
بسم الله الرحمن الرحيم

دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

گروه مهندسی فناوری اطلاعات

گزارش پروژه کارشناسی

رشته مهندسی کامپیوتر گرایش فناوری اطلاعات

عنوان پروژه

پیش‌بینی سلامت روان دانش‌آموزان بر اساس داده‌های مربوط به سبک زندگی آنها با استفاده از روش های داده کاوی

استاد راهنما:

دکتر مرجان کائدی

پژوهشگر:

محمدهادی حقوقی

تیر ۱۴۰۴

دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

گروه مهندسی فناوری اطلاعات

پروژه کارشناسی رشته‌ی مهندسی کامپیوتر گرایش فناوری اطلاعات

آقای محمدهادی حقوقی

تحت عنوان

پیش‌بینی سلامت روان دانش‌آموزان بر اساس داده‌های مربوط به سبک زندگی آنها با استفاده از روش های داده کاوی

در تاریخ / / ۱۳ توسط هیأت داوران زیر بررسی و با نمره به تصویب نهایی رسید.

۱- استاد راهنمای پروژه:

دکتر امضا

۲- استاد داور :

دکتر امضا

امضای مدیر گروه

تقدیم به

با افتخار و تواضع، این پایان‌نامه را به محضر ارزشمند خانواده‌ام، به‌خصوص پدر و مادرگران‌قدرم که همواره در هر لحظه از زندگی، بامحبت، حمایت و ایثار، پشتیبان من بوده‌اند، ارائه می‌نمایم. بدون حضور ایشان، انجام کاری؛ مانند این پایان‌نامه، غیرممکن بوده و اینجا محل قدردانی از زحمات و تلاش‌های بی‌دریغ آنان است.

همچنین، باکمال تشکر و احترام، این اثر را به استاد ارجمندم، سرکار خانم دکتر مرجان کائدی که با لطف، توجه و راهنمایی‌های بی‌دریغشان، مرا در این سفر پژوهشی همراهی نموده‌اند، تقدیم می‌نمایم. ایشان به‌عنوان راهنما، با توجهات علمی و انسانی خود، روحیه و انگیزهٔ من را تقویت نموده و به دستاوردهایم ارزش افزوده‌اند. تلاش‌ها و زحمات ایشان را به یادگار دارم و این اثر را به ایشان اختصاص می‌دهم.

چکیده:

پروژه‌ی حاضر با هدف بررسی وضعیت سلامت روان دانش‌آموزان و تحلیل تأثیر عوامل سبک زندگی بر آن، با بهره‌گیری از داده‌های طرح ملی کاسپین ۵ و با تکیه بر تکنیک‌های نوین داده‌کاوی و یادگیری ماشین اجرا شده است. مجموعه‌داده‌ی مورد استفاده شامل پاسخ‌های ۲۰۰ دانش‌آموز از شهر تبریز در بازه‌ی سنی ۱۳۷۹ تا ۱۳۸۷ به پرسش‌نامه‌ای شامل ۲۳ سؤال در چهار محور اصلی یعنی تغذیه، فعالیت بدنی، اوقات فراغت و سلامت و ناخوشی‌ها می‌باشد. هدف این پروژه استخراج الگوهای پنهان میان متغیرها و شناسایی عوامل مؤثر بر سلامت روان دانش‌آموزان بوده است.

در مراحل ابتدایی، پس از تحلیل ساختار داده و شناسایی چالش‌های موجود، فرآیند پیش‌پردازش داده‌ها آغاز شد. این گام شامل حذف نمونه‌های نامعتبر، حذف ستون‌های کم‌ارزش، نرمال‌سازی داده‌ها، پر کردن مقادیر گمشده، اصلاح واحدهای زمانی، دسته‌بندی متغیرها بر اساس هرم غذایی، و تبدیل متغیرهای کیفی به کمّی بود. همچنین با استفاده از تحلیل‌های آماری و بصری همچون هیستوگرام، نمودارهای جعبه‌ای و ماتریس همبستگی، روابط اولیه‌ی میان ویژگی‌ها بررسی شد.

در ادامه، با طراحی دو مدل شبکه‌ی عصبی، ابتدا هدف به‌صورت رگرسیون مدل‌سازی شد تا امتیاز سلامت روان دانش‌آموزان پیش‌بینی شود. در مرحله‌ی بعد، امتیازات به سه دسته‌ی «بحرانی»، «آسیب‌پذیر» و «مطلوب» تقسیم شده و پروژه در قالب مسئله‌ی دسته‌بندی چندکلاسه ادامه یافت. با آموزش مدل‌ها و ارزیابی آن‌ها بر اساس معیارهایی نظیر دقت، دقت مثبت، حساسیت و امتیاز $F1$ ، مشخص شد که عدم توازن نمونه‌ها در کلاس‌های مختلف به عملکرد نامطلوب مدل در برخی کلاس‌ها منجر شده است.

در پاسخ به این چالش، از روش تقویت داده‌ها به‌صورت دستی بهره گرفته شد. با افزایش داده‌های نمونه‌های کم‌تعداد (به‌ویژه گروه بحرانی) و تنظیم مجدد توزیع کلاس‌ها، عملکرد مدل به‌طور محسوسی ارتقا یافت. همچنین اختلاف دقت بین داده‌های آموزش و اعتبارسنجی کاهش یافته و از بیش‌برازش کاسته شد.

نتایج تحلیل نشان داد که برخی از رفتارهای سبک زندگی نظیر مصرف بیشتر سبزیجات، لبنیات و فعالیت بدنی با وضعیت بهتر سلامت روان همراه است، در حالی که مصرف بیش‌ازحد شیرینی‌جات، چربی‌ها و زمان زیاد اوقات فراغت غیرمولد، اثرات منفی برجای می‌گذارند. همچنین الگوهای رفتاری دانش‌آموزان در ثبت وضعیت سلامت روان خود نیز حاوی نکاتی مهم در خصوص نگاه فرهنگی به مسائل روانی بود.

پروژه‌ی حاضر با تلفیق داده‌کاوی و تحلیل رفتاری می‌تواند بستر مناسبی برای طراحی سامانه‌های هوشمند غربالگری، توصیه‌گرهای سبک زندگی و آموزش‌های سلامت‌محور در مدارس فراهم سازد. همچنین مسیر انجام این پروژه تجربه‌ای ارزشمند در تحلیل داده‌های اجتماعی با حجم محدود و کیفیت متنوع بوده است که قابلیت توسعه در سطح ملی را داراست.

واژگان کلیدی: سلامت‌روان، الگوهای پنهان، پیش‌پردازش، شبکه‌ی عصبی، رگرسیون، آموزش‌مدل، دقت، حساسیت، بیش‌برازش، داده‌کاوی

فهرست مطالب

عنوان **صفحه**

فهرست مطالب

عنوان **صفحه**

فهرست مطالب

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فهرست جدول‌ها

عنوان **صفحه**

مخفف‌ها:

GSHS	Global School Student Health Survey
UNAIDS	Joint United Nations Program on HIV/AIDS
WHO	World Health Organization
DAMA	Data Management Association
CRSIP-DM	Cross-industry standard process for data mining
SVM	Support vector machines
DBSCAN	Density-based spatial clustering of applications with noise
PCA	Principal Component Analysis
ANNs	Artificial Neural Networks
IDE	Integrated Development Environment
CSV	Comma-Separated Values
ReLU	Rectified Linear Unit
MSE	Mean squared error
MAE	Mean Absolute Error
GHQ	Global Healthy Questionnaire

فصل ۱
مقدمه

۱-۱- هدف پروژه

یکی از محورهای ارزیابی سلامتی جوامع مختلف، سلامت روان[1] آن جامعه است[7]. امروزه باتوجه‌به ماشینی شدن زندگی، اختلالات‌روانی، باراقتصادی و اجتماعی سنگینی برسیستم‌های مراقبت‌های بهداشتی سراسر جهان تحمیل می‌کند. باتوجه‌به پیش‌بینی‌های سازمان جهانی بهداشت تا سال ۲۰۲۰ اختلالات روانی پانزده درصد از بارجهانی بیماری‌ها را به خود اختصاص خواهد داد و اختلالاتی همچون افسردگی نیز به‌عنوان دومین علت بار بیماری در سراسر جهان پس از بیماری‌های قلبی عروقی شناخته خواهد شد[2]. همچنین آمارها نشان می‌دهند از هرچهار نفر در جهان یک نفر دچار اختلالات روانی یا عصبی است و درحال حاضر نیز ۴۵۰ میلیون نفر از اختلالات روانی رنج می‌برند[9]. کشور ما نیز از این قاعده مستثنی نیست و شیوع اختلالات روانی ۲۱.۲۳ درصد در مناطق شهری و ۲۰.۹ درصد در مناطق روستایی گزارش شده است.[10] همچنین اختلالات روانی بربخش درخور توجهی از دانش‌آموزان در گروه سنی نوجوانان در سراسر جهان تاثیر گذاشته است؛ به گونه‌ای که مطالعات انجام شده در ۲۷ کشور نشان‌دهنده‌ی شیوع شایان توجهی از اختلالات روانی در دانش‌آموزان نوجوان معادل با ۱۳.۴ درصد هستند[3]. همچنین باتوجه‌به اینکه بسیاری از آسیب‌های روانی در دوران بزرگسالی ناشی از مشکلات دوران نوجوانی است و با تأکید براهمیت قشر دانش‌آموز در سازندگی و پیشرفت جامعه؛ تحلیل داده‌های مربوط به سلامت روان دانش‌آموز در سازندگی پیشرفت جامعه؛ تحلیل داده‌های مربوط به سلامت روان دانش‌آموزان در این گروه به‌منظور نیل به اهداف عالیه‌ی حقوق طبیعی و اجتماعی نوجوانان و مداخله‌ی منطقی و اصولی برای ارتقای سلامت روان جامعه امری ضروری است. این در حالی است که افزایش حجم داده‌ها نیاز به تحلیل و مدیریت دارد و در سطحی بالاتر کشف دانش موجود در آن‌ها و استفاده از تکنیک‌هایی همچون داده‌کاوری[2] در حوزه‌ی سلامت را بیش‌ازپیش نمایان

[1] Mental health
[2] Data mining

می‌کند. در واقع به واسطه‌ی داده‌کاوی در پایگاه داده‌ی سلامت‌روان دانش‌آموزان است که می‌توان امکان کشف‌روابط، روندها و الگوهای مخفی میان داده‌های سلامت روان و دستیابی به دانش‌نوین در این زمنیه را میسر ساخت و گامی مؤثر در راستای افزایش سلامت عمومی برداشت. تحقیقات متعددی در زمینه‌ی استفاده از این تکنینک‌ها در حوزه‌ی سلامت روان انجام شده است. دانیالی و همکارانش(۱۳۹۷) درمطالعه‌ی مقطعی با عنوان «ارتباط خوشه‌های اضطراب و اختلالات روان‌تنی با عادات سبک زندگی در کودکان و نوجوانان؛ مطالعه کاسپین۵[1]» به پیش‌بینی ارتباط اضطراب و اختلالات روان‌تنی با متغیرهای سبک زندگی(عادات‌غذایی، رفتارهای‌بی‌تحرکی و مدت‌زمان خواب) با استفاده از مدل رگرسیون لجستیک[2] پرداخته‌اند. یافته‌ها نشان می‌دهد که خطر سبک زندگی ناسالم(مانند استفاده کم از شیر، میوه و سبزیجات، نداشتن فعالیت بدنی، استفاده زیاد از آب نبات، غذاهای شور، نوشابه و سیگار کشیدن) در کودکان و نوجوانان دارای اضطراب و اختلالات روان‌تنی به طور شایان توجهی از دیگر کودکان و نوجوانان بیشتر است[4]. آلونیو و همکارانش(۲۰۱۸) در مطالعه‌ی مروری سیستماتیک با عنوان«تکنیک‌ها و الگوریتم‌های داده‌کاوی در سلامت روان؛ مروری سیستماتیک» بیان داشتند که به‌کارگیری تکینک‌های تصمیمات بالینی، پیش‌بینی در تشخیص و بهبود کیفیت زندگی بیماران مثمر ثمر واقع گردد[11]. رحمان(۲۰۱۹) در مطالعه‌ای با عنوان«کاربرد داده‌کاوی در داده‌های سلامت روان» با استفاده از تکنیک درخت تصمیم[3] در تحلیل داده‌های سلامت روان بدین نتیجه دست‌یافت که تداخلات کاری و سابقه‌ی خانوادگی از مهم‌ترین عوامل در پیش‌بینی بیماری‌های سلامت روان در سازمان‌ها هستند. درخت تصمیم از روش‌های ساده و قدرتمند آنالیز چند متغیره است که نتایج رده‌بندی را به وسیله‌ی نمودار ساده و درک‌شدنی ارائه می‌دهد[12]. خانم آقاداودیان به همراه جمعی از اساتید دانشگاه علوم پزشکی اصفهان با پژوهش برروی داده‌های پرسش‌نامه‌ی کاسپین۵ و با استفاده از نرم‌افزار رپیدماینر به نتایج قابل توجهی در این حوزه دست یافتند[2]. بررسی‌های آنها نشان می‌دهد که تغذیه مهم‌ترین عامل تاثیرگذار برسلامت روان دانش‌آموزان خواهد بود[2]. اما آنچه که می‌تواند مورد بحث قرار گیرد این است که آیا مراحل داده‌کاوی از جمله: پیش‌پردازش داده‌ها، انتخاب ویژگی‌ها، مدل‌سازی و آموزش‌الگوریتم‌ها، ارزیابی مدل‌ها و تنظیم و بهینه‌سازی مدل‌ها به خوبی انجام شده است؟ آیا می‌توان با استفاده از روش‌های پیشرفته‌تر مانند استفاده از روش‌های مبتنی بر شبکه‌ی عصبی[4] به نتایج‌هایی مشابه در این حوزه دست‌یافت؟

---

[1] Caspian V
[2] Logistic Regression
[3] Decision tree
[4] Neural network

۱-۲- کاربردهای پروژه

با توجه به افزایش مشکلات روانی در میان دانش‌آموزان و تأثیر مستقیم آن بر عملکرد تحصیلی، رفتار اجتماعی، و کیفیت زندگی آنان، ضرورت دارد که از ابزارهای نوین برای شناسایی و تحلیل این مشکلات استفاده شود. پروژه‌ی حاضر با بهره‌گیری از روش‌های داده‌کاوی و الگوریتم‌های هوش مصنوعی، تلاشی در جهت کشف الگوهای پنهان در داده‌های مرتبط با سلامت روان دانش‌آموزان بوده است. یافته‌های این پروژه، نه‌تنها از منظر پژوهشی ارزشمند هستند، بلکه می‌توانند در سطوح مختلف آموزش، روان‌شناسی، و سیاست‌گذاری کاربردهای عملی مهمی داشته باشند.

۱. شناسایی زودهنگام اختلالات روانی

یکی از مهم‌ترین کاربردهای این پروژه، امکان شناسایی زودهنگام اختلالات روانی در میان دانش‌آموزان است. مدل‌های طراحی‌شده می‌توانند با تحلیل داده‌های رفتاری و روان‌شناختی، افرادی را که در معرض خطر مشکلاتی مانند اضطراب، افسردگی، استرس یا اختلالات رفتاری هستند، به سرعت تشخیص دهند. این امر باعث می‌شود مشاوران مدرسه و والدین بتوانند مداخله‌های به‌موقع و مؤثر انجام دهند و از تشدید اختلالات جلوگیری کنند.

۲. طراحی برنامه‌های مشاوره‌ای هدفمند

با استفاده از نتایج این پروژه می‌توان برنامه‌های مشاوره‌ای و روان‌درمانی را به‌صورت دقیق‌تر و متناسب با نیازهای واقعی دانش‌آموزان طراحی کرد. تحلیل الگوهای رفتاری و روان‌شناختی می‌تواند به مدارس کمک کند تا منابع خود را در جهت مسائل شایع‌تر مانند اضطراب امتحان، فشار خانواده یا ضعف در مهارت‌های ارتباطی متمرکز کنند.

۳. کمک به سیاست‌گذاری در حوزه سلامت روان

داده‌ها و تحلیل‌های به‌دست‌آمده از این پروژه می‌توانند پشتوانه‌ای علمی برای تدوین سیاست‌های آموزشی و بهداشتی در سطح مدارس، مناطق آموزشی یا حتی وزارت آموزش و پرورش باشند. برای مثال، اطلاعات به‌دست‌آمده می‌تواند مشخص کند کدام مدارس یا مناطق به خدمات روان‌شناختی[1] بیشتری نیاز دارند.

٤. بهبود عملکرد تحصیلی و انگیزه دانش‌آموزان

با شناسایی و رسیدگی به مشکلات روانی، وضعیت روحی دانش‌آموزان بهبود می‌یابد و این موضوع می‌تواند تأثیر مثبتی بر تمرکز، انگیزه و عملکرد تحصیلی آنان داشته باشد. بنابراین نتایج پروژه، در عین تمرکز بر سلامت روان، به ارتقای کیفیت آموزشی نیز منجر می‌شود.

---

[1] cognitive psychology

۵. توانمندسازی والدین و معلمان

یکی دیگر از کاربردهای مهم این پروژه، فراهم‌کردن اطلاعات و ابزارهای تحلیلی برای والدین و معلمان است. آن‌ها با استفاده از این نتایج می‌توانند رفتارهای نگران‌کننده را بهتر درک کرده، در برابر آن واکنش مناسب‌تری نشان دهند و نقش حمایتی مؤثرتری ایفا کنند.

٦. توسعه سامانه‌های هوشمند غربالگری

پروژه حاضر می‌تواند پایه‌ای برای طراحی و توسعه سامانه‌های هوشمند غربالگری سلامت روان باشد؛ سامانه‌هایی که با جمع‌آوری و تحلیل خودکار داده‌های دانش‌آموز، هشدارهایی در مورد خطرات احتمالی ارائه دهند و فرآیند ارجاع به متخصص را تسهیل کنند. این کاربرد به‌ویژه در مدارس با تعداد بالای دانش‌آموزان بسیار مؤثر خواهد بود.

۷. پشتیبانی از مداخلات فردی و شخصی‌سازی‌شده

با استفاده از تحلیل‌های دقیق داده‌ای، می‌توان برای هر دانش‌آموز، پروفایل[1] سلامت روانی شخصی ایجاد کرد. این پروفایل‌ها به مشاوران کمک می‌کند تا مداخلات دقیق‌تری ارائه دهند که با شرایط، نیازها و ویژگی‌های روان‌شناختی هر فرد هماهنگ باشد، در نتیجه اثربخشی مداخلات افزایش می‌یابد.

۱-۳- ساختار پایان نامه

پایان‌نامه حاضر در ۶ فصل تنظیم‌شده است که به بررسی دقیق و تحلیل موضوع پژوهش پرداخته شده است. ساختار این پایان‌نامه به گونه‌ای طراحی شده است که خواننده را به صورت مرحله‌به‌مرحله از مفاهیم ابتدایی و کلی به جزییات و نتایج نهایی هدایت می‌کند. این ساختار شامل فصل‌های زیر است.

۱. **فصل اول: مقدمه**

در این فصل به معرفی کلی پروژه پرداخته‌ایم. اهداف اصلی و کاربردهای پروژه به طور جامع تشریح کردیم و نقش و اهمیت آن در زمینه تحقیق موردنظر بیان کردم.

۲. **فصل دوم: مفاهیم و ادبیات**

در این فصل به بررسی مفاهیم نظری، دیدگاه‌های علمی، و پژوهش‌های پیشین مرتبط با موضوع تحقیق اختصاص دارد. در این فصل، ابتدا به تبیین مفاهیم کلیدی در حوزه سلامت روان دانش‌آموزان پرداخته می‌شود و نظریه‌های روان‌شناسی مرتبط مانند نظریه‌های رشد هیجانی، اضطراب، دلبستگی، و خودکارآمدی مورد بررسی قرار می‌گیرند. سپس به کاربردهای هوش مصنوعی و داده‌کاوی در حوزه روان‌شناسی و آموزش پرداخته می‌شود، به‌ویژه استفاده از الگوریتم‌های یادگیری ماشین[2] برای تحلیل

---

[1] Profile

[2] Machine learning

داده‌های روانی و رفتاری.

۳. **فصل سوم: شرح پروژه**

در فصل سوم، جزئیات عملی پروژه شامل نحوه‌ی جمع‌آوری داده‌ها، جامعه‌ی آماری، ابزارهای مورد استفاده، مراحل پیش‌پردازش، انتخاب ویژگی‌ها و پیاده‌سازی مدل‌های یادگیری ماشین به تفصیل بیان شده است. همچنین به چالش‌ها و تصمیمات اتخاذ شده در طول فرایند اجرای پروژه اشاره شده است.

٤. **فصل چهارم: نتایج**

این فصل به ارائه‌ی نتایج حاصل از تحلیل‌های آماری، مصورسازی داده‌ها، ارزیابی مدل‌های ساخته‌شده و بررسی عملکرد آن‌ها با استفاده از معیارهایی نظیر دقت، حساسیت و F1-score می‌پردازد. همچنین نتایج به‌دست‌آمده از فاز تقویت داده‌ها و تأثیر آن بر عملکرد مدل‌ها بررسی شده‌اند

٥. **فصل پنجم: نتیجه‌گیری و پیشنهادات**

در این فصل، جمع‌بندی کلی از مسیر انجام پژوهش و نتایج حاصل‌شده ارائه شده است. همچنین پیشنهاداتی برای بهبود روند تحقیق، توسعه‌ی مدل‌ها و کارهای آتی در حوزه‌ی تحلیل داده‌های سلامت روان دانش‌آموزان مطرح گردیده است.

٦. **فصل ششم: تحقیقات پیشین**

آخرین فصل به معرفی و بررسی پژوهش‌های انجام‌شده در زمینه‌ی مشابه، شامل مطالعات داخلی و خارجی در حوزه‌ی سلامت روان نوجوانان و تحلیل داده‌های مرتبط با آن می‌پردازد. مقایسه‌ی روش‌ها و نتایج این پژوهش‌ها با یافته‌های پروژه‌ی حاضر نیز در این فصل گنجانده شده است.

فصل ۲
مفاهیم

۲-۱- مقدمه

در هرپژوهشی، درک صحیح و جامع از مفاهیم مرتبط با حوزه تخصصی کارشده، نقش کلیدی در موفقیت نهایی آن ایفا می‌کند. این بخش با هدف آشنایی خواننده با مفاهیم و مبانی اساسی که در روند انجام این پروژه شده است، تدوین شده است.

سلامت جسم و روان دانش‌آموزان به عنوان آینده‌سازان جامعه، همواره یکی از محورهای کلیدی در حوزه سیاست‌گذاری آموزشی و بهداشتی کشورها بوده است. شیوه‌های زندگی، شامل الگوهای تغذیه، فعالیت بدنی، نحوه گذراندن اوقات فراغت، و تجربه بیماری‌های جسمی یا روانی، از مهم‌ترین عوامل تأثیرگذار بر سلامت نوجوانان محسوب می‌شوند. در این میان، طرح‌های ملی مانند پرسشنامه کاسپین ۵[1] در ایران، با گردآوری داده‌های ارزشمند از وضعیت زندگی و سلامت نوجوانان، بستری مناسب برای تحلیل علمی این ابعاد فراهم آورده‌اند.

از سوی دیگر، با گسترش حجم داده‌های سلامت در سطح مدارس و مراکز تحقیقاتی، روش‌های سنتی تحلیل آماری دیگر پاسخ‌گوی نیازهای پژوهشی دقیق و پیچیده نیستند. در این راستا، داده‌کاوی و به‌ویژه استفاده از شبکه‌های عصبی مصنوعی به‌عنوان یکی از شاخه‌های هوش مصنوعی، ابزارهایی قدرتمند برای کشف الگوهای پنهان در داده‌ها و تحلیل روابط میان متغیرهای مختلف فراهم کرده‌اند. این روش‌ها قابلیت پیش‌بینی، طبقه‌بندی و تفسیر داده‌های پیچیده را دارا هستند و در سال‌های اخیر، جایگاه مهمی در پژوهش‌های حوزه سلامت نوجوانان یافته‌اند.

در این فصل، ابتدا به معرفی دقیق‌تر پرسشنامه کاسپین و ابعاد محتوایی آن پرداخته می‌شود. سپس مفاهیم نظری مربوط به سلامت جسمی و روانی دانش‌آموزان و تأثیر سبک زندگی بر آن بررسی می‌گردد. در

[1] Caspian V

ادامه، اصول پایه داده‌کاوی و شبکه‌های عصبی به زبان ساده تبیین شده و جایگاه آن‌ها در مطالعات سلامت نوجوانان روشن می‌شود. در پایان نیز مروری بر پیشینه پژوهش‌های داخلی و خارجی در این زمینه انجام خواهد شد تا جایگاه تحقیق حاضر در میان مطالعات موجود مشخص گردد.

۲-۲- معرفی طرح پیمایش کاسپین

به‌منظور بررسی و رصد عوامل خطری که سلامتی کودکان را تهدید می‌کند، سازمان جهانی بهداشت نظام مراقبت سلامت دانش‌آموزان را تحت عنوان (GSHS) تدوین کرده است.GSHS یک پیمایش مدرسه محور[1] است که نسبتاً کم‌هزینه است و عمدتاً در بین دانش‌آموزان ۱۳ تا ۱۷ ساله انجام می‌شود. هدف از GSHS ارائه اطلاعات دقیق در مورد رفتارهای بهداشتی و عوامل محافظتی در میان دانش‌آموزان است. این مدل بررسی رفتاری توسط سازمان جهانی بهداشت با همکاری سازمان ملل متحد، یونیسف[2]، یونسکو[3]، سازمان برنامه مشترک ملل متحد در زمینه ایدز[4] و مرکز کنترل و پیشگیری بیماری‌ها بنیان گذاشته شده است و به‌عنوان ابزاری برای بررسی جهانی وضعیت و روند تغییر رفتارهای سلامت و عوامل حفاظت کننده در سنین نوجوانی در سطح بین‌المللی توصیه و ترویج می‌شود. جمهوری اسلامی ایران، مجموعه مطالعاتی را در قالب برنامه نظام مراقبت دانش‌آموزان تحت عنوان برنامه کاسپین (نظارت و پیشگیری از بیماری‌های غیرواگیر در دوران کودکی و نوجوانی در بزرگسالان) به مورد اجرا درآورده است. این برنامه توسط اداره سلامت نوجوانان و مدارس وزارت بهداشت، درمان و آموزش پزشکی با همکاری وزارت آموزش‌وپرورش و حمایت سازمان جهانی بهداشت و یونیسف از سال ۱۳۸۳ هر دو سال یک‌بار اجرا شده است. تاکنون ۵ مرحله پیمایش کاسپین در ایران به اجرا درآمده است[۱۳].

کاسپین۵ در طی سال‌های ۱۳۹۴-۱۳۹۳ در مناطق شهری و روستایی ۳۱ استان اجرا شده است. جامعه آماری مورد بررسی شامل تمامی دانش‌آموزان شاغل به تحصیل در مقاطع دوگانه ابتدایی و متوسطه در مناطق شهری و روستایی استان‌های سراسر کشور است. این مطالعه در ۳۱ استان کشور و بر روی دانش‌آموزان ۷ تا ۱۸ سال انجام شد. واحد آماری مورد بررسی شامل دانش‌آموز شاغل به تحصیل در زمان انجام پرسشگری است. دانش‌آموزان ایرانی ساکن در مناطق شهری و روستایی استان‌های سراسر کشور به شرط داشتن شناسنامه ایرانی وارد مطالعه شدند و دانش‌آموزان خارجی و تابعه سایر کشورها حتی در صورت داشتن مجوز اقامت و یا گویش به زبان فارسی وارد مطالعه نشدند. نمونه‌گیری در سطح دانش‌آموزان از نوع چندمرحله‌ای با

---

[1] School-based
[2] unicef
[3] unesco
[4] Joint United Nations Programme on HIV/AIDS

استفاده از روش‌های نمونه‌گیری خوشه‌ای[1] و طبقه‌ای[2] بود. نمونه‌گیری طبقه‌ای در داخل هر استان و برحسب محل سکونت دانش‌آموز (شهر و روستا) و مقطع تحصیلی (ابتدایی، راهنمایی و دبیرستان) به شیوه متناسب با اندازه[3] با نسبت جنسی برابر انجام شد. نحوه رسیدن به نمونه موردنظر و انتخاب آن‌ها با استفاده از نمونه‌گیری خوشه‌ای در سطح هر استان و با اندازه خوشه‌های برابر صورت گرفت. خوشه‌ها در سطح مدارس تعیین گردید. اندازه هر خوشه ۱۰ نفر برآورد شد یعنی باید در هر خوشه تعداد ۱۰ واحد آماری (شامل ۱۰ دانش‌آموز و والدین او) قرار بگیرد. حجم نمونه محاسبه‌شده برای این مطالعه ۴۸۰ نفر در هر استان بود؛ یعنی ۴۸ خوشه ۱۰ نفری در هر یک از استان‌های کشور و در مجموع با توجه به انجام مطالعه در ۳۱ استان ۱۴۸۸۰ نفر مورد بررسی قرار گرفتند که نهایتاً داده‌های مربوط به ۱۴۲۷۴ نفر مورد تجزیه‌وتحلیل آماری قرار گرفت[۱۳].

۲-۳- مفاهیم کلیدی سلامت دانش‌آموزان

۲-۳-۱- تعریف سلامت از دیدگاه سازمان جهانی بهداشت[4]

سازمان جهانی بهداشت سلامت را چنین تعریف می‌کند: «سلامت حالتی از رفاه کامل جسمی، روانی و اجتماعی است و نه صرفاً فقدان بیماری یا ناتوانی». بر اساس این تعریف، سلامت یک وضعیت ایستا نیست، بلکه یک فرآیند پویا و تعاملی میان ابعاد مختلف زندگی فرد است. از این رو، تحلیل سلامت دانش‌آموزان مستلزم بررسی درهم‌تنیدگی عوامل بدنی، روانی، اجتماعی، محیطی، و سبک زندگی آن‌هاست

۲-۳-۲- سلامت روان و عاطفی دانش‌آموزان

نوجوانی یکی از حساس‌ترین دوره‌های رشد روانی و عاطفی است. این دوره با شکل‌گیری هویت فردی، افزایش حساسیت به قضاوت اجتماعی، و رشد شناختی همراه است. عواملی مانند استرس‌های درسی، فشارهای اجتماعی، تعارض‌های خانوادگی، و استفاده نادرست از رسانه‌های دیجیتال، می‌توانند زمینه‌ساز مشکلاتی همچون اضطراب، افسردگی، کاهش عزت‌نفس، و اختلالات خواب در میان دانش‌آموزان شوند.

سلامت روانی پایدار در این سنین نه تنها باعث افزایش تمرکز، یادگیری مؤثر، و تعاملات اجتماعی مثبت می‌شود، بلکه پیشگیری مؤثری از بروز اختلالات روانی در بزرگسالی نیز به شمار می‌آید.

۲-۳-۳- سلامت جسمی و رشد بدنی

رشد جسمی دانش‌آموزان شامل افزایش قد، وزن، رشد استخوان‌ها و اندام‌ها، و بلوغ جنسی است که

---

[1] cluster sampling
[2] Staritified sampling
[3] Proportional to size
[4] World Health Organization

به‌شدت تحت تأثیر تغذیه، فعالیت بدنی، کیفیت خواب و مراقبت‌های بهداشتی قرار دارد. بیماری‌های شایع این سنین شامل کم‌خونی، چاقی، پوسیدگی دندان، مشکلات بینایی، و اختلالات گوارشی است. شیوه زندگی دانش‌آموزان، به‌ویژه در محیط‌های شهری، اغلب با کم‌تحرکی، تغذیه نامناسب و مصرف بالای فست‌فودها همراه است.

۲-۳-۴- سلامت اجتماعی و روابط بین فردی

روابط اجتماعی نوجوانان، به‌ویژه تعامل با همسالان، معلمان و خانواده، نقش بسیار مهمی در شکل‌گیری رفتارهای سالم یا ناسالم دارد. وجود حمایت اجتماعی کافی، تجربه مشارکت در فعالیت‌های گروهی، و احساس تعلق به مدرسه می‌توانند از عوامل محافظت‌کننده سلامت روان باشند. در مقابل، تجربه‌هایی مانند طرد اجتماعی، خشونت، قلدری مدرسه‌ای، یا فقدان روابط گرم خانوادگی می‌تواند منجر به انزوا، پرخاشگری یا افت تحصیلی شود.

۲-۳-۵- نقش تغذیه در سلامت نوجوانان

تغذیه مناسب یکی از پایه‌های اصلی رشد و سلامت در دوره نوجوانی است. نیازهای تغذیه‌ای در این سن به دلیل رشد سریع، بیشتر از دوران کودکی است. کمبودهای تغذیه‌ای (مانند آهن، کلسیم، و ویتامین‌ها) می‌توانند باعث خستگی، ضعف ایمنی، کاهش تمرکز و یادگیری شوند. از سوی دیگر، مصرف بیش از حد غذاهای فرآوری‌شده، شیرین و پرچرب منجر به افزایش خطر چاقی، دیابت، و مشکلات قلبی‌عروقی می‌گردد.

۲-۳-۶- اهمیت فعالیت بدنی منظم

فعالیت بدنی منظم در سلامت جسمی و روانی نوجوانان نقش مهمی ایفا می‌کند. ورزش علاوه بر تأثیر مثبت بر سیستم قلبی-عروقی و عضلانی، باعث ترشح هورمون‌های شادی‌آور مانند اندورفین[1] شده و در کاهش اضطراب و افسردگی مؤثر است. با این حال، در بسیاری از محیط‌های آموزشی و شهری، فرصت‌های کمی برای ورزش روزانه وجود دارد و دانش‌آموزان بیش از پیش به رفتارهای کم‌تحرک مانند استفاده مداوم از تلویزیون، رایانه یا گوشی‌های هوشمند گرایش یافته‌اند.

۲-۳-۷- نقش محیط مدرسه و آموزش

مدرسه نه تنها محل آموزش رسمی است، بلکه فضایی برای یادگیری مهارت‌های زندگی، شکل‌گیری رفتارهای بهداشتی، و توسعه شخصیت اجتماعی نوجوانان نیز به‌شمار می‌آید. وجود محیط مدرسه‌ای امن، حامی و بدون تبعیض، در ارتقای سلامت روانی و اجتماعی دانش‌آموزان نقش بسزایی دارد. همچنین، آموزش‌های سلامت‌محور، برنامه‌های پیشگیری از اعتیاد، تغذیه سالم و مهارت‌های مقابله با استرس، بخشی از

[1] Endorphins

مسئولیت نهاد آموزش و پرورش در این زمینه است.

۲-۳-۸- تأثیر خانواده بر سلامت نوجوانان

خانواده نخستین و پایدارترین نهاد اجتماعی در زندگی هر فرد است. نگرش‌های والدین نسبت به سلامت، سبک فرزندپروری، وضعیت اقتصادی و فرهنگی خانواده، و الگوهای رفتاری والدین، همگی نقش مستقیم یا غیرمستقیمی در تعیین رفتارهای سلامتی نوجوان دارند. وجود تعامل مثبت والد-فرزند، نظارت آگاهانه، و حمایت عاطفی مستمر از جمله مهم‌ترین پیش‌بینی‌کننده‌های سلامت روانی پایدار در این سنین است.

۲-۳-۹- لزوم پایش علمی سلامت دانش‌آموزان

با توجه به تعدد عوامل مؤثر بر سلامت نوجوانان، ضرورت دارد این وضعیت به‌صورت منظم و علمی پایش شود. استفاده از پرسشنامه‌های استانداردی مانند کاسپین۵[1]، این امکان را فراهم می‌کند تا از طریق داده‌های قابل تحلیل، الگوهای رفتاری و وضعیت سلامت در سطح ملی یا منطقه‌ای ارزیابی شود. تحلیل این داده‌ها با ابزارهای نوینی چون داده‌کاوی[2] و شبکه‌های عصبی می‌تواند نقش مهمی در طراحی مداخلات سلامت‌محور و تصمیم‌گیری‌های آموزشی و بهداشتی ایفا کند.

۲-۴- مفاهیم پایه در داده کاوی

در دنیای امروز، حجم عظیمی از داده‌ها در حوزه‌های مختلف آموزشی، بهداشتی، اقتصادی و اجتماعی تولید و ذخیره می‌شود. صرف داشتن داده، به‌تنهایی ارزشی ندارد، مگر آن‌که بتوان از آن‌ها دانش و بینش استخراج کرد .داده‌کاوی، به‌عنوان یکی از شاخه‌های مهم علم داده[3]، فرآیند کشف الگوهای پنهان، روابط معنی‌دار و دانش مفید از میان داده‌های حجیم و پیچیده است. این علم، در سال‌های اخیر نقش بسزایی در تصمیم‌سازی‌های علمی، سیاست‌گذاری‌های کلان و حتی مداخلات پزشکی و سلامت ایفا کرده است.

۲-۴-۱- تعریف داده کاوی

داده‌کاوی به زبان ساده، به معنای استخراج اطلاعات باارزش از درون پایگاه‌های داده‌ی بزرگ است. این فرایند ترکیبی از تکنیک‌های آمار، یادگیری ماشین، پایگاه داده و هوش مصنوعی است که هدف آن پیش‌بینی، توصیف یا کشف دانش جدید از داده‌ها می‌باشد. بر اساس تعریف انجمن مدیریت داده‌ها، داده‌کاوی "فرآیند تحلیل داده‌ها از زوایای مختلف و استخراج اطلاعاتی است که می‌تواند در تصمیم‌گیری‌های هوشمندانه به کار رود."

---

[1] CaspianV
[2] Data Mining
[3] Data Science

۲-۴-۲- مراحل داده کاوی

فرآیند داده‌کاوی شامل چندین مرحله اصلی است که اغلب در قالب چرخه‌ای با عنوان چرخه‌ی کریسپ[۱] شناخته می‌شود:

۱. درک مسئله[۲]: تعریف اهداف کسب‌وکار یا پژوهش.
۲. درک داده‌ها[۳]: شناخت ویژگی‌ها و ساختار داده‌ها.
۳. پیش‌پردازش داده‌ها[۴]: پاک‌سازی، انتخاب ویژگی، نرمال‌سازی، و تبدیل داده‌ها.
٤. مدل‌سازی [۵]: انتخاب و اجرای الگوریتم‌های داده‌کاوی مانند طبقه‌بندی[۶]، خوشه‌بندی[۷] یا تحلیل انجمنی[۸]
۵. ارزیابی[۹]: سنجش عملکرد مدل‌ها با معیارهایی مانند دقت، حساسیت، F1-scoreو...
٦. پیاده‌سازی و تفسیر[۱۰]: استفاده عملی از مدل ساخته‌شده برای تصمیم‌گیری یا تحلیل.

۲-۴-۳- مهم‌ترین روش‌ها و الگوریتم‌ها در داده‌کاوی

بسته به نوع مسأله، داده‌کاوی از روش‌های گوناگونی بهره می‌برد. مهم‌ترین این روش‌ها عبارت‌اند از:

- طبقه‌بندی: مانند درخت تصمیم[۱۱]، جنگل تصادفی[۱۲]، شبکه عصبی[۱۳]، و ماشین بردار پشتیبان[۱۴] برای پیش‌بینی برچسب داده‌ها.
- خوشه‌بندی : مانند الگوریتم K-Means یا DBSCAN برای دسته‌بندی داده‌های بدون برچسب.
- تحلیل انجمنی[۱۵]: مانند الگوریتم Apriori برای کشف روابط هم‌زمان میان متغیرها (مثلاً قوانین اگر-آنگاه).
- تحلیل رگرسیون[۱۶]: برای پیش‌بینی متغیرهای عددی مانند سن یا شاخص توده بدنی.
- کاهش ابعاد[۱۷]: مانند PCA برای کاهش پیچیدگی داده‌های با ابعاد بالا.

---

[1] CRSIP-DM
[2] Problem Understanding
[3] Data Understanding
[4] Data preprocessing
[5] Modeling
[6] classification
[7] clustering
[8] association
[9] Evaluation
[10] Deployment
[11] Decision tree
[12] Random forest
[13] Neural network
[14] Support vector machines
[15] Association rules
[16] Regression Analysis
[17] Dimensionality Reduction

۲-۴-۴- داده کاوی در حوزه سلامت

یکی از حوزه‌های بسیار مهم و کاربردی داده‌کاوی، سلامت عمومی و بهداشت است. با افزایش دیجیتالی‌شدن داده‌های پزشکی و سلامت، امکان تحلیل‌های گسترده و دقیق برای شناسایی الگوهای بیماری، رفتارهای پرخطر، پیش‌بینی وضعیت بیماران و تخصیص بهینه منابع فراهم شده است.

کاربردهای داده‌کاوی در سلامت شامل موارد زیر است:

- پیش‌بینی ابتلا به بیماری‌ها هم از نوع واگیردار و غیرواگیردار (مانند دیابت یا افسردگی)
- تحلیل روندهای سلامت در جوامع
- کشف روابط میان سبک زندگی و وضعیت جسمی ، روانی
- طراحی مداخلات آموزشی یا بهداشتی برای گروه‌های هدف

۲-۴-۵- داده کاوی در تحلیل سلامت دانش‌آموزان

باتوجه به گستردگی و تنوع داده‌های به‌دست‌آمده از پرسشنامه‌هایی مانند کاسپین۵[1]، استفاده از داده‌کاوی می‌تواند نقش چشم‌گیری در شناسایی الگوهای رفتاری و پیش‌بینی وضعیت سلامت نوجوانان ایفا کند. برای مثال:

- تحلیل سبک زندگی (تغذیه، تحرک، اوقات فراغت) و ارتباط آن با سلامت روانی
- پیش‌بینی دانش‌آموزان در معرض خطر چاقی، اضطراب یا افت تحصیلی با استفاده از طبقه‌بندی‌کننده‌ها
- خوشه‌بندی الگوهای رفتاری برای طراحی مداخلات متناسب با هر گروه

استفاده از روش‌هایی مانند شبکه عصبی، امکان مدل‌سازی پیچیده‌تری را فراهم می‌کند که می‌تواند روابط غیرخطی و چندبعدی بین متغیرها را بهتر درک کند.

۲-۵- شبکه‌های عصبی مصنوعی[2]

شبکه‌های عصبی مصنوعی یکی از پرکاربردترین و مؤثرترین روش‌ها در حوزه‌ی یادگیری ماشین و داده‌کاوی هستند که از ساختار و عملکرد شبکه‌های عصبی بیولوژیکی در مغز انسان الهام گرفته‌اند. این مدل‌ها به‌ویژه در تحلیل داده‌های پیچیده، غیرخطی و چندبُعدی عملکرد بسیار مطلوبی دارند و در حوزه‌هایی مانند سلامت، تشخیص بیماری، روان‌شناسی و علوم رفتاری مورد توجه گسترده قرار گرفته‌اند.

یک شبکه عصبی مصنوعی از مجموعه‌ای از نورون‌های[3] مصنوعی (یا گره‌ها) تشکیل شده است که در

---

[1] CaspianV
[2] Artificial Neural Networks
[3] neuron

لایه‌هایی موسوم به لایه ورودی[1]، لایه‌های میانی یا پنهان[2] و لایه خروجی[3] سازماندهی شده‌اند. هر نرون با نرون‌های لایه قبل و بعد خود از طریق وزن‌ها[4] و توابع فعال‌سازی[5] ارتباط برقرار می‌کند.

شبکه‌های عصبی به‌ویژه در تحلیل الگوهای پنهان در داده‌های سلامت، مانند پیش‌بینی وضعیت روانی یا سبک زندگی افراد، مزیت‌هایی مهم دارند، از جمله:

- یادگیری غیرخطی: امکان شناسایی روابط پیچیده بین متغیرها که به‌راحتی با روش‌های آماری قابل استخراج نیستند.
- قابلیت تعمیم: شبکه پس از آموزش می‌تواند بر داده‌های جدید تعمیم یابد و عملکرد مناسبی داشته باشد.
- سازگاری با داده‌های پرنویز یا ناقص: به‌دلیل توانایی در یادگیری از حجم بالای داده‌ها، عملکرد نسبتاً مقاومی نسبت به خطاها دارد.

در پروژه حاضر، شبکه‌های عصبی برای مدل‌سازی ارتباط میان ویژگی‌های ثبت‌شده در پرسشنامه کاسپین۵[6] (نظیر الگوهای تغذیه، فعالیت بدنی، اوقات فراغت و وضعیت سلامت) استفاده شده‌اند تا الگوهای معنادار و پنهان بین آن‌ها کشف شود. به‌طور خاص، این روش می‌تواند به پیش‌بینی وضعیت سلامت یا شناسایی دسته‌هایی از دانش‌آموزان با ویژگی‌های مشترک کمک کند که در سیاست‌گذاری‌های آموزشی و بهداشتی قابل استفاده خواهد بود.

۲-۶- ابزارها و کتابخانه‌های مورد استفاده

در راستای پیاده‌سازی الگوریتم‌های داده‌کاوی در این پروژه، از زبان برنامه‌نویسی پایتون [7]و مجموعه‌ای از کتابخانه‌های تخصصی و رایج در حوزه‌ی تحلیل داده و یادگیری ماشین استفاده شده است. انتخاب این ابزارها بر اساس سادگی، کارایی، جامعه‌ی کاربری وسیع و سازگاری بالا با داده‌های سلامت صورت گرفته است. در ادامه، ابزارها و کتابخانه‌های مورد استفاده معرفی می‌شوند:

۲-۶-۱- زبان برنامه نویسی پایتون

پایتون یکی از محبوب‌ترین زبان‌های برنامه‌نویسی در حوزه علم داده، یادگیری ماشین و داده‌کاوی است. ساختار ساده، منابع آموزشی گسترده و پشتیبانی از کتابخانه‌های متنوع علمی، این زبان را به گزینه‌ای مناسب برای پروژه‌های تحلیلی در علوم پزشکی و آموزشی تبدیل کرده است.

---

[1] Input layer
[2] Hidden layer
[3] Output layer
[4] Weights
[5] Activation Functions
[6] CaspianV
[7] Python

۲-۶-۲- محیط توسعه

کدنویسی این پروژه در محیط پای‌چارم[1] انجام گرفته است که یکی از قدرتمندترین محیط‌های توسعه برای زبان پایتون است. همچنین از ژوپیتر نوب‌بوک[2] برای مستندسازی، تحلیل مرحله‌به‌مرحله داده‌ها و مصورسازی نتایج بهره گرفته شد که امکان اجرای تعاملی کدها را فراهم می‌کند.

۲-۶-۳- کتابخانه‌های مورد استفاده

- Pandas: کتابخانه‌ای قدرتمند برای مدیریت داده‌های جدولی[3] که در مراحل پیش‌پردازش، پاک‌سازی و تحلیل اولیه داده‌ها مورد استفاده قرار گرفت.
- Numpy: برای انجام محاسبات عددی، عملیات برداری و ماتریسی در تحلیل داده‌ها استفاده شد.
- Scikit-learn: یکی از کتابخانه‌های اصلی و پرکاربرد در یادگیری ماشین که برای اعمال الگوریتم‌هایی مانند طبقه‌بندی[4]، خوشه‌بندی[5]، و همچنین ارزیابی عملکرد مدل‌ها به کار گرفته شد.
- Matplotlib: جهت مصورسازی داده‌ها، ترسیم نمودارهای میله‌ای، پراکندگی، هیستوگرام[6] و تجسم نتایج مدل‌های آماری و داده‌کاوی مورد استفاده قرار گرفت.
- re: در پایتون برای کار با عبارات منظم استفاده می‌شود و امکان جستجو، تطبیق، جایگزینی و استخراج الگوهای خاص در رشته‌ها را فراهم می‌کند..

۲-۷- جمع‌بندی

در این فصل، مبانی نظری و مفهومی مرتبط با پژوهش حاضر مورد بررسی قرار گرفت. ابتدا طرح کشوری کاسپین۵[7] به‌عنوان چارچوب اصلی گردآوری داده‌های این مطالعه معرفی شد. این طرح با تمرکز بر ابعاد مختلف سلامت جسمی، روانی، رفتاری و اجتماعی دانش‌آموزان، بستری مناسب برای تحلیل داده‌محور و کشف روابط میان مؤلفه‌های سبک زندگی فراهم می‌آورد.

در ادامه، مفاهیم کلیدی مرتبط با سلامت دانش‌آموزان تبیین شد. سلامت به‌عنوان مفهومی چندبعدی، دربرگیرنده جنبه‌هایی چون تغذیه، فعالیت بدنی، وضعیت روانی و رفتارهای اجتماعی است که شناسایی و

---

[1] PyCharm
[2] Jupyter Notebook
[3] dataframe
[4] Classification
[5] Clustering
[6] Histogram
[7] CaspianV

تحلیل آن‌ها به‌ویژه در سنین رشد از اهمیت بسزایی برخوردار است. مرور این مفاهیم به روشن شدن بستر نظری پژوهش و هدف‌گذاری دقیق در تحلیل داده‌ها کمک نموده است.

در بخش بعد، اصول و روش‌های داده‌کاوی به‌عنوان ابزار اصلی تحلیل داده‌ها در پروژه تشریح شد. داده‌کاوی با بهره‌گیری از الگوریتم‌های یادگیری ماشین، امکان شناسایی الگوها و ارتباطات پنهان در مجموعه داده‌های گسترده را فراهم می‌سازد. مفاهیمی نظیر طبقه‌بندی، خوشه‌بندی، پیش‌پردازش داده‌ها و ارزیابی مدل‌ها به‌تفصیل مورد بررسی قرار گرفت و کارکردهای آن‌ها در حوزه سلامت و علوم رفتاری تشریح گردید.

همچنین، شبکه‌های عصبی مصنوعی به‌عنوان یکی از تکنیک‌های پیشرفته تحلیل غیرخطی داده‌ها معرفی و به‌عنوان ابزار اصلی مدل‌سازی در این پژوهش مورد بررسی قرار گرفت. توانایی این الگوریتم‌ها در تحلیل روابط پیچیده میان متغیرها، نقش بسزایی در استخراج دانش معتبر از داده‌های سلامت ایفا می‌کند.

در پایان، ابزارها و کتابخانه‌های مورد استفاده در پیاده‌سازی الگوریتم‌ها معرفی شدند. بهره‌گیری از زبان برنامه‌نویسی پایتون[1] و کتابخانه‌هایی نظیر Pandas ، NumPy، Scikit-learn وMatplotlib ، چارچوب فنی مورد نیاز برای اجرای دقیق و علمی الگوریتم‌ها را فراهم نموده است. استفاده از محیط‌های توسعه‌ای همچون PyCharm و Jupyter Notebook نیز به تسهیل فرآیند تحلیل، مستندسازی و تفسیر نتایج کمک شایانی کرده است.

برآیند مطالب ارائه‌شده در این فصل، تأمین مبانی نظری، فنی و مفهومی پژوهش بوده و زمینه‌ساز ورود به فصل بعدی با عنوان «شرح پروژه» است که در آن، مراحل اجرایی، پردازش داده‌ها و نحوه پیاده‌سازی مدل‌ها به تفصیل شرح داده خواهد شد.

[1] Python

فصل ۳
شرح پروژه

۳-۱- مقدمه

در ادامه‌ی بررسی مبانی نظری و مفهومی پژوهش، فصل سوم به تشریح دقیق روش انجام تحقیق اختصاص دارد. این فصل به‌منظور تبیین نحوه‌ی اجرای عملی پروژه، روش گردآوری و پردازش داده‌ها، انتخاب ابزارها، الگوریتم‌ها و مراحل پیاده‌سازی مدل‌های داده‌کاوی نگارش شده است.

پژوهش حاضر با هدف شناسایی الگوهای پنهان میان ابعاد مختلف سلامت دانش‌آموزان، از طریق تحلیل داده‌های حاصل از طرح ملی کاسپین۵ [1]، انجام شده است. این طرح شامل داده‌هایی پیرامون سبک زندگی، تغذیه، فعالیت بدنی، وضعیت روانی و جسمانی دانش‌آموزان سراسر کشور است. با بهره‌گیری از الگوریتم‌های داده‌کاوی، به‌ویژه شبکه‌های عصبی مصنوعی، تلاش شده است تا ارتباطات پنهان میان این متغیرها کشف شده و به مدلی جهت تحلیل و پیش‌بینی وضعیت سلامت دانش‌آموزان دست یابیم.

در این فصل ابتدا به معرفی جامعه آماری و نحوه گردآوری داده‌ها پرداخته می‌شود. سپس فرایند پیش‌پردازش داده‌ها از جمله پاک‌سازی، نرمال‌سازی، تبدیل متغیرها و انتخاب ویژگی‌ها شرح داده خواهد شد. در ادامه، الگوریتم‌های مورد استفاده در تحلیل داده‌ها، ابزارهای پیاده‌سازی، معیارهای ارزیابی مدل‌ها و روند آموزش و تست مدل‌ها معرفی می‌گردند.

این فصل نقش بسیار مهمی در اعتبار علمی پژوهش ایفا می‌کند، چرا که نشان می‌دهد یافته‌های پروژه بر پایه روشی ساختاریافته، قابل تکرار و منسجم به‌دست آمده‌اند.

۳-۲- نوع و روش تحقیق

پژوهش حاضر از نوع کاربردی بوده و با رویکرد تحلیل داده‌های ثانویه و بهره‌گیری از روش‌های داده‌کاوی و یادگیری ماشین انجام شده است. هدف اصلی این تحقیق، کشف الگوها و روابط پنهان میان

---

[1] Caspain V

مؤلفه‌های مختلف سلامت دانش‌آموزان با استفاده از داده‌های جمع‌آوری‌شده در طرح ملی کاسپین۵ می‌باشد.[1]

روش تحقیق حاضر مبتنی بر رویکرد توصیفی تحلیلی است؛ بدین معنا که ضمن توصیف داده‌های موجود، تلاش شده با تحلیل عمیق و مدل‌سازی، الگوهای معنی‌داری استخراج گردد که بتوانند در آینده برای تصمیم‌سازی‌های بهداشتی و آموزشی مورد استفاده قرار گیرند.

۳-۳- جامعه آماری و منابع داده‌ها

جامعه آماری این پژوهش را کلیه دانش‌آموزان شرکت‌کننده در طرح کشوری «کاسپین ۵» تشکیل می‌دهند. این طرح که با عنوان کامل «مطالعه‌ی عوامل خطر بیماری‌های غیرواگیر در کودکان و نوجوانان ایرانی» توسط وزارت بهداشت، درمان و آموزش پزشکی ایران با همکاری وزارت آموزش‌وپرورش اجرا شده، یکی از گسترده‌ترین طرح‌های ملی در حوزه سلامت نوجوانان در کشور به‌شمار می‌رود.

این پیمایش در طی سال‌های ۱۳۹۴-۱۳۹۳ در مناطق شهری و روستایی ۳۱ استان اجرا شده است. جامعه آماری مورد بررسی شامل تمامی دانش‌آموزان شاغل به تحصیل در مقاطع دوگانه ابتدایی و متوسطه در مناطق شهری و روستایی استان‌های سراسر کشور است. این مطالعه در ۳۱ استان کشور و بر روی دانش‌آموزان ۷ تا ۱۸ سال انجام شد. واحد آماری مورد بررسی شامل دانش‌آموز شاغل به تحصیل در زمان انجام پرسشگری است. دانش‌آموزان ایرانی ساکن در مناطق شهری و روستایی استان‌های سراسر کشور به شرط داشتن شناسنامه ایرانی وارد مطالعه شدند و دانش‌آموزان خارجی و تابعه سایر کشورها حتی در صورت داشتن مجوز اقامت و یا گویش به زبان فارسی وارد مطالعه نشدند. نمونه‌گیری در سطح دانش‌آموزان از نوع چندمرحله‌ای با استفاده از روش‌های نمونه‌گیری خوشه‌ای و طبقه‌ای بود. نمونه‌گیری طبقه‌ای در داخل هر استان و برحسب محل سکونت دانش‌آموز (شهر و روستا) و مقطع تحصیلی (ابتدایی، راهنمایی و دبیرستان) به شیوه متناسب با اندازه[2] با نسبت جنسی برابر انجام شد. نحوه رسیدن به نمونه موردنظر و انتخاب آن‌ها با استفاده از نمونه‌گیری خوشه‌ای در سطح هر استان و با اندازه خوشه‌های برابر صورت گرفت. خوشه‌ها در سطح مدارس تعیین گردید. اندازه هر خوشه ۱۰ نفر برآورد شد یعنی باید در هر خوشه تعداد ۱۰ واحد آماری (شامل ۱۰ دانش‌آموز و والدین او) قرار بگیرد. حجم نمونه محاسبه‌شده برای این مطالعه ۴۸۰ نفر در هر استان بود؛ یعنی ۴۸ خوشه ۱۰ نفری در هر یک از استان‌های کشور و در مجموع با توجه به انجام مطالعه در ۳۱ استان ۱۴۸۸۰ نفر مورد بررسی قرار گرفتند که نهایتاً داده‌های مربوط به ۱۴۲۷۴ نفر مورد تجزیه‌وتحلیل آماری قرار گرفت.[۱۳]

---

[1] Caspian V
[2] Proportional to size

طرح کاسپین ۵[1] به‌منظور بررسی الگوی زندگی، رفتارهای تغذیه‌ای، فعالیت‌های بدنی، وضعیت سلامت روان و فیزیکی، و شیوع عوامل خطر در بین نوجوانان ایرانی طراحی شده است. گردآوری داده‌ها در این طرح از طریق پرسشنامه‌های ساخت‌یافته و استاندارد صورت گرفته که دانش‌آموزان در بازه سنی مشخص، به‌صورت خوداظهاری یا با کمک والدین و مسئولان به آن پاسخ داده‌اند.

داده‌های مورد استفاده در این پژوهش شامل پاسخ‌های ۲۳ سؤال کلیدی در پرسشنامه کاسپین ۵ است که ابعاد مختلف سبک زندگی دانش‌آموزان از جمله:

- تغذیه (نظیر مصرف میوه، نوشیدنی‌های قندی، فست‌فود، لبنیات، تنقلات)
- فعالیت بدنی (مانند دفعات ورزش هفتگی یا نوع وسیله‌ی حمل‌ونقل تا مدرسه)
- سلامت عمومی (مانند خواب، اضطراب، دردهای بدنی و روان‌تنی)
- اوقات فراغت و استفاده از رسانه‌ها (مانند زمان استفاده از اینترنت و تلویزیون، زمان انجام تکالیف و مدت زمان خواب روزانه)

را پوشش می‌دهد.

این داده‌ها به‌صورت جدولی و ساخت‌یافته در قالب فایل‌های دیجیتال (معمولاً با فرمت اکسل[2] یا CSV) قابل دسترسی و تحلیل هستند. منبع اصلی داده‌ها نسخه‌ی نمونه ارائه‌شده (توسط پژوهشکده پیشگیری اولیه از بیماری‌های غیرواگیر) می‌باشد. در این پژوهش، از نمونه‌ای استخراج‌شده از این داده‌ها استفاده شده که نماینده‌ای از جامعه‌ی بزرگ‌تر دانش‌آموزی در ایران محسوب می‌شود.

به علت گستردگی جغرافیایی طرح و طراحی نمونه‌گیری خوشه‌ای چندمرحله‌ای آن، داده‌ها از نظر تنوع جنسیتی، سنی، و منطقه‌ای (شهری و روستایی) از پوشش مناسبی برخوردارند. این ویژگی، اعتبار نتایج و قابلیت تعمیم آن‌ها به جامعه‌ی وسیع‌تری از دانش‌آموزان کشور را افزایش می‌دهد.

۳-۴- شرح مجموعه‌ی داده

مجموعه داده‌ی مورد استفاده در این پژوهش با عنوان سمپل[3] دراختیار قرار گرفت. این مجموعه شامل اطلاعات مربوط به تقریباً ۲۰ خوشه‌ی ۱۰ نفری از دانش‌آموزان دختر و پسر شهر تبریز می‌باشد. افراد شرکت‌کننده در این نمونه، متولدین سال‌های ۱۳۷۹ تا ۱۳۸۷ هستند و بازه‌ی سنی مناسبی از دانش‌آموزان مقاطع مختلف تحصیلی را پوشش می‌دهند.

ساختار این مجموعه داده شامل ۲۰۰ ردیف (نمونه) و ۶۷ ستون (ویژگی) است. ردیف‌ها نمایانگر نمونه‌های فردی (دانش‌آموزان) و ستون‌ها بیانگر ویژگی‌ها و متغیرهای مورد بررسی در پرسشنامه هستند.

---

[1] Caspian V
[2] Excel
[3] sample

به‌طور کلی، ستون‌های ابتدایی شامل اطلاعات کلی و ویژگی‌های جمعیت‌شناختی مشترک میان دانش‌آموزان (مانند سن، جنسیت، محل سکونت و ...) هستند و در ادامه، پاسخ‌های هر دانش‌آموز به پرسش‌های مطرح‌شده در پرسشنامه، در ستون‌های اختصاصی مربوط به همان پرسش ثبت شده است.

پرسش‌های مطرح‌شده در پرسشنامه کاسپین ۵[1] در قالب چهار محور اصلی دسته‌بندی شده‌اند که هر یک ابعاد متفاوتی از سبک زندگی و وضعیت سلامت دانش‌آموزان را مورد بررسی قرار می‌دهند:

۱. تغذیه: بخش نخست پرسشنامه به موضوع تغذیه اختصاص دارد. این قسمت با پرسش‌هایی در خصوص زمان صرف وعده‌های غذایی آغاز می‌شود و در ادامه به بررسی الگوی مصرف انواع خوراکی‌ها از جمله تنقلات، نوشیدنی‌های طبیعی و مصنوعی، لبنیات، میوه و سبزیجات، و غذاهای آماده (فست‌فود) می‌پردازد. هدف این بخش، سنجش کیفیت و عادات غذایی دانش‌آموزان و بررسی میزان انطباق آن با الگوهای تغذیه‌ی سالم است.

۲. فعالیت بدنی: در این بخش، میزان فعالیت بدنی دانش‌آموزان مورد ارزیابی قرار می‌گیرد. پرسش‌ها بر محور دفعات و مدت زمان انجام فعالیت‌های فیزیکی در طول هفته و در بازه‌های زمانی مدرسه و خارج از مدرسه تمرکز دارند. همچنین، فعالیت‌هایی همچون پیاده‌روی، ورزش‌های سازمان‌یافته یا بازی‌های فیزیکی نیز در نظر گرفته شده‌اند. این اطلاعات برای تحلیل سبک زندگی فعال یا کم‌تحرک دانش‌آموزان کاربرد دارد.

۳. فعالیت اوقات فراغت: این محور به بررسی نحوه گذران اوقات فراغت دانش‌آموزان اختصاص دارد. سؤالات این بخش میزان استفاده از تلویزیون، رایانه، فضای مجازی، بازی‌های ویدیویی و سایر رسانه‌ها را مورد پرسش قرار می‌دهند. علاوه بر آن، مدت‌زمان اختصاص‌یافته به انجام تکالیف درسی نیز به عنوان یکی از فعالیت‌های عمده در اوقات غیرآموزشی در این بخش گنجانده شده است. این اطلاعات به درک تأثیر استفاده از فناوری بر رفتار و زمان‌بندی روزانه‌ی دانش‌آموزان کمک می‌کند.

٤. آخرین بخش پرسشنامه به ابعاد گوناگون سلامت فردی و روانی دانش‌آموزان می‌پردازد. در این قسمت، سؤالاتی درباره‌ی تجربه‌ی علائم روان‌تنی مانند اضطراب، سردرد، مشکلات خواب، بی‌حوصلگی یا ناراحتی‌های جسمانی مکرر مطرح شده است. هدف این بخش، شناسایی نشانگان اولیه‌ی مشکلات جسمی و روانی در جمعیت نوجوانان است.

پاسخ‌دهی به سؤالات در قالب طیف‌های کیفی (مانند: هرگز، به‌ندرت، گاهی، بیشتر اوقات، همیشه)، یا

[1] Caspain V

ارزش‌های کمی عددی در بازه‌ی ۱ تا ۴ انجام گرفته است. برخی از سؤالات نیز بر اساس مقیاس‌های زمانی (مانند ساعت یا روز در هفته) یا به صورت داده‌های دودویی (بله/خیر) طراحی شده‌اند.

در انتهای پرسشنامه، از دانش‌آموزان خواسته شده است تا برآورد کلی خود از وضعیت سلامت‌شان را در قالب عددی بین ۱ تا ۱۰ ثبت کنند. این خودارزیابی به عنوان یک شاخص ذهنی مهم در تحلیل‌های نهایی پژوهش مورد توجه قرار گرفته است، زیرا احساس فردی نسبت به سلامت می‌تواند مکمل داده‌های عینی ثبت‌شده باشد.

جدول ۳-۱ نمونه کدگذاری پاسخ‌ها در هر بخش پرسشنامه

بخش پرسشنامه	نوع پاسخ‌ها	مثالی از پاسخ‌ها	کدگذاری
تغذیه	کیفی/ عددی/ زمانی/دودویی	هرگز، بندرت، گاهی، بیشتر اوقات، همیشه	۰،۱،۲،۳،۴
فعالیت‌بدنی	عددی/ کیفی	اعشاری در واحد زمان	۷.۳ تبدیل به دقیقه ۴۵۰
اوقات فراغت	عددی/ زمانی	ساعت در روز	۰،۱،۲،۳،۴
سلامت و ناخوش‌ها	کیفی/دودویی	بله/خیر	بله:۱ و خیر:۰
ارزیابی کلی سلامت	عددی	امتیاز ۱ تا ۱۰	مقدار عددی

۳-۵- پیش‌پردازش داده‌ها

پیش‌پردازش داده‌ها یکی از مراحل اساسی و حیاتی در فرآیند تحلیل داده و به‌ویژه در پیاده‌سازی مدل‌های یادگیری ماشین و هوش مصنوعی است. داده‌های خام معمولاً شامل خطاها، مقادیر گمشده، ناهماهنگی‌ها و انواع نویزهایی هستند که می‌توانند دقت و کیفیت نتایج مدل را به‌شدت کاهش دهند. بنابراین، قبل از تحلیل نهایی یا آموزش مدل‌ها، لازم است داده‌ها به‌صورت کامل و استاندارد آماده‌سازی شوند.

پروژه‌ی حاضر با تکیه بر رویکرد سیستماتیک چرخه‌ی کریسپ[1] طراحی و اجرا شده است؛ الگویی ساختارمند و پذیرفته‌شده در تحلیل داده‌ها که پژوهشگر را از درک مسئله تا ارزیابی نهایی مدل هدایت می‌کند. در این پروژه، که تمرکز آن بر تحلیل داده‌های روان‌سلامت و سبک زندگی دانش‌آموزان بر اساس

[1] CRISP-DM

پرسشنامه‌ی کاسپین ۵[1] است، هر مرحله از این چرخه با دقت و حساسیت ویژه‌ای دنبال شده است تا ارتباطات پنهان میان متغیرهای رفتاری، جسمی و ذهنی دانش‌آموزان نمایان گردد[1].

نخستین گام، مواجهه با داده‌هایی بود که در قالب جدول‌هایی نسبتاً بزرگ، حاصل تلاش جمعی از متخصصان حوزه‌ی سلامت، روان‌شناسی و تغذیه بودند. این داده‌ها از طریق پاسخ مستقیم دانش‌آموزان به پرسش‌هایی درباره‌ی تغذیه، فعالیت بدنی، اوقات فراغت و وضعیت روحی ،جسمی جمع‌آوری شده بودند. اما این داده‌های به‌ظاهر ساختاریافته، در نخستین نگاه، چالش‌هایی بنیادین را با خود به همراه داشتند.

از یک‌سو، تنوع ساختار پاسخ‌ها که گاه کیفی (مانند "گاهی"، "بیشتر اوقات") و گاه کمی یا دودویی بودند، پژوهشگر را با معضل استانداردسازی[2] روبه‌رو می‌کرد. از سوی دیگر، عدم توازن در بعضی پاسخ‌ها، مقادیر گمشده، یا ابهام در نوع مقیاس برخی سوالات، نیازمند تفسیر دقیق و تصمیم‌گیری‌های فنی و مفهومی بود.

۳-۵-۱- چرخه‌ی اول

در نخستین گام از فرایند آماده‌سازی داده‌ها، نمونه‌هایی که به لحاظ اطلاعاتی تهی یا ناکامل[3] بودند یعنی دانش‌آموزانی که پاسخ‌های ثبت‌شده‌ی آن‌ها بسیار محدود، ناقص یا نامعتبر بود از مجموعه‌ی داده حذف گردیدند تا کیفیت و یکپارچگی تحلیل حفظ شود. این تصمیم به‌منظور جلوگیری از ایجاد انحراف در نتایج مدل و افزایش دقت پردازش اتخاذ شد.

در ادامه، آن دسته از ستون‌هایی که در تمامی نمونه‌ها دارای مقادیر کاملاً یکسان بودند نیز از داده‌ها حذف شدند. از جمله این ستون‌ها می‌توان به متغیرهایی همچون "university " و "region " اشاره کرد که به دلیل یکنواختی در مقدار و عدم ارائه‌ی اطلاعات متغیر و تمایزآفرین، ارزش تحلیلی خاصی نداشتند. حذف این ستون‌ها به کاهش پیچیدگی مدل و تمرکز بیشتر بر متغیرهای مؤثر کمک شایانی نمود.

```
# Identify columns with the same value in all rows
columns_to_drop = [col for col in data_copy.columns if data_copy[col].nunique() == 1]
# Drop these columns
data_cleaned = data_copy.drop(columns= columns_to_drop)
```

شکل ۳-۱حذف ویژگی‌ها با مقادیر یکسان

در گام بعدی از فرایند پیش‌پردازش، تمرکز بر تبدیل مقادیر عددی اعشاری به فرمت زمانی معنادار قرار گرفت. در برخی از ستون‌ها، زمان صرف وعده‌های غذایی یا انجام فعالیت‌های خاص به‌صورت عددی اعشاری

[1] Caspian V
[2] Standardization
[3] Sparse data

ثبت شده بود؛ برای نمونه، مقدار ۹.۳ در واقع نشان‌دهنده‌ی ساعت ۹:۳۰ صبح است.

برای تبدیل این مقادیر به فرمت صحیح زمانی، ابتدا بخش صحیح عدد به‌عنوان ساعت در نظر گرفته شد. سپس، با کسر این بخش صحیح از مقدار اصلی، بخش اعشاری که نمایانگر دقیقه است به دست آمد. این مقدار اعشاری در عدد ۱۰۰ ضرب شد تا دقیقه‌ی دقیق محاسبه شود. در نهایت، ساعت و دقیقه‌ی استخراج‌شده در قالب یک ساختار زمانی استاندارد ترکیب گردید تا داده‌ها از نظر زمانی نیز تفسیرپذیر و قابل استفاده در تحلیل‌ها باشند.

```
def convert_float_to_time(time):
    new_time = time
    if pd.isnull(time) == False:
        hour = math.floor(time)
        minute = round((time - hour) * 100)
        new_time = f"{hour:02d}:{minute:02d}"
    return new_time
```

شکل ۳-۲تبدیل فرمت اعشاری به فرمت زمانی

از آنجا که تعداد ستون‌های زیادی در این مجموعه‌ی داده نیاز به تبدیل به فرمت زمانی دارند تابعی جهت گرفتن نام ستون و تبدیل تمام مقادیر آنها به زمان ساخته شد.

```
def repair_time(question_num):
    for i in range(M):
        new_time1 = convert_float_to_time(data_copy[question_num][i])
        data_copy.at[i , question_num] = new_time1
    data_cleaned[question_num] = pd.to_datetime(data_copy[question_num] , format = '%H:%M')
```

شکل ۳-۳تابع جهت گرفتن ویژگی و تبدیل به فرمت زمانی

در ادامه‌ی فرایند پیش‌پردازش داده‌ها، برای مواجهه با مقادیر ناشناخته[1] در ستون‌های مرتبط با زمان، از رویکردی مبتنی بر میانگین‌گیری وزنی بهره گرفته شد. بدین منظور، ابتدا میانگین زمان ثبت‌شده در هر ستون محاسبه گردید تا بتوان مقدار جایگزینی منطقی و نزدیک به واقعیت برای داده‌های مفقود پیشنهاد داد.سپس، به‌منظور افزایش دقت و خوانایی داده‌ها، زمان به‌دست‌آمده با روشی خاص گرد شد: اگر مقدار دقایق بیش از ۴۵ دقیقه بود، زمان به ساعت بعدی گرد گردید؛ اگر کمتر از ۱۵ دقیقه بود، به ساعت قبل بازگردانده شد؛ و در صورتی که مقدار دقیقه بین ۱۵ تا ۴۵ قرار داشت، زمان به‌صورت متعارف به نیم‌ساعت تنظیم گردید. این شیوه نه‌تنها داده‌های مفقود را با مقادیری معنادار جایگزین می‌کند، بلکه هم‌راستایی و یکنواختی زمانی داده‌ها را نیز حفظ می‌نماید.

[1] Null

بدین ترتیب تمام ستون‌های بدست‌آمده در فرمت زمان و بدون داده‌های پوچ[1] می‌باشند.

```
def average_of_time(question_num):
    avg = data_cleaned[question_num].mean()
    hour = avg.hour
    minute = avg.minute
    if minute > 45:
        minute = 0
        hour = hour +1
    elif minute < 15:
        minute = 0
    else:
        minute = 30

    avg = pd.to_datetime(f"{hour:02d}:{minute:02d}" , format = '%H:%M')
    return avg
```

شکل ۳-۴جایگذاری داده‌های زمانی مفقود با روش میانگین

در گام بعدی از فرایند پاک‌سازی و آماده‌سازی داده‌ها، به اصلاح ستون‌هایی پرداخته شد که به‌رغم ماهیت عددی، به‌صورت رشته‌هایی شامل یک عدد به‌همراه یک پسوند متنی ذخیره شده بودند. این ساختار نادرست می‌تواند مانع تحلیل دقیق و عددی داده‌ها شود و در فرآیند مدل‌سازی اختلال ایجاد کند.

برای حل این مسئله، تابعی با عنوان "repair_suffix_number" طراحی و پیاده‌سازی شد. این تابع به‌گونه‌ای توسعه یافته که علاوه‌بر شماره‌ی پرسش (یا ستون مورد نظر)، دو پارامتر کلیدی دیگر را نیز دریافت می‌کند:

- suffix : رشته‌ای است که نشان‌دهنده‌ی پسوند چسبیده به عدد می‌باشد.
- substitute: رشته‌ای جایگزین که فاقد ارزش عددی بوده و معادل مقدار صفر در نظر گرفته می‌شود.

این تابع با شناسایی مقادیر دارای پسوند و جداسازی بخش عددی از آن‌ها، زمینه‌ی تبدیل این داده‌ها به مقادیر عددی خالص و قابل تحلیل را فراهم می‌آورد. بدین ترتیب، یکدستی ساختار داده‌ها حفظ شده و

```
def repair_suffix_number(question_num , suffix , substitute):
    data_cleaned[question_num] = data_copy[question_num].str.replace(substitute , f'{0} {suffix}')
    for i in range(len(data_cleaned[question_num])):
        if pd.isna(data_cleaned[question_num][i]) == False:
            data_cleaned[question_num][i] =int((data_cleaned[question_num][i])[0])
Executed at 2024.08.30 19:52:24 in 6ms

repair_suffix_number('h_20' , 'roz' ,'aslan nemikhoram')
```

شکل ۳-۵ تابع تشخیص پسوند و جداکننده‌ی عدد

[1] null

کیفیت تحلیل در مراحل بعدی تضمین می‌شو

در گام بعدی پرسش‌هایی که فرمت دودویی دارند را شناسایی کرده و با دستور زیر پاسخ پرسش‌ها را به صورت درست یا غلط[1] تبدیل شد.

```
data_cleaned['h_21'] = data_cleaned['h_21'].replace({'nemikhoram' : False , 'mikhoram' : True})
Executed at 2024.08.30 19:52:28 in 9ms
```

شکل ۳-۶ تابع تغییر به فرمت دودویی

در گام بعدی، توجه خود را معطوف به آن دسته از پرسش‌های کیفی موجود در مجموعه‌ی داده کردیم که نوع پاسخ‌دهی آن‌ها بر پایه‌ی واژگان توصیفی همچون:"never , rarely , weekly, daily" صورت گرفته بود. از آنجا که تحلیل داده‌های کیفی در قالب مدل‌های عددی نیازمند تبدیل این پاسخ‌ها به مقادیر کمی است، فرآیند نگاشت[2] مفهومی به عددی انجام گرفت.

```
def ordinary_encoding(question_num):
    fill_nan(question_num)
    #data_cleaned[question_num] = data_cleaned[question_num].apply(str)
    categories = [['never', 'rarely', 'Daily', 'weekly']]
    encoder = preprocessing.OrdinalEncoder(categories= categories)
    data = data_cleaned[question_num].values.reshape(-1 , 1)
    data_cleaned[question_num] =  encoder.fit_transform(data)
```

شکل ۳-۷ تابع نگاشت متغیرهای کیفی به مقادیر عددی

در این نگاشت، مقادیر کیفی مذکور به ترتیب به اعداد صحیح بین ۰ تا ۳ تخصیص داده شدند؛ به‌طوری‌که، کمترین میزان بروز رفتار یا عادت (مانند "هرگز") با عدد صفر و بیشترین آن (مانند "روزانه") با عدد ۳ نمایش داده شد. این تبدیل نه‌تنها امکان تحلیل ریاضی و مدل‌سازی دقیق‌تر را فراهم ساخت، بلکه انسجام داده‌ها را نیز افزایش داد.

همچنین، برای تکمیل داده‌های گمشده[3] در این نوع پرسش‌ها، از روش جایگزینی با پرتکرارترین[4] مقدار بهره گرفتیم. بدین معنا که مقدار تکرارشونده‌ی غالب در هر ستون به‌عنوان نماینده‌ی مناسب برای داده‌های مفقود جایگزین گردید. این رویکرد، ضمن حفظ ساختار طبیعی داده‌ها، به کاهش سوگیری در نتایج نهایی کمک شایانی نمود.

در برخی از پرسش‌های موجود در مجموعه‌ی داده، پاسخ‌ها به‌صورت رشته‌هایی مرکب از اعداد و کاراکترها ثبت شده بودند؛ رشته‌هایی که در واقع نشان‌دهنده‌ی بازه‌ای از مقادیر عددی بودند. این ساختار غیر

---

[1] True or false
[2] mapping
[3] Null
[4] mode imputation

استاندارد مانع تحلیل مستقیم و کمی داده‌ها می‌شد و نیاز به پردازش دقیق‌تری داشت.

برای تبدیل این نوع پاسخ‌ها به مقادیری قابل تحلیل، از کتابخانه‌ی قدرتمند re[1] بهره گرفتیم. این ابزار امکان شناسایی و استخراج الگوهای عددی از درون رشته‌ها را فراهم می‌آورد. با استفاده از عبارات منظم، اعداد موجود در هر پاسخ شناسایی شدند و سپس میانگین بازه‌ی عددی (یعنی مقدار میانی بین دو عدد ثبت‌شده) به‌عنوان نماینده‌ی کمی آن پاسخ محاسبه و جایگزین گردید.

```
import re
def extract_number_from_string_to_arr(question_number):
    fill_nan(question_number)
    for i in range(M):
        data_cleaned[question_number][i] = [int(s) for s in re.findall(r'\d+' , data_cleaned[question_number][i])]
```

شکل ۳-۸ تابع استخراج اعداد از میان یک رشته

این شیوه‌ی پردازش، داده‌های بازه‌ای را به‌گونه‌ای معنادار و قابل استفاده در مدل‌های هوش مصنوعی تبدیل کرد و در عین حال دقت تحلیل نهایی را نیز حفظ نمود.

در گام پایانی پیش‌پردازش داده‌ها، تمرکز بر آماده‌سازی برخی از ویژگی‌های کیفی برای تحلیل عددی قرار گرفت. بدین منظور، سه ویژگی کلیدی شامل جنسیت، مقطع تحصیلی و وسیله‌ی رفت‌وآمد به مدرسه با بهره‌گیری از روش کدگذاری وان‌هات[2] به دسته‌هایی مجزا و عددی تبدیل شدند. این روش با تبدیل هر دسته به یک ستون مجزا، امکان تحلیل دقیق‌تر داده‌های کیفی در مدل‌های آماری و یادگیری ماشین را فراهم می‌سازد.

در ادامه، به‌منظور کشف روابط و الگوهای موجود میان متغیرهای مختلف، از ضریب همبستگی پیرسون[3] استفاده شد. این ضریب معیاری آماری برای سنجش میزان و جهت رابطه‌ی خطی بین دو متغیر عددی است. نتایج حاصل از این تحلیل در فصل مربوط به نتایج و تحلیل یافته‌ها به تفصیل ارائه خواهد شد.

از آنجا که ویژگی‌های زمانی در قالب فرمت "time" قابل استفاده مستقیم در محاسبات همبستگی نبودند، تمامی مقادیر زمانی به واحد دقیقه تبدیل شدند. این تبدیل به ما اجازه داد تا بتوانیم روابط میان متغیرهای زمانی و سایر ویژگی‌ها را نیز به‌طور دقیق بررسی نماییم و تحلیل‌های آماری را بر پایه‌ی داده‌هایی یکپارچه و کمی انجام دهیم.

۳-۵-۲- چرخه‌ی دوم

پس از بررسی و تحلیل نتایج حاصل از اولین چرخه‌ی پردازش و مدل‌سازی، تصمیم گرفتیم بار دیگر به

---

[1] Regular expression
[2] One-Hot Encoding
[3] Pearson Correlation Coefficient

مرحله‌ی پیش‌پردازش داده‌ها بازگردیم؛ چراکه تحلیل‌های اولیه نشان دادند که هنوز ظرفیت‌هایی برای بهبود کیفیت داده و افزایش دقت مدل وجود دارد.

در این بازبینی، با مطالعه‌ی دقیق‌تر بر روی ساختار داده‌ها، به این نتیجه رسیدیم که می‌توان برخی از ویژگی‌ها را با تغییر در فرمت یا بازتعریف نحوه‌ی نمایش آن‌ها، معنادارتر و کاربردی‌تر نمود. در همین راستا، گروهی از ویژگی‌ها که ماهیتی مشابه داشتند یا به لحاظ مفهومی مکمل یکدیگر بودند، با هدف افزایش انسجام و کاهش پراکندگی اطلاعات با یکدیگر ادغام شدند.

همچنین، در این مرحله وزن‌دهی برخی از ویژگی‌ها بازنگری شد تا تأثیر آن‌ها در فرآیند مدل‌سازی متناسب با اهمیت واقعی‌شان لحاظ گردد. در کنار این اصلاحات، آن دسته از ویژگی‌هایی که تحلیل‌ها نشان دادند نقشی مؤثر در بهبود عملکرد مدل ندارند و تنها سبب افزایش پیچیدگی بی‌مورد می‌شوند، به‌طور کامل از مجموعه‌ی داده حذف شدند.

در ابتدای این مرحله، تعدادی از این ویژگی‌های غیرضروری را، همان‌گونه که در شکل (۳-۹) نمایش داده شده است، حذف کردیم تا مجموعه‌ی داده‌ای ساده‌تر، معنادارتر و بهینه‌تر برای مرحله‌ی بعدی مدل‌سازی در اختیار داشته باشیم.

```
data_copy.drop(["id2" , "universi" , 'region' , 'cluster' , 'birth_ye' , 'sample_c'], axis=1, inplace=True)
Executed at 2025.05.23 10:07:18 in 13ms
```

شکل ۳-۹حذف ویژگی‌های غیرضروری

در ادامه‌ی فرآیند پیش‌پردازش داده‌ها، توجه خود را معطوف به ویژگی‌های زمانی موجود در پرسش‌نامه کردیم. بسیاری از این پرسش‌ها، زمان انجام فعالیت‌ها را به‌صورت تفکیک‌شده در دو دسته‌ی روزهای تعطیل و روزهای مدرسه دریافت کرده بودند. اگرچه این تمایز در برخی موارد می‌تواند اطلاعات دقیقی به همراه داشته باشد، اما برای هدف این پژوهش که تمرکز بر شناسایی الگوهای کلی در رفتار و سلامت روان دانش‌آموزان است، تجمیع این اطلاعات در قالبی ساده‌تر و تحلیلی‌تر سودمندتر بود.

از این‌رو، تصمیم گرفتیم تا این دو ویژگی مجزا را در هر مورد ادغام کنیم و با محاسبه‌ی میانگین زمانی میان آن‌ها، به یک نماینده‌ی واحد و معنادار از رفتار زمانی دانش‌آموزان در طول هفته دست یابیم. بدین ترتیب، مقادیر حاصل از این میانگین‌گیری جایگزین دو ویژگی اولیه شدند تا هم ساختار داده‌ها بهینه‌تر شود و هم تحلیل مدل با متغیرهای خلاصه‌تر و مؤثرتری صورت گیرد.

```
data_copy['h_18'] = np.floor((data_copy['h_18'] + data_copy['h_19']) / 2)
Executed at 2025.05.23 10:07:19 in 4ms
```

شکل ۳-۰ نمونه‌ای از ادغام دو ویژگی زمانی مربوط به روزهای تعطیل و مدرسه

در بخش پرسش‌های کمی مرتبط با تغذیه، همانند ویژگی‌های زمانی، اطلاعات به‌صورت مجزا برای روزهای تعطیل و مدرسه ارائه شده بود. برای یکپارچه‌سازی و ساده‌سازی این اطلاعات، تصمیم گرفتیم تا با

جمع‌کردن مقادیر این دو بخش، نمایی از رفتار تغذیه‌ای دانش‌آموزان در مقیاس هفتگی به دست آوریم. بدین ترتیب، دو ویژگی مجزا در هر پرسش، در قالب یک ویژگی جدید خلاصه شده و جایگزین آن‌ها گردید.

در ادامه، برای انتخاب گروه‌های غذایی بر مبنای ساختار هرم غذایی، به بازتنظیم ویژگی‌ها پرداختیم. هدف از این مرحله، دسته‌بندی دقیق‌تر و معنادارتر داده‌های مرتبط با تغذیه بود. بر این اساس، ویژگی‌ها در قالب چهار گروه اصلی شامل: شیرینی‌ها، میوه‌ها، لبنیات و روغن‌ها طبقه‌بندی شدند. برای هر دسته، ابتدا مواد غذایی مربوطه شناسایی شده و سپس مقادیر آن‌ها با یکدیگر جمع شد. پس از آن، برای کاهش تأثیر مقادیر خارج از محدوده، میانگین این مقادیر محاسبه گردید و در نهایت، برای هماهنگ‌سازی و مقایسه‌پذیری بهتر، داده‌ها نرمال‌سازی شده و به اعدادی در بازه‌ی ۰ تا ۳ نگاشته شدند.

در نتیجه‌ی این فرایند، مجموعه‌ی نسبتاً گسترده‌ای از پرسش‌های شماره ۳۰ تا ۴۵، که هرکدام یک ویژگی مستقل را تشکیل می‌دادند، در قالب تنها چهار ویژگی اصلی بازتعریف شدند. این گام نه‌تنها موجب کاهش پیچیدگی داده‌ها شد، بلکه به تحلیل دقیق‌تر و مؤثرتر رفتار تغذیه‌ای دانش‌آموزان نیز کمک شایانی کرد

```
sweets_cols = ['h_30', 'h_32', 'h_33', 'h_34', 'h_37', 'h_44_1', 'h_44_2' , 'h_44_3']
data_copy['sweets'] = data_copy[sweets_cols].mean(axis=1)  # normalized score between 0 and 3
```

شکل ۳-۱۱ ادغام پرسش‌ها و ایجاد دسته‌ی شیرینی‌ها

در بخش بعدی، توجه خود را معطوف به فعالیت بدنی دانش‌آموزان می‌کنیم. بر اساس تعریفی که خانم دکتر جهانبخش و همکارانشان ارائه داده‌اند، فعالیت بدنی شامل هر نوع حرکتی است که باعث افزایش ضربان قلب شده و تنفس را تندتر می‌کند.

با توجه به این تعریف، ویژگی مربوط به وسیله‌ی نقلیه‌ی رفت‌وآمد به مدرسه را که پیش‌تر با روش کدگذاری وان‌هات[1] به صورت چند ستون جداگانه نمایش داده شده بود، بازتعریف کردیم. در این بازتعریف، گزینه‌های پیاده‌روی و دوچرخه‌سواری با مقدار عددی ۱ و سایر وسایل نقلیه با مقدار صفر مشخص شدند. این تغییر باعث شد تمامی ستون‌های مرتبط با این ویژگی به مقادیر عددی تبدیل شوند و قابلیت تحلیل بهتر در مدل‌سازی فراهم گردد.

در نهایت، با محاسبه‌ی میانگین این ستون‌ها، ویژگی جدیدی تحت عنوان «فعالیت بدنی[2]» ایجاد کردیم که نمایانگر سطح کلی فعالیت جسمانی دانش‌آموزان در رفت‌وآمد به مدرسه است و در تحلیل‌های بعدی مورد استفاده قرار خواهد گرفت.

---

[1] One-Hot Encoding
[2] physical activity

```
transport_map = {
    'car': 0,
    'bus': 0,
    'School service': 0,
    'Walking': 1,
    'Bicycle': 1
}
data_copy['active_transport'] = data_copy['z_55'].map(transport_map)
```

شکل ۳-۱۲ ارزش‌گذاری و تغییر فرمت ویژگی مربوط به وسیله‌ی نقلیه

در گام بعدی، به بررسی پرسش‌های مرتبط با اوقات فراغت دانش‌آموزان می‌پردازیم. در این بخش با دو چالش عمده مواجه بودیم: اول، اینکه پرسش‌ها به‌صورت جداگانه برای هر روز هفته و همچنین روز جمعه طراحی شده‌اند؛ و دوم، اینکه فعالیت نوشتن تکالیف در روز، به‌عنوان عاملی با اعتبار منفی برای زمان اوقات فراغت در نظر گرفته شده است.

برای رفع این مشکلات، ابتدا مجموع ساعات اختصاص یافته به فعالیت‌های اوقات فراغت در طول هفته، شامل روز جمعه، محاسبه شد. سپس، زمان صرف شده برای نوشتن تکالیف از این مجموع کسر گردید. در نهایت، حاصل به دست آمده بر تعداد روزهای هفته تقسیم شد تا میانگین زمان واقعی صرف شده برای اوقات فراغت به دست آید.

ویژگی نهایی محاسبه شده، نماینده‌ای جامع و دقیق از اوقات فراغت دانش‌آموزان است که جایگزین تمام پرسش‌های مربوط به این بخش در داده‌ها شد و تحلیل‌های بعدی بر اساس این ویژگی انجام خواهد گرفت.

```
data_copy['leisure_time(hour / week)'] = (
    (data_copy['h_57'] * 6 + data_copy['h_61'] * 6 + data_copy['h_63_1'] * 6  - data_copy['h_59']*6)  +
    (data_copy['h_58'] + data_copy['h_62'] + data_copy['h_64_1'] - data_copy['h_60'])
) / 7
```

شکل ۳-۱۳ جایگزین ویژگی اوقات فراغت با پرسش‌های مربوطه

در بخش پایانی، ستون‌های مرتبط با سلامت و ناخوشی‌های دانش‌آموزان را با یکدیگر تجمیع کردیم و ستونی واحد تحت عنوان «ناخوشی‌ها[1]» ایجاد نمودیم. این اقدام به منظور تسهیل تحلیل‌های آماری و کاهش پیچیدگی داده‌ها انجام شد. در نهایت، با استفاده از این ویژگی‌ها، ضریب همبستگی میان متغیرها را محاسبه و بررسی کردیم تا روابط و وابستگی‌های موجود بین آنها را شناسایی کنیم.

```
data_copy['illnesses'] = (
    data_copy['n_82']  + data_copy['n_83'] + data_copy['n_84']  + data_copy['n_85'] +
    data_copy['n_86'] + data_copy['n_87'] + data_copy['n_88'] + data_copy['n_89'])
Executed at 2025.05.19 23:59:12 in 13ms
```

شکل ۳-۱۴ جایگزین ویژگی ناخوشی‌ها با پرسش‌های مربوطه

[1] illnesses

۳-۶- ساخت مدل و شرح الگوریتم‌ها

در عصر حاضر، داده‌ها به عنوان منابعی ارزشمند برای کشف الگوها، پیش‌بینی رفتارها و تصمیم‌سازی‌های هوشمندانه شناخته می‌شوند. به‌ویژه در حوزه‌ی سلامت عمومی و بهداشت نوجوانان، تنوع و پیچیدگی متغیرهای رفتاری، تغذیه‌ای، روانی و محیطی، تحلیل ساده‌ی داده‌ها را ناکافی می‌سازد. از این‌رو، بهره‌گیری از مدل‌ها و الگوریتم‌های گوناگون در علم داده‌کاوی و یادگیری ماشین، نه‌تنها به درک عمیق‌تری از روابط میان متغیرها کمک می‌کند، بلکه می‌تواند روندهای پنهان و غیرخطی را که با روش‌های کلاسیک قابل شناسایی نیستند، آشکار سازد.

در پروژه‌ی حاضر که مبتنی بر داده‌های استخراج‌شده از پرسشنامه‌ی ملی کاسپین۵[1] است، استفاده از مدل‌های مناسب، کلید دستیابی به تحلیل‌های دقیق، پیش‌بینی وضعیت سلامت و شناسایی عوامل خطرآفرین یا حمایتی برای دانش‌آموزان محسوب می‌شود. انتخاب مدل‌های مناسب نه‌تنها به دقت نتایج کمک می‌کند، بلکه اعتبار علمی و کاربردی پژوهش را نیز ارتقا می‌بخشد. از این‌رو، بررسی، مقایسه و انتخاب مدل‌های مختلف در مسیر تحلیل داده‌ها، بخشی ضروری و اساسی از روند پژوهشی به‌شمار می‌رود.

۳-۶-۱- چرخه‌ی اول

با توجه به ماهیت عددی و پیوسته‌ی متغیر هدف در این پروژه که بیانگر ارزیابی سلامت ادراک‌شده توسط دانش‌آموزان در بازه‌ای از ۰ تا ۱۰ است. از مدل‌های مبتنی بر رگرسیون بهره گرفته شد. در این راستا، یکی از پرکاربردترین و قدرتمندترین ابزارهای یادگیری عمیق یعنی شبکه‌ی عصبی مصنوعی به‌کار گرفته شد تا بتواند با بهره‌گیری از ویژگی‌های استخراج‌شده، برآورد دقیقی از وضعیت سلامت ادراک‌شده ارائه دهد.

۳-۶-۱-۱- نرمال‌سازی داده‌ها[2]:

در بسیاری از الگوریتم‌های یادگیری ماشین، به‌ویژه شبکه‌های عصبی، مقیاس ویژگی‌ها تأثیر بسزایی در سرعت و کیفیت آموزش دارد. StandardScaler یکی از متداول‌ترین روش‌های نرمال‌سازی است که هر ویژگی را طوری تغییر می‌دهد که میانگین آن صفر و انحراف معیار آن یک شود.

این روش برای داده‌هایی که توزیع نرمال (یا نزدیک به نرمال) دارند بسیار مؤثر است و باعث می‌شود گرادیان‌ها در شبکه‌ی عصبی بهتر جریان پیدا کنند و فرآیند یادگیری سریع‌تر و پایدارتر شود.

---

[1] Caspian V
[2] Data normalization

```
# Normalize features
scaler = StandardScaler()
X = scaler.fit_transform(X)
```

شکل ۳-۱۵ نرمال‌سازی داده‌ها

۳-۶-۱-۲- تعریف مدل شبکه‌ی عصبی:

با توجه به ماهیت عددی و پیوسته‌ی متغیر هدف در این پروژه که بیانگر ارزیابی سلامت ادراک‌شده توسط دانش‌آموزان در بازه‌ای از ۰ تا ۱۰ است. از مدل‌های مبتنی بر رگرسیون بهره گرفته شد. در این راستا، یکی از پرکاربردترین و قدرتمندترین ابزارهای یادگیری عمیق یعنی شبکه‌ی عصبی مصنوعی به‌کار گرفته شد تا بتواند با بهره‌گیری از ویژگی‌های استخراج‌شده، برآورد دقیقی از وضعیت سلامت ادراک‌شده ارائه دهد.

اینجا از مدل ترتیبی[1] کتابخانه‌ی Keras استفاده شده که اجازه می‌دهد لایه‌ها به‌صورت خطی و پشت‌سرهم تعریف شوند.

- لایه‌ی اول، یک لایه‌ی کامل متصل با ۸ نورون است.که این لایه به اندازه‌ی تعداد ویژگی‌های ورودی (در اینجا ۱۰ ویژگی) ورودی می‌گیرد. همچنین از تابع فعال‌ساز ReLU استفاده شده است که برای جلوگیری از مشکل ناپدید شدن گرادیان و تسریع آموزش، بسیار پرکاربرد است.

- لایه‌ی پنهان دوم دارای ۴ نورون و همان تابع فعال‌ساز ReLU است.

- لایه‌ی خروجی تنها شامل یک نورون بدون تابع فعال‌ساز است. چون هدف ما پیش‌بینی عددی در بازه‌ی پیوسته ۰ تا ۱۰ است، خروجی باید خطی باقی بماند.

```
model = Sequential([
    Dense(8, input_dim=X.shape[1], activation='relu'),
    Dense(4, activation='relu'),
    Dense(1)  # No activation = linear output for regression
])
```

شکل ۳-۱۶ ساخت مدل شبکه‌ی عصبی

۳-۶-۱-۳- تنظیم مدل برای آموزش:

- تابع زیان[2] انتخاب‌شده، میانگین مربعات خطا[3] است که تفاوت بین خروجی مدل و مقدار

[1] Sequential
[2] Loss Function
[3] Mean squared error

واقعی را به‌صورت مربع محاسبه می‌کند. این تابع برای مسائل رگرسیونی همچون مجموعه‌ی داده‌ی کاسپین۵[1] مناسب است چون خطاهای بزرگ‌تر را بیشتر جریمه می‌کند.

- الگوریتم بهینه‌سازی Adam به‌صورت تطبیقی نرخ یادگیری را برای هر پارامتر تنظیم می‌کند. نرخ یادگیری ۰.۰۱ در این بخش از پروژه به کار رفته است.

- علاوه بر تابع زیان، مدل هنگام آموزش میانگین خطای مطلق[2] را نیز به‌عنوان معیار ارزیابی گزارش می‌کند. MAE مقدار مطلق اختلاف پیش‌بینی‌ها با مقادیر واقعی را بدون مربع کردن اندازه می‌گیرد و خواناتر است.

۳-۶-۱-۴- آموزش مدل:

داده‌های آموزش شامل ویژگی‌ها و مقادیر هدف هستند. آموزش مدل در ۱۰۰ دوره‌ی کامل[3] انجام می‌گیرد. در هر دوره، تمام داده‌های آموزشی یک‌بار به مدل ارائه می‌شود و وزن‌ها به‌روزرسانی می‌شوند. داده‌ها در هر مرحله‌ی آموزش به دسته‌های ۱۶تایی تقسیم می‌شوند تا حافظه بهتر مدیریت شود و گرادیان‌ها در دسته‌های کوچک محاسبه شوند. این کار سرعت آموزش را افزایش می‌دهد. ده درصد از داده‌های آموزشی به‌صورت خودکار جدا شده و برای اعتبارسنجی مدل در هر دوره[4] استفاده می‌شود. این کار به ما کمک می‌کند متوجه شویم آیا مدل در حال بیش‌برازش[5] است یا خیر. و در نهایت با verbose = 1 وضعیت هر دوره چاپ می‌شود.

```
model.compile(
    loss='mean_squared_error',
    optimizer=Adam(learning_rate=0.01),
    metrics=['mean_absolute_error']
)

Executed at 2025.05.23 10:17:25 in 23ms
```

شکل ۳-۱۷تنظیم مدل جهت کامپایل

[1] Caspian V
[2] Mean Absolute Error
[3] epoch
[4] epoch
[5] overfitting

```
history = model.fit(
    X_train, y_train,
    epochs=100,
    batch_size=16,
    validation_split=0.1,
    verbose=1
)
```

شکل ۳-۱۸آموزش مدل

۳-۶-۲- چرخه‌ی دوم

الگوریتم‌ها و مدل‌های به‌کاررفته در تحلیل سلامت روان دانش‌آموزان با هدف دسته‌بندی دقیق‌تر و معنادارتر دانش‌آموزان از منظر سلامت روان، در این بخش از پژوهش، تابع هدف از حالت پیوسته‌ی عددی به یک متغیر طبقه‌ای در سه سطح مجزا تقسیم گردید. این سه سطح با توجه به نمره‌ی سلامت ذهنی ادراک‌شده توسط خودِ دانش‌آموزان به‌صورت زیر تعریف شده‌اند:

- گروه بحرانی(Critical): نمرات ۱ تا ۳
- گروه آسیب‌پذیر (Vulnerable): نمرات ۴ تا ۷
- گروه باثبات (Stable): نمرات ۸ تا ۱۰

این تقسیم‌بندی، ضمن تقویت خوانایی تحلیلی داده‌ها، امکان بهره‌گیری از الگوریتم‌های یادگیری نظارتی در قالب دسته‌بندی[1] را فراهم ساخته و بستر پیش‌بینی دقیق‌تر گروه‌های در معرض خطر را فراهم می‌آورد.

```
def categorize(score):
    if score <= 3:
        return 0  # Critical
    elif score <= 7:
        return 1  # Vulnerable
    else:
        return 2  # Stable

df['n_93'] = df['n_93'].apply(categorize)
```

Executed at 2025.05.23 10:19:37 in 6ms

شکل ۳-۱۹تقسیم بندی ویژگی هدف به سه دسته

[1] Classification

۳-۶-۲-۱- آماده‌سازی داده‌ها و پیش‌پردازش

ابتدا با استفاده از تابع categorize، داده‌های ستونی با عنوان n_93 که نشان‌دهنده‌ی سلامت کلی دانش‌آموزان بود، به سه برچسب عددی تبدیل شد.

در ادامه، ۱۱ویژگی مرتبط با عادات تغذیه‌ای، فعالیت بدنی، سبک زندگی و میزان ناخوشی‌های گزارش‌شده به‌عنوان ورودی مدل انتخاب شدند. این ویژگی‌ها با استفاده از روش استانداردسازی[1] نرمال‌سازی شدند تا مدل در مسیر یادگیری با داده‌هایی هم‌مقیاس مواجه شود.

۳-۶-۲-۲- طراحی مدل شبکه‌ی عصبی

```
# Select features and target
features = ['Vegetables', 'tea Intake (estekan/day)', 'sumScore','sweets' , 'fruits' , 'dairy' , 'oils' ,
 'active_transport' , 'physical_activity' , 'leisure_time(hour / week)','illnesses']
target = 'n_93'  # Decimal: 0 (poor) to 10 (good)


X = df[features].values
y = df[target].values

# Normalize features
scaler = StandardScaler()
X = scaler.fit_transform(X)

# Train/test split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42)
```

شکل ۳-۲۰انتخاب ویژگی‌ها و استانداردسازی

برای انجام عمل دسته‌بندی، یک شبکه‌ی عصبی چندلایه[2] طراحی شد. ساختار این شبکه به‌صورت زیر تنظیم شده است:

- **لایه‌ی اول:** ۸ نورون و تابع فعال‌ساز ReLU، که مسئول پردازش اولیه و استخراج ویژگی‌های پنهان است.
- **لایه‌ی دوم:** ۶ نورون با تابع ReLU، که نقش میانجی را بین لایه‌های ورودی و خروجی ایفا می‌کند.
- **لایه‌ی خروجی:** ۳ نورون به تعداد کلاس‌های هدف با تابع فعال‌ساز Softmax که هر نورون احتمال تعلق یک نمونه به هر کلاس را پیش‌بینی می‌کند.

---

[1] StandardScaler
[2] Multi-layer Perceptron

```
model = Sequential([
    Dense(8, input_dim=X.shape[1], activation='relu'),
    Dense(6, activation='relu'),
    Dense(3, activation='softmax')  # 3 classes => softmax
])
```

شکل ۳-۲۱مدل‌سازی شبکه‌عصبی

۳-۶-۲-۳- تنظیم مدل برای آموزش(کامپایل):

مدل با استفاده از تابع زیان sparse_categorical_crossentropy و بهینه‌ساز Adam پیکربندی شد. تابع زیان انتخاب‌شده برای مسائل چندکلاسه با برچسب‌های عددی مناسب بوده و عملکردی دقیق در بهینه‌سازی دارد. Adam به‌عنوان یک الگوریتم بهینه‌سازی تطبیقی، سرعت همگرایی مدل را افزایش داده و مانع از گیر افتادن در کمینه‌های محلی می‌شود.معیار ارزیابی دقت[1] نیز جهت اندازه‌گیری کارایی طبقه‌بندی مدل در نظر گرفته شده است.

```
model.compile(
    loss='sparse_categorical_crossentropy',
    optimizer='adam',
    metrics=['accuracy']
)
```

شکل ۳-۲۲تنظیم مدل برای آموزش

۳-۶-۲-۴- آموزش و ارزیابی مدل:

داده‌ها با نسبت ۲۰/۸۰ به دو مجموعه‌ی آموزش و آزمون تقسیم شدند. پس از طی ۱۰۰ دوره آموزش، خروجی مدل (که به‌صورت احتمالات مربوط به هر کلاس در قالب Softmax است) به برچسب‌های نهایی تبدیل شد.

```
history = model.fit(
    X_train, y_train,
    epochs=100,
    batch_size=16,
    validation_split=0.1,
    verbose=1
)
```

شکل ۳-۲۳آموزش مدل

[1] accuracy

برای ارزیابی مدل، از گزارش طبقه‌بندی[1] استفاده گردید که شاخص‌هایی چون دقت[2]، بازخوانی[3] و امتیاز F1 را برای هر دسته محاسبه می‌کند. این گزارش توان مدل را در تفکیک دقیق میان سه طبقه‌ی سلامت روانی نشان می‌دهد.

```python
from sklearn.metrics import classification_report
import numpy as np

# Predict probabilities
y_pred_probs = model.predict(X_test)

# Convert softmax outputs to class labels
y_pred = np.argmax(y_pred_probs, axis=1)

# True labels (no one-hot encoding)
y_true = y_test  # shape (None,)

# If y_test is still one-hot, convert it:
if len(y_test.shape) > 1:
    y_true = np.argmax(y_test, axis=1)

# Classification report
print(classification_report(y_true, y_pred, target_names=['Critical', 'Vulnerable', 'Stable']))
```

Executed at 2025.05.23 10:34:30 in 94ms

شکل ۳-۲۴ ارزیابی مدل

۳-۶-۲-۵- تقویت داده:

پس از بررسی نتایج و ارزیابی مدل بر این باور رسیدیم که با توجه به محدود بودن حجم نمونه‌های موجود در مجموعه‌داده، و همچنین عدم توازن میان گروه‌های هدف (بحرانی، آسیب‌پذیر و مطلوب)، نیاز به تقویت داده‌ها به‌منظور بهبود عملکرد مدل‌های یادگیری ماشین به‌وضوح احساس می‌شد. در همین راستا، یکی از مؤثرترین روش‌های تقویت داده در داده‌های جدولی، تکنیک SMOTE[4] مورد استفاده قرار گرفت.

با استفاده از این روش داده‌های ما به ۴۵۰ نمونه افزایش یافت از سویی داده‌های مربوط به گروه هدف بحرانی افزایش یافت که نتایج حاصل در بخش مربوطه به تفصیل شرح داده می‌شود.

۳-۷- جمع‌بندی

در این فاز از پژوهش که با موضوعیت شرح پروژه و نحوه‌ی اجرای آن تدوین شد، تلاش گردید تصویری جامع و مرحله‌به‌مرحله از مسیر عملیاتی پروژه ارائه شود. ابتدا با معرفی جامعه‌ی آماری و منبع داده‌ها، شیوه‌ی گردآوری اطلاعات مشخص گردید؛ سپس مراحل پیش‌پردازش داده‌ها شامل پاک‌سازی، نرمال‌سازی، ادغام

---

[1] classification report
[2] Precision
[3] Recall
[4] Synthetic Minority Over-sampling Technique

ویژگی‌ها و بازنمایی مفاهیم کلیدی به‌تفصیل بیان شد.

در ادامه، با تکیه بر اصول داده‌کاوی و چرخه‌ی استاندارد کریسپ[1]، فرآیند مدل‌سازی آغاز شد. در این مسیر، مدل‌های یادگیری ماشین به‌ویژه شبکه‌ی عصبی مصنوعی برای تحلیل و طبقه‌بندی وضعیت سلامت روان دانش‌آموزان طراحی و پیاده‌سازی شدند. به منظور افزایش دقت پیش‌بینی و تفسیرپذیری نتایج، انتخاب ویژگی‌ها به‌صورت هدفمند انجام گرفته و الگوریتم‌های مناسب با نوع داده‌ها برگزیده شد.

همچنین ابزارهای نرم‌افزاری نظیر پایتون، کتابخانه‌هایی چون pandas, scikit-learn, TensorFlow و محیط‌هایی چون Jupyter و PyCharm، بستر اجرای فنی مدل‌ها را فراهم آوردند.

به‌طور کلی، این فاز زمینه‌ساز ورود به تحلیل‌های عمیق‌تر در فاز بعدی پژوهش است؛ جایی که کارایی مدل‌های طراحی‌شده مورد ارزیابی قرار گرفته و تفسیر یافته‌ها با هدف کاربرد در سیاست‌گذاری‌های آموزشی و سلامت روان انجام خواهد شد.

[1] CRISP-DM

فصل ٤
نتایج

۴-۱- مقدمه

در فصل نتایج، تلاش می‌شود تا ثمره‌ی مراحل مختلف تحلیل داده‌ها، از پیش‌پردازش و مدل‌سازی گرفته تا ارزیابی عملکرد الگوریتم‌ها، به شکلی روشن و قابل فهم ارائه شود. این فصل نقش کلیدی در پروژه دارد؛ چرا که یافته‌های آن می‌تواند به درک بهتر وضعیت سلامت روان دانش‌آموزان منجر شود.

در این بخش، ابتدا با استفاده از روش‌های تصویرسازی داده‌ها[1]، توزیع اولیه‌ی ویژگی‌ها، ارتباط میان متغیرها و الگوهای پنهان در داده‌ها نمایان خواهد شد. این گام نه‌تنها به فهم عمیق‌تر داده‌ها کمک می‌کند، بلکه مسیر تصمیم‌گیری برای انتخاب مدل‌های مناسب را نیز هموار می‌سازد.

در ادامه، خروجی مدل‌سازی‌هایی که با استفاده از روش‌های نوین داده‌کاوی و به‌ویژه الگوریتم شبکه‌ی عصبی طراحی شده‌اند، به دقت بررسی می‌شود. عملکرد مدل‌ها از طریق معیارهایی نظیر دقت، خطای میانگین[2] و گزارش طبقه‌بندی[3] ارزیابی می‌گردد تا بتوان میزان موفقیت الگوریتم‌ها در پیش‌بینی وضعیت سلامت روان دانش‌آموزان را مشخص کرد.

در نهایت، تحلیل‌های انجام‌شده در این فصل نه‌تنها اعتبار علمی پروژه را تقویت می‌کند، بلکه بستری برای ارائه‌ی پیشنهادهای کاربردی در حوزه‌ی ارتقای سلامت دانش‌آموزان نیز فراهم می‌آورد.

---

[1] Data Visualization
[2] Mean Square Error
[3] classification report

۴-۲- بررسی اولیه داده

در این بخش ویژگی‌های آماری داده‌ها، مقایسه بین داده‌های جدید با داده‌های اولیه و ارتباط بین ویژگی‌ها مورد بررسی و نمایش گذاشته می‌شود تا در فهم بشتر داده‌ها و یافتن الگوهای پنهان به ما کمک کند. همچنین استفاده از امکانات بصری اقدامات صورت گرفته برای دست‌یابی به مجموعه داده‌ی جدید را برای سایرین قابل درک‌تر می‌نماید.

جدول ۴-۱اطلاعات اولیه مجموعه داده‌ی smaple

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 75 columns):

#	Column	Non-Null Count	Dtype
0	id2	200 non-null	int64
1	universi	199 non-null	object
2	region	199 non-null	object
3	cluster	199 non-null	float64
4	cross	199 non-null	object
5	sex	199 non-null	object
6	birth_ye	199 non-null	float64
7	sample_c	199 non-null	float64
8	h_18	178 non-null	float64
9	h_19	184 non-null	float64
10	h_20	192 non-null	object
11	h_21	187 non-null	object
12	h_22	179 non-null	float64
13	h_23	187 non-null	float64
14	h_24	191 non-null	object
15	h_25	190 non-null	object
16	h_26	193 non-null	float64
17	h_27	189 non-null	float64
18	h_28	193 non-null	object
19	h_29	193 non-null	object
20	h_30	199 non-null	object
21	h_31	198 non-null	object
22	h_32	197 non-null	object
23	h_33	193 non-null	object
24	h_34	196 non-null	object
25	h_35_a	192 non-null	object
26	h_35_b	104 non-null	object
27	h_36	192 non-null	object
28	h_37	197 non-null	object
29	h_38	196 non-null	object
30	h_39	198 non-null	object
31	h_40	196 non-null	object
32	h_41	198 non-null	object
33	h_42	198 non-null	object
34	h_43_1	195 non-null	object
35	h_43_2	196 non-null	object
36	h_44_1	195 non-null	object
37	h_44_2	197 non-null	object
38	h_44_3	195 non-null	object
39	h_45	197 non-null	object
40	z_52	191 non-null	object
41	z_53	198 non-null	object
42	z_54	194 non-null	object
43	z_55	195 non-null	object
44	z_56_1	187 non-null	object
45	z_56_2	186 non-null	object
46	z_56_3	186 non-null	object
47	z_56_4	181 non-null	object
48	z_56_5	189 non-null	object
49	z_56_6	191 non-null	object
50	z_56_7	184 non-null	object
51	h_57	198 non-null	object
52	h_58	197 non-null	object
53	h_59	194 non-null	object
54	h_60	191 non-null	object
55	h_61	191 non-null	object
56	h_62	186 non-null	object
57	h_63_1	185 non-null	float64
58	h_63_2	184 non-null	float64
59	h_64_1	186 non-null	float64
60	h_64_2	188 non-null	float64
61	sumScore	163 non-null	float64
62	tertiles	163 non-null	object
63	n_82	196 non-null	object
64	n_83	191 non-null	object
65	n_84	192 non-null	object
66	n_85	191 non-null	object
67	n_86	197 non-null	object
68	n_87	193 non-null	object
69	n_88	189 non-null	object
70	n_89	192 non-null	object
71	n_90	193 non-null	object
72	n_91	197 non-null	object
73	n_92	197 non-null	object
74	n_93	192 non-null	float64

dtypes: float64(15), int64(1), object(59)
memory usage: 117.3+ KB

جدول ۴-۲اطلاعات اولیه مجموعه داده‌ی تمیز شده در گام نخست

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 196 entries, 0 to 195
Data columns (total 76 columns):

#	Column	Non-Null Count	Dty
0	id2	196 non-null	int64
1	universi	196 non-null	object
2	region	196 non-null	object
3	cluster	196 non-null	int64
4	birth_ye	196 non-null	int64
5	sample_c	196 non-null	int64
6	h_18	196 non-null	int64
7	h_19	196 non-null	int64
8	h_20	196 non-null	int64
9	h_21	196 non-null	bool
10	h_22	196 non-null	int64
11	h_23	196 non-null	int64
12	h_24	196 non-null	int64
13	h_25	196 non-null	bool
14	h_26	196 non-null	int64
15	h_27	196 non-null	int64
16	h_28	196 non-null	int64
17	h_29	196 non-null	bool
18	h_30	196 non-null	int64
19	h_31	196 non-null	int64
20	h_32	196 non-null	int64
21	h_33	196 non-null	int64
22	h_34	196 non-null	int64
23	h_35_a	196 non-null	int64
24	h_36	196 non-null	int64
25	h_37	196 non-null	int64
26	h_38	196 non-null	int64
27	h_39	196 non-null	int64
28	h_40	196 non-null	int64
29	h_41	196 non-null	int64
30	h_42	196 non-null	int64
31	h_43_1	196 non-null	int64
32	h_43_2	196 non-null	float64
33	h_44_1	196 non-null	int64
34	h_44_2	196 non-null	int64
35	h_44_3	196 non-null	int64
36	h_45	196 non-null	int64
37	z_52	196 non-null	int64
38	z_53	196 non-null	bool
39	z_54	196 non-null	int64
40	z_56_1	196 non-null	int64
41	z_56_2	196 non-null	int64
42	z_56_3	196 non-null	int64
43	z_56_4	196 non-null	int64
44	z_56_5	196 non-null	int64
45	z_56_6	196 non-null	int64
46	z_56_7	196 non-null	int64
47	h_57	196 non-null	int64
48	h_58	196 non-null	int64
49	h_59	196 non-null	int64
50	h_60	196 non-null	int64
51	h_61	196 non-null	int64
52	h_62	196 non-null	int64
53	h_63_1	196 non-null	float64
54	h_63_2	196 non-null	int64
55	h_64_1	196 non-null	float64
56	h_64_2	196 non-null	int64
57	sumScore	196 non-null	int64
58	tertiles	196 non-null	int64
59	n_82	196 non-null	int64
60	n_83	196 non-null	int64
61	n_84	196 non-null	int64
62	n_85	196 non-null	int64
63	n_86	196 non-null	int64
64	n_87	196 non-null	int64
65	n_88	196 non-null	int64
66	n_89	196 non-null	int64
67	n_90	196 non-null	int64
68	n_91	196 non-null	int64
69	n_92	196 non-null	int64
70	n_93	196 non-null	int64
71	z_55	196 non-null	object
72	cross_Elemantry	196 non-null	int64
73	cross_intermidiate	196 non-null	int64
74	sex_Boy	196 non-null	int64
75	sex_Girl	196 non-null	int64

dtypes: bool(4), float64(3), int64(66), object(3)

جدول ۴-۳ اطلاعات اولیه مجموعه‌ی داده‌ی تمیزشده در گام دوم

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 196 entries, 0 to 195
Data columns (total 22 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   h_20                    196 non-null    int64
 1   h_24                    196 non-null    int64
 2   h_28                    196 non-null    int64
 3   Vegetables              196 non-null    int64
 4   tea Intake (estekan/day)  196 non-null    float64
 5   sumScore                196 non-null    int64
 6   tertiles                196 non-null    int64
 7   n_90                    196 non-null    int64
 8   n_91                    196 non-null    int64
 9   n_92                    196 non-null    int64
 10  n_93                    196 non-null    int64
 11  sex_Boy                 196 non-null    int64
 12  sex_Girl                196 non-null    int64
 13  sweets                  196 non-null    float64
 14  fruits                  196 non-null    float64
 15  dairy                   196 non-null    float64
 16  oils                    196 non-null    float64
 17  active_transport        196 non-null    int64
 18  physical_activity       196 non-null    float64
 19  leisure_time(hour / week)  196 non-null    float64
 20  illnesses               196 non-null    int64
 21  n_93_cluster            196 non-null    int64
dtypes: float64(7), int64(15)
```

همان‌گونه که در جدول۴-۱ مشاهده می‌کنید تعداد پرسش‌ها ۷۵ و تعداد سطرها ۲۰۰ عدد است و مقادیر بیشتر ستون‌ها مقادیر غیر عددی می‌باشد که برای تحلیل و مدل‌سازی مسئله مناسب نیست. در جدول۴-۲ تعداد ستون‌ها ۷۶ و تعداد سطرها ۱۹۶ عدد می‌باشد که این اختلاف عدد در سطرها به دلیل سطرهای تنک [1]می‌باشد که در ادامه‌ی فرایند تمام این سطرها حذف گردید. اما در این مرحله بیشتر ستون‌ها به مقادیر عددی تبدیل یافت و چون از روش کدگذاری وان‌هات استفاده گردید تعداد ستون‌ها افزایش یافت ولی از سویی دیگر ستون‌هایی که مقادیر تهی[2] زیادی داشتند(ستون b-35 با ۹۶داده‌ی تهی) حذف گردید و تمام ستون‌ها با استفاده از روش‌های میانگین‌گیری داده‌های تهی آنها جایگزین شد که درنهایت ۷۶ ستون برای ما ایجاد شد. در مجموعه‌ی داده‌ی سوم جدول ۴-۳ اینبار تلاش شد با ادغام ویژگی‌ها و ایجاد دسته‌هایی شامل چند ویژگی و حذف برخی از ویژگی‌هایی که اطلاعات نه چندان زیادی به ما می‌دهند تعداد ستون‌ها را کاهش دهیم همان‌گونه که در شکل بالا مشاهده‌می‌کنید تمام ستون‌ها مقادیر عددی دارند که محاسبات و مدل‌سازی‌های ما را درگام بعدی ساده‌تر می‌کنند. در جدول‌های ۴-۴ و ۴-۵ جداولی جهت توصیق آماری داده‌ها ایجاد شده است. در این جداول هدف توصیف توزیع داده‌ها در چارک‌های مختلف، محاسبه‌ی انحراف از معیار[3]، کمینه و بیشینه‌ی ویژگی‌ها می‌باشد.

در جدول۴-۴ نسبت به شکل جدول۴-۵ انحراف از معیار در بیشتر ستون‌ها کاهش یافته است و در پرسش‌هایی که داده‌های آنها مقادیر زمانی بودند و پراکندگی داده‌ها بسیار زیاد بود با ادغام روزهای هفته با هم و وزن‌دهی مناسب این مقدار کاهش یافت و برخی از ستون‌ها که انحراف از معیار بسیار زیادی داشتند و

---

[1] sparse
[2] null
[3] standard deviation

مقیاس‌پذیری داده‌های ما را بهم می‌ریختند حذف شدند مانند ویژگی(h_18 با انحراف از معیار ۶۴).

با ادغام ویژگی‌ها نیز مقدار انحراف از معیار در بین ویژگی‌ها نسب با مجموعه‌ی داده‌ی اول بسیار کاهش یافت برای مثال پرسش‌های 'h_44_3' , 'h_44_2', 'h_44_1', 'h_37', 'h_34', 'h_33', 'h_32', 'h_30' مربوط به دسته‌ی قندی‌ها هستند که میانگین انحراف از معیار آنها برابر است با ۰.۸۷ است اما مقدار بدست آمده در جدول۴-۵ برابر با ۰.۴۳ است و این اتفاق تقریبا برای تمام دسته‌ها نیز رخ داده است و اینکار منجر به کاهش پراکندگی داده‌ها و استاندارد سازی ویژگی ها می‌باشد.

جدول ۴-۴توصیف آماری از مجموعه‌ی داده تمیز شده اول

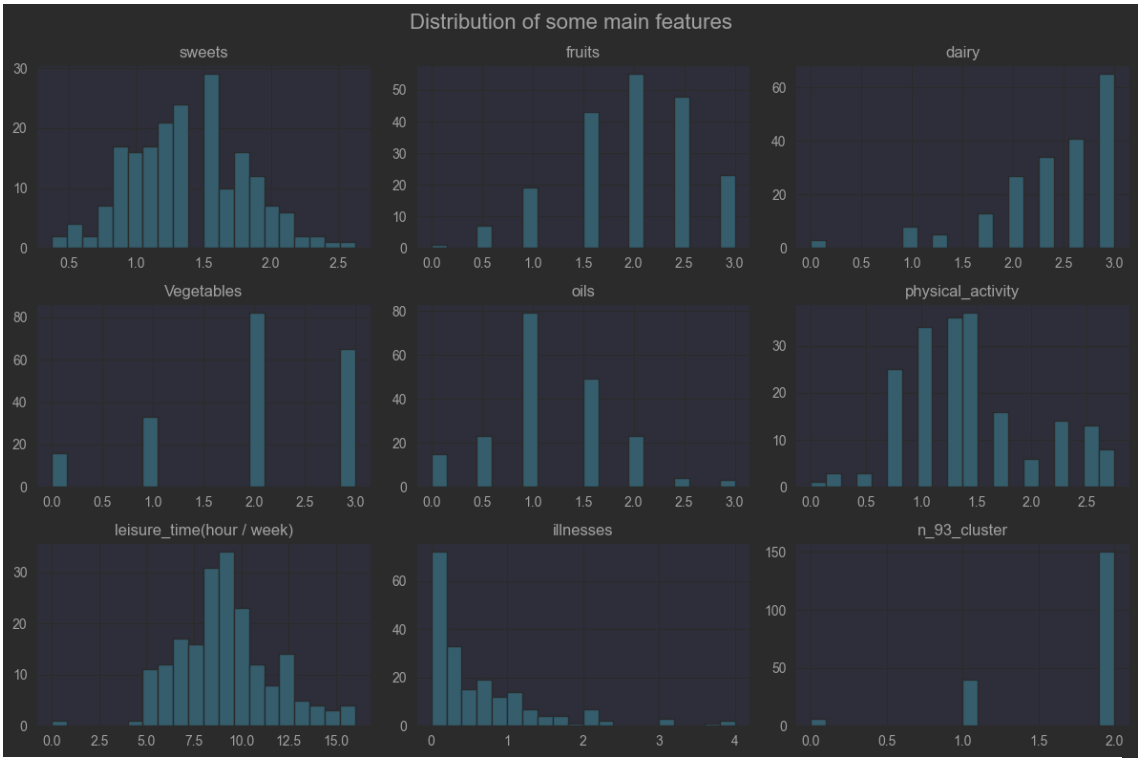
Feature	count	mean	std	min	25%	50%	75%	max
id2	196.0	1112062643.72	6724732.04	1101118501.0	1106118379.25	1112218006.5	1118217803.25	1123128702.0
cluster	196.0	11.9	6.71	1.0	6.0	12.0	18.0	23.0
birth_ye	196.0	82.13	2.83	76.0	80.0	82.0	84.0	87.0
sample_c	196.0	5.46	2.87	1.0	3.0	5.5	8.0	10.0
h_18	196.0	471.84	64.21	360.0	420.0	465.0	540.0	720.0
h_19	196.0	557.3	55.62	420.0	540.0	557.0	600.0	720.0
h_20	196.0	4.85	1.77	0.0	5.0	6.0	6.0	6.0
h_22	196.0	841.04	52.6	660.0	840.0	840.0	870.0	960.0
h_23	196.0	846.15	41.54	720.0	840.0	840.0	870.0	960.0
h_24	196.0	5.11	1.7	0.0	5.0	6.0	6.0	6.0
h_26	196.0	1256.86	54.43	1140.0	1230.0	1260.0	1290.0	1440.0
h_27	196.0	1262.72	56.41	1140.0	1230.0	1260.0	1290.0	1440.0
h_28	196.0	5.36	1.39	0.0	5.0	6.0	6.0	6.0
h_30	196.0	2.01	0.84	0.0	1.0	2.0	3.0	3.0
h_31	196.0	1.26	0.77	0.0	1.0	1.0	2.0	3.0
h_32	196.0	1.33	0.81	0.0	1.0	1.0	2.0	3.0
h_33	196.0	0.34	0.66	0.0	0.0	0.0	0.25	3.0
h_34	196.0	0.85	0.87	0.0	0.0	1.0	2.0	3.0
h_35_a	196.0	2.28	0.83	0.0	2.0	2.0	3.0	3.0
h_36	196.0	1.66	0.91	0.0	1.0	2.0	2.0	3.0
h_37	196.0	1.51	0.88	0.0	1.0	1.0	2.0	3.0
h_38	196.0	2.0	0.91	0.0	1.75	2.0	3.0	3.0
h_39	196.0	2.43	0.8	0.0	2.0	3.0	3.0	3.0
h_40	196.0	2.36	0.79	0.0	2.0	3.0	3.0	3.0
h_41	196.0	2.46	0.9	0.0	2.0	3.0	3.0	3.0
h_42	196.0	1.08	0.78	0.0	1.0	1.0	1.0	3.0
h_43_1	196.0	2.39	0.98	0.0	2.0	3.0	3.0	3.0
h_43_2	196.0	1.53	1.07	0.0	1.5	1.5	1.5	4.0
h_44_1	196.0	2.25	1.03	0.0	2.0	3.0	3.0	3.0
h_44_2	196.0	1.64	0.86	0.0	1.0	2.0	2.0	3.0
h_44_3	196.0	1.09	1.06	0.0	0.0	1.0	2.0	3.0
h_45	196.0	0.87	0.79	0.0	0.0	1.0	1.0	3.0
z_52	196.0	2.83	2.15	0.0	1.0	2.0	4.0	7.0
z_54	196.0	1.56	0.63	0.0	1.0	2.0	2.0	3.0
z_56_1	196.0	1.57	1.32	0.0	0.0	2.0	3.0	4.0
z_56_2	196.0	1.66	1.38	0.0	0.0	2.0	3.0	4.0
z_56_3	196.0	1.75	1.42	0.0	0.0	2.0	3.0	4.0
z_56_4	196.0	1.52	1.36	0.0	0.0	1.0	3.0	4.0
z_56_5	196.0	2.12	1.41	0.0	1.0	2.0	3.0	4.0
h_63_2	196.0	1358.07	62.77	1140.0	1320.0	1380.0	1380.0	1530.0
h_64_1	196.0	9.21	1.48	4.0	8.0	9.0	10.0	14.0
h_64_2	196.0	1387.91	65.66	1140.0	1350.0	1380.0	1440.0	1560.0
sumScore	196.0	20.63	7.99	11.0	14.0	21.0	24.0	49.0
tertiles	196.0	1.22	0.86	0.0	0.0	1.5	2.0	2.0
n_82	196.0	0.75	1.28	0.0	0.0	0.0	1.0	4.0
n_83	196.0	0.59	1.15	0.0	0.0	0.0	1.0	4.0
n_84	196.0	0.42	1.0	0.0	0.0	0.0	0.0	4.0
n_85	196.0	0.3	0.9	0.0	0.0	0.0	0.0	4.0
n_86	196.0	1.24	1.53	0.0	0.0	0.0	3.0	4.0
n_87	196.0	0.84	1.31	0.0	0.0	0.0	1.0	4.0
n_88	196.0	0.51	1.1	0.0	0.0	0.0	0.0	4.0
n_89	196.0	0.31	0.86	0.0	0.0	0.0	0.0	4.0
n_90	196.0	1.54	0.72	0.0	1.0	2.0	2.0	2.0
n_91	196.0	0.65	0.95	0.0	0.0	0.0	1.0	4.0
n_92	196.0	2.35	0.8	0.0	2.0	3.0	3.0	3.0
n_93	196.0	8.44	2.04	0.0	8.0	9.0	10.0	10.0
cross_Elemantry	196.0	0.61	0.49	0.0	0.0	1.0	1.0	1.0
cross_intermidiate	196.0	0.39	0.49	0.0	0.0	0.0	1.0	1.0
sex_Boy	196.0	0.74	0.44	0.0	0.0	1.0	1.0	1.0
sex_Girl	196.0	0.26	0.44	0.0	0.0	0.0	1.0	1.0

جدول ۴-۵توصیف آماری از مجموعه‌ی داده‌ی تمیزشده دوم

Feature	count	mean	std	min	25%	50%	75%	max
h_20	196.0	5.81	1.84	0.0	6.0	7.0	7.0	7.0
h_24	196.0	6.07	1.72	1.0	6.0	7.0	7.0	7.0
h_28	196.0	6.32	1.48	0.0	6.0	7.0	7.0	7.0
Vegetables	196.0	2.0	0.91	0.0	1.75	2.0	3.0	3.0
tea Intake (estekan/day)	196.0	1.53	1.07	0.0	1.5	1.5	1.5	4.0
sumScore	196.0	20.63	7.99	11.0	14.0	21.0	24.0	49.0
tertiles	196.0	1.22	0.86	0.0	0.0	1.5	2.0	2.0
n_90	196.0	1.54	0.72	0.0	1.0	2.0	2.0	2.0
n_91	196.0	0.65	0.95	0.0	0.0	0.0	1.0	4.0
n_92	196.0	2.35	0.8	0.0	2.0	3.0	3.0	3.0
n_93	196.0	8.44	2.04	0.0	8.0	9.0	10.0	10.0
sex_Boy	196.0	0.74	0.44	0.0	0.0	1.0	1.0	1.0
sex_Girl	196.0	0.26	0.44	0.0	0.0	0.0	1.0	1.0
sweets	196.0	1.38	0.43	0.38	1.12	1.38	1.62	2.62
fruits	196.0	1.97	0.66	0.0	1.5	2.0	2.5	3.0
dairy	196.0	2.42	0.63	0.0	2.0	2.67	3.0	3.0
oils	196.0	1.17	0.61	0.0	1.0	1.0	1.5	3.0
active_transport	196.0	0.46	0.5	0.0	0.0	0.0	1.0	1.0
physical_activity	196.0	1.44	0.6	0.0	1.0	1.25	1.75	2.75
leisure_time(hour / week)	196.0	9.16	2.48	0.0	7.43	9.0	10.43	16.0
illnesses	196.0	4.95	6.09	0.0	0.0	3.0	7.0	32.0
n_93_cluster	196.0	1.73	0.51	0.0	2.0	2.0	2.0	2.0

۴-۳- نمایش هسیتوگرام ویژگی‌های اصلی

در این پروژه، که با هدف تحلیل سلامت روان دانش‌آموزان انجام شده، بررسی توزیع ویژگی‌هایی نظیر میزان مصرف گروه‌های غذایی، زمان فعالیت بدنی یا ساعات اوقات فراغت، می‌تواند اطلاعات ارزشمندی درباره‌ی الگوهای رفتاری رایج در میان دانش‌آموزان ارائه دهد. این اطلاعات، علاوه‌بر آن‌که پایه‌ای برای مدل‌سازی‌های دقیق‌تر فراهم می‌کنند، درک ملموس‌تری از شرایط زیستی و روانی شرکت‌کنندگان را نمایش می‌دهند.

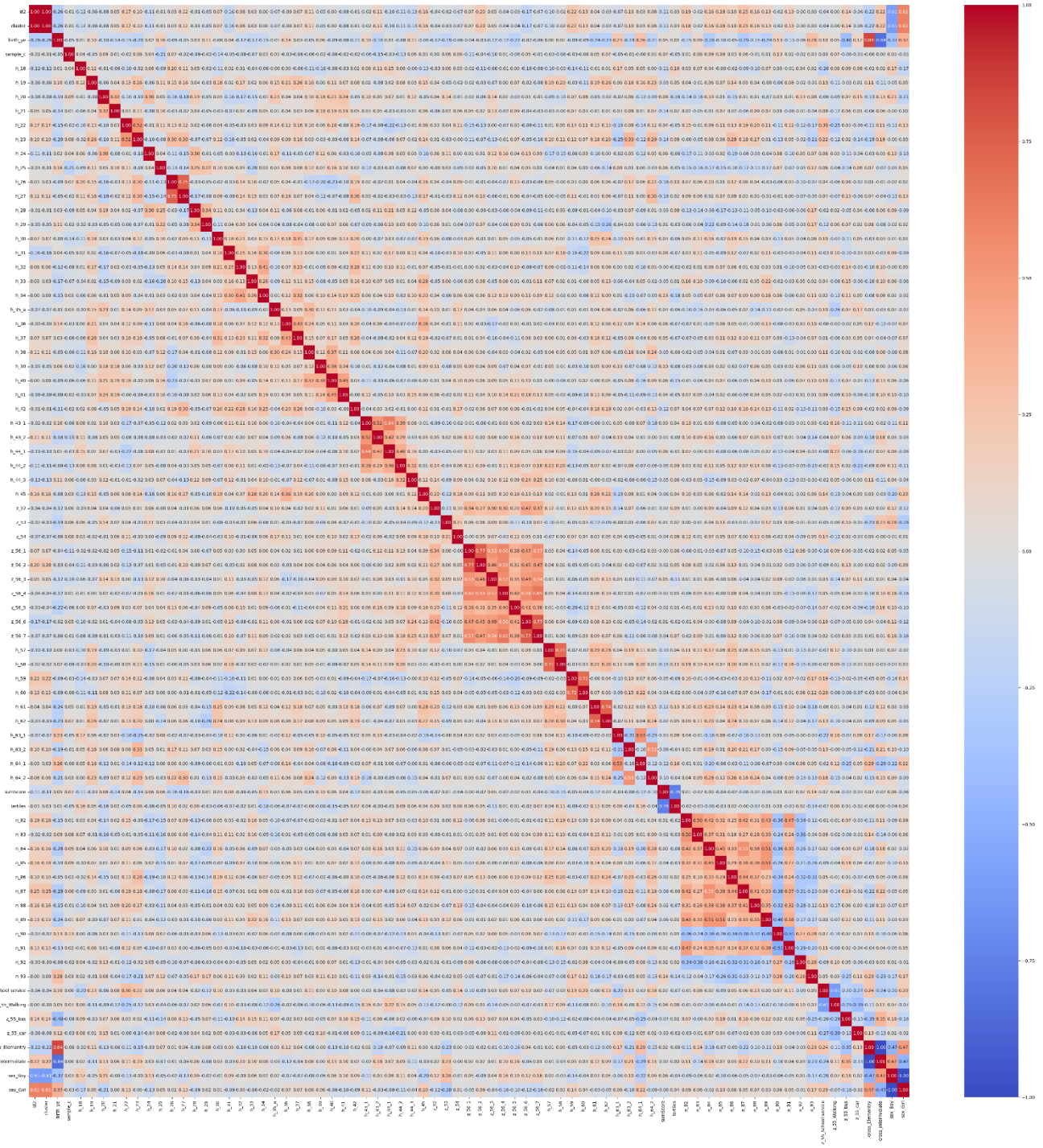


شکل ۴-۱توزیع ویژگی‌های اصلی

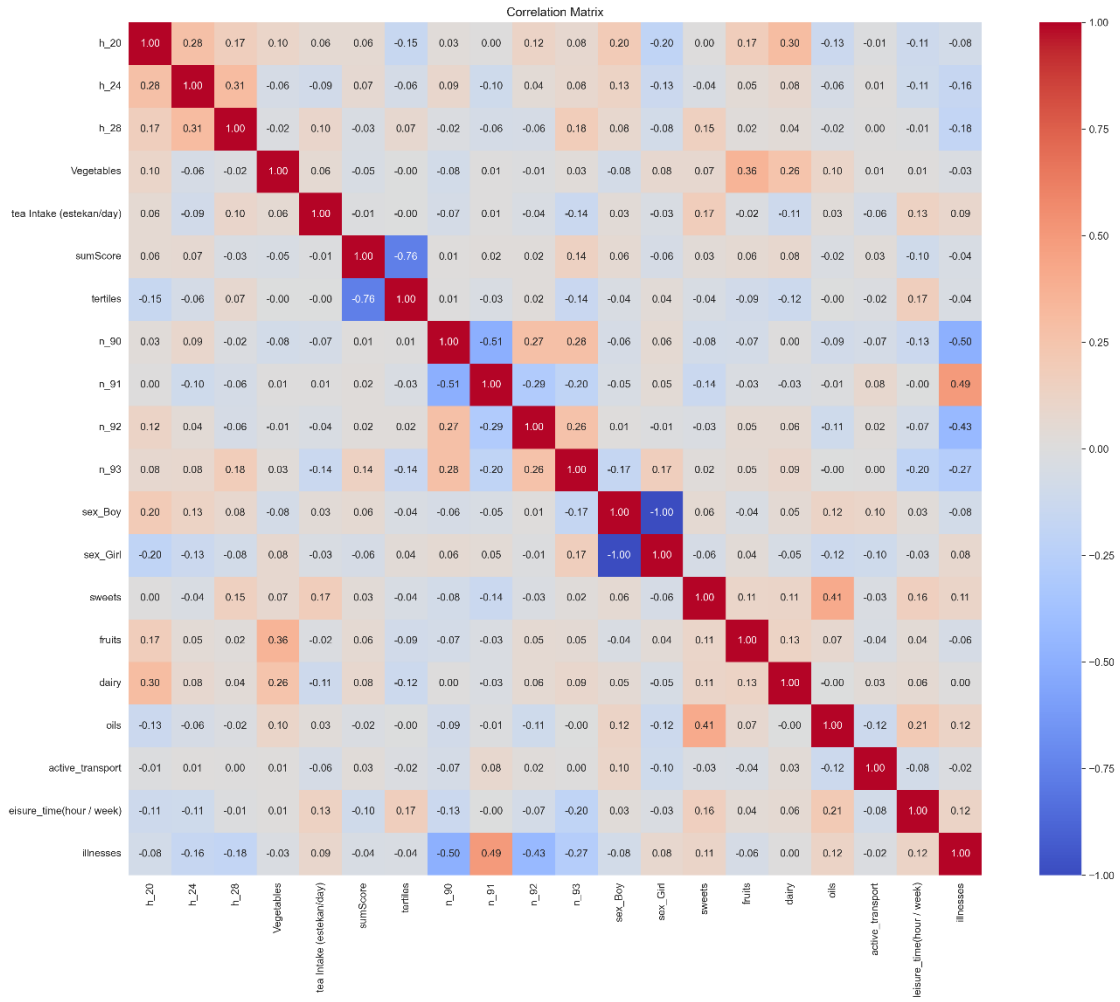
براساس هرم غذایی چربی‌ها و شیرینی‌ها در راس هرم با کمترین ارزش غذایی و سبزیجات و میوه در پایین هرم با بالاترین ارزش غذایی جای گرفته‌اند. همان‌گونه که در شکل۴-۱ مشاهده می‌کنید توزیع میزان مصرف چربی‌ها،شیرینی‌ها، میوه و سبزیجات در بین نوجوانان تقریبا روی میانگین است و این با انحراف از معیار بدست‌آمده در جدول۴-۵ همخوانی دارد اما هرچقدر توزیع داده‌ها در میوه و سبزیجات به سمت بیشینه و در شیرینی‌ها و چربی‌ها به سمت کمینه باشد رژیم غذایی نوجوانان مناسب‌تر است. از سویی می‌تواند نشان‌دهنده‌ی عدم کنترل و نداشتن الگوی تغذیه برای جوانان باشد و این بدان معنا است که انواع مواد غذایی برای نوجوانان قابل دسترس و تنوع غذایی بالایی دارند.

نکته‌ی قابل توجه میزان مصرف بالای لبنیات در بین دانش‌آموزان شهر تبریز است که براساس مطالعات انجام شده شهر تبریز جز شهرهایی هستند که میزان مصرف ویتامینD بالایی دارند. میزان اوقات فراغت دانش‌آموزان در طول هفته از میانگین بیشتر است و می‌تواند اندکی نگران‌کننده باشد زیرا نشان‌دهنده‌ی اهمیت کم‌تر به درس‌خواندن در طول هفته است. اما همین اتفاق منجر شده است که دانش‌آموزان احساس ناخوشی در طول هفته نداشته باشند و از بیماری‌هایی چون سردرد، احساس گیجی، بیخوابی و... به طور مداوم رنج نمی‌برند. از موارد دیگر می‌توان به کم‌تحرکی دانش‌آموزان نیز اشاره کرد.

۴-۴- نمایش هیستوگرام ویژگی‌های اصلی



شکل ۴-۲ماتریس ضریب همبستگی مجموعه داده‌ی اولیه



شکل ۴-۳ماتریس ضریب همبستگی مجموعه‌ی داده‌ی ثانویه

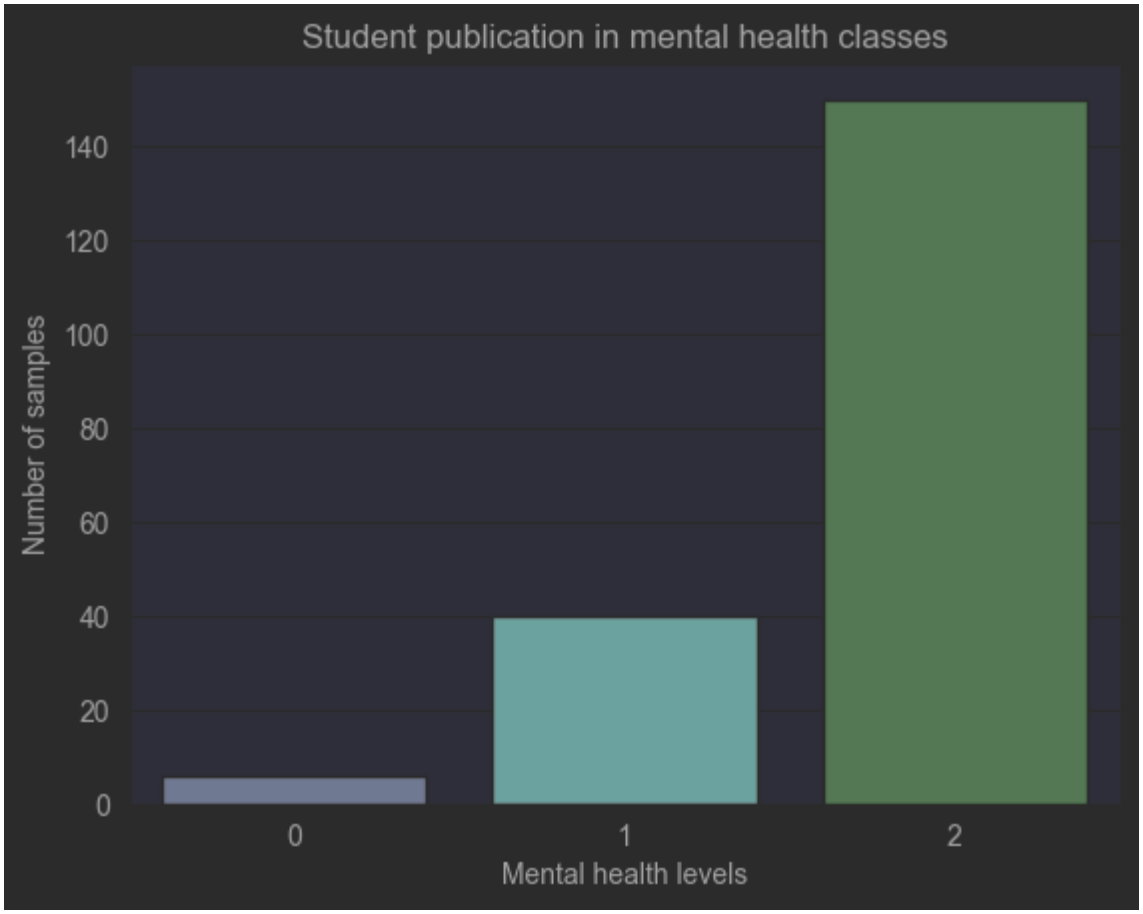
همان‌گونه که در دو شکل بالا مشاهده می‌کنید بسیاری از ویژگی‌ها با ضریب همبستگی پایین در مجموعه‌ی داده‌ی دوم حذف شدند. همچنین با ادغام ویژگی‌ها ضریب همبستگی بین داده‌ها کمی افزایش پیدا کرد اما براساس مشاهدات انجام شده در بخش قبلی و ماتریس همبستگی می‌توان نتیجه‌ی بدست آمده در قسمت قبل که عدم برخورداری از یک الگوی غذایی در بین خانوده‌ها و یا سبک‌زندگی مشخص برای فرزندان می‌باشد را تایید کرد.

به طور کلی نمی‌توان نتیجه گرفت اگر شخصی میزان مصرف سبزیجات بالایی دارد باید میزان مصرف شیرینی کم‌تری داشته باشد ولی اگر بخواهیم براساس هرم غذایی رفتار کنیم باید میزان میوه و سبزیجات در مقابله با شیرینی و چربی‌ها بیشتر باشد. همچنین به‌خاطر تعداد داده‌ی اندک نمی‌توانیم با قطعیت کامل درباره‌ی نتایج بدست‌آمده در تصاویر بالا نظر داد.

در شکل۴-۳ به رابطه مستقیم بین میزان مصرف شیرینی و چربی می‌توان اشاره کرد که نشان‌دهنده‌ی آن است که کسانی که میزان مصرف چربی بالاتری دارند میزان مصرف شیرینی آنها نیز بالاتر است. همچنین رابطه بین ناخوشی‌ها و امتیازی که دانش‌آموزان به وضعیت سلامت خود می‌دهند که نشان‌دهنده‌ی آن است

که داشتن هریک از ناخوشی‌های بالا وزن بیشتری نسبت به سایر دسته‌ها در دادن امتیاز به خود دارند و کسانی که وضعیت سلامتی خود را نامطلوب ارزیابی کرده‌اند از ناخوشی‌های بیشتری رنج می‌برند.

۴-۵- نمایش کمی برای دسته‌بندی هدف



شکل ۴-۴نمایش تعداد دانش‌آموزان در هر دسته

در مجموعه‌داده‌ی اولیه‌ی این پژوهش، هر دانش‌آموز در قالب پرسشی پایانی، نمره‌ای بین ۱ تا ۱۰ را به وضعیت سلامت روان خود اختصاص داده است؛ نمره‌ای که در ظاهر، خودارزیابی ساده‌ای از وضعیت درونی نوجوانان به شمار می‌رود، اما در باطن، حامل پیچیدگی‌های روان‌شناختی، فرهنگی و اجتماعی عمیقی‌ست که نباید از نظر دور بماند.

پرسش بنیادینی که در این‌جا مطرح می‌شود آن است که آیا نوجوانان، در این سن حساس و شکل‌پذیر، توانایی لازم برای تشخیص دقیق شاخص‌های سلامت روان خود را دارند؟ و اگر چنین است، تا چه اندازه در

ارائه‌ی پاسخی صادقانه و بی‌پیرایه، آزادی و امنیت روانی را تجربه می‌کنند؟ در این مطالعه، برای تحلیل دقیق‌تر داده‌ها، نمرات ۱ تا ۴ به‌عنوان وضعیت وخیم، نمرات ۴ تا ۷ به‌عنوان وضعیت آسیب‌پذیر، و نمرات ۷ تا ۱۰ به‌عنوان وضعیت مطلوب طبقه‌بندی شده‌اند؛ اما توزیع این نمرات، حقیقتی تأمل‌برانگیز را نمایان ساخته است.

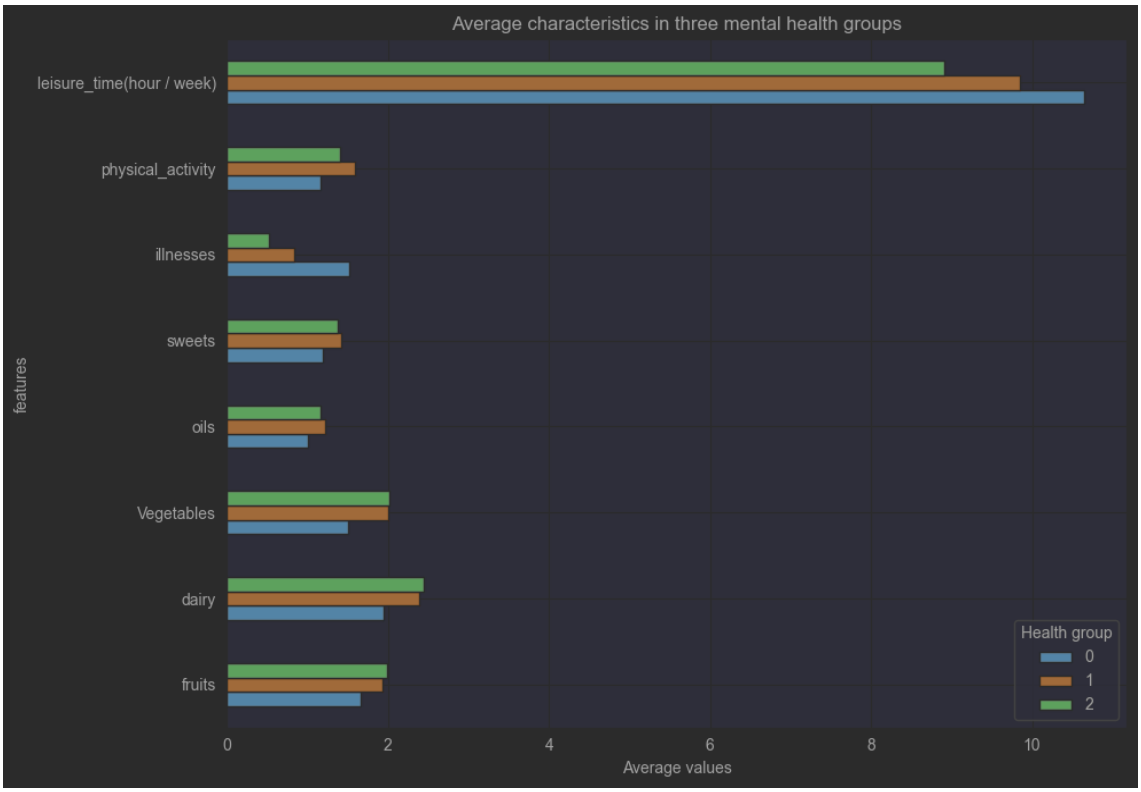
بخش قابل‌توجهی از دانش‌آموزان، خود را در وضعیت مطلوب ارزیابی کرده‌اند؛ این در حالی‌ست که داده‌های رفتاری و پاسخ‌های آنان به سایر پرسش‌ها، در برخی موارد، با این ارزیابی سازگار نیست. به‌نظر می‌رسد که این مسئله، نه از سر ناآگاهی صرف، بلکه بیش از هر چیز، ریشه در باورها و ساختارهای فرهنگی جامعه دارد. در بسیاری از جوامع، به‌ویژه در بسترهای سنتی‌تر، صحبت درباره‌ی بیماری‌های روانی با نوعی انگ اجتماعی همراه است؛ انگی که افراد را به پنهان‌سازی واقعیت وادار می‌کند تا از قضاوت دیگران مصون بمانند.

در واقع، این گرایش ناخودآگاه به نمره‌دهی بالا، گواهی‌ست بر تمایل ذهنی دانش‌آموزان به تلقی خود به‌عنوان فردی سالم، حتی اگر درونی‌ترین احساسات‌شان گواهی غیر از آن بدهد. برای شکستن این چرخه‌ی معیوب، نیازمند فرهنگ‌سازی عمیق، آموزش گسترده و عمومی‌سازی خدمات روان‌درمانی هستیم. تنها در سایه‌ی پذیرش اجتماعی و از میان برداشتن ترس از برچسب خوردن، می‌توان فضایی ایجاد کرد که نوجوانان در آن بتوانند با خود و با دیگران، صادقانه و بی‌واهمه از داوری‌های بیرونی، از رنج‌ها و دغدغه‌های روانی‌شان سخن بگویند.

۴-۶- مقایسه‌ی میانگین ویژگی‌ها در هرگروه

در نمودار۴-۵، میانگین هر یک از ویژگی‌ها بر اساس سه دسته‌ی سلامت روان گروه بحرانی (با رنگ آبی)، آسیب‌پذیر (با رنگ نارنجی) و گروه مطلوب (با رنگ سبز) ترسیم شده است. این نمودار تصویری گویا و ملموس از چگونگی توزیع رفتارها و ویژگی‌های سبک زندگی در میان سطوح مختلف سلامت روان دانش‌آموزان ارائه می‌دهد و نقش مؤثری در درک بهتر روابط میان متغیرهای ورودی و متغیر هدف ایفا می‌کند.

با نگاهی به این نمودار، می‌توان دریافت که افزایش مصرف گروه‌های غذایی مفید همچون سبزیجات، لبنیات و میوه‌ها، ارتباط مستقیمی با بهبود سطح سلامت روان دارد. در مقابل، کاهش میانگین مصرف



شکل ۴–۵ نمودار نوار افقی برای میانگین هر گروه

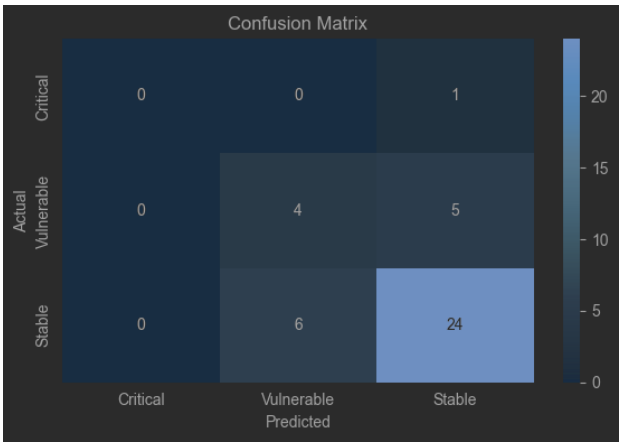
چربی‌ها و شیرینی‌ها در گروه‌های با وضعیت روانی مطلوب، می‌تواند این فرض را تقویت کند که تغذیه سالم نقشی کلیدی در پایداری روانی دانش‌آموزان ایفا می‌کند.

از سوی دیگر، شاخص ناخوشی‌ها نیز در این نمودار تأثیر بسزای خود را نشان داده است؛ به‌گونه‌ای که میانگین بالای این ویژگی در گروه‌های بحرانی، بر وجود پیوندی معنادار میان احساس ناخوشی و افت وضعیت روانی دلالت دارد.

در بررسی سایر شاخص‌ها، نکته‌ای نگران‌کننده نمایان می‌شود: میانگین میزان فعالیت بدنی در میان نوجوانان تبریزی نسبتاً پایین است. این کم‌تحرکی، به‌ویژه در سنین رشد، می‌تواند یکی از عوامل پنهان تضعیف سلامت روان باشد. از دیگر سو، نمودار نشان می‌دهد که افزایش افراطی زمان اوقات فراغت که غالباً صرف فضای مجازی و سرگرمی‌های غیرفعال می‌شود با احساس ناخوشی در ارتباط است و ممکن است یکی از دلایل افت روانی در برخی گروه‌های دانش‌آموزی باشد.

۴-۷- نتایج بدست‌آمده از عملکرد مدل

در گام اول خلاصه‌ای از عملکرد مدل که شامل شاخص‌های کلیدی مانند دقت[1]، دقت مثبت[2]، حساسیت[3] و f1-score است، نشان‌داده می‌شود. در گام بعدی نمودارهای مربوط به دقت و خطای در طول آموزش جهت بررسی مدل گذاشته می‌شود. در هر گام نتایج بدست‌آمده در هر مدل بررسی و با سایر مدل‌ها مقایسه می‌شود.



شکل ۴-۶ ماتریس درهم‌ریختگی پیش‌از تقویت داده‌ها

	precision	recall	f1-score	support
Critical	0.00	0.00	0.00	1
Vulnerable	0.40	0.44	0.42	9
Stable	0.80	0.80	0.80	30
accuracy			0.70	40
macro avg	0.40	0.41	0.41	40
weighted avg	0.69	0.70	0.69	40

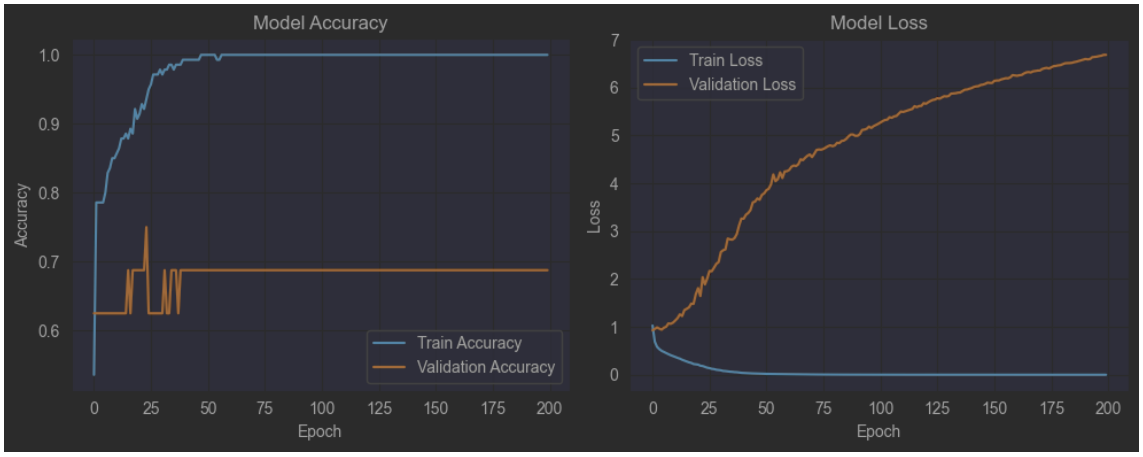
شکل ۴-۷ خلاصه‌ی عملکرد مدل پیش از تقویت داده‌ها

بر اساس نتایج بدست‌آمده در دو شکل بالا مشاهده می‌کنید که در بین داده‌های تست مقادیر مربوط به دسته‌ی "بحرانی" فقط ۱ داده وجود دارد. که همین یک مقدار توسط مدل درست تشخیص داده‌نشده است که باعث شده است دقت مثبت[4] و حساسیت[5] برابر صفر شود و به تبع آن نیز f1-score مقدار صفر را نشان دهد. اما در طرف دیگر ماجرا برای داده‌هایی که در وضعیت سلامتی"مناسب" دارند به دلیل فراوانی داده‌ها مدل توانسته است داده‌های آموزشی را بهتر بیاموزد و پیش‌بینی مناسب تری نسبت به دو دسته‌ی دیگر داشته باشد.

---

[1] Accuracy
[2] Precision
[3] Recall
[4] Precision
[5] Recall

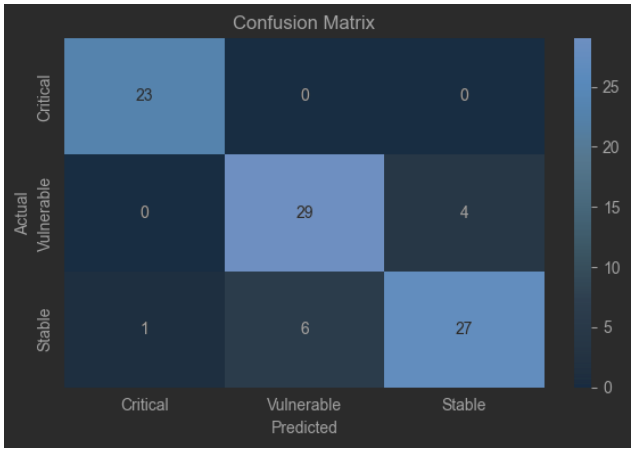
از سویی دقت در داده‌های تست برابر ٪۷۰ است و این بدان معنا است که مدل مقدار کمی یادگیری داشته است ولی تقریبا دارد به صورت رندوم عمل می‌کند و این در میانگین ماکرو که به همه‌ی کلاس‌ها وزن یکسانی داده است مشخص است.



شکل ۴-۸ ارزیابی مدل بر داده‌های آموزشی و اعتبار سنج

همچنین در شکل۴-۸ که دقت داده‌های آموزشی و اعتبار سنج[1] را مشاهده می‌کنید و اختلاف فاحش بین دقت داده‌های آموزشی و اعتبارسنج می‌بینید این نشان‌دهنده‌ی این است که به دلیل حجم کم داده‌ها و حتی سادگی داده‌ها، مدل در همان دوره‌های[2] اول داده‌های آموزشی را حفظ کرده و بیش‌برازش[3] صورت گرفته است. از همین روی در گام بعدی تلاش برای تقویت داده‌ها[4] شده است.

در این گام تعداد داده‌ها به ۴۵۰ نمونه افزایش یافت که تمرکز بر روی دسته‌هایی بود که مقادیر اندکی(مانند دسته‌ی بحرانی) در مجموعه داده‌های اولیه در اختیار داشتیم. که نتایج زیر حاصل شد:



شکل ۴-۹ ماتریس درهم‌ریختگی پس از تقویت داده‌ها

[1] validation
[2] epoch
[3] overfitting
[4] data augmentation

	precision	recall	f1-score	support
Critical	0.96	1.00	0.98	23
Vulnerable	0.83	0.88	0.85	33
Stable	0.87	0.79	0.83	34
accuracy			0.88	90
macro avg	0.89	0.89	0.89	90
weighted avg	0.88	0.88	0.88	90

شکل ۴-۱۰ خلاصه‌ی عملکرد مدل پس از تقویت داده‌ها

همان‌گونه که در شکل بالا مشاهده می‌کنید توزیع داده ها در بخش TP نسبت به شکل۴-۹ توزیع همگون‌تری صورت گرفته است و این نشان دهنده‌ی این است که تقویت داده‌ها در دسته‌های "بحرانی" و "آسیب‌پذیر" صورت گرفته است و انتظار می‌رود معیارهای ارزیابی نیز نسبت به شکل بهبود یافته باشد.

در شکل۴-۱۰ تعداد داده‌های که هر دسته را پشتیبانی[۱] می‌کنند افزایش و مقادیر آنها به یکدیگر نزدیک‌تر شدند از سویی دقت مثبت [۲]و حساسیت[۳] در دسته‌ها افزایش یافته و از سویی اختلاف فاحش بین دقت دسته‌ها وجود ندارد که این می‌تواند نشان‌دهنده‌ی این باشد مدل توانسته روابط پنهان داده‌ها را کشف و الگوهایی را جهت پیش‌بینی مدل یاد بگیرد. از سویی دقت مدل برروی داده‌های تست نیز مقدار ٪۸۸ را نمایش می‌دهد که با تقویت داده نسبت به مدل قبلی افزایش چشمگیری داشته است.

در شکل۴-۱۱ اختلاف خطا بین داده‌های آموزشی و اعتبارسنج کاهش یافته همچنین مدل در همان دوره‌های اول داده‌ها را به طور کامل یاد نمی‌گیرد یا به عبارت دقیق‌تر حفظ نمی‌کند اما هرچقدر مدل به دوره‌های پایانی نزدیک می‌شود و نوسانات در داده‌های اعتبارسنج افزایش می‌یابد نشان‌دهنده‌ی آن است مدل از دوره‌های ۳۰ به بعد دوچار بیش‌برازش شده است و به دلیل سادگی مدل، این نتیجه حاصل می‌شود که با افزایش داده‌ها می‌توانیم از این بیش‌برازش جلوگیری کرد و با توجه به قسمت قبل این افزایش داده‌ها نه تنها در یادگیری مدل اخلالی وارد نکرده است بلکه منجر به یادگیری بهتر مدل، افزایش دقت و کاهش خطا در بین داده‌های آموزشی و اعتبارسنج ما شده است.

شکل ۴-۱۱ ارزیابی مدل بر داده‌های آموزشی و اعتبار سنج با افزایش داده‌ها

۴-۸- جمع‌بندی

در این فصل آنچه بیش از هر چیز خودنمایی می‌کند، چالش ناشی از اندک بودن تعداد نمونه‌های اولیه و عدم توازن بین کلاس‌های هدف است. مدل اولیه‌ی طراحی‌شده با استفاده از یک شبکه‌ی عصبی سه‌لایه، اگرچه توانست الگویی ابتدایی از روابط میان ویژگی‌ها و وضعیت سلامت روان دانش‌آموزان ترسیم کند، اما عملکرد آن، به‌ویژه در دسته‌ی "بحرانی"، به‌دلیل فقدان داده‌های کافی در این کلاس، بسیار محدود بود. شاخص‌های کلیدی همچون دقت مثبت ، حساسیت و F1-score برای این دسته مقدار صفر داشتند که بیانگر ضعف مدل در شناسایی نمونه‌های با وضعیت سلامت وخیم بود.

در گام بعد، با اعمال تکنیک تقویت داده‌ها و استفاده از روش smote، مجموعه‌داده تا حدود ۴۵۰ نمونه گسترش یافت و توازن بهتری میان کلاس‌ها ایجاد شد. این مرحله نقطه‌ی عطفی در فرآیند مدل‌سازی بود، چرا که نتایج حاصل‌شده نشان دادند که مدل توانسته است در تشخیص داده‌های "بحرانی" و "آسیب‌پذیر" عملکرد به‌مراتب بهتری از خود نشان دهد. افزایش معیارهای ارزیابی، به‌ویژه در دسته‌هایی که پیش‌تر نادیده گرفته می‌شدند، گواهی بر این مدعاست.

از سوی دیگر، بهبود دقت مدل از حدود ٪۷۰ به ٪۸۸ در داده‌های تست، و کاهش فاصله‌ی خطا بین داده‌های آموزشی و اعتبارسنج نشان داد که مدل در نسخه‌ی دوم توانست با تعمیم بهتر، الگوهای واقعی‌تر و مؤثرتری را فرا بگیرد .البته، نشانه‌هایی از بیش‌برازش در اواخر فرآیند آموزش مدل نیز دیده شد، که خود حاکی از حساسیت مدل به تعداد داده‌ها و پیچیدگی ساختار آن است.

در نهایت، می‌توان گفت که این پروژه با عبور از یک مسیر گام‌به‌گام شامل تحلیل داده‌های اولیه، مدل‌سازی پایه، تحلیل نتایج اولیه، و سپس تقویت داده و بازآموزی مدل، توانست با وجود محدودیت‌ها، به الگویی قابل قبول برای تحلیل سلامت روان دانش‌آموزان دست یابد. این روند تجربی بر اهمیت داده‌های متوازن، تکنیک‌های پیش‌پردازش دقیق، و انتخاب صحیح مدل‌ها در پروژه‌های داده‌کاوی آموزشی تأکید دارد.

فصل ۵
نتیجه‌گیری و پیشنهادها

۵-۱- نتیجه و جمع‌بندی

پروژه‌ی حاضر با هدف شناسایی و تحلیل عوامل مؤثر بر سلامت روان دانش‌آموزان، بر مبنای داده‌های گردآوری‌شده از پرسش‌نامه‌ی کاسپین۵[1] طراحی و پیاده‌سازی شد. روند پروژه از یک مسیر دقیق و مرحله‌بندی‌شده تبعیت می‌کرد که در گام نخست با بررسی ساختار داده‌ها و تحلیل آماری توصیفی آغاز گردید. در این مرحله، از طریق ابزارهای تصویری نظیر نمودارهای هیستوگرام[2]، نمودار همبستگی و میانگین ویژگی‌ها در هر گروه، شناخت اولیه‌ای از داده‌ها و توزیع آن‌ها حاصل شد.

در ادامه، فرآیند پیش‌پردازش داده‌ها به صورت دقیق انجام گرفت. حذف داده‌های ناقص یا بی‌اثر، یکپارچه‌سازی واحدها و زمان‌ها، ادغام ویژگی‌ها، نرمال‌سازی داده‌های عددی و دسته‌بندی ویژگی‌های کیفی از جمله اقداماتی بودند که منجر به بهبود کیفیت داده‌ها برای مرحله‌ی مدل‌سازی شدند. همچنین، داده‌ها به ویژگی‌های کلیدی کاهش یافتند و با هدف کاهش ابعاد و تمرکز بر متغیرهای مؤثر، انتخاب ویژگی به صورت هدفمند انجام شد.

در فاز نخست مدل‌سازی، یک شبکه‌ی عصبی با ساختار ساده برای مسئله‌ی رگرسیون[3] پیاده‌سازی شد تا سلامت روان به عنوان یک متغیر پیوسته (در بازه‌ی ۱ تا ۱۰) پیش‌بینی شود. اما با توجه به ماهیت محدود داده‌ها و توزیع نامتوازن آن، این مدل نتوانست نتایج دقیقی در بخش‌هایی چون داده‌های بحرانی تولید کند. از این‌رو، در مرحله‌ی بعدی، مسئله به صورت دسته‌بندی سه‌کلاسه (بحرانی، آسیب‌پذیر و مطلوب) بازتعریف شد و مدل شبکه‌ی عصبی با تابع softmax در لایه‌ی نهایی و خطا[4] از نوع sparse-categorical-crossentropy طراحی گردید.

---

[1] Caspian V
[2] histogram
[3] regression
[4] loss

در این مسیر، نتایج اولیه نشان دادند که عملکرد مدل در طبقه‌بندی داده‌های با فراوانی کم (به‌ویژه در کلاس بحرانی) بسیار ضعیف است. به همین جهت، مرحله‌ی جدیدی با عنوان تقویت داده‌ها[1] تعریف شد. با بهره‌گیری از الگوریتم smote ، داده‌ها به ۴۵۰ نمونه افزایش یافتند و به‌ویژه کلاس‌های دارای کمبود داده مورد تقویت قرار گرفتند. نتایج حاصل از این مرحله، بهبود چشمگیری در شاخص‌های ارزیابی مدل همچون دقت کلی[2]، حساسیت[3]، دقت مثبت[4] و F1-score را نشان داد.

مدل نهایی، با دقتی بالغ بر ٪۸۸ در داده‌های تست، توانست عملکردی پایدار و قابل‌قبول از خود نشان دهد. همچنین کاهش اختلاف میان دقت داده‌های آموزشی و اعتبارسنجی نشان از کاهش بیش‌برازش[5] داشت که پیش‌تر به دلیل حجم اندک داده‌ها یکی از چالش‌های اصلی مدل‌سازی بود.

از دیدگاه تحلیلی، نمودارهای میانگین ویژگی‌ها در هر گروه سلامت روان نشان دادند که مصرف بالای میوه، سبزیجات و لبنیات با سلامت روان مطلوب در ارتباط هستند، در حالی که بالا بودن مصرف چربی‌ها و شیرینی‌ها با وضعیت بحرانی مرتبط‌اند. همچنین، زمان بالای اختصاص‌یافته به فعالیت‌های مجازی در اوقات فراغت، و پایین بودن میزان فعالیت بدنی در میان دانش‌آموزان، از جمله عوامل تهدیدکننده‌ی سلامت روان بودند.

در نهایت، می‌توان گفت که پروژه با ترکیبی از داده‌کاوی و تحلیل‌های آماری موفق شد الگوهای پنهان در میان شاخص‌های سبک زندگی و سلامت روان را آشکار سازد و با پیاده‌سازی مدل‌های یادگیری ماشین به راهکاری عملی جهت تحلیل و پایش سلامت روان نوجوانان ایرانی دست یابد. همچنین این نوید را می‌دهد با دسترسی به مجموعه‌ی اصلی داده‌ها می‌تواند گامی موثر در جهت یافتن روابط و الگوهای پنهان در جهت بهبود سلامت نوجوانان با استفاده از ابزارهای داده‌کاوی بردارد.

۵-۲- پیشنهادات

۵-۲-۱ افزایش حجم و تنوع داده‌ها

یکی از مهم‌ترین چالش‌های پروژه، محدودیت در حجم نمونه‌ها و توزیع نامتوازن آن‌ها میان کلاس‌های مختلف (به‌ویژه کلاس بحرانی) بود. این مسئله نه‌تنها دقت مدل را کاهش داد بلکه باعث بیش‌برازش در داده‌های آموزشی شد. در همین راستا پیشنهاد می‌شود در پروژه‌های آتی، مجموعه‌داده‌ی بزرگ‌تری شامل چندین شهر، مقاطع تحصیلی متنوع‌تر و نمونه‌هایی با تنوع فرهنگی بیشتر گردآوری گردد. افزایش داده‌ها به بهبود عملکرد مدل‌های یادگیری ماشین و افزایش اعتبار نتایج کمک شایانی خواهد کرد.

---

[1] data augmentation
[2] accuracy
[3] recall
[4] precision
[5] overfitting

۵-۲-۲ استفاده از مدل‌های پیشرفته‌تر و ترکیبی

در این پروژه از شبکه‌ی عصبی پایه استفاده شد که با وجود سادگی، نتایج قابل قبولی ارائه داد. با این حال، پیشنهاد می‌شود برای مدل‌سازی‌های آتی از الگوریتم‌های پیشرفته‌تری مانند XGBoost، Random Forest، SVM و یا شبکه‌های عمیق‌تر استفاده شود. همچنین، استفاده از روش‌های ترکیبی (ensemble) یا مدل‌هایی با قابلیت attention mechanisms می‌تواند در درک بهتر روابط پنهان میان ویژگی‌ها مفید واقع شود.

۵-۲-۳ طراحی ابزار تعاملی برای تحلیل سلامت روان

با توجه به کارایی تحلیل‌های انجام‌شده، پیشنهاد می‌شود در گام‌های بعدی پروژه، یک ابزار هوشمند تحت وب یا اپلیکیشن موبایل طراحی شود که کاربران (دانش‌آموزان، والدین یا مشاوران مدرسه) بتوانند با وارد کردن اطلاعات سبک زندگی، وضعیت سلامت روان خود یا دیگران را ارزیابی کنند. این ابزار می‌تواند در آینده به عنوان یک سیستم غربالگری سریع برای نهادهای آموزشی و درمانی به‌کار گرفته شود.

۵-۲-۴ گسترش ابعاد تحلیل و وارد کردن متغیرهای روان‌شناختی

پرسش‌نامه‌ی کاسپین، اگرچه داده‌های مفیدی در حوزه‌ی سبک زندگی ارائه می‌دهد، اما فاقد برخی از متغیرهای روان‌شناختی مهم مانند استرس، اضطراب، سطح رضایت از زندگی یا ارتباطات اجتماعی است. پیشنهاد می‌شود در آینده، مجموعه داده‌هایی ترکیبی شامل شاخص‌های روان‌شناختی نیز استفاده گردد تا مدل‌ها تصویر کامل‌تری از سلامت روان ارائه دهند.

۵-۲-۵ بررسی علیت و تحلیل طولی داده‌ها

در این پروژه صرفاً به تحلیل همبستگی‌ها و الگوهای موجود پرداخته شد. اما برای درک بهتر از رابطه علت و معلولی بین سبک زندگی و سلامت روان، توصیه می‌شود داده‌های طولی (longitudinal) و در بازه‌های زمانی مختلف جمع‌آوری و تحلیل شوند. این کار می‌تواند کمک کند تا اثرات طولانی‌مدت یک رفتار خاص مانند تغذیه ناسالم یا کم‌تحرکی، بر وضعیت روانی بررسی گردد.

۵-۲-۶ همکاری با متخصصان حوزه‌های دیگر

اگرچه پروژه از منظر داده‌کاوی و تحلیل آماری قوی ظاهر شده است، همکاری با روان‌شناسان، مشاوران تربیتی، جامعه‌شناسان و متخصصان آموزش می‌تواند به تفسیر بهتر نتایج، تعریف بهتر متغیرها و طراحی پرسش‌نامه‌های دقیق‌تر در نسخه‌های بعدی کمک کند. این نگاه بین‌رشته‌ای، ارزش پژوهش را چندبرابر خواهد کرد.

فصل ۶
تحقیقات پیشین

۶-۱- مقدمه

سلامت روان از موضوعاتی با اهمیت بسیار زیاد می باشد. از آنجایی که بسیاری از بیماری‌ها ناشی از اختلالاتی چون افسردگی، استرس، برنامه‌ی غذایی ناسالم، خواب نامناسب، ناخوشی‌های دوره‌ای و ... می‌باشد. تحقیق و پژوهش‌های بسیاری در این مسیر برای پیشگیری از بیماری‌های ناعلاج انجام شده است. از سویی دیگر هزینه‌های گزاف بیماری‌ها محققان را بر این گماشته که الگوهایی برای سبک زندگی انسان‌ها ارائه دهند که بسیاری از این الگوها با سلامت روان انسان‌ها گره خورده‌اند. از آنجایی که سنین نوجوانی فرصت مناسب برای آموزش و یادگیری انسان‌ها می‌باشد توجه زیادی به آنها شده است و بسیاری از تحقیقات و پژوهش به آنها اختصاص یافته است. از همین روی فرصت مناسبی است که سیری در نتایج بدست آمده توسط محققان صورت گیرد.

۶-۲- بررسی وضعیت سلامت روانی دانش‌آموزان دبیرستانی دختر در سال تحصیلی ۱۳۸۸-۱۳۸۷

در سال ۱۳۸۹ صادقیان به همراه اساتید دانشگاه همدان مطالعه‌ای بر روی سلامت روان دانش‌آموزان دختر دبیرستانی در شهر همدان انجام دادند. ابزار گردآوری شده در این پژوهش پرسشنامه‌ی GHQ-28[1] بود که مقیاس کلی برای افراد بیمار ۲۳ و برای مقیاس‌های فرعی دیگر نقطه‌ی ۷ در نظر گرفته شد. نتایج نشان داد که ۶۰/۲ درصد از واحدهای مورد پژوهش از مقیاس کلی نمره ۲۳ و بالاتر، ۳۶/۷ درصد از مقیاس فرعی خودبیمارانگاری، ۴۶/۵ درصد در مقیاس فرعی اضطراب، ۴۹/۵ درصد از مقیاس اختلال در عملکرد اجتماعی و

[1] Global health questionary

۴۵/۸ درصد از مقیاس افسردگی، نمره ۷ و بالاتر را کسب نموده‌اند. و براساس نتایج بدست آمده سلامت روانی اکثریت واحدهای مورد پژوهش در خطر است که به ترتیب مستعد اختلال در عملکرد اجتماعی، افسردگی، اضطراب و خودبیمارانگاری می‌باشند. پس توجه ویژه مسئولین به این گروه ضروری است[7].

۶-۳- تحلیل همبستگی فعالیت بدنی و سلامت روان در دانش‌آموزان دوره راهنمایی

در این مطالعه میدانی که بر روی ۵۳۲ دانش‌آموز دوره متوسطه از ۱۰ مدرسه در استان H در کشور چین انجام شد،چان به همراه همکارانش به رابطه‌ی بین فعالیت بدنی، سلامت روان و هماهنگی درونی بدن[1] پرداختند. نتایج نشان داد که بین ۲۸٪ تا ۵۴٪ از دانش‌آموزان در جنبه‌های مختلف، درجاتی از مشکلات روانی خفیف را تجربه می‌کنند. بالاترین میزان اختلالات روانی در زمینه‌های اضطراب، فشار تحصیلی و ناپایداری هیجانی گزارش شد.

از نظر میزان فعالیت بدنی، ۵۳٪ از دانش‌آموزان فعالیت کم، ۳۰٪ فعالیت متوسط، و تنها ۱۷٪ فعالیت زیاد داشتند. به‌طور معناداری، پسران نسبت به دختران، و دانش‌آموزان دوره راهنمایی نسبت به دبیرستانی‌ها فعالیت بدنی بیشتری داشتند(همگی با $P<0.01$). به‌ویژه، شدت و مدت زمان ورزش در پسران بیشتر بود.

از نظر آماری، میان نمره فعالیت بدنی و نمرات افسردگی و فشارهای پیرامونی، همبستگی منفی معناداری وجود داشت ($r = -0.103$) برای افسردگی،($P<0.05$) یعنی با افزایش میزان فعالیت بدنی، علائم افسردگی کاهش می‌یابد. هرچند ارتباط مستقیم و معناداری بین فعالیت بدنی و عوامل خودهماهنگی به دست نیامد، اما تحلیل رگرسیون چندمتغیره نشان داد که سه عامل «ناسازگاری بین هارمونی وفعالیت »، «مدت زمان ورزش» و «جنسیت» به‌صورت ترکیبی توانستند ۳۷.۵٪ از تغییرات سلامت روان را پیش‌بینی کنند.

همچنین نمرات میانگین سلامت روان در گروه‌هایی با فعالیت بدنی زیاد، متوسط و کم به‌ترتیب برابر با ۲.۱۶، ۲.۰۷ و ۲.۰۷ بود (بدون تفاوت معنادار کلی)، ولی نمره‌ی افسردگی در این سه گروه به‌صورت معناداری متفاوت بود($P<0.05$) ، به طوری که گروه با فعالیت بدنی بیشتر، افسردگی کمتری داشت[5].

۶-۴- ارتباط بین خوشه‌های اضطراب و اختلال روان‌تنی با عادات سبک زندگی در نوجوانان

یافته‌های این مطالعه که بر روی پرسشنامه‌ی کاسپین۵[2] می‌باشد، با بهره‌گیری از تحلیل خوشه‌ای، چهار گروه روانی را از منظر میزان اضطراب و نشانه‌های روان‌تنی شناسایی کرده است. در میان این گروه‌ها، خوشه‌ای که با اضطراب بالا و اختلالات روان‌تنی فراگیر همراه بود، حدود ۲۰ درصد از شرکت‌کنندگان را در

---

[1] self-harmony
[2] Caspian V

بر می‌گرفت و در آن، دردهای جسمانی، اختلال خواب، احساس بی‌ارزشی، و تنش‌های عصبی به‌مراتب شایع‌تر از دیگر گروه‌ها گزارش شد.

بیشتر از هر چیز، سبک زندگی ناسالم این گروه نگران‌کننده است؛ حذف وعده‌ی صبحانه، مصرف بالای تنقلات شور، شیرینی‌جات و نوشابه، همراه با کاهش مصرف میوه، سبزیجات و لبنیات، تصویری آشنا از تغذیه‌ای آشفته و بی‌تعادل را نشان می‌دهد. این دانش‌آموزان همچنین زمان بیشتری را در برابر صفحه‌های نمایش گذرانده، خواب کوتاه‌تری تجربه کرده، و بیش از دیگران به رفتارهای پرخطر مانند استعمال دخانیات گرایش داشتند.

نتیجه‌ی درخشان این پژوهش آن است که میان سلامت روان و سبک زندگی، پیوندی دوطرفه و پیچیده برقرار است. اضطراب می‌تواند نوجوانان را به‌سوی انتخاب‌های ناسالم سوق دهد، و در مقابل، تغذیه‌ی ضعیف و فعالیت اندک، به تشدید علائم روان‌تنی منجر می‌شود. بر این اساس، سلامت روانی دانش‌آموزان نه تنها نیازمند توجه روان‌شناسان، بلکه در گرو تغییر سبک زندگی، آموزش‌های تغذیه‌ای و کاهش فشارهای محیطی و تحصیلی است. این پژوهش، گامی استوار در جهت طراحی مداخلاتی جامع برای پرورش نسلی شاداب‌تر، سالم‌تر و تاب‌آورتر[4] .

۶-۵- مشارکت کاربر در مراقبت‌های بهداشتی روانی نوجوانان: پروتکلی برای یک بررسی سیستمی

در این مطالعه نویسندگان با رویکردی نظام‌مند، در تلاش برای بهره‌مندی از نوجوانان در راستای تحقیق و بهبود سلامت روانشان هستند. محققان در این پژوهش تجربه‌ی مشارکت نوجوانان در تصمیم‌گیری‌های درمانی، طراحی خدمات روان‌درمانی و ارزیابی اثربخشی آن‌ها را ثبت کرده‌اند. این مطالعه، دریافتی نو از معنای «مشارکت» ارائه می‌دهد؛ مشارکتی نه صوری، بلکه واقعی، مؤثر و مبتنی بر گفت‌وگو و اعتماد متقابل میان متخصص و نوجوان.

از ویژگی‌های بارز این پژوهش، حضور فعال دو نوجوان به‌عنوان هم‌پژوهشگر در تمامی مراحل تحقیق است؛ حضوری که نه تنها به درک مفهومی پروژه افزوده، بلکه تأکیدی است بر این‌که سلامت روان بدون مشارکت صاحبان تجربه، کامل نمی‌شود. آنان در جست‌وجوی منابع، تحلیل متون، نقد مقالات و ترویج یافته‌ها نقش داشتند و از طریق تجربه‌ی زیسته‌ی خود، لایه‌های پنهان و ناگفته‌ی پژوهش‌های بالینی را روشن ساخته‌اند.

نتیجه‌ی این تلاش، نه فقط شناسایی شکاف‌های پژوهشی و نارسایی‌های ساختاری در نظام سلامت روان نوجوانان، بلکه ارائه‌ی مدلی است مبتنی بر شفافیت، احترام، و عدالت در فرایند درمان .پژوهشگران این مقاله بر آن‌اند که با گردآوری شواهد علمی، بستری را فراهم آورند که در آن، نوجوانان نه‌تنها شنیده شوند، بلکه در ساختن آینده‌ی سلامت روان خود، نقش ایفا کنند[3].

۶-۶- جمع‌بندی

در دهه‌های اخیر، توجه به سلامت روان دانش‌آموزان به‌عنوان یکی از ارکان بنیادین توسعه پایدار در نظام‌های آموزشی، بیش از پیش مورد تأکید قرار گرفته است. نتایج پژوهش‌های متعدد در این حوزه بیانگر آن است که اختلالات روانی نظیر اضطراب، افسردگی، احساس بی‌کفایتی و مشکلات رفتاری، سهم قابل‌توجهی از چالش‌های دوران نوجوانی را تشکیل می‌دهند و می‌توانند آثار بلندمدتی بر رشد شخصیتی، عملکرد تحصیلی و کیفیت زندگی اجتماعی افراد برجای گذارند.

مطالعات پیشین به‌طور مستمر بر اهمیت عوامل مختلفی همچون تغذیه سالم، فعالیت بدنی منظم، خواب کافی، تعاملات اجتماعی مثبت، و سبک زندگی متعادل در ارتقای سلامت روان تأکید کرده‌اند. یافته‌ها نشان می‌دهند که سلامت روان، پدیده‌ای چندبعدی و پویا است که تحت تأثیر تعامل پیچیده‌ای از عوامل زیستی، روانی، اجتماعی و محیطی قرار دارد. از این‌رو، بررسی علمی این عوامل و شناخت الگوهای مؤثر در شکل‌گیری وضعیت روانی دانش‌آموزان، ضرورتی انکارناپذیر در تدوین سیاست‌های آموزشی و بهداشتی به شمار می‌آید.

همچنین، رویکردهای نوین پژوهشی در سال‌های اخیر با تأکید بر «مشارکت فعال نوجوانان در فرآیندهای تصمیم‌گیری مرتبط با سلامت روان»، مسیر جدیدی در تولید دانش و ارائه خدمات مبتنی بر نیازهای واقعی این گروه سنی گشوده‌اند. این تغییر نگرش، از مدل‌های سنتی درمان‌محور به سمت مدل‌های مشارکت‌محور، نشانگر درک عمیق‌تری از حقوق نوجوانان و اهمیت درک دیدگاه‌های آنان در بهبود فرآیندهای درمانی و پیشگیرانه است.

در مجموع، مرور مطالعات پیشین آشکار می‌سازد که سلامت روان دانش‌آموزان نه‌تنها یک موضوع پزشکی یا روان‌شناختی صرف، بلکه مسئله‌ای چندرشته‌ای و راهبردی در نظام‌های آموزشی و اجتماعی است. تحلیل دقیق یافته‌های موجود می‌تواند بستر مناسبی برای طراحی مداخلات مؤثر، سیاست‌گذاری علمی و ارتقاء شاخص‌های سلامت روانی در میان نسل نوجوان فراهم آورد.

منابع:

[1] J. Grus, “Data science from scratch first principles with Python,” O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, United States of America, 2019.

[2] M. Jahanbakhsh, J. Aghadavodian, R. Kelishadi, M. Sattari, “Identifying the Relationship between Different Factors Affecting 13 to 18-Year-Old Students' Mental Health in Different Regions of Iran Using Random Forest Technique” in Journal of Advances in Medical and Biomedical Research, May-Jun 2022

[3] P. Viksveen, S.E. Bjønness, S.H. Berg, N.E. Cardenas, J.R. Game, K. Aase, M. Storm, “BMJ open user involvement in adolescents' mental healthcare: protocol for a systematic review” 10.1136/bmjopen-2017-018800 on 21 December 2017.

[4] S.S. Daniali1, R. Riahi, M. Taheri, T. Aminaei, R. Heshmat, M. Qorbani, R. Kelishadi “Association between clusters of anxiety and psychosomatic disorder with lifestyle habits in children and adolescents: the CASPIAN-V study” Electronic Physician (ISSN: 2008-5842), October-December 2018.

[5] Waichun Chan, “Correlation Analysis of Physical Exercise and Mental Health of Middle School Students: An Empirical Study based on 10 Middle Schools in H Province,” Atlantis Press, Diocesan Boys' School, Hong Kong, China, 2018.

[6] R. Kelishadi, G. Ardalan, R. Gheiratmand, R. Majdzadeh, M. Hosseini, M. M. Gouya, E. M. Razaghi, A. Delavari, M. Motaghian, H. Barekati, M. S. Mahmoud-Arabi “Thinness, overweight and obesity in a national sample of Iranian children and adolescents: CASPIAN Study” 05 March 2007

[7] E. Sadeghian, M. Moghadari Kosha, S Gorji “The Study of Mental Health Status in High School Female Students in Hamadan City” Hamadan University of Medical Sciences, Islamic Republic of Iran, Dec 2010

[8] Z. Ahadi, G. Shafiee, M. Qorbani, S. Sajedinejad, R. Kelishadi, S.M. Arzaghi, B. Larijani, R. Heshmat. “An overview on the successes, challenges, and future perspective of a national school-based surveillance program: the CASPIAN study” 20 December 2014

[9] Mental disorders affect one in four people [Internet]. 2001 [Available from: https://www.who.int/whr/2001/media_c entre/press_release/en/.

[10] A.A. Noorbala, S.B. Yazdi, M. Yasamy, K. Mohammad. “Mental health survey of the adult population in Iran. The British Journal of Psychiatry.” 2004; 184(1): 70-3.

[11] S.G. Alonso, I. de La Torre-Díez, S. Hamrioui, M. López-Coronado, D.C. Barreno, L.M. Nozaleda, et al. “Data mining algorithms and techniques in mental health: a systematic review. Journal of Medical Systems.” 2018; 42(9): 161.

[12] S.S. Rahman. “An Application of Data Mining of Mental Health Data. Association for Information Systems Electronic Library (AISeL).” 2019; 3(22): 1-6.

[۱۳] پیمایش-نظام-مراقبت-پیشگیری-از-رفتارها-و-عوامل-مخاطره-آمیز-دانش-آموزان-(کاسپین)/پیمایش-کاسپین-۱۳۹۵-(اطلاعات-کلی)/https://nihr.tums.ac.ir