

عوامل موثر بر وضعیت تحصیلی دانش آموزان

فاطمه سلیمانی(9924803), سحر توکلی(9920613)

دانشکده برق و کامپیوتر، دانشگاه صنعتی اصفهان

fsoleimai697@gmail.com

sahartavakoli2002@gmail.com

مقدمه:

امتحان نیز با دو مقیاس none و completed بررسی میشوند.

نحوه انجام کار:

با استفاده از نمودارهای مختلف، تاثیر عوامل مختلف بر نمره دانش آموزان بررسی شده و روابط بین ویژگی ها و نمره ها بررسی و تحلیل می شوند. در انتها به منظور پیش بینی پاس شدن یا نشدن دانش آموز به طور کلی مدل های مختلف را بررسی و مدل با دقت بیشتر انتخاب میشود.

پردازش:

قبل از انجام هر کاری نیاز هست که داده های پرت و نویزی از دیتاست حذف شوند برای این منظور با کشیدن boxplot مربوط به هر ویژگی داده های پرت را شناسایی کرده و با در نظر گرفتن محدوده ای برخی از داده ها که نویزی شناخته شده اند حذف میشوند. در دیتاست موجود تعداد رکوردها 1000 بوده و به همین دلیل در حذف داده ها باید با دقت عمل کرد. به همین دلیل با بررسی و تست بازه های مختلف در نهایت بازه ی مطلوب برای حذف داده ها به دست آمد. و پس از حذف داده های پرت تعداد کل آن ها به 939 رکورد رسید. در قدم بعد به دلیل نیاز به سبک بندی به منظور انجام برخی بررسی ها دو روش مختلف سبک بندی انتخاب و بررسی شده و در نهایت روش بهینه تر انتخاب شده و در آزمایشات بعدی مورد استفاده قرار می گیرد.

در این مسئله عواملی که بر نتیجه ی 3 امتحان مختلف تاثیر دارد بررسی میشود و با توجه به نتیجه ی به دست آمده میتوان عواملی که تاثیر بیشتری روی نتیجه ی امتحانات دارد را بررسی کرده و بر این اساس برای بهبود نتایج امتحانات دانش آموزان از این فاکتور ها بهره ببریم.

نتیجه ی به دست آمده از این پژوهش، برای استفاده در مراکز آموزشی به منظور بهبود در وضعیت تحصیلی دانش آموزان می تواند بسیار کاربردی باشد.

روش:

داده ها:

دیتاست شامل 1000 رکورد از اطلاعات دانش آموزان است که برای هر دانش آموز اطلاعاتی از قبیل جنسیت، قومیت، وضعیت تحصیلی والدین، نوع ناهار مصرف شده و آمادگی برای امتحان وجود دارد و تاثیر این عوامل بر سه نمره ریاضی و املا و روحوانی بررسی می شود.

وضعیت تحصیلی والدین دانش آموزان در 6 سطح از some collage تا masters degree مقیاس دهی شده ، قومیت یا نژاد دانش آموزان نیز در 5 سطح در قالب گروه های A تا E مقیاس دهی شده است که هر کدام مربوط به نژاد خاصی هستند به عنوان مثال گروه A مربوط به سفید پوستان بریتانیایی هستند و ... نوع ناهار به دو مقیاس standard و free/reduced مقیاس دهی شده و آمادگی قبل از

آزمایشات:

به دلیل تحلیلی بودن ماهیت مسئله و ضرورت کشف ارتباط بین ستون ها مناسب ترین ابزار مورد استفاده رسم نمودارها و بصری سازی خواهد بود، به همین منظور با رسم نمودارهای مختلف سعی بر کشف ارتباط بین feature ها است.

در قدم اول به منظور یک بررسی کلی و مشاهده ی تاثیر هر ستون یک ستون جدید به نام percentage که شامل میانگین این 3 نمره است اضافه می شود .

در قدم بعدی با کشیدن نمودارهای مختلف توزیع نمرات در هر درس به تفکیک و همچنین روی میانگین بررسی می شود که در نتیجه مشاهده می شود در نمره ی writing بهترین عملکرد و بعد از آن در reading و بدترین عملکرد دانش آموزان در درس ریاضی دیده میشود.

در گام بعد تاثیر هر عامل را روی میانگین نمرات و همچنین به تفکیک روی هر نمره بررسی میکنیم:

فاکتور اول جنسیت است که با توجه به نمودارهای رسم شده مشاهده میکنیم خانم ها به طور میانگین و همچنین در نمره ی reading و writing عملکرد بهتری نشان دادند و در امتحان ریاضی آقایان عملکرد بهتری داشتند در نگاهی دیگر با رسم نمودار kdeplot جنبه ی دیگری از این تاثیر را مشاهده میکنیم که در نمره ی reading و writing تفاوت زیادی بین میانگین عملکرد خانم ها و آقایان مشاهده میشود در صورتیکه این اختلاف در نمره ی ریاضی کمتر به چشم می آید.

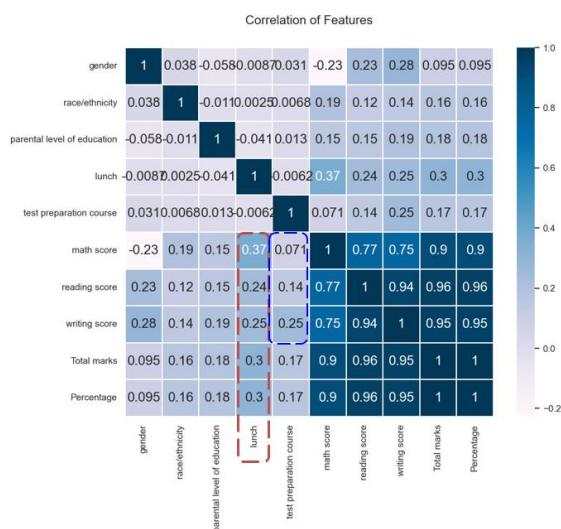
فاکتور بعدی وضعیت ناهار دانش آموزان است که مشاهده میشود افرادی که ناهار استاندارد داشتند علاوه بر میانگین نمرات به تفکیک در هر سه امتحان هم بهتر عمل کردند. با رسم نمودار kdeplot هم دیده میشود که تاثیر ناهار دانش آموزان روی نمره ی ریاضی بسیار چشمگیر تر از دو نمره ی دیگر است هرچند در دو نمره ی reading و writing هم تاثیر زیادی دارد.

فاکتور بعد آمادگی قبل از امتحان دانش آموزان است که افرادی که آمادگی کامل داشتند نسبت به افراد بدون آمادگی عملکرد بهتری در میانگین و به تفکیک روی هر 3 نمره داشتند و هم چنین این فاکتور کمترین تاثیر را روی نمره ی ریاضی و بیشترین تاثیر را روی نمره writing داشته است.

به روش مشابهی که برای عوامل دیگر تشریح شد در بررسی سطح تحصیلات والدین دانش آموزان افرادی که والدین آنها در مقطع masters degree تحصیل کردند در هر 3 نمره عملکرد بهتری نشان دادند هر چند تاثیر خیلی چشمگیری دیده نمیشود.

در بررسی فاکتور آخر یعنی قومیت افراد ، گروه D در writing score عملکرد بهتری نشان دادند و در reading score و math score گروه E بهتر عمل کردند.

پس از این مرحله لازم است نمودار heatmap به منظور بررسی correlation رسم شود که برای اینکار نیاز است داده های categorical استاندارد شده و به عدد تبدیل شوند بعد از اینکار نمودار correlation را رسم کرده و مشاهده میشود که ارتباط reading score و writing score تقریباً به صورت خطی است یعنی هر کس در یکی از ان ها عملکرد خوبی داشته باشد انتظار می رود در دیگری نیز خوب عمل کند و نرخ coef بین ان ها 0.94 است.



با توجه به نمودار بیشترین رابطه مربوط به ناهار و نمره ی ریاضی دانش آموزان است و در کل مشاهده میشود، ناهار تاثیر به نسبت چشمگیری روی نمرات و به خصوص نمره ی ریاضی داشته است ، فاکتور بعدی که بررسی میشود آمادگی قبل از امتحان دانش آموزان است که این عامل بیشترین تاثیر را روی نمره ی writing و حداقل تاثیر را روی نمره ی ریاضی داشته است.

مدل سازی:

در بحث مدل سازی این مسئله در ابتدا یک ستون تحت عنوان pass-or-not به دیتاست اضافه کرده که در صورتیکه میانگین 3 نمره ی reading, writing, math از 60 بیشتر باشد مقدار 1 میگیرد و در غیر اینصورت مقدار 0 به آن نسبت داده میشود سپس ستون های فیچر به عنوان X و ستون pass-or-not به عنوان Y در نظر گرفته میشود و سپس 40 درصد از داده ها به عنوان تست و 60 درصد به عنوان train در نظر گرفته میشود.

به عنوان baseline مدل مورد استفاده مدل DummyClassifier با استراتژی stratified است که به صورت random به داده های تست label میدهد به نحوی که توزیع آنها مشابه توزیع label ها در داده ی آموزشی باشد. که در نتیجه ی آن دقت 58 درصد به دست می آید.

پس از آن 6 مدل مختلف KNN, decision tree, random forest, Guassian Naïve Bayes, SVC, GradientBoostingClassifier را با مقادیر default

بررسی کرده که در نتیجه ی آن دقت 3 مدل KNN, SVC, GNB از مابقی بیشتر بوده بنابراین آن 3 را انتخاب و بر مبنای 4 روش scaling مختلف standardscaler, MinMaxScaler, MaxAbsScaler, RobustScaler دقت این 3 مدل را بررسی کرده و مشاهده میشود روش GNB در هر 4 روش دقت یکسان و بالایی داشته بنابراین GNB انتخاب و بررسی می شود، در مرحله ی بعد پارامترهای بهینه در مدل GNB را به دست آورده و مدل با این پارامترهای بهینه اجرا می شود که در نتیجه ی آن $accuracy=0.74$ می دهد.

از آن جا که این دیتاست بیشتر ماهیت categorical داشته تصمیم بر آن شد مدل random forest با پارامترهای بهینه روی آن تست شود که در نتیجه ی آن $accuracy=0.73$ نتیجه داد که البته دقت آن از روش GNB کمتر است.

نتایج:

- نمره reading, writing به طور خطی و با ضریب 0.94 به هم مرتبط اند. یعنی اگر دانش آموزی در یکی از این امتحان ها عملکرد خوبی داشته باشد در دیگری هم عملکرد خوبی دارد.

- به طور کلی خانم ها عملکرد بهتری داشتند ولی به طور جدا آقایان در ریاضی و خانم ها در reading writing عملکرد بهتری داشتند.
- گروه E عملکرد بهتری در math, reading داشته و گروه D عملکرد بهتری در writing داشته اما به طور کلی گروه E بهتر عمل کرده است.
- نوع ناهار مصرف شده بیشترین تاثیر را روی نمره ها داشته، به خصوص روی نمره ریاضی.
- آمادگی قبل امتحان بیشترین تاثیر را روی نمره writing و کمترین تاثیر را روی نمره ریاضی داشته اما به طور کلی کسانی که آمادگی قبل امتحان داشتند بهتر عمل کردند.
- دانش آموزانی که والدینی با وضعیت تحصیلی بالاتر داشتند بهتر عمل کردند.

جمع بندی و نتیجه گیری:

یکی از سوالات اساسی که این مسئله به منظور پاسخ به آن انجام شده است این است که بهترین راهکار برای بهبود وضعیت تحصیلی دانش آموزان در هر امتحان چیست؟

که با توجه به بررسی های انجام شده میتوان نتیجه گرفت ناهار استاندارد بیشترین تاثیر را روی هر سه امتحان به خصوص امتحان ریاضی داشته است و کسانی که ناهار استاندارد خوردند با اختلاف خوبی نسبت به افرادی که ناهار غیر استاندارد صرف کردند عملکرد بهتری داشته و نمره های بالاتری کسب کردند و همچنین فاکتور موثر دوم آمادگی افراد برای امتحان بوده است که این مورد در امتحان writing بیشترین تاثیر را داشته است و افراد با آمادگی کامل در امتحان writing نمره های بهتری کسب کردند ولی این فاکتور در نمره ی ریاضی حداقل تاثیر را داشته و حتی میتوان از تاثیر آن چشم پوشی کرد.

• <https://www.kaggle.com/code/victorferino/student-s-performance-in-exams-eda-ml#Multiple-Linear-Regression-Model>

در این مسئله سعی بر آن بوده است که تمامی فاکتورهای موثر بر نمرات به طور کامل بررسی شوند اما از آنجاییکه نتیجه ی این مسئله برای مراکز آموزشی و مدارس میتواند بسیار کمک کننده باشد و هدف بیزینسی آن این جامعه هستند، میتوان گفت بررسی عواملی که جنبه ی اکتسابی داشته و افراد در کمتر یا بیشتر کردن تاثیر آن ها نقش دارند اهمیت بیشتری پیدا میکنند مثلاً در این دیتاست عواملی مانند جنسیت یا قومیت افراد قابل تغییر نیستند اما عواملی مانند وضعیت ناهار و آمادگی قبل از امتحان قابل تغییر بوده و اهمیت خیلی زیادی پیدا میکنند چون مراکز آموزشی یا حتی خانواده های دانش آموزان میتوانند از این نتیجه بهره ببرند و از آن برای بهبود وضعیت تحصیلی دانش آموزان خود کمک بگیرند.

منابع:

• http://roycekimmons.com/tools/generated_data/exams

• <https://www.kaggle.com/code/nihar14/analysis-on-factors-affecting-students-scores>

• <https://github.com/AzT3Risk/Students-performance-in-Exams>

• <https://www.kaggle.com/code/kagleo123/student-performance-in-exam-eda-ml-prediction>