**Documentazione**

Progetto realizzato da:

* Domenico Dell’Olio (mat. )  
  mail: |
* Giovanni Pio Delvecchio (mat. )  
  mail: |
* Giuseppe Lamantea (mat. )  
  mail: |

Il software presentato permette, a partire da un dataset di partenza memorizzato tramite il servizio MySQL, la predizione di un attributo numerico target per una certa porzione di esempi descritta dall’utente. Questa tecnica di *data mining* viene implementata tramite l’utilizzo di un **albero di regressione** costruito tramite le informazioni ricavate dalla base di dati a disposizione.

Un albero di regressione è una struttura ad albero utilizzata per la predizione di valori numerici, essa è formata da:

* dei **nodi interni**, detti anche *nodi di split*, che rappresentano gli attributi del dataset
* degli **archi** fra nodi, che rappresentano il valore assunto dall’attributo associato al nodo da cui parte l’arco nella porzione di popolazione rappresentata dal nodo di arrivo
* delle **foglie**, ovvero nodi senza figli, da cui si può generare una predizione del valore dell’attributo target per la porzione del dataset avente le caratteristiche descritte dalla successione di decisioni prese dal nodo radice al nodo foglia in questione

L’algoritmo che si occupa della costruzione dell’albero di regressione è così definito:

1. Determinare il nodo (attributo) in cui effettuare uno split genera porzioni del dataset con varianza dell’attributo target minima
2. Se un nodo generato dallo split rappresenta una porzione abbastanza piccola del dataset, esso sarà un nodo foglia
3. Ripetere il punto 1 sulle porzioni del dataset definite dai figli dello split effettuato, finché non sono state generate tutte le foglie dell’albero

**Progetto base**

La versione base del progetto consiste in un’architettura client/server che permette all’utente di esplorare un albero di regressione creato dal server tramite varie opzioni.  
Il server dovrà essere eseguito su di una macchina con un database MySQL in esecuzione. Il servizio sarà raggiungibile sulla porta 8080, e potrà comunicare con diversi client contemporaneamente.  
I servizi offerti dal server (acceduti dall’utente tramite il client CLI) sono i seguenti:

* caricare un dataset fornendo il nome della tabella in cui è memorizzato nel database
* creare ed esplorare un albero di regressione a partire dal dataset specificato
* salvare l’albero di regressione appena creato nella macchina dove il server viene eseguito
* esplorare un albero di regressione creato in precedenza

Il server effettuerà dei log su console di quanto accade durante la sua esecuzione (ad esempio, connessione di un client, o errore durante la creazione di un albero di regressione). I messaggi di errore rilevanti per l’utente verranno inviati al client.

Il client da riga di comando permette di collegarsi ad un server che sta eseguendo un’istanza del server, il quale indirizzo IP e porta devono essere specificati all’avvio del client (come argomenti di esecuzione del programma). L’utente dovrà fornire il nome del dataset su cui lavorare, e successivamente specificare se creare un nuovo albero di regressione o caricarne uno creato precedentemente. Una volta effettuata la scelta, si avvierà l’esplorazione dell’albero, che viene presentata tramite diverse query testuali. L’utente quindi dovrà specificare il cammino da seguire tramite gli identificativi numerici dei nodi di split. Alla fine dell’esplorazione, gli sarà presentata la predizione del valore dell’attributo target. Sarà quindi possibile ricominciare l’esplorazione sullo stesso albero, o chiudere il programma.  
Il salvataggio dell’albero di regressione verrà effettuato in automatico alla creazione di un nuovo albero. L’utente sarà informato di eventuali errori incontrati dal server durante l’esecuzione delle operazioni richieste.

**Progetto esteso**

L’estensione sviluppata per il progetto consiste in un’interfaccia grafica per desktop sviluppata tramite il framework JavaFX. Le funzionalità del client CLI sono state mantenute, ed è stata aggiunta la possibilità di memorizzare più server a cui connettersi: l’utente potrà quindi aggiungere nuovi server, che saranno mantenuti in memoria anche dopo la chiusura del programma. In generale, il client grafico è più tollerante agli errori da parte dell’utente, e permette l’esplorazione di più alberi su più server in una stessa sessione di utilizzo del programma, a differenza del client da riga di comando.

Il server appartenente al progetto esteso potrà essere sia avviato sulla porta 8080 (porta di default se viene avviato il server senza argomenti), che su una porta specificata dall’utente. Inoltre, i log effettuati dal server vengono salvati in un file di testo consultabile nella macchina che sta eseguendo il server.