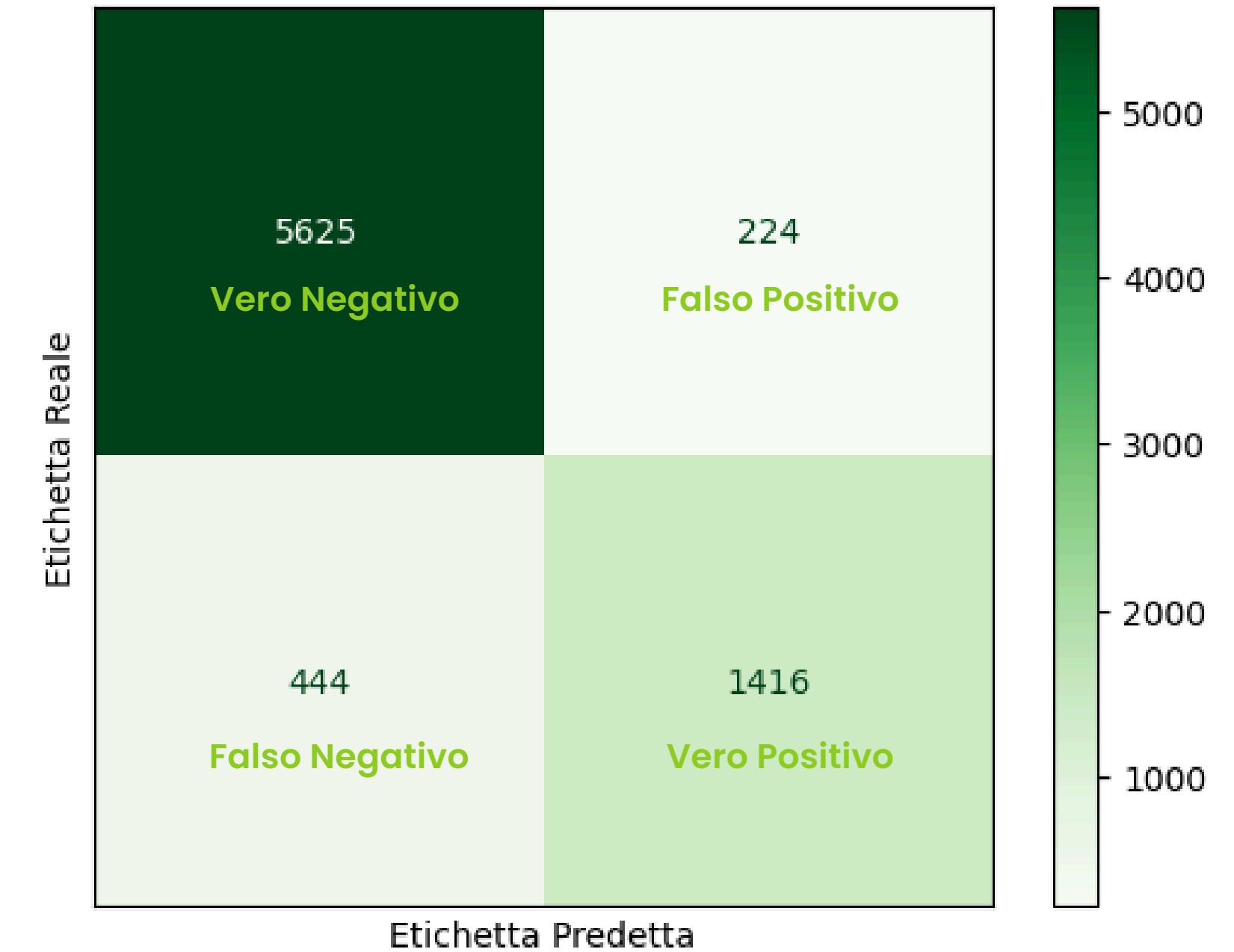
Matrice di confusione

No-Pruning



Accuracy del modello: 0.89

Post-pruning



Accuracy del modello: 0.91

Confronto e conclusioni:

Modello	Precision (1)	Recall (1)	F1-score (1)	Accuracy
No-Pruning	76	79	78	89
Post-Pruning	85	76	80	91

Post-Pruning:

- 1.Superiorità (AUC)
- 2.Migliore capacità discriminatoria
- 3.Semplicità
- 4.Generalizzabilità



Regressione Logistica

Modello lineare utilizzato per la classificazione binaria.
Stima la probabilità di una classe tramite la funzione logistica (sigmoide)

Equilibrio fra le feature:

Gestione dello sbilanciamento

- Abbiamo equilibrato il dataset con l'uso di SMOTE per generare nuovi esempi sintetici della classe minoritaria (solo su training set)

Rimozione colonne ridondanti e check multicollinearità

- Backward Selection
- Filtraggio VIF >10
- Ricalcolo VIF su tutte le feature rimanenti, per assicurarci che fossero "indipendenti".

Split del dataset per allenamento

- Training set → per addestrare il modello
- Test set → per valutare il modello su dati mai visti prima
- SMOTE solo sul training set





Regressione Logistica

Inferenza

Quali variabili influenzano davvero il prestito ?

Punti chiave:

- Selezione delle variabili significative
- Focus su causa-effetto

Esempio osservato:

Maggiore è il tasso d'interesse

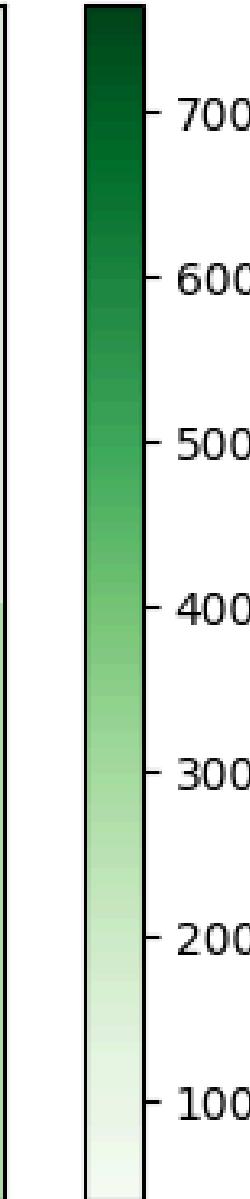
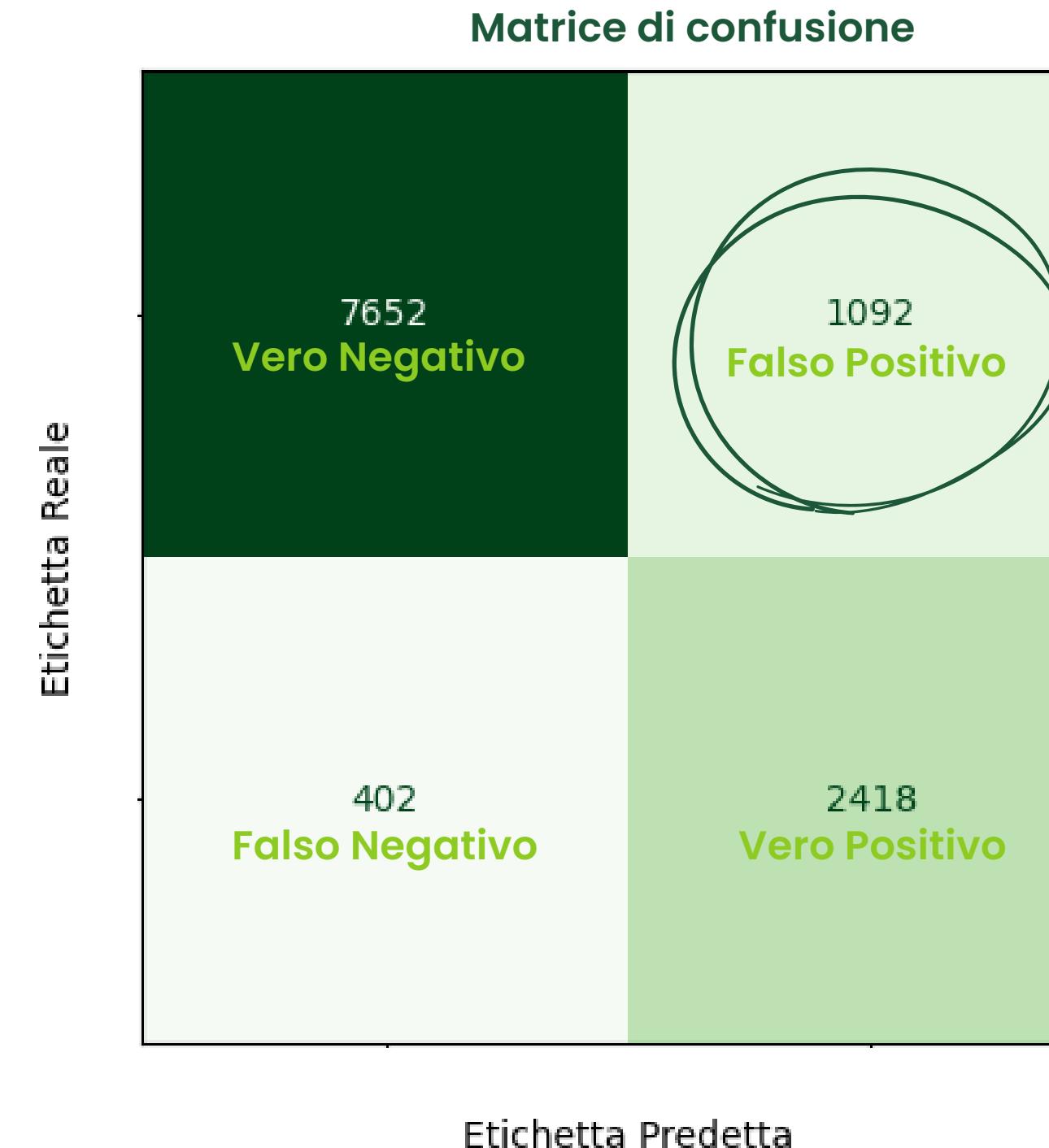
Maggiore è la probabilità di concessione

Perchè?

Evidentemente la banca, di fronte a clienti più rischiosi, approva imponendo condizioni più severe come un tasso più alto

Regessione Logistica

Dopo La Predizione



Troppi falsi positivi

Rimosse progressivamente le feature non significative, migliorando interpretabilità e stabilità del modello.

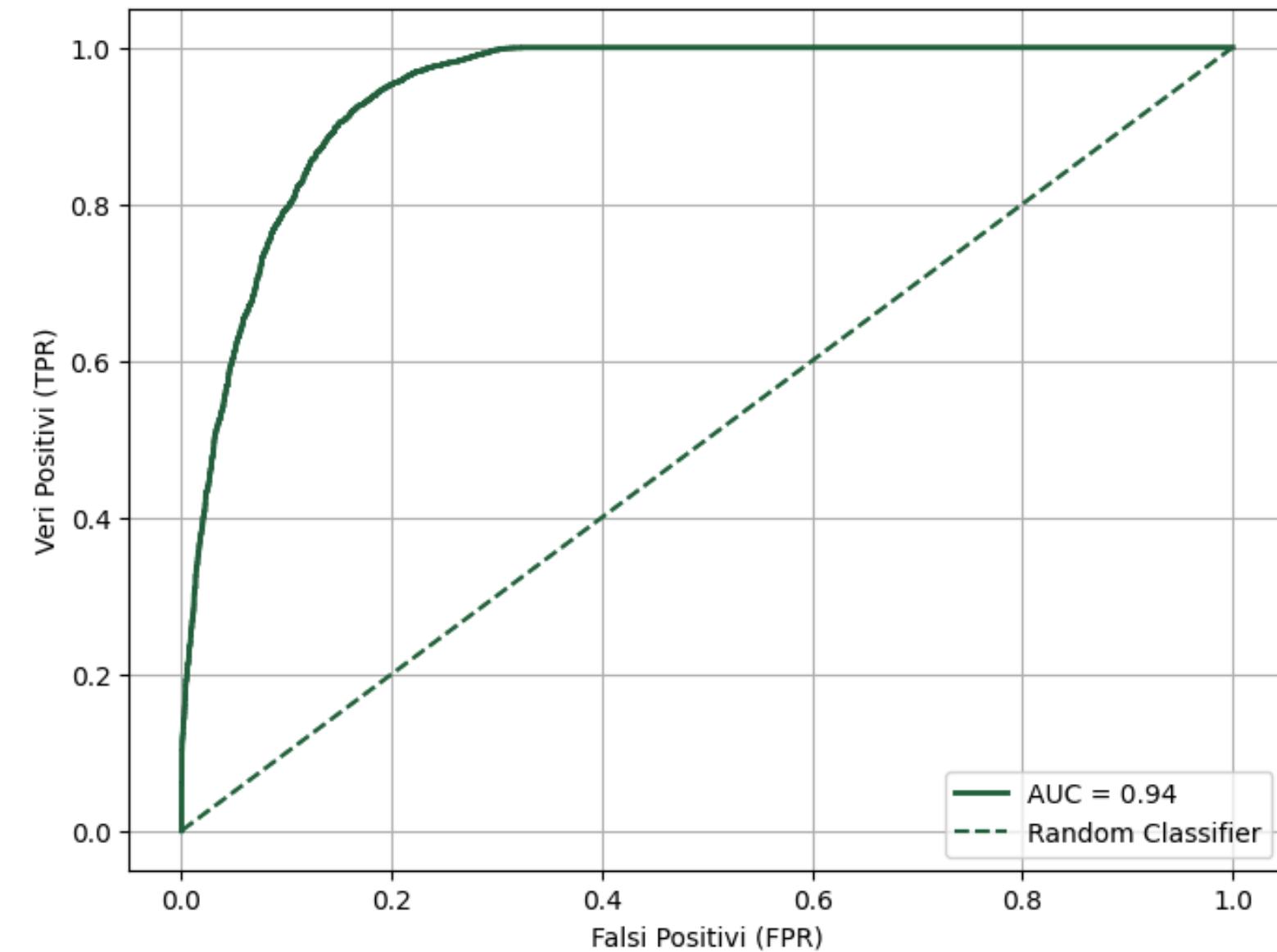
Accuracy del modello: 0.88

Metriche e aggiustamenti

- Bilanciamento tra performance e semplicità
- Gestione efficace dello sbilanciamento con SMOTE

Classe	Precision	Recall	F1-score	Accuracy
Concesso (1)	70	85	77	88

Inclusività



Random Forest

In cosa consiste:

Insieme di **tanti alberi decisionali che votano per decidere se approvare un prestito**

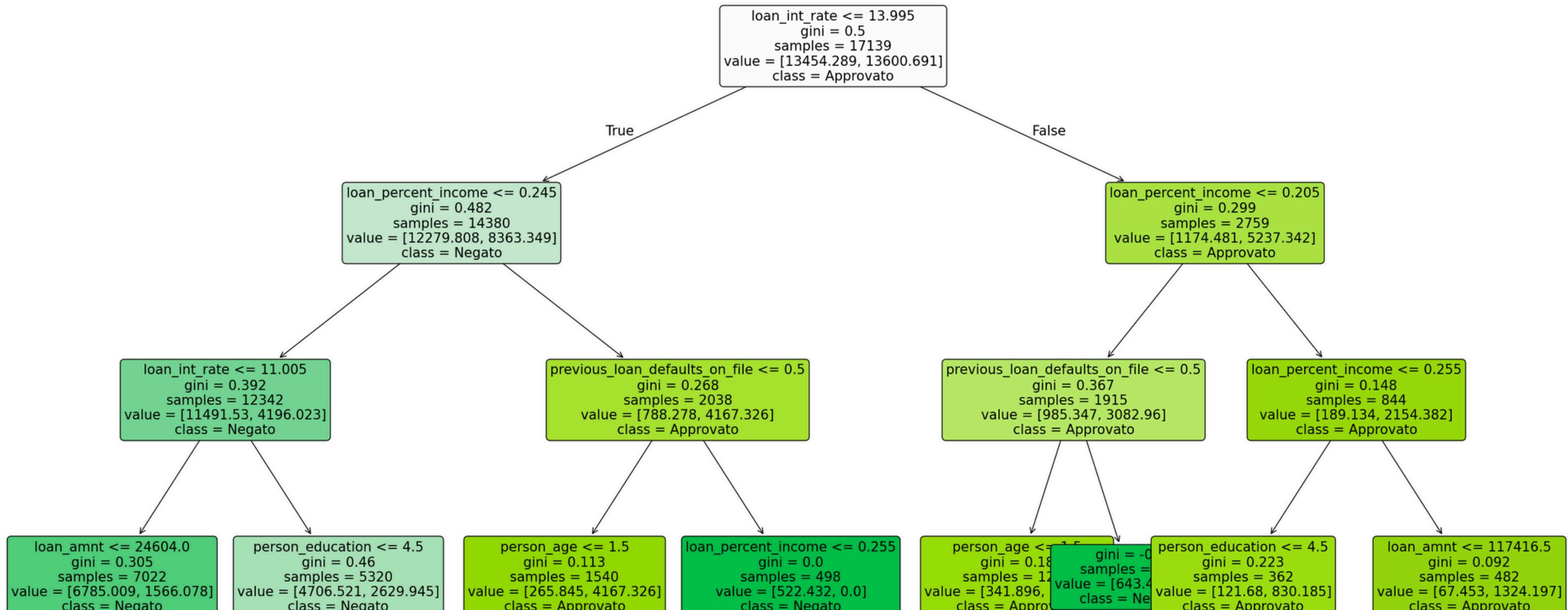
- **Ogni albero vede solo una parte dei dati**
- **Prende decisioni diverse**
- **Unisce i risultati**

Quindi nel pratico:

- **Ogni nodo dell'albero pone una domanda e divide i dati in base alla risposta**
- **Usiamo l'indice numerico Gini per capire quanto la risposta è chiara**

Random Forest

Primo Albero



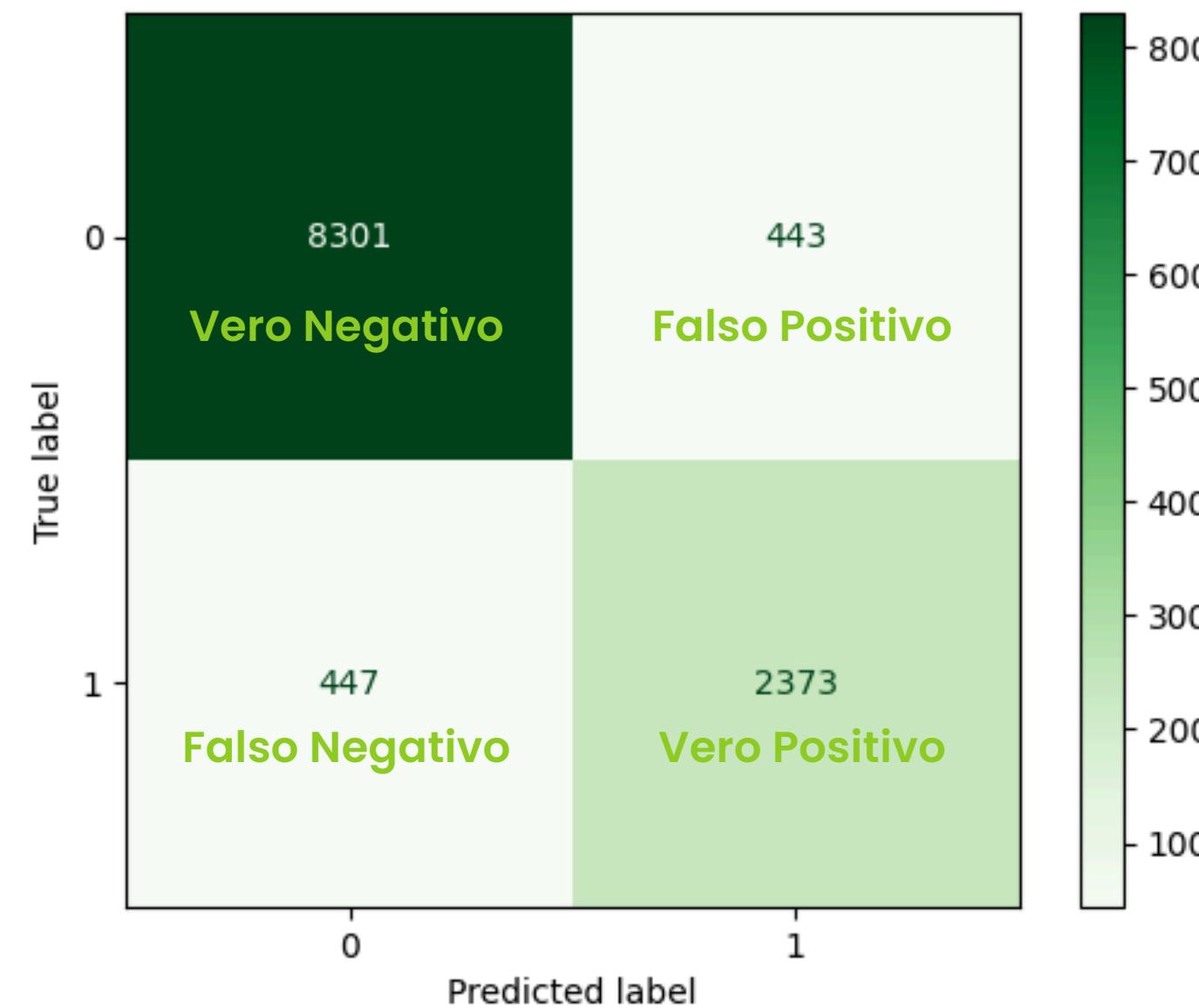
= il nodo tende alla classe Negato (0)
= il nodo tende alla classe Approvato (1)

Più scuro è il colore, più il nodo è "sicuro" della sua decisione.

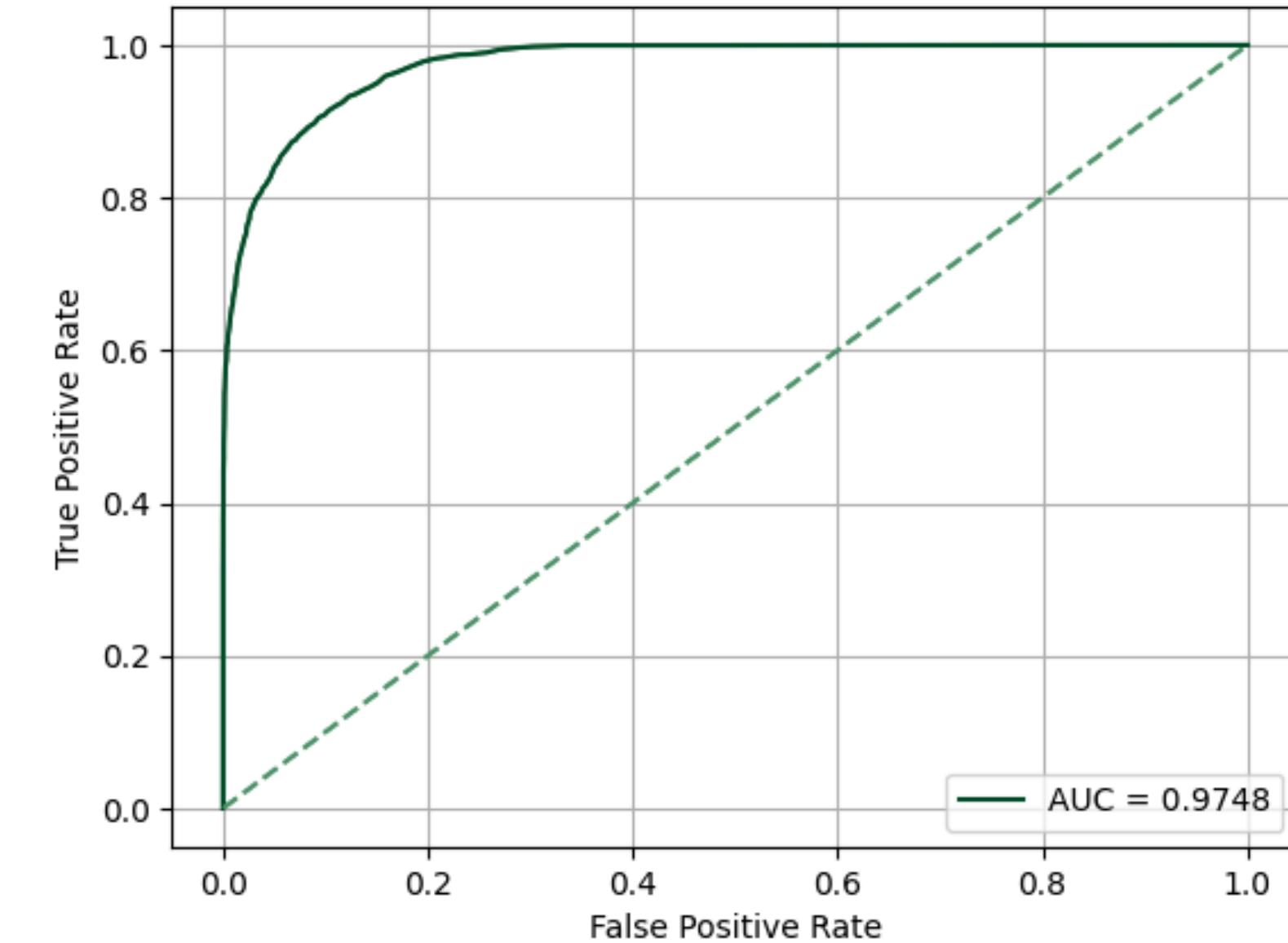
Valutazione Modello

Dopo esseresi allenato e aver bilanciato le classi

Matrice di confusione Random Forest



Curva ROC Random Forest

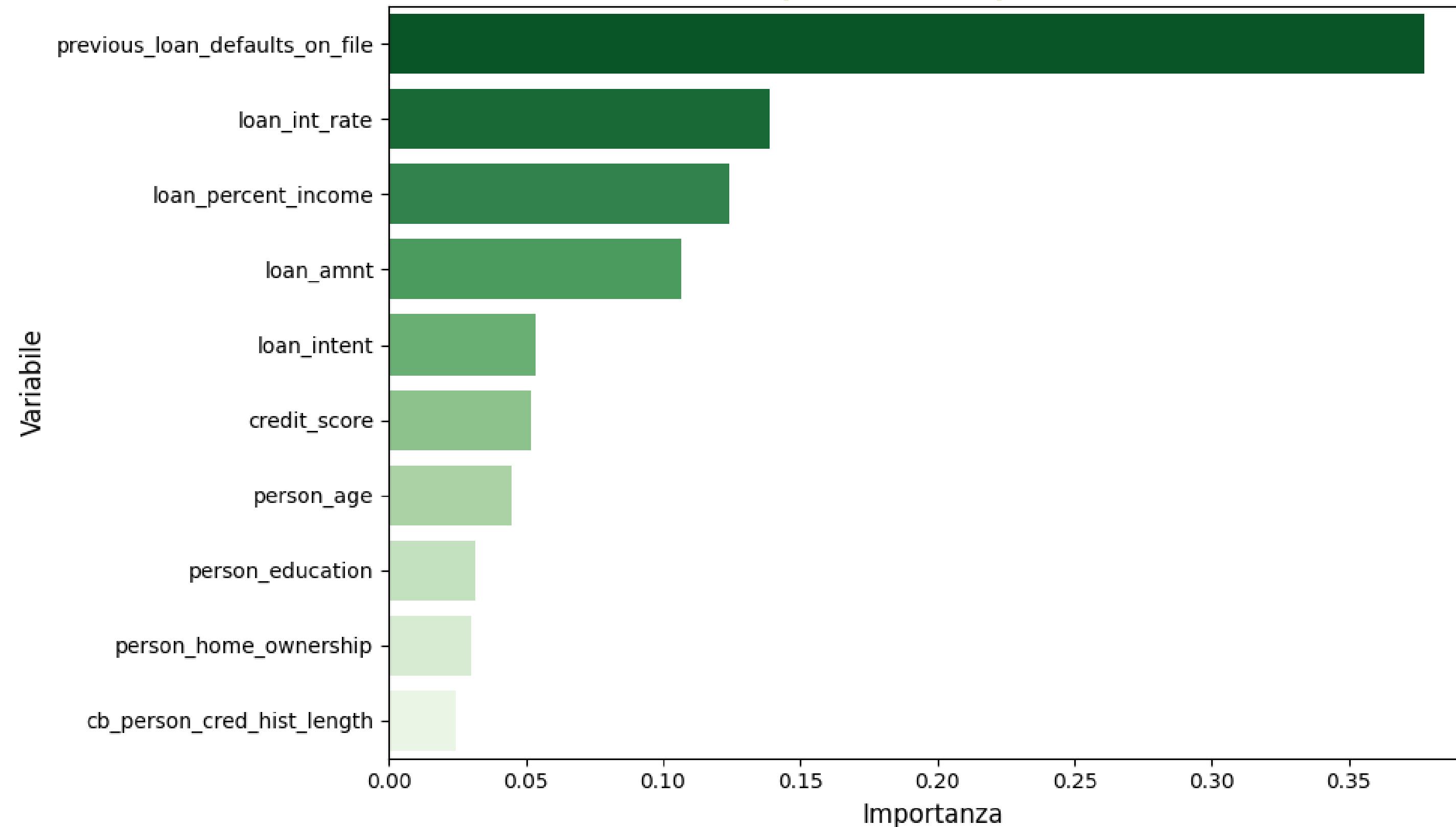


- Diagonale dei **VERI PREDETTI** (positivi e negativi) con somma molto alta.
- Falsi Positivi minori dei Falsi Negativi
Meglio in questo caso : I falsi Negativi sono maggiori, preservando soldi e non creando rischi di perdita inutili

Classe	Precision	Recall	F1-score	Accuracy
Concesso (1)	84	84	84	92

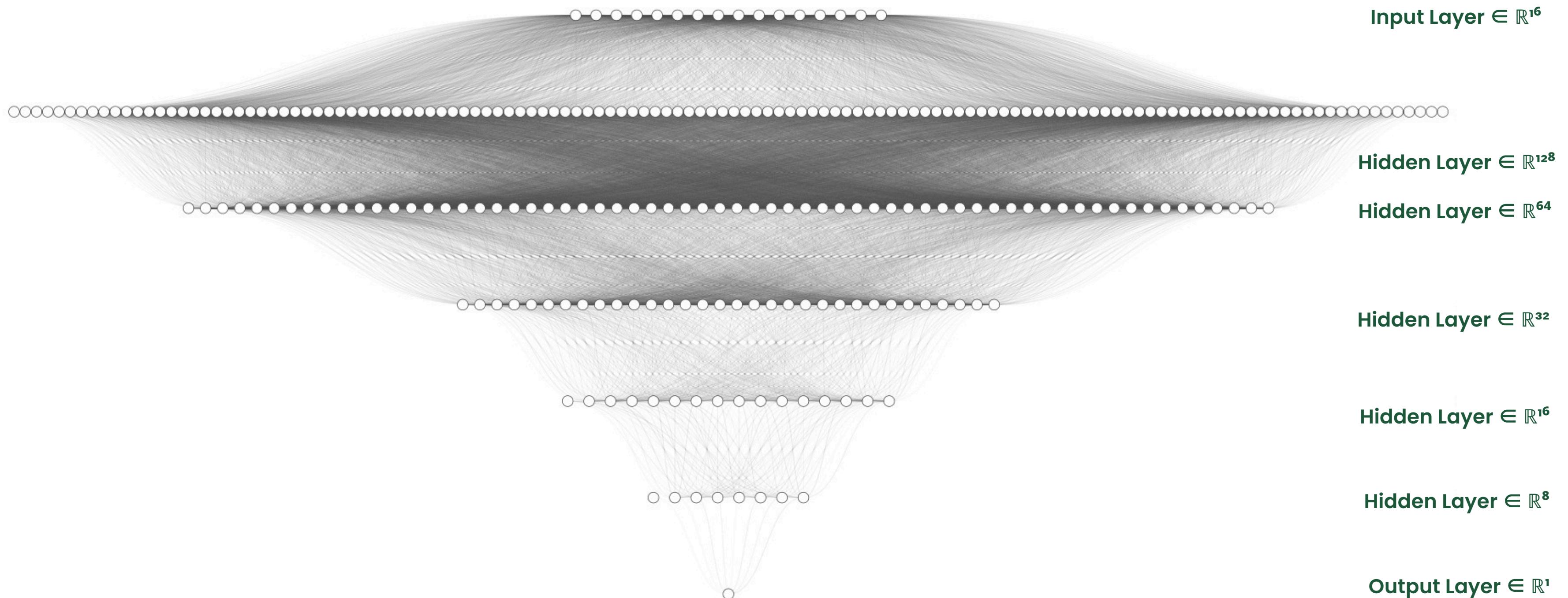
Random Forest

Top 10 Feature Importance



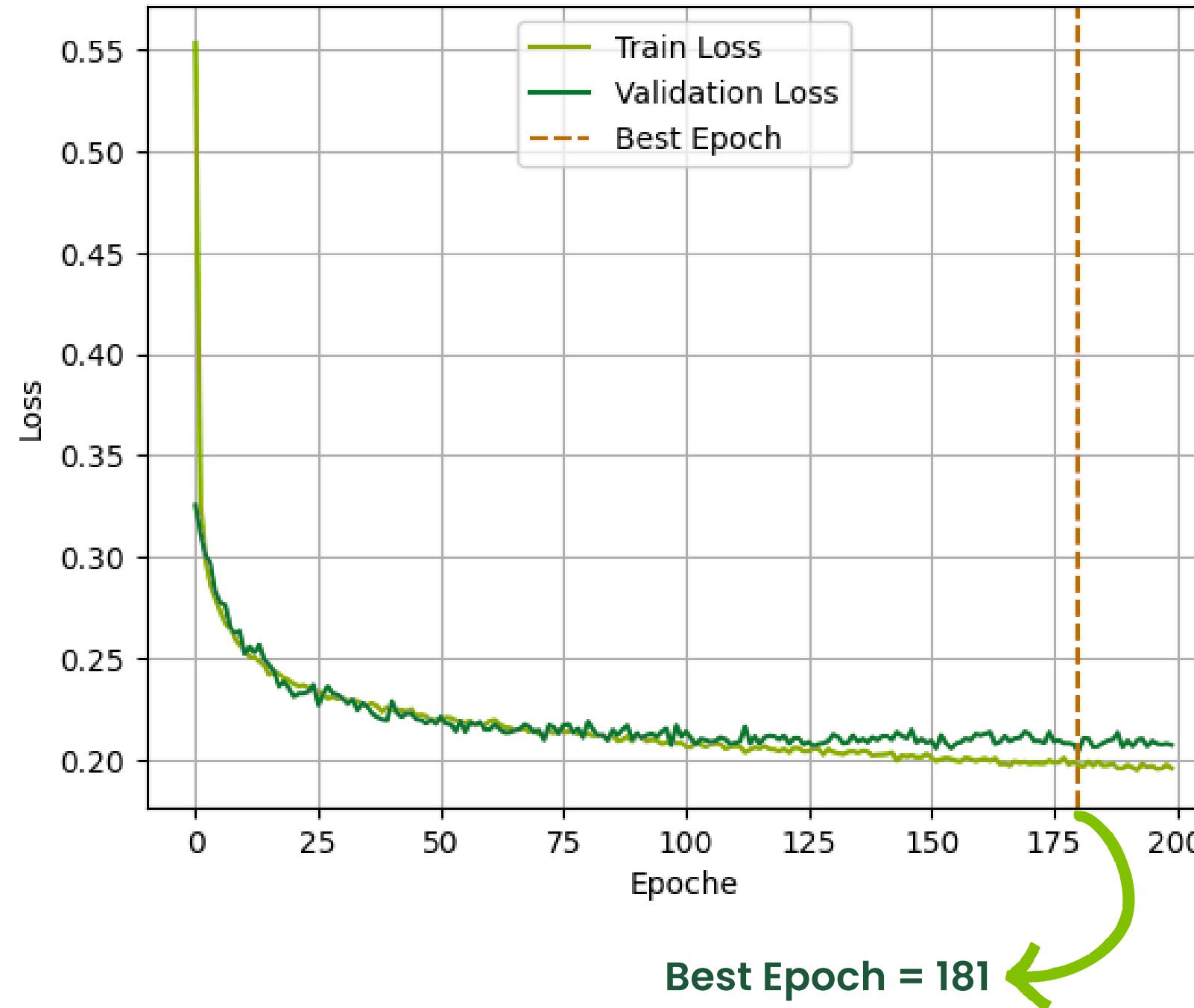
Deep Neural Network

“Quando i dati parlano, la rete neurale ascolta, apprende e decide.”



Funzione Di Costo

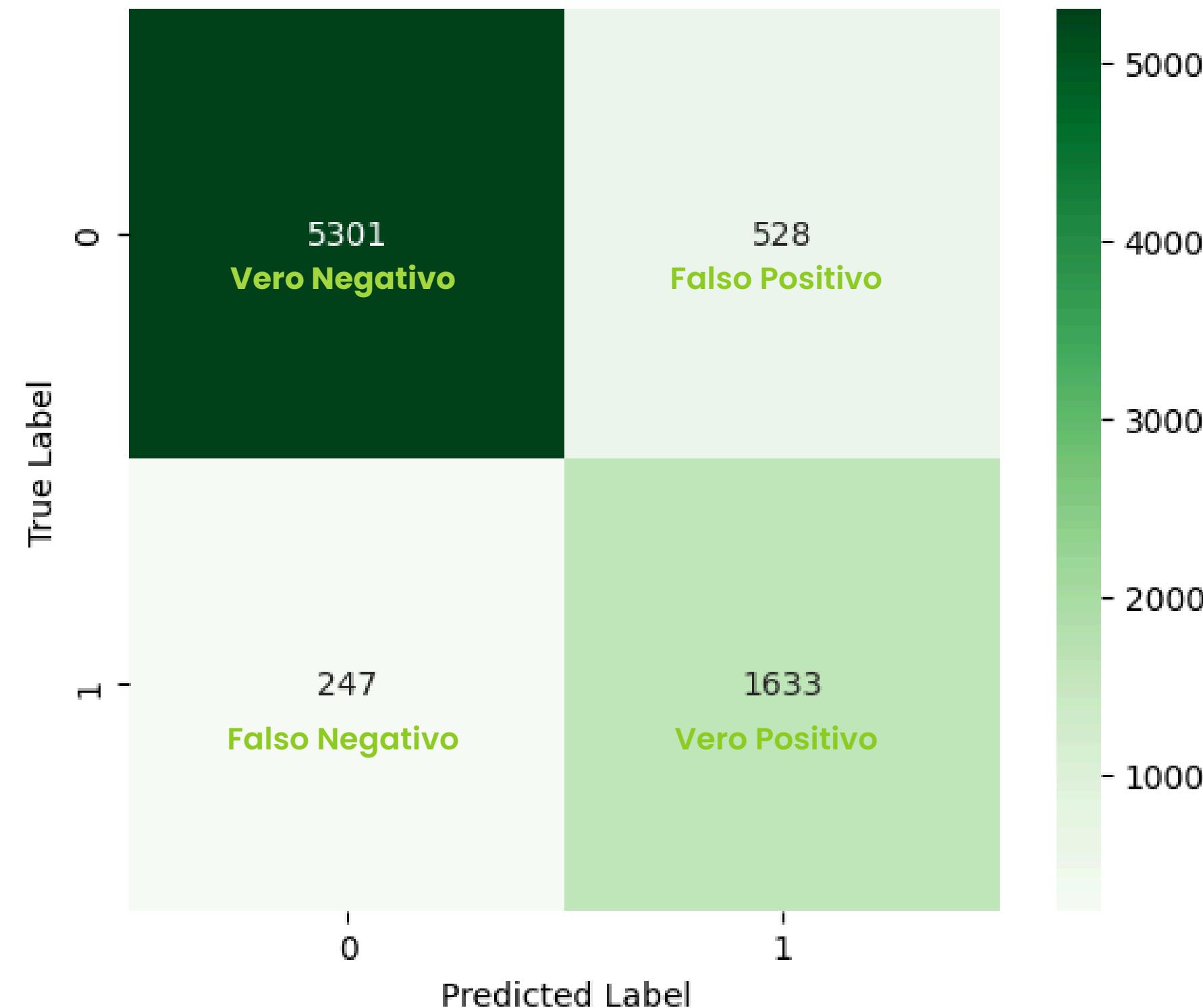
Andamento della Loss durante l'addestramento



Loss = 0.20

**La Bynary Cross-Entropy misura
l'errore nella stima delle probabilità:
più è bassa(vicina allo 0),
più il modello è "sicuro e corretto"
nelle sue previsioni.**

Matrice di Confusione



Il modello classifica correttamente la maggior parte dei casi e non favorisce una classe rispetto all'altra.

Classification Report

Classe	Precision	Recall	F1-score	Accuracy
Concesso (1)	76	87	81	90

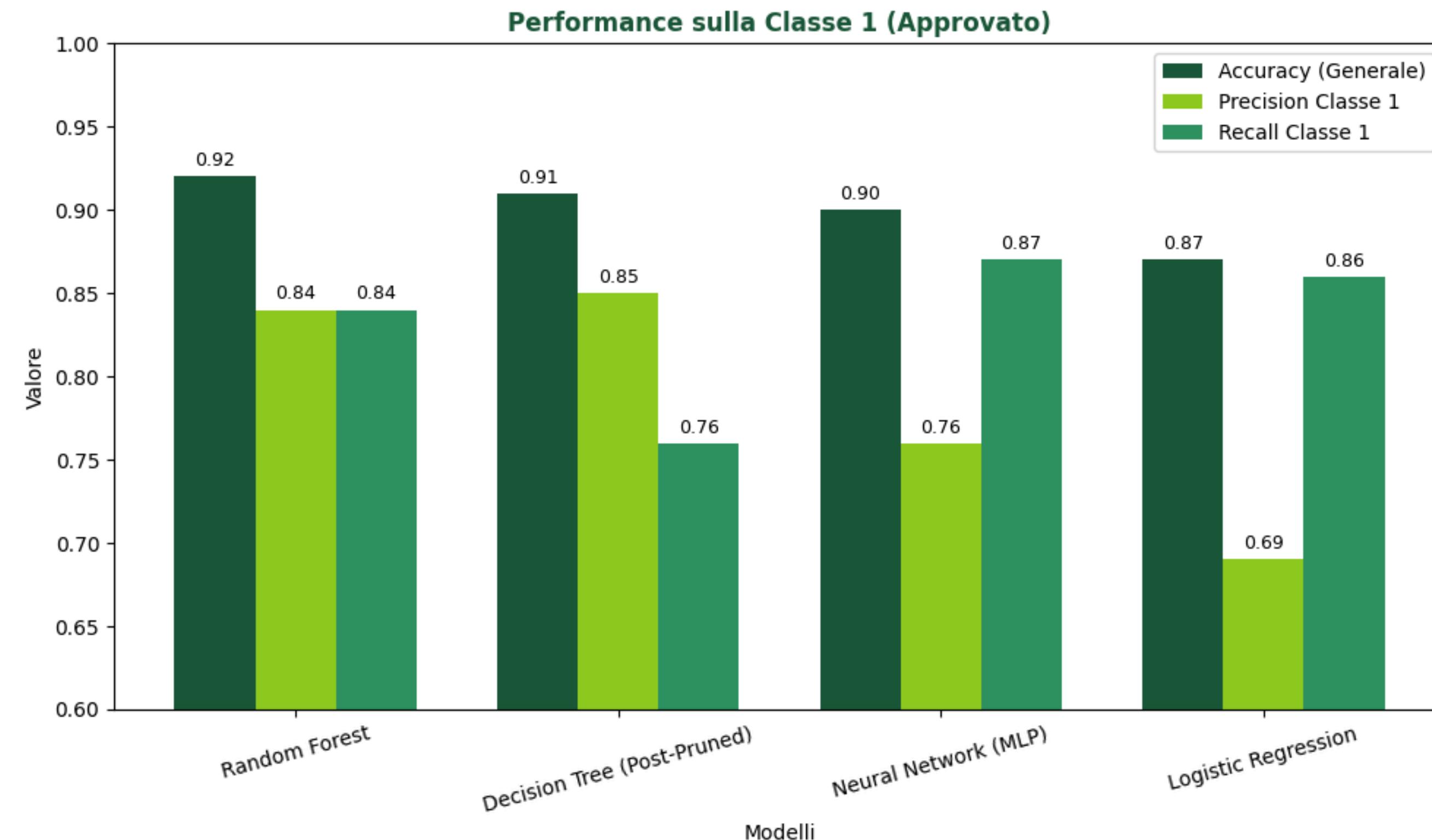
- **Recall alto** (87%): rileva bene i casi concessi;
- **Precision** (76%): qualche falso positivo in più;
- **F1-score** bilanciato: 81%.

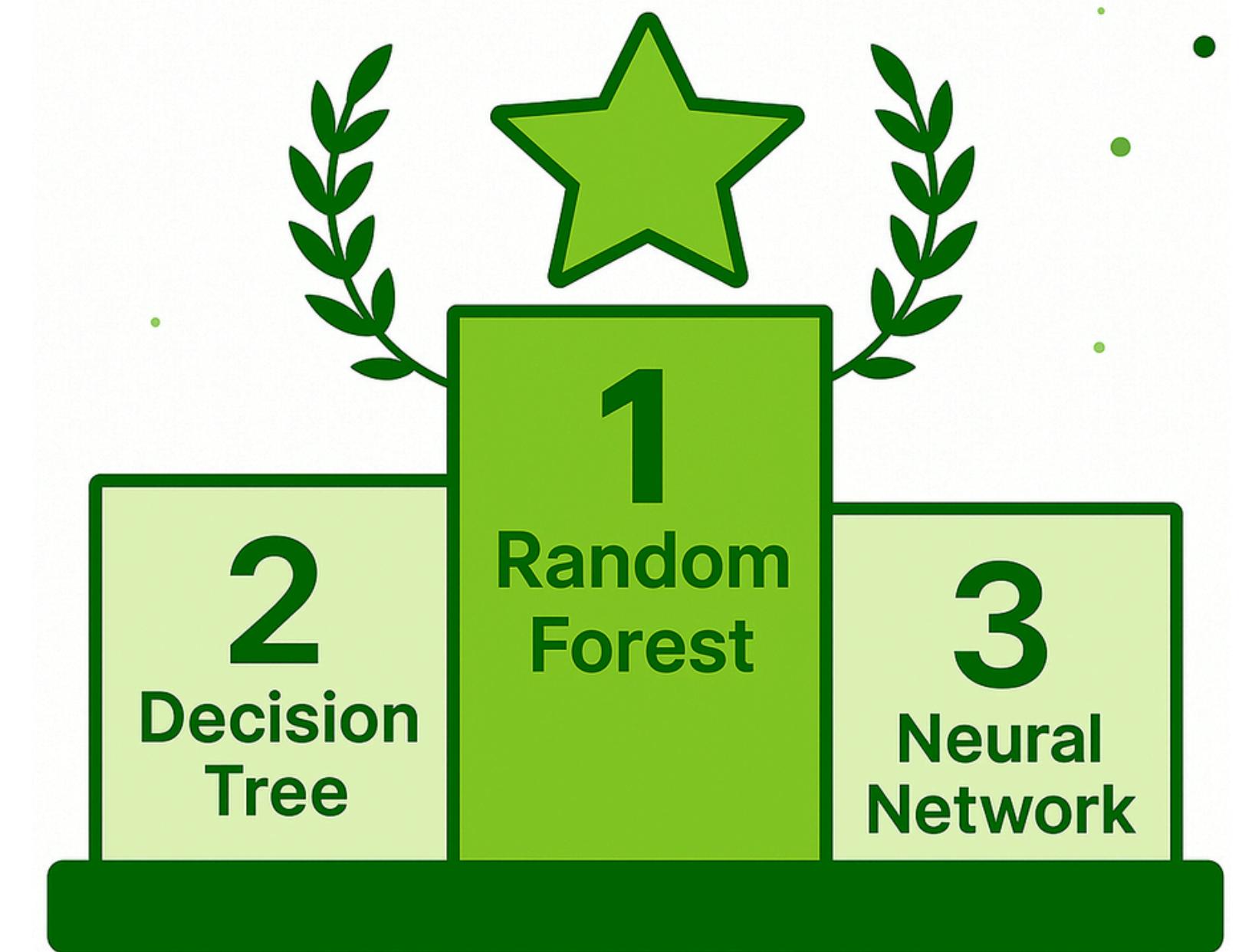


Conclusioni & Prossimi Passi

Qual è il modello migliore?

Confronto fra metriche principali dei modelli





Prossimi passi



Il modello è pronto per fare la differenza.

Il prossimo passo non è tecnico, ma strategico:
Portarlo dove può generare un valore concreto.



Q&A

Grazie per l'attenzione.

Corso: Data & AI – Develhope

- Alessandro Caponigro
- Emanuela Torre
- Francesca Cola