مدل‌های یادگیری ماشین برای پیش‌بینی تشخیص بیماری کبد

میترا منتظری[۱،۲] (M.Sc) ، مهدیه منتظری[*۳] (M.Sc)

۱- *دانشگاه علوم پزشکی کرمان، پژوهشکده آینده‌پژوهی در سلامت، مرکز تحقیقات انفورماتیک پزشکی*

۲- *دانشگاه شهید باهنرکرمان، بخش مهندسی کامپیوتر*

۳- *دانشگاه علوم پزشکی کرمان، پژوهشکده آینده‌پژوهی در سلامت، مرکز تحقیقات مدل‌سازی در سلامت*

چکیده

سابقه و هدف: کبد مهم‌ترین ارگان داخلی بدن می‌باشد که نقش اصلی در متابولیسم بدن دارد. بیماری کبد را نمی‌توان به راحتی در مراحل اولیه کشف کرد زیرا کبد حتی زمانی که قسمتی از آن نیز آسیب‌دیده باشد به درستی کار می‌کند و این خود تشخیص این بیماری را مشکل می‌کند. ابزارهای طبقه‌بندی اتوماتیک به عنوان یک ابزار کمک تشخیص باعث کاهش بار کاری پزشکان می‌گردد. طبقه‌بندی‌هایی که به منظور تشخیص هوشمند بیماری کبد در این پژوهش مورد استفاده قرار گرفته است شامل دسته‌بندهای Trees Random Forest 1NN, Naïve Bayes, SVM, AdaBoost, می‌باشند.

مواد و روش‌ها: داده‌های مورد استفاده از سوابق ۵۸۳ بیمار است که این مجموعه داده در دانشگاه کالیفرنیا در سال ۲۰۱۳ به ثبت رسیده است. برای ارزیابی مدل‌های استفاده شده از اعتبارسنجی ضرب‌دری از نوع k-لایه استفاده شده است. ۵ مدل ماشین یادگیری از نظر ویژگی، حساسیت، سطح زیر منحنی راک و دقت دسته‌بندی مقایسه شدند.

یافته‌ها: میزان دقت این ۵ مدل به ترتیب ۵۵٪، ۷۲٪، ۶۴٪، ۷۰٪ و ۷۱٪ و سطح زیر منحنی راک به ترتیب ۰/۷۲، ۰/۷۲، ۰/۵۹، ۰/۶۷ و ۰/۵ است.

نتیجه‌گیری: مدل Trees Random Forest بهترین مدل ارزیابی گردید که دارای بالاترین میزان دقت می‌باشد. از نظر سطح زیر منحنی راک مدل Trees Random Forest و Naïve Bayes بیش‌ترین سطح زیر منحنی را دارا می‌باشند. لذا به‌کارگیری مدل Trees Random Forest در زمینه تشخیص و پیش‌بینی بیماری کبد پیشنهاد می‌شود. این امر در تحقیقات مرتبط با حوزه‌ی سلامت و به‌خصوص در تخصیص منابع درمانی برای افرادی که پرمخاطره پیش‌بینی می‌شوند از اهمیت بالایی برخوردار است.

واژه‌های کلیدی: بیماری کبد، رده بندی، پیش‌بینی، هوش مصنوعی

مقدمه

کبد یکی از ارگان‌های پشتیبان‌کننده حیات انسان می‌باشد که نقش مهمی در زنده ماندن انسان دارد. به علت مشکلاتی هم‌چون مصرف بیش از حد الکل، استنشاق گازهای مضر، استفاده از غذاهای ناسالم و مصرف بیش از حد داروها، بیماران با مشکل کبدی روز به روز در حال افزایش می‌باشند. یکی از مهم‌ترین مشکلات بهداشتی در جهان بیماری‌هایی است که سهم عمده‌ای از منابع و امکانات را به خود اختصاص داده‌اند که بیماری‌های کبد جزء این دسته می‌باشند.

کبد مهم‌ترین ارگان داخلی بدن می‌باشد که نقش اصلی در متابولیسم بدن دارد. بیماری‌های کبدی معمولاً با التهاب یا آسیب‌دیدگی سلول‌های کبد ایجاد می‌شود که این بیماری

* نویسنده مسئول، تلفن: ۰۳۴۱-۲۱۱۴۵۳۶ نمابر: ۳۲۰۵۴۰۶ - ۰۳۴۱

Email: montazeri@kmu.ac.ir

تاریخ دریافت: ۱۳۹۲/۲/۸ تاریخ پذیرش: ۱۳۹۳/۳/۳

به‌عنوان یکی از ده بیماری کشنده در جهان معرفی شده است. سرطان کبد در مقیاس جهانی به عنوان سومین عامل مرگ و میر و با حدود ۵۶۰/۰۰۰ مورد جدید در هر سال می‌باشد [۱]. موضوع نگران‌کننده این است که بیماری‌های کبد به راحتی تشخیص داده نمی‌شوند و مشکلات بیماران کبدی معمولاً در مراحل اولیه قابل تشخیص نمی‌باشند و اغلب حتی در زمانی که مقداری از کبد دچار آسیب شده باشد تقریباً به درستی کار می‌کند و قادر به حفظ عملکرد طبیعی خود می‌باشد. بنابراین تشخیص به موقع آسیب کبدی یکی از قدم‌های مهم در امر درمان این بیمارن می‌باشد [۲]. تشخیص به موقع بیماری‌های کبدی باعث افزایش نرخ ۱ بقای بیماران می‌گردد. اگرچه پیش‌رفت‌های زیادی در زمینه علوم پزشکی انجام گرفته شده است اما هم‌چنان تشخیص زودهنگام بیماری‌های کبد کار دشواری است. فرآیند تشخیص به موقع این بیماری ارتباط مستقیم با تجربه پزشک دارد. برای به دست آوردن این تجربه نیازمند گذر سال‌های زیادی است. در نتیجه برای کمک به پزشکان در تشخیص زودهنگام این بیماری سیستم‌های تصمیم یار زیادی وجود دارند که حتی در تشخیص موارد جدیدی از این بیماری می‌توانند به پزشک کمک کنند [۳]. یک نمونه از این مدل‌ها، ماشین‌های یادگیری هستند. مدل‌های ماشین یادگیری که کاملاً غیر پارامتری هستند به‌طور فزاینده‌ای در حیطه‌های مختلف علوم به کار می‌روند. هدف اصلی این مدل، تعیین متغیرهای موثر، روابط بین آن‌ها و پیش‌بینی و تخمین می‌باشد که این موضوع در حیطه پزشکی و تحلیل داده‌های بهداشتی به‌علت نوع داده‌های آن نقش مهمی دارد [۴].

در مرجع [۵]، به بررسی استفاده از یادگیری گروهی (Ensemble Learning) برای بهبود دسته‌بندی پرداخته شده است. در روش پیشنهادی از ۳ روش Bagging, Boosting و Random subspace برای تشخیص اختلالات دریچه قلب استفاده شده است. پایگاه داده‌ای که در این روش استفاده شده است شامل ۹۱ مولفه تشخیص بیماری است. این مولفه‌ها با استفاده از روش‌های استخراج ویژگی (Feature Extraction) از سیگنال صوتی صدای قلب ۲۱۵ بیمار گرفته شده است که ۹۵ مورد آن نرمال و ۱۲۰ مورد غیر نرمال گزارش شده است. در مطالعه مرجع [۶] سیستم تشخیص بیماری سریع و تطبیقی ارائه شده است که بر مبنای آموزش بردار کوانتیزاسیون شبکه عصبی مصنوعی است. در این مطالعه، به منظور افزایش میزان موفقیت روش تشخیص بیماری و کاهش زمان تصمیم‌گیری، یک مکانیسم تقویت به شبکه LVQ (Learning Vector Quantization) تعبیه شده است. در واقع پارامترهای مکانیسم یادگیری تقویتی در شبکه عصبی به طور وقفی به روزرسانی می‌گردند. این روش برای تشخیص دو بیماری سرطان سینه و اختلالات تیروئید به کار برده شده است. در مرجع [۷] یک ماشین بردار پشتیبانی (SVM) برای تشخیص اختلالات اورولوژی پیشنهاد شده است. بی‌اختیاری ادرار یکی از بزرگ‌ترین بیماری‌های موثر بین ۱۰٪ و ۳۰٪ از جمعیت بزرگ‌سالان است و انتظار می‌رود در دهه آینده با افزایش هزینه‌های درمان روبه‌رو گردد. انواع مختلفی از اختلالات اورولوژی باعث بی‌اختیاری ادرار می‌گردد که تشخیص دقیق و ارزان این بیماری را تبدیل به یک مساله مهم کرده است. با استفاده از اطلاعات ثبت شده ۳۸۱ بیمار مبتلا به انواع اختلالات اورولوژی نشان داده شده که روش مبتنی بر SVM می‌تواند به دقت طبقه‌بندی به طور متوسط در ۸۴/۲۵٪ دست یابد. با توسعه فن‌آوری‌های بالینی، ویژگی‌های مختلفی برای تشخیص سرطان پستان جمع‌آوری شده است. تعیین موثر یا عدم موثر بودن تمام این ویژگی‌ها کاری چالش‌برانگیز و زمان‌بر است. هدف از این پژوهش، تشخیص سرطان پستان در ویژگی‌های تومور استخراج شده می‌باشد. استخراج و انتخاب ویژگی‌های مناسب برای طبقه‌بندی امری حیاتی هستند که از طریق روش‌های داده کاوی انجام می‌شود. در مرجع [۸] برای استخراج اطلاعات مفید و تشخیص تومور، ترکیبی از ابزار و ماشین بردار پشتیبانی (K-SVM) الگوریتم‌های ارائه شده است. هدف ازالگوریتم K-means شناختن جداگانه الگوهای پنهان از تومورهای خوش‌خیم و بدخیم است. عضویت هر تومور به این الگوها محاسبه و به عنوان یک ویژگی‌های جدید در مدل آموزش در نظر گرفته می‌شود. پس از آن، ماشین بردار پشتیبانی (SVM) برای طبقه‌بندی تومورهای دریافتی استفاده می‌شود و دقت حاصله به ۹۷/۳۸٪ می‌رسد.

بیماری کبد را نمی‌توان به راحتی در مراحل اولیه کشف کرد زیرا کبد حتی زمانی که قسمتی از آن نیز آسیب‌دیده باشد نیز به درستی کار می‌کند و این خود تشخیص این بیماری را مشکل می‌کند [۲]. تشخیص زودرس مشکلات کبدی شانس زنده ماندن بیماران کبدی را افزایش خواهد داد [۹]. روش‌های طبقه‌بندی در سال‌های اخیر در انواع روش‌های تشخیص خودکار بیماری کبد بسیار محبوب شده‌اند [۱۰ و ۹]. بیماری کبد را می‌توان با تجزیه و تحلیل سطح آنزیم‌های خون تشخیص داد [۱۱]. علاوه بر این، در حال حاضر دستگاه‌های تلفن سیار به طور گسترده برای نظارت بر شرایط بدن انسان استفاده می‌شوند. در این‌جا نیز، الگوریتم‌های طبقه‌بندی خودکار می‌توانند بسیار موثر و پر کاربرد باشند. با کمک

ابزارهای طبقه‌بندی خودکار برای بیماری‌های کبدی (با فعال کردن دستگاه سیار یا برقراری از طریق وب)، می‌توان صف بیماران در کارشناسان کبد مانند متخصصان غدد را کاهش داد.

تکنیک‌های طبقه‌بندی یکی از ابزارهای خودکار پیش‌بینی و تشخیص در امور پزشکی می‌باشند. به کمک مدل‌های ماشین یادگیری می‌توان کمک شایانی در زمینه تشخیص و پیش‌بینی بیماری‌ها به جامعه پزشکی نمود [۱۳،۱۲]. مشکلات و بیماری‌های کبد را می‌توان از طریق اندازه‌گیری آیتم‌های مختلف خون مشخص نمود و بیماری‌های را با تجزیه و تحلیل سطوح آنزیم‌ها تشخیص داد [۲]. تاکنون روش‌های مختلفی از ماشین یادگیری ارائه شده است. که در ادامه به بررسی ۵ مدل ماشین یادگیری پرداخته می‌شود:

- **مدل ماشین یادگیری Naïve Bayes**

این مدل اخیراً محبوبیت زیادی پیدا کرده است و به‌صورت فزاینده‌ای در حال استفاده می‌باشد [۱۴]. این مدل یک روش تشخیص الگوی آماری است که پیش‌فرض‌های دقیقی در مورد چگونگی تولید داده‌ها می‌سازد. این مدل برای تخمین پارامترها از مجموعه نمونه‌های آموزشی استفاده می‌کند. در این روش، دسته‌بندی روی نمونه‌های جدید به منظور انتخاب کلاس نمونه جدید از قانون‌های دسته‌بند Bayes استفاده می‌گردد.

دسته‌بند Naïve Bayes ساده‌ترین دسته‌بند در بین این مدل‌هاست که در آن فرض بر این است که همه ویژگی‌های نمونه مستقل از یک‌دیگر هستند. زمانی‌که فرضیه در نظر گرفته به روشنی نادرست باشد Naïve Bayes عمل دسته‌بندی را بسیار خوب انجام می‌دهد که علت این امر این‌گونه توضیح داده شده است که فرضیات دسته‌بندی تنها یک نشانه از برآورد تابع است و تقریب تابع هم‌چنان با دقت پایینی انجام می‌شود در حالی که دقت دسته‌بند بالا است [۱۵].

- **مدل ماشین یادگیری Trees Random Forest**

این مدل، روش یادگیری برای طبقه‌بندی و رگرسیون است که با ساخت بسیاری از درخت‌های تصمیم‌گیری در زمان آموزش و انتخاب بهترین درخت از میان درختان تولید شده کار می‌کند. این الگوریتم برای ایجاد یک جنگل تصادفی توسط لئوبریمن طراحی شد [۱۶]، و عنوان "جنگل‌های تصادفی" علامت تجاری آن می‌باشد. این اصطلاح برای اولین بار از "جنگل‌های تصمیم‌گیری تصادفی" آمده است که توسط قلع کام هو در سال ۱۹۹۵ پیشنهاد گردید [۱۷]. مزیت مهم این روش جلوگیری از یادگیری بیش از اندازه است.

این دسته‌بند یک روش بر مبنای مدل دسته جمعی است. مدل دسته جمعی مدلی است که از چندین مدل درختی استفاده می‌کند و می‌تواند به منظور دسته‌بندی و رگرسیون استفاده گردد. مدل دسته جمعی بر مبنای دقت و میزان اهمیت متغیرها کار می‌کند.

الگوریتم Trees Random Forest با اعمال مراحل زیر اجرا می‌شود:

فاز اول آموزش:

تعداد نمونه‌های آموزشی N و تعداد متغیرها موثر در طبقه‌بندی M نام‌گذاری می‌شود. تعداد متغیرهای ورودی برای تعیین این تصمیم در یک گره از درخت باید خیلی کم‌تر از M باشد.

تولید T مجموعه‌های آموزشی به طوری که تقریباً ۷۰ درصد کل مجموعه داده را شامل شود و ۳۰ درصد باقی‌مانده در هر مجموعه به عنوان داده آزمایشی نام‌گذاری گردد.

یادگیری T درخت توسط T مجموعه آموزشی تولید شده در مرحله قبل.

تخمین میزان خطای کلیه درخت‌ها توسط T نمونه آموزشی.

تکرار الگوریتم به منظور کاهش خطا.

فاز دوم تست:

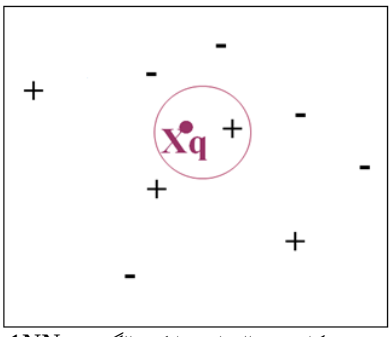
برای تعیین کلاس نمونه جدید، بین درختان تولیدی برای تعیین مقدار متغیر هدف رای‌گیری بر مبنای حداکثر رای انجام می‌شود.

- **مدل ماشین یادگیری 1NN**

الگوریتم دسته‌بند k نزدیک‌ترین همسایه (K-nearest neighbors algorithm) یکی از پرکاربردترین الگوریتم‌های دسته‌بندی است که از آن استفاده‌ی گسترده‌ای در کاربردهای مختلف می‌شود. الگوریتم نزدیک‌ترین همسایه، یک روش تشخیص الگوی آماری است که برای تشخیص کلاس الگوی مورد بررسی، از k الگوی مشابه موجود در نمونه‌های آموزش (k نزدیک‌ترین همسایه) که معمولاً فاصله‌شان تا الگوی مورد بررسی با استفاده از فاصله اقلیدسی وزن‌دهی شده محاسبه می‌شود.

در این روش با داشتن بردار مشخصه، k نزدیک‌ترین بردارهای مشخصه در داده‌های آموزش با استفاده از معیار فاصله اقلیدسی وزن‌دار شده، یافته می‌شوند. شماره دسته‌ی الگوی مورد نظر با توجه به دسته‌های k نزدیک‌ترین همسایه‌ها و از روی فراوانی تعیین می‌شود.

اگر k=1 انتخاب شود، الگوریتم 1-NN شماره دسته‌ی نزدیکترین نمونه را انتخاب خواهد نمود. این روال در شکل ۱ نشان داده شده است و x_q کلاس مثبت را به خود می‌گیرد.



شکل ۱. مثالی از عمل‌کرد الگوریتم 1NN.

- **مدل ماشین یادگیری AdaBoost**

آدابوست مخفف بوستینگ تطبیقی بوده و یک الگوریتم یادگیری ماشین است که توسط یاو فروند و رابرت شاپیر ابداع شد [۱۸]. در واقع آدابوست یک متاالگوریتم است که به منظور ارتقاء عمل‌کرد، همراه دیگر الگوریتم‌های یادگیری استفاده می‌شود. در این الگوریتم، دسته‌بندی در هر مرحله جدید بر مبنای نمونه‌های غلط طبقه‌بندی شده در مراحل قبل، تنظیم می‌گردد. آدابوست نسبت به داده‌های نویزی و پرت حساس است ولی نسبت به مشکل بیش برازش از بیش‌تر الگوریتم‌های یادگیری برتری دارد. طبقه‌بند پایه که در این‌جا استفاده می‌شود فقط کافیست از طبقه‌بند تصادفی (۵۰٪) بهتر باشد و به این ترتیب بهبود عمل‌کرد الگوریتم با تکرارهای بیش‌تر بهبود می‌یابد. حتی طبقه‌بندهای با خطای بالاتر از تصادفی با گرفتن ضریب منفی عمل‌کرد کلی را بهبود می‌بخشند. در الگوریتم آدابوست در هر دور $t = 1, \dots, T$ یک طبقه‌بند ضعیف اضافه می‌شود. در هر فراخوانی بر اساس اهمیت نمونه‌ها، وزن‌ها به روز می‌شود. در هر دور وزن نمونه‌های غلط طبقه‌بندی شده افزایش و وزن نمونه‌های درست طبقه‌بندی شده کاهش داده می‌شود. بنابراین طبقه‌بند جدید تمرکز بر نمونه‌هایی که سخت‌تر یاد گرفته می‌شوند، خواهند داشت. آدابوستی که در این مقاله استفاده شده است از نوع M1 [۲۹] است.

- **مدل ماشین یادگیری SVM**

ماشین بردار پشتیبانی یکی از روش‌های یادگیری با نظارت [۲۰] است که برای طبقه‌بندی [۲۱] و رگرسیون [۲۲] استفاده می‌کنند. این روش از جمله روش‌های نسبتاً جدیدی است که در سال‌های اخیر کارایی خوبی نشان داده است. مبنای کاری دسته‌بندی SVM دسته‌بندی خطی داده‌ها است و در تقسیم خطی داده‌ها سعی می‌کنیم خطی را انتخاب کنیم که حاشیه اطمینان بیش‌تری داشته باشد. حل معادله پیدا کردن خط بهینه برای این داده‌ها است. قبل از تقسیمِ خطی برای این‌که ماشین بتواند داده‌های با پیچیدگی بالا را دسته‌بندی کند داده‌ها را به وسیله تابع phi به فضای با ابعاد خیلی بالاتر [۲۳] می‌بریم. برای این‌که بتوان مساله ابعاد خیلی بالا را با استفاده از این روش‌ها حل کرد از قضیه دوگانی لاگرانژ [۲۴] برای تبدیلِ مساله مینیمم‌سازی مورد نظر به فرم دوگانی آن که در آن به جای تابع پیچیده phi که ما را به فضایی با ابعاد بالا می‌برد، تابعِ ساده‌تری به نامِ تابع هسته که ضرب برداری تابع phi است ظاهر می‌شود استفاده می‌گردد. از توابع هسته مختلفی از جمله هسته‌های نمایی، چندجمله‌ای و سیگموید می‌توان استفاده نمود. یکی از معروف‌ترین خودآموزها مربوط به مرجع [۲۵] است.

برای نشان دادن عمل‌کرد مدل‌های ماشین یادگیری از منحنی راک استفاده می‌شود. این منحنی طرح گرافیکی است که عمل‌کرد یک سیستم طبقه‌بندی دودویی نشان می‌دهد. این منحنی نرخ درست مثبت در مقابل نرخ خطا مثبت را نشان می‌دهد. به نرخ درست مثبت، حساسیت و به نرخ خطا مثبت، ویژگی نیز گفته می‌شود.

علاوه بر این، اعتبارسنجی ضرب‌دری که گاهی تخمین گردشی نیز نامیده می‌شود، نیز یک روش ارزیابی است که مشخص می‌کند نتایج یک تحلیل آماری بر روی یک مجموعه داده تا چه اندازه قابل تعمیم و مستقل از داده‌های آموزشی است. این تکنیک به طور ویژه در کاربردهای پیش‌بینی مورد استفاده قرار می‌گیرد تا مشخص شود مدل مورد نظر تا چه اندازه در عمل مفید خواهد بود. به‌طور کلی یک دور از اعتبارسنجی ضرب‌دری شامل افراز داده‌ها به دو زیرمجموعه مکمل، انجام تحلیل بر روی یکی از آن زیرمجموعه‌ها (داده‌های آموزشی) و اعتبارسنجی تحلیل با استفاده از داده‌های مجموعه دیگر است (داده‌های اعتبارسنجی یا تست). برای کاهش پراکندگی، عمل اعتبارسنجی چندین بار با افرازهای مختلف انجام و از نتایج اعتبارسنجی‌ها میانگین گرفته می‌شود. هنگامی‌که جمع‌آوری داده‌های بیش‌تر سخت، پرهزینه و یا غیرممکن باشد، استفاده از اعتبارسنجی ضرب‌دری کمک می‌کند تا از فرضیات بایاس شده با داده‌های فعلی که قابل تعمیم نیستند، دوری شود. در روش پیشنهادی از اعتبارسنجی ضرب‌دری از نوع k - لایه و نوع یکی- بیرون استفاده شده است. در روش k - لایه داده‌ها به k زیرمجموعه افراز می‌شوند. از این k زیرمجموعه، هر بار یکی برای اعتبارسنجی و k-1 تای دیگر برای آموزش به‌کار می‌روند. این روال k بار تکرار می‌شود و همه داده‌ها دقیقاً k بار برای آموزش و یک بار برای اعتبارسنجی به‌کار می‌روند. در نهایت میانگین نتیجه این k بار اعتبارسنجی به عنوان یک تخمین نهایی برگزیده می‌شود. البته می‌توان از روش‌های دیگر برای

ترکیب نتایج استفاده کرد. به‌طور معمول از اعتبارسنجی ضرب‌دری ۱۰-لایه استفاده می‌شود. در روش یکی- بیرون همان‌طور که از اسم این روش پیداست در هر مرحله یکی از داده‌ها برای اعتبارسنجی بیرون گذاشته می‌شود و بقیه داده‌ها برای آموزش استفاده می‌شوند. این روش در واقع همان روش- k لایه است که در آن k برابر تعداد داده‌ها در نظر گرفته شده است. این روش از نظر محاسباتی بسیار پرهزینه است، زیرا فرآیند آموزش و اعتبارسنجی به تعداد بسیار زیادی تکرار می‌شود.

در مجموعه داده که در این مقاله آورده شده از آیتم‌های مختلف برای تشخیص این بیماری استفاده شده است. مجموعه داده ذکر شده در این پژوهش از سوابق ۵۸۳ بیمار از شمال شرق آندرا پرادش هند، جمع‌آوری شده است که در دانشگاه کالیفرنیا در سال ۲۰۱۳ به ثبت رسیده، می‌باشد [۲۶]. این مجموعه داده دارای متغیرهای Age, Gender, Total Bilirubin , Direct Bilirubin, TotalProteins, Albumin, A/Gratio, SGPT, SGOT, Alkphos می‌باشد. مجموعه داده توسط روش‌های مختلف ماشین یادگیری دسته‌بندی شده است و ویژگی، حساسیت، سطح زیر منحنی راک و دقت دسته‌بندی به عنوان پارامترهایی برای مقایسه این روش‌ها و انتخاب بهترین مدل پیش‌گویی استفاده شده است. در ادامه در بخش بعد روش اجرا شرح داده خواهد شد. نتایج و بحث و نتیجه‌گیری به ترتیب در بخش‌های ۳ و ۴ ارائه می‌شوند.

مواد و روش‌ها

در روش پیشنهادی از مدل‌های ماشین یادگیری برای تشخیص بیماری کبد استفاده شده است. یادگیری ماشین عبارت است از این‌که چگونه می‌توان مدلی طراحی نمود که از طریق تجربه یادگیرد و عمل‌کرد خود را بهتر کند. یادگیری ماشین زمینه جدیدی از هوش مصنوعی است که در حال حاضر دوران رشد و تکامل خود را می‌گذراند. یادگیری ماشین یک زمینه تحقیقاتی بسیار فعال در علوم مختلف است. در واقع یک مدل ماشین یادگیری با یک مسئله جستجو درگیر است که در آن در فضای بسیار بزرگ فرضیه‌ها به دنبال بهترین فرضیه‌ای است که با داده‌های آموزشی و تجربه قبلی سازگار باشد. در روش پیشنهادی از ۵ مدل ماشین یادگیری Naïve Bayes، Trees Random Forest1NN، AdaBoost، SVM جهت تشخیص و پیش‌بینی بیماری کبد استفاده شده است.

نتایج

نرم‌افزار استفاده شده نرم‌افزار Weka و مجموعه داده کبدی مورد استفاده در این مقاله از شمال شرق آندرا پرادش هند توسط آقای Bendi VenkataRamana، پروفسور M. Surendra Prasad Babu و پروفسور N.B. Venkateswarlu جمع‌آوری شده است. این مجموعه داده در سال ۲۰۱۳ در مجموعه داده‌های استاندارد دانشگاه کالیفرنیا [۲۶] ثبت شد. مجموعه داده‌های استاندارد دانشگاه کالیفرنیا که شامل داده‌های استانداردی است که در حیطه علوم مختلف است. این داده‌ها در حیطه پزشکی نیز استانداردسازی شده‌اند و محققین بسیاری از آن‌ها برای تشخیص بیماری استفاده نموده‌اند [۹-۱۷]. مجموعه داده‌های استاندارد دانشگاه کالیفرنیا به صورت رایگان در اختیار کلیه پژوهشگران است و نیاز به گرفتن هیچ‌گونه کسب مجوز ندارد و صرفاً ارجاع به این مجموعه داده کافی است. این مجموعه داده‌ها شامل ۴۱۶ پرونده بیمار کبد و ۱۶۷ پرونده بیمار غیر کبد است. متغیر هدف به دو گروه بیمار کبد یا غیر کبد تقسیم شده است. این مجموعه داده‌ها شامل ۴۴۱ پرونده بیمار مرد و ۱۴۲ پرونده بیمار زن است. این پایگاه داده دارای ۱۰ متغیر تشخیص بیماری می‌باشد که در جدول ۱ نشان داده شده است.

جدول ۱. متغیرهای تشخیص بیماری کبد

شماره ردیف	نوع متغیر تشخیص بیماری
۱.	Age
۲.	Gender
۳.	Total Bilirubin
۴.	Direct Bilirubin
۵.	Total proteins
۶.	Albumin
۷.	A/G ratio
۸.	SGPT
۹.	SGOT
۱۰.	Alkphos

تنظیمات مدل‌های پیشنهادی

با توجه به این که در هر دسته‌بندی، به تعداد نمونه‌هایی که به کلاس مثبت و منفی تعلق دارند به ترتیب با P و N نمایش می‌دهند می‌توان تعریف زیر را بیان کرد:

FP = به نمونه‌هایی که به گروه مثبت تعلق دارند و اشتباه پیش‌بینی شدند.

TP = به نمونه‌هایی که به گروه مثبت تعلق دارند و درست پیش‌بینی شدند.

TN = به نمونه‌هایی که به گروه منفی تعلق دارند و درست پیش‌بینی شدند.

FN = به نمونه‌هایی که به گروه منفی تعلق دارند و اشتباه پیش‌بینی شدند.

در نتیجه روابط ۱ تا ۳ بیان می‌شود:

$$\text{نرخ درست مثبت} = \frac{TP}{P} \qquad (۱)$$

$$\text{نرخ خطای مثبت} = \frac{FP}{N} \quad (۲)$$

$$\text{دقت دسته بندی} = \frac{TP+TN}{P+N} \quad (۳)$$

برای تخمین دقت الگوریتم پنج مدل، از اعتبارسنجی ضربدری استفاده شده است.

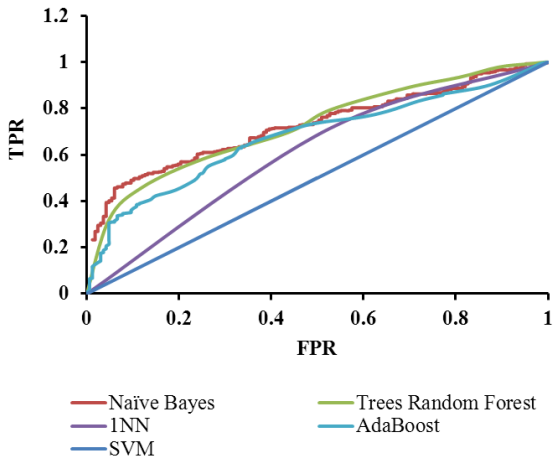
عملکرد مدل‌های پیشنهادی

در این پژوهش ۵ مدل ماشین یادگیری Naïve Bayes، Trees Random Forest1NN، AdaBoost، SVM به منظور تشخیص بیماری کبد استفاده شده است. در مدل دسته‌بندی Trees Random Forest تعداد درختان به کار برده شده ۱۰ درخت است و در مدل AdaBoost تعداد مراحل اجرا ۱۰ مرحله است. ۵ مدل ماشین یادگیری اجرا شده‌اند و نتایج حاصل از اجرای این ۵ مدل در جدول ۲ خلاصه شده است. همان‌طور که در این جدول مشاهده می‌شود، این ۵ مدل از نظر ویژگی، حساسیت، سطح زیر منحنی راک و دقت دسته‌بندی مقایسه شدند و در هر بررسی بهترین مقدار برجسته شده است که از نظر دقت دسته‌بندی مدل ماشین یادگیری Trees Random Forest دارای بالاترین دقت است.

جدول ۲. نتایج حاصل از عملکرد ۵ مدل Naïve Bayes، Trees Random Forest1NN، AdaBoost، SVM

مدل ماشین یادگیری	دقت دسته بندی	سطح زیر منحنی راک	حساسیت	ویژگی
Naïve Bayes	۵۵/۷۴۶۱	۰/۷۲۶	۰/۳۹۹	۰/۰۴۸
Trees Random Forest	۷۲/۲۱۲۷	۰/۷۲۲	۰/۸۸۹	۰/۶۹۵
1NN	۶۴/۴۹۴	۰/۵۹۳	۰/۷۱۴	۰/۵۲۷
AdaBoost	۷۰/۳۲۵۹	۰/۶۷۷	۰/۹۶۹	۰/۹۵۸
SVM	۷۱/۳۵۵۱	۰/۵	۱	۱

شکل ۲ منحنی راک برای ۵ مدل ماشین یادگیری رسم شده است. همان‌طور که در این شکل دیده می‌شود، مدل Naïve Bayes و Trees Random Forest نسبت به ۳ مدل ماشین یادگیری در موقعیت بالاتری قرار دارند و در نتیجه سطح زیر منحنی بیش‌تری دارند. مدل SVM بر روی خط متقارن قرار دارد در نتیجه سطح زیر منحنی آن ۰/۵ است. در نمودار راک مقادیر ۰ تا ۰/۵ بیانگر دسته‌بندی تصادفی و ۰/۵ تا ۱ بیانگر توان‌مندی تشخیص کلی مدل است. در نتیجه عملاً مدل SVM کارایی مطلوبی ندارد و مانند یک دسته‌بند تصادفی عمل می‌کند.



شکل ۲. منحنی مشخصه عملکرد سیستم برای ۵ مدل ماشین یادگیری

بحث و نتیجه‌گیری

کبد یکی از ارگان‌های حیاتی بدن می‌باشد که سلامتی آن برای زنده ماندن انسان لازم و ضروری می‌باشد. به علت مشکلاتی هم‌چون مصرف بیش از حد الکل، استنشاق گازهای مضر، استفاده از غذاهای ناسالم و مصرف بیش از حد داروها، بیماران با مشکل کبدی روز به روز در حال افزایش می‌باشند. ابزارهای طبقه‌بندی اتوماتیک به عنوان یک ابزار کمک تشخیص باعث کاهش بار کاری پزشکان می‌گردد. این پژوهش به ارزیابی الگوریتم‌های طبقه‌بندی منتخب برای طبقه‌بندی مجموعه داده تعدادی از بیماران جهت پیش‌بینی و تشخیص بیماری کبدی می‌پردازد. دسته‌بندهایی که بدین منظور در این پژوهش مورد استفاده قرار گرفته است شامل دسته‌بندهای Naïve Bayes، Trees Random Forest1NN، AdaBoost، SVM می‌باشند. این الگوریتم‌ها بر اساس معیارهای حساسیت، دقت و ویژگی ارزیابی شدند.

مدل یادگیری Naïve Bayes دقت پایینی نسبت به بقیه مدل‌ها دارد ولی دارای سطح زیر منحنی بالاتری نسبت به بقیه مدل‌ها می‌باشد. در واقع این مدل عملکرد خوبی برای نمونه‌های مثبت یعنی کسانی که بیماری کبد دارند داشته است و روی نمونه‌های منفی عملکرد خوبی ندارد که همین مساله باعث شده است عملکرد کلی این مدل (دقت دسته‌بندی) خوبی نداشته باشد. در طراحی مدل‌های یادگیری اگر مدلی دارای دقت ۰/۵ باشد در واقع این مدل به صورت تصادفی کار می‌کند و یادگیری امتیازی برای این مدل به حساب نمی‌آید. مدل Trees Random Forest دارای بالاترین دقت دسته‌بندی است. برای این مدل نیز سطح زیر منحنی راک از همه بیش‌تر است که با توجه به مقدار بالای دقت دسته‌بندی

این مدل این مقدار انتظار می‌رفت. سه مدل آخر نیز نسبت به مدل‌های بالا کارایی پایین‌تری دارند.

در مجموع مدل Trees Random Forest بهترین مدل ارزیابی شد و دارای بالاترین میزان دقت است. از نظر سطح زیر منحنی راک مدل Trees Random Forest و Naïve Bayes بیش‌ترین سطح زیر منحنی دارند. لذا به‌کارگیری مدل Trees Random Forest در زمینه تشخیص و پیش‌بینی بیماری کبد پیشنهاد می‌شود. این امر در تحقیقات مرتبط با حوزه سلامت و به خصوص در تخصیص منابع درمانی برای افرادی که پرمخاطره پیش‌بینی می‌شوند از اهمیت بالایی برخوردار است.

منابع

[1] VenkataRamana B, Surendra Prasad Babu M, Venkateswarlu NB. A critical study of selected classification algorithms for liver disease diagnosis. Int J Data Manage Syst 2011; 3: 101-114.

[2] Lin RH. An intelligent model for liver disease diagnosis. Artif Intell Med 2009; 47: 53-62.

[3] Fathima A, Venkateswara S. An intelligent medical decision support system for diagnosing liver disease, International Conference on Electrical Engineering, Computer Science and Mechanical Engineering, 2013.

[4] Biglarian A, Hajizadeh E, Kazemnejad A. Comparison of artificial neural network and Cox regression models in survival prediction of gastric cancer patients. Koomesh 2010; 11: Pe215-Pe220. (Persian).

[5] Das R, Sengur A. Evaluation of ensemble methods for diagnosing of valvular heart disease. Exp Syst Appl 2010; 37: 5110-5115.

[6] Alkım E1, Gürbüz E, Kılıç E. A fast and adaptive automated disease diagnosis method with an innovative neural network model. Neural Netw 2012; 33: 88-96.

[7] Gil D, Johnsson M. Using support vector machines in diagnoses of urological dysfunctions. Exp Syst Appl 2010; 37: 4713-4718.

[8] Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Exp Syst Appl 2014; 41: 1476-1482.

[9] Ramana BV, Babu MS, Venkateswarlu N. A critical study of selected classification algorithms for liver disease diagnosis. Int J Data Manag Syst 2011; 3: 101-114.

[10] Chuang CL. Case-based reasoning support for liver disease diagnosis. Artif Intell Med 2011; 53: 15-23.

[11] Schiff's Diseases of the Liver, 10 th Edition Copyright © 2007 Lippincott Williams & Wilkins by Schiff, Eugene R.; Sorrell, Michael F.; Maddrey, Willis C.

[12] Montazeri M, Baghshah MS, Enhesari A. Hyper-heuristic algorithm for finding efficient features in diagnose of lung cancer sisease. 2013.

[13] Montazeri M, Naji HR, Montazeri M. Memetic feature selection algorithm based on efficient filter local search. 2013.

[14] Farid DM, Zhang L, Rahman CM, Hossain MA, Strachan R. Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. Exp Syst Appl 2014; 41: 1937-1946.

[15] Friedman ND, Geiger MG. Bayesian network classifiers. Machine Learning 1997; 29: 131-163.

[16] Breiman Leo. Random Forests. Machine Learning 2001; 45: 5-32.

[17] Ho, Tin Kam. Random Decision Forest". Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16. 1995; Pp: 278-282.

[18] Freund Y, Schapire RE. A desicion-theoretic generalization of on-line learning and an application to boosting. J Com Syst Sci 1997; 55: 119-139.

[19] Y. Freund and R. E. Schapire, Experiments with a New Boosting Algorithm, In: L. Saitta, Ed., Proceedings of the 13th International Conference on Machine Learning, Bari, 3-6 July 1996, pp. 148-156.Cortes, Corinna; and Vapnik, Vladimir N.; "Support-Vector Networks", Machine Learning, 20, 1995.

[20] Cortes, Corinna; and Vapnik, Vladimir N.; Support-Vector Networks, Machine Learning, 20, 1995.

[21] Press William H, Teukolsky Saul A, Vetterling William T, Flannery BP. Section 16.5. support vector machines. numerical recipes: the art of scientific computing (3rd ed.). New York: cambridge university press. ISBN 978-0-521-88068-8Aizerman, Mark A.; Braverman, Emmanuel M.; and Rozonoer, Lev I. (1964). "Theoretical foundations of the potential function method in pattern recognition learning". Automation and Remote Control 2007; 25: 821-837.

[22] ACM Website, Press release of March 17th 2009. http://www.acm.org/press-room/news-releases/awards-08-groupa

[23] Boser, Bernhard E.; Guyon, Isabelle M.; and Vapnik, Vladimir N.; A training algorithm for optimal margin classifiers. In Haussler, David (editor); 5th Annual ACM Workshop on COLT, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

[24] Meyer D, Leisch F, Hornik K. The support vector machine under test. Neurocomputing 2003; 55: 169-186.

[26] Bache K, Lichman M. UCI machine learning repository. irvine, CA: university of california, school of information and computer science, 2013. [http://archive.ics.uci.edu/ml].

[27] Beloufa F, Chikh MA. Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm. Comput Methods Programs Biomed 2013; 112: 92-103.

[28] Kahramanli H, Allahverdi N. Design of a hybrid system for the diabetes and heart diseases. Exp Syst Appl 2008; 35: 82-89.

[29] Polat K, Güneş S. A new feature selection method on classification of medical datasets: Kernel F-score feature selection. Exp Syst Appl 2009; 36: 10367-10373.

[30] Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Exp Syst Appl 2014; 41: 1476-1482.

[31] Polat K, Güneş S. Breast cancer diagnosis using least square support vector machine. Digital Signal Process 2007; 17: 694-701.

[32] Fidelis MV, Lopes HS, Freitas AA. Discovering comprehensible classification rules with a genetic algorithm. Evolutionary Computation, 2000. Proceedings of the 2000 Congress on. 2000.

[33] Aizerman M, Braverman A, Emmanuel M, Rozonoer L. Theoretical foundations of the potential function method in pattern recognition learning. Auto Remote Cont 1964; 25: 821-837.

[33] Qasem SN, Shamsuddin SM. Radial basis function network based on time variant multi-objective particle swarm optimization for medical diseases diagnosis. Appl Soft Comput 2011; 11: 1427-1438.

[34] Abdi MJ, Giveki D. Automatic detection of erythemato-squamous diseases using PSO–SVM based on association rules. Eng Appl Artif Intell 2013; 26: 603-608.

Machine learning models for predicting the diagnosis of liver disease

Mitra Montazeri (M.Sc)[1,2], Mahdieh Montazeri (M.Sc) [*3]
1- Medical Informatics Research Center, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran
2 - Computer Engineering Dept., Shahid Bahonar University of Kerman, Kerman, Iran
3 - Research Center for Modeling in Health, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran

(Received: 30 Aug 2013; Accepted: 24 May 2014)

Introduction: The liver is the most important organ of the body has a central role in metabolism. Liver disease cannot be easily discovered in the early stages, because even when the liver is damaged partially, it also can work truly, and this makes it difficult to diagnose. Automatic classification tools as a diagnostic tool can reduce the workload of doctors. Smart ways to detect liver disease classification used in this study consist of classifier and Naïve Bayes, Trees Random Forest, 1NN, AdaBoost, SVM.

Materials and Methods: Our database was 583 patient records which they have been registered at university of California in 2013. For evaluate the proposed models, it is used K-fold cross validation. Five models of machine learning compare base on specificity, sensitivity, accuracy and area under ROC curve.

***Results*:** The accuracy of the five models, respectively, 55%, 72%, 64%, 70% and 71% respectively and area under the ROC curve of 0.72, 0.72, 0.59, and 0.67 is 0.5.

Conclusion: Trees Random Forest model was the best model with the highest level of accuracy. The area under the ROC curve of Trees Random Forest and Naïve Bayes models have the largest area under the curve. Therefore Trees Random Forest model and predict the diagnosis of liver disease is recommended.

Keywords: Liver disease, classification, prediction, Artificial Intelligence

* Corresponding author. Fax: +98 341 2114536; Tel +98 341 2114536
montazeri@kmu.ac.ir

How to cite this article:
Montazeri M, Montazeri M. Machine learning models for predicting the diagnosis of liver disease. koomesh. 2014; 16 (1) :53-59
URL http://koomeshjournal.semums.ac.ir/browse.php?a_code=A-10-1972-2&slc_lang=en&sid=1