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

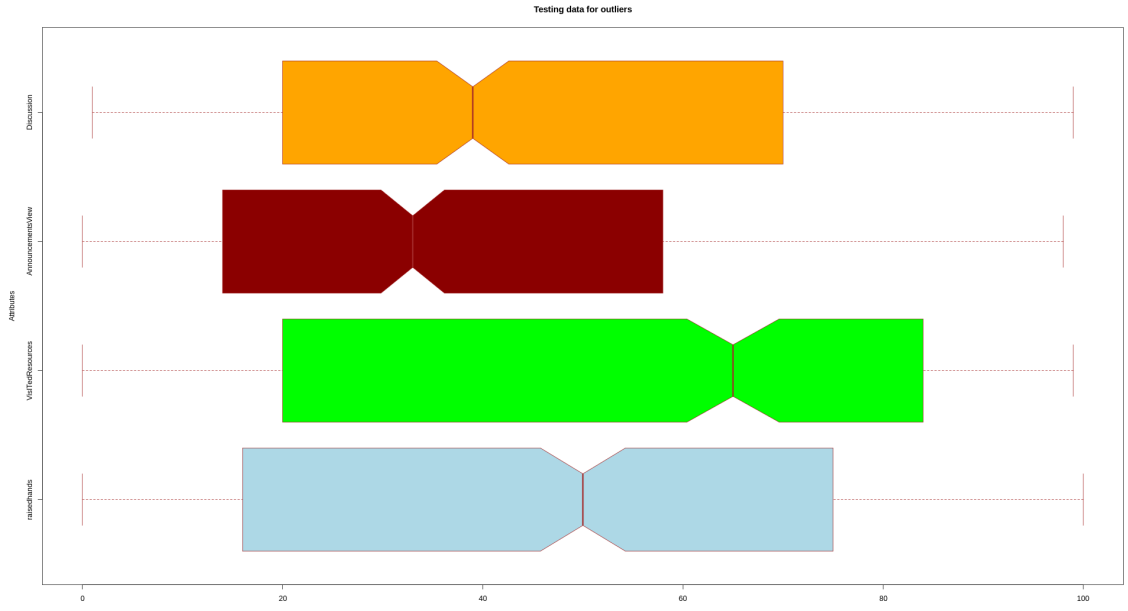
توضیح ویژگی‌های مجموعه‌داده:

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

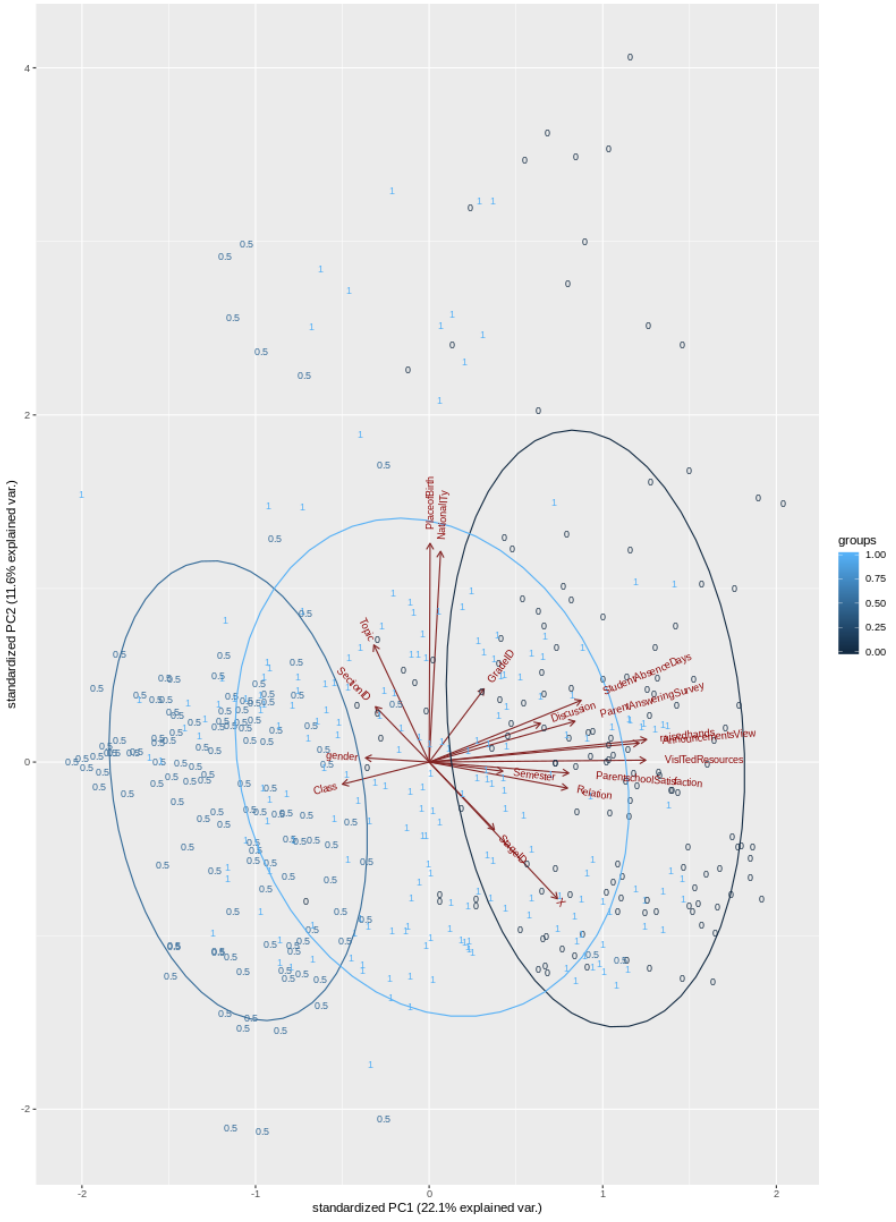
از آن جایی که ما تنها یک جدول در اختیار داریم، عملیات مربوط به این بخش کاربرد زیادی ندارد. فقط برای این که داده‌ی تکراری وجود نداشته باشد، عملیات لازم را بر روی دیتاست اجرا می‌کنیم و در نتیجه مشاهده می‌کنیم که داده‌ی تکراری وجود ندارد.

۳.۲ Data Reduction

در این قسمت ابتدا به بررسی ارتباط بین داده‌ها پرداختیم، با بررسی مقادیر Nationality و PlaceofBirth می‌توانیم نتیجه بگیریم که این دو تا حد زیادی شبیه به یکدیگر هستند. این شباهت را می‌توان با بررسی Cor-relation بین این دو مقدار به دست آورد، پس می‌توانیم یکی از این دو ویژگی را در نظر بگیریم. به همین روش با محاسبه‌ی Correlation برای سایر ویژگی‌ها، به درک خوبی از ارتباط بین صفت‌ها رسیدیم؛ به عنوان مثال با

استفاده از Chi-Square به این نتیجه رسیدیم که صفت StudentAbsenceDays با Class رابطه‌ی معکوس دارد.

در مرحله‌ی بعد، برای کاهش ابعاد داده و یافتن صفت‌های تأثیرگذار در دسته‌بندی کلاس دانش‌آموزان، از روش PCA استفاده کردیم. در نتیجه، ویژگی‌های تأثیرگذار در تعیین کلاس مشخص شدند، که EigenVector های مربوط به این ویژگی‌ها در شکل ۲ مشخص شده‌اند.

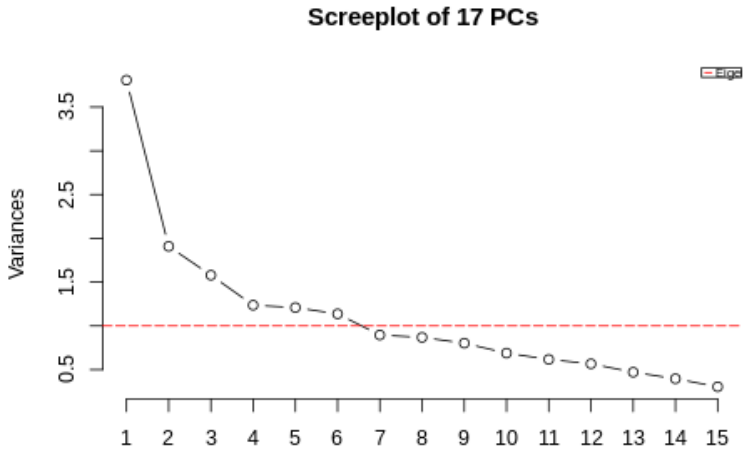


شکل ۲: PCA EigenVectors grouped by Class

همان طور که در شکل ۲ مشخص است، مهم‌ترین صفت‌ها در طبقه‌بندی کلاس دانش‌آموزان به ترتیب raised-hands و VisitedResources و AnnouncementsView و StudentAbsenceDays هستند. (لازم به ذکر

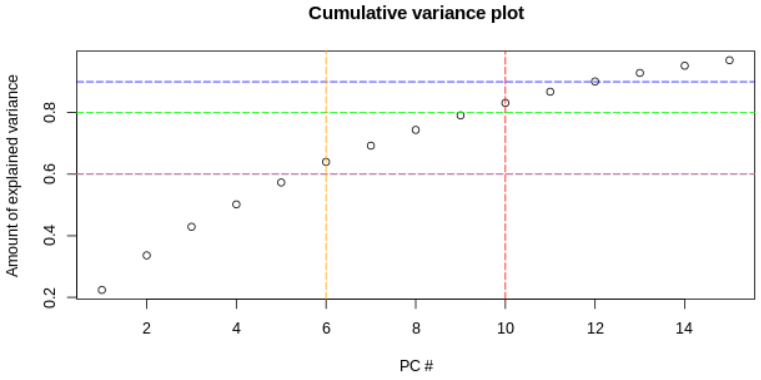
است که مقادیر کلاس دانش‌آموزان به ۰ به عنوان خوب، ۰/۵ به عنوان بد و ۱ به عنوان متوسط نگاشت شده‌اند.)

حال برای این که مشخص کنیم برای نگهداری درصدهای مختلف از تغییرات داده، چند ویژگی را باید حفظ کنیم، مقدار EigenValue ها را رسم می‌کنیم (شکل ۳)، بردارهایی که مقدار آن‌ها کمتر از ۱ باشد، توضیح زیادی درباره‌ی داده‌ها نمی‌دهند، پس می‌توانیم آن‌ها را در نظر نگیریم. به این ترتیب مشاهده می‌شود که فقط نگه‌داری ۶ بردار اصلی کافی است.



شکل ۳: EigenValues

حال برای یافتن این که چه تعداد از بردارهای ویژه، برای حفظ ۶۰، ۸۰ یا ۹۰ درصد تغییرات داده لازم است، باید به نمودار واریانس تجمعی از نتایج PCA توجه کنیم (شکل ۴) متوجه می‌شویم که برای حفظ ٪۶۰، به ۶ ویژگی، برای ٪۸۰، به ۱۰ ویژگی، و برای ٪۹۰ به ۱۳ ویژگی نیاز داریم.



شکل ۴: Cumulative Variances

بهترین بردار ویژه، برداریست که که بیشترین EigenValue را داشته و درصد بیشتری از واریانس را شامل شود. محاسبه‌ی آن با استفاده از ماتریس Correlation یا Covariance انجام می‌گیرد.

یک راه دیگر برای کاهش ابعاد داده، Stepwise Regression است که می‌توانید نحوه‌ی اجرای آن را در کدها ببینید. در کدها از ترکیب روش‌های forward و backward استفاده شده است.

۴.۲ Data Transformation and Data Discretization

با توجه به داده‌های موجود، نیازی به گسسته‌سازی و تجمیع صفت‌ها نبود، اما به دلیل این که بخش زیادی از صفت‌ها مقادیر غیر عددی بودند، نیاز بود تا ابتدا داده‌ها به مقادیر عددی تبدیل شوند و پس از آن نرمال‌سازی بر روی آن‌ها اجرا شود تا برای استفاده در PCA آماده شوند.