مجله‌ی علمی پژوهش در علوم رایانه ، شماره ۱۳ ، بهار ۱۳۹۸، ص ۴۸- ۳۳

ISSN: 2745-3673
http://www.Rcsj.ir

مجله‌ی علمی پژوهش در علوم رایانه (سال چهارم)
شماره ۱۳ / بهار ۱۳۹۸ / ص ۴۸- ۳۳

برسی عملکرد تکنیک‌های داده‌کاوی قواعد انجمنی و K همسایه نزدیک در پیش‌بینی میزان زنده ماندن افراد مبتلابه هپاتیت

محسن قربیان
گروه علوم کامپیوتر، واحد کاشان، دانشگاه آزاد اسلامی، کاشان، ایران.

نام نویسنده مسئول:
محسن قربیان

چکیده

گستردگی تکنیک‌ها داده‌کاوی موجود این امکان را فراهم کرده است تا بتوان با انتخاب تکنیک‌های مناسب، به نتایجی بهتر و با ضریب دقت بالاتر دست‌یافت؛ بنابراین با انتخاب تکنیک بهتر می‌توان به نتایج حاصل از داده‌کاوی صورت گرفته بر روی‌داده‌ها، اطمینان بیشتری حاصل نمود. وجود عوامل مختلف دخیل در بروز بیماری هپاتیت و همچنین ناقص بودن اطلاعات در دسترس را می‌توان از مشکلات اساسی در پیاده‌سازی تکنیک‌های داده‌کاوی دانست که می‌تواند بر روی نتایج نهایی تحقیق اثرگذار باشد. عوامل و نشانه‌های مختلفی در تشخیص بیماری هپاتیت به کار گرفته می‌شوند که از طریق انجام آزمایش‌های گوناگون این عوامل و نشانه‌ها مورد برسی قرار می‌گیرند. در این تحقیق از بین عوامل که منجر به تشخیص بیماری هپاتیت می‌شوند نشانه‌هایی همانند ANOREXIA، LIVER BIG، LIVER FIRM و SPLEEN PALPABLE مورداستفاده قرار می‌گیرند. ازاین‌رو با پیاده‌سازی تکنیک‌های داده‌کاوی قواعد انجمنی و K همسایه نزدیک بافاصله اطمینان ۸۰ درصد و مقدار پشتیبان ۵۰ درصد بر روی اطلاعات افراد مبتلابه هپاتیت و بر اساس معیارهای سنجش accuracy، error rate، specificity، negative prediction value عملکردشان مورد برسی قرار گرفت و نتایج حاصل گویای این است که تکنیک قواعد انجمنی با مقدار 80/62 accuracy، 22/11 error rate، 95/27 specificity و 62/96 negative prediction value عملکرد بهتری را نسبت به تکنیک K همسایه نزدیک از خود نشان داده است. از بین عواملی که به‌عنوان نشانه‌های پیدایش بیماری هپاتیت محسوب می‌شوند عواملی که درنهایت فرد بیمار با دارا بودن آن‌ها می تواند شانس زنده ماندن را پیدا کند پیش‌بینی شدند. ازاین‌رو بر اساس نتایج به دست آمده می‌توان نتیجه گرفت که افرادی که در نتایج مرتبط با آزمایش خود با ویژگی‌های ANOREXIA، LIVER BIG، LIVER FIRM و SPLEEN PALPABLE مثبت مواجه شده است می‌توانند از شانس زنده ماندن برخوردار باشند.

واژگان کلیدی: داده‌کاوی پزشکی، قواعد انجمنی، K همسایه نزدیک، هپاتیت، تکنیک‌های داده‌کاوی.

ISSN: 2745-3673
http://www.Rcsj.ir

مقدمه

به‌کارگیری کامپیوتر در علوم گوناگون و در ابعاد مختلف هم‌زمان با پیشرفت فنّاوری زمینه برای تولید بیش‌ازپیش اطلاعات دیجیتالی فراهم نموده است که این اطلاعات را می‌توان در قالب داده‌هایی با ساختار متناسب با خود در مخازنی تحت عنوان انبار داده ذخیره نمود [1]. ذخیره‌سازی داده‌ها در انبار داده هارا می‌توان فاز اول استفاده از داده‌های تولیدشده قلمداد کرد ازاین‌رو باید این اطلاعات ذخیره‌شده استفاده نمود. پردازش و استفاده از این اطلاعات با این حجم عظیم خارج از توان انسان است زیرا این اطلاعات دارای روابط و الگوهای نهانی است که تنها با استفاده از تکنولوژی‌ها و فناوری‌های ایجادشده به این منظور می‌توان آن‌ها را تحلیل کرد و مورداستفاده قرارداد. نتیجه بکار گیری این تحلیل‌ها را می‌توان برنامه‌ریزی برای آینده دانست که می‌تواند به نتایج مهمی منتهی گردد [2]. بنابراین می‌توان نتیجه گرفت که فناوری تحلیل اطلاعات ازجمله فناوری‌های بسیار مهم و کاربردی است که می‌تواند به فرایند کشف دانش منتهی گردد ازاین‌رو فناوری بکار رفته در تحلیل اطلاعات را داده‌کاوی می‌نامند[3] . حوزه کاربرد داده‌کاوی[1] را نمی‌توان به یک یا دو حوزه محدود کرد زیرا این فناوری را می‌توان در هرجایی که داده‌ای موجود باشد به کاربرد و از آن می‌توان به‌منظور آنالیز و تحلیل داده‌ها استفاده کرد دراین‌بین داده‌هایی که مرتبط با حوزه پزشکی هستند هم ا این امر مستثنا نیستند و می‌توان آن هار را نیز مورد تحلیل و کاوش قرارداد[4]. برای تحلیل اطلاعات مرتبط با حوزه پزشکی باید به این امر دقت کرد که ازآنجایی‌که این داده‌ها با سلامتی انسان‌ها در ارتباط هستند ازاین‌رو اگر تکنیک‌های داده‌کاوی بر روی این داده‌ها پیاده‌سازی شوند نتایج حاصل باید از دقت و صحت مناسبی برخوردار باشند [5]. با بکار گیری تکنیک‌های داده‌کاوی و پیاده‌سازی آن‌ها بر روی‌داده‌ها بیماران می‌توان به اطلاعات و الگوهای نهان موجود در درون آن‌ها پی برد[6]. که نتایج کشف این الگوها می‌تواند در کاهش زمان تشخیص یک بیماری یا پیش‌بینی یک بیماری مورد استفاده قرار گیرد[7] . با توجه به گستردگی تکنیک‌های داده‌کاوی موجود لازم است تا تکنیک‌هایی را انتخاب کرد که بتوانند نتایجی با دقت و صحت مناسب را ارائه دهند ازاین‌رو باید با انتخاب تکنیکی که با تحلیل قوی خود قابلیت ارائه اطلاعاتی را داشته باشد که از صحت و دقت بالاتری برخوردار باشد [8]. برای به‌کارگیری فناوری داده‌کاوی نیز در حوزه‌ی سلامت نیاز است تا تکنیک‌هایی مورداستفاده قرار گیرند که بتوانند دقت و صحت نتایج حاصل را در بالاترین سطح ممکن به نمایش بگذارند تا بتوان به نتایج حاصل اطمینان کرد[9]. با توجه به اهمیتی که نتایج حاصله در پیاده تکنیک‌های داده‌کاوی بر روی‌داده‌های مرتبط به حوزه سلامت، بنابراین نیاز است تا تکنیک‌هایی که انتخاب می‌شوند ازلحاظ ارائه نتایج قابل‌اطمینان باشند [10]. برای انتخاب تکنیکی که بتواند عملکرد مناسب و قابل‌اطمینانی از خود بروز دهد نیاز است تا این تکنیک‌ها مورد مقایسه قرار گیرند. مقایسه تکنیک‌ها از طریق معیارهایی صورت می‌گیرد که به معیارهای سنجش نیز معروف می‌باشند ازاین‌رو تکنیکی که بتواند با استفاده از این معیارهای سنجش نتایج قابل قبولی از خود به نمایش بگذارد را می‌توان به‌عنوان تکنیکی معرفی نمود که نسبت به تکنیک‌های دیگر در تحلیل اطلاعات و قابلت اطمینان نتایج حاصل عملکرد بهتری از خود به نمایش گذاشته است[11]. هدف از این تحقیق بررسی عملکرد دو تکنیک داده‌کاوی K همسایه نزدیک[2] و قواعد انجمنی[3] بر اساس چهار معیار accuracy (معنای صحت) ، error rate (نسبت خطا)، specificity (ویژگی) و negative prediction value (مقادیر منفی پیش بینی شده) است که می‌تواند منجر به انتخاب تکنیکی گردد که عملکرد مناسب‌تری را از خود به نمایش گذاشته است.

۱- پیشینه تحقیق

با گسترش علم و ظهور تکنولوژی‌های پیشرفته راه‌های حل مشکلات روزبه‌روز سهل و محمل تر شده است و این ویژگی منحصر به حوزه‌ی خاصی نیست و می‌توان ادعا کرد که استفاده از تکنولوژی در همه زمینه‌ها به‌عنوان یک اولویت در نظر گرفته می‌شود. حوزه پزشکی یکی از حوضه‌هایی است که انسان همواره با آن درگیر بوده است و متخصصین امر در این حوزه همیشه به دنبال افزایش کیفیت زندگی انسان‌ها بوده‌اند ازاین‌رو این متخصصین همیشه در تلاش بوده‌اند تا با مواردی همچون پیش‌گیری، افزایش سرعت تشخیص بیماری، کاهش هزینه‌های بیماری و کند کردن روند و سرعت رشد یک بیماری به هدف خود که همانا افزایش کیفیت زندگی انسان‌هاست نزدیک شوند. گسترش تکنولوژی در این حوزه نیز اثرات بخصوص خود را گذاشته است و باعث شده است تا با انجام آزمایش‌های گوناگون و یا به‌گونه‌ای دیگر با اختراع دستگاه‌های جراحی پیشرفته در خدمت بهبود کیفیت زندگی انسان باشند. فناوری داده‌کاوی را می‌توان یکی از تکنولوژی‌هایی دانست که با کاوش در اطلاعات مرتبط با بیماران و کشف رابطه میان این اطلاعات به الگوهایی دست یابد که می‌تواند به متخصصین امر در این حوزه یک دید استراتژیک بدهد تا بتوانند بر اساس روابط کشف‌شده میان این اطلاعات به الگوهایی دست یابند که این الگوها می‌تواند آن‌ها را در جهت تشخیص سریع‌تر یا پیش‌بینی یک بیماری کمک کند که نتیجه آن می‌تواند به کاهش هزینه‌های

---

[1] Data Mining
[2] K-Nearest Neighbors
[3] Rule Induction

تحمیلی به بیمار و همچنین افزایش اطمینان به روند بهبود سریع‌تر بیمار منجر گردد. طبیعی است که به دست آوردن اطلاعات با صحت و دقت بالا می‌تواند کمکی شایان به متخصصین نماید ازاین‌رو مهم است تا تکنیکی که انتخاب می‌گردد ازنظر نتیجه حاصل، قابل‌اطمینان باشد. ازاین‌رو در سالیان اخیر متخصصین داده‌کاوی با همکاری متخصصین مربوط به حوزه پزشکی سعی کرده‌اند تا داده‌های مرتبط به بیماران را مورد تحلیل قرار دهند. در سال ۲۰۱۰ آقایان Rajeswari و Reena با استفاده از تکنیک‌های داده‌کاوی اقدام به تجزیه‌وتحلیل اطلاعات مرتبط با بیمارانی کردند که دچار مشکل کبدی بودند ازاین‌رو این دو متخصص با استفاده از نتایج حاصل از پیاده‌سازی تکنیک‌های داده‌کاوی بر روی ۳۴۵ نمونه توانستند میزان تأثیرگذاری مصرف الکل را در بروز اختلال در عملکرد کبد را پیش‌بینی کنند و سرعت رشد بیماری را در بیمارانی که الکل مصرف کرده‌اند را پیش‌بینی کنند[12]. اما در سال ۲۰۰۷ آقای Nguyen به همراه همکاران توانستند که الگوریتم داده‌کاوی جدیدی را معرفی کنند که این الگوریتم با استفاده از الگوهای ایجادشده بین زمانی که افراد در آن دوره زمانی به بیماری هپاتیت مبتلا شده‌اند و همچنین عوامل مؤثر در بروز بیماری توانسته‌اند تا به الگویی دست یابند که این الگو به پزشکان کمک خواهد کرد تا بتوانند زمان تشخیص این بیماری را بهبود ببخشند[13]. در سال ۲۰۰۷ آقای Kim و همکاران با استفاده از تکنیک‌های داده‌کاوی ماشین بردار و درخت تصمیم‌گیری توانستند تا میزان حساسیت کبدهایی که درگیر هستند و امکان تبدیل‌شدن این حساسیت به بیماری مزمن هپاتیت را مورد بررسی قراردادند و سعی کردن تا عملکرد این دو تکنیک را در تشخیص این مهم مورد ارزیابی قرار دهند[14]. در سال ۲۰۱۳ آقای Radwan و همکاران با استفاده از تکنیک داده‌کاوی درخت تصمیم و پیاده‌سازی آن بر روی‌داده‌های بیماران مبتلابه هپاتیت توانستند مدل‌هایی را پیش‌بینی کنند که به‌واسطه آن مدل‌ها توانستند نتیجه استفاده از داروهای ضدویروسی مربوط به هپاتیت را پیش‌بینی کنند که این تحقیق می‌تواند منجر به کاهش هزینه‌های تحمیلی بر بیمار گردد[15]. در سال ۲۰۰۵ آقای Yokoi و همکاران با پیاده‌سازی تکنیک‌های داده‌کاوی بر روی اطلاعات به دست آمده از آزمایش‌های ادرار بیماران مبتلابه هپاتیت و تحلیل نتایج حاصل توانستند به مدل و الگویی دست یابند که می‌تواند به متخصصان امر در حوزه درمان هپاتیت کمک شایانی نماید و روند درمان این بیماری را تسریع بخشد[16]. در سال ۲۰۰۲ آقای Sato و همکاران با پیاده‌سازی تکنیک درخت تصمیم بر روی اطلاعات بیماران مبتلابه هپاتیت و تحلیل نتایج به‌دست‌آمده توانستند به الگوهایی دست یابد. سپس این محققان با ترکیب الگوهای به‌دست‌آمده توانستند به مکانیزمی دست یابند که می‌توانست کار متخصصان را بسیار آسان نماید زیرا متخصصان برای به دست آوردن چنین نتایجی از طریق آزمایش، نیاز به انجام آزمایش های بسیار سخت و طولانی‌مدت داشتند[17].

۲- روش پژوهش

یکی از چالش‌های اساسی برای محققان در فرایند داده‌کاوی انتخاب تکنیکی است که متناسب با حوزه فعالیت مربوطه باشد. به‌عنوان نمونه هنگامی داده‌های مورداستفاده مربوط به حوزه پزشکی و سلامت است اهمیت نتایج حاصل دوچندان خواهد شد زیرا بر اساس این اطلاعات متخصصین امر در مورد نحوه درمان یا تشخیص یک بیماری تصمیم‌گیری می‌کنند که این مهم می‌تواند در سلامت فرد تأثیر مستقیمی بگذارد. ازاین‌رو نیاز است تا تکنیک‌هایی انتخاب شوند که بتوانند بالاترین سطح از اطمینان و درستی را در نتایج به‌دست‌آمده ارائه کنند تا بتوان درست‌ترین تصمیم را با بهترین تحلیل صورت گرفته را همراه نمود و نتیجه را به بهترین شکل ممکن دریافت کرد. در این تحقیق سعی شده است تا دو تکنیک K همسایه نزدیک و قواعد انجمنی را به‌عنوان تکنیک‌های منتخب برگزید که دلیل این امر را می‌توان گستردگی استفاده از این دو تکنیک و همچنین مطلوب بوده کیفیت تحلیل‌های ارائه‌شده از جانب این دو تکنیک دانست. از مدل ایجادشده توسط این تکنیک به‌منظور نمایش حقایق استنتاج شده مربوط به متغیر هدف استفاده می‌شود. مدل‌های ارائه‌شده توسط قواعد انجمنی بسیار ساده و قابل‌فهم می‌باشند و در قالب قوانینی ساده بیان می‌شوند که ازاین‌رو تفسیر این قوانین به‌سادگی صورت می‌گیرد[18] . تکنیک K همسایه نزدیک داده‌ها را در گروه‌هایی تقسیم‌بندی می‌کند که دارای ویژگی‌های نزدیک به یکدیگر می‌باشند و درصورتی‌که نمونه‌ی جدیدی را بخواهد مورد بررسی قرار دهد این نمونه را در قسمت گروه‌هایی قرار می‌دهد که به نمونه مورد نظر بیشترین شباهت را داشته باشد. این تکنیک را می‌توان به‌عنوان یکی از ساده‌ترین تکنیک‌ها در مبحث طبقه‌بندی داده‌ها در نظر گرفت. تکنیک K همسایه نزدیک سعی می‌کند با قرار دادن نمونه‌ای جدید و با استفاده از اطلاعات طبقه‌بندی داده‌های قبلی، با قرار دادن نمونه‌های جدید در طبقه‌بندی نزدیک به آن‌ها سعی در افزایش دقت و تحلیل‌های حاصل نماید[19]. هپاتیت[۴] در لغت به معنی متورم شدن است که این امر می‌تواند دلایل گوناگونی را داشته باشد، مصرف دخانیات، مصرف مشروبات الکلی و یا مواد شیمیایی می‌توانند از دلایل متورم شدن کبد محسوب شوند. هپاتیت دارای انواع گوناگونی است که از این انواع می‌توان به موارد زیر اشاره نمود:

---

[4] Hepatitis

هپاتیت نوع A
هپاتیت نوع B
هپاتیت نوع C

تفاوت انواع هپاتیت را می‌توان در نوع ویروسی که باعث التهاب و متورم شدن کبد می‌شود بیان کرد. بیماری‌های هپاتیت A و هپاتیت B دارای واکسن می‌باشند که انسان می‌تواند با واکسینه کردن خود از ابتلا به این بیماری در امان باشد ولی از طرف دیگر بیماری هپاتیت C فاقد واکسن است و نمی‌توان از طریق واکسینه کردن از ابتلا به آن در امان بود. ازاین‌رو این امکان نیز وجود داد که یک فرد به‌طور هم‌زمان به دو یا سه نوع هپاتیت مبتلا گردد. بیماری هپاتیت C را می‌توان از خطرناک‌ترین انواع بیماری هپاتیت محسوب کرد. راه انتقال این نوع بیماری از طریق خون به‌صورت مستقیم و یا خال‌کوبی، مصرف مواد مخدر و طب سنتی قابل‌انتقال است. [20]. داده‌های مورداستفاده در این تحقیق مربوط به ۱۵۵ بیمار مبتلابه هپاتیت است. این اطلاعات توسط دانشگاه کالیفرنیا در ارواین تهیه شده‌اند که به‌عنوان یکی از معتبرترین دانشگاه‌های ایلات متحده امریکا محسوب می‌شود. اطلاعات موجود در این داده‌ها را می‌توان به قسمت های مختلفی دسته‌بندی کرد. این اطلاعات در مورد افرادی می‌باشند که بر روی آن‌ها آزمایش‌های یکسانی صورت گرفته‌اند و در انتها ضمن مشخص کردن جنسیت آن‌ها میزان مرگ‌ومیر افراد مربوطه را به نمایش گذاشته است. در این تحقیق سعی بر آن است تا با استفاده از تکنیک‌های داده‌کاوی ضمن مقایسه عملکرد تکنیک‌های مطرح‌شده بر اساس معیارهای سنجش تعریف‌شده، با بررسی و تحلیل نتایج حاصل احتمال زنده ماندن فردی که آزمایش‌های مربوطه بر روی آن صورت گرفته است پیش‌بینی شود. دراین‌بین به علت تکثر آزمایش‌های صورت گرفته سعی بر آن شده است تا چهار آزمایش را به‌عنوان متغیرهای این تحقیق در نظر بگیریم و این چهار متغیر تحت لقای کلاس زنده ماندن یا فوت کردن بیمار قرار می‌گیرند. این چهار متغیر را می‌توان به این صورت نمایش داد:

- ANOREXIA (بی اشتهایی)
- LIVER BIG (بزرگی کبد)
- LIVER FIRM (کبد چرب)
- SPLEEN PALPABLE (قابل لمس بودن طهال)

روش انجام تحقیق در (شکل ۱) نشان داده‌شده است.

شکل ۱: مکانیزم پیاده‌سازی تحقیق

۳-۱- پردازش اولیه

به‌منظور استفاده درست و بهینه از تکنیک‌های داده‌کاوی نیاز است تا ساختار اطلاعات واردشده به این تکنیک‌ها استاندارد و مطابق با ساختار در نظر گرفته‌شده برای آن‌ها باشد. از این داده‌های مرتبط با بیماران هپاتیت نیز از این قاعده مستثنا نیستند و می‌بایست دارای ساختاری متناسب با تکنیک‌های دسته‌بندی استفاده‌شده درداده کاوی باشند. پس‌نیاز است تا در اطلاعات خام اولیه تغیراتی ایجاد شود.

۳-۱-۱- تبدیل اطلاعات بیماران به قالب استاندارد

تکنیک‌های داده‌کاوی در نظر گرفته‌شده برای این تحقیق از دسته تکنیک‌های رده‌بندی[۵] می‌باشند ازاین‌رو نیاز است تا اطلاعات بیماران هپاتیت ساختارمند گردند و آن‌ها را در قالب سطر و ستون نمایش داد. از طرفی به دلیل این‌که این اطلاعات دارای ساختار ابتدایی بوده‌اند پس‌نیاز است با اندکی تغیر آن‌ها را به فرم موردنظر تبدیل نمود.

۳-۱-۲- تبدیل نتایج حاصل از آزمایش به YوN

بعد از ساختاردهی ابتدایی به اطلاعات نیاز است تا در مقادیر موجود تغیراتی اعمال شود. اطلاعات بیماران مربوط به آزمایش‌های صورت گرفته بر روی آن‌ها است، از این رو نتایج به‌صورت اعداد نمایش داده‌شده‌اند و از طرفی این اعداد دارای مقادیر ۲ و ۱ می‌باشند ازاین‌رو با توجه به توضیحات مندرج در ارتباط با اطلاعات این بیماران نتایج آزمایش‌ها به‌صورت Yes و No در نظر گرفته‌شده‌اند و ۱ به این

---

5. Classification

معنا است که در صورتی که در مقابل متغیری مقدار ۱ قرارگرفته شده باشد یعنی این‌که جواب آزمایش فرد برای آن متغیر مقدار No بوده است و درصورتی‌که در مقابل متغیری مقدار ۲ قرارگرفته باشد یعنی جواب آن آزمایش Yes بوده است. ازاین‌رو به‌منظور استفاده بهینه و افزایش عملکرد تکنیک‌های داده‌کاوی و خوانایی و درک ساده تر نتایج حاصله به‌جای مقادیر ۲ و ۱ به ترتیب مقادیر Y و N که بیانگر Yes و No می‌باشند را قرار می‌دهیم.

۳-۲- انتخاب ویژگی‌ها و آماده‌سازی داده‌ها

اطلاعات موجود شامل آزمایش‌های گوناگونی است. با توجه به کثرت آزمایش‌های صورت گرفته و گستردگی بازه‌های تعریف‌شده در برخی متغیرها ازاین‌رو نیاز است تا متغیرهای موردنظر را از میان سایر متغیرهای انتخاب کنیم تا بتوانیم تکنیک‌های داده‌کاوی را تنها بر روی متغیرهای انتخاب‌شده پیاده‌سازی کنیم.

۳-۲-۱- انتخاب ویژگی‌ها

تعداد آزمایش‌های صورت گرفته بر روی بیماران مبتلابه هپاتیت ۲۰ مورد است. طبیعی است برسی این ۲۰ مورد آزمایش خارج از هدف تعیین‌شده برای این تحقیق است و ما تنها ۴ مورد را به‌عنوان متغیر انتخاب می‌کنیم و تکنیک‌های داده‌کاوی را بر روی آن‌ها پیاده‌سازی می‌کنیم. دلیل انتخاب این چهار متغیر را می‌توان اهمیت و فراوانی آن‌ها و همچنین تأثیرگذاری بیشتر این آزمایش‌ها درروند درمان بیمار در نظر گرفت. ازاین‌رو ویژگی‌های انتخاب‌شده را می‌توان در جدول ۱ مشاهده کرد.

جدول ۱: انتخاب ویژگی‌ها

Properties	Value	
CLASS	DIE	LIVE
ANOREXIA	N	P
LIVER BIG	N	P
LIVER FIRM	N	P
SPLEEN PALPABLE	N	P

۳-۲-۲- جایگذاری مقادیر ناموجود

از بین ویژگی‌های انتخاب‌شده به‌عنوان متغیر، فیلد[6] مربوط به بعضی از آزمایش‌ها فاقد مقادیر می‌باشند ازاین‌رو عدم وجود این مقادیر می‌تواند بر روی نتایج و تحلیل‌های حاصل از پیاده‌سازی تکنیک‌های داده‌کاوی تأثیرگذار باشند. ازاین‌رو باید برای این فیلدهای خالی مقادیری را در نظر گرفت تا بتوان تأثیر ناشی از عدم وجود بعضی از مقادیر را از بین برد. به این منظور از امکانی که داده‌کاوی در قالب نرم‌افزار پیاده ساز آن یعنی Rapid Miner در اختیار ما قرار داده است استفاده می‌کنیم. امکان در نظر گرفته‌شده راه‌های گوناگونی را به‌منظور پر کردن مقادیر خالی فیلدها به ما پیشنهاد می‌دهد ولی در این تحقیق از روش میانگین‌گیری استفاده می‌شود به این صورت که از تمامی مقادیر موجود در ستون مربوط به ویژگی انتخاب‌شده میانگین گرفته می شود و نتیجه حاصل در خانه‌های خالی از مقدار مربوطه قرار می گیرد. در (جدول ۲) می‌توان مقادیر ناموجود در هر یک از ویژگی‌های انتخاب‌شده را مشاهده کرد.

جدول ۲: مقادیر ناموجود ویژگی‌ها انتخاب‌شده

Properties	Null Value
ANOREXIA	1
LIVER BIG	10
LIVER FIRM	11
SPLEEN PALPABLE	5

۳-۳- مدل‌سازی

مدل‌سازی از داده‌ها را می‌توان به‌عنوان بخشی معرفی نمود که بعد از مرحله آماده‌سازی داده اتفاق خواهد افتاد و در این مرحله تکنیک‌های انتخاب‌شده بر روی‌داده‌هایی با ساختار مناسب و استاندارد متناسب پیاده‌سازی خواهند شد. فرایند مدل‌سازی که تکنیک‌های داده‌کاوی آن را انجام می‌دهند به این صورت خواهد بود که در ابتدای امر این تکنیک‌ها نیاز دارند که به یک دانش قبلی دست یابند و با

---

[6] Field

استفاده از این دانش قبلی خود اقدام به مدل‌سازی کنند یعنی به‌عبارت‌دیگر بتوانند فرایند پیش‌بینی را انجام دهند. در این قسمت داده‌کاوی امکانی را فراهم کرده است تا به‌واسطه آن ما بتوانیم برای تکنیک‌های داده‌کاوی مشخص کنیم که از چه مقدار از داده‌های ورودی به‌عنوان داده‌های یادگیری استفاده کند و این یعنی همان دانش قبلی که تکنیک‌های داد کاوی به آن نیاز دارند و در ادامه تکنیک‌های داده‌کاوی بادانشی که از این قسمت به دست آورده‌اند اقدام به انجام پیش‌بینی بر روی مابقی اطلاعات موجود خواهند نمود. این امر واضح و مبرهن است که هر مقداری دانش تکنیک‌های داده‌کاوی در مورد اطلاعات بالاتر باشد امکان پیش‌بینی بهتر و دقیق‌تری مهیا می‌گردد.

۳-۳-۱- انتخاب تکنیک‌های داده‌کاوی

با توجه به گستردگی تکنیک‌های داده‌کاوی موجود از رو نیاز است تا تکنیک‌های موردنظر خود را انتخاب نمایم تا بتوانیم آن‌ها را بر روی‌داده‌های خود پیاده‌سازی کنیم. ازاین‌رو نیاز است تا داده‌های خود را متناسب با تکنیک‌های در نظر گرفته‌شده آماده‌سازی کنیم و بر اساس آن تکنیک‌ها ساختار قالب‌ها را تعریف نمایم. تکنیک‌های استفاده‌شده در این تحقیق از نو تکنیک‌های دسته‌بندی‌شده‌اند به‌عبارت‌دیگر داده‌های موردنیاز برای پیاده‌سازی این دسته از تکنیک‌ها نیاز است که دارای سطر و ستون باشند.

۳-۳-۲- ساخت مدل

بعد از آماده‌سازی داده‌ها و انتخاب تکنیک‌های متناسب حال نیاز است تا فرایند مدل‌سازی صورت پذیرد. از این مرحله می‌توان به‌عنوان آخرین مرحله از داده‌کاوی یادکرد. در این مرحله با پیاده‌سازی تکنیک‌های داده‌کاوی بر روی‌داده‌های موجود می‌توان به مدل‌های ایجادشده توسط تکنیک‌های داده‌کاوی دست‌یافت. ازاین‌رو بعد از نتایج به‌دست‌آمده لازم است تا این نتایج را در قالب معیارهای سنجشی که از قبل در نظر گرفته‌شده‌اند مورد مقایسه قرارداد تا به این شکل بتوان تکنیکی را که عملکرد بهتری را از خود نشان داده است را معرف نمود. معیارهای تعریف‌شده در این تحقیق عبارت‌اند از accuracy، error rate، specificity، negative prediction value که با استفاده از این چهار معیار سنجش عملکرد دو تکنیک داده‌کاوی K همسایه نزدیک و قواعد انجمنی را می‌تواند ارزیابی کرد.

۳-۳-۲-۱- accuracy

accuracy یا به‌عبارت‌دیگر معیار صحت را می‌توان به‌عنوان یکی از مهم‌ترین و پراستفاده‌ترین معیارهای سنجش یک مدل معرفی کرد. معیار صحت از طریق محاسبه مجموع تعداد مواردی که کلاس آن‌ها توسط مدل به‌درستی پیش‌بینی‌شده شده‌اند به تعداد کل مواردی که توسط مدل پیش‌بینی‌شده‌اند به دست می‌آید. فرمول محاسبه معیار صحت از قرار زیر است[21].

$$accuracy=\frac{TP+TN}{TP+FN+FP+TN} \tag{1}$$

۳-۳-۲-۲- error rate

error rate یا معیار نسبت خطا معکوس عملکرد معیار صحت رفتار می‌کند. معیار نسبت خطا از طریق محاسبه مجموع تعداد مواردی که کلاس آن‌ها توسط مدل به‌اشتباه پیش‌بینی‌شده‌اند به تعداد کل مواردی که توسط مدل پیش‌بینی‌شده‌اند به دست می‌آید. به‌عبارت‌دیگر اگر نتیجه حاصل از معیار صحت را از مقدار عددی ۱ کم کنیم مقدار معیار نسبت خطا به دست خواهد آمد. . فرمول محاسبه معیار نسبت خطا از قرار زیر است[22] .

$$error\ rate=\frac{FN+FP}{TP+FN+FP+TN} \tag{2}$$

۳-۳-۲-۳- specificity

specificity یا ویژگی را به‌عنوان یکی از معیار مهم در سنجش عملکرد تکنیک‌های داده‌کاوی مورداستفاده قرار می‌گیرد. معیار ویژگی از طریق محاسبه نسبت تعداد موارد منفی حقیقی تخمین زده‌شده توسط مدل به مجموع موارد مثبت کاذب و منفی حقیقی پیش‌بینی‌شده به دست می‌آید. فرمول محاسبه معیار ویژگی به‌قرار زیر تعریف می‌گردد.[23]

$$specificity=\frac{TN}{FP+TN} \tag{3}$$

۳-۳-۲-۴- negative prediction value

negative prediction value که به‌اختصار NPV نامیده می‌شود معیار سنجشی است که به نسبت زیاد مورداستفاده قرار می‌گیرد. NPV از طریق محاسبه نسبت موارد منفی حقیقی پیش‌بینی‌شده به مجموع موارد منفی کاذب و منفی حقیقی پیش‌بینی‌شده به دست می‌آید. فرمول NPV به‌قرار زیر است.[24]

$$NPV = \frac{TN}{FN + TN} \quad (4)$$

اجزای تشکیل‌دهنده رابطه های ۱تا۴ هریک به‌گونه‌ای بیان‌کننده حالاتی از پیش‌بینی‌های صورت گرفته می‌باشند که از طریق اعمال تکنیک‌ها بر روی نمونه‌ها حاصل‌شده‌اند. این اجزا را می‌توان به‌صورت زیر تعریف کرد.

TP: تعداد نمونه‌هایی که به‌درستی مثبت تشخیص داده می‌شوند.
TN: تعداد نمونه‌هایی که به‌درستی منفی تشخیص داده می‌شوند.
FP: تعداد نمونه‌هایی که به‌اشتباه مثبت تشخیص داده می‌شوند.
FN: تعداد نمونه‌هایی که به‌اشتباه منفی تشخیص داده می‌شوند.

۳-۵- نتیجه نهایی

هدف از پیاده‌سازی این تحقیق بررسی عملکرد دو تکنیک داده‌کاوی K همسایه نزدیک و قواعد انجمنی است که این مقایسه و ارزیابی با استفاده از چهار معیار سنجش accuracy، error rate، specificity، negative prediction value صورت می‌پذیرد؛ و در ادامه تکنیکی که دارای عملکرد مناسب‌تری باشد انتخاب می‌شود و از طرف دیگر با تحلیل نتایج حاصل از پیاده‌سازی تکنیک‌های داده‌کاوی بر روی‌داده‌های بیماران مبتلابه هپاتیت سعی می‌شود تا عواملی که در صورت مثبت شدن در آزمایش فرد مبتلابه هپاتیت، فرد می‌تواند امید زنده ماندن را داشته باشد، مورد بررسی قرار می‌گیرند.

۴ – نتایج پژوهش

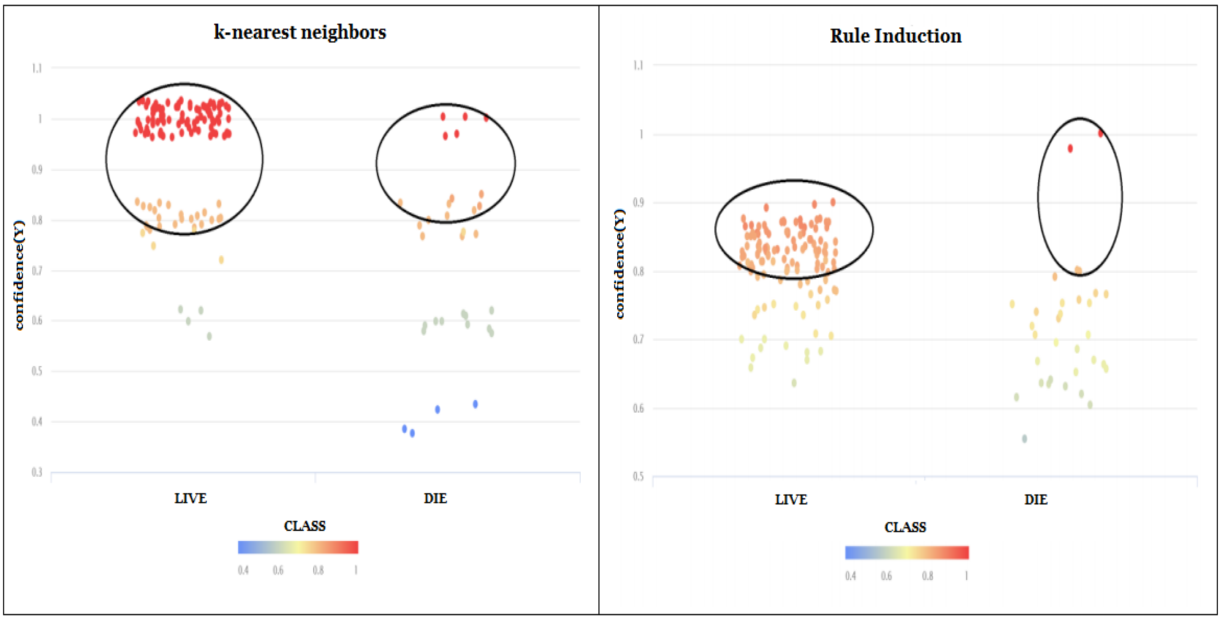
در این بخش در ابتدای امر سعی شده است تا نتایج حاصل از پیاده‌سازی دو تکنیک K همسایه نزدیک و قواعد انجمنی بر روی چهار ویژگی ANOREXIA، LIVER BIG، LIVER FIRM و SPLEEN PALPABLE مورد بررسی قرار گیرد. در این بخش دو قسمت جواب وجود خواهد داشت قسمت اول شامل نتایج عملکرد تکنیک‌های داده‌کاوی پیاده‌سازی شده K همسایه نزدیک و قواعد انجمنی خواهد بود که بر اساس چهار معیار accuracy، error rate، specificity، negative prediction value مورد ارزیابی قرار می‌گیرند و در این قسمت تکنیکی که عملکرد مطلوب‌تری را که از خود نشان داده‌شده باشد معرفی می‌گردد و در قسمت دوم مربوط به نتیجه سعی شده است تا تأثیر مثبت بودن هر یک از چهار ویژگی تعریف‌شده ANOREXIA، LIVER BIG، LIVER FIRM و SPLEEN PALPABLE را در زنده ماند فرد مبتلابه هپاتیت مورد بررسی قرار گیرد. در این قسمت به‌منظور تأثیرگذاری یک ویژگی شروطی لحاظ شده است تا درصورتی‌که یک ویژگی بتواند آن شروط را احراز کند به‌عنوان یک ویژگی در نظر گرفته می‌شود که فرد مبتلابه هپاتیت درصورتی‌که از آن برخوردار باشد می‌تواند امید زنده ماند داشته باشد به‌عبارت‌دیگر درصورتی‌که نتیجه مربوط به آن آزمایش مثبت باشد فرد مبتلا همچنان می‌تواند امید به حیات داشته باشد. ویژگی‌های تعریف‌شده برای این تحقیق مقدار فاصله اطمینان ۸۰ درصد و مقدار پشتیبان ۵۰ درصد است. علت انتخاب مقدار ۸۰ درصد برای فاصله اطمینان را می‌توان به دلیل حساسیت بالای حوزه پزشکی در نظر گرفت و با توجه به مطرح‌شدن موضوع با سلامت انسان نیاز است تا مقدار فاصله اطمینان بالا در نظر گرفته شود تا نتایج حاصل‌شده از دقت و اطمینان کافی برخوردار باشند. مقدار پشتیبان ۵۰ درصد به این معنی است که درصورتی‌که از ۱۵۵ نمونه موجود ویژگی در حداقل ۷۸ مورد مثبت تشخیص داده‌شده است و در این حالت فرد نیز زنده مانده باشد در این صورت است که این ویژگی به‌عنوان مشخصه‌ای در نظر گرفته می‌شود که فرد درصورتی‌که در آزمایش‌های صورت گرفته مقداری مثبت برای این ویژگی به‌دست‌آمده باشد فرد نیز می‌تواند همچنان امید به حیات داشته باشد.

جدول ۳: وضعیت متغیرهای انتخاب‌شده در ۱۵۵ نمونه موجود.

پارامتر بالینی	۱۵۵ نمونه موجود			LIVE		DIE				K همسایه نزدیک		قواعد انجمنی			
	مقادیر ناموجود	منفی	مثبت	منفی	مثبت	منفی	مثبت	ACC	SPEC	ERR	NPV	ACC	SPEC	ERR	NPV
		تعداد	تعداد	تعداد	تعداد	تعداد	تعداد	(٪)	(٪)	(٪)	(٪)	(٪)	(٪)	(٪)	(٪)
ANOREXIA	۱	32	۲12	2۲	۱۰۳	۱۰	۱۹	۷۹,۳	100	۲۰,۶۷	۷۹,۳	۷۶,۷	۹۶,۸	۲۳,۲۱	۷۸,۸
LIVER BIG	۱۰	25	۱۲۰	۲۲	۹۸	۳	۲۲	۸۰,۴	۹۰,۳۳	۱۹,۹۶	۱۲,۵	۸۱,۹	۹۳,۸	۱۸,۰۸	۲۰
LIVER FIRM	۱۱	۵۸	۸۶	۴۷	۷۲	۱۱	۱۴	۷۳,۴	۹۰,۴۴	۲۶,۵۸	۷۳,۵	۷۲,۲	۹۰,۴۴	۲۷,۷۹	۷۲,۴
SPLEEN PALPABLE	۵	۲۹	۱۲۱	۱۰	۱۰۲	۱۲	۱۹	۷۸,۸	۱۰۰	۱۹,۳۷	۸۰,۱	۷۷,۸	۱۰۰	۱۹,۳۷	۸۰,۶

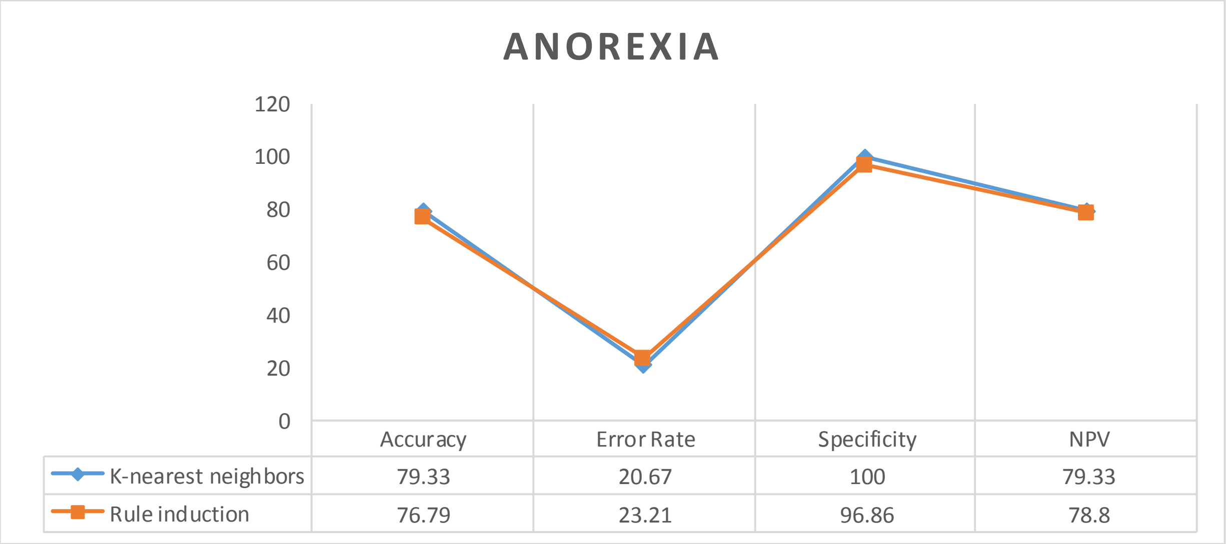
۴-۱- ویژگی ANOREXIA

با استناد به نتایج به‌دست‌آمده درصورتی‌که فرد مبتلابه بیماری هپاتیت درنتیجه آزمایش‌ها خود با ANOREXIA مثبت مواجه شود پیش‌بینی می‌شود فرد بتواند زنده بماند (شکل ۲).



شکل ۲: نمودار عملکرد تکنیک‌های داده‌کاوی در پیش‌بینی زنده ماندن بیمار در صورت مثبت شدن ANOREXIA

با استناد به نتایج به‌دست‌آمده از پیاده سای تکنیک‌های داده‌کاوی و بر اساس ارزیابی‌های صورت گرفته با استفاده از چهار معیار سنجش accuracy، error rate، specificity و negative prediction value می‌توان بیان داشت که تکنیک داده‌کاوی K همسایه نزدیک نسبت به تکنیک قواعد انجمنی در پیش‌بینی زنده ماندن فرد مبتلابه هپاتیت، درصورتی‌که فرد نتیجه ANOREXIA مثبت را در آزمایش خود داشته باشد، دارای عملکرد بهتری است (شکل ۳).

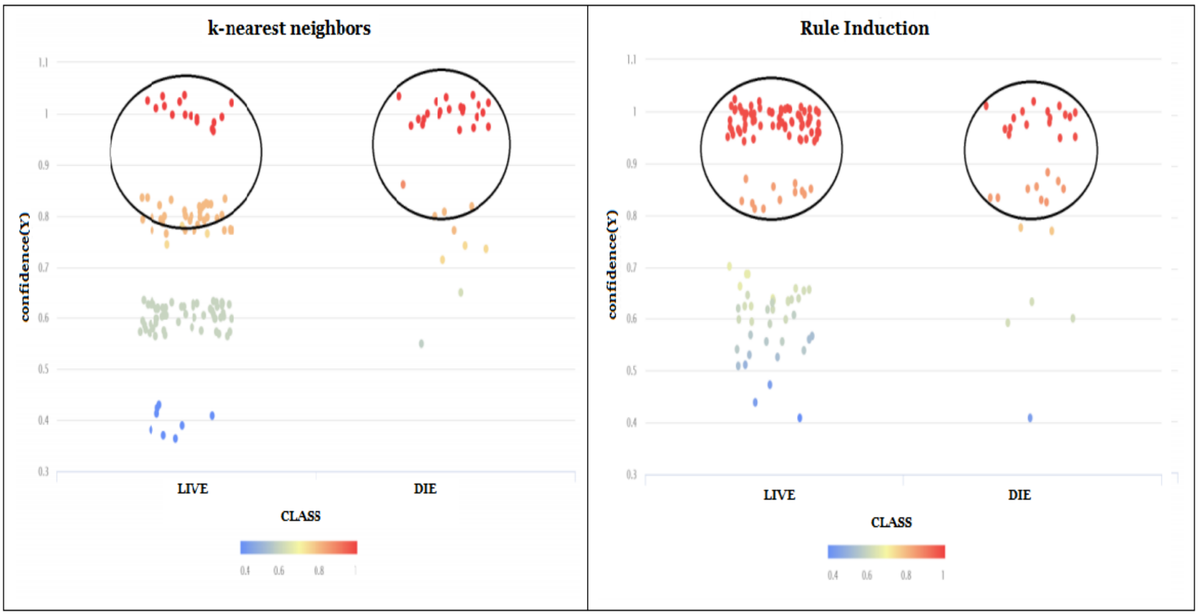


	Accuracy	Error Rate	Specificity	NPV
K-nearest neighbors	79.33	20.67	100	79.33
Rule induction	76.79	23.21	96.86	78.8

شکل ۳: ارزیابی عملکرد تکنیک‌های داده‌کاوی در پیش‌بینی زنده ماندن بیمار در صورت مثبت شدن ANOREXIA

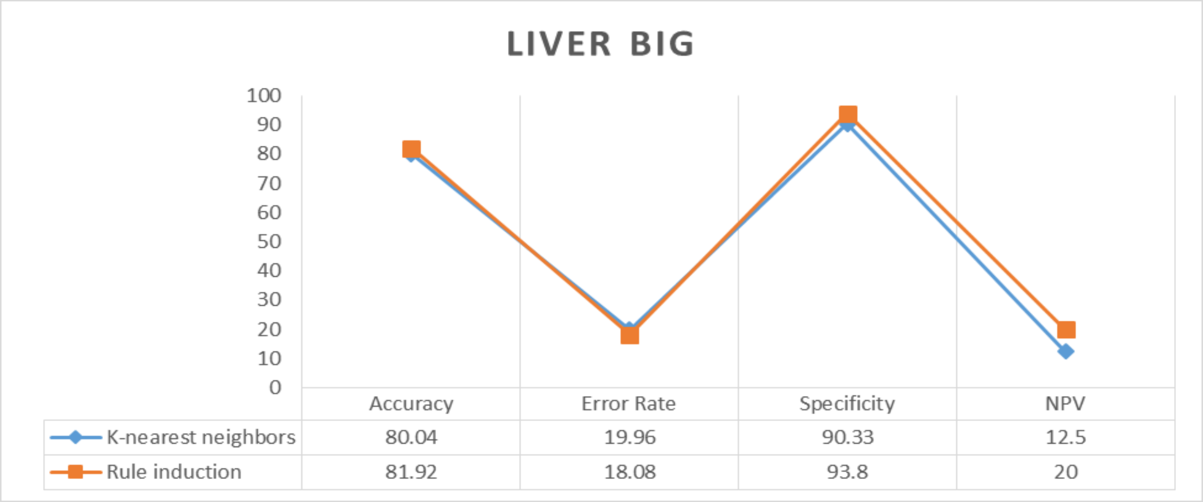
۴-۲- ویژگی LIVER BIG

با استناد به نتایج به‌دست‌آمده درصورتی‌که فرد مبتلابه بیماری هپاتیت درنتیجه آزمایش‌ها خود با LIVER BIG مثبت مواجه شود پیش‌بینی می‌شود فرد بتواند زنده بماند (شکل ۴).



شکل ۴: نمودار عملکرد تکنیک‌های داده‌کاوی در پیش‌بینی زنده ماندن بیمار در صورت مثبت شدن LIVER BIG

با استناد به نتایج به‌دست‌آمده از پیاده سای تکنیک‌های داده‌کاوی و بر اساس ارزیابی‌های صورت گرفته با استفاده از چهار معیار سنجش accuracy، error rate، specificity و negative prediction value می‌توان بیان داشت که تکنیک داده‌کاوی قواعد انجمنی نسبت به تکنیک K همسایه نزدیک در پیش‌بینی زنده ماندن فرد مبتلابه هپاتیت، درصورتی‌که فرد نتیجه LIVER BIG مثبت را در آزمایش خود داشته باشد، دارای عملکرد بهتری است (شکل ۵).

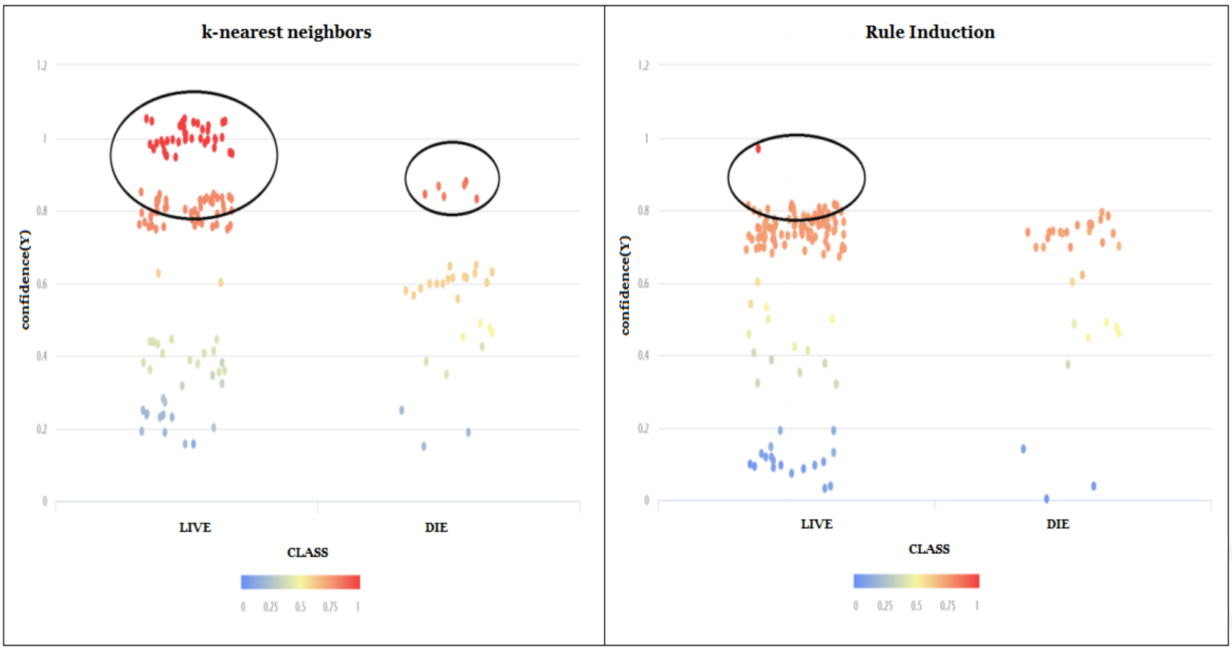


	Accuracy	Error Rate	Specificity	NPV
K-nearest neighbors	80.04	19.96	90.33	12.5
Rule induction	81.92	18.08	93.8	20

شکل ۵: ارزیابی عملکرد تکنیک‌های داده‌کاوی در پیش‌بینی زنده ماندن بیمار در صورت مثبت شدن LIVER BIG

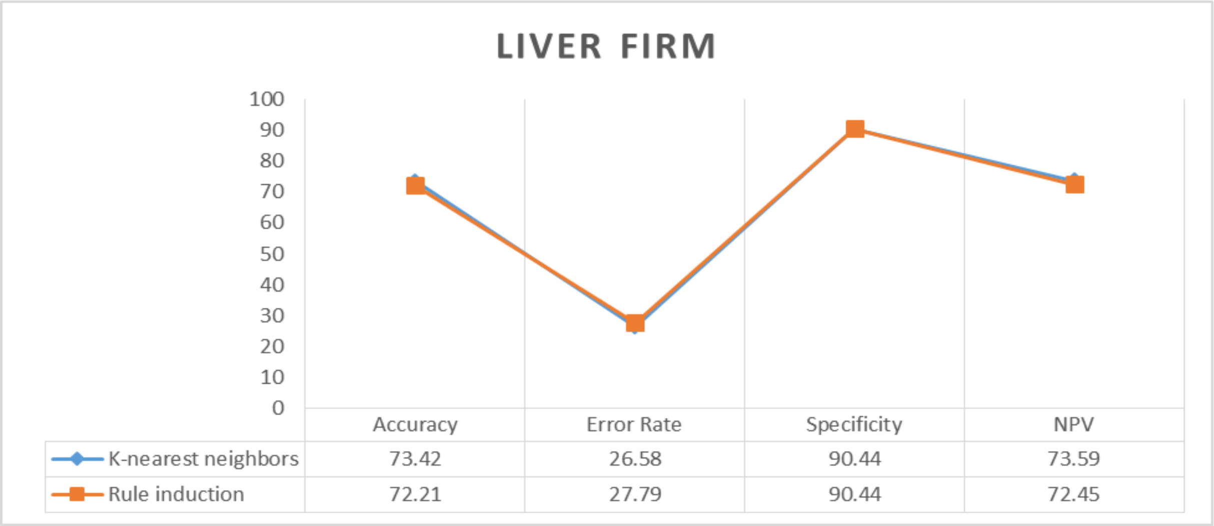
۴-۳- ویژگی LIVER FIRM

با استناد به نتایج به‌دست‌آمده درصورتی‌که فرد مبتلابه بیماری هپاتیت درنتیجه آزمایش‌ها خود با LIVER FIRM مثبت مواجه شود پیش‌بینی می‌شود فرد بتواند زنده بماند (شکل ۶).



شکل ۶: نمودار عملکرد تکنیک‌های داده‌کاوی در پیش‌بینی زنده ماندن بیمار در صورت مثبت شدن LIVER FIRM

با استناد به نتایج به‌دست‌آمده از پیاده سای تکنیک‌های داده‌کاوی و بر اساس ارزیابی‌های صورت گرفته با استفاده از چهار معیار سنجش accuracy، error rate، specificity و negative prediction value می‌توان بیان داشت که تکنیک داده‌کاوی K همسایه نزدیک نسبت به تکنیک قواعد انجمنی در پیش‌بینی زنده ماندن فرد مبتلابه هپاتیت، درصورتی‌که فرد نتیجه LIVER FIRM مثبت را در آزمایش خود داشته باشد، دارای عملکرد بهتری است (شکل ۷).

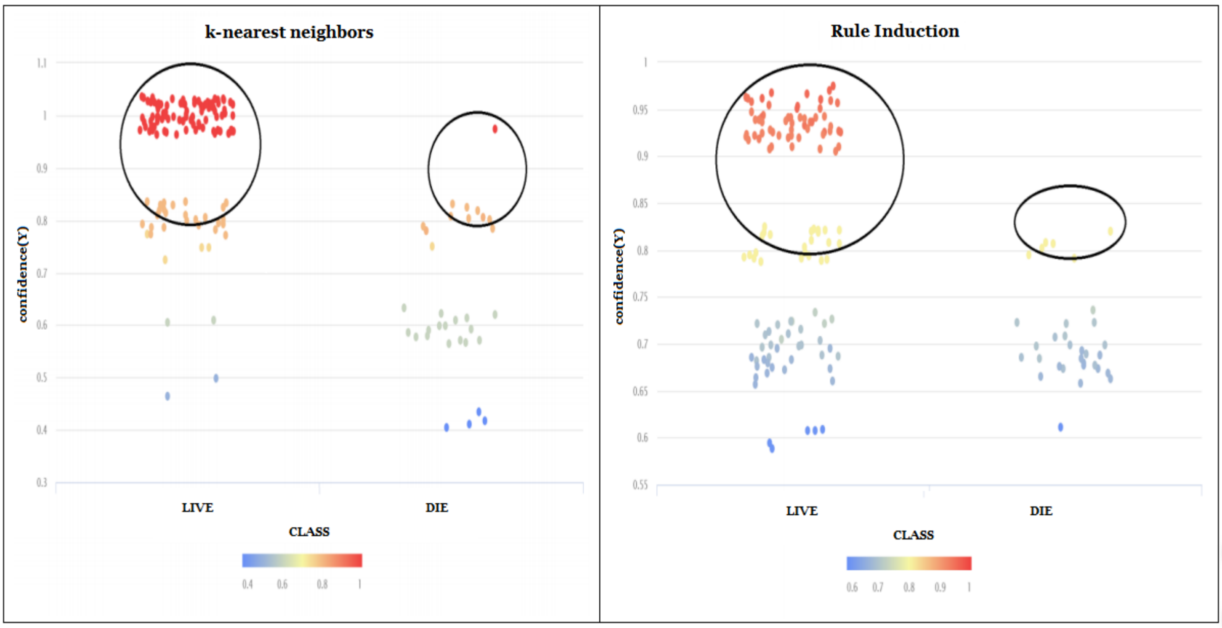


	Accuracy	Error Rate	Specificity	NPV
K-nearest neighbors	73.42	26.58	90.44	73.59
Rule induction	72.21	27.79	90.44	72.45

شکل ۷: ارزیابی عملکرد تکنیک‌های داده‌کاوی در پیش‌بینی زنده ماندن بیمار در صورت مثبت شدن LIVER FIRM

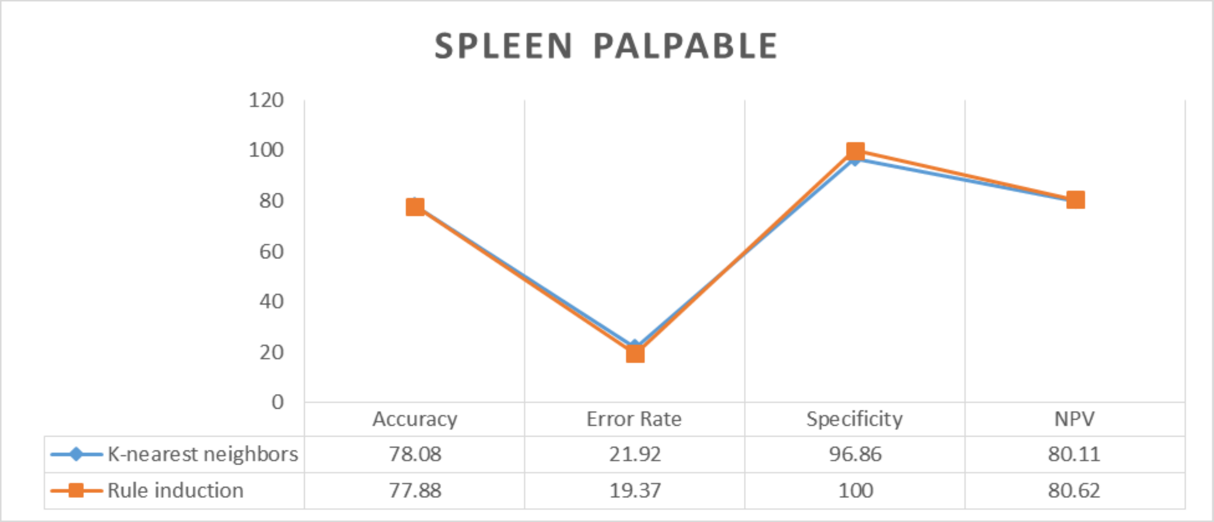
۴-۴- ویژگی SPLEEN PALPABLE

با استناد به نتایج به‌دست‌آمده درصورتی‌که فرد مبتلابه بیماری هپاتیت درنتیجه آزمایش‌ها خود با SPLEEN PALAPABLE مثبت مواجه شود پیش‌بینی می‌شود فرد بتواند زنده بماند (شکل ۸).



شکل ۸: نمودار عملکرد تکنیک‌های داده‌کاوی در پیش‌بینی زنده ماندن بیمار در صورت مثبت شدن SPLEEN PALAPABLE

با استناد به نتایج به‌دست‌آمده از پیاده سای تکنیک‌های داده‌کاوی و بر اساس ارزیابی‌های صورت گرفته با استفاده از چهار معیار سنجش accuracy، error rate، specificity و negative prediction value می‌توان بیان داشت که تکنیک داده‌کاوی قواعد انجمنی نسبت به تکنیک K همسایه نزدیک در پیش‌بینی زنده ماندن فرد مبتلابه هپاتیت، درصورتی‌که فرد نتیجه LIVER BIG مثبت را در آزمایش خود داشته باشد، دارای عملکرد بهتری است (شکل ۹).



	Accuracy	Error Rate	Specificity	NPV
K-nearest neighbors	78.08	21.92	96.86	80.11
Rule induction	77.88	19.37	100	80.62

شکل 9: ارزیابی عملکرد تکنیک‌های داده‌کاوی در پیش‌بینی زنده ماندن بیمار در صورت مثبت شدن SPLEEN PALAPABLE

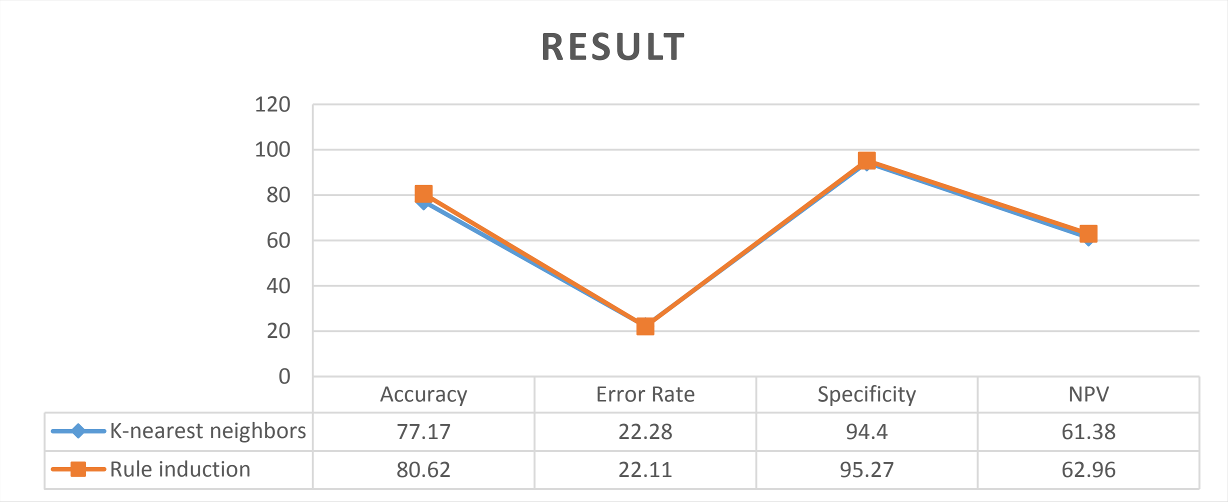
۵- نتیجه‌گیری نهایی

با جمع‌آوری نتایج حاصل از پیاده‌سازی دو تکنیک داده‌کاوی قواعد انجمنی و K همسایه نزدیک بر روی چهار ویژگی ANOREXIA، LIVER BIG، LIVER FIRM و SPLEEN PALPABLE که در (جدول ۳) به نمایش درآمده‌اند، اقدام به میانگین گرفتن از این نتایج می‌شود. با میانگین گرفتن از نتایج حاصل می‌توان به عملکرد کلی تکنیک‌ها در طول فرآیندهای انجام‌شده دست‌یافت. نتایج حاصل از میانگین‌گیری در (جدول ۴) نشان داده‌شده‌اند.

جدول ۴: مقایسه نهایی عملکرد دو تکنیک داده‌کاوی

تکنیک‌های داده‌کاوی	accuracy	rate error	specificity	NPV
همسایه نزدیک K	77.17	22.28	94.4	61.38
قواعد انجمنی	80.62	22.11	95.27	62.96

نتایج به‌دست‌آمده نشان می‌دهد قواعد انجمنی درمجموع ارزیابی‌های صورت گرفته دارای عملکرد بهتری در مقایسه با K همسایه نزدیک است (شکل ۱۰).



شکل ۱۰: ارزیابی نهایی از عملکرد دو تکنیک داده‌کاوی

درمجموع پیاده‌سازی‌های صورت گرفته با استفاده از تکنیک‌های داده‌کاوی K همسایه نزدیک و قواعد انجمنی بر روی‌داده‌های مرتبط با بیماران مبتلابه هپاتیت و بررسی عملکرد این دو تکنیک بر اساس چهار معیار ارزیابی accuracy، error rate، specificity و negative prediction value می‌توان نتیجه گرفت که تکنیک قواعد انجمنی با مقدار accuracy 80/62، error rate 22/11، specificity 95/27 و negative prediction value 62/96 عملکرد بهتری را نسبت به تکنیک K همسایه نزدیک از خود نشان داده است؛ و در مورد نتایج حاصل از بررسی های صورت گرفته بر روی ویژگی‌های موجود در آزمایش افراد مبتلابه هپاتیت می‌توان نتیجه گرفت که افرادی که در نتایج مرتبط با آزمایش خود با ویژگی‌های ANOREXIA، LIVER BIG، LIVER FIRM و SPLEEN PALPABLE مثبت مواجه شده است می‌توانند از شانس زنده ماندن برخوردار باشند.

سپاسگزاری

بدینوسیله نویسنده از داوران ارجمند برای نظرات ارزنده و پیشنهاداتی در راستای بهبود ارائه این مقاله، تشکر میکند.

منابع و مراجع

[1] Ng RT, Pei J. Introduction to the special issue on data mining for health informatics. ACM SIGKDD Explorations Newsletter. 2007; 9(1):1-2.

[2] Soni J, Ansari U, Sharma D, Soni S. Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications. 2011; 17(8):43-8.

[3] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowledge and information systems. 2008; 14(1):1-37.

[4] Padhy N, Mishra D, Panigrahi R. The survey of data mining applications and feature scope. ArXiv preprint arXiv: 12115723. 2012.

[5] Wu X, Zhu X, Wu G-Q, Ding W. Data mining with big data. IEEE transactions on knowledge and data engineering. 2014; 26(1):97-107.

[6] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine. 2005; 34(2):113-27.

[7] Glover S, Rivers PA, Asoh DA, Piper CN, Murph K. Data mining for health executive decision support: an imperative with a daunting future! Health services management research. 2010; 23(1):42-6.

[8] Gharehchopogh FS, Molany M, Mokri FD. Using artificial neural network in diagnosis of thyroid disease: a case study. International Journal on Computational Sciences & Applications (IJCSA) Vol. 2013; 3:49-61.

[9] Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. Advances in knowledge discovery and data mining. 1996.

[10] Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonça A. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC research notes. 2011; 4(1):299.

[11] Sokolova M, Japkowicz N, Szpakowicz S, editors. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. Australasian joint conference on artificial intelligence; 2006: Springer.

[12] Rajeswari P, Reena GS. Analysis of liver disorder using data mining algorithm. Global journal of computer science and technology. 2010.

[13] Ho T-B, Nguyen C-H, Kawasaki S, Le S-Q, Takabayashi K. Exploiting temporal relations in mining hepatitis data. New Generation Computing. 2007;25(3):247.

[14] Uhmn S, Kim D-H, Cho SW, Cheong JY, Kim J, editors. Chronic hepatitis classification using SNP data and data mining techniques. 2007 Frontiers in the Convergence of Bioscience and Information Technologies; 2007: IEEE.

[15] Zayed N, Awad AB, El-Akel W, Doss W, Awad T, Radwan A, et al. The assessment of data mining for the prediction of therapeutic outcome in 3719 Egyptian patients with chronic hepatitis C. Clinics and research in hepatology and gastroenterology. 2013;37(3):254-61.

[16] Abe H, Ohsaki M, Yokoi H, Yamaguchi T, editors. Implementing an integrated time-series data mining environment based on temporal pattern extraction methods: a case study of an interferon therapy risk mining for chronic hepatitis. Annual Conference of the Japanese Society for Artificial Intelligence; 2005: Springer.

[17] Ohsaki M, Sato Y, Yokoi H, Yamaguchi T, editors. A rule discovery support system for sequential medical data, in the case study of a chronic hepatitis dataset. Workshop Notes of the International Workshop on Active Mining, at IEEE International Conference on Data Mining; 2002.

[18] Yin X, Han J, editors. CPAR: Classification based on predictive association rules. Proceedings of the 2003 SIAM International Conference on Data Mining; 2003: SIAM.

[19] Guo G, Wang H, Bell D, Bi Y, Greer K, editors. KNN model-based approach in classification. OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"; 2003: Springer.
[20] Lauer GM, Walker BD. Hepatitis C virus infection. New England journal of medicine. 2001;345(1):41-52.
[21] Lucas P. Bayesian analysis, pattern analysis, and data mining in health care. Current opinion in critical care. 2004; 10(5):399-403.
[22] Witt O, Deubzer HE, Milde T, Oehme I. HDAC family: What are the cancer relevant targets? Cancer letters. 2009; 277(1):8-21.
[23] Cios KJ, Moore GW. Uniqueness of medical data mining. Artificial intelligence in medicine. 2002;26(1-2):1-24.
[24] Steinberg DM, Fine J, Chappell R. Sample size for positive and negative predictive value in diagnostic research using case–control designs. Biostatistics. 2008;10(1):94-105.