دو فصلنامه علمی
فناوری اطلاعات و ارتباطات ایران

سال پانزدهم، شماره ۵۷ و۵۸ ، پاییز و زمستان ۱۴۰۲
صفحات:۱ الی۱۹

Design and implementation of a survival model for patients with melanoma based on data mining algorithms

Farinaz Sanaei*, Seyyed Abdolah Amin Mousavi**, Abbas Toloie Eshlaghi***, Ali Rajabzadeh Ghatari****

*Ph.D. student, Information Technology Management, Department of Information Technology Management, Faculty of Management and Economics, Islamic Azad University, Science and Research Branch, Tehran, Iran
**Assistant Professor, Department of Industrial Management, Faculty of Management, Islamic Azad University, Central Tehran Branch, Tehran, Iran
***Professor, Department of Information Technology Management, Faculty of Management and Economics, Islamic Azad University, Science and Research Branch, Tehran, Iran
****Professor, Industrial Management Department, Faculty of Management and Economics, Tarbiat Modares University, Tehran, Iran

Abstract

Background/Purpose: Among the most commonly diagnosed cancers, melanoma is the second leading cause of cancer-related death. A growing number of people are becoming victims of melanoma. Melanoma is also the most malignant and rare form of skin cancer. Advanced cases of the disease may cause death due to the spread of the disease to internal organs. The National Cancer Institute reported that approximately 99,780 people were diagnosed with melanoma in 2022, and approximately 7,650 died. Therefore, this study aims to develop an optimization algorithm for predicting melanoma patients' survival**.**

Methodology: This applied research was a descriptive-analytical and retrospective study. The study population included patients with melanoma cancer identified from the National Cancer Research Center at Shahid Beheshti University between 2008 and 2013, with a follow-up period of five years. An optimization model was selected for melanoma survival prognosis based on the evaluation metrics of data mining algorithms.

Findings: A neural network algorithm, a Naïve Bayes network, a Bayesian network, a combination of decision tree and Naïve Bayes network, logistic regression, J48, and ID3 were selected as the models used in the national database. Statistically, the studied neural network outperformed other selected algorithms in all evaluation metrics.

Conclusion: The results of the present study showed that the neural network with a value of 0.97 has optimal performance in terms of reliability. Therefore, the predictive model of melanoma survival showed a better performance both in terms of discrimination power and reliability. Therefore, this algorithm was proposed as a melanoma survival prediction model.

Keywords: data mining, prediction, melanoma, disease survival, neural network, decision tree.

فریناز صناعی،سید عبداله امین موسوی و...، دوفصلنامه فناوری اطلاعات و ارتباطات ایران، سال پانزدهم، شماره های ۵۷و۵۸، پاییز و زمستان۱۴۰۲، صفحه ۱ الی ۱۹

ارائه مدلی برای پیش‌بینی بقای بیماران مبتلابه ملانوم بر اساس الگوریتم‌های داده‌کاوی

فریناز صناعی*، سید عبداله امین موسوی**×، عباس طلوعی اشلقی***، علی رجب‌زاده قطری****

*دانشجوی دکتری مدیریت فناوری اطلاعات، گروه مدیریت فناوری اطلاعات، دانشکده مدیریت و اقتصاد، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران

** استادیار، گروه مدیریت صنعتی، دانشکده مدیریت، واحد تهران مرکز، دانشگاه آزاد اسلامی، تهران، ایران

استاد، گروه مدیریت فناوری اطلاعات، دانشکده مدیریت و اقتصاد، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران

استاد، گروه مدیریت صنعتی، دانشکده مدیریت و اقتصاد، دانشگاه تربیت مدرس، تهران، ایران

تاریخ دریافت:۱۴۰۱/۰۷/۱۴ تاریخ پذیرش:۱۴۰۱/۱۱/۲۸

نوع مقاله:پژوهشی

چکیده

مقدمه: ملانوم جزء شایع‌ترین سرطان تشخیصی و دومین علت مرگ ناشی از سرطان در میان افراد است. تعداد مبتلایان به آن در حال افزایش است. ملانوم، نادرترین و بدخیم‌ترین نوع سرطان پوست است.در شرایط پیشرفته توانایی انتشار به ارگان‌های داخلی را دارد و می‌تواند منجر به مرگ شود. طبق برآوردهای انجمن سرطان آمریکا برای ملانوم در ایالات‌متحده برای سال ۲۰۲۲ عبارت‌اند از: حدود ۹۹،۷۸۰ ز افراد مبتلابه ملانوم تشخیص داده شدند و حدود ۷،۶۵۰ نفر در اثر ملانوم جان خود را از دست می‌دهند. لذا هدف از این مطالعه، طراحی بهبود دقت الگوریتم برای پیش‌بینی بقای این بیماران است.

روش پژوهش: روش حاضر کاربردی، توصیفی- تحلیلی و گذشته‌نگر است. جامعه پژوهش را بیماران مبتلابه سرطان ملانوم پایگاه داده مرکز تحقیقات کشوری سرطان دانشگاه شهید بهشتی (۱۳۸۷ تا ۱۳۹۱) که تا ۵ سال مورد پیگیری قرارگرفته بودند، تشکیل داده است. مدل پیش‌بینی بقای ملانوم بر اساس شاخص‌های ارزیابی الگوریتم‌های داده‌کاوی انتخاب شد.

یافته‌ها: الگوریتم‌های شبکه عصبی، بیز ساده، شبکه بیزی، ترکیب درخت تصمیم‌گیری با بیز ساده، رگرسیون لجستیک، J48، ID3 به‌عنوان مدل‌های استفاده‌شده‌ی پایگاه داده کشور انتخاب شدند. عملکرد شبکه عصبی در همه شاخص‌های ارزیابی ازلحاظ آماری نسبت به سایر الگوریتم‌های منتخب بالاتر بود.

نتیجه‌گیری: نتایج مطالعه حاضر نشان داد که شبکه عصبی با مقدار ۰/۹۷ ازلحاظ دقت پیش‌بینی عملکرد بهینه دارد. بنابراین مدل پیش‌بینی کننده بقای ملانوم، هم ازلحاظ قدرت تمایز و هم ازلحاظ پایایی، عملکرد بهتری از خود نشان داد؛ بنابراین، این الگوریتم به‌عنوان مدل پیش‌بینی بقای ملانوم پیشنهاد شد.

واژگان کلیدی: داده‌کاوی، پیش‌بینی، ملانوم، بقای بیماری، شبکه عصبی، درخت تصمیم‌گیری

×نویسنده مسئول:سید عبداله امین موسوی ، a.mousavi@iauctb.ac.ir

۱. مقدمه

سرطان به گروهی از بیماری‌ها اطلاق می‌شود که با رشد کنترل نشده و گسترش سلول‌های غیرطبیعی مشخص می‌گردد. چنانچه این رشد غیرطبیعی مهار نشود به مرگ منجر می‌شود[۱]. ملانوم بدخیم کشنده و تهاجمی‌ترین نوع سرطان پوست است [۲].

خطرناک‌ترین ویژگی ملانوم این است که می‌تواند به‌طور گسترده‌ای در سراسر بدن از طریق عروق لنفاوی و رگ‌های خونی گسترش یابد؛ بنابراین تشخیص زودهنگام یک عامل کلیدی برای پیش‌آگهی این بیماری است[۳].

برآوردهای انجمن سرطان آمریکا برای ملانوم در ایالات‌متحده برای سال ۲۰۲۲ عبارت‌اند از: حدود ۹۹،۷۸۰ ملانوم جدید تشخیص داده می‌شود (حدود ۵۷۱۸۰ نفر در مردان و ۴۲۶۰۰ نفر در زنان). انتظار می‌رود حدود ۷۶۵۰ نفر در اثر ملانوم (حدود ۵۰۸۰ مرد و ۲۵۷۰ زن) جان خود را از دست دهند. نرخ ملانوم طی چند دهه گذشته به‌سرعت در حال افزایش است، اما این با افزایش سن متفاوت بوده است [۴].

درحال‌حاضر با توجه به کشنده بودن ملانوم در ایران، این بیماری عامل مهم مرگ‌ومیر و آسیب‌های اجتماعی خواهد بود، لذا بهداشت سرطان و تصمیم‌گیری مدیریت هوشمند از مهم‌ترین اولویت‌های نظام سلامت کشور است و پیشگیری، غربالگری و درمان اولیه سرطان از اولویت بالایی برخوردار است[۵]. این مهم نیازمند شناخت عوامل مؤثر بر بروز ملانوم بوده و اگر این شناخت در قالب مدل‌های داده‌کاوی ارائه گردد، از دقت وکارآیی بیشتری برخوردار خواهد بود. برقراری نظام ملی ثبت سرطان، به‌طوری‌که دربرگیرنده کلیه‌ی عوامل مؤثر بر بروز ملانوم باشد، می‌تواند در شناخت الگوی حاکم بر بروز ملانوم مؤثر بوده و چنانچه الگوی موردنظر به‌درستی تبیین شود می‌تواند نقش بسزایی در کاهش بروز ملانوم داشته باشد. پیشرفت فنّاوری آسیب‌شناسی پوستی نقش قابل‌توجهی در تشخیص و بالا بردن نرخ بقا در این بیماری داشته است. برای مثال، بیماران مبتلابه ملانوم کمتر یا برابر با ۰/۷۵ میلی‌متر ضخامت، یک پیش‌آگهی خوب داشته است[۶]؛ و بیشتر از ٪۹۳ از آن‌ها میزان بقای ۵ ساله خود رادارند [۷]. بنابراین، تشخیص زودهنگام برای کاهش مرگ‌ومیر مربوط به ملانوم بسیار مهم است[۸]. درپژوهش-حاضر،برای اولین بار در کشور، با اعمال الگوریتم‌های داده‌کاوی‌بر پایگاه‌داده‌ی‌کشور و مقایسه مدل‌های پیش‌بینی، دقیق‌ترین مدل بقا-ی ملانوم استفاده شد تا سیاست‌گذاران بهداشتی، کشورها در کنترل هر چه‌بهتر مرگ‌ومیر سرطان پوست افزایش کیفیت زندگی، امید به زندگی بیماران و مدیریت بهینه تصمیم‌گیری کسب‌وکار و صنعت سلامت یاری نماید[۹]. باپیشرفت‌هایی‌که‌درزمینه‌ی داده‌کاوی صورت گرفته است، ثابت‌شده که روش‌های به‌کاررفته، ازلحاظ قدرت و صحت پیش‌بینی بقا، عملکرد بهتری دارند[۱۰]. همچنین به علت وجود روابط غیرخطی بین متغیرهای پیش‌بینی کننده و متغیر هدف، روش‌های استفاده‌شده در حوزه داده‌کاوی این‌گونه روابط را با خطای کمتری تحلیل می‌کنند[۱۱].

بر اساس مطالعات پژوهشگران، پژوهش مشابه و گسترده‌ای در حوزه استفاده از روش‌های داده‌کاوی در پیش‌بینی بقای ملانوم در سطح کشور و دانشگاه‌های مختلف ازجمله دانشگاه‌های علوم پزشکی صورت نپذیرفته است و مطالعات انجام‌شده بر اساس داده‌های آماده سایر کشورها بوده است.

ازلحاظ نو بودن در جامعه موردبرر سی، لازم به ذکر ا ست تاکنون بر روی بانک اطلاعاتی مرکز تحقیقات سرطان کشوری دانشگاه شهید بهـ شتی مطالعه‌ای جهت برر سی پیش‌بینی بقای ملانوم انجام نـ شده است. در پژوهش‌های دیگر فقط گزارش صحت، حساسیت و ویژگی برای پیش‌بینی مدل‌ها استفاده‌شده است[۱۲]و[۱۳]. درحالی‌که در این پژوهش از مـقدار شـــاخص ترجیح تطبیق یافـته که از اثرات متعادل‌ســازی اســت، اســتفاده‌شــده اســت که جز بارزترین نکات قابل‌تمایز با پژوهش‌های دیگر است.

در این مطالعه بقای بیماری ملانوم ازلحاظ آماری و هوش مصنوعی موردبررسی قرارگرفته است. این دو رویکرد صرفاً جهت مقایسه نتایج روش‌های آماری و هوش مصنوعی نبوده است. بلکه در جهت افزایش آگاهی نسبت به موضوع بقای کلیه بیماران مبتلابه ملانوم با توجه به نقاط قوت هریک از روش‌ها بوده است. ازآنجایی‌که در این پژوهش تعداد متغیرهای بیشتر، در بازه گسترده‌تر بر روی کل بیماران مبتلابه ملانوم در سطح کشور بررسی‌شده است، لذانسبت‌به پژوهش‌های گذشته عملکرد موفقیت‌آمیزتری خواهد داشت.

۲. پیشینه تحقیق

تحقیقات ثابت کرده است که اکثر بیماران مبتلابه ملانوم مخصوصاً در مراحل پایین‌تر، حداقل تا ۵ سال بعد از تشخیص زنده مانده‌اند یا پاسخ آن‌ها به درمان حداقل تا آن فاصله زمانی مثبت بوده است. در صورت عدم درمان و مداوا، پیش‌بینی میزان بقای ۵ ساله بیمار از ٪۹۹ تا ٪۱۴ کاهش می‌یابد [۱۴].

در مطالعه‌ای که توسط دکتر اسماعیلی و همکارانش در سال ۱۳۸۶ در موردبررسی میزان بقا ۵ ساله بیماران مبتلابه ملانوم در ایران برحسب سن، جنس، محل آناتومیکی و نوع کلینیکی تومور از ۷۸۱ بیمار بر روی نمونه‌های ارسال‌شده به مرکز سرطان پوست تهران بین سال‌های ۷۹-۷۵ انجام شد نتیجه‌ی زیر حاصل شد:

میزان بقا زنان بهتر از مردان (۳۸/۵٪ در مقایسه با ۲۳/۵٪) بود.

بهترین میزان بقا مربوط به گروه سن ۵۰-۴۰ سال در هر دو جنس بود (۶۳/۶٪).

بهترین میزان بقا در رابطه با مکان ملانوم مربوط به درگیری صورت بود (۶۳/۶٪).

بهترین میزان بقا در رابطه با نوع کلینکال تومور، مربوط به نوع ملانوم لنتیگو بود (۷۸٪).

درنهایت میزان بقا ۵ ساله بیماران در ایران ۲۸/۵٪ تخمین زده‌شده که این درصد پایین‌تر از کشورهای دیگر بود [۱۵]. در مقام مقایسه، در این تحقیق که در آن ۷۸۱ بیمار و تعداد ۴ متغیر بررسی‌شده است تحقیق جاری با بررسی تعداد بیماران ۴۱۱۸ نفر بررسی‌شده که حدود ۳۳۳۷ بیمار از تحقیق مذکور بیشتر بوده، با توجه به این نکته که در این پژوهش ۱۰ متغیر موردبررسی قرار گرفت که از تحقیق فوق ۶ متغیر بیشتر مورد تحلیل و بررسی قرارگرفته، مطالعه گسترده‌تری انجام‌شده است. دقت شبکه عصبی ۹۷٪ برآورد بقای بیماران مبتلابه ملانوم را بعد از پنج سال داشته است که در مقایسه با عملکرد پژوهش مطرح‌شده میزان صحت و دقت آن سنجیده نشده است. ضمناً ذکر نشده که سلول‌های ملانوسیتی گیرنده‌ای برای هورمون‌های جنسی داشته باشند پس قاعدتاً نباید تفاوتی بین بقاء مردان و زنان وجود داشته باشد. از سیستم‌های هوشمند و داده‌کاوی نیز استفاده‌نشده است.

در سال ۱۳۸۷ عابدی و مقدس زاده درباره تعیین تابع بقا ارزیابی روش‌های مختلف رگرسیون کاکس با رگرسیون چند متغیره، همراه با جدول طول عمر در سرطان پوست، ۱۹۹بیمار مبتلابه سرطان پوست را بررسی کرده، بعد از استخراج اطلاعات موردنظر از پرونده‌های پزشکی، وضعیت سلامت بیماران به‌صورت تلفنی،پیگیری شده و پیشامد نهایی موردبررسی قرار گرفت. نتایج حاصله عبارت‌اند از:توابع بقا برآورد شده توسط روش جدول طول عمر و حد حاصل‌ضرب نشان می‌دهد که نرخ بقا یک‌ساله برابر ۰/۹۴است و نرخ مخاطره در سال اول برابر ۰/۰۶۳ است. مدل مخاطرات متناسب کاکس که بر داده‌های تحقیق برازش یافته بصورت زیر است:

$$h(t)=h_{\circ}(t)\exp(2.07LT_2+1.77kind_RT_2) \quad (1)$$

در مقایسه با پژوهش ما تعداد بیماران بررسی‌شده بسیار کمتر بوده و بررسی روی سرطان پوست غیرملانومی بوده است و در این تحقیق ۱۹۹ بیمار و تعداد ۵ متغیر بررسی‌شده است. تحقیق جاری با بررسی ۴۱۱۸تعداد بیماران مبتلا شده و ۱۰ متغیر (که تأثیر بیشتری درزمینه‌ی بقا دارند)، مطالعه گسترده‌تری انجام داده است. ضمناً در این مطالعه از روش‌های معمول پرکاربرد آماری مانند تحلیل کاکس استفاده کرده‌اند و طراحی مدل پیش‌بینی نداشته‌اند، درحالی‌که در مطالعه ما از روش‌های به‌روز و هوش مصنوعی استفاده‌شده است و ترکیب سیستم‌های هوشمند و نرم‌افزارهای آماری دقت و صحت بالاتری نسبت به سیستم‌های آماری محض دارد.

در مطالعه ای‌که توسط (کرول و همکارانش،۲۰۰۶) و همکارانش در سال ۲۰۰۶ در موردبررسی میزان بقای بیماران مبتلابه ملانوم پوستی بر روی۴۷۵بیمار با توجه به اطلاعات ثبت‌شده در مرکز انکولوژی در کشور ترکیه در بین سال‌های ۲۰۰۳-۱۹۹۱ انجام شد نتایج کلی آن به شرح زیر است:

میانگین میزان بقاء ۵ ساله‌ی بیماران در StageII, Stage III به ترتیب ۶۳/۶٪ و ۳۶/۶٪ بیان شد.

میانگین میزان بقای بیماران با ملانوم متاستاتیک ۹/۹ ماه و میانگین بقای یک‌ساله‌ی آن‌ها ۳۲٪ بود. میانگین میزان بقاء ۵ ساله ۵۰/۵٪ گزارش شد. در مقایسه با تحقیق جاری با بررسی ۴۱۱۸ تعداد بیماران مبتلا شده و ۱۰ متغیر، مطالعه وسیع‌تر و دقیق‌تری انجام‌شده است. هیرابینیش و همکارانش [۱۶] در تحقیقی، راجع به پیش‌بینی سرطان پوست با استفاده از داده‌کاوی و تکنیک‌های آن مطالعه کردند. در این بررسی، سرطان پوست در عصر حاضر، به علت وجود تعداد زیاد تلفات حاصل از سرطان پوست، از تکنیک‌های مختلفی در استخراج داده‌ها برای پیش‌بینی و درمان آن نیز استفاده کردند. در این مقاله الگوریتم‌های اپریوری، درخت تصمیم‌گیری، بیز ساده، شبکه‌های عصبی و الگوریتم‌های متن‌کاوی با ابزارهای منبع باز برای پیش‌بینی سرطان پوست، موردبحث و بررسی قراردادند .همچنین نتایج این پژوهش نشان داد که صحت ۹۳,۳۶LFRS٪ و VF15 ۹۶,۲٪ و شبکه عصبی که ۹۷٪ است. دراین‌پژوهش از دیتاست

MAMEL[1] استفاده‌شده و مقاله به‌صورت مروری مطرح گردیده و جزئیات کار و سایر ملاک‌های ارزیابی انتخاب الگوریتم‌های دسته‌بندی که برای بقای بیماران حائز اهمیت است، بررسی نشده است. تعداد بیماران، متغیرهای بررسی‌شده نامعلوم است. همچنین نیاز به پیاده‌سازی این ابزار و الگوریتم بر روی پلت فرم متمرکز با داده‌های مشابه وجود دارد تا نتایج کارآمدتر شود؛ که در مقایسه با پژوهش ما از الگوریتم‌های بسیار کمتری استفاده‌شده است.

جدول (۱) گویای مقایسه پژوهش‌ها است.

جدول ۱.بررسی پژوهش‌های پیشین با تحقیق موردنظر

پژوهشگران (سال)	پیش بینی بقا	مقایسه روش پژوهش ها	تعداد بیماران،متغیرها پژوهش مقایسه شده	تعداد بیماران،متغیرها پژوهش انجام شده
کرول، ۲۰۰۶	✓	مطالعه از روش‌های معمول پرکاربرد آماری مانند تحلیل کاکس استفاده کرده‌اند. -از سیستم‌های هوشمند و داده‌کاوی استفاده‌نشده است.	تعداد بیماران مورد بررسی=۴۷۵ تعداد متغیرها =۵ دقت وصحت وسایر پارامترهای ارزیابی لحظ نشده است.	تعداد بیماران مورد بررسی =۴۱۱۸ تعداد متغیرها =۱۰ دقت شبکه عصبی=۹۷٪
اسماعیلی، ۲۰۰۷	✓	مطالعه از روش‌های معمول پرکاربرد آماری مانند تحلیل کاکس استفاده کرده‌اند -از سیستم‌های هوشمند و داده‌کاوی استفاده‌نشده است.	تعداد بیماران مورد بررسی=۷۸۱ تعداد متغیرها =۴۰ دقت وصحت وسایر پارامترهای ارزیابی لحظ نشده است.	تعداد بیماران مورد بررسی =۴۱۱۸ تعداد متغیرها =۱۰ دقت شبکه عصبی=۹۷٪
عایدی، ۲۰۰۶	✓	از روش‌های مختلف رگرسیون کاکس با رگرسیون چند متغیره، همراه با جدول طول عمر استفاده کرده اند. -از سیستم‌های هوشمند و داده‌کاوی استفاده‌نشده است.	تعداد بیماران مورد بررسی=۱۹۹ تعداد متغیر =۵ دقت وصحت وسایر پارامترهای ارزیابی لحظ نشده است.	تعداد بیماران مورد بررسی =۴۱۱۸ تعداد متغیرها =۱۰ دقت شبکه عصبی=۹۷٪
هیرا بیتیش، ۲۰۲۰	✓	مطالعه از الگوریتم‌های اپریوری، درخت تصمیم‌گیری، بیز ساده، شبکه‌های عصبی و الگوریتم‌های متن‌کاوی برای پیش‌بینی سرطان پوست استفاده کرده اند که صحت LFRS ۹۳.۳۶٪ و VF۱۵ ۹۶.۳٪ و شبکه عصبی که ۹۸٪ است.	تعداد بیماران مورد بررسی نامشخص تعداد متغیرهانامشخص صحت شبکه عصبی = ۹۸٪	تعداد بیماران مورد بررسی =۴۱۱۸ تعداد متغیرها =۱۰ دقت شبکه عصبی=۹۷٪

۳. روش‌شناسی پژوهش

نوع پژوهش: هدف پژوهش حاضر ازنظر ماهیت کاربردی و ازنظر هدف به روش توصیفی- تحلیلی گذشته‌نگر و ازنظر زمان به‌صورت مقطعی انجام شد.

مراحل انجام پژوهش: با توجه با اهداف مراحل انجام پژوهش به دو مرحله مستقل تقسیم می‌شود:

۱,۳ مرحله اول: استفاده از الگوریتم‌های داده‌کاوی در پایگاه داده بومی (داده‌های سرطانی کل کشور)

۱,۱,۳ جامعه پژوهش: بیماران مبتلابه ملانوم بین سال‌های ۱۳۸۷ تا ۱۳۹۱ که تا ۱۳۹۱ در کل کشور پیگیری شده‌اند.

۳,۱,۲ نمونه پژوهش (شامل حجم نمونه و روش نمونه‌گیری)

تعداد کل بیماران موردبررسی و تحقیق ۷۶۹۰ نفر است. بیمارانی که بعد از تشخیص ملانوم در آن‌ها، کمتر از ۵ سال پیگیری شده‌اند، حذف شدند. همچنین بیمارانی که قبل از رسیدن به پنجمین سال پیگیری بیماری فوت کرده و علت فوت آن‌ها غیر از ملانوم بود نیز حذف شدند؛ که درنهایت به ۴۱۱۸ نفر رسیدند.

۳,۱,۳ ابزار گردآوری داده‌ها

داده‌های پرونده‌ها بر اساس فرم جمع‌آوری داده استخراج شد. در ایجاد این فرم که دربرگیرنده متغیرهای بااهمیت در پیش‌بینی بقای ملانوم و توضیحات مربوط به مقادیر متغیرها است، از مطالعات، پژوهش‌ها و تحقیقات مربوط به‌پیش‌آگهی‌ملانوم‌که توسط پژوهشگر و نهایتاً با مشورت و نظر تخصصی ۵ پزشک متخصص پوست، ایجاد شد. فرم جمع‌آوری داده در جدول (۲) آورده شده است که طبق توضیحات فوق متغیرهای دخیل در بقای ملانوم، بر اساس مطالعه مقالات و راهنماهای بالینی مربوط به ملانوم شناسایی شد و در فرم زیر قرار گرفت.

جدول ۲. چک‌لیست داده‌های ملانوم برای استخراج متغیرهای موردنیاز از پرونده‌های بیماران سرطان کل کشور

نام متغیر	مقادیر متغیر
۱-سن	بر اساس سن یادداشت شده در پرونده
۲-ضخامت تومور	بر اساس ضخامت تومور اندازه‌گیری شده و درج در پرونده I: less than or equal to 1 mm II: 1.01-2 mm III: 2.01-4 mm IV: greather than or equal 4 >= mm
۳-جنس	مرد=۱ زن=۲

[1] Modernizing Medicine to reach these goals

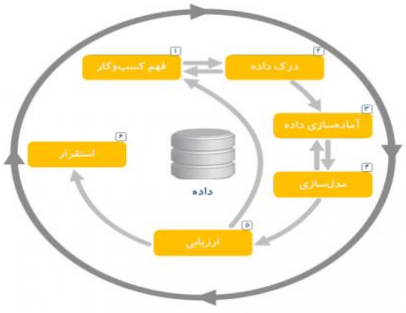
۴-ارتشاح لنفوسیتی	۱=خفیف ۲=متوسط ۳=شدید
۵-وضعیت تأهل	۱=مجرد ۲- متأهل ۳=دوجنسی ۹=نامشخص
۶-درجه بیماری	Grade III=۳ Grade II=۲ Grade I=۱ NOS=۹ Grade IIII=۴
۷-محل تومور	Skin of lip=۰ (پوست لب) Eyelid=۱ (پلک چشم) External ear=۲ (گوش خارجی) Skin of other face=۳ (بقیه جاهای صورت) Skin of scalp and neck=۴ (سرو گردن) Skin of trunk=۵ (تنه) Skin of upper limb and =۶ shoulder(اندام فوقانی و شانه) Skin of lower limb and hip=۷ (اندام تحتانی و لگن) Overlapping lesion of =۸ skin(بینابین نواحی) NOS=۹ (تشخیص داده نشده)
۸-مرحله بیماری (سرطان)	IA=۱ (۱) IB=۲ (۱) IIA=۳ (۲) IIB=۴ (۲) IIC= ۵ (۲) IIIA=۶ (۳) IIIB=۷ (۳) IIIC=۸ (۳) IV=۹ (۴)
۹-رفتار	۰ = خوش‌خیم ۱ = نامشخص ازلحاظ خوش‌خیمی یا بدخیمی ۲ = کارسینوم درجا ۳ = بدخیم اولیه ۶ = بدخیم ثانویه (متاستاتیک) ۹ = نامشخص ازلحاظ اولیه یا ثانویه
۱۰-اشعه درمانی	۱=بله ۲=خیر

۴،۱،۳ روش گردآوری داده‌ها

محقق به‌صورت حضوری به محیط پژوهش مراجعه کرده است. فرم مذکور بر اساس مطالعه متون تحقیقات، پژوهش‌های بقای بیماران مبتلابه ملانوم و نظرخواهی از متخصصین جراحی پوست ایجادشده است نظرخواهی از متخصصان جراحی پوست به روش مصاحبه و به‌صورت جداگانه انجام‌شده، سپس بر اساس فرم استخراج داده، متغیرهای موردنیاز، از پرونده‌هایی که شرایط قرار گرفتن در - نمونه‌گیری را دارا بودند، استخراج شد. برای رعایت مسائل اخلاقی، برخی مشخصات بیماران (متغیرها) محرمانه ماند. خصوصیات متغیرها با توجه به کدگذاری بین‌المللی سرطان‌ها[1] (ICD-O) (که این طبقه‌بندی سرطان‌ها بر اساس ICD-O طبق توصیه (سازمان بهداشت جهانی)[2] WHO صورت گرفته است تا بتوان در مقیاس‌های منطقه‌ای، ملی، بین‌المللی اطلاعات را مقایسه نمود.) کدگذاری شده بود به همراه سایر اطلاعات دموگرافیک بیماران (سن، جنس، محل سکونت) موردبررسی قرار گرفت.

۲،۳ روش تحلیل داده‌ها

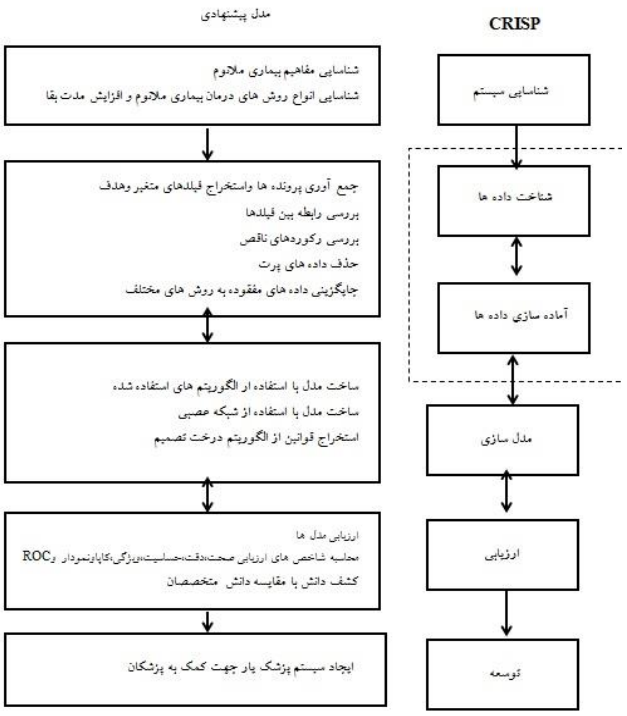
یکی از روش‌های بسیار قوی برای پیاده‌سازی و اجرای پروژه‌های داده‌کاوی متدلوژی CRISP است [۱۷]. در این پژوهش مدل پیشنهادی بر اساس CRISP که شامل پنج گام که در شکل (۱) به نمایش گذاشته‌شده است و به شرح ذیل انجام می‌شود [۱۸]. هریک از این فازها خود شامل زیر بخش‌هایی می‌شوند. حرکت روبه‌جلو و عقب بین فازهای مختلف نیاز است [۱۹]، زیرا ورودی هر فاز به خروجی فاز مرحله قبل وابسته است [۲۰]. در شکل (۲) گام‌های روش ذکرشده با تمام مراحل تحقیق منطبق شده است.



شکل ۱. مراحل اصلی داده‌کاوی در مدل crisp

[1] International Classification of Diseases for Oncology

[2] World Health Organization



شکل ۲. گام‌های روش crisp و مدل پیشنهادی

۱،۳،۲ شناخت سیستم

در این مرحله به شناخت سیستم موردنظر پرداخته می‌شود و سپس اهداف موردنظر و عوامل موفقیت کلیدی سیستم تعیین و بازنگری می‌گردد .طبق نظر متخصصان پوست، با توجه به رشد روزافزون سرطان‌های پوست، هزینه‌های سرسام‌آور درمان این بیماری سرطان‌های دیگر و عوارض شدیدی که روی اعضای حیاتی بدن در درازمدت می‌گذارد [۲۱]، بررسی داده‌های جمع‌آوری شده در رابطه با این بیماری در تشخیص زودرس روش درمان بیماران جدید و مدیریت سیستم تصمیم‌یار پزشکی می‌تواند مفید باشد. بیماران جدید می‌توانند تا حد ممکن از توصیه‌های پزشکی تجویزشده متناسب با بیماران دسته‌ای که در آن قرارگرفته‌اند، بهره ببرند.

۲،۲،۳ شناخت داده‌ها و آماده‌سازی آن‌ها

در این مرحله به جمع‌آوری داده‌های اولیه، توصیف داده‌ها، بازرسی و بررسی داده‌ها و اعتبار سنجی کیفیت داده‌ها پرداخته‌شده است . مطالعه‌ی حاضر از نوع توصیفی -مقطعی بوده و مجموعه داده‌های آن متعلق به مرکز تحقیقات کشوری سرطان وابسته به دانشگاه شهید بهشتی بیمارستان شهدای تجریش تهران است که بیماران مبتلابه ملانوم بین سال‌های ۱۳۸۷ تا ۱۳۹۱ در کل کشور پیگیری شدند.
شناخت و پیش‌پردازش داده‌ها (آماده‌سازی): مجموعه داده ملانوم که در این مرحله از پژوهش استفاده شد،۵ فایل اکسل جداگانه که هرکدام متعلق به یک سال بود (از سال ۱۳۸۷ تا سال ۱۳۹۱) و در هر فایل کلیه بیماران مبتلابه انواع سرطان پوست در آن قرار داشت که طبق استاندارد ICD-O افراد مبتلابه ملانوم در هر فایل جدا و انتخاب‌شده و همگی دریک فایل اکسل[۱] تجمیع شدند. داده‌های پژوهش مربوط به مرکز تحقیقات کشوری سرطان وابسته به دانشگاه شهید بهشتی بیمارستان شهدای تجریش تهران بود که بعد از گردآوری داده توسط پژوهشگران ایجاد شد. این فایل حاوی ۷۶۹۰ رکورد و ۱۷ متغیر بود. برخی از متغیرها مانند شماره پرونده، نام و نام خانوادگی، نام پدر، آدرس بیماران، کد پستی، نام مرکز، شماره تلفن ثابت و موبایل بیماران است که به علت رعایت اصل محرمانگی، این متغیرها از مجموعه داده حذف گردید.
افزایش گرداوری و ثبت روزافزون داده‌های پیچیده در دنیای پزشکی، دلیل استفاده از فرایند داده‌کاوی از پایگاه داده است [۲۲]. به علت وجود داده‌های ازدست‌رفته و تکراری و همچنین متغیرهایی با اسامی نامفهوم و نیاز به مدیریت دقیق این چالش‌ها، مراحل شناخت و آماده‌سازی داده، مهم‌ترین مراحل در داده‌کاوی بوده و بیشترین زمان صرف شده در استخراج دانش از پایگاه داده‌ها به این مراحل اولیه معطوف می‌شوند [۲۳]. درواقع مرحله دوم و سوم از متدلوژی CRISP در این گام بررسی و اجرا می‌گردد.
برای بررسی دقیق‌تر و جزئی‌تر و همچنین اجرای عملیات پیش‌پردازش داده‌ها از قبیل مدیریت مقادیر ازدست‌رفته و پرت، مجموعه داده با محیط نرم‌افزارSPSS وارد گردید. جدول ۳ و ۴ و ۵ توزیع متغیرهای پیش‌بینی کننده و متغیرهای پیش‌بینی کننده پیوسته و متغیر وابسته را نشان می‌دهند.

[۱] Excel

جدول ۳. متغیرهای پیش‌بینی کننده در مجموعه داده ملانوم مرکز تحقیقات کشوری سرطان

اسامی متغیر گروهی	تعداد مقادیر منحصربه‌فرد
جنس	۲
ارتشاح لنفوسیتی	۳
وضعیت تأهل	۴
درجه بیماری	۵
محل تومور	۱۰
مرحله سرطان	۹
رفتار	۶
اشعه درمانی	۲

جدول ۴. متغیرهای پیوسته پیش‌بینی کننده در مجموعه داده ملانوم مرکز تحقیقات کشوری سرطان

اسامی متغیر پیوسته	میانگین	انحراف معیار	بازه (محدوده)
سن	۵۹/۲۵	۱۹/۳۳	۱-۱۰۱
ضخامت تومور	۲/۳۹	۲/۲۴	۰/۰۱-۹/۰۵
تعداد میتوز	۲/۰۶	۱/۹۹۸	۱-۱۳

جدول ۵. توزیع متغیر وابسته پایگاه داده کشوری

دسته	تعداد رکورد	درصد
۰: عدم بقا	۱۴۰۴	۳۴/۱
۱: بقا	۲۷۱۴	۶۵/۹
جمع کل	۴۱۱۸	۱۰۰

متغیر وابسته، متغیری از نوع دوتایی بود که صفر و یک به ترتیب عدم بقا و بقا را نشان می‌دادند. برای تعیین متغیر وابسته، از مجموعه داده از نرم‌افزار SPSSبه نرم‌افزار اکسل وارد شد. متغیرهای «کدگذاری مجدد زمان بقا»، «کدگذاری مجدد وضعیت حیاتی» و «علت مرگ» در تعیین متغیر وابسته دخیل بودند. متغیر «کدگذاری مجدد زمان بقا» از چهار کاراکتر تشکیل می‌شود که دو نویسه اول نشانگر سال پی گیری و دو نویسه بعدی نشانگر ماه پی گیری بودند. ابتدا هر یک از اعداد در ستونی جداگانه تقسیم‌بر صد شدند و سپس با به‌کارگیری دستور ذیل، متغیر وابسته ایجاد شد. شکل (۳)گویای قطعه کد نوشته‌شده است.

```
1-If STR[1] >=5 year and VSR[2]=alive then
2-Record=1
3-Else if STR<5 year and COD[3]=melanom then
4-Record=0
5-Else
6-Ignore the Record
7-End if
```

شکل ۳. نمایش تکه کد برای بررسی بقا

رکوردهایی که در هیچ‌یک از گروه‌های بقا و عدم بقا قرار نگرفته بودند، حذف شدند و تعداد رکوردها به ۴۱۱۸ رکورد رسید. اگر میزان مقادیر ازدست‌رفته مجموعه داده‌ها کمتر از ۱ در صد باشد، در فرآیند استخراج دانش خللی ایجاد نمی‌کند، بین ۱ تا ۵ درصد قابل مدیریت، بین ۵ تا ۱۵ درصد مستلزم اعمال روش‌های پیچیده برای مدیریت و بیشتر از ۱۵ درصد ممکن است به‌صورت جدی بر هر نوع تفسیردانشی تأثیرگذار باشد [۲۴]؛ بنابراین برای مدیریت این مقادیر، به‌جای حذف آن‌ها که باعث از دست رفتن اطلاعات ارزشمندی می‌شود از خط‌مشی‌های زیر استفاده شد. البته لازم به ذکر است که برای اعمال خط‌مشی‌های ذیل باید مقادیر ازدست‌رفته به‌صورت کاملاً تصادفی رخ‌داده باشند [۲۵].

خط‌مشی اول: اگر یک متغیر در بیش از ۵۰ درصد رکوردها با مقادیر ازدست‌رفته مواجه باشد، در مجموعه داده موردمطالعه متغیر اشعه درمانی مقدار ازدست‌رفته‌اش بالای ۵۰ درصد بود که نهایتاً منجر به حذف این متغیر گردید.

خط: مشی دوم: اگر یک متغیر در کمتر از یک درصد رکوردها با مقادیر ازدست‌رفته مواجه باشد، در آن صورت اگر از نوع عددی

[۱] Survival Time Record (STR)

[۲] Vital Status Record(VSR)

[۳] Cause ofDeath (COD)

باشد میانگین مقادیر موجود در آن متغیر جایگزین مقادیر ازدست‌رفته می‌شود و اگر متغیر از نوع اسمی یا ترتیبی باشد در این صورت مد مقادیر موجود در آن متغیر جایگزین مقادیر ازدست‌رفته می‌شود. در مجموعه موردمطالعه برای متغیرهای «وضعیت تأهل» و «تعداد میتوز» از این خط‌مشی استفاده شد.

خط‌مشی سوم: اگر متغیر در کم‌تر از ده درصد از رکوردها با مقادیر ازدست‌رفته مواجه باشد، در آن صورت با توجه به نوع متغیر، مقدار میانگین یا مد در مقادیر موجود آن متغیر در هر کلاس محاسبه‌شده و با توجه به کلاس رکوردی که در متغیر موجود با مقدار ازدست‌رفته مواجه است، مقدار ازدست‌رفته مواجه است، مقدار میانگین یا مد مربوط به همان کلاس جایگزین می‌شود [۲۴]. مقادیر ازدست‌رفته متغیر «مرحله سرطان» از این خط‌مشی به دست آمد.

خط‌مشی چهارم: برای متغیرهایی که بیش از ۱۰ درصد مقادیر ازدست‌رفته داشتند از الگوریتم‌های موجود در طبقه‌بندی برای برآورد مقادیر ازدست‌رفته در آن متغیر و پر کردن آن استفاده شد. در این روش متغیری که دارای اعداد ازدست‌رفته است ابتدا به‌صورت فیلد هدف یا کلاس وارد مطالعه شده و سایر متغیرها به‌صورت ورودی تعریف می‌شوند. پس از به دست آوردن دقت بالا در ایجاد مدل پیش‌بینی، الگوریتم شبیه‌ترین رکورد موجود در پایگاه داده که دارای متغیر ازدست‌رفته نیستند را جایگزین مقادیر ازدست‌رفته متغیر هدف می‌کند. یکی از نکات مثبت این خط‌مشی به‌این‌علت است که به خاطر وجود همبستگی‌هایی که معمولاً بین متغیرها وجود دارد، روش مذکور از این همبستگی‌ها برای ایجاد مدل پیش‌بینی مقادیر ازدست‌رفته استفاده می‌کند. مدل پیش‌بینی دسته‌بندی برای برآورد مقادیر ازدست‌رفته متغیرهای کیفی و مدل پیش‌بینی رگرسیون برای برآورد مقادیر ازدست‌رفته متغیرهای کمی کاربرد دارند مقادیر ازدست‌رفته متغیرهای ارتشاح لنفوسیتی، درجه بیماری، محل تومور، ضخامت تومور با این خط‌مشی جایگذاری شدند. در جدول (۶) کلیه متغیرها و مقادیر ازدست‌رفته به تفکیک و درصد به نمایش درآمده است.

جدول ۶. مقادیر ازدست‌رفته متغیرهای پیش‌بینی کننده پایگاه داده کشوری

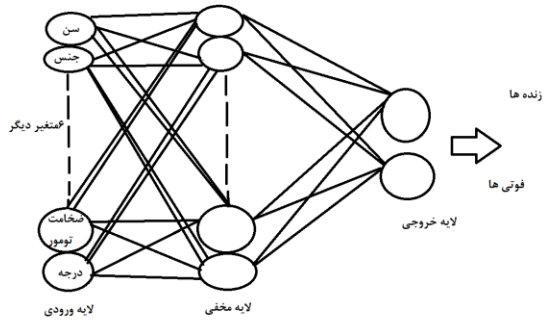
متغیرها	تعداد مقادیر ازدست‌رفته	درصد
جنس	۳۷۱	۹
ارتشاح لنفوسیتی	۱۶۴۷	۴۰
وضعیت تأهل	۲۱	۰/۵
درجه بیماری	۷۱۳	۱۷/۳
محل تومور	۹۸۰	۲۳/۸
مرحله سرطان	۹۱	۲/۲
سن	۰	۰
ضخامت تومور	۶۴۳	۱۵/۶
تعداد میتوز	۲۵	۰/۶
اشعه درمانی	۲۳۰۶	۵۶
رفتار	۸۷۳	۲۱/۲

۳،۲،۳ مدل‌سازی

برای ارائه مدل پیش‌بینی کننده بقای ملانوم با استفاده از مجموعه داده سرطان کل کشور، الگوریتم‌های شبکه عصبی، بیزی ساده، شبکه بیزی، ترکیب درخت تصمیم‌گیری با بیز ساده، الگوریتم رگرسیون لجستیک و الگوریتم ID3 و الگوریتم درخت تصمیم‌گیری j48 بکار گرفته شد. کلیه الگوریتم‌های فوق‌الذکر با استفاده از نرم‌افزار وکا اجرا شدند. دلیل استفاده از این نرم‌افزار،برخورداربودن‌وکا از قابلیت‌های متنوع پیش‌پردازش داده‌ها، در برگرفتن تمام الگوریتم‌های حوزه هوش مصنوعی و همچنین متن‌باز بودن این نرم‌افزار بود. همچنین نرم‌افزار وکا به‌عنوان یک نرم‌افزار دانشگاهی و آکادمیک شناخته می‌شود.

برای ایجاد شبکه عصبی از ساختار معمول آن‌که پرسپترون چندلایه است استفاده‌شده است. ورودی‌های شبکه عصبی را دخل در بقای سرطان پوست و خروجی آن را متغیر هدف یا بقای بیمار تشکیل دادند [۲۶]. بنابراین، شبکه عصبی در پژوهش حاضر از۱۰ ورودی (متغیرهای پیش‌بینی کننده) و یک خروجی (مقادیر متغیر هدف) تشکیل‌شده است. در اکثر پژوهش‌ها بر توانایی شبکه‌های دولایه تأکید شده است. برای تعیین لایه‌های مخفی به پژوهش‌هایی که درزمینه داده‌کاوی در حوزه بالینی انجام‌شده بوداستنادگردید [۲۷]. در اکثر پژوهش‌ها بر توانایی شبکه‌های دولایه تأکید شده است البته

لازم به ذکر است که از این دولایه، لابه خروجی و دیگری لایه مخفی است. ساختار شبکه عصبی پژوهش حاضر در شکل (۴) آورده شده است.



شکل ۴. ساختار کلی شبکه عصبی دسته‌بندی بقای مبتلایان به ملانوم

۴،۲،۳ ارزیابی

پس از مدل‌سازی می‌باید به ارزیابی نتایج حاصل از مدل پرداخت. در اغلب موارد، مقایسه الگوریتم‌های داده‌کاوی در شناسایی این‌که کدام‌یک از آن‌ها در تحلیل مجموعه داده کاوش شده، عملکرد بهتری داشته است، امری ضروری است. بعضی مواقع نیز، اندازه‌گیری عملکرد یک الگوریتم با پارامترها و تنظیم‌های مختلف، مدنظر است [۲۸].

نتایج ارزیابی باعث بهبود مدل شده و مدل را قابل‌استفاده می‌کند. در این مرحله اعتبار مدل بررسی می‌شود، زیرا روش های مدل‌سازی مختلف در شرایط مختلف، رفتارهای متفاوتی از خود نشان می‌دهند. برای مقایسه این روش‌ها با یکدیگر از شاخص‌های متفاوتی استفاده می‌کنند. در اغلب موارد، مقایسه الگوریتم‌های داده‌کاوی در شناسایی این‌که کدام‌یک از آن‌ها در تحلیل مجموعه داده کاوش شده، عملکرد بهتری داشته است، امری ضروری است. بعضی مواقع نیز، اندازه‌گیری عملکرد یک الگوریتم با پارامترها و تنظیم‌های مختلف، مدنظر است یک روش سنجش عملکرد الگوریتم‌ها، اندازه‌گیری صحت آن‌ها ست. برای اندازه‌گیری صحت از فرمول زیر استفاده می‌شود.

$$\text{صحت} = \frac{TP+TN}{TP+TN+FP+EN} = \frac{\text{منفی های صحیح + مثبت های صحیح}}{\text{منفی های اشتباه + مثبت های اشتباه+ منفی های صحیح + مثبت های صحیح}}$$

فرمول فوق، جمع کل بیمارانی که برچسب دسته آن‌ها در اصل مثبت و منفی بوده و توسط الگوریتم نیز به‌درستی، شناسایی‌شده‌اند را بر جمع کل بیماران (اعم از درست یا اشتباه پیش‌بینی‌شده) تقسیم می‌کند؛ بنابراین، این فرمول عملکرد کلی الگوریتم را نشان می‌دهد.

دیگر معیارهای مقایسه بین الگوریتم‌های مختلف شامل: سرعت، قدرت[۱]، مقیاس‌پذیری[۲] و قابلیت فهم الگوریتم‌ها می‌باشند.

منظور از قدرت الگوریتم نحوه تعامل آن با داده‌های ناقص است، مقیاس‌پذیری، درباره مدیریت حجم زیاد داده‌ها توسط الگوریتم است. هراندازه توضیح و درک مدل به‌دست‌آمده از الگوریتم، آسان باشد، به همان اندازه قابلیت فهم آن بالاست.

هنگام مطالعه یک مدل پیش‌بینی در حوزه پزشکی، آشنایی با نحوه تشخیص موارد با برچسب‌های مختلف، مانند پرخطر/کم‌خطر یا سالم/ بیمار توسط مدل مهم است، لذا، اندازه‌گیری حساسیت[۳] و ویژگی[۴] الگوریتم یا الگوریتم‌های تشکیل‌دهنده مدل، اهمیت پیدا می‌کند. در ذیل، فرمول‌های مربوط به حساسیت و ویژگی ذکرشده است.

$$\text{حساسیت} = \frac{Tp}{TP+FN} = \frac{\text{مثبت های صحیح}}{\text{مثبت های صحیح + منفی های اشتباه}}$$

$$\text{ویژگی} = \frac{TN}{TN+FP} = \frac{\text{منفی های صحیح}}{\text{مثبت های اشتباه+ منفی های صحیح}}$$

همچنین شاخص‌های دیگری مانند دقت[۵]، نرخ فراخوانی[۶] و اندازه‌گیری F[۷] وجود دارند که از آن‌ها نیز بهره خواهیم گرفت.

[۱] Robustness
[۲] Scalability
[۳] Sensitivity
[۴] Specificity
[۵] Precision
[۶] Recall
[۷] F-measure

نـرخ فراخـوانی، نسـبت تعـداد مـوارد بازیـابی شـده و مـرتبط بـه تعـداد کـل مـوارد مـرتبط و دقـت نسـبت تعـداد مـوارد بازیـابی شده و مرتبط به تعداد کل موارد بازیابی شده است.

انـدازه‌گیـریF نیـز از میـانگین هارمونیـک نـرخ فراوانـی و دقـت محاسبه می‌شود.

آزمون کاپا برای سنجش میزان انطباق بین تشخیص واقعی با تشخیص سیستم استفاده‌شده است. به‌منظور کسب اطمینان از تصادفی نبودن خروجی سیستم، نتایج حاصل از ارزیابی سیستم با تشخیص واقعی مقایسه گردید و برای این منظور از آزمون آماری کاپا استفاده شد. بیشترین مقدار کاپا یک است که تأثیر شانس در پاسخ‌ها است.

$$kappa = (po - pe)/(1 - pe) \quad (۲)$$

با اســـتفاده از شـــاخص های فوق‌ا لذکر عملکرد الگوریتم های ا ستفاده شده در پایگاه داده سرطان کشوری شهید بهشتی اعمال شدند و مورد مقایسه قرار گرفتند. همچنین ویژگی‌های الگوریتمی که بر مبنای شـاخص‌های فوق نسـبت به سـایر الگوریتم‌ها عملکرد نهایی بهتری داشت استخراج‌شده و با نظر پزشکان متخصص مورد مقایسه قرار گرفت.

در دنیای پزشکی برای فائق آمدن بر محدودیت‌های حساسیت و ویژگی، از منحنی مشخصه عملکرد[۱] استفاده می‌شود. همچنین برای ارزیابی عملکرد مدل‌های ایجادشده توسط داده‌کاوی نیز، این منحنی کاربرد دارد.

AUC روشی برای کمی سازی صحت یک الگوریتم در تشخیص تعلق رکوردهای آموزشی به دسته‌های مختلف است. ازلحاظ مفهوم، AUC عددی است که با محدوده۰/۵(بدون صحت) تا ۱ (صحت کامل) مشخص می‌شود. در حالت کلی، تفسیر AUC با مقادیر مختلف به این صورت است:

- ۰/۵ تا۰/۷: صحت صفرتا پایین
- ۰/۷ تا ۰/۹: صحت متوسط
- بزرگ‌تر از ۰/۹: صحت بالا

هنگام مقایسه یک مدل با AUC بالا نسبت به مدلی با AUC کوچک‌تر، برای انتخاب مدل بهتر، فقط بزرگی مقدار AUC کافی نیست، بلکه باید مقدار آن ازلحاظ آماری معنادار باشد. امکان دارد به‌عنوان‌مثال پیش‌بینی الگوریتم درباره فوتی‌های ناشی از ملانوم با مقدار واقعی آن در مجموعه داده تفاوت داشته باشد. اندازه‌گیری این تفاوت کالیبراسیون نام دارد. هراندازه تفاوت این دو پیش‌بینی کم باشد، نتیجه‌گیری می‌شود که کالیبراسیون مدل دسته‌بندی خوب است. یکی از آزمون‌های مهمی که کالیبراسیون را بررسی می‌کند. آمار کای دو هوسمر ولمشو[۲] است. تفاوت کالیبراسیون با منحنی مشخصه عملکرد در این است که اولی پایایی مدل و دومی قدرت وجه تمایز آن را نشان می‌دهد [۲۸].

بعد از استفاده از الگوریتم‌های مختلف در نرم‌افزار وکا و پیدا کردن الگوریتمی که بتواند دقیق‌تر شانس بقای بیماران مبتلابه ملانوم را تشخیص دهد. مدل مفهومی پیش‌بینی کننده بقای بیماران با استفاده از الگوریتم‌های مورداستفاده در طرح، در قالب یک طرح گرافیکی رسم می‌کنیم. سپس از بین پایگاه داده منتخب، دقیق‌ترین مدل پیش‌بینی کننده بقای افراد مبتلابه ملانوم را با استفاده از آزمون FERIDMANS،ANOVA در نرم‌افزار SPSS تعیین می‌کنیم تا مشخص شود مدل انتخاب‌شده با سایر مدل‌ها ازلحاظ شاخص‌های عملکردی تفاوت معناداری را دارد یا خیر. سپس برای مقایسه دوبه‌دوی مدل‌ها (الگوریتم‌ها) با یکدیگر از آزمون ویکاکسیون[۳] استفاده شد تا فقط مدل‌هایی که ازلحاظ عملکرد تفاوت معناداری داشتند برای مقایسه با مدل‌های تعیین‌شده پایگاه داده انتخاب شوند؛ و در انتها برای تعیین مدل منتخب از آزمون من ویتنی[۴] استفاده می‌کنیم.

۵,۲,۳ توسعه

ساخت مدل، پایان یک پروژه نیست و هدف از پروژه‌های داده‌کاوی کشف دانش و استفاده از دانش کشف‌شده به‌صورت عملی در آینده است .در حقیقت هدف از انجام مراحل مختلف کشف دانش، دست‌یابی به نتایجی است که بتوان از آن‌ها در دنیای واقعی و برای بهبود کارایی سازمان‌ها استفاده کرد.

۳,۳ مرحله دوم: مقایسه عملکرد الگوریتم‌های داده‌کاوی بادانش متخصصین سرطان پوست

[۱] Receiver Operating Characteristic (ROC)

[۲] Hosmer and Lemeshoe" sχ2 statistic

[۳] Wilcoxon

[۴] Maan-Whitney

این مرحله جهت مقایسه عملکرد الگوریتم‌های داده‌کاوی بادانش متخصصین پوست، در پیش‌بینی بقای بیماران مبتلابه ملانوم در پایگاه داده سرطان کل کشور انجام‌شده است. پزشکان متخصص سرطان پوست و پرونده‌های بیماران مبتلابه ملانوم بین سال‌های ۱۳۸۷ تا ۱۳۹۲ که در سال ۱۳۹۸ پیگیری شده‌اند، بر اساس محدودیت‌هایی که داشتند و می‌توانستند در اختیار پژوهشگر قرار دهند،۱۰۰رکورد را پیشنهاد کردند.

چک‌لیستی مبتنی بر صفت‌های بیماران بستری‌شده مبتلابه (ملانوم) بود. این چک‌لیست بر اساس ویژگی‌های بالینی و دموگرافیکی بیماران است. در این چک‌لیست علائم بالینی، بیولوژیکی و پاتولوژیکی بیماران فهرست بندی شده و در مرحله بعد از متخصصان خواسته شد تا احتمال زنده ماندن بیمار را در ۵ سال بعد از تشخیص، با زدن علامتی بر محور مقیاس آنالوگ بصری مشخص کنند. این چک‌لیست از پایان‌نامه رضوی اقتباس و روایی آن نیز تائید شده است.

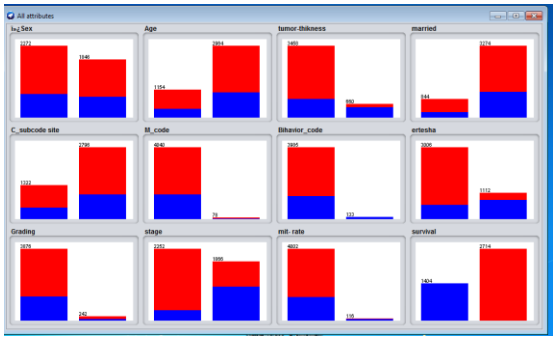
این مرحله برای اعتبار سنجی مراحل قبلی و مقایسه مدل ایجادشده بادانش متخصصان سرطان پوست اجرا شد. رکوردهای استفاده‌شده در این مرحله، قبل از شروع مرحله اول جداشده و در پایگاه داده دیگری نگه‌داری شدند. علت جداسازی این بود که الگوریتم‌های داده‌کاوی که در مراحل قبل برای ایجاد مدل بهینه پیش‌بینی بقای سرطان پوست به‌کاررفته‌اند بر روی این رکوردها عمل آموزش انجام ندهند، چون در این صورت این رکوردها نمی‌توانند به‌درستی میزان خطای مدل پیش‌بینی را مشخص کنند.

مدل‌های بهینه پیش‌بینی که از پایگاه داده‌های کشور انتخاب‌شده‌اند به ترتیب برای پیش‌بینی بقای هر یک از این رکوردها به‌کاررفته و احتمال پیش‌بینی آن‌ها برای هر رکورد ثبت شد. رکوردهای فوق به متخصصان سرطان پوست در قالب چک‌لیست داده شد تا پیش‌بینی آن‌ها نیز به دست آید. سپس، این پیش‌بینی‌ها با استفاده از سطح زیر منحنی مشخصه عملکرد و مقایسه این سطوح با یکدیگر توسط روش دوبه‌دو[1] موردبررسی قرار گرفتند تا درنهایت مدل بهینه پیش‌بینی بقای (ملانوم) انتخاب شود.

بیمارانی که احتمال زنده ماندن آن‌ها ۵ سال بعد از تشخیص سرطان پوست (ملانوم) ۵۰ درصد یا کمتر از آن باشد در دسته فوتی‌ها بیشتر از ۵۰ درصد در دسته زنده مانده‌ها قرار گرفت. علت انتخاب این معیار برابر بودن اولیه و پیش‌فرض وزن پیش‌بینی‌های موارد بقای اشتباه با موارد فوت اشتباه بود. چون در پزشکی، پیش‌بینی‌های موارد بقای اشتباه قابل‌قبول‌تر از پیش‌بینی موارد فوت اشتباه است[۲۸]. لذا برای مدیریت این مسئله منحنی مشخصه عملکرد که قابلیت پیش‌بینی همه موارد بقا و فوت را فراتر از معیارهای پیش‌فرض دارد، برای پیش‌بینی‌های متخصصان ملانوم رسم و با منحنی بهترین الگوریتم که در مراحل قبلی مشخص شده است، مقایسه گردید. برای سنجش پایایی پیش‌بینی متخصصان و الگوریتم‌ها از آزمون Hosmer Lemeshow goodness-of-fit استفاده شد [۲۹].

۴. یافته‌ها

بررسی نتایج حاصل از داده های ۷ الگوریتم داده کاوی که متغیرهای به‌کارگرفته شده در شکل (۵) به نمایش گذاشته شده است، (در نرم‌افزار وکا) مورداستفاده جهت بررسی بقای افراد مبتلابه ملانوم را نشان می‌دهد که از میان هفت الگوریتم برای پیش‌بینی بقای بیماران مبتلابه ملانوم، الگوریتم شبکه عصبی ازلحاظ صحت، حساسیت و ویژگی بالاتر از دیگر الگوریتم‌ها بود.



شکل ۵. نمایش گرافیکی فیلدها

جدول ۷. حساسیت و صحت و ویژگی الگوریتم‌های پیش‌بین بقای بیماران مبتلابه سرطان پوست در پایگاه داده کل کشور

مدل	صحت	حساسیت	ویژگی
شبکه بیزی	۹۰/۸۰	۹۳/۱۴	۸۶/۱۸
J48	۹۳/۲۰	۹۷/۱۶	۸۵/۴۷
شبکه عصبی	۹۳/۵۰	۹۷/۷۹	۹۱/۰۳
رگرسیون لجستیک	۹۲/۷۰	۹۶/۳۵	۸۵/۵۴

[1] pair-wise

ترکیب درخت تصمیم و بیز ساده	۹۰/۳۰	۹۰/۲۳	۹۰/۴۵
بیزی ساده	۹۱/۳۰	۹۳/۶۵	۸۹/۴۵
ID3	۹۱/۳۰	۹۳/۶۲	۸۶/۷۵

از میان هفت الگوریتم مورداستفاده برای پیش‌بینی بقای بیماران مبتلابه ملانوم طبق جدول (۷) الگوریتم شبکه عصبی ازلحاظ صحت، حساسیت و ویژگی بالاتر از دیگر الگوریتم‌ها بود.
همچنین شاخص کاپا (عملکرد) برای الگوریتم‌های استفاده‌شده در پژوهش در پایگاه‌های داده کشور در جدول (۸) است:

جدول ۸. عملکرد کاپا الگوریتم‌های پایگاه داده کشور

الگوریتم	Kappa
شبکه عصبی	۰/۸۵۲
J48	۰/۸۴۴
رگرسیون لجستیک	۰/۸۳۳
شبکه بیزی	۰/۸۰۴
بیزی ساده	۰/۸۰۶
ID3	۰/۸۰۵
درخت تصمیم و بیزی ساده	۰/۷۸۹

طبق آ مار جدول (۷ و ۸) کاملاً می‌توان عملکرد الگوریتم های داده‌کاوی اجراشده در پایگاه داده کل کشور را پیش‌بینی کرد. الگوریتم پایگاه داده کشوری در شاخص‌های صحت، دقت، حساسیت و ویژگی شبکه عصبی تعیین شد. در حوزه پزشکی، پیش‌بینی مواردی که منجر به ناخوشی یا مرگ می‌شوند از اهمیت بیشتری نسبت به پیش‌بینی موارد غیر آن برخوردار است. شاخصی که قدرت الگوریتم را برای پیش‌بینی این موارد برر سی می‌کند، ویژگی ا ست؛ بنابراین مشاهده می‌شود که ویژگی الگوریتم‌های اجراشده در پایگاه داده کشوری کاملاً قابل‌قبول است. همچنین با توجه به این‌که پایگاه داده ک شور متعادل سازی شده لذا محا سبه شاخص کاپا نیز ضروری است که در الگوریتم شبکه عصبی ۰/۸۵۲ بود. کلاً الگوریتم‌های استفاده‌شده در پایگاه داده کشوری بالأخص شبکه عصبی با اطمینان بیشتری، قادر به پیش‌بینی بیماران جدید مبتلا به ملانوم هستند.

جدول ۹. نرخ فراخوانی، ROC اندازه‌گیری F الگوریتم‌های پیش‌بین بقای بیماران مبتلابه ملانوم در پایگاه داده کل کشور

الگوریتم‌ها	نرخ فراخوانی	ROC	دقت	اندازه‌گیری F
شبکه بیزی	۹۰/۸۰	۹۵/۴۰	۹۰/۸۰	۹۰/۸۰
J48	۹۳/۲۰	۹۳/۷۰	۹۳/۲۰	۹۳/۱۰
شبکه عصبی	۹۳/۵۰	۹۶/۱۰	۹۷	۹۳/۴۰
رگرسیون لجستیک	۹۲/۷۰	۹۵/۶۰	۹۲/۷۰	۹۲/۶۰
ترکیب درخت تصمیم و بیز ساده	۹۰/۳۰	۹۵/۷۰	۹۰/۷۰	۹۰/۴۰
بیزی ساده	۹۱/۳۰	۹۵/۷۰	۹۱/۳۰	۹۱/۳۰
ID3	۹۱/۳۰	۹۵/۴۰	۹۱/۲۸	۹۱/۳۰

همان‌طور که در جدول (۹) نشان داده‌شده است نرخ فراخوانی الگوریتم‌های بیزی، درخت تصمیم، شبکه عصبی، رگرسیون لج ستیک، ترکیب درخت ت صمیم و بیز ساده، بیز ساده و ID3 به ترتیب: ۹۰/۸۰، ۹۳/۲۰، ۹۳/۵۰، ۹۲/۷۰، ۹۰/۳۰، ۹۱/۳۰، ۹۱/۳۰ است. همچنین ترتیب پارامتر ROC برای الگوریتم های بیزی، درخت تصمیم، شبکه عصبی، رگر سیون لجستیک، ترکیب درخت ت صمیم و بیز ساده، بیزی ساده و ID3 به ترتیب: ۹۵/۴۰، ۹۳/۷۰، ۹۶/۱۰، ۹۵/۶۰، ۹۵/۷۰، ۹۵/۷۰، ۴۰ /۹۵ و برای پارامتر ا ندازه‌گیریF برای الگوریتم های ذکرشده به ترتیب: ۹۰/۸۰، ۹۳/۱۰، ۹۲/۹۳،۶۰/۴۰، ۹۰/۴۰، ۹۱/۳۰، ۹۱/۳۰ است.

ارزیابی عملکرد متخصصین

سطح معنی‌دار تفاوت بین زیر منحنی مشخصه‌ی عملکرد بیز ساده و متخصص یک ۰/۰۳۹، متخصص دو ۰/۰۲۲، متخصص سه ۰/۰۴،

متخصص چهار ۰/۰۲۳ و متخصص پنج ۰/۰۵۹ بوده است (آلفا=۰۵-/۰). سطح معنی‌دار تفاوت بین مشخصه عملکرد الگوریتم درخت تصمیم‌گیری با بیز ساده با متخصصین سرطان پوست به ترتیب عبارت است از: ۰۵۸/ ۰، ۰/۰۴، ۰/۰۵۹، ۰/۰۴۲ و ۰/۰۷۷ (آلفا: ۰۵-/۰). همچنین سطح معنی‌دار تفاوت بین مشخصه عملکرد شبکه بیزی با متخصصین سرطان پوست به ترتیب: ۰/۰۵۹، ۰/۰۴۲، ۰۶۱-/۰، ۰/۰۴۳ و ۰/۰۷۹ (آلفا=۰۵ / ۰) بود. سطح معنی‌دار تفاوت بین مشخصه عملکرد شبکه عصبی با متخصصین سرطان پوست عبارت است از: ۰۱۹/ ۰، ۰۰۱/ ۰، ۰۰۲/ ۰، ۰۰۲ / ۰ و ۰۳۸/ ۰ (آلفا=۰۵/ ۰). سطح معنی‌دار تفاوت بین مشخصه عملکرد ID3 با متخصصین سرطان پوست به ترتیب: ۰/۰۸۱، ۰/۰۴۲، ۰/۰۶۳، ۰/۰۶۵و۰/۱، (آلفا=۰۵/ ۰). سطح معنی‌دار تفاوت بین مشخصه درخت تصمیم‌گیری با متخصصین سرطان پوست به ترتیب عبارت است از: ۰/۰۸۷، ۰/۰۷، ۰/۰۸۸، ۰۷۱/ ۰ و ۱۰۷/ ۰ (آلفا=۰۵/ ۰). سطح معنی‌دار تفاوت بین مشخصه عملکرد رگرسیون لجستیک با متخصصین سرطان پوست به ترتیب: ۰/۰۹۱، ۰۸/۰۷۳، ۰۸/ ۰، ۰۷۷ /۰ و ۱۱۱/ ۰ (آلفا=۰/۰۵) بود.

در جداول۱۰ و ۱۱،که ماتریس اغتشاش[1] هستند پیش‌بینی‌های انجام‌شده به‌وسیله متخصصان و الگوریتم‌ها به همراه مقایسه آن‌ها با مقادیر واقعی گنجانده‌شده است.

جدول ۱۰. بررسی ماتریس اغتشاش پیش‌بینی‌های انجام‌شده توسط ۵ متخصص سرطان پوست و مقایسه آن‌ها با موارد به وقوع پیوسته ۱۰۰ رکورد پایگاه داده‌های کشوری

	به وقوع پیوسته	متخصص ۱		متخصص ۲		متخصص ۳		متخصص ۴		متخصص ۵	
		زنده	مرده	زنده	مرده	زنده	مرده	زنده	مرده	زنده	مرده
زنده	۶۷	۶۱	۶	۶۳	۴	۶۰	۷	۶۲	۵	۵۷	۱۰
مرده	۳۳	۱۳	۲۰	۱۶	۱۷	۱۲	۲۱	۱۵	۱۸	۹	۲۴

جدول ۱۱. بررسی ماتریس اغتشاش پیش‌بینی‌های انجام‌شده توسط الگوریتم‌های مورداستفاده داده‌کاوی و مقایسه آن‌ها با موارد به وقوع پیوسته ۱۰۰ رکورد پایگاه داده‌های کشوری

	به وقوع پیوسته	شبکه عصبی		J48		شبکه بیزی		بیزی ساده		رگرسیون لجستیک	
		زنده	مرده	زنده	مرده	زنده	مرده	زنده	مرده	زنده	مرده
زنده	۶۷	۶۱	۶	۶۳	۴	۶۰	۷	۶۲	۵	۵۷	۱۰
مرده	۳۳	۱۳	۲۰	۱۶	۱۷	۱۲	۲۱	۱۵	۱۸	۹	۲۴

[1] Confusion matrix

9	58	16	51	11	56	10	57	20	47	۶۷	زنده
23	10	30	3	25	8	24	9	32	1	۳۳	مرده

الگوریتم‌های اجراشده در پایگاه داده بومی بر روی ۱۰۰ رکورد از مجموعه رکوردهای متعلق به همین پایگاه داده که در پایگاه جداگانه قرارگرفته بودند، اجرا گردیدند. سپس پنج نفر از متخصصین سرطان پوست پیش‌بینی‌های خود را در رابطه با اطلاعات مربوط به این ۱۰۰ رکورد انجام دادند. نتایج به‌دست‌آمده از این ۷ الگوریتم و متخصصان سرطان پوست درباره ۱۰۰ رکورد جمع‌آوری‌شده و منحنی‌های مشخص عملکرد آن‌ها رسم گردید. در جدول (۱۲)، بخش زیر منحنی مشخصه‌ی عملکرد[2] الگوریتم‌ها و متخصصان پوست را نشان می‌دهد.

پس از انجام بررسی ماتریس اختلاط، آماره لمشو، مقادیر chi-squared و p-value برای متخصصین سرطان پوست و برای الگوریتم‌های اجرا کرده و در یافته‌ها مقایسه می‌شوند.

ترکیب بیزی ساده و درخت تصمیم		ID3	
زنده	مرده	زنده	مرده
55	12	56	11
7	26	9	24

مقایسه عملکرد الگوریتم‌های داده‌کاوی پایگاه داده کشور با پیش‌بینی‌های متخصصین سرطان پوست

پس از انجام آماره هاسمر- لم شو، مقادیر chi-squared و value-p برای متخصصین سرطان پوست و برای الگوریتم‌های اجراشده طبق جداول زیر به دست آمد.

[2] Areas under the ROC curve (AUC)

جدول ۱۲. سطح زیر منحنی عملکرد الگوریتم‌ها و متخصصین سرطان پوست برای ۱۰۰ رکورد پایگاه داده کشور

الگوریتم‌ها	شاخص عملکرد سطح زیر منحنی
بیزی ساده	۰/۸۴۳
ترکیب درخت تصمیم و بیز ساده	۰/۸۱۱
بیزی شبکه	۰/۸۱۱
شبکه عصبی	۰/۸۷۹
ID3	۰/۷۷۱
J48	۰/۷۶
رگرسیون لجستیک	۰/۷۵۱
متخصص ۱	۰/۷۵۱
متخصص ۲	۰/۶۹۸
متخصص ۳	۰/۷۳۲
متخصص ۴	۰/۷۱۶
متخصص ۵	۰/۷۶۲

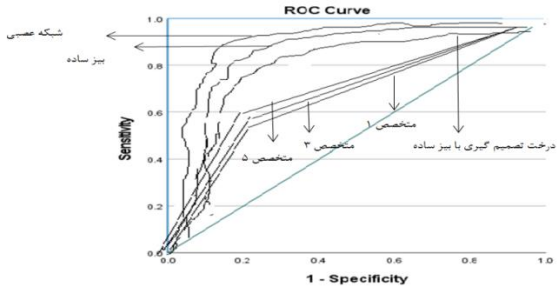
جدول ۱۳. مقدار کای دو و p-Value متخصصین سرطان پوست با استفاده از آزمون هاسمر-لمشو

افراد	کای ۲	p-Value
متخصص ۱	۷/۰۲	۰/۱۲
متخصص ۲	۶/۰۳	۰/۱۳
متخصص ۳	۲/۰۰۲	۰/۷۴
متخصص ۴	۳/۶۳	۰/۱۶
متخصص ۵	۱/۰۰۴	۰/۸۸

جدول ۱۴. مقدار کای دو و p الگوریتم‌های داده‌کاوی با استفاده از آزمون هاسمر-لمشو

الگوریتم	کای ۲	p-Value
بیز ساده	۰/۸	۰/۹
ترکیب درخت تصمیم‌گیری و بیز ساده	۰/۸۶۳	۰/۸۷
شبکه بیزی	۰/۹۴۵	۰/۷۶۴
شبکه عصبی	۰/۷۴۳	۰/۹۷
ID3	۱/۱۲	۰/۷۵
رگرسیون لجستیک	۱/۱۹	۰/۶۵
J48	۱/۱	۰/۷۸

در شکل (۶) منحنی ROC سه الگوریتم برتر ازنظر عملکرد، با سه متخصص برتر مورد مقایسه قرار گرفت است.



شکل ۶. مقایسه‌ی منحنی ROC سه الگوریتم برتر پایگاه داده‌های بومی با سه متخصص برتر

با توجه به شکل بالا مشخص‌گردیدکه سطح زیر منحنیROC الگوریتم‌ها نسبت به متخصصین بیشتر است. در بین متخصصین نیز سطح زیر منحنی متخصص ۵ از سایر متخصصین بیشتر است.

با توجه به اینکه مقدار AUC بین ۰/۹ تا ۱ بوده است. قدرت.
تشخیص آزمون بسیار عالی است؛ بنابراین قدرت تشخیص آزمون با دقت بسیار بالایی انجام‌شده است.
در حالت کلی الگوریتم‌های (مدل‌های) اجراشده در پایگاه داده کشوری ازلحاظ عملکرد در شاخص‌های ارزیابی نسبت به یکدیگر ازلحاظ آماری تفاوت معنی‌داری داشتند:

X^2= ۰/ ۰۰۰۱

P< ۰/۰۰۱

برای مقایسه دوبه‌دوی مدل‌های (الگوریتم‌های) اجراشده در پایگاه داده کشوری با یکدیگر از آزمون‌های ویلکاکسون استفاده شد. همچنین تصحیح بونفرونی[1] برای مقایسه‌های دوبه‌دو مدل‌ها بکار رفت و بنابراین تفاوت معنی‌داری مدل‌ها با یکدیگر در سطح معنی‌داری ۰/۰۰۷ گزارش شد. نتایج نشان دادند که تفاوت مدل‌های شبکه عصبی، بیز ساده، شبکه بیزی و ترکیب درخت تصمیم‌گیری با بیز ساده با ID3,J48 و رگرسیون لجستیک در شاخص‌های ارزیابی ازلحاظ آماری با یکدیگر تفاوت معناداری دارند. درحالی‌که عملکرد الگوریتم‌های شبکه عصبی، بیز ساده و شبکه بیزی و ترکیب درخت تصمیم‌گیری با بیز ساده با یکدیگر ازلحاظ آماری متفاوت نبودند. لذا مدل‌های شبکه عصبی، بیز ساده، شبکه بیزی و ترکیب درخت تصمیم‌گیری با بیز ساده، از پایگاه داده کشوری انتخاب شدند.
برای تعیین مدل نهایی از بین مدل‌های انتخاب‌شده از پایگاه داده کل کشور، آزمون من ویتنی استفاده شد که نتایج نشان داد که عملکرد مدل شبکه عصبی در همه شاخص‌های ارزیابی در پایگاه داده کشوری نسبت به سایر مدل‌ها بالاتر بوده و ازلحاظ آماری تفاوت معنی‌داری دارد. در ذیل نتایج همه مدل‌ها آورده شده است:

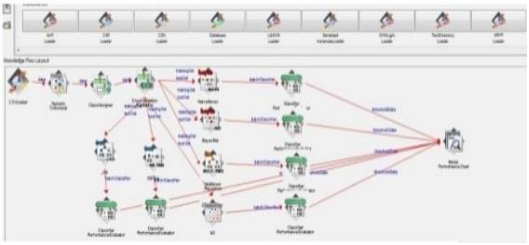
شبکه عصبی: U=۸ و z= -۲/۱۸, p< ۰/۰۰۱

بیز ساده : U=۲۱/۵ و z= -۰/۳۸, p> ۰/۰۵

شبکه بیزی: U=۲۰ و z= -۰/۵۷, p> ۰/۰۵

ترکیب درخت تصمیم با بیز ساده: U= ۱۴ و z=-۱/۳۴ , p> ۰/۰۵

در پایگاه داده کل کشور الگوریتم شبکه عصبی ازلحاظ شاخص‌های ارزیابی‌شده نسبت به سایر الگوریتم‌ها برتری داشت. در شکل ۷ مدل نهایی در نرم‌افزار وکا با توجه به مقایسه الگوریتم‌ها در محیط KnowledgeFlow به اجرا درآمد.



شکل ۷. نرم‌افزار وکا مدل نهایی (اجراشده)

۵. بحث و نتیجه‌گیری

داده‌کاوی روی‌داده‌های پزشکی از اهمیت بالایی برخوردار است و طراحی سیستم‌های تصمیم‌یار هوشمند، جهت یاری‌رساندن به پزشکان درزمینه‌ی تشخیص نوع بیماری یا انتخاب نوع درمان مناسب و مدیریت بهینه تصمیم‌گیری در کسب‌وکار وزارت بهداشت و درمان، با کمک داده‌کاوی می‌تواند کمک شایانی درزمینه‌ی نجات جان انسان‌ها انجام دهد. در همین راستا در پژوهش حاضر الگوریتم شبکه عصبی با بهترین عملکرد به پیش‌بینی بقای بیماران مبتلابه ملانوم پرداخته است. پژوهش حاضر باهدف طراحی مدل پیش‌بینی بقای بیماران مبتلابه ملانوم بر اساس الگوریتم‌های داده‌کاوی با استفاده از روش داده‌کاوی به‌وسیله نرم‌افزار داده‌کاوی وکا و مدیریت هوشمندانه تصمیم‌یار، انجام گردید. از مزایای آن می‌توان به‌دقت و صحت پیش‌بینی بالا و ارائه‌ی راه‌حل‌های مختلف جهت بیشتر شدن شانس بقای افراد مبتلابه سرطان ملانوم اشاره کرد. در این پژوهش از مجموع داده موجود در مرکز سرطان کل کشور دانشگاه شهید بهشتی استفاده گردید؛ و با انتخاب الگوریتم‌های شبکه‌ی عصبی، ID3، ترکیب درخت تصمیم و بیزی ساده، بیزی ساده، رگرسیون لجستیک، شبکه بیزی و j48، طراحی و ارزیابی‌ها صورت گرفت. در این بررسی نشان داده شد که شبکه عصبی در پارامترهای صحت، دقت، حساسیت و ویژگی از همه‌ی الگوریتم‌های ذکرشده بالاتر بود.
اعمال روش‌های داده‌کاوی بر روی‌داده‌های پزشکی می‌تواند به‌عنوان سیستم‌های تصمیم‌یار، در تصمیم‌گیری برای انتخاب نوع درمان و یا تشخیص بیماری‌ها، به متخصصان کمک نماید .این همان فلسفه داده‌کاوی است که با شناخت درست از گذشته، آینده را پیش‌بینی می‌نماید. در این پژوهش با ارائه دقیق‌ترین مدل تصمیم‌یار، گام

[1] Bonferroni

اثربخش و کارآمدی در مدیریت سیستم‌های هوشمند در علم پزشکی برداشته و در علم داده‌کاوی برای بررسی و ارزیابی الگوریتم‌ها و داده‌های خروجی آن‌ها پارامترهایی نظیر صحت، حساسیت، ویژگی، نرخ فراخوانی، شاخص ROC، شاخص کاپا و شاخص F وجود دارد که هر پارامتر برای بررسی یک الگوریتم مورداستفاده حائز اهمیت است. نتایج این پژوهش نشان داد، الگوریتم شبکه عصبی در پیش‌بینی بقای بیماران مبتلابه ملانوم در زمینه‌های صحت، حساسیت، دقت و ویژگی که از روش‌های اصلی ارزیابی یک الگوریتم محسوب می‌شوند، نسبت به سایر الگوریتم‌ها عملکرد بهتری داشت. برای ارزیابی عملکرد الگوریتم‌ها و تحلیل راجع به آن‌ها توجه به این پارامترهای اساسی، حائز اهمیت است.

متقاعد کردن متخصصین بالینی درباره فواید یک مدل حمایت از تصمیم‌گیری بالینی امر مهمی است. برای رسیدن به این هدف، اثبات کارایی مدل ضروری است. الگوریتم شبکه عصبی در پیش‌بینی موارد فوتی بهتر از سایر الگوریتم‌ها و متخصصین عمل کرده است. این نشانگر کاربردی بودن این الگوریتم است. از میان متخصصان هم متخصص ۵ در پیش‌بینی موارد فوتی بهتر بوده است. نتایج آزمون Hosmer Lemeshow goodness-of-fit که برای سنجش کالیبراسیون است. (منظور از کالیبراسیون پایایی تحقیق است) نشان می‌دهد از میان متخصصین، متخصص ۵ به علت بیشتر بودن مقدار p-value (۰/۸۸) و از الگوریتم‌ها، شبکه عصبی با مقدار ۰/۹۷ ازلحاظ پایایی پیش‌بینی مناسب‌تر است؛ بنابراین مدل پیش‌بینی کننده بقای ملانوم، هم ازلحاظ قدرت تمایز و هم ازلحاظ پایایی، عملکرد بهتری از خود نشان داده است. لذا در صورت عدم وجود متخصصین پوست، به‌عنوان تصمیم‌یار هوشمند، جایگزین مناسبی برای آن‌ها خواهند بود.

پیشنهادها برای مطالعات بعدی پژوهشگران

- پیشنهاد می‌شود، پیش‌بینی بقای ملانوم با پیگیری بیش از ۵ سال انجام شود.(۱۰ الی ۱۵ سال)
- پیشنهاد می‌شود، از معماری‌های مختلف الگوریتم شبکه عصبی برای پیش‌بینی بقای ملانوم و مقایسه عملکرد آن‌ها استفاده شود.
- پیشنهاد می‌شود، عملکرد مدل ایجاد در پایگاه داده بومی با پایگاه‌های داده منتخب سایت‌های مختلف مقایسه شود.
- پیشنهاد می‌شود، از الگوریتم‌های درخت تصمیم‌گیری برای غربال کردن متغیرهای کم‌اهمیت و ادغام خروجی آن در شبکه عصبی برای پیش‌بینی بقای ملانوم، استفاده شود.
- پیشنهاد می‌شود، مراکز داده‌کاوی بیماران مختلف در کنار مراکز پژوهشی بیماری ایجاد شود؛ که بعد از فرآیند داده‌کاوی و ایجاد مدل در قالب نرم‌افزار حمایت از تصمیم‌گیری بالینی وارد صنعت کند. تا در راستای پژوهش‌های دانش‌بنیان نیز قابل‌اجرا شود.

مراجع

[1] Zamanian Azodi M, Azizi Jalilian F. Early detection of cancer and proteomics. Journal of Ilam University of Medical Sciences 2013; 21 (1): 112-22. (Persian)

[2] N. R. Aaron, and T. M. Khoshgoftaar, "Modernizing analytics for melanoma with a large-scale research dataset," In 2017 IEEE International Conference on Information Reuse and Integration (IRI), pp. 551 - 558, 2017.

[3] Howlader N, N.A. Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Feuer EJ, Cronin KA. SEER Cancer Statistics Review, 19752016. 2018 December, 11, 2019 2020].

[4] Seyedeh Elham Hosseini Fadafen, Emad Fatemizadeh, "Diagnosis of skin cancer by extracting features from images"Shhab Dansh Institute(2016).

[5] R. Dolfe, and K. Matinzadeh, "Investigating Skin Cancer with Unsupervised Learning", Project Thesis, ETH Royal Institute of Technology 2019.

[6] Kim, R. H. & Meehan, S. A. Immunostain use in the diagnosis of melanomas referred to a tertiary medical center: a 15-year retrospective review (2001–2015). J. Cutan. Pathol. 44, 221–227 (2017).

[7] David A. Kirchar DA, Mark R, et al. Melanoma brain metastasis: mechanisms, models and medicine. J Mol Sci 2016; 17 (7): 1-29.

[8] Wrobel, S. M. Przybylo, and E. Stepien, The Clinical Trial Landscape for Melanoma Therapies. J Clin Med, 2019. 8(3).

[9] K. Q. Kanaan. "Classification of human skin diseases using data mining," International Journal of Advanced Engineering Research and Sci, Vol. 4, no. 1, 2017.

[10] L. Sophia, and M. S. Pallavi. "Predicting Protein in Cancer Diagnosis Using Effective Classification and Feature Selection Technique," IEEE International Conference on Communication and Signal Processing (ICCSP), pp. 156-159, 2018

[11] MK Keleş, Tehnički vjesnik, Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms, 2019, Sarıçam Adana, Turkey

[12] S. Winiarti, H. Yuliansyah, and A. A. Purnama, "Identification of Toddlers' Nutritional Status using Data Mining Approach," International Journal of Advanced Computer Science and Applications, Vol. 9, No. 1, pp. 164 - 169, 2018.

[13] Mahmoodi MS, Mahmoodi SA, Haghighi F, Mahmoodi SM. Determining the stage of breast cancer by data mining algorithms. ijbd. 2014;7(2):36-44.

[14] Dehghan P, Mogharabi M, Zabbah I, Layeghi K & Maroosi A. Modeling Breast cancer using data mining methods. Journal of Health and Biomedical Informatics 2018; 4(4): 266-78[Article in Persian].

[15] Ismaili N, Mohaghegh MA, Safai Faraq Z, Emami Razavi SZ, Chavoshi M, Emami Razavi S H. 5-year survival rate of malignant melanoma in Iran, Modval Scientific Journals, 1386: (11)3:1263-1263

[16] Hira Beenish, Muhammad Fahadad International Conference on Computing and Information Technology, University o f Tabuk, Kingdom o f Saudi Arabia. Volume: 01, Issue: ICCIT- 1441, Page No.: 44 - 47, 9 & 1 9 Sep. 2020

[17] Suma, V., & Hills, S. M. (2020). Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics. Journal of Soft Computing Paradigm (JSCP), 2(03), 153-159.

[18] . Martınez AM, Webb GI, Chen S, Zaidi NA. Scalable learning of Bayesian network classifiers. Journal of Machine Learning Research. 2016;17(44):1-35

[19] Wang, Haoxiang. "IoT based Clinical Sensor Data Management and Transfer using Blockchain Technology." Journal of ISMAC 2, no. 03 (2020): 154-159.

[20] Ameri H, Alizadeh S & Barzegari A. Knowledge extraction of diabetics' data by decision tree method. Health Management
2013; 16(53): 58-72[in Persian]

[21] Ahmed K, Jesmin T, Rahman MZ. Early Prevention and Detection of Skin Cancer Risk using Data Mining. International Journal of Computer Applications 2013;62(4):1 -6.

[22] Gibert K, Sanchez-Marre M, Codina V editors . Choosing the right data mining technique: Classification of methods and intelligent recommendation. Proceeding the IEMSs fifth biennial meeting international congress on environmental modeling and software ;2015.

[23] Witten I. Frank E.Hall M. Data mining : Practical machine Learning tools and techniques: Morgan Kaufinann: 2011.

[24] R. Arulmurugan, K. R. Sabarmathi, and H. Anandakumar, "Classification of sentence level sentiment analysis using cloud machine learning techniques," Cluster Computing, vol. 22, no. S1, pp. 1199–1209, Sep. 2017.

[25] Sarafi Nejad A, Saeid AH, Rose I.M, Rowhanimanesh AR. Modeling a Data Mining Decision Tree and Propose a New Model for the Diagnosis of Skin Cancer by Immunohistochemical Staining Methods. Journal of Health and Biomedical Informatics 2014; 1(1): 54-62.[Persian].

[26] U.-O. Dorj, K.-K. Lee, J.-Y. Choi, M. J. M. T. Lee, and Applications, "The skin cancer classification using deep convolutional neural network," pp. 1-16, 2018.

[27] Yu CT, Chao CM, Cheng BW. Prediction of surv using three artificial intelligence techniques. Journal of Theoretical and applied Information Technology. 2014;60(1):179-83. 99.

[28] Razavi AR Applications of knowledge discovery in quality registries - predicting recurrence of breast

cancer and analyzing non-compliance with a clinical guideline: Institutionen för medicinsk teknik; 2015.

[29] Gopalakrishnan, V. et al. Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. Science 359, 97–103 (2018).