

# Sistema Ibrido per lo Screening dell'Autismo basato su Machine Learning e Ontologie

Membri:

Cognome e nome: Storelli Leonardo, Matricola:758472

Link GitHub: <https://github.com/GongoTheBongo/Progetto-ICON-a.a-23-24.git>

A.A: 2023/2024

## Indice:

<b>1. INTRODUZIONE .....</b>	<b>3</b>
<b>2. IL DATASET E ANALISI ESPLORATIVA .....</b>	<b>4</b>
2.1 Descrizione dei Dati .....	4
2.2 Analisi Grafica (EDA) .....	4
<b>3. ONTOLOGIA (RAPPRESENTAZIONE DELLA CONOSCENZA).....</b>	<b>5</b>
3.1 Modellazione del Dominio .....	5
3.2 Interrogazione Semantica (Querying) .....	6
<b>4. APPRENDIMENTO SUPERVISIONATO: ANALISI DEGLI ALGORITMI ..</b>	<b>7</b>
4.1 K-Nearest Neighbors (KNN) .....	7
4.2 Random Forest (RF) e Feature Importance .....	8
4.3 Support Vector Machine (SVM) - BEST MODEL .....	9
<b>5. ESPERIMENTO DI SELEZIONE FEATURE (ALL vs TOP-3) .....</b>	<b>11</b>
5.1 Configurazione dell'Esperimento .....	11
5.2 Risultati e Discussione.....	11
<b>6. APPRENDIMENTO NON SUPERVISIONATO (CLUSTERING).....</b>	<b>12</b>
6.1 Analisi K-Means .....	12
<b>7. ARCHITETTURA SOFTWARE E IMPLEMENTAZIONE .....</b>	<b>13</b>
<b>8. CONCLUSIONI E SVILUPPI FUTURI.....</b>	<b>13</b>

# 1. INTRODUZIONE

- **Obiettivo del progetto:** Sviluppare un sistema di supporto decisionale (DSS) per la diagnosi precoce dei disturbi dello spettro autistico (ASD). Il sistema mira a classificare i pazienti sulla base delle risposte al test AQ-10 e dati demografici.
- **Approccio Metodologico:** Il progetto adotta un approccio ibrido che combina:
  - **Machine Learning Supervisionato:** Per la predizione della diagnosi (Classificazione).
  - **Machine Learning Non Supervisionato:** Per l'individuazione di pattern nascosti nei dati (Clustering).
  - **Ingegneria della Conoscenza:** Uso di un'ontologia OWL per rappresentare semanticamente il dominio e permettere interrogazioni complesse.

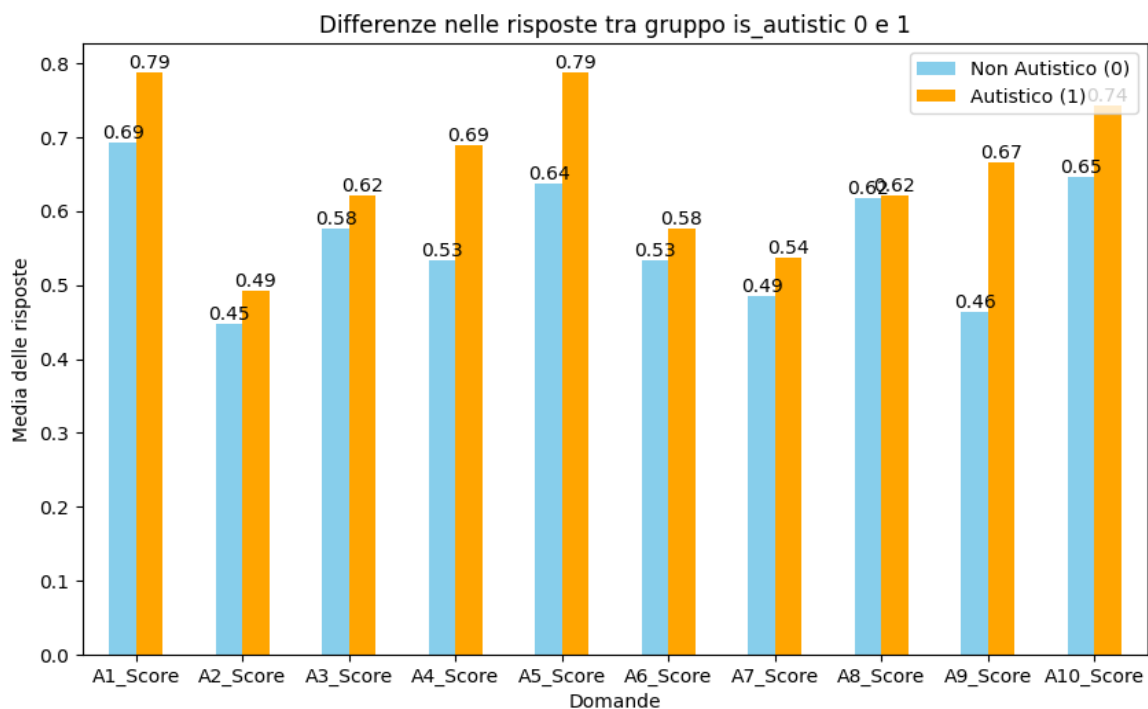
## 2. IL DATASET E ANALISI ESPLORATIVA

### 2.1 Descrizione dei Dati

- **Fonte:** Il dataset (Autism-Dataset.csv) raccoglie dati di screening relativi a pazienti, includendo:
  - **Feature Comportamentali:** 10 domande con risposta binaria (A1\_Score... A10\_Score).
  - **Dati Demografici:** Età, genere, etnia, paese di residenza.
  - **Storia Clinica:** Casi di ittero alla nascita (jundice), casi di PDD in famiglia.
  - **Target:** La variabile Class/ASD (o is\_autistic) che indica la diagnosi finale.
- **Preprocessing e Pulizia:**
  - Conversione delle variabili categoriche tramite **One-Hot Encoding**.
  - Pulizia della colonna target is\_autistic per garantire valori binari (0/1).
  - **Gestione dei Dati Mancanti (Missing Values):** Per garantire che i modelli di Machine Learning possano elaborare correttamente tutti i record, è stata adottata una strategia di imputazione numerica:
    - **Target (Class/ASD):** Le righe prive di diagnosi o con etichette non valide sono state rimosse dal dataset, poiché non utilizzabili per l'addestramento supervisionato.
    - **Feature (Domande e Demografiche):** Per le variabili predittive, i valori mancanti (NaN) generati durante la conversione numerica sono stati sostituiti con la **moda** della relativa colonna. Questa scelta permette di preservare la distribuzione statistica generale dei dati minimizzando la perdita di informazioni.

### 2.2 Analisi Grafica (EDA)

- Per comprendere quali domande discriminano meglio tra soggetti autistici e neurotipici, è stato generato un grafico delle medie delle risposte.



*Descrizione: Sull'asse X sono presenti le domande (A1-A10); l'asse Y mostra la media delle risposte (frequenza del valore "1"). Le barre colorate distinguono tra gruppo diagnosticato (Autistico) e controllo. Si evidenzia visivamente che le domande A1, A4, A5 e A9 mostrano le discrepanze maggiori tra i due gruppi.*

### 3. ONTOLOGIA E RAPPRESENTAZIONE SEMANTICA DEL DOMINIO

La componente di Ingegneria della Conoscenza di questo progetto mira a superare i limiti della semplice analisi numerica, fornendo una struttura semantica formale che descrive le relazioni tra pazienti, test di screening e diagnosi. L'ontologia è stata sviluppata utilizzando il linguaggio standard **OWL (Web Ontology Language)**.

#### 3.1 Modellazione del Dominio (Tassonomia)

L'ontologia (ontologia.owl) formalizza la conoscenza medica contenuta nel dataset attraverso una tassonomia di classi che rispecchia il flusso logico dello screening. La struttura è articolata su tre entità principali:

1. **Paziente:** È la classe centrale che rappresenta il soggetto sottoposto a screening. Funge da nodo aggregatore per tutte le informazioni demografiche, cliniche e per i risultati diagnostici. Ogni individuo di questa classe corrisponde univocamente a un record del dataset.

2. **Test:** Rappresenta l'evento specifico di screening (il questionario AQ-10 compilato). Questa modellazione permette di separare concettualmente la persona dall'atto medico, consentendo teoricamente di associare più test allo stesso paziente in momenti diversi.
3. **Domanda:** Rappresenta i singoli item del questionario. Questa classe permette di modellare nel dettaglio il contenuto del test, collegando ogni sessione di screening alle specifiche domande somministrate.

## 3.2 Proprietà e Relazioni

Le proprietà definiscono gli attributi dei concetti e le relazioni logiche tra di essi, trasformando l'elenco di classi in un grafo connesso.

### 3.2.1 Data Properties (Attributi dei Dati)

Queste proprietà mappano i valori grezzi del dataset sugli individui dell'ontologia, tipizzandoli (es. intero, booleano, stringa):

- **Attributi del Paziente:**
  - age: L'età anagrafica del soggetto.
  - gender: Il genere dichiarato.
  - ethnicity: L'etnia di appartenenza.
  - jundice: Valore booleano che indica la presenza di ittero alla nascita (noto fattore di rischio).
  - isAutistic: La proprietà target fondamentale che rappresenta la diagnosi finale (Positivo/Negativo).
  - used\_app\_before: Indica se l'utente ha familiarità con applicazioni di screening.
- **Attributi del Test:**
  - punteggio (screening score): Il risultato numerico calcolato dal test (valore da 0 a 10).
  - CompilatoreTest: Specifica chi ha compilato il questionario (es. il paziente stesso, un genitore, un operatore sanitario).
  - IdTest: Un identificativo univoco per la tracciabilità dell'esame.

### 3.2.2 Object Properties (Relazioni Semantiche)

Le *Object Properties* sono i "ponti" che collegano le istanze delle diverse classi, permettendo la navigazione del grafo di conoscenza:

- **did\_test (Ha effettuato il test):** Questa relazione collega un individuo della classe Paziente all'individuo della classe Test corrispondente. È il legame semantico che permette di risalire dal risultato clinico alla persona.
- **has\_question (Include la domanda):** Collega un Test alle istanze della classe Domanda, definendo la struttura del questionario somministrato.

### 3.3 Strumenti di Sviluppo e Popolamento

- **Editor:** La struttura concettuale (T-Box) è stata progettata e verificata utilizzando **Protégé**, l'ambiente di riferimento per l'editing di ontologie semantiche.
- **Popolamento (A-Box):** Il popolamento dell'ontologia è avvenuto in modo automatico mappando le righe del file CSV in triple RDF/OWL. Ogni paziente nel dataset è stato istanziato come individuo nell'ontologia, con le relative proprietà valorizzate dai dati reali.

### 3.4 Interrogazione Semantica

L'utilizzo della libreria owlready2 ha permesso di integrare il motore inferenziale direttamente nel flusso applicativo. Il sistema non si limita a leggere i dati, ma esegue interrogazioni strutturate (simili a query SPARQL) per estrarre conoscenza.

Le principali tipologie di interrogazione implementate includono:

1. **Filtraggio Diagnostico:** Il sistema interroga la base di conoscenza per estrarre il sottoinsieme di pazienti con diagnosi positiva (proprietà isAutistic vera), isolando la coorte di interesse per l'analisi clinica.
2. **Navigazione Relazionale:** Sfruttando la proprietà did\_test, il sistema naviga dal nodo Paziente al nodo Test per recuperare i dettagli dell'esame (come il punteggio o il compilatore) senza dover effettuare join manuali come nei database relazionali.
3. **Introspezione dello Schema:** Il sistema è in grado di analizzare dinamicamente la struttura dell'ontologia stessa, elencando a runtime le classi e le proprietà disponibili, garantendo flessibilità in caso di evoluzione del modello dati.

## 4. APPRENDIMENTO SUPERVISIONATO: ANALISI DEGLI ALGORITMI

In questa fase, sono stati confrontati tre algoritmi di classificazione per predire la diagnosi di ASD.

- **Pipeline di Addestramento:** Per garantire risultati robusti, ogni modello utilizza una `ImbPipeline` che include:
  - **SMOTE:** Sovracampionamento nel training set per bilanciare le classi (evitando che il modello ignori la classe sbilanciata).
  - **GridSearchCV:** Ricerca automatica degli iperparametri ottimali con Cross-Validation stratificata (5-fold).

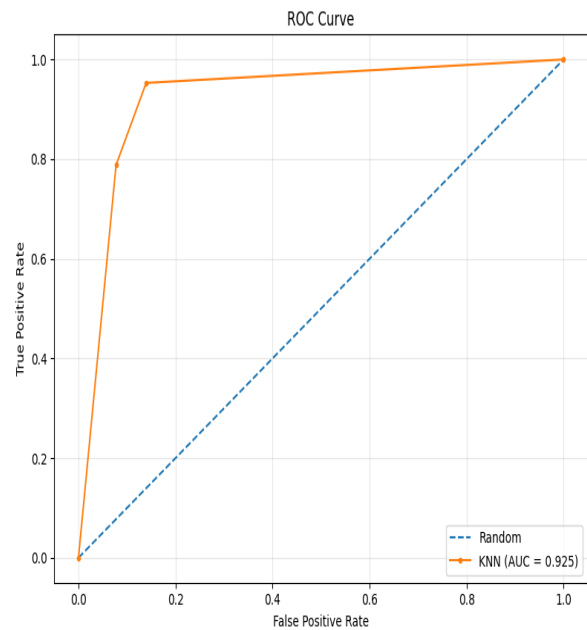
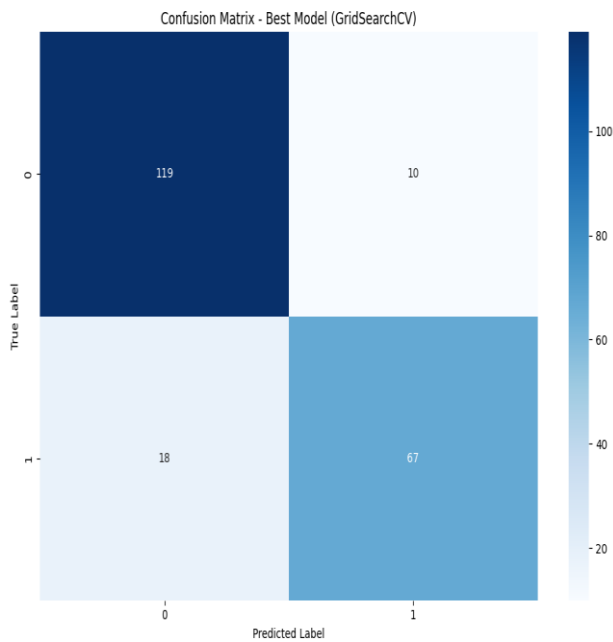
### 4.1 K-Nearest Neighbors (KNN)

- **Funzionamento:** Classifica un paziente in base alla maggioranza dei "vicini" più simili nello spazio delle feature.
- **Configurazione:** si parte da una griglia di partenza formata in questo modo:

```
# Definizione della griglia di iperparametri da testare
param_grid = {
    'knn__n_neighbors': list(range(1, 21)),
    'knn__weights': ['uniform', 'distance'],
    'knn__metric': ['euclidean', 'manhattan', 'minkowski']
}
```

- 
- Poi tramite la ricerca di iperparametri tramite `GridSearchCV` i nuovi parametri (sono quelli ottimali) saranno:
  - `Knn__n_neighbors=2`
  - `Knn__weights: uniform`
  - `Knn__metric: manhattan`
- **Performance:** Il modello ottiene una buona accuratezza (~92%).

I grafici mostrano la Heatmap della matrice di confusione (con i valori TP, TN, FP, FN) e la curva ROC con il relativo valore AUC.



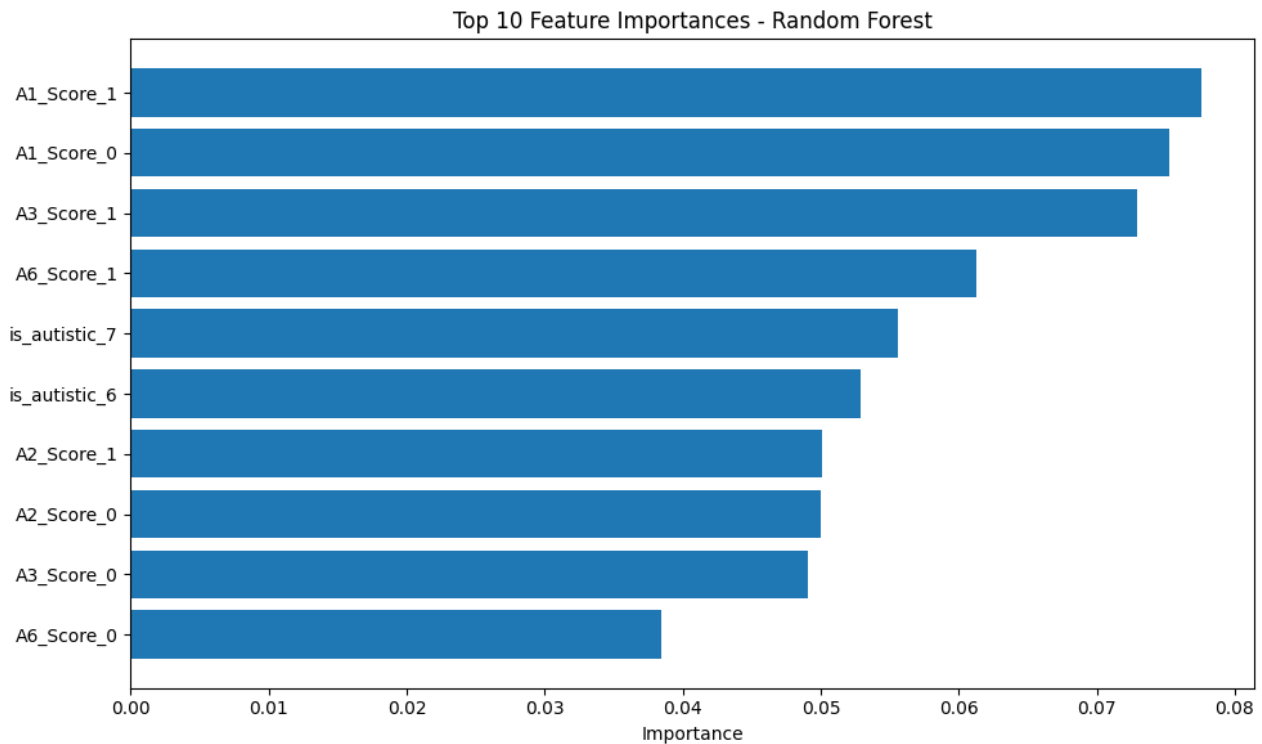
## 4.2 Random Forest (RF) e Feature Importance

- **Funzionamento:** Un insieme (ensemble) di alberi decisionali che votano per la classe finale. È robusto contro l'overfitting.
- **Configurazione:** Si parte da una griglia di partenza formata in questo modo:

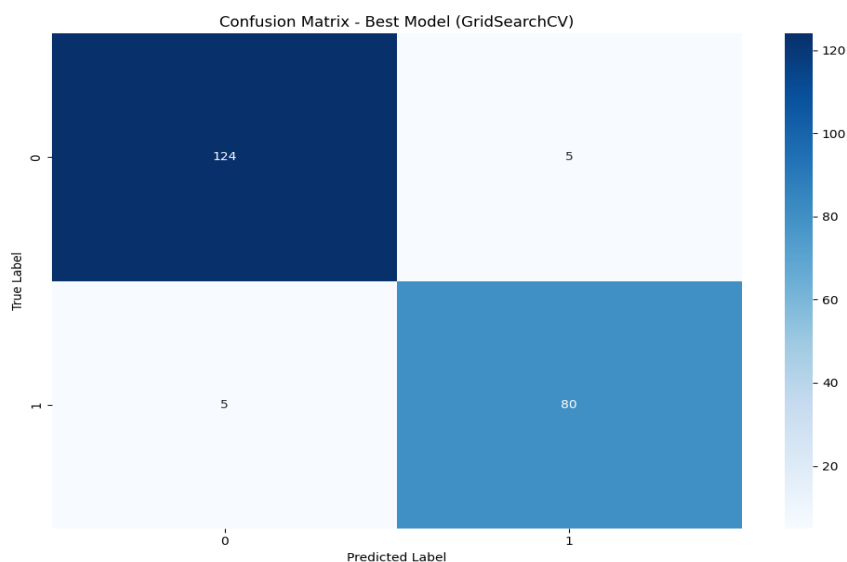
```
# Definizione della griglia di iperparametri da testare
param_grid = {
    'rf_n_estimators': [100, 200],
    'rf_max_depth': [16, 18, 20, None],
    'rf_min_samples_leaf': [1, 2, 4]
}
```

- Sempre utilizzando la GridSearchCV i nuovi parametri saranno:
  - Rf\_\_n\_estimators:100
  - Rf\_\_max\_depth:None(significa che la profondità ottimale è maggiore di 20)
  - Rf\_\_min\_samples\_leaf:2
- **Performance:** Accuratezza elevata (~94%) e ottima stabilità.
- **Feature Importance:** Una caratteristica chiave di RF è la capacità di calcolare l'importanza di ogni variabile. L'analisi ha rivelato che le domande **A1\_1** e **A1\_0**,

insieme al punteggio totale di screening, sono i predittori più forti. il Grafico sottostante elenca le feature dall'alto verso il basso in ordine di importanza. Le barre più lunghe corrispondono a A1\_Score\_1 e A1\_Score\_0. Dalla precedente analisi sul dataset si può notare che la domanda A\_1 rientra anche qui nelle domande con maggiore importanza, cosa che non succede per le altre.



La matrice di confusione risultante dal miglior modello Random Forest.



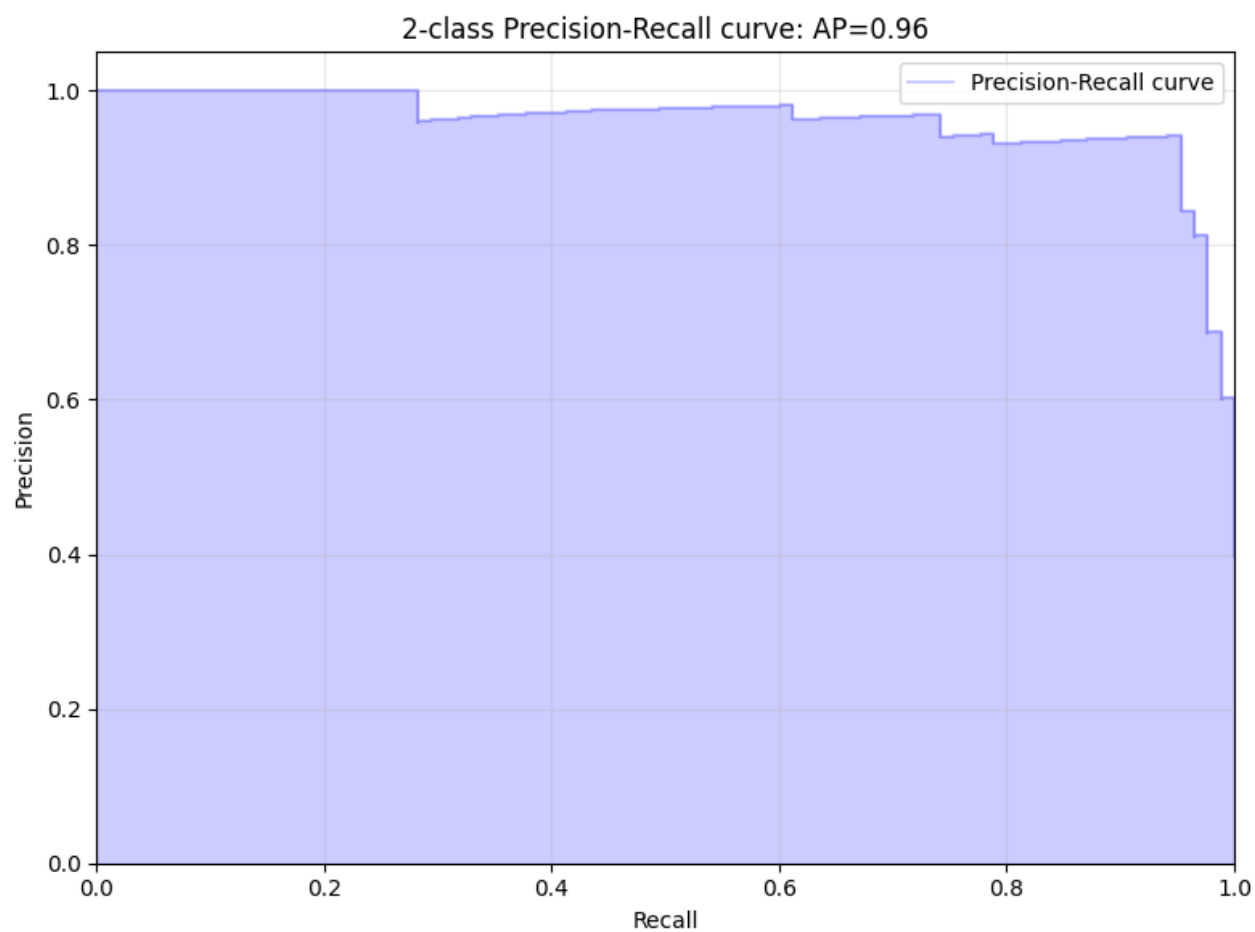
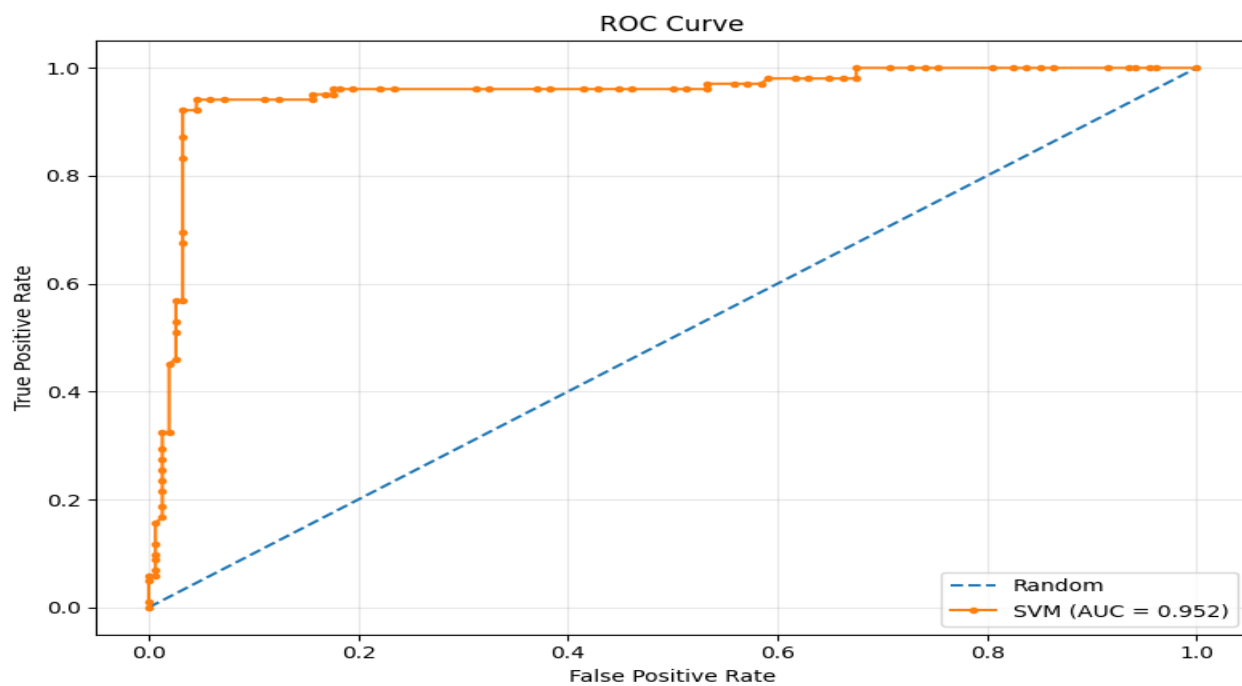
## 4.3 Support Vector Machine (SVM) - BEST MODEL

- **Funzionamento:** Cerca l'iperpiano che separa meglio le classi massimizzando il margine.
- **Configurazione:** la griglia di partenza è formata in questo modo:

```
# Definizione della griglia di iperparametri da testare
param_grid = {
    'svm__C': [0.1, 1, 10, 100],
    'svm__gamma': [1e-4, 1e-3, 0.01],
    'svm__kernel': ['rbf', 'linear']
}
```

- 
- Utilizzando la GridSearchCV per trovare gli iperparametri ottimali otteniamo:
  - Svm\_\_C: 0.1
  - Svm\_\_gamma: 1e-4
  - Svm\_\_kernel: linear.
- **Performance:** Si conferma il modello migliore del progetto.
  - **Accuracy:** ~95.8%
  - **Precision & Recall:** Entrambe superiori al 94%, indicando pochissimi falsi positivi e falsi negativi.

I grafici sottostanti della curva ROC mostrano un AUC molto vicino a 1.0 (eccellente capacità discriminativa) e la curva Precision-Recall che rimane alta.



## 5. ESPERIMENTO DI SELEZIONE FEATURE (ALL vs TOP-3)

Basandosi sui risultati della "Feature Importance" del Random Forest (Sezione 4.2), è stato condotto un esperimento per verificare se fosse possibile ridurre il test a sole 3 domande.

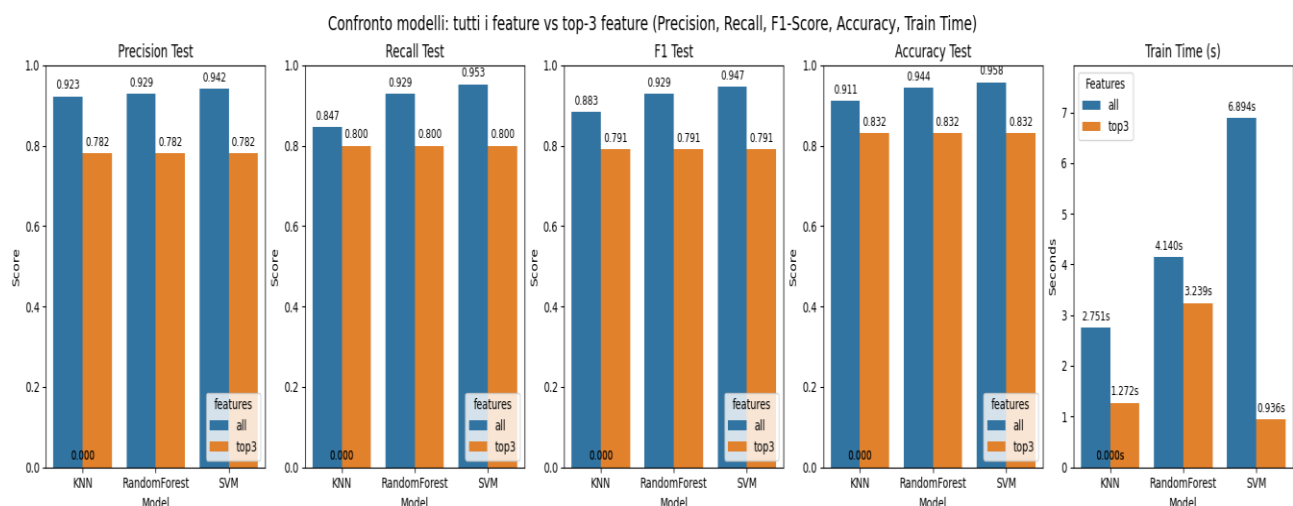
### 5.1 Configurazione dell'Esperimento

- **Obiettivo:** Creare un modello "leggero" usando solo le top-3 feature che sono: A1\_score\_1, A1\_score\_0 e A3\_score\_1 e confrontarlo con il modello completo.

### 5.2 Risultati e Discussione

- Il confronto ha evidenziato un **crollo significativo delle prestazioni** utilizzando solo 3 feature.
  - L'accuratezza media è scesa dal **95%** (All features) al **83%** (Top-3).
  - Anche Precision e Recall sono peggiorate drasticamente (scendendo sotto l'80%).
  - Si nota anche una rilevante diminuzione dei tempi di esecuzione, soprattutto nell'SVM

Il grafico mostra per ogni modello (KNN, RF, SVM) due barre affiancate (una per "All", una per "Top-3") relative a metriche come Accuracy e F1-Score, evidenziando visivamente il calo.



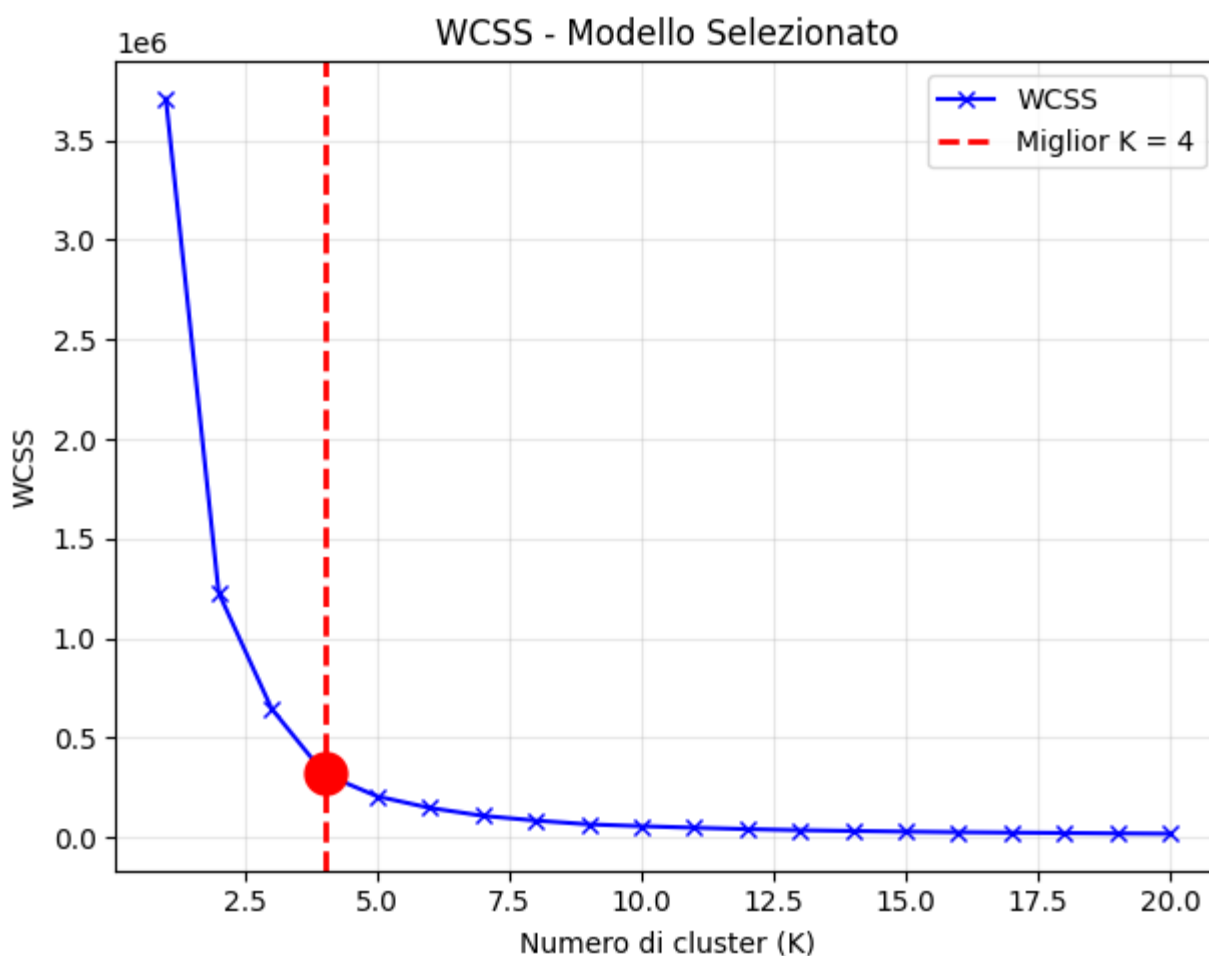
- **Conclusion:** Sebbene alcune domande siano dominanti, la diagnosi dell'autismo è complessa e richiede l'interazione di tutte le informazioni raccolte. Un test ridotto non è sufficientemente affidabile nonostante la sua "leggerezza" rispetto al dataset

originale. Può risultare utile in contesti con poche risorse computazionali a disposizione (a scapito dell'accuratezza).

## 6. APPRENDIMENTO NON SUPERVISIONATO (CLUSTERING)

### 6.1 Analisi K-Means

- **Obiettivo:** Esplorare il dataset senza usare le etichette di diagnosi per trovare raggruppamenti naturali (cluster) di pazienti.
- **Metodologia:** Utilizzo dell'algoritmo **K-Means**.
- **Determinazione di K (Elbow Method):** Per evitare di scegliere arbitrariamente il numero di gruppi, è stato usato l'algoritmo KneeLocator. Questo analizza la curva WCSS (Within-Cluster Sum of Squares) per trovare il punto di "gomito", ovvero il compromesso ottimale tra compattezza dei cluster e numero di gruppi.



*Descrizione:* Grafico che mostra la curva decrescente (WCSS sull'asse Y, K sull'asse X). Una linea tratteggiata verticale indica il numero ottimale di cluster identificato automaticamente (in questo caso K=4).

## 7. CONCLUSIONI E SVILUPPI FUTURI

- **Sintesi:** Il sistema ha dimostrato l'efficacia del Machine Learning per lo screening dell'ASD, con l'SVM che raggiunge prestazioni eccellenti (>95%). L'integrazione con l'ontologia fornisce un livello semantico utile per l'interrogazione dei dati.
- **Lezioni apprese:** La riduzione delle feature (Top-3), sebbene allettante per velocità, non è praticabile senza perdere accuratezza diagnostica.