بررسی الگوریتم های داده کاوی جهت پیش بینی و تشخیص بیماری قلبی

فاطمه جمالو، ریحانه ابراهیمی، عسل خسروی

گروه علوم کامپیوتر، دانشکده علوم ریاضی، دانشگاه الزهراء، تهران، ایران

چکیده

در زمینه تشخیص و پیش بینی بیماری قلبی، با توجه به وابستگی عملکرد هر یک از اندام های بدن به عوامل متفاوت و گوناگون، ما با انبوهی از داده ها روبرو هستیم که برای تجزیه و تحلیل آنها، بکار گیری ابزارها و تکنیک های مناسب داده کاوی ما را یاری خواهد نمود. در این مقاله با تجزیه و تحلیل مدلهای مختلف داده کاوی از جمله درخت تصمیم، جنگل تصادفی، طبقه بندی ساده بیزین، K نزدیکترین همسایگی، خوشه بندی K میانگین، خوشه بندی مبتنی بر چگالی، ماشین های بردار پشتیبان و شبکه عصبی مصنوعی به پیش بینی بیماری قلبی می پردازیم و در نهایت با مقایسه معیارهای ارزیابی متناسب هر مدل، بهترین و موثرترین مدل را ارائه می دهیم. این مدل ها می توانند برای شناسایی افراد مبتلا به بیماری قلبی و پیشگیری از درمان های پر هزینه، مفید باشد.

کلمات کلیدی: داده کاوی، شبکه عصبی مصنوعی، درخت تصمیم، K نزدیکترین همسایگی، خوشه بندی K میانگین، ماشین های بردار پشتیبان، طبقه بندی ساده بیزین، خوشه بندی مبتنی بر چگالی و جنگل تصادفی

۱. مقدمه

قلب یکی از اعضای مهم و حیاتی بدن انسان است. پیش بینی بیماری و تشخیص عملکرد نامناسب این عضو مهم بعضاً بسیار پر هزینه می باشد. در سال های اخیر نیز بیماری قلبی یکی از بیماریهای زمینه ای مهم بوده که استعداد بدن برای ابتلا به بیماری های دیگر را بالا می برد، لذا در جوامع پزشکی سعی می شود از طریق سایر علوم از قبیل آمار و کامپیوتر، انبوه داده ها و اطلاعات سوابق بیماران و پرونده های پزشکی جهت شناسایی قوانین حاکم بر ایجاد بیماری قلبی را ساماندهی و تجزیه و تحلیل نمایند. بنابراین پیچیدگی اطلاعات پزشکی و وجود ابزار های داده کاوی باعث می شود که داده کاوی بر روی داده های پزشکی بسیار مهم تلقی شود. مطالعات زیادی در خصوص تشخیص بیماری قلبی با استفاده از تکنیکهای مختلف داده کاوی انجام شده است. در مقاله تحت عنوان بررسی متدهای کلاس بندی و انتخاب ویژگی برای پیش بینی بیماری قلبی بکارگیری روشهای یادگیری ماشین جهت تشخیص بیماران قلبی، الگوریتم های مختلف را به منظور یافتن بهترین صحت پیاده سازی و نتایج را مقایسه نموده است. در این مقاله بیشتر بدنبال مواردی بودند که مدل، فرد را بیمار تشخیص نداده در صورتیکه واقعیت چیزی دیگری است (False-Negative) یعنی داده جدید را درست برچسب بزنند و بایاس تخمین و همچنین خطا را کم کنند. این مقاله همانند مقاله های دیگر از مدل CRISP[1] استفاده کرده است و عملکرد خوب الگوریتم های یادگیری عمیق را نشان می دهد و نتیجه می گیرد که ویژگیهای کلسترول و فشار خون بر اساس

[1] Cross-Industry Standard Process

ویژگی سن، معیار زیاد خوبی نمی باشد ولی ویژگی های تالاسمی و درد قفسه سینه تاثیر بیشتری برای تشخیص داشته است. استنباط اولیه در این مقاله این بوده است که مدل درخت تصمیم به خوبی پاسخگو می باشد در حالیکه مدل رگرسیون منطقی[2] هم عملکرد موثرتری را نشان داده است. [1] در مقاله دیگری که از روش جدید FRC index[3] استفاده شده است، به ما نشان می دهد دو الگوریتم جنگل تصادفی[4] و شبکه عصبی پرسپترون چند لایه[5] دارای صحت نزدیک به هم هستند و اگر بخواهیم الگوریتم RF را به الگوریتم MLP ترجیح دهیم باید بین شفافیت و صحت یک سازش وجود داشته باشد. تجزیه و تحلیل عملکرد انجمنی می تواند به بهتر فهمیدن و پیاده سازی کردن الگوریتم درخت تصمیم کمک کند. [2] در مقاله دیگر الگوریتم های درخت تصمیم، ماشین بردار پشتیبان، K نزدیکترین همسایگی، شبکه عصبی مصنوعی و رگرسیون منطقی روی دیتاست UCI پیاده سازی شده است. دیتاست استفاده شده گسسته و کلاس برچسب هایش باینری بوده و در نهایت به این نتیجه رسیده اند که الگوریتم درخت تصمیم با کمترین محاسبه و بطور صریح توانسته برچسب را تشخیص دهد. [3] در مقاله دیگری مدل CRT[6] استفاده شده و ۱۳ ویژگی را به ۱۱ ویژگی کاهش داده است. همچنین ویژگی قند خون ناشتا و نوار قلب را لحاظ نکرده است. با بررسی نتایج شبیه سازی، توانایی این مدل برای تشخیص بیماری قلبی با بهترین صحت و مناسبترین همگرایی در داده ها را نشان داده است که در مقایسه با بقیه متدهای استفاده شده، در CRT نتایج منطقی تری بدست آمده است. [4] در پژوهش دیگری با استفاده از وابستگی مقادیر ویژگی ها، آنها را برای پیش بینی بیماری قلبی تجزیه و تحلیل کرده است. ۱۳ ویژگی را مورد بررسی قرار داده و براساس وابستگی به آنها وزن و رتبه داده است و از روی پیاده سازی متدهای کلاس بندی مختلف به این نتیجه رسیده است که ۱۰ ویژگی Thal, CP, CA, Oldpeak, Exang, Thalach, Slope, Age, Sex, Resteeg در بیماری قلبی به هم مرتبط هستند و پیشنهاد نموده که برای مطالعه روی تاثیر تکنیک های Big data و بالا بردن صحت از این دسته ویژگی ها استفاده شود. [5] در مقاله ای دیگر عملکرد مدل های کلاس بندی (که توسط تکنیک های یادگیری ماشین ایجاد شده) و بکارگیری متدهای مختلف انتخاب ویژگی بررسی شده است و این نتیجه حاصل گردیده، که این امکان وجود دارد که مدل های بهتر با صحت بالاتر و کاملتری برای پیش بینی بیماری قلبی میتوان ایجاد کرد. با بکارگیری متدهای انتخاب ویژگی و استخراج ویژگی روی داده های ترکیبی، بهترین مدل با صحت مناسب و مطلوب حاصل شده است. [6] در مقاله ای دیگر استفاده ترکیبی از مدل های چارچوبی را نشان می دهد که شامل چندین مرحله آماده سازی داده، آموزش و تست می باشد. در واقع مدلی ترکیبی از درخت تصمیم، رگریسون منطقی و ماشین های بردار پشتیبان را ارائه می دهد تا به پیش بینی دقیق تری از بیماری های قلبی برسد. بعلاوه با مقایسه بین مدل ارائه شده و الگوریتم های مختلف به مدلی موثر که در عین حال قادر است ریسک و احتمال ابتلا به بیمار قلبی را بررسی کند، دست یافته است. [7] در مقاله ای دیگر مدلی ترکیبی از KNN[7] و ژنتیک پیشنهاد شده تا طبقه بندی مطلوبتری ارائه گردد. همانطور که میدانیم الگوریتم ژنتیک سعی در بهینه سازی راه حلی دارد که برای یک مساله در نظر می گیریم.

---

[2] Logistic Regression

[3] Feature Ranking Cost

[4] Random Forest، که از این لحظه به بعد به جهت خلاصه نویسی از مخفف معتبر RF استفاده میکنیم

[5] Multilayer perceptron، که از این لحظه به بعد به جهت خلاصه نویسی از مخفف معتبر MLP استفاده میکنیم

[6] Classification and Regression Tree

[7] K Nearest Neighbor، که از این لحظه به بعد به جهت خلاصه نویسی از مخفف معتبر KNN استفاده میکنیم

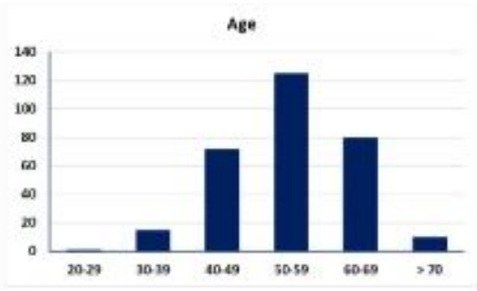
[8] در پژوهشی دیگر از شناسایی الگو و روش های داده کاوی در پیش بینی مدل ها در حوزه تشخیص بیماری های قلبی صحبت شده است. آزمایش ها با استفاده از الگوریتم طبقه بندی NB [8]، DT [9]، KNN و ANN [10] انجام شده و نتایج ثابت می کند که تکنیک NB نسبت به دیگر تکنیک های استفاده شده از لحاظ عملکرد پیشی گرفته است. [9] در مقاله ای دیگر به معرفی یک روش کارآمد برای استخراج الگوهای خاص در بیماری های قلبی پرداخته است. در واقع با استفاده از الگوریتم K-means داده های جمع آوری شده را خوشه بندی کرده و ویژگی های مرتبط به هم را با استفاده از الگوریتم MFIA [11] برای تشخیص حمله قلبی استخراج کرده است. بدین صورت که موارد مکرر که بدلیل تکرار الگویشان مقادیر بزرگتر و احتمال رخداد بیشتری دارند تاثیر بسزایی هم در تشخیص بیماری دارند. [10]

دربخش دوم این مقاله به ارائه مجموعه داده ها و معرفی ویژگی ها می پردازیم. در بخش سوم چندین مدل متفاوت داده کاوی نظیر درخت تصمیم، جنگل تصادفی، طبقه بندی ساده بیزین، K نزدیکترین همسایگی،خوشه بندی K میانگین، خوشه بندی مبتنی بر چگالی، ماشین های بردار پشتیبان و شبکه عصبی را معرفی می نماییم. در بخش چهارم با ارزیابی و لحاظ نمودن پارامترهای متناسب هر الگوریتم، بهترین میزان پارامتر استخراج شده و بهترین مدل را برای پیش بینی بیماری قلبی نشان می دهیم.

۲. شناخت، آماده سازی و جمع آوری داده

تعداد ۳۰۳ نمونه داده بیماری قلبی با ۱۴ ویژگی به همراه نمودارشان در این پژوهش مورد بررسی قرار می گیرد. در این نمودارها محور افقی، داده های اسمی یا بازه برای داده های عددی و محور عمودی، تعداد نمونه ها برای هر ویژگی خاص را بیان می کند. [11]

- ویژگی سن، برای محدوده سنی بین ۲۹ تا ۷۷ سال تعریف شده است.



شکل ۱: نمودار هیستوگرام ویژگی سن

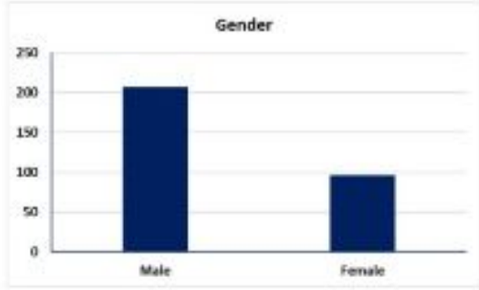
[8] Naive Bayes Network، که از این لحظه به بعد به جهت خلاصه نویسی از مخفف معتبر NB استفاده میکنیم.

[9] Decision Tree، که از این لحظه به بعد به جهت خلاصه نویسی از مخفف معتبر DT استفاده میکنیم.

[10] Artificial Neural Network، که از این لحظه به بعد به جهت خلاصه نویسی از مخفف معتبر ANN استفاده میکنیم.

[11] Maximal Frequent Itemset Algorithm

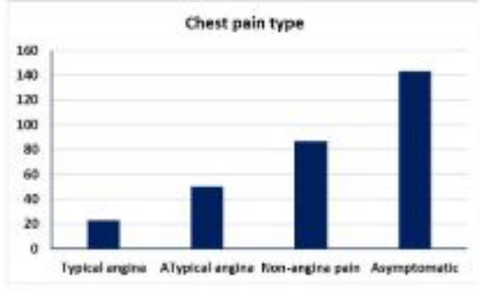
- ویژگی جنسیت، برای دو جنس مرد و زن تعریف شده است.



شکل ۲: نمودار هیستوگرام ویژگی جنسیت

- ویژگی درد قفسه سینه[12]، در چهار سطح تعریف می شود:

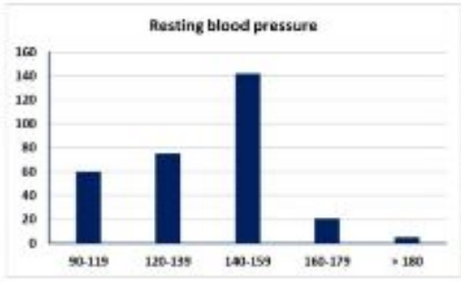
- آنژین معمول[13] عبارت است از شروع درد قفسه سینه به دنبال فعالیت، که به شکل سنگینی، فشار، له شدگی و خفگی احساس می شود و با استراحت بهبود می یابد و ممکن است با انتشار به ریشه گردن، فک، دندان ها و درد بین دو کتف همراه باشد.
- آنژین غیرمعمول[14] عبارت است از علائمی نظیر تنگی نفس، تهوع، خستگی و ضعف احساس شود (این حالت در افراد مسن و دیابتی شایع تر است).
- درد غیر آنژین[15] عبارت است از علائمی نظیر رفلاکس معده و مری، که می تواند درد قلبی را تشدید کند نظیر آمبولی ریه، التهاب دنده ها، دردهای دیواره قفسه سینه، درد ناشی از رومانورم عصبی
- بدون علامت[16] انفاکتوس حاد بدون علامت می باشد و احتمال حوادث کرونری را افزایش می دهد.



شکل ۳: نمودار هیستوگرام ویژگی درد قفسه سینه

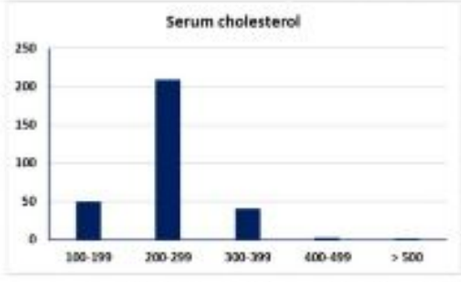
[12] Chest pain type
[13] Atypical angina
[14] Typical angina
[15] No-angina pain
[16] Asymptomatic

- ویژگی فشار خون حالت سکون[17]، این ویژگی در شرایط مختلف بصورت زیر تعریف می شود:
  - در کلینیک ۱۴۰/۹۰ mmHg
  - در منزل هنگام بیداری ۱۳۸/۸۵ mmHg
  - بالاتر از این مقادیر بیانگر ابتلای فرد به بحران فشار خون (HTN)
  - در منزل هنگام خواب ۱۲۰/۷۵ mmHg



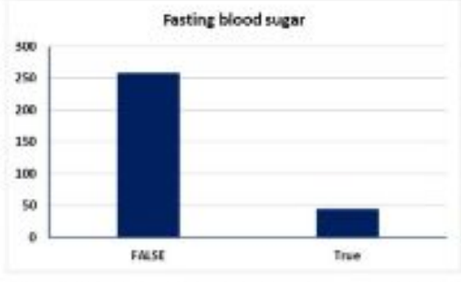
شکل ۴: نمودار هیستوگرام ویژگی فشار خون حالت سکون

- ویژگی کلسترول سرم[18]، میزان کلسترول کل بطور نرمال باید کمتر از ۲۰۰ mg/dl در افراد بالغ باشد. میزان کلسترول بین mg/dl ۲۰۰ تا ۲۳۹ بعنوان محدوده ایمن در نظر گرفته می شود. بیشتر از ۲۴۰ mg/dl بعنوان کلسترول بالا و پر خطر در نظر گرفته می شود. کمتر از ۱۰۰ mg/dl بعنوان نرمال در نظر گرفته می شود.



شکل ۵: نمودار هیستوگرام ویژگی کلسترول سرم

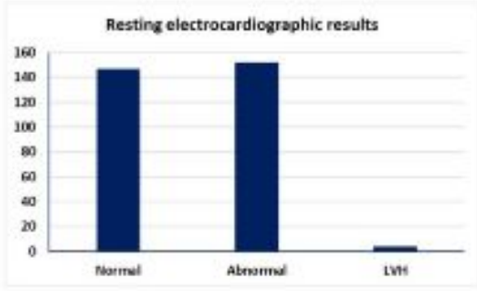
- ویژگی قند خون ناشتا[19]، میزان قند خون در حالت نرمال زیر ۱۰۰ mg/dl بوده و بین ۱۲۵-۱۰۰ mg/dl بعنوان اختلال گلوکز ناشتا نامیده می شود که این شرایط پیش دیابت و مقادیر بالاتر از ۱۲۶ mg/dl بعنوان دیابت شناخته می شود.



شکل ۶: نمودار هیستوگرام ویژگی قند خون ناشتا

[17] Resting blood pressure
[18] Serum cholesterol
[19] Fasting blood sugar

• ویژگی الکتروکاردیوگرافی حالت سکون[20] (نوار قلب)، درالکتروکاردیوگرافی بطور نرمال باید چند آیتم در نظر گرفته شود: نظیر وجود موج P, QRS, T، ریتم نرمال و سینوسی، ضربان حدود ۷۰ بار در دقیقه و محور قلب نرمال (محدوده بین ۳۰- تا ۹۰+ درجه)، که وجود در هر کدام از آیتم های فوق الکتروکاردیوگرافی غیر نرمال را نمایش می دهد. هیپرتروفی بطن چپ (LVH) در الکتروکاردیوگرافی، انحراف محور قلب به سمت چپ، R های بلند در لیدهای سمت چپ قلب، افسردگی ST و الگوی تنش را نمایش می دهد.



شکل ۷: نمودار هیستوگرام ویژگی الکتروکاردیوگرافی حالت سکون

• ویژگی ماکزیمم ضربان قلب[21]، میزان ضربان قلب بیشتر از ۱۰۰ بار در دقیقه بعنوان تاکی کاردی در نظر گرفته می شود. تاکی کاردی علل مختلفی دارد که شایع ترین آن استرس و تحریک عصب سمپاتیک، فشار خون بالا، بیماری کرونری و نارسایی قلبی می باشد.

شکل ۸: نمودار هیستوگرام ویژگی ماکزیمم ضربان قلب

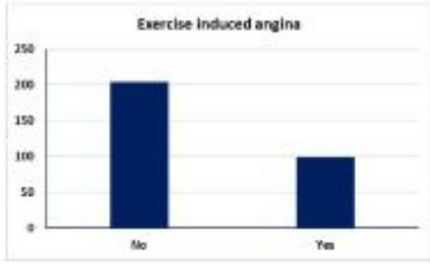
• ویژگی آنژین ناشی از ورزش[22]، اگر درد موضعی قلب بدنبال فعالیت و ورزش اتفاق بیافتد به معنای وجود تنگی در عروق کرونری و وجود بیماری کرونری می باشد زیرا بدنبال فعالیت میزان نیاز عضله قلب به اکسیژن و خونرسانی افزایش یافته و بدلیل وجود تنگی در عروق خونرسانی کامل برای قلب تامین نشده و درد آنژیمی اتفاق می افتد.

---

20 Resting electrocardiography

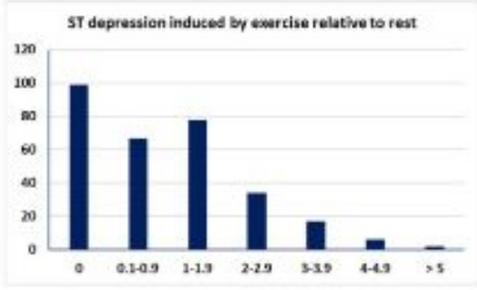
21 Maximum heart rate achieved

22 Exercise induced angina



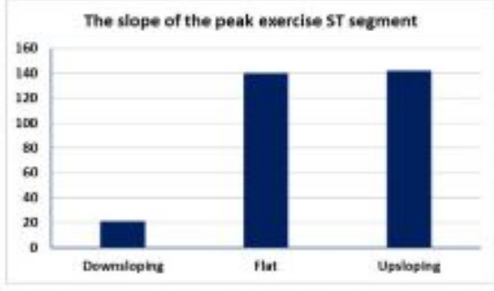
شکل ۹: نمودار هیستوگرام ویژگی آنژین ناشی از ورزش

- ویژگی افسردگی ST ناشی از فعالیت نسبت به حالت سکون[۲۳]، اگر افسردگی ST بدنبال ورزش اتفاق بیافتد نشانگر وجود تنگی در عروق کرونری است که با استراحت کردن این افسردگی ST در الکتروکاردیوگرافی از بین می رود.



شکل ۱۰: نمودار هیستوگرام ویژگی افسردگی ST ناشی از فعالیت نسبت به حالت سکون

- ویژگی شیب ST در اوج فعالیت[۲۴]، در تست ورزش هنگامی که HR بیمار به عدد هدف مورد نظر برسد با نگاه به الکتروکاردیوگرافی می توانیم وجود یا عدم وجود تنگی عروق کرونری را مشخص کنیم. اگر در الکتروکاردیوگرافی تغییرات نواری مبنی بر افسردگی ST بصورت شیب منفی و یا صاف مشاهده کنیم (حداقل یک خانه پایین بیاید و دو خانه تداوم داشته باشد) یعنی تست ورزش مثبت است. اما اگر شیب مثبت باشد مورد قبول نبوده و تست مثبت نیست.

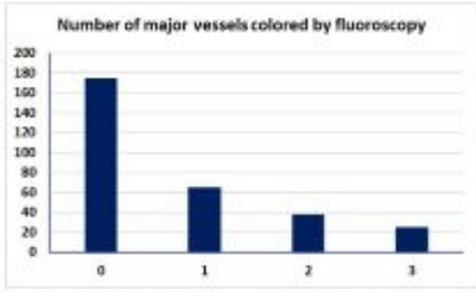


شکل ۱۱: نمودار هیستوگرام ویژگی شیب ST در اوج فعالیت

[23] ST depression induced by exercise relative to rest

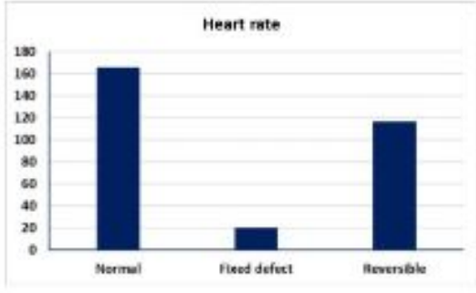
[24] The slope of the peak exercise ST segment

- ویژگی تعداد عروق اصلی رنگ شده توسط فلوروسکوپی[۲۵]، افزایش تعداد عروق فلوروسکوپی شده، شدت درگیری سه دریچه قلب را نشان داده و علامت وجود بیماری عروق کرونری شدید است.



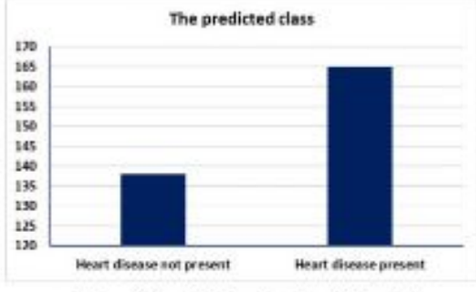
شکل ۱۲: نمودار هیستوگرام ویژگی تعداد عروق اصلی رنگ شده توسط فلوروسکوپی

- ویژگی تالاسمی[۲۶]، در سه سطح تعریف می شود: نرمال، ضعف ثابت و ضعف برگشت پذیر



شکل ۱۳: نمودار هیستوگرام ویژگی تالاسمی

- ویژگی هدف[۲۷]، هدف تعیین کلاس پیش بینی بیمار قلبی است.



شکل ۱۴: نمودار هیستوگرام ویژگی هدف

[25] Number of major vessels colored by fluoroscopy
[26] Heart rate
[27] The predicted class

جدول ۱: انواع ویژگی ها و مقدار آن

شماره	ویژگی	نوع	مقدار
۱	سن	عدد	رنج ۲۹ - ۷۷ سال میانگین : ۵۴.۴
۲	جنسیت	اسمی	۰- زن (۹۶) ۱- مرد (۲۰۷)
۳	درد قفسه سینه	اسمی	۰= آنژین معمول (۲۳) ۱= آنژین غیرمعمول (۵۰) ۲- درد غیرآنژین (۸۷) ۳- بدون علامت (۱۴۳)
۴	فشار خون ایستا	عدد	۹۴ - ۲۰۰ میانگین : ۱۳۱.۶
۵	کلسترول سرم	عدد	۱۲۶ - ۵۶۴ میانگین : ۲۴۶.۳
۶	قند خون ناشتا	اسمی	۰- نداشتن (۲۵۸) ۱- داشتن (۴۵)
۷	الکتروکاردیوگرافی ایستا (نوار قلب)	اسمی	۰= نرمال (۱۴۷) ۱= دارای موج غیر نرمال ST-T (۱۵۲) ۲= نشانه بروز احتمالی LVH (۴)
۸	ماکزیمم ضربان قلب حاصل	عدد	۷۱ - ۲۰۲ میانگین : ۱۴۹.۶
۹	آنژین ناشی از ورزش	اسمی	۰- ندارد (۲۰۴) ۱- دارد (۹۹)
۱۰	افسردگی ST ناشی از فعالیت نسبت به ایستا	عدد	۰ - ۶.۲ میانگین : ۱.۰
۱۱	شیب اوج فعالیت بخش ST	اسمی	۰- شیب منفی (۲۱) ۱- صاف (۱۴۰) ۲- شیب مثبت (۱۴۲)
۱۲	تعداد عروق اصلی رنگ شده توسط فلوروسکوپی	اسمی	۰= (۱۷۵) ۱- (۶۵) ۲- (۳۸) ۳- (۲۵)
۱۳	تالاسمی	اسمی	۱- نرمال (۱۶۶) ۲- ضعف ثابت (۲۰) ۳- ضعف برگشت پذیر (۱۱۷)
۱۴	هدف	اسمی	۰= بیمار قلبی نیست (۱۳۸) ۱= بیمار قلبی هست (۱۶۵)

به منظور شناخت کافی روی داده ها و ویژگی های آنها، با یکی از متخصصین بیمارستان قلب شهید رجایی مذاکره صورت گرفت و به منظور آماده سازی داده روی تمام مقادیر ویژگی های عددی، نرمال سازی z-score را طبق فرمول (۱) پیاده سازی کردیم. در این فرمول، y نمونه نرمال سازی شده، x نمونه قبل از نرمال سازی، μ میانگین داده ها و δ انحراف معیار داده ها می باشد.

$$y = \frac{x-\mu}{\delta} \quad (1)$$

۳. متدولوژی

در این مقاله از ۸ الگوریتم داده کاوی جهت تجزیه و تحلیل مقادیر ویژگی های بیماری قلبی بشرح زیر استفاده شده است:

- درخت تصمیم (Decision Tree)

یکی از پرکاربرد ترین الگوریتم‌های داده کاوی، الگوریتم درخت تصمیم است. این الگوریتم، مجموعه داده را به زیردسته های کوچکتر تجزیه می کند و یک درخت تصمیم مرتبط به صورت تدریجی را تشکیل می دهد و ساختار آن مانند یک درخت است با این وجه تمایز که از ریشه به سمت پایین (برگ) رشد کرده است. بالاترین گره تصمیم گیری در یک درخت که مطابق با بهترین پیش بینی کننده است، گره ریشه نام دارد. در واقع با مجموعه ای از شرط های منطقی برای پیش بینی یک ویژگی روبرو هستیم. درخت های تصمیم گیری می توانند برای داده های گسسته و پیوسته به کار روند و در این الگوریتم برای اینکه بدانیم روی کدام ویژگی درخت را بشکنیم، از ویژگی استفاده می شود که حداکثر مقدار را در فرمول (۲) دارد. همچنین این الگوریتم جزو دسته الگوریتم های نظارتی است و برای ارزیابی می توان از معیارهای یادآوری، صحت و دقت استفاده نمود.

$$\text{Gain Ratio} = \frac{\text{Information Gain}}{\text{Split } \textit{Entropy}} = \frac{\text{H(s)(before)} - \sum_{j=1}^{K} \frac{|S_p|}{|S|} \text{H}(S_p)\text{(after)}}{-\sum_{j=1}^{K} p_j \log_2 p_j} \quad (2)$$

$$\text{Entropy: H(S)} = -\text{p}(+)\log_2 \text{p}(+) - \text{p}(-)\log_2 \text{p}(-) \quad (3)$$

که در آن p مثبت تعداد داده های با برچسب مثبت بروی کل داده ی آن شاخه و p منفی تعداد داده های با برچسب منفی بروی کل داده ی آن شاخه، $\frac{|S_p|}{|S|}$ تعداد داده ی بعد از تقسیم بروی داده های قبل از تقسیم. در واقع وزن آن شاخه بعد از تقسیم یک ویژگی است.

- جنگل تصادفی (Random Forest)

یک روش یادگیری ترکیبی برای دسته‌بندی، رگرسیون می‌باشد، که بر اساس ساختاری متشکل از شمار بسیاری درخت تصمیم، بر روی داده های آموزش و پیش‌بینی هر درخت به شکل مجزا، عمل می نماید. جنگل‌های تصادفی برای درخت های تصمیم که در مجموعه آموزشی دچار بیش برازش[۲۸] می‌شوند، مناسب هستند. عملکرد جنگل تصادفی معمولاً بهتر از درخت تصمیم است و سریعتر به زیر درخت خالص[۲۹] می رسد، اما این بهبود عملکرد تا حدی به نوع داده بستگی دارد. عملکرد جنگل تصادفی به این صورت است که ویژگی ها را بصورت تصادفی انتخاب می کند که این تاثیر داده های نویز و ویژگی های بی ربط را کم می کند و باعث کاهش رخداد نفرین بعد می شود و سپس بر اساس آنتروپی و فرمول (۲) و (۳) درخت ایجاد می کند.

[28] Overfitting
[29] Pure Subset

• طبقه بندی ساده بیزین (Naïve Bayes)

بیزین یک طبقه‌بندی کننده احتمالاتی ساده است که مجموعه‌ای از احتمالات را با تجمیع فرکانس و ترکیب ارزش مجموعه داده‌ها با استفاده از فرمول (۴) و (۵) محاسبه می‌کنند. این الگوریتم از قضیه بیز استفاده می‌کند و تمام ویژگی‌های مستقل یا غیر وابسته به آن‌ها را با مقدار متغیر کلاسه بندی شده در نظر میگیرد. روش بیز یک روش احتمالاتی است که برای طبقه‌بندی کلاس‌ها از داده‌ها استفاده می‌کند و مبتنی بر احتمالات شرطی است. مزیت کلاسه بندی ببزین این است که نیاز به مقدار کمی از داده‌های آموزشی برای تخمین پارامترها (میانگین‌ها و واریانس متغیرها) لازم برای کلاسه بندی دارد. از آنجا که متغیرها مستقل فرض می‌شوند، تنها واریانس متغیرها برای هر کلاس باید تعیین شود و نه کل. این روش می‌تواند برای هر دو مساله طبقه‌بندی دوتایی و چند کلاسه مورد استفاده قرار گیرد. این الگوریتم برای ایجاد مدل‌هایی با قابلیت‌های پیش‌بینی استفاده می‌شود:

Bayes' Theorem:

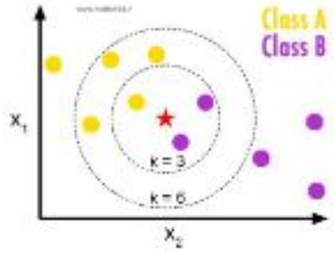
$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \qquad (4)$$

برای محاسبه احتمال A به B، الگوریتم تعداد مواردی را محاسبه می‌کند که در آن A و B با هم اتفاق می‌افتند و آن را به تعداد موارد تقسیم می‌کنند که در آنها **B** تنها رخ می‌دهد. P (B | A) احتمال پیشین شرطی B بر A است. P (B) احتمال قبلی B است.

$$\text{Posterior} = \frac{\text{Likelihood} * \text{Prior}}{\text{Evidence}} \qquad (5)$$

• K نزدیکترین همسایگی (KNN)

الگوریتم نزدیک‌ترین همسایگی برای مسائل طبقه‌بندی و رگرسیون قابل استفاده است. این الگوریتم، از "تشابه ویژگی" برای پیش بینی مقادیر نقاط داده جدید استفاده می کند. بدین معنی است که به نقطه داده جدید بر اساس میزان مطابقت آن با نقاط مجموعه داده، یک مقدار اختصاص می دهد. ابتدا باید مقدار K را انتخاب کنیم، K می تواند هر عدد صحیحی باشد، با کمک هر یک از متد های اقلیدسی فرمول های (۶) و (۷) و (۸) ، فاصله بین داده تست و ویژگی های تمام داده ها را محاسبه می کنیم. سپس بر اساس مقدار فاصله، آنها را به صورت صعودی مرتب می کنیم. در نهایت، K تا نزدیک ترین همسایگی داده جدید را در یک کلاس قرار می دهیم. اگر مقدار K به تعداد داده ها نزدیک شود دقت آزمون و آموزش کم می شود و مدل بصورت کلی داده ها را تفسیر می کند و اگر مقدار K بسیار کم باشد بیش برازش اتفاق می افتد در نتیجه دقت آموزش افزایش و دقت آزمون کاهش می یابد.



شکل ۱۵: کلاس بندی نقاط

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}\ ;\ \text{Euclidian} \tag{6}$$

$$\sum_{i=1}^{k}|x_i - y_i|\ ;\ \text{Manhattan} \tag{7}$$

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^p\right)^{\frac{1}{p}}\ ;\ p \geq 3 \tag{8}$$

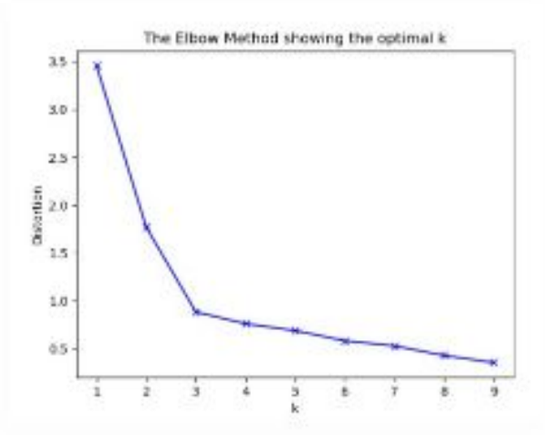
که در آن x_i بردار ویژگی داده ی جدید، y_i بردار ویژگی داده ی قدیم است.

- خوشه بندی K میانگین (K-means)

روش تحلیل خوشه‌ای برای تعیین ویژگی‌های مجموعه داده‌ها و متغیر هدف آنالیز می‌شود. که معمولا برای تعیین نحوه اندازه‌گیری شباهت استفاده می‌شود. الگوریتم K-means ساده‌ترین الگوریتم یادگیری برای حل مشکلات خوشه بندی است. این فرآیندی ساده و آسان است که داده‌ها را به تعدادی از خوشه‌ها طبقه‌بندی می‌کند. ما سعی میکنیم که خوشه ها تا حد امکان از هم دور باشند سپس، هر نقطه متعلق به مجموعه داده ی ارائه شده را برداشته و به نزدیک‌ترین مرکز دسته مرتبط می کنیم. و هر بار برای داده های هر خوشه میانگین ویژگی ها را گرفته، سپس مرکز دسته جدید به دست می آید. هنگامی که به K تا مرکز دسته می‌رسیم، پیوند جدیدی بین داده‌ها و نزدیک‌ترین مرکز دسته انجام می‌شود. این حلقه به این دلیل ایجاد می‌شود که تغییرات کلیدی این حلقه تا زمانی که هیچ تغییری انجام نشود، موقعیت مکان را تغییر می‌دهد. مراحل الگوریتم K-means عبارت است از:

۱. از x_i ، $i=1,2,\ldots..n$ که باید در K تا خوشه تقسیم‌بندی شود، استفاده کنید.
۲. مشخص کردن نزدیک‌ترین مرکز خوشه به هر نقطه داده با استفاده از فاصله اقلیدسی یا منهتن
۳. موقعیت هر خوشه را نسبت به میانگین هر نقطه داده (μ_j) متعلق به آن خوشه تعیین کنید
۴. مراحل ۲ و ۳ را تا زمانی که دسته ها و مرکز دسته ها تغییری نکنند، تکرار کنید

$$J = min\sum_{i=1}^{n}\sum_{j=1}^{k}||x_i - \mu_j||^2 \tag{9}$$



شکل ۱۶: نمودار Elbow

فرمول (۹) هزینه خوشه بندی را بیان می کند که برای تعیین K مناسب (تعداد خوشه) از آن استفاده می شود. مطابق شکل (۱۶) محور افقی K و محور عمودی j را نشان می دهد. برای K کمتر، j (Distortion) افزایش پیدا می کند. یعنی مدل، داده ها را کلی تر تفسیر می کند و هر چه مقدار K به تعداد داده ها نزدیکتر شود، j کاهش پیدا می کند و دقت خوشه بندی افزایش پیدا می کند که در واقع هر داده را یک خوشه در نظر می گیرد در صورتیکه این اصلا بهینه نیست. الگوریتم K-means برای معیارهای شباهت مشترک که در بالا ذکر شد، همگرا خواهد شد.

- خوشه بندی مبتنی بر چگالی (DBSCAN)

الگوریتم های خوشه بندی مبتنی بر چگالی یکی از روش های اصلی برای خوشه بندی در داده کاوی می باشد. الگوریتم DBSCAN، پایه روش های خوشه بندی مبتنی بر چگالی است که علی رغم مزایای فراوان مشکلاتی نظیر سخت بودن تعیین پارامترهای ورودی مانند شعاع همسایگی و MinPts را دارد.

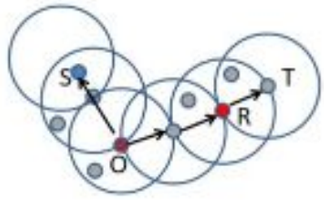
الگوریتم DBSCAN، قابلیت تشخیص خوشه ها با چگالی های متفاوت و داده های نویز را دارد و همچنین بر خلاف K-means، خوشه های تودرتو و چسبیده به هم را نیز به خوبی تشخیص می دهد. این الگوریتم نیاز به دو پارامتر MinPts و EPS دارد. پارامتر MinPts حداقل تعداد نقاط موجود در یک خوشه و پارامتر EPS شعاع همسایگی را مشخص می کند.

نقاطی که دارای حداقل تعداد MinPts با شعاع حداکثر EPS هستند نقاط مرکزی[30] و نقاطی که کمتر از MinPts در شعاع همسایگی خودشان نقطه وجود داشته باشد ولی به اندازه EPS تا نقطه مرکزی فاصله داشته باشند را نقاط مرزی[31] می نامند و نقاطی را که فاقد شرایط نقاط مرزی و مرکزی باشند را نویز می نامند.

[30] Core point
[31] Border point

نقاطی که در شعاع EPS از نقطه ی مرکزی قرار دارند نسبت به نقاط مرکزی، رابطه ی دسترسی پذیر مبتنی بر چگالی مستقیم[۳۲] دارند. هر دو نقطه ای که بواسطه یک نقطه مرکزی رابطه Density Reachable برقرار کند رابطه Density Connectivity دارند و درون یک خوشه هستند. هر دو نقطه مرکزی که در شعاع EPS هم قرار دارند نیز در یک خوشه می باشند.

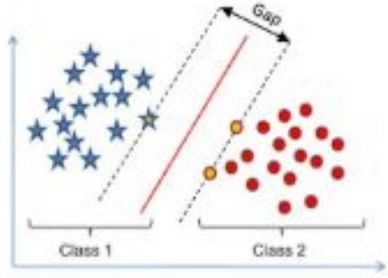
در شکل ۱۷ دیده می شود که S از O، و T از R، رابطه Directly Density Reachable دارند پس Density Connected O.R.S.T هستند.



شکل ۱۷: خوشه بندی مبتنی بر چگالی

- ماشین های بردار پشتیبان (SVM)[۳۳]

یکی از روشهای یادگیری با نظارت است که از آن برای طبقه بندی استفاده می شود. داده ها، مجموعه ای از نقاط در فضای n بعدی هستند که با مرز تصمیم و خطوط حاشیه مناسب، مرز دسته های آنها مشخص می شود. با جابجایی یکی از آنها، خروجی طبقه بندی ممکن است تغییر کند.



شکل ۱۸: مرز تصمیم و کلاس بندی

یک راه حل مناسب برای انجام این کار، ساخت یک مرز بهینه، محاسبه فاصله ی مرزهای به دست آمده با بردارهای پشتیبان هر دسته و در نهایت انتخاب مرزی است که از دسته های موجود، مجموعاً بیشترین فاصله را داشته باشد. در شکل ۱۸، خط میانی، تقریب خوبی از این مرز را نشان می دهد که از هر دو دسته فاصله ی زیادی دارد. این عمل تعیین مرز و انتخاب خط بهینه به آسانی با محاسبات ریاضی قابل انجام است.

---

[32] Directly Density Reachable

[۳۳] Supporting Vector Machine، که از این لحظه به بعد به جهت خلاصه نویسی از مخفف معتبر SVM استفاده میکنیم

$$\max. W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \quad (10)$$

$$\text{Subject to } C \geq \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (11)$$

$$K(X, Y) = \exp\left(-\frac{||X - Y||^2}{2\sigma^2}\right) \quad (12)$$

فرمول (۱۰) و (۱۱) مساله دوگان برای پیدا کردن بزرگترین مارجین یا حاشیه و یا کوچک کردن بردار عمود بر مرز تصمیم است. C در واقع خطا، α ضریب لاگرانژ و بجای ضرب داخلی $x_i^T x_j$ از توابع کرنل میتوان استفاده کرد که معروفترین آن فرمول (۱۲) RBF می باشد. در این فرمول σ شعاع کرنل است.

- شبکه عصبی مصنوعی (Artificial Neural Network)

در شبکه عصبی Neuron Transmitter ها پیام را انتقال می دهند. در شبکه‌های عصبی مصنوعی (ANN) عددها انتقال پیدا کرده و روی یال های وزن دار و جهت دار شروع به حرکت می کنند که این وزن ها در اولین مرتبه بصورت تصادفی تعیین می شود و پس از یادگیری مدل که در اثر تصحیح و اشتباه است، این وزن ها تغییر می کند. هر نرون مجموع "وزن یال x عدد روی نرون قبلی" را بعنوان ورودی دریافت می کند. حال نرون ها می توانند پیام را رد و بدل کنند. یک تابع فعالسازی[۳۴] مشخص می کند که پیامی انتقال پیدا کند یا نکند. تابع Sigmoid، فرمول (۱۳)، برای ورودی خالص z به صورت زیر تعریف می‌شود:

$$\varphi(z) = \frac{1}{1 + e^{-z}} \quad (13)$$

در این تابع اگر 0 دریافت شود 0 و اگر عددی مثبت دریافت شود 1 و اگر عددی منفی دریافت شود -1 را انتقال می دهد. شبکه‌های عصبی مصنوعی (ANN) پتانسیل تشخیص الگوهای پنهان در داده‌ها را دارند که برای پیش‌بینی بازار سهام بسیار موثر می باشد. MLP[۳۵] یک نوع شبکه عصبی است که با استفاده از الگوریتم پس انتشار[۳۶] آموزش داده می‌شود، شامل چندین لایه از واحدهای محاسباتی است که به روش شبکه پیشران[۳۷] متصل می‌شوند. الگوریتم پس انتشار از قاعده کاهش گرادیان و فرمول (۱۴) استفاده می کند.

این یک اتصال مستقیم از واحدهای پایین‌تر به یک واحد در لایه بعدی را تشکیل می‌دهد. ساختار پایه یMLP از یک لایه ورودی، یک یا چند لایه پنهان و یک لایه خروجی تشکیل شده است.

در واقع هدف شبکه عصبی یافتن بهترین وزن برای یال های بین نرونی است که با استفاده از فرمول (۱۵) خروجی برای محاسبه تغییرات وزن برای هر یک از ارتباطات بین نرونی استفاده می شود و تغییرات وزن محاسبه میگردد.

$$\Delta\omega_{ij} = -\eta \frac{\partial E}{\partial \omega_{ij}} = -\eta \delta_j o_i \quad (14)$$

$$\Delta\omega_{ij}(t+1) = (1-\alpha)\eta\delta_j o_i + \alpha\Delta\omega_{ij}(t) \quad (15)$$

---

[34] Activation function

[35] Multi-layer perceptron

[36] Back propagation

[37] Feed forward

۴. ارزیابی

پس از اجرای مدل، بایستی به ارزیابی نتایج حاصله پرداخت. ارزیابی نتایج باعث بهبود مدل شده و آنرا قابل استفاده می کند. شاخص های مختلفی از قبیل دقت (Precision)، یادآوری (Recall) و صحت (Accuracy) برای ارزیابی الگوریتم های کلاس بندی و خوشه بندی وجود دارند که طبق فرمول های (۱۶) تا (۱۸) محاسبه می شوند. برای محاسبه میزان شاخص ها میتوان از ماتریس درهم ریختگی (Confusion matrix) شکل (۱۹) استفاده کرد. حالت ایده ال ماتریس این است که بیشتر داده های مرتبط با مشاهدات، روی قطر اصلی ماتریس قرار گرفته باشند و مابقی مقادیر ماتریس صفر یا نزدیک به صفر باشند.

TN: بیانگر تعداد رکوردهایی است که مدل بدرستی تشخیص منفی داده و در واقعیت هم برچسب آنها منفی بوده است.

TP: بیانگر تعداد رکوردهایی است که مدل بدرستی تشخیص مثبت داده و در واقعیت هم برچسب آنها مثبت بوده است.

FP: بیانگر تعداد رکوردهایی است که مدل به غلط تشخیص منفی داده ولی در واقعیت برچسب آنها مثبت بوده است.

FN: بیانگر تعداد رکوردهایی است که مدل به غلط تشخیص مثبت داده ولی در واقعیت برچسب آنها منفی بوده است.

		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

شکل ۱۹: ماتریس درهم ریختگی

میزان صحت مدل های تولید شده با استفاده از فرمول (۱۶) برای داده های آموزش (Train) و آزمون (Test) طبق جدول ۲ می باشد.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (16)$$

دقت (Precision)، در واقع نسبت تمام مواردی که درست تشخیص دادیم به تمام مواردی که گفتیم بیمار هستند، با استفاده از فرمول (۱۷) بدست می آید.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (17)$$

یادآوری (Recall)، در واقع نسبت تمام مواردی که درست تشخیص دادیم به تمام مواردی که در واقعیت بیمار هستند، با استفاده از فرمول (۱۸) بدست می آید.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (18)$$

زمانیکه داده های ما بالانس نیستند یعنی نتایج مثبت بیش از ۱،۵ برابر نتایج منفی است یا بلعکس باشد، از فرمول (۱۹) برای بالانس کردن داده ها استفاده می شود.

$$\text{Fscore} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (19)$$

از آنجاییکه در این مقاله نسبت داده های مثبت و منفی بسیار کمتر از ۱،۵ می باشد لذا داده های ما نیاز به بالانس کردن ندارند پس میتوان معیار Fscore را نادیده گرفت.

جدول ۲: عملکرد الگوریتم ها

MODEL	Precision	Recall	Accuracy	Fscore	Silhouette
DT	%72. 13	%64. 7	%66. 39	%68.21	---
RF	%79. 68	%76.12	%76.23	%77. 86	---
NB	%81. 8	%73. 5	%76.92	%77.40	---
KNN	%82. 66	%91.17	%84.42	%86.71	---
K means	---	---	%81. 18	---	%16. 82
DBSCAN	---	---	%30. 69	---	%14.02
SVM	%80. 76	%94. 02	%84. 42	%86. 89	---
ANN	%70. 6	%58. 5	%64. 47	%64. 0	---

در این پژوهش برای پیاده سازی مدل DT، RF، NB، KNN و SVM ۴۰ ٪ داده ها را به آزمون و ۶۰ ٪ داده ها برای آموزش لحاظ نمودیم و در ارزیابی نتایج حاصل از پیاده سازی، به این نتیجه رسیدیم که الگوریتم های نظارتی عملکرد بهتری دارند. مقایسه نتایج عملکرد الگوریتم ها در معیارهای ارزیابی بدین شرح می باشد: در معیار ارزیابی دقت (Precision) پیش بینی الگوریتم KNN (درصد ۸۲/۶۶٪)، در معیار ارزیابی یادآوری(Recall) پیش بینی الگوریتم SVM (۹۴/۰۲ ٪) و در معیار ارزیابی صحت (Accuracy) پیش بینی الگوریتم های SVM و KNN (هر دو ۸۴/۴۲٪) مطلوبتر بودند. همچنین در الگوریتم های بدون نظارت با معیار Silhouette، پیش بینی الگوریتم K-means (۱۶/۸۲ ٪) بالاترین میزان را دارا می باشد.

۵. نتیجه گیری و ارائه پیشنهاد

هدف این مقاله، ارائه مدل پیش بینی کننده بیماری قلبی بوده، از این رو به معرفی و ارائه الگوریتم های متفاوت از جمله DT, RF NB, KNN, K-means, DBSCAN, SVM, ANN پرداختیم و با مقایسه نتایج آنها، با استفاده از معیارهای ارزیابی معرفی شده، نشان دادیم الگوریتم KNN و SVM عملکرد بهتری نسبت به سایر الگوریتم ها دارند، بدین معنی که این دو الگوریتم برروی داده های این تحقیق نتایج معنادارتر و مطلوبتری برای پیش بینی بیماری قلبی ارائه می نمایند. در الگوریتم های بدون نظارت K-means وDBSCAN که در هر دو چسبندگی درون خوشه ای و واریانس بین خوشه ای مناسبی مشاهده گردید، الگوریتم K-means دارای عملکرد موثرتری می باشد.

همچنین برای تحقیقات آتی پیشنهاد می شود، ترکیبی از متدهای Feature Selection بهمراه الگوریتم های یادگیری ماشین استفاده شود. از سوی دیگر میتوان برای بهینه کردن مدل پیش بینی به الگوریتم ژنتیک هم توجه نمود. همچنین بجای بررسی کردن ۱۳ ویژگی، میتوان برای ویژگی ها Rank (وزن / رتبه) در نظر گرفت و فقط ویژگی های مرتبط تر و موثرتر را انتخاب و بررسی نمود.

۶. مراجع

[1] N. Pereira, "Using Machine Learning Classification Methods to Detect the Presence of Heart Disease," Technological University Dublin, Dublin, 2019.

[2] H. Meshref, "Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach," *IJACSA,* vol. 10, 12 Nov 2019.

[3] M. Abdar, S. R.Niakan Kalhori, T. Sutikno and I. M. Ibnu Subroto, "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases," *IJECE,* vol. 5, 6 Nov 2015.

[4] M. Subhi Al-batah, "Testing the Probability of Heart Disease using Classification and Regression Tree Model," 2013.

[5] S. Chellammal and R. Sharmila, "Recommendation of Attributes for Heart Disease Prediction using Correlation Measure," *IJRTE,* vol. 8, no. 2S3, 2019.

[6] R. Spencer, F. Thabtah, N. Abedelhamid and M. Thompson, "Exploring feature selection and classification methods for predicting heart disease," *Digital Health,* vol. 6, pp. 1-10, 28 Feb 2020.

[7] T. Mythili, D. Mukherji, N. Padalia and A. Naidu, "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)," *International Journal of Computer Applications,* vol. 68, Apr 2013.

[8] A. j. M., B. Deekshatulu and P. Chandra, "Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm," in *CIMTA*, India, 2013.

[9] T. Peter and K. Somasundaram, "An empirical study on prediction of heart disease using classification data mining techniques," in *ICAESM*, India, 2012.

[10] S. B.Patil and Y. Kumaraswamy, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction," *IJCSNS,* vol. 9, pp. 1-8, Feb 2009.

[11] "https://www.kaggle.com/ronitf/heart-disease-uci,".