1 Descrizione dati iniziali

Il dataset che abbiamo analizzato contiene dati sulle carriere accademiche degli studenti del corso di laurea di informatica dell'universit degli studi di Firenze e il loro voto conseguito al test di ingresso. In particolare, le informazioni presenti sono:

- Coorte: Anno di immatricolazione
- Crediti totali: Numero crediti complessivi dello studente
- Crediti con voto: Numero di crediti assegnati allo studente per esami con votazione in trentesimi(tutti tranne inglese)
- Voto medio: Media pesata dei voti degli esami sostenuti

In seguito ci sono sei coppie di attributi che indicano:

- Nome dell'esame
- Data in cui lo studente ha sostenuto l'esame

Gli esami sono Algoritmi e strutture dati (ASD), Architetture degli elaboratori(ARC), Programmazione(PRG), Analisi I(ANI), Matematica discreta e logica(MDL) e Inglese.

- Punteggio conseguito al test di ingresso.

Per quanto riguarda gli attributi relativi ai voti degli esami i valori che questi assumono sono: il voto conseguito dallo studente nel caso in cui abbia sostenuto l'esame; oppure zero se lo studente non ha sostenuto l'esame durante la sessione estiva del proprio anno di immatricolazione.

2 Gestione dei dati

Abbiamo deciso di effetture le seguenti operazioni sul dataset:

- eliminazione degli studenti che hanno sostenuto solo inglese
- riportare tutti gli attributi relativi alle date degli esami nel formato YYYY-MM-DD

Il motivo per cui abbiamo deciso di eseguire l operazione del primo punto che gli studenti in questione non avendo voti in trentesimi non ci sono pari significativi per analisi. Per quanto riguarda il secondo punto invece, nel caso in cui uno studente non avesse sostenuto un particolare esame, erano presenti valori pari a zero in alcuni casi e valori pari a "0000-00-00" in altri. Abbiamo quindi deciso di trattare questa incosisistenza dei dati ponendo i valori pari a "0000-00-00".

Per effetturare queste due operazioni abbiamo importato il dataset in database tramite l'operazione di import riportata nel codice.

CODICE IMPORT

Per quanto riguarda la gestione dell incosisistenza relativa alle date degli esami abbiamo risolto il problema specificando le query ripotate nel codice

```
update dmo.studenti set data_ARC = '0000-00-00' where data_ARC = '0';
update dmo.studenti set data_ASD = '0000-00-00' where data_ASD = '0';
update dmo.studenti set data_PRG = '0000-00-00' where data_PRG = '0';
update dmo.studenti set data_ANI = '0000-00-00' where data_ANI = '0';
update dmo.studenti set data_MDL = '0000-00-00' where data_MDL = '0';
update dmo.studenti set data_INGLESE = '0000-00-00' where data_INGLES
```

3 Analisi dei dati

Per analizzare i dati a disposizione abbiamo utilizzato il linguaggio R per determinare la correlazione tra i diversi attributi. In figura riportata la tabella con i valori di correlazione.

Come possibile notare dalla tabella, ci sono alcune ovvie correlazioni che coinvolgono gli attributi crediti totali e crediti con voto. Per quanto riguarda l'attributo corte si pu notare come questo abbia una scarsa correlazione con gli altri attributi. Le correlazioni pi evidenti e significative sono quelle che coinvolgono l' attributo crediti totali (e quindi anche crediti con voto) con il voto che gli studenti hanno ottenuto per i diversi esami sostenuti. Infatti il valore di queste correlazioni compreso tra 0.53 per quanto riguarda l'attributo di crediti totali e quello di Algoritmi e strutture dati e 0.73 per quanto riguarda crediti totali e il voto di Architetture degli elaboratori. Tale risultato pu essere interpretato in maniera abbastanza intuitiva, infatti mostra che se uno studente ha sostenuto pi esami(e quindi ha pi crediti) ha in generale ottenuto dei voti migliori.

L'attributo test correlato maggiormente con l'attributo voto medio. Si pu notare inoltre che il valore di correlazione significativamente pi alto di quello che l'attributo test ha con i voti dei singoli esami, con l'unica eccezione di MDL. Questi valori lasciano suggerire che il punteggio conseguito al test d'ingresso di ogni studente sia un buon parametro per valutarne l'andamento generale piuttosto che i voti di un singolo esame. Inoltre l'attributo test mostra una discreta correlazione con i crediti totali(e quindi anche con i crediti con voto).

Per quanto riguarda le correlazioni tra i voti dei diversi esami possibile notare che il valore pi alto quello tra Architetture degli elaboratori e Programmazione che circa pari a 0.43. Mentre invece si ha una correlazione quasi nulla, ossia circa 0.10, tra il voto di Algoritmi e strutture dati e il voto di Analisi 1.

In figura X viene riportata la matrice di correlazione.

In figura Y viene riportato lo scatter plot dei attributi del dataset.

Per applicare gli algoritmi di clustering stato utilizzato il software Weka. Tuttavia prima di applicare tali algoritmi stato necessiaria un ulteriore fase di preprocessing nella quale sono stati normalizzati tutti gli attributi del dataset in una scala di valori compresi tra zero e uno in modo da evitare problemi dovuto alle diverse scale di valori dei attributi.