

750AA - Statistica II

Indicazioni per il progetto personale di analisi dei dati

Ciascun progetto è un lavoro personale e deve essere svolto (principalmente) su uno degli argomenti visti nel corso:

1. regressione;
2. classificazione (o eventualmente clustering);
3. analisi e previsione di serie storiche.

Tabelle di dati

Sarà necessario reperire le proprie tabelle di dati da fonti note, certificate e verificabili. Per fonti note, certificate e verificabili si intende dire che i dati siano reperibili presso un sito web di provata affidabilità (ad esempio, ma non solo, Istat, Eurostat, etc). Non sono ammessi come fonti di dati quei siti (ad esempio kaggle) che si prefiggono scopi formativi/didattici, che quindi propongono dataset opportunamente scelti e la cui analisi è già fornita in dettaglio.

Inoltre i progetti non sono lavori di gruppo, ed è importante che ogni candidata/o abbia una tabella di dati differente da quella degli altri. Avere una tabella in comune con altri invalida il progetto di tutti coloro che hanno le tabelle in comune.

Nella scelta dei dati, preferire ove possibile dati grezzi ed evitare di considerare fattori o tabelle già, in qualche modo manipolati (dati normalizzati, indici di riferimento, etc.).

La qualità dei dati scelti è parte della valutazione finale, perché mostra la capacità di riconoscere su quali tabelle di dati i metodi studiati siano più efficaci. Consiglio di cimentarsi immediatamente a reperire, scaricare e manipolare tabelle di dati.

Redazione delle relazioni

Ogni relazione deve avere una lunghezza limitata (circa 7 pagine ognuna). I vincoli di spazio non devono andare a danno della chiarezza espositiva, della solidità delle argomentazioni e della visibilità delle immagini.

Le conclusioni sui dati devono essere basate soprattutto sulle analisi svolte, ed eventualmente, in caso di fenomeni anomali, su fatti incontrovertibili (ad esempio, crisi economica del 2008, lockdown del 2020, etc) e non su considerazioni

personali poco oggettive e spesso opinabili. Ricordarsi di motivare con argomenti quantitativi le proprie scelte. Evitare una sterile ripetizione di schemi appresi a lezione.

Ogni relazione deve prevedere una breve introduzione alla tabella (cosa rappresentano i dati) e una breve descrizione dei dati stessi che metta in prospettiva e interesse aziendale i dati che si intendono studiare, allo scopo di identificare un problema che sarà poi risolto (eventualmente) dall'analisi svolta. Lo studio di dati "veri" ha senso quando c'è un obiettivo finale. Quale che sia la parte di progetto che si svolge, la candidata o il candidato devono immaginare una ambientazione plausibile per i propri dati e sulla base di questa fissare uno o più obiettivi le cui conclusioni possano poi essere ricavate (solo ed esclusivamente) dall'analisi dei dati. L'idea in altre parole è di "calarsi" nella parte di chi esamina un problema, per fornire a management/collaboratori il quadro più chiaro possibile del problema.

Detto questo, non si deve dimenticare che la relazione è una prova d'esame, dunque le considerazioni contenute nella relazione devono essere esplicitamente supportate da quanto appreso durante il corso.

Il documento principale del progetto dovrà contenere i link ai siti presso i quali le tabelle sono state tratte. Lo scopo è da una parte di garantire che i dati non siano stati "fabbricati" ai fini dell'analisi che si intende svolgere, dall'altra di permettere di verificare in breve tempo la correttezza del procedimento. All'interno del file del progetto devono essere presenti i link precisi alle tabelle considerate. Per link preciso si intende che selezionando tale link, il browser si apre alla tabella dei dati. La mancanza dei link o l'eccessiva laboriosità nel recuperare i dati invalidano il progetto.

Consegna delle relazioni

Ogni relazione sarà consegnata attraverso un modulo MS Forms, il cui indirizzo sarà comunicato agli iscritti ad ogni prova.

Ogni relazione dovrà essere consegnata esclusivamente in formato PDF (solo PDF, non sono accettati file tipo word o simili).

Il modulo richiederà anche l'upload della tabella di dati utilizzata, che dovrà essere contenuta in un file CSV (solo CSV, non sono accettati file tipo excel o simili).

Infine, si richiede di caricare anche il codice utilizzato per l'analisi. È Questo potrà essere inserito nella relazione come appendice, la cui lunghezza non rientra nel computo massimo delle pagine specificato in precedenza.