۱. این پایگاه داده شامل پنج تراکنش را در نظر بگیرید. اطلاعات سود اقلام در دست است. فرض کنید min-sup = 60%. می خواهیم همه itemsetهای مکرری را پیدا کنیم که میانه سود آنها بزرگتر مساوی ۵ است. با استفاده از الگوریتم FP-Growth، itemsetهای مکرر با این شرایط را پیدا کنید.

TID	Items
100	A, B, C, E, F
200	A, C, D, G
300	B, C, F
400	A, B, C, D, F
500	B, E, F, G

Item	Profit
A	7
B	3
C	1
D	6
E	-2
F	10
G	2

۲. می خواهیم برای دانشجویان یک درس پیش بینی کنیم که در درس قبول می شوند یا خیر (?Pass). این پیش بینی را بر اساس نمره میان ترم و نمره تمرین ها انجام می دهیم.

Midterm	**HW**	**Pass?**
C	B	N
C	A	Y
B	B	N
B	A	Y
A	B	Y
A	A	Y

داده های زیر را به عنوان داده های آموزشی در نظر بگیرید.

الف) entropy(pass) را محاسبه کنید.

ب) اگر از معیار Information Gain به عنوان تابع سنجش استفاده کنیم، کدام صفت برای آزمون در ریشه درخت انتخاب می شود؟

پ) کل درخت تصمیم متناظر با این مجموعه آموزشی را رسم کنید.

۳. ۱۰ نمونه تست در اختیار ما قرار گرفته است که می خواهیم آنها را در دو کلاس + و – رده بندی کنیم. برای این منظور رده بند M را ایجاد کرده ایم که نتیجه آن، احتمال تعلق هر نمونه به کلاس + است. نتیجه اعمال این رده بند در جدول روبرو آمده است. همچنین به ازای هر نمونه، برچسب صحیح آن در ستون gold label نشان داده شده است.

Data Item	Gold Label	P(+|M)
1	+	0.9
2	+	0.79
3	+	0.66
4	+	0.62
5	+	0.44
6	-	0.33
7	-	0.84
8	-	0.58
9	-	0.4
10	-	0.37

الف) نمودار ROC را برای رده بند داده شده رسم کنید.

ب) فرض کنید برای تعیین برچسب هر نمونه از این قانون استفاده می کنیم که اگر p(+|M) بیشتر از ۰.۵ بود، نمونه را در کلاس + در نظر می گیریم و در غیر این صورت در کلاس -.

معیارهای زیر را برای این مجموعه نتایج محاسبه کنید.

I. Precision
II. Recall
III. Accuracy
IV. Sensitivity
V. Specificity

۴. این مجموعه داده را در نظر بگیرید:

{0, 1, 2, 6, 10, 13, 15, 20, 21, 23, 25}

الف) با استفاده از خوشه بندی سلسله مراتبی پایین به بالا و معیار complete link داده ها را خوشه بندی کنید. دندروگرام خوشه بندی را رسم کنید.

ب) می خواهیم داده ها را به سه خوشه تقسیم کنیم. با توجه به بخش الف این سه خوشه چه خواهند بود؟

ب) می خواهیم با استفاده از روش DB-Scan همین داده ها خوشه بندی کنیم. ϵ و min_points را به گونه ای تعیین کنید که نتیجه خوشه بندی به شکل زیر باشد. نقاط core و نقاط border هر خوشه را مشخص کنید.

$C_1 = \{0, 1,2\}, C_2 = \{10, 13, 15\}, C_3 = \{20, 21, 23, 25\}$

$Outlier: 6$

۵. معیارهای خارجی (external) برای ارزیابی نتایج خوشه بندی را در نظر بگیرید. در این معیارها نتیجه ارزیابی کاربر خبره در دست است (ground truth) و می خواهیم نتیجه خوشه بندی را با خوشه هایی که خبره تشخیص داده است مقایسه کنیم. نتیجه خوشه بندی را C و خوشه های خبره را T در نظر بگیرید و به سوالات زیر پاسخ دهید.

الف) با داشتن نتیجه خوشه بندی C و خوشه های خبره T، معیار ارزیابی کیفیت Q(C, T) باید چهار خاصیت اصلی داشته باشند. این چهار خاصیت کدامند؟

ب) معیار purity برای ارزیابی خوشه بندی را در نظر بگیرید. آیا این معیار این چهار خاصیت را دارد؟ تک تک بررسی کنید.

پ) جدول زیر از نتایج یک خوشه بندی را در نظر بگیرید. معیارهای purity و maximum matching را برای این خوشه بندی محاسبه کنید.

$C\backslash T$	T_1	T_2	T_3	T_4	Sum
C_1	15	20	50	15	100
C_2	50	30	70	100	250
C_3	10	20	150	20	200
C_4	25	50	30	45	150
m_j	100	120	300	180	700