چهارمین کنفرانس
دانشجویان مهندسی کامپیوتر و فناوری اطلاعات و ارتباطات
4th Conference on
Computer Engineering, Information Technology and Communications Students

ارزیابی و مقایسه الگوریتم‌های داده‌کاوی برای پیش‌بینی ریسک حمله قلبی

علی وطن‌خواه[1]، خاطره رجب[2]

۱- کارشناسی ارشد مهندسی کامپیوتر، گروه آموزشی مهندسی برق و کامپیوتر، دانشگاه ملی مهارت، تهران، ایران،

ایمیل: alivatankhah1988@gmail.com

۲- کارشناسی مهندسی کامپیوتر، گروه آموزشی مهندسی کامپیوتر، دانشگاه ملی مهارت، تهران، ایران،

ایمیل: khatererajab@gmail.com

چکیده

مقدمه: بیماری‌های قلبی‌عروقی از مهم‌ترین تهدیدهای سلامت انسان هستند و تشخیص زودهنگام ریسک حمله قلبی می‌تواند هزینه‌ها و عوارض درمانی را کاهش دهد. هدف این پژوهش، بررسی و مقایسه عملکرد الگوریتم‌های داده‌کاوی درخت تصمیم، جنگل تصادفی، نزدیک‌ترین همسایه (KNN) و بیز ساده در پیش‌بینی ریسک حمله قلبی است.

روش‌ها: از مجموعه داده‌ای واقعی از وب‌سایت Kaggle شامل ۲۹۹۹ رکورد و ۱۹ ویژگی (سن، فشار خون، کلسترول، دیابت، سابقه خانوادگی و غیره) با برچسب «ریسک حمله قلبی» استفاده شد. پس از پیش‌پردازش (انتخاب ویژگی، تبدیل به دودویی، تعیین برچسب و نرمال‌سازی)، مدل‌سازی در نرم‌افزار RapidMiner و با تقسیم داده به ۷۰ درصد آموزش و ۳۰ درصد آزمون انجام شد و عملکرد مدل‌ها با شاخص دقت و منحنی ROC مقایسه گردید.

یافته‌ها: الگوریتم درخت تصمیم با دقت ۶۳.۷۸ درصد بهترین عملکرد را داشت و در منحنی ROC نیز به حالت ایده‌آل نزدیک‌تر بود. جنگل تصادفی با دقت ۶۳.۵۶ درصد و KNN با دقت ۵۷.۶۷ درصد در رتبه‌های بعدی قرار گرفتند. مدل بیز ساده نیز دقت ۴۷.۳۳درصد نشان دادکه ضعیف‌ترین عملکرد را در میان الگوریتم‌های بررسی‌شده نشان داد.

نتیجه‌گیری: الگوریتم درخت تصمیم نسبت به سایر مدل‌ها دقت و پایداری بالاتری در پیش‌بینی ریسک حمله قلبی داشت و به‌عنوان مناسب‌ترین مدل شناسایی شد. این نتایج می‌تواند در توسعه ابزارهای پشتیبان تصمیم پزشکی مورد استفاده قرار گیرد.

واژگان کلیدی: حمله قلبی، داده‌کاوی، یادگیری ماشین، درخت تصمیم، جنگل تصادفی.

۱. مقدمه

زندگی انسان‌ها وابسته به عملکرد مناسب قلب است و اگر عملکرد قلب به صورت مناسب صورت نگیرد روی سایر قسمت‌های بدن از قبیل ذهن، کلیه و غیره اثر خواهد گذاشت [۱]. پژوهش‌هایی که تاکنون انجام شده حاکی از این است که پیش‌بینی ابتلا به بیماری حمله قلبی در مراحل اولیه دشوار بوده؛ بنابراین توسعه نرم‌افزاری که از فاکتورهای شناخته‌شده این بیماری در انتخاب الگوریتم‌های تخصصی بهره‌برداری کند، می‌تواند آسیب‌پذیری بیماری‌های قلبی را با توجه به علائم اولیه پیش‌بینی کرده و هزینه‌های درمان را نیز کاهش دهد [۲]. به دلیل اهمیت این بیماری، روش‌ها و ابزارهای مختلفی برای بررسی عملکرد قلب در علم پزشکی ابداع شده که به پزشکان توانایی تشخیص نوع بیماری و پیش‌بینی احتمال بروز آن در آینده را می‌دهد [۳]. از آن جمله تحلیل داده‌های بیماران و استفاده از الگوریتم‌های مختلف

داده‌کاوی در پیش‌بینی و تشخیص امراض قلبی است که امروزه در علم انفورماتیک سلامت امری مرسوم و متداول گردیده است [۴]. به‌طور کلی در بسیاری از مطالعات به اثبات رسیده است که از الگوریتم‌های یادگیری ماشینی و تجزیه و تحلیل الگوریتم‌های داده‌کاوی مختلف می‌توان برای پیش‌بینی بیماری‌های قلبی استفاده نمود [۵-۱۱]. با استفاده ازنرم‌افزار RapdMinder، امکان آشنایی دقیق‌تر با ویژگی‌های داده‌ها و آماده‌سازی آن‌ها برای مراحل بعدی فراهم می‌گردد. لذا هدف اصلی ما در این بحث استفاده از الگوریتم‌های داده‌کاوی مانند Decision Tree، Random Forest، KNN و ve BayesNai برای پیش‌بینی حملات قلبی است.

۲. جمع‌آوری اطلاعات

این پروژه، با استفاده از داده‌های واقعی استخراج‌شده از وب‌سایت کگل می‌باشد. روش تحقیق در این پروژه از نوع داده‌کاوی و مبتنی بر رویکرد یادگیری ماشین می‌باشد. این تحقیق به‌صورت تجربی انجام شده و نتایج حاصل از این الگوریتم‌ها با یکدیگر مقایسه شده تا مناسب‌ترین مدل برای پیش‌بینی ریسک حمله قلبی انتخاب شود. مجموعه داده مورد استفاده شامل اطلاعات مربوط به ریسک حمله قلبی بوده و در قالب داده‌های ساختاریافته ارائه شده است. در این مجموعه داده، ۲۹۹۹ رکورد و ۱۹ ویژگی وجود دارد که بر اساس داده‌های موجود استخراج شده‌اند. همچنین این پایگاه داده دارای یک برچسب هدف مرتبط با وضعیت ریسک حمله قلبی می‌باشد که به‌عنوان متغیر خروجی در فرآیند داده‌کاوی و مدل‌سازی مورد استفاده قرار گرفته است.

مراحل روش تحقیق شامل جمع‌آوری مجموعه داده، پیش‌پردازش داده‌ها، انتخاب و پیاده‌سازی الگوریتم‌های داده‌کاوی، آموزش مدل‌ها و در نهایت ارزیابی و مقایسه نتایج حاصل می‌باشد. در این راستا، مجموعه داده پیش‌پردازش‌شده به محیط نرم‌افزار RapidMinder وارد شده و فرایند مدل‌سازی با بهره‌گیری از الگوریتم‌های منتخب طبقه‌بندی آغاز می‌شود. داده‌ها به ۷۰ درصد آموزش و ۳۰ درصد آزمون تقسیم شدند.جدول (۱)

جدول ۱-ویژگی‌های مجموعه داده تحقیق

نام ویژگی	محدوده داده
Patient ID	رشته
Age	عددی
Sex	Male-Female
Cholesterol	عددی
Blood Pressure	عدد اعشاری
Heart Rate	عددی
Diabetes	۱و۰
Family History	۱و۰
Smoking	۱و۰
Obesity	۱و۰
Alcohol Consumption	۱و۰

Diet	رشته
Previous Heart Problems	۰و۱
Stress Level	عددی
Physical Activity Days Per Week	عددی
Sleep Hours Per Day	عددی
Country	رشته
Continent	رشته
Heart Attack Risk	۰و۱

جدول (۲)و جدول (۳)به ترتیب توزیع آماری متغیرهای کمّی (سن، فشار خون سیستولیک و دیاستولیک، کلسترول) و ویژگی‌های جمعیتی و بالینی (جنسیت، سابقه خانوادگی، دیابت، سیگار، چاقی) جامعه مورد مطالعه را به تفکیک دو گروه بدون ریسک و دارای ریسک حمله قلبی نشان می‌دهند. با توجه به نزدیک بودن مقادیر میانگین و میانه در متغیرهای کمّی، توزیع داده‌ها نسبتاً متقارن بوده و فاقد انحراف چشمگیر است، هرچند مقدار نسبتاً بالای انحراف معیار سن (حدود ۲۱.۵ در هر دو گروه) بیانگر تنوع بالای جمعیت نمونه از نظر سنی است. مقایسه دو گروه در جدول(۲) نشان می‌دهد میانگین و میانه فشار خون و کلسترول تفاوت محسوسی با یکدیگر ندارند. به همین ترتیب، در جدول (۳)نیز بیشتر متغیرهای جمعیتی و بالینی نظیر جنسیت، سابقه خانوادگی، سیگار و چاقی توزیع نسبتاً یکسانی در دو گروه دارند، به‌جز ابتلا به دیابت که فراوانی آن در گروه دارای ریسک (۶۷.۷۲ درصد) نسبت به گروه بدون ریسک (۶۴.۹۱ درصد) بیشتر است و می‌تواند مؤید ارتباط این عامل با افزایش احتمال بروز حمله قلبی باشد. به‌طور کلی، نبود تفاوت چشمگیر در بیشتر متغیرهای بررسی‌شده میان دو گروه نشان می‌دهد این عوامل به‌تنهایی قادر به تفکیک دقیق افراد در معرض ریسک نیستند و ضرورت استفاده از مدل‌های چندمتغیره داده‌کاوی برای پیش‌بینی دقیق‌تر ریسک حمله قلبی را برجسته می‌سازد.

جدول ۲-توزیع آماری متغیرهای کمّی جمعیت مورد مطالعه بر اساس وضعیت ریسک حمله قلبی

ویژگی ها	بدون ریسک (n = ۱۹۵۲)				
	Mean	Median	StdDev	Min	Max
سن (سال)	۵۴.۰۵	۵۵	۲۱.۵۴	۱۸	۹۰
فشار خون سیستولیک (mmHg)	۱۳۵.۲	۱۳۵	۲۶.۵	۹۰	۱۸۰
فشار خون دیاستولیک (mmHg)	۸۵.۵۵	۸۶	۱۴.۹۱	۶۰	۱۱۰
کلسترول (dL/mg)	۲۶۰.۰۱	۲۵۶	۸۰.۷۲	۱۲۰	۴۰۰

ویژگی ها	دارای ریسک (n = ۱۰۴۷)				
	Mean	Median	StdDev	Min	Max
سن (سال)	۵۳.۹۴	۵۳	۲۱.۵۶	۱۸	۹۰
فشار خون سیستولیک (mmHg)	۱۳۵.۱۷	۱۳۵	۲۶.۴۲	۹۰	۱۸۰
فشار خون دیاستولیک (mmHg)	۸۴.۵۲	۸۴	۱۴.۷۴	۶۰	۱۱۰
کلسترول (dL/mg)	۲۶۱.۳۱	۲۵۷	۸۱.۲۱	۱۲۰	۴۰۰

جدول۳-ویژگی‌های جمعیتی و بالینی تفصیلی جمعیت مورد مطالعه بر اساس وضعیت ریسک حمله قلبی

ویژگی‌ها	بدون ریسک		دارای ریسک		مجموع	
	n	٪	n	٪	n	٪
جنسیت						
مرد	۱۳۵۵	۶۹.۴۲	۷۲۳	۶۹.۰۵	۲۰۷۸	۶۹.۲۹
زن	۵۹۷	۳۰.۵۸	۳۲۴	۳۰.۹۵	۹۲۱	۳۰.۷۱
سابقه خانوادگی بیماری قلبی						
دارد	۹۶۲	۴۹.۲۸	۴۹۶	۴۷.۳۷	۱۴۵۸	۴۸.۶۲
ندارد	۹۹۰	۵۰.۷۲	۵۵۱	۵۲.۶۳	۱۵۴۱	۵۱.۳۸
ابتلا به دیابت						
دارد	۱۲۶۷	۶۴.۹۱	۷۰۹	۶۷.۷۲	۱۹۷۶	۶۵.۸۹
ندارد	۶۸۵	۳۵.۰۹	۳۳۸	۳۲.۲۸	۱۰۲۳	۳۴.۱۱
مصرف سیگار						
دارد	۱۷۵۳	۸۹.۸۱	۹۲۷	۸۸.۵۴	۲۶۸۰	۸۹.۳۶
ندارد	۱۹۹	۱۰.۱۹	۱۲۰	۱۱.۴۶	۳۱۹	۱۰.۶۴
چاقی						
دارد	۹۵۶	۴۸.۹۸	۵۰۶	۴۸.۳۳	۱۴۶۲	۴۸.۷۵

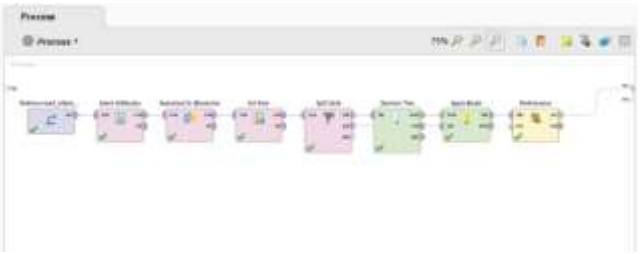
۵۱.۲۵	۱۵۳۷	۵۱.۶۷	۵۴۱	۵۱.۰۲	۹۹۶	ندارد

پس از حذف داده‌های زائد و پرت (که در این مجموعه داده مشاهده نشد)، مجموعه داده در صفحه پردازش قرار گرفت و عملگر انتخاب ویژگی برای انتخاب سن، فشار خون، کلسترول و ریسک حمله قلبی (برچسب) به‌کار گرفته شد. سپس عملگر تبدیل عدد به دودویی برای کاهش پیچیدگی داده‌ها اجرا شد و برچسب باینری به مقادیر درست/غلط تبدیل گردید. آنگاه نقش برچسب (ریسک حمله قلبی) تعیین و در نهایت عملگر نرمال‌سازی روی ویژگی‌های سن و کلسترول اعمال شد تا مقادیر ویژگی‌ها به یک مقیاس مشترک برده شوند.شکل(۱)

شکل ۱- برچسب‌دار شدن ویژگی Result برای پیش‌بینی حمله قلبی

۳. مدل‌سازی با الگوریتم درخت تصمیم

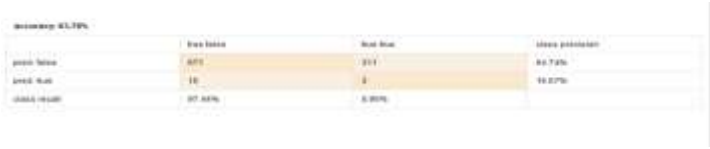
الگوریتم درخت تصمیم یکی از تکنیک‌های پرکاربرد در داده‌کاوی، سیستم‌هایی است که طبقه‌بندی‌کننده‌ها را ایجاد می‌کنند [۱۲]. این الگوریتم با تقسیم داده‌ها به زیرمجموعه‌های کوچک‌تر و همگن‌تر، ساختاری درختی از تصمیم‌ها ایجاد می‌کند. دقت درخت تصمیم در پروژه ما ۶۳.۷۸ درصد می‌باشد. این تقسیم‌بندی به مدل اجازه می‌دهد تا ابتدا الگوها را یاد بگیرد و سپس عملکرد خود را روی داده‌های جدید و ناشناخته ارزیابی کند.شکل(۲)



شکل ۲- محیط گرافیکی مدل الگوریتم درخت تصمیم

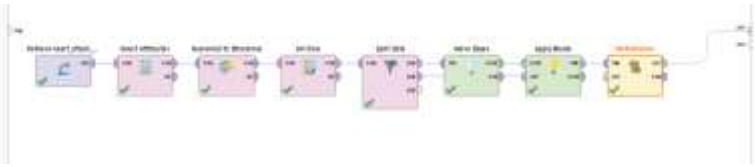
در جدول (۴) نتیجه صحت و دقت الگوریتم درخت تصمیم آورده شده است. دقت کل این الگوریتم ۶۳.۷۸ درصد به‌دست آمده است.

جدول ۴- نتیجه دقت و صحت الگوریتم درخت تصمیم



۴. مدل‌سازی با الگوریتم بیز ساده

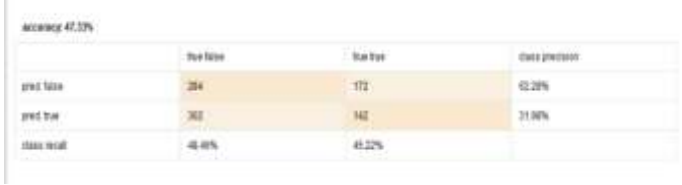
الگوریتم بیز ساده (Naive Bayes) یکی از روش‌های دسته‌بندی احتمالاتی مبتنی بر قضیه بیز است که با فرض استقلال شرطی بین ویژگی‌های ورودی، احتمال تعلق هر نمونه به هر یک از کلاس‌های خروجی را محاسبه می‌کند. این الگوریتم بر اساس توزیع پیشین کلاس‌ها (Prior Probability) و توزیع شرطی هر ویژگی نسبت به کلاس هدف، برچسب نهایی هر نمونه را با انتخاب کلاسی که بیشترین احتمال پسین را دارد تعیین می‌کند. با وجود سادگی محاسباتی و سرعت بالا در پیاده‌سازی، این الگوریتم در مسائلی که ویژگی‌ها همبستگی بالایی با یکدیگر دارند، معمولاً از دقت پایین‌تری برخوردار است.شکل(۳)



شکل ۳- محیط گرافیکی مدل الگوریتم بیز ساده

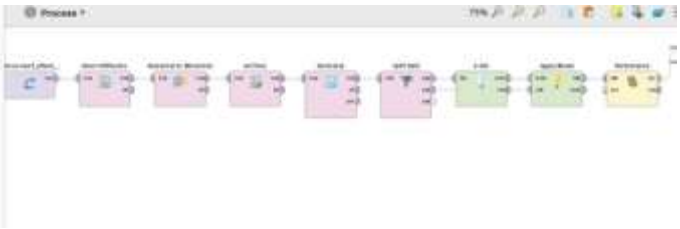
نتایج ارزیابی نشان داد که الگوریتم بیز ساده با دقت کلی ۴۷.۳۳ درصد، پایین‌ترین عملکرد را در میان مدل‌های بررسی‌شده در این پژوهش داشت. بررسی ماتریس درهم‌ریختگی نیز نشان داد این مدل در تشخیص هر دو کلاس (با ریسک و بدون ریسک) عملکرد ضعیف و نامتعادلی داشته و توانایی تفکیک‌پذیری محدودی نسبت به سایر الگوریتم‌ها از خود نشان داد.جدول(۵)

جدول۵- نتایج صحت و دقت به دست آمده از الگوریتم بیز ساده



۵. مدل‌سازی با الگوریتم نزدیک‌ترین همسایه

الگوریتم نزدیک‌ترین همسایه یک الگوریتم یادگیری ماشین نظارت‌شده متداول است که می‌تواند برای حل مسائل طبقه‌بندی و رگرسیون استفاده شود [۱۳]. در این الگوریتم، هر نمونه جدید بر اساس کلاس K نمونه نزدیک موجود در داده‌های آموزشی، دسته‌بندی می‌شود. نزدیک‌ترین همسایه به دلیل سادگی و قابلیت تفسیر بالا، یکی از الگوریتم‌های پرکاربرد در پیش‌بینی ریسک و مسائل پزشکی محسوب می‌شود.شکل(۴) و جدول (۶)



شکل ۴- محیط گرافیکی مدل الگوریتم نزدیک‌ترین همسایه

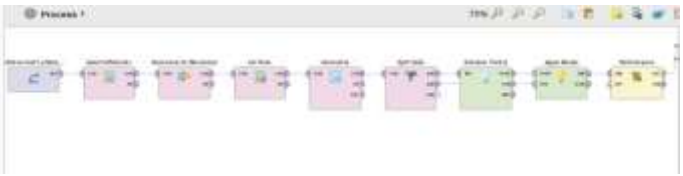
جدول ۶- نتایج صحت و دقت به‌دست‌آمده از الگوریتم نزدیک‌ترین همسایه

accuracy: 57.67%

	true false	true true	class precision
pred. false	442	237	65.10%
pred. true	144	77	34.84%
class recall	75.42%	24.52%	

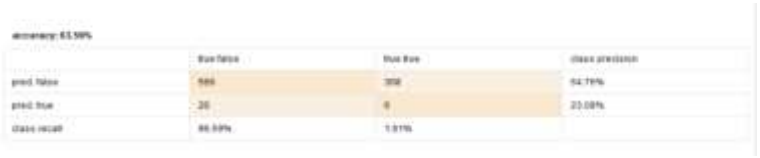
۶. مدل‌سازی با الگوریتم جنگل تصادفی

جنگل تصادفی یک الگوریتم یادگیری ماشین مبتنی بر تجمیع درخت‌های تصمیم است که با ساخت چندین درخت روی نمونه‌ها و ویژگی‌های تصادفی، پیش‌بینی نهایی را از طریق رأی‌گیری اکثریت تعیین می‌کند. در مقایسه با الگوریتم‌های سنتی، جنگل تصادفی دارای ویژگی‌های خوبی است [۱۴]؛ به همین دلیل در مسائل پیش‌بینی ریسک و داده‌کاوی پزشکی کاربرد گسترده‌ای دارد.شکل(۵) و جدول (۷)



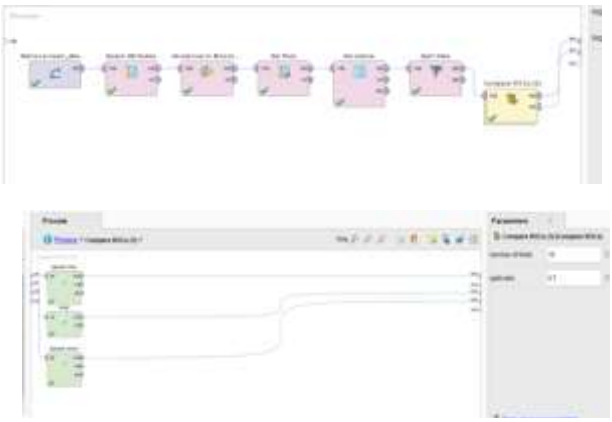
شکل ۵- محیط گرافیکی مدل الگوریتم جنگل تصادفی

جدول ۷- نتایج صحت و دقت به‌دست‌آمده از الگوریتم جنگل تصادفی



۷. مقایسه منحنی الگوریتم‌ها (ROC)

منحنی مشخصه عملکرد سیستم، ابزاری نموداری برای بررسی فرضیات پژوهش از آزمون‌های منحنی ROC است [۱۵]. هرچه مساحت زیر این نمودار به عدد ۱ نزدیک‌تر باشد، نشان‌دهنده دقت بالاتر مدل در تشخیص صحیح کلاس‌های مورد نظر است. سه الگوریتم درخت تصمیم، نزدیک‌ترین همسایه و جنگل تصادفی در صفحه مقایسه منحنی تنظیم شدند و با اعتبارسنجی ۱۰قسمتی (۷۰٪ آموزش و ۳۰٪ آزمون در هر تکرار)، عملکرد هر روش محاسبه و میانگین آن‌ها به‌عنوان نتیجه اجرا نمایش داده شد.شکل(۶)



شکل ۶- محیط گرافیکی مقایسه منحنی ۳ الگوریتم

نمودار نشان می‌دهد که مدل درخت تصمیم (خط آبی) با عملکردی متمایز، به بالاترین سطح کارایی دست یافته و تقریباً به حالت ایده‌آل (نزدیک‌شدن به گوشه بالا سمت چپ) نزدیک شده است که نشان‌دهنده دقت بسیار بالای این مدل در تفکیک کلاس‌های ریسک حمله قلبی است. در مقابل، مدل‌های جنگل تصادفی (خط قرمز) و نزدیک‌ترین همسایه (خط سبز) در جایگاه‌های پایین‌تری قرار دارند که نشان می‌دهد توان تفکیک‌کنندگی آن‌ها نسبت به درخت تصمیم کمتر است. اگرچه هر سه مدل عملکردی بهتر از حالت تصادفی (قطر نمودار) دارند، اما مدل جنگل تصادفی نسبت به نزدیک‌ترین همسایه برتری نسبی دارد. با این حال، با توجه به انحراف معیار و ناحیه سایه‌روشن اطراف نمودارها، درخت تصمیم با پایداری و دقت بیشتری کلاس‌های مثبت و منفی را شناسایی کرده است. در مجموع، این نمودار نشان می‌دهد

که برای داده‌های مورد مطالعه، مدل درخت تصمیم نسبت به دو مدل دیگر، قدرت پیش‌بینی و قابلیت اطمینان بالاتری ارائه می‌دهد.شکل (۷) و جدول (۸).

شکل ۷- نمودار مقایسه منحنی ۳ الگوریتم پیاده‌شده

جدول ۸- نتایج الگوریتم‌های مورد استفاده برای تشخیص ریسک حمله قلبی

نام روش	مقدار Accuracy
الگوریتم درخت تصمیم	٪۶۳.۷۸
الگوریتم جنگل تصادفی	٪۶۳.۵۶
الگوریتم K_NN	٪۵۷.۶۷

۸. نتیجه‌گیری

در این تحقیق، الگوریتم‌های درخت تصمیم، قوانین انجمنی، بیز ساده، نزدیک‌ترین همسایه و جنگل تصادفی بر روی داده‌های ریسک حمله قلبی پایگاه داده Kaggle پیاده‌سازی و با معیارهای دقت و ماتریس درهم‌ریختگی ارزیابی شدند. نتایج نشان داد مدل درخت تصمیم با دقت ۶۳.۷۸ درصد بهترین عملکرد را داشته است؛ با این حال، بررسی دقیق‌تر توزیع ویژگی‌ها نشان می‌دهد این دقت تفاوت چشمگیری با نرخ کلاس اکثریت (۶۵.۱ درصد نمونه‌های بدون ریسک) ندارد.

برای درک علت این محدودیت، توزیع ویژگی‌های کلیدی از جمله سن، کلسترول، فشار خون، دیابت و سابقه خانوادگی میان دو گروه دارای ریسک و بدون ریسک مقایسه شد. این بررسی نشان داد میانگین این ویژگی‌ها در دو گروه تقریباً یکسان است (برای مثال، میانگین سن ۵۳.۹ در برابر ۵۴.۰ سال و میانگین کلسترول ۲۶۱.۳ در برابر ۲۶۰.۰ در گروه‌های دارای ریسک و بدون ریسک). این هم‌پوشانی بالا میان توزیع ویژگی‌ها در دو کلاس، توضیح می‌دهد که چرا مدل‌های داده‌کاوی، صرف‌نظر از نوع الگوریتم، نتوانستند مرز تصمیم‌گیری قوی‌ای بین دو کلاس بیابند؛ به عبارت دیگر، محدودیت اصلی از ماهیت داده ناشی می‌شود، نه از انتخاب یا تنظیم الگوریتم‌ها.این یافته یک نکته مهم روش‌شناختی برای پژوهش‌های آتی در این حوزه دارد: پیش از اتکا به الگوریتم‌های پیچیده‌تر، ارزیابی قدرت تفکیک‌کنندگی ویژگی‌ها (مثلاً از

طریق آزمون‌های آماری یا تحلیل اهمیت ویژگی) گامی ضروری است. همچنین محدودیت‌های سخت‌افزاری و پردازشی، حجم داده‌های قابل استفاده و امکان مقایسه کامل تمامی روش‌ها را در این پژوهش محدود کرد. با این حال، نتایج حاصل توانست الگوهای کلی داده را استخراج و محدودیت‌های این مجموعه داده خاص را آشکار کند؛ یافته‌ای که می‌تواند پایه‌ای مناسب برای انتخاب دقیق‌تر مجموعه داده و طراحی ویژگی‌های بالینی معنادارتر در پژوهش‌های آتی پیش‌بینی ریسک حمله قلبی فراهم نماید.

۹. مراجع

1. Chaitrali SD, Sulabha SA. A data mining approach for prediction of heart disease using neural networks. International Journal of Computer Engineering and Technology 2012;3(3):30-40.

2. Premsmith J, Ketmaneechairat H. A Predictive Model for Heart Disease Detection Using Data Mining Techniques. Journal of Advances in Information Technology 2021;12(1):14-20.

3. Srinivas K, Rani BK, Govrdhan A. Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering 2010;2(02):250-5.

4. Alizadehsani R, Khosravi A, Roshanzamir M, Abdar M, Sarrafzadegan N, Shafie D, et al. Coronary artery disease detection using artificial intelligence techniques: A survey of trends, geographical differences and diagnostic features 1991-2020. Comput Biol Med 2021;128:104095.

5. Salve I, Kumar D, Borikar A. Heart Disease Prediction using Data Mining Techniques. Journal of Algebraic Statistic 2022;13(3):2250-2259.

6. Hung PS, Tseng YH, Tasi CF, Chen J, et al. Artificial Intelligence-Enabled ECG Algorithm for the Prediction and Localization of Angiography-Proven Coronary Artery Disease. IJERPH 2022;10(2).

7. Chen JIZ, Hengjinda P. Early prediction of coronary artery disease (CAD) by machine learning method: a comparative study. IRO Journals 2021;3(01):17-33.

8. Imamovic D, Babovic E, Bijedic N. Prediction of mortality in patients with cardiovascular disease using data mining methods. IS(INFOTEH) 2020:1-4.

9. Vindhya L, Beliray PA, Sravani CR, Divya DR. Prediction of Heart Disease Using Machine Learning Techniques. IJRESM 2020;3(8):325-326.

10. Choi Y, Boo Y. Comparing logistic regression models with alternative machine learning methods to predict the risk of drug intoxication mortality. IJERPH 2020;17(3):89.

11. Niaksu O. CRISP data mining methodology extension for medical domain. Baltic Journal of Modern Computing 2015;3(2):92.

12. Charbuty B, Abdulazeez A. Classification Based on Decision Tree Algorithm for Machine Learning. JASTT 2021;2(01):20-8.

13. Chirici G, Mura M, McInerney D, et al. A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications. Remote Sensing of Environment 2016;176:282-294.

14. Liu Y, Wang Y, Zhang J. New Machine Learning Algorithm: Random Forest. ICICA 2012:246-52.

15. Martínez-Camblor P, Pardo-Fernández JC. The Youden index in the generalized receiver operating characteristic curve context. Int J Biostat 2019;15(1).