Seoul Bike Sharing Demand

Machine Learning - Progetto

Autori
Kristian Kovacev - matricola 885839
Paolo Mascheroni - matricola: 886220

Introduzione

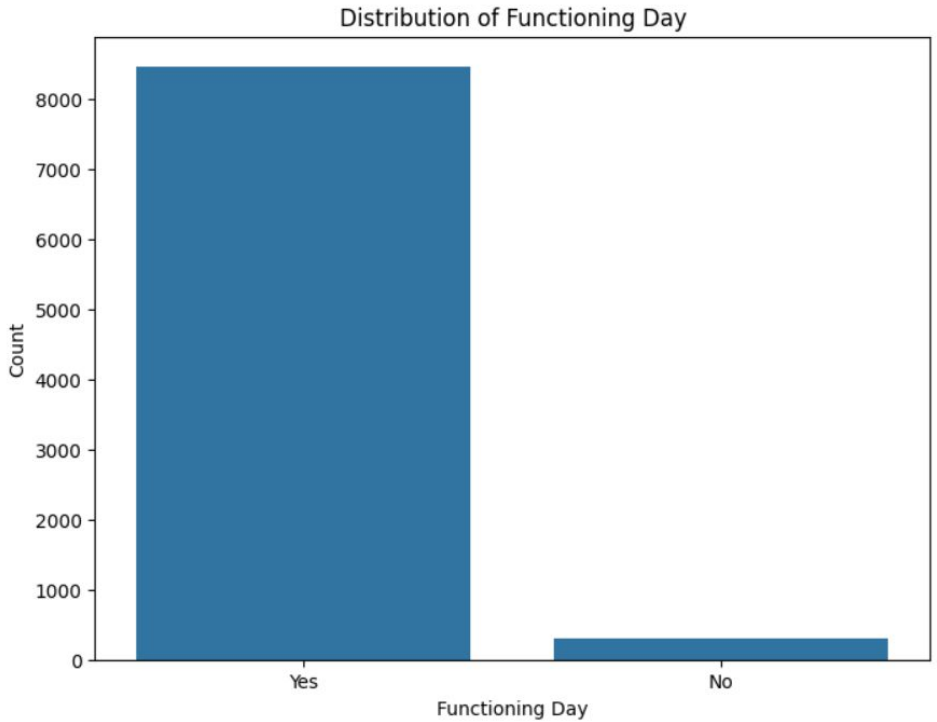
- Dataset: Seoul Bike Sharing Demand
- Piattaforma: UCI Machine Learning Repository
- 8760 istanze
- 13 features, 1 target

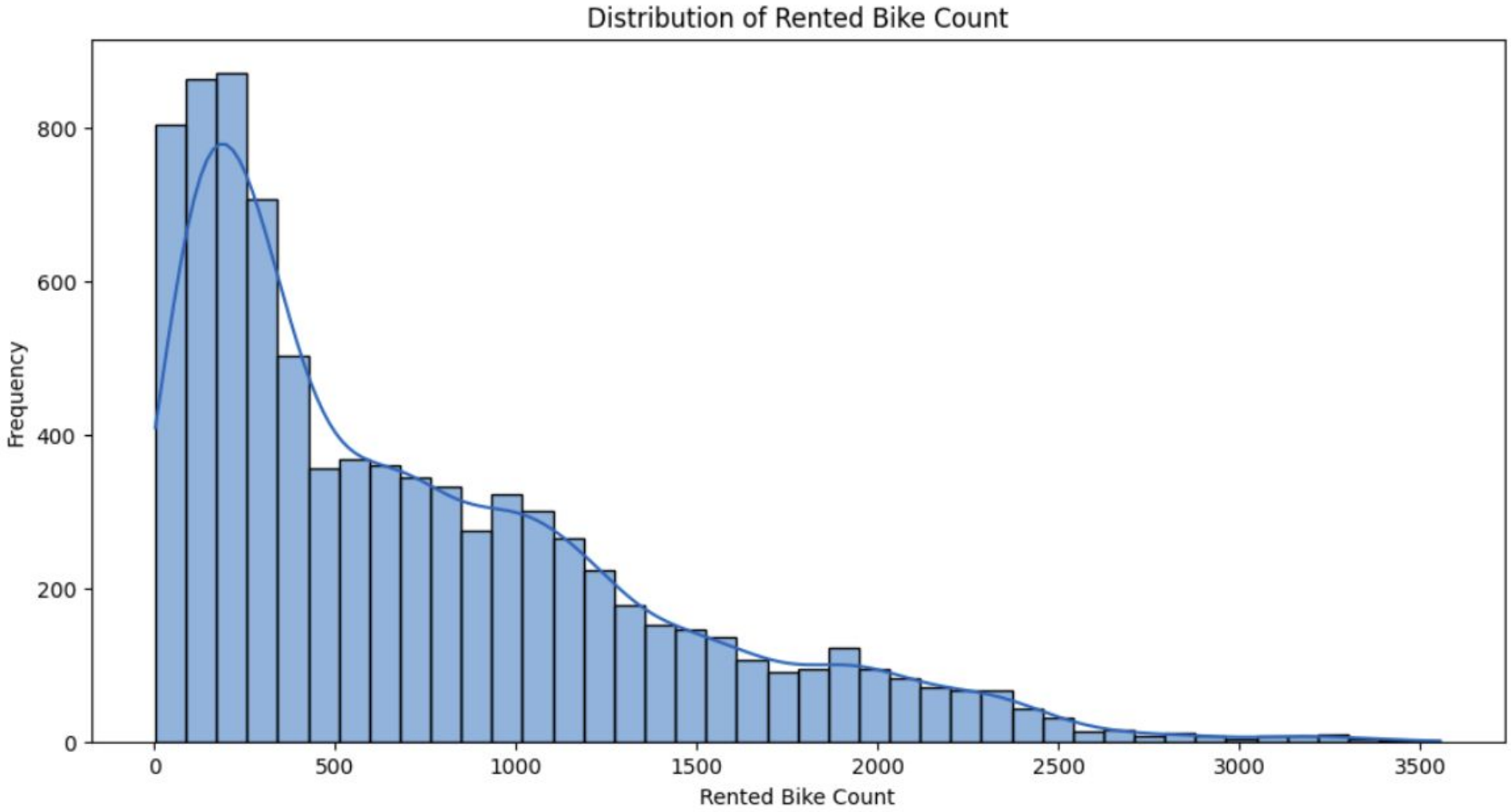
Analisi esplorativa del dataset

Nome Attributo	Tipo	Descrizione
Date	Date	Data in cui è stata effettuata la rilevazione
Hour	Integer	Ora della giornata in cui è stata effettuata la rilevazione
Rented Bike Count	Integer	Numero di bici noleggiate in quell'ora
Temperature	Continuous	Temperatura dell'aria in C°
Humidity	Integer	Umidità dell'aria in percentuale
Wind Speed	Continuous	Velocità del vento in m/s
Visibility	Integer	Indice di visibilità ai 10 metri

Nome Attributo	Tipo	Descrizione
Dew point temperature	Continuous	Punto di rugiada dell'aria (in C°)
Solar Radiation	Continuous	Radiazione solare in Mj/m2
Rainfall	Integer	Quantità di pioggia in mm
Snowfall	Integer	Quantità di neve in cm
Seasons	Categorical	stagione tra [Primavera, Estate, Autunno, Inverno]
Holiday	Binary	'No Holiday' se il giorno non è festivo 'Holiday' se il giorno è festivo

Variabile target




Distribution of Rented Bike Count
Frequency
0
200
400
600
800
0
500
1000
1500
2000
2500
3000
3500
Rented Bike Count

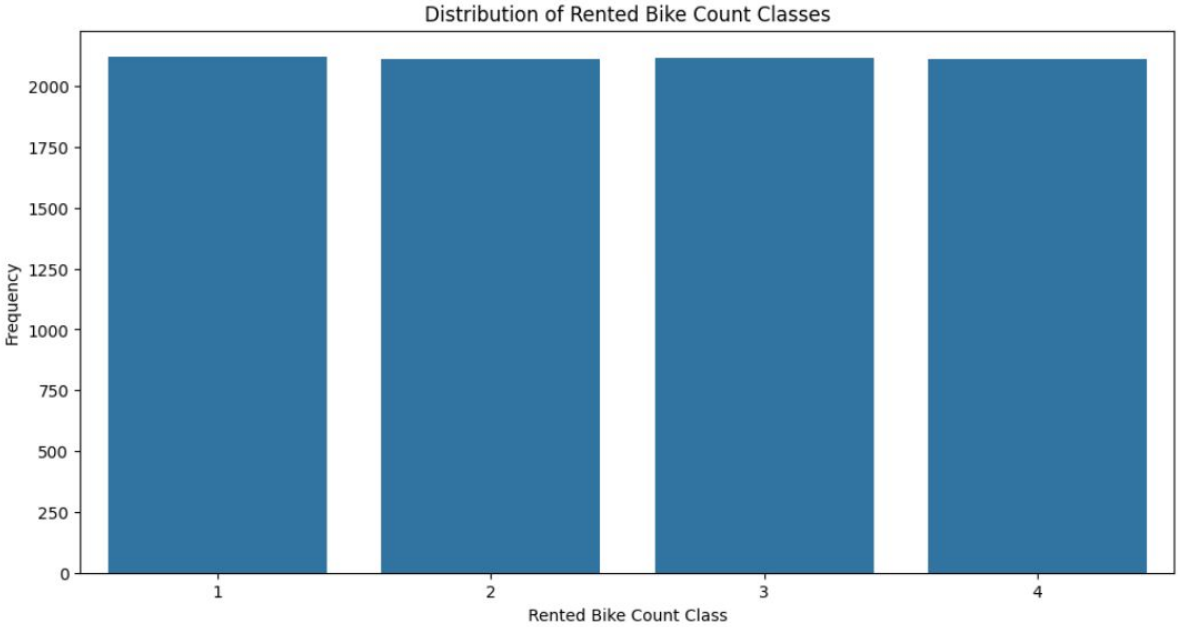
Nuova variabile target: Rented Bike Count

Quindi la nuova variabile target **Rented_Bike_Count_Class** è così definita:

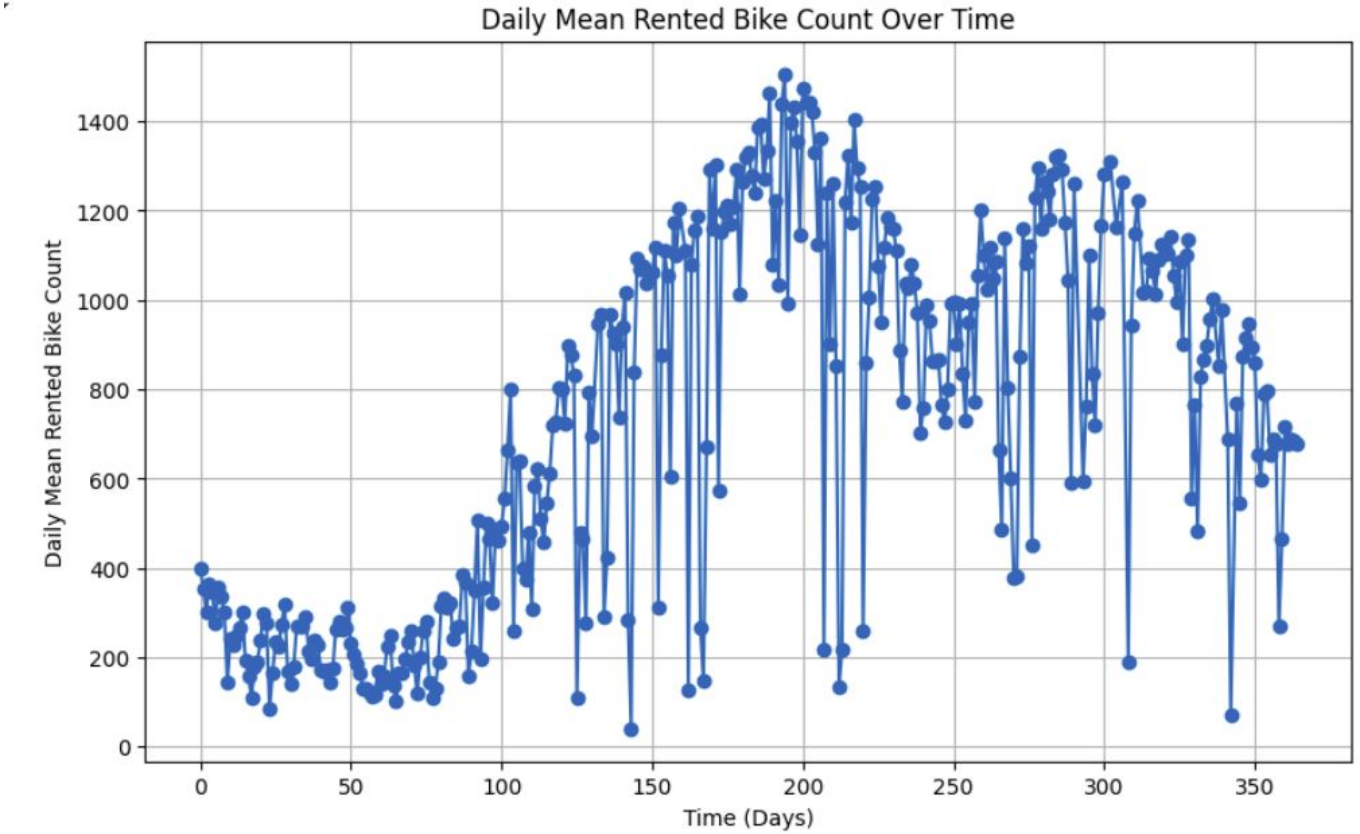
- 1 se Rented Bike Count è in [0,214)
- 2 se Rented Bike Count è in [214, 542)
- 3 se Rented Bike Count è in [542, 1084)
- 4 se Rented Bike Count è in [1084, 3556]

	Rented Bike Count
count	8465.000000
mean	729.156999
std	642.351166
min	2.000000
25%	214.000000
50%	542.000000
75%	1084.000000
max	3556.000000

Distribuzione nuova variabile target



Eliminazione della data



Encoding delle feature categoriche e intere

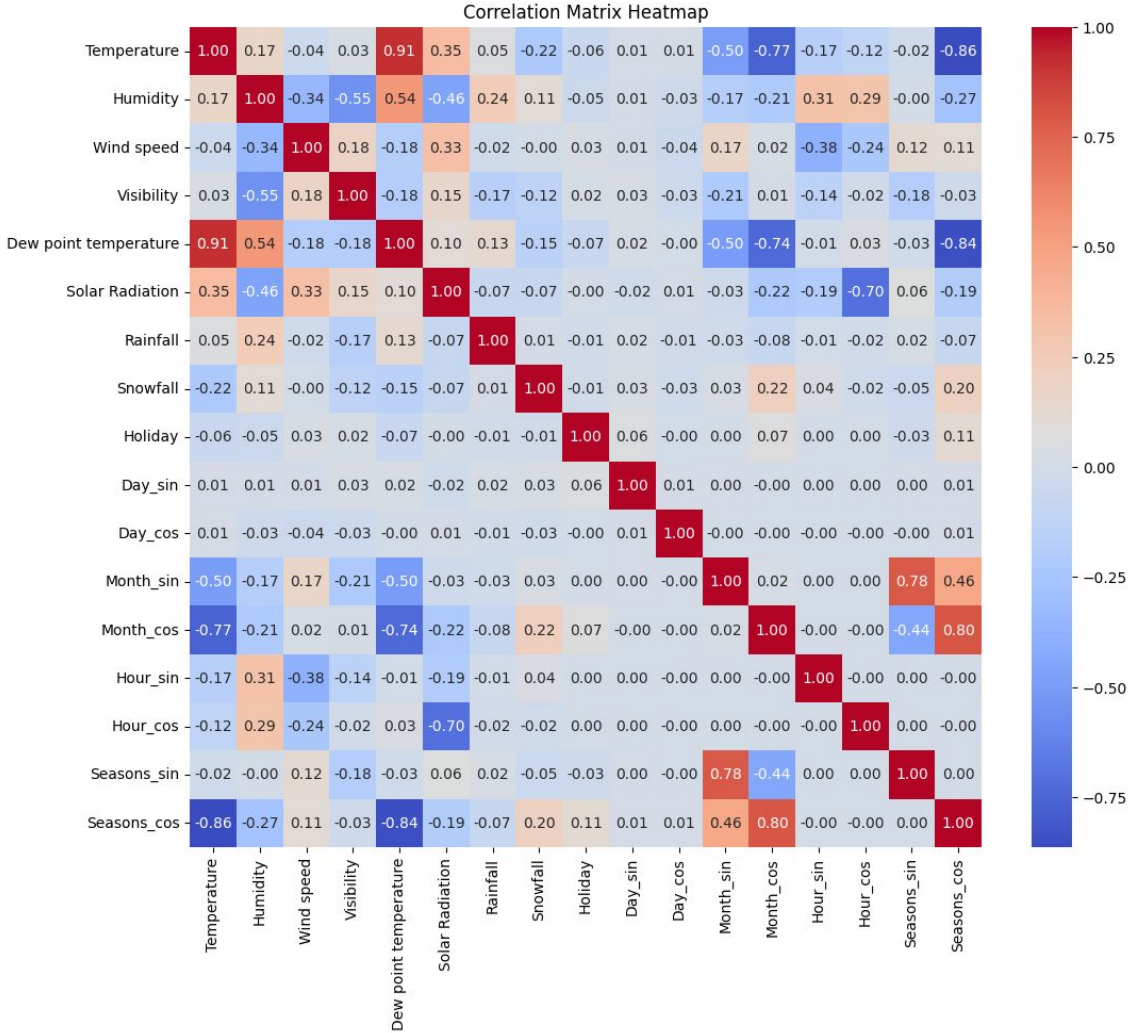
- **Holiday** → 0 se il valore è ‘No Holiday’
  1 se il valore è ‘Holiday’

- **Day, Month, Hour, Seasons** → Encoding ciclico
  Utilizziamo un encoding ciclico sostituendo ognuna delle feature sopracitate con due nuove feature, una componente seno e una coseno, secondo la formula:

$$x_{sin} = sin(\frac{2\pi x}{P})$$

$$x_{cos} = cos(\frac{2\pi x}{P})$$

Dove P è il periodo di oscillazione (24 per le ore, 12 per i mesi, 7 per i giorni della settimana, 4 per le stagioni).


Correlation Matrix Heatmap
Temperature 1.00 0.17 -0.04 0.03 0.91 0.35 0.05 -0.22 -0.06 0.01 0.01 -0.50 -0.77 -0.17 -0.12 -0.02 -0.86
Humidity 0.17 1.00 -0.34 -0.55 0.54 -0.46 0.24 0.11 -0.05 0.01 -0.03 -0.17 -0.21 0.31 0.29 -0.00 -0.27
Wind speed -0.04 -0.34 1.00 0.18 -0.18 0.33 -0.02 -0.00 0.03 0.01 -0.04 0.17 0.02 -0.38 -0.24 0.12 0.11
Visibility 0.03 -0.55 0.18 1.00 -0.18 0.15 -0.17 -0.12 0.02 0.03 -0.03 -0.21 0.01 -0.14 -0.02 -0.18 -0.03
Dew point temperature 0.91 0.54 -0.18 -0.18 1.00 0.10 0.13 -0.15 -0.07 0.02 -0.00 -0.50 -0.74 -0.01 0.03 -0.03 -0.84
Solar Radiation 0.35 -0.46 0.33 0.15 0.10 1.00 -0.07 -0.07 -0.00 -0.02 0.01 -0.03 -0.22 -0.19 -0.70 0.06 -0.19
Rainfall 0.05 0.24 -0.02 -0.17 0.13 -0.07 1.00 0.01 -0.01 0.02 -0.01 -0.03 -0.08 -0.01 -0.02 0.02 -0.07
Snowfall -0.22 0.11 -0.00 -0.12 -0.15 -0.07 0.01 1.00 -0.01 0.03 -0.03 0.03 0.22 0.04 -0.02 -0.05 0.20
Holiday -0.06 -0.05 0.03 0.02 -0.07 -0.00 -0.01 -0.01 1.00 0.06 -0.00 0.00 0.07 0.00 0.00 -0.03 0.11
Day_sin 0.01 0.01 0.01 0.03 0.02 -0.02 0.02 0.03 0.06 1.00 0.01 0.00 -0.00 0.00 0.00 0.00 0.01
Day_cos 0.01 -0.03 -0.04 -0.03 -0.00 0.01 -0.01 -0.03 -0.00 0.01 1.00 -0.00 -0.00 -0.00 -0.00 -0.00 0.01
Month_sin -0.50 -0.17 0.17 -0.21 -0.50 -0.03 -0.03 0.03 0.00 0.00 -0.00 1.00 0.02 0.00 0.00 0.78 0.46
Month_cos -0.77 -0.21 0.02 0.01 -0.74 -0.22 -0.08 0.22 0.07 -0.00 -0.00 0.02 1.00 -0.00 -0.00 -0.44 0.80
Hour_sin -0.17 0.31 -0.38 -0.14 -0.01 -0.19 -0.01 0.04 0.00 0.00 -0.00 0.00 -0.00 1.00 -0.00 0.00 -0.00
Hour_cos -0.12 0.29 -0.24 -0.02 0.03 -0.70 -0.02 -0.02 0.00 0.00 -0.00 0.00 -0.00 -0.00 1.00 0.00 -0.00
Seasons_sin -0.02 -0.00 0.12 -0.18 -0.03 0.06 0.02 -0.05 -0.03 0.00 -0.00 0.78 -0.44 0.00 0.00 1.00 0.00
Seasons_cos -0.86 -0.27 0.11 -0.03 -0.84 -0.19 -0.07 0.20 0.11 0.01 0.01 0.46 0.80 -0.00 -0.00 0.00 1.00
Temperature
Humidity
Wind speed
Visibility
Dew point temperature
Solar Radiation
Rainfall
Snowfall
Holiday
Day_sin
Day_cos
Month_sin
Month_cos
Hour_sin
Hour_cos
Seasons_sin
Seasons_cos
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Dataset dopo preprocessing

Nome Attributo	Tipo	Descrizione
Humidity	Continuous	Umidità dell'aria in percentuale
Wind Speed	Continuous	Velocità del vento in m/s
Visibility	Continuous	Indice di visibilità ai 10 metri
Dew point temperature	Continuous	Punto di rugiada dell'aria (in C°)
Solar Radiation	Continuous	Radiazione solare in Mj/m2
Rainfall	Continuous	Quantità di pioggia in mm
Snowfall	Continuous	Quantità di neve in cm

Nome Attributo	Tipo	Descrizione
Holiday	Binary	0 se il giorno non è festivo 1 se il giorno è festivo
Day_sin	Continuous	Componente seno della variabile Day
Day_cos	Continuous	Componente coseno della variabile Day
Hour_sin	Continuous	Componente seno della variabile Hour
Hour_cos	Continuous	Componente coseno della variabile Hour
Month_sin	Continuous	Componente seno della variabile Month
Month_cos	Continuous	Componente coseno della variabile Month

Split del dataset

60%	20%	20%
TRAIN	VALIDATION	TEST

SVM - Ottimizzazione iperparametri

Grid search per trovare migliore combinazione di iperparametri sul validation set

Kernel

Ricerca kernel tra Linear, RBF e Poly.

Il miglior kernel selezionato in base alla validazione è stato **RBF**.

C e Gamma

Scelto il kernel RBF abbiamo fatto una ricerca tra:

- C: 10, 25, 50, 75, 100, 120
- Gamma: 0.2, 0.1, 0.05, 0.01, 0.005, 0.001

Ottenendo come valori ottimali: C = 100 e gamma = 0.05

Infine, **k-fold cross validation** su pochi valori vicini ai migliori trovati → C = 105 e gamma = 0.05

Reti Neurali

- Early stopping con validation set → Evita overfitting
- Dropout sui layer → migliora generalizzazione del modello

Ottimizzazione dei parametri:

- Costo computazionale alto → impossibile provare una grid search
- Diversi valori provati per gli iperparametri:

  (funzione di attivazione, numero layers, dropout, batch_size…)

- media delle metriche su 3 addestramenti per confrontare i modelli

  Miglior modello → 3 layer nascosti, funzione Relu, dropout 0.2, learning rate 0.001. batch_size 64

Risultati dei modelli

Modello	Accuracy	Precision	Recall	F1-Score
SVM	0.8240	0.8227	0.8212	0.8221
Reti Neurali	0.8406	0.8400	0.8406	0.8400

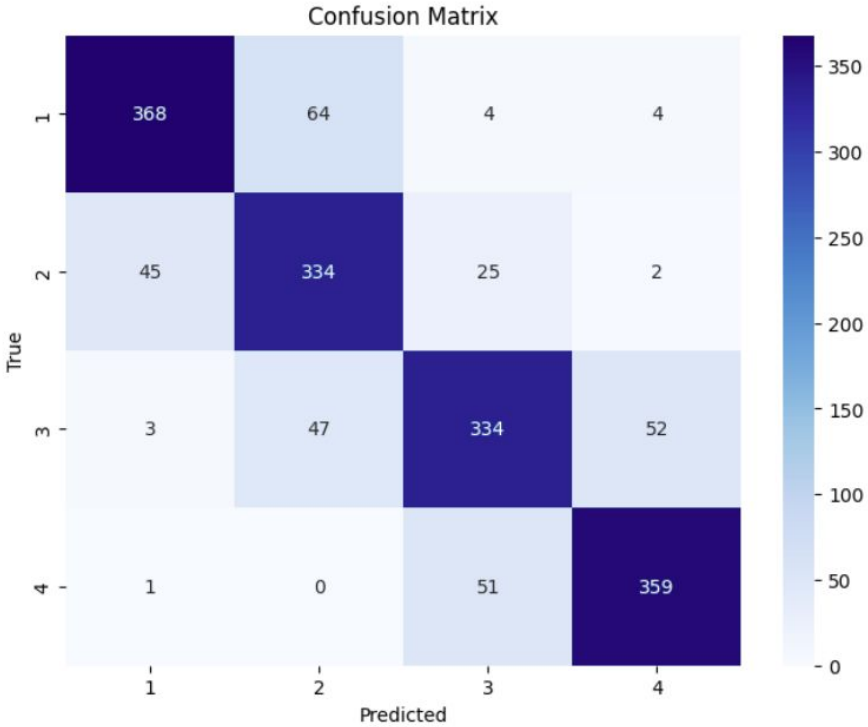
Classe	Precision	Recall	F1-Score
1	0.8832	0.8421	0.8641
2	0.75223	0.8232	0.7817
3	0.81179	0.7736	0.7918
4	0.8614	0.8719	0.8711

SVM

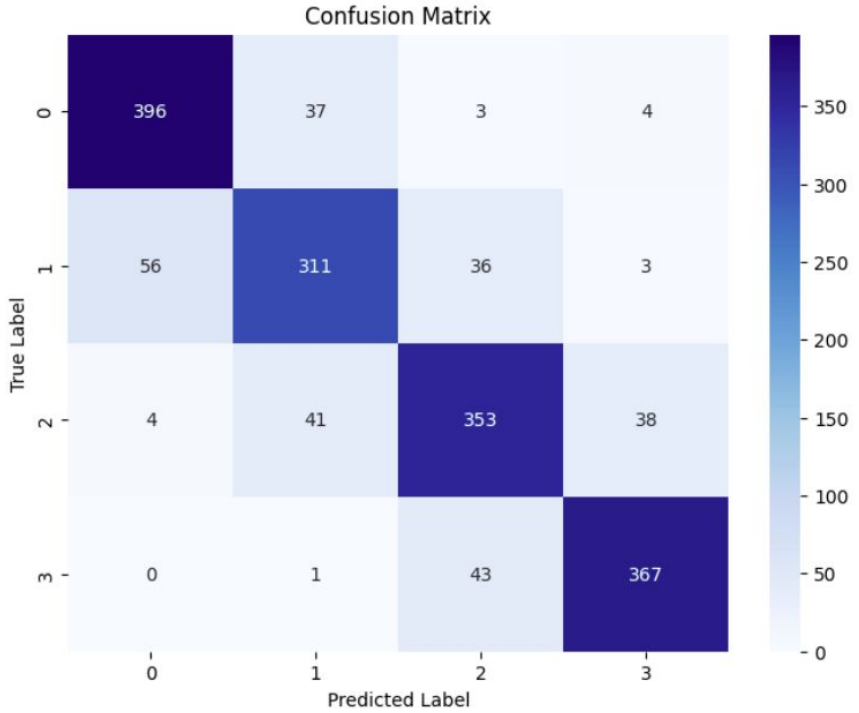
Classe	Precision	Recall	F1-Score
1	0.8906	0.8805	0.8853
2	0.7855	0.7926	0.7889
3	0.8145	0.7876	0.8008
4	0.8695	0.9017	0.8851

RETE NEURALE

Confusion Matrix



SVM



RETE NEURALE