مجله علمی مطالعات پژوهشی در مهندسی برق آینده، ۱۴۰۳، دوره ۲، ص ص ۷-۱۷

مجله علمی مطالعات پژوهشی

در مهندسی برق آینده

www.journalhi.com/pow/

ارزیابی سیستم‌های تشخیص کووید-۱۹ با استفاده از یادگیری ماشین

مجتبی عزیزی* [۱]، سعید علیخانی [۲]

۱- دانشکده شیمی و مهندسی شیمی، دانشگاه صنعتی مالک اشتر، تهران، ایران، صندوق پستی ۳۴۵۴-۱۶۷۶۵.

۲- کارشناسی ارشد هوش مصنوعی و رباتیکز، دانشکده برق و کامپیوتر، دانشگاه آزاد اسلامی واحد تهران شمال، ایمیل: sa.alikhani75@gmail.com

چکیده

داده‌کاوی فرایند تجزیه وتحلیل حجم عظیمی از داده‌ها برای کشف اطلاعات است؛ این فرایند، توسط شرکت‌ها برای تبدیل داده‌های خام به اطلاعات مفید در نظر گرفته می‌شود و همچنین کاوش و تجزیه‌وتحلیل انبوهی از اطلاعات برای به‌دست آوردن الگوها و روندهای معنی‌دار، مورد استفاده قرار می‌گیرد. کاوش در داده به شرکت‌ها در حل مشکلات، کاهش خطرات و استفاده از فرصت‌های جدید کمک می‌کند. این شاخه از علم داده، نام خود را از شباهت‌های جست‌وجوی اطلاعات ارزشمند در یک پایگاه داده بزرگ با استخراج از کوه برای یافتن سنگ معدن الهام گرفته است. ویروس کرونا بیش از ۱۰۰ میلیون نفر را مبتلا کرده و منجر به مرگ تقریباً سه میلیون نفر در سراسر جهان شده است. برای کاهش این گسترش همه‌گیری بی‌سابقه، استفاده از تکنیک‌های هوش مصنوعی برای توسعه ابزارهایی که از پزشکان در کارهای مختلف حمایت می‌کنند، توجه روزافزونی به‌خود جلب کرده است. علی‌رغم نتایج امیدوارکننده برای کار تشخیصی (یعنی تشخیص COVID-19)، توسعه مدل‌های پیش‌آگهی، یا برای پیش‌بینی پذیرش بستری در ICU یا سایر پیامدها (ازجمله مرگ) یا طبقه‌بندی بیماران بر اساس خطر، تاکنون پیشرفت‌های هوش مصنوعی در زمینه تشخیص COVID-19 عقب مانده است. در فصل اول، به بررسی تکنیک‌های مربوط به داده‌کاوی پرداخته شده است. سپس در فصل دوم، با استفاده از مجموعه داده مربوط به COVID-19 که از سایت Kaggel گرفته شده، عملیات داده‌کاوی بر روی این مجموعه داده انجام و نتایج آن نیز گزارش شده است. سپس، مقالات مربوط به تشخیص COVID-19 در فصل سوم با هم مورد مقایسه قرار گرفته‌اند.

کلمات کلیدی: داده‌کاوی؛ تشخیص؛ COVID-19؛ هوش مصنوعی؛ Kaggel

Evaluation of COVID-19 Diagnosis Systems Using Machine Learning

Dr. Mojtaba Azizi [1,*], Saeed Alikhani [2]

1- Faculty of Chemistry and Chemical Engineering, Malek Ashtar University of Technology, P.O. Box 16765-3454, Tehran, Iran.
2- Department of Computer Engineering, North Tehran Branch, Islamic Azad University, Tehran, Iran. Email: sa.alikhani75@gmail.com

Abstract

Data mining is the process of analyzing large volumes of data to uncover meaningful information. This process is employed by companies to transform raw data into valuable insights, as well as to explore and analyze vast amounts of information to identify significant patterns and trends. Data mining aids companies in solving problems, reducing risks, and capitalizing on new opportunities. This branch of data science derives its name from the analogy of searching for valuable information in a large database, similar to mining ore from a mountain. The COVID-19 virus has infected over 100 million people and resulted in nearly three million deaths worldwide. To curb the spread of this unprecedented pandemic, increasing attention has been directed toward the use of artificial intelligence techniques to develop tools that assist physicians in various tasks. Despite promising results for diagnostic tasks (such as COVID-19 detection), the development of prognostic models—whether for predicting ICU admissions, other outcomes (including mortality), or classifying patients based on risk—has so far lagged behind in the context of COVID-19 diagnosis.
In the first chapter, the relevant data mining techniques are explored. In the second chapter, data mining operations are conducted on a COVID-19 dataset obtained from Kaggle, and the results are reported. Finally, in the third chapter, studies related to COVID-19 diagnosis are compared.

Keywords: Data mining; Diagnosis; COVID-19; Artificial intelligence; Kaggle

۱- مقدمه

الگوریتم یک روش برای پیاده‌سازی حل مسئله است، یعنی روشی برای جست‌وجوی الگو در داده‌ها. اما با گسترش سیستم‌های پایگاهی و ذخیره حجم بالای اطلاعات در این سیستم‌ها به ابزاری قوی‌تر نیاز است تا بتوان داده‌ها را پردازش کرد و اطلاعات را در اختیار کاربران قرار داد. در واقع، ما به روش‌هایی نیاز داریم که اصطلاحاً به کشف دانش بپردازد. یعنی روشی که با کمترین دخالت کاربر، به‌صورت خودکار الگوها و رابطهٔ منطقی بین آنها را بیان کند.

الگوریتم‌های داده‌کاوی

با استفاده از تکنیک‌های داده‌کاوی، سرعت انجام محاسبات و فضای مورد نیاز در حافظه (RAM) بهبود قابل ملاحظه‌ای پیدا می‌کند. به‌طور کلی می‌توان انواع تکنیک‌های داده‌کاوی را در یکی از ۳ دسته‌ای که در ادامه می‌آید و یا ترکیبی از آنها، قرار داد.

۱- **دسته‌بندی**[1]: در این نوع یادگیری، بر اساس ویژگی‌های تعریف شده به داده‌ها برچسب زده می‌شود و آنها را در کلاس‌های مختلف قرار می‌دهند. این الگوریتم، قادر است مدل برچسب‌گذاری را یاد بگیرد و با استفاده از سیستم یادگیری هوشمند، به نمونه‌های جدید برچسب بزند و آنها را تفکیک کند. این تفکیک، نوعی یادگیری به حساب می‌آید و الگوریتم بعد از این یادگیری، می‌تواند مدل خود را بر روی داده‌های جدید اعمال کند.

۲- **خوشه‌بندی**[2]: در این مورد، الگوریتم داده‌ها را بر اساس ذات آنها گروه‌بندی می‌کند. مثلاً مشتریان یک فروشگاه اینترنتی را بر اساس شباهت‌هایی که دارند (سن، جنس، میزان تحصیلات و...)، به خوشه‌های مختلف تقسیم می‌کند.

۳- **یادگیری تقویتی**[3]: در این نوع از یادگیری، الگوریتم، به‌وسیله تبادل اطلاعات و عملیات با محیط اطراف، به‌طور پیوسته به کشف اطلاعات و یادگیری اقدام می‌کند. به‌عنوان مثال، الگوریتمی را در نظر بگیرید که به‌وسیله تعامل با محیط و شبیه‌سازی آن به‌صورت هوشمند، به طراحی انواع مختلف فرم‌های سبد خرید می‌پردازد تا بهترین طراحی را برای مشتریان ایجاد کرده و در نهایت میزان فروش و سود را افزایش دهد.

روش انتخاب الگوریتم‌های داده‌کاوی

برای اینکه پژوهشگران یا تحلیلگران بدانند چگونه یک الگوریتم را انتخاب کنند، راهنمای مشخصی وجود ندارد و همین امر برای آنها یک چالش در علم داده‌کاوی محسوب می‌شود. انتخاب یک الگوریتم مشخص، کاری بسیار پیچیده است که گاهی برای نتیجه بهتر، از چندین الگوریتم استفاده کرده و پردازش‌ها را با الگوریتم‌های مختلف تکرار می‌کنند.

جناب ویگر که یکی از فعالان علم داده‌کاوی می‌باشد، در مقاله‌ای توصیه کرده؛ درصورت امکان، بهتر است قبل از انجام پردازش روی مجموعه داده‌های حقیقی، الگوریتم را روی یک مجموعه داده ورودی به‌صورت آزمایشی پیاده کنید تا عملکرد الگوریتم برای حل یک نوع مسئله خاص، سنجیده شود.

چندین عامل تأثیرگذار در انتخاب الگوریتم مناسب

اولین گزینه هدف مسئله است که در آن مسائلی از قبیل: گرفتن دلایل چرایی،کاوش داده‌ها و ماهیت مسئله که قصد حل کردن موضوع را دارد، در نظر گرفته می‌شود.

مورد بعدی در انتخاب یک الگوریتم، ساختار داده می‌باشد که در تعیین الگوریتم مورد استفاده نقش بسیار مهمی دارد؛ زیرا ارتباط بین داده‌ها، متغیرها و روشی که داده‌ها بر اساس آن ذخیره می‌شوند را مشخص می‌کند.

سومین گزینه، نتایج مورد انتظار است که هر فرایند داده‌کاوی به یک خروجی به‌عنوان راه حل مسئله نیاز دارد تا با توجه به نوع نتایج مورد انتظار، الگوریتمی را انتخاب کند که قادر به تولید آن نتایج باشد. زیرا نتیجه یک فرایند داده‌کاوی، موفقیت یا شکست آن را تعیین می‌کند؛ چراکه الگوریتمی که استفاده می‌شود باید یک الگوی مورد انتظار در خروجی ارائه دهد تا نتایج مورد نظر بر اساس وظیفه‌ای که در داده‌کاوی نیاز به انجام آن است، تولید کند. همچنین باید آن‌قدر سریع باشد تا در حجم انبوهی از داده‌ها به جست‌وجو بپردازد.

مورد بعدی، وظایفی است که ضمن فرایند داده‌کاوی انجام می‌شوند.

مورد پنجم، آشنایی داده‌کاوان با الگوریتم است که برای انجام این کار به درک کاملی از الگوریتم نیاز دارند. اگر از یک الگوریتم، بدون درک جزئیات عملکرد آن و تنها با داشتن درک کلی از ویژگی‌هایش استفاده کند، امکان دارد الگوریتم انتخاب شده برای وظایفی که نیاز به انجام است، مناسب نباشد.

در انتخاب یک الگوریتم باید به پیکربندی پایگاه داده مورد استفاده، توجه داشته باشند. در واقع، باید الگوریتمی انتخاب کنند که پایگاه داده آن یکپارچه باشد تا از هزینه‌های اضافی جلوگیری شود.

مورد آخر، داشتن یک مدل‌سازی فرایند است. زیرا با استفاده از مدل ساخته شده، داده‌کاو می‌تواند تصمیم بگیرد که در هر مرحله باید از چه روش‌ها و الگوریتم‌هایی استفاده شود و سپس به انتخاب الگوریتم مناسب برای هر مرحله بپردازد.

مهم‌ترین الگوریتم‌های داده‌کاوی

کلاس‌بندی و خوشه‌بندی، روش‌هایی هستند که برای تحلیل داده‌ها به‌کار می‌روند. در این روش‌ها برخی از الگوریتم‌هایی که استفاده می‌شوند، معرفی می‌شوند.

K-means

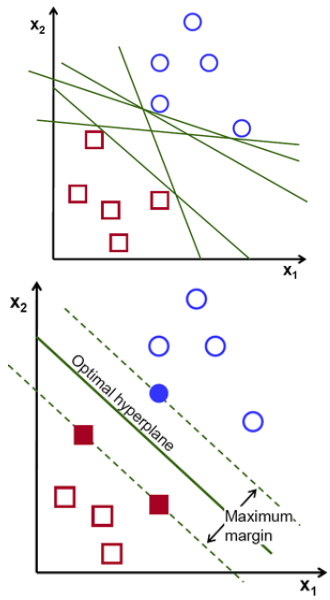
در این روش، ابتدا تعداد دلخواه K نقطه را به‌طور تصادفی از میان نقاط موجود انتخاب کرده و به‌عنوان مرکز خوشه‌ها[4] در نظر می‌گیریم. در واقع k تعداد خوشه‌ها نیز محسوب می‌شود. سپس، فاصله هر نقطه را تا مرکز خوشه‌ها به‌دست می‌آوریم. نقاط نزدیک به هر مرکز خوشه، متعلق به آن خوشه هستند. بنابراین، نوع خوشه‌بندی و موقعیت هر نقطه تغییر می‌کند. در مراحل بعدی میانگین نقاط، به‌عنوان مرکز خوشه در نظر گرفته شده و این روند آن‌قدر تکرار می‌شود تا موقعیت نقاط ثابت شود و خوشه‌ها تغییری نکنند. هر خوشه در داده‌کاوی، مجموعه‌ای از نقاطی است که بیشترین ویژگی‌های مشابه را در مجموعه داده‌های ورودی دارند. از این الگوریتم، برای خوشه‌بندی داده‌ها استفاده شده و یکی از اصلی‌ترین الگوریتم‌های داده‌کاوی محسوب می‌شود.

ماشین بردار پشتیبان

الگوریتم ماشین بردار پشتیبان[5] (SVM)، کاربردهای زیادی درحوزه یادگیری ماشین دارد و کاربرد آن در تحلیل داده‌هایی است که برای روش‌های کلاس‌بندی و رگرسیون[4] مورد استفاده قرار می‌گیرند. مجموعه‌ای از نقاط در فضای داده‌ای موجود، مسئول مرزبندی و دسته‌بندی داده‌ها هستند. هر ماشین بردار پشتیبان، با استفاده از معیار خود که بردارهای پشتیبان هستند، دسته‌بندی نقاط را انجام می‌دهد. این الگوریتم، برای توصیف کلاس‌بندی داده‌ها به‌کار می‌رود.

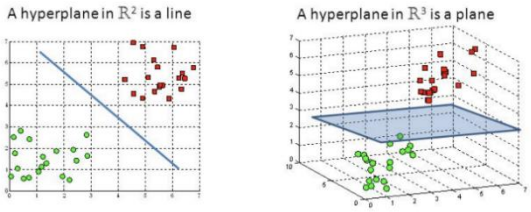
هدف از الگوریتم ماشین بردار پشتیبان، یافتن یک ابر صفحه در یک فضای N بُعدی (N- تعداد ویژگی‌ها) است که به‌طور مشخص نقاط داده را طبقه‌بندی می‌کند.

[1] Classification
[2] Clustering
[3] Reinforcement Learning
[4] Centroid
[5] Support Vector Machines(SVM)



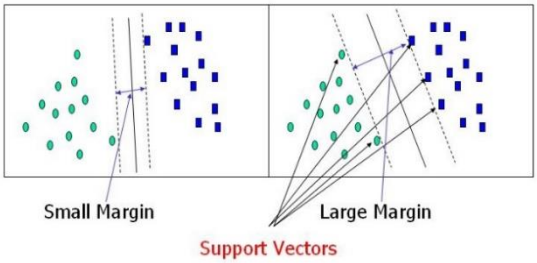
شکل ۱ طبقه‌بندی داده‌ها با استفاده از ماشین بردار پشتیبان.

برای جدا کردن دو دسته از نقاط داده، ابرصفحه‌های ممکن زیادی وجود دارد که می‌توان انتخاب کرد. هدف ما این است که؛ صفحه‌ای را پیدا کنیم که حداکثر حاشیه (یعنی حداکثر فاصله بین نقاط داده هر دو کلاس) را داشته باشد. به حداکثر رساندن فاصله حاشیه، مقداری تقویت را فراهم می‌کند تا بتوان نقاط داده آینده را با اطمینان بیشتری طبقه‌بندی کرد.



شکل ۲ یافتن ابرصفحه برای طبقه‌بندی داده‌ها با استفاده از ماشین بردار پشتیبان.

ابرصفحه‌ها مرزهای تصمیم‌گیری هستند که به طبقه‌بندی نقاط داده کمک می‌کنند. نقاط داده‌ای که در دو طرف ابرصفحه قرار می‌گیرند را می‌توان به کلاس‌های مختلف نسبت داد. همچنین، ابعاد ابرصفحه به تعداد ویژگی‌ها بستگی دارد. اگر تعداد ویژگی‌های ورودی ۲ باشد، آنگاه ابرصفحه فقط یک خط است. اگر تعداد ویژگی‌های ورودی ۳ باشد، ابرصفحه به یک صفحه دوبُعدی تبدیل می‌شود. تصور زمانی که تعداد ویژگی‌ها از ۳ بیشتر شود، دشوار خواهد شد.



شکل ۳ ابرصفحه با بیشترین فاصله بین صفحات.

بردارهای پشتیبان، نقاط داده‌ای هستند که به ابرصفحه نزدیک‌تر هستند و بر موقعیت و جهت ابرصفحه تأثیر می‌گذارند. با استفاده از این بردارهای پشتیبانی، حاشیه طبقه‌بندی کننده را به حداکثر می‌رسانیم. حذف بردارهای پشتیبانی موقعیت هایپرپلان را تغییر می‌دهد. اینها نکاتی هستند که به ما در ساخت SVM کمک می‌کنند.

در رگرسیون لجستیک، خروجی تابع خطی را می‌گیریم و مقدار را در محدوده [۰,۱] با استفاده از تابع سیگموئید تصمیم‌گیری می‌کنیم. اگر مقدار تصمیم‌گیری شده بزرگ‌تر از یک مقدار آستانه (۰.۵) باشد، به آن برچسب ۱ و در غیر این صورت، یک برچسب ۰ به آن اختصاص می‌دهیم. در SVM، خروجی تابع خطی را می‌گیریم و اگر آن خروجی بزرگ‌تر از ۱ باشد، شناسایی می‌کنیم. آن را با یک کلاس و اگر خروجی ۱- باشد، با کلاس دیگری شناسایی می‌کنیم. ازآنجاکه مقادیر آستانه در SVM به ۱ و ۱- تغییر می‌کنند، این محدوده تقویتی مقادیر یک و منفی یک را به‌دست می‌آوریم که به‌عنوان حاشیه عمل می‌کند.

ما در الگوریتم SVM، در پیِ به حداکثر رساندن حاشیه بین نقاط داده و هایپرپلن هستیم. تابع ضرری که به حداکثر رساندن حاشیه کمک می‌کند.

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases}$$

$$C(x, f(x)) = max\ (0, 1 - y * f(x))$$

اگر مقدار پیش‌بینی شده و مقدار واقعی علامت یکسانی داشته باشند، بهای تمام شده صفر است. اگر آنها نباشند، سپس ارزش ضرر را محاسبه می‌کنیم. ما همچنین یک پارامتر تنظیم به تابع هزینه اضافه می‌کنیم. هدف، پارامتر تنظیم تعادل بین حداکثر کردن و ضرر است. پس از افزودن پارامتر تنظیم، توابع هزینه به‌صورت زیر است:

$$min_w \lambda \|w\|^2 + \sum_{i=1}^{n} max\ (0, 1 - y * f(x))$$

اکنون که تابع ضرر را داریم، مشتقات جزئی را با توجه به وزن‌ها در نظر می‌گیریم تا شیب‌ها را پیدا کنیم. با استفاده از گرادیان‌ها می‌توانیم وزن‌های خود را به‌روز کنیم.

$$\frac{\delta}{\delta w_k} \lambda \parallel w \parallel^2 = 2\lambda w_k$$

$$\frac{\delta}{\delta w_k} \left(1 - y_i \langle x_i, w \rangle\right)_+ = \begin{cases} 0, & \text{if } y_i \langle x_i, w \rangle \geq 1 \\ -y_i x_{ik}, & \text{else} \end{cases}$$

هنگامی که طبقه‌بندی اشتباهی وجود ندارد، یعنی مدل ما به‌درستی کلاس نقطه داده را پیش‌بینی می‌کند و فقط باید گرادیان را از پارامتر تنظیم به‌روز کنیم.

$$w = w - \alpha \cdot (2\lambda w)$$

هنگامی که یک طبقه‌بندی اشتباه وجود دارد، یعنی مدل ما در پیش‌بینی کلاس نقطه داده اشتباه می‌کند و در این صورت، از دست دادن را به‌همراه پارامتر تنظیم برای انجام به‌روزرسانی گرادیان لحاظ می‌کنیم.

$$w = w + \alpha \cdot (y_i \cdot x_i - 2\lambda w)$$

Naive Bayes

الگوریتم نیو بیز، ازجمله الگوریتم‌های کلاس‌بندی است که بر مبنای تکنیک‌های دسته‌بندی احتمالی است. این الگوریتم آماری از قاعده بیز در ریاضیات استفاده کرده و با تعیین متغیرهای مستقلی اقدام به مشخص کردن احتمال وقوع و دسته‌بندی داده‌ها می‌کند. این، تنها یکی از الگوریتم‌های خانواده بیز است که در تحلیل داده‌ها به‌کار می‌رود. الگوریتم مذکور، در کلاس‌بندی و بازیابی متن کاربرد زیادی دارد و قابلیت پیش‌بینی رفتار کاربران را برای کسب‌وکارها فراهم می‌کند.

روش‌های ساده بیز، مجموعه‌ای از الگوریتم‌های یادگیری نظارت‌شده بر اساس اعمال قضیه بیز با فرض «ساده‌انگیز» استقلال شرطی بین هر جفت ویژگی با توجه به مقدار متغیر کلاس هستند.

با وجود مفروضات ظاهراً بیش از حد ساده شده، طبقه‌بندی کننده‌های ساده بیز در بسیاری از موقعیت‌های دنیای واقعی، معروف به طبقه‌بندی اسناد و فیلتر هرزنامه، به‌خوبی عمل کرده‌اند. آنها به مقدار کمی از داده‌های آموزشی برای تخمین پارامترهای لازم نیاز دارند.

فراگیرای و طبقه‌بندی کننده‌های ساده بیز می‌توانند در مقایسه با روش‌های پیچیده‌تر، بسیار سریع باشند. جداسازی توزیع‌های ویژگی شرطی کلاس به این معنی است که؛ می‌توان هر توزیع را به‌طور مستقل به‌عنوان یک توزیع یک‌بُعدی تخمین زد. این امر، به‌نوبه خود در کاهش مشکلات ناشی از ابعاد، مؤثر خواهد بود.

Apriori

آپریوری[1] الگوریتم محبوبی است که می‌تواند داده‌های مرتبط با هم را پیدا کرده و میزان وابستگی را در هر دسته مشخص کند. این الگوریتم کلاسیک، با استفاده از قوانین وابستگی آیتم‌های ورودی را دریافت کرده (که به‌عنوان مثال در یک پایگاه داده این آیتم‌ها ممکن است تراکنش‌های مشتریان باشد)، سپس دسته‌بندی را انجام می‌دهد. این الگوریتم تا جایی ادامه پیدا می‌کند که بین دسته‌بندی‌های مختلف، آیتم مشابه دیگری وجود نداشته باشد.

رتبه‌بندی صفحه

این الگوریتم، همان‌طور که از نامش پیدا است، برای رتبه‌بندی صفحات وب‌سایت‌ها به‌کار می‌رود. موتورهای جست‌وجوی گوگل از این الگوریتم برای شناسایی میزان اهمیت صفحات وب و رتبه‌بندی آنها در نمایش به کاربران استفاده می‌کنند. بنابراین، یکی دیگر از کاربردهای این الگوریتم را می‌توان در حوزه سئو دانست. با استفاده از آمار تعداد لینک‌های ورودی به یک سایت و میزان کیفیت آنها، به بررسی و مقایسه وب‌سایت‌ها می‌پردازد.

رگرسیون

این الگوریتم، ازجمله روش‌های آماری برای تعیین روابط میان داده‌ها است که با استفاده از داده‌های پیشین، مدل‌های ریاضیاتی را استخراج کرده و برای پیش‌بینی ارزش داده‌هایی که در آینده تولید می‌شوند، به‌کار می‌برد. این دسته از الگوریتم‌ها انواع مختلفی مانند خطی، چندگانه و غیره دارند و با تکیه بر منطق ریاضیاتی، در بررسی و مدل‌سازی متغیرهایی برای تحلیل داده‌ها بسیار کاربردی هستند. این الگوریتم‌ها برای کلاس‌بندی داده‌ها نیز به‌کار می‌روند.

گذشته از این، اگر برای شناسایی و پیش‌بینی متغیر وابسته فقط از یک متغیر مستقل استفاده شود، مدل را رگرسیون خطی ساده (Simple Linear Regression) می‌گویند. فرم مدل رگرسیون خطی ساده به‌صورت زیر است:

$$Y=\beta 0+\beta 1X+\epsilon$$

برای مثال، فرض کنید کارخانه‌ای می‌خواهد میزان هزینه‌هایش را بر اساس ساعت کار برآورد کند. شیب خط حاصل از برآورد نشان می‌دهد به‌ازای یک ساعت افزایش ساعت کاری چه میزان بر هزینه‌هایش افزوده خواهد شد. از طرفی، عرض از مبدأ خط رگرسیون نیز هزینه ثابت کارخانه را حتی در زمانی که ساعت کاری نیست، نشان می‌دهد. این هزینه را می‌توان به‌عنوان هزینه‌های ثابت مانند دستمزد نگهبانان و هزینه روشنایی فضای کارخانه فرض کرد.

گاهی مدل رگرسیونی را بدون عرض از مبدأ در نظر می‌گیرند و $\beta 0=0$ محسوب می‌کنند. این کار به این معنی است که؛ با صفر شدن مقدار متغیر مستقل، مقدار متغیر وابسته نیز باید صفر در نظر گرفته شود. زمانی که محقق مطمئن باشد که خط رگرسیون باید از مبدأ مختصات عبور کند، این‌گونه مدل در نظر گرفته می‌شود. فرم مدل رگرسیونی در این حالت به‌صورت زیر است:

$$Y=\beta 1X+\epsilon$$

باید توجه داشت که منظور از رابطه خطی در مدل رگرسیون، وجود رابطه خطی بین ضرایب است نه بین متغیرهای مستقل. برای مثال این مدل $y=\beta 0+\beta 1x2+\epsilon y=\beta 0+\beta 1x2+\epsilon$ را نیز می‌توان مدل خطی در نظر گرفت، درحالی‌که مدل $y=\beta 0x\beta 1+\epsilon y=\beta 0x\beta 1+\epsilon$ دیگر خطی نیست و به مدل نمایی شهرت دارد.

به‌عنوان مثال تعدادی مقدار برای متغیر وابسته بر اساس مقدار $x=65$ وجود دارد که شکل توزیع فراوانی آنها به‌صورت نرمال با میانگین $\beta 0+\beta 1\times 65\beta 0+\beta 1\times 65$ است. همچنین، برای نقطه ۹۰ نیز مقدار پیش‌بینی یا برآورد برای متغیر وابسته به‌صورت $\beta 0+\beta 1\times 90\beta 0+\beta 1\times 90$ خواهد بود. در هر دو حالت، واریانس خطا یا واریانس مقدارهای پیش‌بینی شده (پهنای منحنی زنگی شکل) ثابت است.

به‌منظور برآورد پارامترهای رگرسیون خطی ساده، کافی است تابع مجموع مربعات خطا را کمینه کرد.

شبکه‌های عصبی

یکی از بهترین الگوریتم‌های داده‌کاوی در حل مسائل پیچیده، الگوریتم شبکه‌های عصبی[2] است که علاوه‌بر داده‌کاوی، در حوزه‌هایی مانند یادگیری ماشین و یادگیری عمیق نیز بسیار مورد بحث است. این الگوریتم نیز با یافتن شباهت‌های بین داده‌ها اقدام به برچسب‌گذاری و کلاس‌بندی آنها کرده و مدل‌های مختلفی را جهت تحلیل داده‌ها ارائه می‌دهد. الگوریتم‌های شبکه عصبی علاوه‌بر حوزه کسب‌وکار در پیش‌بینی نرخ بازار سهام و مسائل اقتصادی نیز مورد توجه هستند.

KNN

الگوریتم نزدیک‌ترین همسایه[3] با گرفتن هر داده جدید، آن را با داده‌های قبلی مقایسه کرده و در دسته‌ای قرار می‌دهد که داده‌های جدید و قدیم بیشترین شباهت را داشته باشند. در واقع، در دسته‌ای قرار می‌گیرد که شباهت بیشتری با داده‌های اطراف و به عبارتی همسایگان نزدیکش داشته باشد. این الگوریتم غیرپارامتری است و فرضیات تحلیلی خود را بر مبنای مدل قبلی توزیع داده‌ها قرار نمی‌دهد. الگوریتم مورد نظر، ازجمله روش‌های کلاس‌بندی داده‌ها است.

فرایند داده‌کاوی

داده‌کاوی به‌صورت کلی و عمومی در ۶ مرحله اصلی انجام می‌شود؛ در ابتدا داده‌های مورد نیاز (داده‌های هدف) جمع‌آوری شده و مورد پردازش و پاک‌سازی قرار می‌گیرند، یعنی داده‌های اضافه حذف شده و تنها داده‌های مورد نیاز وارد سیستم می‌شوند.

در مرحله‌ی بعد، الگوی میان داده‌ها کشف و ارزیابی و سپس الگوریتم و روش‌های داده‌کاوی بر روی داده‌ها انجام خواهد شد.

در نهایت نیز اطلاعات به‌دست‌آمده از فرایند داده‌کاوی در قالب فرمت‌های قابل درک برای انسان مانند نمودار، تصویر، گزارش و... ارائه شده

[1] Apriori

[2] Neural Network

[3] K–Nearest Neighbors

و دانش مورد نظر که از میان انبوه داده‌های خام استخراج شده است در اختیار سازمان قرار خواهد گرفت.

در شکل ۴ به ترتیب فرایند مربوط به داده‌کاوی آورده شده است.

شکل ۴ فرایند مربوط به داده‌کاوی.

فرایند داده‌کاوی، شامل چندین گام است. این فرایند از داده‌های خام آغاز می‌شود و تا شکل‌دهی دانش جدید ادامه دارد. فرایند بازگشتی داده‌کاوی شامل گام‌های زیر است:

پاک‌سازی داده[1]: پاک‌سازی یا تمیز کردن داده‌ها به فرایندی جهت تشخیص، حذف و اصلاح داده‌های نادرست از مجموعه جداول، رکوردها، یا بانک‌های اطلاعاتی همچنین شناسایی قسمت‌های ناقص و نادرست داده‌ها و سپس اصلاح و جایگزینی آنها اشاره دارد. هدف از پاک‌سازی داده‌ها، استخراج اطلاعات دقیق و درست است. چراکه اطلاعات نادرست می‌تواند منجر به نتیجه‌گیری غلط شود و کسب‌وکار شما را با مشکل روبه‌رو کند.

یکپارچه‌سازی داده[2]: یکپارچه‌سازی اطلاعات، یک بینش نسبتاً جدید در رابطه با مشتریان، محصولات، کانال‌های بازاریابی و... ایجاد کرده و بستر مناسب برای نگرش جامع و کامل به عناصر اصلی کسب‌وکار را در یک سازمان فراهم می‌کند. بدون یکپارچه‌سازی داده‌ها نمی‌توانید در بازار رقابتی امروز حرف زیادی برای گفتن داشته باشید.

انتخاب داده[3]: در بخش انتخاب، باید داده‌های مرتبط با تحلیل داده‌ها انتخاب شده و از مجموعه داده‌ها برای انجام تحلیل‌ها بازیابی شوند. یک انتخاب اصولی و درست می‌تواند منجر به بهبود یادگیری استقرایی از جهات گوناگون ازجمله سرعت یادگیری و ظرفیت تعمیم شود.

تبدیل داده[4]: گاهی برای اینکه دقت تجزیه‌وتحلیل را بالا ببریم باید در داده‌های خامی که برای تحلیل در دسترس ما قرار دارند، تغییراتی ایجاد کنیم. یکی از این تغییرات، فرایند تبدیل داده‌ها است. تبدیل داده‌ها، روش‌هایی بر پایه ریاضی است که برای متغیرهایی به‌کار می‌رود که از شاخص‌های آماری نرمال بودن، خطی بودن، پراکندگی یکسان و... پیروی نمی‌کنند.

تبدیل داده نوعی روش تثبیت داده نیز به‌شمار می‌رود. در این فاز، داده‌های انتخاب شده به فرم دیگری تبدیل می‌شود. این کار به‌سادگی، درستی و دقت بیشتر داده‌کاوی کمک می‌کند.

داده‌کاوی: در این بخش از روش‌های هوشمندانه برای استخراج الگوهای مهم و اثرگذار از میان داده‌ها استفاده می‌شود.

ارزیابی الگو[5]: در این قسمت، الگوهای به‌دست‌آمده در بخش قبل از جنبه‌های گوناگون مانند دقت، صحت، قابلیت تعمیم و... مورد بررسی و ارزیابی قرار می‌گیرد.

ارائه دانش[6]: داده‌کاوی، در نهایت به ارائه دانش ختم می‌شود. دانش به‌دست‌آمده در این بخش به شیوه‌ای مشخص و قابل فهم به کاربر ارائه می‌شود. البته برای اثرگذاری بیشتر، روش‌های بصری‌ساز نیز مورد استفاده قرار می‌گیرد که با وجود این روش‌ها، کاربران در درک و تفسیر نتایج داده‌کاوی موفق‌تر خواهند بود

۲- بیان مساله

همه‌گیری کووید-۱۹ کل جهان را تحت‌تأثیر قرار داد. سیستم‌های مراقبت‌های بهداشتی، برای چنین درخواست شدید و طولانی جهت تخت‌های ICU، متخصصان، تجهیزات حفاظت شخصی و منابع مراقبت‌های بهداشتی آمادگی نداشتند. برزیل، اولین مورد COVID-19 را در ۲۶ فوریه ثبت کرد و در ۲۰ مارس به انتقال و همه‌گیری جامعه رسید. در این فصل، به پیاده‌سازی روش‌های مربوط به داده‌کاوی به‌منظور تشخیص هرچه بهتر کووید-۱۹ پرداخته می‌شود.

داده‌های کووید-۱۹

برای به‌دست آوردن اطلاعات دقیق به‌منظور پیش‌بینی و آماده‌سازی بهتر سیستم‌های مراقبت‌های بهداشتی و جلوگیری از فروپاشی سازمانی با توجه به نیاز ظرفیت بالای تخت‌های ICU (به فرض در دسترس بودن منابع انسانی، PPE و متخصصان)، استفاده از داده‌های بالینی – فردی، به‌جای داده‌های اپیدمیولوژیک و جمعیتیِ تعریف شده است، ضروری به نظر می‌رسد.

هدف‌های تحقیقاتی این مجموعه داده به دو صورت زیر شرح داده شده است.

هدف اول:

پیش‌بینی پذیرش در ICU موارد تأیید شده COVID-19، بر اساس داده‌های موجود. آیا می‌توان پیش‌بینی کرد که کدام بیماران به حمایت بخش مراقبت‌های ویژه نیاز دارند؟ هدف، ارائه دقیق‌ترین پاسخ به بیمارستان‌های ثالثیه و سه‌ماهه است تا بتوان منابع ICU را ترتیب داد و یا برای انتقال بیمار برنامه‌ریزی کرد.

هدف دوم:

پیش‌بینی شود که موارد تأیید شده COVID-19 در ICU بستری نشود. بر اساس نمونه فرعی از داده‌های به‌طور گسترده در دسترس، آیا می‌توان پیش‌بینی کرد که کدام بیماران به حمایت بخش مراقبت‌های ویژه نیاز دارند؟

هدف این است که؛ به بیمارستان‌های محلی و موقت پاسخ کافی ارائه شود تا پزشکان خط‌مقدم بتوانند با خیال راحت این بیماران را ترخیص کرده و از راه دور آنها را پیگیری کنند.

این مجموعه داده، حاوی داده‌های ناشناس از بیمارستانی در برزیل است. همه داده‌ها با پیروی از بهترین شیوه‌ها و توصیه‌های بین‌المللی، ناشناس شدند. داده‌ها با ستونی مطابق با Min Max Scaler تمیز و مقیاس‌بندی شده‌اند تا بین ۱- و ۱ قرار گیرند.

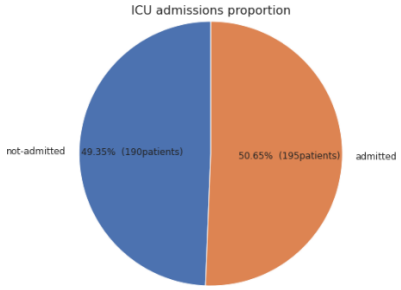
[1] Data Cleaning
[2] Data Integration
[3] Data Selection
[4] Data Transformation
[5] Pattern Evaluation
[6] Knowledge Representation

بیشترین، کمترین، واریانس و... مجموعه داده به‌صورت شکل ۵ نشان داده شده است.

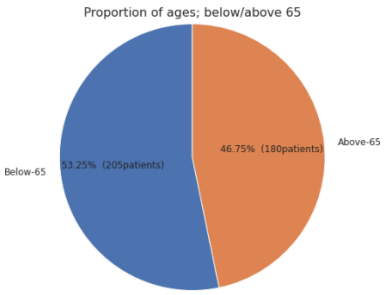
	count	mean	std	min	25%	50%	75%	max
PATIENT_VISIT_IDENTIFIER	1925.0	192.000000	111.168431	0.0	96.0	192.000000	288.000000	384.0
AGE_ABOVE65	1925.0	0.467532	0.499074	0.0	0.0	0.000000	1.000000	1.0
GENDER	1925.0	0.368831	0.482613	0.0	0.0	0.000000	1.000000	1.0
DISEASE GROUPING 1	1920.0	0.108333	0.310882	0.0	0.0	0.000000	0.000000	1.0
DISEASE GROUPING 2	1920.0	0.028125	0.165373	0.0	0.0	0.000000	0.000000	1.0
...	...	...	...	...	...	...	...	...
HEART_RATE_DIFF_REL	1240.0	-0.817800	0.270217	-1.0	-1.0	-0.989822	-0.662529	1.0
RESPIRATORY_RATE_DIFF_REL	1177.0	-0.719147	0.446600	-1.0	-1.0	-1.000000	-0.634409	1.0
TEMPERATURE_DIFF_REL	1231.0	-0.771327	0.317694	-1.0	-1.0	-0.975824	-0.594677	1.0
OXYGEN_SATURATION_DIFF_REL	1239.0	-0.886982	0.296772	-1.0	-1.0	-0.980333	-0.880155	1.0
ICU	1925.0	0.267532	0.442787	0.0	0.0	0.000000	1.000000	1.0

229 rows × 8 columns

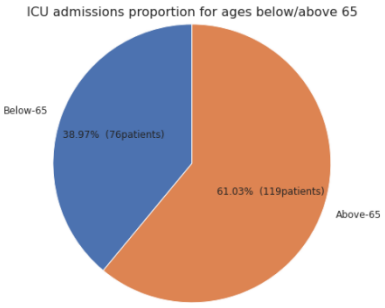
شکل ۵ - نسبت پذیرش در ICU به شرح زیر است.



شکل ۶

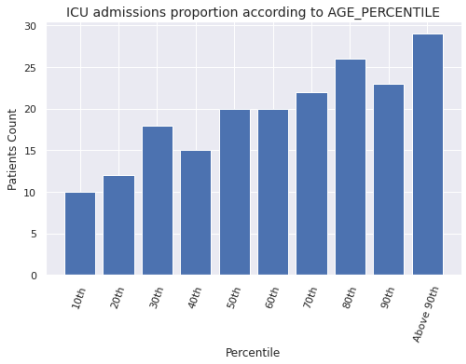


شکل ۷



شکل ۸

میزان پذیرش در ICU بر اساس سنین مختلف به‌صورت نمودار بلوکی در شکل ۹ نشان داده شده است. همان‌طور که در این بلوک دیاگرام می‌بینیم، هرچه سنین بالاتر می‌رود احتمال بستری بیماران در ICU بیشتر می‌شود.



شکل ۹

در جدول ۱ توزیع تعداد پذیرش‌های ICU نشان داده شده است.

جدول ۱ - توزیع تعداد پذیرش‌های ICU

	icu_count	gender_count	above65_count
0	1410	863	569
1	515	352	331

توزیع همه داده‌ها به‌صورت شکل ۱۰ نشان داده شده است.

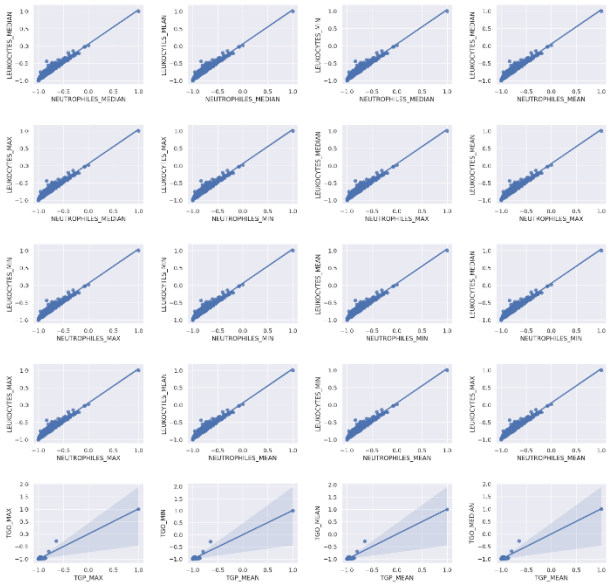
شکل ۱۰

پیش‌پردازش بر روی داده‌ها

مشکل: یکی از چالش‌های اصلی کار با داده‌های مراقبت‌های بهداشتی این است که؛ نرخ نمونه‌گیری در انواع مختلف اندازه‌گیری‌ها متفاوت است. به‌عنوان مثال، نمونه‌برداری از علائم حیاتی بیشتر (معمولاً ساعتی) نسبت به آزمایشگاه‌های خون (معمولاً روزانه) انجام می‌شود.

نکات و ترفندها: منطقی است فرض کنیم بیماری که اندازه‌گیری در یک پنجره زمانی ثبت نشده، ازنظر بالینی پایدار است و به‌طور بالقوه علائم حیاتی و آزمایشگاه خون مشابه پنجره‌های مجاور را ارائه می‌دهد. بنابراین، ممکن است با استفاده از ورودی بعدی یا قبلی، مقادیر ازدست‌رفته را پر کنید. هنگام انتخاب الگوریتم خود به مسائل چندخطی و واریانس صفر در این داده‌ها توجه می‌شود.

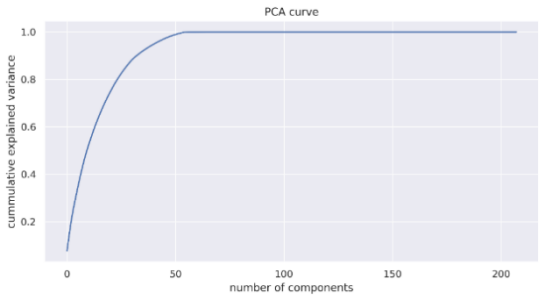
در شکل ۱۱ همبستگی مربوط به ویژگی‌های مجموعه داده نشان داده شده است.



شکل ۱۱

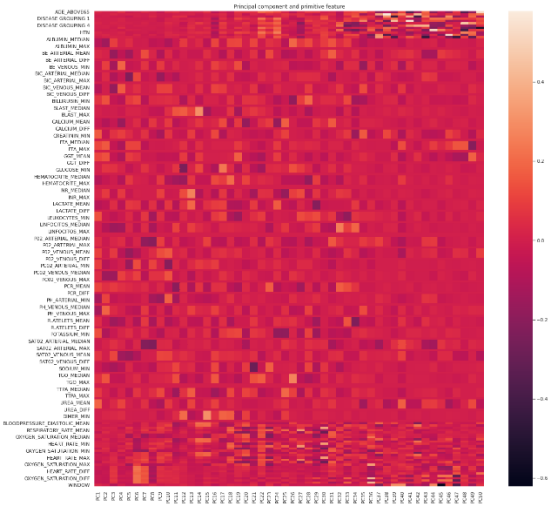
سپس، داده‌ها نرمال‌سازی می‌شوند.

بعد از آن، عملیات PCA بر روی مجموعه داده انجام می‌شود که به‌صورت شکل ۱۲ نشان داده شده است.



شکل ۱۲

نمایش هیت‌مپ مربوط به مجموعه داده بعد از اعمال PCA به‌صورت شکل ۱۳ نشان داده شده است.



شکل ۱۳

آموزش و ارزیابی مجموعه داده

آموزش مدل با استفاده از مدل‌های مختلف انجام شده که دقت مدل‌ها بر روی مجموعه داده به شرح زیر است.

	LogisticRegression	LinearSVC	KNeighbors	RandomForest	GradientBoosting	ExtraTree	XGBoost
cross_val_score	0.711035	0.715126	0.685097	0.715132	0.702105	0.719984	0.707789
auc	0.510139	0.511377	0.541672	0.522965	0.514538	0.540955	0.512392
precision	0.370370	0.372881	0.382653	0.423077	0.354839	0.524390	0.367647
recall	0.058480	0.064327	0.219298	0.096491	0.096491	0.125731	0.073099
accuracy	0.711039	0.710227	0.685065	0.712662	0.700487	0.725649	0.707792
f1_score	0.101010	0.109726	0.278810	0.157143	0.151724	0.202830	0.121951

شکل ۱۴

دقت دسته‌بندی بر روی داده‌های اعتبارسنجی مجموعه داده در زیربخش‌های زیر ارائه شده است.

RandomForestClassifier

دقت مدل ۷۲ درصد است.

KNeighborsClassifier

ماتریس درهم‌ریختگی مدل به‌صورت زیر است و دقت مدل ۶۶ درصد است.

۱۸۸	۳۴
۶۸	۱۸

LogisticRegression

دقت مدل ۶۹ درصد است.

ماتریس درهم‌ریختگی مدل به‌صورت زیر است.

۲۱۱	۱۱
۸۴	۲

XGBClassifier

دقت مدل ۷۱ درصد است.

ماتریس درهم‌ریختگی مدل به‌صورت زیر است.

۲۱۳	۹
۸۰	۶

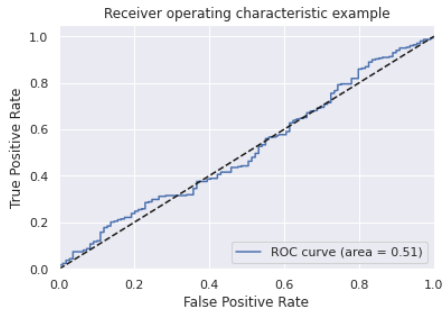
LinearSVC

دقت مدل ۶۹ درصد است.

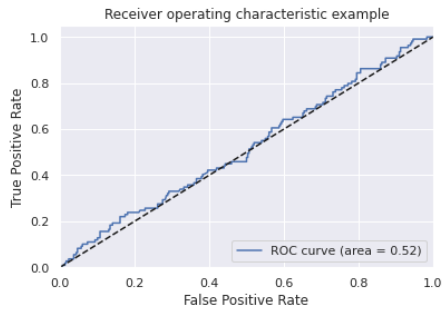
ماتریس درهم‌ریختگی مدل به‌صورت زیر است.

۲۱۱	۱۱
۸۴	۲

نمودار ROC به‌صورت زیر است.

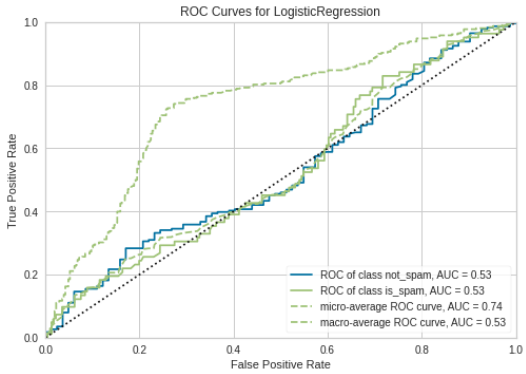


شکل ۱۵



شکل۱۶

نمودار ROC برای دسته‌های مختلف مجموعه داده به‌صورت زیر است.



شکل ۱۷

۳- روش حل

شیوع کنونی همه‌گیری کووید-۱۹ منجر به مرگ میلیون‌ها نفر در سراسر جهان، فشار بر سیستم‌های مراقبت‌های بهداشتی کشورهای توسعه‌یافته و فروپاشی اقتصاد کشورهای کم‌درآمد شده است. اگرچه اکثریت آنها معمولاً خفیف تا متوسط هستند، اما بیماری کووید-۱۹ می‌تواند تا مرز ذات‌الریه شدید (یعنی سندرم تنفسی حاد شدید) و مرگ نیز پیشرفت کند. در زمان بستری شدن در بیمارستان، ارزیابی وضعیت بیماران COVID-19 برای مدیریت بالینی آنها حیاتی است. به‌ویژه هنگامی که با منابع و پرسنل محدود بیمارستانی سروکار داریم. استفاده از روش‌های هوش مصنوعی به‌منظور تشخیص و دسته‌بندی کووید-۱۹ به‌کار گرفته شده است که در این فصل به بررسی مقاله‌های ارائه شده در این زمینه پرداخته می‌شود.

مقالات ارائه شده

- نشانگرهای زیستی ذات‌الریه شدید COVID-19 در هنگام پذیرش

با استفاده از داده‌کاوی با استفاده از مجموعه داده‌های آزمایش خون آزمایشگاهی رایج در تحقیقات اپیدمیولوژیک COVID-19، هوش مصنوعی یک رویکرد منحصربه‌فرد برای پیش‌بینی در مورد شدت بیماری برای مدیریت بیماران COVID-19 است. بااین‌حال، محدودیت هوش مصنوعی، خطر بالای سوگیری است. در این مقاله [۱] مهارت داده‌کاوی و یادگیری ماشین، دو شکل پیشرفته هوش مصنوعی برای پیش‌بینی ذات‌الریه شدید COVID-19 بر اساس آزمایش‌های معمول آزمایشگاهی، بررسی شده است. نمونه‌ای متشکل از ۴۰۰۹ بیمار کووید-۱۹ بر اساس هیپوکسی خون در بدو بستری به دو گروه شدید ($PaO2 < 60$ میلی‌متر جیوه، ۴۸۹ مورد) و غیرشدید ($PaO2 \geq 60$ میلی‌متر جیوه، ۳۵۲۰ مورد) تقسیم شدند و مجموعه داده‌های آزمایشگاهی آنها توسط نرم‌افزار R تجزیه‌وتحلیل شد و میز کار WEKA پس از بررسی، داده‌ها برای انتخاب مؤثرترین ویژگی‌ها ازجمله هموگرام، $pCO2$، تعادل پایه اسید خون، زمان پروترومبین، بیومارکرهای التهابی و گلوکز پردازش شدند. بهترین تناسب متغیرها با موفقیت توسط پرسپترون چندلایه، یک الگوریتم شبکه عصبی پیش‌رو که تشخیص ماشینی کووید-۱۹ شدید را با دقت ۹۶.۵ درصد انجام می‌دهد و یا توسط نرم‌افزار C4.5، یک الگوریتم یادگیری نظارت شده بر اساس هدف، تأیید شد. متغیر از پیش تعریف شده (شدت) که درخت تصمیم را با دقت ۸۹.۴ درصد ایجاد کرد. در نهایت، یک ماتریس همبستگی دومتغیره پیچیده پیرسون همراه با خوشه‌بندی سلسله‌مراتبی پیشرفته (دندروگرام) برای کشف دانش انجام شد. ساختار پنهان مجموعه‌های داده الگوهای تغییر مربوط به ایجاد پنومونی ناشی از COVID-19 را نشان داد که شامل پروتئین واکنش‌دهنده لنفوسیت به C و نسبت پروتئین لکوسیت به C، نوتروفیل، pH و $pCO2$ بود. رویکردهای داده‌کاوی به نوسانات خونی مرتبط با ذات‌الریه شدید COVID-19 نه‌تنها می‌تواند پیامدهای بالینی نامطلوب را پیش‌بینی کند، بلکه اهداف درمانی احتمالی را نیز آشکار می‌کند.

- پیش‌بینی پذیرش ICU برای بیماران COVID-19: رویکرد یادگیری ماشینی بر اساس داده‌های شمارش کامل خون

در مقاله [۲] توسعه مدل‌های یادگیری ماشینی (ML) برای پیشرفت COVID-19 را مورد بحث قرار داده است. همچنین به‌طور خاص، به وظیفه پیش‌بینی پذیرش در بخش مراقبت‌های ویژه (ICU) در ۵ روز آینده پرداخته شده است. در این مقاله، سه مدل ML را بر اساس ۴۹۹۵ آزمایش شمارش کامل خون (CBC) توسعه داده شده است. این مقاله سه مدل ML را پیشنهاد می‌کند که ازنظر تفسیرپذیری متفاوت هستند: دو مدل کاملاً قابل تفسیر و یک جعبه سیاه. ما یک 0.81 AUC و ۰.۸۳ را برای مدل‌های قابل تفسیر (به ترتیب درخت تصمیم و رگرسیون لجستیک) و 0.88 AUC برای مدل جعبه سیاه (یک مجموعه) گزارش شده است. این امر نشان می‌دهد که می‌توان داده‌های CBC و روش‌های ML را برای پیش‌بینی مقرون‌به‌صرفه پذیرش ICU بیماران COVID-19 مورد استفاده قرار داد. به‌ویژه، ازآنجاکه CBC را می‌توان به‌سرعت از طریق معاینات معمول خون به‌دست آورد، مدل‌های ارائه شده نیز می‌توانند در تنظیمات محدود به منابع اعمال شوند و برای دریافت نشانه‌های سریع در بستری بیماران مورد استفاده قرار گیرد.

- نقش هوش مصنوعی در مدیریت بیماران بحرانی COVID-19

شیوع بیماری کروناویروس ۲۰۱۹ (COVID-19) چالش بزرگی را برای سیستم مراقبت های بهداشتی در سراسر جهان ایجاد کرده است. یکی از مهم‌ترین نکات این چالش، مدیریت بیماران COVID-19 است که به مراقبت‌های تنفسی حاد نیاز دارند. در این مقاله [۳] برای کشف یک مدل مبتنی بر هوش مصنوعی جهت بهبود مراقبت‌های حیاتی از بیماران COVID-19 بررسی انجام شد. در یک مطالعه توصیفی، تمامی تحقیقات منتشر شده موجود در پایگاه‌های اطلاعاتی PubMed، Web of Science، Google scholar و سایر اطلاعات بازیابی شد. بر اساس این مطالعات، یک مدل سه‌مرحله‌ای ورودی، فرایند و خروجی ایجاد شد. مدل سه‌مرحله‌ای کاربرد هوش مصنوعی در بخش مراقبت‌های ویژه (ICU) تکمیل شد. ورودی شامل داده‌های بالینی، پاراکلینیکی، پزشکی شخصی (OMICS) و اپیدمیولوژیک بود. این فرایند، شامل هوش مصنوعی (به‌عنوان مثال شبکه عصبی مصنوعی، یادگیری ماشینی، یادگیری عمیق و سیستم‌های خبره) بود. خروجی که تصمیم‌گیری ICU بود، تشخیص، درمان، طبقه‌بندی خطر، پیش‌آگهی و مدیریت را شامل می‌شد. تلاش‌های سیستم مراقبت‌های بهداشتی برای شکست COVID-19 می‌تواند توسط یک سیستم تصمیم‌گیری مبتنی بر هوش مصنوعی پشتیبانی شود که آنها را دوبرابر کرده و به مدیریت بسیار کارآمدتر این بیماران، به‌ویژه بیمارانی که در بخش مراقبت‌های ویژه COVID-19 هستند، کمک می‌کند.

نتیجه:

تلاش‌های سیستم مراقبت‌های بهداشتی برای شکست COVID-19 می‌تواند توسط یک سیستم تصمیم‌گیری مبتنی بر هوش مصنوعی پشتیبانی شود که آنها را دوبرابر کرده و به مدیریت بسیار کارآمدتر این بیماران، به‌ویژه بیمارانی که در بخش مراقبت‌های ویژه COVID-19 هستند، کمک می‌کند.

- مقایسه تکنیک‌های داده‌کاوی برای پیش‌بینی مرگ‌ومیر داخل بیمارستانی در بیماران مبتلا به کووید-۱۹

در مقاله [۴] موارد ارائه شده در این مقاله به‌صورت زیر آورده شده است.

مقدمه:

اپیدمی کووید-۱۹ درحال‌حاضر سیستم‌های مراقبت بهداشتی در سراسر جهان را با تردیدها و چالش‌های غیرمنتظره زیادی در تصمیم‌گیری پزشکی و به اشتراک‌گذاری مؤثر منابع پزشکی مواجه کرده است. مدل‌های پیش‌بینی مبتنی بر یادگیری ماشین (ML) می‌توانند به‌طور بالقوه برای غلبه بر این عدم قطعیت‌ها سودمند باشند.

هدف:

هدف این مطالعه، آموزش چندین الگوریتم ML برای پیش‌بینی مرگ‌ومیر در بیمارستان COVID-19 و مقایسه عملکرد آنها برای انتخاب بهترین الگوریتم است. در نهایت، عوامل مؤثر با استفاده از برخی از روش‌های انتخاب ویژگی امتیاز بالاتری گرفتند.

روش‌ها:

با استفاده از یک رجیستری تک مرکزی، پرونده ۱۳۵۳ بیمار تأیید شده کووید-۱۹ بستری در بیمارستان آیت‌الله طالقانی شهر آبادان بررسی شد. ما از شش تکنیک امتیازدهی ویژگی و نُه الگوریتم معروف ML استفاده کردیم. برای ارزیابی عملکرد مدل‌ها، معیارهای به‌دست‌آمده ماتریس درهم‌ریختگی محاسبه شد.

یافته‌ها:

شرکت‌کنندگان در این مطالعه، ۱۳۵۳ بیمار بودند که جنسیت مذکر بالاتر از زنان (۷۴۲ در مقابل ۶۱۱) و میانگین سنی ۵۷.۲۵ سال (بین‌چارکی ۱۰۰-۱۸) بود. پس از امتیازدهی ویژگی، از بین ۵۴ متغیر، تعداد مطلق نوتروفیل/*لنفوسیت* و از دست دادن چشایی و بویایی، سه پیش‌بینی‌کننده اصلی بودند. از سوی دیگر، شمارش پلاکت، منیزیم و سردرد کمترین اهمیت را برای پیش‌بینی مرگ‌ومیر کووید-۱۹ به‌دست آوردند. نتایج تجربی نشان داد که؛ الگوریتم شبکه بیزی با دقت ۸۹/۳۱ درصد و حساسیت ۶۴/۲ درصد در پیش‌بینی مرگ‌ومیر موفق‌تر بوده است.

نتیجه‌گیری:

ML سطح معقولی از دقت را در پیش‌بینی ارائه می‌دهد. بنابراین، استفاده از مدل‌های پیش‌بینی مبتنی بر ML سیستم‌های بهداشتی پاسخ‌گوتر را تسهیل می‌کند و برای شناسایی به‌موقع بیماران آسیب‌پذیر برای اطلاع‌رسانی قضاوت مناسب توسط ارائه‌دهندگان مراقبت‌های بهداشتی مفید خواهد بود.

- پیش‌بینی مرگ‌ومیر بیماران COVID-19 بر اساس تکنیک‌های داده‌کاوی

اگر ویروس کرونا (COVID-19) به‌موقع پیش‌بینی، مدیریت و کنترل نشود، سیستم‌های بهداشتی هر کشور و مردم آن با مشکلات جدی مواجه خواهند شد. مدل‌های پیش‌بینی می‌توانند در مدیریت منابع سلامت و جلوگیری از شیوع و مرگ ناشی از COVID-19 مفید باشند. در مقاله [۵] با هدف پیش‌بینی مرگ‌ومیر در بیماران مبتلا به کووید-۱۹ بر اساس تکنیک‌های داده‌کاوی انجام شد. برای انجام این مطالعه، ابتدا عوامل مرگ‌ومیر بیماران COVID-19 بر اساس مطالعات مختلف شناسایی شد. این عوامل توسط پزشکان متخصص تأیید شد. بر اساس فاکتورهای تأیید شده، داده‌های بیماران کووید-۱۹ از ۸۵۰ پرونده پزشکی استخراج شد. برای پیش‌بینی از مدل‌های درخت تصمیم (J48)، MLP، KNN، جنگل تصادفی و داده‌کاوی SVM استفاده شد. مدل‌ها بر اساس دقت، صحت، ویژگی، حساسیت و منحنی ROC ارزیابی شدند. بر اساس نتایج، مؤثرترین عامل مورد استفاده برای پیش‌بینی مرگ بیماران کووید-۱۹، تنگی نفس بود. بر اساس ROC (۱۰۰۰)، دقت (۹۹.۲۳٪)، صحت (۹۹.۷۴٪)، حساسیت (۹۸.۲۵٪) و ویژگی (۹۹.۸۴٪)، جنگل تصادفی بهترین مدل در پیش‌بینی مرگ‌ومیر نسبت به سایر مدل‌ها بود. پس از جنگل تصادفی، مدل‌های KNN5، MLP و J48 به ترتیب در رتبه‌های بعدی قرار گرفتند. تجزیه‌وتحلیل داده‌های بیماران کووید-۱۹ می‌تواند ابزاری مناسب و کاربردی برای پیش‌بینی میزان مرگ‌ومیر این بیماران باشد. با توجه به حساسیت علم پزشکی به حفظ جان انسان‌ها و کمبود نیروی انسانی متخصص در نظام سلامت، استفاده از مدل‌های پیشنهادی می‌تواند شانس درمان موفقیت‌آمیز را افزایش داده، از مرگ زودهنگام پیشگیری کند و هزینه‌های مربوط به درمان‌های طولانی‌مدت را نیز برای بیماران و بیمارستان‌ها کاهش دهد.

- مدل‌های پیش‌بینی بالینی برای COVID-19: مطالعه سیستماتیک

موارد ارائه شده در مقاله [۶] به‌صورت زیر آورده شده است.

زمینه:

COVID-19 یک بیماری تنفسی به‌سرعت در حال ظهور است که توسط SARS-CoV-2 ایجاد می‌شود. با توجه به انتقال سریع SARS-CoV-2 از انسان به انسان، بسیاری از سیستم‌های مراقبت‌های بهداشتی در خطر فراتر رفتن از ظرفیت‌های مراقبت بهداشتی خود هستند. به‌ویژه ازنظر آزمایش‌های SARS-CoV-2، بیمارستان و بخش مراقبت‌های ویژه (ICU)، تخت‌ها و هواکش‌های مکانیکی. الگوریتم‌های پیش‌بینی‌کننده به‌طور بالقوه می‌توانند با شناسایی افرادی که به احتمال زیاد آزمایش SARS-CoV-2 مثبت را دریافت می‌کنند، در بیمارستان و یا در ICU بستری می‌شوند، فشار بر سیستم‌های مراقبت بهداشتی را کاهش دهند.

هدف:

هدف از این مطالعه، توسعه و ارزیابی مدل‌های پیش‌بینی بالینی است که با استفاده از یادگیری ماشین و بر اساس داده‌های بالینی جمع‌آوری شده و تخمین می‌زنند که بیماران احتمالاً آزمایش SARS-CoV-2 مثبت دریافت می‌کنند، به بستری شدن در بیمارستان و یا مراقبت شدید نیاز دارند.

روش‌ها:

با استفاده از یک رویکرد سیستماتیک برای توسعه و بهینه‌سازی مدل، انواع مختلفی از مدل‌های یادگیری ماشین، ازجمله رگرسیون لجستیک، شبکه‌های عصبی، ماشین‌های بردار پشتیبان، جنگل‌های تصادفی و تقویت گرادیان، آموزش و مقایسه شده است. برای ارزیابی مدل‌های توسعه‌یافته، یک ارزیابی بر روی داده‌های دموگرافیک، بالینی و تجزیه‌وتحلیل خون از گروهی از ۵۶۴۴ بیمار انجام شده است. علاوه‌بر این، تعیین شده است که کدام ویژگی‌های بالینی تا چه میزان برای هریک از وظایف بالینی ذکر شده با استفاده از توضیحات علّی پیش‌بینی‌کننده هستند.

نتایج:

نتایج تجربی این مقاله نشان می‌دهد که؛ مدل‌های پیش‌بینی‌کننده، بیمارانی را شناسایی کرده‌اند که تست SARS-CoV-2 پیشینی با حساسیت ۷۵٪ و ویژگی ۴۹٪مثبت بوده است. بیمارانی که SARS-CoV-2 مثبت هستند و نیاز به بستری در بیمارستان با ۰.۹۲ ناحیه زیر منحنی مشخصه اپراتور گیرنده دارند: (AUC؛ ۹۵٪ فاصله اطمینان ([0.81-0.98 [CI) و بیمارانی که SARS-CoV-2 مثبت هستند، نیاز به بستری دارند. مراقبت‌های ویژه با (AUC 95% 0.98) فاصله اطمینان (CI: 0.95-1.00).

- پیش‌بینی یادگیری عمیق احتمال پذیرش و مرگ‌ومیر در ICU در بیماران COVID-19 با استفاده از متغیرهای بالینی

موارد ارائه شده در مقاله [۷] به‌صورت زیر آورده شده است.

زمینه:

این مطالعه با هدف توسعه یک مدل یادگیری عمیق و یک سیستم امتیاز ریسک با استفاده از متغیرهای بالینی برای پیش‌بینی پذیرش در بخش مراقبت‌های ویژه (ICU) و مرگ‌ومیر در بیمارستان در بیماران COVID-19 انجام شد.

روش‌ها:

این مطالعه گذشته‌نگر شامل ۵۷۶۶ فرد تحت بررسی برای COVID-19 بین ۷ فوریه ۲۰۲۰ تا ۴ مه ۲۰۲۰ بود. اطلاعات دموگرافیک، بیماری‌های مزمن، علائم حیاتی، علائم و آزمایش‌های آزمایشگاهی هنگام پذیرش جمع‌آوری شد. یک مدل شبکه عصبی عمیق و یک سیستم امتیاز ریسک برای پیش‌بینی پذیرش ICU و مرگ‌ومیر در بیمارستان ساخته شد. عملکرد پیش‌بینی از ناحیه مشخصه عملکرد گیرنده زیر منحنی (AUC) استفاده کرد.

نتایج:

مهم‌ترین پیش‌بینی‌کننده‌های ICU پروکلسی‌تونین، لاکتات دهیدروژناز، پروتئین واکنش‌گر C، فریتین و اشباع اکسیژن بود. بالاترین پیش‌بینی‌کننده‌های مرگ‌ومیر عبارت بودند از؛ سن، لاکتات دهیدروژناز، پروکلسی‌تونین، تروپونین قلبی، پروتئین واکنش‌گر C و اشباع اکسیژن. سن و تروپونین پیش‌بینی‌کننده‌های منحصربه‌فرد برای مرگ‌ومیر بودند، اما بستری در ICU نبودند. مدل یادگیری عمیق بستری در ICU و مرگ‌ومیر را به ترتیب با AUC 0.780 (٪۹۵ فاصله اطمینان [۰.۷۶۰-۰.۷۸۵]) و ۰.۸۴۴ (٪۹۵ CI [0.839-0.848]) را پیش‌بینی کرد. امتیاز ریسک مربوطه به ترتیب AUC 0.728 (95% فاصله اطمینان (0.726-0.729 [CI]) و ۰.۸۴۸ (٪۹۵ CI [0.847-0.849]) را به‌همراه داشت.

نتیجه‌گیری:

یادگیری عمیق و امتیاز خطر ناشی از آن این پتانسیل را دارد که پزشکان خط‌مقدم را با ابزارهای کمّی برای طبقه‌بندی بیماران به‌طور مؤثرتر در شرایط حساس به زمان و محدودیت منابع در اختیار پزشکان خط‌مقدم قرار دهد.

- پیش‌بینی فردی پیامدهای نامطلوب COVID-19 با MLHO

برآوردهای دقیق از پیامدهای نامطلوب کووید-۱۹ می‌توانست به تخصیص بهتر منابع مراقبت‌های بهداشتی و اقدامات پیشگیرانه هدفمند کارآمدتر، ازجمله بینش در اولویت‌بندی بهترین نحوه توزیع واکسیناسیون منجر شود. در مقاله [۸] MLHO توسعه داده شده است، یک چارچوب یادگیری ماشینی سرتاسر که از ویژگی تکراری و انتخاب الگوریتم برای پیش‌بینی نتایج سلامت استفاده می‌کند. MLHO برای پیش‌بینی خطر بستری شدن در بیمارستان، بستری شدن در ICU، نیاز به تهویه مکانیکی و مرگ و انتخاب ویژگی و مدل را اجرا می‌کند. این پیش‌بینی را بر اساس داده‌های سوابق پزشکی گذشته بیماران (قبل از ابتلای آنها به COVID-19) استوار می‌کند. معماری MLHO یک کالیبراسیون مدل موازی و نتیجه‌گرا را امکان‌پذیر می‌کند، که در آن الگوریتم‌های یادگیری آماری مختلف و بردارهای ویژگی‌ها به‌طور هم‌زمان برای بهبود پیش‌بینی نتایج سلامت آزمایش می‌شوند. ما با استفاده از داده‌های بالینی و جمعیت‌شناختی از گروه بزرگی از بیش از ۱۳۰۰۰ بیمار مبتلا به کووید-۱۹ مثبت، چهار پیامد نامطلوب را با استفاده از حدود ۶۰۰ ویژگی که نشان‌دهنده سوابق بهداشتی و جمعیت‌شناسی بیماران قبل از COVID-19 بود، مدل‌سازی کردیم. میانگین AUC ROC برای پیش‌بینی مرگ‌ومیر ۰.۹۱ بود. درحالی‌که، عملکرد پیش‌بینی بین ۰.۸۰ تا ۰.۸۱ برای ICU، بستری شدن در بیمارستان و تهویه بود. به‌طور گسترده دسته‌هایی از ویژگی‌هایی را که در مدل‌سازی مورد استفاده قرار گرفتند و تأثیر نسبی آنها برای پیش‌بینی هر نتیجه را توصیف می‌کنیم. نتایج ما نشان داد؛ درحالی‌که متغیرهای جمعیت‌شناختی (یعنی سن) پیش‌بینی‌کننده‌های مهم پیامدهای نامطلوب پس از عفونت COVID-19 هستند، ادغام سوابق بالینی گذشته برای یک مدل پیش‌بینی قابل اعتماد حیاتی است. با گسترش همه‌گیری COVID-19 در سراسر جهان، چارچوب‌های یادگیری ماشینی قابل انطباق و تفسیر (مانند MLHO) برای بهبود آمادگی ما برای رویارویی با امواج احتمالی کووید-۱۹ در آینده و همچنین سایر بیماری‌های عفونی جدید که ممکن است ظهور کنند، بسیار مهم است.

- مقایسه الگوریتم‌های یادگیری ماشین برای پیش‌بینی پذیرش و مرگ‌ومیر در ICU در COVID-19

ازآنجاکه پیش‌بینی مسیر COVID-19 چالش برانگیز است، مدل‌های یادگیری ماشینی می‌توانند به پزشکان در شناسایی افراد پرخطر کمک کنند. در این مقاله [۹] عملکرد ۱۸ الگوریتم یادگیری ماشینی را برای پیش‌بینی پذیرش و مرگ‌ومیر در ICU در بین بیماران COVID-19 مقایسه می‌کند. با استفاده از داده‌های بیماران COVID-19 از پایگاه داده مراقبت‌های بهداشتی Mass General Brigham (MGB)، مدل‌هایی را با استفاده از بیمارانی که در فاصله مارس تا آوریل ۲۰۲۰ به بخش اورژانس (ED) مراجعه می‌کردند (n=۳۵۹۷) توسعه داده شده و اعتبار داخلی آنها را تأیید کرده است و آنها با استفاده از زمانی مجزا تأیید شده‌اند. مثلاً افرادی که بین ماه مه تا آگوست ۲۰۲۰ به ED مراجعه کردند (n=۱۷۱۱) نشان می‌دهد که مدل‌های مبتنی بر مجموعه در پیش‌بینی پذیرش ۵ روزه ICU و مرگ‌ومیر ۲۸ روزه ناشی از COVID-19 بهتر از سایر مدل‌ها عمل می‌کنند. CRP، LDH، و اشباع O2 برای مدل‌های بستری در ICU مهم بودند درحالی‌که eGFR <60 ml/min/1.73m2 و درصد نوتروفیل و لنفوسیت مهم‌ترین متغیرها برای پیش‌بینی مرگ‌ومیر بودند. پیاده‌سازی چنین مدل‌هایی می‌تواند به تصمیم‌گیری بالینی برای شیوع بیماری‌های عفونی آینده ازجمله COVID-19 کمک کند.

۴- نتایج

در جدول ۲ نتایج مقالات مورد بررسی این مقاله با یکدیگر مقایسه شده‌اند، در این میان و براساس نتایج گزارش شده و مدل های مورد بررسی در این مقالات الگوریتم جنگل تصادفی (Random Forest) با دقت ۹۹٪ از بالاترین عملکرد در میان مدل های بررسی شده برخوردار بوده است.

جدول ۲ مقایسه مقالات بررسی شده.

مقاله	سال	روش(های) ارائه شده	مجموعه داده	دقت
[1]	2021	داده‌کاوی و یادگیری ماشینی، دو شکل پیشرفته هوش مصنوعی را برای پیش‌بینی ذات‌الریه شدید COVID-19 به کار گرفته شده است.	بیمار کووید-19 بر اساس هیپوکسی خون در بدو بستری به دو گروه شدید (489 مورد) و غیر شدید (3520 مورد).	بهترین دقت 96/5
[2]	2021	سه مدل ML ارائه شده است.	4995 آزمایش.	88
[3]	2020	شبکه عصبی مصنوعی؛ یادگیری عمیق و سیستم‌های خبره	بیماران COVID-19 است که به مراقبت‌های تنفسی حاد نیاز دارند.	-
[4]	2021	استفاده از الگوریتم‌های یادگیری ماشین	پرونده 1353 بیمار کووید-19 بستری در بیمارستان آیت‌الله طالقانی شهر آبادان	شبکه بیزی با دقت 89/31
[5]	2021	مدل‌های درخت تصمیم، MLP، KNN، جنگل تصادفی و داده‌کاوی SVM استفاده شده است.	پیش‌بینی مرگ بیماران کووید-۱۹ تنگی نفس.	جنگل تصادفی بهترین مدل با دقت تقریباً 99
[6]	2020	استفاده از انواع مختلفی از مدل‌های یادگیری ماشین.	تجزیه‌وتحلیل خون از گروهی از 5644 بیمار.	تقریباً 95 درصد.
[7]	2020	یک مدل شبکه عصبی عمیق و یک سیستم امتیاز ریسک برای پیش‌بینی پذیرش ICU و مرگ‌ومیر	5766 فرد تحت بررسی برای COVID-19.	تقریباً 95 درصد
[8]	2021	استفاده از معماری MLHO یک کالیبراسیون مدل موازی و نتیجه‌گرا را امکان‌پذیر می‌کند.	پیش‌بینی بر اساس داده‌های سوابق پزشکی گذشته بیماران.	پیش‌بینی مرگ‌ومیر 0.91
[9]	2021	مقایسه عملکرد 18 الگوریتم یادگیری ماشینی	پیش‌بینی پذیرش و مرگ‌ومیر در ICU در بین بیماران COVID-19	-

۵- نتیجه گیری

به جرعت می‌توان به کارگیری یادگیری ماشین در تشخیص کووید- ۱۹ را به عنوان یک سیستم کمک تشخیصی با سرعت و دقت قابل اتکا در عین حال پیچیدگی و هزینه کمتر در پیاده‌سازی نسبت به سایر روشها نظیر یادگیری عمیق توصیه کرد. الگوریتم های مختلف و روش‌های بررسی شده در این مقاله می‌توانند راهنمای استفاده از آنها در سناریوهایی با ویژگی‌های مشابه کووید-۱۹ باشند. با اتکا به تجربه کووید-۱۹ و اهمیت تشخیص زودهنگام و اقدام به موقع در بیماری‌های این‌چنینی توسعه سیستم های کمک تشخیصی جهت تحلیل آزمایشات و تصاویر پزشکی مختلف مرتبط می‌تواند آمادگی سیستم های درمانی در مقابله با این بیماری ها را دوچندان نماید.

مراجع

[1] M. Pulgar-Sánchez *et al.*, "Biomarkers of severe COVID-19 pneumonia on admission using data-mining powered by common laboratory blood tests-datasets," *Computers in biology and medicine,* vol. 136, p. 104738, 2021.

[2] L. Famiglini, G. Bini, A. Carobene, A. Campagner, and F. Cabitza, "Prediction of ICU admission for COVID-19 patients: a Machine Learning approach based on Complete Blood Count data," in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 2021: IEEE, pp. 160-16.

[3] S. Rahmatizadeh, S. Valizadeh-Haghi, and A. Dabbagh, "The role of artificial intelligence in management of critical COVID-19 patients," *Journal of Cellular and Molecular Anesthesia,* vol. 5, no. 1, pp. 16-22, 2020.

[4] M. Shanbehzadeh, A. Orooji, and H. Kazemi-Arpanahi, "Comparing of data mining techniques for predicting in-hospital mortality among patients with covid-19," *Journal of Biostatistics and Epidemiology,* vol. 7, no. 2, pp. 154-173, 2021.

[5] K. Moulaei, F. Ghasemian, K. Bahaadinbeigy, R. E. Sarbi, and Z. M. Taghiabad, "Predicting mortality of COVID-19 patients based on data mining techniques," *Journal of Biomedical Physics & Engineering,* vol. 11, no. 5, p. 653, 2021.

[6] P. Schwab, A. D. Schütte, B. Dietz, and S. Bauer, "Clinical predictive models for COVID-19: systematic study," *Journal of medical Internet research,* vol. 22, no. 10, p. e21439, 2020.

[7] X. Li *et al.*, "Deep learning prediction of likelihood of ICU admission and mortality in COVID-19 patients using clinical variables," *PeerJ,* vol. 8, p. e10337, 2020.

[8] H. Estiri, Z. H. Strasser, and S. N. Murphy, "Individualized prediction of COVID-19 adverse outcomes with MLHO," *Scientific reports,* vol. 11, no. 1, pp. 1-9, 2021.

[9] S. Subudhi *et al.*, "Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19," *NPJ digital medicine,* vol. 4, no. 1, pp. 1-7, 2021.