CORSO DI BIG DATA – MODULO ANALISI PER I BIG DATA

Prova scritta del 15 febbraio 2024

Si consideri il data set Titanic composto dai file `train.csv` e `test.csv`, allegati al presente compito, costituiti da 891+418 record che descrivono i passeggeri del Titanic e li etichettano come sopravvissuti o meno sulla base delle seguenti caratteristiche:

- PassengerId
- Survived (1/0, solo train)
- Pclass (1, 2, 3)
- Name
- Sex ('male', 'female')
- Age (frazionaria se meno di 1; se stimata è nella forma xx.5)
- SibSp (numero di familiari; Sibling: fratelli/sorelle; Spouse: moglie/marito)
- Parch (numero familiari; Parent: padre/madre; Child: figli/figlie; Parch=0 per piccoli accompagnati solo dalla tata)
- Ticket
- Fare
- Cabin
- Embarked (C = Cherbourg, Q = Queenstown, S = Southampton)

1. Individuare eventuali dati mancanti e farne l'imputazione secondo criteri che rispecchino la diversa stratificazione sociale e distribuzione di genere dei passeggeri ovvero rimuoverle se troppo sparse.

   punti ___/ 4

2. Eseguire la feature selection per individuare le caratteristiche più rilevanti del data set attraverso una *tecnica embedded* che impieghi un classificatore come modello. La scelta del modello embedded è lasciata al candidato, tenendo conto che il problema è di classificazione binaria.

   punti ___/ 8

3. Implementare un classificatore con un metodo di ensemble che lavori sulle feature selezionate e addestrarlo con una K-fold cross-validation, individuando anche una opportuna griglia di iperparametri. Misurare le performance in termini di Accuracy e AUC.

   punti ___/ 8

4. Implementare una rete neurale convoluzionale in Tensorflow che esegua la classificazione binaria sulle feature selezionate, individuando una griglia di iperparametri e implementando l'early stopping. Confrontare i risultati con i modelli precedenti usando sempre Accuracy e AUC.

   punti ___/ 10

TOTALE: punti ___/ 30

Data: _____ Allievo: _____ Matricola: _____

Regole della prova scritta

Di seguito si riportano le regole da seguire e le caratteristiche della prova ai fini della valutazione:

1. La durata complessiva della prova è pari a due ore e prevede una serie di quesiti che approfondiscono diversi aspetti dello stesso problema: ognuno sarà libero di dedicare ad ogni quesito il tempo che vorrà.
2. La prova si svolge ***interamente*** al calcolatore.
3. L'ambiente di sviluppo predisposto è un environmet conda/mamba dotato di editor Spyder e dei seguenti pacchetti:
    a. Pandas
    b. Matplotlib
    c. Seaborn
    d. Numpy
    e. Scipy
    f. Scikit-learn
    g. Pyspark
    h. Keras
    i. Tensorflow
4. Ai fini dell'avvio dell'ambiente, aprire il prompt dei comandi e digitare i due seguenti comandi:
    ```
    $ mamba activate spyder-env
    $ spyder
    ```
5. Sarà consentito consegnare dopo la prima ora di prova.
6. Sarà necessario spegnere e consegnare i dispositivi mobili (smartphone, smartwatch e tablet) alla cattedra prima dello svolgimento della prova.
7. La navigazione internet dalle postazioni sarà bloccata, in generale, e consentita solo verso i siti di documentazione delle librerie ed il repository delle slide in pdf.
8. Il docente distribuirà copia digitale del compito ed eventuali data set direttamente dalla propria postazione ovvero tramite penna USB e allo stesso modo raccoglierà gli elaborati di programmazione.
9. Il candidato consegnerà comunque il presente foglio datato e con l'indicazione del nome e del numero di matricola.
10. Ai fini del calcolo del voto finale della prova, il *valore massimo* di ciascun quesito è riportato in calce allo stesso. La prova riceverà una valutazione pari alla somma dei voti riportati in ciascun quesito. Si precisa che il docente attribuirà ad ogni quesito una votazione *non binaria*, cioè non tutto il valore oppure 0, ma valuterà la correttezza formale dell'elaborato, il rigore metodologico dell'approccio teorico e l'originalità delle soluzioni proposte per attribuire una votazione nel range definito dal valor massimo del quesito.

Data: _____ Allievo: _____ Matricola: _____