

بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِيْمِ



معاونت پژوهش و فن آوری

بنام خدا

مشور اخلاق پژوهش

بایاری از مذاقند بجان و اعتقاد باین که حالم محضر خداست. او بحواره نافر بر اعمال انسان بجهت مثوب پاس داشت تمام پلند داش و پژوهش و نظر بر آیت جایکده و انجاه، داعلای فرمک و تمن بشری، دادجیان و اعصابی بینت ملی و اصحابی و انجاه آزاد اسلامی تهدیدی کردیم اصول زیر را در احجام فایت های پژوهشی مد نظر قرار داده و از آن تحقیک نگیم:

- ۱- اصل حقیقت جویی. تلاش درستای پی جویی حقیقت و وفاداری به آن و دوری از هرگونه پسنان سازی حقیقت.
- ۲- اصل رعایت حقوق. الترام برعایت کامل حقوق پژوهشگران و پژوهیدگران (انسان، حیوان و ببات) و سایر صاحبان حق.
- ۳- اصل مالکیت مادی و معنوی. تهدید بر عایت کامل حقوق مادی و معنوی دانشجو و کارکنان پژوهش.
- ۴- اصل منافع ملی. تهدید بر عایت مصالح ملی و دلخواشی شد و توسعه کشود کیه بکاران پژوهش.
- ۵- اصل رعایت انصاف و امانت. تهدید ب اختیاب از هرگونه جانب داری غیر علی و خناخت از اموال، تجزیات و منابع داشتار.
- ۶- اصل رازداری. تهدید به صیانت از اسرار و اطلاعات محیان افراد، سازمان یا کشور و کیه افزاد و نهادهای مرتبط با تحقیق.
- ۷- اصل احترام. تهدید بر عایت حريم هدایت ادب احترام تحقیقات و رعایت جانب تقدیر خود از ای از هرگونه حرمت گشتن.
- ۸- اصل ترویج. تهدید بر عایت داشتیج آن بکاران علی و دانشجویان به غیر از مواردی که منع قانونی دارد.
- ۹- اصل براست. الترام بر است جویی از هرگونه فشار غیر حرفه ای و اعلام موضع نسبت به کسانی که حوزه علم و پژوهش را بشتابه ای غیر علمی می آیند.



دانشگاه آزاد اسلامی
واحد علوم تحقیقات تهران

تعهدنامه اصالت رساله یا پایان نامه

اینجانب مهندام پدرام دانش آموخته مقطع کارشناسی ارشد ناپیوسته در رشته مهندسی کامپیوتر گرایش نرم افزار که در تاریخ از پایان نامه خود تحت عنوان " ارایه‌ی یک روش بهبود یافته برای پیش‌بینی ضرورت بستری شدن بیماران کووید ۱۹ در بخش مراقبت‌های ویژه با استفاده از تکنیک‌های ترکیبی داده کاوی " با کسب نمره و درجه دفاع نموده‌ام بدین وسیله متعهد می‌شوم:

- ۱) این پایان نامه حاصل تحقیق و پژوهش انجام شده توسط اینجانب بوده و در مواردی که از دستاوردهای علمی و پژوهشی دیگران (اعم از پایان نامه، کتاب، مقاله و ...) استفاده نموده‌ام، مطابق ضوابط و رویه موجود، نام منبع مورد استفاده و سایر مشخصات آن را در فهرست مربوطه ذکر و درج کرده‌ام.
- ۲) این پایان نامه قبلاً برای دریافت هیچ مدرک تحصیلی (هم سطح، پایین‌تر یا بالاتر) در سایر دانشگاه‌ها و مؤسسات آموزش عالی ارائه نشده است.
- ۳) چنان‌چه بعد از فراغت از تحصیل، قصد استفاده و هرگونه بهره‌برداری اعم از چاپ کتاب، ثبت اختراع و ... از این پایان نامه داشته باشم، از حوزه معاونت پژوهشی واحد مجوزهای مربوطه را اخذ نمایم.
- ۴) چنان‌چه در هر مقطع زمانی خلاف موارد فوق ثابت شود، عواقب ناشی از آن را می‌پذیرم و واحد دانشگاهی مجاز است با اینجانب مطابق ضوابط و مقررات رفتار نموده و در صورت ابطال مدرک تحصیلی ام هیچ‌گونه ادعایی نخواهم داشت.

نام و نام خانوادگی:

تاریخ و امضاء:



دانشگاه آزاد اسلامی
واحد علوم تحقیقات تهران
دانشکده مکانیک، برق و کامپیوتر
پایاننامه برای دریافت درجه کارشناسی ارشد
در رشته مهندسی کامپیوتر گرایش نرم افزار

ارایه‌ی یک روش بهبودیافته برای پیش‌بینی ضرورت بسترهای شدن بیماران کووید ۱۹ در بخش مراقبت‌های ویژه با استفاده از تکنیک‌های ترکیبی داده‌کاوی

استاد راهنمای:
خانم دکتر مریم رستگارپور

نگارش

بهار ۱۴۰۲

با آرزوی سلامتی و بهرهمندی از لحظات شیرین و مملو از موفقیت برای استاد گرامی، خانم دکتر مریم رستگارپور، مراتب تشکر و سپاسگزاری خود را از راهنمایی داهیانه و همکاری سخاوتمندانه‌ی ایشان نثار پیشگاه آن بزرگوار می‌نمایم. همچنین از همراهی‌ها و پشتیبانی پدر و مادر مهربانم صمیمانه سپاسگزارم.

فهرست مطالب

۱	چکیده.....
۲	فصل اول: کلیات تحقیق.....
۲	۱-۱- مقدمه.....
۵	۱-۲- بیماری کووید-۱۹.....
۶	۱-۳- روش‌های تشخیص کووید-۱۹.....
۹	۱-۴- مانیتورینگ بیمار و ملاحظات مراکز درمانی.....
۱۱	فصل دوم: مرور تحقیقات پیشین.....
۱۱	۱-۲- مقدمه.....
۱۳	۲-۲- روش یادگیری ترکیبی.....
۱۵	۱-۲-۱- روش‌های ترکیب خروجی یادگیرندها.....
۱۷	۱-۲-۲- استعمال یا وابستگی یادگیرندها.....
۱۸	۱-۳-۲- پیش‌بینی کامپیوتری احتمال بسترهای شدن در بخش مراقبت‌های ویژه.....
۱۹	۱-۳-۱- انتخاب یا استخراج ویژگی‌های مناسب.....
۲۰	۲-۳-۲- انتخاب و تعلیم مدل‌های یادگیرنده.....
۲۲	۲-۳-۳- ارزیابی عملکرد مدل‌های یادگیرنده.....
۲۴	فصل سوم: روش پیشنهادی، اصول و الگوریتمها.....
۲۴	۱-۳- مقدمه.....
۲۴	۲-۳- روش‌هایی برای پیش‌پردازش دادگان.....
۲۴	۱-۲-۱- روش K-نزدیک‌ترین همسایگی ($K-nn$).....
۲۵	۱-۲-۲- خریب همبستگی پیرسون (PCC).....

۲۶	۳-۲-۳- معیار کولبک-لیبلر (KL).....
۲۷	۳-۲-۴- روش کاهش ویژگی PCA
۲۹	۳-۲-۵- روش کاهش ویژگی $t\text{-SNE}$
۳۱	۳-۲-۶- حذف ویژگی بازگشته (RFE).....
۳۲	۳-۳- مدل‌های یادگیرنده.....
۳۲	۳-۳-۱- ماشین بردار پشتیبانی (SVM).....
۳۴	۳-۳-۲- رگرسیون لجستیک
۳۵	۳-۳-۳- درخت تصمیم، جنگل تصادفی
۳۸	۳-۳-۴- $XGBoost$
۴۰	۴-۳- برحی تکنیک‌های کمکی استفاده شده در فرآیند تعلیم و تصمیم‌گیری
۴۰	۴-۴-۱- بوت استرب
۴۱	۴-۴-۲- توقف زودهنگام
۴۲	۴-۴-۵- فلوچارت کلی روش پیشنهادی
۴۴	فصل ۴: شبیه‌سازی و تجزیه و تحلیل داده‌ها
۴۴	۴-۱- مقدمه
۴۴	۴-۲- معرفی پایگاه دادگان
۴۶	۴-۲-۱- آماده‌سازی دادگان
۴۷	۴-۲-۲- مشخصات آماری ویژگی‌های ثبت شده
۵۰	۴-۲-۲-۱- دسته بنایی ویژگی‌ها
۵۲	۴-۲-۳- کاهش بعد ویژگی‌ها و نمایش دادگان
۵۳	۴-۳- کاهش بعد ویژگی‌ها و تعلیم مدل
۵۳	۴-۳-۱- ایجاد مدل پایه برای مقایسه
۵۵	۴-۳-۲- مدل جنگل تصمیم تصادفی
۷۱	۴-۳-۳- مدل $XGBoost$
۶۲	۴-۴- مقایسه مدل‌ها
۶۵	فصل ۵: بحث و نتیجه گیری
۶۵	۱-۵- مقدمه

۶۶	۲-۵ مزایای روش پیشنهادی
۶۷	۳-۵ نتایج تحقیق
۶۷	۴-۵ کارهای آینده
۶۹	منابع:

فهرست شکل‌ها

- شکل ۱-۲: منحنی ROC برای ارزیابی کلی مدل
۲۳
- شکل ۱-۳: مراحل اصلی روش کاهش بعد PCA
۲۹
- شکل ۲-۳: مثالی از اعمال روش کاهش ویژگی t-SNE بر روی بانک اطلاعاتی MNIST
۳۰
- شکل ۳-۳: یک ماشین بردار پشتیبان برای طبقه‌بندی ویژگی‌هایی با دو بعد
۳۳
- شکل ۴-۳: رگرسیون لجستیک برای طبقه‌بندی ویژگی‌هایی با دو بعد
۳۴
- شکل ۵-۳: مثال بسیار ساده‌ای از یک درخت تصمیم
۳۶
- شکل ۶-۳: نمودار یک جنگل تصادفی
۳۷
- شکل ۷-۳: طرح کلی برای پاپ لاین اولیه تحقیق.
۴۳
- شکل ۸-۱: اطلاعات غیر قابل استفاده در دادگان برای تعلیم مدل
۴۵
- شکل ۸-۲: توزیع افراد منتقل شده به بخش مراقبت‌های ویژه
۴۷
- شکل ۸-۳: ماتریس همبستگی متغیرهای فیزیولوژیک ثبت شده پس از نرمالیزه شدن
۴۸
- شکل ۸-۴: نمایش ارتباط متغیرهای جنسیت، سن، HTN و وضعیت سیستم ایمنی با احتمال بستری در ICU
۵۰
- شکل ۸-۵: توزیع متغیرهای فیزیولوژیکی که به صورت بیشینه گزارش شده‌اند.
۵۱
- شکل ۸-۶: توزیع متغیرهای فیزیولوژیکی که به صورت میانگین نشان داده شده است.
۵۲
- شکل ۸-۷: تحلیل مولفه‌های اصلی در روش PCA
۵۲
- شکل ۸-۸: نمایش توزیع دادگان پس از کاهش بعد با استفاده از روش t-SNE
۵۳
- شکل ۸-۹: عملکرد مدل‌های پایه (رگرسیون لجستیک و ماشین بردار پشتیبان) در حالت‌های مختلف
۵۵
- شکل ۸-۱۰: منحنی ROC مربوط به عملکرد مدل ماشین بردار پشتیبان در بهترین حالت
۵۵
- شکل ۸-۱۱: عملکرد مدل پایه‌ی جنگل تصادفی با و بدون اعمال کاهش ابعاد ورودی و متوازن‌سازی کلاس‌ها.
۵۶
- شکل ۸-۱۲: متغیرهای فیزیولوژیکی که بیشترین تاثیر را در عملکرد مدل جنگل تصمیم تصادفی دارند.
۵۷
- شکل ۸-۱۳: توزیع برخی مقادیر کمینه‌ی متغیرهای فیزیولوژیکی ثبت شده.
۵۸
- شکل ۸-۱۴: عملکرد مدل پایه‌ی جنگل تصمیم تصادفی با و بدون اعمال بوت‌استرپ و متوازن‌سازی کلاس‌ها
۶۰
- پارامترهای بهینه برای مدل پایه بدون نمونه‌های بوت‌استرپ محاسبه شده‌اند.
- شکل ۸-۱۵: عملکرد مدل پایه‌ی جنگل تصمیم تصادفی با و بدون اعمال بوت‌استرپ و متوازن‌سازی کلاس‌ها
۶۱
- پارامترهای بهینه برای مدل پایه با در نظر گرفتن نمونه‌های بوت‌استرپ محاسبه شده‌اند.
- شکل ۸-۱۶: منحنی ROC مربوط به عملکرد مدل XGBoost
۶۲
- شکل ۸-۱۷: عملکرد مدل XGBoost با و بدون اعمال کاهش ابعاد ورودی و متوازن‌سازی کلاس‌ها
۶۲
- شکل ۸-۱۸: مقایسه‌ی عملکرد مدل‌های بررسی شده در این تحقیق
۶۳
- شکل ۸-۱۹: مقایسه‌ی عملکرد مدل‌های بررسی شده در این تحقیق در مراحل مختلف ثبت ویژگی‌ها
۶۴

فهرست جداول

- | | |
|----|--|
| ۴۸ | جدول ۱-۴: فهرست ویژگی‌های به ترتیب همبستگی بر اساس شاخص پیرسون |
| ۴۹ | جدول ۲-۴: فهرست کامل ویژگی‌های با ضریب همبستگی بالاتر از ۹۹% |
| ۵۹ | جدول ۳-۴: ابرپارامترهای تنظیم شده در مدل جنگل تصمیم تصادفی |

چکیده

با گذشت بیش از سه سال از شروع همه‌گیری جهانی بیماری کووید-۱۹ و پس از پیدایش سویه‌های مختلف ویروس و همچنین خطر بروز نمونه‌های مشابه در آینده، پیش‌بینی هزینه‌هایی که بیماری‌های عفونی بر سیستم بهداشت و درمان تحمیل می‌کنند اهمیت خاصی در بهبود برنامه‌های آگاهی رسانی عمومی و همچنین مدیریت ظرفیت‌های درمانی و نگهداری از مبتلایان یافته‌اند. از منظر هزینه‌های مراقبتی و برنامه‌ریزی‌های کلان بیمارستانی، پیش‌بینی نیاز افراد مبتلا به بستری شدن در بخش مراقبت‌های ویژه بیمارستان نیز حائز اهمیت بوده و توجه محققان را به خود معطوف نموده است. لذا، در این تحقیق پیشنهاد شده است که برای تشخیص سریع این‌که کدامیک از مبتلایان ممکن است دچار عالیم مراحل حاد این بیماری و نیازمند به مراقبت‌های ویژه شوند می‌توان از روش‌های مبتنی بر یادگیری ماشین استفاده نمود. در این تحقیق، همچنان نشان داده شد که می‌توان قدرت پیش‌بینی‌کنندگی این روش‌ها را از طریق ترکیب چندین مدل و ساخت یک مدل یادگیری گروهی بهبود بخشد. به طور ویژه، بر روی داده‌های آزمایشگاهی جمع آوری شده از ۳۸۴ مراجعه به بیمارستان سیریولبانز در سائوپلو نشان داده شد که مدل‌های جنگل تصادفی و XGBoost به عنوان مثال‌هایی از الگوریتم‌های یادگیری گروهی می‌توانند با دقت، ۰/۹۴، امتیاز اف۱، ۰/۸۶ و مساحت زیر منحنی، ۰/۹۸ احتمال نیاز به بستری در بخش مراقبت‌های ویژه را پیش‌بینی نمایند. همچنین نشان داده شد که اگر فقط اطلاعات اولین مراجعه‌ی هر بیمار مبنای تصمیم‌گیری قرار گیرد، مدل XGBoost توسعه یافته در تحقیق حاضر می‌تواند با دقت ۰/۸۶ و مساحت زیر منحنی برابر با ۰/۸۸ نیاز به بستری در بخش مراقبت‌های ویژه را پیش‌بینی کند که این نتایج در مقایسه با بهترین مدل موجود (با عملکردی با دقت، ۰/۷۳ و مساحت زیر منحنی، ۰/۷۳) بهبود قابل توجیه را نشان می‌دهد.

فصل اول: کلیات تحقیق

۱-۱- مقدمه

در سال‌های اخیر و به دنبال پیشرفت‌های چشمگیر در زمینه‌ی جمع‌آوری و ذخیره‌ی دادگان حیاتی، توسعه‌ی مدل‌های یادگیرنده‌ی دقیق و همچنین طراحی پردازنده‌های قوی دیجیتال، الگوریتم‌های هوش مصنوعی و روش‌های یادگیری ماشین با اقبال گسترده‌ای در کاربردهای تشخیصی پزشکی مواجه شده‌اند. شبکه‌های عصبی عمیق^۱ به طور عام و شبکه‌های پیچشی^۲، به طور خاص، کاربردهای عملی بسیاری در زمینه‌ی تصویرگیری پزشکی و پردازش خودکار این تصاویر پیدا کرده‌اند. تشخیص و تفکیک بافت‌های سرطانی در ریه با استفاده از تصاویر سی‌تی^۳ [۱]، تشخیص ضایعات پوستی در تصاویر کلینیکی [۲]، بررسی پیشرفت بیماری دیابت با استفاده از پردازش تصاویر شبکیه [۳]، تشخیص بیماری‌های قلبی و عروقی مانند پلاک‌های کلسیمی در تصاویر MRI و آلتراسوند درون رگی [۴]، مثال‌هایی از این کاربردها هستند. از سوی دیگر، شبکه‌های عصبی بازگشته و همچنین شبکه‌های برخوردار از مکانیسم توجه به خود، مانند ترانسفومرها، در جهت بهبود الگوریتم‌های پردازش سیگنال‌های حیاتی و همچنین مستندسازی، مدیریت و تفسیر خودکار داده‌های بالینی متنی به کار گرفته شده‌اند [۵ و ۶]. از میان مهمترین کاربردهای این الگوریتم‌ها می‌توان به ارزیابی خطر ایست قلب ناگهانی با استفاده از پردازش سیگنال ECG^۷ [۷] و بررسی امکان پیش‌بینی حملات صرع از طریق ویژگی‌های استخراج

^۱ Deep neural network

^۲ Convolutional

^۳ Computed tomography (CT)

^۴ Electrocardiogram

شده از سیگنال‌های EEG^۵ [۸] اشاره کرد. این در حالی است که به منظور پردازش و دسته‌بندی داده‌های جدولی^۶، همچنان بهترین روش‌های ممکن، از میان الگوریتم‌های مبتنی بر یادگیری درخت تصمیم انتخاب و ارزیابی می‌شوند.

در هر یک از کاربردهای مذکور، در حالت ایده‌آل، الگوریتم یا مدل انتخابی با استفاده از یک مجموعه‌ی دادگان بسیار-بعدی و با تکیه بر تشخیص‌های قبلی پزشک تعلیم می‌یابند (یادگیری با ناظارت). سپس، این سیستم تصمیم‌گیری کامپیوتری که با دقت و سرعت بالایی قادر به کلاس‌بندی دادگان ورودی، تشخیص نمونه‌های بیمار و تشخیص نوع و درجه‌ی ناهنجاری هر نمونه است، در اختیار پزشک یا سیستم درمانی قرار می‌گیرد. علاوه بر تسهیل تشخیص بیماری، بهینه سازی برنامه ریزی برای مراقبت‌های بهداشتی و کلینیکی و مانیتورینگ روند درمانی از مهمترین مزایای وجود چنین سیستم‌هایی خواهد بود. به علاوه، توسعه‌ی روش‌های خودکار تشخیصی، قابلیت تکرار ارزیابی‌های کمی را نیز بهبود داده و می‌تواند کاهش نرخ اشتباهات تشخیصی را به دنبال داشته باشد. اما، در کاربردهای دنیای واقعی، طراحی و توسعه‌ی چنین سیستمی بسیار چالش برانگیز بوده و با محدودیت‌هایی از قبیل ناکافی بودن تعداد نمونه‌های تعلیم، دشواری در دسترسی به دادگان با کیفیت، جامع و بی‌طرف، و همچنین تعداد بالای متغیرهای ورودی مواجه است. حتی در صورت دستیابی به مجموعه‌ی دادگان مناسب، تعلیم یک مدل تصمیم‌گیرنده‌ی دقیق و سریع، نیازمند امکانات پیشرفته‌ی سخت‌افزاری و نرم‌افزاری بوده و در بسیاری از موارد تضمینی نیست که مدل، قدرت تعمیم بالایی در مواجهه با دادگان جدید داشته باشد. بنابراین، در بسیاری از تحقیقات منتشر شده در زمینه‌ی کاربرد روش‌های هوش مصنوعی در پردازش و به ویژه طبقه‌بندی دادگان پزشکی، هدف اصلی معطوف به بهبود کیفیت دادگان، بهبود روش‌های استخراج ویژگی‌های مناسب، و نیز افزایش دقت و حساسیت مدل‌های یادگیرنده بوده است. در نتیجه، توسعه‌ی یک پایپلاین^۷ بهینه، قابل اعتماد، مقاوم در برابر نویز و آسان برای بهره برداری و نگهداری، همچنان مهمترین چالش در هر پروژه‌ی داده‌کاوی، به ویژه در کاربردهای پزشکی خواهد بود.

با توجه به اهمیت توسعه‌ی نرم افزارها و روش‌های داده‌کاوی مناسب در کاربردهای پزشکی، در این تحقیق، یک نرم‌افزار بهینه و قابل اعتماد برای پردازش داده‌های بالینی که به صورت جدولی جمع‌آوری و ثبت شده‌اند، ارایه می‌شود. در این داده‌های جدولی برای هر مشاهده در یک زمان خاص، یک ردیف در نظر گرفته شده و کمیت‌های اندازه‌گیری شده به صورت ستون‌هایی از ویژگی‌ها ذخیره می‌شوند. با توجه به این تعریف، اگرچه این تحقیق معطوف به داده‌کاوی برای یک کاربرد پزشکی خاص است، نرم افزار و روش پیشنهادی با اعمال تغییراتی اندک، قابل تعمیم به کاربردهای مشابه نیز خواهد بود.

^۵ Electroencephalogram

^۶ Tabular data

^۷ Pipeline

در زمان تعریف این پروژه، بیماری کووید-۱۹ اثرات منفی جبران ناپذیری را بر زندگی میلیون‌ها نفر در سراسر گذاشته بود و ارایه روش‌های سریع و ارزان برای تشخیص افراد مبتلا، تشخیص احتمال نیاز به بسترهای شدن در بخش مراقبت‌های ویژه‌ی بیمارستان و همچنین تشخیص احتمال نجات فرد مبتلا به عنوان یکی از مهمترین و ضروری‌ترین اولویت‌های تحقیقاتی در بخش داده‌کاوی بالینی در نظر گرفته می‌شد. دو سال پس از آغاز همه‌گیری، پس از بیش از ۲۶۰ میلیون مورد بیماری، بیش از ۵ میلیون مرگ و اثرات منفی اقتصادی، اجتماعی و روانی، در تاریخ ۴ آذر ماه ۱۴۰۰ سازمان بهداشت جهانی اولین مورد از سویه‌ی اومیکرون را تایید کرد. براساس تحلیل اطلاعات جمع آوری شده درباره این سویه‌ی جدید به نظر می‌رسد که سویه‌های بعدی سرعت انتشار بیشتری از سویه‌های قبلی دارد و راحت‌تر از سد دفاعی بدن عبور می‌کند [۹]. هر چند برخی شرکت‌های داروسازی تولید نسخه‌ی بهروز شده واکسن کووید را برای مقابله با سویه‌ی اومیکرون آغاز کردند [۱۰]، شیوع سریع و گسترده‌ی ویروس که در بسیاری موارد نیز بدون علامت‌های معمول رخ می‌داد، ریشه‌کنی آن را دشوار نموده و تحقق ایده‌ی ایمنی جمعی را منوط به فرض واکسیناسیون تمام جمعیت جهان [۱۱] می‌کرد.

امروز و با گذشت بیش از سه سال از شروع این همه‌گیری جهانی و پس از پیدایش سویه‌های مختلف ویروس و همچنین خطر بروز نمونه‌های مشابه در آینده، پیش‌بینی هزینه‌هایی که بیماری‌های عفونی بر سیستم بهداشت و درمان تحمیل می‌کنند می‌تواند در بهبود برنامه‌های آگاهی رسانی عمومی و همچنین مدیریت ظرفیت‌های درمانی و نگهداری مبتلایان موثر واقع شود. همچنان جان بسیاری از مردم، به ویژه سالمندان، در خطر بوده و یافتن راه‌های مؤثر برای تشخیص زودهنگام این بیماری در افراد، از اولویت‌های تحقیقاتی به شمار می‌رود. از منظر هزینه‌های مراقبتی و برنامه‌ریزی‌های کلان بیمارستانی، پیش‌بینی نیاز افراد به بسترهای شدن در بخش مراقبت‌های ویژه‌ی بیمارستان نیز حائز اهمیت بوده و توجه محققان را به خود معطوف نموده است. لذا، در این تحقیق پیشنهاد می‌شود که برای تشخیص سریع این‌که کدامیک از مبتلایان ممکن است دچار علایم مراحل حاد این بیماری و نیازمند به مراقبت‌های ویژه شوند روش‌های مبتنی بر یادگیری ماشین به کار گرفته شوند.

با توجه به کاربردهای روش‌های یادگیری ماشین در سایر زمینه‌های تشخیصی و همچنین دانش موجود در زمینه‌ی اپیدمیولوژی بیماری‌های عفونی، پیشنهاد می‌شود که می‌توان قدرت پیش‌بینی کنندگی این روش‌ها را از طریق ترکیب چندین مدل و ساخت یک مدل تجمیعی بهبود بخشد. لذا، در این تحقیق، نتایج تعلم و ارزیابی چند مدل تجمیعی گزارش شده و با یکدیگر مقایسه می‌شوند. به طور ویژه، مدل‌های جنگل تصادفی و XGBoost به عنوان مثال‌هایی از الگوریتم‌های یادگیری گروهی بر روی داده‌های آزمایشگاهی که در سال

سال ۲۰۲۰ از ۳۸۴ مراجعه به بیمارستان سیریولبانز در سائوپلو میلادی جمع‌آوری شده و در اختیار عموم قرار گرفته است، تعلیم داده شده و ارزیابی می‌شوند. مقایسه‌ی این مدل‌ها با مدل‌های استانداردی مانند ماشین بردار پشتیبان و رگرسیون لجستیک و همچنین مدل‌های توسعه یافته در سایر مقالات نشان می‌دهد که این الگوریتم‌های یادگیری گروهی می‌توانند، به ویژه پس از اولین مراجعه به بیمارستان، با دقت بالاتری احتمال نیاز به بستری در بخش مراقبت‌های ویژه را پیش‌بینی نمایند. با توجه به متغیرهای بیولوژیک مورد استفاده و همچنین تعیین‌پذیری روش حل مساله در این تحقیق، می‌توان از این مدل‌ها با اندکی تغییر در پیش‌بینی احتمال نیاز به بستری در بخش مراقبت‌های ویژه در سایر بیماری‌های مرتبط با ویروس خانواده کرونا و همچنین آنفلانزای فصلی نیز استفاده نمود.

در ادامه‌ی این فصل، ابتدا به تاریخچه‌ی مختصری از بیماری کووید-۱۹ اشاره می‌شود و سپس علایم و سیگنال‌های حیاتی که برای تشخیص این بیماری و شدت آن اندازه‌گیری می‌شوند، معرفی خواهند شد. در فصل ۲ گزارشی از تحقیقاتی که در زمینه‌ی کاربردهای روش‌های یادگیری ماشین در تشخیص کووید-۱۹ منتشر شده‌اند، ارایه می‌شود. فصل ۳ در برگیرنده‌ی اصول و الگوریتم‌هایی است که در تحقیق حاضر مورد استفاده قرار گرفته‌اند. در فصل ۴، نتایج تعلیم و ارزیابی این الگوریتم‌ها گزارش شده و در فصل ۵، یافته‌های این تحقیق جمع‌بندی و ارایه می‌شوند.

۱-۲- بیماری کووید-۱۹

اولین بار در ماه دسامبر سال ۲۰۱۹، چندین مورد ابتلا به نوعی ذات‌الریهی ناشناخته در شهر ووهان^۸ از استان هوی^۹ چین گزارش شد [۱۲]. پس از آن، شیوع این بیماری تنفسی به سرعت در سراسر استان هوی، کشور چین و سایر کشورهای جهان گسترش یافت. به دنبال آزمایشاتی مانند توالی یابی کل ژنوم^{۱۰}، مشخص شد که عامل این بیماری یک نوع ویروس جدید از خانواده‌ی ویروسی کرونا است که مانند سایر اعضای این خانواده، موجب عفونت دستگاه تنفسی در پرندگان و پستانداران می‌شود [۱۳]. تا قبل از همه‌گیری اخیر شش نوع ویروس کرونای انسانی به طور رسمی شناسایی شده بود: HCoV-HKU1، HCoV-OC43، HCoV-NL63 و MERS-CoV، HCoV-229E (SARS-CoV)، HCoV-229E (SARS-CoV)، HCoV-229E (SARS-CoV) [۱۴]. تا قبل از همه‌گیری سارس در سال ۲۰۰۳ توجه خاصی به خانواده کرونا ویروس معطوف نمی‌شد. تا این‌که در پی همه‌گیری MERS در سال ۲۰۱۲ و اخیراً با شیوع کووید-۱۹، همه‌گیری‌های ناشی از شیوع این ویروس،

^۸ Wuhan

^۹ Hubei

^{۱۰}. Whole genome sequencing

نگرانی‌های جهانی را برانگیخت. این خانواده از ویروس‌ها به شدت بیماری‌زا هستند و از خفash‌ها به درختان نخل یا شترهای یک کوهانه و در نهایت به انسان سرایت می‌کنند [۱۵].

ابتدا در ۱۳ ژانویه‌ی ۲۰۲۰، سازمان بهداشت جهانی نام ویروس جدید ۲۰۱۹^{۱۱} را برای این ویروس پیشنهاد داد و سپس، در ۷ فوریه، ۲۰۲۰، چین به طور رسمی از عنوان بیماری تنفسی ویروس کرونای جدید برای اشاره به این بیماری استفاده کرد. چند روز بعد، در ۱۱ فوریه، ۲۰۲۰، سازمان بهداشت جهانی، عنوان بیماری کرونا ویروس ۲۰۱۹^{۱۲} (به اختصار: کووید-۱۹) را بر روی این بیماری گذاشت [۱۶].

کووید-۱۹ عمدهاً با علایمی مانند التهاب ریه شناخته می‌شود و می‌تواند باعث آسیب به دستگاه گوارش، کبد و سیستم عصبی شود [۱۵]. به علاوه، این عفونت ویروسی می‌تواند تب، سرفه، سردرد و سایر علائم مشاهده شده در انواع آنفلانزا را به دنبال داشته باشد.

این ویروس، عمدهاً از طریق دستگاه تنفسی پخش می‌شود. انتقال آئروسل از انسان به انسان بدون شک منبع اصلی سرایت است که عمدهاً از طریق قطرات، دست‌ها یا سطوح آلوده اتفاق می‌افتد. ذرات ویروسی که در ترشحات سیستم تنفسی فرد آلوده وجود دارند، دیگران را از طریق تماس مستقیم با غشاها مخاطی [۱۵] با دوره نهفتگی متوسط بین ۲ تا ۱۲ روز آلوده می‌کنند [۱۷].

بررسی مقالات منتشر شده در اولین سال همه‌گیری کووید-۱۹ نشان می‌داد که ویروس کرونای بومی انسانی، می‌توانند تا ۹ روز بر روی سطوحی مانند فلز، شیشه یا پلاستیک باقی بمانند [۱۵]. همچنین، این نوع ویروس‌ها می‌توانند به طور موثر در عرض ۱ دقیقه با استفاده از یک محلول ضدغونی سطوح، مانند اتانول، پراکسید هیدروژن، یا هیپوکلریت سدیم غیرفعال شوند [۱۸]. به علاوه، بیشتر شواهد موجود این فرضیه را تأیید می‌کنند که فاصله اجتماعی ۱/۵ متری برای جلوگیری از انتقال ویروس از طریق هوای تنفسی کافی است. به نظر می‌رسد که انتقال ویروس تقریباً ۸ روز پس از ظهور علائم همچنان امکان پذیر است [۱۵].

۱-۳- روش‌های تشخیص کووید-۱۹

علایم بالینی کووید-۱۹ از عفونت بدون علامت گرفته تا نارسایی شدید تنفسی، در بین افراد مختلف، متفاوت است. اما، بررسی مقالات منتشر شده در اولین سال همه‌گیری کووید-۱۹ نشان می‌دهد که اکثریت قریب به اتفاق افراد با علائم و الگوهای بالینی شدیدتر، از یک یا چند بیماری زمینه‌ای پزشکی همزمان، مانند فشار خون بالا، دیابت و اختلالات قلبی عروقی، رنج می‌برند [۱۵]. به علاوه، نتایج این تحقیقات نشان می‌داد سن

^{۱۱} Novel coronavirus (2019-ncov)

^{۱۲} Coronavirus disease 2019

^{۱۳} COVID-19

بالا و ضعف سیستم ایمنی، منجر به بروز علایم شدیدتر و همچنین افزایش موارد مرگ و میر در میان سالمدان می‌شود.

از آنجا که تظاهرات بالینی کووید-۱۹ مشابه سایر عفونت‌های ویروسی است، تشخیص افتراقی آن از سایر عفونت‌های ویروسی ضروری خواهد بود. در حال حاضر، عملدهی روش‌های تشخیصی کووید-۱۹ مبتنی بر روش‌های زیست‌شناسی مولکولی مانند واکنش زنجیره‌ای پلیمراز رونویسی معکوس^{۱۴} (RT-PCR)، هستند. اگرچه تست‌های PCR بسیار دقیق بوده و حتی در صورت وجود مقدار ناچیزی از ویروس در نمونه نیز قادر به تشخیص کووید-۱۹ هستند، استفاده از این روش نسبتاً پیچیده‌ی آزمایشگاهی، زمان‌بر بوده و هزینه‌های قابل توجهی را به بیمار و سیستم درمانی تحمیل می‌کند. بنابراین، در بسیاری از مراکز پزشکی و بهداشتی، این خدمات آزمایشگاهی در دسترس نبوده و ارایه نمی‌شود. به علاوه، علیرغم این‌که درصد تشخیص تست-RT-PCR بسیار بالا است، ممکن است نتایج مثبت کاذب به دلیل آلودگی سواب، به ویژه در بیماران بدون علامت گزارش شود.

یکی دیگر از روش‌های آزمایشگاهی تشخیص بالینی کووید-۱۹، مبتنی بر آشکارسازی و شناسایی اسید نوکلئیک ویروس است که به شدت تحت تاثیر میزان/تعداد ویروس در بیماران و روش‌های جمع‌آوری نمونه قرار می‌گیرد. در بسیاری از موارد ممکن است نتیجه‌ی این آزمایش به صورت منفی کاذب گزارش شود. در چندین مطالعه نیز پیشنهاد شده است که شاخص‌های هماتولوژی نیز می‌توانند به عنوان شاخص‌هایی برای تشخیص و همچنین ارزیابی میزان پیشرفت بیماری مورد استفاده قرار گیرند. به عنوان نمونه، در یکی از این تحقیقات، گزارش شده است که در میان بیمارانی که به دلیل اختلالات تنفسی ناشی از کووید-۱۹ در بیمارستان بستری شده‌اند، ناهنجاری‌هایی در تعداد گلبول‌های سفید و لینفوцит‌های خون و همچنین افزایش غلظت آنزیم‌های آلانین آمینوتранسفراز و آسپارتات ترانس آمیناز مشاهده شده است [۲۰]. در تحقیق مشابهی که بر روی ۱۰۹۹ بیمار بستری شده انجام شده، گزارش شده است که در ۸۳٪ مبتلایان، غلظت لینفوцит موجود در خون، در ۳۶٪ موارد غلظت پلاکت خون، و در ۳۴٪ موارد غلظت گلبول‌های سفید به طور معناداری پایین‌تر از حد طبیعی بوده است [۲۱]. در یک مطالعه‌ی دیگر، مقدار خفیفتری از کاهش غلظت پلاکت خون و افزایش غلظت آنزیم‌های آسپارتات ترانس آمیناز^{۱۵} و لاکتات دهیدروژنаз^{۱۶} گزارش شده است [۲۲]. در مراحل پیشرفت‌های بیماری، بالارفتن شاخص‌های تشخیصی التهاب مانند کاهش غلظت نشانگر خونی پروکلسی‌تونین^{۱۷} و افزایش غلظت پروتئین واکنشی سی^{۱۸} مشاهده شده است که می‌تواند بروز ناهنجاری‌هایی

^{۱۴} Reverse transcription polymerase chain reaction

^{۱۵} Aspartate transaminase

^{۱۶} Lactate dehydrogenase

^{۱۷} Procalcitonin

^{۱۸} C-reactive protein

در قلب، کبد و گردن خون را توضیح دهد [۲۱]. در تحقیق مشابهی [۲۲]، ارتباط پیشرفت بیماری با سطح اشبع اکسیژن خون و غلظت پروتئین واکنشی سی در مبتلایان به کووید-۱۹ مورد بررسی قرار گرفته است. نتایج این تحقیق نشان می‌دهد که در سطح خفیف بیماری که سطح اشبع اکسیژن خون هنوز در محدوده طبیعی اندازه‌گیری می‌شود، غلظت پروتئین واکنشی سی به طور متوسط $1/1 \text{ mg/dL}$ اندازه‌گیری شده و در شرایط کاهش شدید سطح اشبع اکسیژن خون، این مقدار به $6/6 \text{ mg/dL}$ افزایش می‌یابد. تحقیق دیگری، همبستگی معناداری را بین غلظت پروتئین واکنشی سی و خطر مرگ در اثر بیماری نشان می‌دهد [۲۳]. در ۷٪ بیماران مبتلا به کووید-۱۹ که بر اثر التهاب شدید و ناگهانی عضلات قلب جان خود را از دست داده‌اند، کاهش پروتئین تروپونین^{۱۹} نیز مشاهده شده است [۲۴]. به نظر می‌رسد که غلظت تروپونین می‌تواند شاخص تعیین‌کننده‌ای در پیش‌بینی حالت‌های شدید بیماری و حتی احتمال مرگ ناشی از این بیماری باشد. در اغلب بیماران بستری شده در مراکز درمانی، غلظت دی-دایمر^{۲۰} و فریتین^{۲۱} نیز بالاتر از حد طبیعی گزارش شده است [۱۵].

علیرغم تحقیقات متعددی که بر روی نتایج آزمایش خون بیماران صورت گرفته، در مورد قدرت پیش‌بینی‌کنندگی شاخص‌های مذکور اجتماعی وجود ندارد و میزان کاربردی بودن این شاخص‌ها در مطالعات مختلف، متفاوت گزارش شده‌اند. به علاوه، از آنجا که در بیشتر این تحقیقات معمولاً فقط یک یا چند شاخص هماتولوژی به عنوان مبنای تشخیص در نظر گرفته می‌شوند، کاربرد نتایج این تحقیقات محدود بوده و تصویر کاملی از عالیم تشخیصی این بیماری، مثلاً در مقایسه با آنقولانزا، به دست نمی‌آید.

ایده‌ی استفاده از تصاویر پزشکی ریه (گرفته شده با تکنیک‌هایی مانند اشعه X، سی‌تی اسکن یا آلتراسوند) به عنوان یک روش تشخیصی جایگزین روش‌های آزمایشگاهی نیز در مطالعات متعددی مورد بررسی قرار گرفته است [۲۵]. نتایج این تحقیقات نشان می‌دهد که ویروس‌های متعلق به این خانواده تظاهرات قابل توجهی را در تصاویر رادیوگرافی نشان می‌دهند [۲۶]. با توجه به این‌که تصویربرداری از قفسه سینه یک روش تشخیصی معمول و نسبتاً کم هزینه بوده و نتایج آن نیز در زمان کوتاهی آماده می‌شود، به نظر می‌رسد که می‌توان آن را جایگزین روش پرهزینه و کمیاب آزمایش PCR کرد. اما، استفاده از این روش چالش‌هایی را نیز در پی دارد. به عنوان مثال، تفسیر این تصاویر به منظور تشخیص بیماری کووید-۱۹ نیازمند حضور مداوم و مستمر پزشک‌های تعلیم یافته در مراکز بهداشتی و درمانی است. این ضرورت، نه تنها بار بیش از حدی را به پزشک تحمیل می‌کند و موجب خستگی و در پی آن افزایش احتمال اشتباه در تشخیص می‌شود، بلکه ارایه‌ی خدمات تشخیصی در مناطق دور افتاده را محدود می‌کند. برای حل این مشکل، در چندین تحقیق،

^{۱۹} Troponin

^{۲۰} D-dimer

^{۲۱} Ferritin

پیشنهاد شده است که با استفاده از الگوریتم‌های هوش مصنوعی و روش‌های یادگیری ماشین، سیستم‌های خودکار تشخیصی طراحی و تعلیم داده شوند [۲۷، ۲۸ و ۲۹]. ادعا می‌شود که در صورت تعلیم یک مدل طبقه‌بندی کننده با دقت و حساسیت بالا می‌توان در مدت زمان بسیار کوتاهی و بدون نیاز به حضور فرد متخصص، بیماری کووید-۱۹ را تشخیص داد. در بیشتر این تحقیقات، از یک شبکه‌ی عصبی عمیق و یا ترکیبی از چند شبکه به عنوان مدل یادگیرنده استفاده شده و عملکرد این مدل‌ها با استفاده از شاخص‌های دقت و حساسیت در تشخیص نمونه‌های بیمار از افراد سالم گزارش شده‌اند. اگرچه نتایج این تحقیقات با استفاده از مجموعه‌ی دادگان مستقل، اغلب، دقت بالای ۹۵٪ را در تشخیص نشان می‌دهد [۲۸]، به علت نبود یک مجموعه‌ی استاندارد دادگان، نمی‌توان این نتایج را به صورت منصفانه قضاوت و ارزیابی کرد. به علاوه، حتی با بهره‌گیری از دانش افراد متخصص نیز، تشخیص کووید-۱۹ با استفاده از تصاویر اشعه‌ی X، سی‌تی اسکن و یا آلتراسوند ریه، نه تنها دقت و حساسیت محدودی را به ویژه در مراحل ابتدایی ابتلا به بیماری، نشان می‌دهد، بلکه همواره یک تاخیر زمانی ۲ تا ۳ روزه بین شروع بیماری و مشخص شدن علایم در تصاویر پزشکی وجود دارد. با این حال، یک نتیجه‌ی مشترک در تمام این ارزیابی‌ها این بوده که استفاده از تصاویر سی‌تی اسکن و آلتراسوند ریه تمایز کافی بین کووید-۱۹ و سایر عفونت‌های ویروسی که مشکلات تنفسی در پی دارند، برقرار نمی‌کند. به عبارتی استفاده از این تصاویر، در رد وجود عفونت مربوط به کووید-۱۹ بهتر از تشخیص آن از سایر مشکلات تنفسی عمل می‌کنند [۳۰]. در میان این روش‌های تصویرگیری، به نظر می‌رسد تصاویر سی‌تی، حساسیت پایین و تشخیص بالایی در دسته‌بندی بیماران بدون علامت مبتلا به کووید-۱۹ را به دنبال دارند.

۱-۴- مانیتورینگ بیمار و ملاحظات مراکز درمانی

در حال حاضر، هیچ داروی ثبت شده‌ای برای درمان بیماری کووید-۱۹ وجود ندارد و اثربخشی واکسن‌هایی که برای سویه‌های قبلی طراحی شده‌اند نیز مورد تردید است. بنابراین، از منظر مدیریت بیماری و همچنین تشخیص منابع تشخیصی و درمانی مراکز بهداشتی و درمانی، هدف اصلی، مانیتورینگ و درمان علایم و تلاش برای جلوگیری از نارسایی تنفسی است. به علاوه، اطمینان از ایزوله سازی بیمار به منظور جلوگیری از انتقال به سایر بیماران، اعضای خانواده و ارائه‌دهندگان مراقبت‌های بهداشتی نیز ضروری است. در حالی که در موارد خفیف، ایزوله شدن در خانه بهترین گزینه بوده و تخته‌های بیمارستانی را برای موارد شدید در دسترس قرار می‌دهد، تشخیص نیاز به بستری شدن و همچنین نیاز به نگهداری در بخش مراقبت‌های ویژه اهمیت بالایی پیدا می‌کند. یک روش نسبتاً پرهزینه در تعیین نیاز به نگهداری در بخش مراقبت‌های ویژه، آشکارسازی و

اندازه‌گیری تراکم در مناطق گراند گلس^{۲۲} موجود در تصاویر رادیولوژی ریه است [۲۶]. روش پیشنهادی دیگری که مورد توجه تحقیق حاضر است، بررسی الگوی تغییرات نشانگرهای خونی کووید-۱۹ با استفاده از اصول داده‌کاوی و پردازش داده‌های جدولی است که در بخش‌های بعد به تفصیل توضیح داده می‌شود.

^{۲۲} Ground-glass opacity

فصل دوم: مرور تحقیقات پیشین

۱-۲ - مقدمه

پس از شیوع کووید-۱۹ در چین و گسترش سریع آن به سایر کشورها، محققان بسیاری در سراسر جهان به تحقیق در مورد ابعاد مختلف این بیماری پرداختند. علاوه بر تحلیل پیامدهای اجتماعی و اقتصادی این همه‌گیری، مطالعات متعددی بر روی روش‌های پیشگیری، تشخیص و درمان این بیماری صورت گرفت. به موازات تحقیقات متعدد پزشکی، پژوههای گوناگونی نیز تعریف شدند که کاربردهای روش‌های داده‌کاوی و یادگیری ماشین و همچنین الگوریتم‌ها و ابزارهای هوش مصنوعی و هوش مصنوعی افزوده^{۳۳} را در تشخیص و پیش‌بینی مورد ارزیابی قرار می‌دهند [۳۱]. بر این اساس، مدل‌های متنوعی برای سنجش اهمیت و قابل اعتماد بودن تصویرگیری پزشکی و همچنین تحلیل دادگان آزمایشگاهی در تشخیص کووید-۱۹ ارائه شدند [۳۲]. مدل‌های دیگری به منظور تشخیص زودهنگام بیماری تعليم یافتند [۳۳] و روش‌هایی نیز جهت پیش‌بینی احتمال بهبود یا مرگ و همچنین احتمال بستری شدن در بخش مراقبت‌های ویژه ارائه شدند [۳۴ و ۳۵]. از دیدگاه روش‌شناسی، اغلب این مدل‌های پیش‌بینی‌کننده یا بر روی دادگان تصویری ریه تعليم یافته‌اند یا بر روی نشانگرهای خونی و محیطی تشخیص کووید-۱۹ که به صورت داده‌های جدولی تحلیل می‌شوند. در پردازش تصاویر سی‌تی، اشعه‌ایکس و همچنین آلتراسوند ریه، به طور معمول، شبکه‌های عصبی عمیق به تنها‌یی و یا در ترکیب با مدل‌های تصمیم‌گیرنده‌ای مانند رگرسیون خطی یا لجستیک^{۳۴} و همچنین ماشین بردار

^{۳۳} Augmented artificial intelligence

^{۳۴} Logistic regression

پشتیبان (SVM)^{۲۵} به کار گرفته شده‌اند. به عنوان مثال، اردکانی و همکاران [۳۲] کاربرد یادگیری عمیق در تصویربرداری سی‌تی را بررسی کردن، که در حال حاضر یک روش بسیار سریع و موثر برای تشخیص ابتلای مراجعین به کووید-۱۹ به شمار می‌رود. این در حالیست که در تحلیل و مدل‌سازی دادگان آزمایشگاهی یا دادگان محیطی/اجتماعی مانند دما، احتمال ارتباط با فرد آلوده و ... (یعنی دادگان جدولی) تا پایان سال ۲۰۲۰، الگوریتم درخت تصمیم بیشترین موارد استفاده را داشته است. پس از آن، سایر طبقه‌بندی‌کننده‌ها، مانند ماشین بردار پشتیبان، مدل بیز ساده^{۲۶}، و مدل K-نزدیک‌ترین همسایگی نیز به منظور تشخیص کووید-۱۹ به کار گرفته شده‌اند [۳۶ و ۳۷]. در این بازه‌ی زمانی، در حل مسائل پیچیده‌تری مانند تعیین شاخص‌های پیش‌بینی‌کننده‌ی خطر مرگ و میر، احتمال بهبود و یا پیش‌بینی احتمال بستره شدن در بخش مراقبت‌های ویژه، علاوه بر روش‌های یاد شده، الگوریتم جنگل تصادفی^{۲۷} نیز مورد استفاده قرار گرفته است [۳۸]. پس از آن و در پی انتشار مجموعه دادگان ثبت شده از بیماران و مراکز درمانی بیشتر، بررسی کاربرد روش‌های یادگیری گروهی/ترکیبی^{۲۸}، چه در مورد تشخیص بیماری از روی دادگان بالینی و چه در مورد تعیین شاخص‌های پیش‌بینی‌کننده‌ی خطر مرگ و میر، مورد استقبال بیشتری قرار گرفتند. به عنوان مثال، [۳۵] از ترکیب طبقه‌بندی‌کننده‌های رگرسیون خطی، ماشین بردار پشتیبان و ماشین یادگیری افراطی^{۲۹} برای پیش‌بینی روند افزایش مرگ و میر در اثر کووید-۱۹ استفاده نمود. در این تحقیق که بر روی دادگان جمع آوری شده از ۷۹ کشور انجام شده، پارامترهای ورودی عبارتند از: تعداد نمونه‌های تایید شده، تعداد و در دسترس بودن ابزار تست، ساختار سنی جامعه‌ی مورد مطالعه، ظرفیت سیستم بهداشتی و درمانی، سطح رفاه اقتصادی، شرایط آب و هوایی و همچنین عوامل خطر (مانند بیماری‌های زمینه‌ای، استعمال دخانیات و ...).

در تحقیقاتی که با موضوع تشخیص ضرورت بستره شدن بیماران کووید-۱۹ در بخش مراقبت‌های ویژه نیز صورت گرفته، اغلب مدل‌های معرفی شده مبنی بر اصول یادگیری ترکیبی هستند. بنابراین، در ادامه‌ی این فصل، ابتدا تاریخچه‌ی مختصه‌ی درباره‌ی این مدل‌ها ارائه شده و سپس، مطالعاتی که با هدف بررسی کاربرد این روش‌ها در تشخیص ضرورت بستره شدن بیماران کووید-۱۹ در بخش مراقبت‌های ویژه صورت گرفته‌اند، اشاره می‌شود.

^{۲۵} Support vector machine

^{۲۶} Naive Bayes

^{۲۷} Random forest

^{۲۸} Ensemble learning

^{۲۹} Extreme learning machine

۲-۲- روش یادگیری ترکیبی

یادگیری ترکیبی که با عنوانین یادگیری تجمعی و یادگیری گروهی نیز شناخته می‌شود، اصطلاحی کلی برای اشاره به روش‌های متعدد را برای تصمیم‌گیری در حل مسائل طبقه‌بندی یا رگرسیون ترکیب می‌کنند. ایده‌ی اصلی در این روش‌ها که معمولاً در برخورد با وظایف یادگیری با سرپرستی (تحت ناظارت) مورد استفاده قرار می‌گیرند، این است که با ترکیب چند مدل، که به طور جداگانه روی مجموعه‌ی دادگان ورودی-خروجی تعلم داده شده‌اند، خطاها را که مدل یادگیرنده‌ی منفرد احتمالاً توسط سایر مدل‌ها جبران می‌شود. در نتیجه، عملکرد کلی، یعنی پیش‌بینی نهایی گروه، بهتر از یک مدل منفرد خواهد بود. این مدل می‌تواند ترکیب هر نوع الگوریتم یادگیری ماشینی (به عنوان مثال درخت تصمیم، شبکه عصبی، مدل رگرسیون خطی و غیره) باشد. به علاوه، با توجه به قابلیت پردازش موازی در پردازنده‌های امروزی، تعلم این مدل‌های ترکیبی به مدت زمان بیشتری نسبت به تعلم یک مدل منفرد نیاز ندارد و به صورت بهینه قابل پیاده‌سازی و اجرا است.

در توضیح عملکرد چشمگیر این مدل‌های یادگیری ترکیبی در بهبود قدرت پیش‌بینی‌کنندگی مدل می‌توان به عوامل زیر اشاره نمود [۳۹]:

الف- اجتناب از بیش‌برازش^۰: در برخورد با مجموعه دادگان کم‌تعداد، یک مدل منفرد با تعداد بالای پارامترهای یادگیرنده، مستعد یافتن فرضیه‌های مختلف است که هر یک می‌توانند تمام داده‌های آموزشی (داده‌های تعلم) را به طور کامل پیش‌بینی کنند، در حالی که پیش‌بینی‌های ضعیفی برای نمونه‌های دیده نشده (داده‌های آزمایش) ارائه می‌دهند. حال اگر، یک الگوریتم یادگیری گروهی، به عنوان مثال، میانگین این فرضیه‌های مختلف را مبنای تصمیم‌گیری قرار دهد، خطر انتخاب یک فرضیه نادرست کاهش یافته و بنابراین، عملکرد کلی پیش‌بینی بهبود می‌پابد.

ب- مزیت محاسباتی: در حالی که ممکن است یک مدل یادگیرنده‌ی منفرد در جستجوی مجموعه‌ی پارامترهای یادگیرنده‌ی بهینه در یک نقطه‌ی بهینه محلی^۱ گیر کند، روش‌های گروهی، با ترکیب چندین یادگیرنده، خطر افتادن در یک نقطه‌ی کمینه‌ی محلی را کاهش می‌دهند.

ج- بازنمایی ویژگی‌های دادگان: در مسائل یادگیری ماشین، بسیار ممکن است که فرضیه بهینه، که ارتباط بین دادگان ورودی و خروجی را توصیف می‌کند، خارج از فضای هر مدل واحد باشد. بنابراین، با ترکیب مدل‌های مختلف، فضای جستجوی فرضیه بهینه، ممکن است گسترش یابد و از این رو، تناسب بهتری با فضای داده حاصل می‌شود.

^۰ Overfitting

^۱ Local optima

علاوه بر مزیّت‌هایی که در بالا به آن اشاره شد، مدل‌های یادگیری ترکیبی پتانسیل خوبی برای چیره شدن بر مشکل عدم توازن تعداد نمونه‌های کلاس‌های مختلف در مسئله‌ی طبقه‌بندی دارند. در بسیاری از تحقیقاتی که با دادگان پزشکی سر و کار دارند، تعداد نمونه‌های بیمار، به ویژه در مراحل حاد بیماری، بسیار کمتر از تعداد نمونه‌های سالم است. این موضوع، اغلب، موجب سوگیری مدل به سمت شناخت نمونه‌های سالم (یعنی گروه/کلاسی که نمونه‌های بیشتری دارد) شده و موقع ارزیابی، تعداد منفی‌های کاذب سایر کلاس‌ها را افزایش می‌دهد. این مسئله در مواردی که تشخیص زودهنگام و یا تشخیص دقیق شدت بیماری اهمیت بالایی دارند، قابلیت اطمینان و کارایی مدل را پایین می‌آورد. برای غلبه بر این مشکل، پیشنهاد می‌شود که در حل چنین مسائلی از روش یادگیری ترکیبی به گونه‌ای استفاده شود که هریک از مدل‌هایی که در مدل گروهی به کار می‌روند روی یک زیرمجموعه از دادگان تعلیم یابند که در آن تعداد نمونه‌های کلاس‌های مختلف متوازن است [۴۰]. به علاوه، پیش از این اثربخشی ترکیب تکنیک کم-نمونه برداری تصادفی و روش‌های یادگیری ترکیبی مانند بگینگ و بوستینگ^{۳۲} در مواجهه با موضوع عدم توازن تعداد نمونه‌ها در کاربردهای غیرپزشکی نشان داده شده است [۴۱].

به منظور ساخت یک مدل یادگیری گروهی، ابتدا باید روش تعلیم هر یک از مدل‌های شرکت‌کننده در گروه و همچنین روش ترکیب خروجی مدل‌های شرکت‌کننده را تعیین نمود. همچنین در انتخاب مدل‌هایی که با یکدیگر ترکیب شده و مدل گروهی را شکل می‌دهند باید به گونه‌ای عمل کرد که ۱) تک‌تک این مدل‌ها با دقت بالاتر از یک مدل تصادفی قادر به حل مساله‌ی مورد نظر باشند و ۲) مدل‌ها به صورت متنوع انتخاب شوند [۴۲]. در مطالعات مختلف، تکنیک‌های زیر برای تحقق این دو شرط پیشنهاد شده‌اند [۳۹]:

الف - دستکاری ورودی: در این رویکرد، هر مدل شرکت‌کننده در مدل گروهی با استفاده از یک زیرمجموعه‌ی کوچک‌تر از مجموعه‌ی دادگان آموزش می‌بیند، به طوری که مدل‌های پایه، ورودی‌های متنوعی خواهند داشت.

ب - الگوریتم یادگیری دستکاری شده: در این تکنیک، هر مدل پایه یا به روش‌های مختلف و یا با تغییر ابرپارامترها تعلیم می‌باید. به این صورت، مدل‌های مختلف مسیرهای همگرایی متفاوتی را در طول تعلیم طی می‌کنند و شرط تنوع مدل‌های شرکت‌کننده در یادگیری گروهی تحقق می‌یابد.

ج - پارتیشن‌بندی: تنوع را می‌توان با تقسیم مجموعه داده اصلی به زیرمجموعه‌های کوچکتر و سپس استفاده از هر زیر مجموعه برای آموزش یک مدل پایه‌ی متفاوت نیز به دست آورد. تفاوت این روش با تکنیک دستکاری ورودی در عدم وجود همپوشانی میان زیرمجموعه‌های مورد استفاده برای تعلیم است. به علاوه، پارتیشن‌بندی دادگان را می‌توان به دو شکل پارتیشن‌بندی افقی و عمودی پیاده‌سازی

نمود. در پارتيشن‌بندی افقی، مجموعه داده اصلی را بر اساس تعداد مشاهدات به چندین مجموعه تقسیم می‌کنیم که هر مشاهده، کل مجموعه ویژگی‌ها را با خود به همراه دارد. به این صورت، مدل‌های شرکت‌کننده در یادگیری گروهی، با استفاده از نمونه‌های متفاوت تعلیم می‌یابند [۴۳]. در پارتيشن‌بندی عمودی، عملیات تقسیم دادگان به زیرمجموعه‌ها بر روی ویژگی‌ها اعمال می‌شود به طوری که هر مدل، تمام نمونه‌های موجود در دادگان اصلی را جهت تعلیم مشاهده می‌کند اما هر نمونه تمام ویژگی‌های ثبت شده را با خود به همراه ندارد و فقط روی مجموعه‌ی کوچک‌تری از ویژگی‌ها آموزش می‌یابند [۴۴].

د- هیبریداسیون گروه^{۳۳}: این رویکرد در هنگام ساخت مجموعه، حداقل دو استراتژی را از میان تکنیک‌های بالا پیاده‌سازی می‌کند. به عنوان مثال در الگوریتم جنگل تصادفی به عنوان یکی از روش‌های یادگیری گروهی، همان ابتدا و در هنگام ساخت هر درخت، نمونه‌ها^{۳۴} دستکاری می‌شوند. به علاوه در الگوریتم یادگیری نیز با انتخاب تصادفی زیرمجموعه‌ای از ویژگی‌ها در هر گره، تنوع ایجاد می‌گردد.

۱-۲-۲- روش‌های ترکیب خروجی یادگیرنده‌ها

پس از انتخاب مدل‌های پایه و همچنین روش تعلیم این مدل‌ها، نوبت به بررسی استراتژی‌های موجود برای ترکیب خروجی مدل‌های شرکت‌کننده در گروه می‌رسد. از آنجا که مسئله‌ی مورد بررسی در این تحقیق، یک مسئله‌ی طبقه‌بندی دودویی است، در ادامه فقط به روش‌هایی اشاره می‌شود که در این نوع طبقه‌بندی‌کننده‌ها کاربرد دارند.

یکی از راه حل‌های طبیعی برای ترکیب خروجی‌های چند مدل، وزن‌دهی هر مدل پایه براساس رای اکثریت^{۳۵} است که در آن، کلاس انتخابی برای هر نمونه، کلاسی است که بیشترین آرا را دارد. یکی دیگر از رویکردهای وزن‌دهی، تعیین وزنی متناسب با عملکرد مدل بر روی مجموعه‌ی دادگان مورد استفاده برای اعتبارسنجی^{۳۶} است. در روش‌های ترکیبی بیزین^{۳۷} نیز وزن‌دهی بر اساس درست‌نمایی احتمالاتی^{۳۸} به دست آوردن مدل به شرط استفاده از کل مجموعه داده صورت می‌گیرد [۳۹].

^{۳۳} Ensemble hybridization

^{۳۴} Instances

^{۳۵} Majority voting

^{۳۶} Validation

^{۳۷} Bayesian

^{۳۸} Probabilistic likelihood

استفاده از روش‌های فرایادگیری^{۳۹} یکی دیگر از استراتژی‌هایی است که در ترکیب خروجی‌های مدل‌های شرکت کننده در گروه به کار می‌رود. در روش فرایادگیری، خروجی تک تک مدل‌های شرکت کننده به عنوان ورودی‌های یک مدل فرایادگیرنده عمل می‌کنند که خروجی نهایی را تولید می‌کند. روش‌های فرایادگیری در مواردی که مدل‌های پایه خاصی عملکرد متفاوتی در زیرفضاهای مختلف دارند، به خوبی کار می‌کنند. به عنوان مثال، زمانی که مدل‌های پایه به طور مداوم به درستی نمونه‌های خاصی را طبقه‌بندی می‌کنند یا به طور مداوم در طبقه‌بندی یک سری از نمونه‌ها اشتباه می‌کنند. از میان کاربردی‌ترین تکنیک‌های فرایادگیری می‌توان به روش انباشتن^{۴۰} اشاره کرد [۴۵]. در این روش، ابتدا یک مجموعه فراداده ایجاد می‌شود که شامل همان تعداد نمونه موجود در مجموعه داده اصلی است. با این تفاوت که به جای استفاده از ویژگی‌های ورودی اصلی، از خروجی مدل‌های یادگیرنده به عنوان ورودی مدل فرآگیر استفاده می‌شود. این مدل فرآگیر برای پیش‌بینی همان برچسب‌های اصلی نمونه‌ها که پیشتر برای تعلیم مدل‌های یادگیرنده نیز استفاده شده بودند، آموزش داده می‌شود. در هنگام ارزیابی عملکرد مدل نهایی نیز، هر نمونه‌ی آزمایش ابتدا به همه‌ی مدل‌های یادگیرنده داده شده و خروجی این مدل‌ها جمع‌آوری می‌شود. سپس، این بردار خروجی به عنوان ورودی به مدل فرایادگیرنده اعمال شده و نتیجه‌ی نهایی محاسبه می‌شود. برای اطمینان از عدم وقوع بیش‌برازش، توصیه شده است که مجموعه‌ی دادگان اصلی بر اساس تعداد نمونه‌ها به دو زیر مجموعه تقسیم شوند که یکی برای تعلیم مدل‌های پایه و دیگری برای تعلیم مدل فرآگیر مورد استفاده قرار می‌گیرند [۳۹].

۲-۲-۱- استقلال یا وابستگی یادگیرنده‌ها

یکی دیگر از جنبه‌های مهم در یادگیری گروهی، استقلال یا وابستگی مدل‌های یادگیرنده در حین آموزش و آزمایش است. اگرچه می‌توان مدل‌های یادگیرنده را مستقل از یکدیگر طراحی نمود و تعلیم داد، تکنیک‌هایی نیز وجود دارند که توصیه می‌کنند در هر تکرار، دانش به دست آمده در هر مدل پایه، با سایرین به اشتراک گذاشته شود. به این صورت که خروجی هر مدل یادگیرنده در هر تکرار به خروجی سایر مدل‌های یادگیرنده در تکرارهای قبل وابسته است. این تکنیک به عنوان مثال در روش‌های AdaBoost و تقویت گرادیان^{۴۱} مورد استفاده قرار می‌گیرد.

ایده‌ی اصلی تکنیک AdaBoost تمرکز بر نمونه‌هایی است که قبلاً هنگام آموزش یک مدل پایه‌ی جدید به اشتباه طبقه‌بندی شده بودند [۴۶]. میزان اهمیت هر یک از این نمونه‌ها توسط وزنی تعیین می‌شود که به هر نمونه در مجموعه‌ی آموزش اختصاص داده می‌شود. به این صورت که در اولین تکرار، همه‌ی نمونه‌ها وزن یکسانی دارند، پس از آن و در هر تکرار، وزن نمونه‌هایی که به اشتباه طبقه‌بندی شده افزایش می‌یابد، در حالی

^{۳۹} Meta-learning

^{۴۰} Stacking

^{۴۱} Gradient boosting

که وزن نمونه‌هایی که به درستی طبقه‌بندی شده‌اند کاهاش می‌یابد. علاوه بر این، وزن‌هایی نیز به مدل‌های پایه‌ی شرکت کننده در یادگیری گروهی اختصاص داده می‌شود که متناسب با عملکرد هر مدل در انجام تسک^{۴۲} است. از آنجا که وزن نمونه‌ها در هر تکرار به عملکرد همه‌ی مدل‌های شرکت کننده وابسته است، این روش تکراری مجموعه‌ای از یادگیرنده‌گان پایه را فراهم می‌کند که یکدیگر را تکمیل می‌کنند. از آنجا که این روش، ممکن است منجر به بیش‌برازش، به ویژه در مجموعه دادگان کوچک شود، نسخه‌های تغییریافته‌ای از آن نیز در مطالعات دیگری پیشنهاد شده‌اند. از میان این نسخه‌ها، می‌توان به AdaBoost با حاشیه‌ی نرم^{۴۳} و Modest AdaBoost اشاره کرد که با در نظر گرفتن تنظیم‌کننده‌هایی^{۴۴}، قابلیت تعمیم مدل را بهبود می‌دهند [۳۹].

در روش تقویت گرادیان [۴۷]، الگوریتم تعلیم مدل گروهی برای کمینه کردن یکتابع هزینه‌ی کل طراحی می‌شود. ایده‌ی اولیه‌ی این تکنیک این است که تقویت^{۴۵} را می‌توان به عنوان یک الگوریتم بهینه سازی بر روی یکتابع هزینه‌ی مناسب تفسیر کرد. به این صورت که در یک الگوریتم تعلیم تکراری و برای پیدا کردن نقطه‌ی بهینه در فضای تابع هزینه، باید در هر تکرار، مدل‌های یادگیرنده‌ای را انتخاب و به مدل اولیه اضافه کرد، به گونه‌ای که برایند عملکرد همه‌ی یادگیرنده‌ها در جهت منفی گرادیان تابع هزینه باشند. یادگیرنده‌ها در این روش، اغلب، درختان تصمیم کم عمق هستند که به تنها یک عملکرد مطلوبی در حل مسئله‌ی طبقه‌بندی ندارند. بنابراین، انتخاب مناسب تعداد درختان (یعنی تعداد تکرارها) در این روش از اهمیت بالایی برخوردار است. در حالی که تعداد پایین درختان، اثر منفی در دقت مدل دارد، تعداد بالای مدل‌های یادگیرنده نیز می‌تواند موجب بیش‌برازش شود. لذا، انتخاب مناسب‌ترین تعداد تکرار معمولاً با استفاده از یک مجموعه اعتبارسنجی برای ارزیابی عملکرد کلی پیش‌بینی انجام می‌شود. به علاوه، برای کاهاش احتمال بیش‌برازش مدل، روش‌هایی مانند تقویت تصادفی گرادیان [۴۸ و ۴۹] معرفی شده‌اند. در روش افزایش گرادیان تصادفی، درختان تصمیم (یا به طور کلی، یادگیرنده‌های پایه) به طور متواالی و با استفاده از زیرمجموعه‌های کوچک که به طور تصادفی از مجموعه داده اصلی نمونه‌برداری شده‌اند، آموزش می‌یابند. از سوی دیگر، روش XGBoost^{۴۶} که در سال‌های اخیر کاربردهای بسیاری در حل مسائل مربوط به یادگیری ماشین پیدا کرده است، با معرفی یک سری تصحیحات و بهینه‌سازی‌های الگوریتمی، مقیاس‌پذیری^{۴۷} روش‌های تقویت گرادیان را به طور چشمگیری افزایش داده است. این روش، اولاً، الگوریتم‌هایی را برای فرایند تقسیم یک گره به چندین زیرگره

^{۴۲} Task

^{۴۳} Soft margin AdaBoost

^{۴۴} Regularizer

^{۴۵} Boosting

^{۴۶} Extreme gradient boosting

^{۴۷} Scalability

(در درخت تصمیم به عنوان مدل یادگیرنده) ارائه می‌کند که ۱) داده‌های خلوت^{۴۸} را با جهت‌های پیش‌فرض گره‌ها مدیریت می‌کند، ۲) داده‌های وزن‌دار را با استفاده از عملیات ادغام و هرس آدرس‌دهی می‌کند، ۳) به طور مؤثر بر روی تمام تقسیم‌های ممکن شمارش می‌کند تا آستانه‌ی تقسیم بهینه شود. به علاوه، XGBoost به تابع هزینه، یک گزینه‌ی تنظیم کننده نیز اضافه می‌کند که احتمال بیش‌برازش را به خوبی کاهش می‌دهد. از آنجا که در هنگام پیاده‌سازی XGBoost می‌توان از پلتفرم‌های توزیع شده و همچنین امکانات پردازش موازی استفاده نمود، این مدل‌ها می‌توانند سریع‌تر از مدل‌های دیگر اجرا شوند.

۲-۳- پیش‌بینی کامپیوتری احتمال بسترهای بخش مراقبت‌های ویژه

در اواسط سال ۲۰۲۰، در یک تحقیق جامع [۵۰]، مجموعه‌ای از ۶۴۶ مقاله‌ی مرتبط با شرایط بسترهای بیماران مبتلا به کووید-۱۹ در بخش مراقبت‌های ویژه^{۴۹} مورد بررسی قرار گرفته‌اند که در پایان، تنها ۳۷ مقاله و در مجموع، اطلاعات ۲۴۹۸۳ بیمار، مرتبط با این موضوع در نظر و مورد ارزیابی نهایی قرار گرفته‌اند. در پایان‌بندی این تحقیق آمده است که نرخ بسترهای بیماران در بخش مراقبت‌های ویژه ۳۲٪ برآورد می‌شود. به علاوه، تخمین زده شده که ۳۹٪ از افراد بسترهای در ICU تا زمان گزارش این تحقیق، جان خود را از دست داده‌اند. نرخ بالای مرگ و میر گزارش شده در این پژوهش، به مدیران مراکز بهداشتی این هشدار را می‌دهد که بسیاری از بیماران یا بسیار دیر به ICU انتقال داده شده‌اند و یا این‌که از توجه و مراقبت کافی در زمان بسترهای بودن برخوردار نبوده‌اند. نتایج این تحقیق و تحقیقات مشابهی که بر ضرورت تشخیص زمان بهینه‌ی انتقال بیمار به بخش مراقبت‌های ویژه تاکید می‌کردند، محققان حوزه‌های هوش مصنوعی، یادگیری ماشین و داده‌کاوی را بر آن داشت که تحقیقات خود را بر روی تحلیل و مدل‌سازی داده‌های ثبت شده از بیماران بسترهای شده در بیمارستان و سپس انتقال یافته به بخش مراقبت‌های ویژه آغاز کنند. هدف اغلب این تحقیقات این است که احتمال نیاز به انتقال بیمار به ICU، به صورت سریع و کم‌هزینه و با مطالعه‌ی تغییرات تصاویر پزشکی و یا نشانگرهای خونی مبتلایان به کووید-۱۹ برآورد شود.

تحقیقاتی که تاکنون در زمینه‌ی تشخیص بیماری کووید ۱۹ و یا پیش‌بینی احتمال بسترهای بیمار در بخش مراقبت‌های ویژه انجام شده‌اند نشان می‌دهند با دسترسی به نتایج معاینات بالینی افراد مبتلا و به کارگیری الگوریتم‌های یادگیری ماشین می‌توان با دقت مناسبی، آینده‌ی وضعیت بیمار از نظر بهبودی یا تشدید عالیم بیماری را پیش‌بینی نمود. از آنجا که دادگان مورد استفاده در تحقیق حاضر، از نوع دادگان جدولی و ثبت شده از طریق آزمایشات کلینیکی هستند، به مقالاتی که الگوریتم‌های یادگیری ماشین را بر تصاویر و یا داده‌های

^{۴۸} Sparse

^{۴۹} Intensive care unit (ICU)

اجتماعی/محیطی اعمال کرده‌اند، اشاره نمی‌شود. به علاوه، بیشترین تمرکز این بخش، بر مقالاتی معطوف است که مسئله‌ی برآورده احتمال بسترهای شدن بیمار در بخش مراقبت‌های ویژه را مورد مطالعه قرار داده‌اند. در مقالات موری [۵۱ و ۵۲]، خلاصه‌ی خوبی از تحقیقات در این دو حوزه جمع‌آوری و منتشر نموده‌اند. روش‌های ارائه شده در اغلب این پژوهش‌ها مشتمل بر دو بخش استخراج/انتخاب ویژگی‌ها و تعلیم مدل یادگیرنده هستند که در ادامه‌ی این بخش به آن‌ها پرداخته می‌شود.

۲-۱-۳- استخراج یا انتخاب ویژگی‌های مناسب

مرحله‌ی انتخاب ویژگی که با هدف کاهش بُعد مسئله صورت می‌گیرد، منجر به کاهش پیچیدگی‌های محاسباتی (از نظر مدت زمان تعلیم، تعداد پارامترهای مدل و ذخیره‌ی مدل تعلیم یافته) شده و امکان تحلیل و شناخت بهتر متغیرهای مؤثر را در اختیار قرار می‌دهد. در اغلب مطالعاتی که بر روی دادگان جمع‌آوری شده از مبتلایان به کووید-۱۹ انجام شده، از یکی از روش‌های فیلتر یا wrapper استفاده شده است. در روش فیلتر که اغلب به صورت مهندسی ویژگی‌های موجود پیاده‌سازی می‌شود، مشخصات آماری ویژگی‌ها و یا همبستگی آن‌ها با خروجی مطلوب مدل مبنای حذف یا حفظ متغیرهای موجود قرار می‌گیرند. واریانس متغیرها، آنتروپی، ضریب جینی^{۵۰}، بهره‌ی اطلاعات^{۵۱}، میزان همبستگی^{۵۲} متغیرها و شاخص Chi-Square از جمله شاخص‌هایی هستند که برای انتخاب ویژگی‌های مناسب در تحلیل دادگان پزشکی مربوط به بیماری کووید ۱۹ مورد استفاده قرار گرفته‌اند [۵۳، ۵۴ و ۵۵].

در یکی از این پژوهش‌ها، روش مهندسی ویژگی‌ها که نیازمند دانش مقدماتی نسبت به مجموعه‌ی دادگان است، با wrapper جایگزین شده که در آن، مدل طبقه‌بندی‌کننده روی زیرمجموعه‌های کوچکی از مجموعه دادگان اصلی تعلیم یافته و عملکرد آن ارزیابی می‌شود. سپس، زیرمجموعه‌ای که بهترین عملکرد را به دنبال داشته است به عنوان بردار ویژگی برای تعلیم مدل اصلی مورد استفاده قرار می‌گیرد [۵۶]. بر روی دادگان غیر پزشکی و در هنگام استفاده از مدل‌های مانند درخت تصمیم و مدل بیز ساده، نشان داده شده که این روش در مقایسه با روش فیلتر برای انتخاب ویژگی‌ها، منجر به تعلیم مدل یادگیرنده (در مسئله‌ی طبقه‌بندی دادگان ورودی) با دقت بالاتری می‌شود [۵۷]. در یکی از پژوهش‌های مربوط به تشخیص کامپیوتری کووید-۱۹، از این روش برای انتخاب ویژگی‌های مناسب برای تعلیم یک مدل طبقه‌بندی‌کننده‌ی XGBoost استفاده شده است که با تعلیم تکراری مدل مشابه و با استفاده از بردارهای ویژگی وزن‌دار، تعداد ویژگی‌ها ابتدا از ۱۶۵ به ۵۰ و سپس به ۲۰ ویژگی کاهش یافته است [۵۶].

۵۰ Gini

۵۱ Information gain

۵۲ Correlation

در ادبیات تحقیق یادگیری ماشین و کاربردهای آن در حل مسائل مهندسی، علاوه بر روش‌های فیلتر و wrapper از الگوریتم ژنتیک، روش‌های جستجوی مستقیم و معکوس ترتیبی، جستجوی مستقیم و معکوس ترتیبی تعمیم یافته [۵۸] و ترکیب الگوریتم‌های خوش‌بندی و رده‌بندی ویژگی‌ها [۵۹] نیز برای انتخاب بهترین ویژگی‌ها استفاده می‌شود که در تحقیقات مربوط به تشخیص بیماری و یا ارزیابی شدت کووید-۱۹ به کار گرفته نشده‌اند. به علاوه، روش‌های استخراج ویژگی مانند Matrix Factorization، یادگیری منیفلد^{۵۳} و الگوریتم‌های مبتنی بر Autoencoder نیز هنوز راهی به این تحقیقات نیافته‌اند و بررسی تأثیر آن‌ها بر عملکرد مدل، به صورت بالقوه، موضوع مناسبی برای تحقیق است.

۲-۳-۲- انتخاب و تعلیم مدل‌های یادگیرنده

در بخش تعلیم مدل یادگیرنده در تحقیقاتی که به منظور تشخیص کووید-۱۹ یا تعیین احتمال نیاز به استری شدن در بخش مراقبت‌های ویژه انجام گرفته‌اند، الگوریتم‌های مبتنی بر درخت تصمیم، رگرسیون لجستیک و الگوریتم جنگل تصمیم تصادفی با استفاده از ویژگی‌های بالینی و نتایج آزمایش‌های خون بیماران تعلم داده شده‌اند [۵۴، ۵۵ و ۵۶]. در یکی از این پژوهش‌ها و با استفاده از داده‌های جمع‌آوری شده در دو بیمارستان در چین، گزارش شده است که با استخراج ویژگی‌هایی مانند میزان آنزیم آلامین آمینوترانسفراز در کبد، درصد درد ماهیچه و میزان هموگلوبین خون می‌توان با دقیق حدود ۰/۷، وقوع موارد حاد را پیش‌بینی نمود [۵۳]. در تحقیق دیگری که با هدف پیش‌بینی وقوع حالت‌های حاد در افراد مبتلا به کووید-۱۹ با استفاده از الگوریتم XGBoost انجام شده است، ویژگی‌هایی مانند سن، وجود آسیب‌ها و بیماری‌های کلیوی، افزایش لاکراتات دهیدروژناز (LDH)، تندنفسی و هیپرگلیسمی به عنوان ویژگی‌های اصلی پیش‌بینی کننده معرفی شده‌اند [۶۰]. نتایجی که در تحقیق [۶۱] منتشر شده نیز نشان می‌دهد که با تعلیم یک مدل XGBoost و این بار با استفاده از یک مجموعه دادگان مربوط به ۳۷۵ بیمار (۲۰۱ بیهود یافته) در بیمارستان دانشگاه تونگجی در ووهان، احتمال مرگ ناشی از این بیماری با دقت ۹۳٪ پیش‌بینی شده است. در این پژوهش نیز ویژگی‌هایی مانند LDH، لنفوسيت، و پروتئين واکنشی سی به عنوان ویژگی‌های مهم برای پیش‌بینی معرفی شده‌اند که این نتایج هماهنگی خوبی با نتایج مطالعات پزشکی، که در فصل ۱ به آن‌ها اشاره شد، دارند. در تحقیقات مشابهی، همین الگوریتم با استفاده از ویژگی‌های دیگری مانند نرخ تنفس، ضربان قلب، شاخص توده بدنی، میزان نیتروژن و کراتینین خون، بر روی دادگان جمع‌آوری شده در بیمارستان‌های مختلف تعلم داده شده و پیش‌بینی احتمال مرگ با دقت‌های بالاتر از ۹۰٪ گزارش شده است [۶۲، ۶۳ و ۶۴].

علاوه بر مدل‌های مبتنی بر XGBoost، از مدل ماشین بردار پشتیبان و رگرسیون لجستیک نیز برای پیش‌بینی حالت‌های حاد و نیز احتمال مرگ مبتلایان استفاده شده است [۶۵، ۶۶، ۶۷ و ۶۸]. در این پژوهش‌ها ویژگی‌هایی

^{۵۳} Manifold learning

از نشانگرهای سرمی (به عنوان مثال، کلسیم، اسید لاکتیک و آلبومین، گلوتاتیون ، لنفوسیت‌های T بالغ و پروتئین تام) برای تعلیم و یا تخمین پارامترهای مدل استفاده شده است.

در تشخیص بیماری کووید-۱۹ با استفاده از دادگان مربوط به نتایج آزمایش خون ۲۷۹ مراجعه کننده (۱۷۷ مبتلا)، در مقایسه با الگوریتم‌های مختلف مانند درخت تصمیم، K-نزدیکترین همسایگی، ماشین بردار پشتیبان مدل بیز ساده و رگرسیون لجستیک، الگوریتم جنگل تصمیم تصادفی با دقت ۰/۸۲، حساسیتی برابر با ۰/۹۲ و تشخیص ۰/۶۵ بهترین عملکرد را نشان داده است [۶۹]. نتایج این تحقیق همچنین نشان می‌دهد که این روش نسبت به نقص دادگان و یا عدم تعادل در تعداد نمونه‌های مشاهده شده در هر کلاس نیز حساسیت کمتری دارد. این الگوریتم در پیش‌بینی پذیرش و مرگ و میر مبتلایان به کووید-۱۹ در ICU نیز در ترکیب با سایر مدل‌ها موفق عمل کرده است. در تحقیقی که با هدف مقایسه ۱۸ مدل یادگیری ماشین در پیش‌بینی انتقال بیمار به بخش مراقبت‌های ویژه روی دادگان ۳۵۹۷ مراجعه کننده به مرکز درمانی Mass General Brigham (MGB) در بوستون امریکا صورت گرفته، نشان داده شده است که الگوریتم جنگل تصمیم تصادفی به تنها یکی یا در ترکیب با طبقه‌بندی کننده‌های Adaboost، بگینگ، ExtraTrees و XGBoost می‌تواند با میانگین اف ۱^{۰۴} بالاتر از ۰/۸۰ احتمال پذیرش در ICU را به درستی پیش‌بینی کند [۷۰]. در همین تحقیق عملکرد الگوریتم درخت تصمیم و رگرسیون لجستیک با میانگین امتیاز برابر اف ۱ با ۰/۷۸ و ۰/۷۷ گزارش شده است. در این پژوهش نیز، سطح اشتعاع اکسیژن خون و لاکتات دهیدروژناز (LDH) ویژگی‌های مهم برای تعیین نیاز به بستری شدن در ICU معرفی شده‌اند.

در تحقیق دیگری که بر روی دادگان جمع‌آوری شده از ۱۳۶۱ بیمار بستری شده در بیمارستان San Raffaele شهر میلان در ایتالیا صورت گرفته، نشان داده شده که یک ماشین بردار پشتیبان پس از تنظیم مناسب ابرپارامترهایش، می‌تواند عملکرد خوبی در پیش‌بینی نیاز بیماران به بستری شدن در ICU داشته باشد [۷۱]. در گزارش عملکرد این مدل، معیارهای دقت طبقه‌بندی و مساحت زیر منحنی^{۵۵} (AUC) به ترتیب برابر با ۰/۸۲ و ۰/۸۵ اعلام شده است. به علاوه، عملکرد این مدل با مدل‌هایی مانند درخت تصمیم، XGBoost و پرسپترون چندلایه مقایسه شده که به ترتیب، مساحت‌های زیر منحنی ۰/۸۱ و ۰/۷۱ را به دنبال داشته‌اند. در میان این مدل‌ها، درخت تصمیم گیری، بیشترین حساسیت (یعنی ۰/۶۶) را داشته و پرسپترون چندلایه بالاترین درجه‌ی تشخیص (۰/۹۰) را نشان داده است.

در پژوهش دیگری [۷۲]، یک مدل گروهی سلسله مراتبی برای پیش‌بینی عوارض جانبی در طول بستری، از جمله احتمال بستری در ICU، بر روی مجموعه کوچکی (۲۲۹ بیمار) از ویژگی‌های آزمایشگاهی و بالینی

^{۵۴} F1-score

^{۵۵} Area under the curve

موجود در زمان بستری تعلیم داده شده است. با توجه به کوچکی مجموعه دادگان، این مدل، عملکرد نسبتاً پایینی با میانگین حساسیت ۰/۳۴ و دقت متوسط ۰/۶۴ داشته است.

بر روی یک مجموعه دادگان بزرگ‌تر (۱۹۸۲ بیمار) در [۷۳]، یک مدل جنگل تصمیم تصادفی بر روی داده‌های آزمایشگاهی، ویژگی‌های بالینی و ECG در زمان پذیرش، تعلیم داده شده که احتمال انتقال به ICU را با حساسیت ۰/۷۲، تشخیص ۰/۷۶، دقت ۰/۷۶ و سطح زیر منحنی ۰/۸۰. در حالی که این نتایج، تعادل خوبی بین حساسیت و تشخیص را نشان می‌دهد، این مدل از تعداد قابل توجهی از ویژگی‌ها، از جمله پارامترهای آزمایشگاهی، علائم، و نتایج الکتروکاردیولوژیک استفاده می‌کند. بنابراین، ممکن است استفاده از آن در کاربردهای بیمارستانی، به ویژه در موقعیت‌هایی که مرکز بهداشتی با محدودیت منابع و نیروی انسانی مواجه است، توجیه‌پذیر نباشد. به علاوه، از آنجا که نویسنده‌گان به جزئیاتی در مورد روش‌های استخراج ویژگی و پیش‌پردازش دادگان اشاره نکرده‌اند، تکرار پذیری مطالعه، امکان پذیر به نظر نمی‌رسد.

در [۷۴] یک مدل جنگل تصمیم تصادفی برای پیش‌بینی نیاز به بستری در ICU براساس داده‌های آزمایشگاهی، بالینی و جمعیت‌شناختی ۱۰۴۰ بیمار توسعه داده شده و حساسیت، تشخیص و سطح زیر منحنی در ارزیابی عملکرد آن ترتیب برابر با ۰/۹۱، ۰/۸۷ و ۰/۹۶ گزارش شده است. علیرغم این که ارزیابی کمی نشان‌دهنده‌ی موفقیت آمیز بودن طراحی و توسعه‌ی این مدل است، باید توجه داشت که نویسنده‌گان این مقاله، عملیات پیش‌پردازش و استخراج ویژگی را پیش از تقسیم نمونه‌ها به مجموعه‌های تعلیم، ارزیابی و آزمایش انجام داده‌اند و لذا، قابل تصور است که نتایج گزارش شده تحت تاثیر نشست داده‌ها و تخمین بیش از حد^۵ قرار داشته باشد.

در پژوهش [۷۵] نیز یک مدل رگرسیون لجستیک با استفاده از اطلاعات ۷۲۵ بیمار تعلیم داده شده که احتمال انتقال به ICU را با حساسیت ۰/۸۶ و تشخیص ۰/۷۱ پیش‌بینی می‌کند. مساحت زیر منحنی عملکرد این مدل برابر با ۰/۸۷ گزارش شده است. مشابه با رویکرد اتخاذ شده در [۷۳]، مدل گزارش شده بر اساس چندین دسته ویژگی، از جمله نشانگرهای خونی-شیمیایی، علائم، و تصاویر رادیولوژی تعلیم یافته است.

۲-۳-۳-۲- ارزیابی عملکرد مدل‌های یادگیرنده

همان‌گونه که در بخش قبل نیز اشاره شده است، در اغلب مطالعات این حوزه، به منظور ارزیابی عملکرد مدل، شاخص‌های دقت (Acc)، حساسیت (Se)، تشخیص (P)، امتیاز اف ۱ و مساحت زیر منحنی گزارش می‌شوند. از آنجا که هدف اصلی این پژوهش‌ها، توسعه و ارزیابی مدل‌های طبقه‌بندی‌کننده است، با داشتن مقادیر مثبت درست (TP)، منفی درست (TN)، مثبت کاذب (FP) و منفی کاذب (FN) این شاخص‌ها به صورت روابط

۲-۱-۲-۴ محاسبه می‌شوند:

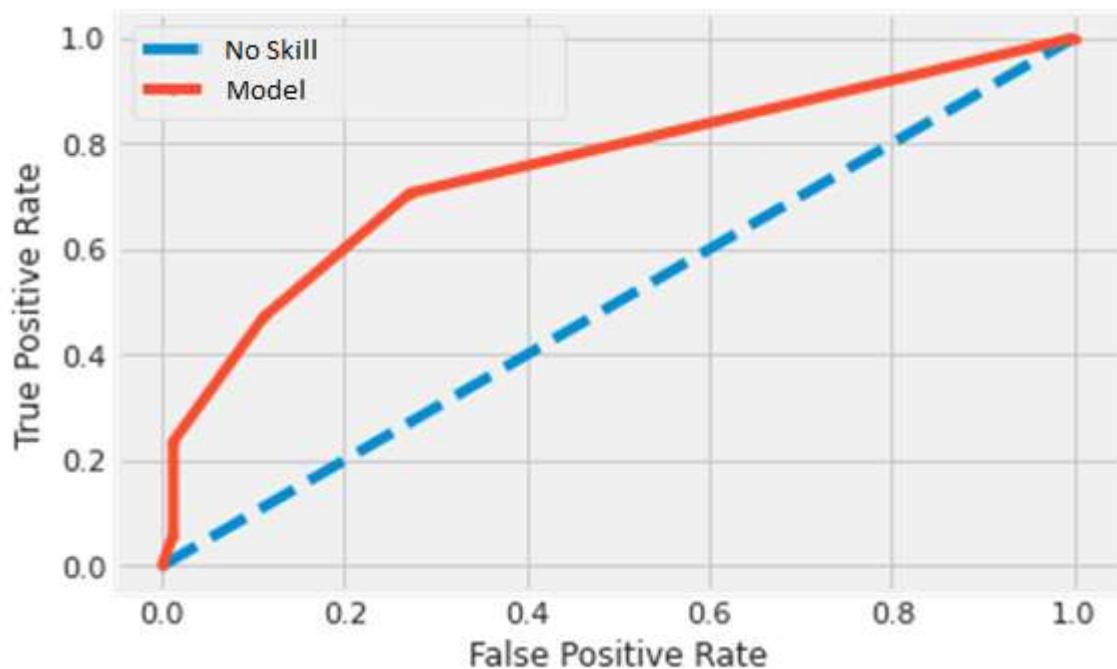
$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (1-2)$$

$$Se = \frac{TP}{TP + FN} \quad (2-2)$$

$$P = \frac{TP}{TP + FP} \quad (3-2)$$

$$F1-score = \frac{2TP}{2TP + FP + FN} \quad (4-2)$$

با توجه به اینکه در این پژوهه با یک مساله‌ی طبقه‌بندی دو کلاسه و با مجموعه دادگان نا متوازن رو به رو هستیم، منحنی‌های ROC (شکل ۲-۱) نیز جهت ارزیابی عملکرد مدل رسم می‌شوند. در حالت ایده‌آل، تلاش بر این است که نرخ نمونه‌هایی به درستی مثبت تشخیص داده شده (TP) از نرخ نمونه‌هایی که به اشتباه مثبت (FP) تشخیص داده شده‌اند بیشتر شوند و منحنی به گوشه بالا و سمت چپ متمایل گردد. مساحت زیر این منحنی یکی دیگر از شاخص‌هایی است که در تحقیق حاضر نیز مورد گزارش شده و افزایش آن در تنظیم ابزارها مورد توجه بوده است.



شکل ۲-۱: منحنی ROC برای ارزیابی کلی مدل

فصل سوم: روش پیشنهادی، اصول و الگوریتم‌ها

۱-۳ - مقدمه

در این فصل روش پیشنهادی این تحقیق و همچنین روش‌هایی که به عنوان روش‌های استاندارد برای مقایسه عملکرد مدل‌های پیشنهادی پیاده‌سازی شده‌اند، شرح داده می‌شوند. پیش از نمایش و توضیح فلوچارت نهایی روش پیشنهادی، مجموعه مفاهیم و الگوریتم‌های که در ساخت این روش‌ها به کار برده شده‌اند معرفی می‌شوند. این فصل به چهار قسمت اصلی تقسیم شده است. در ابتدا روش‌هایی که در فرآیند پاکسازی، آماده‌سازی و پرکردن جاهای خالی در دادگان به کار رفته شده‌اند توضیح داده می‌شوند. در قسمت دوم الگوریتم‌های تصمیم‌گیرنده از جمله الگوریتم‌های ترکیبی داده‌کاوی معرفی شده‌اند. در قسمت سوم برخی ابزارهای کمکی در این الگوریتم‌ها آورده شده است و در قسمت آخر فلوچارت کلی روش پیشنهادی ارائه می‌شود.

۲-۳ - روش‌هایی برای پیش‌پردازش دادگان

۳-۱-۱- روش K-نزدیک‌ترین همسایگی (K-nn)^{۵۷}

روش K-نزدیک‌ترین همسایگی، یک تکنیک یادگیری ماشینی است که به صورت گستردۀ برای کارهای طبقه‌بندی و رگرسیون استفاده می‌شود. این روش زیرمجموعه‌ای از روش‌های یادگیری مبتنی بر نمونه‌ها است،

^{۵۷} K-nearest neighbors

به این معنی که مدل از داده‌های آموزشی ساخته می‌شود و سپس بر اساس شباهت بین نقطه-داده جدید و مثال‌های موجود در نقطه-داده‌های آموزشی، پیش‌بینی صورت می‌پذیرد. الگوریتم K-نر迪کترین همسایگی براساس شناسایی k عدد نرديکترین نقطه-داده در مجموعه آموزشی به نقطه-داده جدید (مورد پرس و جو) کار می‌کند. «نرديک» بودن، در این فضای معمولاً با استفاده از یک معیار فاصله، مانند فاصله اقلیدسی^{۵۸} یا فاصله منهتن^{۵۹} اندازه‌گیری می‌شود. هنگامی که نرديکترین همسایگان شناسایی شدند، الگوریتم، کلاس یا مقدار نقطه-داده جدید را بر اساس کلاس اکثربین در بین این نرديکترین همسایگان و یا مقدار ميانگين نرديکترین همسایگان پیش‌بینی می‌کند. یکی از نقاط قوت الگوریتم K-nn، سادگی و انعطاف‌پذیری آن است. می‌توان از آن برای کارهای طبقه‌بندی و رگرسیون استفاده کرد. این روش همچنین می‌تواند داده‌ها را با توزیع‌های دلخواه و مرزهای تصمیم متنوع مدیریت کند. علاوه بر این، الگوریتم هیچ پیش‌فرضی در مورد توزیع داده‌ها ندارد، که می‌تواند در برخی موارد مفید باشد. با این حال، الگوریتم K-nn می‌تواند به انتخاب k و معیار فاصله استفاده شده حساس باشد. مقدار کوچک k ممکن است منجر به بیش‌برازش^{۶۰} شود، در حالی که مقدار زیاد k ممکن است منجر به کم‌برازش^{۶۱} گردد. انتخاب معیار فاصله نیز به ویژگی‌های داده‌ها بستگی دارد و با توجه به مساله و نوع دادگان، برخی از معیارها ممکن است مناسب‌تر از سایرین باشند. یکی دیگر از محدودیت‌های الگوریتم K-nn مقیاس‌پذیری آن است. با افزایش اندازه مجموعه داده، زمان مورد نیاز برای محاسبه نرديکترین همسایه‌ها نیز افزایش می‌یابد، که می‌تواند الگوریتم را برای مجموعه داده‌های بزرگ غیرعملی کند. با وجود این محدودیت‌ها، الگوریتم K-nn همچنان یک تکنیک یادگیری ماشینی محبوب و پرکاربرد است، به ویژه در کاربردهایی که تفسیرپذیری و سادگی مهم هستند.

در تحقیق حاضر از روش K-نرديکترین همسایگی برای پیش‌بینی و پر کردن فضاهای خالی در مجموعه دادگان استفاده شده است و جزئیات بیشتر درباره نحوه اجرا و نتیجه آن در فصل ۴ توضیح داده شده است.

۲-۲-۳- ضریب همبستگی پیرسون (PCC)^{۶۲}

همبستگی پیرسون معیاری از همبستگی خطی بین دو متغیر است که معمولاً در آمار و تجزیه و تحلیل داده‌ها استفاده می‌شود. در زمینه ماتریس‌ها، همبستگی پیرسون اغلب برای اندازه‌گیری رابطه بین ستون‌ها یا ردیف‌های یک ماتریس استفاده می‌شود.

^{۵۸} Euclidean distance

^{۵۹} Manhattan distance

^{۶۰} Overfitting

^{۶۱} Underfitting

^{۶۲} Pearson correlation coefficient

برای محاسبه همبستگی پیرسون بین دو سطر از یک ماتریس، ابتدا میانگین و انحراف معیار هر ستون یا سطر را محاسبه می‌کنیم. سپس کوواریانس بین دو سطر را محاسبه می‌کنیم که میانگین حاصل ضرب انحرافات هر مقدار از میانگین آن است. در نهایت کوواریانس را بر حاصل ضرب انحرافات استاندارد دو ستون یا سطر تقسیم می‌کنیم تا ضریب همبستگی پیرسون،^{۶۳} به دست آید (رابطه ۱-۳).

$$r = \frac{\text{cov}(X, Y)}{\text{std}(X) \times \text{std}(Y)}$$

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n - 1)} \quad (1-3)$$

کوواریانس بین دو سطر یا ستون: $\text{cov}(X, Y)$

انحراف معیار استاندارد سطر یا ستون Y و X : $\text{std}(Y)$ و $\text{std}(X)$

ضریب همبستگی پیرسون از ۱ تا -۱ متغیر است، که -۱ نشان دهنده همبستگی منفی کامل (یعنی وقتی یک متغیر افزایش می‌یابد، متغیر دیگر کاهش می‌یابد)، ۰ نشان دهنده عدم همبستگی، و ۱ نشان دهنده همبستگی مثبت کامل است (یعنی زمانی که یک متغیر افزایش می‌یابد، متغیر دیگر نیز بیشتر می‌شود). بزرگی ضریب همبستگی نشان دهنده قدرت رابطه بین دو متغیر است.

۲-۳-۳- معیار کولبک-لیبلر (KL)^{۶۴}

معیار کولبک-لیبلر که با عنوان آنتروپی نسبی نیز شناخته می‌شود، روشی برای اندازه‌گیری تفاوت بین دو توزیع احتمال است و بین دو توزیع احتمال $P(x)$ و $Q(x)$ به صورت رابطه (۲-۳) تعریف می‌شود [۷۶]:

$$KL(P||Q) = \int P(x) \log \left(\frac{P(x)}{Q(x)} \right) dx \quad (2-3)$$

که در آن انتگرال از کل فضای نمونه x گرفته می‌شود. معیار کولبک-لیبلر مقدار اطلاعات از دست رفته را هنگامی که از توزیع احتمال $(x) Q$ به عنوان تقریبی ازتابع احتمال $(x) P$ استفاده می‌شود، اندازه‌گیری می‌کند.

این یک کمیت غیر منفی است و تقریباً در همه جا برابر با صفر است اگر و فقط اگر $P(x) = Q(x)$. معیار کولبک-لیبلر کاربردهای مهم بسیاری در یادگیری ماشین دارد؛ انتخاب مدل، انتخاب ویژگی و بهینه‌سازی مدل‌های تولیدی، مثال‌هایی از این کاربردها هستند. به علاوه، این معیار، می‌تواند به عنوان یک تابع ضرر برای آموزش مدل‌های تولیدی، مانند خودرمزنگذار متغیر (VAE)^{۶۴} و شبکه‌های مولد رقابتی (GAN)^{۶۵} استفاده

^{۶۳} Kullback-Leibler

^{۶۴} Variational autoencoder

^{۶۵} Generative adversarial network

شود. همچنین، معیار کولبک-لیبلر یک مفهوم کلیدی در استنتاج بیزی^{۶۶} است که در آن برای اندازه‌گیری تفاوت بین توزیع‌های قبلی و پسین استفاده می‌شود.

۴-۲-۳-روش کاهش ویژگی PCA

برای درک این موضوع که چرا از اساس نیاز به کاهش ابعاد ویژگی‌ها وجود دارد باید به این نکته توجه کرد که با افزایش خطی تعداد ویژگی‌ها یا ابعاد، مقدار عملیات محاسباتی که برای تعمیم دقیق نیاز است، به طور نمایی افزایش می‌یابد. در این شرایط، کاهش ابعاد داده‌ها به چند دلیل می‌تواند مفید باشد:

- ساده‌سازی داده‌ها: تجزیه و تحلیل و تفسیر داده‌های با ابعاد بالا ممکن است دشوار باشد. با کاهش تعداد ابعاد، کار با داده‌ها و تجسم آنها آسان‌تر می‌شود.
- بهبود بهره‌وری محاسباتی: پردازش و تجزیه و تحلیل داده‌های با ابعاد بالا می‌تواند از نظر محاسباتی پر هزینه باشد. بنابراین، کاهش ابعاد داده‌ها می‌تواند الزامات محاسباتی را به میزان قابل توجهی کاهش دهد و کارایی تجزیه و تحلیل را بهبود بخشد.
- دقت بهتر: داده‌های با ابعاد بالا می‌تواند منجر به بیش‌برازش و کاهش دقت مدل‌های یادگیری ماشین شود. حالتی که در دنیای یادگیری به «نفرین ابعاد»^{۶۷} معروف است. کاهش ابعاد داده‌ها می‌تواند با حذف نویز و ویژگی‌های تکراری به کاهش این مشکل کمک کند.
- بهبود تعمیم مدل^{۶۸}: با کاهش ابعاد داده‌ها، ممکن است بتوان الگوها و روابط زیربنایی را شناسایی کرد که به راحتی در داده‌هایی با ابعاد بالا آشکار نمی‌شوند. این موضوع می‌تواند به تعمیم بهتر و بهبود عملکرد پیش‌بینی در مدل‌های یادگیری ماشین منجر شود.

به طور خلاصه، کاهش ابعاد داده‌ها می‌تواند تجزیه و تحلیل و تفسیر را آسان‌تر کند، کارایی محاسباتی را بهبود بخشد، دقت و عملکرد تعمیم مدل‌های یادگیری ماشین را افزایش دهد و داده‌ها را به طور کلی قابل مدیریت تر کند.

به طور عمومی دو روش برای کاهش ابعاد وجود دارد:

- حذف مستقیم برخی از ویژگی‌های کم اهمیت
- استخراج ویژگی، به معنی کشف و حفظ ویژگی‌های مهم و حذف بقیه ویژگی‌ها

تکنیک PCA یکی از روش‌های کاهش ابعاد مبتنی بر استخراج ویژگی است [۷۶]. در این روش ابتدا روابط مهم‌تر بین داده‌ها شناسایی شده و بر اساس این روابط، یکتابع تبدیل بر داده‌ها اعمال می‌شود. در مرحله

^{۶۶} Bayesian inference

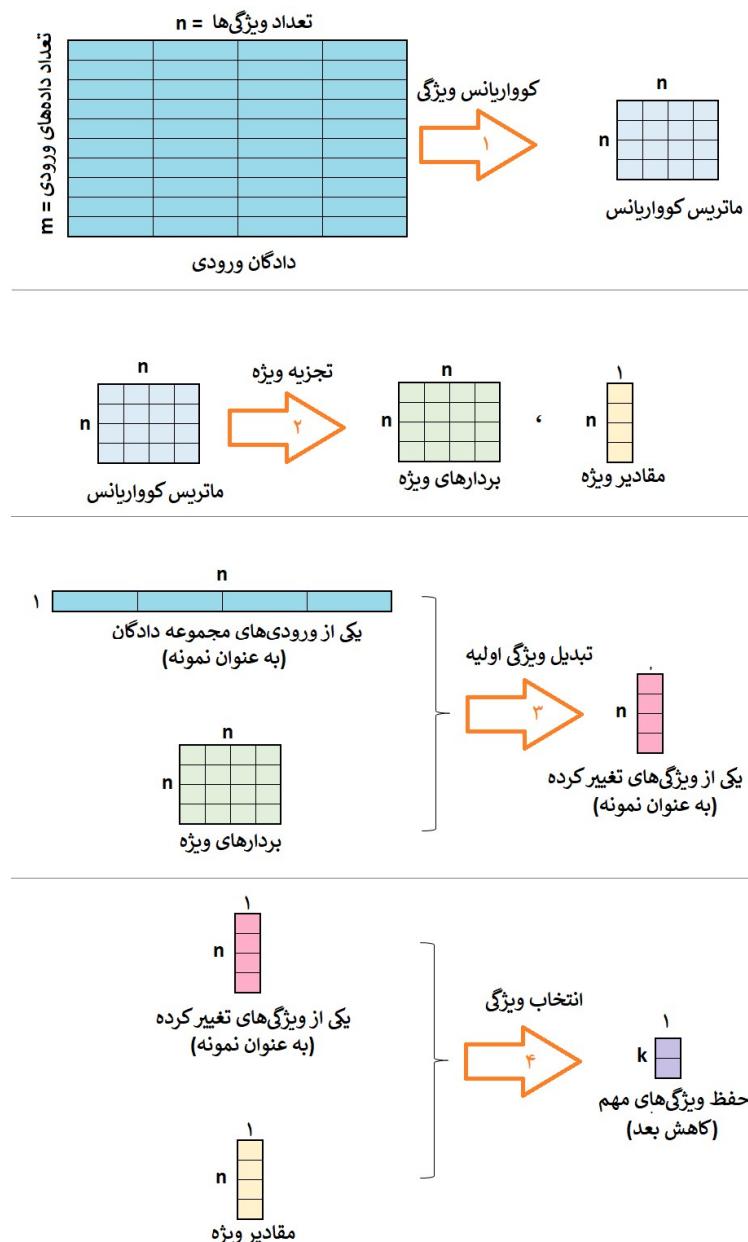
^{۶۷} Principal component analysis

^{۶۸} Curse of dimensionality

^{۶۹} Improved generalization

بعدی اهمیت این روابط به صورت کمی ارزش‌گذاری شده و به این ترتیب در مرحله‌ی آخر امکان حفظ ویژگی‌هایی با مهتمرین روابط و حذف سایر ویژگی‌ها وجود خواهد داشت. این روش از چهار مرحله‌ی اصلی تشکیل شده است (شکل ۱-۳): کوواریانس ویژگی، تجزیه ویژه ماتریسی، تبدیل ویژگی‌های اولیه و انتخاب ویژگی بر اساس واریانس [۷۶].

- کوواریانس ویژگی: تشخیص روابط بین ویژگی‌ها از طریق تشکیل ماتریس کوواریانس.
- تجزیه ویژه ماتریسی: اعمال یک تبدیل خطی یا تجزیه ماتریسی برای به دست آوردن بردارها و مقادیر ویژه.
- تبدیل ویژگی اولیه: مجموعه داده‌ها با استفاده از بردارهای ویژه تولید شده در مرحله‌ی قبل به یک سری ویژگی‌های ثانویه تبدیل می‌شوند. در واقع، در این مرحله جدول ویژگی‌های جدیدی تولید می‌شود که هر کدام ترکیبی خطی از ویژگی‌های اولیه هستند.
- انتخاب ویژگی‌ها بر اساس واریانس: اهمیت مؤلفه‌های اصلی با استفاده از مقادیر ویژه تعیین می‌گردد و در پی آن، تنها مؤلفه‌های مهم حفظ می‌شوند.



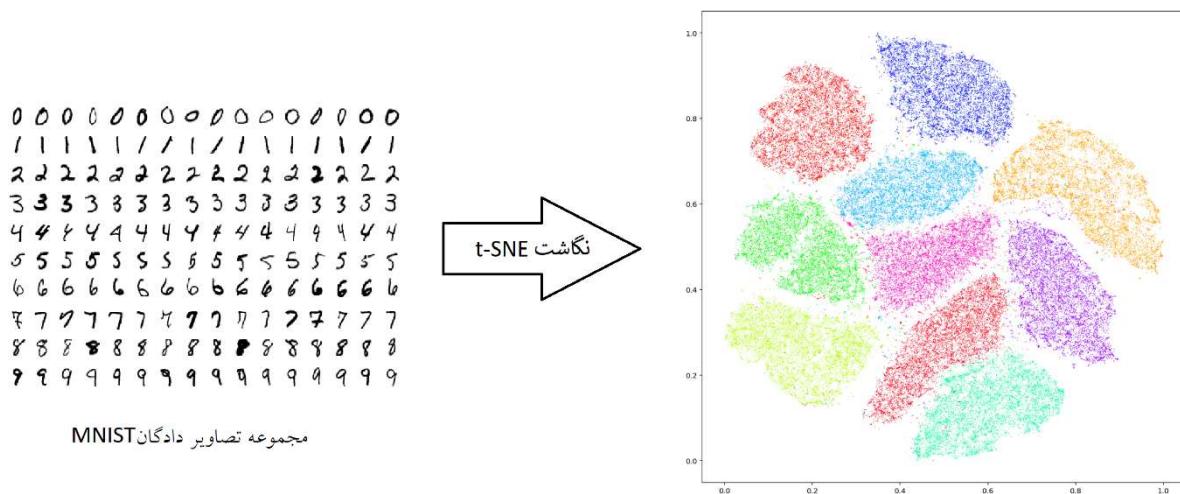
شکل ۳: مراحل اصلی روش کاهش بعد PCA

۷. t-SNE ویژگی کاهش

روش t-SNE، یک تکنیک دیگر برای کاهش ابعاد است که معمولاً برای تصویر کردن داده های با ابعاد بالا در فضایی با ابعاد پایین تر (معمولًاً دو بعدی یا سه بعدی) استفاده می شود. t-SNE به ویژه برای کاوش و تفسیر مجموعه داده های پیچیده، مانند تصاویر یا داده های ژنومی مفید است که در آنها روابط بین نقاط داده ممکن

۷. t-Distributed Stochastic Neighbor Embedding

است غیر خطی بوده و در فضایی با ابعاد بالا به سختی قابل تجسم باشند [78]. شکل ۲-۳ مثالی از استفاده از الگوریتم t-SNE برای کلاس بنده تصاویر مربوط به ارقام انگلیسی را نشان می‌دهد.



شکل ۲-۳: مثالی از روش کاهش ویژگی t-SNE، بانک اطلاعاتی **MNIST** که مجموعه‌ای هفتادهزارتاپی از ارقام دستنویس است با استفاده از الگوریتم t-SNE به ۱۰ کلاس (یک کلاس به ازای هر رقم) نگاشت شده است. هر تصویر شامل $28 \times 28 = 784$ پیکسل (ویژگی) است، با استفاده از الگوریتم t-SNE تعداد ویژگی‌ها به ۲ عدد (x, y) کاهش یافته و تفسیر مجموعه بسیار آسان شده است.

ایده‌ی اصلی پشت t-SNE ایجاد یک توزیع احتمال است که شباهت‌های بین نقاط داده با ابعاد بالا به تصویر می‌کشد و سپس یک نگاشت به توزیعی با ابعاد پایین‌تر ایجاد می‌کند که این شباهت‌ها را تا حد امکان حفظ کند. این الگوریتم با بهینه‌سازی مکرر یک تابع هزینه، کار می‌کند و سعی می‌کند که تفاوت بین توزیع‌های احتمالی با ابعاد بالا و کم بعد را به حداقل رساند.

این الگوریتم ابتدا توزیع احتمال را بر روی جفت نقاط داده با ابعاد بالا بر اساس شباهت آنها محاسبه می‌کند، که در آن شباهت به عنوان احتمال شرطی تعریف می‌شود که یک نقطه، نقطه دیگری را به عنوان همسایه خود با توجه به پارامتر مقیاس معین انتخاب کند. سپس با استفاده از پارامتر مقیاس متفاوت، توزیع احتمال مشابه را بر روی جفت نقاط داده با ابعاد پایین ایجاد می‌کند. سپس الگوریتم توزیع احتمال کم بعدی را بهینه می‌کند تا اختلاف بین دو توزیع را به حداقل برساند.

یکی از مزایای t-SNE نسبت به سایر تکنیک‌های کاهش ابعاد، مانند PCA، این است که می‌تواند روابط پیچیده و غیرخطی بین نقاط داده را حفظ کند. با این حال، t-SNE همچنین از نظر محاسباتی گران است و در ضمن می‌تواند نسبت به انتخاب ابرپارامترها حساس باشد، مانند پارامتر پیچیدگی که مقیاس توزیع احتمال را کنترل می‌کند.

به طور خلاصه، t-SNE یک تکنیک کاهش ابعاد قدرتمند است که می‌تواند برای تجسم داده‌های با ابعاد بالا در فضایی با ابعاد پایین‌تر استفاده شود، در حالی که شباهت‌های بین نقاط داده را تا حد امکان حفظ می‌کند.

۶-۲-۳- حذف ویژگی بازگشتی^{۷۹} (RFE)

حذف ویژگی بازگشتی (RFE) یکی دیگر از تکنیک انتخاب ویژگی است که در یادگیری ماشین برای شناسایی مهم‌ترین ویژگی‌ها در یک مجموعه داده استفاده می‌شود [۷۹]. این روش یکی از روش‌های مبتنی بر تکرار است که زیرمجموعه‌ای از مهم‌ترین ویژگی‌ها را براساس عملکرد آنها در یک الگوریتم یادگیری ماشین دلخواه انتخاب می‌کند. این الگوریتم کم‌اهمیت‌ترین ویژگی‌ها را در هر تکرار حذف می‌کند و این فرآیند تا رسیدن به تعداد مشخص و از پیش تعیین شده‌ای از ویژگی‌ها تکرار می‌شود.

مراحل مربوط به RFE عبارتند از:

- ۱) انتخاب یک الگوریتم یادگیری ماشینی: اولین قدم انتخاب یک الگوریتم یادگیری ماشینی مناسب برای مسئله مورد نظر است. برخی از الگوریتم‌های معروف که به خوبی با RFE کار می‌کنند عبارتند از رگرسیون خطی، رگرسیون لجستیک، درخت‌های تصمیم‌گیری و ماشین‌های بردار پشتیبانی.
- ۲) مقداردهی اولیه پارامترها: تعیین پارامترهای RFE مثلاً تعیین ابعاد زیرمجموعه‌های ویژگی‌ها یا تعیین این که چه تعداد ویژگی در هر تکرار حذف شوند.
- ۳) آموزش الگوریتم: الگوریتم یادگیری ماشین بر روی مجموعه اولیه و کامل ویژگی‌ها آموزش داده می‌شود و عملکرد آن با استفاده از یک معیار مناسب مانند دقت، مساحت زیر منحنی، یا امتیاز اف ۱ ارزیابی می‌شود.
- ۴) رتبه‌بندی ویژگی‌ها: ویژگی‌ها با توجه به امتیاز اهمیت آنها رتبه بندی می‌شوند. امتیاز اهمیت را می‌توان با استفاده از روش‌های مختلفی مانند ضرایب در رگرسیون خطی، اهمیت ویژگی‌ها در درخت‌های تصمیم یا وزن‌ها در ماشین‌های بردار پشتیبان محاسبه کرد.
- ۵) حذف کم‌اهمیت‌ترین ویژگی: ویژگی یا ویژگی‌های با کمترین امتیاز اهمیت در این مرحله حذف می‌شوند و سپس تعلیم الگوریتم یادگیری ماشین این بار با مجموعه ویژگی کاهش یافته انجام می‌شود.
- ۶) ارزیابی عملکرد: عملکرد الگوریتم را در مجموعه ویژگی‌های کاهش یافته با استفاده از همان متريک مورد استفاده در مرحله ۳ ارزیابی می‌شود.
- ۷) تکرار مراحل چهار تا شش: مراحل چهارم تا ششم آن قدر تکرار می‌شود تا به تعداد مشخصی از ویژگی‌های از پیش تعریف شده برسد، یا تا زمانی که عملکرد الگوریتم شروع به تنزل کند.

(۸) مجموعه ویژگی های بهینه: مجموعه ویژگی های باقیمانده، زیرمجموعه ای از همه ویژگی ها است که بهترین عملکرد را برای مشکل داده شده ارائه می دهد.

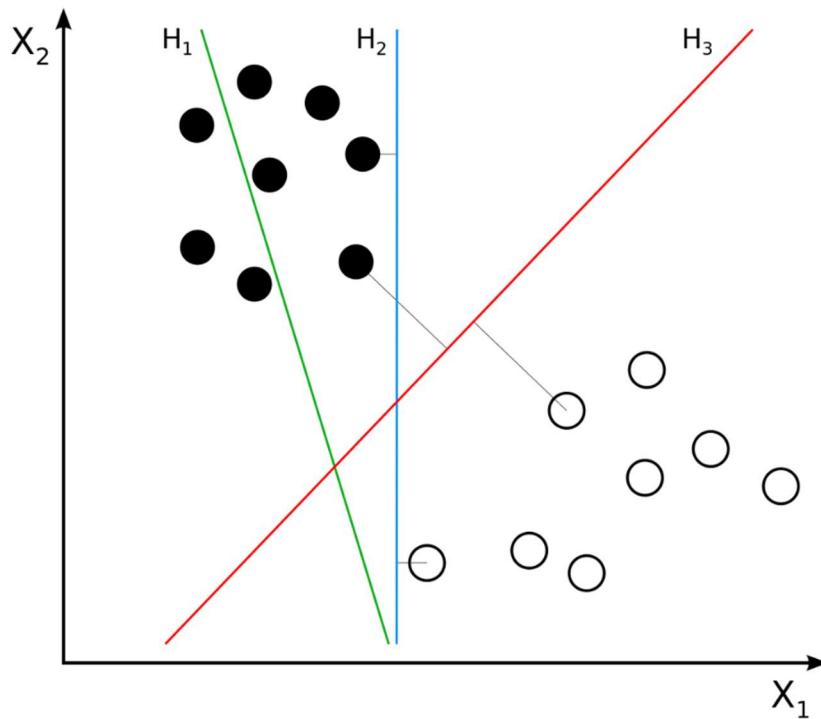
روش RFE مزایای متعددی نسبت به سایر تکنیک های انتخاب ویژگی دارد. این روش، ساده و مؤثر بوده و نیازی به دانش قبلی در مورد مجموعه ای داده یا ویژگی ها ندارد. همچنین ابعاد مجموعه ای دادگان را کاهش می دهد، که می تواند به بهبود عملکرد الگوریتم یادگیری ماشین و کاهش بیش برازش کمک کند. با این حال، RFE می تواند از نظر محاسباتی گران باشد، به خصوص برای مجموعه داده های بزرگ با ویژگی های زیاد.

۳-۳-۳- مدل های یادگیرنده

۱-۳-۳- ماشین بردار پشتیبانی (SVM)

ماشین بردار پشتیبانی یک الگوریتم یادگیری ماشینی نظارت شده است که برای طبقه بندی و تحلیل رگرسیون استفاده می شود. اساس این روش یافتن ابر صفحه های بهینه است که داده ها را در دو یا چند کلاس دسته بندی می کند.

در طبقه بندی ویژگی هایی با دو بعد، ابر صفحه به یک خط تبدیل می شود که داده ها را با به حداقل رساندن حاشیه - که فاصله بین ابر صفحه و نزدیک ترین نقاط داده از هر کلاس است - به دو کلاس جدا می کند. نقاط داده ای که نزدیک ترین به ابر صفحه هستند، بردارهای پشتیبان نامیده می شوند، از این رو به آن ماشین بردار پشتیبان می گویند. شکل ۲-۳ یک ماشین بردار پشتیبان دودویی را به تصویر می کشد.



شکل ۳-۳: یک ماشین بردار پشتیبان برای طبقه‌بندی ویژگی‌هایی با دو بعد. در این حالت ابرصفحه به یک خط (H_3) تبدیل شده است که داده‌ها را با به حداقل رساندن حاشیه در این مثال به دو کلاس تقسیم می‌کند. H_1 کلاس‌ها را از هم جدا نمی‌کند. H_2 کلاس‌بندی را انجام می‌دهد، اماً با یک حاشیه کوچک. H_3 کلاس‌ها را با حداقل حاشیه جدا می‌کند.

ماشین بردار پشتیبان همچنین می‌تواند طبقه‌بندی چند کلاسه را با استفاده از رویکرد یک در مقابل همه یا یک در مقابل یک انجام دهد، که در آن طبقه‌بندی‌کننده‌های دودویی متعددی برای طبقه‌بندی هر کلاس در برابر بقیه آموزش داده می‌شوند. این الگوریتم همچنین می‌تواند برای تجزیه و تحلیل رگرسیون با یافتن ابرصفحه‌ای که به بهترین شکل با نقاط داده مطابقت دارد، استفاده شود، جایی که فاصله بین ابرصفحه و نقاط داده به حداقل می‌رسد. ماشین بردار پشتیبان از یکتابع هسته برای تبدیل داده‌ها به فضایی با ابعاد بالاتر استفاده می‌کند، جایی که می‌توان داده‌ها را راحت‌تر از هم جدا کرد. برخی از توابع هسته که معمولاً مورد استفاده قرار می‌گیرند، تابع پایه خطی، و یا توابع چند جمله‌ای یا شعاعی (RBF) هستند.

ماشین بردار پشتیبان یک الگوریتم قدرتمند و همه‌کاره است که می‌تواند مشکلات طبقه‌بندی و رگرسیون خطی و غیرخطی را مدیریت کند. با این حال، ممکن است از نظر محاسباتی گران و حساس به انتخاب ابرپارامترها، مانند تابع هسته، پارامتر تنظیم و ضریب هسته باشد. بنابراین، تنظیم دقیق این پارامترها برای دستیابی به عملکرد مطلوب ضروری است.

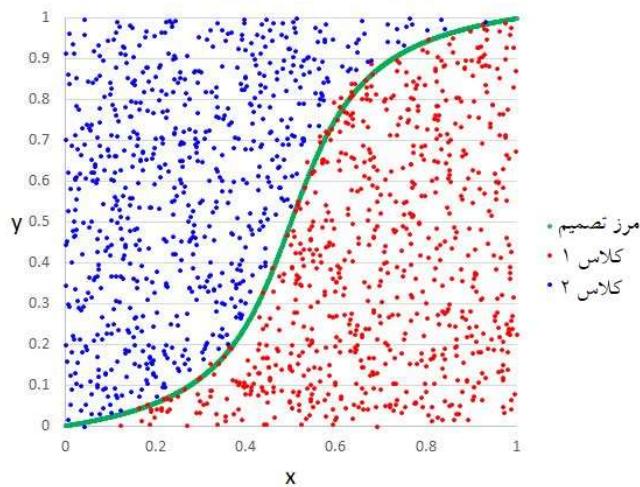
۳-۲- رگرسیون لجستیک

رگرسیون لجستیک الگوریتم آماری و یادگیری دیگری است که برای طبقه بندی دودویی و چند کلاسه استفاده می شود. این روش، احتمال وقوع یک رویداد یا نتیجه را با برآش یک تابع لجستیک به داده ها پیش‌بینی می کند. لازم به یادآوری است که تابع لجستیک دودویی که در ریاضیات با عنوان سیگموئید نیز شناخته می شود، یک تابع ریاضی است که هر مقدار ورودی را به مقدار خروجی بین 0° و 1° نگاشت می کند و با رابطه $3-3$ توصیف می شود.

$$f(x) = \frac{1}{1 + e^{-\beta x}} \quad (3-3)$$

مدل رگرسیون لجستیک فرض می کند که رابطه بین متغیرهای مستقل (ویژگی ها) و متغیر وابسته (هدف) را می توان به عنوان یک تابع لجستیک مدل سازی کرد. در این حالت، تابع لجستیک رابطه هر ورودی را با خروجی بین 0° و 1° نگاشت می کند، که نشان دهنده احتمال تعلق هدف به یک کلاس خاص است.

شکل ۳-۴ یک مثال ساده از یک رگرسیون لجستیک دودویی را نمایش می دهد.



شکل ۳-۴: رگرسیون لجستیک یک الگوریتم یادگیری ماشینی است که برای مسائل طبقه بندی استفاده می شود. این الگوریتم مقادیر احتمالی را با کمک یک تابع سیگموئید پیش بینی می کند که به دو کلاس (طبقه بندی دودویی) یا بیشتر (طبقه بندی چند کلاسه) نگاشت می شوند.

در صورت طبقه بندی چند کلاسه، به جای تابع سیگموئید از تابع بیشینه هموار^{۷۲} استفاده می شود.

در رگرسیون لجستیک، پارامترهای مدل با استفاده از روش برآورد حداکثر درستنمایی^{۷۳} (MLE) برآورده می‌شوند. هدف این مدل به حداکثر رساندن احتمال داده‌های مشاهده شده با توجه به پارامترهای مدل برآورده شده است. فرآیند بهینه سازی شامل به حداقل رساندن یکتابع هزینه است که تفاوت بین احتمالات پیش‌بینی شده و مقادیر هدف واقعی را اندازه گیری می‌کند.

رگرسیون لجستیک می‌تواند روابط خطی و غیرخطی بین ویژگی‌ها و هدف را با استفاده از اصطلاحات چند جمله‌ای یا تعاملی یا با استفاده از تبدیل‌های غیرخطی ویژگی‌ها مدیریت کند.

اگرچه رگرسیون لجستیک یک الگوریتم ساده و قابل تفسیر است، زمانی که کلاس‌ها نامتعادل هستند یا زمانی که داده‌ها دارای روابط غیرخطی هستند که توسط تابع لجستیک دیده نمی‌شوند، ممکن است عملکرد خوبی نداشته باشد. علاوه بر این، رگرسیون لجستیک فرض می‌کند که رابطه بین ویژگی‌ها و هدف خطی است، که ممکن است در عمل همیشه این‌طور نباشد.

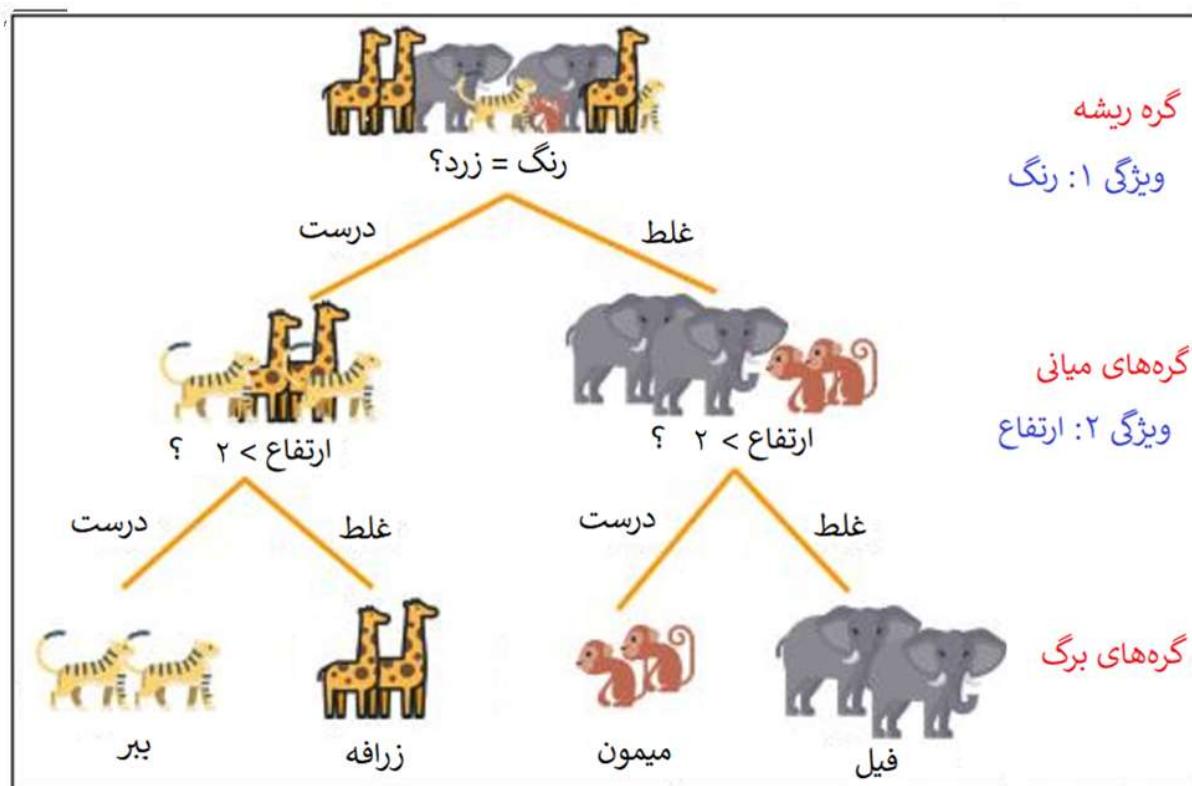
۳-۳-۳- درخت تصمیم، جنگل تصادفی^{۷۴}

درختان تصمیم و جنگل‌های تصادفی الگوریتم‌های یادگیری ماشینی هستند که به طور گسترده در زمینه‌های مختلف مورد استفاده قرار می‌گیرند. در این الگوریتم‌ها تصادفی‌سازی نقش مهمی در بهبود عملکرد نهایی ایفا می‌کند.

درخت تصمیم یک ساختار سلسله مراتبی است که از گره‌ها و شاخه‌ها تشکیل شده است. گره‌ها تصمیمات یا اقدامات را نشان می‌دهند و شاخه‌ها آن‌ها را به ترتیبی به هم متصل می‌کنند که منجر به یک نتیجه نهایی می‌شود. در یادگیری ماشینی، از درخت‌های تصمیم برای پیش‌بینی با یادگیری از مجموعه‌ای از داده‌های آموزشی استفاده می‌شود. هر گره در درخت یک ویژگی یا ویژگی داده را نشان می‌دهد و شاخه‌ها مقادیر احتمالی آن ویژگی را نشان می‌دهند. نتیجه نهایی درخت توسط گره برگ که پس از دنبال کردن شاخه‌ها از گره ریشه به آن می‌رسد تعیین می‌شود. شکل ۳-۵ مثال بسیار ساده‌ای از یک درخت تصمیم را نمایش می‌دهد که براساس دو ویژگی رنگ و ارتفاع نوع یک حیوان را از بین نمونه‌های موجود تعیین می‌کند.

^{۷۳} Maximum likelihood estimation

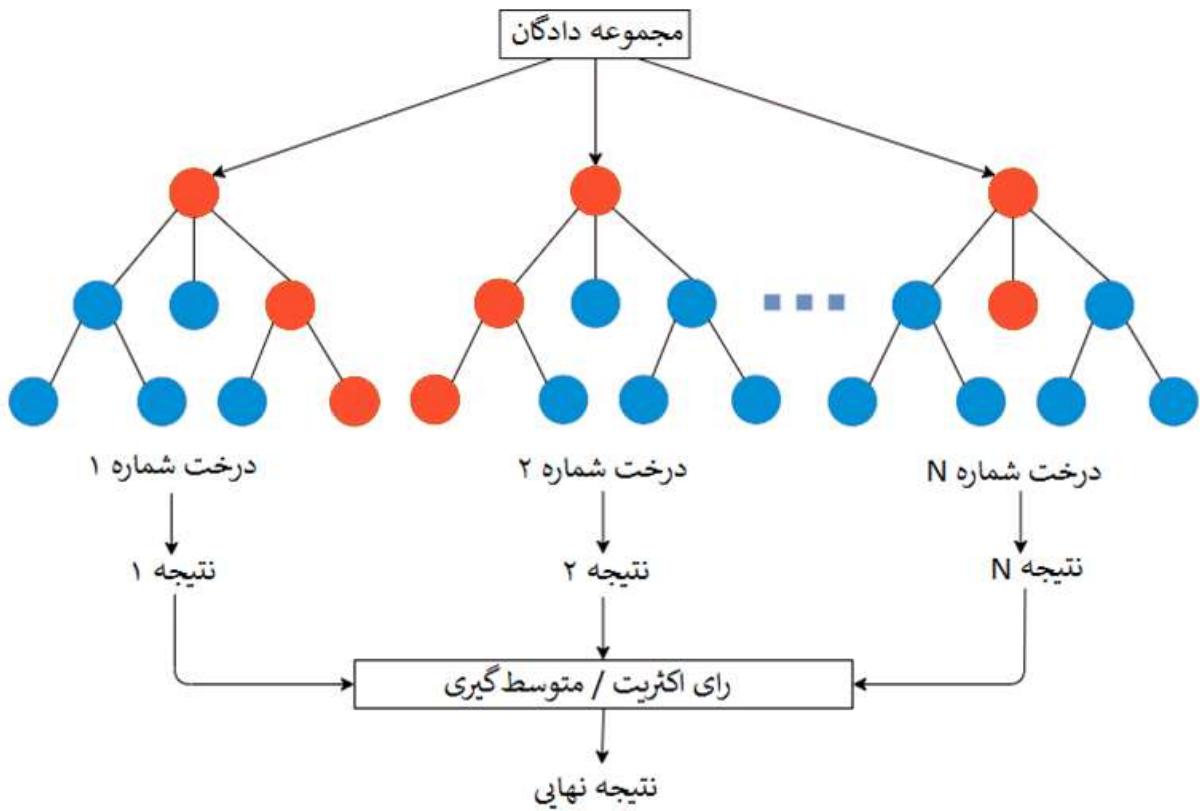
^{۷۴} Random Tree, Random Forest



شکل ۳-۵: مثال بسیار ساده‌ای از یک درخت تصمیم که براساس دو ویژگی رنگ و ارتفاع، نوع یک حیوان را از بین نمونه‌های موجود تعیین می‌کند.

یک درخت تصمیم را می‌توان با استفاده از یک زیرمجموعه تصادفی از ویژگی‌های موجود در داده‌ها ساخت. این تصادفی‌سازی خطر بیش‌برازش را کاهش می‌دهد، بیش‌برازش زمانی رخ می‌دهد که یک مدل بیش از حد پیچیده است و نویز یا نوسانات تصادفی موجود در داده‌های آموزشی را یاد می‌گیرد. با انتخاب تصادفی زیرمجموعه‌ای از ویژگی‌ها، یک درخت تصادفی مجبور می‌شود روی مرتبط‌ترین ویژگی‌ها تمرکز کند و از بیش‌برازش اجتناب نماید.

جنگل تصادفی مجموعه‌ای تصادفی از درختان تصمیم است. به جای ساختن یک درخت تصمیم واحد، یک جنگل تصادفی، چندین درخت را با استفاده از زیرمجموعه‌های تصادفی مختلف از ویژگی‌ها و داده‌های آموزشی ایجاد می‌کند (شکل ۳-۶). پیش‌بینی نهایی جنگل با ترکیب پیش‌بینی‌های همه درختان مشخص می‌شود [۸۰]. این رویکرد خطر بیش‌برازش را کاهش می‌دهد و دقت و پایداری مدل را بهبود می‌بخشد.



شکل ۶-۳: جنگل‌های تصادفی از محبوب‌ترین روش‌های ترکیبی داده‌کاوی هستند که از مجموعه‌ای تصادفی از درختان تصمیم ساخته می‌شوند. پیش‌بینی نهایی جنگل با ترکیب پیش‌بینی‌های همه درختان مشخص می‌شود (بگینگ)

الگوریتم جنگل تصادفی چندین مزیت نسبت به سایر الگوریتم‌های یادگیری ماشین دارد. بسیار مقیاس‌پذیر است و می‌تواند مجموعه داده‌های بزرگ با ویژگی‌های ابعادی بالا را مدیریت کند. همچنین معیارهای اهمیت ویژگی (وزن دار کردن ویژگی‌ها) را ارائه می‌دهد که می‌تواند در انتخاب ویژگی و درک روابط اساسی در داده‌ها مفید باشد. به علاوه، می‌تواند داده‌های از دست رفته و ویژگی‌های نویزی را مدیریت کند و در برابر موارد پرت مقاوم است.

جنگل‌های تصادفی از محبوب‌ترین روش‌های ترکیبی داده‌کاوی هستند که با موفقیت برای طیف گسترده‌ای از مشکلات از جمله طبقه‌بندی تصویر، تشخیص اشیا، پردازش زبان طبیعی و تشخیص پزشکی استفاده شده‌اند. آن‌ها همچنین در داده‌کاوی و تجزیه و تحلیل داده‌های اکتشافی برای کشف الگوها و روابط پنهان در مجموعه داده‌های بزرگ استفاده می‌شوند.

به طور خلاصه، درختان تصمیم و جنگل‌های تصادفی الگوریتم‌های یادگیری ماشینی قدرتمندی هستند که می‌توانند دقت و ثبات پیش‌بینی‌ها را بهبود بخشنند. با ترکیب تصادفی‌سازی و یادگیری گروهی، آن‌ها می‌توانند بر محدودیت‌های درختان تصمیم‌گیری سنتی غلبه کنند و رویکردی مقیاس‌پذیر و انعطاف پذیر برای یادگیری ماشین ارائه دهند.

همانند جنگل تصادفی، XGBoost هم یکی از الگوریتم‌های داده‌کاوی ترکیبی است که در یادگیری ناظارت شده استفاده می‌شود. این الگوریتم که خصوصاً برای حل مسائل رگرسیون و طبقه‌بندی استفاده می‌شود. مانند الگوریتم جنگل تصادفی، روشی مبتنی بر درخت تصمیم است. البته این الگوریتم از تقویت گرادیان برای بهبود مکرر عملکرد مدل استفاده می‌کند [۴۹].

در ادامه به برخی از ویژگی‌های کلیدی XGBoost اشاره شده است:

۱. سرعت و مقیاس پذیری: XGBoost به دلیل سرعت و مقیاس‌پذیری شناخته شده است. تا جایی که این الگوریتم می‌تواند به طور موثر مجموعه داده‌های بزرگ را با میلیون‌ها ردیف و هزاران ویژگی مدیریت کند.

۲. تنظیم‌کننده: XGBoost یک پارامتر تنظیم داخلی^{۷۶} برای جلوگیری از بیش‌برازش فراهم می‌کند. وجود این پارامتر در هنگام برخورد با داده‌های با ابعاد بالا حائز اهمیت است.

۳. توابع هدف قابل تنظیم: XGBoost امکان تعریف تابع هدف برای مدل را فراهم می‌کند، که کنترل بیشتری بر رفتار مدل می‌دهد.

۴. مدیریت دادگان از دست رفته: XGBoost می‌تواند مقادیر از دست رفته یا ناموجود در داده‌های ورودی را مدیریت کند و این موضوع این الگوریتم را مناسب تحلیل مجموعه داده‌های دنیای واقعی می‌کند.

۵. هرس درختان: XGBoost شامل مکانیزمی برای هرس درختان در طول فرآیند آموزش برای جلوگیری از بیش‌برازش است.

۶. توقف زودهنگام^{۷۷}: XGBoost می‌تواند عملکرد مدل را بر روی یک مجموعه اعتبارسنجی در طول آموزش ناظارت کند و هنگامی که بهبود مدل متوقف شد به طور خودکار فرآیند آموزش را متوقف کند. بخش ۳-۴-۲ را مشاهده کنید.

۷. پردازش موازی: XGBoost از پردازش موازی پشتیبانی می‌کند که به آن اجازه می‌دهد از چندین GPU یا CPU برای سرعت بخشیدن به روند آموزش استفاده کند.

به طور کلی، XGBoost یک الگوریتم یادگیری ماشینی قدرتمند و انعطاف‌پذیر است که می‌تواند طیف گسترده‌ای از مشکلات رگرسیون و طبقه‌بندی را مدیریت کند. سرعت، مقیاس‌پذیری و مقاومت

^{۷۵} Extreme Gradient Boosting

^{۷۶} Regularizer

^{۷۷} Early stopping

آن نسبت به جاهای خالی در مجموعه‌های داده در دنیای واقعی، آن را به انتخابی محبوب برای بسیاری از متخصصان یادگیری ماشین تبدیل کرده است.

با توجه به شباهت‌های ذکر شده بین الگوریتم XGBoost و جنگل تصادفی، جا دارد که به برخی تفاوت‌های کلیدی این دو الگوریتم نیز اشاره کنیم.

۱. نوع الگوریتم: جنگل تصادفی یک الگوریتم بگینگ است، در حالی که XGBoost یک الگوریتم تقویت کننده یا بوستینگ است. روش‌های بگینگ چندین مدل را به طور موازی بر روی نمونه‌های فرعی مختلف داده‌های آموزشی ایجاد می‌کنند و پیش‌بینی‌های آن‌ها را ترکیب می‌کنند (شکل ۶-۳)، در حالی که روش‌های تقویت کننده، یک مدل واحد را به صورت تکراری، با تمرکز بر دشوارترین مثال‌ها در هر تکرار، می‌سازند.

۲. انتخاب ویژگی: جنگل تصادفی یک زیرمجموعه تصادفی از ویژگی‌ها را برای تقسیم هر گره تخصیص می‌دهد، در حالی که XGBoost ویژگی‌ها را بر اساس اهمیت آنها برای مدل فعلی انتخاب می‌کند.

۳. معاوضه بایاس-واریانس: جنگل تصادفی واریانس مدل را با میانگین‌گیری بیش از چندین مدل کاهش می‌دهد، در حالی که XGBoost با بهبود مکرر پیش‌بینی‌های مدل، هم سوگیری و هم واریانس را کاهش می‌دهد.

۴. مدیریت مجموعه داده‌های نامتعادل: XGBoost می‌تواند مجموعه داده‌های نامتعادل را بهتر از جنگل تصادفی با تنظیم عملکرد یک تابع هزینه و دادن وزن بیشتر به کلاس‌های اقلیت مدیریت کند.

۵. سرعت: XGBoost سریعتر از جنگل تصادفی است، به خصوص در مجموعه داده‌های بزرگ، به دلیل توانایی آن در موازی کردن فرآیند آموزش و بهینه‌سازی محاسبات.

به طور کلی، جنگل تصادفی یک الگوریتم ساده‌تر و کم هزینه‌تر و مقاوم‌تر به برخی تغییرات است که روی طیف وسیعی از مجموعه‌های داده به خوبی کار می‌کند، در حالی که XGBoost یک الگوریتم پیچیده‌تر و قدرتمندتر است که می‌تواند به دقت بالاتری دست یابد، اما برای استفاده مؤثر به تنظیم دقیق و تخصص بیشتری نیاز دارد. انتخاب بین دو الگوریتم بستگی به مشکل خاص در دست، اندازه و پیچیدگی مجموعه داده، و مبادله بین دقت و هزینه محاسباتی دارد.

۳-۴- برخی تکنیک‌های کمکی استفاده شده در فرآیند تعلیم و تصمیم‌گیری

۳-۱- بوت استرپ^{۷۸}

بوت استرپ یک تکنیک آماری است که در قدم اول چندین مجموعه داده مشابه با مجموعه‌ی دادگان اصلی را ایجاد می‌کند. ساخت این مجموعه‌ها معمولاً از طریق نمونه‌برداری مجدد از مجموعه داده اصلی و در صورت لزوم نمونه‌برداری همراه با جایگزینی انجام می‌پذیرد. تکنیک بوت استرپ در یادگیری ماشینی برای تخمین ویژگی‌های آماری یک مدل یا ارزیابی پایداری مدل نیز استفاده می‌شود.

در بوت استرپ، تعدادی داده‌ی تصادفی با استفاده از نمونه‌برداری و در صورت لزوم جایگزینی در مجموعه اصلی وارد می‌شوند و به این ترتیب مجموعه جدید ساخته می‌شود. با تکرار چند باره این فرآیند، چندین نمونه بوت استرپ به دست می‌آید. تعیین تعداد نمونه‌های بوت استرپ معمولاً^{۷۹} تابعی از اندازه مجموعه داده اصلی و سطح دقت مورد نظر است.

به عنوان مثال، اگر مجموعه داده ای با ۱۰۰۰ مشاهده داشته باشیم، می‌توانیم با جایگزینی تصادفی ۱۰۰۰ مشاهده، از مجموعه داده اصلی، ۱۰۰۰ نمونه بوت استرپ ایجاد کنیم. هر بوت استرپ در این مثال شامل چندین نمونه از مجموعه داده اصلی و چندین نمونه است که در مجموعه اصلی اصلاً وجود ندارند. همان‌گونه که گفته شد، نمونه‌های بوت استرپ برای تخمین ویژگی‌های آماری یک مدل استفاده می‌شوند. برای مثال، اگر بخواهیم میانگین یک متغیر را تخمین بزنیم، می‌توانیم میانگین هر نمونه بوت استرپ را محاسبه کنیم و سپس میانگین میانگین‌های نمونه بوت استرپ را بگیریم. این به ما تخمینی از میانگین می‌دهد که دقیق‌تر از یک تخمین واحد از مجموعه داده اصلی است.

بوت استرپ همچنین برای ارزیابی پایداری پیش‌بینی‌های مدل استفاده می‌شود. به عنوان مثال، اگر قرار باشد که یک مدل رگرسیون خطی را به مجموعه داده اصلی برآذش کنیم، می‌توانیم چندین نمونه بوت استرپ نیز ایجاد کنیم و مدل رگرسیون خطی را به هر نمونه بوت استرپ اعمال کنیم. سپس می‌توانیم پیش‌بینی مدل‌های مختلف را مقایسه کنیم تا بینیم برآوردها چقدر پایدار هستند. اگر تخمین‌ها در نمونه‌های بوت استرپ مشابه باشند، می‌توانیم به پایداری مدل اطمینان بیشتری داشته باشیم.

بوت استرپ یک تکنیک قدرتمند در یادگیری ماشینی است که می‌تواند برای طیف گسترده‌ای از مسائل و مدل‌ها استفاده شود. می‌توان از آن برای تخمین ویژگی‌های آماری مدل‌ها، ارزیابی پایداری تخمین‌های مدل و نیز ارزیابی عملکرد مدل‌ها استفاده نمود. با این حال، مهم است که در نظر داشته باشیم که نمونه‌های بوت

استرپ مستقل نیستند و این می‌تواند بر دقت تخمین‌ها تأثیر بگذارد. بنابراین، استفاده از تکنیک‌های دیگر مانند اعتبارسنجی متقابل^{۷۹} برای ارزیابی دقت تخمین‌های مدل مهم است.

۴-۲- توفیق زودهنگام^{۸۰}

توقف زودهنگام تکنیکی است که در یادگیری ماشین برای جلوگیری از بیش‌برازش و بهبود عملکرد تعمیم یک مدل استفاده می‌شود. این شامل توقف آموزش یک مدل زمانی است که شروع به بیش‌برازش با داده‌های آموزشی می‌کند، در حالی که آموزش باید به طور سنتی تا زمان همگرایی یا تعداد ثابتی از تکرارها ادامه می‌یافتد.

دلیل توقف زودهنگام این است که با افزایش تعداد تکرارها یا دوره‌ها در طول آموزش، مدل شروع به حفظ داده‌های آموزشی و تطبیق با نویز موجود در داده‌ها می‌کند، به جای این‌که الگوهای زیربنایی را که قابلیت تعمیم به داده‌های جدید و دیده نشده را دارند، بیاموزد. این منجر به بیش‌برازش و عملکرد ضعیف در اعتبارسنجی یا داده‌های آزمایشی می‌شود.

برای اجرای توقف زودهنگام، ابتدا داده‌ها را به مجموعه‌های آموزشی^{۸۱}، اعتبارسنجی^{۸۲} و تست^{۸۳} تقسیم می‌کنیم. مجموعه آموزشی برای برازش مدل، مجموعه اعتبارسنجی برای نظارت بر عملکرد مدل در طول آموزش و مجموعه آزمون برای ارزیابی عملکرد نهایی مدل استفاده می‌شود.

در طول آموزش، عملکرد مدل در مجموعه اعتبارسنجی در فواصل زمانی منظم، معمولاً پس از هر دوره یا تعداد ثابتی از تکرارها، نظارت می‌شود. اگر عملکرد اعتبارسنجی شروع به کاهش کند، به عنوان مثال، هزینه مجموعه اعتبارسنجی شروع به افزایش کند، آموزش متوقف می‌شود و مدلی با بهترین عملکرد اعتبارسنجی ذخیره می‌شود.

معیار توقف را می‌توان بر اساس چندین عامل مانند تفاوت بین هزینه آموزشی و اعتبارسنجی یا نرخ تغییر هزینه در مجموعه اعتبارسنجی تعیین کرد. در عمل، یک رویکرد متداول این است که زمانی که هزینه اعتبارسنجی برای تعداد ثابتی از دوره‌ها بهبود نمی‌یابد، تمرین متوقف شود. این زمان، که به عنوان پارامتر صبر^{۸۴} شناخته می‌شود، بر اساس مسئله‌ی مورد نظر و یا منابع محاسباتی موجود تنظیم می‌شود.

توقف زودهنگام را می‌توان برای الگوریتم‌های مختلف یادگیری ماشین، مانند شبکه‌های عصبی، درخت‌های تصمیم‌گیری و ماشین‌های بردار پشتیبانی اعمال کرد. این یک تکنیک ساده و در عین حال موثر است که

^{۷۹} Cross validation

^{۸۰} Early stopping

^{۸۱} Training set

^{۸۲} Validation set

^{۸۳} Test set

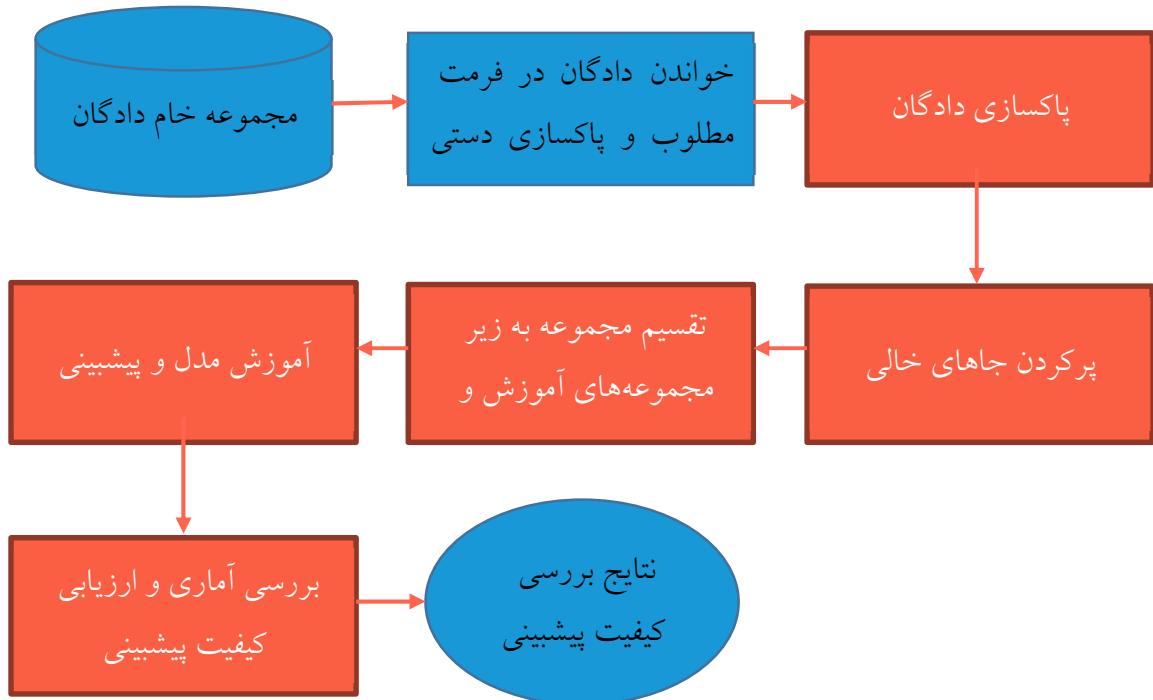
^{۸۴} Patience parameter

می‌تواند عملکرد تعمیم یک مدل را به طور قابل توجهی بهبود بخشد، زمان آموزش را کاهش دهد و منابع محاسباتی را ذخیره کند. با این حال، توجه به این نکته مهم است که توقف زودهنگام بهترین عملکرد را در مجموعه تست تصمین نمی‌کند و ممکن است در برخی موارد منجر به کم برآزش شود. بنابراین، برای بهبود عملکرد کلی مدل باید در ترکیب با تکنیک‌های دیگر مانند منظم‌سازی^{۸۵} و اعتبارسنجی متقابل استفاده شود.

۳-۵- فلوچارت کلی روش پیشنهادی

در شکل ۷-۳ نمایی کلی از روش پیشنهادی شامل مراحل آماده‌سازی دستی و سپس ماشینی دادگان، به کار گرفتن مدل و بررسی آماری و ارزیابی روش پیشنهادی ارائه شده است. توضیح این مراحل با جزئیات بیشتر و بررسی نتایج میانی و نهایی روش پیشنهادی موضوع فصل چهارم این تحقیق خواهد بود. به طور خلاصه در این تحقیق و پس از شیوه سازی‌هایی که در فصل ۴ گزارش می‌شوند، الگوریتم‌های زیر برای قرار گرفتن در خانه‌های این فلوچارت انتخاب شده‌اند:

- پر کردن جاهای خالی دادگان ثبت نشده: ۱) روش padding، ۲) ترکیب روش‌های جلوسو و بازگشتی، ۳) روش k-نزدیک‌ترین همسایگی
- ابعاد مورد استفاده در تقسیم مجموعه اطلاعات: ۱) بخش آموزش: ۷۰٪ از مشاهدات، ۲) بخش اعتبارسنجی: ۲۰٪ از دادگان بخش آموزش که در ۵ گروه اعتبارسنجی متقابل مورد استفاده قرار می‌گیرند. ۳) بخش تست: ۳۰٪ از مشاهدات
- استخراج ویژگی: ۱) حذف ویژگی‌های غیر تأثیرگذار با توجه به بالا بودن ضریب همبستگی آنها با سایر ویژگی‌ها. ۲) روش تحلیل مؤلفه‌های اصلی (PCA) و ۳) حذف ویژگی بازگشتی (RFE)
- مدل‌های برگزیده: الگوریتم جنگل تصادفی و XGBoost
- مدل‌های مقایسه: رگرسیون لجستیک و ماشین بردار پشتیبان ROC
- بهینه‌سازی ابرپارامترها: اعتبارسنجی متقابل با معیار افزایش سطح زیر منحنی
- تکنیک‌های کمکی: بوت استرپ، وزن‌دهی به کلاس با نمونه‌های کمتر
- معیارهای ارزیابی عملکرد مدل: دقت، معیار اف ۱ و مساحت زیر منحنی که با استفاده از دو روش توضیح داده شده در فصل ۴ محاسبه می‌شوند.



شکل ۳-۷: طرح کلی برای پایپ لاین تحقیق. این پایپ لاین به عنوان مرجع در نظر گرفته شده و جزئیات و تغییرات در فصل ۴ توضیح داده شده‌اند

فصل ۴: شبیه‌سازی و تجزیه و تحلیل داده‌ها

۴-۱- مقدمه

در فصل ۳، روش پیشنهادی این تحقیق و همچنین روش‌های استاندارد برای مقایسه‌ی عملکرد مدل پیاده‌سازی شده‌اند، به طور مفصل معرفی گردیدند. در این فصل به معرفی پایگاه داده‌ی مورد استفاده و نتایج به دست آمده از پیاده‌سازی این روش‌ها پرداخته می‌شود. این نتایج در سه بخش پیش‌پردازش دادگان، انتخاب (استخراج) ویژگی و تعلیم طبقه‌بندی کننده‌ها ارایه می‌گردند. به منظور مقایسه‌ی عملکرد مدل‌ها نیز شاخص‌های دقت (Acc)، امتیاز اف ۱ و سطح زیر منحنی که در فصل ۲ معرفی شده‌اند، گزارش می‌شوند. تمامی مراحل، به زبان برنامه‌نویسی پایتون و با استفاده از کتابخانه‌های SciPy، NumPy، Pandas و Scikit-learn طراحی و پیاده‌سازی شده‌اند و کدهای مربوطه در سایت گیت‌هاب به طور عمومی به اشتراک گذاشته شده‌اند.

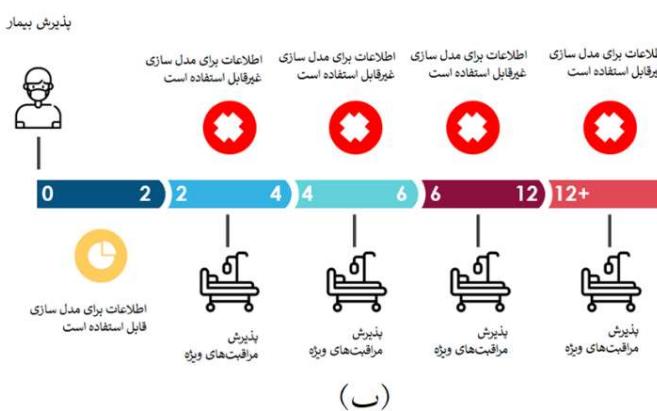
۴-۲- معرفی پایگاه دادگان

داده‌های مورد استفاده در این تحقیق از ۳۸۴ مراجعه به بیمارستان سیریولبانز در سائوپلو بزرگ جمع‌آوری شده است و به صورت رایگان در اختیار عموم قرار گرفته و از پایگاه ایترنی Kaggle قابل دسترسی است [۷۶].

متغیرهای حیاتی و نتایج آزمایش خون و غلظت اکسیژن خون برای بیماران مبتلا به کووید ۱۹ در پنجره‌های زمانی ۲ ساعته از لحظه پذیرش تا ۱۲ ساعت بعد از پذیرش اندازه‌گیری و ثبت شده‌اند.

این مجموعه داده‌ها در فرمت استاندارد مايكروسافت اکسل **.xlsx** ذخیره شده است و کل مجموعه داده‌ها شامل یک جدول با تعداد ۱۹۲۶ سطر (۱۹۲۵ سطر اطلاعات بیماران و یک سطر عنوان ستون‌ها) و ۲۳۱ ستون است.

پایگاه Kaggle بر این موضوع تاکید کرده است که اطلاعات حیاتی بیمارانی که دارای پرچم هدف هستند (در بخش مراقبت‌های ویژه بستری شده‌اند) نباید در مدل سازی برای پیش‌بینی استفاده شود. زیرا ممکن است که نیاز به بستری شدن در بخش مراقبت‌های ویژه قبل از جمع‌آوری اطلاعات تشخیص داده شده باشد (شکل ۴-۱). [۷۶]



شکل ۴-۱: اطلاعات حیاتی بیمارانی که در حال حاضر در بخش مراقبت‌های ویژه بستری شده‌اند دیگر نباید در مدل سازی برای پیش‌بینی استفاده شود.

ویژگی‌هایی که در هر مراجعه ثبت شده‌اند شامل ۵۴ ویژگی است: ۳ ویژگی از اطلاعات جمعیتی بیمار، ۹ ویژگی از سوابق بیماری، ۳۶ ویژگی از نتایج آزمایش خون و ۶ ویژگی از اندازه‌گیری علائم حیاتی هستند. مقادیر هر یک از این ۵۴ ویژگی براساس ماهیت اندازه‌گیری، یا به صورت یک عدد حقیقی و یا به صورت یک عدد دو دویی ذخیره شده‌اند. در صورتی که مقدار ویژگی یک عدد حقیقی باشد، داده موجود در هر ستون

بر اساس مقادیر بیشینه و کمینه در محدوده [۱، ۱] نرمال‌سازی شده‌اند. ضمناً این داده‌های حقیقی به صورت مقادیر بیشینه، کمینه، میانگین و تغییرات دامنه‌ی هر یک از ویژگی‌های اندازه‌گیری شده ثبت شده‌اند که در مجموع تعداد ستون‌های مجموعه داده را به ۲۳۱ می‌رسانند. این ویژگی‌ها در فصل ۴ معرفی می‌شوند.

به علاوه، اطلاعات حیاتی بیماران تنها در صورتی در مدل‌سازی برای پیش‌بینی استفاده می‌شود که این بیماران در حال حاضر در بخش مراقبت‌های ویژه بستری نشده باشند. با در نظر گرفتن این موضوع مراجعانی که در بدو ورود در بخش مراقبت‌های ویژه بستری شده‌اند از فرآیند مدل‌سازی حذف می‌شوند و عملاً تعداد مراجعات قابل بررسی به ۳۵۳ مورد کاهش می‌یابد.

همان‌گونه که در بالا نیز ذکر شده است، اطلاعات بیماران در بازه‌های ۲ ساعته و در پنج مرحله جمع‌آوری شده است. طبیعتاً در این بین برخی از بیماران در مرحله اول و برخی در مرحله دوم تا پنجم به بخش مراقبت‌های ویژه منتقل شده‌اند. برخی هم هرگز در بخش مراقبت‌های ویژه بستری نشده‌اند. از طرف دیگر اطلاعات کسانی که در مراقبت‌های ویژه بستری شده‌اند، دیگر قابل استفاده نخواهد بود. این موضوع باعث می‌شود که سری زمانی در مورد هر بیمار طول متفاوتی داشته باشد و انتخاب و آموزش مدل پیچیده می‌شود. برای طراحی یک پایپ لاین استاندارد برای مقایسه، تنها از اطلاعات بیمارانی استفاده شده است که شامل سری کامل زمانی هستند (شکل ۱-۴-الف). بدین معنی که این بیماران در چهار مرحله اولی در بخش عادی بستری بوده‌اند و تنها در مرحله آخر (پنجم) یا به بخش مراقبت‌های ویژه منتقل شده‌اند و یا مخصوص شده‌اند. با اعمال این محدودیت تعداد مراجعات قابل بررسی باز هم کاهش می‌یابد و به ۲۵۵ مورد می‌رسد.

۴-۲-۱- آماده‌سازی دادگان

یکی از چالش‌های کار با این مجموعه –و اغلب دادگان پزشکی– وجود جاهای خالی در اطلاعات جمع‌آوری شده است. در این وضعیت که اطلاعات بیمار در یک یا چند تا از پنجره‌های زمانی جمع‌آوری نشده است، پیشنهاد جمع‌آورنده‌گان اطلاعات این است که جای خالی این اطلاعات می‌تواند با اطلاعات اندازه‌گیری شده در پنجره زمانی قبل پر شود [۷۶]. علاوه بر این روش پدینگ^{۸۶}، روش‌های زیر نیز جهت پر کردن جاهای خالی پیاده‌سازی شدند که بعداً در بخش ارزیابی عملکرد مدل، تاثیر این روش‌ها نیز گزارش می‌شود.

(الف) روش‌های ساده‌ی آماری:

در این روش‌ها جاهای خالی با استفاده از یکی از ابزارهای آماری توصیفی (مثلًاً میانگین، میانه یا متداول‌ترین) در امتداد هر ستون جایگزین می‌شوند.

متداول‌ترین مقدار(Most Frequent) : پر تکرارترین مقدار در امتداد هر ستون در جای خالی نوشته می‌شود. در صورتی که بیش از یک مقدار واجد این شرایط باشند مقدار کمتر نوشته خواهد شد.

میانگین (mean): میانگین مقادیر موجود در هر ستون در جای خالی نوشته می‌شود.

میانه (median): میانه مقادیر موجود در هر ستون در جای خالی نوشته می‌شود.

ب) روش چند متغیره یا تکرارشونده:

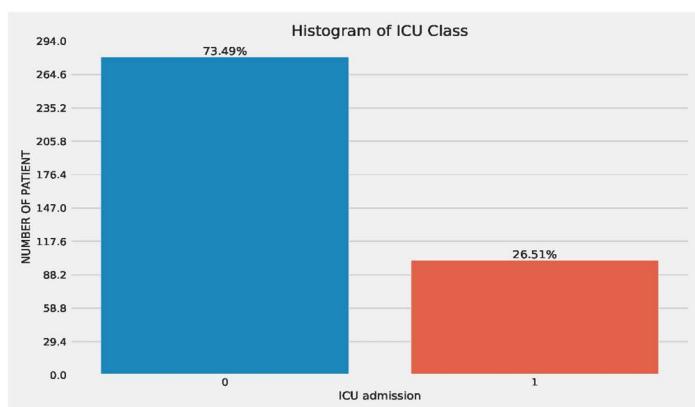
رویکرد پیچیده‌تری برای پیش‌بینی مقادیر خالی استفاده از روش‌های چند متغیره است. به این صورت که ویژگی شامل جاهای خالی به صورت تابعی از دیگر ویژگی‌ها تخمین زده می‌شود و سپس این تابع تخمینی برای پر کردن جاهای خالی مورد استفاده قرار می‌گیرد. تخمین تابع به صورت حلقه‌ای تکرارشونده برای حصول دقت بالاتر اجرا می‌شود.

ج) روش K-نرديكتريين همسايگي:

در روش پياده‌سازی از فاصله اقلیدسي برای تعیین نرديكتريين همساييه استفاده شده است. اين فاصله طبیعتاً برای همسايگي در هر ویژگي به صورت جداگانه حساب شده است و در روش پياده‌سازی شده همه فاصله‌ها به صورت يكناخت (بدون وزن‌دهی) محاسبه شده‌اند. در صورتی که در يك نمونه بيش از یک ویژگي مفقود شده باشد، به ازاي هر ویژگي ناموجود، همساييه‌های مختلفی برای اين نمونه پيدا خواهد شد. در صورتی که تعداد همساييه‌ها از حد مشخصی كمتر باشد و فاصله با داده‌های تعلم نامشخص، ميانگين داده‌های تعلم برای پر کردن ویژگي خالی استفاده می‌شود.

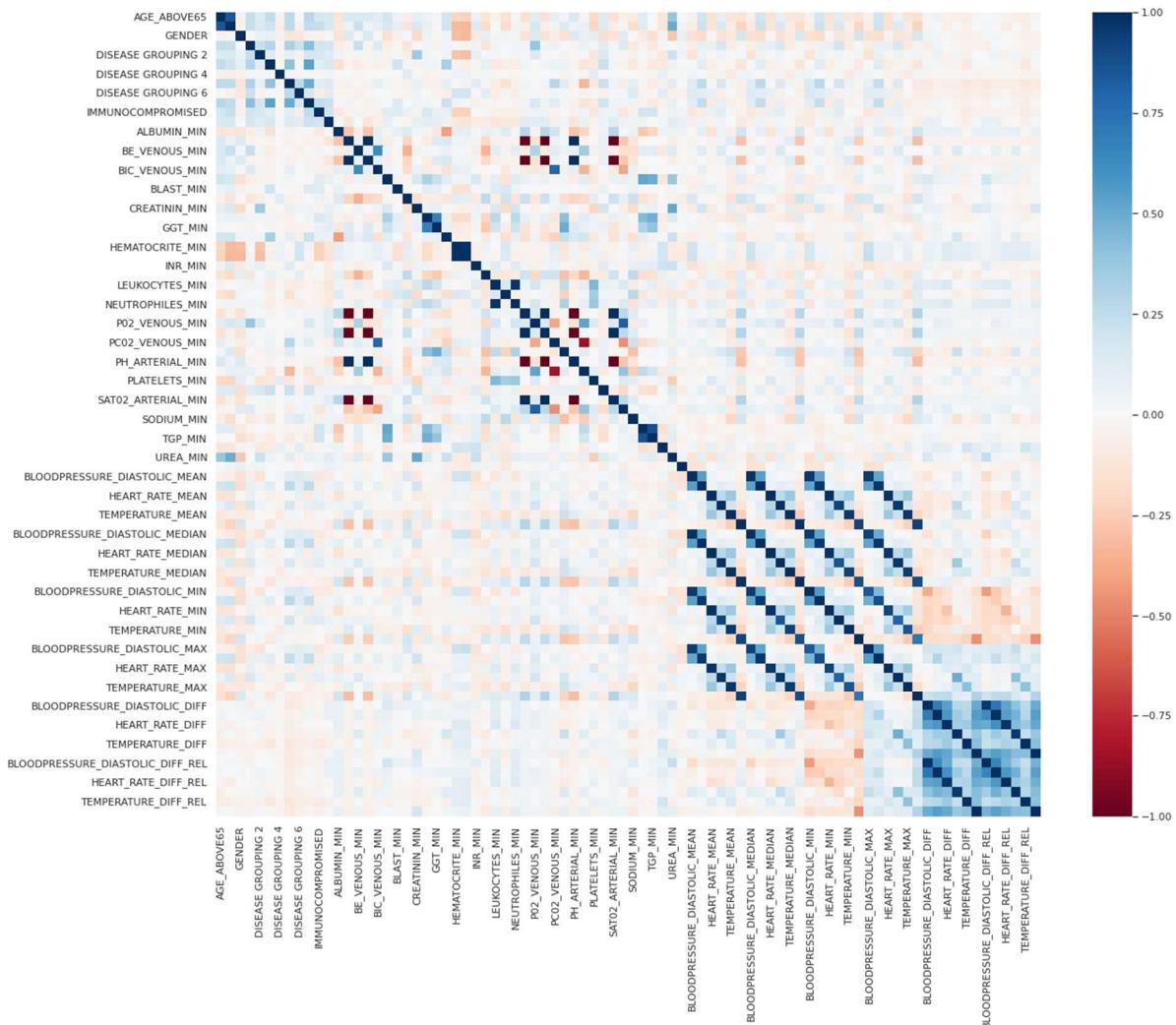
۴-۱-مشخصات آماری ویژگی‌های ثبت شده

پس از پر کردن جای خالی داده‌های ثبت نشده با استفاده از ترکيب روش‌های جلوسو و بازگشتی و حذف ردیف‌های تکراری، توزیع دو کلاس (منتقل شده/منتقل نشده به بخش مراقبت‌های ویژه) به صورت زیر مشاهده شد که در آن ۲۶/۵۱٪ نمونه‌ها به هر حال به ICU انتقال داده شده‌اند.



شکل ۴-۲: توزیع افراد منتقل شده به بخش مراقبت‌های ویژه

برای ایجاد درک بهتری نسبت به همبستگی ویژگی‌های ثبت شده، ماتریس همبستگی این ویژگی‌ها به روش همبستگی پیرسون محاسبه و در شکل ۳-۴ نشان داده شده است.



شکل ۳-۴: ماتریس همبستگی متغیرهای فیزیولوژیک ثبت شده پس از نرمالیزه شدن در بازه [-۱، ۱].

مرتب‌سازی اعداد همبستگی به ترتیب نزولی نشان می‌دهد که مقادیر کمینه‌ی ثبت شده برای ویژگی فیزیولوژیک مرتبط با فشار خون و ... بیشترین همبستگی را با یکدیگر دارند (جدول ۱-۴).

جدول ۱-۴: فهرست ویژگی‌های به ترتیب همبستگی بر اساس شاخص پیرسون

	Feature_1	Feature_2	Pearson Correlation
1128	BE_ARTERIAL_MIN	PH_ARTERIAL_MIN	1
1300	BIC_ARTERIAL_MIN	SAT02_ARTERIAL_MIN	1
1125	BE_ARTERIAL_MIN	PC02_ARTERIAL_MIN	1
1291	BIC_ARTERIAL_MIN	P02_ARTERIAL_MIN	1
1123	BE_ARTERIAL_MIN	P02_ARTERIAL_MIN	1
...

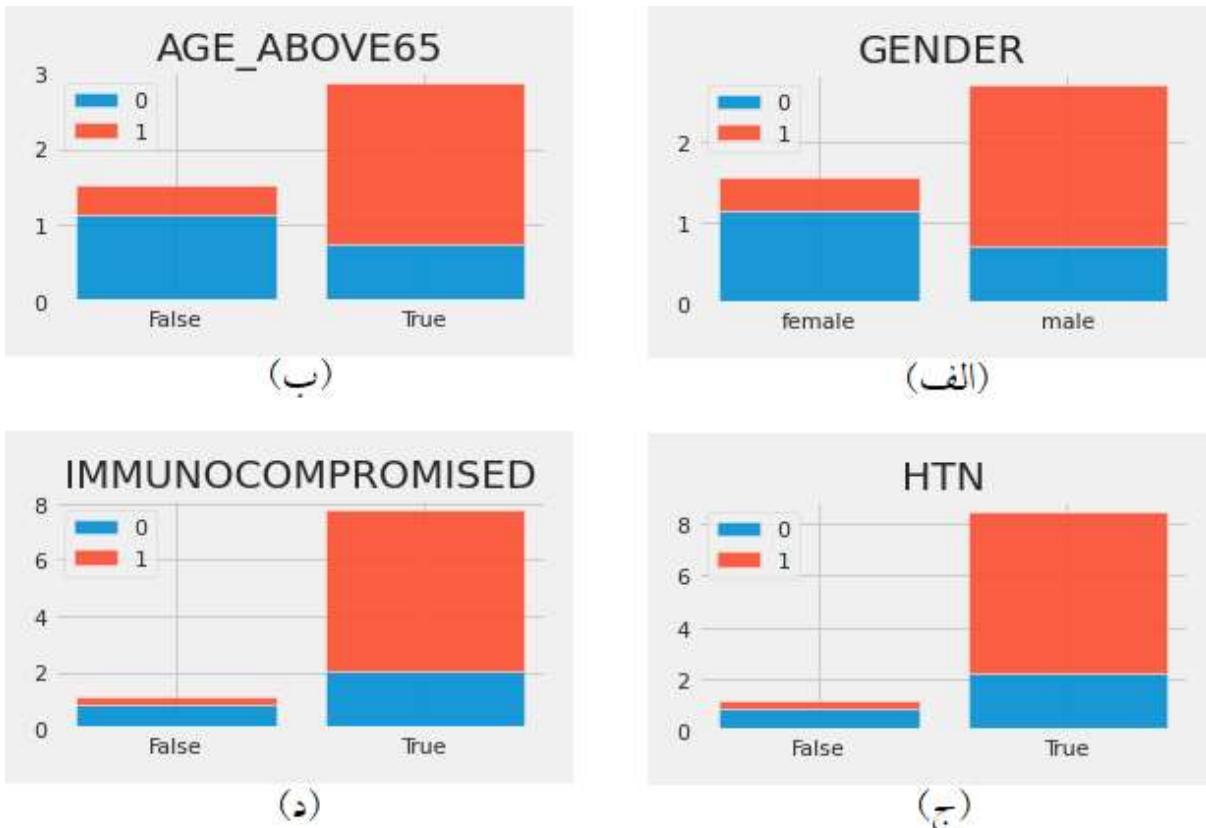
546	DISEASE GROUPING 4	SODIUM_MIN	0.00028
1398	BIC_VENOUS_MIN	BLOODPRESSURE_DIASTOLIC_MEDIAN	0.000155
1362	BIC_VENOUS_MIN	BLAST_MIN	0.000153
520	DISEASE GROUPING 4	BIC_VENOUS_MIN	0.000113
47	AGE_ABOVE65	DIMER_MIN	0.000008
3486 rows × 3 columns			

با توجه به معنی آماری ضریب همبستگی پیرسون، می‌توان آن دسته از ویژگی‌هایی که ضریب همبستگی بالاتر از ۹۹٪ دارند را حذف نمود که این اطلاعات بعداً (در صورت رضایت‌بخش نبودن نتایج) در مرحله‌ی کاهش تعداد ویژگی‌ها به کار گرفته خواهد شد. فهرست کامل این ویژگی‌ها در جدول ۲-۴ نشان داده شده است.

جدول ۲-۴: فهرست کامل ویژگی‌های با ضریب همبستگی بالاتر از ۹۹٪

	Feature_1	Feature_2	Pearson Correlation
1128	BE_ARTERIAL_MIN	PH_ARTERIAL_MIN	1
1300	BIC_ARTERIAL_MIN	SAT02_ARTERIAL_MIN	1
1125	BE_ARTERIAL_MIN	PC02_ARTERIAL_MIN	1
1291	BIC_ARTERIAL_MIN	P02_ARTERIAL_MIN	1
1123	BE_ARTERIAL_MIN	P02_ARTERIAL_MIN	1
2812	PC02_ARTERIAL_MIN	SAT02_ARTERIAL_MIN	1
1132	BE_ARTERIAL_MIN	SAT02_ARTERIAL_MIN	1
2640	P02_ARTERIAL_MIN	PH_ARTERIAL_MIN	1
1296	BIC_ARTERIAL_MIN	PH_ARTERIAL_MIN	1
1107	BE_ARTERIAL_MIN	BIC_ARTERIAL_MIN	1
1293	BIC_ARTERIAL_MIN	PC02_ARTERIAL_MIN	1
2644	P02_ARTERIAL_MIN	SAT02_ARTERIAL_MIN	1
2808	PC02_ARTERIAL_MIN	PH_ARTERIAL_MIN	1
2637	P02_ARTERIAL_MIN	PC02_ARTERIAL_MIN	1
3064	PH_ARTERIAL_MIN	SAT02_ARTERIAL_MIN	1
6466	TEMPERATURE_DIFF	TEMPERATURE_DIFF_REL	0.999672
6551	OXYGEN_SATURATION_DIFF	OXYGEN_SATURATION_DIFF_REL	0.998998
4256	HEART_RATE_MEAN	HEART_RATE_MEDIAN	0.996628
4171	BLOODPRESSURE_SISTOLIC_MEAN	BLOODPRESSURE_SISTOLIC_MEDIAN	0.996262
4086	BLOODPRESSURE_DIASTOLIC_MEAN	BLOODPRESSURE_DIASTOLIC_MEDIAN	0.991894
4426	TEMPERATURE_MEAN	TEMPERATURE_MEDIAN	0.990118

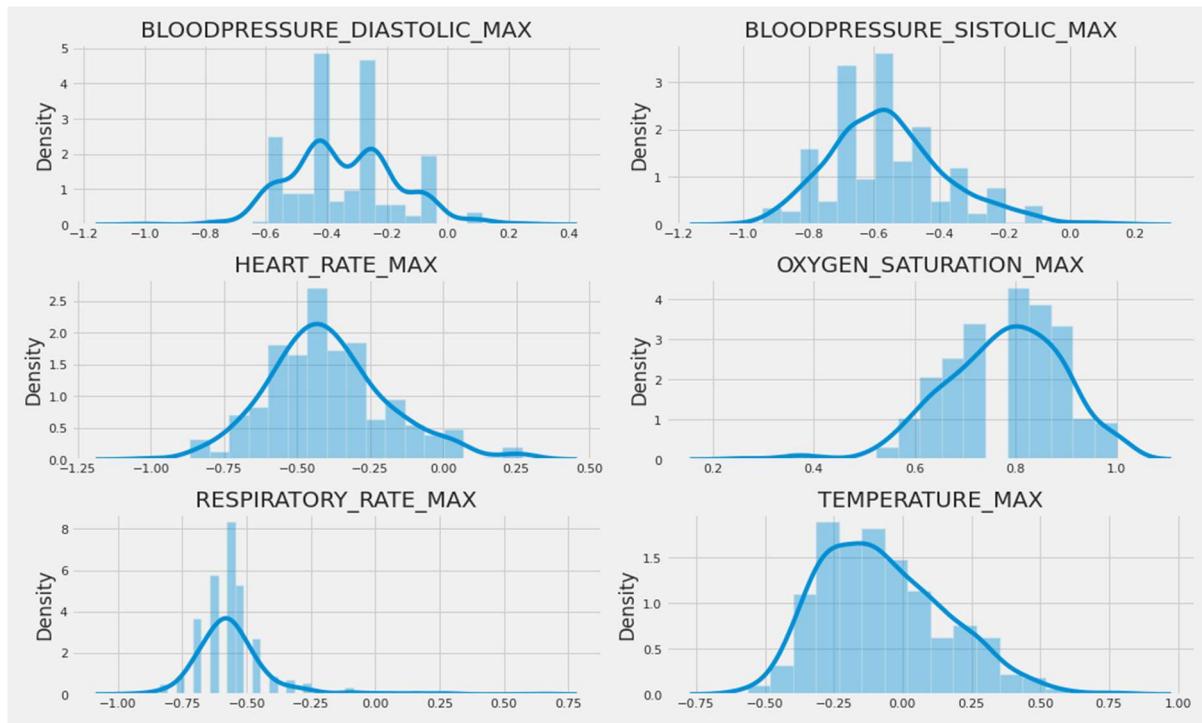
همچنین در محاسبه‌ی همبستگی این متغیرها با خروجی‌های مطلوب مدل، یعنی بردار برچسب‌ها (اعداد باينري نشان‌دهنده‌ی بستری با عدم بستری در ICU) در نظر گرفته نشده. به نظر می‌رسد که سن بالای ۶۵ سال، متغیر HTN و وضعیت سیستم ایمنی با احتمال بستری در ICU رابطه‌ی مستقیمی داشته باشند (شکل ۴-۴).



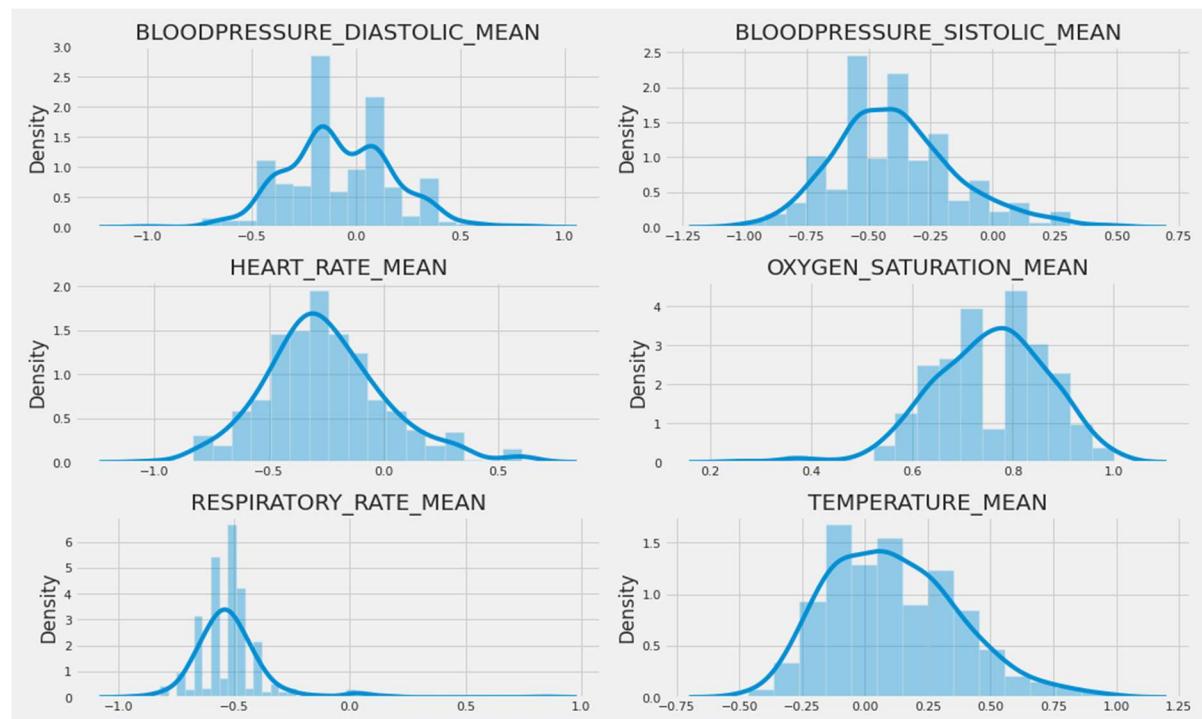
شکل ۴-۴: نمایش ارتباط متغیرهای الف- جنسیت، ب- سن بالای ۶۵ سال، ج- HTN و د- وضعیت سیستم ایمنی با احتمال بستری در ICU. در این کدگذاری ۰ به معنای عدم انتقال و ۱ به معنای انتقال بیمار به بخش مراقبت‌های ویژه است. به عنوان مثال، در شکل ب، تعداد افرادی که سن آنها بالای ۶۵ بوده و به بخش ICU منتقل شده‌اند قابل توجه است. روند مشابهی در شکل د نیز دیده می‌شود. به این معنی که اگر سیستم ایمنی فرد ضعیف بوده، با احتمال بسیار بالایی در نهایت به ICU منتقل شده است.

۴-۲-۲- دسته بندی ویژگی‌ها

در ادامه، ویژگی‌ها به صورت دستی و براساس این‌که کمینه، بیشینه، میانگین، تفاوت و میانه‌ی یک متغیر فیزیولوژیک را اندازه‌گیری می‌کنند دسته‌بندی شدند تا توزیع آن‌ها در مجموعه دادگان بررسی شود. به عنوان مثال در شکل‌های ۴-۴ و ۴-۵، این توزیع‌ها برای بیشینه و همچنین میانگین متغیرها و تفاوت عددی آن‌ها از یک اندازه‌گیری تا اندازه‌گیری بعدی نشان داده شده است. شایان ذکر است که اعداد ثبت شده به گونه‌ای نرمالیزه شده‌اند که در بازه‌ی [-۱ و ۱] قرار گیرند.



شکل ۴-۵: توزیع متغیرهای فیزیولوژیکی که به صورت بیشینه گزارش شده‌اند.



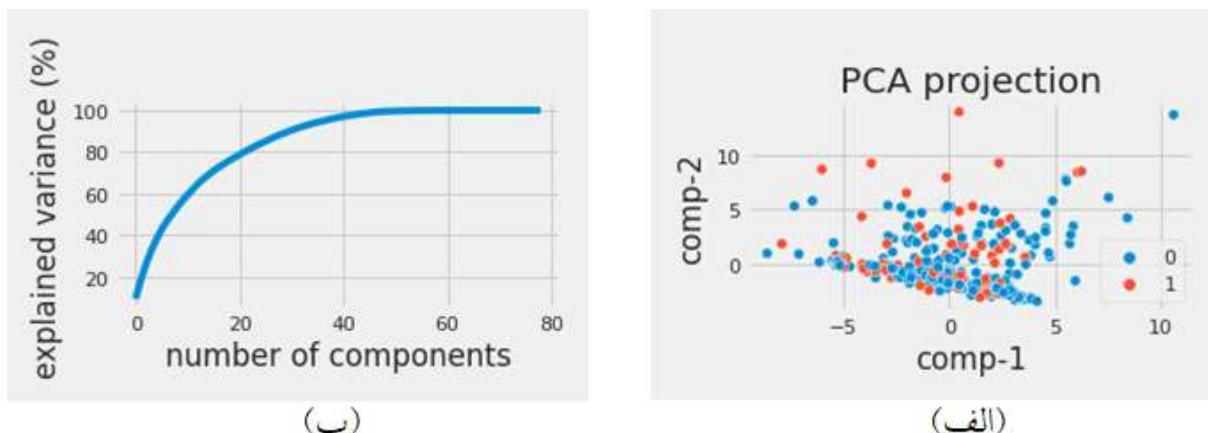
شکل ۴-۶: توزیع متغیرهای فیزیولوژیکی که به صورت میانگین عددی آنها از یک اندازه‌گیری تا اندازه‌گیری بعدی نشان داده شده است.

۴-۲-۳- کاهش بعد ویژگی‌ها و نمایش دادگان

با توجه به نتایج بخش قبل و همچنین به منظور ایجاد تصویر بهتری از توزیع دادگان در فضاهایی با بعد پایین‌تر، در این بخش روش‌های تحلیل مؤلفه‌های اصلی و t-SNE روی دادگان اعمال می‌شوند.

الف- تحلیل مؤلفه‌های اصلی (PCA):

برای نمایش دادگان در فضای با بعد پایین‌تر ابتدا آن‌ها را به گونه‌ای استانداردسازی (هنجرسازی) کردیم که میانگین هر ویژگی صفر شود. سپس با استفاده از روش محاسبه‌ی ماتریس کوواریانس، بردارها و مقادیر ویژه محاسبه شده، سپس دو مؤلفه اصلی استخراج گردیده و توزیع دادگان در فضا در شکل ۴-۷-الف نشان داده شده است که عدم تعادل تعداد نمونه‌ها در آن به خوبی مشاهده می‌شود. به منظور بررسی تأثیر هر مؤلفه روی واریانس، مقدار تجمعی واریانس‌های تک‌تک مؤلفه‌ها محاسبه و در شکل ۴-۷-ب رسم شده است. به خوبی مشاهده می‌شود که در صورت استفاده از یک روش خطی مانند PCA برای کاهش ابعاد ورودی، حداقل ۵۰ مؤلفه‌ی اصلی (تقریباً ۶۳٪ مؤلفه‌ها) باید لحاظ شوند. شایان ذکر است که پس از حذف ویژگی‌هایی که همبستگی بالایی دارند، مجموع کل ویژگی‌ها به ۷۹ کاهش یافته است.

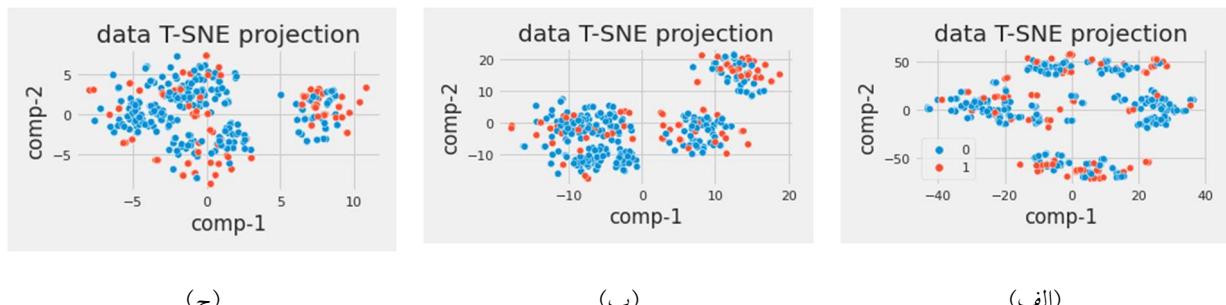


شکل ۴-۷: سمت راست: نمایش توزیع نمونه‌ها در فضای دو بعدی متشکل از دو مؤلفه اصلی محاسبه شده به روش PCA، سمت چپ: مقدار تجمعی واریانس‌های تک‌تک مؤلفه‌ها. در این کدگذاری ۰ به معنی عدم انتقال و ۱ به معنی انتقال بیمار به بخش مراقبت‌های ویژه است.

ب) روش t-SNE

در ادامه برای ارزیابی توزیع ویژگی‌ها و توازن تعداد نمونه‌ها، از نمایش داده‌ها در یک فضای n -بعدی از این روش غیرخطی مبتنی بر تعریف همسایگی در فضای n -بعدی استفاده شده است. در این پیاده‌سازی ابتدا یکتابع توزیع احتمالات بر روی هر دو نقطه در فضای n -بعدی در نظر گرفته

می‌شود به گونه‌ای که اینتابع برای نقاط مشابه (بر اساس یک معیار خاص مشابهت) عدد بالاتری را دریافت می‌کند و برای نقاط غیرمشابه، عدد احتمال کمتری را در نظر می‌گیرد. سپس، روش t-SNE تابع توزیع احتمال مشابهی را در فضای با بعد پایین‌تر در نظر گرفته و سپس واگرایی این دو توزیع احتمالی را که با معیار کولبک-لیبلر بر حسب موقعیت دو نقطه در فضا سنجیده می‌شود، کمینه می‌کند [۸۲]. در این پیاده‌سازی، از معیار فاصله‌ی اقلیدسی برای سنجش فاصله/شباهت میان دو نقطه استفاده می‌شود. برای رسم شکل ۴-۸، تلاش شده که با تغییر پارامترهای این الگوریتم، توزیع نمونه‌ها در فضای دو بعدی به صورت خوش‌های مجزا دیده شوند که ظاهراً این امر با استفاده از این روش و تغییر این پارامترها امکان‌پذیر نیست. در این بررسی‌ها هم از داده‌های هنجار شده و هم غیر هنجار شده استفاده گردید اما تغییری در توزیع خوش‌های مشاهده نشد.



شکل ۴-۸: نمایش توزیع دادگان پس از کاهش بعد با استفاده از روش t-SNE با تغییر پارامتر پیچیدگی (Perplexity) به ازای مقادیر ۱۰ (سمت راست)، ۵۰ (وسط) و ۱۰۰ (سمت چپ). در این کدگذاری ۰ به معنای عدم انتقال و ۱ به معنای انتقال بیمار به بخش مراقبت‌های ویژه است.

۴-۳- کاهش بعد ویژگی‌ها و تعلیم مدل

۴-۳-۱- ایجاد مدل پایه برای مقایسه

پیش از طراحی و تعلیم مدل‌های مبتنی بر یادگیری گروهی، مجموعه اطلاعات به سه بخش آموزش، اعتبارسنجی و تست تقسیم شده و مدل‌های استانداری مانند طبقه‌بندی کننده رگرسیون لجستیک و مدل ماشین بردار پشتیبان پیاده‌سازی شدند.

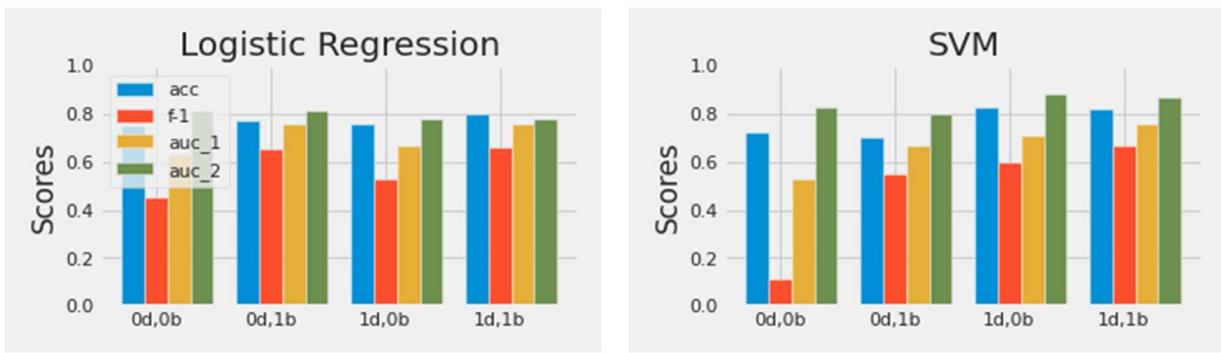
هر دو مدل در چهار حالت مختلف اجرا شده و نتایج با هم مقایسه شده‌اند (جدول ۴-۳):

- بدون کاهش ابعاد و بدون متوازن سازی وزن‌ها (0d, 0b)
- بدون کاهش ابعاد و با متوازن سازی وزن‌ها (0d, 1b)
- با کاهش ابعاد و بدون متوازن سازی وزن‌ها (1d, 0b)
- با کاهش ابعاد و با متوازن سازی وزن‌ها (1d, 1b)

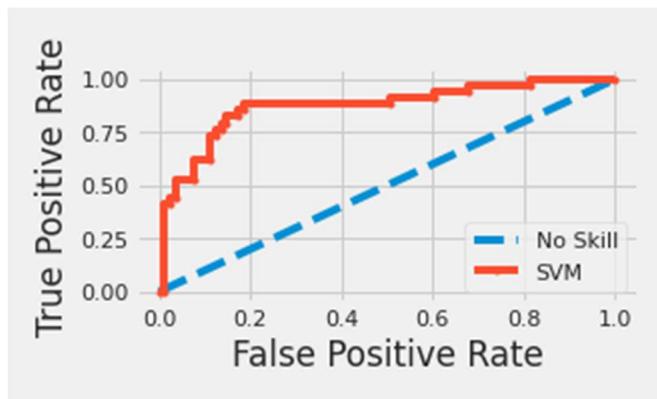
ورودی این مدل‌ها، اطلاعات جمع‌آوری شده از بیماران در چهار پنجره زمانی است که به صورت یک بردار یک بعدی تغییر فرم داده شده است. پرچم هدف یک داده دوتایی است. عدد ۰ به معنی عدم بستره شدن بیمار در بخش مراقبت‌های ویژه است و عدد ۱ به معنی بستره شدن بیمار خواهد بود.

واضح است که این نوع پیش‌بینی به نحوی کم کیفیت‌ترین حالت پیش‌بینی است. حال ایده‌آل این است که مدل با دیدن اطلاعات بیمار در اولین پنجره زمانی –یعنی سریع‌ترین زمان– پیش‌بینی دقیقی درباره احتمال بستره شدن بیمار در آینده انجام دهد. با این وجود در این طراحی اولیه اطلاعات تا آخرین پنجره زمانی قبل از بستره شدن بیمار برای آموزش استفاده خواهد شد تا به عنوان مرجعی برای مقایسه‌های آینده استفاده شود. نتایج اولین پیش‌بینی با شرایط فوق اگرچه دارای دقتی متوسط (بیش از ۷۰٪) است، مقدار بسیار پایین امتیاز اف ۱ به وضوح نشان می‌دهد که عدم تعادل در تعداد نمونه‌های بستره و غیر بستره در بخش مراقبت‌های ویژه، منجر به سوگیری طبقه‌بندی‌کننده به سمت تشخیص همه نمونه‌ها به عنوان سالم شده است. وزن‌گذاری اطلاعات ورودی بر اساس تعداد موارد، نتایج پیش‌بینی را به طرز قابل توجهی بهبود بخشدید که تاییدی بر این عدم تعادل است. علاوه بر آن، تاثیر کاهش ابعاد نیز قابل توجه است. ارزیابی کمی نتایج پیش‌بینی‌ها در شکل ۴-۹ قابل مشاهده است. در تهیه‌ی این شکل، ابتدا مدل‌ها بدون کاهش بعد ویژگی‌های ورودی و بدون وزن‌گذاری کلاس‌ها تعلیم یافتند (حالت ۰d, ۰b). سپس، بدون کاهش بعد ویژگی‌های ورودی، این بار وزنی متناسب با معکوس فراوانی اعضا به هر کلاس اختصاص داده شده و عملکرد مدل ارزیابی گردید (حالت ۱d, ۱b). در ادامه و برای بررسی اثر کاهش ابعاد ورودی، روش PCA اعمال شده و ۵۰ مؤلفه‌ی اول، به منظور تعلیم و ارزیابی مدل مورد استفاده قرار گرفتند.

همانگونه که در شکل ۴-۹ دیده می‌شود، مدل ماشین بردار پشتیبان پس از اعمال کاهش بعد و متوازن سازی کلاس‌ها، بهترین عملکرد را نشان می‌دهد (دقت، ۸۲٪، امتیاز اف ۱، ۶۷٪ و مساحت زیر منحنی ۰،۷۶). مساحت زیر منحنی ۰،۸۷. منحنی ROC این مدل در شکل ۱۰-۴ نشان داده شده است. شایان ذکر است که در تمامی شبیه سازی‌ها، روش اعتبارسنجی متقابل و با در نظر گرفتن ۵ تکرار به منظور یافتن بهترین ابرپارامترها اعمال شده است. به علاوه، ابتدا مساحت زیر منحنی بر روی برچسب‌های پیش‌بینی شده برای دادگان ورودی محاسبه شده است (auc_1). این در حالیست که در مقالاتی که با موضوع پیش‌بینی بیماری کووید-۱۹ و همچنین ارزیابی احتمال بستره در بخش مراقبت‌های ویژه منتشر شده‌اند، این مقدار به صورت متفاوتی محاسبه می‌شود. در این روش، به جای پیش‌بینی برچسب کلاس هر نمونه، احتمال تعلق نمونه به یکی از کلاس‌ها محاسبه شده و سپس منحنی ROC براساس این احتمال رسم شده و مساحت زیر آن اندازه‌گیری می‌شود (auc_2).



شکل ۹-۴: عملکرد مدل‌های پایه (رگرسیون لجستیک و ماشین بردار پشتیبان) در حالت‌های مختلف با و بدون اعمال کاهش ابعاد ورودی و متوازن سازی تعداد نمونه‌های کلاس‌ها. دقت، امتیاز اف ۱ و مساحت زیر منحنی به عنوان معیارهای ارزیابی عملکرد مدل با رنگ‌های متفاوت نشان داده شده‌اند.



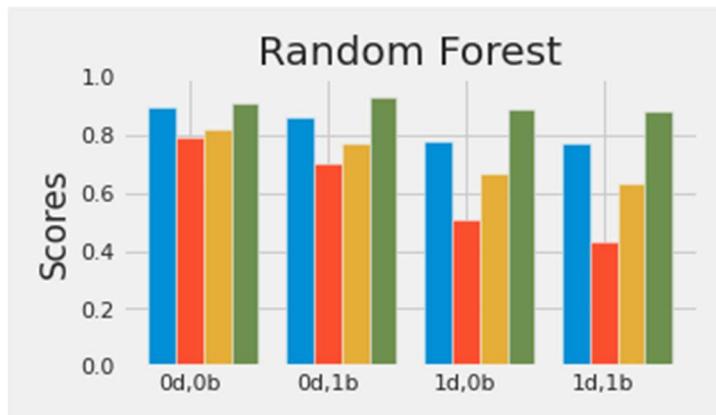
شکل ۱۰-۴: منحنی ROC مربوط به عملکرد مدل ماشین بردار پشتیبان در بهترین حالت (با کاهش بعد ویژگی‌های ورودی و متوازن سازی کلاس‌ها)

شایان ذکر است که در تمامی شبیه‌سازی‌های ذکر شده، از روش padding برای پر کردن جای خالی دادگان ثبت نشده، استفاده شده است. پیش از انتخاب این روش، تمامی روش‌های ذکر شده در بخش ۱-۲-۴ پیاده‌سازی شده و آزمایشات ۱۰۰ بار تکرار شدند. با دقت در میانگین و انحراف معیار نتایج حاصله، مشخص شد که میانگین عملکرد مدل با تغییر روش پر کردن جاهای خالی تغییر چندانی نمی‌کند، اماً انحراف معیار شاخص‌های ارزیابی با اعمال روش padding پایین‌تر بوده که مدل را مقاوم‌تر و قابل اعتماد‌تر می‌کند.

۴-۳-۲-۳- مدل جنگل تصمیم تصادفی

به منظور بررسی عملکرد مدل‌های مبتنی بر یادگیری گروهی، ابتدا آزمایشات بخش قبل بر روی یک مدل پایه‌ی جنگل تصمیم تصادفی تکرار شد. همانگونه که در شکل ۱۱-۴ نشان داده شده است، برخلاف روش‌های رگرسیون لجستیک و ماشین بردار پشتیبان، کاهش بعد ویژگی‌ها با استفاده از روش PCA و همچنین وزن‌دهی به کلاسی که اعضای کمتری دارد، منجر به نتایج ضعیفتری در مقایسه با مدل پایه می‌شود. مقایسه‌ی این

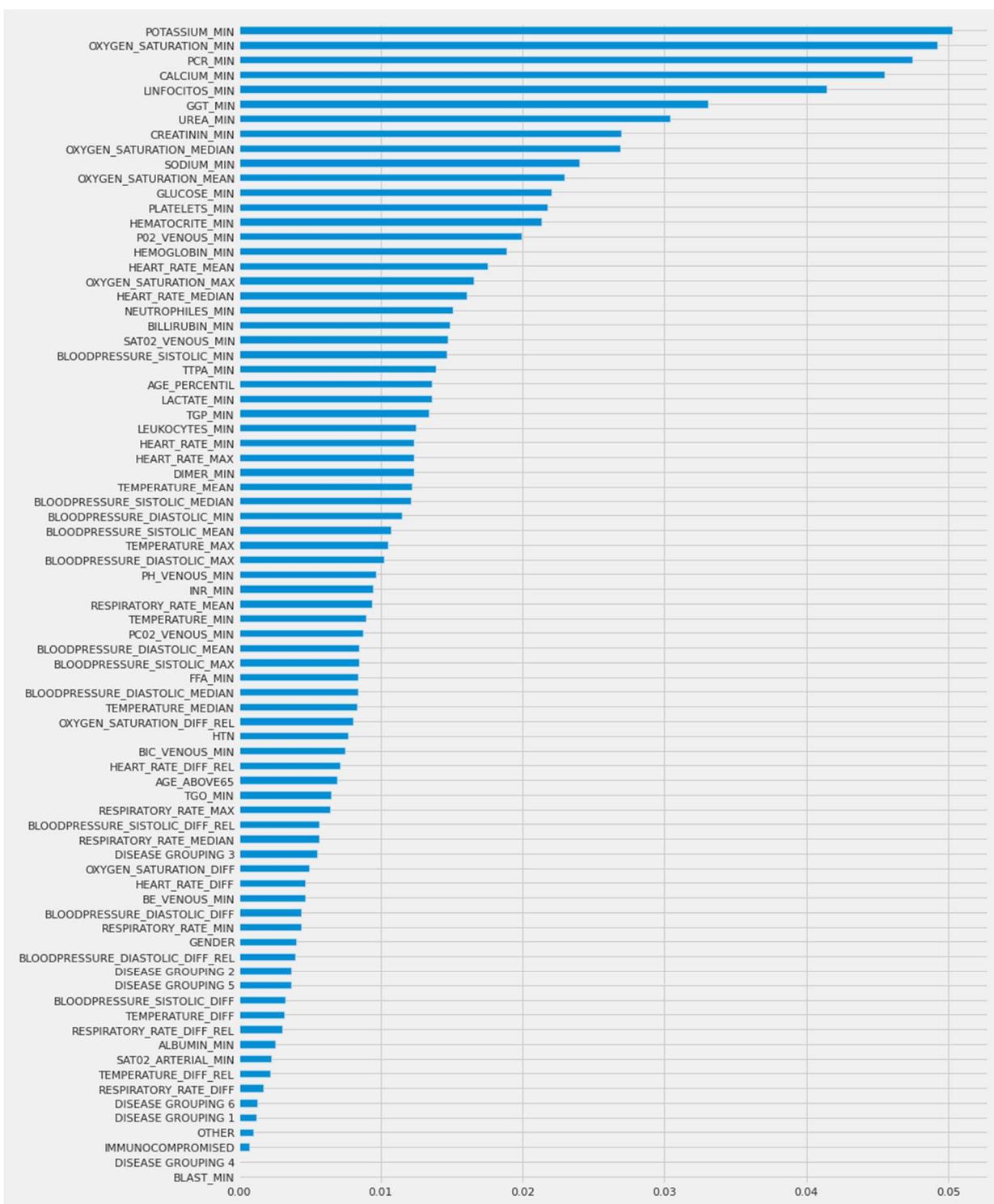
شکل با نتایج نشان داده شده در شکل ۴-۹ به خوبی نشان می‌دهد که این مدل پایه‌ی یادگیری گروهی، حتی بدون تنظیم ابرپارامترها، نتایج بهتری را در مقایسه با دو مدل پیشین به دنبال دارد (دقت، ۰/۹، امتیاز اف ۱، ۰/۷۹، مساحت زیرمنحنی ۱، ۰/۸۲ و مساحت زیرمنحنی ۲، ۰/۹۱).



شکل ۱۱-۴: عملکرد مدل پایه‌ی جنگل تصمیم تصادفی در حالت‌های مختلف با و بدون اعمال کاهش ابعاد ورودی و متوازن‌سازی تعداد نمونه‌های کلاس‌ها. دقت، امتیاز اف ۱ و مساحت زیر منحنی ۱ و ۲ به عنوان معیارهای ارزیابی عملکرد مدل به ترتیب، با رنگ‌های آبی، قرمز، زرد و سبز نشان داده شده‌اند.

در ادامه، همین مدل پایه و بدون کاهش بعد ویژگی‌ها و بدون متوازن‌سازی تعداد اعضای کلاس‌ها برای بررسی‌های بیشتر انتخاب شد. شکل ۱۲-۴ ویژگی‌های ورودی را به ترتیب اثرگذاری بر عملکرد مدل نشان می‌دهد. قابل توجه است که اغلب این ویژگی‌های مهم، مقادیر کمینه‌ی متغیرهای فیزیولوژیکی مانند غلظت پتاسیم، میزان اکسیژن اشباع خون، غلظت کلسیم، تعداد لنفوسيت‌ها و ... هستند. شکل ۱۳-۴ توزیع برخی مقادیر کمینه‌ی متغیرهای فیزیولوژیکی ثبت شده را نشان می‌دهد. به عنوان مثال، با توجه به شکل ۱۲-۴، مهمترین ویژگی در تعیین کلاس هر بردار ورودی کمینه‌ی اکسیژن اشباع خون شناسایی شده است. حال، در شکل ۱۳-۴، توزیع این متغیر دیده می‌شود. سپس، از این ترتیب، برای کاهش بعد ویژگی‌های ورودی استفاده شد و مدل یک بار دیگر و این بار بر روی بردار کاهش یافته‌ی ویژگی‌ها که متشکل از ۸۰٪ ویژگی‌ها به ترتیب تاثیرگذاری آن‌هاست، تعلیم داده شد و ارزیابی گردید. در این حالت، نتایج به صورت دقت، ۰/۸۶، امتیاز اف ۱، ۰/۷۱، مساحت زیر منحنی ۱، ۰/۷۸ و مساحت زیر منحنی ۲، ۰/۹۲ حاصل گردید.

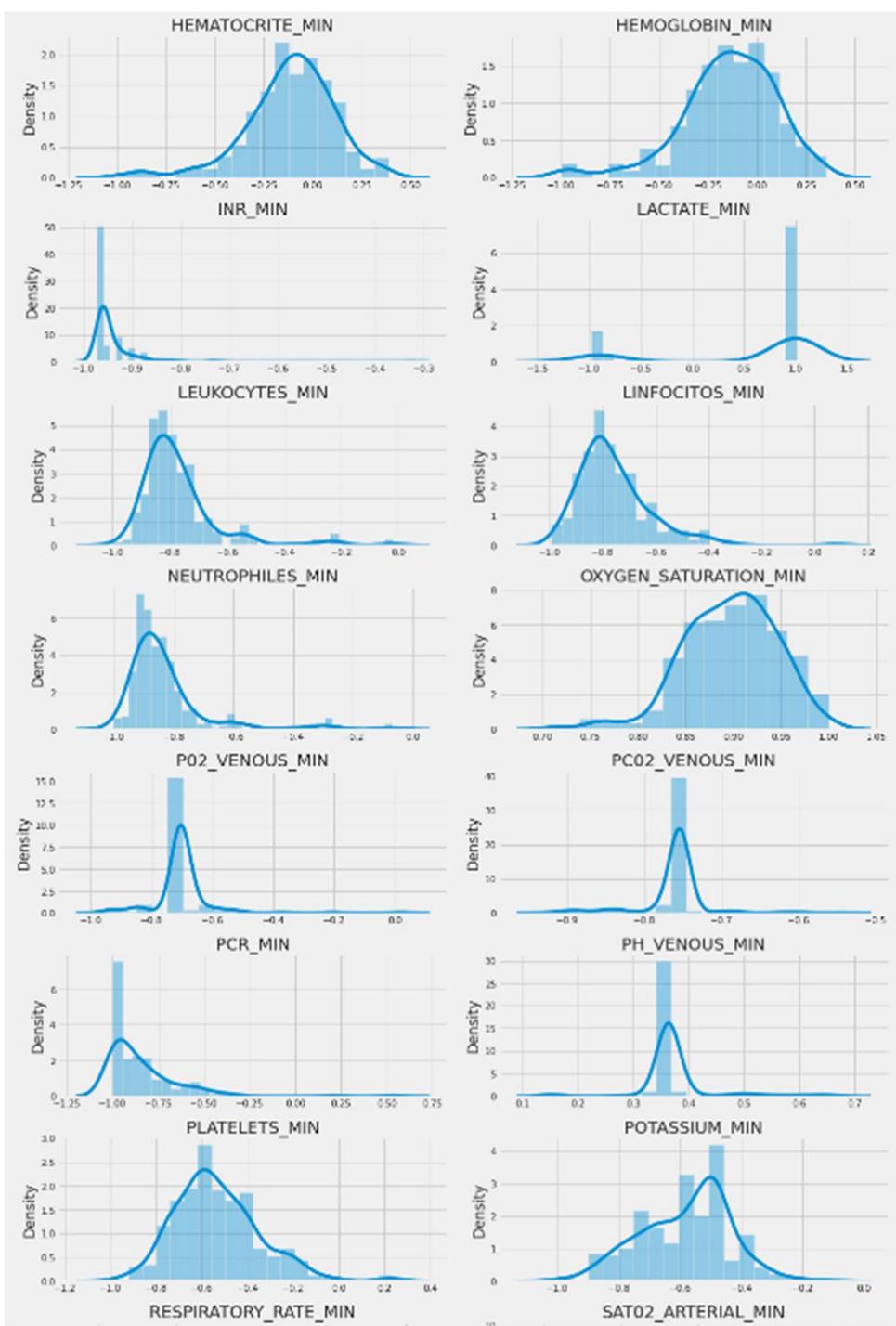
نزدیکی مقادیر این معیارهای ارزیابی به آنچه در بالا گزارش شد، نشان می‌دهد که اغلب ویژگی‌های ورودی در عملکرد مدل تاثیر دارند و ممکن است کاهش ابعاد ورودی، ایده‌ی مناسبی برای بهبود عملکرد مدل نباشد. به ویژه این‌که، همان‌گونه که در بخش ۱-۲-۴ توضیح داده شد، ویژگی‌هایی که همبستگی بالایی با یکدیگر داشته‌اند، از همان ابتدا حذف شده و در این شبیه‌سازی‌ها وارد نشده‌اند.



شکل ۱۲-۴: متغیرهای فیزیولوژیکی که بیشترین تاثیر را در عملکرد مدل جنگل تصمیم تصادفی دارند.

برای اطمینان از این موضوع، روش (RFE) نیز برای استخراج ویژگی‌ها پیاده‌سازی گردید که در نهایت بهبودی در نتایج حاصل نشد (دقت، ۰/۸۸، امتیاز اف ۱، ۰/۷۶، مساحت زیر منحنی ۱، ۰/۸۱ و مساحت زیر منحنی ۲،

۰/۹۲). نکته‌ی جالب توجه در این شبیه‌سازی این بود که ویژگی‌هایی که در شکل ۱۲-۴ به عنوان ویژگی‌های تاثیرگذار نشان داده شده‌اند، در این روش نیز رتبه‌های بهتری به دست آوردن.



شکل ۱۳-۴: توزیع برخی مقادیر کمینه‌ی متغیرهای فیزیولوژیکی ثبت شده.

شبیه‌سازی‌هایی که تاکنون در این بخش گزارش شده‌اند مربوط به مدل پایه‌ی جنگل تصمیم تصادفی بوده که در آن ابرپارامترها تنظیم نشده‌اند. لذا، در ادامه‌ی این تحقیق، فضای زیر به عنوان فضای ابرپارامترها تعریف

و از ترکیب روش سیستماتیک اعتبارسنجی متقابل (با در نظر گرفتن ۱۰ تکرار) با روش Grid Search، برای انتخاب بهترین پارامترها استفاده شد. در این بهینه سازی، افزایش سطح زیر منحنی ROC به عنوان معیار برای انتخاب بهترین مجموعه ابرپارامترها در نظر گرفته شد.

جدول ۴-۳: ابرپارامترهای تنظیم شده در مدل جنگل تصمیم تصادفی

مقادیر انتخابی	نام پارامتر
۵۰۰، ۱۰۰ و ۱۰	تعداد تخمین زننده‌های مدل (تعداد درخت‌های جنگل)
ضریب جینی و آنتروپی	معیار ارزیابی کیفیت فرایند تقسیم یک گره به چندین زیرگره
None و ۱۰ و ۵	عمق درخت تصمیم
مجدور تعداد ویژگی‌ها و لگاریتم تعداد ویژگی‌ها در مبنای ۲	تعداد ویژگی‌های در نظر گرفته شده در ارزیابی فرایند تقسیم یک گره به چندین زیرگره.

پارامترهای بهینه به شرح زیر به دست آمد و نتایج به صورت دقت، $0/87$ ، امتیاز اف $1/73$ و مساحت زیر منحنی $1/79$ و مساحت زیر منحنی $2/93$ حاصل شد:

تعداد تخمین زننده‌های مدل (تعداد درخت‌های جنگل) = 500

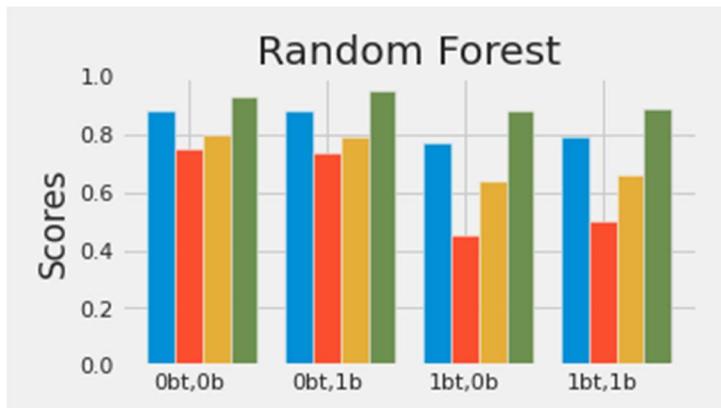
معیار ارزیابی کیفیت فرایند تقسیم یک گره به چندین زیرگره = آنتروپی

تعداد ویژگی‌های در نظر گرفته شده در ارزیابی فرایند تقسیم گره = مجدور تعداد ویژگی‌ها

عمق درخت تصمیم = None (به این معنی که عمق درخت تصمیم محدودیت ندارد و بسته به عملکرد مدل می‌تواند تغییر کند).

در ادامه، اثر اعمال بوت استرپ و همچنین وزن دهنی به کلاس‌ها در هنگام تعلم مدل نیز مطالعه شد. همان‌گونه که در شکل ۱۴-۴ دیده می‌شود، بهترین نتایج در حالتی حاصل شد که نمونه‌های بوت استرپ استفاده نمی‌شوند. وزن دار کردن کلاس‌ها در این حالت، تاثیر چشمگیری بر روی نتایج ندارد. در این حالت، نتایج بهترین عملکرد به قرار زیر حاصل شد:

دقت، $0/88$ ، امتیاز اف $1/75$ ، مساحت زیر منحنی $1/80$ و مساحت زیر منحنی $2/93$.



شکل ۱۴-۴: عملکرد مدل پایه‌ی جنگل تصمیم تصادفی در حالت‌های مختلف با و بدون اعمال بوت‌استرپ (bt) و متوازن‌سازی تعداد نمونه‌های کلاس‌ها (b). دقت، امتیاز اف ۱ و مساحت زیر منحنی ۱ و ۲ به عنوان معیارهای ارزیابی عملکرد مدل به ترتیب، با رنگ‌های آبی، قرمز، زرد و سبز نشان داده شده‌اند. پارامترهای بهینه برای مدل پایه بدون نمونه‌های بوت‌استرپ محاسبه شده‌اند.

البته از آنجا پارامترهای بهینه برای مدل پایه‌ی بدون نمونه‌های بوت‌استرپ محاسبه شده بودند، همین آزمایش، این بار با مدل پایه‌ی مشتمل بر استفاده از نمونه‌های بوت‌استرپ تکرار شد که نتایج آن در شکل ۱۵-۴ نشان داده شده است. در این حالت، پارامترهای بهینه به شرح زیر به دست آمد و نتایج به صورت دقت، ۰/۸۸، امتیاز اف ۱، ۰/۷۴ و مساحت زیر منحنی ۱، ۰/۷۹ و مساحت زیر منحنی ۲، ۰/۹۵ حاصل شد که از بهترین نتایج حالت بدون نمونه‌های بوت‌استرپ بهتر بود:

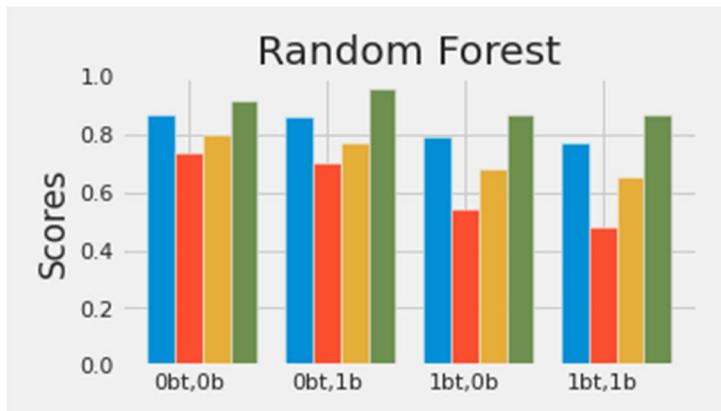
تعداد تخمین‌زننده‌های مدل (تعداد درخت‌های جنگل) = ۱۰۰

معیار ارزیابی کیفیت فرایند تقسیم یک گره به چندین زیرگره = آنتروپی

تعداد ویژگی‌های در نظر گرفته شده در ارزیابی فرایند تقسیم گره = مجدور تعداد ویژگی‌ها

عمق درخت تصمیم = ۱۰

اما همان‌گونه که در شکل ۱۵-۴ نشان داده شده است، پس از چند بار تکرار آزمایش، باز هم در حالت عدم استفاده از نمونه‌های بوت‌استرپ نتایج بهتری حاصل می‌شود.



شکل ۴-۱۵: عملکرد مدل پایه‌ی جنگل تصمیم تصادفی در حالت‌های مختلف با و بدون اعمال بوت استرپ (bt) و متوازن‌سازی تعداد نمونه‌های کلاس‌ها (b). دقت، امتیاز اف ۱ و مساحت زیر منحنی ۱ و ۲ به عنوان معیارهای ارزیابی عملکرد مدل به ترتیب، با رنگ‌های آبی، قرمز، زرد و سبز نشان داده شده‌اند. پارامترهای بهینه برای مدل پایه با در نظر گرفتن نمونه‌های بوت استرپ محاسبه شده‌اند.

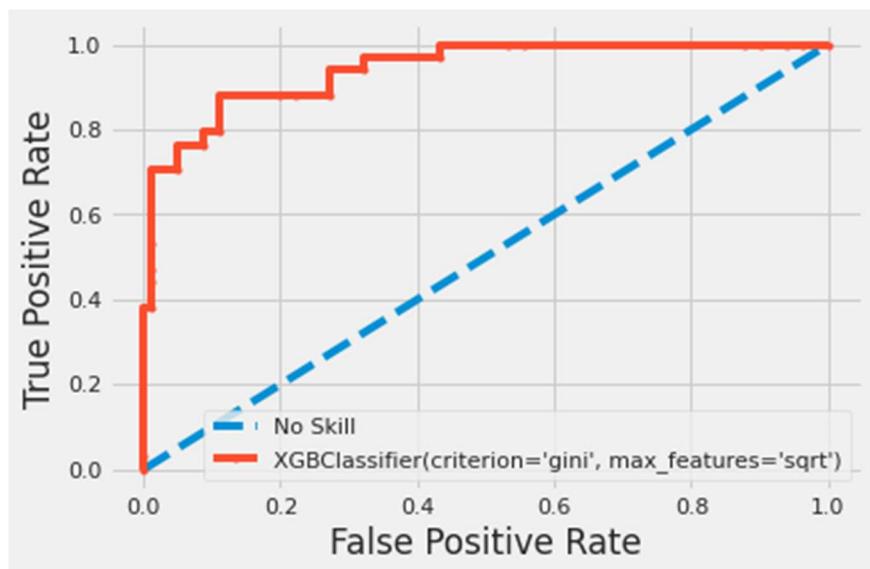
۴-۳-۳- مدل XGBoost

در این بخش، نتایج شبیه‌سازی مدل XGBoost گزارش شده و با نتایج بخش قبل (الگوریتم جنگل تصمیم تصادفی) مقایسه می‌شود.

مدل XGBoost پس از تنظیم پارامترهای بهینه، نتایجی به صورت دقت، $0/90$ ، امتیاز اف $1/81$ و مساحت زیر منحنی $1/85$ و مساحت زیر منحنی $2/94$ به دنبال داشت. پارامترهای بهینه در این مدل به شرح زیر به دست آمده است:

$$\begin{aligned} \text{تعداد تخمین‌زننده‌های مدل (تعداد درخت‌های جنگل)} &= 100 \\ \text{معیار ارزیابی کیفیت فرایند تقسیم یک گره به چندین زیرگره} &= \text{ضریب جینی} \\ \text{تعداد ویژگی‌های در نظر گرفته شده در ارزیابی فرایند تقسیم گره} &= \text{مجدور تعداد ویژگی‌ها} \\ \text{عمق درخت تصمیم} &= 3 \end{aligned}$$

همان‌گونه که در منحنی ROC شکل ۴-۱۶ دیده می‌شود، این مدل در مقایسه با سایر مدل‌ها، بهترین تعادل را بین همهٔ معیارهای ارزیابی عملکرد مدل نشان می‌دهد. شکل ۴-۱۷ نیز نشان می‌دهد که این الگوریتم با وزن‌دار کردن کلاس‌ها بهبود نمی‌یابد و همچنین کاهش بعد ویژگی‌های ورودی به روش PCA، عملکرد آن را بدتر می‌کند.



شکل ۱۶-۴: منحنی ROC مربوط به عملکرد مدل XGBoost

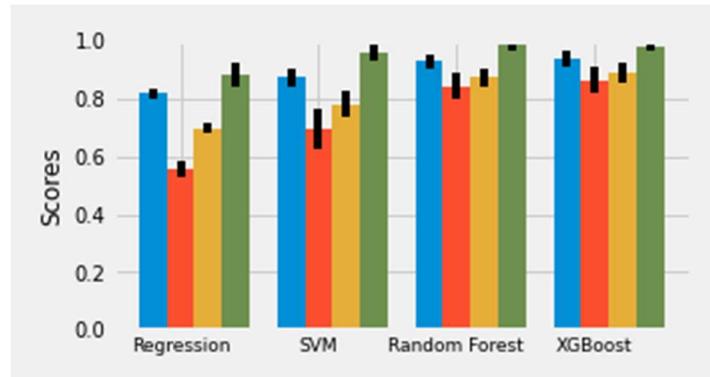


شکل ۱۷-۴: عملکرد مدل XGBoost در حالت‌های مختلف با و بدون اعمال کاهش ابعاد ورودی و متوازن سازی تعداد نمونه‌های کلاس‌ها. دقت، امتیاز اف ۱ و مساحت زیر منحنی ۱ و ۲ به عنوان معیارهای ارزیابی عملکرد مدل به ترتیب، با رنگ‌های آبی، قرمز، زرد و سبز نشان داده شده‌اند.

۴-۴- مقایسه مدل‌ها

در بخش قبل، عملکرد مدل‌های رگرسیون لجستیک، ماشین بردار پشتیبان، جنگل تصمیم تصادفی و XGBoost در طبقه‌بندی دادگان ورودی به دو مجموعه‌ی بستری یا عدم بستری در بخش مراقبت‌های ویژه بررسی شدند. به علاوه، اثر کاهش بعد دادگان ورودی و سایر پارامترها و استراتژی‌های تعلیم نیز مطالعه شدند. در این بخش، تعلیم و ارزیابی این مدل‌ها با در نظر گرفتن مقادیر بهینه‌ی به دست آمده در بخش قبل، ۱۰۰ بار و با

در نظر گرفتن random seed مختلف تکرار شده و نتایج آن در شکل ۱۸-۴ نشان داده شده است. مقادیر انحراف معیار نیز جهت مقایسه رسم شده است.

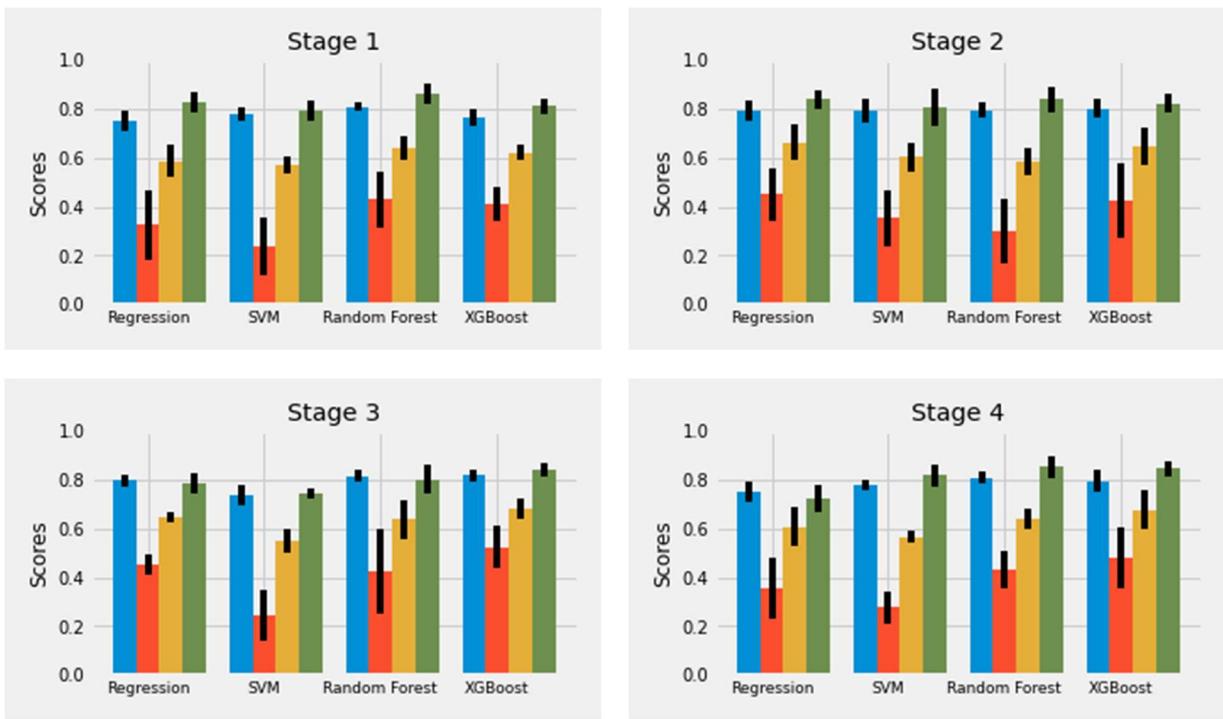


شکل ۱۸-۴: مقایسه‌ی عملکرد مدل‌های بررسی شده در این تحقیق. دقت، امتیاز اف ۱ و مساحت زیر منحنی ۱ و ۲ به عنوان معیارهای ارزیابی عملکرد مدل به ترتیب، با رنگ‌های آبی، قرمز، زرد و سبز نشان داده شده‌اند.

با توجه به اینکه در مطالعات مختلف از مجموعه دادگان متفاوت استفاده شده است، نمی‌توان به طور مستقیم عملکرد مدل‌های پیشنهادی در این تحقیق را با آن‌ها مقایسه نمود. تنها مقاله‌ای که تا کنون روی این مجموعه دادگان متشرشده است [۵۵]، نتایج را به صورت دقت، $0/۹۷$ ، امتیاز اف ۱، $0/۹۶$ ، سطح زیر منحنی ۲، $0/۹۷$ گزارش داده است. در تحقیق مذکور از مدل XGBoost با شرط توقف سریع (early stopping) استفاده شده است. در مقایسه با این مدل، در تحقیق حاضر، پارامترها به گونه‌ای بهینه شده‌اند که مساحت زیر منحنی ۲، بیشینه شود، بنابراین، نتایج حاصل شده به صورت دقت، $0/۹۴$ ، امتیاز اف ۱، $0/۸۶$ و مساحت زیر منحنی ۲، $0/۹۸$ ، برتری مدل پیشنهادی را نسبت به مدل قبلی نشان می‌دهد. هرچند که برای افزایش امتیاز اف ۱ و یا افزایش همزممان این امتیاز و مساحت زیر منحنی ۲، می‌توان پارامترهای مدل را به صورت جداگانه و یا به صورت شبکه‌ای دو بعدی تنظیم و بهینه نمود.

در شبیه‌سازی‌هایی که در بخش‌های قبل گزارش شد، هر اندازه‌گیری به عنوان یک مشاهده در نظر گرفته شد که مجموع تعداد مشاهدات را پس از حذف ردیف‌های تکراری به ۳۸۱ مشاهده می‌رساند.

در ادامه‌ی این تحقیق، همه‌ی مراحل پیشین و این بار فقط با احتساب اولین اندازه‌گیری برای هر بیمار (stage1) و دومین (stage2) الی چهارمین اندازه‌گیری (stage4)، تکرار شده و نتایج در شکل ۱۹-۴ گزارش شده است. تعداد مشاهدات در هر اندازه‌گیری ۱۵۶ نمونه است.



شکل ۱۹-۴: مقایسه‌ی عملکرد مدل‌های بررسی شده در این تحقیق در مراحل مختلف ثبت ویژگی‌های اندازه‌گیری شده. دقت، امتیاز اف ۱ و مساحت زیر منحنی ۱ و ۲ به عنوان معیارهای ارزیابی عملکرد مدل به ترتیب، با رنگ‌های آبی، قرمز، زرد و سبز نشان داده شده‌اند.

نتایج استفاده از متغیرهای اندازه‌گیری شده در اولین ثبت را می‌توان با نتایج گزارش شده در [۵۵] مقایسه نمود (دقت، ۰/۷۳، امتیاز اف ۱، ۰/۷۲ و مساحت زیر منحنی ۲، ۰/۷۳). مدل پیشنهادی پژوهش حاضر با دقت ۰/۸۶ و مساحت زیر منحنی ۲ برابر با ۰/۸۸ قادر به تشخیص نیاز به بستری در بخش مراقبت‌های ویژه پس از اولین اندازه‌گیری است. همان‌گونه که در شکل ۱۹-۴ نیز نشان داده شده، با پیش رفتن مرحله‌ی اندازه‌گیری، عملکرد مدل، به ویژه با معیار مساحت زیر منحنی ۲، بهتر می‌شود که نشان می‌دهد، تفاوت ویژگی‌های اندازه‌گیری شده در گروه نیازمند به بستری و گروه بی نیاز از بستری در بخش مراقبت‌های ویژه محسوس‌تر است. اما، باید توجه داشت که مساحت زیر منحنی ۲، در مرحله‌های ۳ و ۴ تفاوت چندانی نشان نمی‌دهند. علت این رفتار، نقص روش جمع‌آوری دادگان است که به ازای هیچ بیماری، هر دو اندازه‌گیری موجود نیست و معمولاً جای خالی اندازه‌گیری ثبت نشده با مقادیر اندازه‌گیری شده در مرحله‌ی قبل (یا بعد) پر شده است. از آنجا که در [۵۵] با پیشرفت مراحل، از اطلاعات ثبت شده در مراحل قبل نیز استفاده می‌شود، نمی‌توان نتایج تحقیق حاضر را در سایر مراحل، با آن مقایسه نمود.

فصل ۵: بحث و نتیجه گیری

۱-۵ - مقدمه

در زمان تعریف این پروژه، بیماری کووید-۱۹ اثرات منفی جبران ناپذیری را بر زندگی میلیون‌ها نفر در سراسر گذاشته بود و ارایه روش‌های سریع و ارزان برای تشخیص افراد مبتلا، تشخیص احتمال نیاز به بستری شدن در بخش مراقبت‌های ویژه‌ی بیمارستان و همچنین تشخیص احتمال نجات فرد مبتلا به عنوان یکی از مهمترین و ضروری‌ترین اولویت‌های تحقیقاتی در بخش داده‌کاوی بالینی در نظر گرفته می‌شد. دو سال پس از آغاز همه‌گیری، پس از بیش از ۲۶۰ میلیون مورد بیماری، بیش از ۵ میلیون مرگ و اثرات منفی اقتصادی، اجتماعی و روانی، در تاریخ ۴ آذر ماه ۱۴۰۰ سازمان بهداشت جهانی اولین مورد از سویه اوامیکرون را تایید کرد. امروز و با گذشت بیش از سه سال از شروع این همه‌گیری جهانی و پس از پیدایش سویه‌های مختلف ویروس و همچنین خطر بروز نمونه‌های مشابه در آینده، پیش‌بینی هزینه‌هایی که بیماری‌های عفونی بر سیستم بهداشت و درمان تحمیل می‌کنند در بهبود برنامه‌های آگاهی‌رسانی عمومی و همچنین مدیریت ظرفیت‌های درمانی و نگهداری مبتلایان موثر واقع شود. همچنان جان بسیاری از مردم، به ویژه سالمندان، در خطر بوده و یافتن راه‌های مؤثر برای تشخیص زودهنگام این بیماری در افراد، از اولویت‌های تحقیقاتی به شمار می‌رود. از منظر هزینه‌های مراقبتی و برنامه‌ریزی‌های کلان بیمارستانی، پیش‌بینی نیاز افراد به بستری شدن در بخش مراقبت‌های ویژه‌ی بیمارستان نیز حائز اهمیت بوده و توجه محققان را به خود معطوف نموده است. لذا، در این تحقیق پیشنهاد شده است که برای تشخیص سریع این‌که کدامیک از مبتلایان ممکن است دچار علایم مراحل حاد این بیماری و نیازمند به مراقبت‌های ویژه شوند روش‌های مبتنی بر یادگیری ماشین به کار گرفته

شوند. در این تحقیق، همچنان نشان داده شد که می‌توان قدرت پیش‌بینی‌کنندگی این روش‌ها را از طریق ترکیب چندین مدل و ساخت یک مدل یادگیری گروهی بهبود بخشد. به طور ویژه، بر روی داده‌های آزمایشگاهی جمع آوری شده از ۳۸۴ مراجعه به بیمارستان سیریولبانز در سائوپلو نشان داده شد که مدل‌های جنگل تصادفی و XGBoost به عنوان مثال‌هایی از الگوریتم‌های یادگیری گروهی می‌توانند با دقت، ۰/۹۴، امتیاز اف ۱، ۰/۸۶ و مساحت زیر منحنی، ۰/۹۸ احتمال نیاز به بستری در بخش مراقبت‌های ویژه را پیش‌بینی نمایند. همچنان نشان داده شد که اگر فقط اطلاعات اولین مراجعه‌ی هر بیمار مبنای تصمیم‌گیری قرار گیرد، مدل XGBoost توسعه یافته در تحقیق حاضر می‌تواند با دقت ۰/۸۶ و مساحت زیر منحنی برابر با ۰/۸۸ نیاز به بستری در بخش مراقبت‌های ویژه را پیش‌بینی کند که این نتایج در مقایسه با مدل پیشنهادی [۵۵] (یعنی دقت، ۰/۷۳، امتیاز اف ۱، ۰/۷۲ و مساحت زیر منحنی، ۰/۷۳) بهبود قابل توجهی را نشان می‌دهد.

۲-۵ مزایای روش پیشنهادی

در این پژوهش از روش‌های یادگیری گروهی مانند الگوریتم جنگل تصادفی و XGBoost استفاده شده که در آن‌ها به جای تعلیم یک مدل پیش‌بینی کننده‌ی واحد، مجموعه‌ای از یادگیرنده‌ها تعلیم می‌یابند. لذا، اولین مزیت این روش‌های ترکیبی، عملکرد بهتر آن‌ها مخصوصاً در حل مساله‌ی طبقه‌بندی است. دومین مزیت این روش‌ها، مقاومت آن‌ها نسبت به تغییر شرایط تعلیم و ارزیابی، مانند تغییر random seed است. به این صورت که با تغییر شرایط شبیه‌سازی، واریانس پیش‌بینی‌ها و معیارهای ارزیابی عملکرد مدل کمتر از مدل‌های تک یادگیرنده است. به علاوه، الگوریتمی مانند XGBoost، به گونه‌ای طراحی و بهینه‌سازی شده است که با ایجاد امکان پردازش موازی و همچنان مدیریت دسترسی به حافظه، پیچیدگی و زمان محاسبه را نیز کم می‌کند. همچنان، با تنظیم پارامترهای این مدل و مخصوصاً تعیین نحوه‌ی انتخاب و تقسیم دادگان تعلیم به زیرمجموعه‌هایی که از دادگان، به خوبی می‌توان احتمال بیش‌برازش مدل بر روی دادگان تعلیم را به صفر رساند. این روش، الگوریتم‌هایی را برای فرایند تقسیم یک گره به چندین زیرگره (در درخت تصمیم به عنوان مدل یادگیرنده) ارائه می‌کند که ۱) داده‌های خلوت ۸۷ را با جهت‌های پیش‌فرض گره‌ها مدیریت می‌کند، ۲) داده‌های وزن‌دار را با استفاده از عملیات ادغام و هرس آدرس‌دهی می‌کند، ۳) به طور مؤثر بر روی تمام تقسیم‌های ممکن شمارش می‌کند تا آستانه‌ی تقسیم بهینه شود. علاوه بر این، همانگونه که در نتایج فصل ۴ بحث شد، این مدل، نسبت به نحوه‌ی پر کردن جاهای خالی در دادگان حساس نیست. هر دو مدل الگوریتم جنگل تصادفی و XGBoost که در این تحقیق به کار گرفته شده‌اند، نیازی به نرمال‌سازی دادگان ورودی و الگوریتم‌های پیش‌پردازش اضافی ندارند.

۵- نتایج تحقیق

در این تحقیق، یک مدل پیش‌بینی کننده طراحی و توسعه یافته که با دریافت اطلاعات کلینیکی و آزمایشگاهی بیماران بیماران مبتلا به کووید-۱۹، نیاز یا عدم نیاز به بستری آنها در بخش مراقبت‌های ویژه را پیش‌بینی می‌کند. اگرچه این پژوهش به طور خاص برای کووید-۱۹ طراحی و پیاده‌سازی شده است، می‌توان نتایج و همچنین مراحل پیاده‌سازی آن را برای سایر بیماری‌های ویروسی از خانواده‌ی ویروس کرونا نیز به کار برد. در این تحقیق از مجموعه دادگانی استفاده شده که ۴ بار و با فواصل زمانی ۱۲ ساعت از بیماران بستری شده در بیمارستان، ثبت شده‌اند. یکی از مهمترین یافته‌های این تحقیق، این است که فقط با داشتن اولین اندازه گیری، یعنی بلافارسله پس از مراجعت به بیمارستان، می‌توان با دقت خوبی پیش‌بینی کرد که آیا این بیمار نیاز به بستری شدن در بخش مراقبت‌های ویژه در ساعت پیش‌رو را پیدا خواهد کرد یا خیر. همچنین نشان داده شد که با افزایش تعداد دفعات اندازه گیری، دقت پیش‌بینی بهبود می‌یابد. از دیدگاه روش‌شناسی نیز نشان داده شد که در مقایسه با روش‌های تک یادگیرنده مانند رگرسیون لجستیک و ماشین بردار پشتیبان، روش‌های یادگیری گروهی مانند الگوریتم جنگل تصادفی و XGBoost، نه تنها عملکرد بهتری دارند، بلکه واریانس کمتری را در ۱۰۰ تکرار با random seed متفاوت نشان می‌دهند.

به علاوه، در فصل ۴ این تحقیق، فهرستی از متغیرهای فیزیولوژیک ارایه شده است که بیشترین تاثیر را در پیش‌بینی نیاز به بستری شدن در بخش مراقبت‌های ویژه دارند. اغلب این متغیرها، کمینه‌ی بیومارکرهای خونی هستند که از میان آنها می‌توان به کمینه‌ی غلظت پتاسیم، میزان اکسیژن اشباع خون، غلظت کلسیم، تعداد لنفوцит‌ها و غلظت کراتینین اشاره کرد. با توجه به این فهرست، ممکن است بتوان با اجتناب از ثبت متغیرهای فیزیولوژیک کم تاثیر در تصمیم‌گیری، در برخی هزینه‌های بیمارستانی و آزمایشگاهی صرفه‌جویی نمود.

۶- کارهای آینده

مجموعه دادگان نسبتاً کوچک و البته غیر بومی را می‌توان به عنوان مهمترین محدودیت این تحقیق معرفی کرد. در صورت دسترسی به دادگان جمع آوری شده در یکی از بیمارستان‌های ایران، می‌توان از این دادگان به عنوان مجموعه‌ی دادگان out-of-distribution استفاده نمود و تعییم پذیری مدل پیشنهادی را ارزیابی کرد. به علاوه، در مجموعه دادگان فعلی، برخی از متغیرهای مهم آزمایشگاهی مانند غلظت فریتین و پروتئین واکنش سی اندازه گیری نشده‌اند که در تحقیقات قبلی به ویژگی‌های تاثیرگذار در تشخیص کامپیوتری کووید-۱۹ معرفی شده‌اند.

به غیر از توسعه و بهبود مجموعه‌ی دادگان، می‌توان مدل را به گونه‌ای طراحی نمود که از توالی زمانی ویژگی‌ها نیز در پیش‌بینی برچسب‌های خروجی استفاده کند. در این تحقیق، به دلیل ناکافی و تکراری بودن متغیرهای فیزیولوژیک ثبت شده در زمان‌های مختلف، تحقق این امر امکان‌پذیر نبود.

منابع:

- [1] D. Ardila *et al.*, “End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography,” *Nature medicine*, vol. 25, no. 6, pp. 954–961, 2019.
- [2] Y. Liu *et al.*, “A deep learning system for differential diagnosis of skin diseases,” *Nature medicine*, vol. 26, no. 6, pp. 900–908, 2020.
- [3] R. Gargeya and T. Leng, “Automated identification of diabetic retinopathy using deep learning,” *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017.
- [4] G. Litjens *et al.*, “State-of-the-art deep learning in cardiovascular image analysis,” *JACC: Cardiovascular imaging*, vol. 12, no. 8 Part 1, pp. 1549–1565, 2019.
- [5] S. Saadatnejad, M. Oveisí, and M. Hashemi, “LSTM-based ECG classification for continuous monitoring on personal wearable devices,” *IEEE journal of biomedical and health informatics*, vol. 24, no. 2, pp. 515–523, 2019.
- [6] R. Zhu, X. Tu, and J. X. Huang, “Utilizing BERT for biomedical and clinical text mining,” in *Data Analytics in Biomedical Engineering and Healthcare*, Elsevier, 2021, pp. 73–103.
- [7] G. Petmezas *et al.*, “Automated atrial fibrillation detection using a hybrid CNN-LSTM network on imbalanced ECG datasets,” *Biomedical Signal Processing and Control*, vol. 63, p. 102194, 2021.
- [8] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, “Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals,” *Computers in biology and medicine*, vol. 100, pp. 270–278, 2018.
- [9] S. S. A. Karim and Q. A. Karim, “Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic,” *The Lancet*, vol. 398, no. 10317, pp. 2126–2128, 2021.
- [10] B. Pfizer, *Pfizer and BioNTech provide update on omicron variant*. 2022.
- [11] “WHO forecasts coronavirus pandemic will end in 2022,” *POLITICO*, Dec. 22, 2021. <https://www.politico.eu/article/who-forecasts-coronavirus-pandemic-will-end-in-2022/> (accessed Jan. 04, 2023).

- [12] E. Mahase, *Coronavirus: covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate*. British Medical Journal Publishing Group, 2020.
- [13] R. Lu *et al.*, “Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding,” *The lancet*, vol. 395, no. 10224, pp. 565–574, 2020.
- [14] S. J. Fong, N. Dey, and J. Chaki, “An introduction to COVID-19,” in *Artificial intelligence for coronavirus outbreak*, Springer, 2021, pp. 1–22.
- [15] G. Pascarella *et al.*, “COVID-19 diagnosis and management: a comprehensive review,” *Journal of internal medicine*, vol. 288, no. 2, pp. 192–206, 2020.
- [16] “WHO Director-General’s remarks at the media briefing on 2019-nCoV on 11 February 2020.” <https://www.who.int/director-general/speeches/detail/who-director-general-s-remarks-at-the-media-briefing-on-2019-ncov-on-11-february-2020> (accessed Jan. 11, 2023).
- [17] S. A. Lauer *et al.*, “The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application,” *Annals of internal medicine*, vol. 172, no. 9, pp. 577–582, 2020.
- [18] G. Kampf, D. Todt, S. Pfaender, and E. Steinmann, “Persistence of coronaviruses on inanimate surfaces and their inactivation with biocidal agents,” *Journal of hospital infection*, vol. 104, no. 3, pp. 246–251, 2020.
- [19] W. Zhang *et al.*, “Molecular and serological investigation of 2019-nCoV infected patients: implication of multiple shedding routes,” *Emerging microbes & infections*, vol. 9, no. 1, pp. 386–389, 2020.
- [20] W. Guan *et al.*, “Clinical characteristics of coronavirus disease 2019 in China,” *New England journal of medicine*, vol. 382, no. 18, pp. 1708–1720, 2020.
- [21] H. Yun, Z. Sun, J. Wu, A. Tang, M. Hu, and Z. Xiang, “Laboratory data analysis of novel coronavirus (COVID-19) screening in 2510 patients,” *Clinica Chimica Acta*, vol. 507, pp. 94–97, 2020.
- [22] B. E. Young *et al.*, “Epidemiologic features and clinical course of patients infected with SARS-CoV-2 in Singapore,” *Jama*, vol. 323, no. 15, pp. 1488–1494, 2020.
- [23] Q. Ruan, K. Yang, W. Wang, L. Jiang, and J. Song, “Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China,” *Intensive care medicine*, vol. 46, no. 5, pp. 846–848, 2020.
- [24] E. Driggin *et al.*, “Cardiovascular considerations for patients, health care workers, and health systems during the COVID-19 pandemic,” *Journal of the American College of cardiology*, vol. 75, no. 18, pp. 2352–2371, 2020.
- [25] E. A. Akl *et al.*, “Use of chest imaging in the diagnosis and management of COVID-19: a WHO rapid advice guide,” *Radiology*, vol. 298, no. 2, pp. E63–E69, 2021.
- [26] W. Alsharif and A. Qurashi, “Effectiveness of COVID-19 diagnosis and management tools: A review,” *Radiography*, vol. 27, no. 2, pp. 682–687, 2021.
- [27] A. K. Das, S. Ghosh, S. Thunder, R. Dutta, S. Agarwal, and A. Chakrabarti, “Automatic COVID-19 detection from X-ray images using ensemble learning with convolutional neural network,” *Pattern Analysis and Applications*, vol. 24, no. 3, pp. 1111–1124, 2021.
- [28] M. Ilyas, H. Rehman, and A. Naït-Ali, “Detection of covid-19 from chest x-ray images using artificial intelligence: An early review,” *arXiv preprint arXiv:2004.05436*, 2020.

- [29] G. D. Rubin *et al.*, “The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society,” *Radiology*, vol. 296, no. 1, pp. 172–180, 2020.
- [30] N. Islam *et al.*, “Thoracic imaging tests for the diagnosis of COVID-19,” *Cochrane Database of Systematic Reviews*, no. 3, 2021.
- [31] A. Kumar, P. K. Gupta, and A. Srivastava, “A review of modern technologies for tackling COVID-19 pandemic,” *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 4, pp. 569–573, 2020.
- [32] A. A. Ardakani, A. R. Kanafi, U. R. Acharya, N. Khadem, and A. Mohammadi, “Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks,” *Computers in biology and medicine*, vol. 121, p. 103795, 2020.
- [33] W. Yang *et al.*, “The role of imaging in 2019 novel coronavirus pneumonia (COVID-19),” *European radiology*, vol. 30, pp. 4874–4882, 2020.
- [34] L. Wang *et al.*, “Real-time estimation and prediction of mortality caused by COVID-19 with patient information based algorithm,” *Science of the total environment*, vol. 727, p. 138394, 2020.
- [35] S. Cui, Y. Wang, D. Wang, Q. Sai, Z. Huang, and T. C. E. Cheng, “A two-layer nested heterogeneous ensemble learning predictive method for COVID-19 mortality,” *Applied Soft Computing*, vol. 113, p. 107946, 2021.
- [36] A. S. Albahri *et al.*, “Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review,” *Journal of medical systems*, vol. 44, pp. 1–11, 2020.
- [37] L. J. Muhammad, M. M. Islam, S. S. Usman, and S. I. Ayon, “Predictive data mining models for novel coronavirus (COVID-19) infected patients’ recovery,” *SN Computer Science*, vol. 1, no. 4, p. 206, 2020.
- [38] J. Sarkar and P. Chakrabarti, “A machine learning model reveals older age and delayed hospitalization as predictors of mortality in patients with COVID-19,” *MedRxiv*, p. 2020.03. 25.20043331, 2020.
- [39] O. Sagi and L. Rokach, “Ensemble learning: A survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.
- [40] V. Nikulin, G. J. McLachlan, and S. K. Ng, “Ensemble approach for the classification of imbalanced data,” in *AI 2009: Advances in Artificial Intelligence: 22nd Australasian Joint Conference, Melbourne, Australia, December 1-4, 2009. Proceedings 22*, Springer, 2009, pp. 291–300.
- [41] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, “A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, 2011.
- [42] H. Deng, G. Runger, E. Tuv, and M. Vladimir, “A time series forest for classification and feature extraction,” *Information Sciences*, vol. 239, pp. 142–153, 2013.
- [43] N. V. Chawla, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, “Learning ensembles from bites: A scalable and accurate approach,” *The Journal of Machine Learning Research*, vol. 5, pp. 421–451, 2004.
- [44] L. Rokach, “Genetic algorithm-based feature set partitioning for classification problems,” *Pattern Recognition*, vol. 41, no. 5, pp. 1676–1700, 2008.
- [45] D. H. Wolpert, “Stacked generalization,” *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

- [46] Y. Freund and R. E. Schapire, “Experiments with a new boosting algorithm,” in *icml*, Citeseer, 1996, pp. 148–156.
- [47] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of statistics*, pp. 1189–1232, 2001.
- [48] J. H. Friedman, “Stochastic gradient boosting,” *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.
- [49] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
- [50] S. M. Abate, S. Ahmed Ali, B. Mantfardo, and B. Basu, “Rate of Intensive Care Unit admission and outcomes among patients with coronavirus: A systematic review and Meta-analysis,” *PloS one*, vol. 15, no. 7, p. e0235653, 2020.
- [51] Z. A. A. Alyasseri *et al.*, “Review on COVID-19 diagnosis models based on machine learning and deep learning approaches,” *Expert systems*, vol. 39, no. 3, p. e12759, 2022.
- [52] F. Shi *et al.*, “Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19,” *IEEE reviews in biomedical engineering*, vol. 14, pp. 4–15, 2020.
- [53] X. Jiang *et al.*, “Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity,” *Computers, Materials & Continua*, vol. 63, no. 1, pp. 537–551, 2020.
- [54] F. S. Heldt *et al.*, “Early risk assessment for COVID-19 patients from emergency department data using machine learning,” *Scientific reports*, vol. 11, no. 1, pp. 1–13, 2021.
- [55] M. Ezz, M. K. Elbashir, and H. Shabana, “Predicting the need for icu admission in covid-19 patients using xgboost,” *Computers, Materials and Continua*, pp. 2077–2092, 2021.
- [56] R. Aznar-Gimeno *et al.*, “A clinical decision web to predict ICU admission or death for patients hospitalised with COVID-19 using machine learning algorithms,” *International Journal of Environmental Research and Public Health*, vol. 18, no. 16, p. 8677, 2021.
- [57] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” *Artificial intelligence*, vol. 97, no. 1–2, pp. 273–324, 1997.
- [58] F. J. Ferri, P. Pudil, M. Hatef, and J. Kittler, “Comparative study of techniques for large-scale feature selection,” in *Machine intelligence and pattern recognition*, Elsevier, 1994, pp. 403–413.
- [59] A. U. Haq, D. Zhang, H. Peng, and S. U. Rahman, “Combining multiple feature-ranking techniques and clustering of variables for feature selection,” *Ieee Access*, vol. 7, pp. 151482–151492, 2019.
- [60] A. Vaid *et al.*, “Machine learning to predict mortality and critical events in a cohort of patients with COVID-19 in New York City: model development and validation,” *Journal of medical Internet research*, vol. 22, no. 11, p. e24018, 2020.
- [61] L. Yan *et al.*, “A machine learning-based model for survival prediction in patients with severe COVID-19 infection,” 2020.
- [62] E. Rechtman, P. Curtin, E. Navarro, S. Nirenberg, and M. K. Horton, “Vital signs assessed in initial clinical encounters predict COVID-19 mortality in an NYC hospital system,” *Scientific reports*, vol. 10, no. 1, pp. 1–6, 2020.
- [63] D. Bertsimas *et al.*, “COVID-19 mortality risk assessment: An international multi-center study,” *PloS one*, vol. 15, no. 12, p. e0243262, 2020.

- [64] X. Guan *et al.*, “Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study,” *Annals of Medicine*, vol. 53, no. 1, pp. 257–266, 2021.
- [65] A. L. Booth, E. Abels, and P. McCaffrey, “Development of a prognostic model for mortality in COVID-19 infection using machine learning,” *Modern Pathology*, vol. 34, no. 3, pp. 522–531, 2021.
- [66] L. Sun *et al.*, “Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19,” *Journal of Clinical Virology*, vol. 128, p. 104431, 2020.
- [67] H. Yao *et al.*, “Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests,” *Frontiers in cell and developmental biology*, p. 683, 2020.
- [68] C. Hu *et al.*, “Early prediction of mortality risk among patients with severe COVID-19, using machine learning,” *International journal of epidemiology*, vol. 49, no. 6, pp. 1918–1929, 2020.
- [69] D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi, and F. Cabitza, “Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study,” *Journal of medical systems*, vol. 44, pp. 1–12, 2020.
- [70] S. Subudhi *et al.*, “Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19,” *NPJ digital medicine*, vol. 4, no. 1, p. 87, 2021.
- [71] L. Famiglini, A. Campagner, A. Carobene, and F. Cabitza, “A robust and parsimonious machine learning method to predict ICU admission of COVID-19 patients,” *Medical & Biological Engineering & Computing*, pp. 1–13, 2022.
- [72] T. W. Campbell *et al.*, “Predicting prognosis in COVID-19 patients using machine learning and readily available clinical data,” *International journal of medical informatics*, vol. 155, p. 104594, 2021.
- [73] F.-Y. Cheng *et al.*, “Using machine learning to predict ICU transfer in hospitalized COVID-19 patients,” *Journal of clinical medicine*, vol. 9, no. 6, p. 1668, 2020.
- [74] F. T. Fernandes, T. A. de Oliveira, C. E. Teixeira, A. F. de M. Batista, G. Dalla Costa, and A. D. P. Chiavegatto Filho, “A multipurpose machine learning approach to predict COVID-19 negative prognosis in São Paulo, Brazil,” *Scientific reports*, vol. 11, no. 1, p. 3343, 2021.
- [75] G. Wu *et al.*, “Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: an international multicentre study,” *European Respiratory Journal*, vol. 56, no. 2, 2020.
- [76] S. Kullback and R. A. Leibler, “On information and sufficiency,” *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [77] I. T. Jolliffe, *Principal component analysis for special types of data*. Springer, 2002.
- [78] G. E. Hinton and S. Roweis, “Stochastic neighbor embedding,” *Advances in neural information processing systems*, vol. 15, 2002.
- [79] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Machine learning*, vol. 46, pp. 389–422, 2002.
- [80] L. Breiman, “Random forests,” *Machine learning*, vol. 45, pp. 5–32, 2001.
- [81] “COVID-19 - Clinical Data to assess diagnosis.” <https://www.kaggle.com/datasets/Sírio-Libanes/covid19> (accessed Dec. 24, 2021).
- [82] L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE.,” *Journal of machine learning research*, vol. 9, no. 11, 2008.

$\forall \xi$