مجله محاسبات نرم
SOFT COMPUTING JOURNAL
تارنمای مجله: scj.kashanu.ac.ir
دانشگاه کاشان
University of Kashan

بهبود کارایی الگوریتم‌های یادگیری ماشین در تشخیص بیماری‌های قلبی با بهینه‌سازی داده‌ها و ویژگی‌ها✧

هادی ویسی[1*]، استادیار، حمیدرضا قایدشرف[2]، دانشجوی دکتری، مرتضی ابراهیمی[3]، استادیار
[1و2و3] دانشکده علوم و فنون نوین، دانشگاه تهران، تهران، ایران.

اطلاعات مقاله

تاریخچه مقاله:
دریافت ۱۵ فروردین ماه ۱۳۹۹
پذیرش ۰۶ مرداد ماه ۱۳۹۹

کلمات کلیدی:
پیش‌بینی بیماری‌های قلبی
دسته‌بندی
الگوریتم‌های یادگیری ماشینی
شبکه‌های عصبی

چکیده

قلب یکی از مهم‌ترین اعضای بدن بوده و بیشترین علت مرگ‌ومیر در دنیا و ایران، بیماری‌های قلبی است. ازاین‌رو تشخیص زودهنگام و بموقع، یکی از ارکان مهم برای جلوگیری و کاهش مرگ‌ومیر ناشی از این بیماری است. هدف از این پژوهش، ایجاد مدل‌های تشخیص بیماری‌های قلبی با استفاده از روش‌های یادگیری ماشینی است. مدل‌ها بر روی مجموعه داده‌های قلب کلیولند دانشگاه کالیفرنیا، ایروین ایجاد شده است. با توجه به روش پیشنهادی پژوهش، پس از پردازش کامل داده‌ها که شامل شناسایی داده‌های پرت، نرمال‌سازی، گسسته‌سازی و انتخاب ویژگی می‌باشد، با توجه به ماهیت الگوریتم‌ها، داده‌ها به دو شکل داده‌های عددی نرمال‌شده و گسسته‌شده به بازه‌های بهینه، تغییر یافته است. همچنین ورودی الگوریتم‌های مورد استفاده، یک بار ویژگی‌های پردازش‌شده و بار دیگر ویژگی‌های ایجادشده توسط الگوریتم تحلیل مؤلفه‌های اصلی می‌باشد. از طرفی با استفاده از روش‌های جست‌وجوی تصادفی با اعتبارسنجی متقابل و جست‌وجوی شبکه‌ای از طریق Talos Scan پارامترهای مناسب هر الگوریتم انتخاب و مدل‌ها ایجاد و ارزیابی شده است. در بین الگوریتم‌های درخت تصمیم، جنگل تصادفی، ماشین بردار پشتیبان و XGBoost، بیشترین صحت مربوط به ماشین بردار پشتیبان به میزان ۹۲/۹٪ و در بین شبکه‌های عصبی بیشترین صحت به میزان ۹۴/۶٪، مربوط به شبکهٔ عصبی پرسپترون چندلایه است.

۱. مقدمه

در طول تاریخ، زندگی انسان‌ها و حیات آن‌ها تحت‌تأثیر بیماری‌های مختلفی بوده است؛ در بین این بیماری‌ها در علوم پزشکی، به بیماری‌های قلبی توجه زیادی شده است [۱]. طبق آمار سازمان جهانی بهداشت بیماری‌های قلبی‌عروقی، عامل اصلی مرگ‌ومیر در دنیا و کشورهای مختلف صنعتی بوده که نزدیک به ۴۷٪ از آن را به خود اختصاص داده است و تقریباً هر ۳۴ ثانیه یک بار، یک نفر در دنیا به‌علت این بیماری، جان خود را از دست می‌دهد [۲]. از طرف دیگر، در ایران طبق گزارش وزارت بهداشت، درمان و آموزش پزشکی کشور ۱۱ تا ۱۵٪ مرگ‌ومیر در کشور به‌علت بیماری‌های قلبی‌عروقی می‌باشد [۳]. بر اساس اطلاعات حاصل از یک مطالعهٔ همه‌گیرشناسی که در سال ۱۳۸۸ در ایران انجام شد، مشخص شد که از بین ۳۲۱۵۷۰

✧ نوع مقاله: پژوهشی
* نویسنده مسئول
پست‌های الکترونیک: h.veisi@ut.ac.ir (ویسی)
h.ghaedsharaf@ut.ac.ir (قایدشرف)
mo.ebrahimi@ut.ac.ir (ابراهیمی)

نحوه ارجاع به مقاله: ویسی، هادی، قایدشرف، حمیدرضا، ابراهیمی، مرتضی، «بهبود کارایی الگوریتم‌های یادگیری ماشین در تشخیص بیماری‌های قلبی با بهینه‌سازی داده‌ها و ویژگی‌ها»، مجله محاسبات نرم، جلد ۸، شماره ۱، ص ۷۰ـ۸۵، بهار و تابستان ۱۳۹۸.

فوتی، تعداد ۸۲۳۰۷ مورد ناشی از بیماری عروق کرونر قلبی بوده که این عدد معادل ۲۵/۶٪ از مرگ‌ومیر است. متخصصان قلب و عروق عوامل جسمانی زیادی از جمله فشارخون بالا، سطح بالای کلسترول مضر، بیماری دیابت، کمبود فعالیت بدنی، چاقی و وراثت را علت بروز بیماری عروق کرونر دانسته اما این موارد حداکثر ۵۰٪ از بروز این بیماری را پیش‌بینی می‌کنند و عوامل ذکرشده به‌تنهایی قادر به پیش‌بینی و علل بروز بیماری عروق کرونر قلب نیستند [۴]. بیماری‌های قلبی از آن دسته بیماری‌هایی هستند که یک فرد ممکن است این بیماری را داشته باشد اما به‌دلیل عدم آگاهی لازم از بیماری خود، بیماری تشدید و در نتیجه روند درمان سخت یا غیرقابل کنترل گردد. ازاین‌رو تشخیص و درمان بموقع این بیماری بسیار حائز اهمیت است. یکی از روش‌های تشخیص بیماری عروق کرونر قلب، آنژیوگرافی است اما با توجه به اینکه این‌گونه روش‌ها پرهزینه بوده، محققان همیشه به‌دنبال روش‌های جایگزین برای افزایش آگاهی افراد و تشخیص بموقع بیماری‌های قلبی بوده‌اند [۵]. تاکنون روش آنژیوگرافی عمومی‌ترین روش ارزیابی بیماری عروق کرونر قلبی بوده و از آن برای تشخیص این بیماری استفاده شده است، اما از آنجایی که این روش تهاجمی بوده، علاوه بر هزینه‌های گزافی که برای بیمار و کادر درمانی دارد، می‌تواند خطرهای زیادی همچون مرگ، سکتهٔ قلبی و مغزی را به همراه داشته باشد [۴]. امروزه از الگوریتم‌های یادگیری ماشینی و شبکه‌های عصبی برای ساخت مدل‌های پزشکی با اهداف مختلف استفاده می‌شود که در تمامی صنعت‌ها به‌ویژه صنعت سلامت، گردآوری، ذخیره و تجزیه و تحلیل داده‌ها و در نتیجه مدل‌سازی بیماری‌ها می‌تواند کمک بزرگی در حوزهٔ سلامت به بیماران و پزشکان کند. بیماری عروق کرونر از شایع‌ترین بیماری‌های قلبی بوده و وضعیتی است که کلسترول، کلسیم و دیگر چربی‌ها در شریان‌هایی که خون را به قلب می‌رسانند، انباشه شده و منجر به انسداد شریان قلب و مانع رسیدن مقدار کافی خون به قلب می‌شود. در نتیجه قلب با کمبود اکسیژن مواجه شده و طی آن فرد در قفسهٔ سینه خود، احساس درد می‌کند. تشخیص بموقع این بیماری، با هدف پیشگیری و کنترل بیماری، امری ضروری بوده و به همین علت از علم کامپیوتر و روش‌های هوش مصنوعی برای کمک به بیماران و تصمیم‌های پزشکی، استفاده‌های زیادی شده است [۲ و ۶].

۱ـ۱. پیشینهٔ پژوهش

پژوهش [۷] با هدف پیش‌بینی حملهٔ قلبی با استفاده از روش‌های داده‌کاوی و الگوریتم‌های یادگیری ماشینی انجام شده است. هدف از این پژوهش، آگاه‌سازی بیماران از وضعیت خود بوده تا بتوانند اقدام مناسب را برای پیشگیری و درمان انجام دهند. این پژوهش بر روی ۷۱۱ بیمار قلبی در یکی از بیمارستان‌های ایران انجام شده است. این مجموعه داده‌ای دارای ۲۸ ویژگی بوده که با استفاده از الگوریتم رگرسیون لجستیک[1]، مدلی برای پیش‌بینی خطر ابتلا به بیماری قلبی ساخته شده است و در بهترین حالت به میزان صحت ۹۴/۹٪ رسیده‌اند.

پژوهش [۸] با هدف پیش‌بینی بیماری‌های قلبی بر روی مجموعه داده‌های بیماران قلبی کلیولند انجام شده است. در این پژوهش با استفاده از الگوریتم‌های درخت تصمیم[2] و بیز ساده[3] مدل‌های پیش‌بینی ایجادشده است، که در بین مدل‌های ساخته‌شده، بیشترین صحت مربوط به الگوریتم بیز ساده و به میزان ۸۵/۰۳٪ می‌باشد.

پژوهش [۹] بر روی داده‌های بیماران قلب داده‌های قلب کلیولند دانشگاه کالیفرنیا – ایروین، برای پیش‌بینی بیماری قلبی انجام شده است. در این پژوهش از ۱۴ ویژگی معروف این داده‌ها (به‌کاررفته در اکثر پژوهش‌ها) استفاده شده و مدل‌های پیش‌بینی با استفاده از الگوریتم‌های درخت تصمیم، شبکهٔ عصبی مصنوعی[4] و بیز ساده ایجاد شده است که در بین الگوریتم‌های مورد استفاده، عملکرد تمامی روش‌ها به‌جز روش بیز ساده، با استفاده از روش انتخاب ویژگی، بهبود یافته است. این در حالی است که بیشترین صحت مربوط به الگوریتم بیز ساده به میزان ۸۲/۹۱٪ است.

پژوهش [۱۰] با هدف طراحی سیستم خبره برای پیش‌بینی

1. Logistic Regression
2. Decision Tree
3. Naïve Bayes
4. Neural Networks

بیماری قلبی انجام شده است. این پژوهش بر روی داده‌های بیماران قلبی کلیولند بوده که مدل‌های مختلفی از الگوریتم‌های درخت‌های تصمیم، ماشین بردار پشتیبان[1] و یک چهارچوب ابتکاری ترکیبی به نام M2-BagWeight ایجاد شده که در نهایت بیشترین صحت به میزان ۸۷/۳۷٪ مربوط به روش ترکیبی است.

پژوهش [۱۱] با هدف پیش‌بینی بیماری‌های قلبی بر روی مجموعه داده‌های بیماران قلبی کلیولند انجام شده است. در این پژوهش با استفاده از الگوریتم‌های شبکهٔ عصبی و درخت تصمیم (C4.5) مدل‌های پیش‌بینی ایجاد شده اما برای بهبود صحت مدل‌ها، از یک روش ترکیبی استفاده شد که در بین مدل‌های ساخته‌شده، بیشترین صحت مربوط به مدل ترکیبی و به میزان ۷۸/۱۴٪ است.

پژوهش [۱۲] به هدف پیش‌بینی بیماری‌های قلبی انجام شده است. در این پژوهش با استفاده از شبکه‌های عصبی با ساختار پرسپترون چندلایه (MLP)[2] مدل‌هایی بر روی مجموعه داده‌های قلب کلیولند ایجاد شد که بیشترین صحت برای داده‌های آزمون، ۸۳/۳۳٪ به دست آمده است.

پژوهش [۱۳] با هدف تعیین عوامل خطرزای بیماری‌های عروق کرونر با استفاده از روش‌های استخراج قوانین بر روی مجموعه داده‌های بیماران قلبی Z-Alizadeh-Sani انجام شده است. این مجموعه داده‌ای شامل ۳۰۳ نفر از مرکز آموزشی، درمانی قلب و عروق شهید رجایی تهران است. در این پژوهش از الگوریتم‌های آپریوری[3] و آپریوری پیش‌بینی‌کننده[4] برای استخراج عوامل خطرزای این بیماری استفاده شده است. میزان اطمینان مجموعه قوانین تولیدی از الگوریتم آپریوری، ۱۰۰٪ و میزان صحت پیش‌بینی‌شده با استفاده از الگوریتم آپریوری پیش‌بینی‌کننده، ۹۹/۳۷٪ بوده است.

۲. مواد و روش‌ها

در این بخش، روش‌ها و الگوریتم‌های مورد استفاده در این پژوهش معرفی شده است. روش مورد استفاده در بیشتر پژوهش‌های داده‌کاوی و یادگیری ماشینی به‌خصوص در صنعت سلامت، روش CRISP[5] بوده و طبق گزارشی که در اکتبر سال ۲۰۱۴ توسط گرگوری پیاتتسکی[6] در سایت KD Nuggets منتشر شد، محبوب‌ترین روش طبق این نظرسنجی، CRISP می‌باشد [۱۴]. این روش شامل شش مرحلهٔ شناخت کسب‌وکار، شناخت داده‌ها، آماده‌سازی داده‌ها، مدل‌سازی، ارزیابی مدل و توسعه و پیاده‌سازی می‌باشد. این روش یک فرایند چرخشی بوده و در برخی از مراحل امکان بازگشت به عقب وجود دارد [۱۵].

۲.۱. الگوریتم‌های مورد استفاده

در این قسمت، الگوریتم‌های مورد استفاده در پژوهش معرفی می‌شوند. شایان ذکر است که الگوریتم‌های این بخش به دو دسته الگوریتم‌های استفاده‌شده در بخش پردازش داده‌ها و الگوریتم‌های مورد استفاده در بخش مدل‌سازی تقسیم می‌شود.

۲.۱.۱. پردازش داده‌ها

با توجه به ماهیت برخی الگوریتم‌های یادگیری ماشین، بهتر است داده‌های ورودی به الگوریتم طوری تغییر شکل دهند که با الگوریتم مورد نظر سازگار شود. یکی از کارهای انجام‌شده در این پژوهش، گسسته‌سازی برخی از ویژگی‌های عددی است. مراحل گسسته‌سازی ویژگی‌ها در بخش ۴.۲.۳ قابل مشاهده است. روش‌های استفاده‌شده در این مرحله، استفاده از روش‌های گسسته‌سازی بهینه مبتنی بر روش MDLP[7] و گسسته‌سازی مبتنی بر بصری‌سازی داده‌ها (استفاده از تراکم جمعیتی یکسان) است. روش MDLP یک روش با ناظر است که متناسب با کلاس داده‌ها می‌تواند مناسب‌ترین بازه‌های عددی را تعیین کند [۱۶]. ایدهٔ اصلی این الگوریتم یک شرط است که می‌گوید گسسته‌سازی بازه‌ها زمانی قابل قبول است که حاصل Gain (S,A) که بهرهٔ اطلاعات برای ویژگی A می‌باشد، از آستانهٔ به‌دست‌آمدهٔ T از رابطهٔ (۱) بیشتر باشد [۱۷].

1. Support Vector Machine
2. Multilayer Perceptron
3. Apriori
4. Predictive Apriori

5. Cross-industry Standard Process
6. Piatetsky
7. Minimal Description Length Principle

$$C=(k * En(S)- k_1*En(S_1) -k_2 * En(S_2))$$
$$T = \frac{1}{n}\times \log_2(n-1) + \frac{1}{n}\times[\log_2(3^k-2) - C] \quad (۱)$$

که در آن، n تعداد عناصر تقسیم‌شده در یک بازه و k تعداد کلاس‌ها در یک گروه و En آنتروپی است. Gain (S,A) یکی از معروف‌ترین معیارهایی است که برای ساخت درخت تصمیم از آن استفاده می‌شود و خود برای انتخاب، از معیار آنتروپی استفاده می‌کند. در رابطهٔ (۲) می‌توان نحوهٔ محاسبهٔ این شاخص را مشاهده کرد.

$$Gain\ (S,A)= En(S) - En_A(S) \quad (۲)$$

که S مجموعه داده‌های آموزشی است. رابطه‌های (۳) و (۴) نحوهٔ محاسبهٔ آنتروپی S و آنتروپی ویژگی A را نشان می‌دهند.

$$En\ (S) = -\sum_{i=1}^{c} P_i\times \log_2(P_i) \quad (۳)$$

$$En_A\ (S) = \sum_{j=1}^{v} \frac{|S_j|}{|S|}\times En(S_j) \quad (۴)$$

که در آن، C تعداد کلاس‌های داده آموزشی S و P_i احتمال این را که نمونه‌ای از داده‌ها متعلق به کلاس i ام باشد، نشان می‌دهد. از طرفی، V تعداد اعضای ویژگی A و S_j قسمتی از داده‌های اولیه است که مقدار ویژگی آن‌ها V_j می‌باشد [۱۸]. در هر مجموعه داده‌ای، ویژگی‌های زیادی وجود دارد که تمامی برای هدف مسئله مناسب نیستند و می‌بایست مجموعه‌ای از ویژگی‌ها که ما را به اهداف تعیین‌شده نزدیک‌تر می‌کند، انتخاب شود. این ویژگی می‌تواند توسط یک فرد خبره و یا توسط الگوریتم انتخاب شود. یکی از شناخته‌شده‌ترین الگوریتم‌های کاهش و انتخاب ویژگی، الگوریتم تحلیل مؤلفه‌های اصلی (PCA)[1] می‌باشد. تحلیل مؤلفهٔ اساسی روشی آماری است که یک ترکیب خطی بین متغیرها ایجاد می‌کند. به‌عبارتی روشی برای تغییر نگاشت ویژگی‌ها از یک دستگاه مختصات به دستگاه مختصات دیگری بوده و به‌طور گسترده در کاهش بعد و استخراج ویژگی استفاده می‌شود. تحلیل مؤلفهٔ اساسی بر روی ماتریس کوواریانس اعمال شده تا بتواند زیرفضایی با بردارهای متعامد پیدا کند که در آن زیرفضا داده‌های نزدیک به هم را جدا و پراکنده می‌سازد. همچنین داده‌های تبدیل‌شده به‌صورت خطی از یکدیگر مستقل بوده و به‌عنوان مؤلفه‌های اصلی محسوب می‌شوند [۱۹ و ۲۰].

۲.۱.۲. الگوریتم‌های مدل‌سازی

- الگوریتم درخت تصمیم

درخت تصمیم روشی مرسوم و محبوب در طبقه‌بندی است. امروزه در حوزه‌های مختلف به‌خصوص حوزهٔ پزشکی از این الگوریتم به‌صورت گسترده استفاده می‌شود. ساختار درخت تصمیم در یادگیری ماشین، یک مدل پیش‌بینی‌کننده بوده و از پرکاربردترین روش‌های یادگیری ماشینی محسوب می‌شود. در یک درخت تصمیم برگ‌ها نشان‌دهندهٔ دسته‌بندی و شاخه‌ها و گره‌های میانی ویژگی‌های مختلف برای رسیدن به یک کلاس را نشان می‌دهند. درخت تصمیم را می‌توان به کمک مجموعه‌ای از شروط یا قوانین نمایش داد [۲۱ و ۲۲].

- الگوریتم جنگل تصادفی[2]

الگوریتم جنگل تصادفی یک روش دسته‌بندی بوده و به‌عنوان یکی از روش‌های محبوب و قدرتمند برای مسائلی با ابعاد بزرگ و پیچیده محسوب می‌شود. این الگوریتم گروهی از درخت‌های تصمیم بوده که هر درخت به هر نمونه یک رأی می‌دهد و مجموعه‌ای از درختان، بر اساس این آرا، دسته‌ای که بیشترین رأی را دارد، به‌عنوان کلاس داده‌ها انتخاب می‌کنند [۲۳].

- الگوریتم XGBoost

این الگوریتم از دسته الگوریتم‌های گرادیان تقویتی[3] بوده که عملکرد بسیار خوبی در دسته‌بندی، رگرسیون و رتبه‌بندی دارد و به‌دلیل پیش‌بینی دقیق، سرعت زیاد و پشتیبانی از اجرای چندمنظوره و توزیع‌شدهٔ آن، در مسائل دسته‌بندی بسیار محبوب است [۲۴].

- الگوریتم ماشین بردار پشتیبان

دسته‌ای از الگوریتم‌های یادگیری ماشینی با ناظر بوده که از آن‌ها برای کلاس‌بندی و رگرسیون بر پایهٔ نظریهٔ یادگیری آماری استفاده می‌شود [۲۵]. ماشین بردار پشتیبان هر داده را با توجه به کلاسشان به یک فضای جدید برده، به‌طوری که داده‌ها

1. Principal Component Analysis

2. Random Forest
3. Gradient Boosting Algorithms

به‌صورت خطی (ابرصفحه) قابل دسته‌بندی باشند. سپس با یافتن خطوط پشتیبان (صفحات پشتیبان در فضای چندبعدی)، معادلهٔ خطی را که بتواند بیشترین فاصله بین دسته‌ها ایجاد کند، پیدا می‌کند [۲۶].

- شبکهٔ عصبی پرسپترون چندلایه

شبکهٔ عصبی پرسپترون چندلایه یکی از محبوب‌ترین شبکه‌های عصبی پیش رو می‌باشد. این شبکه‌ها دارای سه دسته لایهٔ ورودی، مخفی و خروجی هستند. در شبکه‌های پرسپترون چندلایه، داده‌های ورودی به لایهٔ ورودی وارد شده تا مقادیر ورودی به نرون‌های ورودی محاسبه شود. پس از آن برای اعمال پردازش، آن‌ها را به لایه‌های مخفی ارسال و در نهایت برای محاسبهٔ مقدار خروجی شبکه، به نرون‌های لایهٔ خروجی وارد می‌شوند. این شبکه بر مبنای الگوریتم پس‌انتشار خطا آموزش می‌بیند به این صورت که خروجی‌های محاسبه‌شده از شبکه با خروجی‌های واقعی مقایسه و مقدار خطا محاسبه و به عقب برگشت داده می‌شود تا خطاهای لایه‌های میانی هم محاسبه گردد. در نهایت وزن‌های شبکه به‌روزرسانی و سعی می‌شود تابع هزینهٔ خطا در گام‌های مختلف زمانی کمینه گردد [۲۷].

۲.۲. معرفی داده‌های پژوهش

در این پژوهش، از مجموعه داده‌های بیماران قلبی کلیولند از مخزن داده‌ای وب‌سایت مرجع UCI استفاده شده است. این مجموعه داده‌ای به‌صورت عموم در دسترس و قابل استفاده است. این مجموعه شامل ۲۸۲ نمونه از افراد سالم و بیمار بوده و ۷۶ ویژگی دارد که در واقع ۳۰ ویژگی آن قابل استفاده می‌باشد؛ زیرا دیگر ویژگی‌ها به‌علت داشتن مقادیر ناموجود بیش از حد یا عدم تعادل داده‌ای در هر دسته، عملاً قابل استفاده نیست. در جدول (۱) مجموعه ویژگی‌های قابل استفاده توضیح داده شده است [۲۸]. شایان ذکر است در صورتی که ویژگی نیاز به گسسته شدن داشته باشد، بعد از اعمال مراحل گسسته‌سازی طبق شکل (۶) و تعیین بازهٔ مناسب، محدودهٔ هر ویژگی در جدول (۱) آورده شده است. شایان ذکر است محدودهٔ پزشکی ویژگی Chol در [۲۹]، ویژگی‌های TrestBPs و Trestbpd در [۳۰] و ویژگی Thalrest در [۳۱] به دست آمده است.

جدول (۱): معرفی ویژگی‌های این مجموعه داده‌ای به‌همراه بازه‌های گسسته‌شده

نام ویژگی و توضیحات آن	نوع ویژگی و محدودهٔ گسسته‌شده
Age (سن افراد)	(−∞,۵۱), (۵۱,۵۹], (۵۱,+∞)
Sex (جنسیت)	۰: زن ۱: مرد
Famhist (سابقهٔ خانوادگی)	۰: خیر ۱: بله
Cigs_Use (مصرف سیگار)	۰: خیر ۱: بله
CP (نوع درد قفسهٔ سینه)	1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic
Chol (سروم کلسترول خون) (mg/dl)	(-∞ ,۲۰۰) , [۲۰۰,۲۳۹] , (۲۳۹,+∞)
FBS قند خون (mg/dl)	۰: ۱۲۰ به بالا ۱: کمتر از ۱۲۰
TrestBPs (فشار خون سیستولی)	(−∞,۹۰), [۹۰,۱۲۰] (۱۲۰,۱۴۰], (۱۴۰,+∞)
Trestbpd (فشارخون دیاستولی)	(-∞,۶۰), [۶۰,۸۰] (۸۰,۹۰], (۹۰,+∞)
Tpeakbps (بیشترین فشارخون سیستولی در تست ورزش)	(−∞,۱۶۰] , (۱۶۰,۱۸۰], (۱۸۰,+∞)
Tpeakbpd (بیشترین فشار خون دیاستولی در تست ورزش)	(−∞,۷۵] , (۷۵,۸۰], (۸۰,+∞)
HTN (فشارخون بالا)	۰: خیر ۱: بله
RestECG (نتایج نوار قلب در حال استراحت)	۰: نرمال ۱: موج غیر قلبی ۲: افزایش ضخامت بطن چپ توسط معیارهای Estes
Exang (آنژین ناشی از ورزش)	۰: خیر ۱: بله
Thalrest (ضربان قلب در حال استراحت)	(−∞ ,۶۰), [۶۰,۱۰۰], (۱۰۰,+∞)
Thalach (حداکثر ضربان قلب)	(−∞,۱۴۸), [۱۴۸,+∞)
OldPeak (افت فاصلهٔ ST در نوار قلب ناشی از ورزش نسبت به استراحت)	(−∞,۱/۸), [۱/۸,+∞)
Slope (انحراف در شیب ST در زمان آزمون ورزش)	1: Upsloping 2: Flat 3: Downsloping
Thal (اسکن تالیوم)	3: normal 6: fixed defect 7: reversable defect

ادامه جدول (۱): معرفی ویژگی‌های این مجموعه داده‌ای به‌همراه بازه‌های گسسته‌شده

نام ویژگی و توضیحات آن	نوع ویژگی و محدودهٔ گسسته‌شده
CA تعداد عروق اصلی درگیر در فلورسکوپی	{۰و۱و۲و۳}
Class کلاس داده‌ها	وضعیت آنژیوگرافی ۰ : سالم (کمتر از ۵۰٪ تنگی قطر) ۱ : بیمار (بیش از ۵۰٪ تنگی قطر)
از ویژگی‌های lmt، Ladprox، Laddist، Cxmain، OM1، Rcaprox، Rcadist برای ساخت سطوح مختلف بیماری قلبی و در نتیجه تعیین سالم و بیمار بودن یک فرد استفاده شده است. در نتیجه در این پژوهش برای ساخت مدل، از ویژگی‌های ذکرشده استفاده نشده است.	

۳. روش پیشنهادی پژوهش

در این قسمت مراحل انجام‌شده در بخش‌های مختلف پژوهش توضیح داده می‌شود. بخش ۱.۳. این پژوهش مرحلهٔ شناخت کسب‌وکار و داده‌ها را توضیح می‌دهد. بخش ۳,۲ مرحلهٔ آماده‌سازی و پردازش داده‌ها را به‌تفصیل بیان می‌کند. این قسمت شامل مراحل انتخاب ویژگی، مدیریت داده‌های ناموجود، شناسایی و مدیریت داده‌های پرت، نرمال‌سازی و گسسته‌سازی داده‌ها و همچنین مصورسازی داده‌ها به‌منظور شناخت بهتر ویژگی‌هاست. بخش ۳.۳. مراحل مدل‌سازی، تفسیر و ارزیابی نتایج را نشان می‌دهد.

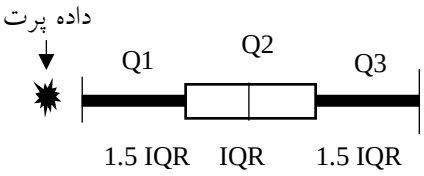
۱.۳. مرحلهٔ شناخت داده‌ها

در این قسمت از پژوهش، ضمن کسب آگاهی لازم از قلب و بیماری‌های قلبی، می‌بایست مجموعه داده‌ها را از منظر پزشکی و دید فنی بررسی کرد تا داده‌ها جهت ایجاد مدل برای پیش‌بینی بیماری‌های قلبی مناسب باشند. با توجه به مطالعه‌های انجام‌شده و پژوهش‌های زیادی که بر روی مجموعه داده‌های بیماران قلبی کلیولند انجام شده است، می‌توان پی برد که این مجموعه داده‌ای، انتخاب مناسبی برای ایجاد مدل با هدف پیش‌بینی بیماری‌های قلبی است. همچنین با توجه به اینکه داده‌های مورد استفاده در پژوهش داده‌های واقعی هستند که برچسب بیمار و یا سالم بودن افراد از طریق پزشکان و آزمایش‌های پزشکی زده شده است در نهایت بعد از ساخت مدل‌های پیش‌بینی و رسیدن به صحت قابل قبول برای مدل‌ها، می‌توان نتیجه گرفت که نتایج به‌دست‌آمده علاوه بر دیدگاه فنی از دید پزشکی هم قابل قبول بوده و می‌تواند برای پیش‌بینی بیماری‌های قلبی مورد استفاده قرار گیرد.

۲.۳. مرحلهٔ آماده‌سازی داده‌ها

۱.۲.۳. دادهٔ پرت تک‌متغیره

در دادهٔ پرت تک‌متغیره بدون درنظرگرفتن ویژگی‌های دیگر با توجه به نمودار جعبه‌ای، دادهٔ پرت شناسایی می‌شود. نمودار جعبه‌ای نموداری است که تغییرات داده‌ها را در یک ویژگی توصیف می‌کند. این نمودار بر اساس شاخص‌های آماری کوچکترین مقدار، چارک اول (Q1)، میانه، چارک سوم (Q3) و بزرگ‌ترین مقدار ساخته می‌شود. با استفاده از این نمودار که در شکل (۱) نشان داده شده، فاصلهٔ بین چارک اول و چارک سوم و همچنین میانهٔ جعبه یعنی چارک دوم قابل مشاهده است. حال با توجه به رابطهٔ (۵)، مقدار IQR محاسبه و با توجه به رابطه‌های (۶) و (۷) حد بالا و پایین این نمودار مشخص می‌شود.



شکل (۱) : نمودار جعبه‌ای و جزییات آن

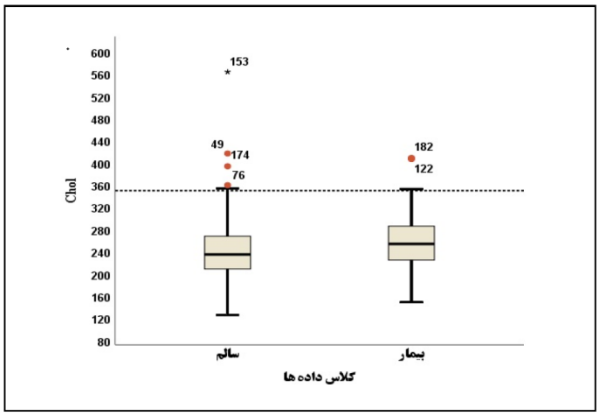
$$IQR = Q3 - Q1 \quad (۵)$$

$$\text{حد پایین} : Q1 - (1.5 * IQR) \quad (۶)$$

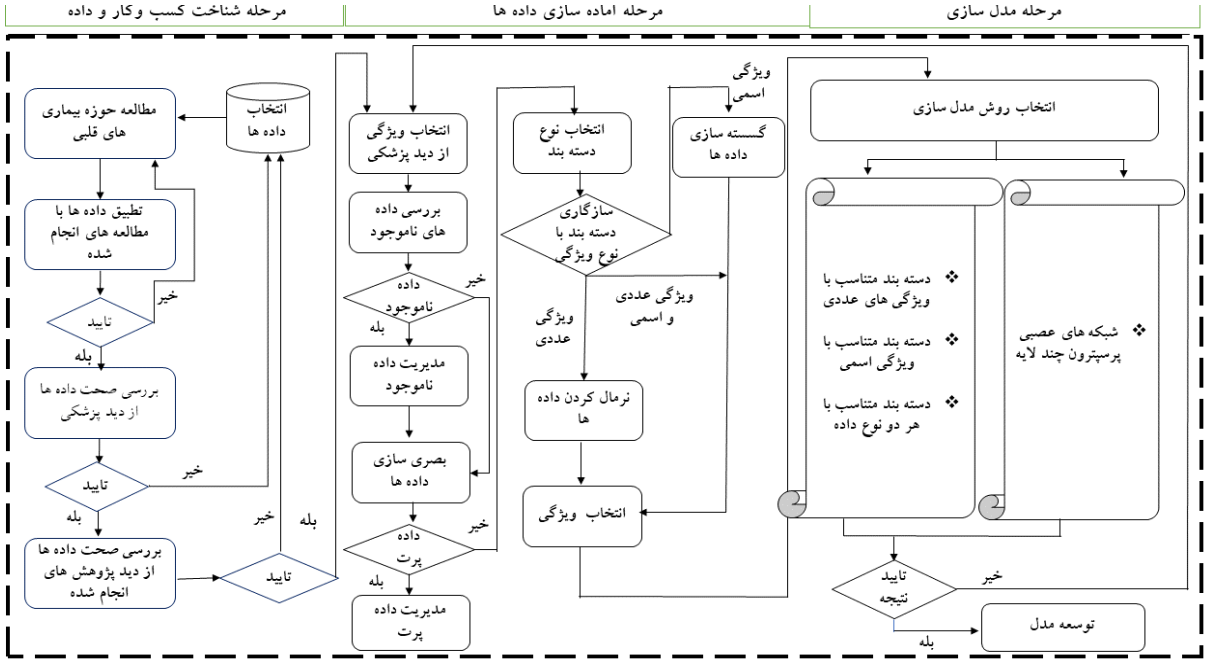
$$\text{حد بالا} : Q3 + (1.5 * IQR) \quad (۷)$$

هر داده‌ای که از حد بالا بیشتر و از حد پایین کمتر باشد داده پرت تلقی می‌شود. در شکل (۱)، داده‌های پرت روی تک‌ویژگی نشان داده شده است [۳۲]. با توجه به اینکه برخی از ویژگی‌ها از نوع ویژگی‌های پزشکی هستند، سیاست درنظرگرفته برای تبدیل داده‌های پرت به داده‌های نرمال این است که اگر داده‌ای دارای محدودهٔ پزشکی باشد و دادهٔ مورد نظر درون کلاس فرد بیمار قرار داشت، این داده تغییر نکند؛

زیرا می‌تواند جزء عوامل خطرزای بیماری قلبی باشد و به ساخت مدلی دقیق‌تر کمک کند، اما اگر فرد درون کلاس سالم قرار داشته باشد، داده‌های پرت به نزدیک‌ترین محدودهٔ نرمال خود تغییر می‌یابند. گفتنی است که در این پژوهش، ابتدا داده پرت تک‌متغیره بر روی هر ویژگی شناسایی و مدیریت شده و پس از آن، طبق بخش ۲.۲.۳ دادهٔ پرت چندمتغیره شناسایی و مدیریت می‌شود. شکل (۲) نمودار جعبه‌ای ویژگی کلسترول را نشان می‌دهد. با توجه به اینکه این ویژگی دارای محدودهٔ پزشکی است و داده‌های پرت درون محدوده بالای این ویژگی قرار گرفته است، برای افراد بیمار این داده‌ها یک عامل خطر تلقی شده و تغییر نخواهند کرد اما داده‌های پرت افراد سالم که تعداد چهار داده شناسایی شده است، به نزدیک‌ترین مقدار مجاز خود، یعنی حد بالا و حد پایین تغییر داده می‌شود. شکل (۳) مراحل مختلف پژوهش را نشان می‌دهد.



شکل (۲): نمودار جعبه‌ای کلسترول برای داده‌های پرت تک‌متغیره



شکل (۳): مراحل شناخت کسب‌وکار و داده، آماده‌سازی داده‌ها و ساخت مدل‌های پیش‌بینی بیماری‌های قلبی

۲.۲.۳. دادهٔ پرت چندمتغیره

برای شناسایی دادهٔ پرت چندمتغیره درون داده‌ها ویژگی وابسته را کلاس داده‌ها و ویژگی‌های مستقل، مجموعه‌ای از ویژگی‌های عددی در نظر گرفته شده است. پس از آن فاصلهٔ ماهالانوبیس[1] داده‌ها [۳۳] نسبت به کلاس محاسبه و از طریق نمودار جعبه‌ای و نمودار هیستوگرام، داده‌های پرت شناسایی و از مجموعه داده‌ها حذف شده است. با توجه به اینکه فاصلهٔ ماهالانوبیس از ماتریس کواریانس داده‌ها استفاده می‌کند، معیاری از فاصله هریک ازمشاهدات در فضای چندبعدی از مرکز میانگین تمام مشاهدات است. در نتیجه می‌تواند فاصلهٔ مناسبی برای تشخیص داده‌های پرت چندمتغیره باشد. شکل (۴) نمودار جعبه‌ای فاصلهٔ

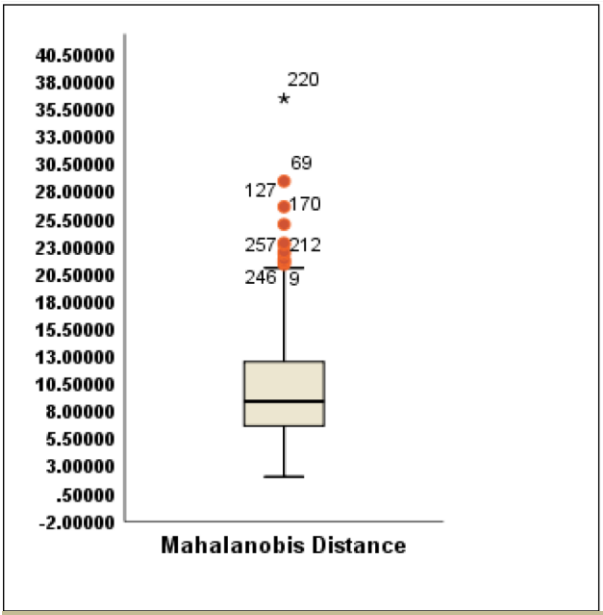
1. Mahalanobis

ماهالانوبیس داده‌ها را نسبت به کلاس نشان می‌دهد. همان طور که در بخش ۱.۲.۳ گفته شده، بعد از بررسی دادهٔ پرت درون هر ویژگی به بررسی دادهٔ پرت کلی (چندمتغیره) پرداخته شده است که در مجموع از بین تمام داده‌های پردازش‌شده مرحلهٔ ۳,۲,۱ تعداد ۸ داده (۸ فرد بیمار یا سالم) به‌عنوان دادهٔ پرت چندمتغیره شناسایی کرده و پس از آن حذف شده است.

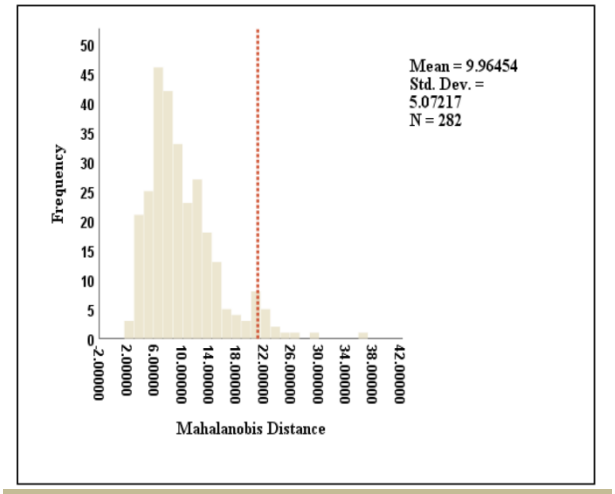
شکل (۵) میزان پراکندگی داده‌ها را بر اساس فاصلهٔ ماهالانوبیس داده‌ها نسبت به کلاس نشان می‌دهد. فاصلهٔ ماهالانوبیس هر داده با استفاده از رابطهٔ (۸) محاسبه می‌شود.

$$Mi^2 = (X_i - \mu_i)^T C^{-1} (X_i - \mu_i) \qquad (۸)$$

که در آن، M^2 فاصلهٔ ماهالانوبیس داده i، X_i بردار متغیرها برای نمونهٔ i ام و μ بردار میانگین مقادیر متغیرهای مستقل است. در این رابطه C ماتریس کواریانس داده‌های آموزشی (متغیرهای مستقل) بوده و منظور از T ترانهاده کردن عبارت داخل پرانتز می‌باشد. فاصلهٔ ماهالانوبیس مانند فاصلهٔ اقلیدسی است. با این تفاوت که فاصلهٔ ماهالانوبیس، توسط ماتریس کوواریانس نرمال‌سازی شده است [۳۴]. همان طور که در شکل (۵) قابل مشاهده است، داده‌هایی که فاصلهٔ ماهالانوبیس آن‌ها نسبت به کلاس بیشتر از مقدار ۲۱/۵ بوده، به‌عنوان دادهٔ پرت شناسایی شده و از مجموعه داده‌ها حذف شده است.



شکل (۴): نمودار جعبه‌ای فاصلهٔ ماهالانوبیس داده‌ها برای شناسایی داده‌های پرت چندمتغیره



شکل (۵): میزان پراکندگی داده‌ها بر اساس فاصلهٔ ماهالانوبیس نسبت به کلاس داده‌ها

۳.۲.۳. نرمال‌سازی داده‌های عددی

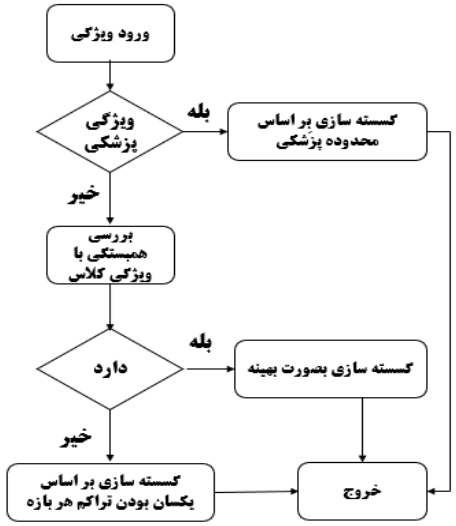
نرمال‌سازی داده‌ها یکی از روش‌های مقیاس‌بندی داده‌ها در فرآیند استفاده از الگوریتم‌های یادگیری ماشین می‌باشد و می‌تواند با اهداف مختلف بر روی ویژگی‌های عددی اعمال شود. در پژوهش‌هایی که در حوزه سلامت انجام می‌شود یکی از مسائل دسترسی به داده‌ها، امنیت و حفظ حریم خصوصی افراد است که با نرمال‌سازی داده‌ها تا حدودی می‌توان به این مهم دست یافت [۳۵]. نرمال‌سازی در فرآیند مدل‌سازی و عملکرد مدل بسیار حائز اهمیت بوده و در سرعت یادگیری مدل، تأثیر به سزایی دارد همچنین بیشتر در مواقعی کاربرد دارد که بازه‌های داده‌ها با یکدیگر متفاوت و هدف کاهش تأثیر منفی بازه‌های مختلف اعداد، بر روی عملکرد مدل می‌باشد. در این پژوهش برای نرمال‌سازی داده‌های عددی از روش نرمال‌سازی MIN-MAX استفاده شده است. مجموعه داده‌های هر ویژگی طبق رابطه (۹) به بازه‌هایی بین New_{max} و New_{min} که به‌ترتیب ۱ و ۰ می‌باشد، نگاشت شده است.

$$X_i = New_{min} + \left(New_{max} - New_{min}\right) * \frac{X_i - X_{min}}{X_{max} - X_{min}} \qquad (۹)$$

که X_{min} مقدار کمینهٔ داده قبل از نرمال‌سازی، X_{max} مقدار بیشینهٔ داده قبل از نرمال‌سازی، New_{min} مقدار کمینه بعد از نرمال‌سازی و New_{max} مقدار بیشینه بعد از نرمال‌سازی است. همچنین X_i به مقداری در بازهٔ New_{min} و New_{max} نگاشت می‌گردد [۳۶ و ۳۷].

۴.۲.۳. گسسته‌سازی داده‌های عددی

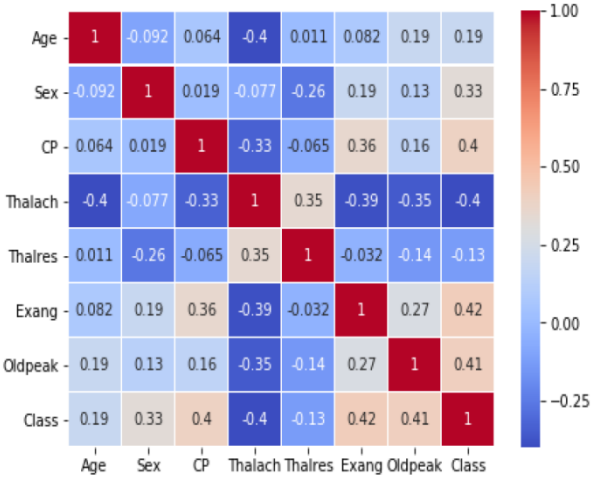
با توجه به ماهیت برخی از الگوریتم‌های مدل‌سازی به‌کاررفته در این پژوهش، می‌بایست ویژگی‌های عددی گسسته و به بازه‌های مناسب تبدیل شوند. با توجه به اینکه برخی از ویژگی‌ها خود دارای محدودهٔ پزشکی بوده، فرایند گسسته‌سازی طبق شکل (۶) انجام شده است. نتایج حاصل از گسسته‌سازی در جدول (۱) قابل مشاهده است.



شکل (۶): فرایند گسسته‌سازی ویژگی‌های عددی

۵.۲.۳. مصورسازی داده‌ها

یکی از مهم‌ترین مباحث در فرایند مدل‌سازی انتخاب ویژگی و شناخت همبستگی بین ویژگی‌هاست. شکل (۷) میزان همبستگی برخی از ویژگی‌های مهم این مجموعهٔ داده‌ای را نشان می‌دهد.



شکل (۷): نمایش همبستگی ویژگی‌ها

دیده می‌شود که ویژگی‌های Thal، Exang، Oldpeak، Thalach و CP بیشترین همبستگی با کلاس داده‌ها را دارد. نحوهٔ محاسبهٔ همبستگی دو ویژگی x و y طبق رابطهٔ (۱۰) محاسبه می‌شود. در این رابطه، Corr همبستگی دو ویژگی، Cov کوواریانس دو ویژگی، ρ انحراف‌معیار، E امید ریاضی و μ میانگین یک ویژگی است.

$$Corr(x,y) = \frac{Cov(x,y)}{\rho_x \rho_y} = \frac{E\left[(x-\mu_x)\times(y-\mu_y)\right]}{\rho_x \rho_y} \qquad (۱۰)$$

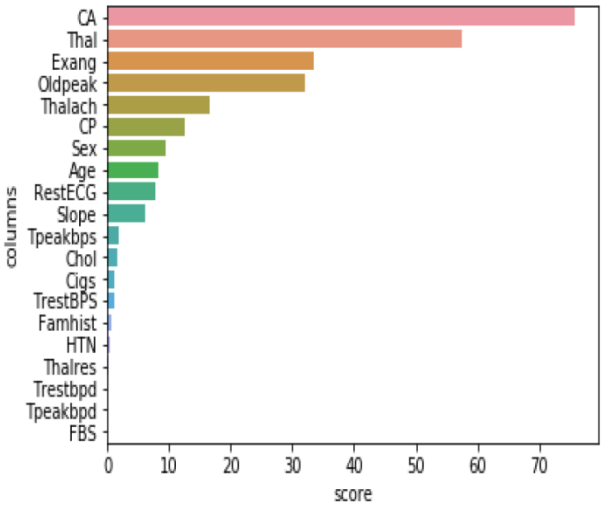
ضریب همبستگی عددی بین ۱ و ۱- می‌باشد و میزان رابطه و نوع رابطه بین دو ویژگی (مستقیم یا معکوس بودن) را نشان می‌دهد. در صورتی که دو ویژگی با یکدیگر رابطه نداشته باشند، همبستگی صفر خواهد بود [۳۸]. از دیگر معیارهای بررسی ویژگی‌ها آزمون کای دو پیرسون است که نمرهٔ هر ویژگی را نسبت به ویژگی کلاس محاسبه می‌کند. این آزمون از طریق رابطهٔ (۱۱) محاسبه می‌شود که در آن، X^2 نتیجهٔ آزمون، O فراوانی‌های مشاهده‌شده و E فراوانی‌های مورد انتظار است. این آزمون مناسب ویژگی‌های اسمی بوده و کار اصلی آن، بررسی معناداری تفاوت بین فراوانی‌های مشاهده‌شده و مورد انتظار می‌باشد که در این پژوهش بر روی مجموعه داده‌های گسسته‌شده، اعمال شده است [۳۹].

$$X^2 = \sum_{K=1}^{n} \frac{(O_K - E_k)}{E_k} \qquad (۱۱)$$

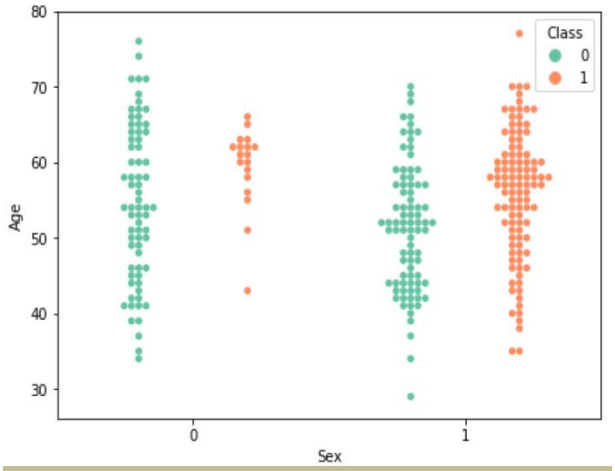
شایان ذکر است حضور یک ویژگی با درجه همبستگی یا نمرهٔ بالا نسبت به کلاس داده‌ها، در مدل‌سازی همواره منجر به بهبود عملکرد مدل نمی‌شود ولی احتمال بهبود کارایی مدل را افزایش می‌دهد. شکل (۸) نمره‌های هر ویژگی را با استفاده از آزمون کای دو پیرسون نشان می‌دهد. می‌توان مشاهده کرد که نتایج به‌دست‌آمده از این آزمون با نتایج حاصل از معیار همبستگی در شکل (۷) سازگار است.

شکل (۹) فراوانی افراد سالم و بیمار را به تفکیک جنسیت در سنین مختلف نشان می‌دهد. با توجه به شکل دیده می‌شود که مردان به نسبت زنان در سنین پایین‌تری دچار بیماری قلبی شده و همان طور که شکل (۱۰) هم نشان می‌دهد، در هر دو

جنسیت با افزایش سن، احتمال ابتلا به بیماری قلبی زیاد می‌شود.

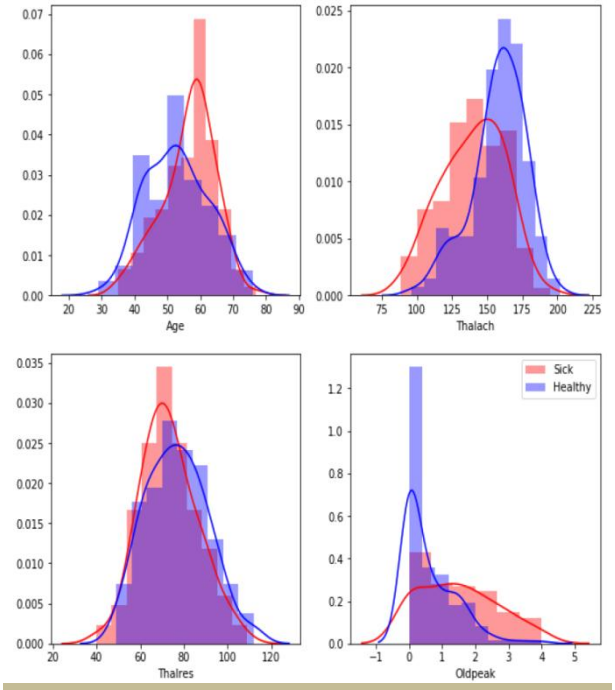


شکل (۸): رتبه‌بندی هر ویژگی نسبت به کلاس با توجه به آزمون کای دو پیرسون

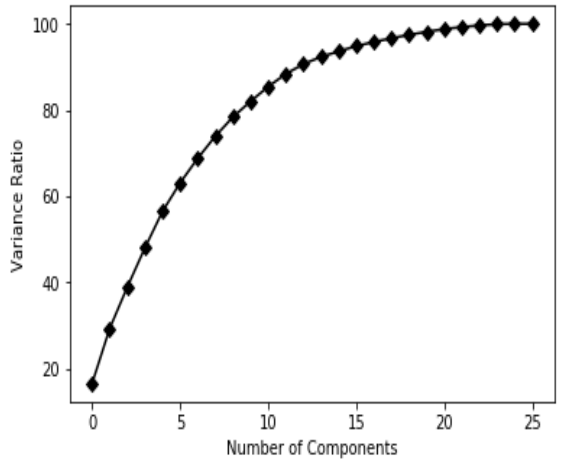


شکل (۹): فراوانی افراد سالم و بیمار به تفکیک جنسیت در سنین مختلف (۰: کلاس سالم و زنان و ۱ : کلاس بیمار و مردان)

شکل (۱۰) نشان می‌دهد که ویژگی Oldpeak برای افراد بیمار معمولاً از مقدار ۱/۸ بیشتر بوده که در فاز گسسته‌سازی هم این ویژگی به مقادیر کمتر و بیشتر از ۱/۸ تفکیک شده است. شکل (۱۱) جمع تجمعی نرخ واریانس را بر اساس تعداد مؤلفه‌های PCA نشان می‌دهد. دیده می‌شود که درنظرگرفتن ۱۰ تا ۱۵ مؤلفهٔ PCA برای کاهش ابعاد مناسب بوده؛ به همین علت در فاز مدل‌سازی با مؤلفه‌های PCA همین تعداد در نظر گرفته شده است.



شکل (۱۰): توزیع فراوانی ویژگی‌های Age، Thalres، Thalach، Oldpeak بر اساس کلاس داده‌ها



شکل (۱۱): جمع تجمعی نرخ واریانس براساس تعداد مؤلفه‌های PCA

به‌طور خلاصه با توجه به عکس شمارهٔ (۳) مرحلهٔ آماده‌سازی داده‌ها قبل از مرحلهٔ مدل‌سازی نشان داده شده است. این مرحله شامل مجموعه اقداماتی است که داده‌ها را به‌شکل مناسب برای ورود به مرحلهٔ مدل‌سازی آماده می‌کند. برای مثال برای بهبود عملکرد الگوریتم‌های مدل‌سازی و انتخاب بهتر پارامترهای ورودی به آن‌ها (مانند انتخاب مناسب تعداد ویژگی‌های ورودی به الگوریتم، انتخاب مناسب تعداد مؤلفه‌های

PCA و همچنین دستیابی به بازه‌های گسسته‌شدهٔ بهینه برای ویژگی‌های عددی) از تکنیک‌های مصورسازی داده‌ها و انتخاب ویژگی استفاده شده است. همچنین اقدامات دیگر انجام‌شده در مرحلهٔ آماده‌سازی داده‌ها همگی برای بالا بردن کارایی مدل‌های پیش‌بینی‌کنندهٔ بیماری‌های قلبی است. در ادامه مرحلهٔ مدل‌سازی و پیاده‌سازی شرح داده می‌شود.

۳.۳. مرحلهٔ مدل‌سازی و پیاده‌سازی

در این بخش در ابتدا برای انجام مدل‌سازی از الگوریتم‌های یادگیری ماشین درخت تصمیم، جنگل تصادفی، XGBoost و ماشین بردار پشتیبان استفاده شده و در جدول (۲) برخی از جزئیات الگوریتم‌های ذکرشده به‌همراه دادهٔ استفاده‌شده در هر الگوریتم به‌تفکیک آورده شده است. پس از آن مدل‌سازی با شبکه‌های عصبی پرسپترون چندلایه انجام شده است. شایان ذکر است برای آموزش تمامی الگوریتم‌های یادگیری ماشین از یک مجموعه واحد داده و برای مقایسهٔ نتایج الگوریتم‌ها، از ۲۰٪ داده‌ها به‌عنوان یک مجموعه دادهٔ آزمون واحد استفاده شده است. همچنین ورودی الگوریتم‌ها بر اساس جدول (۲) ابتدا ویژگی‌های گسسته‌شده – نرمال‌شده بوده و پس از آن از مؤلفه‌های الگوریتم PCA استفاده شده که برای هر دو حالت نتایج بر روی داده‌های آزمون ارزیابی و در جدول (۳) و (۵) آمده است. همچنین برای به‌دست‌آوردن پارامترهای نزدیک به بهینه برای الگوریتم‌های یادگیری ماشین، از جست‌وجوی تصادفی با استفاده از اعتبارسنجی متقابل و برای شبکهٔ عصبی MLP، از جست‌وجوی شبکه‌ای با روش Talos Scan [۴۰] استفاده شده است. ارزیابی مدل‌ها با استفاده از رابطه‌های (۱۲) تا (۱۵) انجام می‌شود. در بین الگوریتم‌های یادگیری ماشین بیشترین صحت مربوط به الگوریتم ماشین بردار پشتیبان به میزان ۹۲/۹٪ بوده و و در بین شبکه‌های عصبی بیشترین صحت به میزان ۹۴/۶٪ مربوط به شبکهٔ عصبی پرسپترون چندلایه است.

$$\text{صحت} = \frac{TP+TN}{TP+FN+FP+TN} \quad (۱۲)$$

$$\text{دقت} = \frac{TP}{TP+FP} \quad (۱۳)$$

$$\text{بازخوانی} = \frac{TP}{TP+FN} \quad (۱۴)$$

$$\text{معیار F1} = \frac{2\ Precision * Recall}{Precision + Recall} \quad (۱۵)$$

که در آن، TP[1]، TN[2]، FP[3] و FN[4] به‌ترتیب پیش‌بینی درست (کلاس واقعی مثبت)، پیش‌بینی درست (کلاس واقعی منفی)، پیش‌بینی غلط (کلاس واقعی منفی) و پیش‌بینی غلط (کلاس واقعی مثبت) است. رابطهٔ (۱۵) میانگین هارمونیک[5] معیارهای دقت و بازخوانی می‌باشد [۴۱].

در جدول (۳) D1 مجموعه داده‌ها با ویژگی‌های گسسته‌شده یا نرمال‌شده و D2 مجموعه ویژگی‌های ایجادشده توسط الگوریتم تحلیل مؤلفه‌های اصلی می‌باشد. همچنین DT معرف الگوریتم درخت تصمیم، RF معرف الگوریتم جنگل تصادفی، XG معرف الگوریتم XGBoost و SVM معرف الگوریتم ماشین بردار پشتیبان است. همچنین تمامی نتایج جدول (۳) و (۵) برحسب درصد محاسبه شده است. با توجه به جدول (۳) بیشترین میزان صحت در بین الگوریتم‌های یادگیری ماشینی مربوط به الگوریتم ماشین بردار پشتیبان و به میزان ۹۲/۹٪ می‌باشد.

در ادامه برای بهبود نتایج و بالابردن صحت مدل‌سازی، از شبکه‌های عصبی پرسپترون چندلایه استفاده شده است. یکی از مسائلی که در ارتباط با شبکه‌های عصبی وجود دارد، مسئلهٔ رخداد بیش‌برازش[6] مدل در شبکه‌هاست. به‌عبارت دیگر می‌بایست مدل ایجادشده به‌گونه‌ای باشد که تلاش برای کم کردن نرخ خطای داده‌های آموزش، موجب افزایش خطای داده‌های آزمون نگردد و مستلزم این است که زمان توقف شبکه و تعداد پارامترهای هر شبکه به‌درستی تنظیم شود.

1. True Positive
2. True Negative
3. False Positive
4. False Negative
5. Harmonic
6. Overfitting

جدول (۲): برخی از جزئیات الگوریتم‌های DT، RF، XG و SVM

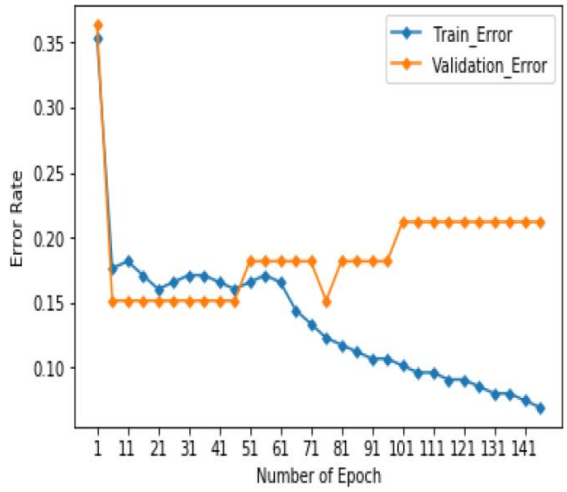
الگوریتم	مقادیر برخی متغیرهای بهینه حاصل از جست‌وجوی تصادفی با اعتبارسنجی متقابل
DT	• Criterion : شاخص جینی[1] • Max_Depth (حداکثر عمق درخت): ۱۷ • Min_Samples_Leaf (حداقل تعداد نمونهٔ لازم در برگ‌های درخت): ۱۵ • بیشتر مناسب ویژگی‌های اسمی (گسسته‌شده)
RF	• N_Estimators (تعداد درختان): ۱۰۰ • Criterion : شاخص جینی • Max_Depth (حداکثر عمق درخت): ۱۹ • Min_Samples_Leaf (حداقل تعداد نمونهٔ لازم در برگ‌های درخت): ۱۱ • نحوهٔ برچسب‌زنی نمونه‌های آزمون: بر اساس رأی اکثریت درخت‌ها برای یک نمونهٔ ورودی • بیشتر مناسب ویژگی‌های اسمی (گسسته‌شده)
XG	• Booster: درخت گرادیان تقویتی[2] • نرخ یادگیری: ۰/۶ • مناسب برای هم ویژگی‌های اسمی و هم عددی
SVM	• Kernel (هسته): خطی • C[3] (متغیر تنظیم): ۱ • loss function (تابع هزینه): Squared Hinge • بیشتر مناسب ویژگی‌های عددی

جدول (۳): نتایج مدل‌های ساخته‌شده با استفاده از الگوریتم‌های درخت تصمیم، جنگل تصادفی، XGBoost و ماشین بردار پشتیبان بر روی داده‌های آزمون

الگوریتم‌ها	داده‌ها	صحت	دقت	بازخوانی	معیار F1
DT	D1	۸۷/۵	۸۸/۱	۸۶/۱	۸۶/۸
DT	D2	۸۹/۳	۹۰/۶	۸۷/۶	۸۸/۶
RF	D1	۹۱/۱	۹۰/۶	۹۱/۱	۹۰/۸
RF	D2	۸۵/۷	۸۵/۲	۸۵/۲	۸۵/۲
XG	D1	۹۱/۱	۹۱	۹۰/۴	۹۰/۷
XG	D2	۸۷/۵	۸۷/۳	۸۶/۸	۸۷
SVM	D1	۹۲/۹	۹۳/۳	۹۲	۹۲/۵
SVM	D2	۹۲/۹	۹۳/۳	۹۲	۹۲/۵

برای دستیابی به تعداد تکرار مناسب، از مجموعه داده‌های آموزش ۱۵٪ داده‌ها را به‌عنوان دادهٔ ارزیابی انتخاب و با استفاده از شبکهٔ عصبی پرسپترون چندلایه، خطای داده‌های آموزش و ارزیابی محاسبه و محدودهٔ تعداد تکرار مناسب برای توقف الگوریتم آموزش، مشخص شده است.

معمولاً در شبکه‌های عصبی با افزایش تکرار آموزش، خطای داده‌های آموزش و داده‌های ارزیابی کاهش می‌یابد اما این مسئله تا زمانی صادق است که در بین گام‌های زمانی آموزش، هم خطای داده‌های آموزش و هم خطای داده‌های ارزیابی کاهش یابد و هنگامی که خطای داده آموزش کاهش و خطای داده‌های ارزیابی زیاد شود، می‌بایست الگوریتم آموزش متوقف گردد. در شکل (۱۲) مشاهده می‌شود که در محدودهٔ ۴۵ تا ۵۰ گام زمانی (تکرار) خطای آموزش و خطای ارزیابی کاهش یافته اما پس از آن با افزایش تکرار، خطای آموزش کاهش و خطای ارزیابی افزایش می‌یابد و می‌بایست الگوریتم آموزش متوقف شود. همچنین علاوه بر گام‌های تکرار، معماری شبکه در رخداد بیش‌برازش مدل بسیار مؤثر است؛ به‌طوری که می‌بایست تعداد لایه‌ها و نرون‌های هر لایه، طوری انتخاب شود که افزایش تعداد پارامترهای شبکه موجب کاهش کارایی شبکه بر روی داده‌های آزمون نگردد. جدول (۴) معماری شبکه‌های عصبی MLP را نشان می‌دهد. همچنین در جدول (۵) نتایج به‌دست‌آمده از شبکه‌های عصبی MLP بر روی داده‌های آزمون، آورده شده است.



شکل (۱۲): مقایسهٔ خطای داده‌های آموزش و ارزیابی در تکرارهای مختلف

1. Gini Index
2. Gradient Tree Boosting
3. Regularization Parameter

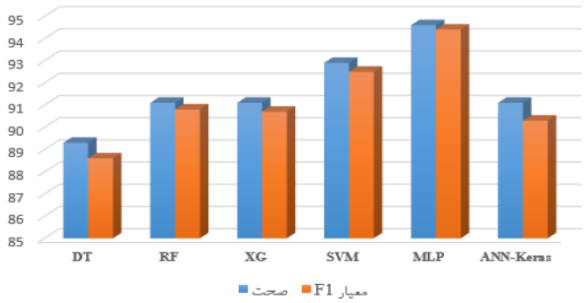
جدول (۴): معماری شبکه‌های عصبی MLP استفاده‌شده در این پژوهش

شبکه	تعداد لایه‌های مخفی	تعداد نرون‌های لایه‌ها	توابع فعال‌ساز لایه‌ها
MLP	یک لایهٔ مخفی	لایهٔ مخفی: ۱۰ لایهٔ خروجی: ۱	لایه مخفی: Relu لایهٔ خروجی:Sigmoid
MLP	دو لایهٔ مخفی	لایهٔ مخفی اول: ۱۲ لایهٔ مخفی دوم: ۸ لایهٔ خروجی: ۲	هر دو لایهٔ مخفی:Relu لایهٔ خروجی: Softmax

جدول (۵): نتایج به‌دست‌آمده از شبکه‌های عصبی MLP بر روی داده‌های آزمون

شبکهٔ عصبی	داده‌ها	صحت	دقت	بازخوانی	معیار F1
MLP یک لایهٔ مخفی	D1	۹۴/۶	۹۴/۸	۹۴/۱	۹۴/۴
MLP یک لایهٔ مخفی	D2	۸۷/۵	۸۷/۳	۸۶/۸	۸۷
MLP دو لایهٔ مخفی	D1	۹۱/۱	۹۳/۴	۸۹/۱	۹۰/۳
MLP دو لایهٔ مخفی	D2	۸۹/۳	۸۸/۹	۸۸/۹	۸۸/۹

با توجه به جدول (۵)، بیشترین صحت به میزان ۹۴/۶٪ مربوط به شبکهٔ عصبی پرسپترون چندلایه می‌باشد که شبکه بر روی داده‌های نرمال‌شده، آموزش دیده است. شکل (۱۳) مقایسهٔ تمام الگوریتم‌های به‌کاررفته در این پژوهش را نشان می‌دهد که برای هر الگوریتم بیشترین میزان صحت کسب‌شده، در نظر گرفته شده است.



شکل (۱۳): بهترین نتایج به‌دست‌آمده از تمام الگوریتم‌های استفاده‌شده در پژوهش

۴. مقایسهٔ روش پیشنهادی با دیگر پژوهش‌ها

در جدول (۶) بهترین نتیجهٔ به‌دست‌آمده از مدل‌های ایجادشده در این پژوهش با نتایج حاصل از پیشینهٔ پژوهش به‌صورت اختصار آورده شده و بهبود نتیجهٔ مدل‌سازی پیش‌بینی بیماری قلبی بر روی داده‌های قلب UCI در این پژوهش به‌وضوح دیده می‌شود. شایان ذکر است در تمامی پژوهش های ذکرشده در جدول (۶) از داده‌های قلب UCI استفاده شده و در این پژوهش هم در ابتدا ۲۰٪ داده‌ها به‌عنوان داده‌های آزمون به‌صورت کاملاً تصادفی انتخاب و برای ارزیابی تمامی مدل‌ها استفاده شده است و چون دسترسی به داده‌های آزمون پژوهش‌های مورد مقایسه امکان پذیر نیست، ممکن است برخی نمونه‌های داده‌های آزمون این پژوهش با داده‌های آزمون دیگر پژوهش‌ها متفاوت باشد. همچنین می‌بایست اشاره کرد که در قسمت‌های مختلف مراحل پردازش و آماده‌سازی داده‌ها علاوه بر زبان برنامه‌نویسی پایتون از ابزارهای SPSS Statistics و رپیدماینر[1] استفاده شده و در بخش مدل‌سازی از زبان برنامه‌نویسی پایتون برای ایجاد مدل‌های مختلف پیش‌بینی بیماری‌های قلبی استفاده شده است.

جدول (۶): مقایسهٔ بهترین نتایج حاصل از پژوهش‌های قبلی با بهترین نتیجهٔ حاصل از این پژوهش

	پژوهش‌ها	بهترین صحت ٪
۱	پژوهش ما	۹۴/۶
۲	[۸] Venkatalakshmi et al	۸۵/۰۳
۳	[۹] Shafique et al	۸۲/۹۱
۴	[۱۲] Kazemi et al	۸۳/۳۳
۵	[۱۰] Bashir et al	۸۷/۳۷

۵. نتیجه‌گیری

در این پژوهش به‌علت اهمیت تشخیص زودهنگام بیماری‌های قلبی، مدل‌هایی برای پیش‌بینی این بیماری با استفاده از الگوریتم‌های یادگیری ماشینی مختلف و شبکه‌های عصبی ایجاد شده است. در روش پیشنهادی این پژوهش با توجه به اینکه زمانی مدل‌های پیش‌بینی می‌توانند عملکرد قابل قبولی داشته باشند که داده‌ها به بهترین شکل ممکن پردازش و آماده مدل‌سازی شده باشند، سعی شده پردازش‌های مختلفی بر روی داده‌ها انجام گیرد و متناسب با هر الگوریتم مدل‌سازی داده‌ها

1. Rapidminer

پردازش و آماده شوند. به همین دلیل، منطبق با اقدامات ذکرشده در بخش ۲.۳ این پژوهش مجموعه داده‌ای به دو صورت داده‌های نرمال‌شده و داده‌های گسسته‌شده، تغییر شکل یافته است. همچنین برای ساخت مدل‌ها یک بار از داده‌های نرمال‌شده-گسسته‌شده و بار دوم از ویژگی‌های ایجادشده توسط الگوریتم تحلیل مؤلفه‌های اصلی استفاده شده است که در الگوریتم درخت تصمیم صحت دسته‌بندی هنگامی که از مؤلفه‌های اصلی به جای ویژگی‌های گسسته‌شده استفاده شده است، بهبود یافته است. در نهایت برای ایجاد مدل‌های پیش‌بینی‌کنندهٔ بیماری‌های قلبی از الگوریتم‌های مدل‌سازی مختلفی با درنظرگرفتن بهینه‌سازی پارامتر، استفاده شده است. بیشترین صحت در بین الگوریتم‌های یادگیری ماشین، مربوط به الگوریتم ماشین بردار پشتیبان به میزان ۹۲/۹٪ می‌باشد. در بخش دیگر این پژوهش برای ارتقای صحت مدل‌های پیش‌بینی‌کنندهٔ بیماری از شبکهٔ عصبی MLP استفاده شده است که بیشترین صحت مربوط به شبکهٔ عصبی MLP با تک‌لایهٔ مخفی، به میزان ۹۴/۶٪ می‌باشد. با توجه به اینکه در این‌گونه شبکه‌ها با افزایش تعداد پارامترها، می‌بایست داده به اندازهٔ کافی باشد، دیده می‌شود وقتی که از یک لایهٔ مخفی با ۱۰ نرون استفاده شود، بیشترین صحت به میزان ۹۴/۶٪ به دست می‌آید. شایان ذکر است در شبکه‌های عصبی، پیدا کردن پارامترهای بهینه برای هر شبکه کار بسیار پیچیده و بااهمیتی است. برای همین از جست‌وجوی شبکه‌ای با روش Scan Talos استفاده شده است. در انتها اگر بخواهیم به برخی از نتایج پزشکی حاصل‌شده از این پژوهش اشاره کنیم، می‌توان گفت که بیماری‌های قلبی با سن افراد ارتباط مستقیم داشته و با بالارفتن سن خطر ابتلا به بیماری‌های قلبی بیشتر می‌شود. همچنین سن بروز بیماری‌های قلبی در مردان پایین‌تر از زنان است و زنان به نسبت مردان در سنین بالاتری دچار بیماری قلبی می‌شوند و ویژگی‌های CA، CP، Oldpeak، Exang، Thalو Thalach از ویژگی‌های بسیار مهم در پیش‌بینی و بروز علائم بیماری‌های قلبی هستند.

مراجع

[1] Nahar, J., Imam, T., Tickle, K.S. and Chen, Y.-P. P., "Association rule mining to detect factors which contribute to heart disease in males and females", *Expert Syst. Appl.*, Vol. 40, No. 4, pp. 1086–1093, 2013.

[2] Nikhil Kumar, K. D. M. and Koushik, K. V. S., "Prediction of Heart Diseases using Data Mining and Machine Learning Algorithms and Tools", *nternational J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, Vol. 3, No. 3, pp. 887–898, 2018.

[3] Mahmoodi, M.S., "Designing a Heart Disease prediction System using Support Vector Machine", *J. Heal. Biomed. Informatics*, Vol. 4, No. 1, 2017.

[4] Hassanzadeh, M., Zabbah, I. and Layeghi, K., "Diagnosis of Coronary Heart Disease using Mixture of Experts Method", J. Heal. Biomed. Informatics, Vol. 5, No. 2, 2018.

[5] Dekamin, A. and Sheibatolhamdi, A., "Research Paper: A Data Mining Approach for Coronary Artery Disease Prediction in Iran", *J. Adv. Med. Sci. Appl. Technol. Adv. Med. Sci. Appl. Technol.*, Vol. 3, No. 31, pp. 29–38, 2017.

[۶] حسین نژادگرگری، میهن، اصغری، زینب، ولیان خروانق، علی، «**نقش داده‌کاوی در سلامت**»، کنفرانس بین‌المللی پژوهش در علوم و تکنولوژی، تهران، مؤسسه سرآمد همایش کارین، ۱۳۹۴.

[7] Soleimani, P. and Neshati, A., "Applying the Regression Technique for Prediction of the Acute Heart Attack", *World Acad. Sci. Eng. Technol. Int. J. Medical, Heal. Biomed. Bioeng. Pharm. Eng.*, Vol. 9, No. 11, pp. 763–767, 2015.

[8] Venkatalakshmi, B. and Shivsankar, M.V., "Heart Disease Diagnosis using Predictive Data Mining", *Int. J. Innov. Res. Sci. Eng. Technol.*, Vol. 3, No. 3, pp. 1873–1877, 2014.

[9] Shafique, U., Majeed, F., Qaiser, H. and Mustafa, I.U., "Data mining in healthcare for heart diseases", *Int. J. Innov. Appl. Stud.*, Vol. 10, No. 4, p. 1312, 2015.

[10] Bashir, S., Qamar, U. and Khan, F.H., "WebMAC: A web based clinical expert system", *Inf. Syst. Front.*, Vol. 20, No. 5, pp. 1135–1151, Oct. 2018.

[11] Maji, S. and Arora, S., "Decision Tree Algorithms for Prediction of Heart Disease", in *Information and Communication Technology for Competitive Strategies*, 2019, pp. 447–454.

[12] Kazemi, M., Mehdizadeh, H. and Shiri, A., "Heart disease forecast using neural network data mining

technique", *J. ilam Univ. Med. Sci.*, Vol. 25, No. 1, pp. 20–32, May 2017.

[13] Ghaedsharaf, H., Sadredini, M.H., Khayami, R. and Babaei Beigi, M.A., "Extract effective factors Incidence of coronary artery disease using association rules", in *The 1st National Conference on Recent Advances in Engineering and Modern Sciences*, 1397.

[14] Piatetsky, G., "Latest KDnuggets Poll asked", 2014.[Online] Available: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html.

[15] Niaksu, O., "CRISP Data Mining Methodology Extension for Medical Domain", *Balt. J. Mod. Comput.*, Vol. 3, pp. 92–109, 2015.

[16] IBM Knowledge Center, "Overview (OPTIMAL BINNING command)." [Online] Available: https://www.ibm.com/support/knowledgecenter/en/SSLVMB_23.0.0/spss/data_validation/syn_optimal-binning_overview.html.

[17] Mathias, "Discretizing a continuous variable using Entropy", 2013. [Online] Available: http://clear-lines.com/blog/post/Discretizing-a-continuous-variable-using-Entropy.aspx.

[18] Badr El Din Ahmed, A. and Sayed Elaraby, I., "PER: A prediction for Student's Performance Using Decision Tree ID3 Method", *India - World J. Comput. Appl. Technol.*, Vol. 2, No. 2, pp. 43–47, 2014.

[19] Wang, Y., Ma, X. and Qian, P., "Wind Turbine Fault Detection and Identification Through PCA-Based Optimal Variable Selection", *IEEE Trans. Sustain. Energy*, Vol. 9, No. 4, pp. 1627–1635, 2018.

[20] Wold, S., Esbensen, K. I. M. and Geladi, P., "Principal Component Analysis", Vol. 2, pp. 37–52, 1987.

[21] Han, J., Kamber, M. and Pei, J., *Data Mining: Concepts and Techniques*. 2012.

[۲۲] محجوبی، جواد، اعتماد شهیدی، امیر فرشاد، «تخمین **ارتفاع امواج ناشی از باد در نکاء به کمک درختان تصمیم رگرسیونی**»، اولین کنفرانس داده‌کاوی ایران، تهران، دانشگاه صنعتی امیرکبیر، مؤسسهٔ پژوهشی داده‌پردازان گیتا، ۱۳۸۶.

[23] Kumar, S. and Sahoo, G., "A Random Forest Classifier based on Genetic Algorithm for Cardiovascular Diseases Diagnosis", Vol. 30, No. 11, pp. 1723–1729, 2017.

[24] Mitchell R. and Frank E., "Accelerating the XGBoost algorithm using GPU computing", *PeerJ Comput. Sci.*, Vol. 3, p. e127, 2017.

[25] Belaid, S. and Mellit, A., "Prediction of daily and mean monthly global solar radiation using support vector machine in an arid climate", *Energy Convers. Manag.*, Vol. 118, pp. 105–118, 2016.

[۲۶] یادگاری، وحید، متین‌فر، احمدرضا، «**شناسایی حملات منع سرویس وب با استفاده از آنتروپی و الگوریتم ماشین بردار پشتیبان**»، مجله پدافند الکترونیکی، شماره ۶، صفحات ۷۹ـ۸۹، ۱۳۹۷.

[27] Ansari, H.R., Zarei, M.J., Sabbaghi, S. and Keshavarz, P., "A new comprehensive model for relative viscosity of various nanofluids using feed-forward back-propagation MLP neural networks", *Int. Commun. Heat Mass Transf.*, Vol. 91, No. December 2017, pp. 158–164, 2018.

[28] "UCI Machine Learning Repository: Heart Disease Data Set." [Online] Available: https://archive.ics.uci.edu/ml/datasets/heart+Diseas.

[29] "Cholesterol levels by age: Differences and recommendations." [Online] Available: https://www.medicalnewstoday.com/articles/315900.php.

[30] "Blood Pressure: Blood pressure chart." [Online] Available:http://www.bloodpressureuk.org/BloodPressureandyou/Thebasics/Bloodpressurechart.

[31] "Heart rate: What is a normal heart rate?" [Online] Available:https://www.medicalnewstoday.com/articles/235710.php.

[32] Barbato, G., Barini, E.M., Genta, G. and Levi, R., "Features and performance of some outlier detection methods", *J. Appl. Stat.*, Vol. 38, No. 10, pp. 2133–2149, 2011.

[33] Leys, C., Klein, O., Dominicy, Y. and Ley, C., "Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance", *J. Exp. Soc. Psychol.*, Vol. 74, pp. 150–156, Jan. 2018.

[۳۴] مهرگینی، بهزاد، معماریان، حسین، «**ارزیابی کارایی روش فاصلهٔ ماهالانوبیس در تفکیک رخساره‌های نفتی، در یکی از میادین هیدروکربوری ایران**»، چهاردهمین کنفرانس ژئوفیزیک ایران، تهران، انجمن ژئوپلتیک ایران، ۱۳۸۹ .

[35] Saranya, C. and Manikandan, G., "A study on normalization techniques for privacy preserving data mining", *Int. J. Eng. Technol.*, Vol. 5, No. 3, pp. 2701–2704, 2013.

[36] Liu, Z., "Procedia A method of SVM with Normalization in Intrusion Detection", *Procedia Environ. Sci.*, Vol. 11, pp. 256–262, 2011.

[37] Sajedi, H. and Taslimi, M., "Author gender identification from text using Bayesian Random Forest", *Signal Data Process.*, Vol. 16, No. 1, 2019.

[38] Benesty, J., Chen, J. and Huang, Y., "On the Importance of the Pearson Correlation Coefficient in Noise Reduction", *IEEE Trans. Audio. Speech. Lang. Processing*, Vol. 16, No. 4, pp. 757–765, May 2008.

[39] Seong, J.H. and Seo, D.H., "Wi-Fi fingerprint using radio map model based on MDLP and euclidean distance based on the Chi squared test", *Wirel. Networks*, Vol. 9, pp. 1–9, 2018.

[40] Maria Jensen, "Using Talos for Feature Hyperparameter Optimization?" [Online]

Available:https://neurospace.io/blog/2019/04/using-talos-for-feature-hyperparameter-optimization/.

[41] Desai, S.D., Dessai, I.F. and Kulkarni, L., "Intelligent Heart Disease Prediction System Using Probabilistic Neural Network", *J. Adv. Comput. Theory Eng.*, Vol. 2, No. 3, pp. 38–44, 2013.