Data Mining e Scoperta di Conoscenza

Progetto 5

Realizzazione di un task per la "Valutazione del rischio di mancato pagamento"

Si realizzi un modello di classificazione il cui scopo è, a partire da un set di 30.000 ordini di acquisto effettuati on-line, predire se la probabilità di mancato pagamento è alta ("high risk") oppure bassa ("low risk"). In dettaglio:

- Si scelga un metodo di classificazione particolarmente appropriato per il problema assegnato, giustificando la scelta. Dopo aver addestrato il modello predittivo sulle 30.000 istanze di training *dmc2005_train.txt e si valutino* le performances classificando 20.000 istanze di test contenute in *dmc2005_class.txt*, assegnando ad ogni istanza la corretta classe di rischio.
- Si confronti il modello di classificazione scelto con altri modelli di classificazione standard e si valutino gli indici di bontà del modello scelto sia rispetto all'accuratezza predittiva (influenzata dal modello di costo proposto) sia rispetto all'efficienza della costruzione del modello e della predizione.

Si descriva nel dettaglio la metodologia adottata (pre-processing, analisi dei risultati) giustificando le scelte fatte.

I datasets e la descrizione dettagliata degli attributi ivi contenuti, vengono forniti contestualmente al progetto.

NOTE PER L'ESECUZIONE DEL PROGETTO

1. Scrivi un rapporto di circa 10 pagine in cui
   a. Descrivi analiticamente l'algoritmo che hai implementato.
   b. Commenti le parti essenziali del codice Java che hai scritto, e metti in un'appendice l'intero codice
   c. commenti e illustri graficamente e quantitativamente gli esperimenti effettuati.
2. Prepare delle slides Powerpoint (non più di 10 slides) in cui riassumi gli esiti del progetto