پروژه درس مبانی داده کاوی (د. شاکری) نیمسال دوم ۹۶۲

پروژه پیشنهادی شامل چندین فاز بوده که باید طبق ترتیب زمانی تعیین شده انجام شوند. با وجود اینکه فازهای اولیه برای تمام تیم ها مشترک است اما باید به طور جداگانه در هر تیم انجام گیرد.

فاز (۲) آماده سازی داده و کاوش آن (مهلت تحویل ۹۷/۳/۲)

الف) فایل genes-leukemia.csv داده شده همراه با این دستورالعمل را گرفته و تبدیل به فایل Weka با نام genes-leukemia.arff کنید. برای این کار می توانید یا از یک ویرایشگر متن مانند emacs استفاده کنید و یا یک دستور Weka پیدا کنید که فایل cvs. را به arff. تبدیل کند.

ب) مشخصه هدف، CLASS نام دارد. طبقه بند J48 را روی این مجموعه داده با حالت تست "use training set" اعمال کنید.

ج) با استفاده از فایل genes-leukemia.arff دو زیر مجموعه تولید کنید:

- زیرمجموعه genes-leukemia-train.arff حاوی ۳۸ نمونه (s1...s38) داده
- زیرمجموعه genes-leukemia-test.arff حاوی ۳۴ نمونه باقیمانده (s39...s72)

د) طبقه بند J48 را روی genes-leukemia-train.arff آموزش داده و "use training set" را به عنوان حالت تست استفاده کنید. به چه درخت تصمیمی می رسید؟ دقت آن چقدر است؟

ه) حال genes-leukemia-test.arff را به عنوان مجموعه تست مشخص کنید. در این حالت به چه درخت تصمیمی می رسید؟ دقت آن نسبت به مرحله (د) چقدر است؟

م) اکنون فیلد "Source" را از طبقه بند حذف کرده و مراحل (د) و (ه) را تکرار کنید. چه مشاهده می کنید؟ آیا دقت روی مجموعه تست افزایش می یابد؟ اگر چنین است چرایی آن را تحلیل کنید.

و) کدام طبقه بند بالاترین دقت را روی مجموعه تست تولید می کند؟ پاسخ خود را از نظر تئوری عملکرد طبقه بند تحلیل کنید.