Università degli Studi di Palermo

Dipartimento di Ingegneria
Direttore: prof. Antonino Valenza

dipartimento di ingegneria unipa

CORSO DI BIG DATA – MODULO ANALISI PER I BIG DATA

**Prova scritta del 17 gennaio 2024**

Si consideri il data set `housing.csv`, allegato al presente compito, costituito da 20640 record che descrivono il valore medio di uno stabile in California sulla base delle seguenti caratteristiche:

- longitude
- latitude
- housing_median_age
- total_rooms (numero totale di camere nello stabile)
- total_bedrooms (numero totale di camera da letto nello stabile)
- population (numero di abitanti dello stabile)
- households (numero di famiglie nello stabile)
- median_income
- median_house_value
- ocean_proximity

1. Individuare eventuali dati mancanti, farne l'imputazione e procedere al trattamento delle feature categoriche.

   punti ___/ 4

2. Eseguire la feature selection per individuare le caratteristiche più rilevanti del data set attraverso una *tecnica embedded* che impieghi un regressore come modello. Infatti, sappiamo che la regressione lineare attribuisce coefficienti piccoli o nulli alle feature che influenzano poco la predizione.

   punti ___/ 8

3. Implementare una regressione Lasso ed una Ridge per il data set così processato e confrontarne le prestazioni in termini di RMSE e di $R^2$.

   punti ___/ 8

4. Implementare una piccola rete neurale convoluzionale in Tensorflow che esegua la regressione e confrontare i risultati con i modelli precedenti usando sempre RMSE e $R^2$.

   punti ___/ 10

TOTALE: punti ___/ 30

---

**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

Data: _____ Allievo: _____ Matricola: _____

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: _____