مجله پژوهش های معاصر در علوم و تحقیقات
سال اول، شماره ۷، بهمن ۱۳۹۸
ISSN: ۲۶۷۶-۵۷۶٤
www.jocrisar.ir

طراحی سیستم پیش بینی بیماری قلبی-عروقی با استفاده از الگوریتم های یادگیری ماشین

محمد کاظمی[۱]، امیرعباس کاظمی[۲] امیرحسین باب الحکمی[۳]، نگار عباسی[۴]

[۱] دانشجوی کارشناسی ارشد، گروه کامپیوتر، دانشگاه آزاد اسلامی مشهد
[۲] دانشجوی کارشناسی ارشد، گروه کامپیوتر، دانشگاه فردوسی مشهد
[۳] دانشجوی دکتری، گروه کامپیوتر دانشگاه فردوسی مشهد
[۴] کارشناسی ارشد، گروه آمار دانشگاه پیام نور مشهد

چکیده

بیماری‌های قلبی یکی از شایع‌ترین بیماری‌هاست که در حال حاضر تعداد افراد مبتلا به این نوع بیماری ها در حال افزایش می باشد. این در حالی است در صورتی که مراقبت‌های لازم برای بیمار در زمان مناسب صورت نگیرد، می‌تواند باعث مرگ بیمار شود. از این‌رو تشخیص دقیق در مرحله معاینه اولیه به همراه درمان مناسب می‌تواند منجر به اجتناب از افزایش میزان مرگ‌ومیر ناشی از بیماری قلبی گردد. برای رسیدن به این مهم می‌توان از تکنیک‌های موجود در زمینه داده‌کاوی بهره گرفت. داده‌کاوی داده‌های مفیدی را از مجموعه داده های موجود استخراج می‌کند که منجر به پیش‌بینی یا دسته‌بندی اطلاعات از طریق خوشه‌بندی، کلاس‌بندی و یا کشف الگوهای پنهان می‌شود. تاکنون تحقیقات زیادی با استفاده از مدل‌های داده‌کاوی در تشخیص بیماری‌های مختلف مانند بیماری‌های قلبی و عروقی انجام شده است..

در این مقاله قصد داریم با استفاده از رویکردی مبتنی بر انتخاب ویژگی به عنوان یک گام پیش پردازش، مدلی باهدف تشخیص بیماری قلبی ارائه گردد. راهکار پیشنهادی دارای ۳ گام اصلی می باشد که گام۱) پیش پردازش داده ها با هدف رفع مقادیر Null و پرت در مجموعه داده ها، گام۲) انتخاب ویژگی های موثر با بهره وری از ۲ روش ضریب همبستگی پیرسون و تجزیه و تحلیل مولفه های اصلی که سعی در حذف ویژگی هایی که با صفت هدف رابطه خاصی ندارند و رفتار این ویژگی مستقل از صفت هدف می باشند، است. و در گام۳) با استفاده از ۳ الگوریتم J۴۸، شبکه بیزین و SVM مدلی برای پیش بینی بیماری قلبی ساخته می شود. نتایج بدست آمده نشان می دهد الگوریتم J۴۸ با دقت ۰,۸۹ دارای بالاترین دقت است.

واژه‌های کلیدی: تشخیص بیماری قلبی، انتخاب ویژگی، ضریب همبستگی پیرسون، تحلیل مولفه های اساسی، طبقه بندی

۱. مقدمه

بیماری‌های قلبی یکی از شایع‌ترین بیماری‌ها می‌باشند که در حال حاضر تعداد افراد مبتلا به بیماران قلبی در حال افزایش است. به طور کلی این بیماری دارای علائم مختلفی می باشد اما از مهمترین آن می توان به درد در قفسه سینه، ناحیه فک، گردن، نفس تنگی و ... اشاره کرد. همچنین مشکلات قلبی می تواند باعث بروز بیماری عروق کرونر قلب، نارسایی قلبی و سکته مغزی شود. با این وجود، اگرچه عامل اصلی مرگ و میر در دهه اخیر در دنیا شناخته شده است، به عنوان کنترل پذیرترین و قابل پیشگیری ترین بیماری نیز شناخته می شود. نتایج بدست آمده نشان می دهد، شناسایی و تشخیص به موقع بیماری قلبی علاوه بر امکان جلوگیری از پیشرفت و بروز عوارض ناشی از آن، می تواند در درمان صحیح و کامل بیماری موثر باشد. برای تشخیص به موقع بیماری قلبی تست های مختلفی مانند نوار قلب، تست استرس و آنژیو قلب وجود دارد ولی این نوع آزمایش ها هزینه ی زیادی در بر دارند و نمی توان آنها را به طور گسترده استفاده کرد. تشخیص در مراحل ابتدایی کمک شایانی به روند درمان افراد میکند و نسبت به روش های متداول پزشکی مقرون به صرفه تر است. امروزه داده کاوی و کشف دانش به کمک ما آمده است تا با جمع آوری اطلاعاتی مرتبط، به پیش بینی بیماری قلبی بپردازیم. یکی از عملکردهای پیش گویانه در داده کاوی، دسته بندی است. دسته بندی، فرآیند یافتن مدلی است که با تشخیص دسته ها و یا مفاهیم داده می تواند دسته ناشناخته اشیا دیگر را پیش گویی کند[۱] هدف از پیش بینی، کمک به روند تشخیص و بهبود سلامت بیماران است. تاکنون مطالعات زیادی پیرامون تکنیک های داده کاوی بر روی انباره بیماران قلبی و عروقی انجام شده است، که تکنیک های مختلف داده کاوی مانند درخت تصمیم و شبکه های عصبی را به کار برده اند که (a) پیش بینی حمله قلبی و سندرم کرونری حاد با استفاده از شبکه عصبی داده کاوی[۴]–[۲](b) ارزیابی معیارهای مهم برای پیش بینی بقا بیمار با استفاده از شبکه های بیزین[۵](c) تشخیص ایسکمی قلبی در رکوردهای طولانی مدت الکتروکاردیوگرام[۶] (d) ارزیابی فاکتورهای ریسکی مرتبط با بیماری قلبی با استفاده از درخت تصمیم و قوانین تلازمی[۸] ,[۷] و (e) کشف روابط بین عوامل خطرزای قلبی با استفاده از درخت تصمیم[۹] از جمله پژوهش های انجام شده در این زمینه است.

در این مقاله نیز قصد داریم با بهره وری از تکنیک های داده کاوی و تحلیل ویژگی های مختلف مدلی باهدف پیش بینی بیماری قلبی ارائه شود. گام های ارائه شده شامل ۳ فاز اصلی می باشند. در ابتدا (۱) پیش پردازش های لازم باهدف رفع مقادیر Null انجام گرفته است. برای رفع مقادیر Null از تکنیک پیش بینی و طبقه بندی نمونه ها استفاده شده است. به این صورت که نمونه هایی که مقادیر Null دارند را به عنوان داده های تست و سایر نمونه ها را به عنوان داده های آموزش در نظر گرفته ایم. (۲) در گام دوم به بررسی ویژگی های موثر در پیش بینی بیماری قلبی پرداخته خواهد شد و با بهره وری از ۲ روش ضریب همبستگی پیرسون و PCA صفاتی که باهدف مسئله ارتباطی ندارند، شناسایی و حذف خواهند شد. (۳) و درگام نهایی از ۳ الگوریتم درخت تصمیم، شبکه بیزین و SVM باهدف ساخت مدل طبقه بندی، پیش بینی بیماری قلبی استفاده خواهد شد. در ادامه این مقاله، در بخش۲ مقالات مرتبط به هدف مسئله مورد بررسی قرار خواهند گرفت. در بخش ۳ راهکار پیشنهادی و گام های ارائه شده به صورت دقیق‌تر مورد بررسی قرار خواهند. در بخش ۴ نتایج آزمایشات و ارزیابی‌های انجام شده ارائه و درانتها نیز نتایج بدست آمده تحلیل و کارهای آینده معرفی می‌شوند.

۲. پیشینه

در این بخش از مقاله، برخی از پژوهش های مرتبط در پیش بینی بیماری قلبی مورد بررسی قرار خواهد گرفت. قلب یکی از مهمترین اجزاء بدن می باشد که وظیفه انتشار خوب به سایر بخش های بدن را بعهده دارد. بافت قلب از ۳ لایه داخلی (آندوکارد)، لایه عضلانی میانی (میوکارد) و لایه خارجی (پریکارد) تشکیل شده است و هر یک از این لایه وظایف مخصوص به خود را دارند. لایه اندکارد پوشش حفره های دهلیز و بطن، میوکارد ضخیم ترین لایه قلب است که ادامه حیات انسان به آن وابسته می باشد و پریکارد پوشش پیوندی یا آبشامه قلب را می سازد[۱]. در یک تقسیم بندی کلی، بیماری قلبی و عروقی

شامل بیماری عروق کرونری، بیماری دریچه قلب، بیماری های مادرزادی قلب، بیماری عضله قلب و ... می باشد که از این بین بیماری شرایین کرونری از شایعترین بیماری قلبی است. این بیماری عمدتا به صورت آنژین صدری و انفارکتوس قلبی است که از مهمترین دلایل بروز آن می توان به آتروسکلروزیس، ترومبوزیس، اسپاسم و انوریسم کرونری اشاره کرد. آتروسکلروزیس که به معنی اختلال عمومی در شریان هاست، از شایعترین علل رخداد این بیماری می باشد. این اختلال به وسیله پلاک های زردرنگ از کلسترول و چربی ها مشخص می شود و مانع از رساندن اکسیژن به باقت می شود. به عبارتی عروق کرونری تغذیه کننده بافت های قلبی می باشند. وسعت این بیماری به آناتومی توزیع رگ های مسدود بستگی دارد که به عبارتی عروق کرونری تغذیه کننده بافت های قلبی می باشند. باتوجه به پیچیدگی های بسیاری که در تشخیص وجود بیماری قلبی یا نوع آن دیده می شود، وجود یک ابزار کمکی که بتواند دانش ارزشمند را در اختیار پزشکان قرار بدهد بسیار احساس می شود تا با بهره وری از این ابزار، بتوان تاثیر داروهای ناسازگار را کاهش و در پیش بینی رفتار آینده بیماران براساس سابقه ثبت شده و تشخیص بیماری قبلی را استفاده کرد.

داده کاوی یک ابزار کاربردی در شناسایی اطلاعات مفید و ناشناخته از حجم عظیم داده هایت. از این اطلاعات می‌توان برای پیش بینی آینده موقعیت‌ها و به عنوان یک کمک برای فرآیند تصمیم گیری استفاده کرد. دانش مفید را می‌توان با استفاده از داده کاوی در برنامه کاربردی مراقبت پزشکی مثل سیستم پشتیبانی تصمیم، پیدا نمود. داده بزرگ تحویل داده شده توسط سازمان‌های مراقبت پزشکی بسیار پیچیده و بسیار بزرگ هستند و نمی‌توان آنها را با تکنیک‌های معمول مدیریت و تحلیل نمود. داده کاوی روشی برای تغییر این مقدار داده به اطلاعات مفید برای پشتیبانی تصمیم ارائه می‌کند. داده بزرگ کاوی در مراقبت پزشکی درباره یادگیری مدل‌هایی برای پیش بینی بیماری بیمار است. به‌عنوان نمونه، داده کاوی می‌تواند به سازمان‌های بیمه سلامت در تشخیص سوء استفاده‌ها یا تظاهر، به مؤسسات پزشکی در تصمیم گیری در مورد مدیریت روابط مشتریان، به پزشکان در تشخیص درست درمان و به بیماران در دریافت خدمات پزشکی بهتر و اقتصادی‌تر، کمک نماید. این تحلیل پیشگویانه در خدمات پزشکی بسیار استفاده می‌شود. در این مقاله نیز قصد داریم مدلی باهدف پیش بینی بیماری قلبی ارائه شود. اما در این فرآیند از چند تکنیک فیچرسلکشن باهدف آنالیز پارامترهای مختلف استفاده شده است. که به وسیله آن بتوان ویژگی های نامرتبط را قبل از آنکه تاثیر منفی بر روی مدل پیش بینی بگذارد شناسایی و حذف کنیم. همچنین براساس خروجی این روش ها از میزان تاثیر هریک از پارامترهای مختلف آگاهی پیدا کرد. با این وجود باتوجه به اهمیت این بیماری، مطالعات مختلفی در این زمینه انجام شده است و در ادامه به مرور برخی از این مقالات پرداخته خواهد شد.

در مطالعه ای برای پیش بینی بیماری قلبی از آزمون اکوکاردیوگرافی قفسه سینه استفاده میشود که این داده ها در بازه زمانی ۲ سال تهیه شده است.داده ها شامل ۱۵ متغیر است.چهار آزمایش برای الگوریتم های طبقه بندی طراحی شد و آزمایشها بر روی مجموعه دادههای آموزشی کامل شامل ۷۳۳۹ نمونه انجام شد. برای کلیه آزمایش ها دو موقعیت در نظر گرفته شد که حاوی ۱۵ ویژگی و دیگری شامل ۸ ویژگی انتخابشده است. درمجموع هشت مدل توسعه داده شد.به عنوان نتیجه طبقه بندی J۴۸ را بر روی مجموعه داده با ۸ ویژگی انتخابشده با دقت ۹۵،۵۶ ٪ بعنوان بهترین الگوریتم دسته بندی معرفی نمودند[۲]. در پژوهشی دیگر از الگوریتم های J۴۸ ، شبکه بیزین و بیزین ساده و الگوریتم Simple CART برای طبقه بندی و توسعه یک مدل برای تشخیص حملات قلبی در مجموعه داده بیماران استفاده شد. هدف این پژوهش، پیشبینی احتمال حمله قلبی در مجموعه داده بیمار با استفاده از تکنیک های داده کاوی و تعیین مدلی که بالاترین درصد صحت پیشبینی را به همراه دارد. مجموعه داده بیماران از داده های پزشکان در آفریقای جنوبی جمع آوری و از ۱۱ویژگی آن در آزمایشات استفاده گردید[۴]. مقاله ای در سال ۲۰۱۴ با عنوان پیش بینی بستری شدن بدلیل بیماری قلبی با استفاده از روش یادگیری نظارت شده ارائه گردید. داده های مورد استفاده در آن پژوهش از مرکز پزشکی بوستون در بازه زمانی ۵ سال انتخاب شدند که به عنوان نتیجه اعلام گردید که ۸۲ ٪ بیماران مورد مطالعه در سال بعدی در بیمارستان بستری می شوند[۳] در حوزه های دیگر و با استفاده از الگوریتم ژنتیک و شبکه RBF در سال ۲۰۱۵ پژوهشی صورت پذیرفت که با استفاده از پایگاه داده موجود در سایت معتبر UCI و با ۱۴ ویژگی و ۳۰۳ نمونه است. الگوریتم های طبقه بندی روی داده ها اعمال و به عنوان نتیجه ارائه شد که نتایج

خوبی از دو الگوریتم Naïve Bayes با ۸۳ ٪ و J۴۸ با ۷۷ ٪ دقت دریافت شده است[۵]. در مقاله][۶]، مدلی برای توسعه یک شبکه عصبی مصنوعی برای مدل های تشخیص بیماری های عروق کرونری قلب، با استفاده از مجموعه ای از عوامل سنتی و ژنتیکی این بیماری است. پایگاه داده اصلی برای شبکه های عصبی مصنوعی شامل بالینی، آزمایشگاهی، کاربردی، آنژیوگرافی عروق کرونر و ژنتیکی است. بهترین دقت با توپولوژی شبکه های عصبی پرسپترون چند لایه از دو لایه پنهان برای مدل شامل شده توسط هر دو عوامل خطر CHD ۱۲ ژنتیکی و غیر ژنتیکی به دست آمد. مجموعه داده استفاده شده در این مقاله، شامل ۹۸۷ بیمار است و تحت آنژوگرافی برای تشخیص عروق کرونر قرار گرفته اند. تشخیص بیماری عروق کرونر قلب نیز از هر دو نتایج بالینی و کرونر گرافی کسب گردید و این نتایج با استفاده از آزمایش و نوع ژنتیک اجازه ایجاد یک پایگاه داده از بیماران مبتلا را فراهم می کند که پس از آن برای تشخیص CHD با استفاده از شبکه های عصبی مورد استفاده قرار بگیرد. مدل شبکه عصبی با استفاده از پروسپترون چند لایه ایجاد و دقت مدل های الگوریتم ژنتیک بهبود یافته شده با استفاده از پارامترهای مختلف بهینه سازی از جمله تعداد نرون در لایه پنهان، تعداد ورودی به شبکه های عصبی و ضریب شیب فعالسازی مورد استفاده قرار گرفته که امکان بهینه سازی را ایجاد می کند. در مقاله[۸] از یک الگوریتم یادگیری ترکیبی برای شبکه عصبی پرسپترون چند لایه فازی از ترکیب دو الگوریتم فراابتکاری جستجوی گرانشی و بهینه سازی ازدحام ذرات استفاده می کنند. ۵ مجموعه داده مربوط به بیماری قلبی، سرطان سینه، هپاتیت، اختلالات کبدی و سرطان ریه در این مطالعه بررسی شده اند. که دقت بدست آمده از اجرای راهکار پیشنهادی بر روی مجموعه داده بیماری قلبی برابر ۷۹,۹۶٪ است.

۳. راهکار پیشنهادی

راهکار پیشنهادی ارائه شده برای این پژوهش در ۳ گام تعریف شده است: (۱) پیش پردازش داده‌ها (۲) تعیین مجموعه ویژگی‌های مؤثر و شناسایی شده با ۲ روش پیرسون و PCA و (۳) ارائه مدل پیش بینی که از ۳ الگوریتم درخت تصمیم، شبکه بیزین و SVM استفاده شده است. در ادامه گام ۲ که باهدف انتخاب ویژگی‌های موثر و کاهش ابعاد دیتاست و گام ۳ که شامل طبقه بندی و آموزش داده‌ها است به صورت دقیق تر مورد بررسی قرار خواهد گرفت.

۳,۱- ضریب همبستگی پیرسون

هدف از استفاده از این روش، شناسایی صفات کم اهمیت برای مسئله هدف و شناسایی مواردی که با یکدیگر همپوشانی دارند است[۱۰]. برای کسب این اطلاعات لازم است ۲ گام انجام شود که شامل:

- **ضریب همبستگی صفات با صفت هدف:** اگر ضریب همبستگی بین ۰,۲ تا ۰,۲ باشد، یعنی ویژگی مستقل است و تاثیری در مسئله هدف ندارد و حذف می‌شود.
- **ضریب همبستگی دو به دوی صفات (بجز صفت هدف):** اگر ضریب همبستگی بین ۲ ویژگی بیش از ۰,۸ باشد یعنی با یکدیگر همپوشانی دارند و می‌توان یکی از صفات را حذف کرد (صفت با ضریب همبستگی کمتر با صفت هدف حذف خواهد شد).

براساس فازهایی که تعیین شد، مجموعه ویژگی‌های منتخب توسط این روش باقیمانده با صفت هدف ارتباط مستقیم (ضریب همبستگی مثبت) و یا ارتباط معکوس (ضریب همبستگی منفی) دارند. همچنین این ویژگی‌ها نسبت به یکدیگر مستقل می‌باشند.

۳,۲- تحلیل مولفه های اصلی

دومین روش Feature Selection در گام انتخاب ویژگی ها، الگوریتم PCA که یکی از روش های پرکاربرد در کاهش ابعاد داده ورودی است، استفاده خواهد شد. تحلیل مولفه‌های اصلی در تعریف ریاضی یک تبدیل خطی متعامد است که داده را به دستگاه مختصات جدید می‌برد به طوری که بزرگترین واریانس داده بر روی اولین محور مختصات، دومین بزرگترین واریانس

بر روی دومین محور مختصات قرار می‌گیرد و همین طور برای بقیه. تحلیل مولفه‌های اصلی می‌تواند برای کاهش ابعاد داده مورد استفاده قرار بگیرد، به این ترتیب، مولفه‌هایی از مجموعه داده را که بیشترین تأثیر در واریانس را دارند حفظ می‌کند.

۳,۳ - الگوریتم‌های طبقه بندی

بعد از شناسایی ویژگی‌های مؤثر و حذف ویژگی‌های کم اهمیت، در انتها داده‌ها را برای طبقه بندی و ساخت مدل پیش بینی با استفاده از روش طبقه بندی درخت تصمیم، شبکه بیزین و SVM آموزش خواهیم داد و براساس نتایج بدست آمده، بهترین مدل از نظر دقت پیش بینی به عنوان مدل نهایی ارائه و مشخص خواهد گردید. در ادامه این چند الگوریتم به صورت خلاصه معرفی خواهند شد.

۳,۳,۱ - درخت تصمیم

درخت تصمیم یکی از روش‌های قوی و متداول برای دسته بندی و پیش بینی است. در واقع درخت‌های تصمیم بالا به پایین رایج‌ترین تکنیک دسته بندی هستند و از مهم‌ترین دلایل رایج بودنشان می‌توان به شفاف بودن، قابل فهم، انعطاف پذیری و پردازش نسبتاً سریع ساختار آنها نام برد. پیش بینی به دست آمده از درخت در قالب یک سری قواعد توضیح داده می‌شود. در این درخت هر گره داخلی شامل سوالی بر مبنای یک صفت مشخص و یک فرزند برای هر پاسخ ممکن بوده و هر برگ با یکی از کلاس‌های ممکن برچسب گذاری می‌شود[۱۱]

۳,۳,۲ - شبکه بیزین

شبکه بیزین یک مدل گرافیکی برای روابط بین مجموعه‌ای از انواع پارامترهای متغیر است. این مدل گرافیکی ساختار S یک گراف بدون دور جهت دار است و همه گره‌ها در S تناظر یک به یک با ویژگی‌های مجموعه داده دارند. کمان‌ها نشان دهنده تاثیرات بین ویژگی‌ها هستند. زمانی که دسته بندی بیزین برای مجموعه داده‌های بزرگ به کار برده شود دارای سرعت و دقت بالایی است. طبقه‌بندی بیزین ساده نیز بر اساس نظریه بیز است. این روش برای مقادیر ویژگی‌های پیوسته ورودی و خروجی مناسب است و در دنیای واقعی برای تشخیص دست خط کاربرد دارد. این روش مناسب‌ترین روش برای انتخاب یک مدل در مقایسه با داده‌های موجود بر اساس شرایط احتمالی است[۱۲]

۳,۳,۳ - ماشین بردار پشتیبان

به مجموعه ای از نقاط در فضای n بعدی داده ها، بردار پشتیبان گفته می شود که مرز بندی دسته ها را نشان داده و دسته بندی و مرزبندی آنها را انجام می دهد و با جابجایی یکی از این دو مورد ممکن است تغییر کند. SVM یا ماشین بردار پشتیبان، با معیار قرار دادن بردار های پشتیبان بهترین دسته بندی و تفکیک بین داده ها را انجام می دهد همچنین در SVM مبنای یادگیری ماشین و ساخت مدل، داده های قرار گرفته شده در بردارهای پشتیبان می باشد. هدف الگوریتم SVM یافتن بهترین مرز در بین داده ها بوده و بیشترین فاصله ممکن از تمام دسته ها را در نظر می گیرد و به سایر نقاط داده ها حساس نمی باشد.

۴ - نتایج آزمایشات

در این بخش از مقاله نتایج بدست آمده از اجرای راهکار پیشنهادی بر روی دیتاست ارائه و مورد بررسی قرار خواهد گرفت. برای پیاده سازی عملیات پیش پردازش، Feature Selection، طبقه بندی و ساخت مدل پیش بینی از زبان برنامه نویسی R که یکی از زبان‌های پرکاربرد در زمینه داده‌کاوی می‌باشد، استفاده گردید.

۴,۱ - مجموعه داده‌ها

داده‌های مورد بررسی در این پژوهش از مجموعه پایگاه داده در این مقاله از مجموعه داده تشخیص بیماری قلبی Cleveland که در سایت مرجع UCI قرارگرفته، استفاده‌شده است. علائم زیادی از بیماری قلبی وجود دارد و یافتن الگوهایی از داده بیماری قلبی در تشخیص دلایل آتی این بیماری کمک می‌کند[۱۳]. پایگاه داده شامل ۳۰۳ نمونه که ۶ مورد آن دارای مقادیر Null است. این مجموعه داده‌های اولیه ۷۶ صفت خام دارد درحالی‌که همه آزمایش‌ها فقط بر روی ۱۴ صفت از آن‌ها انجام‌شده

است. بنابراین، این پایگاه داده شامل ۱۳ علامت بیماری و یک صفت تشخیص است که صفت هدف به وجود بیماری قلبی بر اساس علائم موجود در بیمار اشاره دارد که یک مقدار عددی (کمتر از ۵۰٪ تنگی عروق) یا ۱ (بیشتر از ۵۰٪ تنگی عروق) است.

۴,۲- معیارهای ارزیابی

پس از اجرای مدل سازی باید به ارزیابی نتایج حاصل پرداخت. نتایج ارزیابی باعث بهبود مدل می‌شود و مدل را قابل استفاده می‌نماید. شاخص‌های مختلفی مانند شفافیت (Specificity)، حساسیت (Sensitivity)، دقت (Precision) و صحت (Accuracy) برای ارزیابی روش‌های دسته بندی وجود دارند که طبق روابط ۱ تا ۴ محاسبه می‌گردند. برای محاسبه میزان شاخص‌ها می‌توان از ماتریس اغتشاش (Confusion Matrix) استفاده کرد. این ماتریس ابزار مفیدی برای تحلیل چگونگی عملکرد روش دسته بندی در تشخیص داده‌ها با مشاهدات دسته‌های مختلف است. حالت ایده آل این است که بیشتر داده‌های مرتبط با مشاهدات روی قطر اصلی ماتریس قرار گرفته باشند و مابقی مقادیر ماتریس صفر یا نزدیک صفر باشند. همچنین برای ارزیابی دقیق کارایی از تکنیک K-fold cross Validation استفاده شده است.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (۱)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (۲)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}\frac{TP}{TP+FP} \quad (۳)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (۴)$$

۴,۳- نتایج ضریب همبستگی پیرسون

یکی از مهمترین گام های ارائه شده در پیش پردازش داده ها، آنالیز و تعیین ویژگی های موثر می باشد. از مهمترین اهداف مدنظر در این گام، حذف ویژگی های غیرمرتبط با هدف مسئله که تشخیص بیماری قلبی است، می باشد. وجود ویژگی های نامناسب و مخرب می تواند در ساخت مدل پیش بینی تاثیر منفی داشته باشد و دقت مدل را کاهش دهد. همچنین بررسی این پارامترها می تواند اطلاعات مفیدی از نحوه ارتباط این پارامترها با یکدیگر و میزان اثرگذاری بر ویژگی هدف را دراختیار پزشکان قرار بدهد. همانطور که در راهکار پیشنهادی ارائه گردید برای این بخش از ۲ الگوریتم پیرسون و PCA استفاده شده است. در جدول ۱ نتایج بدست آمده از اجرای مجموعه داده کلیولند با تکنیک های معرفی شده می باشد که باتوجه به مقدار ضریبی دریافتی رتبه بندی شده است. باتوجه به رتبه های تخصیص داده شده توسط این ۲ الگوریتم، ۵ ویژگی Trestbps, Chol, Fbs، Gender و Restecg بصورت مشترک به عنوان ویژگی های نامطلوب انتخاب و از لیست پارامترهای موجود حذف گردید.

جدول۱ : نتایج حاصل از آنالیز ویژگی های مختلف در تشخیص بیماری قلبی

Name	Pearson (Rank)	PCA (Rank)	میزان تاثیر
Age	۰,۲۸ (۸)	۰,۲۸ (۸)	Weak
Gender	۰,۱۹ (۹)	۰,۱۱ (۱۱)	Weak
CP	۰,۴ (۶)	۰,۲۹ (۷)	Moderate
Trestbps	۰,۱۵ (۱۱)	۰,۱۶ (۹)	Independence
Chol	۰,۰۸ (۱۲)	۰,۰۹ (۱۲)	Independence
Fbs	۰,۰۲ (۱۳)	۰,۰۸ (۱۳)	Independence
Restecg	۰,۱۶ (۱۰)	۰,۱۵ (۱۰)	Independence
Thalach	۰,۴۶ (۵)	۰,۳۹ (۲)	Strong
Exang	۰,۴۳ (۳)	۰,۳۳ (۵)	Strong
Oldpeak	۰,۴۲ (۴)	۰,۴ (۱)	Strong
Slope	۰,۳۳۹۲ (۷)	۰,۳۵ (۳)	Moderate
Ca	۰,۴۶ (۲)	۰,۳۱ (۶)	Strong
Thal	۰,۵۳ (۱)	۰,۳۴ (۴)	Strong

۴,۴- نتایج مدل طبقه بندی

بعد از شناسایی ویژگی‌های مؤثر، حذف ویژگی‌های کم اهمیت و خوشه بندی داده ها در گام نهایی با استفاده از الگوریتم های طبقه بندی J۴۸، شبکه بیزین، SVM مدلی باهدف تشخیص وجود بیماری قلبی ساخته خواهد شد و براساس نتایج بدست آمده، بهترین مدل از نظر دقت پیش بینی به عنوان مدل نهایی ارائه و مشخص خواهد گردید. مجموعه داده ورودی برای این گام شامل ۳۰۳ نمونه و ۹ ویژگی (۸ ویژگی موثر و صفت هدف) است.

باتوجه به دقت بدست آمده از آموزش داده های بیماری قلبی-عروقی نشان متوسط مدل های پیش بینی، الگوریتم J۴۸ که یکی از روش های درخت تصمیم است، دارای بالاترین دقت که برابر ۰,۸۹ نمونه های دیتاست را به درستی پیش بینی کرده است.

جدول ۲: نتایج تکمیلی الگوریتم های طبقه بندی

	صحت	دقت	حساسیت	شفافیت
J۴۸	۰,۸۹	۰,۹۴	۰,۹۲	۰,۹۲
Bayesian Network	۰,۷۹	۰,۸۸	۰,۸۵	۰,۸۶
SVM	۰,۸۳	۰,۸۸	۰,۸۶	۰,۸۵

۵. نتیجه گیری

محیط مراکز بهداشتی و درمانی به تعداد زیاد داده در رابطه با بیماران قلبی و عروقی دارند اما یک فقدان ابزار تحلیل موثر برای کشف رابطه های مخفی بین بیماران قلبی و عروقی وجود دارد. پس داده کاوی می تواند به عنوان ابزار موثری در پیدا کردن اطلاعات پنهان بین بیماران باشد. تکنیک های داده کاوی برای پزشکان این کار را ممکن می سازد که اطلاعات تشخیص انواع بیماران با شرایط یکسان را پیش بینی نماید. از این رو در این مقاله به ارائه روشی به منظور پیش بینی بیماری قلبی و عروقی پرداخته شده است. گام های ارائه شده شامل ۳ فاز اصلی می باشند. در ابتدا پیش پردازش های لازم باهدف رفع مقادیر Null انجام گرفته است. برای رفع مقادیر Null از تکنیک پیش بینی و طبقه بندی نمونه ها استفاده شده است. به این صورت که نمونه هایی که مقادیر Null دارند را به عنوان داده های تست و سایر نمونه ها را به عنوان داده های آموزش در نظر گرفته ایم. در گام دوم به بررسی ویژگی های موثر در پیش بینی بیماری قلبی پرداخته شد و با بهره وری از ۲ روش ضریب همبستگی پیرسون و PCA صفاتی که باهدف مسئله ارتباطی ندارند، شناسایی و حذف گردید. در گام۳ با استفاده از ۳ الگوریتم J۴۸، شبکه بیزین و SVM مدل پیش بینی باهدف تعیین بیماری قلبی بیمار ساخته شد که الگوریتم J۴۸ با دقت ۸۹٪ دارای بالاترین دقت می باشد.

در کارهای آینده می توان از سایر روش های پیش بینی کننده استفاده کرد و به مقایسه با مدل پیشنهادی پرداخت. همچنین می توان با جمع آوری نمونه های بیشتر از مراکز مختلف بهداشتی فاکتورهای ریسکی که باعث شده افراد سابقه دار دوباره به این بیماری دچار شوند، را شناسایی نمود. موجود بودن اطلاعات دموگرافیک بیشتر و دقیق تر از بیماران از طریق مصاحبه حضوری از جمله سبک زندگی و تغذیه افراد و عامل مهم تر، قرارگیری فرد در شرایط استرس زا، نقش عوامل اجتماعی و روحی و روانی را در ایجاد بیماری قلبی-عروقی تا چه اندازه می تواند دخیل باشد، را بررسی و تحلیل نمود.

فهرست و مراجع

[۱] N. S. Chandra Reddy, S. Shue Nee, L. Zhi Min, and C. Xin Ying, "Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction," *Int. J. Innov. Comput.*, vol. ۹, no. ۱, ۲۰۱۹.

[۲] T. Karayilan and Ö. Kiliç, "Prediction of Heart disease using neural network," in *۲nd International Conference on Computer Science and Engineering, UBMK ۲۰۱۷*, ۲۰۱۷, pp. ۷۱۹–۷۲۳.

[۳] M. Raihan *et al.*, "Risk Prediction of Ischemic Heart Disease Using Artificial Neural Network," in *۲nd International Conference on Electrical, Computer and Communication Engineering, ECCE ۲۰۱۹*, ۲۰۱۹, pp. ۱–۵.

[۴] S. Radhimeenakshi and G. M. Nasira, "Prediction of Heart Disease using Neural Network with Back Propagation," *Int. J. Data Min. Tech. Appl.*, vol. ۴, no. ۱, pp. ۱۹–۲۲, ۲۰۱۵.

[۵] S. K. Sen, "Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms," *Int. J. Eng. Comput. Sci.*, vol. ۶, no. ۶, Jun. ۲۰۱۷.

[۶] S. Ansari *et al.*, "A review of automated methods for detection of myocardial ischemia and infarction using electrocardiogram and electronic health records," *IEEE Rev. Biomed. Eng.*, vol. ۱۰, pp. ۲۶۴–۲۹۸, ۲۰۱۷.

[۷] H. Amin *et al.*, "Predictive Analysis of Heart Disease Using K-Means and Apriori Algorithms," *J. Appl. Sci. Comput.*, vol. VI, pp. ۲۱۸۳–۲۱۸۹, ۲۰۱۹.

[۸] M. Mirmozaffari, A. Alinezhad, and A. Gilanpour, "Data Mining Apriori Algorithm for Heart Disease Prediction," *Int. J. Comput. Commun. Instrum. Eng.*, vol. ۴, no. ۱, ۲۰۱۷.

[۹] P. C. M.A.JABBAR ,DEEKSHATULU, "Intelligent heart disease prediction system using random forest and evolutionary approach," *J. Netw. Innov. Comput.*, vol. ۴, pp. ۱۷۵–۱۸۴, ۲۰۱۶.

[۱۰] A. H. Babolhakami, B. Behkamal, T. Dehghani, K. Etminani, and M. Naghibzadeh,

"Protein's number of beta-sheets prediction using structural features," in ۲۰۱۷ ۷th *International Conference on Computer and Knowledge Engineering, ICCKE* ۲۰۱۷, ۲۰۱۷, vol. ۲۰۱۷–Janua, pp. ٤٥٥–٤٦٠.

[۱۱] T. Smayra, Z. Charara, G. Sleilaty, G. Boustany, L. Menassa-Moussa, and G. Halaby, "Classification and Regression Tree (CART) model of sonographic signs in predicting thyroid nodules malignancy," *Eur. J. Radiol. Open*, vol. ٦, pp. ٣٤٣–٣٤٩, ۲۰۱۹.

[۱۲] U. Sidiq, S. Mutahar Aaqib, and R. A. Khan, "Diagnosis of Various Thyroid Ailments using Data Mining Classification Techniques," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, pp. ۱۳۱–۱۳٦, ۲۰۱۹.

[۱۳] UCI, "UCI Machine Learning Repository: Heart Disease Data Set," *Uci*, ۲۰۱۹. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Heart+Disease٪۰Ahttp://archive.ics.uci.edu/ml/datasets/Heart+Disease. [Accessed: ۲۳-Nov-۲۰۱۹].