شناسایی عوامل مؤثر و پیش‌بینی بیماری خشکی چشم با استفاده از الگوریتم‌های داده‌کاوی در محیط RapidMiner

علی وطن‌خواه [۱]، یاسمن عباسی [۲]

۱- کارشناسی ارشد مهندسی کامپیوتر، گروه آموزشی مهندسی برق و کامپیوتر، دانشگاه ملی مهارت، تهران، ایران

ایمیل: alivatankhah1988@gmail.com

۲- کارشناسی مهندسی کامپیوتر، گروه آموزشی مهندسی برق و کامپیوتر، دانشگاه ملی مهارت، تهران، ایران

ایمیل: yasamanabc1382@gmail.com

چکیده

خشکی چشم یکی از شایع‌ترین بیماری‌های سطح چشم است که به دلیل ماهیت چندعاملی و علائم متنوع، تشخیص آن همواره با چالش همراه بوده است. افزایش استفاده از وسایل دیجیتال، تغییر سبک زندگی و عوامل محیطی موجب افزایش شیوع این بیماری در سال‌های اخیر شده است. ازاین‌رو، استفاده از روش‌های هوشمند برای پیش‌بینی و تشخیص زودهنگام بیماری می‌تواند نقش مهمی در کاهش عوارض و کمک به تصمیم‌گیری پزشکان داشته باشد. هدف از این پژوهش، مقایسه عملکرد الگوریتم‌های داده‌کاوی در پیش‌بینی بیماری خشکی چشم با استفاده از داده‌های بالینی و دموگرافیک بیماران در محیط RapidMiner است. در این مطالعه از مجموعه‌داده Dry Eye Disease شامل ۲۰۰۰۰ نمونه و ۲۶ ویژگی استفاده شد. پس از انجام مراحل پیش‌پردازش شامل بررسی داده‌های گمشده، حذف داده‌های پرت و انتخاب ویژگی، چهار الگوریتم Decision Tree، Random Forest، K-Nearest Neighbor و Gradient Boosted Trees پیاده‌سازی شدند. عملکرد مدل‌ها با استفاده از معیارهای Accuracy، Precision و Recall مورد ارزیابی قرار گرفت. نتایج نشان داد که ویژگی‌های قرمزی و احساس خارش در چشم بیشترین تأثیر را در پیش‌بینی بیماری خشکی چشم دارند. همچنین الگوریتم Random Forest با دقت ۶۸/۶۸ درصد نسبت به سایر الگوریتم‌های بررسی‌شده عملکرد بهتری در پیش‌بینی بیماری ارائه کرده است. یافته‌های پژوهش نشان می‌دهد که روش‌های داده‌کاوی می‌توانند به عنوان ابزاری مناسب در غربالگری اولیه و توسعه سامانه‌های پشتیبان تصمیم در حوزه چشم‌پزشکی مورد استفاده قرار گیرند.

کلمات کلیدی: خشکی چشم، داده‌کاوی، یادگیری ماشین، RapidMiner، Random Forest، Decision Tree، K-Nearest Neighbor ، Gradient Boosted Trees

۱. مقدمه

بیماری خشکی چشم یکی از شایع‌ترین بیماری‌های سطح چشم است که در اثر اختلال در هموستاز لایه اشکی ایجاد می‌شود و با ناپایداری فیلم اشکی، افزایش اسمولاریته، التهاب سطح چشم و اختلالات عصبی-حسی همراه است [۱]. این بیماری با علائمی مانند احساس خشکی، سوزش، خارش، احساس وجود جسم خارجی، قرمزی چشم، تاری دید و خستگی چشم شناخته می‌شود و در صورت عدم تشخیص و درمان مناسب می‌تواند کیفیت زندگی، عملکرد بینایی و فعالیت‌های روزمره بیماران را به‌طور قابل توجهی کاهش دهد [۱,۲].

بر اساس گزارش TFOS DEWS II، خشکی چشم یک بیماری چندعاملی است و شیوع آن در کشورهای مختلف بین ۵ تا ۵۰ درصد گزارش شده است. این اختلاف به عواملی نظیر شرایط آب‌وهوایی، سبک زندگی، سن، جنسیت، بیماری‌های زمینه‌ای و تفاوت در معیارهای تشخیصی بستگی دارد [۲]. در سال‌های اخیر، افزایش استفاده از رایانه، تلفن همراه و سایر نمایشگرهای دیجیتال، همراه با کاهش نرخ پلک زدن و قرار گرفتن طولانی‌مدت در محیط‌های خشک یا دارای تهویه، موجب افزایش بروز این بیماری به‌ویژه در افراد جوان شده است [۲].

تشخیص بیماری خشکی چشم معمولاً بر اساس ترکیبی از علائم بالینی، آزمون‌های تخصصی و پرسشنامه‌های استاندارد انجام می‌شود. آزمون‌هایی مانند Schirmer Test، TBUT[1] و رنگ‌آمیزی سطح چشم از متداول‌ترین روش‌های تشخیص هستند اما هر یک دارای محدودیت‌هایی بوده و نتایج آن‌ها ممکن است تحت تأثیر شرایط محیطی یا وضعیت بیمار قرار گیرد [۳]. به همین دلیل، توسعه روش‌های هوشمند برای تحلیل داده‌های بیماران و کمک به تشخیص بیماری، به یکی از موضوعات مورد توجه پژوهشگران تبدیل شده است.

در دهه‌های اخیر، پیشرفت فناوری اطلاعات و افزایش حجم داده‌های پزشکی، زمینه را برای استفاده از روش‌های داده‌کاوی و یادگیری ماشین در حوزه سلامت فراهم کرده است. داده‌کاوی با استخراج الگوهای پنهان از میان حجم زیادی از داده‌ها، می‌تواند روابط میان عوامل خطر، ویژگی‌های بالینی و وضعیت بیماری را شناسایی کرده و در پیش‌بینی بیماری‌ها مورد استفاده قرار گیرد [۴,۵]. به همین دلیل، کاربرد الگوریتم‌های طبقه‌بندی در تشخیص بیماری‌هایی مانند دیابت، بیماری‌های قلبی، سرطان و بیماری‌های چشمی به‌طور چشمگیری افزایش یافته است.

در مطالعات مختلف، استفاده از الگوریتم‌های یادگیری ماشین در تشخیص بیماری‌های چشمی نتایج امیدوارکننده‌ای داشته است. الگوریتم‌هایی مانند Decision Tree، Random Forest، Support Vector Machine، K-Nearest Neighbor و روش‌های مبتنی بر Gradient Boosting به دلیل توانایی در شناسایی الگوهای پیچیده و مدیریت داده‌های چندبعدی، کاربرد گسترده‌ای در مسائل طبقه‌بندی پزشکی پیدا کرده‌اند [۹–۵]. با این حال، عملکرد این الگوریتم‌ها به ویژگی‌های مجموعه‌داده، کیفیت پیش‌پردازش و انتخاب ویژگی‌ها وابسته است و مقایسه آن‌ها بر روی داده‌های مختلف می‌تواند به انتخاب مناسب‌ترین مدل برای هر مسئله کمک کند.

[1] Tear Break-Up Time

هدف پژوهش حاضر، شناسایی مهم‌ترین عوامل مؤثر بر بیماری جهت بهبود عملکرد مدل‌های پیش‌بینی و درک بهتر عوامل خطر، بررسی و مقایسه عملکرد چهار الگوریتم Decision Tree، Random Forest، K-Nearest Neighbor و Gradient Boosted Trees در پیش‌بینی بیماری خشکی چشم با استفاده از مجموعه‌داده Dry Eye Disease و تعیین مناسب‌ترین الگوریتم بر اساس معیارهای استاندارد ارزیابی است.

۲. مبانی نظری

۱.۲. بیماری خشکی چشم

بیماری خشکی چشم یکی از شایع‌ترین بیماری‌های سطح چشم است که به‌عنوان یک بیماری چندعاملی سطح چشم شناخته می‌شود. این بیماری با از بین رفتن هموستاز لایه اشکی مشخص شده و عواملی مانند ناپایداری فیلم اشکی، افزایش اسمولاریته، التهاب سطح چشم، آسیب سلول‌های اپیتلیال و اختلالات عصبی-حسی در ایجاد و پیشرفت آن نقش دارند [۱].

فیلم اشکی ساختاری سه‌لایه شامل لایه لیپیدی، لایه آبی و لایه موسینی است که به ترتیب توسط غدد میبومین، غدد اشکی و سلول‌های جامی ملتحمه تولید می‌شوند. این لایه‌ها با همکاری یکدیگر موجب حفظ رطوبت سطح چشم، تأمین اکسیژن و مواد مغذی قرنیه، ایجاد سطح اپتیکی یکنواخت و محافظت از چشم در برابر عوامل بیماری‌زا می‌شوند. هرگونه اختلال در عملکرد این لایه‌ها می‌تواند منجر به ناپایداری فیلم اشکی و در نهایت بروز بیماری خشکی چشم شود [۱].

خشکی چشم به‌طور کلی به دو گروه اصلی تقسیم می‌شود:

الف) خشکی چشم ناشی از کمبود ترشح اشک؛ در این نوع بیماری، غدد اشکی قادر به تولید مقدار کافی اشک نیستند. بیماری‌هایی مانند سندرم شوگرن، افزایش سن و برخی داروها از مهم‌ترین عوامل ایجاد این نوع خشکی چشم محسوب می‌شوند [۱].

ب) خشکی چشم ناشی از تبخیر بیش از حد اشک؛ این نوع، شایع‌ترین فرم بیماری است و عمدتاً در اثر اختلال عملکرد غدد میبومین ایجاد می‌شود. کاهش کیفیت لایه لیپیدی موجب افزایش سرعت تبخیر اشک و ناپایداری فیلم اشکی خواهد شد [۲].

۲.۲. عوامل مؤثر بر بیماری خشکی چشم

بروز بیماری خشکی چشم تحت تأثیر عوامل متعددی قرار دارد و معمولاً نتیجه تعامل عوامل فردی، محیطی و بیماری‌های زمینه‌ای است. بر اساس گزارش TFOS DEWS II، افزایش سن یکی از مهم‌ترین عوامل خطر این بیماری محسوب می‌شود؛ به‌گونه‌ای که با افزایش سن، عملکرد غدد اشکی و غدد میبومین کاهش یافته و احتمال ابتلا به خشکی چشم افزایش می‌یابد. همچنین شیوع بیماری در زنان بیش از مردان گزارش شده است که این موضوع به تغییرات هورمونی، به‌ویژه در دوران یائسگی، نسبت داده می‌شود [۲].

عوامل محیطی مانند قرار گرفتن طولانی‌مدت در معرض باد، گردوغبار، هوای خشک، استفاده از سیستم‌های تهویه، آلودگی هوا و کاهش رطوبت محیط نیز از جمله عوامل مؤثر در ایجاد بیماری هستند. علاوه بر این، استفاده طولانی از رایانه، تلفن همراه و سایر نمایشگرهای دیجیتال باعث کاهش تعداد پلک زدن و افزایش تبخیر اشک شده و خطر ابتلا به خشکی چشم را افزایش می‌دهد [۲].

از دیگر عوامل خطر می‌توان به استفاده از لنزهای تماسی، مصرف برخی داروها مانند آنتی‌هیستامین‌ها و داروهای ضدافسردگی، بیماری‌های خودایمنی، دیابت، جراحی‌های انکساری چشم و کمبود ویتامین A اشاره کرد [۳].

۳.۲. روش‌های تشخیص بیماری خشکی چشم

تشخیص بیماری خشکی چشم معمولاً بر اساس ترکیب علائم بیمار، معاینات بالینی و آزمون‌های تخصصی انجام می‌شود. از مهم‌ترین آزمون‌های تشخیصی می‌توان به Schirmer Test برای اندازه‌گیری میزان ترشح اشک، Tear Break-Up Time برای بررسی پایداری فیلم اشکی و رنگ‌آمیزی فلورسئین برای ارزیابی آسیب سطح قرنیه اشاره کرد [۳].

با وجود کاربرد گسترده این آزمون‌ها، نتایج آن‌ها ممکن است تحت تأثیر شرایط محیطی، مهارت معاینه‌کننده و وضعیت لحظه‌ای بیمار قرار گیرد. به همین دلیل، در سال‌های اخیر استفاده از روش‌های مبتنی بر داده‌کاوی و یادگیری ماشین به‌عنوان ابزارهای کمکی در تشخیص و پیش‌بینی بیماری مورد توجه پژوهشگران قرار گرفته است [۴].

۴.۲. داده‌کاوی در پزشکی

با گسترش فناوری اطلاعات و استقرار سامانه‌های الکترونیکی در مراکز درمانی، حجم عظیمی از داده‌های پزشکی شامل اطلاعات دموگرافیک، سوابق بیماری، نتایج آزمایش‌ها، تصاویر پزشکی و اطلاعات درمانی تولید می‌شود. بخش قابل‌توجهی از این داده‌ها بدون استفاده باقی می‌مانند، درحالی‌که استخراج دانش از آن‌ها می‌تواند نقش مهمی در تشخیص بیماری‌ها، پیش‌بینی پیامدهای درمان و پشتیبانی از تصمیم‌گیری بالینی داشته باشد. در چنین شرایطی، داده‌کاوی به عنوان یکی از مهم‌ترین شاخه‌های علم داده، امکان کشف الگوها و روابط پنهان موجود در داده‌ها را فراهم می‌کند [۴,۵].

داده‌کاوی فرآیندی است که با استفاده از روش‌های آماری، یادگیری ماشین و هوش مصنوعی، اطلاعات ارزشمند را از میان حجم زیادی از داده‌ها استخراج می‌کند. این فرآیند معمولاً شامل مراحل جمع‌آوری داده، پیش‌پردازش، انتخاب ویژگی، مدل‌سازی، ارزیابی و استخراج دانش است [۵]. در علوم پزشکی، داده‌کاوی کاربردهای گسترده‌ای از جمله پیش‌بینی بیماری‌ها، طبقه‌بندی بیماران، شناسایی عوامل خطر، تحلیل سوابق درمانی و توسعه سامانه‌های پشتیبان تصمیم‌گیری بالینی دارد [۴].

در سال‌های اخیر، به‌کارگیری داده‌کاوی در چشم‌پزشکی نیز رشد قابل‌توجهی داشته است. پژوهشگران از الگوریتم‌های یادگیری ماشین برای تشخیص بیماری‌هایی مانند رتینوپاتی دیابتی، گلوکوم، دژنراسیون ماکولا و خشکی چشم استفاده کرده‌اند. این روش‌ها با تحلیل هم‌زمان متغیرهای مختلف، قادر به شناسایی الگوهایی هستند که ممکن است در بررسی‌های معمول بالینی به‌سادگی قابل تشخیص نباشند [۴].

یکی از مهم‌ترین مزایای داده‌کاوی، توانایی تحلیل داده‌های چندبعدی و تولید مدل‌های پیش‌بینی‌کننده با قابلیت تعمیم مناسب است. البته کیفیت نتایج این روش‌ها به عواملی مانند کیفیت داده‌ها، نحوه پیش‌پردازش، انتخاب ویژگی‌ها و انتخاب الگوریتم مناسب وابسته است [۵,۶]. به همین دلیل، انجام مراحلی مانند حذف داده‌های پرت، مدیریت داده‌های گمشده و انتخاب ویژگی‌های مؤثر، پیش از آموزش مدل‌ها ضروری است.

در پژوهش حاضر نیز، پس از انجام مراحل پیش‌پردازش و انتخاب ویژگی، از الگوریتم‌های طبقه‌بندی برای پیش‌بینی بیماری خشکی چشم استفاده شده است. این رویکرد علاوه بر افزایش دقت مدل، موجب کاهش پیچیدگی محاسباتی و بهبود قابلیت تفسیر نتایج می‌شود.

۵.۲. الگوریتم‌های مورد استفاده

در این پژوهش، چهار الگوریتم پرکاربرد طبقه‌بندی شامل Decision Tree، Random Forest، K-Nearest Neighbor و Gradient Boosted Trees برای پیش‌بینی بیماری خشکی چشم مورد استفاده قرار گرفتند. انتخاب این الگوریتم‌ها به دلیل کاربرد گسترده آن‌ها در مسائل طبقه‌بندی پزشکی، قابلیت پیاده‌سازی در نرم‌افزار RapidMiner و امکان مقایسه عملکرد آن‌ها در شرایط یکسان انجام شده است.

۱.۵.۲. الگوریتم درخت تصمیم

درخت تصمیم یکی از رایج‌ترین روش‌های طبقه‌بندی است که داده‌ها را بر اساس مجموعه‌ای از قوانین تصمیم به شاخه‌های مختلف تقسیم می‌کند. در این الگوریتم، هر گره داخلی نشان‌دهنده یک ویژگی، هر شاخه بیانگر نتیجه یک تصمیم و هر گره برگ بیانگر کلاس نهایی است. مهم‌ترین مزیت درخت تصمیم، سادگی، سرعت اجرا و قابلیت تفسیر بالا است؛ با این حال، این الگوریتم ممکن است در داده‌های پیچیده دچار بیش‌برازش شود [۷].

۲.۵.۲. الگوریتم جنگل تصادفی

جنگل تصادفی یک روش یادگیری گروهی است که از ترکیب تعداد زیادی درخت تصمیم مستقل تشکیل می‌شود. هر درخت با استفاده از نمونه‌گیری تصادفی از داده‌ها و انتخاب تصادفی ویژگی‌ها ساخته می‌شود و تصمیم نهایی بر اساس رأی اکثریت درخت‌ها اتخاذ می‌شود. این ویژگی موجب کاهش واریانس مدل، افزایش پایداری و کاهش احتمال بیش‌برازش می‌شود. به همین دلیل، Random Forest یکی از موفق‌ترین الگوریتم‌های طبقه‌بندی در مسائل پزشکی محسوب می‌شود [۷].

۳.۵.۲. الگوریتم نزدیک‌ترین همسایه

الگوریتم KNN یک روش یادگیری مبتنی بر نمونه است که نمونه جدید را بر اساس شباهت آن با نزدیک‌ترین نمونه‌های موجود در مجموعه آموزش طبقه‌بندی می‌کند. در این الگوریتم، فاصله بین نمونه‌ها معمولاً با معیار فاصله اقلیدسی محاسبه می‌شود. اگرچه KNN ساختار ساده‌ای دارد، اما عملکرد آن به انتخاب مقدار مناسب K و مقیاس‌بندی ویژگی‌ها وابسته است [۸].

۴.۵.۲. الگوریتم Gradient Boosted Trees

Gradient Boosted Trees یکی از روش‌های پیشرفته یادگیری گروهی است که مدل را به‌صورت مرحله‌ای ایجاد می‌کند. در این روش، هر درخت جدید با هدف کاهش خطای مدل قبلی ساخته می‌شود و در نهایت مجموعه‌ای از درخت‌ها یک مدل قوی را تشکیل می‌دهند. این الگوریتم توانایی مناسبی در مدل‌سازی روابط پیچیده و غیرخطی دارد، اما تنظیم پارامترهای آن نسبت به Random Forest حساس‌تر است [۹].

در این پژوهش، عملکرد چهار الگوریتم مذکور با استفاده از مجموعه‌داده Dry Eye Disease و در محیط RapidMiner مورد ارزیابی قرار گرفت تا مناسب‌ترین مدل برای پیش‌بینی بیماری خشکی چشم بر اساس معیارهای استاندارد ارزیابی تعیین شود.

۳. مطالعات مرتبط

بررسی مطالعات پیشین نشان می‌دهد که بیشتر پژوهش‌های انجام‌شده بر توسعه روش‌های هوش مصنوعی برای تشخیص بیماری یا شناسایی عوامل خطر مؤثر بر خشکی چشم متمرکز بوده‌اند. در پژوهش حاضر، علاوه بر مقایسه عملکرد چهار الگوریتم طبقه‌بندی در محیط RapidMiner، از روش Weight by Chi-Squared Statistic برای شناسایی مهم‌ترین شاخص‌های مؤثر بر بیماری استفاده شده است. این رویکرد علاوه بر ارائه یک مدل پیش‌بینی، امکان تعیین میزان اهمیت هر ویژگی را نیز فراهم می‌کند که می‌تواند در تحلیل عوامل خطر و طراحی سامانه‌های پشتیبان تصمیم سودمند باشد.

در سال‌های اخیر، استفاده از روش‌های داده‌کاوی و یادگیری ماشین در تشخیص و پیش‌بینی بیماری‌های چشمی مورد توجه پژوهشگران قرار گرفته است. هدف اصلی این مطالعات، توسعه مدل‌هایی با دقت بالا برای کمک به تشخیص زودهنگام بیماری‌ها و کاهش وابستگی به روش‌های سنتی تشخیص است.

Beam و Kohane با بررسی کاربرد یادگیری ماشین در پزشکی بیان کردند که الگوریتم‌های داده‌کاوی می‌توانند با تحلیل حجم بالای داده‌های سلامت، الگوهای پنهان را شناسایی کرده و نقش مؤثری در توسعه سامانه‌های پشتیبان تصمیم‌گیری بالینی ایفا کنند [۴].

علاوه بر این، Breiman در مقاله کلاسیک خود، الگوریتم Random Forest را معرفی کرد و نشان داد که استفاده از روش‌های یادگیری گروهی موجب افزایش دقت مدل، کاهش حساسیت نسبت به نویز و بهبود قابلیت تعمیم می‌شود. این ویژگی‌ها سبب شده است که Random Forest در بسیاری از مسائل طبقه‌بندی پزشکی به یکی از پرکاربردترین الگوریتم‌ها تبدیل شود [۷].

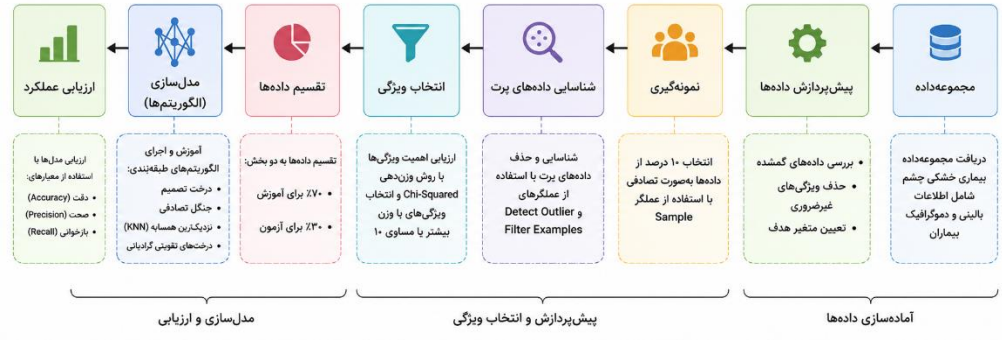
همچنین Lu و همکاران در یک مطالعه مروری جامع، کاربرد هوش مصنوعی در تشخیص، غربالگری و مدیریت بیماری خشکی چشم را بررسی کردند. نتایج این پژوهش نشان داد که الگوریتم‌های یادگیری ماشین و یادگیری عمیق با تحلیل داده‌های بالینی و تصاویر سطح چشم می‌توانند دقت تشخیص بیماری را افزایش داده و به‌عنوان ابزار کمکی در تصمیم‌گیری بالینی مورد استفاده قرار گیرند [۱۰].

Britten-Jones و همکاران در یک مطالعه مروری و فراتحلیل، عوامل خطر مرتبط با بیماری خشکی چشم را بررسی کردند. نتایج نشان داد که عواملی مانند افزایش سن، جنسیت مؤنث، کیفیت پایین خواب، استفاده طولانی‌مدت از نمایشگرهای دیجیتال و برخی بیماری‌های سیستمیک از مهم‌ترین عوامل خطر این بیماری محسوب می‌شوند [۱۱].

مطالعات فوق نشان می‌دهد که انتخاب الگوریتم مناسب و انجام صحیح مراحل پیش‌پردازش و انتخاب ویژگی، تأثیر مستقیمی بر عملکرد مدل‌های پیش‌بینی دارد. بر همین اساس، در پژوهش حاضر نیز با استفاده از مراحل پیش‌پردازش، انتخاب ویژگی و مقایسه چهار الگوریتم مختلف در محیط RapidMiner، تلاش شده است مناسب‌ترین مدل برای پیش‌بینی بیماری خشکی چشم تعیین شود.

۴. مواد و روش پژوهش

پژوهش حاضر از نظر هدف، کاربردی و از نظر روش اجرا، یک مطالعه تجربی در حوزه داده‌کاوی و یادگیری ماشین است. هدف اصلی این پژوهش، شناسایی عوامل مؤثر، مقایسه عملکرد الگوریتم‌های مختلف طبقه‌بندی در پیش‌بینی بیماری خشکی چشم و تعیین مناسب‌ترین مدل بر اساس معیارهای استاندارد ارزیابی است. تمامی مراحل پیش‌پردازش داده‌ها، آموزش مدل‌ها و ارزیابی عملکرد آن‌ها در محیط نرم‌افزار RapidMiner Studio انجام شده است.



شکل ۴-۱- مراحل انجام پژوهش

۱.۴. معرفی مجموعه‌داده

در این پژوهش از مجموعه‌داده Dry Eye Disease که در پایگاه Kaggle منتشر شده است، استفاده شد [۱۲]. این مجموعه‌داده شامل اطلاعات دموگرافیک، ویژگی‌های بالینی، سبک زندگی و علائم مرتبط با بیماری خشکی چشم است و به‌منظور توسعه مدل‌های یادگیری ماشین برای پیش‌بینی ابتلا به این بیماری تهیه شده است.

این مجموعه‌داده شامل ۲۰۰۰۰ نمونه و ۲۶ ویژگی است و متغیر هدف نیز وضعیت ابتلا یا عدم ابتلا به بیماری خشکی چشم را مشخص می‌کند. در جدول ۴-۱ ویژگی‌های مجموعه داده خشکی چشم شرح داده شده است.

جدول ۴-۱- ویژگی‌های مجموعه‌داده

نام ویژگی	توضیحات
Gender	جنسیت فرد
Age (سال)	سن فرد
Sleep duration (ساعت)	میانگین ساعت خواب در شبانه‌روز
Sleep quality	میزان کیفیت خواب
Stress level	میزان استرس فرد
Blood pressure	فشار خون ثبت‌شده
Heart rate (ضربه در دقیقه)	ضربان قلب در دقیقه
Daily steps (قدم)	تعداد قدم‌های روزانه
Physical activity	شاخص سطح فعالیت فیزیکی
Height (سانتی‌متر)	قد بر حسب سانتی‌متر

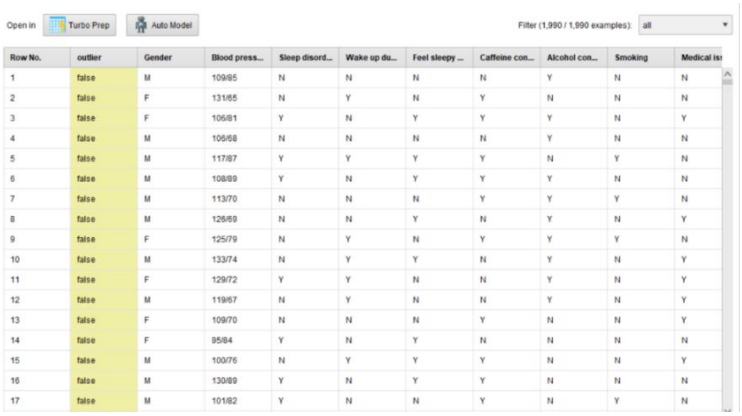
Weight (کیلوگرم)	وزن بر حسب کیلوگرم
Sleep disorder	آیا فرد دچار اختلال خواب است
Wake up during night	آیا فرد شب بیدار می‌شود
Feel sleepy during day	احساس خواب‌آلودگی در روز
Caffeine consumption	مصرف کافئین
Alcohol consumption	مصرف الکل
Smoking	سیگار کشیدن
Medical issue	وجود مشکل پزشکی دیگر
Ongoing medication	مصرف دارو در حال حاضر
Smart device before bed	استفاده از دستگاه هوشمند قبل از خواب
Average screen time (ساعت)	میانگین ساعت استفاده از صفحه نمایش در روز
Blue-light filter	استفاده از فیلتر نور آبی
Discomfort Eye-strain	احساس خستگی چشم
Redness in eye	قرمزی چشم
Itchiness / Irritation in eye	احساس خارش یا تحریک در چشم
Dry Eye Disease	وجود بیماری خشکی چشم

۲.۴. پیش‌پردازش داده‌ها

پیش از آموزش مدل‌های داده‌کاوی، مجموعه‌داده مورد بررسی قرار گرفت تا کیفیت داده‌ها افزایش یابد. ابتدا داده‌ها از نظر وجود مقادیر گمشده بررسی شدند. نتایج نشان داد که مجموعه‌داده فاقد داده‌های گمشده بوده و نیازی به جایگزینی یا حذف نمونه‌ها از این نظر وجود نداشت.

با توجه به حجم نسبتاً زیاد داده‌ها، به‌منظور کاهش زمان پردازش و افزایش سرعت اجرای الگوریتم‌ها، از عملگر Sample در نرم‌افزار RapidMiner استفاده شد و ۱۰ درصد از داده‌ها به‌صورت تصادفی برای ادامه تحلیل انتخاب شدند. پس از نمونه‌گیری، مجموعه‌داده شامل ۲۰۰۰ نمونه بود.

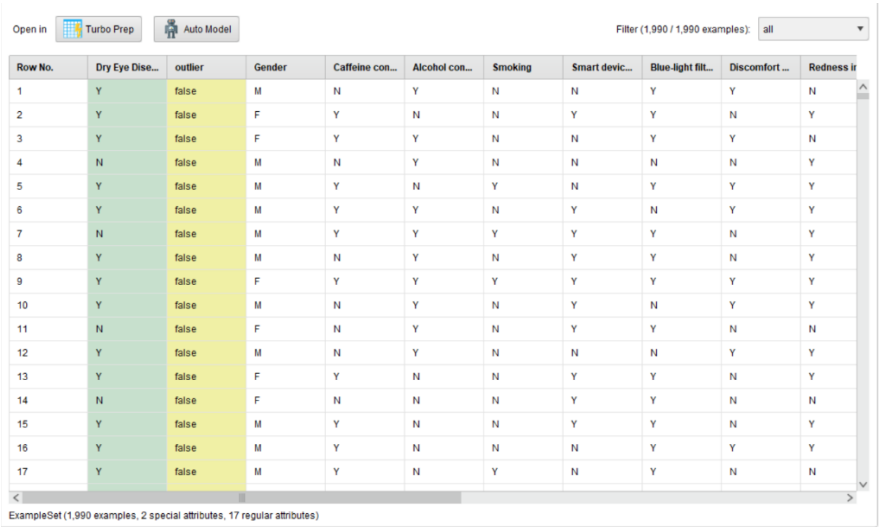
در مرحله بعد، داده‌های پرت با استفاده از عملگر Detect Outlier و معیار فاصله اقلیدسی شناسایی شدند. برای این منظور تعداد همسایه‌ها برابر ۱۰ در نظر گرفته شد و نمونه‌هایی که به‌عنوان داده پرت تشخیص داده شدند، با استفاده از عملگر Filter Examples از مجموعه‌داده حذف شدند. در شکل ۴-۱ مشاهده می‌شود که پس از حذف داده‌های پرت، تعداد نمونه‌های قابل استفاده به ۱۹۹۰ نمونه کاهش یافت.


Open in Turbo Prep Auto Model Filter (1,990 / 1,990 examples): all

Row No.	outlier	Gender	Blood press...	Sleep disord...	Wake up du...	Feel sleepy ...	Caffeine con...	Alcohol con...	Smoking	Medical is
1	false	M	109/85	N	N	N	N	Y	N	N
2	false	F	131/65	N	Y	N	Y	N	N	N
3	false	F	105/81	Y	N	Y	Y	Y	N	Y
4	false	M	106/68	N	N	N	N	Y	N	N
5	false	M	117/87	Y	Y	Y	Y	N	Y	N
6	false	M	108/89	Y	N	Y	Y	Y	N	N
7	false	M	113/70	N	N	N	Y	Y	Y	N
8	false	M	126/69	N	N	Y	N	Y	N	Y
9	false	F	125/79	N	Y	N	Y	Y	Y	N
10	false	M	133/74	N	Y	Y	N	Y	N	Y
11	false	F	129/72	Y	Y	N	N	Y	N	Y
12	false	M	119/67	N	Y	N	N	Y	N	Y
13	false	F	109/70	N	N	N	Y	N	N	Y
14	false	F	95/84	Y	N	Y	N	N	N	N
15	false	M	100/76	N	Y	Y	Y	N	N	Y
16	false	M	130/89	Y	N	Y	Y	N	N	N
17	false	M	101/82	Y	N	N	Y	N	Y	N

شکل ۴-۱- نمایی از داده‌ها پس از حذف داده‌های پرت

پس از آن، ویژگی‌هایی که نقش مستقیمی در فرآیند پیش‌بینی نداشتند، از طریق عملگر Select Attributes حذف شدند. در این مرحله ویژگی‌های Blood Pressure، Height، Weight، Sleep Disorder، Wake Up During، Night، Medical Issue، Ongoing Medication و همچنین ستون مربوط به Outlier از مجموعه‌داده حذف شدند. در ادامه نیز با استفاده از عملگر Set Role، متغیر Dry Eye Disease به‌عنوان متغیر هدف تعیین شد. در شکل ۴-۲ مشاهده می‌شود که ویژگی Dry Eye Disease به عنوان متغیر هدف برای پیش‌بینی این بیماری مشخص شده است.



Open in Turbo Prep Auto Model Filter (1,990 / 1,990 examples): all

Row No.	Dry Eye Dise...	outlier	Gender	Caffeine con...	Alcohol con...	Smoking	Smart devic...	Blue-light filt...	Discomfort ...	Redness in
1	Y	false	M	N	Y	N	N	Y	Y	N
2	Y	false	F	Y	N	N	Y	Y	N	Y
3	Y	false	F	Y	Y	N	N	Y	Y	N
4	N	false	M	N	Y	N	N	N	N	Y
5	Y	false	M	Y	N	Y	N	Y	Y	Y
6	Y	false	M	Y	Y	N	Y	N	Y	Y
7	N	false	M	Y	Y	Y	Y	Y	N	Y
8	Y	false	M	N	Y	N	Y	Y	N	Y
9	Y	false	F	Y	Y	Y	Y	Y	Y	Y
10	Y	false	M	N	Y	N	Y	N	Y	Y
11	N	false	F	N	Y	N	Y	Y	N	N
12	Y	false	M	N	Y	N	N	N	Y	Y
13	Y	false	F	Y	N	N	Y	Y	N	Y
14	N	false	F	N	N	N	Y	Y	N	N
15	Y	false	M	Y	N	N	Y	Y	N	Y
16	Y	false	M	Y	N	N	N	Y	Y	Y
17	Y	false	M	Y	N	Y	N	Y	N	N

ExampleSet (1,990 examples, 2 special attributes, 17 regular attributes)

شکل ۴-۲- برچسب دار شدن ویژگی Dry Eye Disease برای پیش بینی خشکی چشم

۳.۴. انتخاب ویژگی

یکی از اهداف این پژوهش، شناسایی مهم‌ترین عوامل مؤثر بر بیماری خشکی چشم بود. به همین منظور، از عملگر Weight by Chi-Squared Statistic برای تعیین میزان اهمیت ویژگی‌ها استفاده شد. این روش با محاسبه میزان وابستگی هر ویژگی به متغیر هدف، وزن هر ویژگی را تعیین می‌کند؛ به‌طوری‌که ویژگی‌هایی با وزن بیشتر، تأثیر بیشتری در پیش‌بینی بیماری دارند.

پس از محاسبه وزن ویژگی‌ها، از عملگر Select by Weights استفاده شد و تنها ویژگی‌هایی که وزن آن‌ها بیشتر یا مساوی ۱۰ بود، برای مرحله مدل‌سازی انتخاب شدند. همانطور که نتایج در شکل ۴-۳ نشان داد که ویژگی‌های Redness، Itchiness، Discomfort Eye Strain، Daily Steps، Sleep Duration و Age بیشترین اهمیت را در پیش‌بینی بیماری خشکی چشم داشته و به‌عنوان ورودی الگوریتم‌های طبقه‌بندی انتخاب شدند.

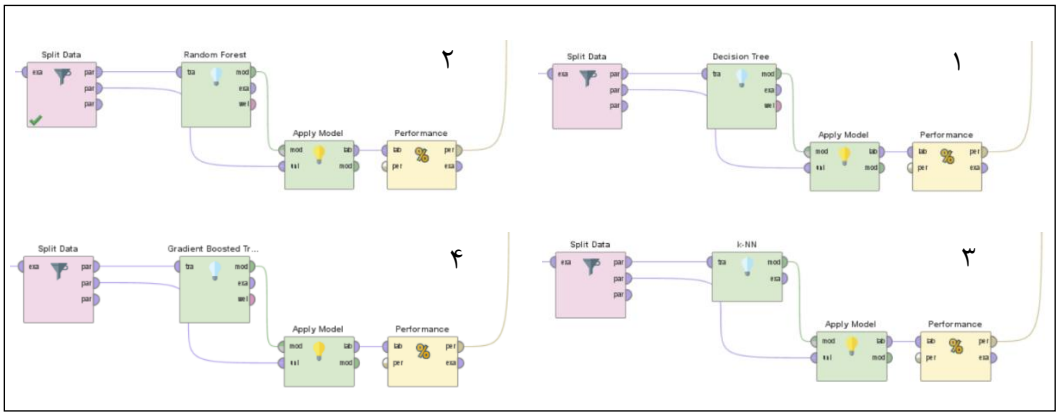
Gender	4.016
Physical ...	7.509
Heart rate	9.352
Age	11.754
Sleep du...	12.479
Daily ste...	13.620
Discomf...	17.781
Itchiness...	18.681
Rednes...	42.580

attribute	weight
Caffeine ...	0.253
Stress le...	0.275
Smart de...	0.563
Smoking	0.580
Alcohol c...	1.685
Sleep qu...	2.419
Blue-ligh...	2.965
Average ...	3.083

شکل ۴-۳- ویژگی‌های وزن‌دهی‌شده و وزن آن‌ها

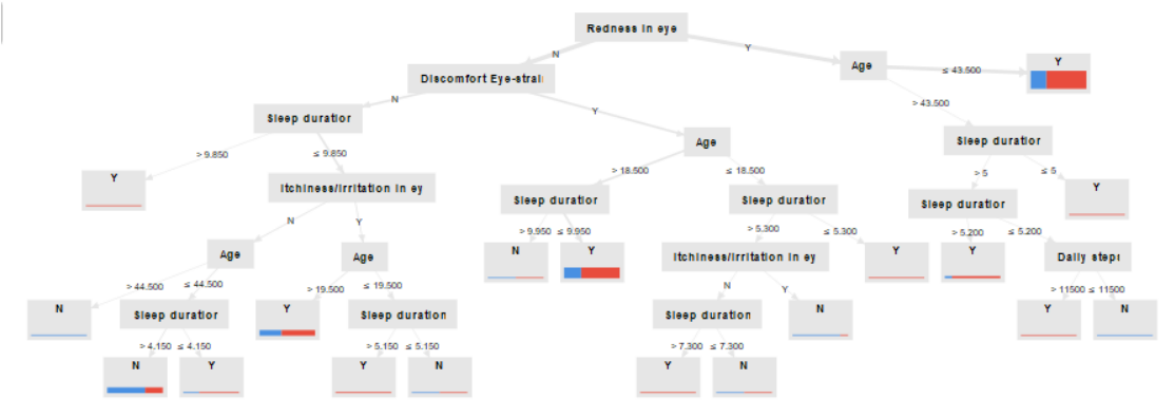
۴.۴. مدل‌سازی و ارزیابی الگوریتم‌ها

پس از انتخاب ویژگی‌های مؤثر، مجموعه‌داده با استفاده از عملگر Split Data به دو بخش آموزش و آزمون تقسیم شد؛ به‌طوری‌که ۷۰ درصد داده‌ها برای آموزش مدل‌ها و ۳۰ درصد برای ارزیابی عملکرد آن‌ها در نظر گرفته شد. در این پژوهش، چهار الگوریتم طبقه‌بندی شامل Decision Tree (۱)، Random Forest (۲)، K-Nearest Neighbor (۳) و Gradient Boosted Trees (۴) در محیط RapidMiner پیاده‌سازی و اجرا شدند. در شکل ۴-۴ مشاهده می‌شود که هر یک از الگوریتم‌ها با استفاده از داده‌های آموزشی آموزش داده شده و سپس عملکرد آن‌ها بر روی داده‌های آزمون مورد ارزیابی قرار گرفت.



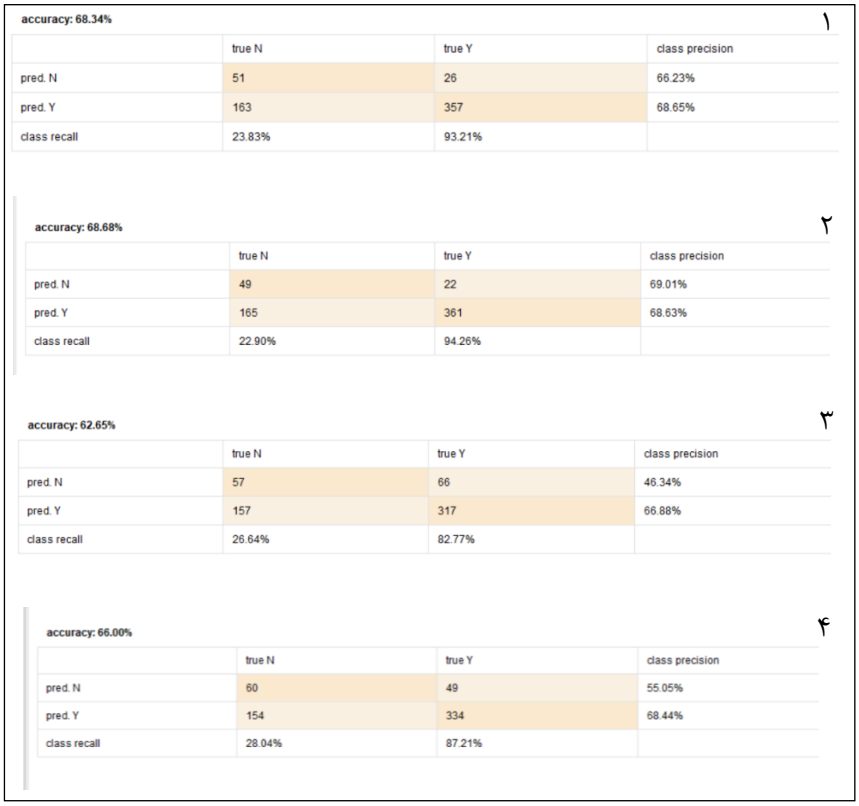
شکل ۴-۴- یادگیری مدل با استفاده از چهار الگوریتم طبقه‌بندی

در تحلیل انجام‌شده با استفاده از مدل Decision Tree، مشاهده شد که اولین ویژگی انتخاب‌شده برای ایجاد شاخه، متغیر Redness in Eye (قرمزی چشم) است. این موضوع نشان می‌دهد که الگوریتم درخت تصمیم، ویژگی قرمزی چشم را مناسب‌ترین متغیر برای ارائه بیشترین اطلاعات در خصوص ابتلا یا عدم ابتلا به بیماری خشکی چشم تشخیص داده است. همان‌طور که در شکل ۵-۴ مشاهده می‌شود، مدل درخت تصمیم ابتدا بر اساس این ویژگی داده‌ها را تقسیم کرده و سپس با استفاده از سایر ویژگی‌ها، فرآیند طبقه‌بندی بیماران را ادامه داده است.



شکل ۴-۵- نمودار گرافیکی الگوریتم درخت تصمیم

به‌منظور مقایسه عملکرد مدل‌ها، از معیارهای Accuracy، Precision و Recall استفاده شد. معیار Accuracy بیانگر نسبت نمونه‌های طبقه‌بندی‌شده صحیح به کل نمونه‌ها است، در حالی که معیارهای Precision و Recall به‌ترتیب میزان صحت پیش‌بینی‌های انجام‌شده و توانایی مدل در شناسایی صحیح نمونه‌های هر کلاس را نشان می‌دهند. مقادیر این معیارها از خروجی عملگر Performance در نرم‌افزار RapidMiner استخراج و برای مقایسه الگوریتم‌ها مورد استفاده قرار گرفت که در شکل ۴-۶ قابل مشاهده است.



۱

accuracy: 68.34%

	true N	true Y	class precision
pred. N	51	26	66.23%
pred. Y	163	357	68.65%
class recall	23.83%	93.21%	

۲

accuracy: 68.68%

	true N	true Y	class precision
pred. N	49	22	69.01%
pred. Y	165	361	68.63%
class recall	22.90%	94.26%	

۳

accuracy: 62.65%

	true N	true Y	class precision
pred. N	57	66	46.34%
pred. Y	157	317	66.88%
class recall	26.64%	82.77%	

۴

accuracy: 66.00%

	true N	true Y	class precision
pred. N	60	49	55.05%
pred. Y	154	334	68.44%
class recall	28.04%	87.21%	

شکل ۴-۶- نتایج به‌دست‌آمده از یادگیری مدل با چهار الگوریتم طبقه‌بندی

۵. یافته‌های پژوهش

به‌منظور تعیین مهم‌ترین عوامل مؤثر در پیش‌بینی بیماری خشکی چشم، وزن ویژگی‌های مجموعه‌داده با استفاده از روش Weight by Chi-Squared Statistic محاسبه شد. نتایج نشان داد که تمامی ویژگی‌های موجود در مجموعه‌داده نقش یکسانی در پیش‌بینی بیماری ندارند و برخی از آن‌ها تأثیر بیشتری بر تشخیص بیماری دارند.

بر اساس نتایج حاصل از وزن‌دهی ویژگی‌ها، متغیر Redness با وزن ۴۲.۵۸۰ بیشترین اهمیت را در پیش‌بینی بیماری خشکی چشم به خود اختصاص داد. پس از آن، ویژگی‌های Itchiness (18.681)، Discomfort Eye Strain (17.781)، Daily Steps (13.620)، Sleep Duration (12.479) و Age (11.754) به‌ترتیب بیشترین وزن را کسب کردند و به‌عنوان مهم‌ترین شاخص‌های مؤثر در مدل‌سازی انتخاب شدند.

نتایج بیانگر آن است که علائم بالینی مرتبط با سطح چشم، به‌ویژه قرمزی و خارش چشم، نقش تعیین‌کننده‌ای در پیش‌بینی بیماری خشکی چشم دارند. همچنین، متغیرهای مرتبط با سبک زندگی مانند مدت زمان خواب و میزان فعالیت روزانه نیز در کنار سن، از عوامل اثرگذار در پیش‌بینی این بیماری محسوب می‌شوند.

همچنین نتایج حاصل از مدل‌سازی با چهار الگوریتم طبقه‌بندی نشان داد که هر چهار الگوریتم توانایی پیش‌بینی بیماری خشکی چشم را دارند، اما میزان عملکرد آن‌ها در شاخص‌های ارزیابی متفاوت است. مقایسه نتایج نشان داد که الگوریتم Random Forest با دقت ۶۸/۶۸ درصد بهترین عملکرد را در مقایسه با سایر الگوریتم‌ها ارائه کرده است. این الگوریتم علاوه بر دستیابی به بیشترین مقدارAccuracy، در معیارهای Precision و Recall نیز نتایج مطلوب‌تری نسبت به سایر مدل‌ها به دست آورد.

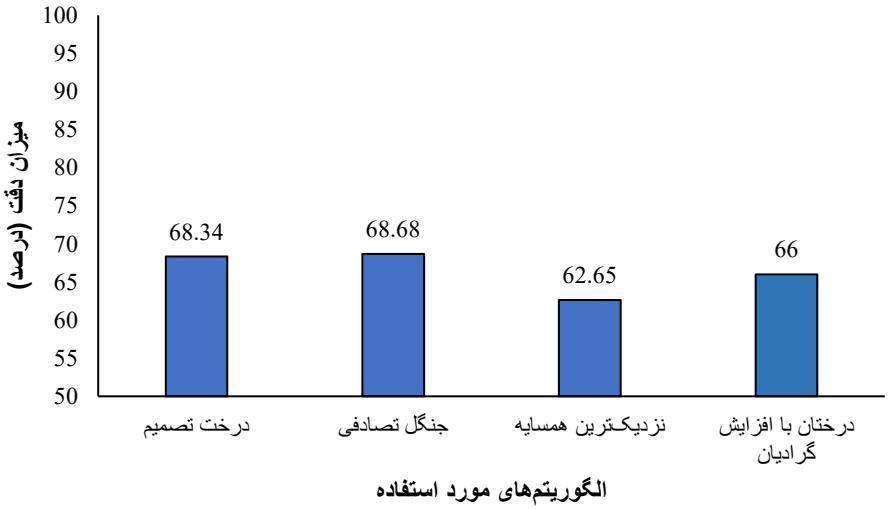
در مقابل، الگوریتم Decision Tree با وجود ساختار ساده و قابلیت تفسیر بالا، عملکرد ضعیف‌تری نسبت بهRandom Forest داشت. الگوریتم K-Nearest Neighbor نیز اگرچه توانست بخش قابل توجهی از نمونه‌ها را به‌درستی طبقه‌بندی کند، اما نسبت به Random Forest دقت کمتری داشت. همچنین، الگوریتم Gradient Boosted Trees عملکرد قابل قبولی ارائه کرد، اما در مجموعه‌داده مورد استفاده برتری محسوسی نسبت به Random Forest مشاهده نشد. در جدول ۴-۲ نتایج این الگوریتم‌ها براساس معیارهای ارزیابی دقت، صحت کلاس مثبت و منفی و بازیابی کلاس مثبت و منفی مقایسه شده است.

جدول ۴-۲- مقایسه نتایج مدل‌سازی با چهار الگوریتم طبقه‌بندی

الگوریتم	Accuracy	Precision (+)	Precision (-)	Recall (+)	Recall (-)
Decision Tree	۶۸/۳۴٪	۶۸/۶۵٪	۶۶/۲۳٪	۹۳/۲۱٪	۲۳/۸۳٪
Random forest	۶۸/۶۸٪	۶۸/۶۳٪	۶۹/۰۱٪	۹۴/۲۶٪	۲۲/۹۰٪
KNN	۶۲/۶۵٪	۶۶/۸۸٪	۴۶/۳۴٪	۸۲/۷۷٪	۲۶/۶۴٪
Gradient Boosted Trees	۶۶٪	۶۸/۴۴٪	۵۵/۰۵٪	۸۷/۲۱٪	۲۸/۰۴٪

۶. مقایسه عملکرد الگوریتم‌ها

به‌منظور مقایسه بهتر عملکرد مدل‌های طبقه‌بندی، نتایج حاصل از معیار Accuracy به‌صورت نمودار ستونی در شکل ۶-۱ نمایش داده شده است. مقایسه الگوریتم‌ها نشان می‌دهد که در مجموعه‌داده مورد استفاده، Random Forest بهترین عملکرد را در پیش‌بینی بیماری خشکی چشم داشته است. پس از آن، Gradient Boosted Trees، K-Nearest Neighbor و Decision Tree به‌ترتیب در رتبه‌های بعدی قرار گرفتند. این نتایج نشان می‌دهد که استفاده از روش‌های مبتنی بر یادگیری گروهی می‌تواند موجب بهبود عملکرد مدل‌های پیش‌بینی در مسائل پزشکی شود.



شکل ۶-۱- مقایسه دقت الگوریتم‌های مورد استفاده در تشخیص خشکی چشم

۷. بحث

نتایج این پژوهش نشان داد که استفاده از روش انتخاب ویژگی پیش از مرحله مدل‌سازی، نقش مهمی در شناسایی متغیرهای اثرگذار و کاهش ویژگی‌های کم‌اهمیت دارد. انتخاب تعداد محدودی از ویژگی‌های مؤثر، علاوه بر کاهش پیچیدگی مدل، موجب می‌شود فرآیند یادگیری بر روی متغیرهایی متمرکز شود که ارتباط بیشتری با متغیر هدف دارند. چنین رویکردی در بسیاری از مطالعات داده‌کاوی پزشکی نیز به‌عنوان یکی از مراحل اصلی توسعه مدل‌های پیش‌بینی معرفی شده است.

بررسی ویژگی‌های منتخب نشان داد که بخش قابل توجهی از آن‌ها به علائم بالینی بیماری مربوط هستند. این موضوع با گزارش TFOS DEWS II همخوانی دارد؛ به‌گونه‌ای که علائمی مانند قرمزی چشم، احساس ناراحتی و تحریک سطح چشم از نشانه‌های مهم بیماری خشکی چشم معرفی شده‌اند و می‌توانند در فرآیند تشخیص مورد استفاده قرار گیرند. علاوه بر این، حضور متغیرهایی مانند مدت خواب و میزان فعالیت روزانه در میان ویژگی‌های منتخب، نشان می‌دهد که عوامل مرتبط با سبک زندگی نیز می‌توانند در کنار علائم بالینی در پیش‌بینی بیماری نقش داشته باشند. این یافته با مطالعات اخیر که ارتباط کیفیت خواب، خستگی و سبک زندگی را با خشکی چشم گزارش کرده‌اند، قابل توجیه است.

در مقایسه الگوریتم‌های داده‌کاوی، اختلاف عملکرد مدل‌ها نشان می‌دهد که انتخاب الگوریتم مناسب می‌تواند بر کیفیت پیش‌بینی تأثیر قابل توجهی داشته باشد. عملکرد بهتر الگوریتم Random Forest را می‌توان به ساختار مبتنی بر یادگیری گروهی آن نسبت داد. این الگوریتم با ایجاد مجموعه‌ای از درخت‌های تصمیم و ترکیب نتایج آن‌ها، حساسیت کمتری نسبت به تغییرات داده‌ها داشته و معمولاً نسبت به یک درخت تصمیم منفرد، قابلیت تعمیم بالاتری دارد. این ویژگی سبب شده است که Random Forest در بسیاری از مسائل طبقه‌بندی پزشکی و زیست‌پزشکی به‌عنوان یکی از الگوریتم‌های قابل اعتماد شناخته شود.

به طور کلی، یافته‌های این پژوهش نشان می‌دهد که ترکیب روش‌های انتخاب ویژگی با الگوریتم‌های داده‌کاوی، علاوه بر بهبود فرآیند پیش‌بینی، می‌تواند به شناسایی عوامل اثرگذار بر بیماری نیز کمک کند. چنین رویکردی می‌تواند در طراحی سامانه‌های هوشمند غربالگری و ابزارهای پشتیبان تصمیم در حوزه چشم‌پزشکی مورد استفاده قرار گیرد؛ با این حال، ارزیابی این مدل‌ها بر روی داده‌های بالینی واقعی و جمعیت‌های متنوع‌تر، برای بررسی قابلیت تعمیم نتایج ضروری است.

۸. نتیجه‌گیری

در این پژوهش، از روش‌های داده‌کاوی برای شناسایی عوامل مؤثر و پیش‌بینی بیماری خشکی چشم در محیط RapidMiner استفاده شد. نتایج نشان داد که به‌کارگیری روش انتخاب ویژگی پیش از مرحله مدل‌سازی، امکان تمرکز بر مهم‌ترین متغیرهای مرتبط با بیماری را فراهم کرده و زمینه را برای توسعه مدل‌های پیش‌بینی کارآمدتر مهیا می‌کند.

مقایسه الگوریتم‌های مورد بررسی نیز نشان داد که هر یک از مدل‌ها توانایی مناسبی در طبقه‌بندی داده‌ها دارند، اما میزان عملکرد آن‌ها یکسان نیست. این موضوع بیانگر آن است که انتخاب الگوریتم مناسب، یکی از عوامل مؤثر در طراحی سامانه‌های هوشمند تشخیص بیماری است و استفاده از مدل‌های مبتنی بر یادگیری گروهی می‌تواند در برخی مجموعه‌داده‌ها نتایج مطلوب‌تری ایجاد کند.

یافته‌های این پژوهش نشان می‌دهد که روش‌های داده‌کاوی می‌توانند علاوه بر کمک به پیش‌بینی بیماری، در شناسایی عوامل اثرگذار بر خشکی چشم نیز کاربرد داشته باشند. نتایج حاصل از این مطالعه می‌تواند به‌عنوان مبنایی برای توسعه سامانه‌های پشتیبان تصمیم در حوزه چشم‌پزشکی و انجام پژوهش‌های آینده بر روی داده‌های بالینی گسترده‌تر مورد استفاده قرار گیرد.

با توجه به نتایج به‌دست‌آمده، پیشنهاد می‌شود در پژوهش‌های آینده از مجموعه‌داده‌های بزرگ‌تر و چندمرکزی، شامل اطلاعات بیماران از مراکز درمانی مختلف، استفاده شود. همچنین، ترکیب داده‌های بالینی با تصاویر پزشکی و به‌کارگیری روش‌های یادگیری عمیق می‌تواند موجب بهبود عملکرد سامانه‌های پیش‌بینی بیماری خشکی چشم شود. بررسی روش‌های انتخاب ویژگی پیشرفته، تنظیم بهینه پارامترهای الگوریتم‌ها و ارزیابی مدل‌ها بر روی داده‌های واقعی بیماران نیز از دیگر مسیرهای پیشنهادی برای تحقیقات آینده است.

۹. مراجع

1. Craig, J. P., Nichols, K. K., Akpek, E. K., Caffery, B., Dua, H. S., Joo, C. K., Liu, Z., Nelson, J. D., Nichols, J. J., Tsubota, K., & Stapleton, F. (2017). TFOS DEWS II Definition and Classification Report. The Ocular Surface, 15(3), 276–283.

2. Stapleton, F., Alves, M., Bunya, V. Y., Jalbert, I., Lekhanont, K., Malet, F., Na, K. S., Schaumberg, D., Uchino, M., Vehof, J., Viso, E., Vitale, S., & Jones, L. (2017). TFOS DEWS II Epidemiology Report. The Ocular Surface, 15(3), 334–365.

3. Wolffsohn, J. S., Arita, R., Chalmers, R., Djalilian, A., Dogru, M., Dumbleton, K., Gupta, P. K., Karpecki, P., Lazreg, S., Pult, H., Sullivan, B. D., Tomlinson, A., Tong, L., Villani, E., Yoon, K. C., Jones, L., & Craig, J. P. (2017). TFOS DEWS II Diagnostic Methodology Report. The Ocular Surface, 15(3), 539–574.

4. Beam, A. L., & Kohane, I. S. (2018). Big Data and Machine Learning in Health Care. JAMA, 319(13), 1317–1318.

5. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.

6. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques (4th ed.). Morgan Kaufmann.

7. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

8. Cover, T. M., & Hart, P. E. (1967). Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory, 13(1), 21–27.

9. Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29(5), 1189–1232.

10. Lu, Y., Wang, Z., & Chen, X. (2024). Artificial intelligence for dry eye disease: A systematic review and meta-analysis. Journal of Clinical Medicine, 13(5).

11. Britten-Jones, A. C., Wang, M. T. M., Samuels, I., Jennings, C., & Craig, J. P. (2024). Risk factors for dry eye disease: A systematic review and meta-analysis. The Ocular Surface, 32.

12. Kaggle. (2024). Dry Eye Disease Dataset. Available at: https://www.kaggle.com/