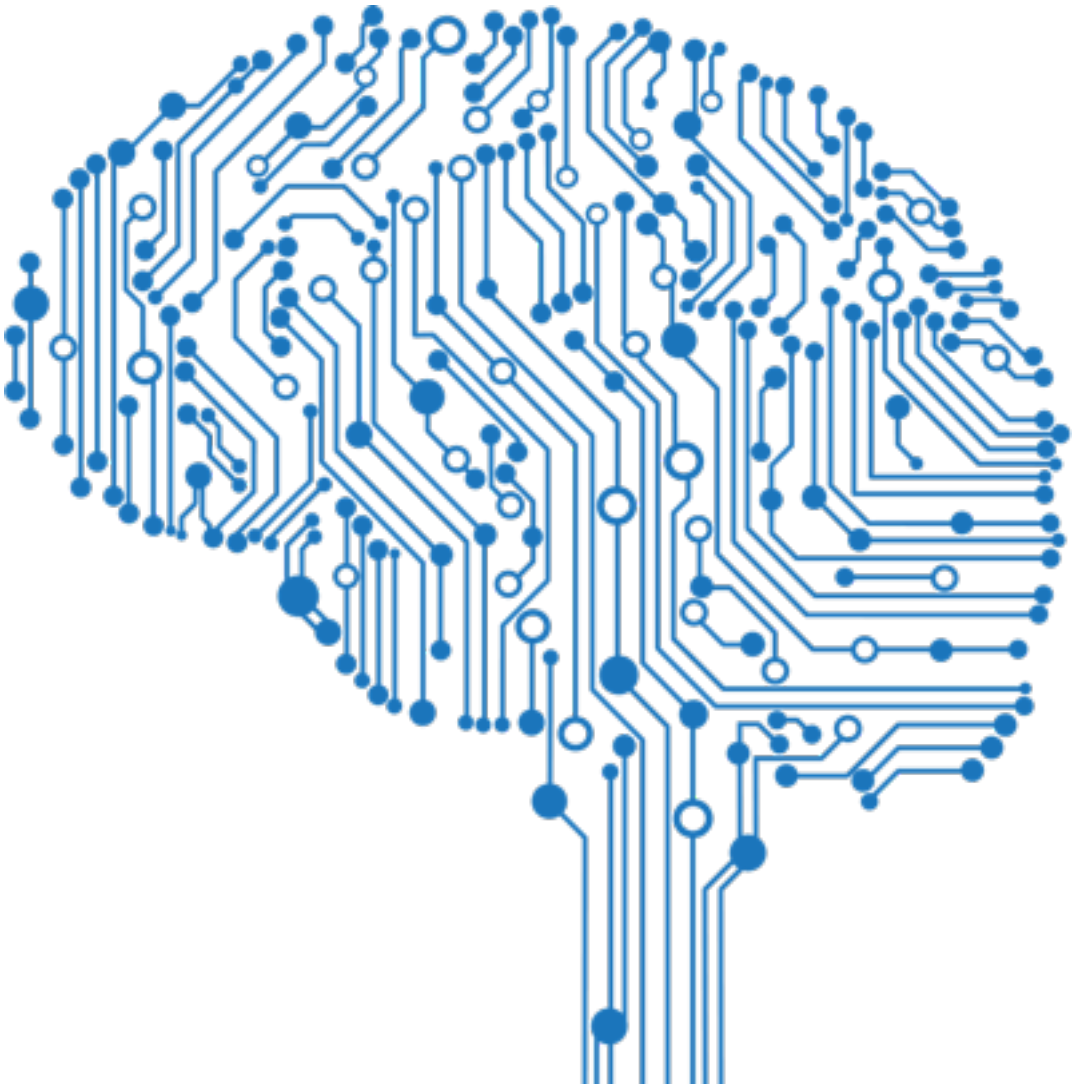
PROGETTO DI INGEGNERIA DALLA CONOSCENZA



GRUPPO:
MAFFIONE GIUSEPPE
MOSCHESE MICHELE
VANNELLA ANGELO MICHELE

1. INTRODUZIONE:

Lo scopo che ha portato alla nascita dal progetto è quello di poter predire in futuro una diagnosi di autismo, in modo tale da ottimizzare i costi e i tempi che oggi sono richiesti per una diagnosi.
Per tale proposito sono state adottato tecniche di apprendimento SUPERVISIONATO e NON SUPERVISIONATO. Inoltre è stata modellata anche un ontologia di dominio che permette di usufruire una rappresentazione formale e concettualizzata della realtà in esame, con il fine ultimo di renderla relazionabile con realtà ontologiche già esistenti.

I dati presi sono esame sono disponibili nei link qui sotto segnati:

- https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children
- https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Adolescent
- https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult

L'utilizzo dei tre dataset è dettato dalla ricerca di completezza della predizione in modo tale da coprire uno spetto più ampio. Tutti e tre i dataset hanno le medesime features:

Attribute	Type
Age	Number
Gender	String
Ethnicity	String
Born with jaundice	Boolean (yes or no)
Family member with PDD	Boolean (yes or no)
Who is completing the test	String
Country of residence	String
Used the screening app before	Boolean (yes or no)
Screening Method Type	Integer (0,1,2,3)
Question 1 Answer	Binary (0, 1)
Question 2 Answer	Binary (0, 1)
Question 3 Answer	Binary (0, 1)
Question 4 Answer	Binary (0, 1)
Question 5 Answer	Binary (0, 1)
Question 6 Answer	Binary (0, 1)
Question 7 Answer	Binary (0, 1)
Question 8 Answer	Binary (0, 1)
Question 9 Answer	Binary (0, 1)
Question 10 Answer	Binary (0, 1)
Screening Score	Integer

2. APPRENDIMENTO SUPERVISIONATO

L'**apprendimento supervisionato** è una tecnica di apprendimento automatico che mira a istruire un sistema informatico in modo da consentirgli di elaborare automaticamente previsioni sui valori di uscita di un sistema rispetto ad un input sulla base di una serie di esempi ideali, costituiti da coppie di input e di output, che gli vengono inizialmente forniti. Con **SUPERVISIORE**, intendiamo i sagnali di output già noti all'interno del nostro dataset. In questo caso utilizzeremo tecniche di classificazione.

Nel nostro caso è stato necessario un algoritmo che bilanci il dataset, in nostro soccorso abbiamo utilizzato la funzione SMOTE, che ridimensiona la classe degli esempi maggiori e minori.

Per ogni algoritmo implementato sono statati prodotti 4 tipologie di grafici differenti:

- ROC Curve
- Precision-Recall Curve
- Bar Chart di varianza e deviazione standard
- Matrice di Confusione

Per la maggior parte degli algoritmi è stata usata la tecnica della cross-validation per rilevare possibili problemi di sovra-adattamento. A tal proposito si riportano i valori del punteggio medio (cross-val-score), della varianza, dev. Standard su cinque iterate.

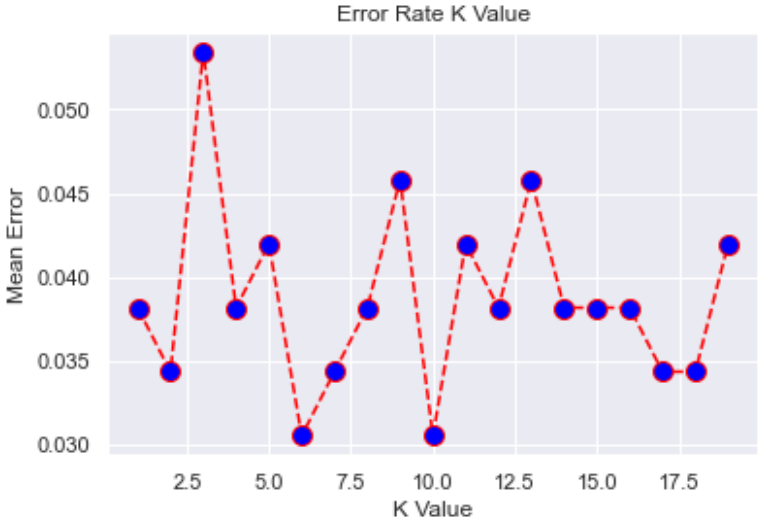
2.1. K-NEAREST-NEIGHBOUR

Il KNN è un algoritmo di classificazione atto nel trovare la classe di appartenenza di un dato in ingresso, si esegue semplicemente un loop su ogni esempio di training. Per ogni esempio viene calcolata la distanza tra esempio e ingresso usando la metrica desiderata (ad esempio la distanza euclidea). Infine, si prendono i **k** esempi con distanza minore e si contano le classificazioni di questi esempi. Vince la classificazione con maggior numero di esempi tra i **k** trovati. In caso di parità solitamente si prende la classe appartenente all'esempio più vicino in assoluto tra i **k** trovati. In caso di ulteriore parità (magari ci sono due o più esempi di classe diversa con stessa distanza minima) dovrei utilizzare regole ulteriori (ad esempio la classe con maggior numero di esempi tra tutti gli esempi di training o altro).

Da quanto appena detto appare evidente la semplicità di questo algoritmo (un semplice loop). Inoltre, è altrettanto evidente che in caso si desideri arricchire il dataset di training con nuovi esempi, non è necessario eseguire una fase di training vera e propria, ma l'algoritmo sarà subito pronto a eseguire predizioni.

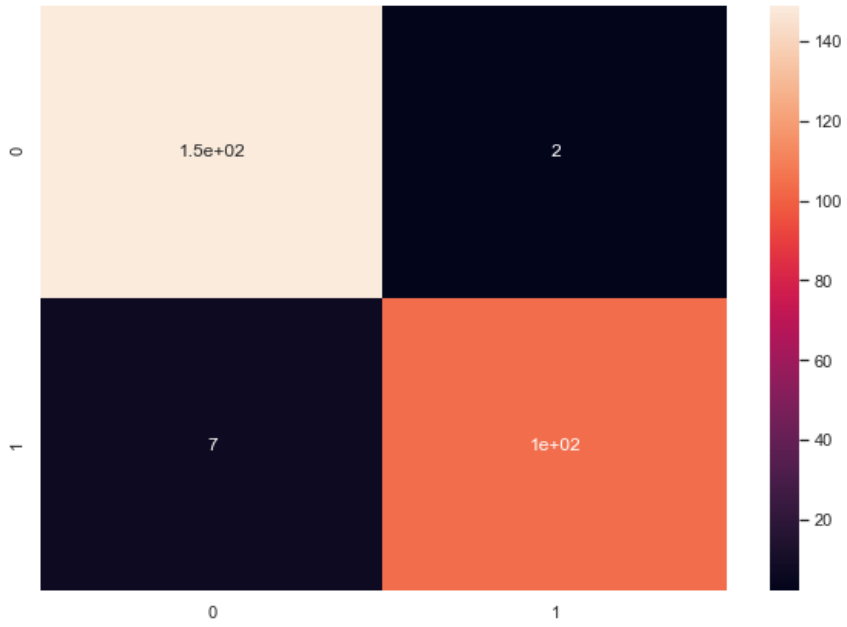
D'altro canto, se il training set è grande, lo sforzo computazionale potrebbe aumentare troppo (una cosa è un loop eseguito su 1.000 esempi, una cosa diversa è un loop eseguito su 1.000.000 di esempi…). Inoltre, oltre al valore di **k**, influisce sul risultato anche la metrica utilizzata, che quindi va scelta in maniera opportuna.

Nel nostro specifico caso abbiamo utilizzato la funzione SMOTE.



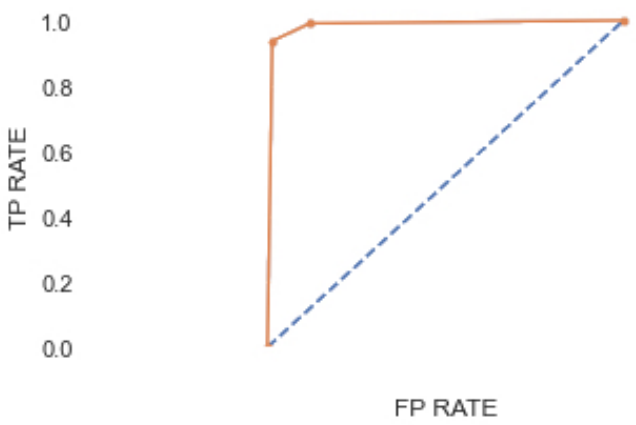
In base alle caratteristiche del nostro Dataset, i vicini suggeriti per la minimizzazione dell'errore sono il 6 e successivamente il 10 con un errore minimo commesso vicino allo 0.030

In valori ottenuti con i restati grafici sono:

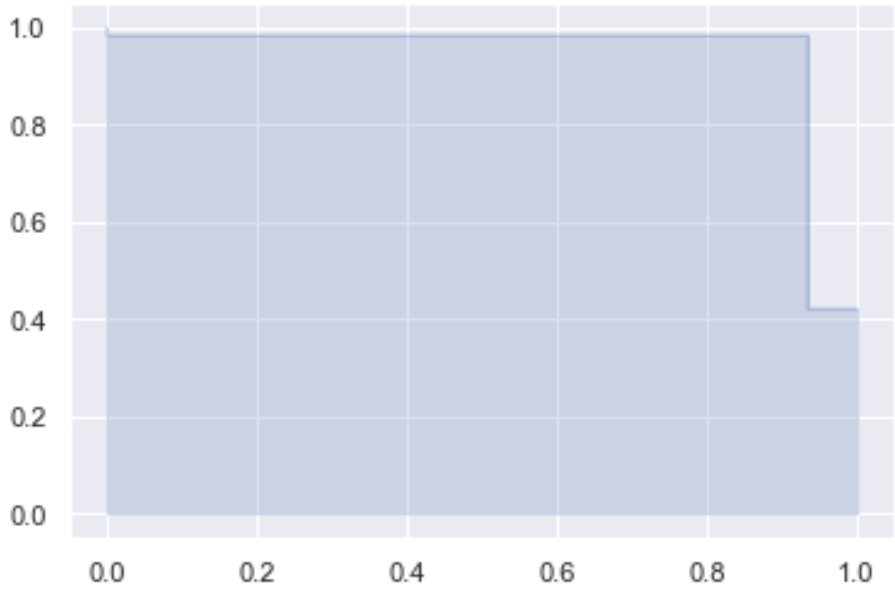


Clasification report:

	precision	recall	f1-score	support
0	0.96	0.99	0.97	151
1	0.98	0.94	0.96	111
accuracy			0.97	262
macro avg	0.97	0.96	0.96	262
weighted avg	0.97	0.97	0.97	262

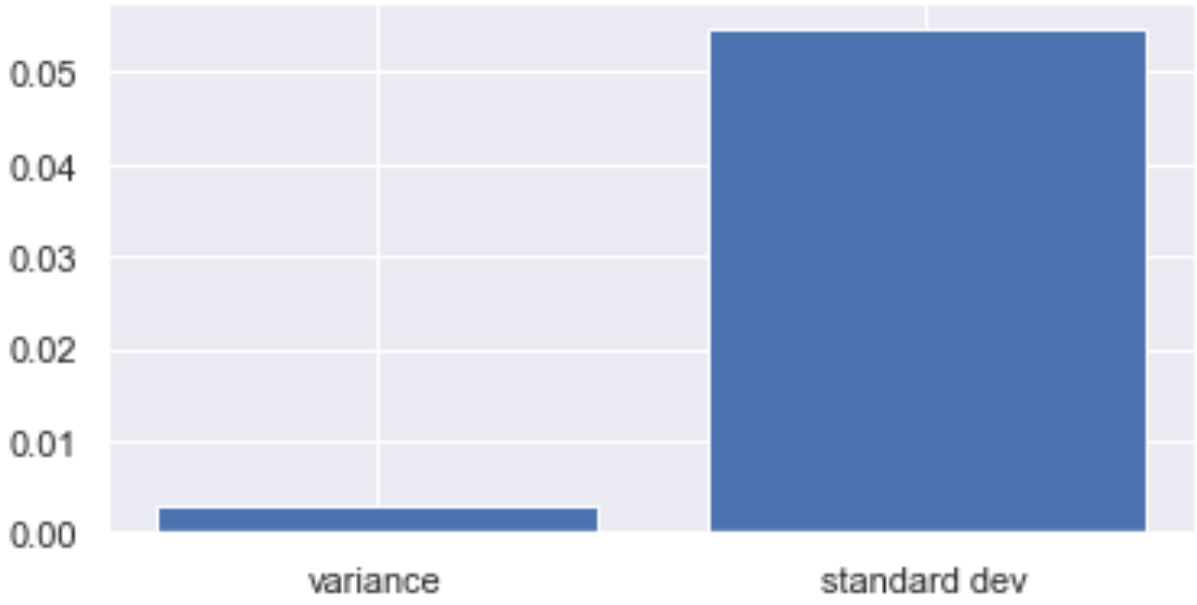


AUC: 0.985



Average Precision = 0.945

Accuracy = 0.965



cv_scores mean:0.8960
cv_score variance:0.0029
cv_score dev standard:0.05468

2.2. RANDOM FOREST

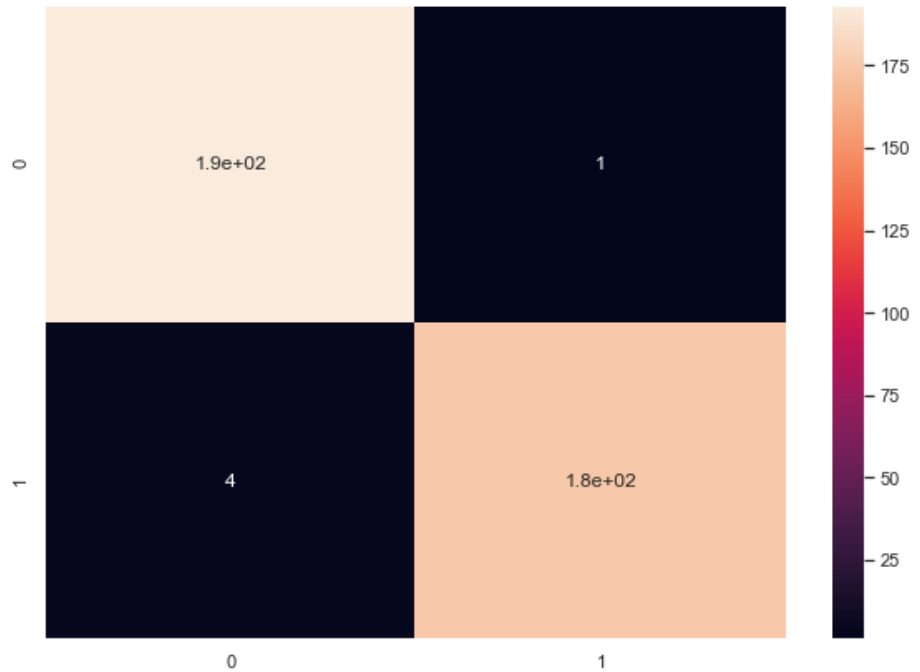
L'algoritmo Random Forest, o delle foreste casuali in italiano, è un algoritmo di **apprendimento supervisionato**.

Rappresenta un tipo di modello ensemble, che si avvale del bagging come **metodo di ensemble** e l'albero decisionale come **modello individuale**.

Ciò significa che una foresta casuale combina molti alberi decisionali in un unico modello. Individualmente, le previsioni fatte dagli alberi decisionali potrebbero non essere accurate, ma combinate insieme, le previsioni saranno in media più vicine al risultato.

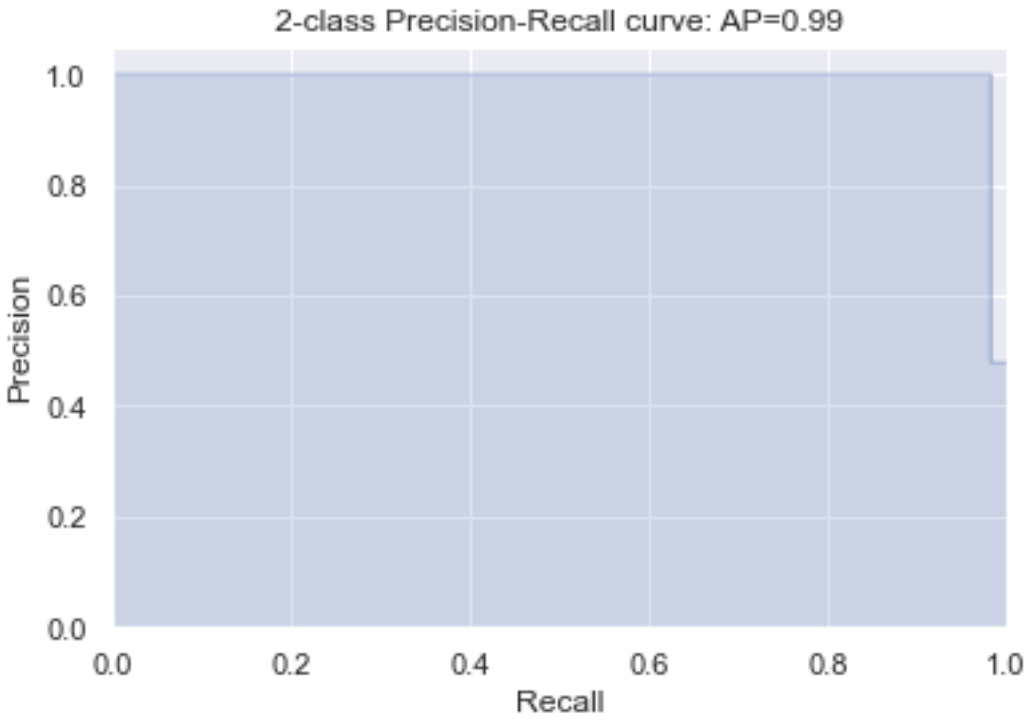
Il risultato finale restituito dal Random Forest altro non è che la media del risultato numerico restituito dai diversi alberi nel caso di un **problema di regressione**, o la classe restituita dal maggior numero di alberi nel caso la Random Forest sia stata utilizzata per risolvere un **problema di classificazione**.

Nella sua applicazione abbiamo natato che la sua classificazione era quasi vicini al 1



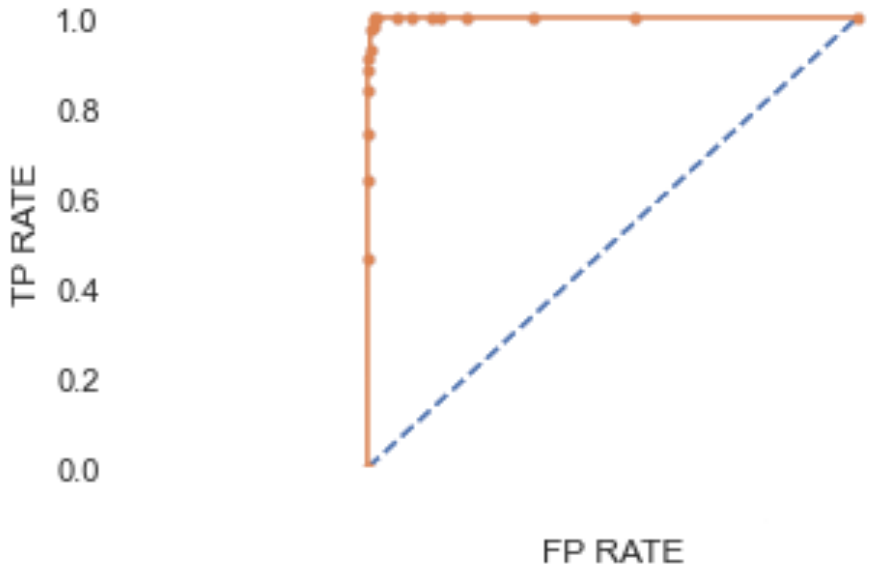
Clasification report:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	194
1	0.99	0.99	0.99	179
accuracy			0.99	373
macro avg	0.99	0.99	0.99	373
weighted avg	0.99	0.99	0.99	373

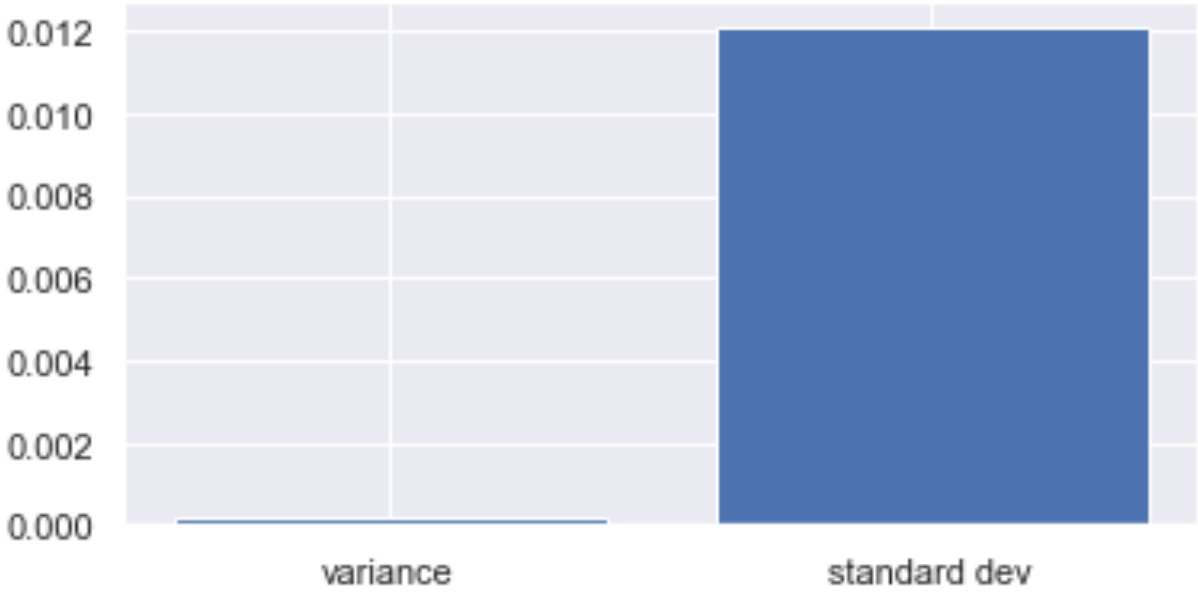


AP = 0.99

Accuracy = 0.999998



AUC: 0.9999



cv_scores mean:0.9911549423500453

cv_score variance:6.0706994739654e-05

cv_score dev standard:0.007791469356909132

2.3. Support Vector Machines

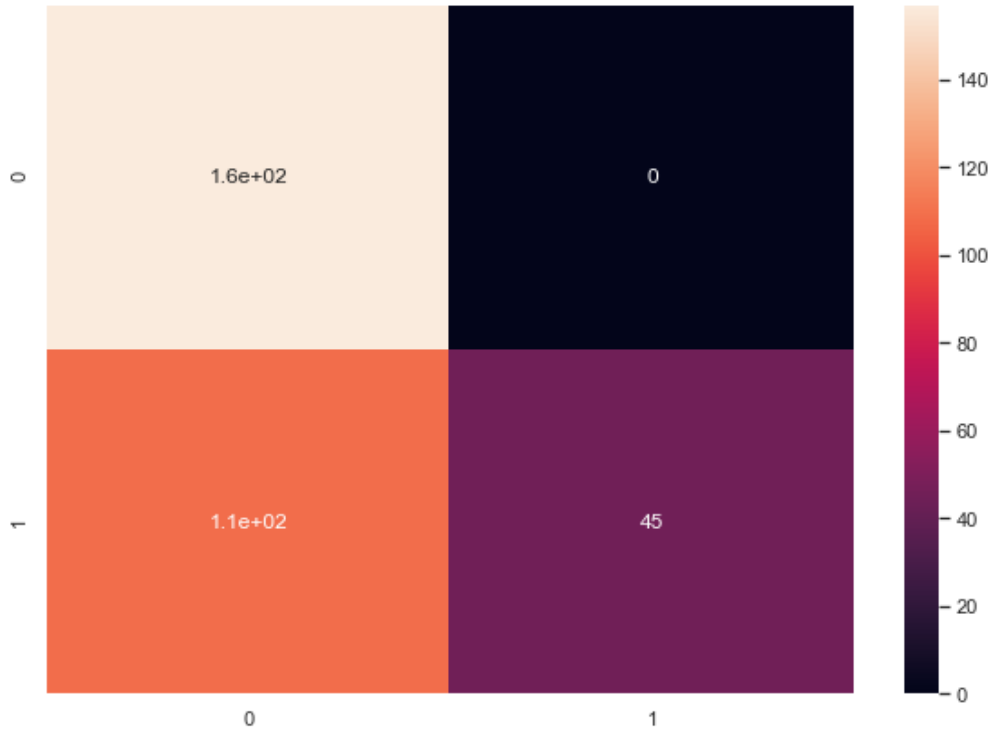
L'SVM è basato sull'idea di trovare un iperpiano che divida al meglio un set di dati in due classi. Per comprenderne il funzionamento è bene definire alcuni concetti chiave:

Iperpiano o **limite di decisione lineare:** per un'attività di classificazione con solo due dimensioni spaziali x_1 e x_2 (o x e y), un iperpiano è raffigurato come una linea che separa e classifica un insieme di dati. A 3 dimensioni, un iperpiano è rappresentato da un piano (il rettangolo in blu) come puoi vedere nell'immagine seguente. Con più di tre dimensioni viene definito genericamente "iperpiano".

Support Vector: definiti anche **vettori di supporto** in italiano, essi sono i punti dati più vicini all'iperpiano. Tali punti dipendono dal set di dati che si sta analizzando e se vengono rimossi o modificati alterano la posizione dell'iperpiano divisorio. Per questo motivo, possono essere considerati gli elementi critici di un set di dati.

Margine: è definito come la distanza tra i vettori di supporto di due classi differenti più vicini all'iperpiano. Alla metà di questa distanza viene tracciato l'iperpiano, o retta nel caso si stia lavorando a due dimensioni.

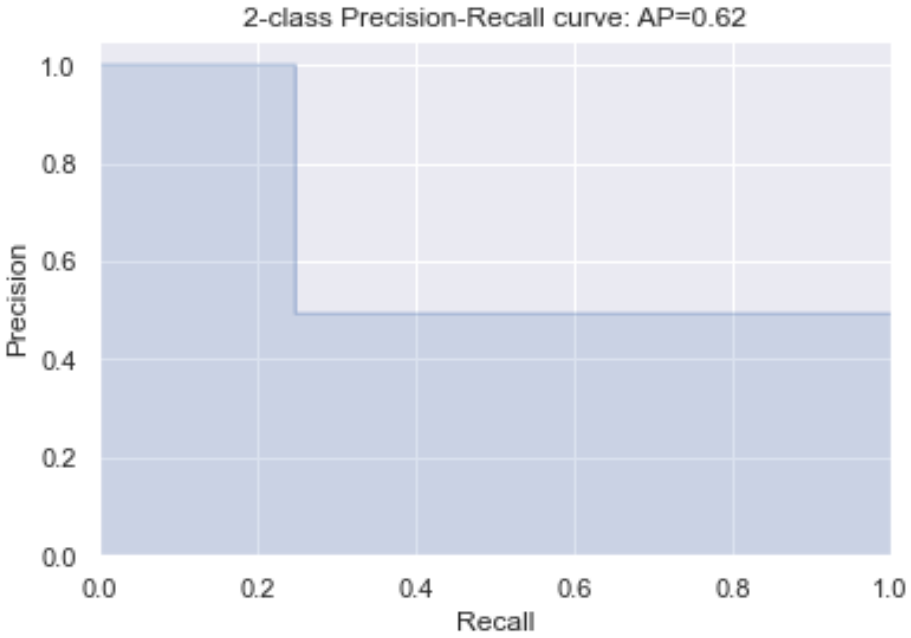
Nel nostro caso, i grafici ottenuti sono:



Clasification report:

precision recall f1-score support

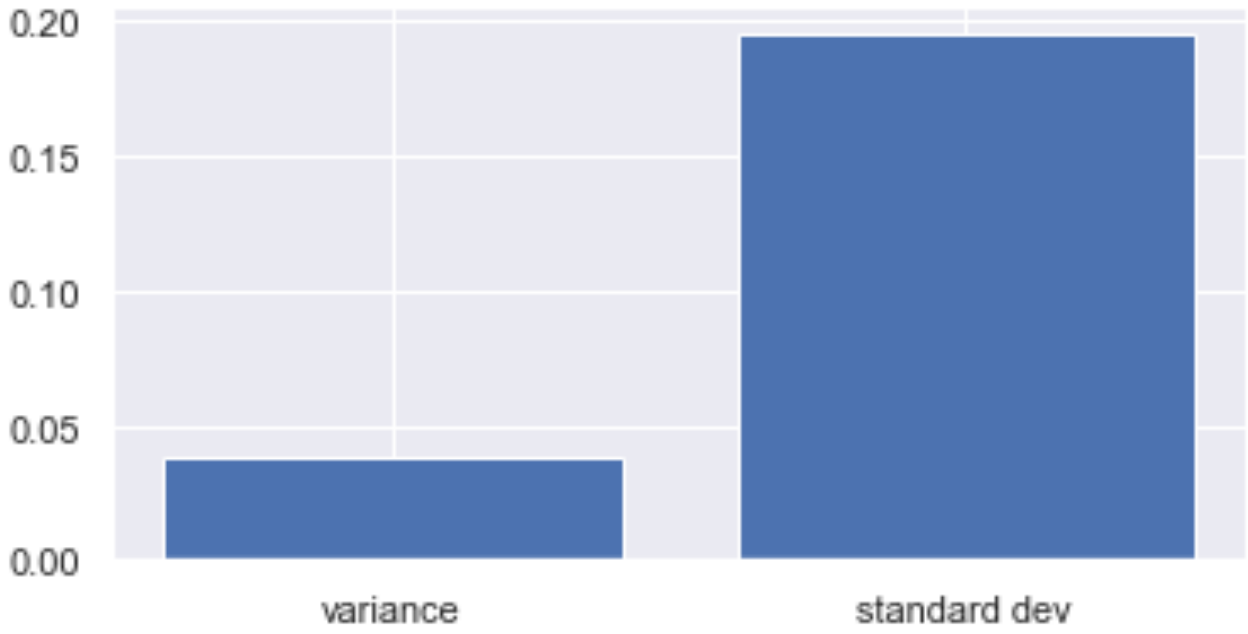
0	0.58	1.00	0.73	157
1	1.00	0.25	0.40	154
accuracy			0.63	311
macro avg	0.79	0.62	0.56	311
weighted avg	0.79	0.63	0.56	311



ù

average precision = 0.62

accuracy =0.639



cv_scores mean:0.6338709677419355

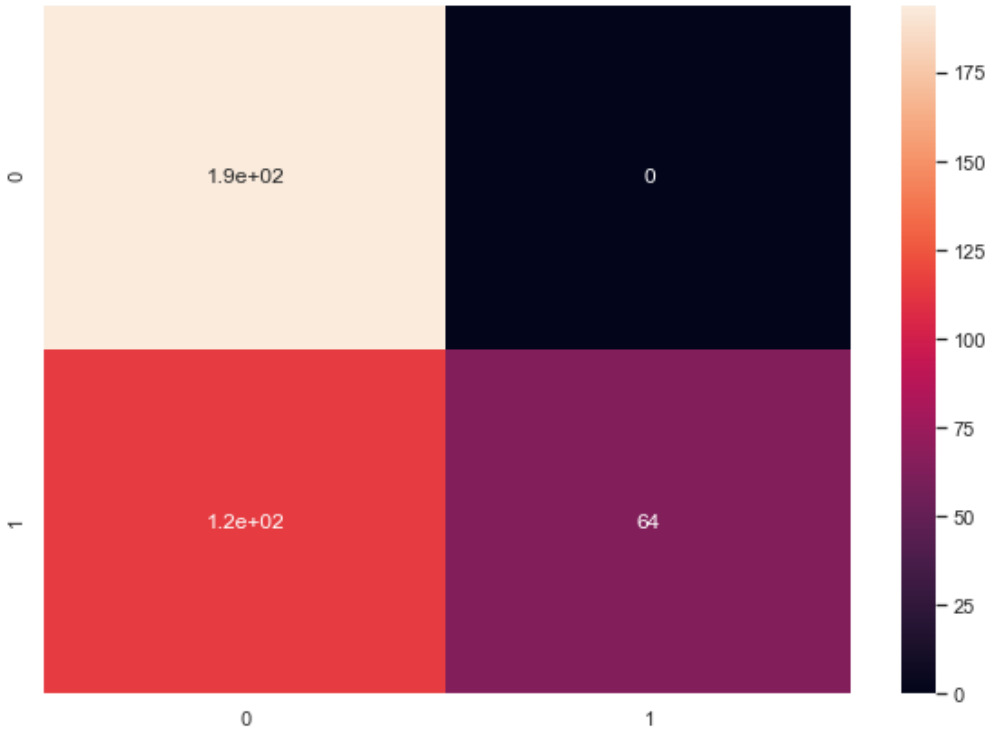
cv_score variance:0.0323533298387101

cv_score dev standard:0.17987031394510353

2.4. Multinomilan Naive Bayes

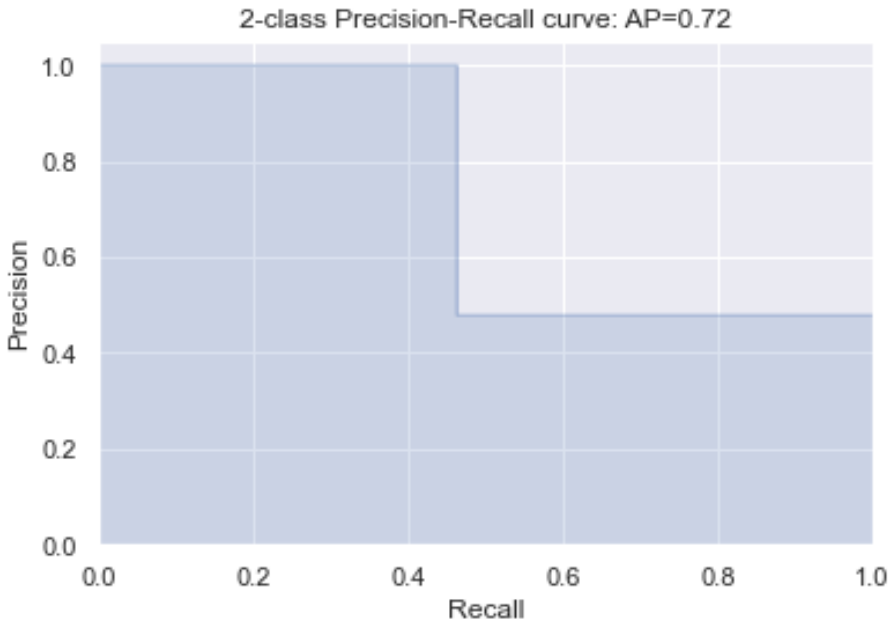
E' un classificatore basato su calcoli probabilistici basati sul teorema di Bayes con l'aggiunta di un forte indipendenza tra le feature.
Applicandolo ad un evento multidimensionale viene rappresentato come la frequenza di eventi generati da una distribuzione polinomiale ($p_1,\ldots,p_n$) dove p_i è la probabilità che si verifichi l'evento.

Nel nostro caso i grafici prodotti da questo algoritmo sono:



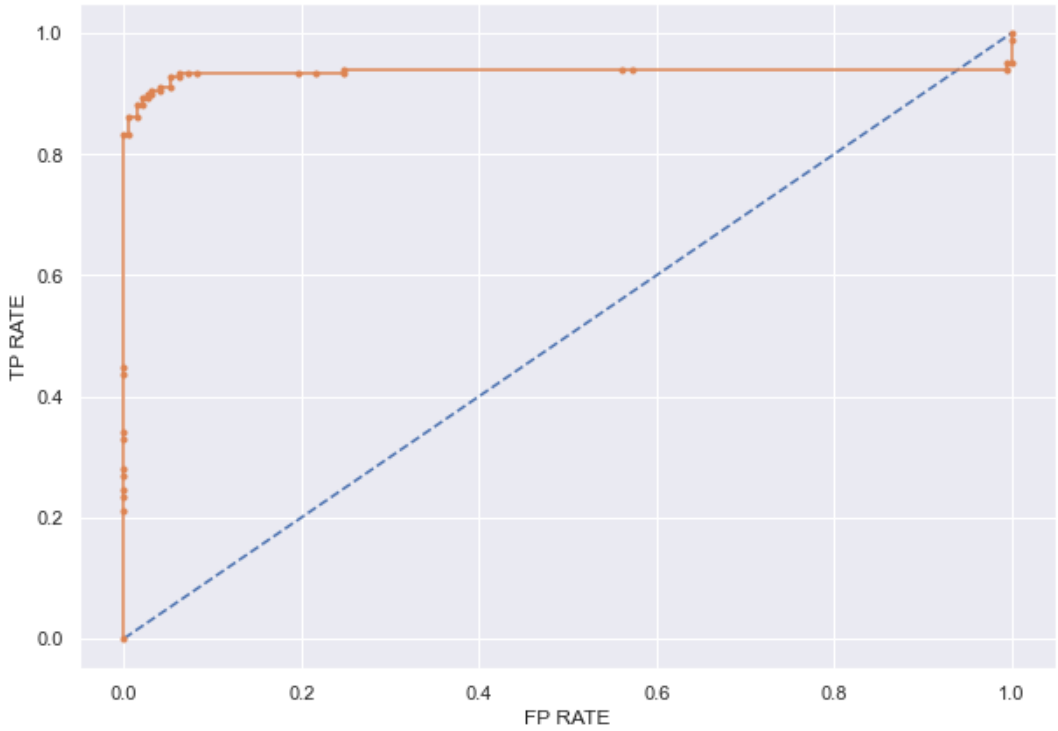
Clasification report:

	precision	recall	f1-score	support
0	0.67	1.00	0.80	194
1	1.00	0.46	0.63	179
accuracy			0.74	373
macro avg	0.83	0.73	0.72	373
weighted avg	0.83	0.74	0.72	373

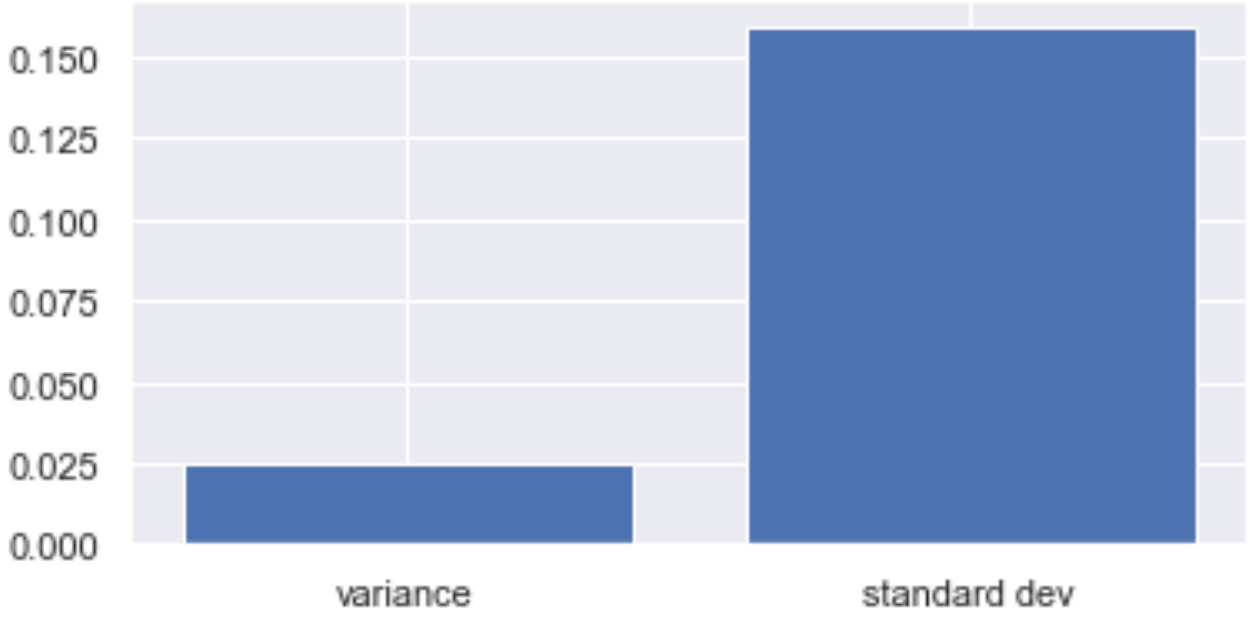


average precision = 0.72

accuratezza = 0.766



AUC: 0.960



cv_scores mean:0.8226195102992616
cv_score variance:0.021942193507301526
cv_score dev standard:0.14812897592065344

2.5. Neural Network

Le reti neurali si basano principalmente sulla simulazione di neuroni artificiali opportunamente collegati.

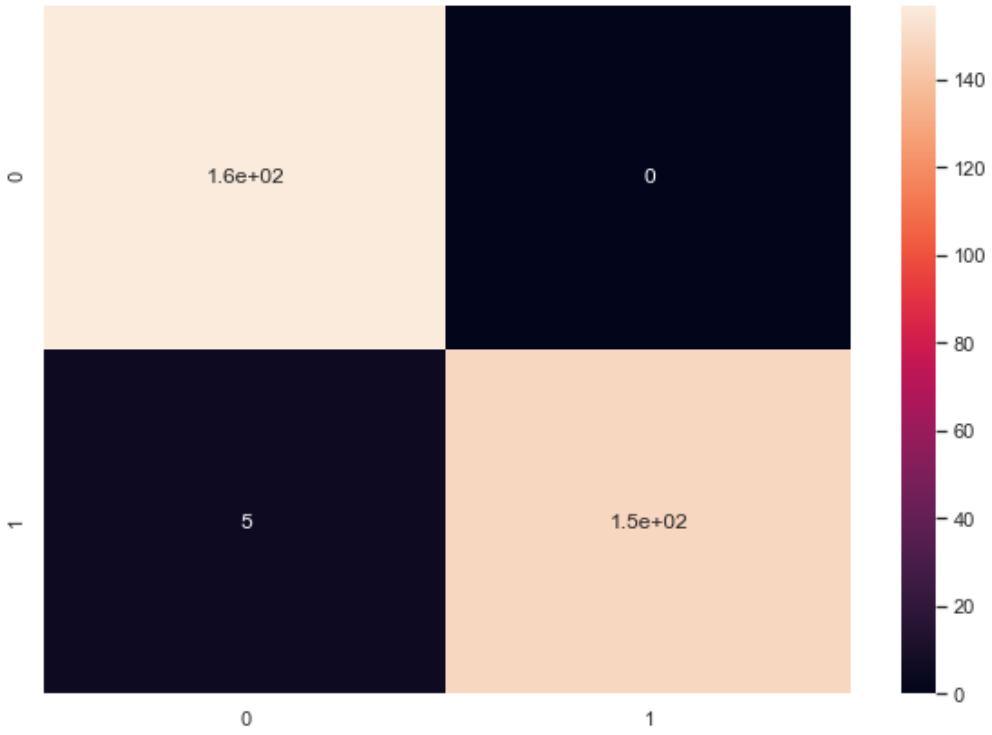
I suddetti neuroni ricevono in ingresso degli stimoli e li elaborano. L'elaborazione può essere anche molto sofisticata ma in un caso semplice si può pensare che i singoli ingressi vengano moltiplicati per un opportuno valore detto peso, il risultato delle moltiplicazioni viene sommato e se la somma supera una certa soglia il neurone si attiva attivando la sua uscita. Il peso indica l'efficacia sinaptica della linea di ingresso e serve a quantificarne l'importanza, un ingresso molto importante avrà un peso elevato, mentre un ingresso poco utile all'elaborazione avrà un peso inferiore. Si può pensare che se due neuroni comunicano fra loro utilizzando maggiormente alcune connessioni allora tali connessioni avranno un peso maggiore, fino a che non si creeranno delle connessioni tra l'ingresso e l'uscita della rete che sfruttano "percorsi preferenziali". Tuttavia è sbagliato pensare che la rete finisca col produrre un unico percorso di connessione: tutte le combinazioni infatti avranno un certo peso, e quindi contribuiscono al collegamento ingresso/uscita.

I singoli neuroni vengono collegati alla schiera di neuroni successivi, in modo da formare una rete di neuroni. Normalmente una rete è formata da tre strati. Nel primo abbiamo gli ingressi (I), questo strato si preoccupa di trattare gli ingressi in modo da adeguarli alle richieste dei neuroni. Se i segnali in ingresso sono già trattati può anche non esserci. Il secondo strato è quello nascosto (H, *hidden*), si preoccupa dell'elaborazione vera e propria e può essere composto anche da più colonne di neuroni. Il terzo strato è quello di uscita (O) e si preoccupa di raccogliere i risultati ed adattarli alle richieste del blocco successivo della rete neurale. Queste reti possono essere anche molto complesse e coinvolgere migliaia di neuroni e decine di migliaia di connessioni.

Nel nostro caso la nostra rete prevede una struttura sequenziale a tre livelli:

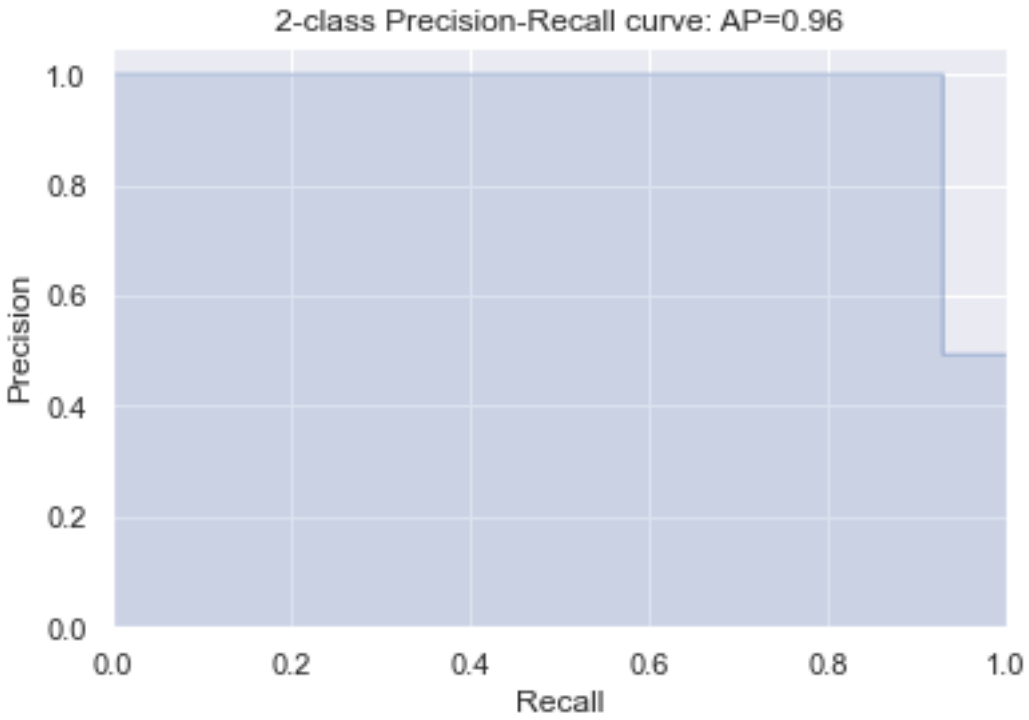
uno di input(costituito da 60 ingressi), uno nascosto(costituito da 30 neuroni artificiali) e uno di outputche restituisce un valore tra 0 e 1 in base alla appartenenza dell'esempio alla categoria presa sotto esame.

I risultati ottenuti sono:



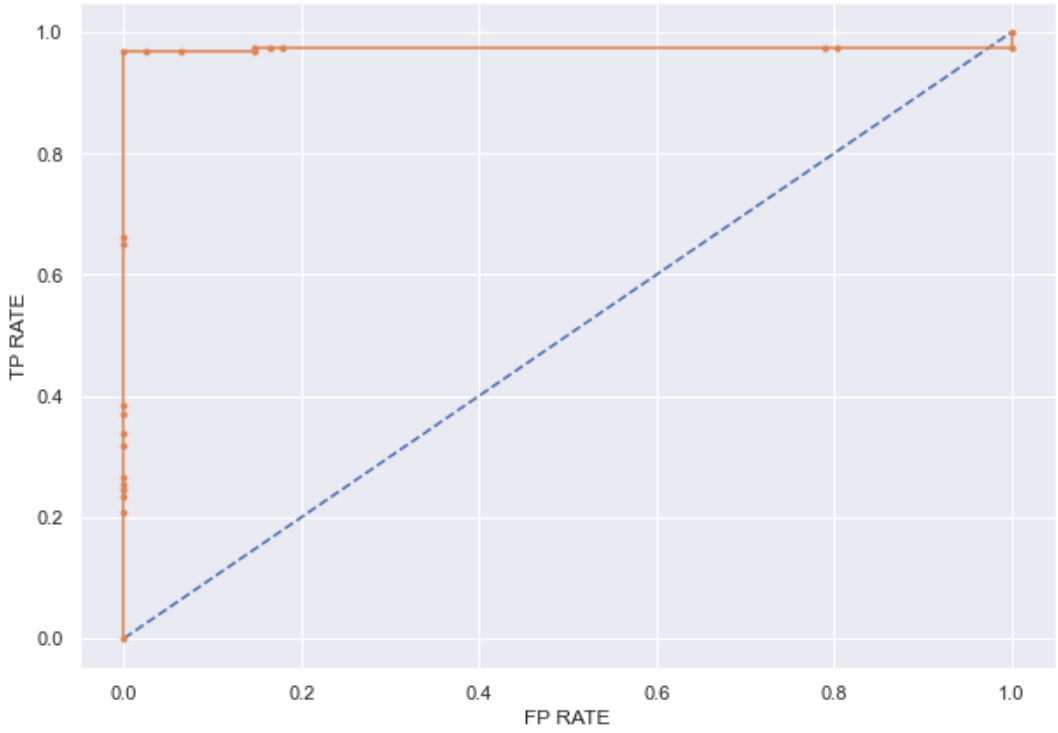
Clasification report:

	precision	recall	f1-score	support
0	0.97	1.00	0.98	157
1	1.00	0.97	0.98	154
accuracy			0.98	311
macro avg	0.98	0.98	0.98	311
weighted avg	0.98	0.98	0.98	311

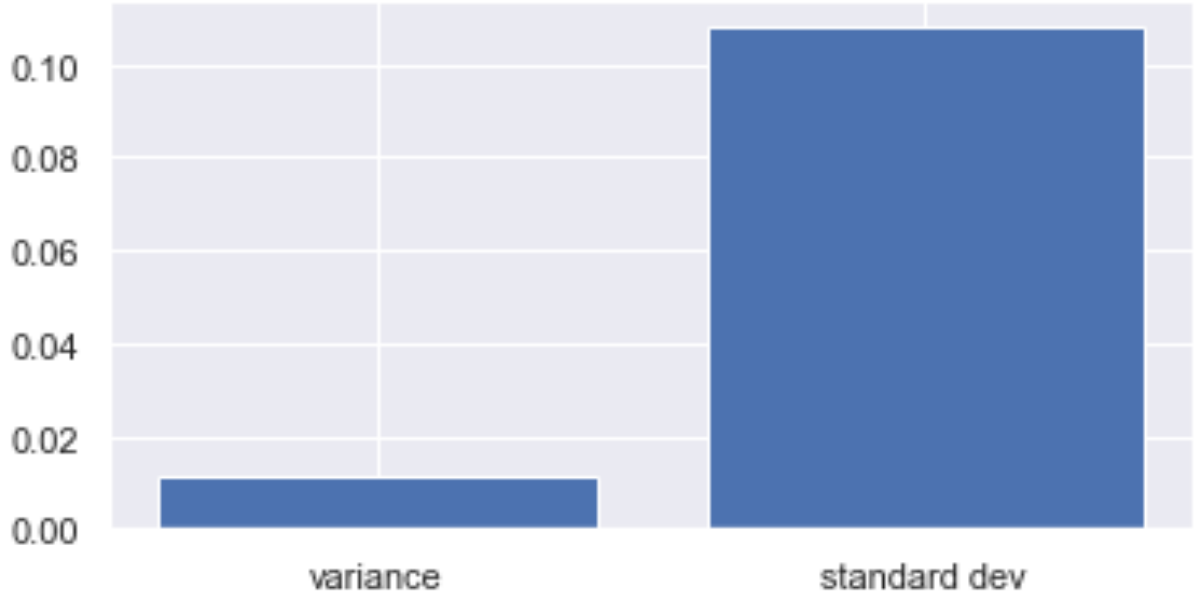


AP = 0.96

ACCUARCY = 0.974



AUC: 0.955



cv_scores mean:0.9395161271095276

cv_score variance:0.014633195519323294

cv_score dev standard:0.12096774578094482

3. APPRENDIMENTO NON SUPERVISIONATO

L' apprendimento non supervisionato e una tecnica in cui non è necessario supervisionare o condividere i dati etichettati con il modello. Al contrario, l'algoritmo del modello comprenderà automaticamente e inizierà ad imparare dai dati senza guida. Il modello utilizzerà i dati non etichettati per identificare nuovi modelli e informazioni dovute alla progettazione del loro algoritmo. Con questo metodo, possiamo trovare informazioni nuove e non identificate in precedenza. Questo tipo di comportamento di apprendimento è simile a quello degli esseri umani. Immaginate come analizziamo e osserviamo l'ambiente circostante per raccogliere i dati e capire e riconoscere le cose. Allo stesso modo, le macchine con algoritmi di apprendimento non supervisionati scoprono schemi per trovare risultati utili. Per esempio, il sistema può identificare la differenza tra cani e gatti comprendendo le caratteristiche e le caratteristiche di entrambi gli animali.

3.1. K-MEANS

Il K-Means è un algoritmo di **apprendimento non supervisionato** che trova un numero fisso di cluster in un insieme di dati.

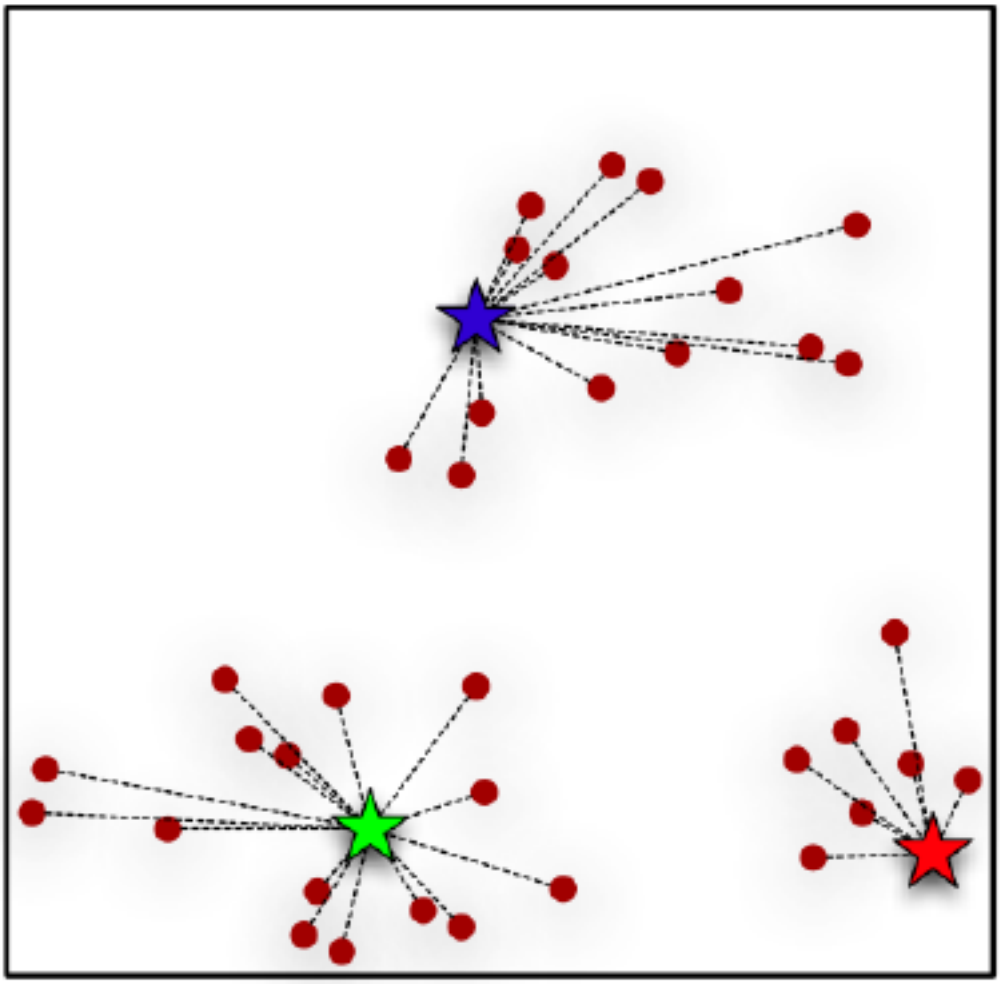
I **cluster** rappresentano i gruppi che dividono gli oggetti a seconda della presenza o meno di una certa somiglianza tra di loro, e vengono scelti a priori, prima dell'esecuzione dell'algoritmo.

Ognuno di questi cluster raggruppa un particolare insieme di oggetti, che vengono definiti **data points**.

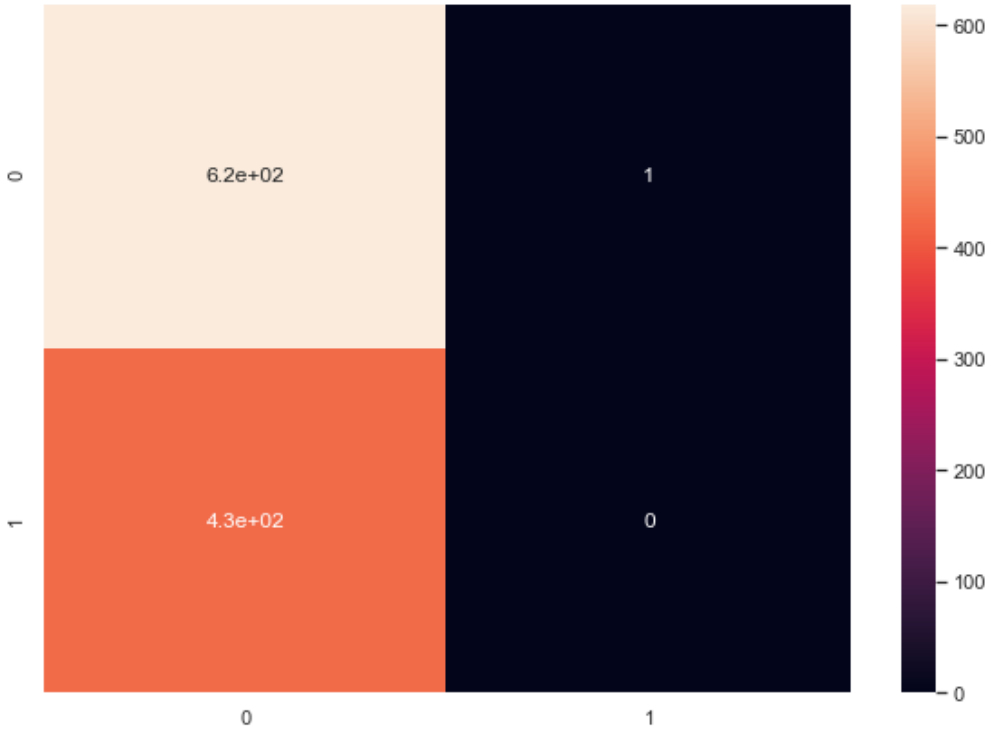
L'insieme dei data points analizzati definisce il **set di dati**, che rappresenta l'insieme di tutte le istanze analizzate dall'algoritmo.

Quando si utilizza un algoritmo K-Means, per ogni cluster si definisce un **centroide**, ossia un punto (immaginario o reale) al centro di un cluster.

Nell'immagine sotto il centroide è rappresentato dalle tre stelle colorate di blu, rosso e verde per intenderci, mentre i data points sono gli elementi che compongono i cluster, ossia i puntini rossi vicino alle stelle.

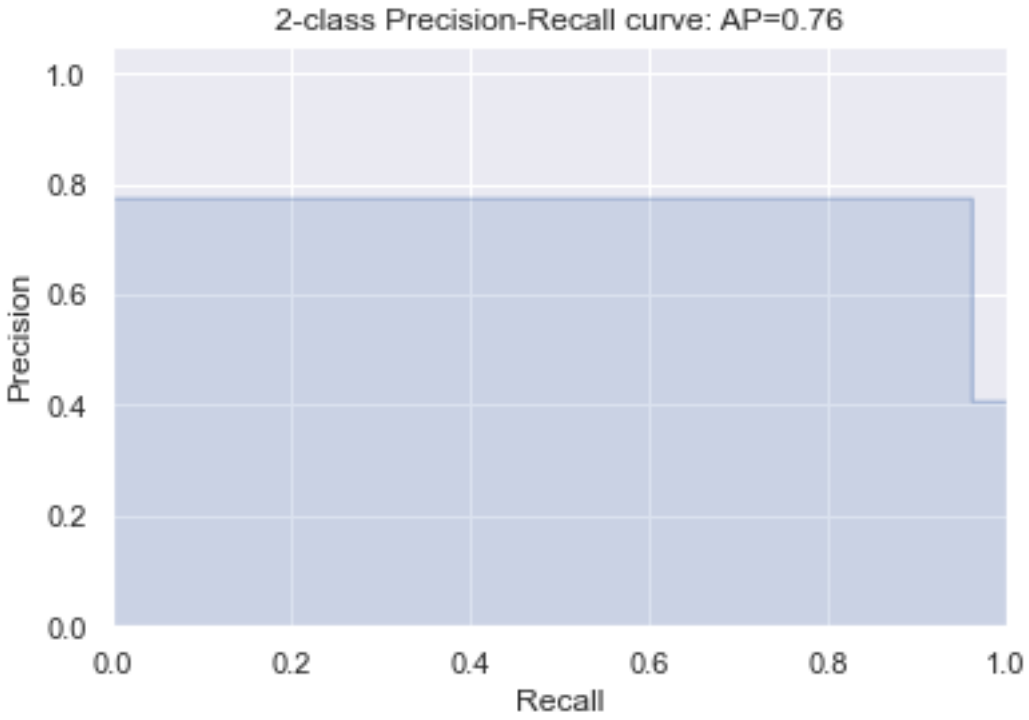


I risultati ottenuti nel nostro caso sono:



Clasification report:

	precision	recall	f1-score	support
0	0.59	1.00	0.74	620
1	0.00	0.00	0.00	427
accuracy			0.59	1047
macro avg	0.30	0.50	0.37	1047
weighted avg	0.35	0.59	0.44	1047



AP = 0.76

ACCURACY = 0.86

4. ONTOLOGIE

Un' ontologia è la specificazione dei significati dei simboli in un sistema informatico. La specifica formale è importante per l'interoperabilità semantica, ovvero l'abilità di basi di conoscenza differenti di operare insieme ad un livello semantico tale che i significati dei simboli sono rispettati. L'ontologia viene descritta come "una specificazione di una concettualizzazione". Una rappresentazione formale di un insieme di conoscenze è una concettualizzazione, ossia un insieme di oggetti, concetti e relazioni fra di essi che esistono in una particolare area d'interesse.

4.1. DOMINIO

Il disturbo dello spettro autistico (ASD) è una condizione del neuro sviluppo associata a costi sanitari significativi e la diagnosi precoce può ridurli in modo

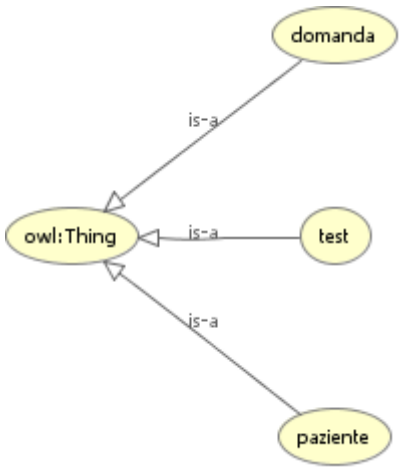
significativo. Sfortunatamente, i tempi di attesa per una diagnosi di ASD sono lunghi e le procedure non sono convenienti. L'impatto economico dell'autismo e l'aumento del numero di casi di ASD in tutto il mondo rivela un'urgente necessità di sviluppare metodi di screening efficaci e di facile attuazione. Pertanto, è imminente uno screening dell'ASD efficiente in termini di tempo e accessibile per aiutare gli operatori sanitari e informare le persone sull'opportunità di perseguire una diagnosi clinica formale. La rapida crescita del numero di casi di ASD in tutto il mondo necessita di set di dati relativi ai tratti comportamentali. Tuttavia, tali set di dati sono rari e rendono difficile eseguire analisi approfondite per migliorare l'efficienza, la sensibilità, la specificità e l'accuratezza predittiva del processo di screening dell'ASD. In questo set di dati, registriamo dieci caratteristiche comportamentali (AQ-10) più dieci caratteristiche individuali che si sono dimostrate efficaci nel rilevare i casi di ASD dai controlli nella scienza del comportamento.

4.1.2 ANALISI DOMINIO

Dato il set di dati utilizzato e le informazioni sugli attributi, abbiamo deciso di dividere gli attributi del dataset in 'ciò che deve essere rappresentato', 'ciò che caratterizza ciò che deve essere rappresentato', 'ciò che può essere ricavato ' e 'ciò che può essere scartato'. Alla prima categoria appartengo le entità che abbiamo rappresentato nell'ontologia, alla seconda appartengono le proprietà delle entità rappresentate, mentre l'attributo 'Screening Score' e 'Used the screening app before' appartengono alla categoria 'ciò che può essere ricavato' e l'attributo 'Country of residence' appartiene alla categoria 'ciò che può essere scartato', decisione prese visto l'ambito del dominio osservato. Nel dominio osservato ci sono dieci domande comportamentali, che non sono definite all'interno della documentazione del dataset, quindi abbiamo deciso di considerare queste domande, domande standard ovvero tutte uguali nei vari test, dato che hanno stesso codominio (Booleano) e la loro variazione è in funzione del metodo utilizzato. Abbiamo deciso di rappresentare come classi dell'ontologia:

- Paziente: La classe Paziente, rappresenta l'insieme dei pazienti sottoposti al test. A questa classe abbiamo associato diverse proprietà:
    - Age: rappresenta l'età del paziente al momento del test.
    - Gender: rappresenta il genere del paziente.
    - isAutistic: vera se paziente è stata già riscontrata una forma di autismo.
    - Ittero: vera se il paziente è nato con una forma di ittero.
    - PDDparent: vera se una parente stretto del paziente presenta forme di PDD.

    - Etnia: rappresenta l'etnia del paziente.
    - IdPaziente: rappresenta il paziente in modo univoco.
- Test: La classe Test, rappresenta l'insieme dei test svolti da tutti i pazienti. A questa classe sono associati tre proprietà:
    - Metodo: rappresenta la categoria del metodo utilizzato, che varia in base alla categoria di età del paziente (infante, bambino, adolescente, adulto).
    - CompilatoreTest: rappresenta chi ha compilato svolto il test con il paziente
    - IdTest: identifica univocamente un test
- Domanda: rappresenta la classe delle domande svolte. Abbiamo associato a questa classe due proprietà:
    - Risposta: vera se è vera la risposta del test.
    - NumQuestion: numero della domanda all'interno del test



Tra le classi abbiamo definito delle relazioni che rappresentano come queste interagiscono tra di loro. Le relazioni definite sono:

- Did_Test(Paziente) -> Test : permette di individuare i test svolti da un paziente.
- Has_Question(Test) ->Domanda: permette di individuare le risposte date dal paziente.

Informazioni come 'Screening Score' e 'Used the screening app before' possono essere derivate dall'ontologia realizzata con semplici query in Sparql

```
PREFIX esame:<http://www.semanticweb.org/miche/ontologies/2021/10/untitled-onto
select (count(?quest) as ?c) ?idP
where{
    ?p esame:IdPaziente ?idP.
    ?p esame:did_test ?t.
    ?t esame:has_question ?quest.
    ?quest esame:Risposta ?answ.
  filter(?answ = True)
}
group by ?idP
```

QUERY RESULTS

Table | Raw Response

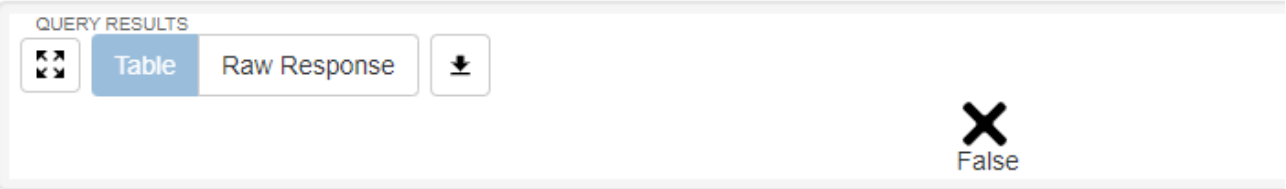
Showing 1 to 3 of 3 entries

	c	idP
1	"6"^^xsd:integer	"0102BC"
2	"7"^^xsd:integer	"01AB"
3	"6"^^xsd:integer	"0201FG"

Query svolta utilizzando Apache Jena Fuseki per derivare dall'ontologia Screening Score, con relativo risultato, che mostra il numero di domande la cui risposta è 'Vero' raggruppate

per paziente.

```
ask
where{
    ?p esame:did_test ?d.
    ?p esame:IdPaziente ?idP
  filter(?idP = "0201BC")
}
```



Query per derivare ‘Used the screening app before’, vera se il paziente ha già effettuato un test.

```
select ?numQ ?idP
where{
    ?p esame:did_test ?t.
    ?p esame:IdPaziente ?idP.
    ?p esame:isAutistic ?isA.
    ?t esame:has_question ?quest.
    ?quest esame:Risposta ?ans.
    ?quest esame:NumQuestion ?numQ
  filter(?ans = True && ?isA = True)
}
```

QUERY RESULTS

Table | Raw Response

Showing 1 to 12 of 12 entries

Search:

	numQ	idP
1	"10"^^xsd:integer	"0102BC"
2	"1"^^xsd:integer	"0102BC"
3	"3"^^xsd:integer	"0102BC"

	numQ	idP
1	"10"^^xsd:integer	"0102BC"
2	"1"^^xsd:integer	"0102BC"
3	"3"^^xsd:integer	"0102BC"
4	"5"^^xsd:integer	"0102BC"
5	"7"^^xsd:integer	"0102BC"
6	"9"^^xsd:integer	"0102BC"
7	"2"^^xsd:integer	"0201FG"
8	"3"^^xsd:integer	"0201FG"
9	"4"^^xsd:integer	"0201FG"
10	"6"^^xsd:integer	"0201FG"
11	"8"^^xsd:integer	"0201FG"
12	"9"^^xsd:integer	"0201FG"

Query che permette di conosce le domande la cui risposta è vera dei pazienti autistici.