UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI **INGEGNERIA INFORMATICA, MODELLISTICA, ELETTRONICA E SISTEMISTICA**
DIMES

# Anomaly detection per frodi bancarie

A.A. 2022/2023

Progetto del corso di Data Mining

Studentessa Gaia Assunta Bertolino Mat. 242590

# Step

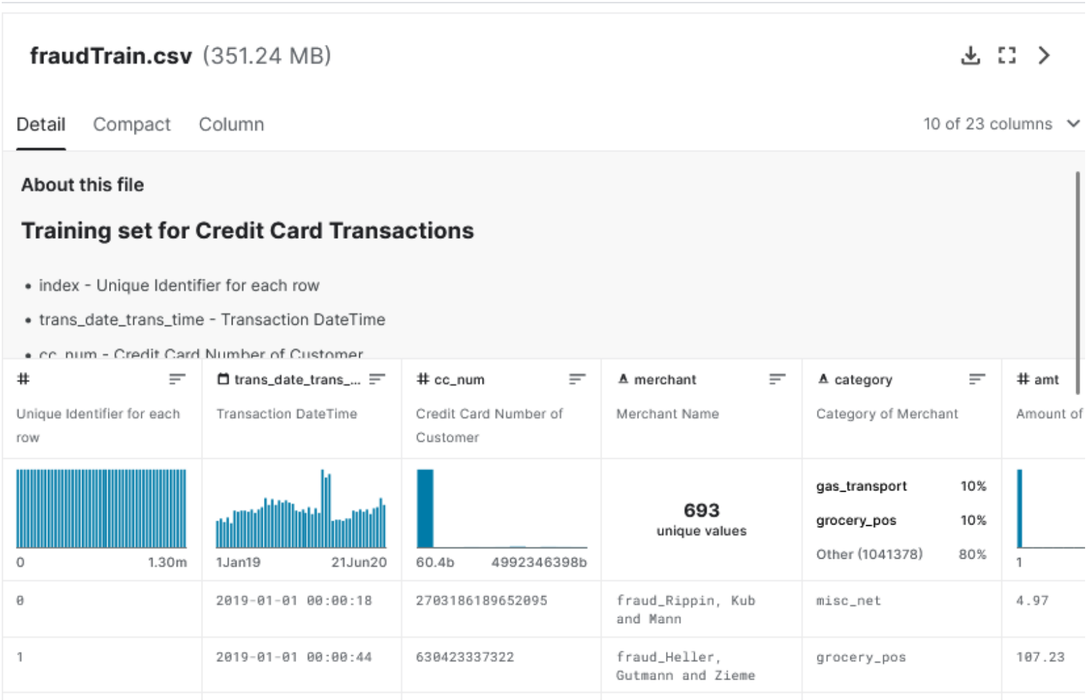
# 01 Introduzione al problema

Dataset contente transazioni con carta di credito

>> ***Obiettivo:*** trovare un modello in grado di individuare le transazioni fraudolente

↓

***ANOMALY DETECTION***




*Fonte*: https://www.kaggle.com/datasets/kartik2112/fraud-detection?select=fraudTrain.csv

# 02 Pre-processing

**Osservazioni sul dataset**

Numero totale di records: 1.852.394

- Training set: 1.296.675 (70%)
- Test set: 555.719 (30%)

**Informazioni sugli attributi**

- Venditore
- Acquisto
- Proprietario

**Trasformazioni degli attributi e riduzione del dataset**

+

- Date
- Time
- Age

x

- trans_date_trans_time
- merchant
- first
- dob
- unix_time
- zip
- street
- cc_num
- lat
- long

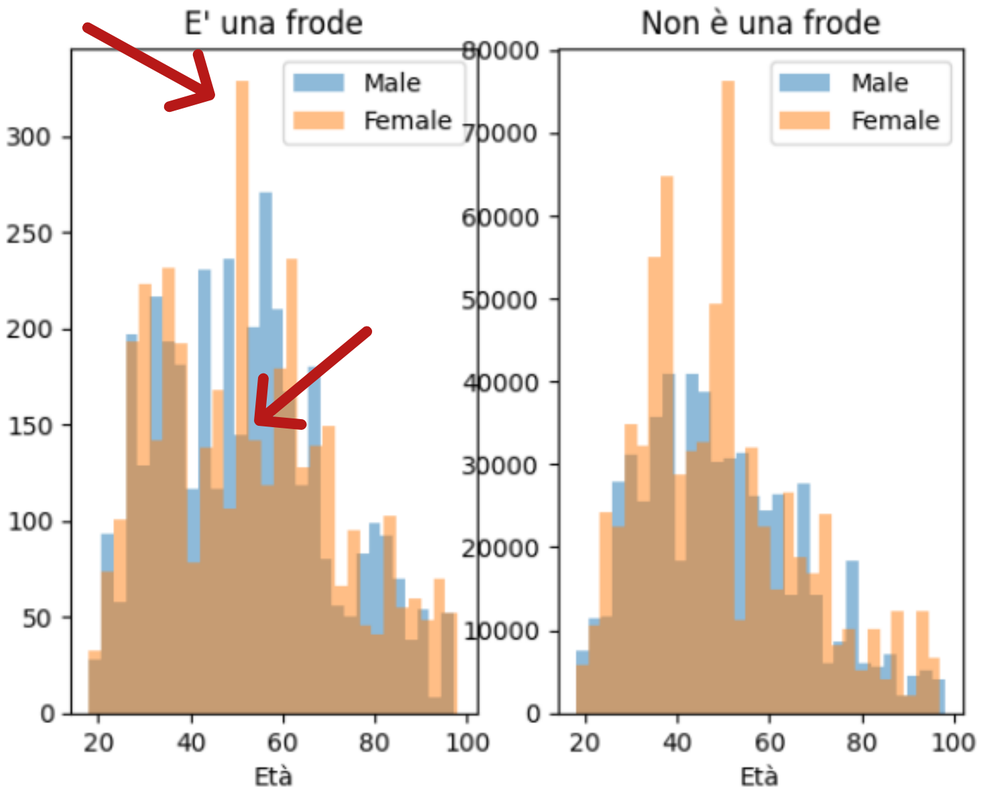
# 03 Valutazioni sugli attributi

1. Le donne sono molto più soggette ad essere derubate

Genere

| gender | is_fraud | count | prob |
|---|---|---|---|
| F | 0 | 706128 | 0.994738 |
| M | 0 | 583041 | 0.993574 |
| | 1 | 3771 | 0.006426 |
| F | 1 | 3735 | 0.005262 |

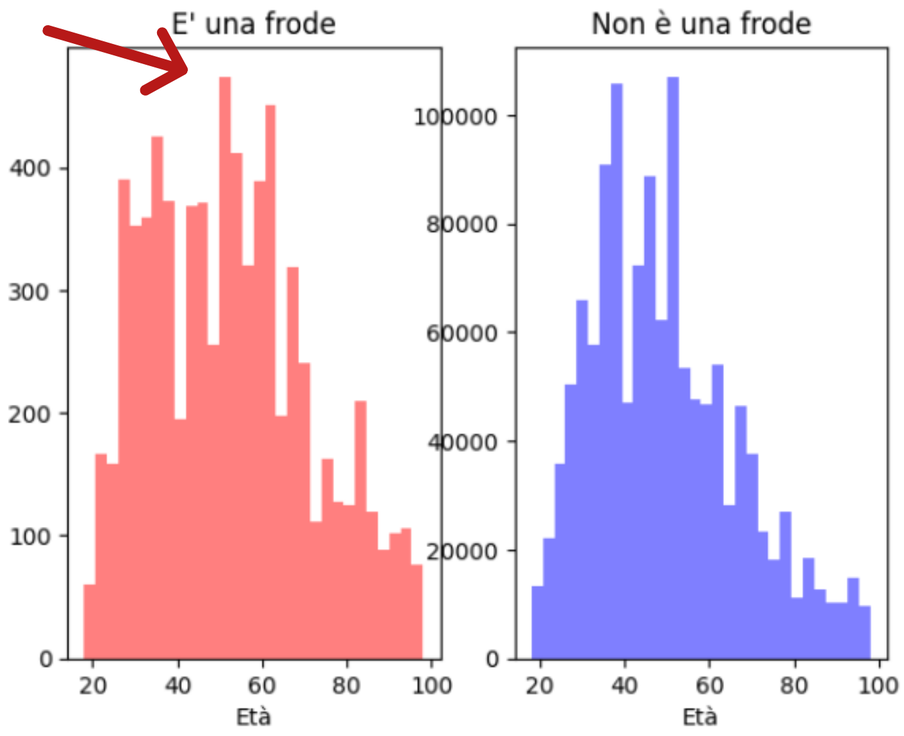
Sovrapposizione fra genere e età



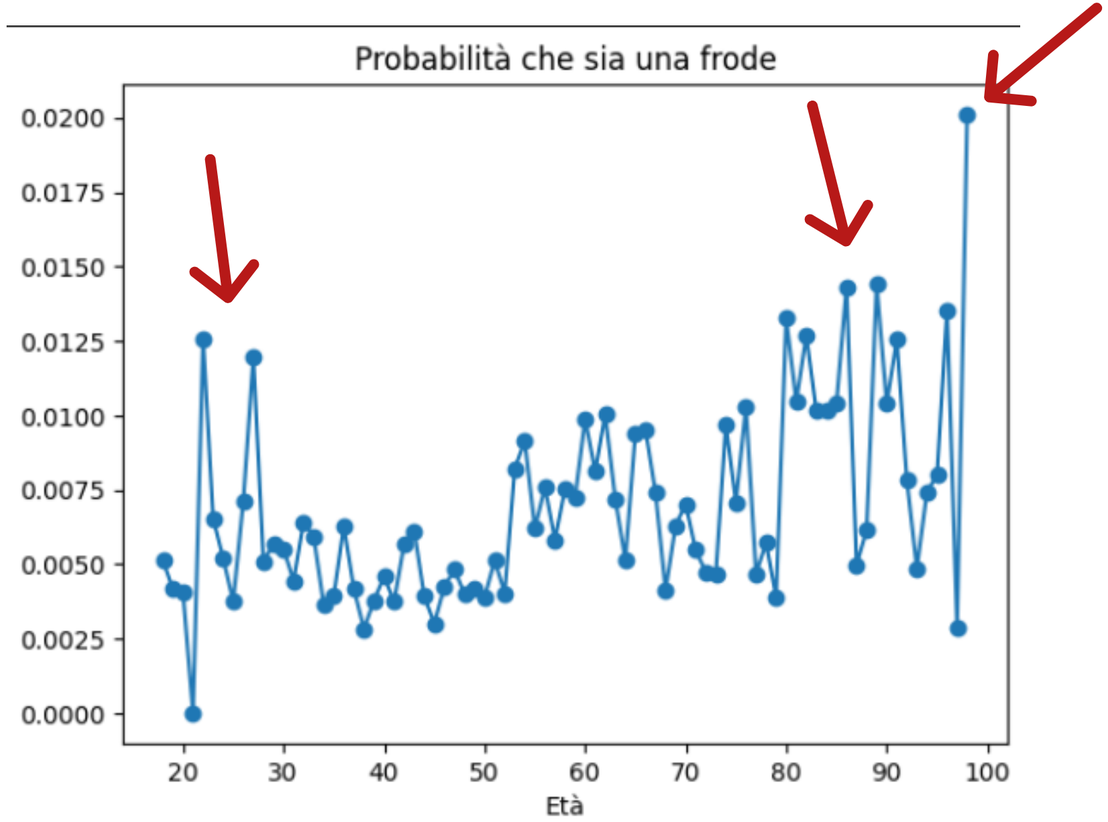
# 03 Valutazioni sugli attributi

2. Gli anziani sono più soggetti ad essere derubati

Età



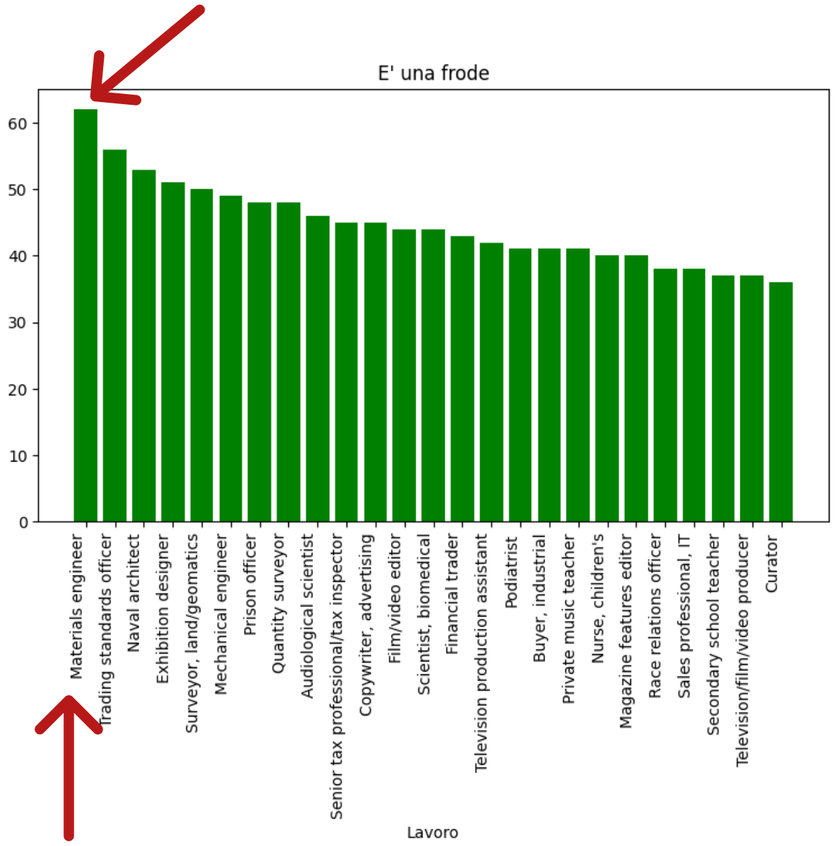

Legame probabilità-età

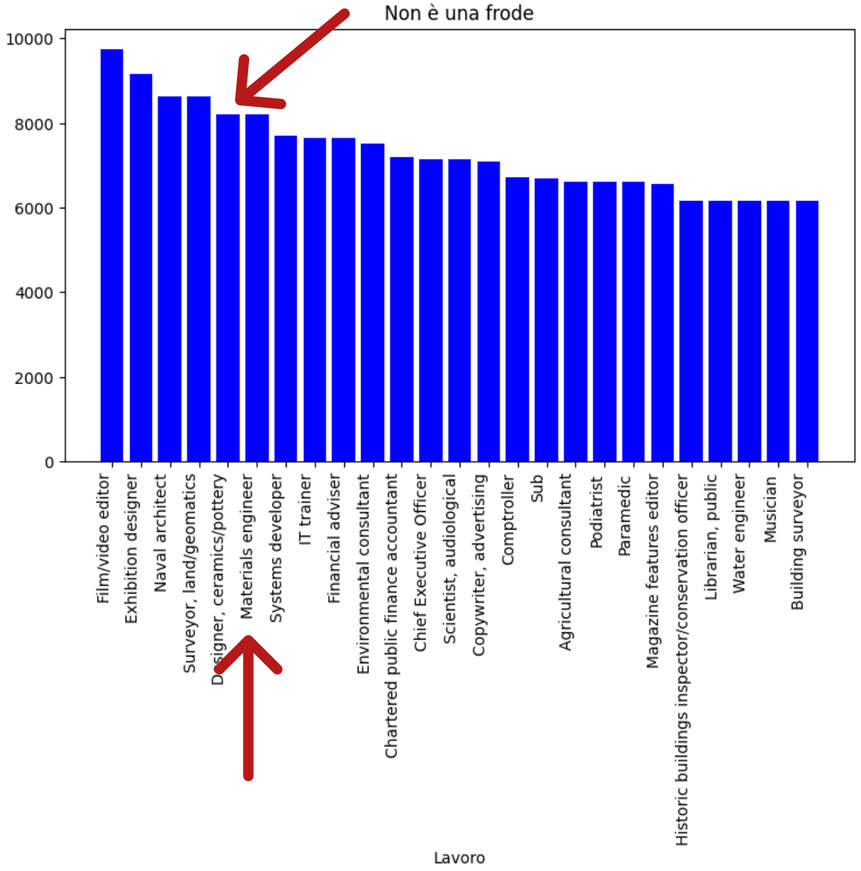


# 03 Valutazioni sugli attributi

Lavoro

3. Ci sono lavori in cui si è più soggetti ad essere derubati

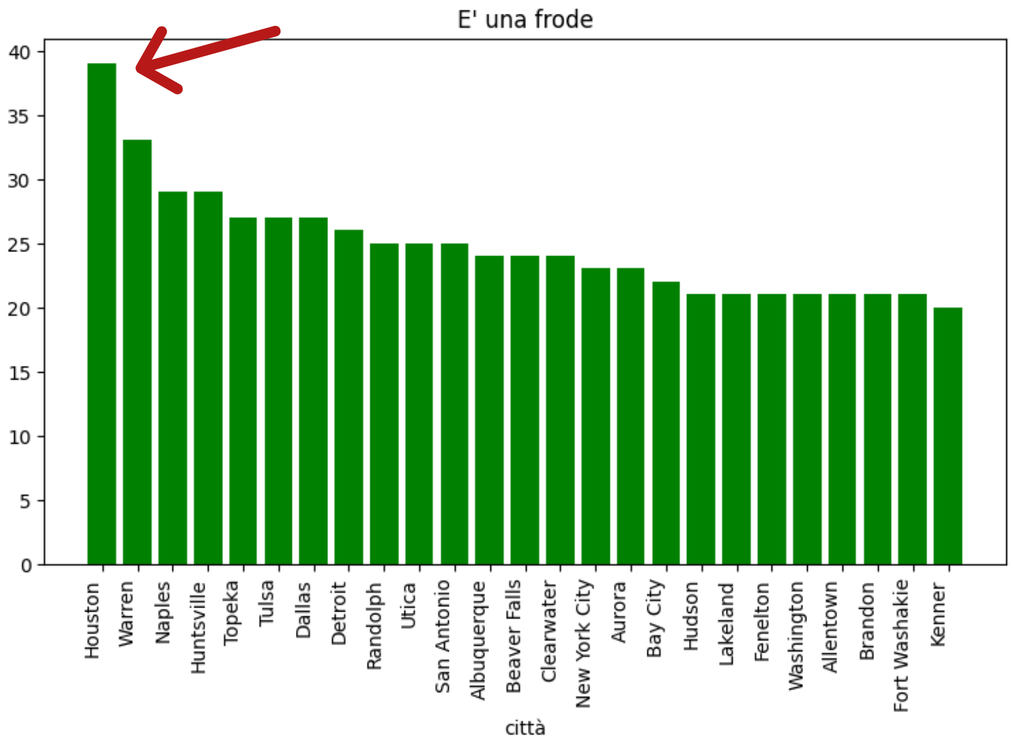


# 03 Valutazioni sugli attributi

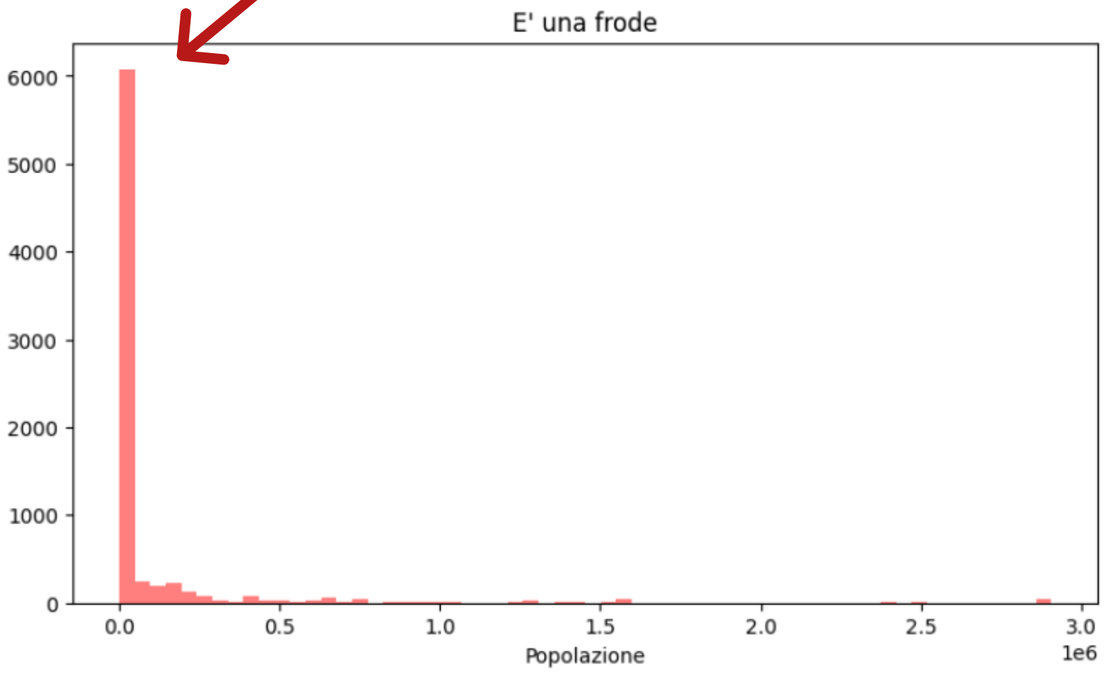
4. Ci sono città in cui si è più soggetti ad essere derubati

Città




5. Nelle città più popolose si è più soggetti ad essere derubati

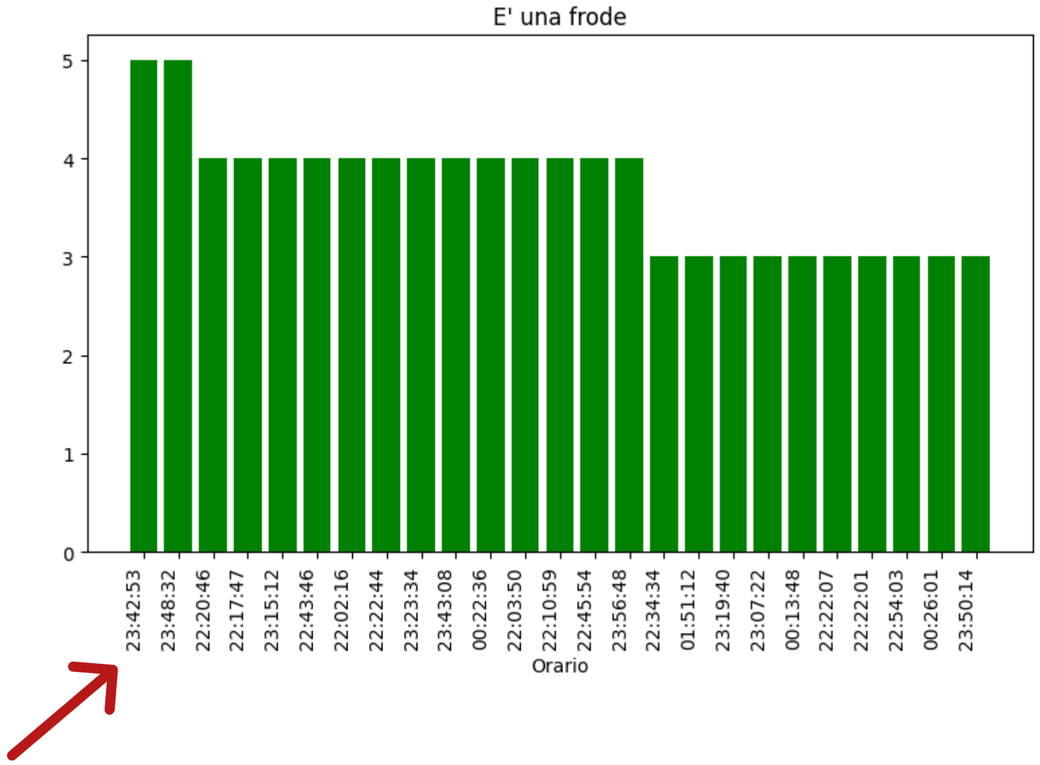
Popolazione



# 03 Valutazioni sugli attributi

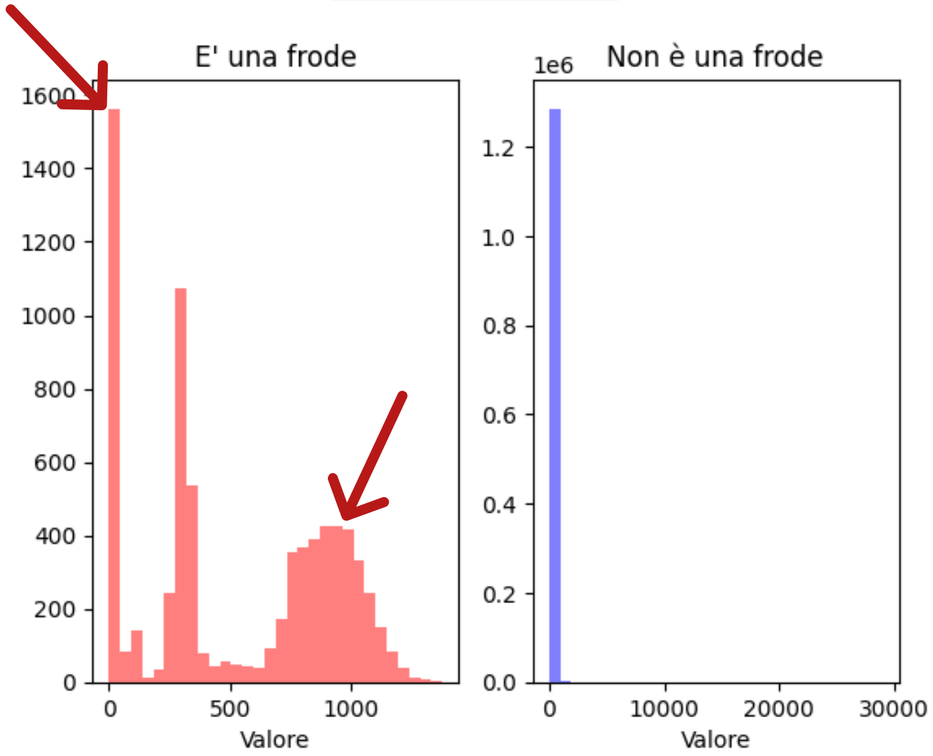
6. Negli orari più tardi è più facile compiere transazioni fraudolente

Orario




7. Transazioni fraudolente più piccole sono più facili da fare

Ammontare

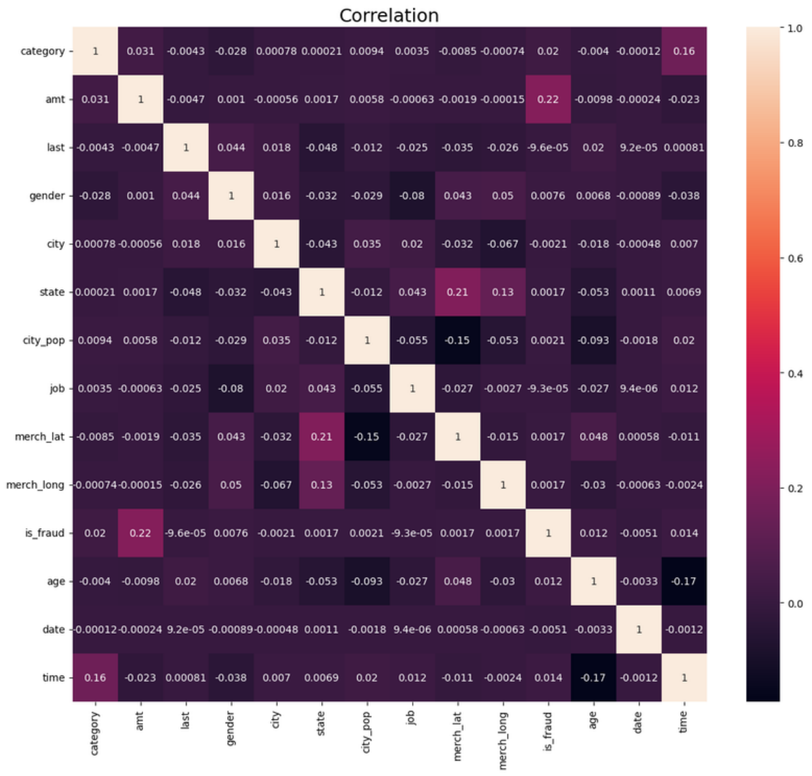


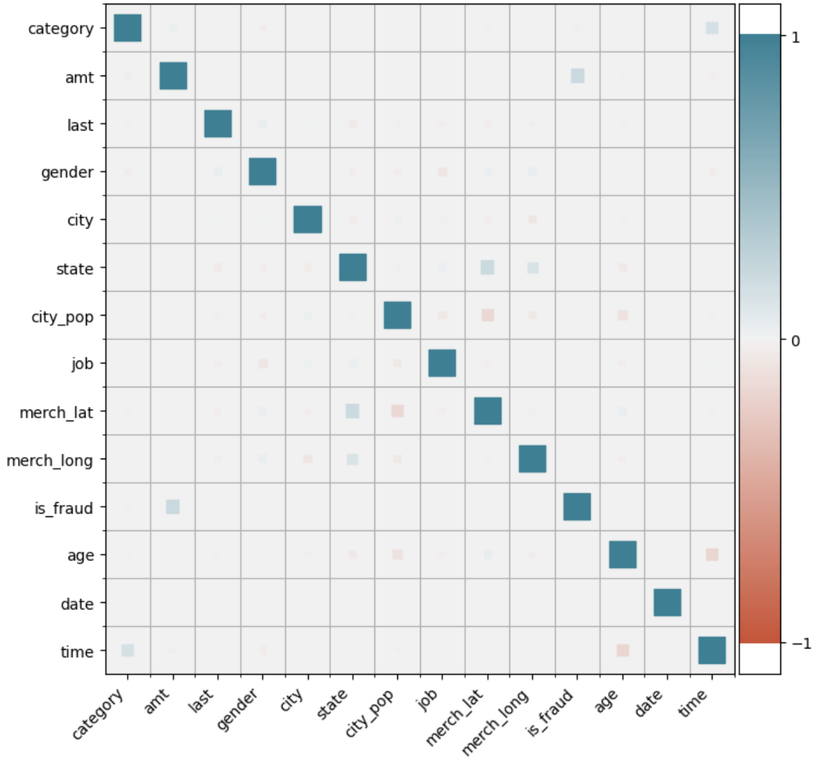
# 03 Valutazioni sugli attributi

One-hot encoding

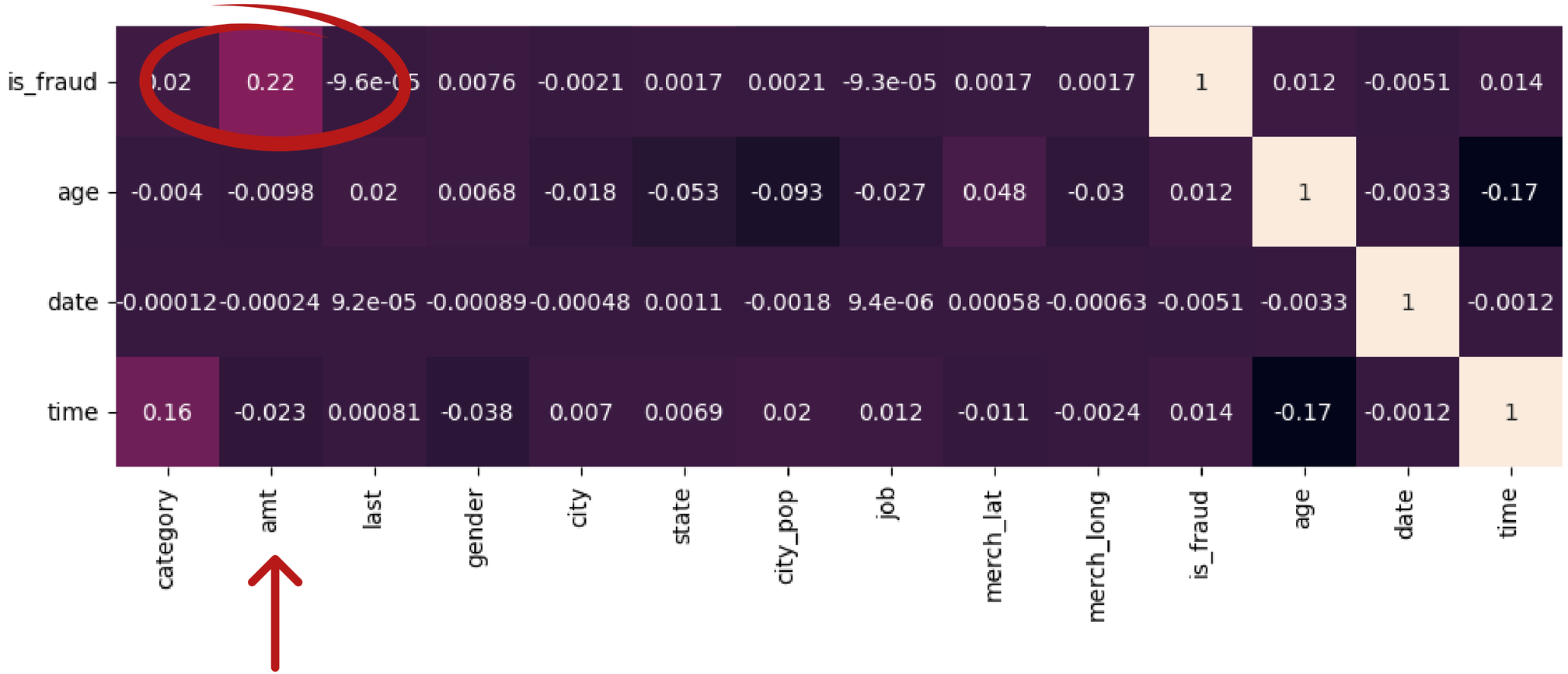
Applicazione della funzione di correlazione

Rappresentazione grafica tramite heatmap



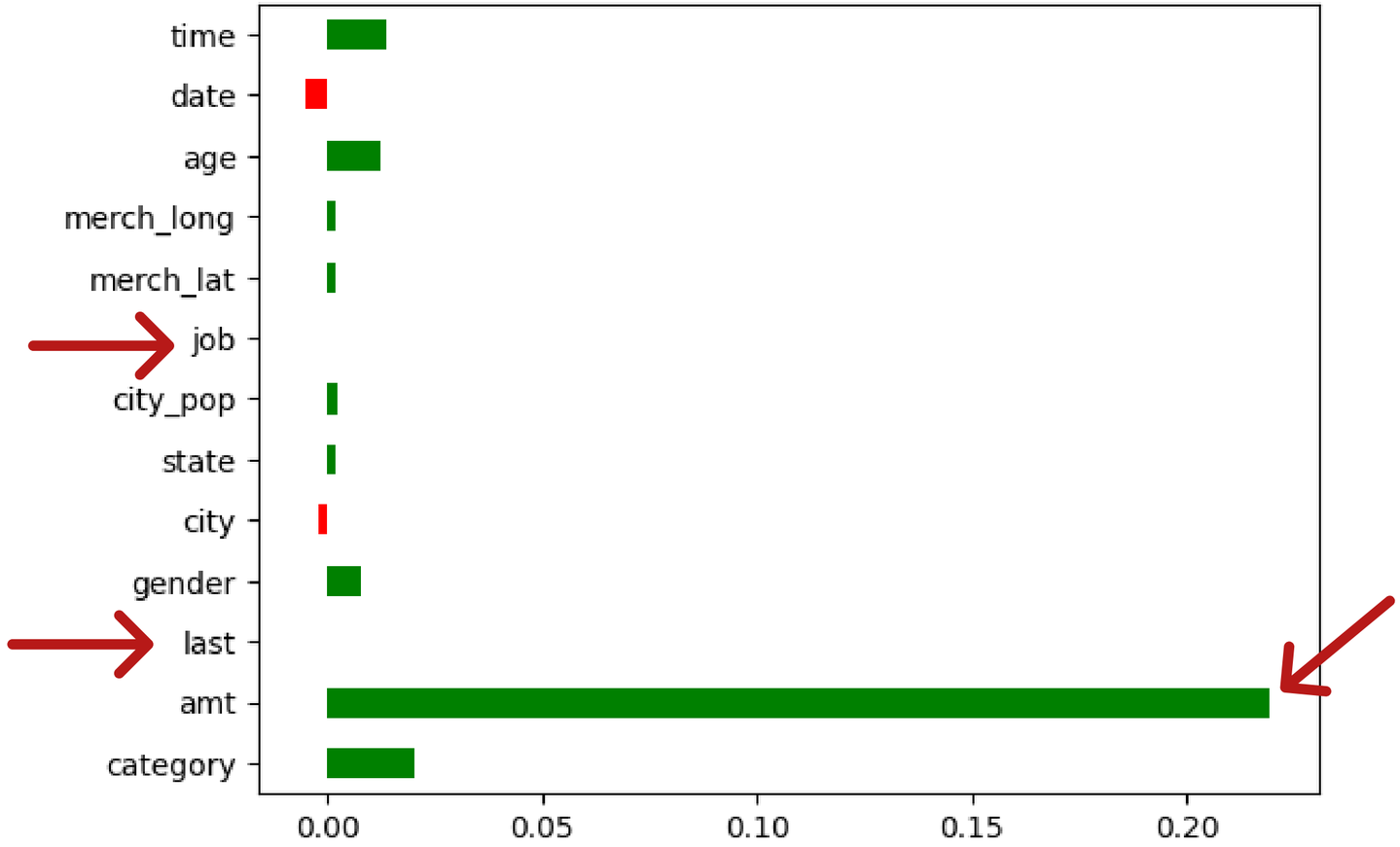
# 03 Valutazioni sugli attributi



# 03 Valutazioni sugli attributi

Rilevanza dell'attributo *amt*

Irrilevanza degli attributi *job* e *last*

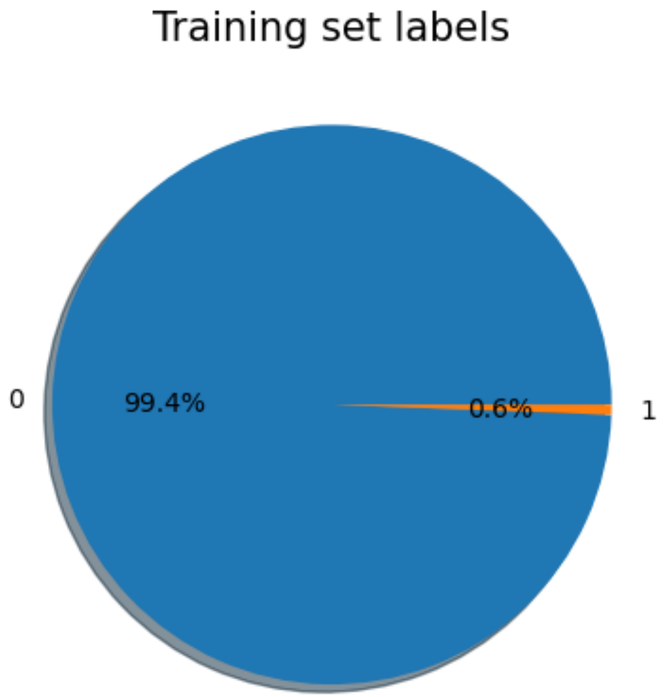


# 04 Applicazione del machine learning

Suddivisione fra training set effettivo e validation set

Numero totale di records: 1.296.675

- Training set: 868.772 (67%)
- Validation set: 427.903 (33%)



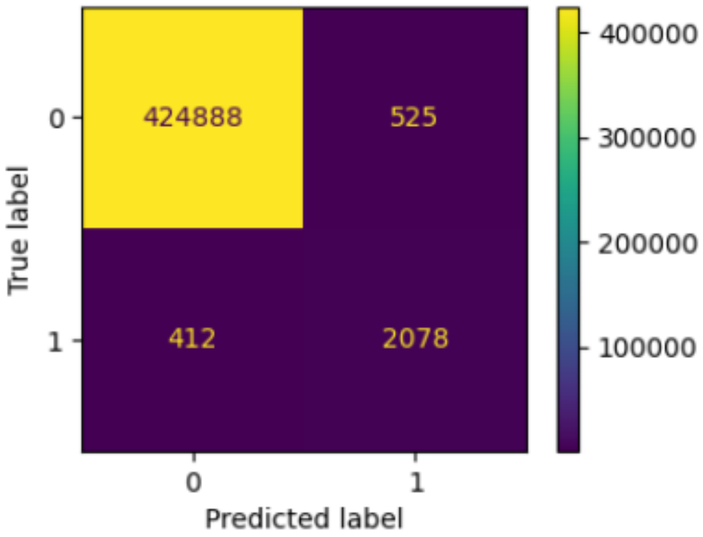

Applicazione degli algoritmi di Machine Learning

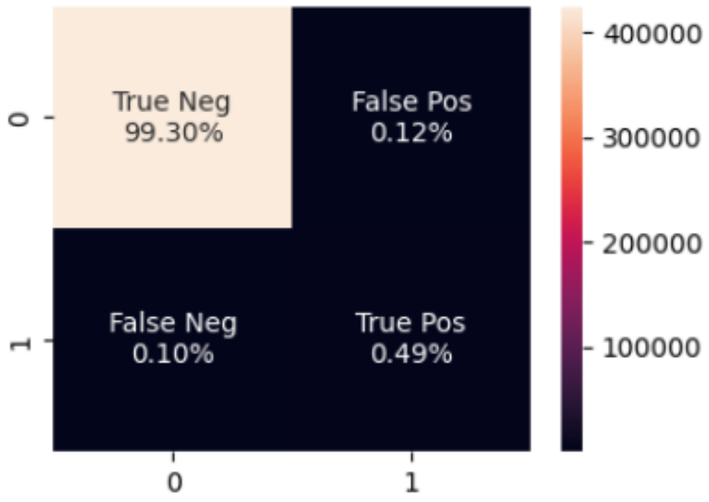
- Decision tree
- Random forest
- Isolation forest
- Local Outlier Factor
- DBSCAN
- K-Nearest Neighbors

# 04 Applicazione del machine learning

Decision Tree

- *Accuracy 99.78%*
- *Precision 79.83%*
- *Recall 83.45%*
- ***F1** 0.9075*
- *AUC 0.9167*

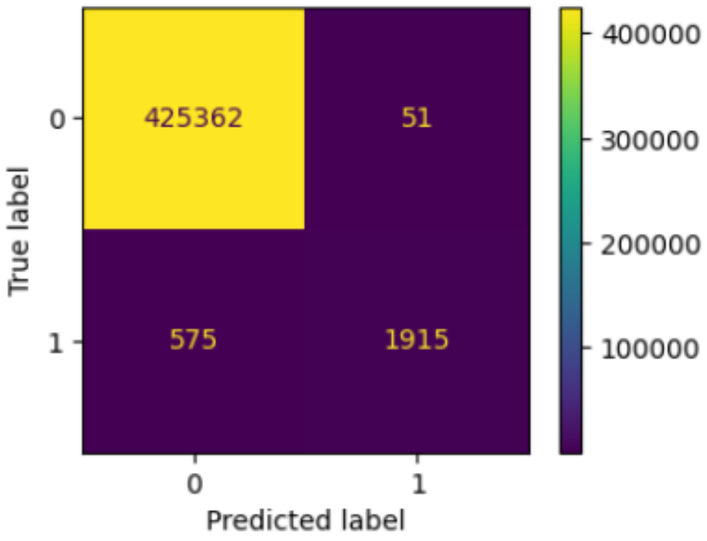


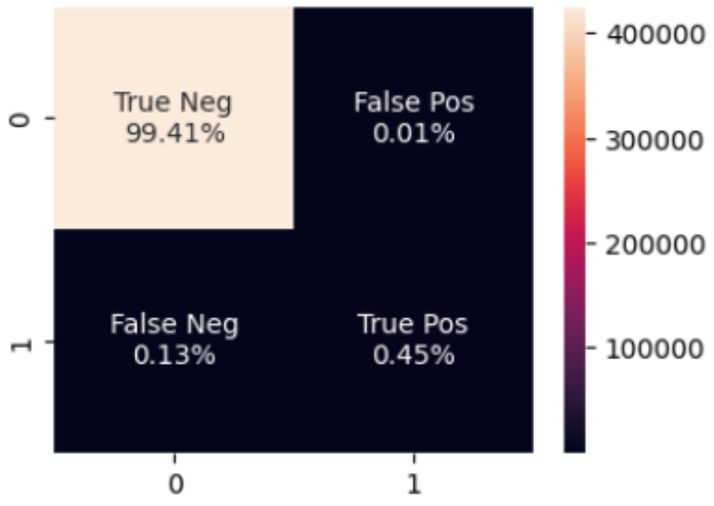



# 04 Applicazione del machine learning

Random Forest

- *Accuracy 99.85%*
- *Precision 97.41%*
- *Recall 76.91%*
- ***F1** 0.9294*
- *AUC 0.8845*

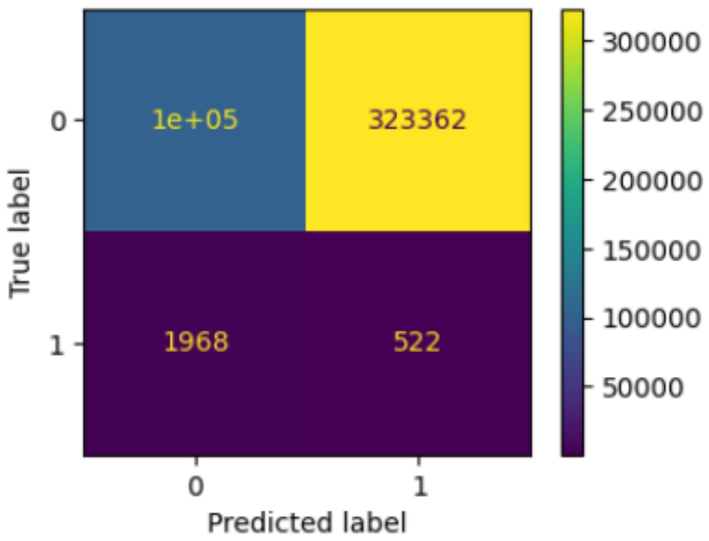


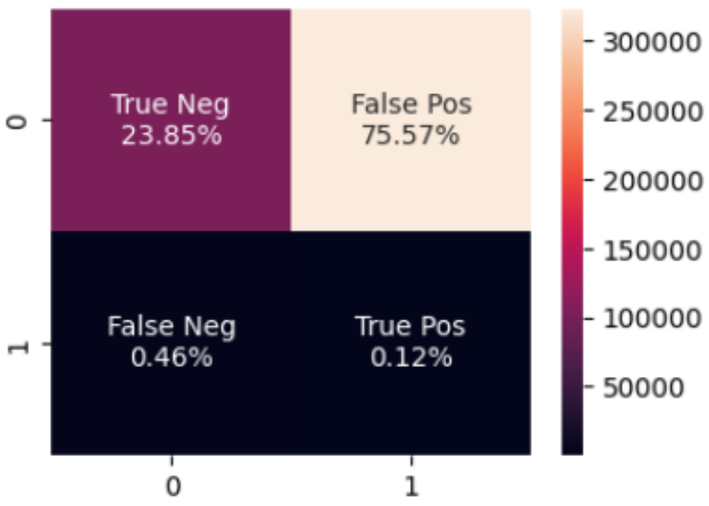



# 04 Applicazione del machine learning

Isolation Forest

- *Accuracy 23.97%*
- *Precision 0.16%*
- *Recall 20.96%*
- ***F1** 0.1944*
- *AUC 0.2248*



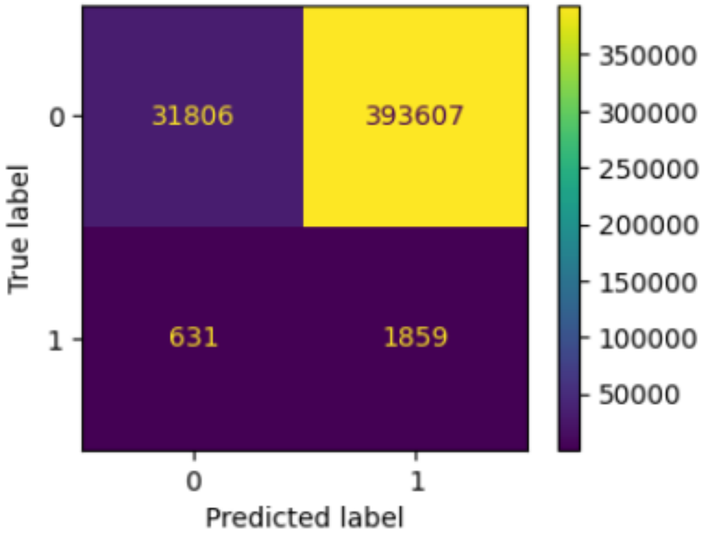



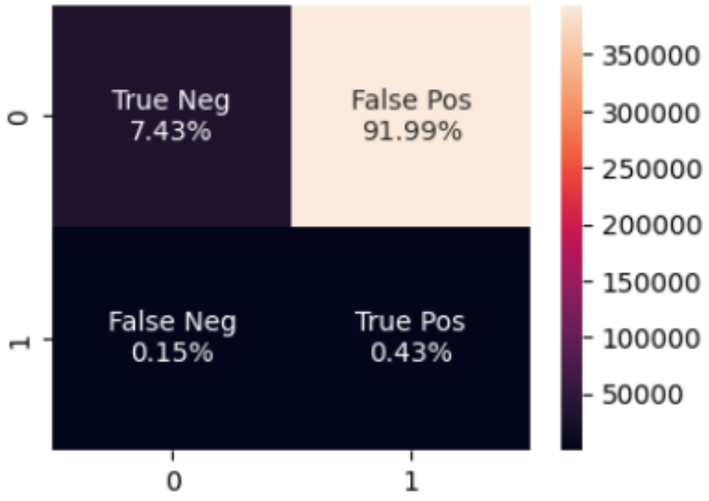

*OVERFITTING*

# 04 Applicazione del machine learning

Local Outlier Factor

- *Accuracy 7.87%*
- *Precision 0.47%*
- *Recall 74.66%*
- ***F1** 0.0741*
- *AUC 0.4107*



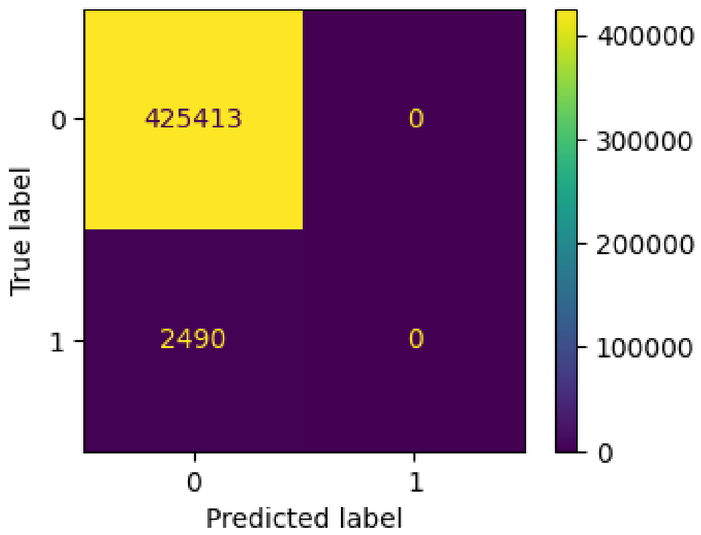



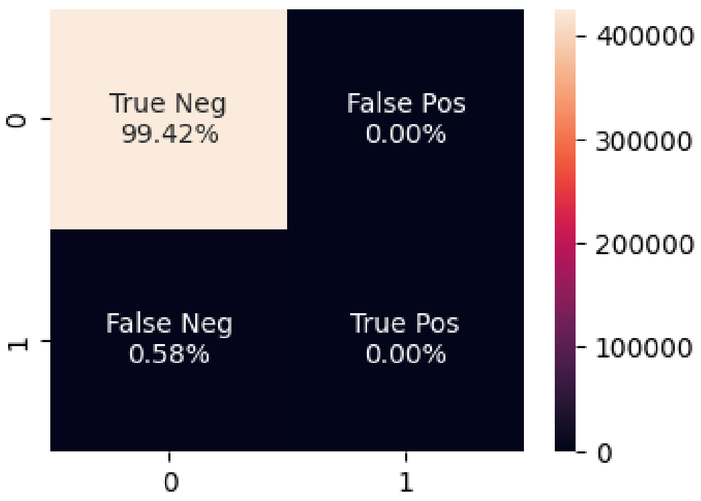

*OVERFITTING*

# 04 Applicazione del machine learning

DBSCAN

- *Accuracy 99.42%*
- *Precision 0.0%*
- *Recall 0.0%*
- ***F1** 0.4985*
- *AUC 0.5000*



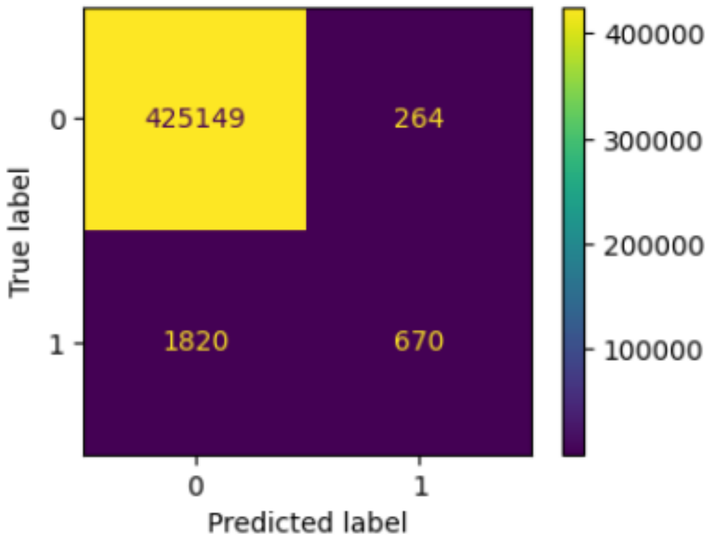



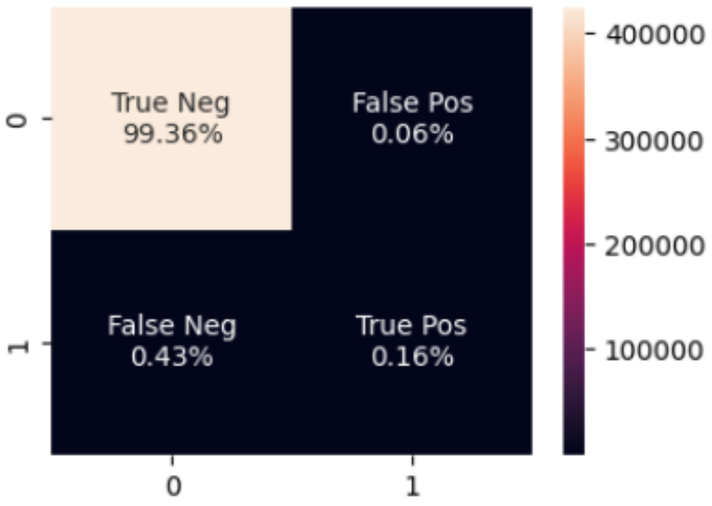

*UNDERFITTING*

# 04 Applicazione del machine learning

K-Nearest Neighbors

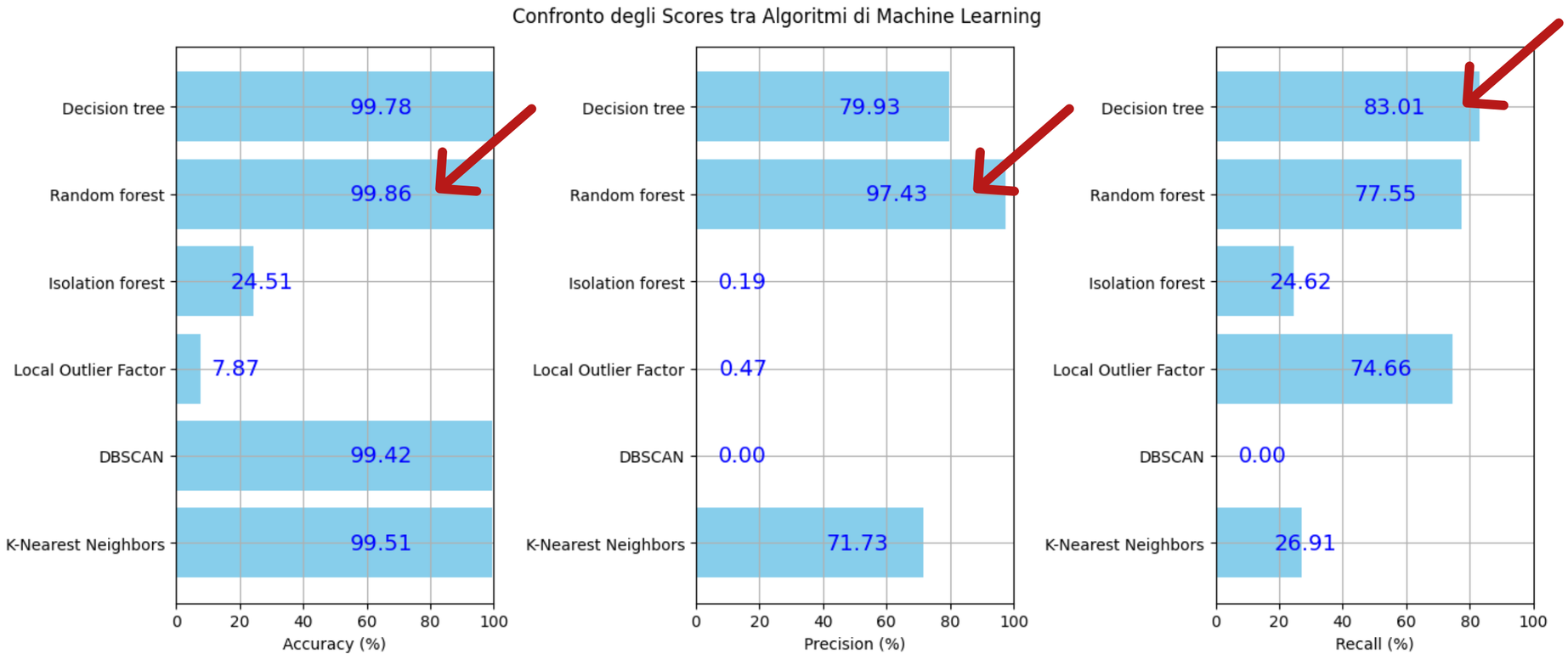
- *Accuracy 99.51%*
- *Precision 71.73%*
- *Recall 26.91%*
- ***F1** 0.6945*
- *AUC 0.6342*



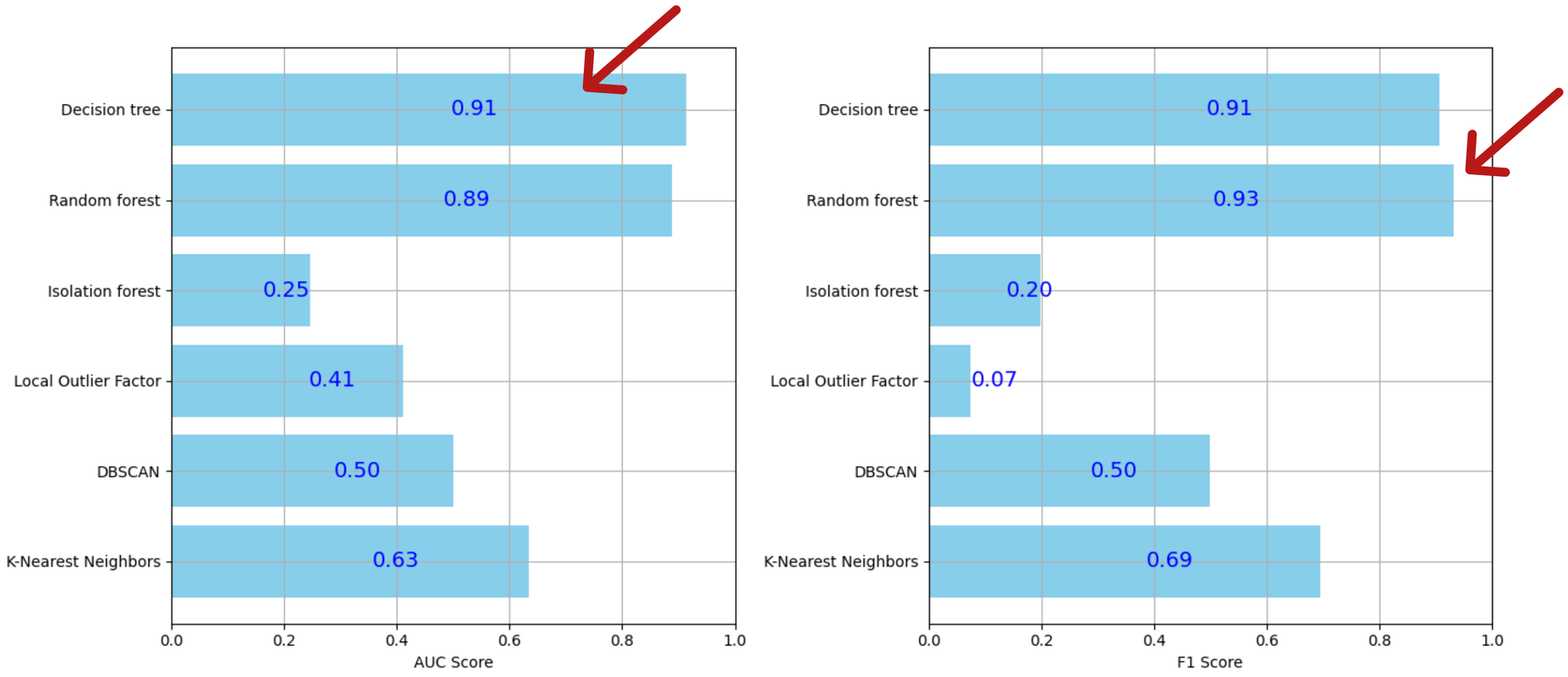



# 04 Applicazione del machine learning

## Criteri di valutazione



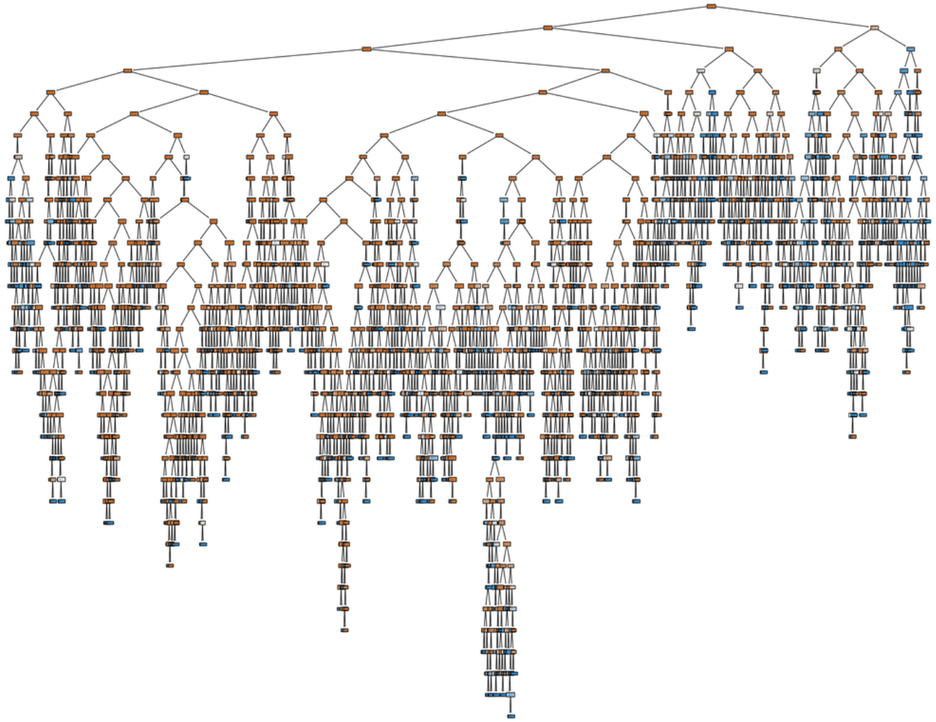
# 04 Applicazione del machine learning




Algoritmo scelto: Decision Tree

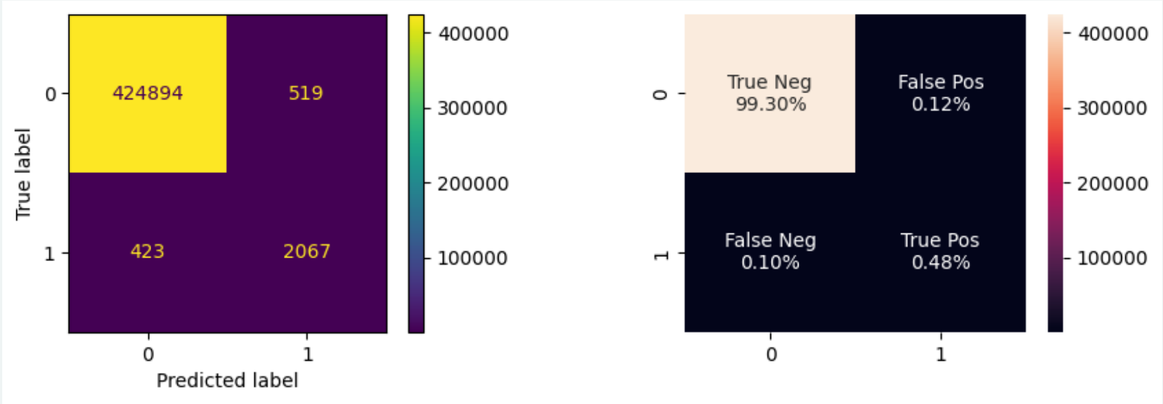
# 04 Applicazione del machine learning

Albero di decisione generato



Scores:

- Accuracy: 99.78%
- Precision: 79.93%
- Recall: 83.45%
- AUC: 0.9167
- F1: 0.9075



# 05 Tuning

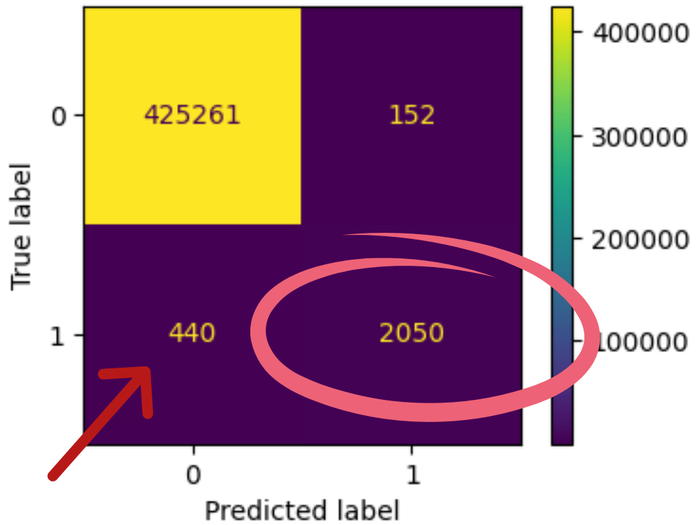
Tuning su accuracy

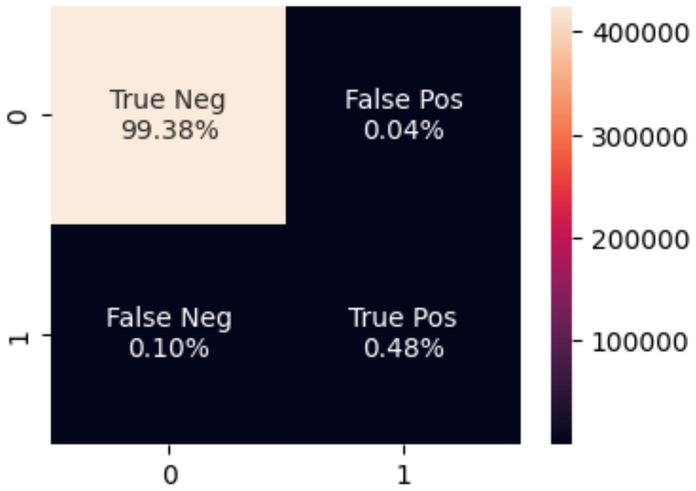
## Modello:

- Criterion: ‘entropy’
- max_depth: 10
- min_samples_leaf: 10
- random_state: 30

## Scores:

- Accuracy: 99.86%
- Precision: 93.1%
- Recall:82.33%
- AUC: 0.9115
- F1: 0.9366



# 05 Tuning

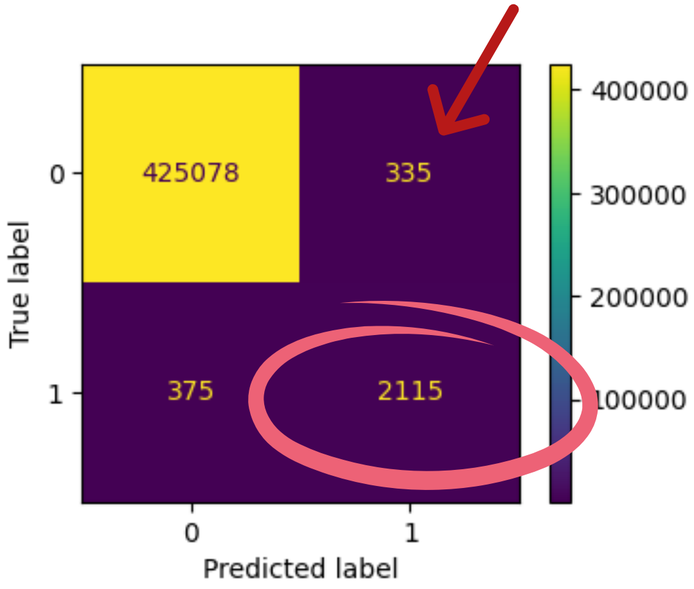
Tuning su recall

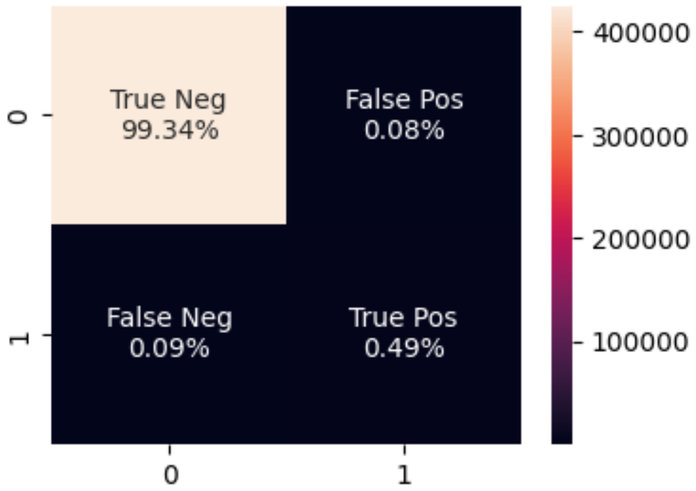
## Modello:

- Criterion: ‘entropy’
- max_depth: 20
- min_samples_leaf: 5
- random_state: 30

## Scores:

- Accuracy: 99.83%
- Precision: 86.33%
- Recall: 84.94%
- AUC: 0.9243
- F1: 0.9277



# 05 Tuning

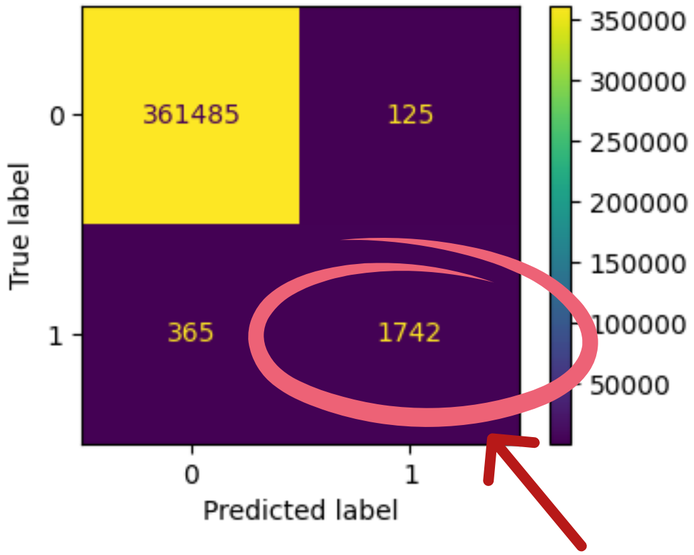
Tuning su f1

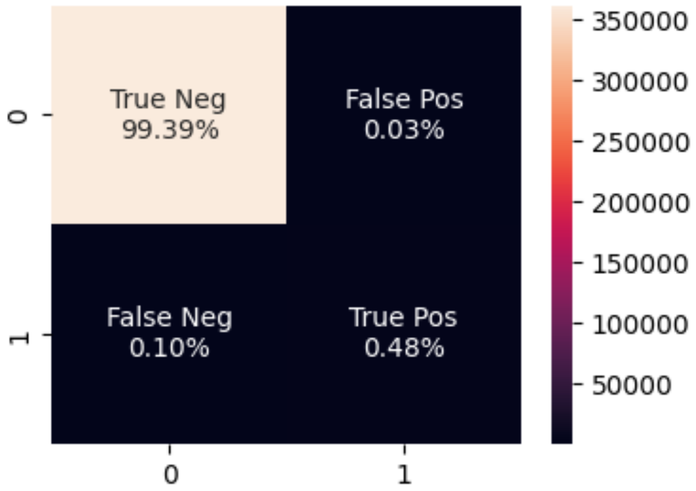
## Modello:

- Criterion: ‘entropy’
- max_depth: 10
- min_samples_leaf: 10
- random_state: 30

## Scores:

- Accuracy: 99.87%
- Precision: 93.30%
- Recall: 82.68%
- AUC: 0.9132
- F1: 0.9380



# 05 Conclusioni sul test set

## Miglior modello:

**decision tree** con tuning su **recall** score

- *criterion='entropy'*
- *max_depth=20*
- *min_samples_leaf=5*
- *random_state=30*

## Scores sul test set:

- Accuracy: 99.84%
- Precision: 78.97%
- Recall: 79.86%
- AUC: 0.8989
- F1: 0.8953