Una breve introduzione al Machine Learning

Davide Giosué Lippolis

13/01/2021

Machine learning o analisi statistica?

Mai più minimi quadrati

Machine Learning	Statistica
Permette generalizzazione	Utile a scovare relazioni tra le variabili
Big Data	Small Data
Funziona anche quando il numero di features è grande	Il rapporto tra numero di features e quantità di dati dev’essere piccolo
Alta accuratezza predittiva	Spiegabilità

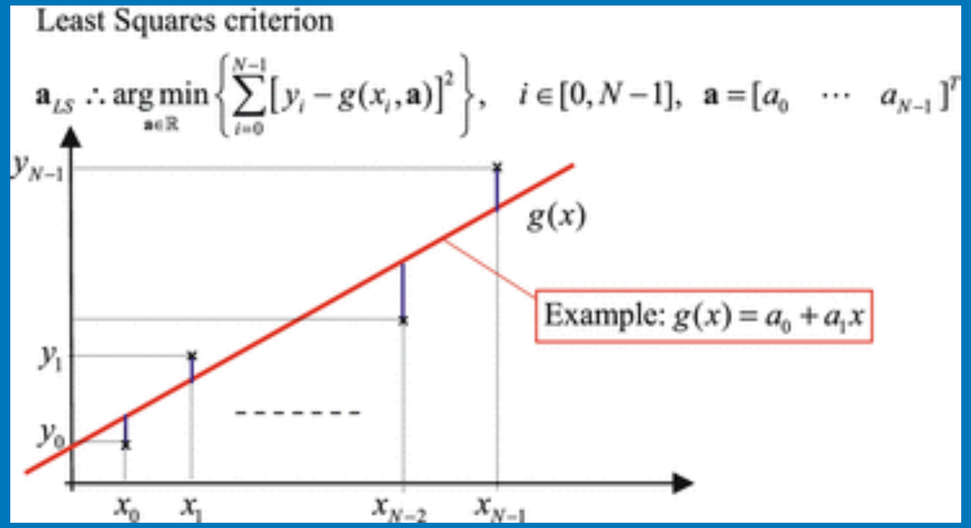
Cos’è il Machine Learning

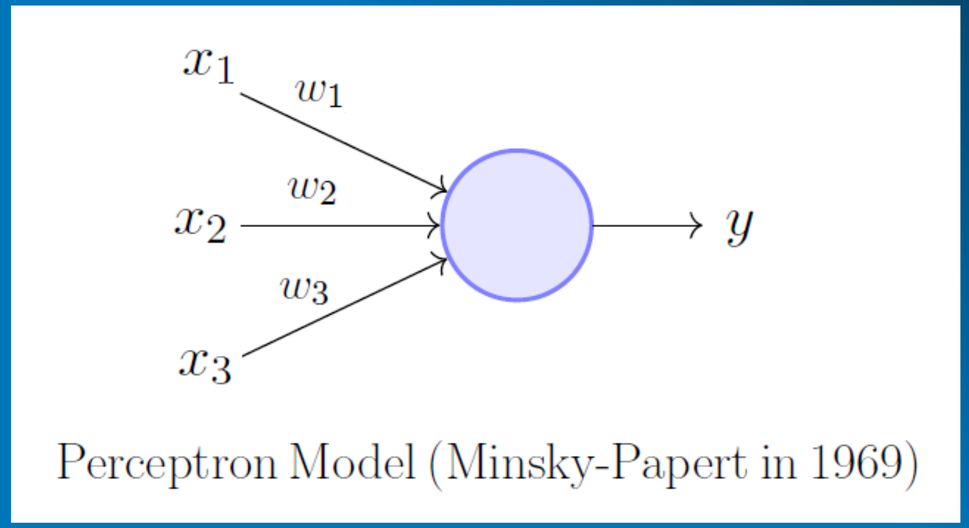
“Il Machine Learning è un'applicazione dell'intelligenza artificiale (AI) che fornisce ai sistemi la capacità di apprendere e migliorare automaticamente dall'esperienza senza essere programmati esplicitamente.
L'apprendimento automatico si concentra sullo sviluppo di programmi per computer in grado di accedere ai dati e utilizzarli per apprendere da soli”.[1]

“La Computational Statistics punta alla creazione di algoritmi per implementare metodi statistici su calcolatori, includendo anche quelli impensabili prima dell’avvento dei computer (per esempio Bootstrap, Simulation), insieme alla possibilità di risolvere problemi intrattabili analiticamente”.[1]

1. “Machine Learning", Tom Mitchell, McGraw Hill, 1997
2. “Computational statistics or statistical computing, is that the question?”, Carlo Lauro, The statistical software newsletter, 1996

Mai più minimi quadrati?





Perceptron Model (Minsky-Papert in 1969)

$$y = f\left(W \cdot x + w_0\right) \implies y = w \cdot x + w_0$$

$$\frac{\partial}{\partial W} MSE = \frac{\partial}{\partial W} \sum_i^N (y_{GTi} - y_i)^2 = 0$$

$$a_{LS} = argmin\left(\sum_i^N (y_i - a_0 - a_1 x_i)^2 \right)$$

$$w_{ML} = argmin\left(\sum_i^N (y_i - w_0 - w x_i)^2 \right)$$

Le quattro classi di ML

1. Supervised learning
   Necessita che i dati di training siano provvisti di label (ground truth).

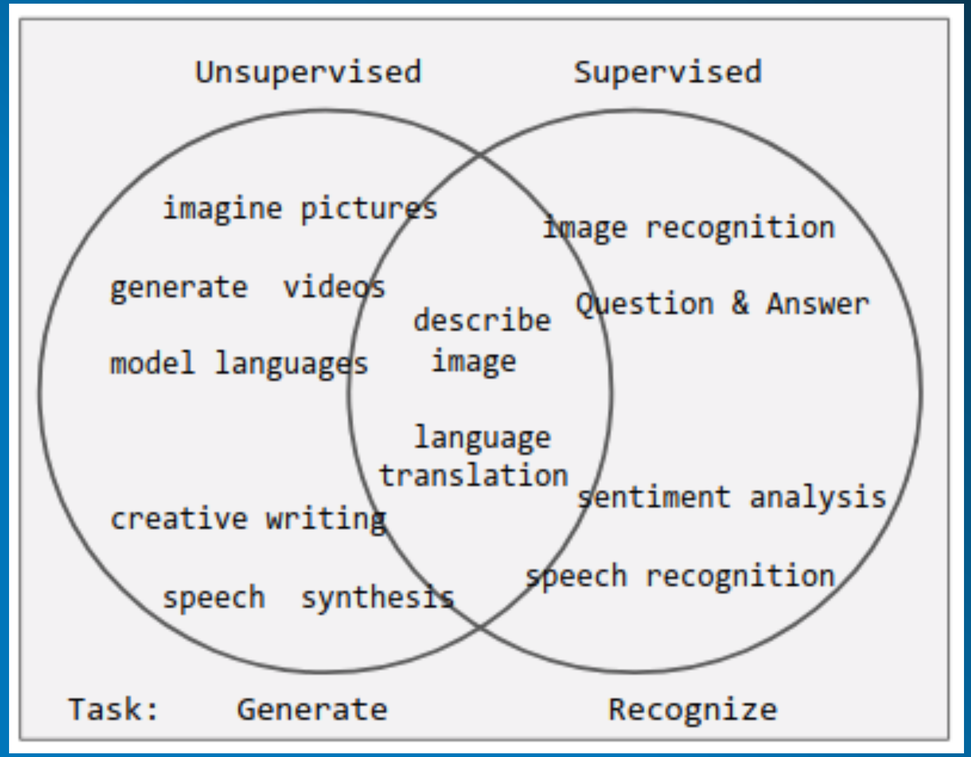
   - Regressione
   - Classificazione

2. Semi-supervised learning
   Metodi compositi che utilizzano sia tecniche di apprendimento supervisionato che non supervisionato.
   Molto importanti per applicazioni mediche.

3. Unsupervised learning
   Non “necessitano” di ground truth.

   - Clustering
   - Dimensionality reduction e feature extraction

4. Reinforcement learning
   Metodi basati sulla teoria decisionale e sul concetto di feedback.



https://en.wikipedia.org/wiki/File:Task-guidance.png

I principali algoritmi di ML

Supervised (regressione)	Supervised (classificazione)	Unsupervised
Regressione lineare	Logistic Regression	K - means
Regressione polinomiale	Support Vector Machines	Gaussian Mixtures
Regressione iterativa (stepwise)	K - nearest neighbors	Principal Component Analysis
Metodi Bayesiani (Hierarchical methods, Processi Gaussiani)	Alberi decisionali (Random Forests)	Independent Component Analysis
Regressioni penalizzate Ridge / Lasso / Elastic Net	Deep neural networks	Autoencoders

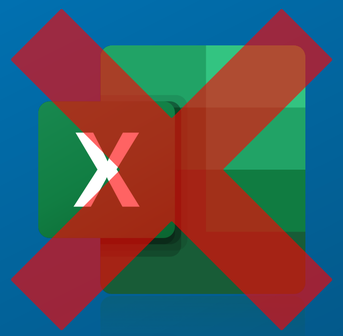
Pipeline analisi dati

- Data collection
- Data cleaning (cleansing)
- Feature extraction (labeling, feature selection, dimensional reduction)
- Training del modello
- Validazione modello (testing)
- Visualizzazione risultati

Data collection

La sfida più importante che impongono i Big Data è quella dello storage.

Analisi di immagini o video ad alta risoluzione possono richiedere quantità di dati da processare dell’ordine dei Terabyte.



Data cleaning

La maggior parte del tempo macchina è utilizzato per il training del modello.

La maggior parte del tempo uomo è utilizzato per il preprocessing del dataset.

Alcune pratiche comuni comprendono:

- Rimozione di dati superflui
- Rimozione di dati duplicati
- Rimozione refusi, tipi di dati errati
- Sostituzione dei dati mancanti

Feature Extraction (Selection)

Sono una componente fondamentale dell'analisi dei Big Data, permettendo di ridurre drasticamente la dimensione dello spazio delle features.
In particolare:

- Migliorano l'accuratezza dei modelli utilizzati
- Aumentano la velocità di training
- Rendono i modelli utilizzati maggiormente comprensibili (anche da un punto di vista visuale)
- Riducono il rischio di overfitting

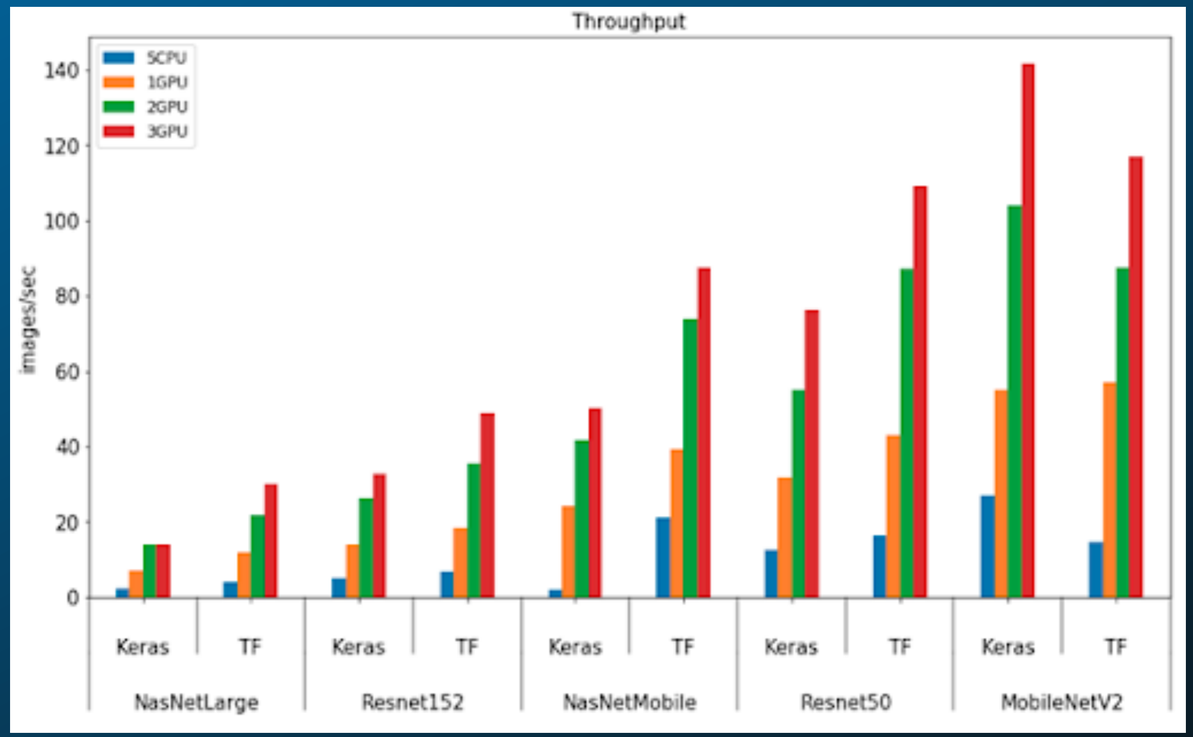
Definizione e Training del modello

Il processo di training è lo stadio computazionalmente più complesso. Tramite esso viene effettuato il tuning dei parametri del modello al fine di massimizzare la sua predittività.

La scelta del modello dipende dal problema in questione. L'esperienza di chi effettua l'analisi gioca un ruolo fondamentale in questo stadio.

Oltre all'accuratezza dei risultati, alcuni parametri importanti da considerare sono:

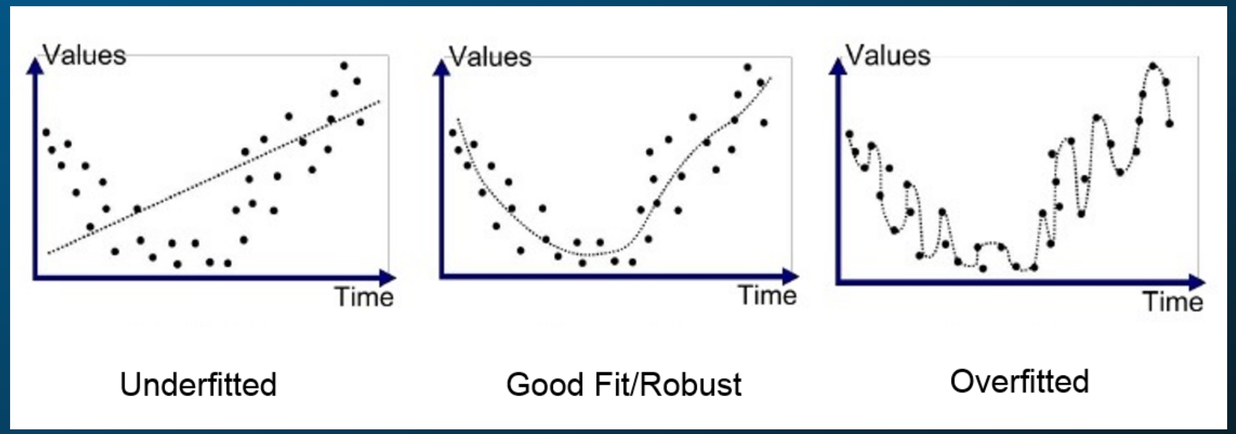
- Tempo a disposizione
- Potenza di calcolo a disposizione
- Complessità del problema



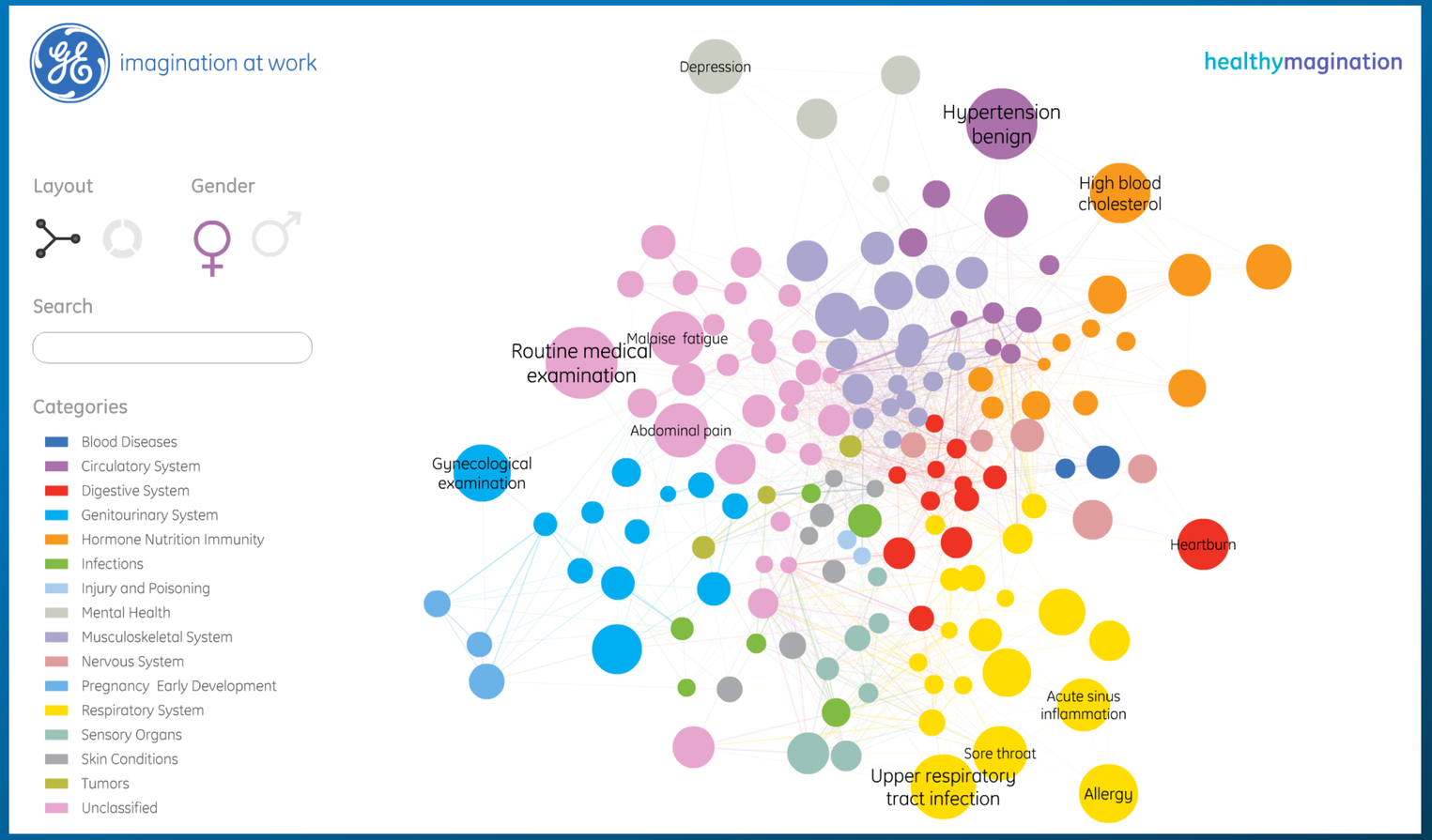
Testing delle performance del modello

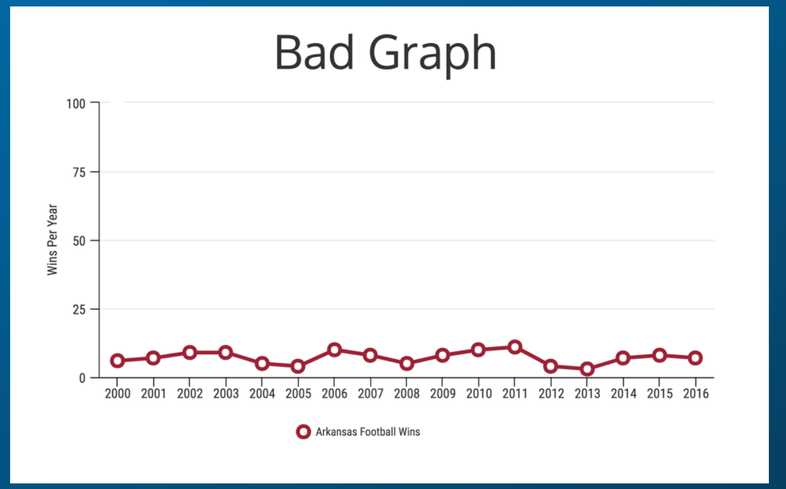
- Scelta delle metriche più importanti in base all'applicazione
- Definizione delle soglie di accettazione del modello
- Testing del modello su un campione indipendente di dati
- Valutazione overfitting o underfitting

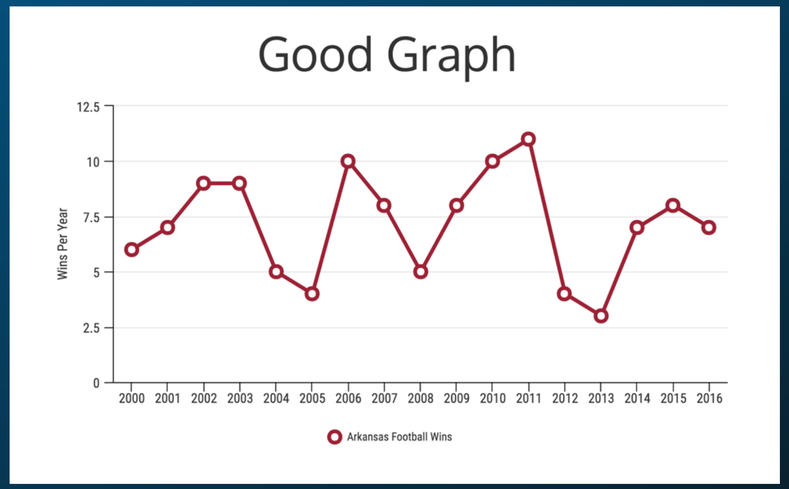
		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$



Data visualization

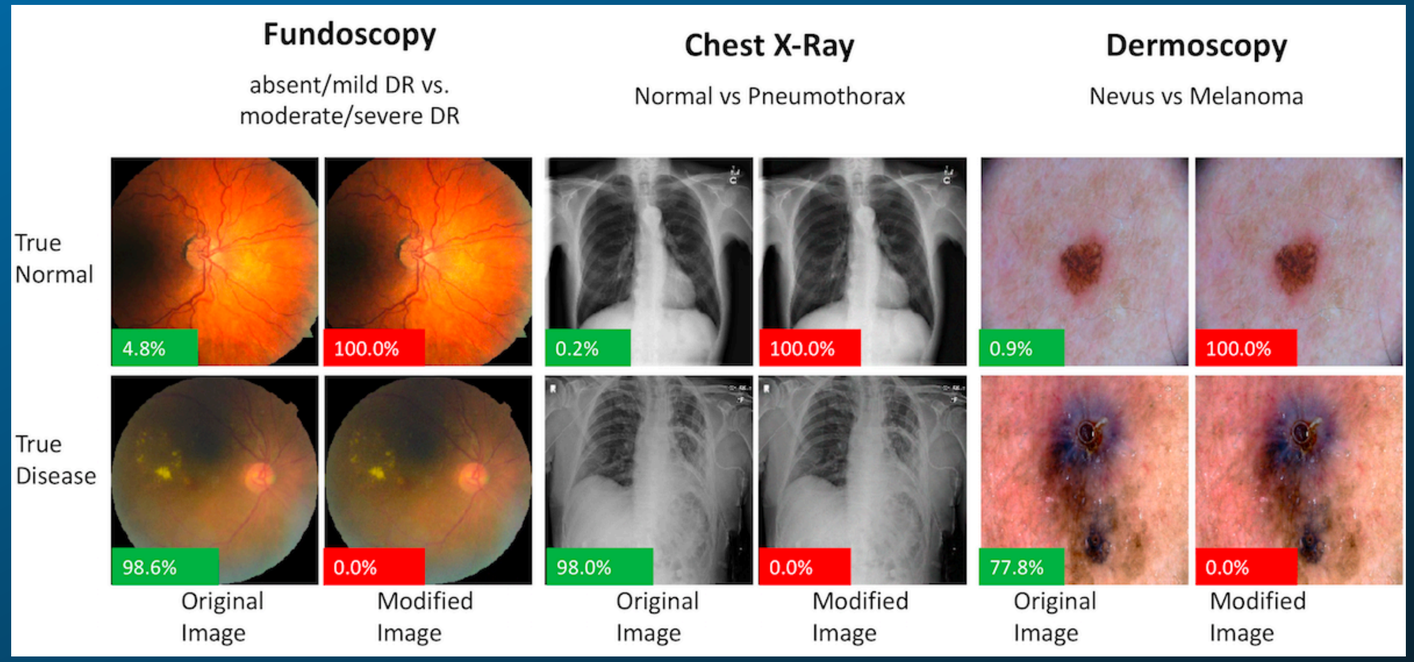






Limiti e criticità del Machine Learning

- Necessità di raccolta dati massiva per risoluzione di problemi sufficientemente complessi
- Annotazione dei dati costosa e spesso non accurata
- Scarsa spiegabilità dei risultati ottenuti
- Incapacità di riconoscere bias nei dati
- Difficile trasferimento di conoscenza tra problemi concettualmente simili



Q & A