**COS’È L’APPRENDIMENTO SUPERVISIONATO?**

L’**apprendimento supervisionato** è un tipo di *Machine Learning* in cui il modello impara a **prevedere un valore (*output*)** osservando dei dati **etichettati (*input* + risposta)**.

Per fare un esempio al fine di capirne il suo significato, consideriamo un paio di misure dei valori dei parametri che potrebbero essere coinvolti nel nostro progetto:

Misura → **[12.4, 7.3, 91.2, 34.1, 22.7, 10.0, 65.5]**

Etichetta → “***medium***”.

Misura → **[12.5, 25.3, 18.7, 10.2, 5.1, 33.4, 22.0]**

Etichetta → “***low***”.

Misura → **[85.2, 72.1, 91.3, 88.7, 79.5, 95.4, 90.1]**

Etichetta → “***high***”.

Il modello riceve in *input* molti esempi del genere e impara a generalizzare: ossia, impara a capire a quale classe (*low*, *medium*, *high*) appartiene una nuova misura mai vista prima. **L’obiettivo è quindi quello di riuscire a realizzare un algoritmo di Intelligenza Artificiale che, data una misura in *input*, restituisca l’etichetta corretta.**

**LA *RANDOM FOREST***

La tecnica di apprendimento automatico utilizzata in questo algoritmo, che si basa proprio sull’approccio supervisionato, è la ***Random Forest***: si tratta di un insieme di **alberi decisionali** (*Decision Trees*) che, combinando l’*output* di ciascuno, si raggiunge un unico risultato.

Un **albero di decisione** è un modello che prende decisioni in base a **domande sui dati**.

Immagine che contiene linea, diagramma, design

Il contenuto generato dall'IA potrebbe non essere corretto.

Gli alberi decisionali iniziano con un **nodo radice** (il nodo radice rappresenta un intero set di dati da cui parte il tutto per l’appunto)per poi diramarsi nei vari sotto-nodi arrivando fino ai ***Leaf-Node*** (nodi foglia) che rappresentano il risultato finale dell’intero albero decisionale. I nodi interni rappresentano funzionalità che dividono i dati in sottoinsiemi secondo metriche e specifiche peculiari e, a seguire, i nodi foglia sono quei sottoinsiemi di dati che non possono essere suddivisi ulteriormente. Forniamo un piccolo esempio di albero decisionale secondo le misure adottate nel progetto:

*feature\_3 > 50?*

*/ \*

*sì no*

*/ \*

*high feature\_1 > 30?*

*/ \*

*sì no*

*low medium*

come si può notare, ogni nodo fa una **domanda** e ogni foglia dà una **risposta finale** (la classe).

Pertanto, l’algoritmo esegue questi passaggi:

* Costruisce tanti alberi (nel nostro caso sono 100);
* Ogni albero è leggermente diverso (usa solo un sottoinsieme dei dati e delle feature);
* Le previsioni finali sono decise a maggioranza.

Giusto per rendere l’idea, è come avere un gruppo di esperti che votano e decidono insieme.

Collegandoci al nostro codice

1. Prima si etichettano i dati

Immagine che contiene testo, Carattere, schermata, design

Il contenuto generato dall'IA potrebbe non essere corretto.

1. Addestro il modello usando per l’appunto una *Random Forest* da 100 alberi su dati già etichettati (X = misure, y = classi)

Immagine che contiene testo, schermata, Carattere, linea

Il contenuto generato dall'IA potrebbe non essere corretto.

1. Dai nuovi dati in *input* e il modello prevede la classe corretta.



Se immaginassimo di avere 3 Dataset fatti in questa maniera

Immagine che contiene schermata, tastiera

Il contenuto generato dall'IA potrebbe non essere corretto.

il modello vede queste etichette e impara delle regole, ad esempio: se *feature2* è maggiore di 50 e *feature7* è maggiore di 25 allora forse è “*high*” se invece *feature2* è minore di 20 e *feature1* è minore di 25 allora forse è “*low*”.

* **COME FUNZIONA IL *TRAINING*?**

Crea 100 alberi (o quanto scritto nel codice con *n\_estimators*=...).

Per ogni albero:

Sceglie un sottoinsieme casuale dei dati.

Usa un sottoinsieme casuale delle feature.

Crea un albero con domande tipo: "feature\_3 > 45?"

Salva gli alberi nel modello finale.

* **COME FUNZIONA LA PREVISIONE?**

Una nuova misura arriva (es. [23.1, 47.5, 12.0, 66.1, 42.3, 19.0, 88.5])

Il dato viene dato in *input* a tutti gli alberi.

Ogni albero fa la sua previsione (es: 60 alberi dicono "*high*", 30 "*medium*", 10 "*low*").

La classe più votata vince: "*high*".