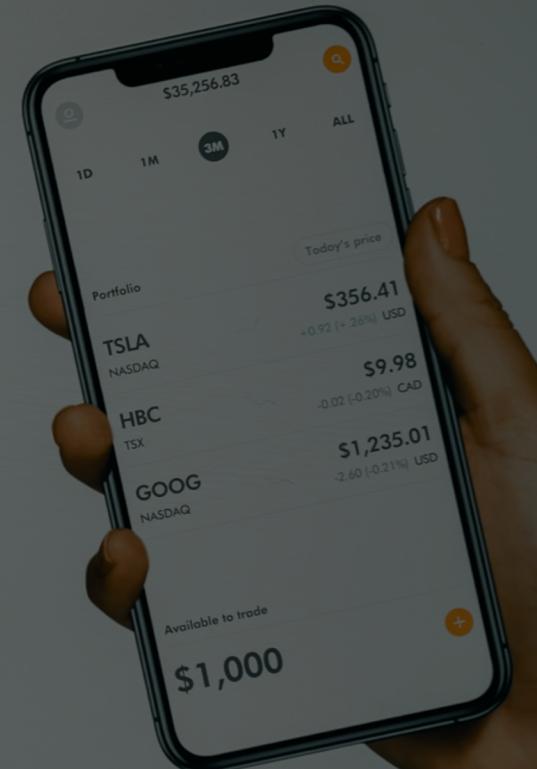


Progetto Big Data e Machine Learning

Progetto N° 62

Bernovschi Denis e Giacomo Licci



Introduzione

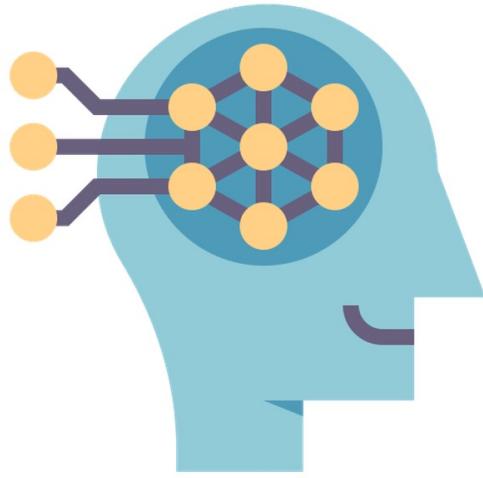
L'analisi si è svolta sulla base di dati proveniente dall'Università Politecnica delle Marche, inherente alle nuove iscrizioni nei percorsi triennali. In particolare l'obiettivo è analizzare il percorso di studi precedente e la correlazione con i percorsi universitari intrapresi. Infine si pone in analisi i risultati dei percorsi scelti.

Analisi

«La prima delle analisi si è concentrata sull'analisi del **percorso intrapreso**, i possibili risultati e l'influenza relativa al percorso precedente»

Analisi

«Abbiamo inoltre analizzato il tempo medio tra la conclusione del percorso precedente e l'inizio della carriera universitaria, valutando possibili relazioni tra essi e i vari cambiamenti di percorso in seguito all'iscrizione»



Modello Predittivo

Split Utilizzati

- ❖ SPLIT SEMPLICE
- ❖ SPLIT STRATIFIED SHUFFLE
- ❖ SPLIT STRATIFIED KFOLD
- ❖ SPLIT REPEATED STRATIFIED KFOLD

Modelli Utilizzati

- ❖ **Regressione Logistica:** Ci permette di generare un risultato che, di fatto, rappresenta una probabilità che un dato valore di ingresso appartenga a una determinata classe.
- ❖ **Regressione Lineare:** Quando tra due variabili c'è una relazione di dipendenza, si può cercare di prevedere il valore di una variabile in funzione del valore assunto dall'altra.

Modelli Utilizzati

- ❖ **KNN** (K-Nearest-Neighbor): è un approccio alla classificazione dei dati che stima la probabilità che un punto (dato) sia membro di un gruppo o dell'altro a seconda del gruppo in cui si trovano i punti (dati) più vicini (VOTING)
- ❖ **SVM** (Support Vector Machine)...l'obbiettivo è quello di trovare un iperpiano in uno spazio N-dimensionale con N, numero caratteristiche che classifica distintamente i dati in ingresso. È di tipo OVO (ONE versus ONE)
- ❖ **DECISION TREE**: L'apprendimento del Decision Tree o l'induzione degli alberi di decisione è uno degli approcci di modellizzazione predittiva... per passare dalle osservazioni su un elemento (rappresentato nei rami) alle conclusioni sul valore obiettivo dell'elemento (rappresentato nelle foglie)

Conclusioni

- ❖ I modelli citati precedentemente non sono in grado di classificare correttamente e quindi di conseguenza predire con un'accuratezza elevata, il motivo principale riscontrato è dovuto alla elevata sparsità del dataset di partenza.
- ❖ All'interno dello sviluppo del progetto sono stati testati diversi algoritmi di split al fine di ridurre questa problematica, ed ottenere migliori risultati, nonostante ciò non vi si registrano ulteriori miglioramenti ai risultati già riportati nella relazione



Rapid Miner

«A dimostrazione delle conclusioni appena citate, abbiamo utilizzato un’altro tool per dimostrare effettivamente la loro veridicità»

- ❖ KNN
- ❖ DECISION TREE
- ❖ SVM (versione beta)

Conclusioni

- ❖ Gli studenti provenienti da altre scuole più specializzate tendono a scegliere pochi percorsi specifici molto spesso coerente con il loro indirizzo di scuola superiore.
- ❖ Gli studenti di indirizzi scientifici prediligono diverse facoltà
- ❖ Il voto del percorso scolastico non influisce particolarmente il rendimento universitario.
- ❖ Un altro dato confortante deriva dall'immediata iscrizione all'università da parte di studenti appena diplomati

Grazie per l'attenzione