ICON 2024/2025

Progetto di Classificazione delle Specie di Pinguini

Pietro Gadaleta - Matricola: 774511 - p.gadaleta6@studenti.uniba.it

A.A. 2024-25

https://github.com/pietrogad/ICON24-25

Indice

1 Introduzione

1.1 Contesto Biologico

I pinguini sono uccelli marini non volatili che presentano significative variazioni morfologiche tra specie diverse. Queste differenze, che includono dimensioni del becco, lunghezza delle pinne e massa corporea, sono adattamenti evolutivi ai diversi ambienti e diete. La classificazione accurata delle specie è fondamentale per studi ecologici come la tassonomia, ovvero lo studio della classificazione degli organismi viventi.

La classificazione incontra, però, diverse difficoltà introdotte dall'ibridazione, creando così esemplari borderline, e convergenza evolutiva, ovvero specie non imparentate che sviluppano caratteristiche simili.

L'utilizzo di tecniche di machine learning migliora i sistemi di classificazione biologica andando così a scovare microdifferenze che l'essere umano non riesce a percepire.

1.2 Obiettivi del Progetto

Il progetto si propone di indentificare il modello che meglio si adatta alla classificazione delle specie tramite utilizzo delle loro caratteristiche morfologiche

1.3 Argomenti di interesse

Di seguito un elenco di argomenti utilizzati all'interno del progetto:

- Apprendimento supervisionato;
- Reti neurali e apprendimento profondo;
- Naive Bayes

2 Dataset e Preprocessing

2.1 Dataset

Il dataset utilizzato contiene 333 osservazioni di pinguini appartenenti a tre specie: **Adelie**,**Gentoo** e **Chinstrap** Per ogni esemplare sono state registrate quattro caratteristiche morfologiche:

- **culmen_length_mm**: Lunghezza del becco
- **culmen_depth_mm**: Profondità del becco
- **flipper_length_mm**: Lunghezza della pinna
- **body_mass_g**: Massa corporea

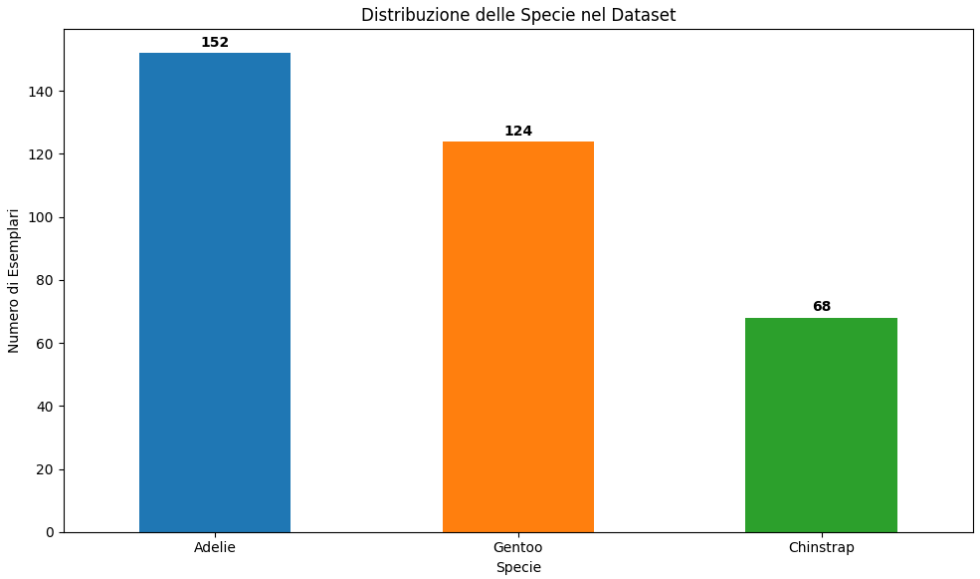

Figura 1: Distribuzione delle specie nel dataset dopo la pulizia

La Figura 1 mostra la distribuzione delle classi nel dataset:

- **Adelie**: 146 esemplari (43.8%);
- **Gentoo**: 119 esemplari (35.7%);
- **Chinstrap**: 68 esemplari (20.4%).

Come è possibile osservare dai risultati registrati, il dataset ha un visibile problema di classi sbilanciate (la specie Chinstrap è rappresentata da solo 68 esemplari contro i 146 delle Adelie e i 124 dei Gentoo) e ciò implica che l'utilizzo della sola accuracy come metrica di valutazione può essere fuorviante. Per far fronte a tale problematica si è deciso di:

- Utilizzare, in aggiunta all'accuracy, anche l'F1-score come metrica di valutazione;
- Applicare la Stratified K-Fold Cross Validation (implicito nello shuffle casuale) in modo da minimizzare il fenomeno del overfitting.

2.2 Preprocessing

Le attività di preprocessing si sono definite sull'utilizzo di quattro fasi:

1. **Gestione dei valori mancanti**: I valori mancanti, indicati nel dataset come NaN, sono stati sostituiti con la media della rispettiva colonna:

$$\text{valore\_sostituito} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{per} \quad x_i \neq \text{NaN}$$

   Questa scelta mantiene la distribuzione originale dei dati senza ridurre la dimensione del dataset.

2. **Normalizzazione**: Poiché nel Dataset potrebbero esserci degli outliers si è deciso di applicare il metodo di normalizzazione Z-Score:
$$z = \frac{x - \mu}{\sigma}$$
con:
   - μ: La media delle feature da normalizzare;
   - σ: La deviazione standard sulla stessa feature.

3. **Codifica delle etichette**: Conversione delle specie, rappresentato come dato categorico, nella corrispondente rappresentazione numerica.
   - Adelie $\rightarrow$ 0
   - ChinStrap $\rightarrow$ 1
   - Gentoo $\rightarrow$ 2

3 Metodi e Modelli utilizzati

3.1 K-Fold Cross Validation

Prima di procedere con l'implementazione, il dataset è stato sottoposto ad un processo di K-fold Cross Validation. I parametri usati per l'implementazione sono:

- n_splits = 5 (indica il numero di split)
- random_state = 42 (indica il seed di randomizzazione)
- shuffle = True

3.2 Decision Tree

Un **Decision Tree** è un modello basato su una struttura ad albero rovesciato composto da:

- **Nodi interni**: etichettati con condizioni booleane basate sulle feature di input. Ogni nodo ha due figli associati agli esiti *true/false*;
- **Nodi foglia**: contengono una stima puntuale per la feature target:
  - Classe (per problemi di classificazione)
  - Valore reale (per problemi di regressione)

Scelte progettuali:

- **Profondità massima**: Limitata a 3 per prevenire overfitting

3.3 Naive Bayes

Il classificatore Naive Bayes si basa sul teorema di Bayes, assumendo indipendenza condizionale tra le features.
Fondamenti matematici:
$$P(y|x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i|y)}{P(x_1, \ldots, x_n)}$$
Dove:

- $P(y)$: Probabilità a priori della classe

- $P(x_i|y)$: Probabilità della feature data la classe

3.4 Artificial Neural Network (ANN)

Le ANN o Artificial Neural Network sono modelli ispirati alle reti neurali biologiche. Tali modelli si basano su strutture non-lineari di dati statistici organizzati come strumenti di modellazione.
In generale la struttura delle ANN si basa sull'utilizzo di:

- Input Layer: layer iniziale della rete che riceve i dati in ingresso. Ogni nodo rappresenta una feature del dataset;
- Hidden Layer: layer intermedi tra input e output, ogni nodo combina gli input, applica una funzione di attivazione e passa il risultato al layer successivo;
- Dropout Layer: layer usato durante l'addestramento per ridurre l'overfitting, disattiva casualmente una percentuale di neuroni impedendo che la rete si adatti troppo ai dati di training;
- Output Layer: ultimo layer, che produce il risultato finale della ANN, risultato comoe una probabilità (classificazione) o un valore numerico (regressione).

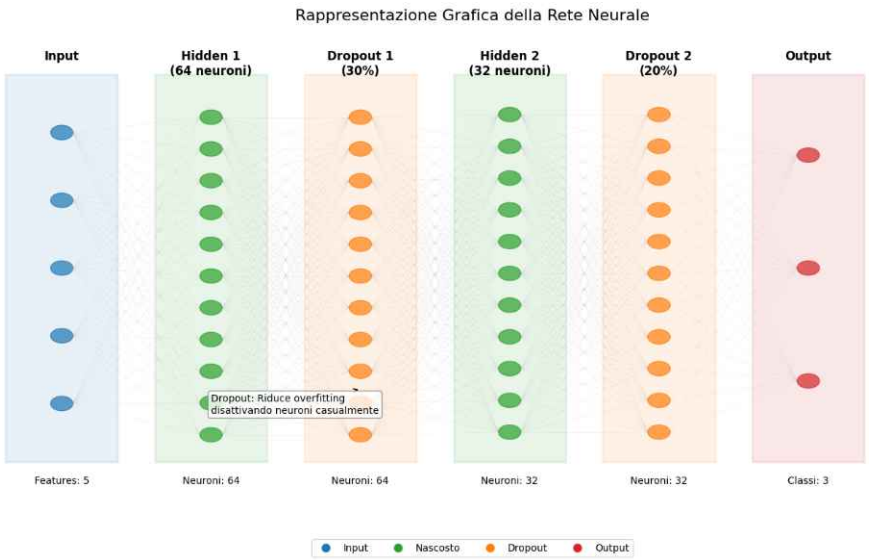

Figura 2: Architettura della ANN

Funzione di attivazione:

- ReLU: $f(x) = \max(0, x)$
- Softmax: $\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$

Loss function:

$$\text{logloss}(p, a) = -\log p[a]$$

4 Metodologia di Valutazione

4.1 Metriche di Valutazione

- **Accuracy**: Percentuale di classificazioni corrette

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **F1-Score (macro)**: Media armonica di precision e recall, bilanciata tra le classi

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5 Risultati

5.1 Analisi Comparativa dei Modelli

La Tabella 1 presenta i risultati comparativi dei tre modelli implementati, con particolare attenzione a due aspetti fondamentali:

Tabella 1: Sintesi delle prestazioni (media)

Modello	Accuracy	F1-Score (macro)
Albero Decisionale	0.9656	0.9638
Naive Bayes	0.9688	0.9652
ANN	**0.9742**	**0.9711**

5.1.1 Considerazioni sulla Tabella

Dall' analisi dei risultati si può notare come **la ANN si sia dimostrato il modello più efficace**, superando sia il Decision Tree che il Naive Bayes in termini di accuratezza e F1-score.

Questo successo è da attribuire alla capacità delle ANN di catturare relazioni non lineari tra le caratteristiche morfologiche; infatti, risulta essere meno sensibile alle variazioni casuali nei dati.

Il Decision Tree ed il Naive Bayes, pur mostrando prestazioni simili tra loro, hanno rilevato alcune limitazioni; infatti, faticano a classificare correttamente classi minoritarie come la specie Chinstrap, le cui condizioni morfologiche si trovano in una situazione borderline tra le altre due specie.

Nota fondamentale: La differenza assoluta tra il miglior modello (ANN) e il peggiore (Decision Tree) è solo 0.86% in accuracy, confermando che tutte e tre gli approcci sono validi per questo specifico problema biologico.

5.2 Analisi delle Matrici di Confusione

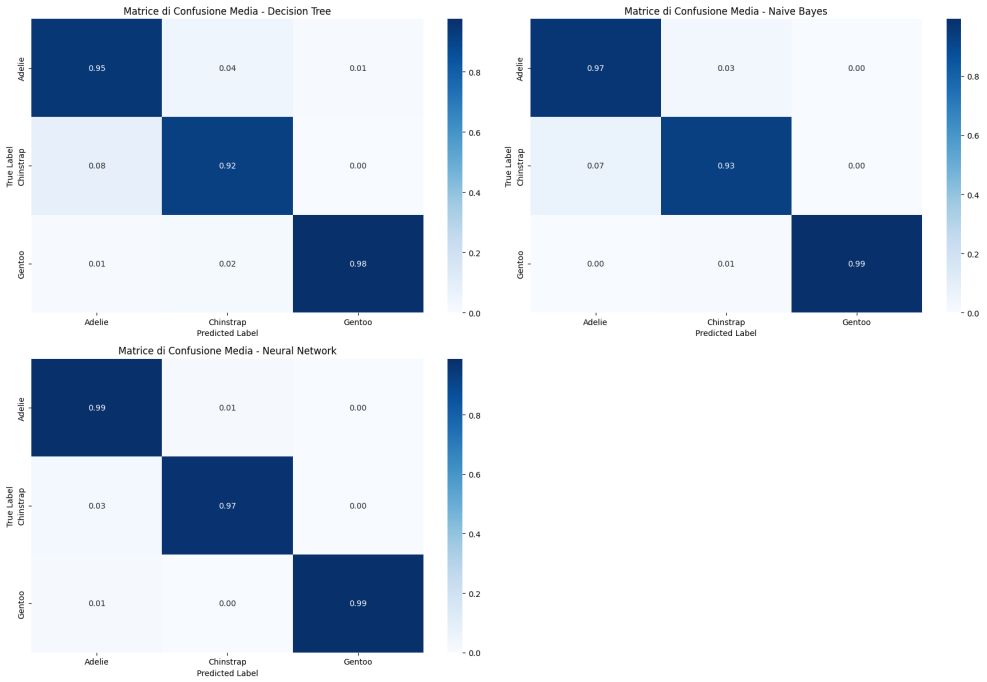

Figura 3: Confusion Matrix normalizzate

Tutti i modelli dimostrano capacità nell'identificare i pinguini *Gentoo*. Questi esemplari vengono riconosciuti correttamente in quasi tutti i casi, con tassi di successo che sfiorano il 100%. Questo risultato non sorprende se consideriamo le marcate differenze morfologiche dei Gentoo rispetto alle altre specie: le loro pinne notevolmente più lunghe e la massa corporea più imponente costituiscono caratteristiche distintive.

Problemi principali emergono nella distinzione tra *Adelie* e *Chinstrap*. I modelli mostrano una certa confusione tra queste due specie. I Chinstrap, infatti, vengono scambiati per Adelie più frequentemente del contrario.

Questo errore suggerisce che:

- Gli esemplari "tipici" di Chinstrap sono ben riconosciuti
- Gli esemplari "atipici" o borderline tendono ad assomigliare più alle Adelie che non ai Gentoo

Questa ambiguità è biologicamente corretta, infatti Adelie e Chinstrap condividono caratteristiche morfologiche simili rispetto ai Gentoo.

5.3 Intepretazione delle caratteristiche per i modelli

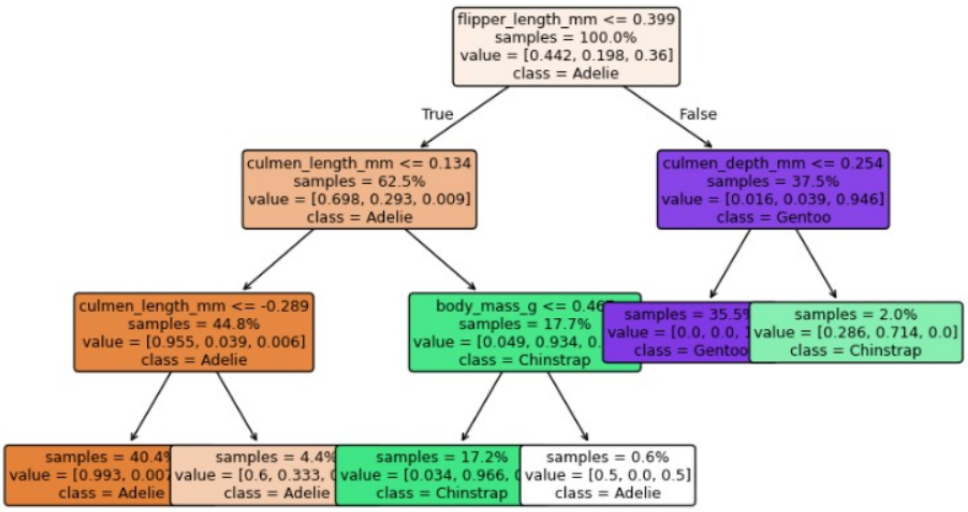

Figura 4: Struttura del Decision Tree

La struttura del DT è rappresentata da nodi e rami come descritto nei paragrafi precedenti. Andando più approfonditamente possiamo vedere come vengono strutturati i nodi:

- Condizione di split: La regola booleana basata su una feature di input e una soglia;
- Samples: Indica la percentuale di campioni del dataset totale che sono "passati" attraverso quel nodo;
- Value: Questa è una lista di numeri che rappresenta la distribuzione delle classi per i campioni presenti in quel nodo;
- Class: La classe più frequente tra i campioni che si trovano in quel nodo.

  In particolare ci concentriamo sulle condizioni di split, per ogni nodo l'algoritmo cerca la migliore caratteristica e la migliore soglia su quella caratteristica per dividere i dati. L'obiettivo è rendere i nodi figli risultanti il più "puri" possibile, cioè con campioni appartenenti in larga parte a una singola classe. Per fare ciò, l'algoritmo valuta diverse possibili divisioni utilizzando un criterio di purezza. Il criterio più comune per gli alberi decisionali di classificazione è l'impurezza di Gini

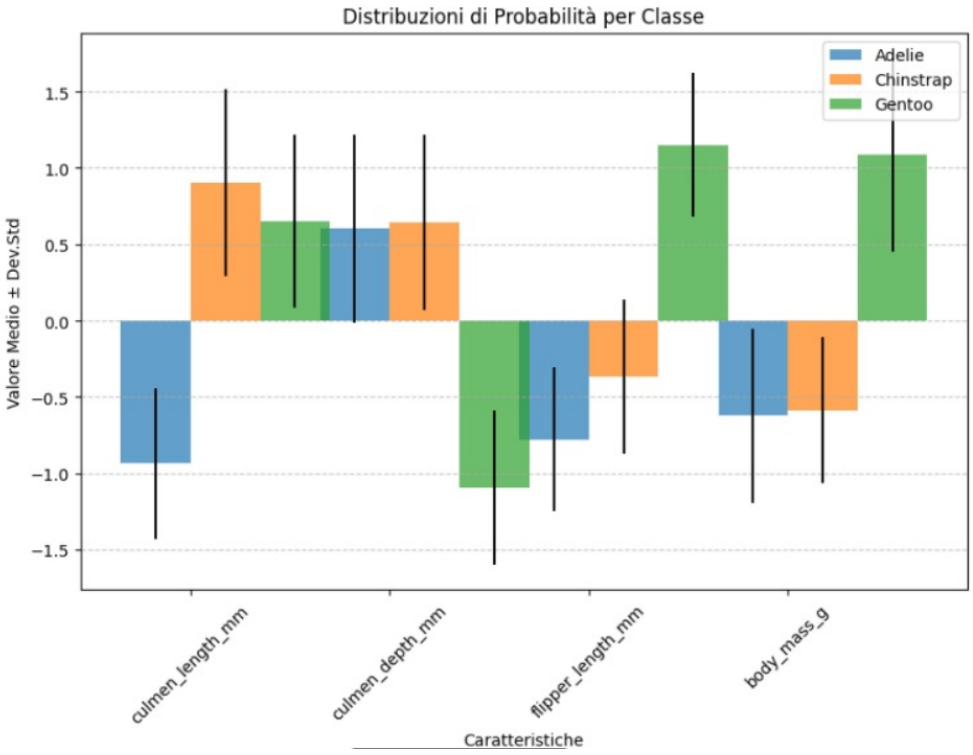

Figura 5: Distribuzione del Naive Bayes

Questo grafico visualizza le distribuzioni di probabilità delle diverse caratteristiche (features) per ciascuna classe di pinguino, come sono state stimate dal modello Naive Bayes durante la fase di addestramento. Il grafico si struttura nella seguente maniera:

- Asse X (Caratteristiche): Mostra le quattro caratteristiche morfologiche dei pinguini utilizzate nel dataset:
  - culmen_length_mm (lunghezza del becco);
  - culmen_depth_mm (profondità del becco);
  - flipper_length_mm (lunghezza della pinna);
  - body_mass_g (massa corporea);

  Queste caratteristiche sono state normalizzate utilizzando il metodo Z-Score, il che spiega la presenza di valori sia positivi che negativi sull'asse Y. Questo processo rende le diverse misurazioni confrontabili.
- Asse Y (Valore Medio ± Dev.Std):
  - Barre colorate: Il centro di ogni barra rappresenta il valore medio (μ) della caratteristica specifica per quella determinata classe di pinguino;
  - Linee verticali nere: Queste linee indicano la deviazione standard (σ) dalla media.

Per ogni caratteristica e per ogni classe, il modello stima una distribuzione di probabilità. Questo grafico mostra le medie e le deviazioni standard di queste distribuzioni stimate.

Quando un nuovo pinguino deve essere classificato, il modello attinge alle distribuzioni di probabilità che ha stimato per determinare quanto sia probabile che le misurazioni di quel pinguino si presentino data ciascuna delle possibili specie.

5.3.1 Confronto tra i modelli

- **Naive Bayes**: i Chinstrap vengono scambiati per Adelie nel 7% dei casi, mentre l'errore dal lato opposto è solo del 3%. Questo suggerisce che il modello tende a "assorbire" gli esemplari borderline di Chinstrap nella categoria Adelie.

- **Decision Tree**: presenta la performance più debole in questo confronto, con un tasso di confusione leggermente più alto (8% Chinstrap→Adelie; 4% Adelie→Chinstrap). L'albero commette anche qualche raro errore nel classificare Adelie come Gentoo (1%).
- **ANN**: si distingue nettamente, infatti riduce al minimo la confusione tra le due specie simili: solo 1% delle Adelie viene scambiato per Chinstrap e solo il 3% dei Chinstrap viene classificato come Adelie.

6 Conclusioni

Possiamo concludere, quindi, che per il contesto preso in considerazione l'approccio basato sull'utilizzo delle **ANN** permette di ottenere **risultati globalmente migliori**, rispetto ai metodi che utilizzano un approccio con Decision Tree e Naive Bayes.

Come si può anche notare dall'analisi della Precision, Recall e F1-score, l'ANN risulta avere **performance stabili a 0.97**, che sono superiori a quelle dei restanti metodi, le quali si aggirano **intorno a 0,96**. La coerenza tra gli errori dei modelli e le effettive relazioni evolutive tra le specie valida l'approccio complessivo.