A.A. 2023/2024

Progetto – case study modello di regressione

Corso di Statistica – Ingegneria informatica (cod. 21060)

Docenti: Rodolfo Metulini, Alice Giampino, Lorenzo Leoni

Ad ogni gruppo sarà assegnato dai docenti uno dei seguenti dataset (o parte di esso):

1. **Agrimonia**, contiene i dati di qualità dell'aria, meteorologia e attività agricola dal gennaio 2016 a dicembre 2021 come meglio dettagliato in Fassò et al, 2023, https://www.nature.com/articles/s41597-023-02034-0

2. **Forest Fires**, contiene dati utili a modellare gli incendi delle foreste in funzione di una serie di indicatori presenti nel dataset. Per maggiori dettagli: "A Data Mining Approach to Predict Forest Fires using Meteorological Data." In J. Neves, M. F. Santos and J. Machado Eds.,"New Trends in Artificial Intelligence", Proceedings of the 13th EPIA 2007 Portuguese Conference on Artificial Intelligence, December, Guimaraes, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9. Available at: http://www3.dsi.uminho.pt/pcortez/fires.pdf

3. **House price data**. Il dataset contiene un grande numero di variabili di diverso genere (quantitative, categoriche, etc…) riferite a 2930 case vendute nello stato dell'IOWA tra il 2006 e il 2010. Maggiori informazioni a questo link: https://jse.amstat.org/v19n3/decock.pdf

4. **Basketball teams**, contiene i dati sulle statistiche di squadra per singola stagione delle leghe professionistiche americane di basket, dal 1937 al 2021. Il dataset è disponibile da Kaggle. Al seguente url maggiori dettagli https://www.kaggle.com/datasets/open-source-sports/mens-professional-basketball/

Consegna

Utilizzare tecniche di regressione lineare o generalizzata in R Studio (o altro software a vostro piacere) per studiare una data variabile risposta (Y) in funzione delle altre variabili esplicative a disposizione nel dataset (covariate, X).

Identificare i modelli più adatti a descrivere Y in base alle conoscenze acquisite sui modelli di regressione lineari e generalizzati, sui criteri di valutazione e confronto tra modelli (significatività delle variabili esplicative, adattamento del modello ai dati, trasformazioni dei dati, diagnostica sui residui) e eventuali altri criteri (AIC, BIC, stepwise, cross validation, regolarizzazione).

Discutere e commentare le scelte effettuate.

Per finire, commentare ed interpretare i risultati ottenuti. Classiche domande di interesse sono: Quali covariate sono statisticamente significative? Qual è la performance del modello?

N.B.: Maggiori dettagli, quali scelta delle variabili Y ed X, definizione del subset di analisi, obiettivo di ricerca, etc … verranno forniti durante le lezioni del Giovedì.

Output da consegnare

- ✓ un report finale (preferibilmente svolto in R Markdown e consegnato in .html, o in power point /pdf) con breve descrizione del lavoro svolto (dalle 3 alle 7 pagine, inclusi codici ed eventuali tabelle e figure), le analisi svolte, i risultati principali e relativi commenti (come previsto dalla consegna). Il file deve essere rinominato usando la seguente sintassi: GYYYYY_Report.html, dove YYYYY indica il nome del gruppo (es. Favignana, Giglio, …). Nella prima pagina del report devono essere indicati i nomi, cognomi e le matricole di tutti i componenti del gruppo, oltre che il nome del gruppo;

- ✓ La presentazione del case study in ppt/pdf con la sintassi GYYYY_presentazione.pdf

Modalità di consegna e scadenza

Ogni gruppo dovrà inviare il report finale di cui sopra entro 7 giorni prima la data di appello (lunedì 8 Gennaio 2023 ore 23.59, per chi presenterà il 15 Gennaio) per posta a rodolfo.metulini@unibg.it e a.giampino@campus.unimib.it (ad entrambe le mail).

La presentazione può essere inviata invece anche il giorno stesso.

Discussione dei case studies

la discussione si svolgerà nelle date di appello ufficiale, cioè Lunedì 15 gennaio 2023 oppure Venerdì 9 febbraio 2023. Ciascun gruppo può decidere una delle due date, e dovrà comunicare la presenza ai docenti almeno una settimana prima della data (attraverso il foglio condiviso).

Ciascun gruppo deve essere composto da minimo 3 e massimo 4 persone. Tutti i componenti del gruppo dovranno intervenire nella discussione orale del progetto, che dovrà essere calibrata per durare 10 minuti.

Contatto

Ad ogni gruppo verrà associato un docente di riferimento (si veda il foglio condiviso).

Per ricevimenti e supporto, accordarsi via mail con il referente del gruppo.