I modelli di Machine Learning forniscono sempre delle previsioni che hanno un carattere probabilistico. Anche quando riportiamo dei risultati in forma assoluta (es: la diagnosi fornita dal modello è positiva, vi è la malattia) dobbiamo ricordarci che, dietro le quinte, il modello ha fornito una probabilità superiore ad una soglia che abbiamo fissato, e, in base a questa soglia, abbiamo deciso che il modello ha detto: "SI"

Ovviamente, non possiamo attenderci che il modello produca sempre risultati corretti. Anche se ci limitiamo alla valutazione su un insieme di dati di test, in generale su molti di questi dati il modello fornirà una previsione corretta, ma su alcuni, si spera pochi, produrrà una previsione sbagliata.

E' importante misurare le prestazioni del modello

**PRECISION E RECALL**

Precision e recall sono due comuni classificazioni statistiche. La precision può essere vista come una misura di esattezza o fedeltà mentre la recall (recupero) è una misura di completezza. Sono due indicatori usati nel machine learning per valutare la qualità di un modello decisionale o di un modello predittivo. Sono spesso correlate inversamente. Quando miglioro la precisione, peggiora la sensibilità del modello. E viceversa. Il più delle volte devo cercare un punto intermedio di equilibrio. Se poi il modello è sia preciso c

Nell’information Retrival la precision è definita come il numero di documenti attinenti recuperati da una ricerca fratto il numero totale di docuemnti recuperati. La recall è definita come il numero di documenti attinenti recuperati da una ricerca diviso il numero totale di documenti attinenti esistenti che dovrebbero essere stati recuperati.

In un processo di classificazione statistica:

* ­la **precisione** per una classe è il numero di veri positivi (il numero di oggetti etichettati correttamente come appartenenti alla classe) diviso il numero totale di elementi etichettati come appartenenti alla classe (la somma di veri positivi e falsi positivi, che sono oggetti etichettati erroneamente come appartenenti alla classe). È il rapporto tra il numero delle previsioni corrette di un evento (classe) sul totale delle volte che il modello lo prevede. Quando un modello è preciso per una classe, ogni volta che prevede l’evento sbaglia raramente. I falsi positivi sono pochi. Potrebbe però non prevedere tutti gli eventi ossia non essere selettivo/sensibile. I falsi negativi potrebbero essere molti anche se il modello è preciso.

La sua definizione fa sì che basti avere **una** previsione positiva corretta per avere una precisione del 100%, oppure **nessuna previsione falsa positiva** e comunque la precisione sarebbe del 100%.

* **Recall** in questo contesto è definito come il numero di veri positivi diviso il numero totale di elementi che effettivamente appartengono alla classe (per esempio la somma di veri positivi e falsi negativi, che sono oggetti che non sono stati etichettati come appartenenti alla classe ma dovrebbero esserlo). Misura la sensibilità del modello, rapporto tra le previsioni corrette per una classe sul totale dei casi in cui si verifica effettivamente. Quando un modello è sensibile per una classe, lo prevede ogni volta che si verifica. Potrebbe però prevederlo anche quando non si verifica. I falsi positivi potrebbero esser molti. Indica il rapporto di istanze positive correttamente individuate dal sistema di Machine Learning.

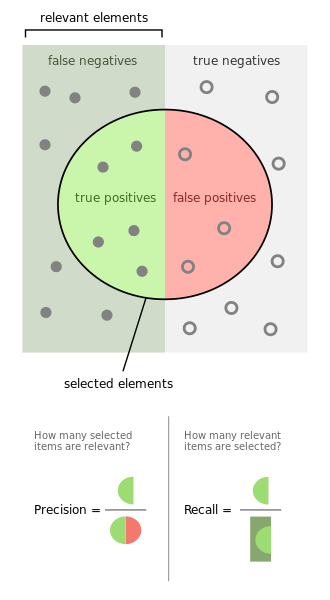
La precisione può essere vista come una misura della qualità e il richiamo come una misura della quantità. Una precisione maggiore significa che un algoritmo restituisce risultati più rilevanti di quelli irrilevanti e un richiamo elevato significa che un algoritmo restituisce la maggior parte dei risultati rilevanti (indipendentemente dal fatto che vengano restituiti anche quelli irrilevanti)

La precisione e il richiamo non sono metriche particolarmente utili se usati singolarmente. Ad esempio, è possibile avere un richiamo perfetto semplicemente recuperando ogni singolo oggetto. Allo stesso modo, è possibile avere una precisione quasi perfetta selezionando solo un numero molto piccolo di elementi estremamente probabili.

Spesso esiste una relazione inversa tra precisione e richiamo, dove è possibile aumentarne uno a costo di ridurre l'altro

**Definizione (classificazione)**

In un'attività di [classificazione](https://en.wikipedia.org/wiki/Classification_(machine_learning)) , la precisione per una classe è il *numero di veri positivi* (cioè il numero di elementi correttamente etichettati come appartenenti alla classe positiva) *diviso per il numero totale di elementi etichettati come appartenenti alla classe positiva* (cioè la somma di veri positivi e [falsi positivi](https://en.wikipedia.org/wiki/Type_I_and_type_II_errors) , che sono elementi etichettati erroneamente come appartenenti alla classe). Il richiamo in questo contesto è definito come il *numero di veri positivi diviso per il numero totale di elementi che effettivamente appartengono alla classe positiva* (cioè la somma di veri positivi e [falsi negativi](https://en.wikipedia.org/wiki/Type_I_and_type_II_errors) , che sono elementi che non sono stati etichettati come appartenenti alla classe positiva ma avrebbe dovuto essere).

In un processo di classificazione, un valore di precisione di 1.0 per la classe C significa che ogni oggetto che è stato etichettato come appartenente alla classe C vi appartiene davvero (ma non dice niente sul numero di elementi della classe C che non sono stati etichettati correttamente) mentre un valore di recupero pari ad 1.0 significa che ogni oggetto della classe C è stato etichettato come appartenente ad essa (ma non dice niente sul numero di elementi etichettati non correttamente con C).

In un processo di classificazione, i termini vero positivo, vero negativo, falso positivo e falso negativo sono usati per confrontare la classificazione di un oggetto (l'etichetta di classe assegnata all'oggetto da un classificatore) con la corretta classificazione desiderata (la classe a cui in realtà appartiene l'oggetto).

Per le attività di classificazione, i termini *veri positivi* , *veri negativi* , *falsi positivi* e *falsi negativi* (vedere [Errori di tipo I e II](https://en.wikipedia.org/wiki/Type_I_and_type_II_errors) per le definizioni) confrontano i risultati del classificatore sottoposto a test con giudizi esterni attendibili. I termini ***positivo*** e ***negativo*** *si* riferiscono alla previsione del classificatore (a volte nota come *aspettativa* ), mentre i termini ***vero*** e ***falso*** *si* riferiscono al fatto che tale previsione corrisponda al giudizio esterno (a volte noto come *osservazione* ).

Precisione e recupero sono definite come:

Precision =

Recall =

**La domain knowledge** è fondamentale per designare la **win condition** del nostro lavoro:

**Esempio.**Supponiamo che un programma per computer per riconoscere i cani (l' elemento **rilevante** ) nelle fotografie identifichi otto cani in un'immagine contenente dieci gatti e dodici cani, e degli otto identifichi come cani, cinque sono effettivamente cani (veri positivi), mentre gli altri tre sono gatti (falsi positivi). Sette cani sono stati persi (falsi negativi) e sette gatti sono stati correttamente esclusi (veri negativi). La precisione del programma è quindi 5/8 (veri positivi / tutti positivi) mentre il suo richiamo è 5/12 (veri positivi / elementi rilevanti).

**Esempio**. Un antifurto è molto preciso se ogni vuolta che suona c'è un ladro in casa. Non suona se passa il gatto o un topo.  
Un antifurto potrebbe non suonare quando c'è un ladro in casa (falso negativo). Il suo livello di precisione resterebbe comunque elevato perché la precisione non misura le volte che il modello non prevede l'evento (falsi negativi). In questo caso, il modello è preciso ma poco sensibile. Un antifurto è sensibile se suona ogni volta che c'è un ladro in casa. Potrebbe però suonare anche se passa un gatto o un topo. Un antifurto sensibile scatta al minimo momento in casa. Seleziona tutti i casi in cui c'è il ladro, ma anche quelli in cui passa il gatto (falso positivo). In questo caso, il modello è sensibile ma non preciso.

**Esempio.** Immaginiamo di avere un classificatore allenato a riconoscere e distinguere video sicuri per bambini di tenera età. Configuriamo il problema con: safe, classe positiva and danger, classe negativa

E’ preferibile optare per un sistema che sacrifichi qualche video sicuro ma che blocchi tutti quelli pericolosi, a costo quindi di classificare pericoloso qualche video sciuro (tanti False Negative).

Invece tutti i video sicuri saranno effettivamente tali, perché vogliamo avere zero Falsi Positivi. In questo modo avremo una low recall, con high precision.

Di solito, i punteggi di precisione e richiamo non vengono discussi isolatamente. Invece, i valori di una misura vengono confrontati per un livello fisso sull'altra misura (ad esempio, la *precisione a un livello di richiamo di 0,75* ) o entrambi vengono combinati in una singola misura. Esempi di misure che sono una combinazione di precisione e richiamo sono la [misura F](https://en.wikipedia.org/wiki/Precision_and_recall#F-measure) (la media [armonica](https://en.wikipedia.org/wiki/Harmonic_mean) ponderata di precisione e richiamo), o il [coefficiente di correlazione di Matthews](https://en.wikipedia.org/wiki/Matthews_correlation_coefficient) , che è una [media geometrica](https://en.wikipedia.org/wiki/Geometric_mean) delle varianti corrette per caso: i [coefficienti di regressione](https://en.wikipedia.org/wiki/Regression_coefficient)[Informedness](https://en.wikipedia.org/wiki/Informedness) ( DeltaP ') e [Markedness](https://en.wikipedia.org/wiki/Markedness" \o "Marcatura) (DeltaP).

La Precisione peggiora se vi sono tanti falsi positivi. Il Recall peggiora se vi sono tanti falsi negativi.

**Misura F**

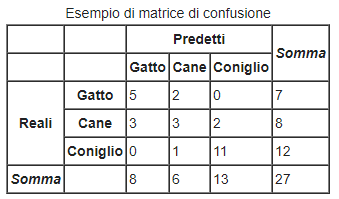
Una misura che combina precision e recall è la media armonica di precision e recall, la misura F

Questa misura è approssimativamente la media delle due quando sono vicine, ed è più in generale la [media armonica](https://en.wikipedia.org/wiki/Harmonic_mean) , che, nel caso di due numeri, coincide con il quadrato della [media geometrica](https://en.wikipedia.org/wiki/Geometric_mean) diviso per la [media aritmetica](https://en.wikipedia.org/wiki/Arithmetic_mean).

Spesso è conveniente fondere Precision e Recall in una sola metrica chiamata F1 score (media armonica – reciproco della media aritmetica dei reciproci). Rispetto ad una media convenzionale, quella armonica attribuisce un peso maggiore ai valori piccoli. Questò fa sì che un classificatore ottenga un alto punteggio F1 solo quando precisione e recupero sono entrambi alti.

**MATRICE DI CONFUSIONE**

Per valutare un modello di machine Learning per l’analisi di grandi dataset. Si può vedere se si è verificato un errore o se le previsioni sono più o meno corrette, per determinare quanto un modello predittivo di classificazione sia accurato ed efficace.

Detta anche tabella di errata classificazione restituisce una rappresentazione dell’accuratezza di una classificazione. Ogni colonna della matrice rappresenta i valori predetti, ogni riga i valori reali. L'elemento sulla riga i e sulla colonna j è il numero di casi in cui il classificatore ha classificato la classe "vera" i come classe j. Attraverso questa matrice è osservabile se vi è "confusione" nella classificazione di diverse classi. Attraverso l'uso della matrice di confusione è possibile calcolare il coefficiente kappa.

Esempio: si distinguono tre classi: gatto, cane e coniglio. Nelle righe si scrivono i valori veri, reali. Mentre nelle colonne quelli predetti, stimati dal sistema. Nell'esempio si può notare che dei 7 gatti reali, il sistema ne ha classificati 2 come cani. Allo stesso modo si può notare come dei 12 conigli veri, solamente 1 è stato classificato erroneamente. Gli oggetti che sono stati classificati correttamente sono indicati sulla diagonale della matrice, per questo è immediato osservare dalla matrice se il classificatore ha commesso o no degli errori.

Immagine che contiene tavolo

Descrizione generata automaticamenteLa matrice di confusione fornisce tutti i dati per valutare accuratamente se sono più presenti i FP o i FN

Una considerazione importante nello sviluppo del modello: come dicevamo prima, ogni modello setta una soglia di decisione (es: se la probabilità è > 0.7 decido: POSITIVO). Possiamo ridurre il numero di FP alzando la soglia di decisione, ma in questo modo aumentiamo anche i FN. E' un "tradeoff".