LINEE GUIDA PER LA PREPARAZIONE DEL PROGETTO PRELIMINARE ALL'ESAME

Il progetto può essere svolto individualmente oppure in team fino ad un massimo di tre persone, con un impegno individuale approssimativo di cinque giornate di lavoro. Il dataset può essere scelto dal candidato da siti noti come UCI (https://archive.ics.uci.edu/ml), Kaggle (https://www.kaggle.com/datasets), Amazon (http://aws.amazon.com/datasets), oppure da altri elencati qui https://bit.ly/30mpWdF oppure dataset di borsa (accessibili con package come fix_yahoo_finance, alpha_vantage, iexfinance, quandl), oppure anche dataset di organizzazioni/aziende che ne abbiano autorizzato l'utilizzo.
In base al dataset prescelto, il problema può riguardare la previsione di una variabile continua o discreta, nel primo caso il problema è di regressione, mentre nel secondo è di classificazione; in alternativa il dataset può riguardare un problema di recommendation dove comunque la variabile da predire è discreta (binaria 0/1 oppure un voto es. da 1 a 5).
Svolgere le seguenti attività implementando un file jupyter in python con le librerie software necessarie:

1a) descrizione del problema e comprensione dei dati, significato delle variabili, numero istanze, feature e descrizione della variabile da predire (continua, se discreta binaria o multiclasse e relative classi) etc.
1b) analisi esplorativa dei dati con distribuzioni, medie, stddev, percentili, numero valori distinti delle feature, feature scarsamente utili perché identificatori o perché ad elevata/scarsa variabilità, valori mancanti, grafici a dispersione, calcolo e visualizzazione della correlazione tra coppie di feature etc. In caso di mancanza o insufficienti correlazioni, si può ripetere l'analisi anche con feature non lineari.

2) normalizzazione/standardizzazione (necessarie ad esempio per evitare la prevalenza fittizia di feature con domini più grandi e difficoltà poi di interpretazione del modello o delle feature più rilevanti), discretizzazione, binarizzazione delle feature, individuazione con lasso - nel caso di regressione - delle feature più rilevanti ed eventuali collinearità (nella classificazione utilizzare la regolarizzazione con norma L1 per individuare le feature più rilevanti), oppure in caso di scarsa efficacia predittiva utilizzare anche feature non lineari come al punto 1b; l'efficacia predittiva con solo le feature più rilevanti può essere confrontata generando modelli che usano tutte le feature, anche usando altri algoritmi trattati a lezione.

3) generazione di diversi modelli di learning usando tutti gli algoritmi visti a lezione che siano applicabili al caso di studio prescelto impiegando random search e grid search in k (nested) cross fold validation allo scopo di individuare gli iperparametri migliori in modo corretto. Nel caso le classi del problema prescelto siano sbilanciate, applicare uno o più metodi per il trattamento di tali casi.

4) valutazione dei modelli di regressione, classificazione, recommendation con le metriche appropriate, selezione dei 2-3 modelli ritenuti migliori con motivazione della scelta, tra cui anche lo scarto quadratico medio dell'errore dei relativi iperparametri ottenuto in k cross fold validation. Valutazione dei modelli a regime con calcolo degli intervalli di confidenza predittivi fissata la confidenza del 95%.

5) Designare dall'analisi precedente il modello migliore, se si differenzia dagli altri 2-3, e spiegare /interpretare la conoscenza appresa attraverso l'analisi dei parametri appresi (coefficienti degli iperpiani), sia nella regressione, sia nella classsificazione, vale a dire quali feature sono più positivamente/negativamente correlate ed in che misura con la variabile da predire.

> *Ad esempio nella previsione dei consumi energetici trattata in aula e in lab, quanto i consumi dipendono dalla temperatura e quanto dal giorno della settimana, oppure nel caso di studio dei prezzi delle case quali sono le variabili che più influiscono sul prezzo ed in che misura lo aumentano/diminuiscono. Altro esempio: da un problema di classificazione volto a determinare la fascia delle spese di manutenzione previste di automobili, emerge che all'aumentare dell'età delle vetture non si rilevano aumenti significativi della fascia di spesa, invece l'incremento della mancanza di manutenzione negli anni precedenti comporta l'aumento delle spese di manutenzione per l'anno successivo. I coefficienti devono essere denormalizzati/destandardizzati per poterli interpretare nei domini originali (argomento trattato a lezione nella regressione che vale anche nella classificazione).*

Da ogni fase si può tornare a ripetere quelle precedenti in caso di esito insoddisfacente, dove per insoddisfacente intendiamo che l'efficacia del modello predittivo (F1-Measure nella classificazione oppure R^2 o MAE nella regressione) sia statisticamente equivalente, con confidenza del 99%, a quella di un modello casuale. Se per il dataset prescelto è nota la migliore efficacia, allora confrontarla in modo corretto con il proprio esito. La relazione deve contenere i 5 punti sopra indicati e può essere incorporata passo dopo passo nel file jupyter oppure in un pdf a parte di 4-5 pagine.
Il candidato/team deve consegnare: sorgente, pdf ed html del file jupyter commentato, la relazione ed i dati utilizzati, con almeno 5 giorni di anticipo rispetto alla prova orale, salvo diversa indicazione in **almaesami**, mediante condivisione con bitbucket o GitHub con relativo link per l'esecuzione del codice in colab o binder.