

# Advanced Machine Learning Project - Un aiuto per Airbnb

Approcci di Machine Learning per  
prevedere la destinazione dei  
nuovi iscritti statunitensi

Dario Carolla - 807547  
Matteo Gaverini - 808101  
Paolo Mariani - 800307





# OVERVIEW

**Airbnb**



Piattaforma  
digitale per  
affittare camere

**Obiettivo**



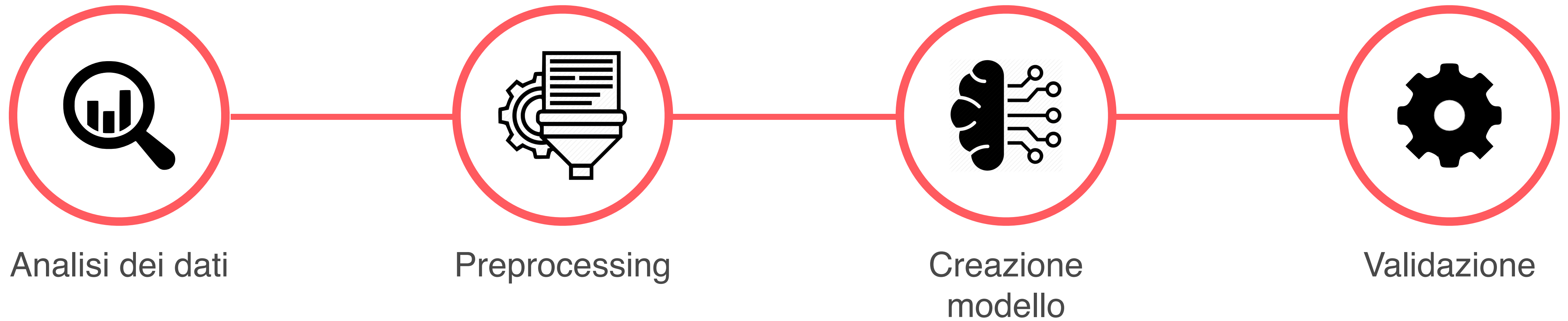
Determinare la  
destinazione del  
viaggio dei nuovi  
clienti

**Dati**



Informazioni sugli  
utenti e log di  
sessione

# WORKFLOW



# ANALISI DATI

DATI FORNITI DA AIRBNB



## **train/test\_user.csv**

Train: 213451 osservazioni - 16 variabili

Test: 62096 osservazioni - 14 variabili

Attributi descrittivi riguardanti gli utenti che hanno effettuato la registrazione alla piattaforma (età, sesso, tipo dispositivo utilizzato etc.).

Intervallo temporale: 01-01-2010 —> 30-09-2014

countries.csv

- Date di iscrizione e primo utilizzo
- Dati anagrafici
- Dispositivi utilizzati
- Canale di contatto
- Variabile target

sessions.csv

age\_gender\_bkts.csv

# ANALISI DATI

DATI FORNITI DA AIRBNB



## **train/test\_user.csv**

Train: 213451 osservazioni - 16 variabili

Test: 62096 osservazioni - 14 variabili

Attributi descrittivi riguardanti gli utenti che hanno effettuato la registrazione alla piattaforma (età, sesso, tipo dispositivo utilizzato etc.).

Intervallo temporale: 01-01-2010 —> 30-09-2014

countries.csv

- **Date di iscrizione e primo utilizzo**
- Dati anagrafici
- Dispositivi utilizzati
- Canale di contatto
- Variabile target

sessions.csv

age\_gender\_bkts.csv

# ANALISI DATI

DATI FORNITI DA AIRBNB



## train/test\_user.csv

Train: 213451 osservazioni - 16 variabili

Test: 62096 osservazioni - 14 variabili

Attributi descrittivi riguardanti gli utenti che hanno effettuato la registrazione alla piattaforma (età, sesso, tipo dispositivo utilizzato etc.).

Intervallo temporale: 01-01-2010 —> 30-09-2014

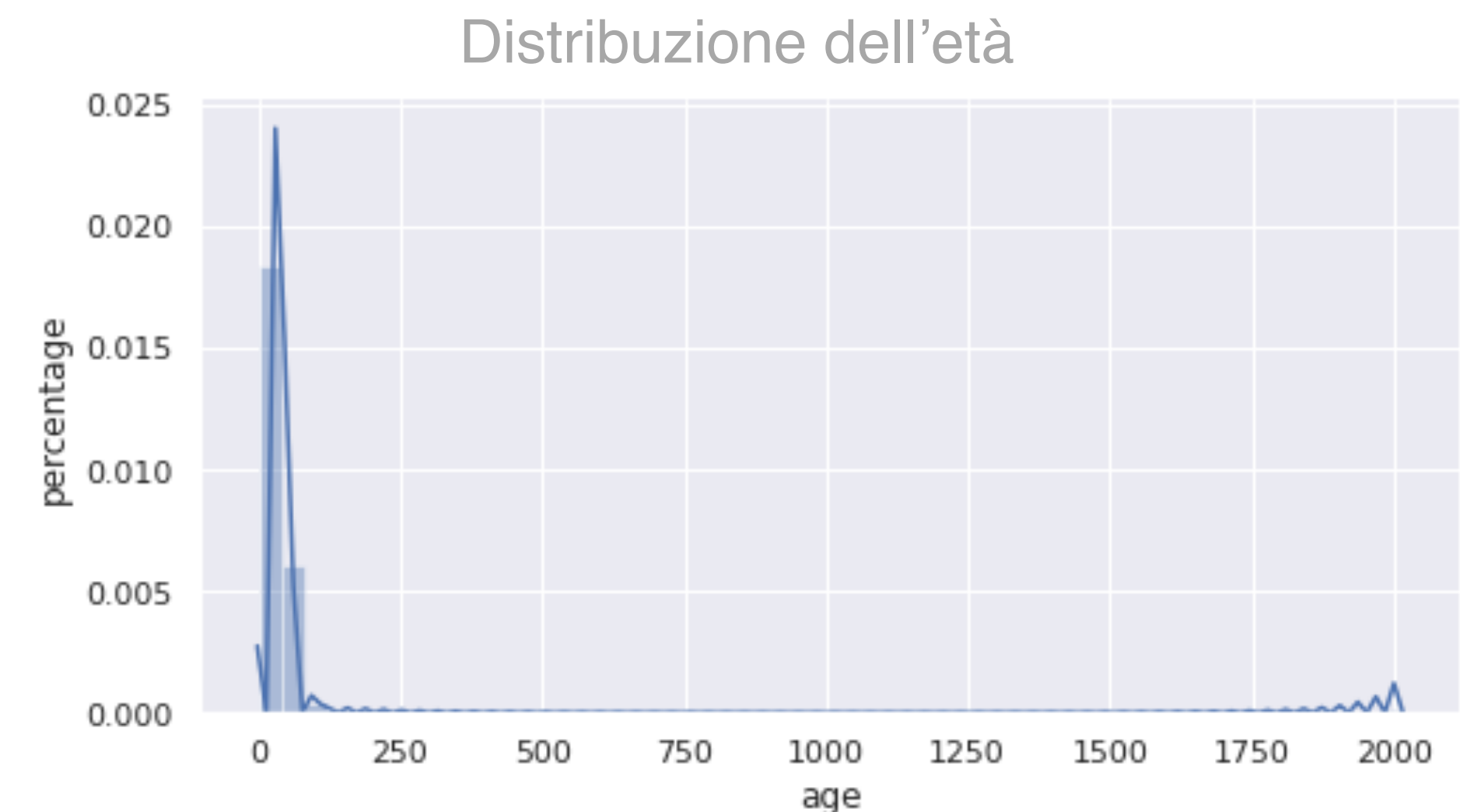
countries.csv

- Date di iscrizione e primo utilizzo
- **Dati anagrafici**
- Dispositivi utilizzati
- Canale di contatto
- Variabile target

sessions.csv

age\_gender\_bkts.csv

% NA 'age': 42.41%



# ANALISI DATI

DATI FORNITI DA AIRBNB



## **train/test\_user.csv**

Train: 213451 osservazioni - 16 variabili

Test: 62096 osservazioni - 14 variabili

Attributi descrittivi riguardanti gli utenti che hanno effettuato la registrazione alla piattaforma (età, sesso, tipo dispositivo utilizzato etc.).

Intervallo temporale: 01-01-2010 —> 30-09-2014

countries.csv

- Date di iscrizione e primo utilizzo
- Dati anagrafici
- **Dispositivi utilizzati**
- Canale di contatto
- Variabile target

sessions.csv

age\_gender\_bkts.csv

# ANALISI DATI

DATI FORNITI DA AIRBNB



## train/test\_user.csv

Train: 213451 osservazioni - 16 variabili

Test: 62096 osservazioni - 14 variabili

Attributi descrittivi riguardanti gli utenti che hanno effettuato la registrazione alla piattaforma (età, sesso, tipo dispositivo utilizzato etc.).

Intervallo temporale: 01-01-2010 —> 30-09-2014

countries.csv

- Date di iscrizione e primo utilizzo
- Dati anagrafici
- Dispositivi utilizzati
- **Canale di contatto**
- Variabile target

sessions.csv

% NA *'first\_affiliate\_tracked'*: 2.28 %

age\_gender\_bkts.csv



# ANALISI DATI

DATI FORNITI DA AIRBNB



## train/test\_user.csv

Train: 213451 osservazioni - 16 variabili

Test: 62096 osservazioni - 14 variabili

Attributi descrittivi riguardanti gli utenti che hanno effettuato la registrazione alla piattaforma (età, sesso, tipo dispositivo utilizzato etc.).

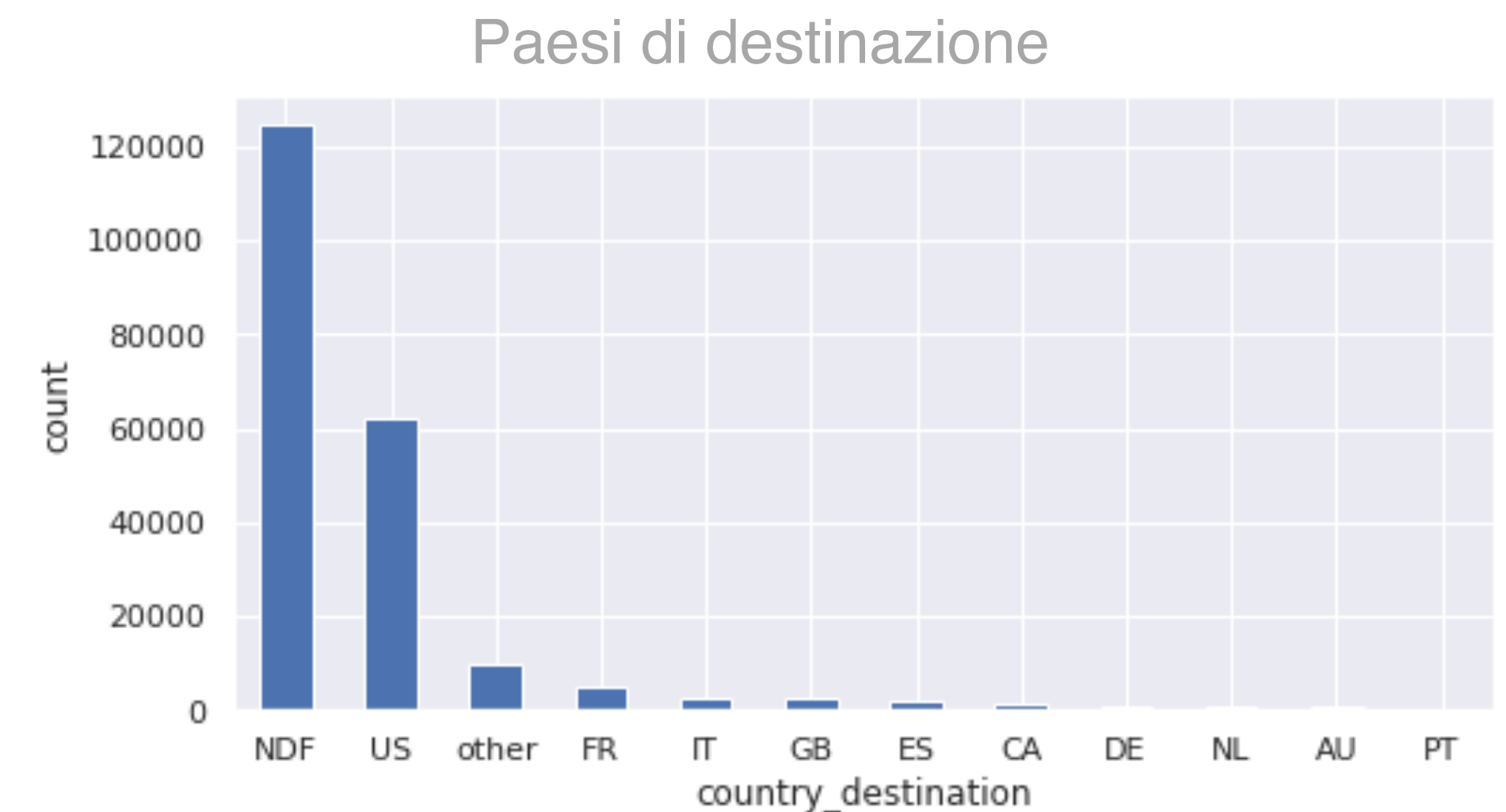
Intervallo temporale: 01-01-2010 —> 30-09-2014

countries.csv

- Date di iscrizione e primo utilizzo
- Dati anagrafici
- Dispositivi utilizzati
- Canale di contatto
- **Variabile target**

sessions.csv

age\_gender\_bkts.csv



# ANALISI DATI

DATI FORNITI DA AIRBNB

Paesi di destinazione



**train/test\_user.csv**

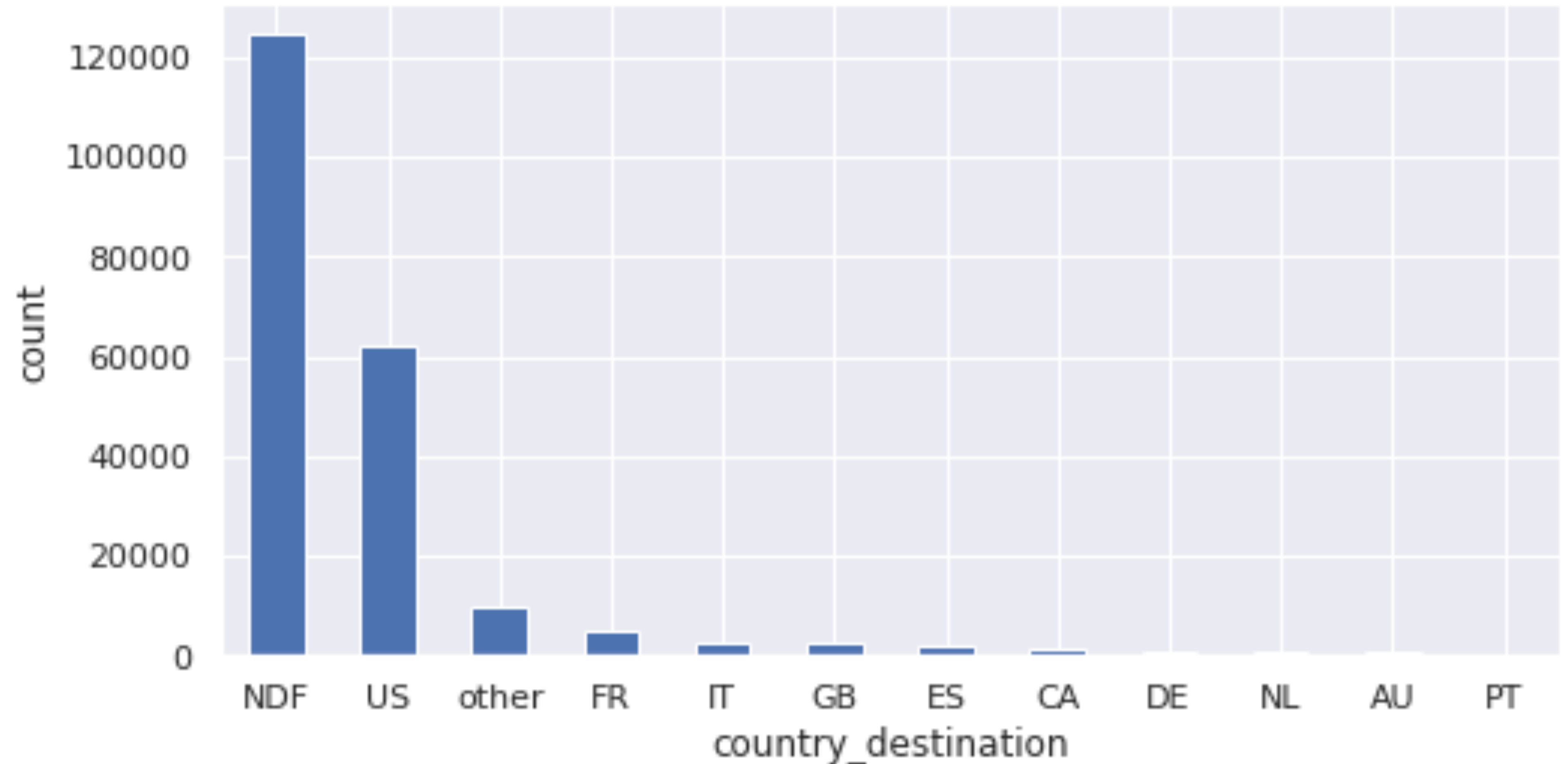
Train: 213451 osservazioni

Test: 62096 osservazioni

countries.csv

sessions.csv

age\_gender\_bkts





# ANALISI DATI

DATI FORNITI DA AIRBNB

train/test\_user.csv

Informazioni relative ai possibili paesi di destinazione:

- Posizione geografica
- Lingua ufficiale
- Superficie



**countries.csv**

11 osservazioni  
7 variabili

sessions.csv

age\_gender\_bkts.csv

# ANALISI DATI

DATI FORNITI DA AIRBNB

train/test\_user.csv

Attributi relativi ai log di utilizzo del sito da parte degli utenti registrati.

Intervallo temporale: 01-01-2014 —> 30-09-2014

countries.csv

- Tipologie di azioni svolte
- Dispositivo utilizzato
- Durata azione



**sessions.csv**

10567737 osservazioni  
6 variabili

age\_gender\_bkts.csv



# ANALISI DATI

DATI FORNITI DA AIRBNB

train/test\_user.csv

Attributi relativi ai log di utilizzo del sito da parte degli utenti registrati.

Intervallo temporale: 01-01-2014 —> 30-09-2014

countries.csv

- **Tipologie di azioni svolte**
- Dispositivo utilizzato
- Durata azione

% NA *'action'*: 0.7 %

% NA *'action\_type'*: 10.6 %



**sessions.csv**

10567737 osservazioni  
6 variabili

age\_gender\_bkts.csv

# ANALISI DATI

DATI FORNITI DA AIRBNB

train/test\_user.csv

Attributi relativi ai log di utilizzo del sito da parte degli utenti registrati.

Intervallo temporale: 01-01-2014 —> 30-09-2014

countries.csv

- Tipologie di azioni svolte
- **Dispositivo utilizzato**
- Durata azione



**sessions.csv**

10567737 osservazioni  
6 variabili

age\_gender\_bkts.csv



# ANALISI DATI

DATI FORNITI DA AIRBNB

train/test\_user.csv

Attributi relativi ai log di utilizzo del sito da parte degli utenti registrati.

Intervallo temporale: 01-01-2014 —> 30-09-2014

countries.csv

- Tipologie di azioni svolte
- Dispositivo utilizzato
- **Durata azione**

% NA '*secs\_elapsed*': 1.2 %



**sessions.csv**

10567737 osservazioni  
6 variabili

age\_gender\_bkts.csv

# ANALISI DATI

DATI FORNITI DA AIRBNB

train/test\_user.csv

Statistiche descrittive riguardo la popolazione di ogni destinazione suddivisa per range di età e genere.

countries.csv

sessions.csv



**age\_gender\_bkts.csv**

420 osservazioni

5 variabili



# PREPROCESSING

`train/test_user.csv`

Variable Trasformation →

Per le date estrazione della stagionalità tramite seno e coseno.

*'age'* : sostituzione dei valori corrispondenti all'anno di nascita con l'età al momento dell'iscrizione;  
sostituzione dei valori esterni all'intervallo [18,100] con valore costante *'-1'*.

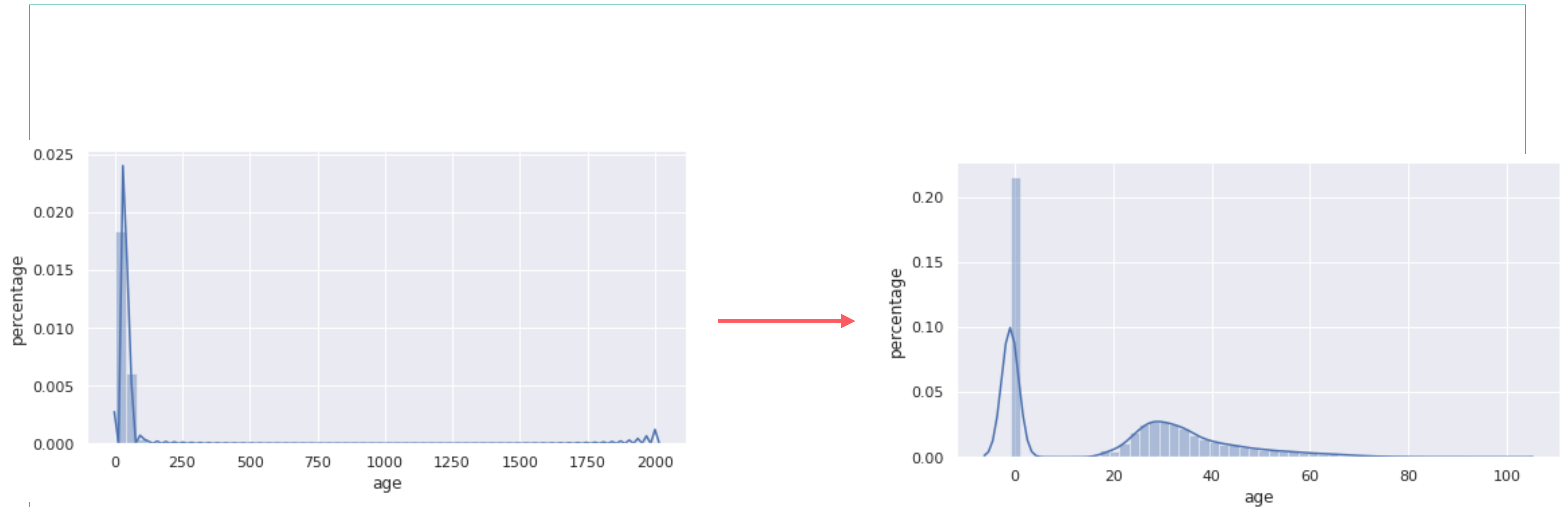
Missing Replacement →

*'age'* : sostituzione dei missing values con valore costante *'-1'*.

*'first\_affiliate\_tracked'* : sostituzione condizionata dei missing values rispetto al valore di *'affiliate\_channel'*.

# PREPROCESSING

train/test\_user.csv



# PREPROCESSING

`sessions.csv`

Missing replacement



*'action'* : sostituzione NA con stringa *'message'* poiché per tali osservazioni il valore di *'action\_type'* è sempre *'message\_post'*.

*'action\_type'*: sostituzione condizionata degli NA con la moda che l'attributo assume in relazione a specifici valori di *'action'*.

*'secs\_elapsed'*: sostituzione NA con mediana.

Variabili derivate



Il totale delle azioni effettuate, le azioni più frequenti ed il relativo conteggio, l'ultima azione effettuata in ordine cronologico.

Variabile binaria che indica se l'utente ha richiesto la traduzione di un contenuto.

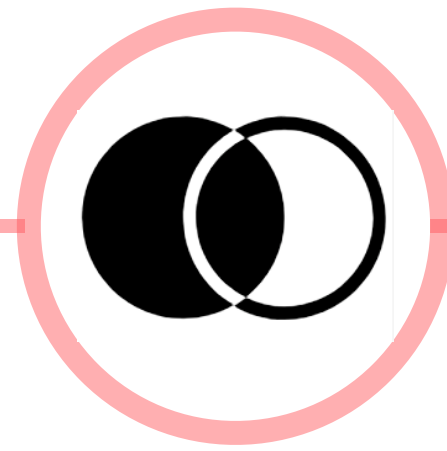
Totale, minimo, massimo, media, dev. std. della durata delle azioni svolte. Scarto tra il tempo totale delle azioni dell'utente e la media dell'intero dataset.



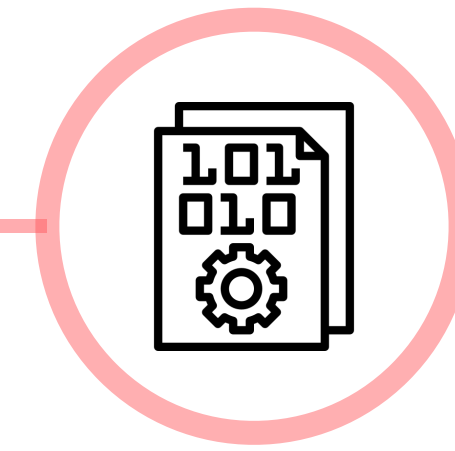
# PREPROCESSING



Analisi di  
correlazione



Unione dei  
Dataset

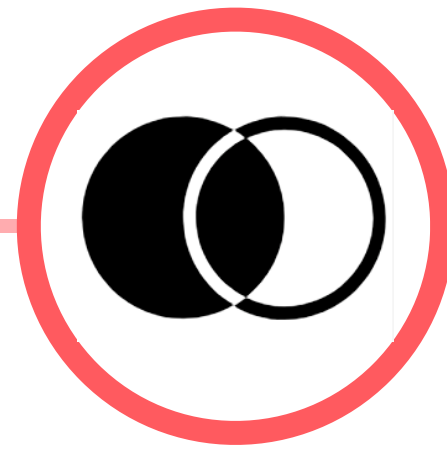


Codifica e  
normalizzazione

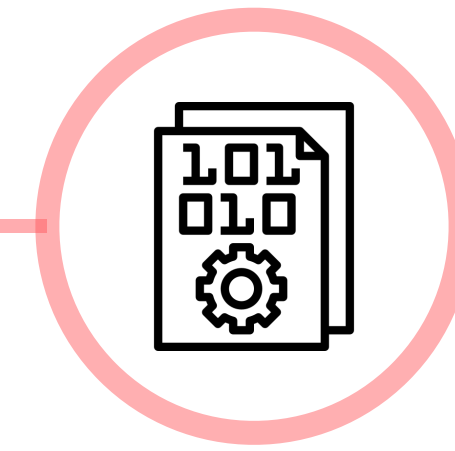
# PREPROCESSING



Analisi di  
correlazione



Unione dei  
Dataset



Codifica e  
normalizzazione

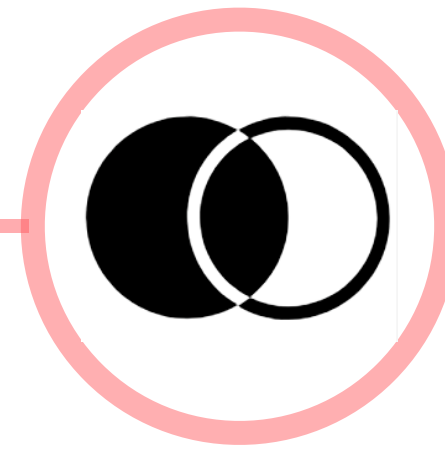
Left join tra *train/test\_user* e *sessions* su id dell'utente

Sostituzione NA generati con -1 per 140265 utenti

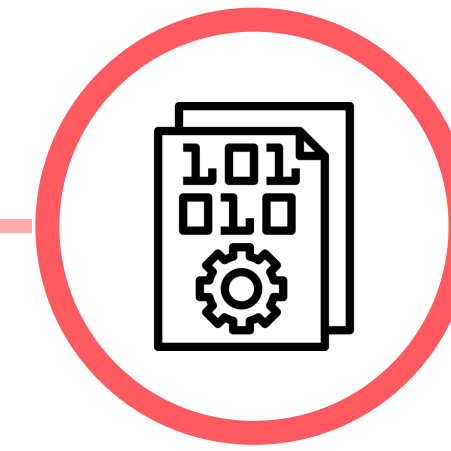
# PREPROCESSING



Analisi di  
correlazione



Unione dei  
Dataset



Codifica e  
normalizzazione

Feature categoriche:

- One-hot encoding per le variabili con meno di dieci possibili valori
- Label encoding per le variabili con più di dieci possibili valori

Feature numeriche:

- Min/max scaling

Dataset finale:

55 variabili

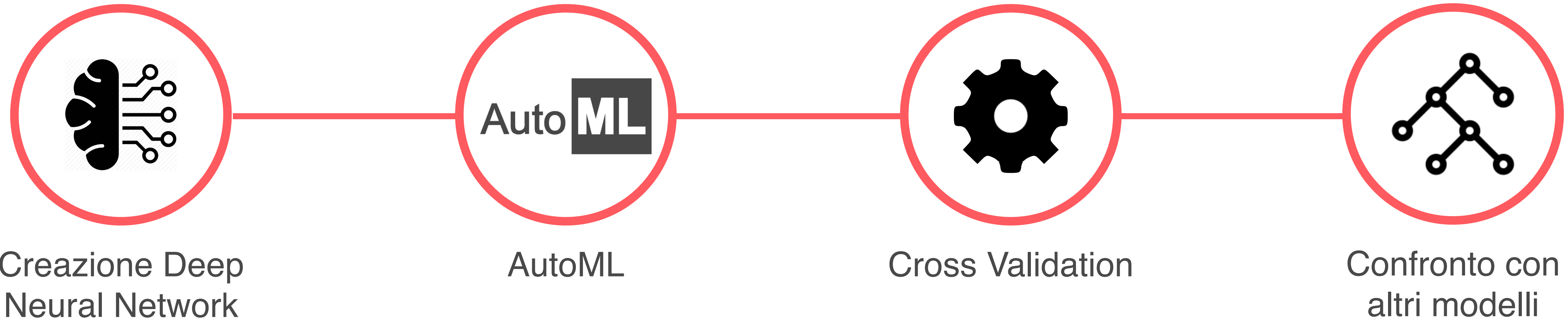
275547 osservazioni

Diviso in train e test (Kaggle)



# CREAZIONE DEL MODELLO

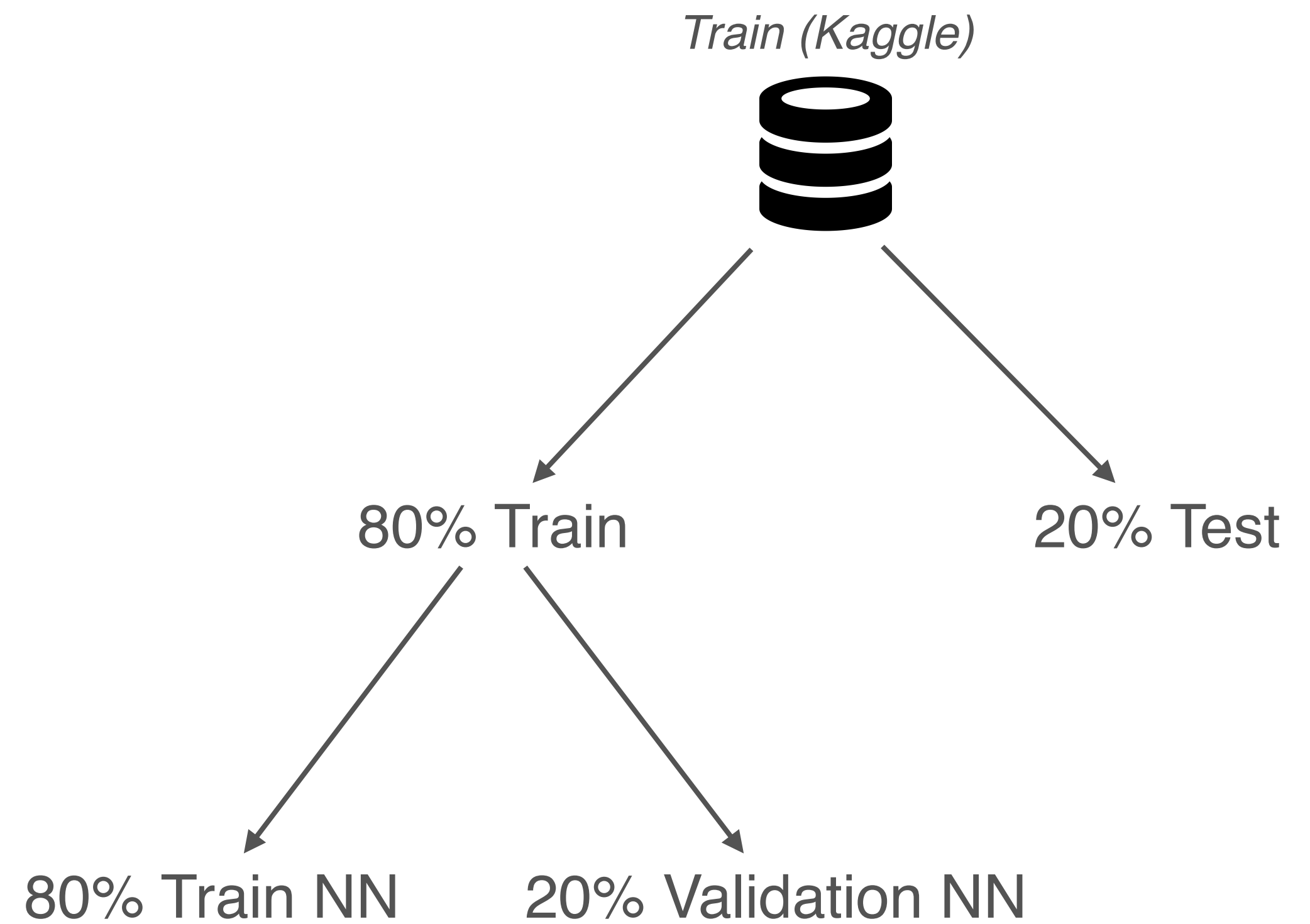
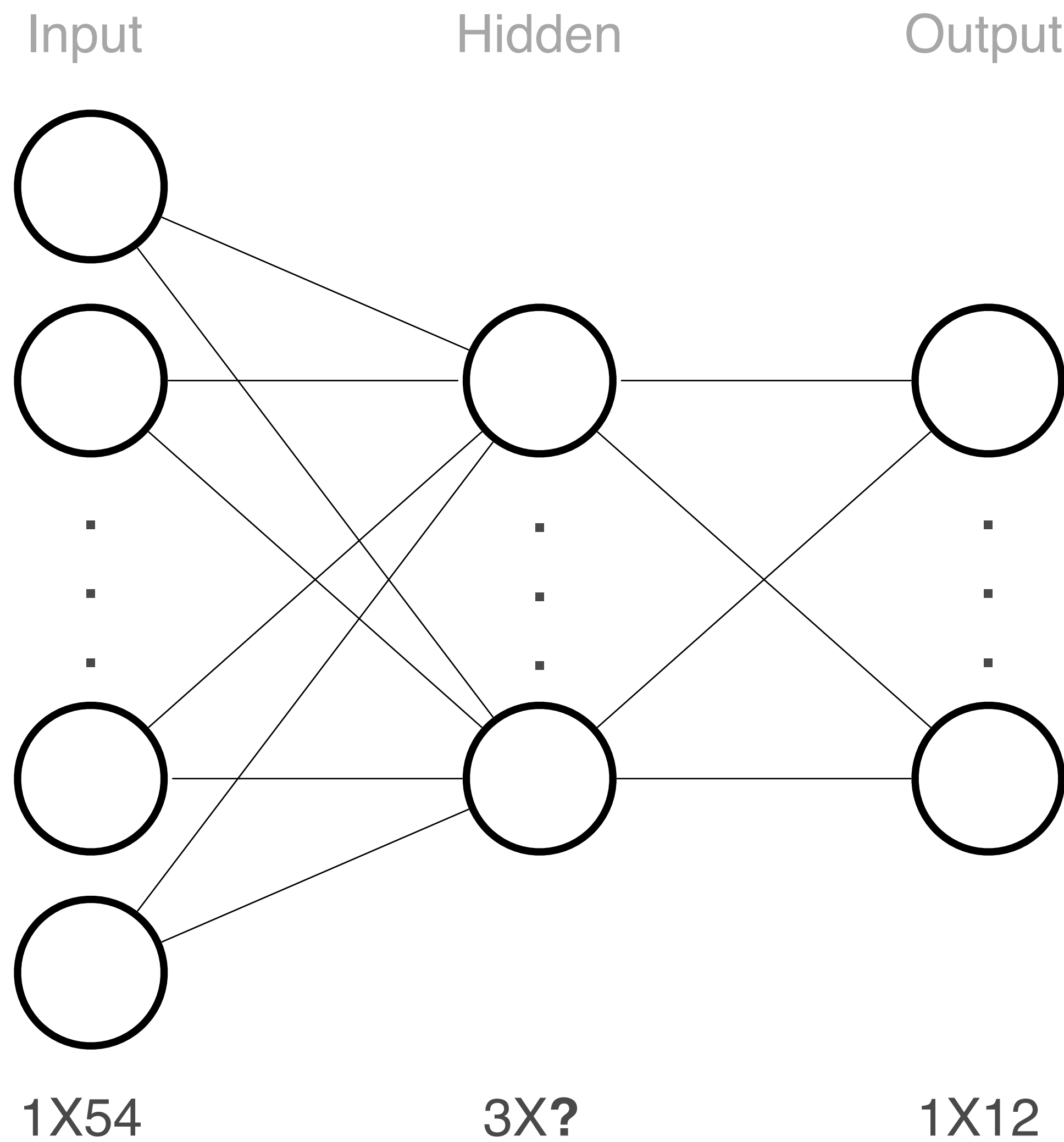
DEEP NEURAL NETWORK



Risorse utilizzate:  
Google Colaboratory - Macchina virtuale  
con 25 GB di RAM e 68 GB di Storage

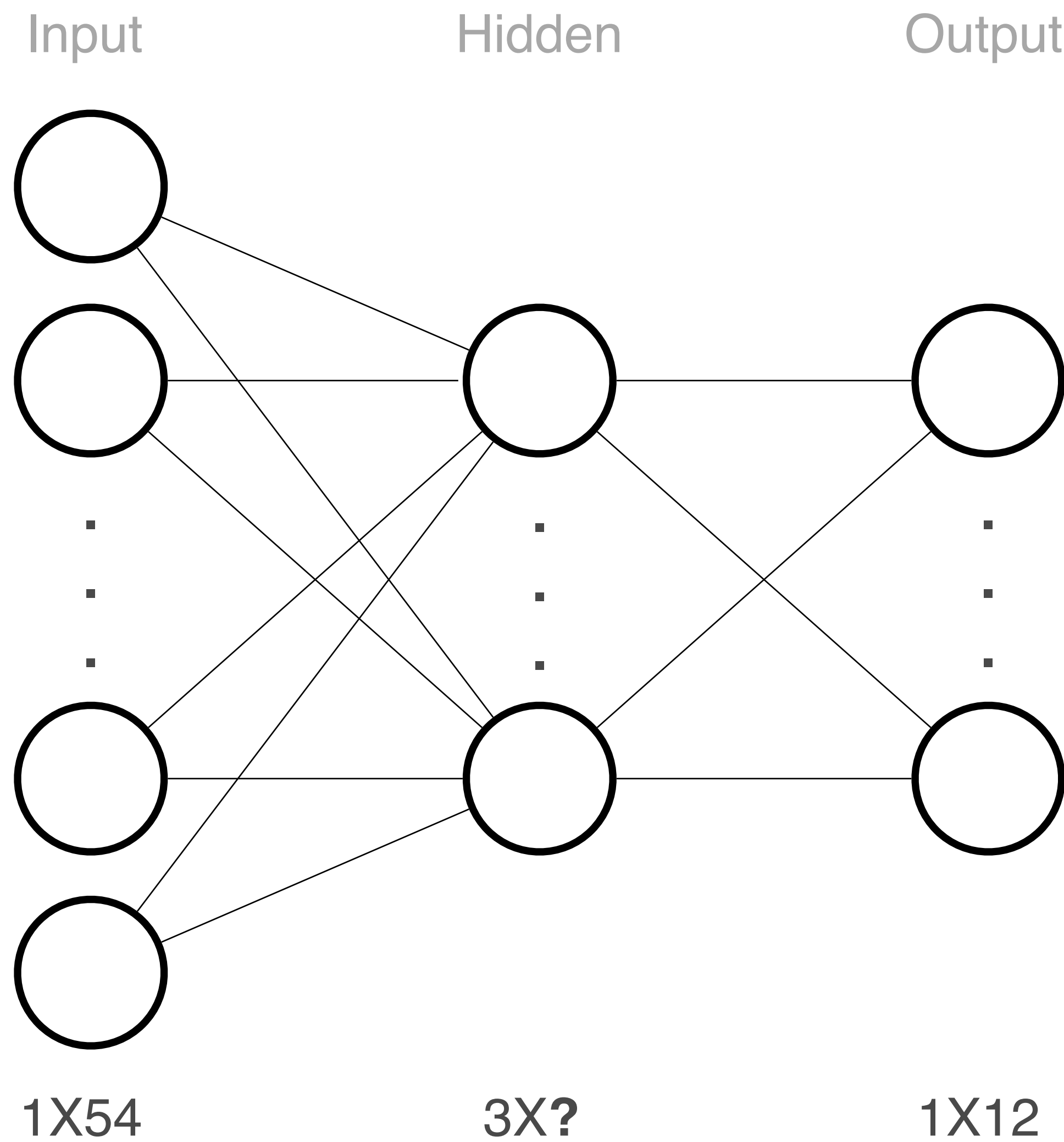
# STRUTTURA NN

DEEP NEURAL NETWORK



# STRUTTURA NN

DEEP NEURAL NETWORK

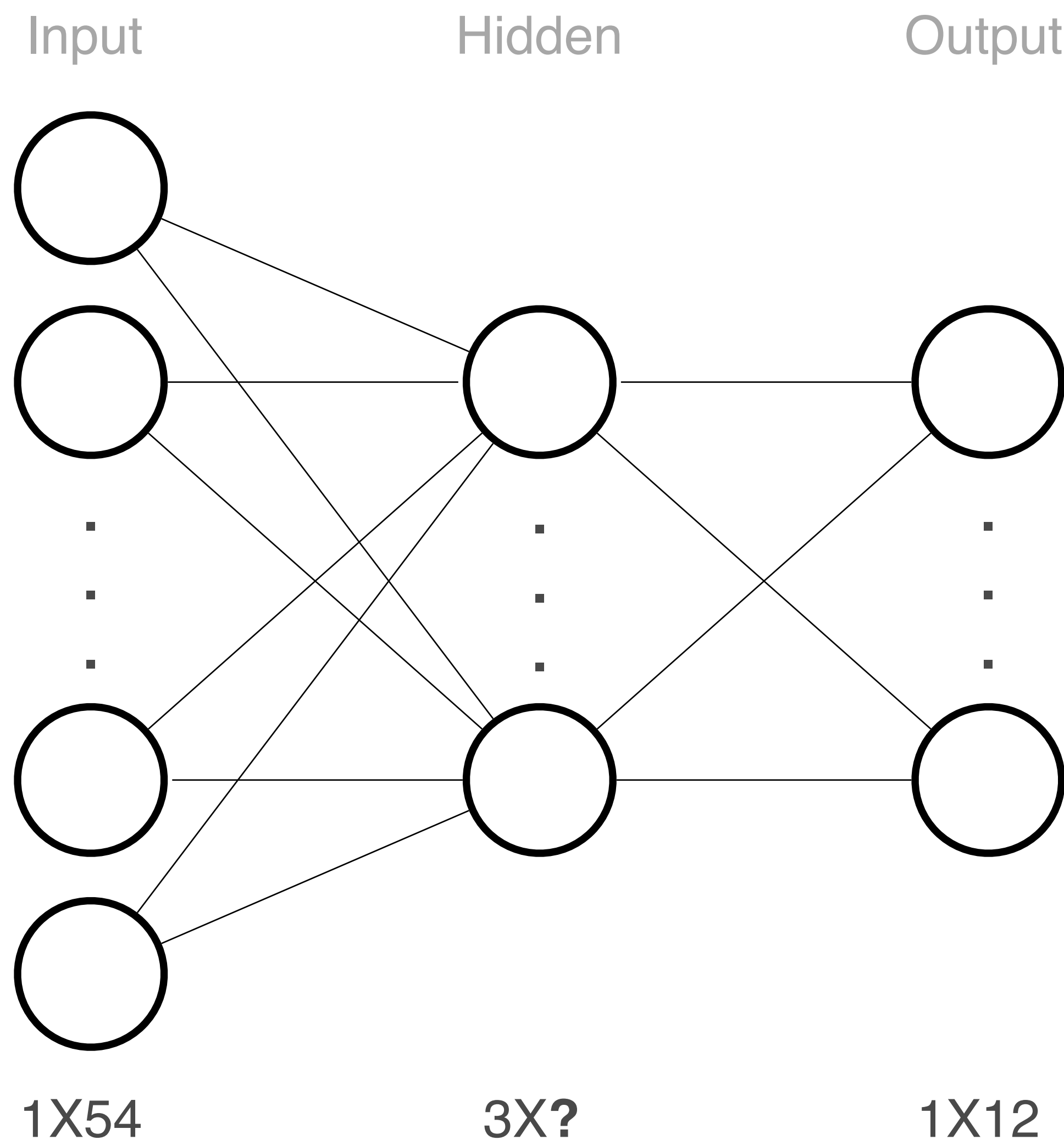


Struttura della rete:

- Neuroni di input: 54
- Neuroni di output: 12
- Numero di strati nascosti: 3

# STRUTTURA NN

DEEP NEURAL NETWORK



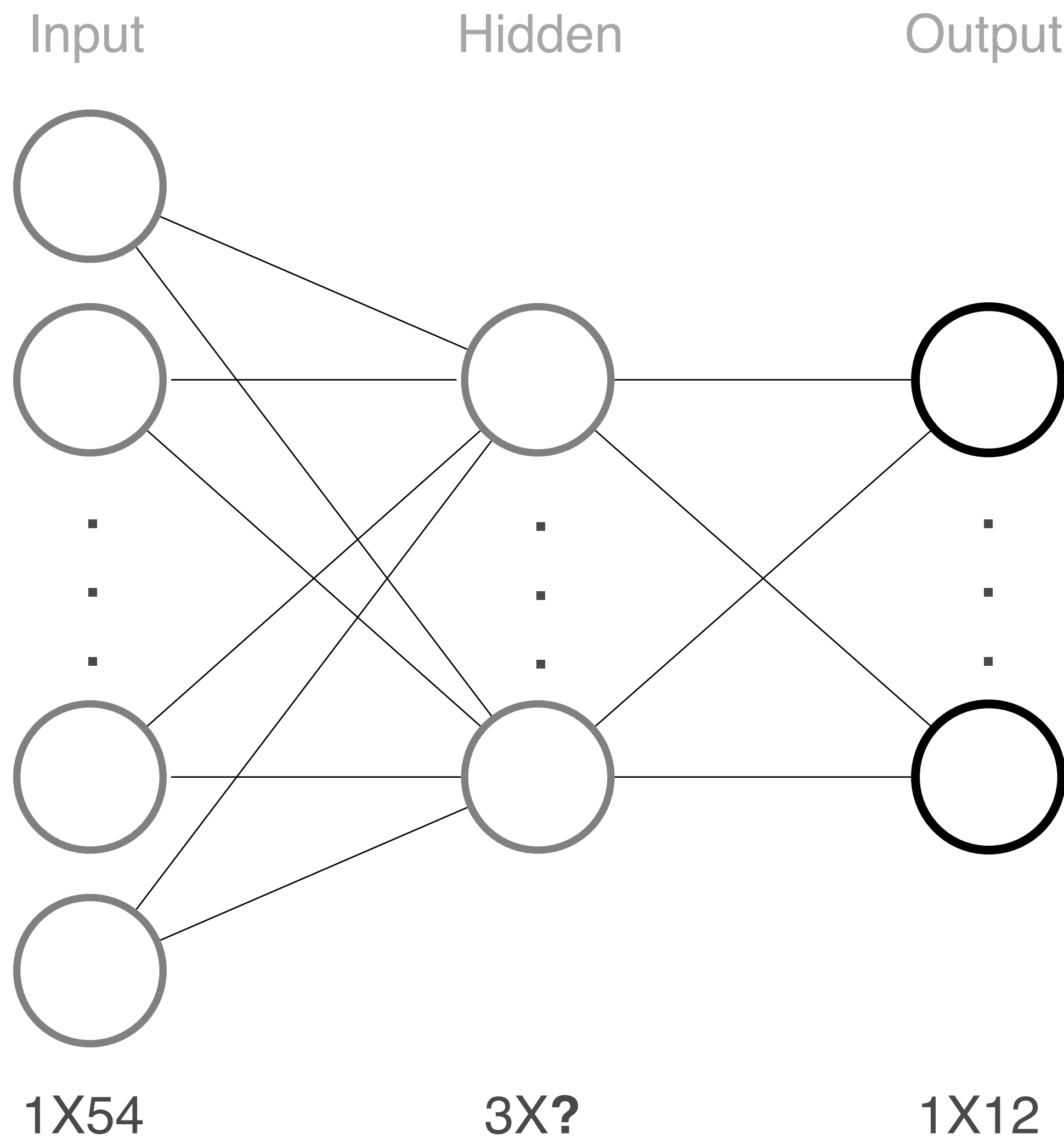
Funzione di Loss: ***Categorical Cross Entropy***

- Classificazione multiclasse
- Appartenenza ad una sola classe



# STRUTTURA NN

DEEP NEURAL NETWORK

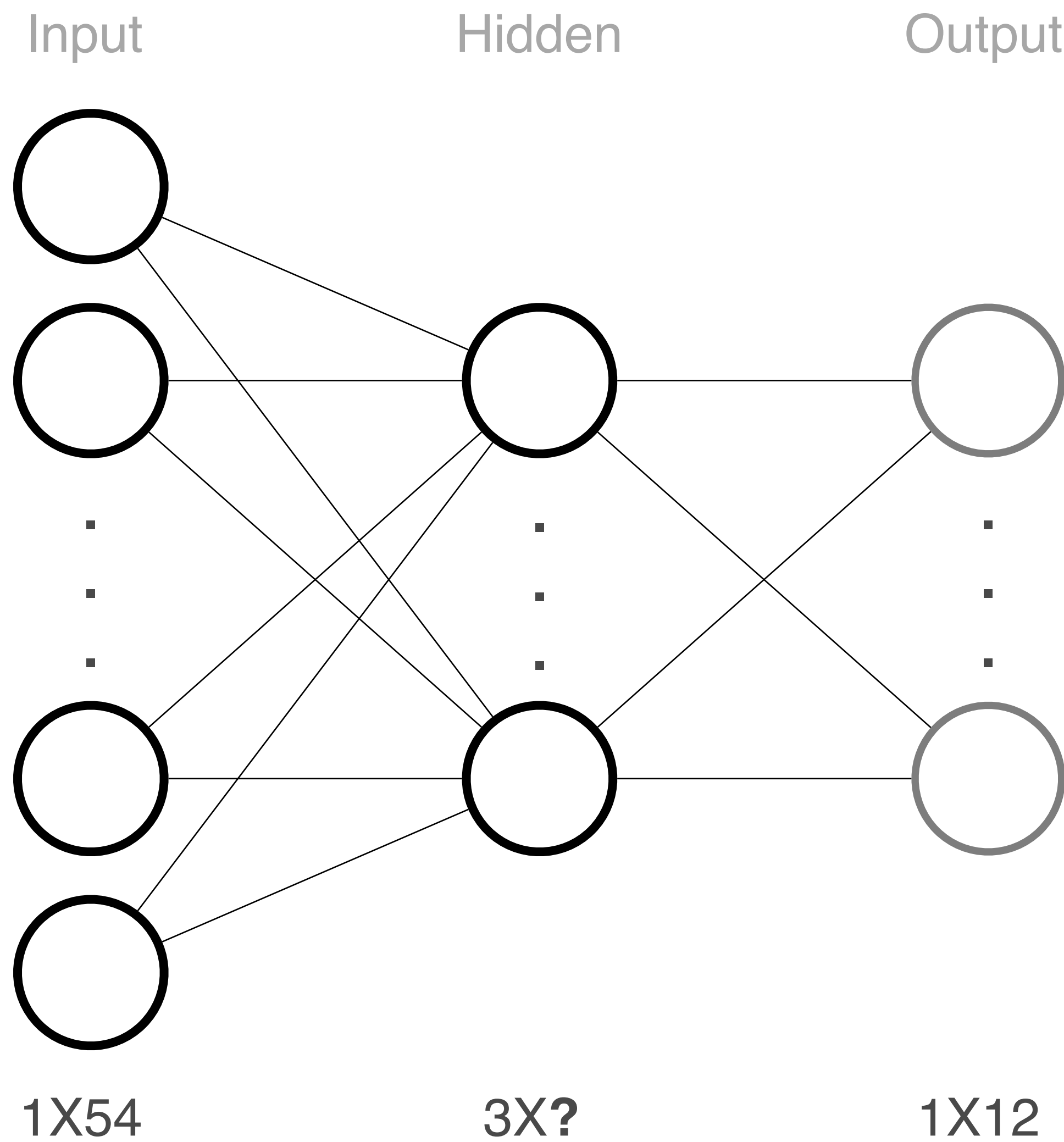


Funzione di attivazione: ***Softmax***

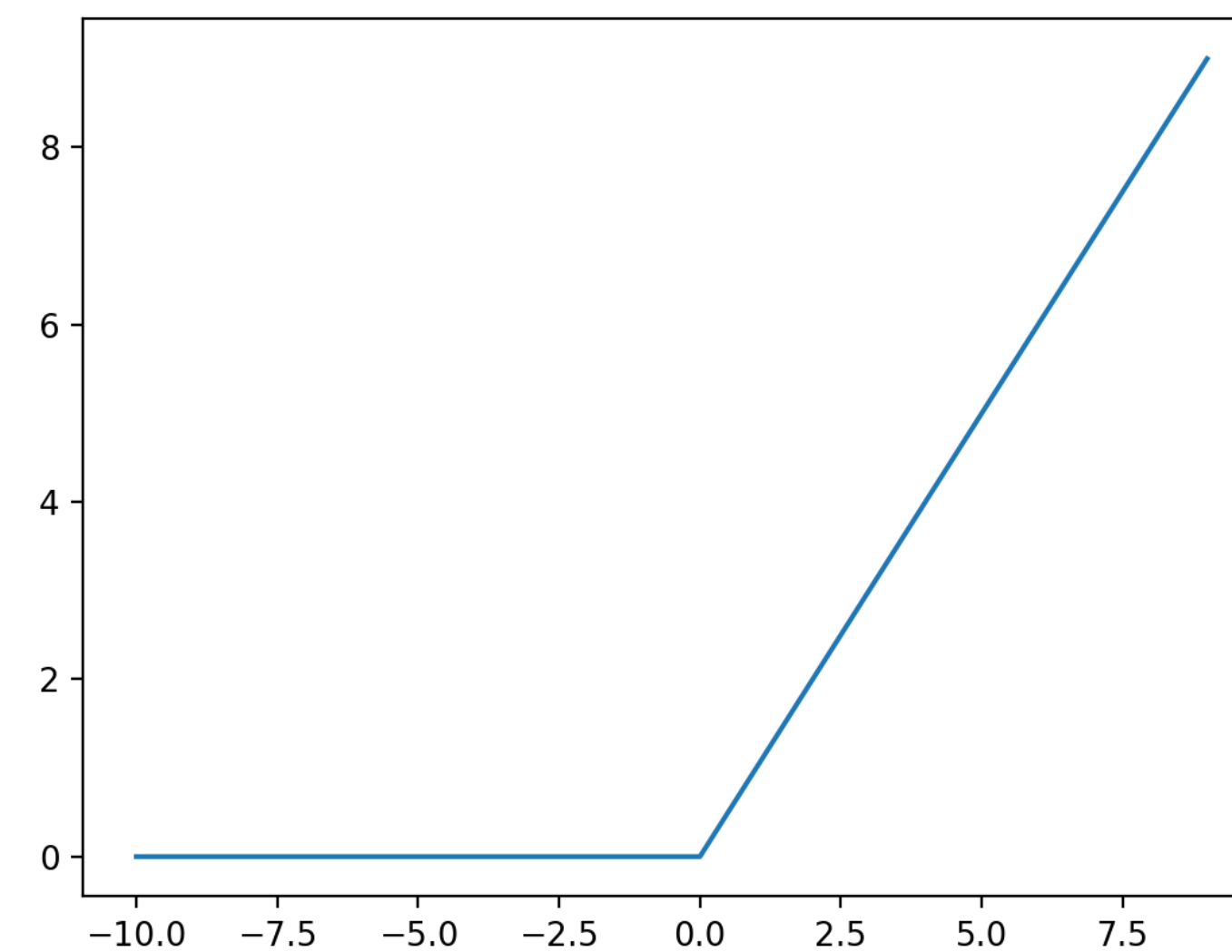
- Classificazione multiclasse
- Restituisce una distribuzione di probabilità

# STRUTTURA NN

DEEP NEURAL NETWORK

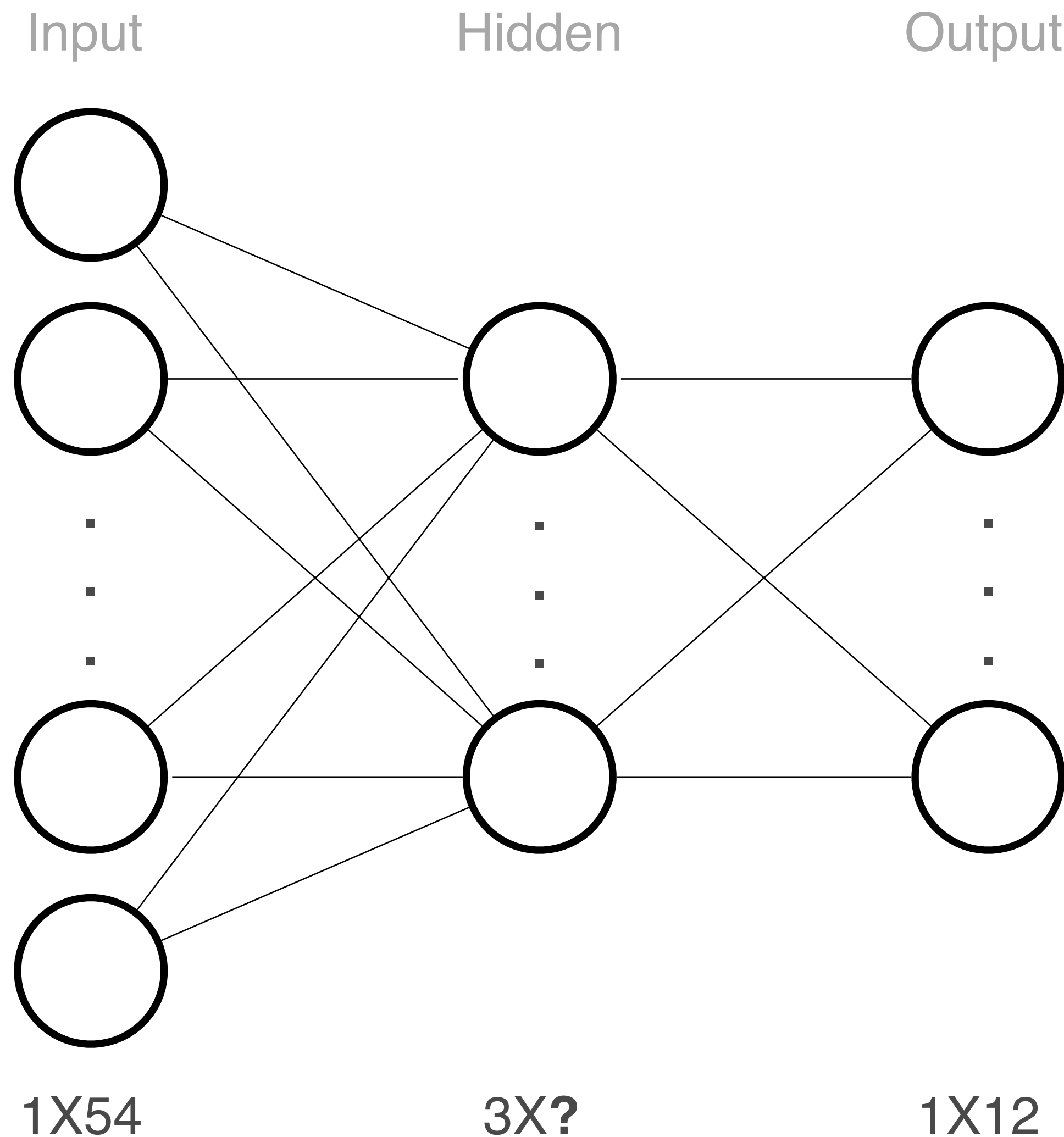


Funzione di attivazione: *Relu*



# STRUTTURA NN

DEEP NEURAL NETWORK

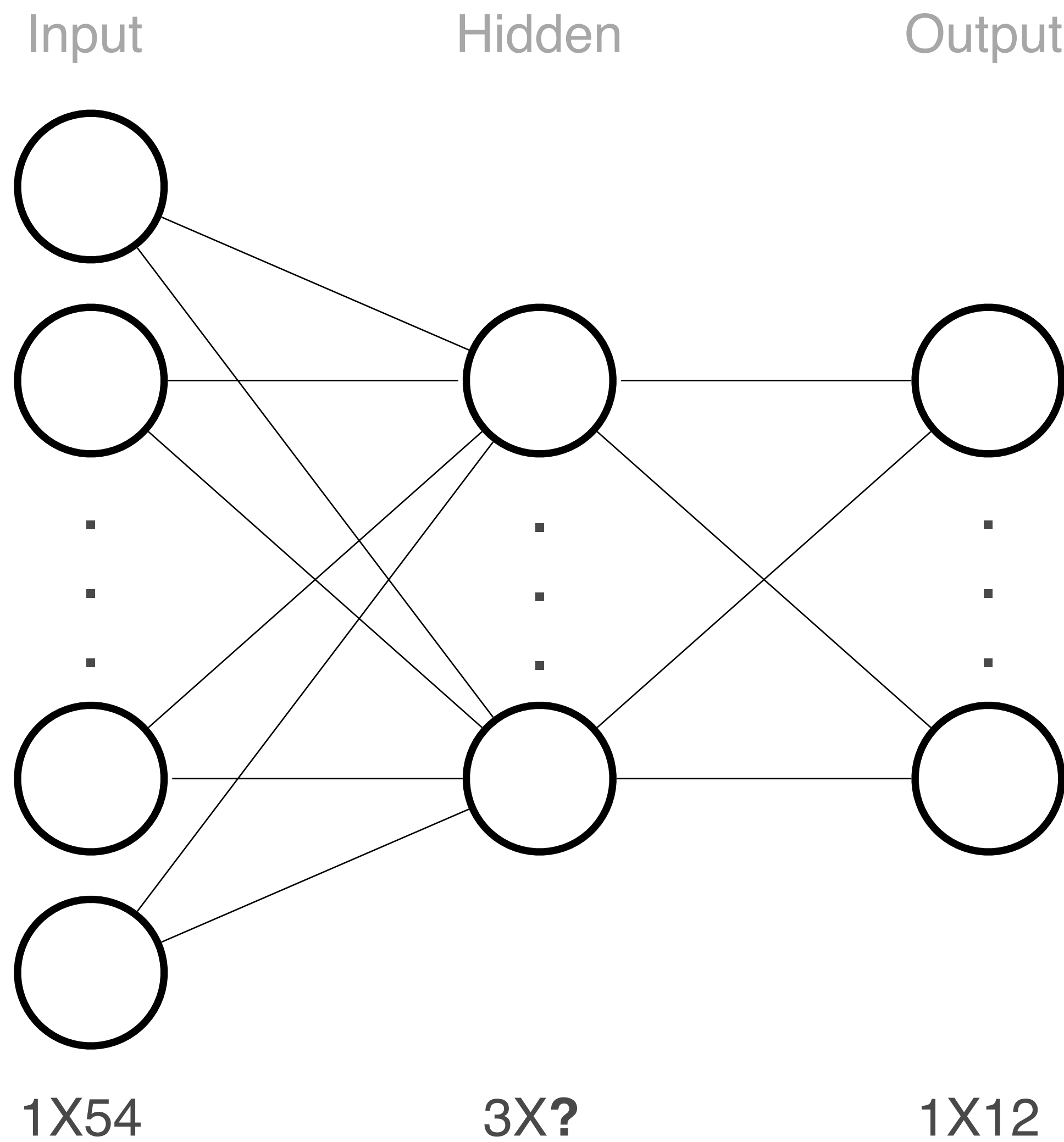


Ottimizzatore: ***Adam***

- Computazionalmente efficiente
- Richiede poca memoria
- Adatto per problemi con elevato numero di parametri e dati

# STRUTTURA NN

DEEP NEURAL NETWORK



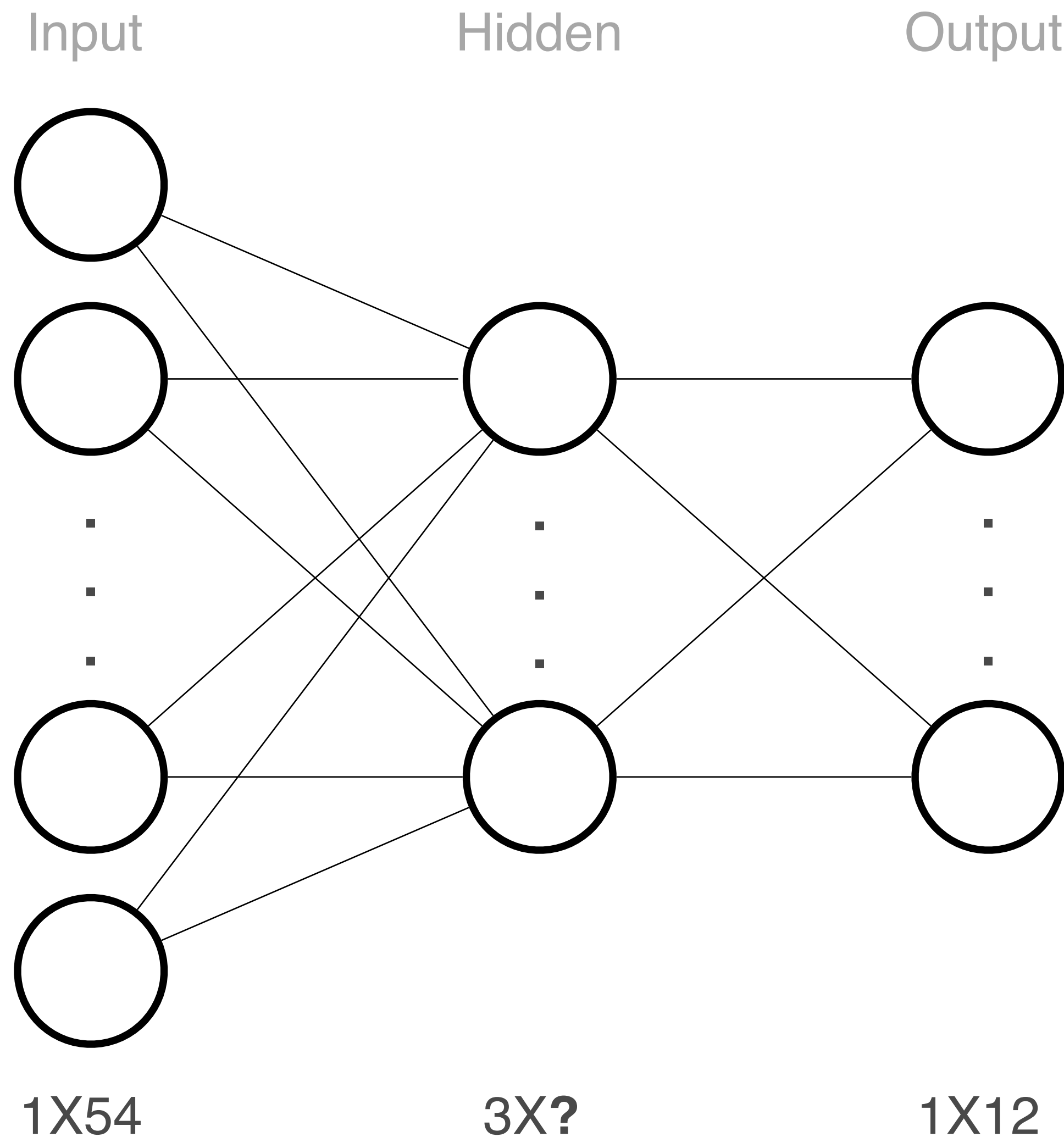
Metrica: ***Top-5 accuracy***

- L'esito corretto deve essere tra i primi cinque



# STRUTTURA NN

DEEP NEURAL NETWORK

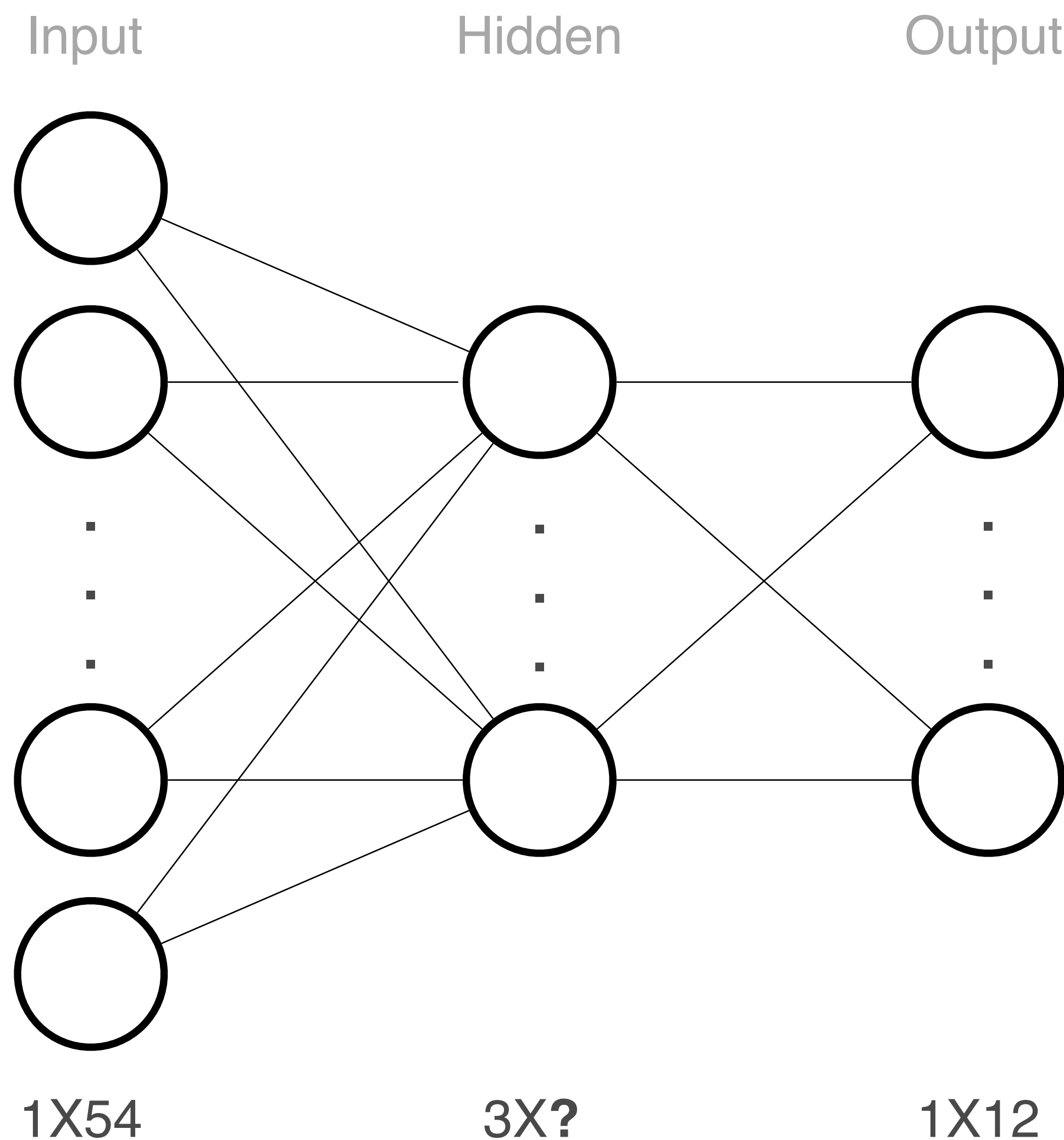


Fitting del modello:

- Batch size: 128
- Numero massimo di epoche: 200
- Callback: ***EarlyStopping***
  - Per evitare *Overfitting*
  - Monitora il valore della validation loss
  - Massimo 8 iterazioni senza miglioramenti
  - Ripristina i pesi migliori

# STRUTTURA NN

DEEP NEURAL NETWORK



## Iperparametri ottimizzati

- *Dropout rate*: [0.05, 0.4]
- *Learning rate*: [0.001, 0.01]
- Numero di neuroni:
  - Primo strato nascosto: [1, 54]
  - Secondo strato nascosto: [1, 54]
  - Terzo strato nascosto: [1, 54]

# AutoML

Processo che permette di ottimizzare una **funzione obiettivo** determinando la combinazione ideale degli iperparametri.

$$DCG_k = \sum_{i=1}^k \frac{2^{rel_i} - 1}{\log_2(i + 1)},$$

$$nDCG_k = \frac{DCG_k}{IDCG_k}$$

# AutoML

Processo che permette di ottimizzare una funzione obiettivo determinando la combinazione ideale degli iperparametri.





# AutoML

Processo che permette di ottimizzare una funzione obiettivo determinando la combinazione ideale degli iperparametri.



# AutoML

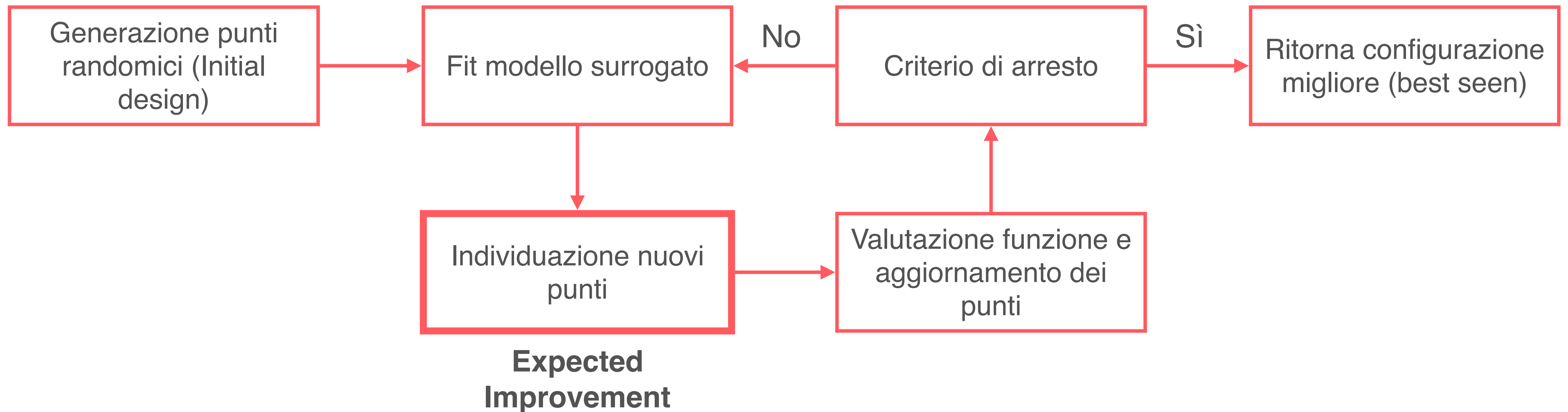
Processo che permette di ottimizzare una funzione obiettivo determinando la combinazione ideale degli iperparametri.

## Gaussian Process



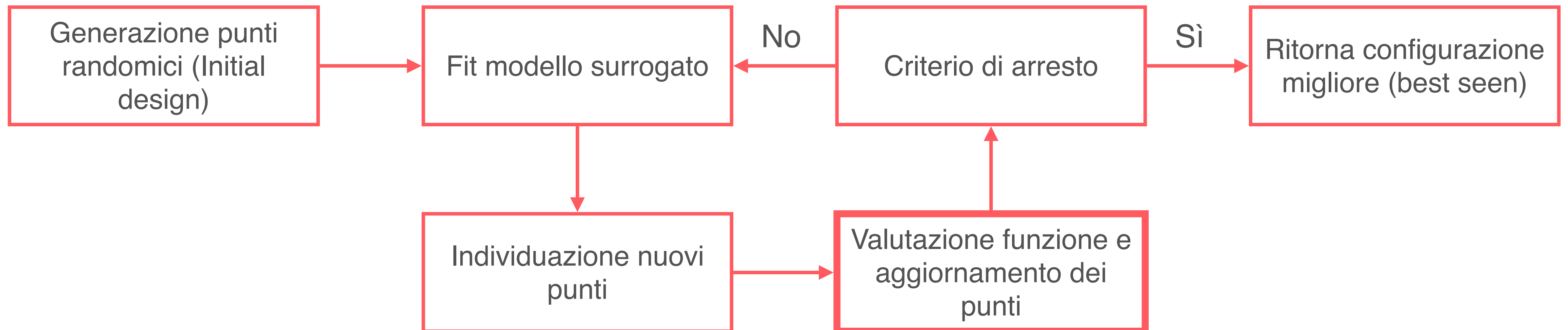
# AutoML

Processo che permette di ottimizzare una funzione obiettivo determinando la combinazione ideale degli iperparametri.



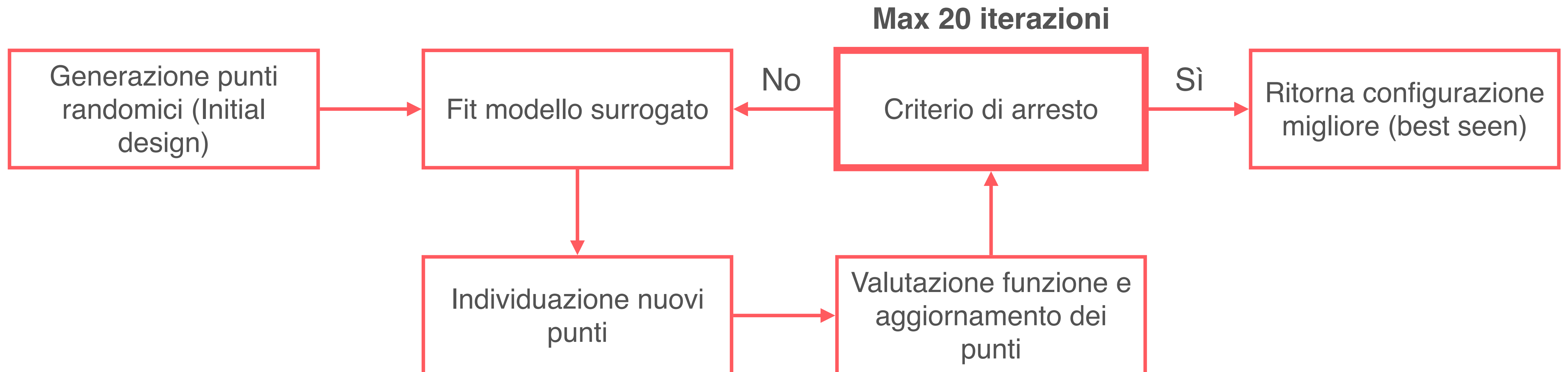
# AutoML

Processo che permette di ottimizzare una funzione obiettivo determinando la combinazione ideale degli iperparametri.



# AutoML

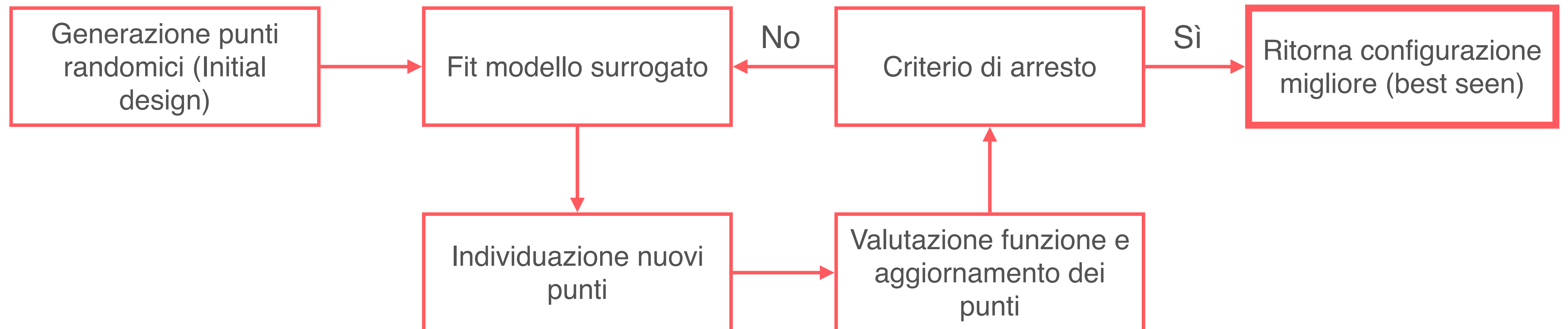
Processo che permette di ottimizzare una funzione obiettivo determinando la combinazione ideale degli iperparametri.





# AutoML

Processo che permette di ottimizzare una funzione obiettivo determinando la combinazione ideale degli iperparametri.

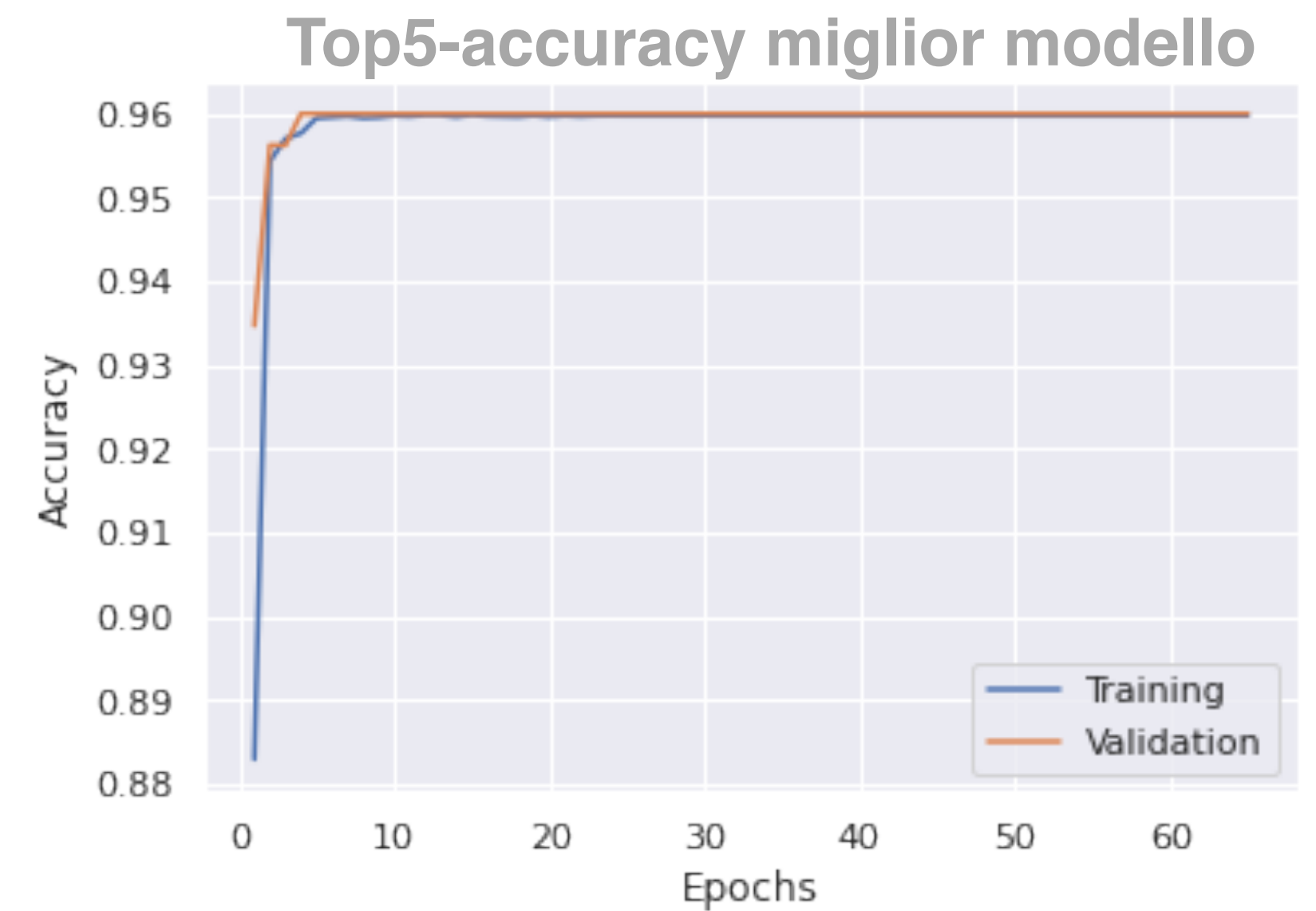
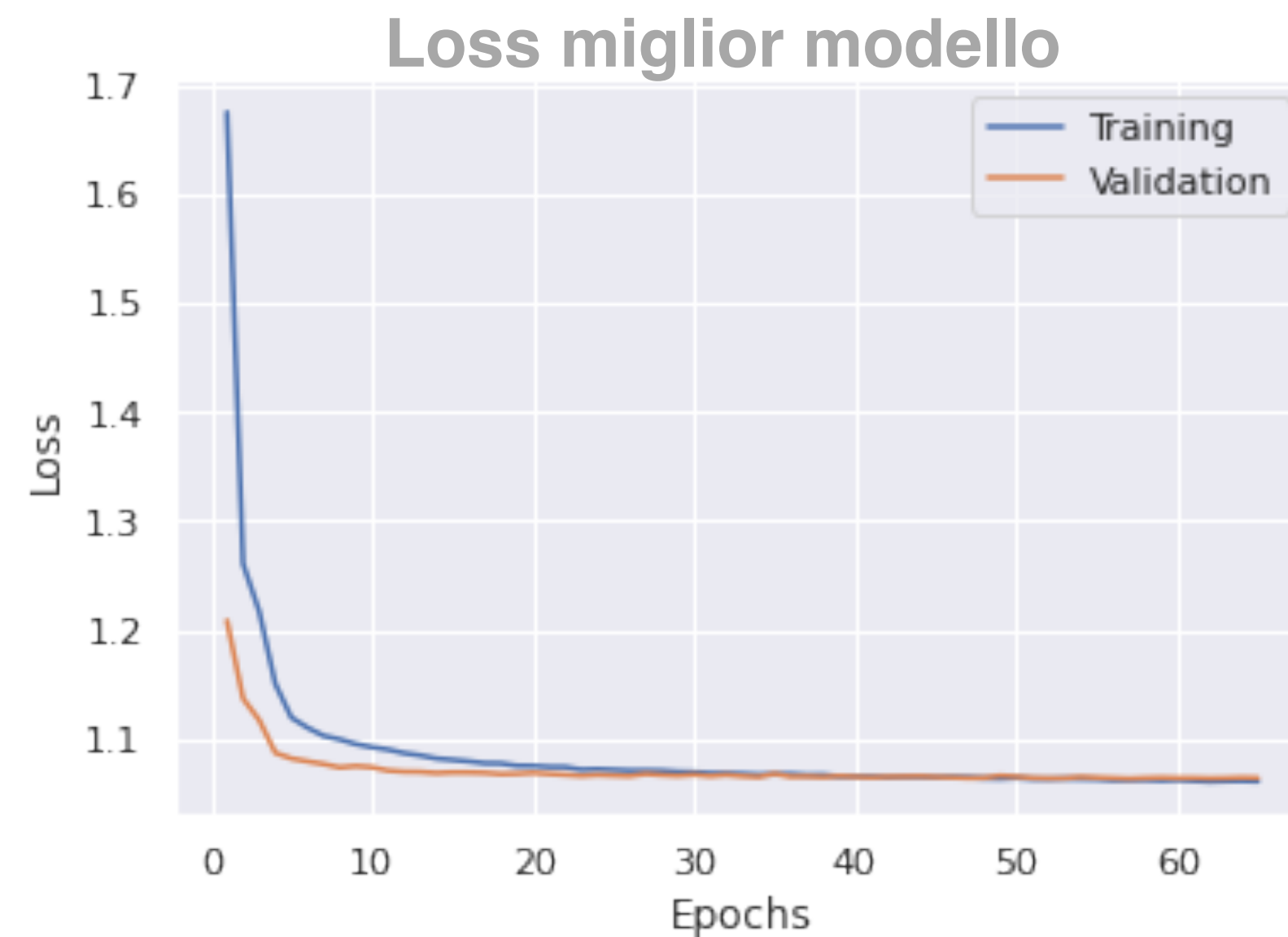


# RISULTATI NN

DEEP NEURAL NETWORK

## Iperparametri ottenuti:

- *Dropout rate*: 0.1
- *Learning rate*: 0.004
- Numero di neuroni:
  - Primo strato nascosto: 30
  - Secondo strato nascosto: 16
  - Terzo strato nascosto: 3

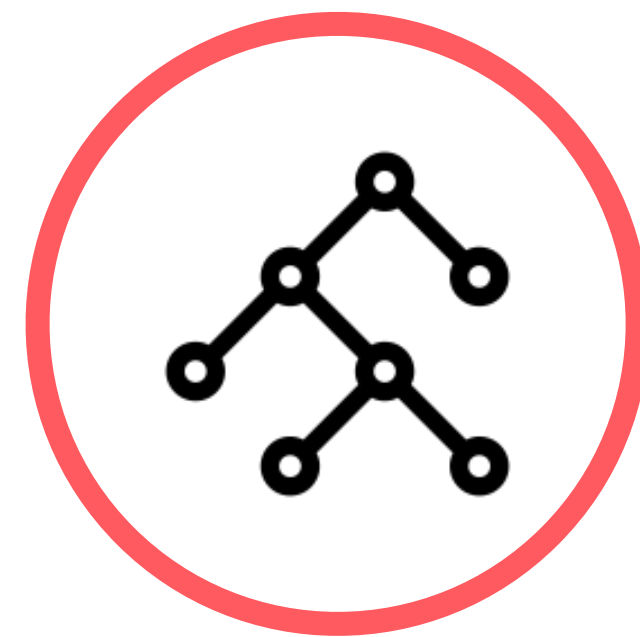


NDCG per il miglior modello con 10-CV: 82.53 +/- 0.16

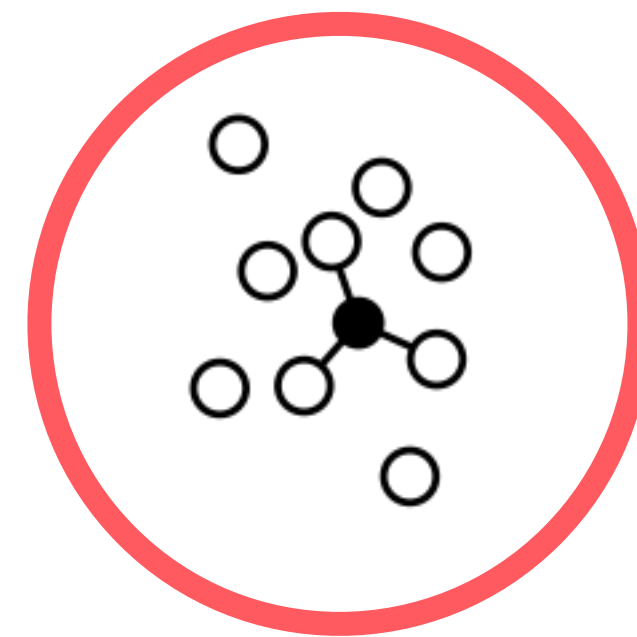
## ALTRI MODELLI



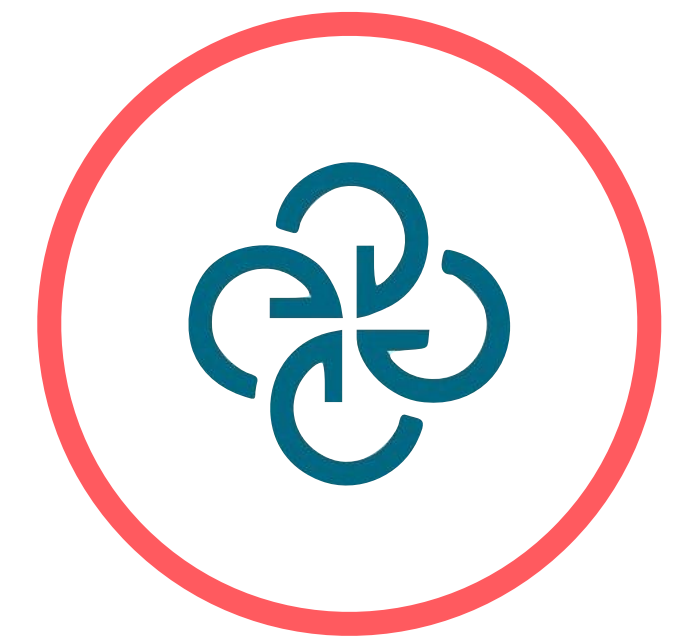
XGBoost



Random  
Forest



KNN



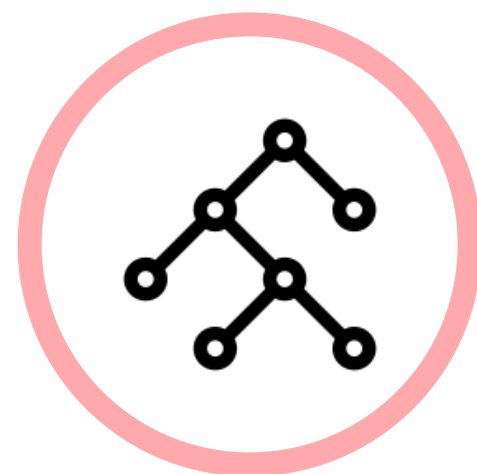
Ensemble  
model

# ALTRI MODELLI

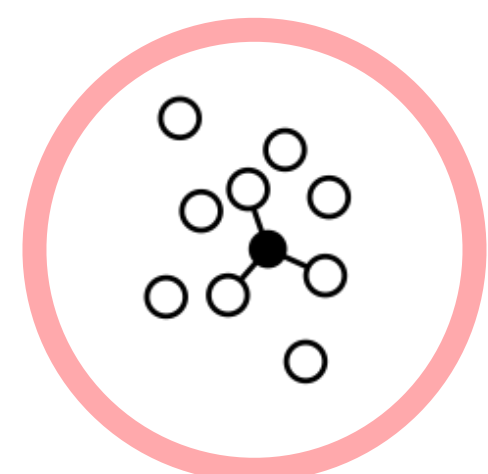
XGBoost



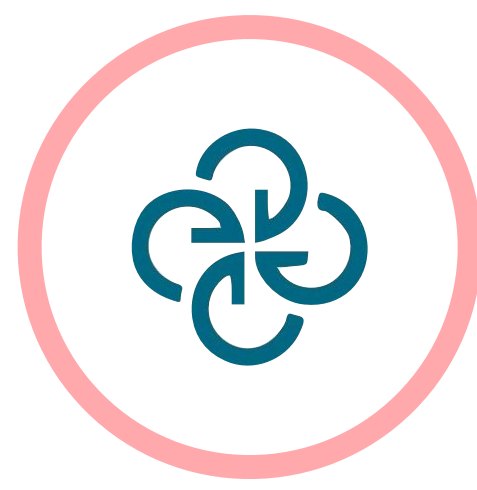
XGBoost



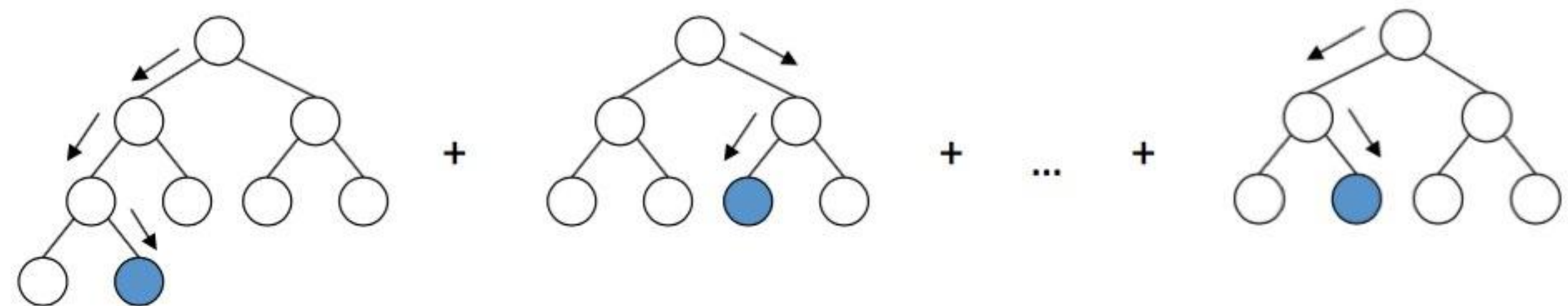
Random Forest



KNN



Ensemble model



Sequenza di decision tree che tramite la procedura di gradient boosting permette di migliorare i risultati precedenti.

Iperparametri ottimizzati con AutoML:

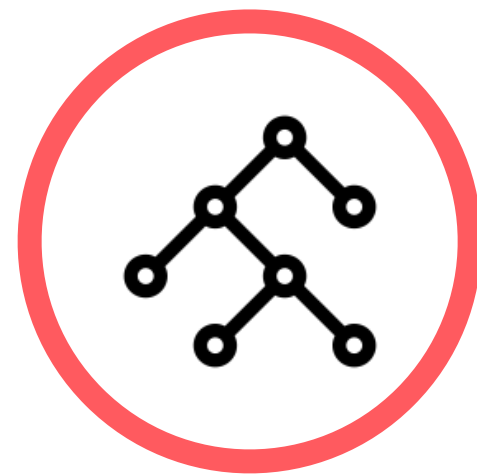
- Learning Rate  $[0, 1] \rightarrow 0.37$
- Numero di alberi  $[2, 10] \rightarrow 7$
- Profondità massima  $[3, 10] \rightarrow 7$
- Alpha regularizer  $[0, 1] \rightarrow 0.47$
- Gamma regularizer  $[0, 2] \rightarrow 2$

# ALTRI MODELLI

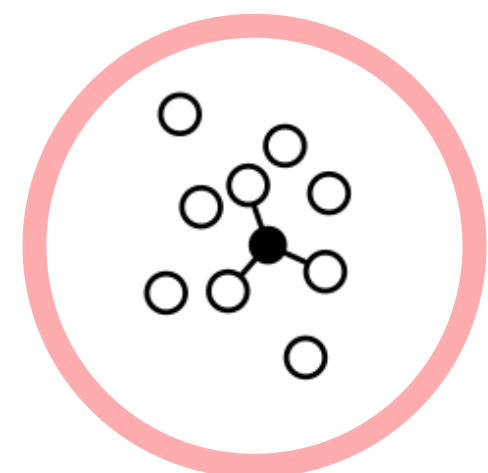
Random Forest



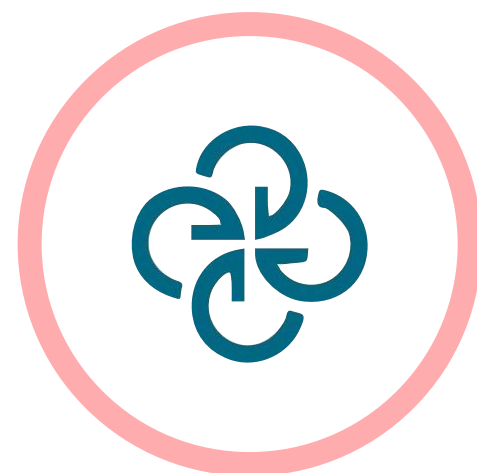
XGBoost



Random Forest



KNN



Ensemble model

Iperparametri ottimizzati con AutoML:

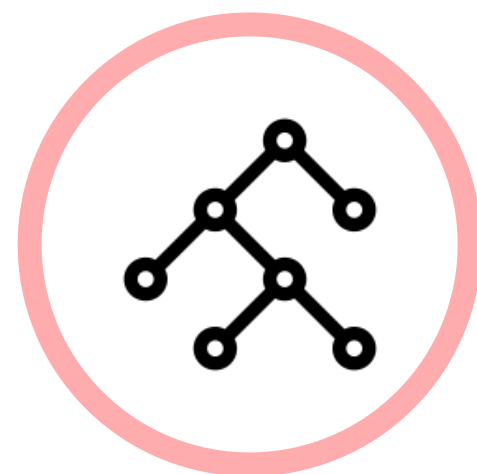
- Numero di alberi [100, 500]  $\rightarrow$  452
- Profondità massima [5, 30]  $\rightarrow$  14

# ALTRI MODELLI

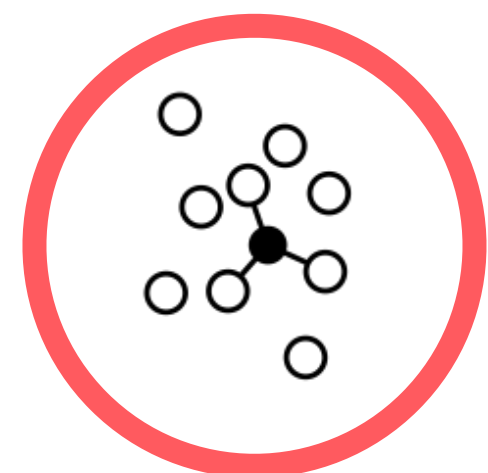
KNN



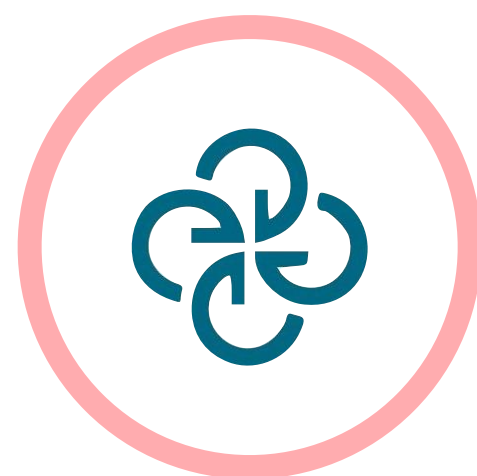
XGBoost



Random Forest



KNN



Ensemble model

Distanza utilizzata: *Minkowski*

Iperparametri ottimizzati con AutoML:

- K [50, 200]  $\rightarrow$  117

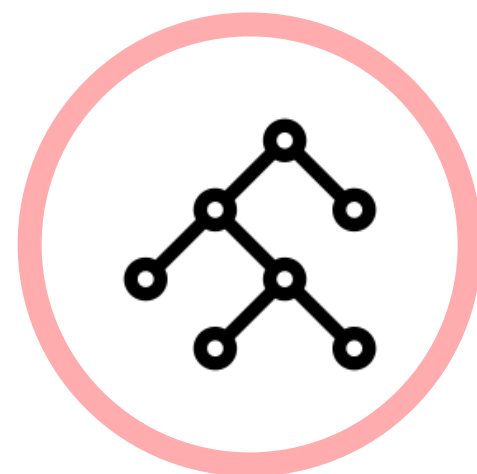


# ALTRI MODELLI

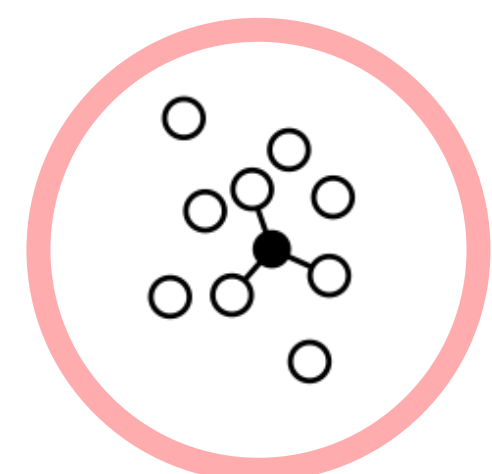
Ensemble Model



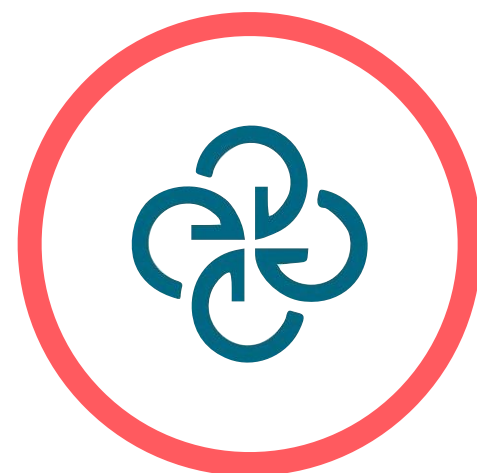
XGBoost



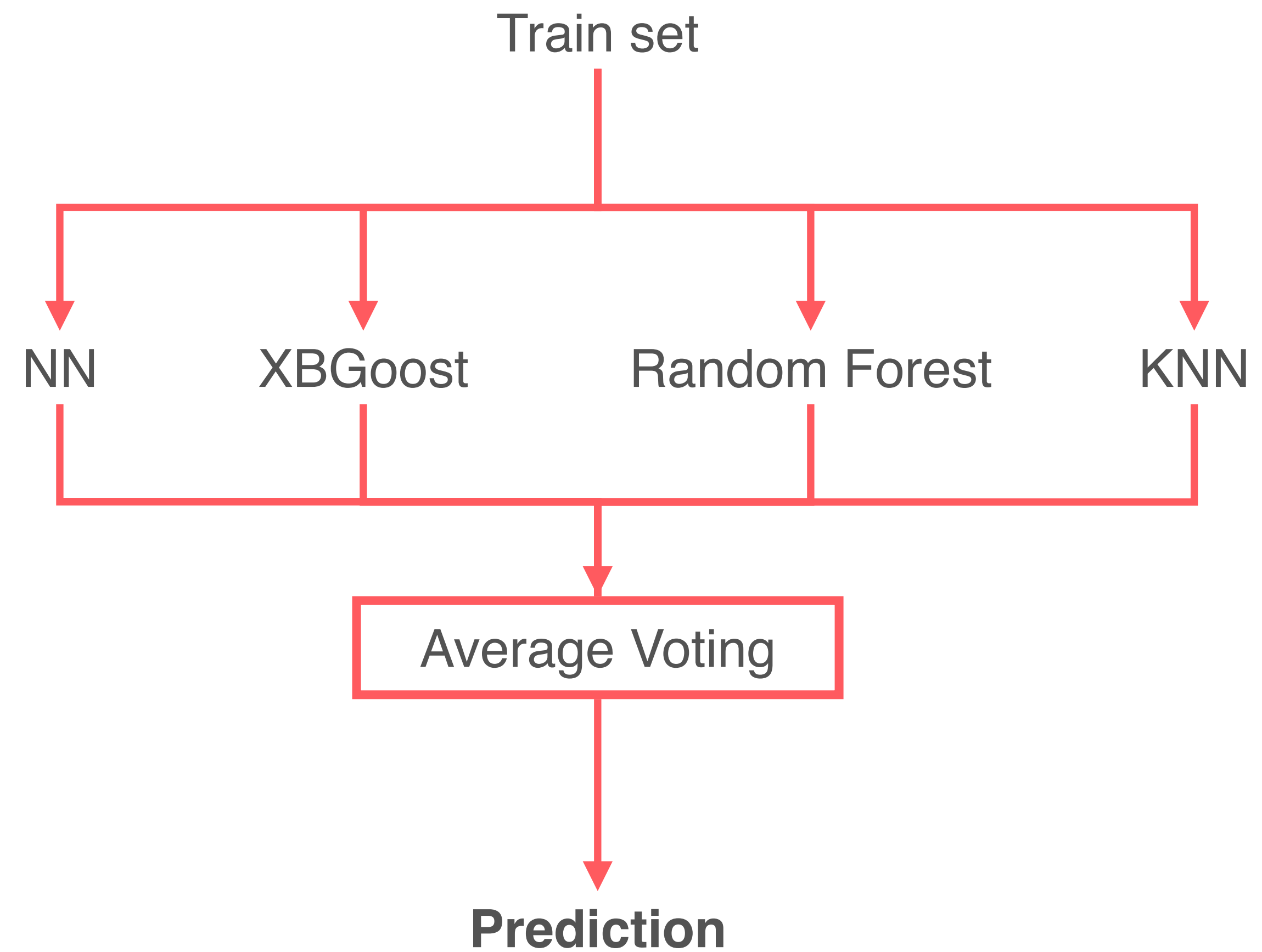
Random Forest



KNN



Ensemble model



# RISULTATI

Classificatore	Tempo	Cross Validation (NDCG +/- SD)	Kaggle Score (NDCG)
FCNN	1506	82.53 (+/- 0.16)	87.02%
XGBoost	842	82.85 (+/- 0.16)	87.37%
Random Forest	1557	82.81 (+/- 0.16)	87.42%
K-Nearest Neighbors	2451	81.48 (+/- 0.20)	86.05%
Ensemble	5942	82.64 (+/- 0.17)	87.14%

# RISULTATI

Classificatore	Tempo	Cross Validation (NDCG +/- SD)	Kaggle Score (NDCG)
FCNN	1506	82.53 (+/- 0.16)	87.02%
XGBoost	842	82.85 (+/- 0.16)	87.37%
Random Forest	1557	82.81 (+/- 0.16)	87.42%
K-Nearest Neighbors	2451	81.48 (+/- 0.20)	86.05%
Ensemble	5942	82.64 (+/- 0.17)	87.14%

# RISULTATI

Classificatore	Tempo	Cross Validation (NDCG +/- SD)	Kaggle Score (NDCG)
FCNN	1506	82.53 (+/- 0.16)	87.02%
XGBoost	842	82.85 (+/- 0.16)	87.37%
Random Forest	1557	82.81 (+/- 0.16)	87.42%
K-Nearest Neighbors	2451	81.48 (+/- 0.20)	86.05%
Ensemble	5942	82.64 (+/- 0.17)	87.14%

# CONCLUSIONI

L'approccio Deep Learning non risulta essere il migliore in questo contesto.

I risultati sono fortemente influenzati dallo sbilanciamento dei dati.

## CONCLUSIONI

L'approccio Deep Learning non risulta essere il migliore in questo contesto.

I risultati sono fortemente influenzati dallo sbilanciamento dei dati.

## SVILUPPI FUTURI

Migliorare la classificazione delle label più rare cercando di eliminare l'effetto dello sbilanciamento del dataset.

Incrementare le risorse computazionali per poter aumentare il numero di iperparametri da ottimizzare.





**Grazie per l'attenzione**

