تمرین سوم – هرس درخت و مقادیر مفقودشده

مسئله: در جدول زیر، 20 داده آموزشی به همراه خروجی مورد نظر هر یک از آنها داده شده است. هر یک از سطرهای این جدول مشخص کننده یکی از نمونه‌های آموزشی با 4 مشخصه f_1، f_2، f_3 و f_4 به همراه خروجی y می‌باشد. هر یک از چهار مشخصه می‌توانند یکی از مقادیر عددی 0، 1 یا 2 را اختیار کنند. خروجی نیز با توجه به مقدار مشخصه‌ها می‌تواند یکی از مقادیر عددی 0، 1 یا 2 را داشته باشد. مقادیری که با ? مشخص شده‌اند، مقادیر مفقودشده (Missing Values) هستند.

f_1	f_2	f_3	f_4	Y
2	2	1	0	**2**
2	0	1	1	**0**
0	0	2	0	**0**
1	2	0	1	**2**
0	2	?	0	**0**
1	2	1	0	**1**
1	1	1	2	**0**
1	1	0	1	**1**
0	1	0	1	**2**
0	?	1	1	**0**
0	0	0	1	**0**
2	2	2	0	**2**
1	2	1	2	**1**
2	2	2	2	**0**
0	1	2	2	**2**
2	1	1	2	**1**
2	2	0	1	**2**
1	1	2	1	**1**
1	0	0	1	**1**
0	2	2	1	**1**

فرض کنید قصد داریم یک درخت تصمیم‌گیری برای این داده‌ها تولید نماییم. درخت تصمیم‌گیری مورد نظر را بدست آورده و ترسیم نمایید. مراحل بدست آوردن درخت به همراه محاسبات مرتبط با آن را به طور دقیق تشریح نمایید. معیار انتخاب مشخصه را هر یک از چهار حالت زیر در نظر بگیرید و برای هر حالت درخت را به طور جداگانه ترسیم نمایید.

الف) Information Gain

ب) Gini Index

ج) Gain Ratio

د) Twoing Criteria

در هر قسمت نیز برای هرس از الگوریتم Error-Based Pruning با میزان معناداری مطلوب $\alpha=0.2$ استفاده نمایید. در تولید درخت اولیه از یک هرس پیش‌دستانه نیز استفاده کنید. برای این منظور، زمانی که در یک گره سه عنصر یا کمتر حضور داشتند، یک برگ ایجاد نمایید. فرض کنید که مقادیر ویژگی‌های مفقودشده در محاسبهٔ معیار انشعاب تأثیری نداشته و صرفاً در انتها به صورت خطی معیار محاسبه‌شده را کاهش می‌دهند.