**Progetto Big Data**

a.a. 2019-2020

Sessione Invernale

Il progetto consiste nell’analisi del data set [US\_Accidents\_May19.csv](https://www.dropbox.com/s/jr9h0mg2fxcm10n/us-accidents.zip?dl=0), che riporta 2.25 milioni di record di incidenti stradali, registrati su base di singola contea, raccolti continuativamente da febbraio 2016, utilizzando diverse sorgenti di dati incluse due API per la raccolta di flussi di eventi legati al traffico stradale. Di seguito si riporta la [pagina del fornitore del data set](https://smoosavi.org/datasets/us_accidents) con la descrizione dettagliata dei campi.

Si richiede lo svolgimento di due tipi di analisi, utilizzando uno qualunque dei framework studiati, tra MongoDB, Cassandra, Apache Hadoop, con Hive e Pig, e Spark.

1. Analisi correlativa del numero di incidenti per stato e per orario della giornata, oltre alla determinazione del numero degli incidenti medio per ora del giorno, calcolato su tutti gli stati americani. Si richiede la visualizzazione dei dati nella forma voluta dal gruppo di lavoro.
2. Analisi predittiva della severità dell’evento sulla base del campo Severity che classifica gli incidenti in quattro classi con le etichette numeriche da 1 a 4. A questo fine si suddivida il data set in due parti: il training set e il test set. Il test set sarà estratto dal data set complessivo selezionando esattamente un record ogni cinque, mentre il training set sarà composto dai record rimanenti, con una percentuale complessiva 80%-20%. Si utilizzi un qualunque modello, a scelta del gruppo di lavoro, per addestrare un classificatore *C*(*x*) 🡪 {1, 2, 3, 4} sul training set e testare le performance sul test set.

Le misure di performance della classificazione saranno: Accuracy, balanced Accuracy, *F1*-micro e *F1*-macro, definite secondo le equazioni che seguono. Si assuma che il classificatore mappi ognuno degli *N* campioni di input *x* in una classe *Ck*, k = 1, 2, …, *C*.



ACC e bACC differiscono perché nel primo caso si assume che il data set sia bilanciato, cioè tutte le classi hanno un numero praticamente uguale di campioni, mentre nell’altro caso si tiene conto del fatto che alcune classi potrebbero avere un numero molto piccolo di elementi.

Per comprendere il significato della misura F1 è necessario introdurre i concetti di precision *P* e recall *R* che partono dal risultato di una classificazione binaria, riportato tramite la “matrice di confusione”.

Per una classificazione binaria *Ci*-vs-all, si può creare la seguente tabella dei risultati, detta matrice di confusione.

|  |  |  |
| --- | --- | --- |
| **Predizione →**  **Reference ↓** | *Ci* | Altre classi |
| *Ci* | *TPi* | *FNi* |
| Altre classi | *FPi* | *TNi* |

TP 🡪 true positives, FP 🡪 false positives, TN 🡪 true negatives, FN 🡪 false negatives.

Per un classificatore ideale la matrice di confusione è diagonale.

Per ogni classe, la precisione *Pi* indica la capacità di predire correttamente il numero più elevato possibile di campioni della classe *Ci* a prescindere dal numero di falsi positivi predetti, mentre la recall *Ri* indica la capacità di discriminare correttamente *Ci* dalle altre classi. La misura *F1* è una misura cumulativa di performance che indica un risultato migliore quanto più è alta.



I valori *F1*-micro e *F1*-macro estendono il concetto di *F1* a più classi, effettuando una media di performance secondo due schemi: a singolo campione (micro) a singola classe (macro).





Ogni singolo progetto sarà consegnato in forma di Notebook e riporterà sia il codice utilizzato, sia la descrizione testuale di tutto il procedimento adottato.

In particolare dovranno essere motivate tutte le scelte relative a:

* Preprocessing dei dati (cleaning, inputazione dei dati mancanti, filtraggio etc)
* Scelta dell’architettura software per Big Data adottata
* Descrizione del processo di analisi correlativa
* Scelta della metodologia di visualizzazione dei dati
* Scelta delle feature rilevanti per la classificazione
* Scelta del modello di addestramento
* Discussione del bilanciamento o meno delle classi
* Discussione dei risultati ottenuti