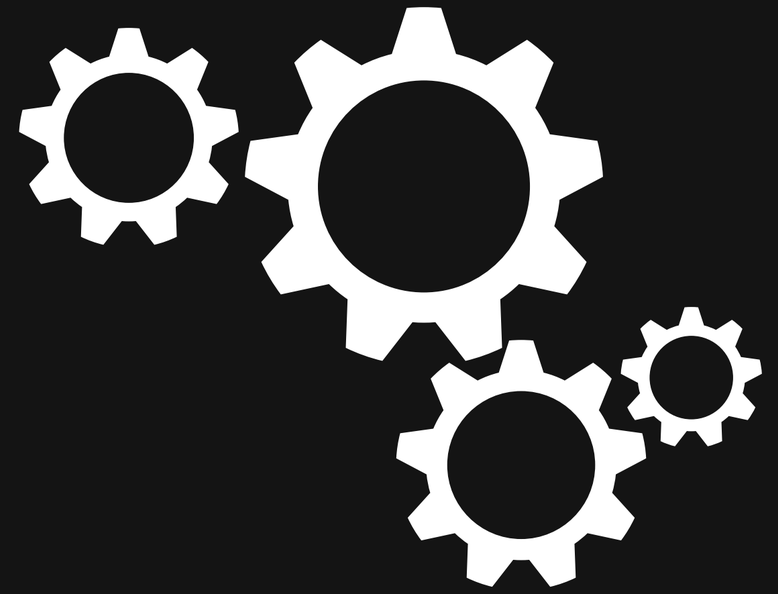
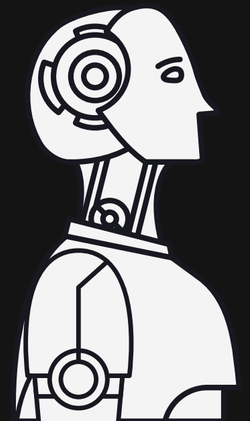
USARE LA BWT PER LA GENERAZIONE DI FEATURES PER MIGLIORARE ALGORITMI DI MACHINE LEARNING PER LA CLASSIFICAZIONE DI SEQUENZE BIOLOGICHE CHE PRESENTANO GENE FUSION





VINCENZO TARANTINO

BARTOLOMEO MAZZEO

Indice

Machine learning

Il machine learning è un insieme di tecniche che vengono usate per addestrare le macchine tramite i dati per effettuare classificazione o regressione. Ne esistono diversi tipi:

- Supervisionato: Il modello impara da dati etichettati
- Non supervisionato: Il modello trova pattern nei dati senza etichette (clustering)
- Apprendimento per rinforzo: Il modello migliora le sue decisioni in base a ricompense

Applicazioni del machine in Bionformatica

Il Machine Learning (ML) viene utilizzato nella bioinformatica per analizzare grandi quantità di dati biologici, in particolare sequenze genomiche, con l'obiettivo di individuare pattern e caratteristiche rilevanti. Grazie a questi modelli, è possibile apprendere dai dati di addestramento e applicare le conoscenze acquisite per classificare nuove sequenze.

GENE FUSION

Il genoma è l'insieme completo del DNA di un organismo, che contiene tutte le informazioni genetiche necessarie per il suo sviluppo e funzionamento. Esso è composto da geni, ovvero segmenti di DNA che codificano per proteine. In alcuni casi, può verificarsi una fusione genica, un fenomeno in cui due geni distinti si uniscono, portando alla produzione di proteine anomale.
Queste proteine possono alterare i processi cellulari e contribuire allo sviluppo di malattie, come il cancro.

Il Machine Learning è utile per individuare sequenze che presentano quest’anomalia

DATI A DISPOSIZIONE

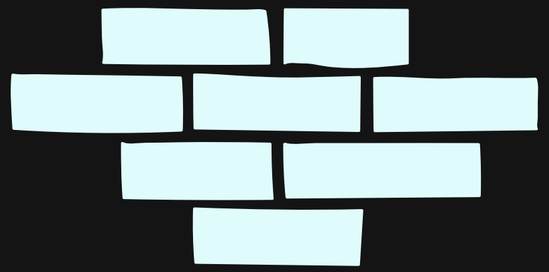
File all_transcipts.fasta ⟶ Sequenze senza gene fusion

File fusim_bench.fasta ⟶ Sequenze con gene fusion

MANIPOLAZIONE DATASET PER OTTENERE FEATURES

- Adattamendo della lunghezza delle sequenze senza gene fusion a quelle con gene fusion
- Bilanciamento del numero di sequenze senza gene fusion a quelle con gene fusion

Suddivisione delle sequenze lunghe (in all_transcripts) in blocchi di lunghezza simile a quelle di fusim_bench e selezione di più blocchi della stessa sequenza



APPLICAZIONE BWT

La Burrows-Wheeler Transform (BWT) è un algoritmo utilizzato per la compressione dei dati e la manipolazione di sequenze testuali.
La BWT trasforma una stringa S in una nuova rappresentazione che raggruppa caratteri simili, migliorando la compressione e la ricerca di pattern. Il processo si basa su tre passaggi principali:

1. Generazione delle rotazioni cicliche di S.
2. Ordinamento lessicografico delle rotazioni.
3. Estrazione dell'ultima colonna della matrice ordinata, che costituisce la BWT di S

ESTRAZIONE B-MERS

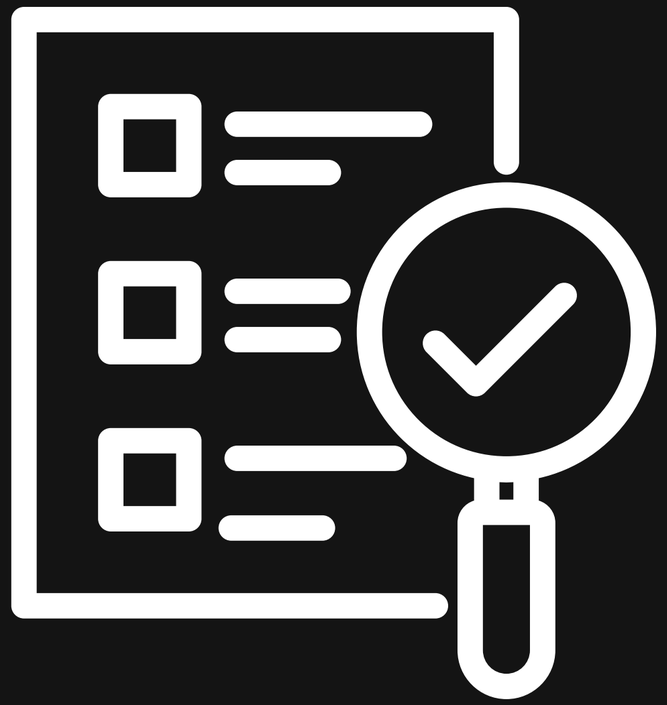
Avere la BWT delle sequenze è utile perché possiamo usarla per estrarre le features che verranno usate per addestrare i modelli di ML. Le features che estraiamo sono i b-mers, sono sottosequenze di lunghezza variabile estratte direttamente dalla rappresentazione BWT.

Per estrarli:

- Si individuano i caratteri ripetuti consecutivamente nella BWT.
- Per ciascuna ripetizione, si seleziona l'intervallo corrispondente nelle rotazioni ordinate.
- Si identifica il prefisso comune a tali rotazioni e si costruisce il b-mer concatenando il carattere ripetuto con il prefisso.

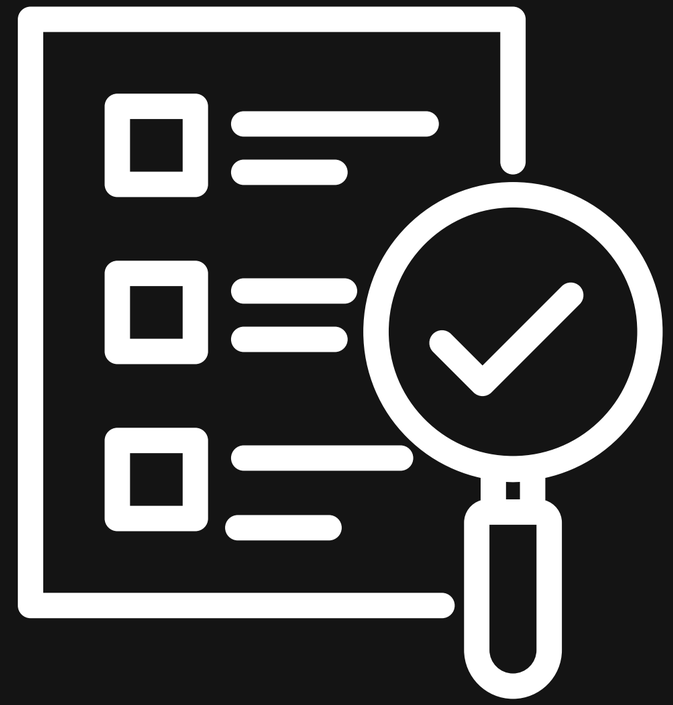
CREAZIONE DATASET FINALE (1)

A questo punto creiamo i vettori numerici per ogni sequenza costrunedo un dataset che ha come righe tutti gli id delle sequenze, sulle colonne i b-mers e per ogni cella il numero di volte che un bmer j-esimo compare nella sequenza i-esima.



CREAZIONE DATASET FINALE (2)

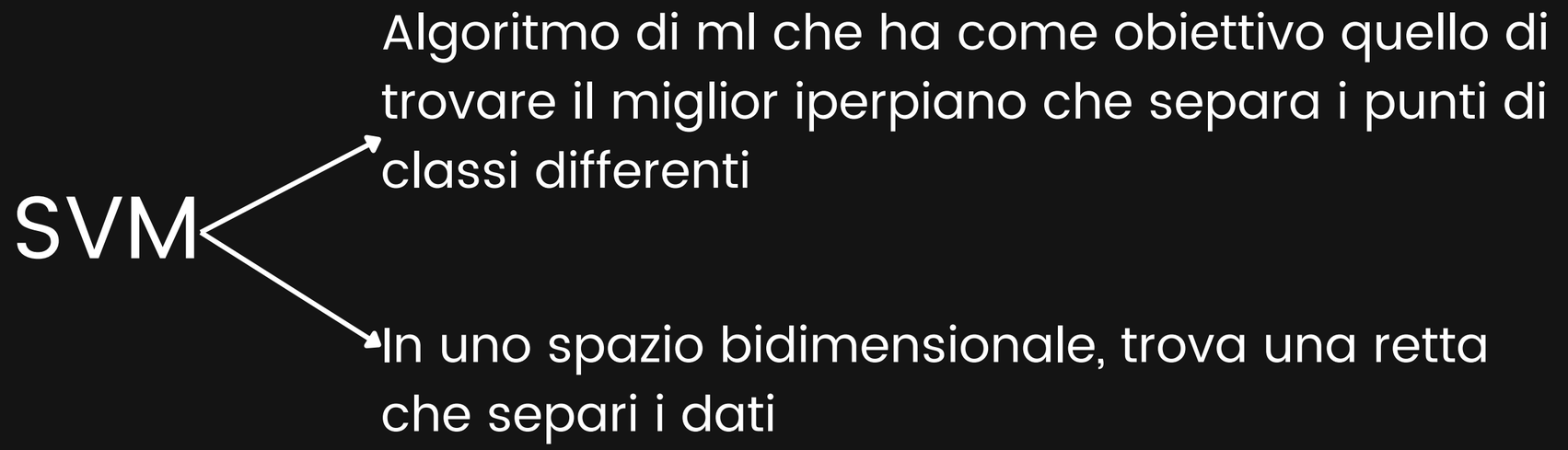
Per ridurre ulteriormente il numero di b-mers considerati applichiamo l'ECCD per considerare solo quelli che hanno più impatto sul livello di incertezza del dataset



- Calcolo entropia globale (elementi bilanciati nelle due classi)
- Calcolo entropia condizionata per ogni bmer
- Entropia globale - entropia condizionata

ADDESTRAMENTO MODELLI - SVM

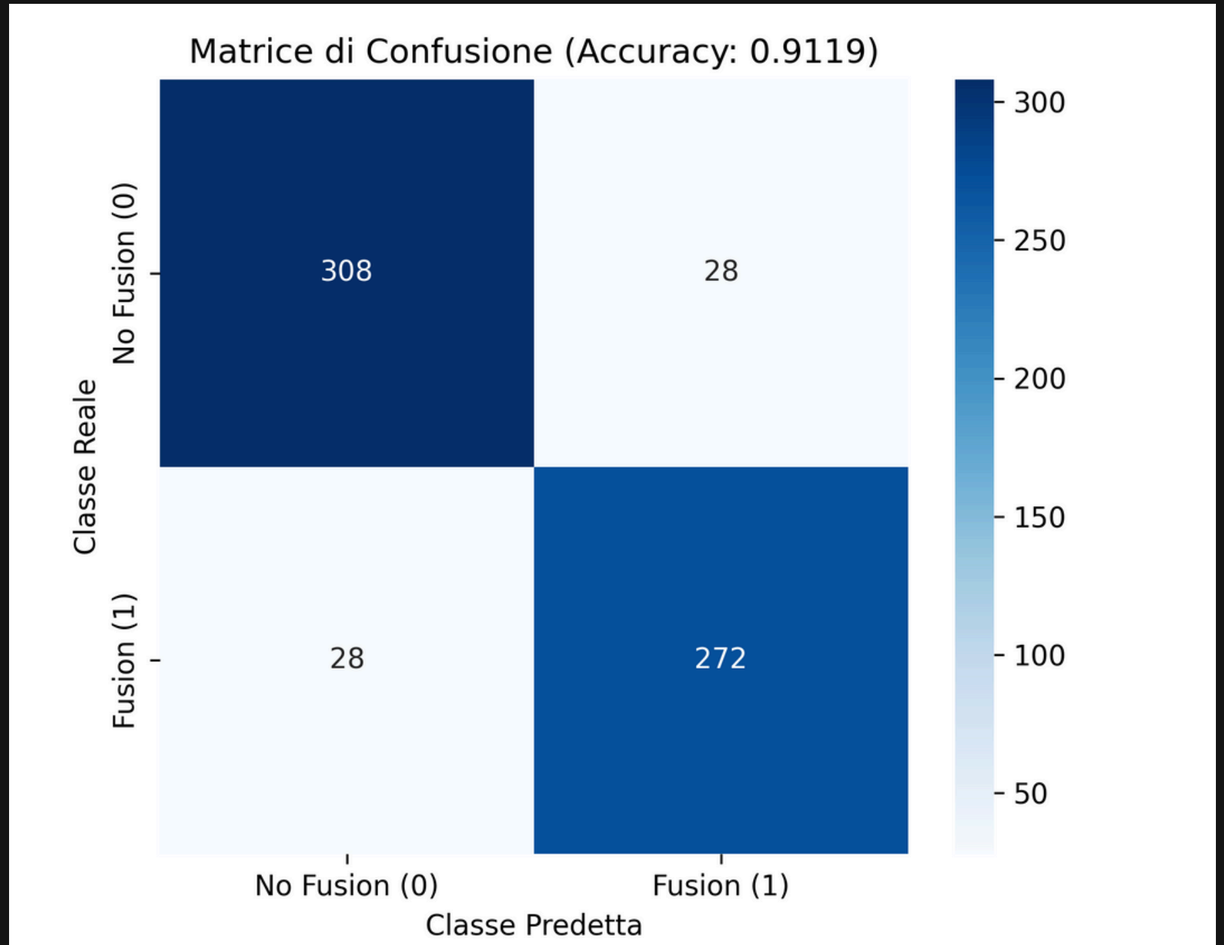
I dati sono stati divisi in dati di train e test e forniti ad una SVM



Risultati ottenuti - SVM

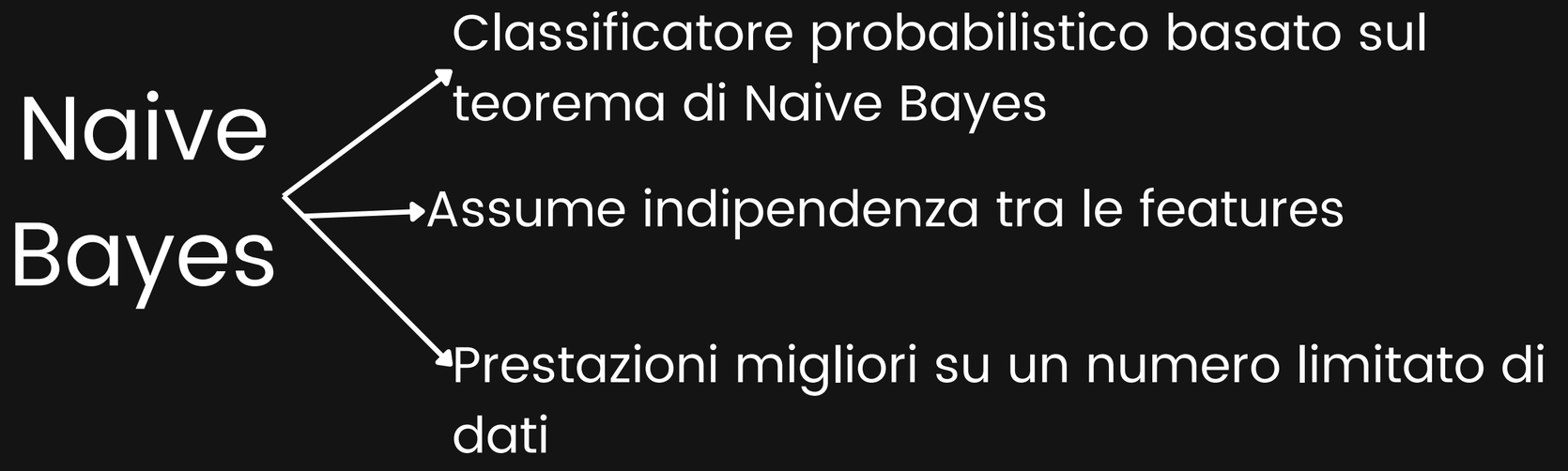
Accuracy: 91.19%

Il modello ha classificato correttamente circa il 91% dei dati



ADDESTRAMENTO MODELLI - NAIVE BAYES

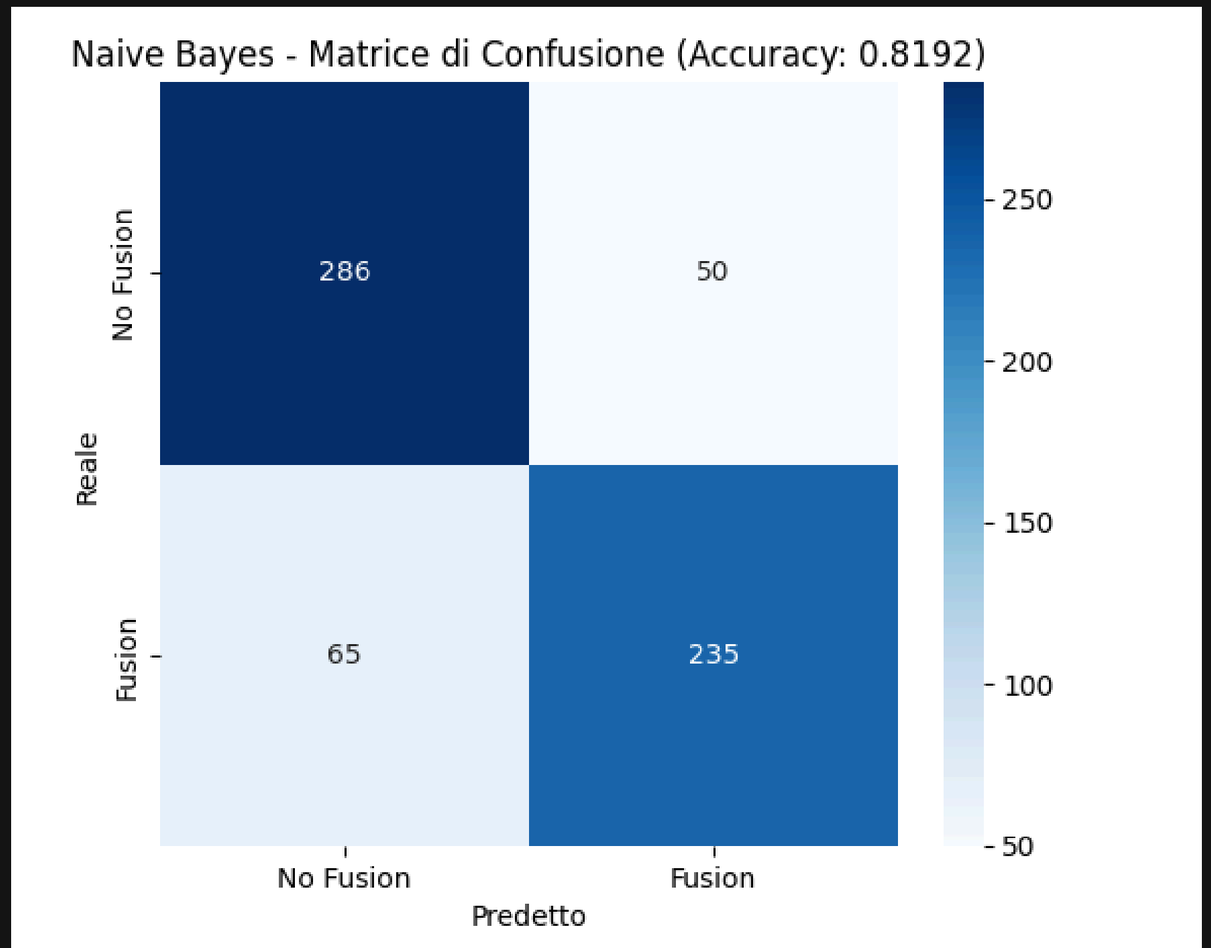
I dati sono stati divisi in dati di train e test e forniti ad un Naive Bayes



Risultati ottenuti - NAIVE BAYES

Accuracy: 81.92%

Il modello ha classificato correttamente circa il 82% dei dati



ADDESTRAMENTO MODELLI - RANDOM FOREST

I dati sono stati divisi in dati di train e test e forniti ad una Random Forest

Random forest

Questo tipo di modelli usano più gli alberi decisionali per effettuare predizioni e classificazioni

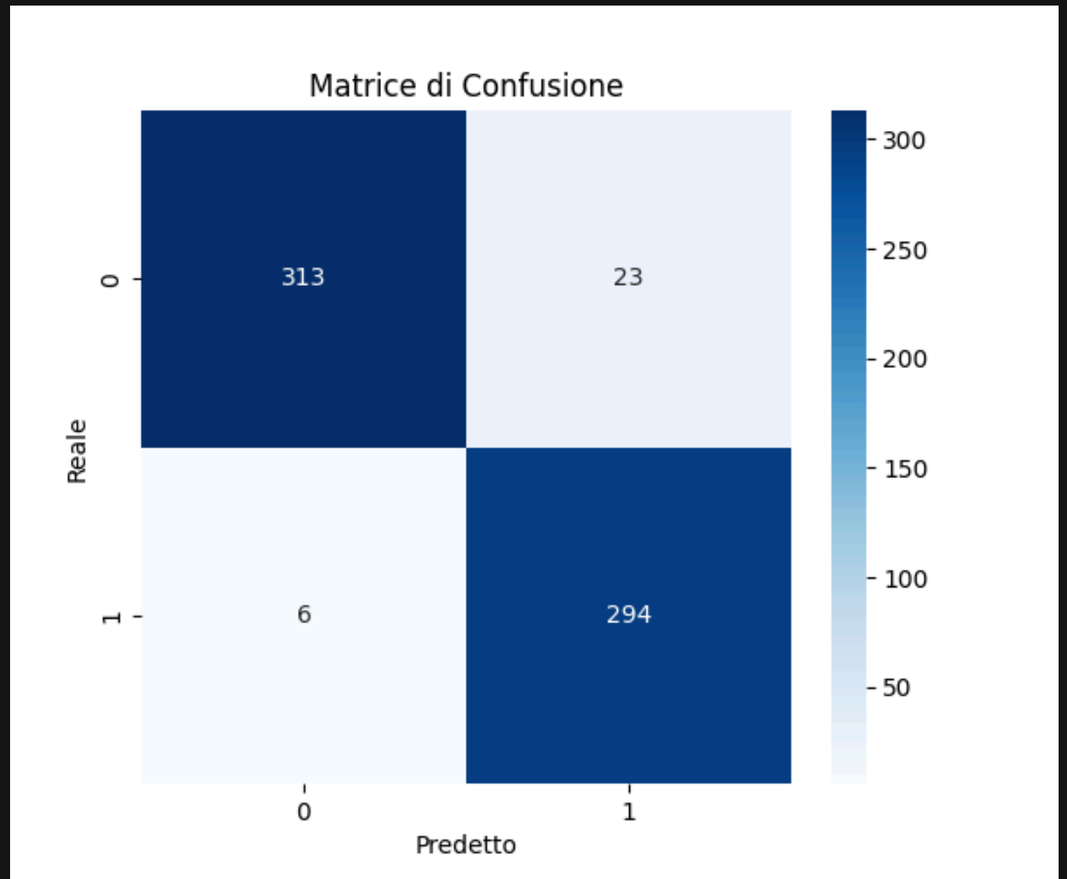
Ogni albero analizza un sottoinsieme casuale di features ed effettua le sue scelte (ad esempio in base al numero di volte che compare un b-mer)

Il risultato finale sarà ottenuto combinando le decisioni prese dai vari alberi tenendo conto della maggioranza

Risultati ottenuti - RANDOM FOREST

Accuracy: 95.44%

Il modello ha classificato correttamente circa il 95% dei dati



ADDESTRAMENTO MODELLI - RETI NEURALI

I dati sono stati divisi in dati di train e test e forniti ad una Rete neurle

Reti Neurali

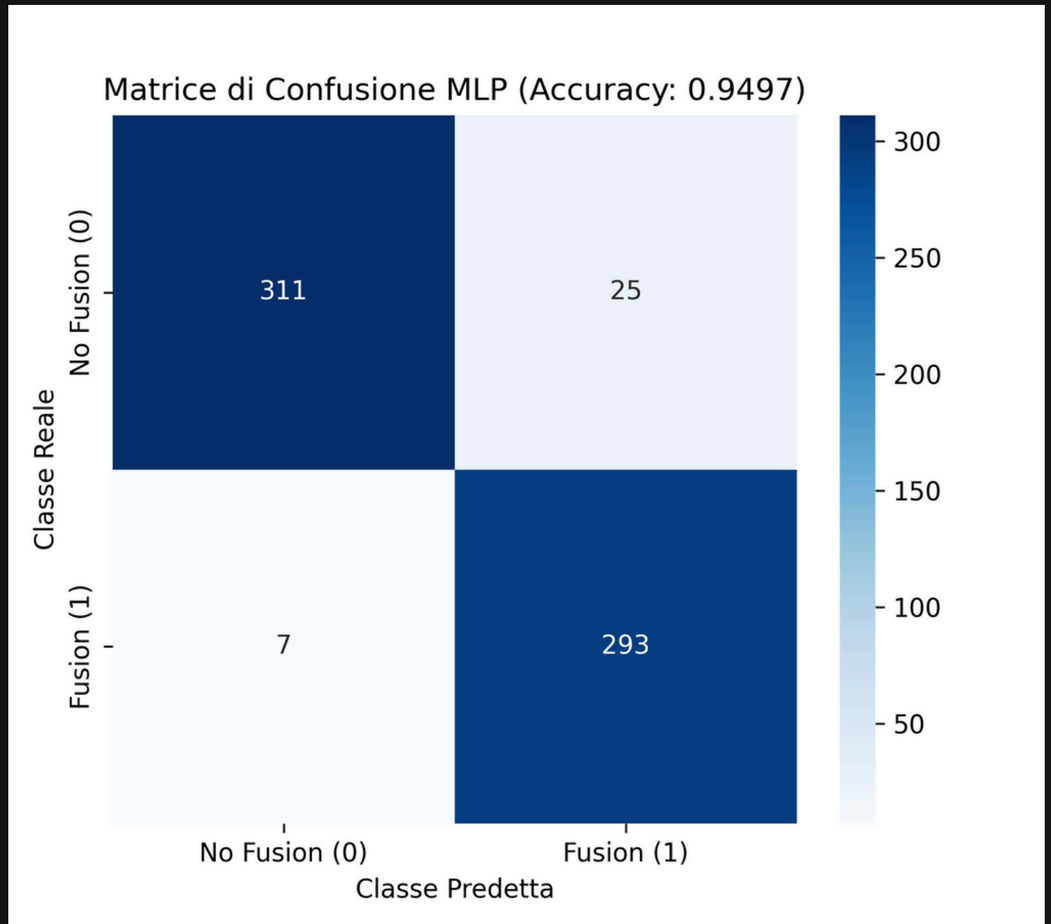
Cercano di simulare il ragionamento umano tramite neuroni artificiali divisi in strati.
Il primo riceve le informazioni (input), il secondo le elabora per individuare pattern (nascosto), il terzo assegna la decisione (strato di output)

I risultati possono essere influenzati dal numero di neuroni usati nello strato nascosto, dalla funzione di attivazione e dal numero di epoche (quante volte il modello viene addestrato)

Risultati ottenuti - RETI NEURALI

Accuracy: 94.97%

Il modello ha classificato correttamente circa il 95% dei dati



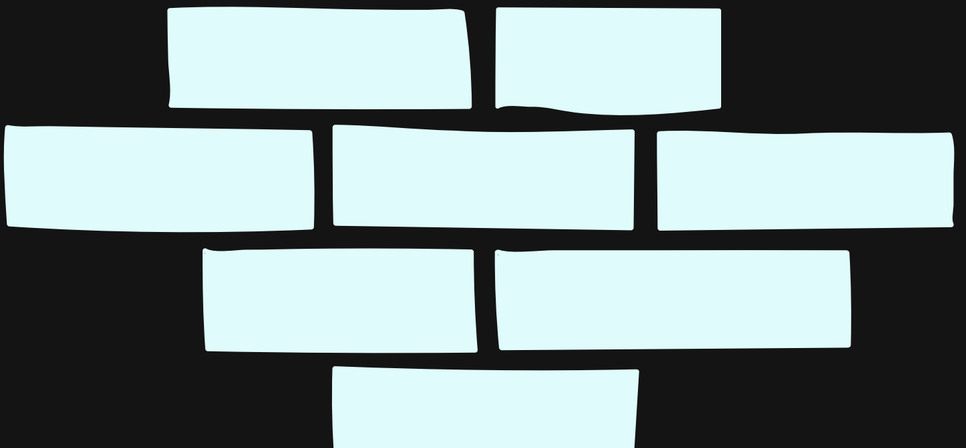
CONCLUSIONI

I risultati ottenuti dimostrano che la Burrows-Wheeler Transform (BWT) ha permesso di estrarre features significative per la classificazione delle sequenze genomiche con e senza gene fusion. Tutti i modelli di Machine Learning testati hanno raggiunto buone prestazioni, confermando l'efficacia dei b-mers generati tramite la trasformazione BWT.
In risultati migliori sono stati ottenuti usando random forest e reti neurali.

Modello	Accuracy	Precision	Recall	F1-score
SVM	0.9119	0.91	0.91	0.91
Naive Bayes	0.8192	0.82	0.78	0.80
Random Forest	0.9544	0.96	0.95	0.95
Rete Neurale	0.9497	0.95	0.95	0.95

Sviluppi futuri: approccio diverso nella selezione delle features e classificazioni che includono più classi

GRAZIE PER L’ATTENZIONE



v.tarantino7@gmail.com
b.mazzeo3@gmail.com