**Report Reale Mutua**

1. **Introduzione**

Questo breve report ha lo scopo di spiegare come configurare il processo di validazione e quali sono gli algoritmi utilizzati al suo interno.

1. **Configurazione**

Sarà necessario settare i seguenti parametri nel file MainApp.py per eseguire lo script:

* dataset\_path: path del dataset per cui si vuol effettuare la cross\_validation
* n\_fold: numero di fold utilizzati dalla cross validation
* risk\_levels: lista dei livelli di rischio per cui si vuol effettuare la cross validation

1. **Scripting**

Il codice è organizzato in 4 diversi file presenti nella directory Run:

* MainApp.py: rappresenta il core del codice. Contiene tutte le chiamate a metodi per permettere il corretto processo di validazione;
* SKLearn.py: contiene i metodi utili per effettuare la cross validation;
* ReadDataset: contiene i metodi per la lettura del dataset di Reale Mutua;
* txt2xls: permette la creazione di un file .xls riassuntivo dell’esito della cross validation.
  1. **Algoritmi utilizzati ed organizzazione della logica**

L’idea di base è combinare l’esito di diversi classificatori per effettuare la classificazione.

* + 1. **Lettura del dataset**

Il dataset viene letto grazie al metodo ReadDataset.read\_csv().

Esso restituisce l’header, il dataset e il target scelto per la classificazione.

E’ importante notare che dal dataset vengono rimossi tutti i 4 livelli di rischio, verrà restituito solamente quello scelto come target (la scelta avviene tramite passaggio di parametro al metodo).

* + 1. **Label Encoding**

Gli algoritmi di classificazione forniti da scikit-learn consentono di lavorare solamente con valori numerici.

Per tal motivo è necessario un associazione numero-stringa, la quale viene effettuata dal metodo SKLearn.LabelEncoder().

Essa restituisce un dizionario contenete tali associazioni, cosi da poter facilmente passare dal valore numerico alla stringa corrispondente.

Ho scelto l’utilizzo di un Label Encoding creato ad hoc per questo dataset. Tale scelta è dettata dal label encoder fornito da scikit-learn, il quale reordina tutte le stringhe secondo un ordine alfabetico.

Algoritmi di classificazione come Decision Tree possono soffrire di tale ordinamento, in quanto si basano su criteri di split come x<=y e x>y.

Ad esempio immaginiamo un encoding [Ancona, Informatica, Palermo] = [1, 2, 3].

La decisione 2<=3 non avrebbe senso logico, difatti il confronto si vorrebbe tra due città e non tra città e professione.

* + 1. **Decision Tree**

All’interno della directory dtree\_*yyyy* (dove *yyyy* rappresenta l’anno del dataset in esame), sono presenti dei .pdf in cui son presenti gli alberi di decisione per ogni livello di rischio.

Essi servono per capire come lavora l’algoritmo e quali attributi esso considera più discriminanti rispetto agli altri.

* + 1. **Cross Validation**

Il metodo SKLearn.cross\_validation() effettua la Stratified Cross Validation sul dataset, al fine di valutare il funzionamento del processo di classifiazione.

Diversi step sono stati fatti che hanno portato al corrente work flow del metodo.

1. Il primo passo è stato effettuare una semplice classificazione sul dataset, applicando diversi classificatori e configurazioni.

Tale semplicità non ha portato però a buoni risultati, specialmente per il livello di rischio ‘cost\_caused\_claim’, che è associato a 4 classi differenti.

Guardando al significato delle 4 classi (range1, range2, range3 e range4), si è pensato di trasformare il problema in binario, ovvero raggruppare range1, range2 e range3 e separare range4.

Questa scelta è dettata dall’interesse di predire con elevata precisione gli incidenti che causano un grande danno economico (range4) piuttosto che gli incidenti meno onerosi.

1. Dato il caso di studio, ovvero il lavorare con delle polizze assicurative, è normale avere un dataset non bilanciato, cioè avere la maggior parte delle polizze che non hanno avuto incidente.

Una soluzione a questa problematica è l’utilizzo dell’oversampling e dell’operatore di Bagging.

In particolare l’oversampling è utilizzato in due versioni (immaginiamo di dover classificare una entry con la classe A o B, dove B è la classe meno popolosa):

* Se il parametro *percentage* del metodo cross\_validation() non è settato, di default l’oversampling lavora per avere un egual numero di elementi appartenenti alla classe A e B nel training set.
* Altrimenti, è possibile settare una percentuale per l’oversampling. Ciò significa che è possibile avere un numero di elementi della classe B pari al X% rispetto alla dimensione del dataset.

Le entry con cui viene fatto oversampling vengono presi in modo randomico dal dataset di partenza.

1. Dopo il processo di cui sopra è stato notato, analizzando le matrici di confusione relative ai vari tentativi di classificazione, che il richiamo associato alla classe meno popolosa è sempre piuttosto alto (circa 75%).

Per raffinare il processo di classificazione sono stati messi insieme più classificatori, ovvero:

* Vengono ordinati diversi classificatori in base al valore del richiamo (dal più alto al più basso)
* Eseguo la classificazione
* Interseco i risultati dei diversi algoritmi, per trovare quanti classificatori hanno predetto un entry come classe B
* Se trovo un numero di classificatori X > n\_right (dove n\_right è un parametro a scelta) che hanno predetto tale entry come classe B, allora assegnerò a tale entry classe B, altrimenti classe A.

Utilizzando tutte le strategie sopra citate, sono stati notati netti miglioramenti nella predizione, fino a raggiungere per i quattro livelli di rischio un f1-score tra il 60-70%.

La configurazione che ha portato ad i migliori risultati è stata la combinazione di 7 diversi algoritmi di classificazione, con il parametro n\_right=4 ed un oversampling del 25%.

1. **Risultati**

I risultati sono organizati all’interno della cartella statistics\_ *yyyy* (dove *yyyy* rappresenta l’anno del dataset in esame).

Vengono generati due tipi di file:

* Statistics\_risk\_level\_yyyy.txt: file .txt in cui sono salvate la matrice di confusione, richiamo, precisione e f1-score relativo ad ogni run dello script.

Vengono indicati anche i parametri (come percentuale di oversampling, numero di classificatori utilizzati...) con cui è stato fatto girare lo script;

* Statistics\_yyyy.xls: file .xls che racchiude tutte le statistiche relative a quell’anno cosi da essere piu leggibili. Ogni pagina del file exel è relativa ad un livello di rischio.

1. **Testing**

La directory Testing contiene una versione degli script sopra citati che permette il testing delle varie configurazioni in maniera del tutto automatica.

E’ possibile visualizzare l’esito delle varie configurazioni tramite il file statistics\_yyy.xls presente nella cartella statistics\_yyyy al fine di valutare la miglior configurazione possibile da poter inserire poi all’interno degli script presenti nella directory Run.