



UNIVERSITÀ  
DEGLI STUDI  
FIRENZE



# Data Mining per la Previsione del GPA

Regole Associative e  
Apprendimento Supervisionato

Isabella Cerasuolo

CdL Data Science, Calcolo Scientifico &  
Intelligenza Artificiale  
a.a. 2024-2025

# Struttura del progetto

## Regole Associate

- Association rules con algoritmo Apriori con supporto, confidenza e lift

## Apprendimento Supervisionato

- Decision Trees
- k-NN
- ANN

## Confronto tra modelli

## Conclusioni



# Student Performance dataset

- **StudentID**
- **Age** [Num, Discr]
- **Gender** [Cat, Nom]
- **Ethnicity** [Cat, Nom]
- **ParentalEducation** [Cat, Ord]
- **StudyTimeWeekly** [Num, Cont]
- **Absences** [Num, Discr]
- **Tutoring** [Cat, Nom]
- **ParentalSupport** [Cat, Ord]
- **Extracurricular** [Cat, Nom]
- **Sports** [Cat, Nom]
- **Music** [Cat, Nom]
- **Volunteering** [Cat, Nom]
- **GPA**
- **GradeClass** [Cat, Ord]

# Regole associative: Apriori

- Tecnica di data mining che permette di identificare combinazioni ricorrenti di eventi o azioni in un dataset.
- Obiettivo: identificare le combinazioni di variabili che si verificano frequentemente tra gli studenti, fornendo insight utili per capire quali fattori influenzano maggiormente la performance accademica
- Minsup=15%; Minconf=70%

- **Age:**

- \_15\_16
- \_17\_18

- **StudyTimeWeekly:**

- \_s=[0,5]
- \_m=[6,10]
- \_l=[11,20]

- **Absences:**

- \_s=[0,10]
- \_m=[11,15]
- \_l=[16, 20]
- \_xl>20

- **GradeClass**

- \_Good=[0,2]
- \_Bad=[3,4]



# Regole associative: Apriori

#	Pattern (Antecedente $\Rightarrow$ Consequente)	Supporto	Confidenza	Lift
1	(StudyTimeWeekly_s) $\Rightarrow$ (GradeClass_Bad)	0.186	0.746	1.10
2	(StudyTimeWeekly_m) $\Rightarrow$ (GradeClass_Bad)	0.152	0.706	1.04
3	(Absences_s) $\Rightarrow$ (GradeClass_Good)	0.268	0.760	2.37
4	(GradeClass_Good) $\Rightarrow$ (Absences_s)	0.268	0.837	2.37
5	(Absences_l) $\Rightarrow$ (GradeClass_Bad)	0.163	0.933	1.37
6	(Absences_xl) $\Rightarrow$ (GradeClass_Bad)	0.281	0.960	1.41
7	(ParentalSupport_1) $\Rightarrow$ (GradeClass_Bad)	0.150	0.734	1.08
8	(Age_15_16, Gender) $\Rightarrow$ (GradeClass_Bad)	0.178	0.706	1.04

# Regole associative: Apriori

#	Pattern (Antecedente $\Rightarrow$ Consequente)	Supporto	Confidenza	Lift
1	(StudyTimeWeekly_s) $\Rightarrow$ (GradeClass_Bad)	0.186	0.746	1.10
2	(StudyTimeWeekly_m) $\Rightarrow$ (GradeClass_Bad)	0.152	0.706	1.04
3	(Absences_s) $\Rightarrow$ (GradeClass_Good)	0.268	0.760	2.37
4	(GradeClass_Good) $\Rightarrow$ (Absences_s)	0.268	0.837	2.37
5	(Absences_l) $\Rightarrow$ (GradeClass_Bad)	0.163	0.933	1.37
6	(Absences_xl) $\Rightarrow$ (GradeClass_Bad)	0.281	0.960	1.41
7	(ParentalSupport_1) $\Rightarrow$ (GradeClass_Bad)	0.150	0.734	1.08
8	(Age_15_16, Gender) $\Rightarrow$ (GradeClass_Bad)	0.178	0.706	1.04

# Apprendimento supervisionato

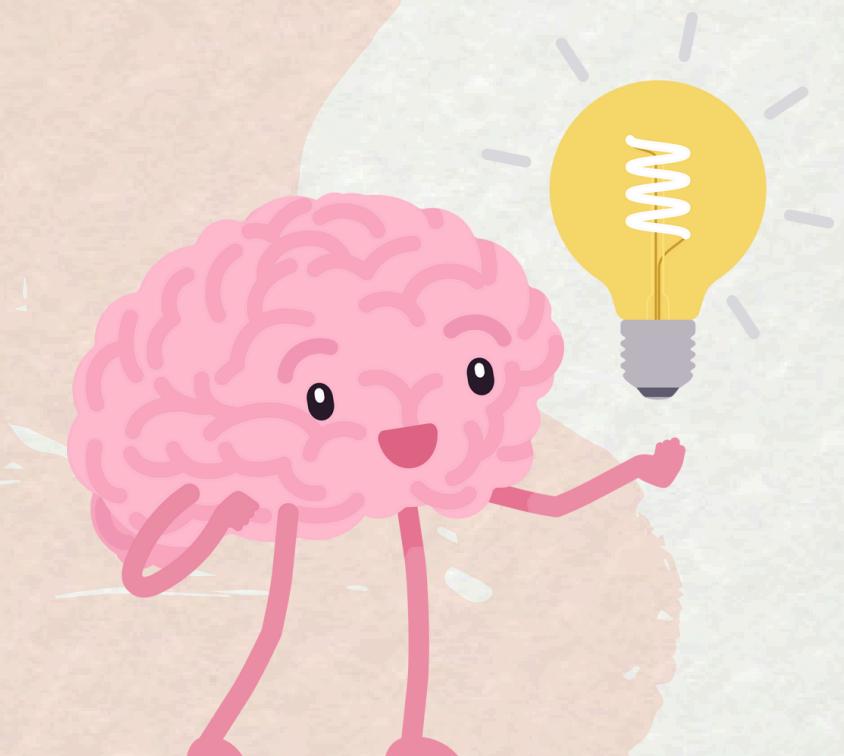
- Obiettivo: Prevedere il rendimento accademico degli studenti utilizzando modelli di Machine Learning per identificare pattern nei dati e migliorare la capacità predittiva.
- 80% training e 20% test
- Variabile target= **GradeClass** (modelli di classificazione, non di regressione)
- Utilizzo di SMOTE per mitigare lo squilibrio delle classi e migliorare la capacità di generalizzazione dei modelli (68% nella classe 0 e 32% nella classe 1)

## Tecniche usate:

- Decision Tree
- k-NN
- ANN Multistrato

## Metriche di valutazione

- Confusion Matrix
  - Accuracy
  - Error Rate

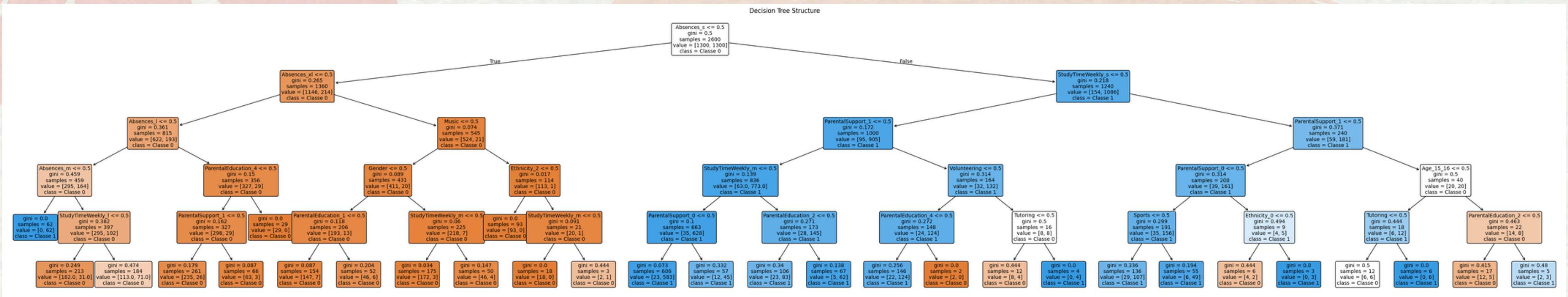


# Decision Tree

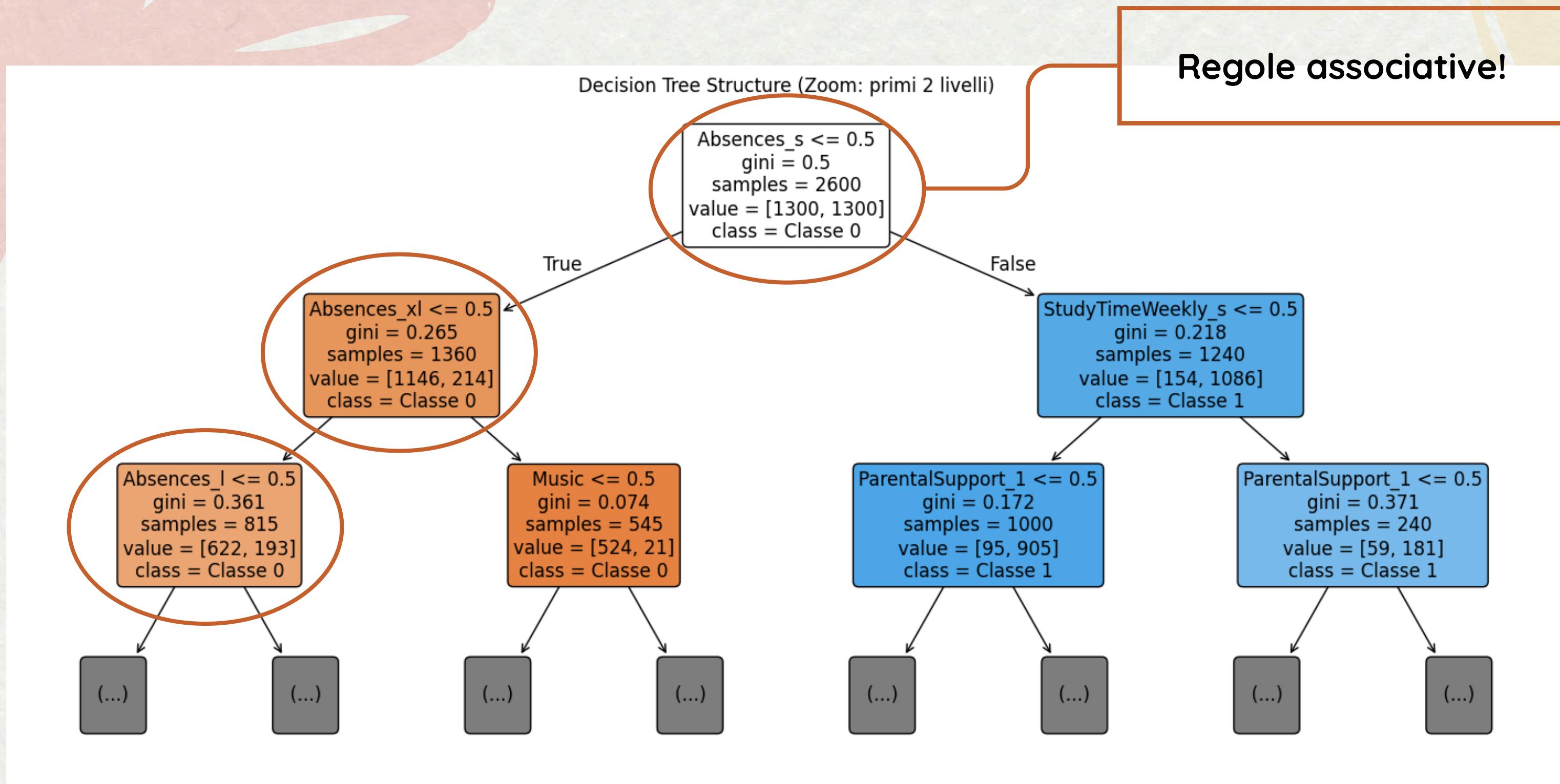


- Suddivide iterativamente il dataset in sottoinsiemi più omogenei rispetto alla variabile target.
- Struttura:
  - **Nodo radice:** intero dataset.
  - **Nodi interni:** ciascun nodo valuta una condizione su una delle feature.
  - **Rami:** Ogni ramo corrisponde a un possibile esito della condizione.
  - **Foglie:** rappresentano la classe predetta
- Feature split, Gini Index
- Profondità massima, 5

# Decision Tree

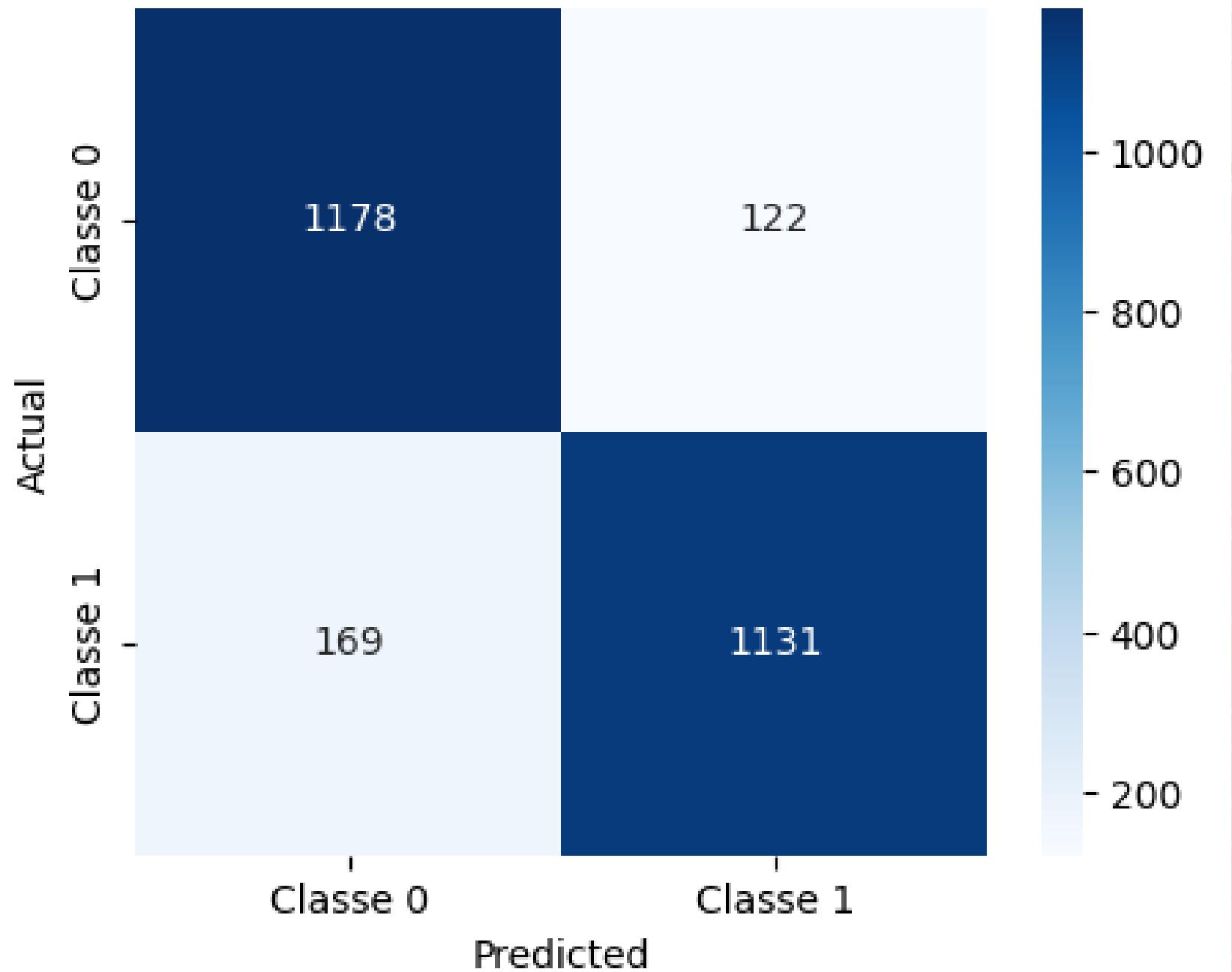


# Decision Tree



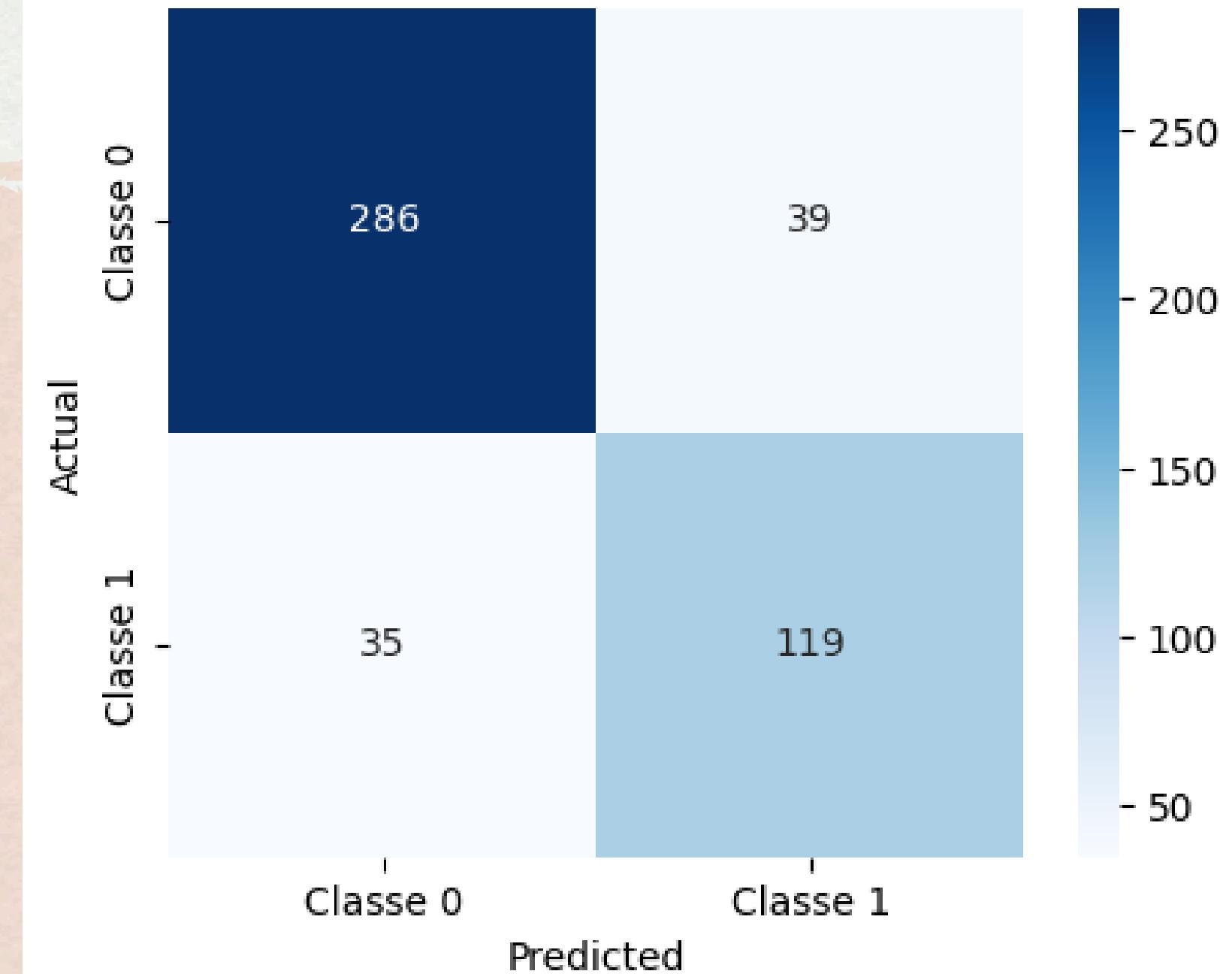
# Decision Tree

Confusion Matrix - Decision Tree - Training



Training Accuracy, 89%

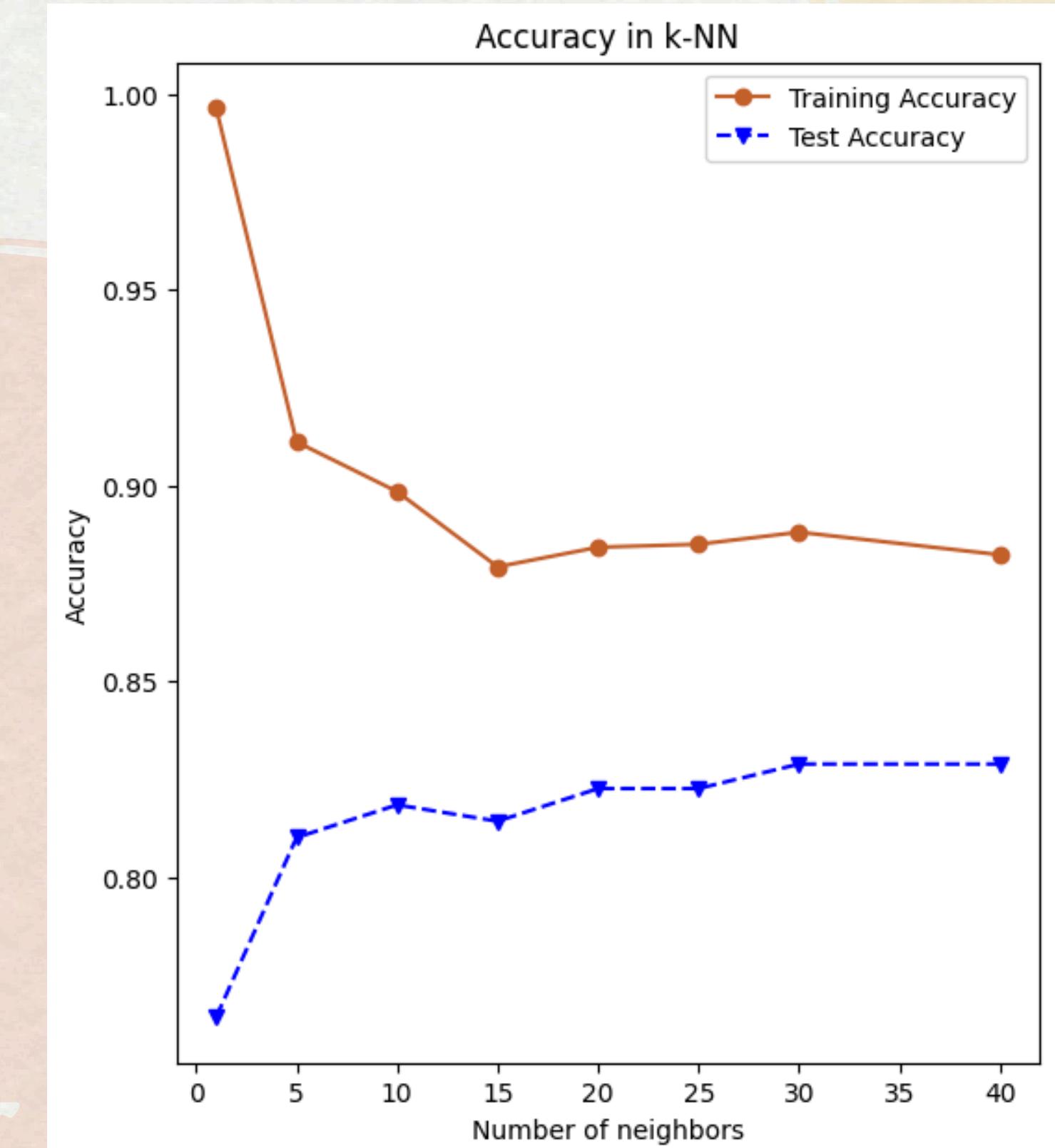
Confusion Matrix - Decision Tree - Test



Test Accuracy, 85%

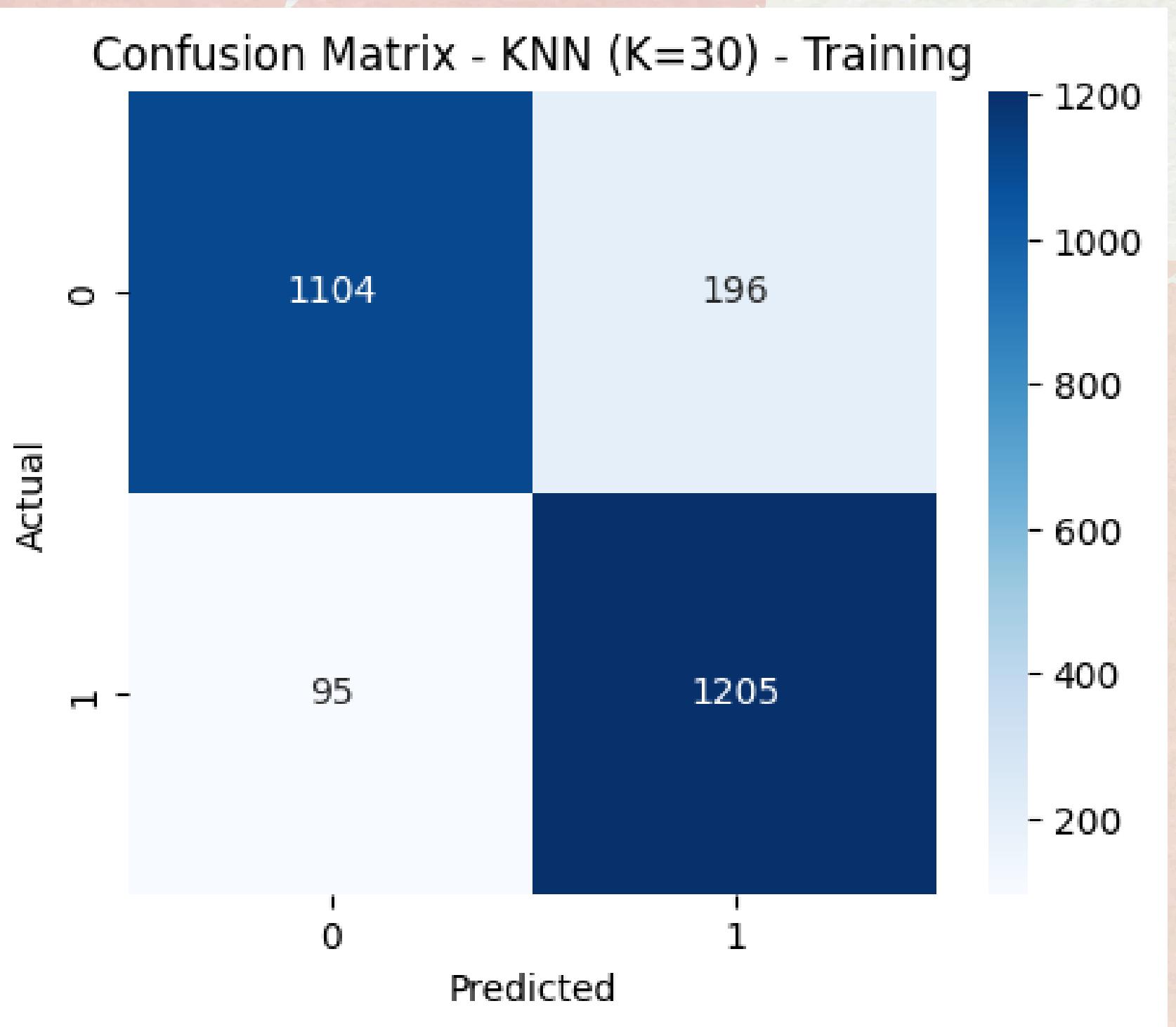
# k-NN

- Classificare studenti in base a caratteristiche binarie (0/1)
- **Distanza di Hamming**: ideale per variabili binarie/categoriche
- Test su **diversi valori di k per trovare il migliore**
- Scelta del parametro k
  - k troppo basso → Overfitting (sensibile al rumore)
  - k troppo alto → Underfitting (poca separazione tra classi)
  - **k = 30** ottimizza bias-varianza
- Vantaggi
  - **Semplice**
  - **Flessibile**
  - **Intuitivo** (vicini = classificazione)

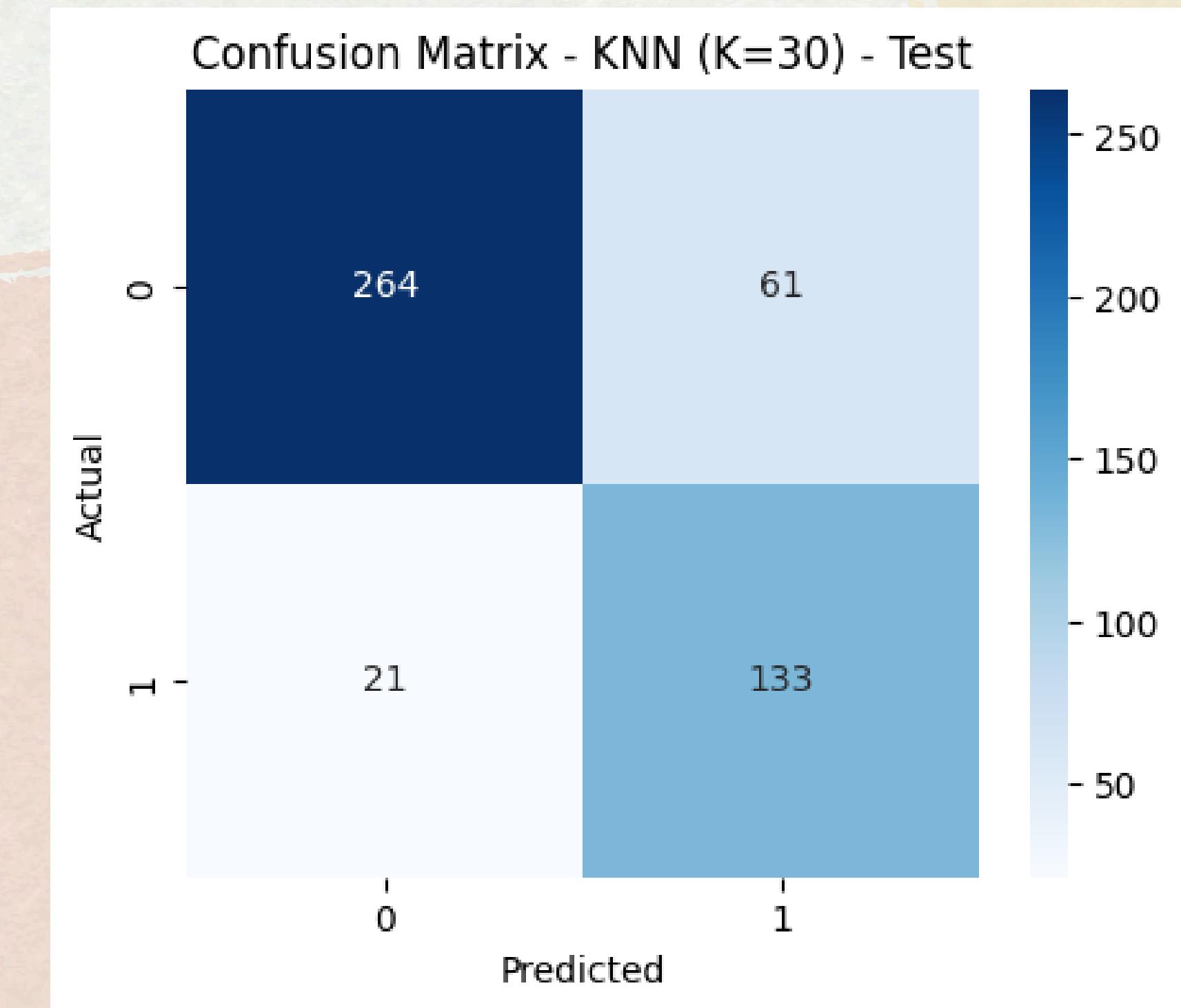


# k-NN

Confusion Matrix - KNN (K=30) - Training



Confusion Matrix - KNN (K=30) - Test

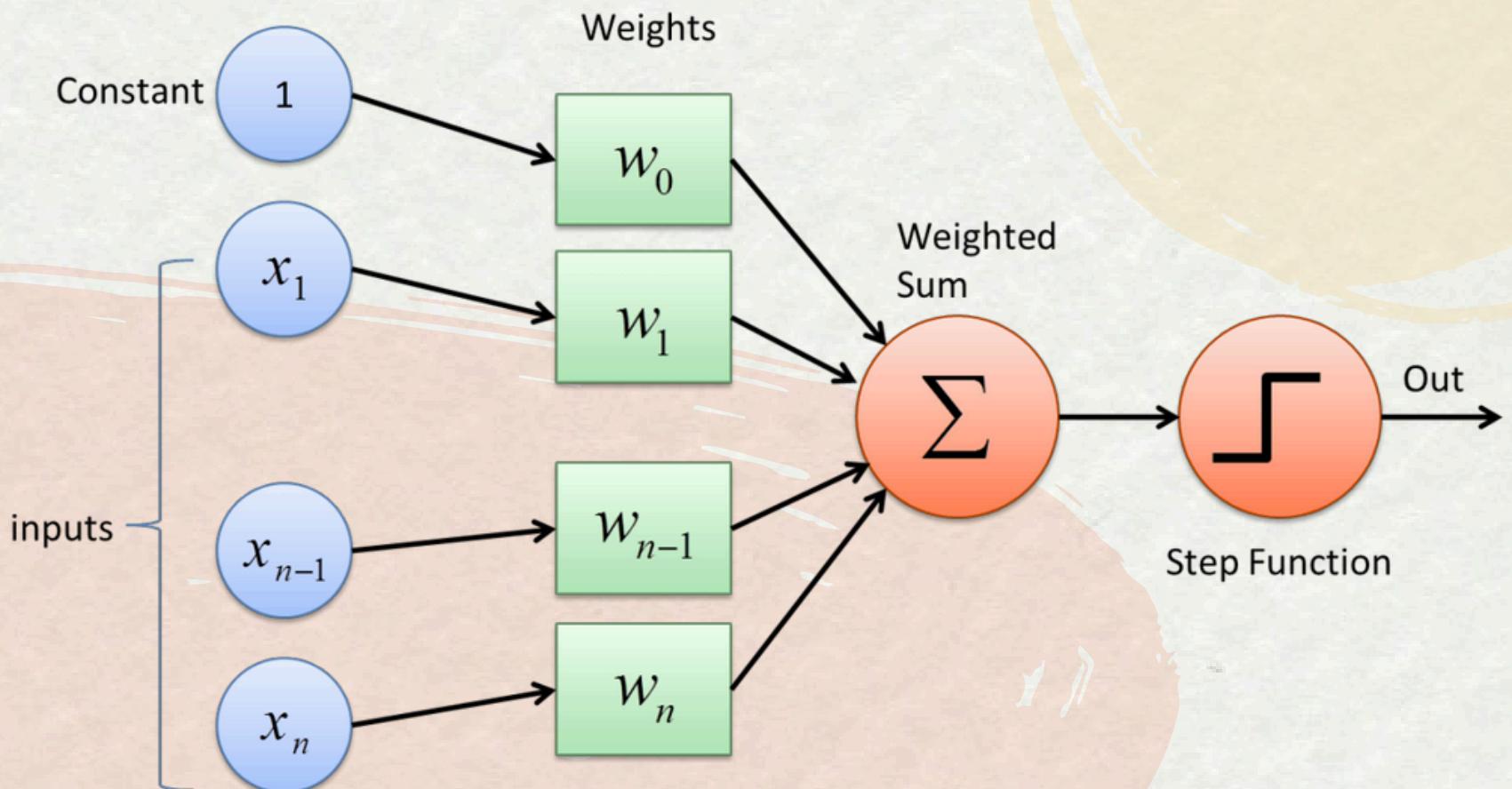


Accuracy Training, 89%

Accuracy Test, 83%

# ANN - Artificial Neural Network

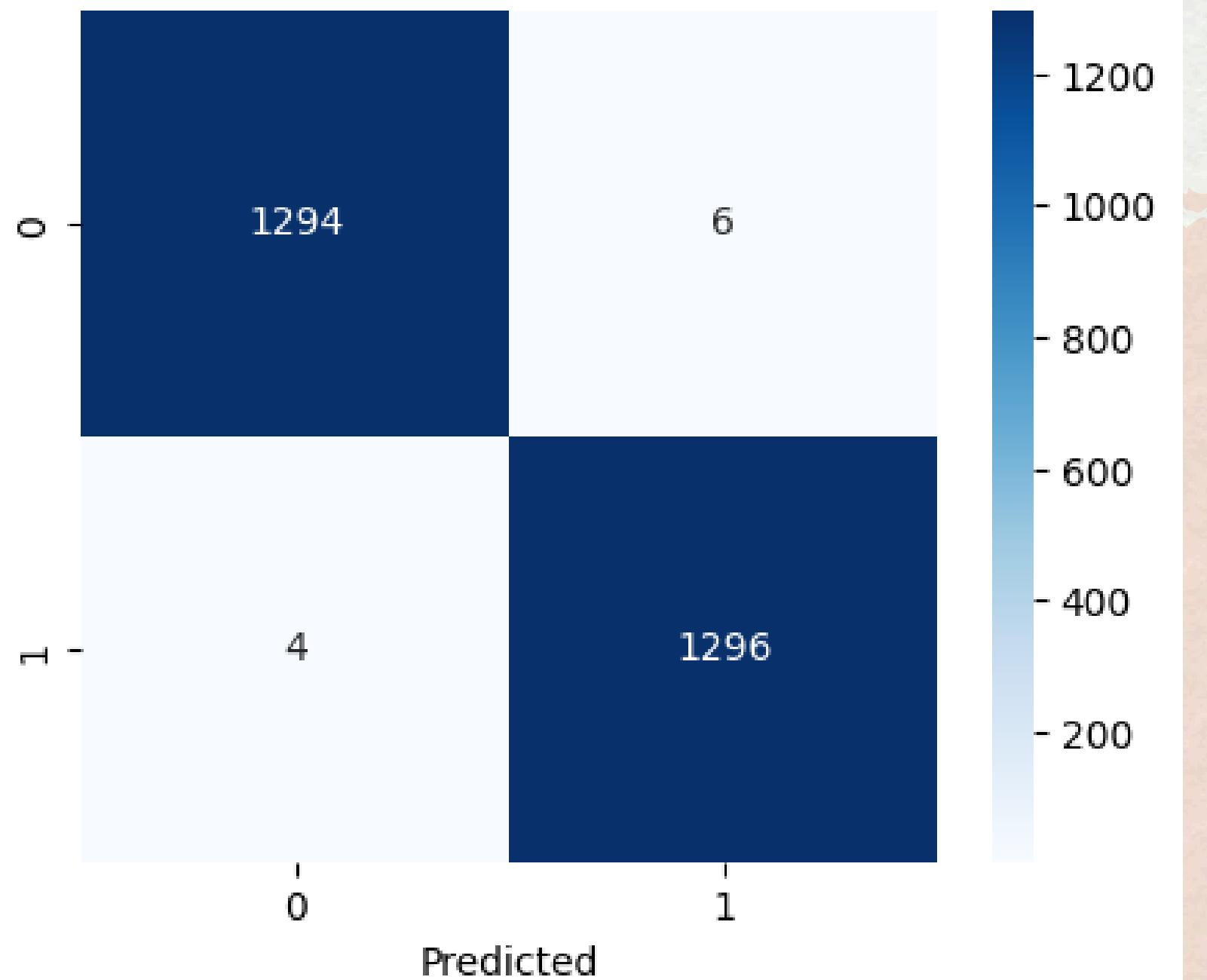
- Modello ispirato al cervello umano, formato da **neuroni artificiali organizzati in strati**.
- Ogni **neurone** :
  - Elabora **input**
  - Applica una **funzione di attivazione (ReLU)**:  
 $x > 0 \rightarrow$  invariato,  $x < 0 \rightarrow 0$ )
  - Trasmette il risultato allo strato successivo
- **Il Perceptron:**
  - Neurone base
    - Calcola una **somma pesata** degli input + **bias**
    - Applica una funzione di attivazione.
- Multi-Layer Perceptron (**MLP**)
  - Combinazione di Perceptron in strati: **Input → Nascosto (50 neuroni) → Output**
  - Impara ottimizzando i pesi tramite **backpropagation**



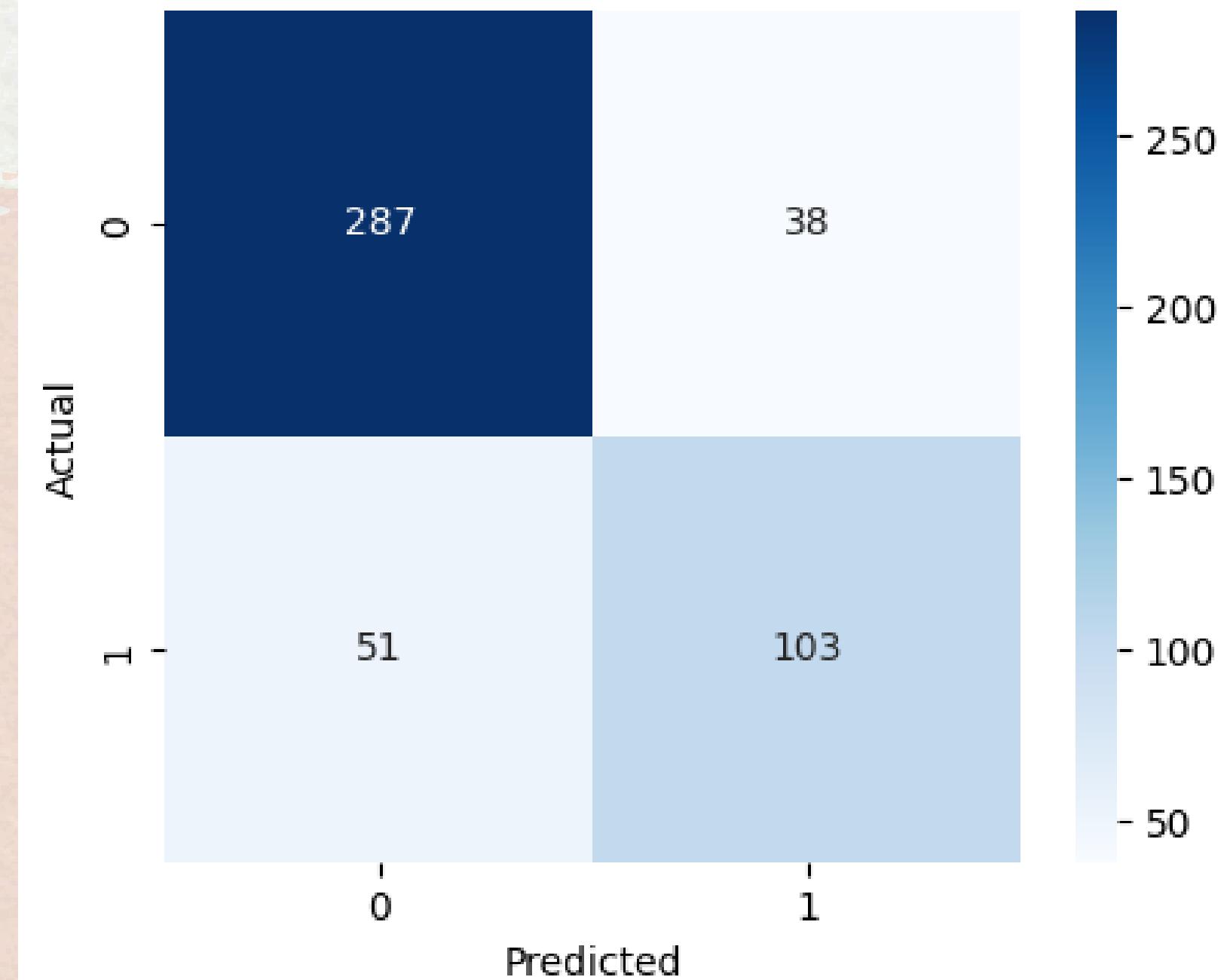
$$y = f \left( \sum_{i=1}^n w_i x_i + b \right)$$

# ANN- Artificial Neural Network

Confusion Matrix - ANN Classifier - Training



Confusion Matrix - ANN Classifier - Test



Accuracy Training, 99,6%

Accuracy Test, 81%

# Confronto tra modelli

Model	Hyperparameters	Training Accuracy	Test Accuracy
<b>Decision Tree (best)</b>	max_depth=5, criterion='gini', random_state=42	0.888077	0.845511
<b>KNN (n_neighbors=30)</b>	n_neighbors=30, metric='hamming'	0.888077	0.828810
<b>ANN</b>	hidden_layer_sizes=(50), activation='relu', so...	0.996154	0.814196

- **Decision Tree (max\_depth=5)**
  - Buona generalizzazione, basso overfitting
  - Leggera difficoltà nel riconoscere la classe GradeClass\_Good
- **k-NN (k=30)**
  - Prestazioni inferiori, overfitting moderato
  - Scarsa discriminazione della classe GradeClass\_Good
- **ANN**
  - Ottima capacità di apprendere pattern complessi
  - Rischio di overfitting più elevato

# Conclusioni

- **Frequent pattern mining** (Apriori) ha evidenziato associazioni significative tra:
  - **Assenze, tempo di studio, supporto familiare e rendimento.**
- **Modelli predittivi** (supervised learning) confermano l'importanza delle stesse variabili, mostrando buona accuratezza nella stima delle performance.
- La **combinazione** di regole associative e modelli predittivi:
  - Offre una **visione completa** dei fattori che influenzano il rendimento.





**THANK YOU!**