توضیحات فاز اول پروژه‌ی داده‌کاوری

محمد کاهانی، علی گلدانی

عنوان پایگاه داده: SAP

۲۳ آذر ۱۳۹۸

۱ شرح کلی دیتاست

SAP یک دیتاست آموزشی می‌باشد که از یک سیستم مدیریت آموزش به نام Kalboard-۳۶۰ گردآوری شده است. این سیستم با هدف تسریع فرایند یادگیری، با استفاده از تکنولوژی‌های روز طراحی شده است و امکان دسترسی همگام‌سازی شده به منابع آموزشی را از طریق هر دستگاه متصل به اینترنت، به کاربران می‌دهد.

این مجموعه داده از طریق یک ابزار ثبت فعالیت کاربران تهیه شده است. این ابزار رابط کابری تجربه (xAPI) نام دارد. این ابزار امکان نظارت بر روی نحوه‌ی پیشرفت یادگیری کاربران را از طریق ثبت فعالیت‌هایی مانند خواندن یک مقاله یا مشاهده‌ی یک ویدیو فراهم می‌کند. این ابزار به مسئولین کمک می‌کند تا بتوانند با تحلیل روند استفاده و پیشرفت کاربران، یک تجربه‌ی آموزشی را توصیف کنند. این مجموعه‌داده شامل ۴۸۰ دانش‌آموز و ۱۶ ویژگی در رابطه باآن‌هاست. این ویژگی‌ها در ۳ دسته‌ی کلی طبقه‌بندی می‌شوند:

۱. اطلاعات جمعیتی مانند جنسیت و ملیت

۲. ویژگی‌های مربوط به سابقه‌ی تحصیلی مانند مقطع تحصیلی، سطح نمره و کلاس دانشجو

۳. ویژگی‌های رفتاری مانند تعداد دفعاتی که دانش‌آموز دستش را در کلاس بالا برده‌است، میزان استفاده از منابع درسی، پاسخگویی به پرسشنامه‌ها توسط والدین و رضایت از مدرسه

دانش‌آموزان بر اساس نمره‌ی کل آن‌ها به سه بازه‌ی عددی تقسیم می‌شوند:

- **سطح پایین**: نمرات در بازه‌ی ۰ تا ۶۹
- **سطح متوسط**: نمرات در بازه‌ی ۷۰ تا ۸۹
- **سطح بالا**: نمرات در بازه‌ی ۹۰ تا ۱۰۰

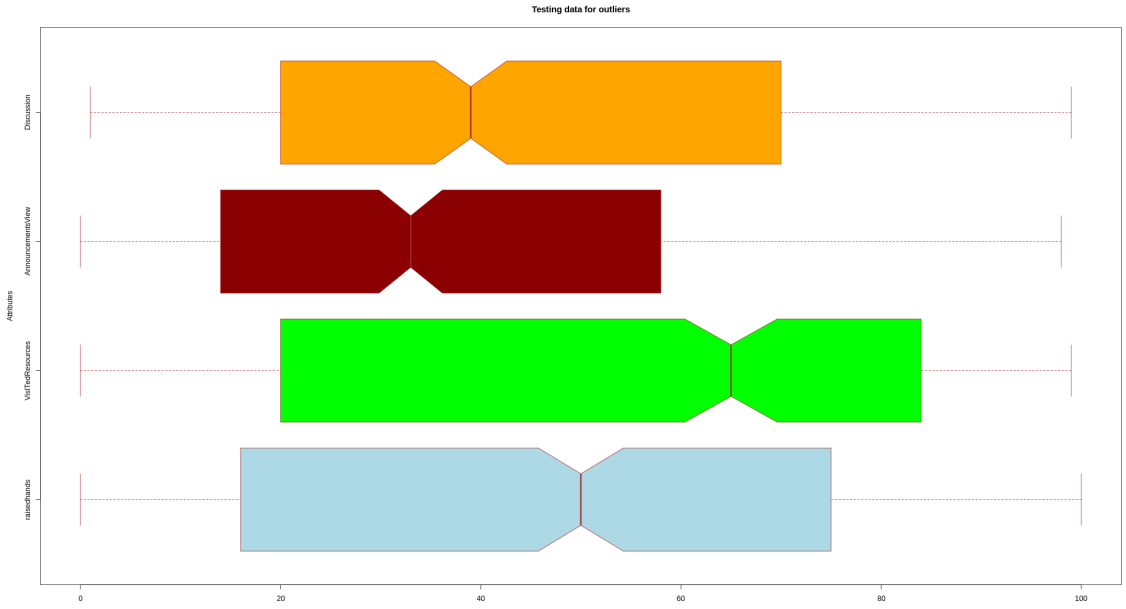
توضیح ویژگی‌های مجموعه‌داده:

ویژگی	توصیف
gender	جنسیت دانش‌آموز
Nationality	ملیت دانش‌آموز
PlaceofBirth	محل تولد دانش‌آموز
StageId	مقطع تحصیلی دانش‌آموز
GradeId	سال تحصیلی دانش‌آموز
SectionId	کلاسی که دانش‌آموز در آن است
Topic	موضوع درس
Semester	نیم‌سال تحصیلی
Relation	پدر یا مادر مسئول دانشجو
raisedhands	میزان دفعاتی که دانش‌آموز دستش را در کلاس بالا برده است
VisitedResources	تعداد دفعاتی که دانش‌آموز منابع درس را مشاهده کرده است
AnnouncementsView	تعداد دفعاتی که دانش‌آموز اعلانات جدید را چک می‌کند
Discussion	تعداد دفعاتی که دانش‌آموز در گروه‌های مباحثه شرکت کرده‌است
ParentAnsweringSurvey	این که پدر یا مادر دانش‌آموز به پرسش‌نامه‌های مدرسه پاسخ داده است یا خیر
ParentschoolSatisfaction	این که پدر یا مادر از مدرسه رضایت داشته‌اند یا خیر
StudentAbsenceDays	این که دانش‌آموز بیش از ۷ جلسه غیبت داشته‌است یا خیر
Class	طبقه‌بندی دانش‌آموز بر اساس نمره‌ی کل

۲ پیش‌پردازش داده‌ها

۱.۲ Data Cleaning

در این دیتاست، مجموعا ۱۸ ویژگی داریم که ۴تای آن‌ها مقادیر عددی هستند و سایر ویژگی‌ها Nominal یا Categorical هستند. با رسم نمودار BoxPlot بر روی این مقادیر عددی (شکل ۱)، متوجه می‌شویم که داده‌ها مقدار نویز ندارند، پس نیازی به انجام عملیات برای حذف نویز در آن‌ها نیست.



شکل ۱: BoxPlot Diagram

۲.۲ Data Integration

از آن جایی که ما تنها یک جدول در اختیار داریم، عملیات مربوط به این بخش کاربرد زیادی ندارد. فقط برای این که داده‌ی تکراری وجود نداشته باشد، عملیات لازم را بر روی دیتاست اجرا می‌کنیم و در نتیجه مشاهده می‌کنیم که داده‌ی تکراری وجود ندارد.

۳.۲ Data Reduction

۴.۲ Data Transformation and Data Discretization