گروه مهندسی کامپیوتر

میانترم – "مبانی داده‌کاوی" محمدرضا پژوهان

نمودار BoxPlot را برای لیست داده‌های زیر رسم کنید.

10 20 30 60 65 70 75 80 90 100 110 120 150 220 245

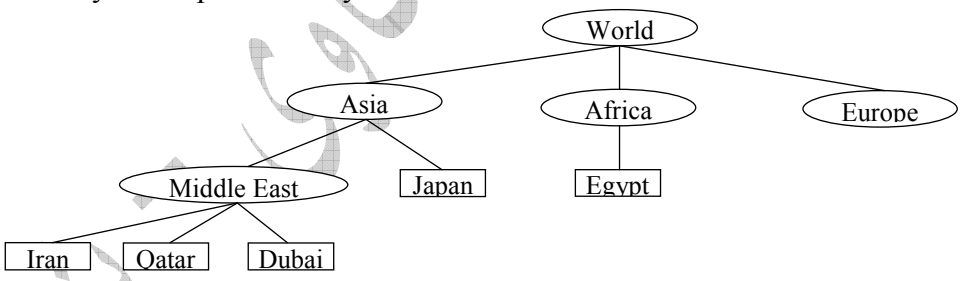
با داشتن فراوانی داده‌های قد و وزن مربوط به ۱۵۰۰ نفر به شرح جدول ذیل، آیا این دو صفت با هم Correlation دارند یا نه؟

وزن / واقعی	لاغر	معمولی	چاق
کوتاه	۱۰۰	۱۵۰	۵۰
متوسط	۱۵۰	۶۰۰	۱۵۰
بلند	۵۰	۲۰۰	۵۰

برای داده‌های مربوط به اطلاعات بیماران که در جدول ذیل همراه با فرضیاتی داده شده، Distance(ID=1,ID=3) و Distance(ID=3,ID=5) را محاسبه کنید. وزن کلیه ستونها را یکسان فرض نموده و از نرمال‌سازی به بازه [0,1] (با استفاده از روش min-max) و تابع فاصله اقلیدسی استفاده نمایید.

فرضیات:

- Age: min=10 , max=90
- Education level: "<Diploma" , "Diploma" , "B.Sc." , "M.Sc." , "PhD" , ">PhD"
- Nationality Concept Hierarchy:



ID	Age	Sex	Nationality	Education
1	37	M	Iran	>PhD
2	42	F	Qatar	Diploma
3	23	M	Japan	B.Sc.
4	55	M	Iran	M.Sc.
5	17	F	Egypt	<Diploma

با توجه به جدول دادگان زیر،

(Day)	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

الف) با فرض اولویت ۱: Wind ، ۲: Humidity ، ۳: Temperature و ۴: Outlook درخت تصمیم با این دادگان آموزشی بسازید (نیازی به محاسبه Information Gain نیست). نودهای درخت را با مستطیل، برگها را با دایره و مقادیر یالها را روی یالها نشان دهید.

ب) می‌خواهیم مدل بیز-ساده برای پیش‌بینی را ایجاد کنیم. کلیه احتمالات مورد نیاز را محاسبه نمایید. در صورت مواجه شدن با مقدار صفر به کمک Laplacian Correction مشکل را حل و مقادیر را اصلاح کنید.

پ) ازمدل بیز ساده استفاده نمایید و پیش‌بینی کنید که اگر یک روز بارانی (Rain) و گرم (Hot) با رطوبت بالا (High) و وزش باد ملایم (Weak) باشد، در این روز امکان برگزاری بازی تنیس هست یا نه؟

پ) با استفاده از روش KNN امکان یا عدم امکان بازی در روز داده شده در قسمت قبل را پیش‌بینی نمایید. (معیار فاصله را فاصله شهری (Manhattan Distance)، K=3 و داده‌های Nominal را بدون هر Concept Hierarchy در نظر بگیرید.

وزن برای Temperature را 2 و برای دیگر صفات را 1 در نظر بگیرید.

Temperature از نوع Ordinal بوده و مقادیر درجه حرات به ترتیب شامل ابتدا Hot، سپس Mild و نهایتاً Cool می‌باشد. بقیه صفات همگی Categorical می‌باشند.

موفق و پیروز باشید – محمدرضا پژوهان