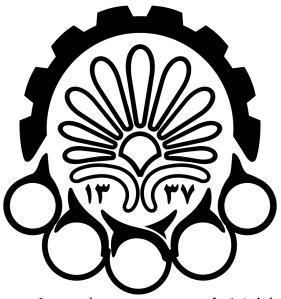

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

تمرین سری اول درس داده‌کاوی

آماده‌سازی و پیش‌پردازش داده

استاد درس:

دکتر شاکری

تاریخ انتشار: شنبه ۱۳ اسفند ماه ۱۴۰۱

مهلت تحویل : پنجشنبه ۲۵ اسفند ماه ۱۴۰۱ – ساعت ۱۰ صبح

تمرین ۱. جدول زیر را در نظر بگیرید.

سن	درصد چربی
23	9.5
23	26.5
27	7.8
27	17.8
39	31.4
41	25.9
47	27.4
49	27.2
50	31.2
52	34.6
54	42.5
54	28.8
56	33.4
57	30.2
58	34.1
58	32.9
60	41.2
61	35.7

قرار است بر اساس سن و درصد چربی بدن نمونه‌های انتخاب شده آزمایشی انجام گردد.

(آ) میانگین، انحراف معیار و انحراف معیار مطلق هر یک از ویژگی‌های سن و درصد چربی را محاسبه کنید.

(ب) داده‌های دو ویژگی را نرمال‌سازی کنید.

(ج) براساس انحراف معیار و انحراف معیار مطلق، به استانداردسازی مقادیر دو ویژگی بپردازید.

تمرین ۲. توضیح دهید که هریک از روش‌های زیر داده‌ها را به چه بازه‌ای انتقال می دهد؟

(آ) نرمال‌سازی

(ب) استانداردسازی بر اساس انحراف معیار

(ج) استانداردسازی بر اساس انحراف معیار مطلق

تمرین ۳. در نرمال‌سازی داده‌ها آموختیم که داده ها به بازه $[0, 1]$ منتقل می‌گردند. با تعمیم این روش تابعی را معرفی کنید که داده به بازه دلخواه $[a, b]$ منتقل گردد.

تمرین ۴. مجموعه داده زیر را درنظر بگیرید.

$$\{5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215\}$$

مجموعه داده فوق را به کمک هر یک از روش‌های زیر در سه دسته مجزا قرار دهید.

(آ) روش گسسته سازی بسامدی [1]

(ب) روش گسسته سازی بخشی [2]

تمرین ۵. چهار روش برای مدیریت مقادیر از دست رفته بیان شد؛ نحوه اعمال هر از این روش‌ها را به طور کامل توضیح دهید و زمان مناسب استفاده از این روش‌ها را نیز بیان کنید.

[1] Equal-frequency partitioning
[2] Equal width partitioning

تمرین ۶. جدول زیر را درنظر بگیرید.

سرطان	سابقه خانوادگی سرطان	سن
+	−	پیر
+	+	میانسال
+	−	میانسال
+	+	پیر
+	+	جوان
−	−	میانسال
−	−	جوان
−	+	جوان
+	−	پیر
−	−	جوان
−	−	میانسال
+	+	جوان
−	−	میانسال

بهره اطلاعاتی را برای هر دو ویژگی سن و سابقه خانوادگی سرطان محاسبه کنید و توضیح دهید بر اساس کدام ویژگی است که می توان تفکیک بهتری روی داده ها انجام داد؟