

Identification of At-Risk Data via Kernelized SVMs

Michele Mastroberti

Sicurezza delle Architetture
Orientate ai Servizi

Obiettivi del Progetto

- Identificare e classificare dati sensibili con rischio differenziato
- Utilizzare SVM per analizzare la separabilità dei dati in funzione della loro vicinanza ai confini decisionali
- Confrontare l'efficacia di kernel lineari e non lineari

Preprocessing

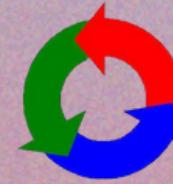
- Standardizzazione delle variabili numeriche (z-score scaling).
- One-Hot Encoding per trasformare le variabili categoriali in valori numerici.
- Suddivisione del dataset:
 - 80% training set
 - 20% test set

Metodologia per l'Assegnazione dei Livelli di Rischio



Identificazione dei Support Vectors

I punti più vicini al confine decisionale della SVM vengono identificati come critici e assegnati ai primi livelli di rischio



Assegnazione Iterativa

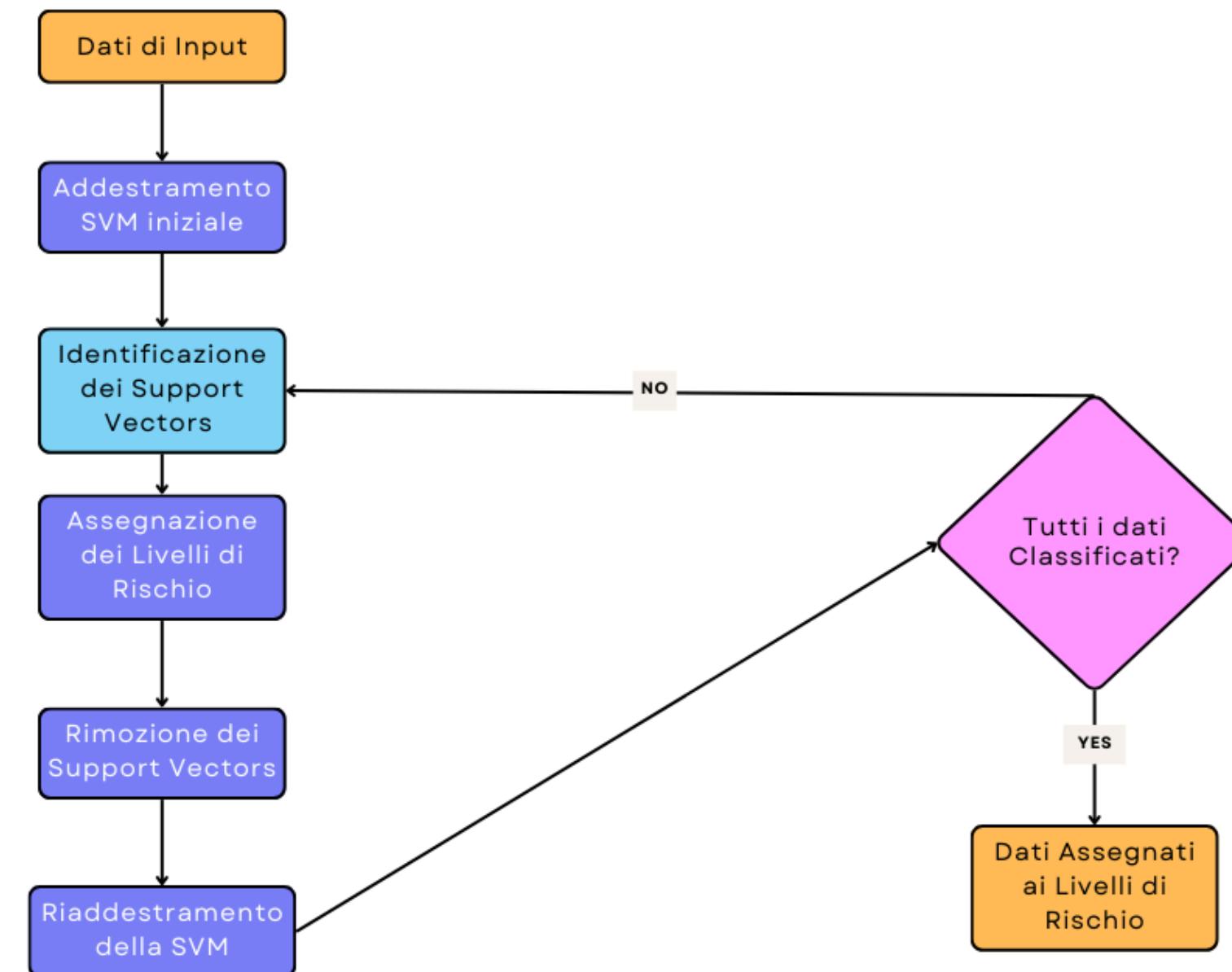
Dopo ogni iterazione, i support vectors vengono rimossi e la SVM viene riaddestrata, assegnando i livelli successivi



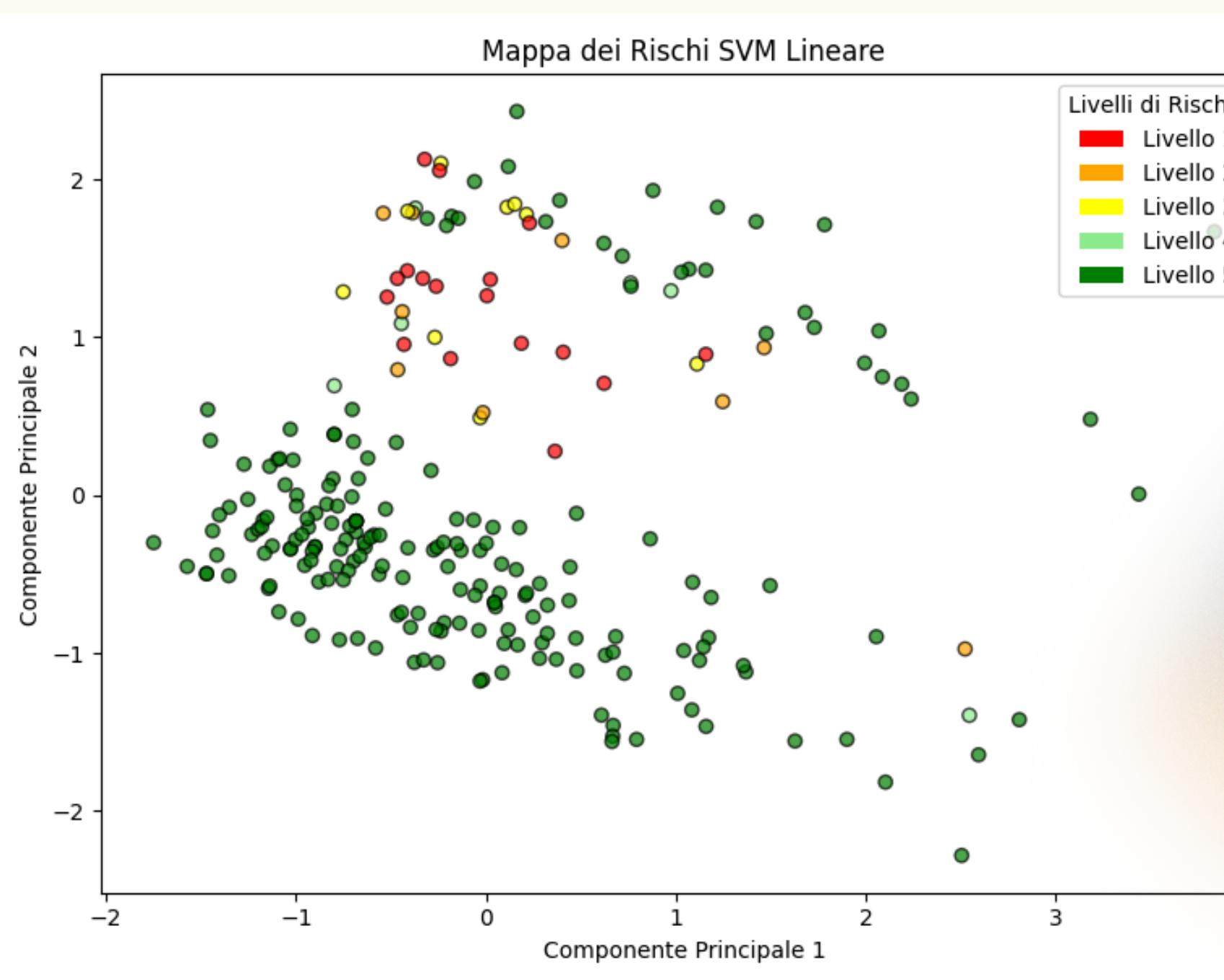
Scala dei Livelli di Rischio

I dati vengono classificati su 5 livelli, dal più alto (1 - rosso) al più basso (5 - verde), migliorando la comprensione del rischio

Processo di Assegnazione del Rischio



Analisi del Modello SVM Lineare



Caratteristiche del kernel lineare:

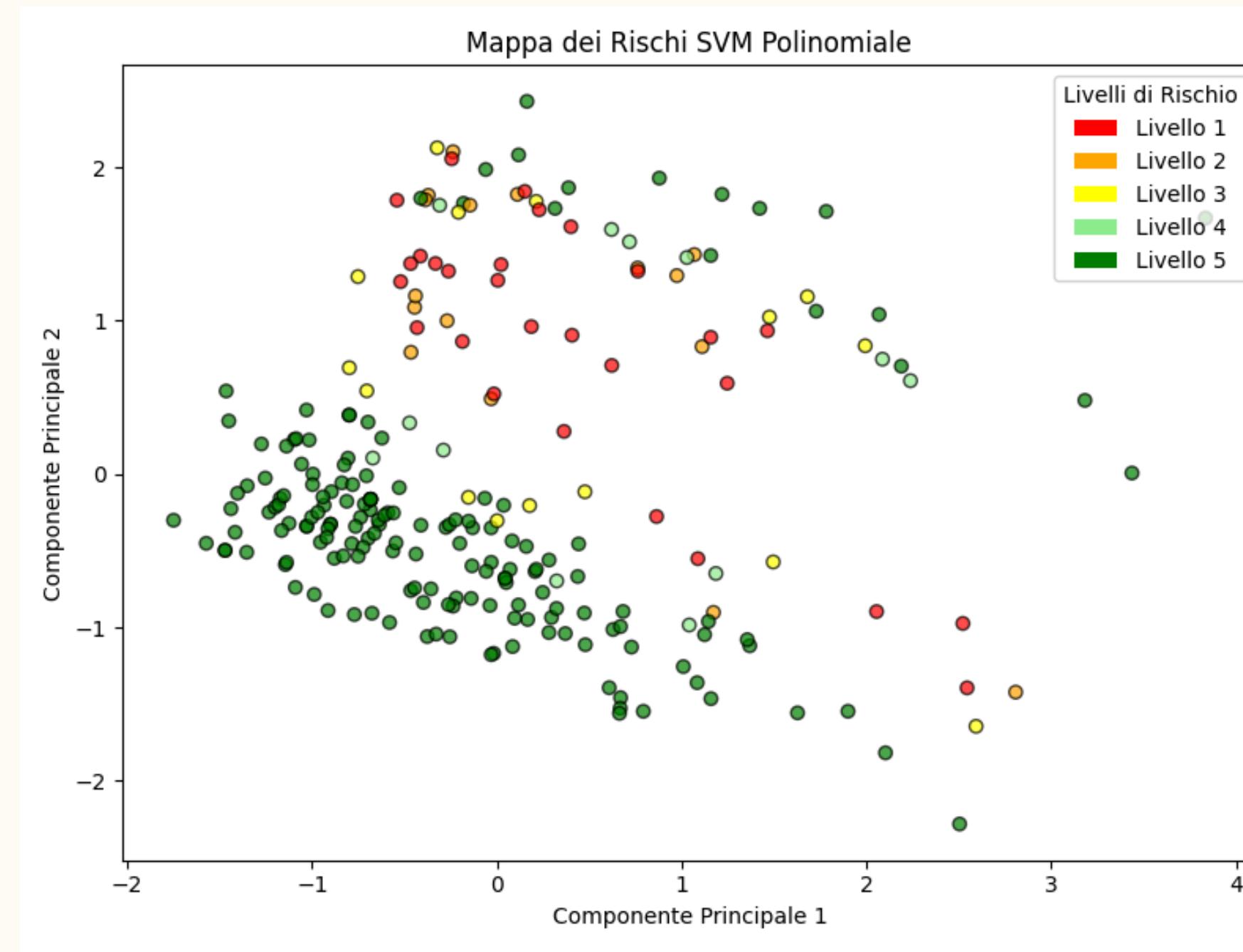
- Separazione dei dati tramite un iperpiano lineare
- Adatto a problemi con dati ben separabili linearmente

Vantaggi e limitazioni:

- ✓ Elevata efficienza computazionale
- ✗ Non cattura relazioni non lineari

Analisi Modello SVM

Polinomiale



Caratteristiche del kernel polinomiale:

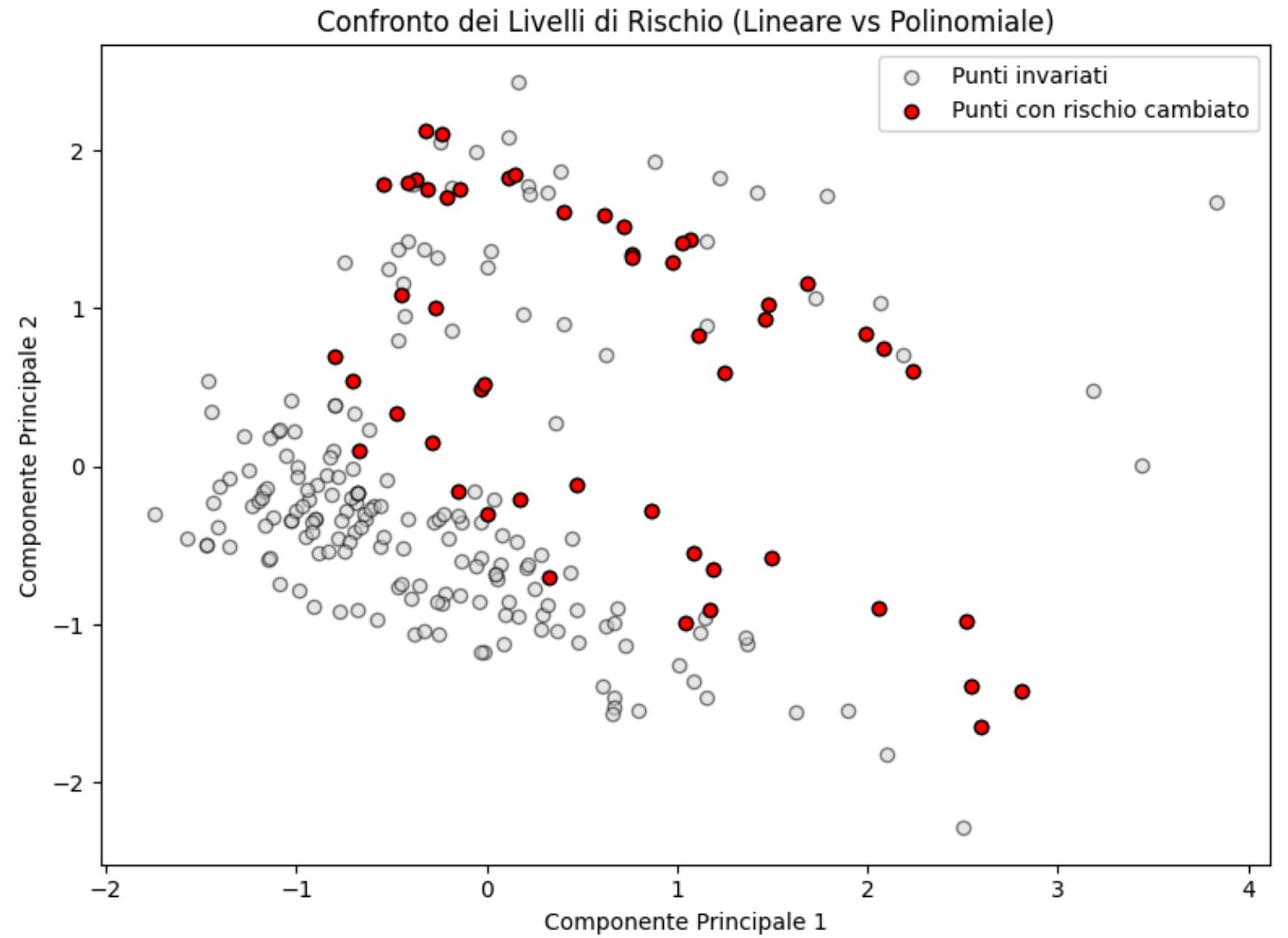
- Permette di gestire relazioni non lineari tra le variabili
- Maggiore capacità di adattamento ai dati complessi

Vantaggi e limitazioni:

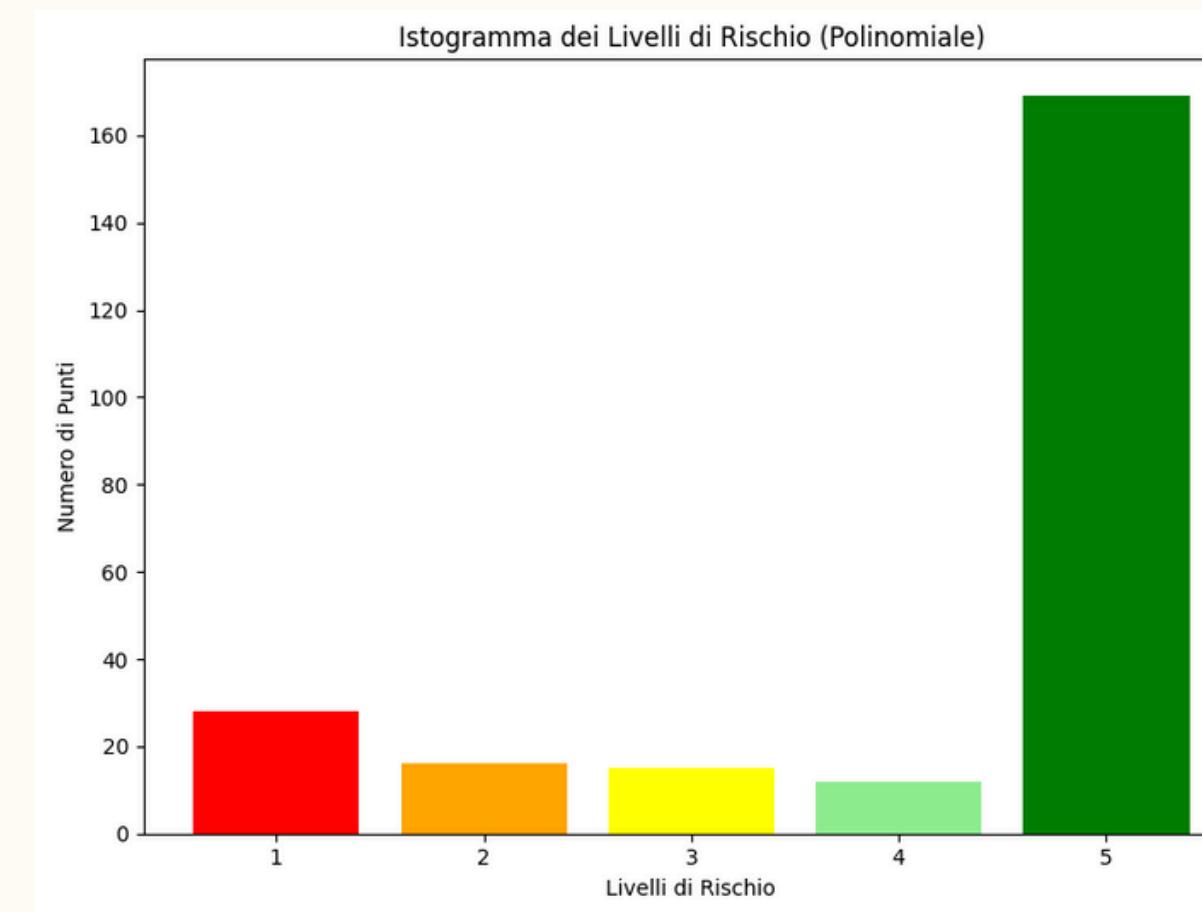
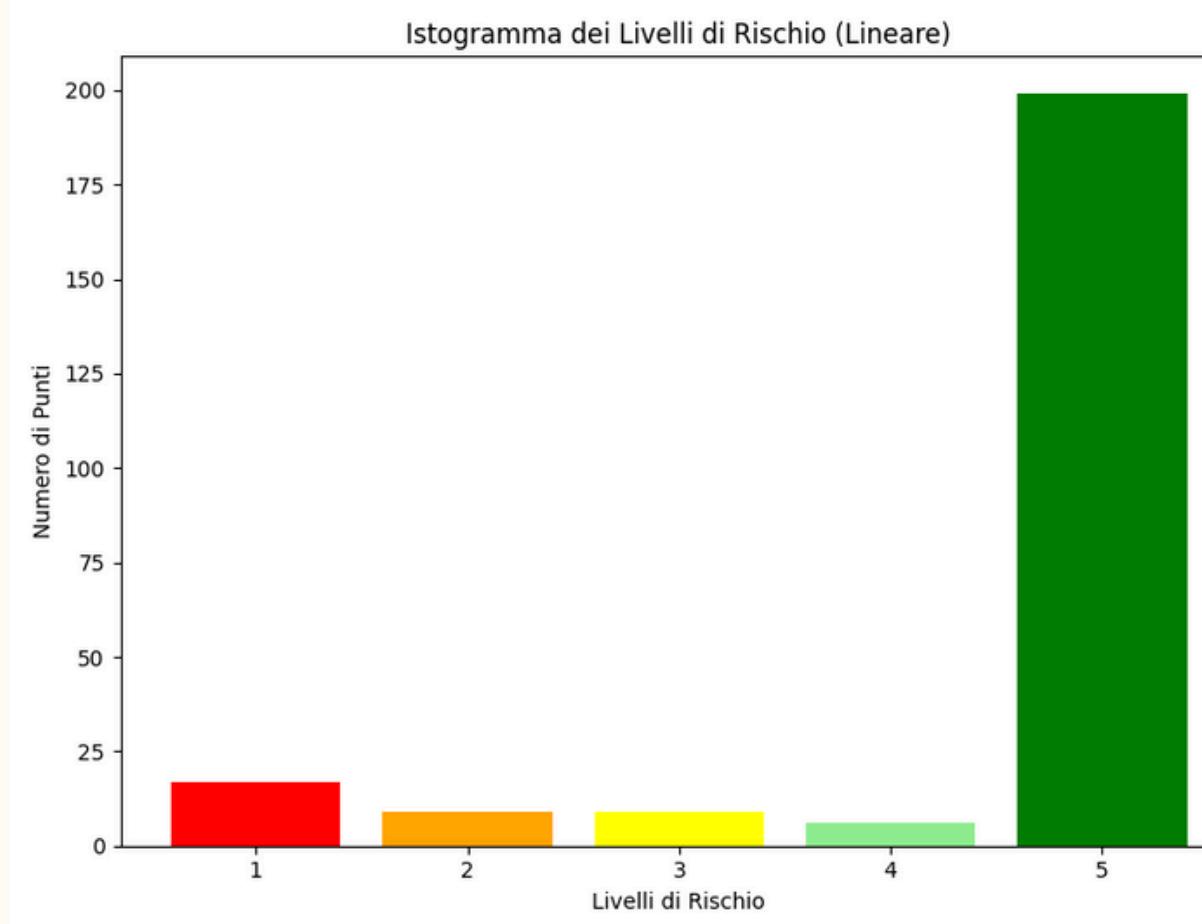
- Maggiore accuratezza nella separazione dei dati
- Maggiore costo computazionale rispetto alla versione lineare

Analisi dei Cambiamenti nella Classificazione

- Punti invariati (grigi) → Classificazione identica tra i due modelli.
- Punti con rischio cambiato (rossi) → Identificati più chiaramente dal kernel polinomiale.



Sintesi dei Risultati



Il modello SVM lineare offre una separazione più semplice e interpretabile, risultando più efficiente computazionalmente.

Il modello SVM polinomiale, invece, cattura meglio le relazioni non lineari, permettendo una maggiore capacità discriminativa nei livelli di rischio.

References

- L. Mauri, B. Apolloni, E. Damiani.
Robust ML model ensembles via
risk-driven anti-clustering of
training data, *Information Sciences*,
vol. 633, pp. 122–140, 2023