# ارزیابی کارایی الگوریتم های مبتنی بر قانون و درخت تصمیم در تشخیص بیماری قلبی

**مریم کالیان برازجانی**
**کارشناسی ارشد کامپیوتر نرم افزار**
maryamkamalian@ymail.com

## چکیده

امروزه در دانش پزشکی شاهد جمع آوری داده های فراوان در مورد بیماری های مختلف هستیم . تحقیق روی این داده ها و بدست آوردن نتایج و الگو های مفید در رابطه با بیماری ها یکی از اهداف استفاده از این داده ها است. بیماری قلبی با توجه به شیوع و سهمی که در مرگ و میر انسانها دارد از اهمیت بالایی برخوردار است. در این تحقیق ارزیابی دو الگوریتم درخت تصمیم و مبتنی بر قانون که برای پیش بینی بیمار قلبی یا عدم بیماری قلبی به کار رفته اند انجام شده است، نتایج ارزیابی نشان میدهد که در این تحقیق الگوریتم مبتنی بر قانون از نظر دقت ومتریک از درخت تصمیم برتر می باشد.

کلمات کلیدی: درخت تصمیم،مبتنی بر قانون،بیماری قلبی،دقت

## ۱.مقدمه

قلب یک پمپ عضلانی توخالی است که بدون توقف خون را به سراسر بدن پمپ می کند. اگرچه قلب بزرگتر از مشت دست نیست ، ولی در طول دوران عمر در حدود ۳۰۰ میلیون لیتر خون را پمپ می کند. رگ های بزرگ خون که به قلب متصل اند خون را به ریه ها وسراسر بدن می برند و باز می گردانند. بیماری قلبی مرگ سلول‌های عضلانی قلب در اثر کاهش یا توقف جریان خون سرخرگ‌های قلب می باشد . بیماری قلبی امروزه مهم ترین عامل مرگ و میر انسان ها می باشد. در ابتدای قرن بیستم ۱۰ درصدکل مرگ و میرها به علت بیماری های قلبی بود[2]. در انتهای همین قرن موارد مرگ و میر ناشی از بیماری های قلبی به ۲۵ درصد افزایش یافت و پیش بینی می شود با توجه به روند کنونی تا سال ۲۰۲۵میلادی بیشتر از ۳۵ تا ۶۰ درصد موارد مرگ و میر در جهان از

بیماری های قلبی ناشی شود[2]. رشد چشم گیر بیماریهای قلبی و اثرات و عوارض آن ها و هزینه های بالایی که بر جامعه وارد می کند، باعث شده که جامعه پزشکی به دنبال برنامه هایی جهت بررسی بیشتر، پیشگیری و شناسایی زود هنگام و درمان موثر آن باشد .از این رو با استفاده از داده کاوی و کشف دانش در سیستم مراکز قلب می توان دانش ارزشمند را ایجاد کرد که این دانش کشف شده می تواند باعث بهبود کیفیت سرویس به وسیله مدیران مرکز شود تا رفتار آینده بیماران قلبی را از روی سابقه داده شده پیش بینی کنند. همچنین در این کاربرد به دنبال این نیستیم که تعیین کنیم چه کسانی دارای بیماری قلبی هستند بلکه به دنبال این هستیم که چه عواملی در بروز این بیماری نقش بیشتر داشته است . در داده کاوی پزشکی پژوهش های با رویکرد پیش بینی انجام شده است[1].

تفاوت داده کاوی با روشهای آماری در این است که در علم آمار ما به دنبال اثبات فرضیه مورد نظر هستیم اما در داده کاوی بر خلاف علم آمار به دنبال پیشگویی هستیم نه کشف یا اثبات. بدین معنا که با استفاده از روش های داده کاوی به دنبال تایید آنچه از قبل وجود دارد نیستند بلکه به دنبال مشخص کردن الگوهای از قبل شناخته نشده هستند[2]. هدف این تحقیق بررسی عوامل بیماری زا و اینکه چه عواملی باعث بروز بیماری قلبی یا عدم بیماری قلبی می شود هست که الگوریتم های درخت تصمیم و مبتنی بر قانون برای بررسی مورد استفاده قرار گرفته اند.

## ۲–تحقیقات پیشین

پیش بینی بیماری قلبی با استفاده از داده کاوی در تحقیقات زیادی مورد بررسی قرار گرفته است. در یکی از تحقیقات سه مدل با استفاده از الگوریتم های درخت تصمیم و شبکه عصبی و نایو بیس مورد بررسی قرار گرفته که در ارزیابی مدلها مدل نایو بیس از دو مدل دیگر برتر بود[1]. در این تحقیق با بهبود دادن در مجموعه داده ها و طبقه بندی کردن داده ها و با استفاده از دو الگوریتم درخت تصمیم و مبتنی بر قانون نتایج را بهبود دادیم[7].

## ۳–درخت تصمیم c4.5

درخت تصمیم یکی از مشهورترین و قدیمی ترین روشهای ساخت مدل دسته بندی است. در الگوریتم های دسته بندی مبتنی بر درخت تصمیم دانش خروجی به صورت یک درخت از حالات مختلف مقادیر ویزگی ها ارائه می شود.درخت های تصمیم بر اساس قواعد تصمیم گیری برای پیش بینی و دسته بندی مورد استفاده قرار میگیرند. در مواردی که می خواهیم نتیجه دسته بندی را در قالب دسته هایی مانند وام پر ریسک در مقابل وام های کم ریسک و یا خرید یا عدم خرید بیان کنیم استفاده از درخت تصمیم بسیار کارآمد خواهد بود[7].یکی از مزایای درخت تصمیم ایجاد امکانی برای شناخت بهتر فیلدهای با اهمیت است. زیرا در درخت تصمیم به طور خودکار فیلدهای با اهمیت بیشتر به گره های بالایی درخت انتقال می یابند و همچنین درخت تصمیم فیلدهای کم اهمیت را کنار می گذارد[7].مدل درخت تصمیم به صورت مجموعه ای از قواعد اگر و آنگاه ظاهر می شود که در بسیاری از موارد اطلاعات را در شکل بسیار کامل نمایش می دهد[7].

### ۳.۱–الگوریتم c4.5

این الگوریتم با قابلیت یادگیری از انواع داده های ورودی چه اسمی و چه عددی یک درخت تصمیم را به عنوان مدل خروجی می سازد. در روش c4.5 باید پارامتر information gain را جهت اعمال روش انتروپی انتخاب کنیم[9].

## ۴–الگوریتم Rule base (RIPPER)

این روش استخراج مجموعه ای از قوانین به صورت شرط است که می توانیم این شروط را به طور مستقیم از مجموعه داده های آموزشی استخراج کنیم[11].یک قانون به صورت کلی زیر نمایش داده می شود:

If (a1 op v1) and (a2 op v2) and … then class=ci

که در آن ai نشان دهنده نام یک صفت خاصه. Vi یک مقدار مشخص. متغیر class به صفت خاصه کلاس و ci به یکی از مقادیر کلاس ها اشاره

می کنند. متغیر op از میان عملگرهای مقایسه ای انتخاب می شود[11].
در این بخش سمت چپ هر قانون را با واژه شرط یا شروط می شناسیم و هر شرط کامل( همراه با تخمین کلاس ) را به عنوان یک قانون بیان میکنیم. همانطور که در شکل کلی هر قانون مشاهده می کنید قسمت ابتدایی و سمت چپ هر قانون مجموعه ای از تست هایی است که بر روی صفات خاصه انجام شده است. میان این شروط از عملگر منطقی and استفاده می شود.قسمت نتیجه گیری قانون ( سمت راست ) همیشه تعیین یک کلاس از مجموعه برچسب های موجود در داده های اصلی است[11].چنانجه مجموعه شروط سمت چپ یک قانون برای نمونه ای از داده ها صادق باشد در واقع قانون مزبور آن نمونه را پوشش می دهد. بنابراین می توان گفت یکی از اهداف ما یافتن قوانینی است که تعداد نمونه های بیشتری از مجموعه داده های آموزشی را پوشش دهد. بدین ترتیب امیدوار خواهیم بود که مدل با تعداد قوانین کمتری قابل توصیف خواهد بود[11].
حال فرض کنید برای یک نمونه آزمایشی بیش از یک قانون وجود دارد که مقادیر صفات خاصه نمونه آزمایشی شروط این قوانین را ارضا می کند. استراتژی های متفاوتی وجود دارد که می توان با کمک آنها این مشکل را حل کرد. ولی به طور کلی می توان گفت در همه این راه حل ها یک نوع اولویت گذاری میان قوانین مطرح است. برای مثال در برخی از روش ها ترتیب میان قوانین از اهمیت ویژه ای برخوردار است. بدین ترتیب اولین قانونی که نمونه آزمایشی را پوشش میدهد به عنوان قانون اصلی جهت تخمین برچسب کلاس انتخاب می شود. ترتیب اولویت ها می تواند بر اساس اولویت کلاس ها تعیین شود. راه حل دیگر استفاده از قانونی با اندازه بزرگتر است اندازه یک قانون بر اساس تعداد شروط مشخص خواهد شد.

## ۵–داده های مورد نظر

مجموعه داده به کار گرفته شده در این تحقیق از سایت archive.ics.uci.edu استخراج شده است که مربوط به ۳ شهر مختلف

بوده[10]. سلولند از ایالات متحده آمریکا. هانگاریان از کشور مجارستان و سوییتزرلند از کشور سویس. ویزگی های مربوطه در ابتدا ۷۵ ویزگی برای هر رکورد بوده که به ۱۴ ویزگی که از همه مهمتر بوده خلاصه شده است. ۱۳ ویزگی از ویزگی های ورودی و ۱ ویزگی نشان دهنده وجود بیماری قلبی و یا عدم بیماری قلبی لست که با صفر و یک نشان داده می شود[10].

## ۶–نمونه گیری و اجرای الگوریتم

مجموعه داده از ۲۷۰ رکورد تشکیل شده که ۷۰ درصد را برای قسمت آموزش و ۳۰ درصد را برای قسمت آزمایش استفاده کرده ایم.هر دو الگوریتم را ۱۰ بار و با سیدهای مختلف (تغییر اعداد تصادفی ) انجام دادیم که هر بار دقت های متفاوتی ایجاد شده است[12].

## ۷–متریک های ارزیابی

- دقت مدل ( که دقت بر اساس پیش بینی های درست و نادرست به دست می آید)
- tp (وجود بیماری قلبی را درست پیش بینی کرده است)
- tn(عدم بیماری قلبی را درست پیش بینی کرده است)
- که در این دو متریک تعداد بالا بهتر است.
- fp(وجود بیماری قلبی را اشتباه پیش بینی کرده است)
- fn (عدم بیماری قلبی را اشتباه پیش بینی کرده است)

## ۸–ارزیابی

داده های مورد نظر را با ۷۰ درصد برای آموزش و ۳۰ درصد برای آزمایش وارد مدل ها کردیم و هر کدام از مدل ها را ۱۰ بار با سیدهای مختلف اجرا کرده و دقت های بدست آمده را وارد جدول کردیم تا با استفاده از فرمول ارزیابی بررسی کنیم که آیا مدلی برتر از دیگری هست یا خیر.

**جدول ۱:دقت های بدست آمده از الگوریتم ها**

| | درخت تصمیم | مبتنی بر قانون | اختلاف |
|---|---|---|---|
| **۱** | ۷۲.۷۸ | ۷۹.۰۳ | ۶.۲۵ |
| **۲** | ۶۴.۰۳ | ۶۶.۸۱ | ۲.۷۸ |
| **۳** | ۶۶.۳۹ | ۷۲.۷۸ | ۶.۳۹ |
| **٤** | ۶۵.۸۳ | ۷۰.۶۹ | ۴.۸۶ |
| **٥** | ۶۳.۰۶ | ۷۲.۹۲ | ۹.۸۶ |
| **٦** | ۶۶.۶۷ | ۷۵.۲۸ | ۸.۶۱ |
| **۷** | ۷۲.۷۸ | ۷۴.۱۷ | ۱.۳۹ |
| **۸** | ۷۱.۸۱ | ۷۳.۳۳ | ۱.۵۲ |
| **۹** | ۵۹.۳۱ | ۶۴.۰۳ | ۴.۷۲ |
| **۱۰** | ۷۱.۵۳ | ۷۶.۵۳ | ۵ |
| **میانگین** | ۶۷.۴۱۹ | ۷۵.۵۵۷ | ۵.۱۳۸ |
| **انحراف معیار** | ۴.۶۴ | ۴.۴۴۳ | ۲.۷۹۱ |

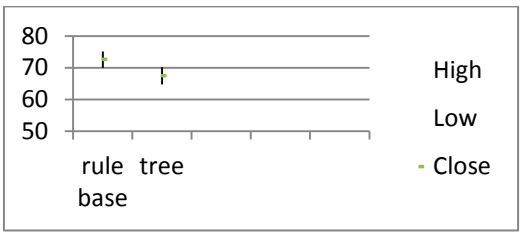
تعداد تکرار n=10 که با تغییر ورودی تولید کننده اعداد تصادفی انجام شده است.از فرمول

(۱)

$$mean+-t(s)/\sqrt{n}$$

استفاده میکنیم. برای بدست آوردن بازه دقت هر کدام از الگوریتم ها در اینجا مقدار t را برای ۹۰ درصد دقت ۱.۸۳۳ قرار می دهیم.

بازه الگوریتم درخت تصمیم (۶۴.۷۳ تا ۷۰.۱۰۸) و بازه الگوریتم مبتنی بر قانون (۶۹.۹۸۲ تا ۷۵.۱۳۲ ) می باشد




**شکل ۱: بازه های دقت الکوریتم ها**

مشاهده می شود که حدود ۰.۱ یکی از الگوریتم ها در دیگری است به همین خاطر از آنالیز زوج [۱]هم استفاده میکنیم و این بار در فرمول مقادیر میانگین و انحراف معیار اختلاف ها را جایگذاری می کنیم.
بازه بدست آمده ۳.۴۶۷ تا ۶.۷۵۵ است به دلیل اینکه صفر در این بازه نیست می توان نتیجه گرفت که الگوریتم مبتنی بر قانون از درخت تصمیم در این تحقیق بهتر عمل کرده است.

## ۸.۱ –بررسی دیگر متریک ها

**جدول ۳: بررسی تشخیص های درست و غلط**

| | درخت تصمیم | مبتنی بر قانون |
|---|---|---|
| **تعداد پیش بینی های** tp[۲] | ۱۷ | ۲۰ |
| **تعداد پیش بینی های** fp[۳] | ۲۲ | ۱۸ |
| **تعداد پیش بینی های** tn[۴] | ۲۴ | ۳۲ |
| **تعداد پیش بینی های** fn[۵] | ۱۸ | ۱۱ |

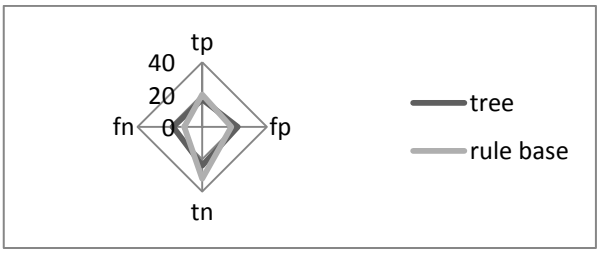
## ۹– نتیجه گیری

در این تحقیق از الگوریتم های درخت تصمیم و مبتنی بر قانون برای بررسی وجود بیماری و عدم بیماری قلبی استفاده کردیم هر دو الگوریتم را با هم مقایسه کردیم و مشاهده شد که الگوریتم مبتنی بر قانون از لحاظ دقت برتر از الگوریتم درخت تصمیم بود و همچنین الگوریتم مبتنی بر قانون در تشخیص صحیح بیماری (tp) و عدم بیماری (tn)عملکرد بهتری نسبت به

---

[۱] paired
[۲] True positive
[۳] False positive
[۴] True negative
[۵] False negative

درخت تصمیم داشت و همچنین در مواردی که بیماری نبوده و به اشتباه بیماری پیش بینی شده (fp) و مواردی که بیماری بوده و به اشتباه عدم بیماری پیش بینی شده (fn) مبتنی بر قانون برتر از درخت تصمیم پاسخگو بوده است. نمودار kiviat این مقایسه ها را بهتر نشان می دهد در اینجا هر چه شکل به خط نزدیکتر باشد الگوریتم بهتر عمل کرده است.




**نمودار ۲: مقایسه کارایی الگوریتم ها**

## ۱۰–مراجع

*[1] Palaniappan, Sellappan, and Rafiah Awang. "Intelligent heart disease prediction system using data mining techniques." In 2008 IEEE/ACS international conference on computer systems and applications, pp. 108-115. IEEE, 2008.*

*[2] Pintus, Roberta, Pier Paolo Bassareo, Angelica Dessì, Martino Deidda, Giuseppe Mercuro, and Vassilios Fanos. "Metabolomics and cardiology: toward the path of perinatal programming and personalized medicine." BioMed research international 2017 (2017).*

*[3] Le Boudec, Jean-Yves. Performance evaluation of computer and communication systems. Epfl Press, 2011.*

*[4] Obenshain, Mary K. "Application of data mining techniques to healthcare data." Infection Control & Hospital Epidemiology 25, no. 8 (2004): 690-695.*

*[5] Liu, Ying, Jayaprakash Pisharath, Wei-keng Liao, Gokhan Memik, Alok Choudhary, and Pradeep Dubey. "Performance evaluation and characterization of scalable data mining algorithms." In 16th IASTED international conference on parallel and distributed computing and systems (PDCS). MIT, Cambridge, pp. 620-625. 2004.*

*[6] Oprea, Cristina. "Performance evaluation of the data mining*

*classification methods." Information society and sustainable development 2344 (2014): 249-253.*
*[7] Ganganwar, Vaishali. "An overview of classification algorithms for imbalanced datasets." International Journal of Emerging Technology and Advanced Engineering 2, no. 4 (2012): 42-47.*
*[8]Patil, Tina R., and S. S. Sherekar. "Performance analysis of Naive Bayes and J48 classification algorithm for data classification." International journal of computer science and applications 6, no. 2 (2013): 256-261.*
*[9] Ganganwar, Vaishali. "An overview of classification algorithms for imbalanced datasets." International Journal of Emerging Technology and Advanced Engineering 2, no. 4 (2012): 42-47.*
*[10] Blake, C.L., Mertz, C.J,."UCI Machine Learning Databases", http://mlearn.ics.uci.edu/databases/heart-disease/,2004.*
[۱۱] اسماعیلی،مهدی،مفاهیم و تکنیک های داده کاوی،تهران،آتی نگر،۱۳۹۱
[۱۲] اسماعیلی،مهدی،آموزش گام به گام داده کاوی با رپیدماینر،تهران،آتی نگر،۱۳۹۵