گزارش تمرین داده‌کاوی

# 1. مقدمه

در این پروژه تمرینی، یک دیتاست فرضی شامل اطلاعات جمعیت‌شناختی افراد ایجاد شد و با استفاده از الگوریتم‌های KNN و KMeans مورد تحلیل قرار گرفت. هدف بررسی امکان طبقه‌بندی و خوشه‌بندی افراد بر اساس ویژگی‌های فردی و سطح درآمد آن‌ها بود.

# 2. ساخت دیتاست و پیش‌پردازش

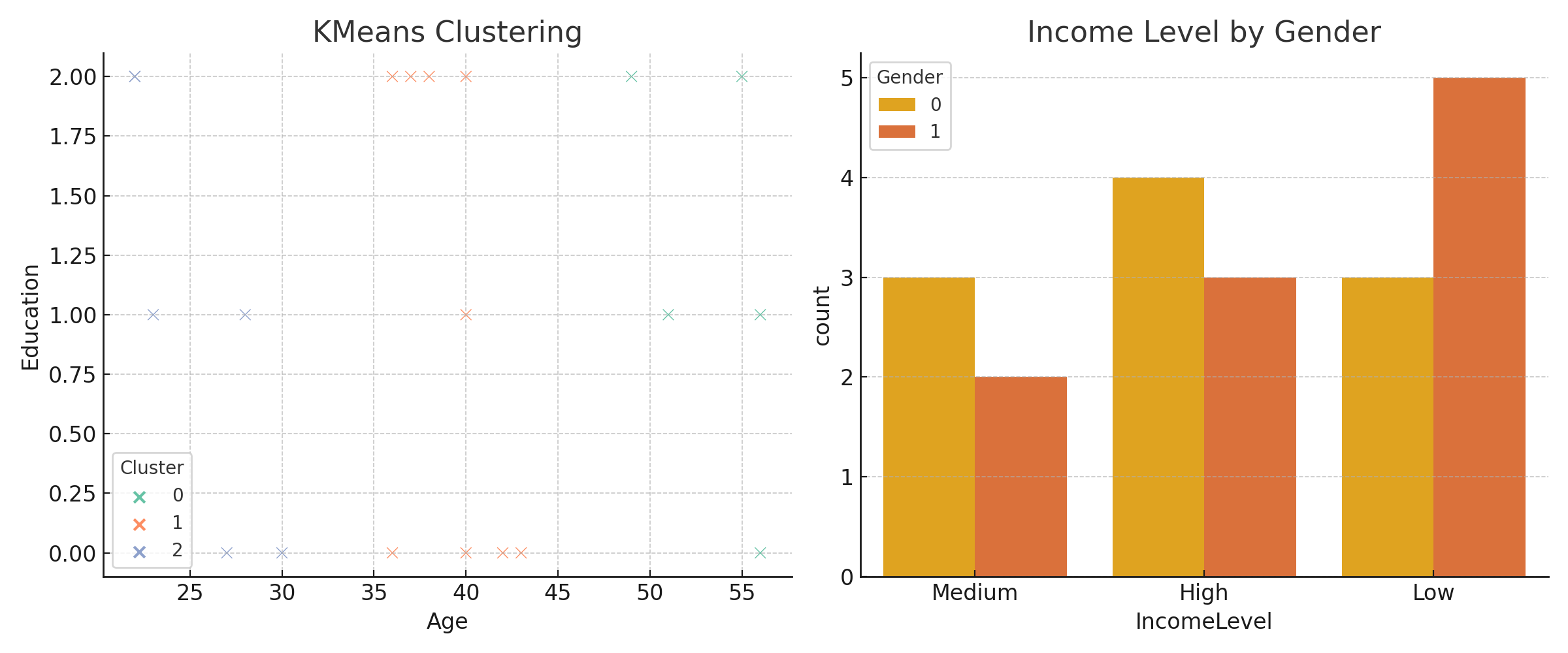
دیتاست شامل ستون‌های ID، سن، جنسیت، سطح تحصیلات، شغل و سطح درآمد بود. با استفاده از LabelEncoder داده‌های متنی به عددی تبدیل شدند.

# 3. مدل‌سازی

برای طبقه‌بندی از الگوریتم KNN با تعداد همسایگان k=3 استفاده شد. دقت مدل برابر با 33.3 درصد به دست آمد. همچنین از الگوریتم KMeans با تعداد خوشه 3 برای خوشه‌بندی استفاده شد.

# 4. نمودارها

نمودار سمت چپ، خروجی الگوریتم KMeans را نمایش می‌دهد. نمودار سمت راست توزیع سطح درآمد بر حسب جنسیت را نشان می‌دهد.



# 5. نتیجه‌گیری

با توجه به محدود بودن داده‌ها، دقت الگوریتم KNN پایین بود. همچنین خوشه‌بندی اولیه الگوریتم KMeans توانست الگوهایی ابتدایی از دسته‌بندی افراد ارائه دهد. برای بهبود نتایج، نیاز به داده‌های واقعی و گسترده‌تر می‌باشد.

# 6. نمونه‌ای از داده‌ها

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Age | Gender | Education | Job | IncomeLevel | IncomeLevelEncoded | Cluster |
| 1 | 30 | 1 | 0 | 1 | Medium | 2 | 2 |
| 2 | 40 | 0 | 2 | 3 | Medium | 2 | 1 |
| 3 | 56 | 0 | 0 | 0 | High | 0 | 0 |
| 4 | 40 | 0 | 1 | 0 | High | 0 | 1 |
| 5 | 36 | 1 | 0 | 0 | Low | 1 | 1 |

# 7. دقت الگوریتم KNN

دقت مدل KNN برابر است با: 33.33٪