Lavoro di gruppo

07/02/2023

Come compito conclusivo di questa attività di PLS dovrete provare voi a costruire un albero decisionale su un dataset nuovo.

Suddivisione in gruppi

Innanzituto dovrete costruire dei gruppi da minimo **4** a massimo **6** persone. Per eventuali necessità (es. non lasciare studenti singoli, o classi numerose) si possono fare anche gruppi con più di 6 o meno di 4 persone. Una volta creati i gruppi dovete scegliere uno dei dataset presenti nella sezione in fondo. Per ogni dataset vi indichiamo quale variabile risposta analizzare, sta a voi scegliere invece quali variabili indipendenti considerare.

Analisi su R

Una volta scelto il dataset, procedete alla costruzione dell'albero decisionale con R. Utilizzate gli script delle lezioni in laboratorio per recuperare i codici. Nel vostro lavoro dovranno essere presenti i seguenti punti:

- **Analisi preliminare sulle variabili utilizzate per la costruzione dell'albero**: dovete conoscere le variabili che includete nel modello, per cui fate dei grafici descrittivi e calcolate degli indici di sintesi in modo da conoscere bene il vostro dataset. La prima cosa da controllare ogni volta è il formato delle variabili.
- **Rappresentazione grafica dell'albero**: una volta costruito il vostro albero è importante visualizzarlo graficamente per poter interpretare l'influenza delle variabili scelte sul risultato finale.
- **Valutare i risultati finali**: costruito il modello occorre valutare la sua bontà di previsione. Per farlo dovete calcolare l'indice di *accuracy* come fatto in laboratorio. Potete provare diversi modelli (con diverse variabili) per trovare quello migliore. Ricordate il principio di parsimonia, per cui a parità di accuracy la scelta migliore è sempre il modello più semplice, cioè con meno variabili.
- **Valutare importanza delle variabili esplicative**: infine studiare quali variabili sono più importanti nel vostro albero tramite il calcolo della *variable importance*. Ricordatevi sempre di commentare e di provare a interpretare perché una variabile è importante o meno nella stima finale. Anche qui, una volta valutata l'importanza delle variabili potete provare a tornare indietro e costruire un albero più semplice con solo le variabili più "importanti".

Nota: non esiste un procedimento unico e corretto per la costruzione dei modelli di previsione. Si va molto spesso a tentativi, il che vuol dire che potete fare diverse prove finché non trovate un risultato che vi soddisfa. L'importante è che alla fine specifichiate quale modello avete scelto.

Presentazione

Una volta fatte le analisi su R dovrete preparare una presentazione che riassuma il vostro lavoro. Per farlo vi chiediamo di preparare un massimo di 10 slides (compresa una di introduzione e una di conclusione e facendo due slides per ogni punto richiesto nella sezione precedente). Cercate di essere sintetici, ma anche di non trascurare i risultati principali e le scelte che avete fatto nella costruzione dell'albero. Ricordatevi infatti che dovete raccontare il vostro lavoro a chi non sa cosa avete fatto e non conosce gli strumenti che avete utilizzato. Devono essere presenti nella vostra presentazione **tutti** i punti riportati nella sezione precedente.

Elenco dataset

1. Titanic

Potete scaricare il dataset *Titanic* da Kaggle al link link (dovete iscrivervi al sito, basta semplicemente l'account Google per farlo). Utilizzate solamente il file *train.csv*, ignorate sia *test.csv* che *gender_submission.csv*.

Questo dataset raccoglie i nomi delle persone imbarcate sul Titanic. Dovrete capire quali variabili discriminano maggiormente i passeggerri che sono sopravvissuti (**survived** = 1) e quelli che invece non lo sono (**survived** = 0) durante il famoso disastro del 1912. La variabile da prevedere è dunquee *survived*.

Suggerimento: potrebbe essere utile stabilira se una persona sia sposata o meno usando il nome del passeggero. Creare dunque una variabile con "Mr.", "Mrs.", "Miss" o altro e vedere se può essere una variabile discrimante.

2. Home loan approval

Potete scaricare il dataset *Home loan approval* da Kaggle al link link (dovete iscrivervi al sito, basta semplicemente l'account Google per farlo). Utilizzate solamente il file *train.csv*, ignorate sia *test.csv* che *gender_submission.csv*.

Questo dataset contiene 614 richieste di prestiti che sono state erogate o meno da parte di una banca. Lo scopo è prevedere in maniera automatica se erogare o meno un prestito a un soggetto basandosi sulle sue caratteristiche anagrafiche, di reddito e anche sulle cifre che esso richiede. La variabile risposta è *Loan_status*.

Suggerimento: convertite la variabile risposta *Loan_status* in una dicotomica 1/0 in modo da facilitare l'interpretazione. Con 1 indicate se il prestito è stato erogato, con 0 se non è stato erogato.

3. Bank Customer Churn Dataset

Potete scaricare il dataset *Bank Customer Churn Dataset* da Kaggle al link link (dovete iscrivervi al sito, basta semplicemente l'account Google per farlo).

Lo scopo di questo lavoro è prevedere se un cliente di una certa banca cambierà o meno banca (churn). La vairiabile risposta da studiare è *churn* che assume valore 1 se il cliente se ne è andato, 0 se è rimasto. Provare a costruire un albero e capire quali variabili possono essere più indicative di un possibile cambiamento.

4. Mushroom Dataset

Questo dataset *Mushroom Dataset* è scaricabile dalla repository UCI al link link (senza alcuna iscrizione, cliccarte su "data folder" e scaricate il file .zip, al cui interno troverete il set di dati e metadati chiamato "primary_data.csv").

L'obiettivo è prevedere la velenosità o meno (rispettivamente p="poisonous", e="edible") per un campione di funghi, basandosi sulle caratteristiche osservate. Le guide affermano chiaramente che non esiste una regola semplice per determinare la commestibilità di un fungo. Provate a trovare un insieme di regole attraverso la costruzione di un albero di classificazione, carcando di capire quali sono le caratteristiche più importanti per identificare un fungo velenoso.

5. Heart Failure Dataset

Questo dataset *Heart Failure Dataset* è scaricabile dalla repository UCI al link link (senza alcuna iscrizione, cliccarte su "data folder" e scaricate il file "heart_failure_clinical_records_dataset.csv").

Questo dataset contiene 299 documentazioni mediche di pazienti che hanno avuto un infarto. Lo scopo è prevedere se questo evento sia stato o meno letale per un soggetto basandosi su 13 delle sue caratteristiche clinico-anagrafiche.