مقاله اصیل

مدل‌سازی تاثیر برخی از متغیرها بر شدت کووید ۱۹ با الگوریتم CART در دانشگاه علوم پزشکی مشهد

مصطفی بسکابادی *[۱]، منور افضل آقایی[۲]، نسرین تلخی [۱]، زهرا جمالیان بهمن جانسفلی [۲]، احسان موسی فرخانی[۲]، حبیب الله اسماعیلی [۱]

[۱] گروه آمارزیستی، دانشکده بهداشت، دانشگاه علوم پزشکی مشهد، مشهد، ایران.
[۲] گروه اپیدمیولوژی، دانشکده بهداشت، دانشگاه علوم پزشکی مشهد، مشهد، ایران.
* نویسنده مسول: مصطفی بسکابادی، گروه آمار، دانشکده بهداشت، دانشگاه علوم پزشکی مشهد، مشهد، ایران Boskabadym1@mums.ac.ir .

دریافت: شهریور ۱۴۰۱؛ **پذیرش:** ۱۴۰۱ آبان

چکیده

مقدمه: با توجه به اینکه هنوز ویروس جدید کرونا (کووید ۱۹) شیوع دارد یکی از دغدغه‌های مهم در سلامت جامعه، عوامل تاثیرگذار بر شدت این بیماری است. در این مطالعه، با استفاده از الگوریتم CART ، تعیین الگو وضعیت شدت بیماران مبتلا به ویروس کووید ۱۹ در جمعیت تحت پوشش دانشگاه علوم پزشکی مشهد مورد بررسی قرار می‌گیرد. **روش مطالعه:** این پژوهش از نوع مقطعی-تحلیلی می‌باشد. در این مطالعه، تمام افراد مراجعه کرده بابت بیماری کووید ۱۹ که در پیک دوم بیماری و پیک چهارم دارای پرونده الکترونیک سلامت فعال در سامانه سینا می‌باشند وارد مطالعه می‌شوند. تجزیه و تحلیل داده ها با استفاده از نرم افزار آماری JMP نسخه ۱۳ انجام شده است. در بخش مدل سازی از روش‌های داده‌کاوی و الگوریتم CART استفاده می‌شود. **یافته ها:** نتایج توصیفی مطالعه به ما نشان داد ۶ درصد بیماران دارای PCR مثبت، مبتلا به شکل شدید بیماری کرونا می‌باشند. متغیر سن اهمیت زیادی در شدت بیماری داشت و نقطه بحرانی برای شدت بیماری، سن ۶۰ سال بوده بطوریکه شدت بیماری را از حدود ۳درصد زیر ۶۰ سال به حدود ۱۸ درصد بالای ۶۰ سال افزایش می‌دهد. متغیرهای بیماری قلبی، کلیوی، تنفسی، چربی خون، دیابت دیگر متغیرهای با اهمیت بوده اند. **نتیجه گیری:** نتایج مدل رگرسیون و طبقه بندی درختی نشان می‌دهد که از بین متغیرهای برای سن زیر ۶۰ سال به ترتیب اهمیت بیماری قلبی، سن، دیابت، بیماری تنفسی و چربی خون و برای سن بالای ۶۰ سال به ترتیب اهمیت سن، بیماری قلبی، بیماری کلیوی، بیماری تنفسی و دیابت موثر می‌باشند. با توجه به سطح زیر منحنی ROC مدل برازش داده شده عملکرد خوبی برای داده‌های بیماری کرونا دارد بطوریکه پیش‌بینی شدت بیماری را تا ۶ برابر افزایش می‌دهد.

کلمات کلیدی: کووید-۱۹، الگوریتم CART ، منحني راک، پیش بیني

۱. مقدمه

در ۳۱ دسامبر ۲۰۱۹ ، موارد متعددي از پنومونيي با اتيولوژي نامشخص، در ووهان چين گزارش شد. اين بيماري پس از مشاهده در چين به سرعت در سطح جهان گسترش يافت و بصورت پاندمی درآمد (۱). طبق گزارش سازمان مرکز مدیریت بیماریها، شیوع بیماری در جهان (تا تاریخ ۳۰ خرداد ماه ۱۴۰۱)، بیش از ۵۴۴ میلیون و در ایران بیش از ۷ میلیون نفرگزارش شده است. موارد مرگ و میر در جهان بیش از ۶ میلیون نفر بوده که در ایران این رقم به بیش از ۱۴۱ هزارنفررسیده است. طيف بيماري از وضعيت بدون علامت تا پنومونی شديد و مرگ متفاوت است. بر اساس شواهد، در ۸۰ درصد مبتلایان، شکل خفیف بیماری دیده می‌شود، اما بيماري شديد در ۱۵ تا ۲۰ درصد موارد ديده می‌شود، که در اين وضعيت نياز به بستری شدن، ونتيلاسيون و مرگ افزايش می‌يابد (۲).

مطالعات انجام شده نشان داده است فاکتورهایی مانند سن، جنس، بیماری هایی چون دیابت، پر فشاری خون، بیماری قلبی عروقی، بیماری کلیوی، بیماری ریوی و بیماریهای نقص سیستم ایمنی با شدت بیماری کووید ارتباط داشته است، بطوری که افراد با سابقه این بیماریها شکل شدید بیماری را بیشتر از خود نشان می‌دادند(۳-۷). بر همین اساس بررسی‌های انجام شده در امریکا نشان داده است قریب ۴۳ درصد از افراد بستری شده، سن بالای ۶۵ سال داشته اند،که بیش از ۷۶درصد ازآنها فوت نموده اند. همچنین داده هاي به دست آمده از کشور چین و ایتالیا نشان داده که بیشترین میزان مرگ و میر در سنین ۶۰ سال و بالاتر بوده است(۸). در ارتباط با سایر متغیرها در سایر مطالعات گزارش گردیده است که در افراد با نمایه توده بدنی بالای ۳۵ مرگ و میر بیشتر می‌باشد. در مردان بالای ۷۵ سال که مبتلا به بیماری قلبی می‌باشند در مقایسه با زنان رخداد شکل شدید بیماری بیشتر بوده است. شانس بیماری شدید، در افرادی با نقص سیستم ایمنی بیشتر می‌باشد (۳-۷).

آگاهی از این نکته که کدام ویژگی افراد میتواند بروز بیماری شدید را موجب شود بسیار مهم است. با پیش‌بینی و تخمین پیامدهای شدید میتوان به بیماران و سیاستمداران کمک کرد. زیرا منابع در سیستم سلامت محدود هستند و به منظور بهترین استفاده از منابع و خدمات، شناسایی این افراد در کمک به موقع به آنها ضروری بنظر می‌رسد، بطوری

که استفاده از مدلهای پیش‌بینی درسطح اول مراقبتهای بهداشتی، به مراقبین سلامت در آموزش دقیق تر و حساس سازی این افراد میتواند کمک کننده باشد و مانع از افزایش بار مراجعات به مراکز درمانی و صرفه جویی در هزینه ها در نهایت ارتقاء سلامتی افراد می گردد.

در سال‌های اخیر با توجه به دسترسی گسترده به مقادیر بسیار عظیمی از داده ها و نیاز قریب الوقوع برای تبدیل داده به اطلاعات مفید و دانش، روش‌های داده‌کاوی توجه زیادی را درعلوم مختلف از جمله علوم پزشکی به خود جلب نموده است. داده‌کاوی به مفهوم "کشف مدل ها و ایده‌های بزرگ، برای تصمیم گیری و تشخیص الگو و پیش‌بینی" است. داده‌کاوی در حوزه سلامت کاربردهای فراوانی دارد که از جمله آن ها می‌توان به مواردی مانند تشخیص بیماری ها، دسته بندی بیماران در مدیریت بیماری، پیدا کردن الگوهایی برای تشخیص زمان بقا بیماران و غیره اشاره کرد(۹، ۱۰). الگوریتم‌های داده‌کاوی در مطالعات بیماری کووید ۱۹ نیز مورد استفاده قرار گرفته و تحلیل‌های ارزشمندی را ارائه داده اند(۱۱-۱۳). با توجه به اهم ت موضوع کرونا و تغ رات انجام شده در ش وه زندگي و ارائه تقس م بندی جدید جهت سلامت جامعه در سال‌های اخیر و ضرورت بررسی موارد موثر در ایجاد شدت این بیماری، در این تحقیق سعی می‌شود از دیدگاه آماری و داده‌کاوی تحلیل اپیدمی شدت بیماری ویروس کووید ۱۹ مورد بررسی قرار گیرد. لذا هدف اصلی این مقاله، تعیین مدل مناسب و تحلیل به روش داده‌کاوی برای شدت بیماری کووید ۱۹ بر اساس برخی از مشخصه‌های مراجعه کنندگان دارای PCR مثبت به مراکز بهداشت و درمان در جمعیت تحت پوشش دانشگاه علوم پزشکی مشهد است.

۲. روش مطالعه

این مقاله یک پژوهش از نوع مقطعی-تحلیلی می‌باشد. در این پژوهش ۳۰۳۶۴ نفر از افراد مراجعه کننده به مراکز بهداشتی دانشگاه علوم پزشکی مورد بررسی و تحلیل قرار می‌گیرند. جمع آوری داده ها و اطلاعات بیماران از سامانه سینا معاونت بهداشتی دانشگاه علوم پزشکی مشهد در پیک دوم (اواخر خرداد ماه ۹۹ تا اواخر شهریور ماه ۹۹) و پیک چهارم(۱۴ فروردین ۱۴۰۰ تا اخر اردیبهشت ماه ۱۴۰۰) صورت گرفته است. معیار ورودی بیماران در این پژوهش افراد دارای علائم بیماری کرونا و مراجعه کننده به مراکز بهداشتی بوده که بعد از تجویز پزشک مرکز به آزمایش PCR ، آزمایش PCR مثبت داشته و دارای پرونده الکترونیک سلامت فعال می‌باشند و معیار خروج بیماران فردی که پرونده الکترونیک فعال سلامت در مراکز بهداشتی نداشته و اطلاعات ناقصی دارند. روش آماری در این تحقیق استفاده از مدل‌های داده‌کاوی است. برای این مجموعه داده با توجه به حجم بالای داده‌ها روش درخت رگرسیون و طبقه بندی Classification and Regression Trees (CART) جهت برازش مدل مناسب پیشنهاد می‌شود. درخت تصمیم یکی از روش‌های ناپارمتری طبقه کردن است که در داده‌کاوی بسیار مورد استفاده قرار می‌گیرد. با توجه به نوع متغیر به دو دسته طبقه بندی درختی برای متغیر گسسته و رگرسیون درختی برای متغیر پیوسته تقسیم می‌شود. روش درخت رگرسیون و طبقه بندی (CART) توسط بریمن و همکاران (۱۹۸۴) معرفی شده است برای توضیحات تکمیلی از نوع ساخت مدل درختی CART به بسکابادی و همکاران (۹) مراجعه شود. در سال‌های اخیر با توجه به دسترسي گسترده به مقادیر بسیار عظیمي از داده‌ها و نیاز قریب الوقوع براي تبدیل داده به اطلاعات مفید و دانش، روش‌های داده کاوي توجه زیادي را در علوم مختلف از جمله علوم پزشکی به خود جلب نموده است. داده‌کاوی عبارت است از "کشف روش ها و الگوهای ویژه در پایگاه داده های بزرگ، برای هدایت تصمیم گیری در مورد فعالیت های آینده". در ایران نیز با توسعه آمارهای ثبتی لزوم استفاده از روش‌های داده‌کاوی بیشتر مورد توجه محققین قرار گرفته است. تجزیه و تحلیل داده‌ها را در این تحقیق با استفاده از JMP نسخه ۱۳ انجام می‌شود.

۳. یافته ها

در این بخش جزییات همه متغیرهای مختلف بصورت گزارش توصیفی بررسی می‌شود. داده‌های این تحقیق حاصل تمام افراد مراجعه کننده در مراکز بهداشتی دانشگاه علوم پزشکی مشهد بوده که آزمایش PCR مثبت داشته و در سامانه سینا ثبت گردیده‌اند. ابتدا به پاک‌سازی داده ها پرداخته شده و داده‌های غیر متعارف حذف می‌شوند. کل افراد باقیمانده در مطالعه ۳۰۳۶۴ نفر می‌باشند. در این جمعیت ۲۸۵۷۷ نفر معادل ۹۶ درصد بیماران خفیف بوده و ۱۷۸۷ نفر معادل ۶ درصد بیماران با شدت بالا که بستری بیمارستان و ICU داشته‌اند ثبت شده است. میانگین و انحراف معیار سن افراد ثبت شده به ترتیب برابر ۴۱/۷ و ۱۶/۵ می‌باشد. همچنین با توجه به این مجموعه داده نقطه بحرانی (cutoff) برای شدت بیماری کرونا در مدلسازی، سن ۶۰ سال محاسبه شده است بطوریکه شدت بیماری را از حدود ۳درصد زیر ۶۰ سال به حدود ۱۸ درصد بالای ۶۰ سال افزایش می‌دهد. میانگین و انحراف معیار BMI بیماران به ترتیب برابر ۲۵/۹ و ۵/۲ می‌باشد. در این جمعیت ۱۶۲۶۴ نفر جمعیت زنان معادل ۵۴ درصد و ۱۴۱۰۰ نفر جمعیت مردان معادل ۴۶ درصد می‌باشد. متغیرهای دیگر بررسی شده در این تحقیق بطور خلاصه در جدول ۱ نشان داده شده است.

طبق روش‌های متداول داده‌کاوی مجموعه داده‌ها را به دو دسته بصورت ۷۵ درصد داده‌های آموزشی و ۲۵ درصد داده‌های آزمایشی بصورت تصادفی تقسیم می‌کنیم. این تقسیم بندی به دلیل این است که با داده‌های آموزشی مدل مناسب را برازش می‌دهیم. با داده‌های آزمایشی مدل را مورد ارزیابی قرار میدهیم. با توجه به اینکه متغیر سن اهمیت زیادی در تحلیل دارد و تاثیرگذارترین متغیر در این مجموعه داده‌ها با نقطه بحرانی ۶۰ سال است، در این تحقیق این متغیر را به دو دسته زیر ۶۰ سال و بالای ۶۰ سال تقسیم کرده و برای هر دسته جداگانه به مدلسازی درختی خواهیم پرداخت. این تکنیک از پیچیدگی زیاد مدل درختی کاسته و قدرت تفسیر پذیری مدل را برای مدیریت و کنترل جامعه هدف بالا میبرد. با استفاده از روش درخت رگرسیون و طبق بندی (CART) مدل نهایی درختی در ادامه بدست می‌آوریم.

۱.۳. برازش مدل برای گروه سنی مراجعین زیر ۶۰ سال

با روش بهینه‌سازی مقدار ضریب تعیین (R2) مدل، برای مدل رگرسیون درختی ۱۰ بار تقسیم شاخه درخت مناسب می‌باشد. بنابراین با توجه به روش درختی CART با تقسیمات دوتایی ۲۰ گره در مدل خواهیم داشت. با توجه به نمودار شکل ۲ با مدل درختی برازش داده شده در شکل ۱ ، قدرت تصمیم گیری تشخیص شدت بیماری و پیش‌بینی موارد جدید برای فردی که مراجعه می‌کند تا ۶ برابر افزایش می‌یابد.

در بررسی کارایی مدل با توجه به شکل ۳ ، مساحت زیر منحنی مشخصه محرکه گیرنده (ROC) Curve برای تعیین میزان صحت مدل رده بندی درختی برابر ۷۳/۷ درصد برای داده‌های آموزشی بوده که نشان دهنده توان نسبتا بالای مدل رده بندی درختی در تعیین عوامل موثر بر تشخیص شدت بیماری است. همچنین مقدار مساحت ۷۲/۲ درصد برای داده‌های آزمایشی نیز نشان از قدرت پیش‌بینی به نسبت خوب این مدل است.

۲.۳. برازش مدل برای گروه سنی مراجعین بالای ۶۰ سال

با روش بهینه‌سازی مقدار ضریب تعیین مدل، برای مدل رگرسیون درختی ۷ بار تقسیم شاخه درخت مناسب می‌باشد. بنابراین با توجه به روش درختی CART با تقسیمات دوتایی ۱۴ گره در مدل خواهیم داشت.
با توجه به نمودار شکل ۵ با مدل درختی برازش داده شده در شکل ۴ ، قدرت تصمیم گیری تشخیص شدت کرونا برای فردی که مراجعه میکند ۱/۷ و برای فرد سالم ۲/۱ برابر افزایش می‌یابد.
در بررسی کارایی مدل با استفاده از مساحت زیر منحنی مشخصه محرکه گیرنده برای تعیین میزان صحت مدل رده بندی درختی (شکل ۶)، مقدار مساحت ۷۰/۳ درصد بر روی داده‌های آموزشی و ۶۵/۳۰ درصد بر روی داده‌های آزمایش به دست آمد که نشان دهنده توان تقریبا بالای مدل رده بندی درختی در تعیین عوامل موثر بر تشخیص شدت بیماری کرونا است.

۴. بحث

در این بخش به بحث در مورد مطالعات مرتبط و سپس تفسیر مدل های درختی برازش داده شده در بخش قبل به مجموعه داده های تعیین کننده شدت بیماری کرونا می‌پردازیم. در مطالعات انجام شده گذشته نشان داده شده است فاکتورهایی مانند سن، جنس، بیماری هایی چون دیابت، پر فشاری خون، بیماری قلبی عروقی، بیماری کلیوی، بیماری ریوی و بیماریهای نقص سیستم ایمنی با شدت بیماری کووید ارتباط داشته است، بطوری که افراد با سابقه این بیماریها شکل شدید بیماری را بیشتر از خود نشان می‌دادند(۳-۷). بر همین اساس بررسی‌های انجام شده در امریکا نشان داده است قریب ۴۳ درصد از افراد بستری شده، سن بالای ۶۵ سال داشته اند،که بیش از ۷۶درصد ازآنها فوت نموده اند. همچنین داده هاي به دست آمده از کشور چین و ایتالیا نشان داده که بیشترین میزان مرگ و میر در سنین ۶۰ سال و بالاتر بوده است(۸).
در این مقاله نیز مشابه مطالعات گذاشته با مجموعه داده های جمع آوری شده از دانشگاه علوم پزشکی مشهد و استفاده از مدل‌های داده‌کاوی نیز اهمیت متغیرهای سن، بیماری های قلبی، کلیوی، تنفسی، چربی خون و دیابت را نشان میدهد. روش مدل سازی در این مقاله اطلاعات ارزشمندی برای نوع تاثیر متغیرهای تاثیرگذار بیماری کرونا به ما میدهد که در ادامه با تفسیر مدل بیان می شود.

۱.۴. تفسیر مدل درختی برای گروه سنی مراجعین زیر ۶۰ سال

افراد در گروه سنی کمتر از ۶۰ سال با احتمال ۵۴.۹۶ درصد، شدید بودن بیماری کرونا رو تجربه نکردند و تنها با ۴۶.۳ درصد احتمال، مراجعه کننده ها با شدید بودن این بیماری مواجه بودند و این نتیجه زمانی حاصل شده است که از سایر عوامل یا متغیرها اطلاعی در دست نباشد.
از آنجایی که درنظر گرفتن اطلاعات بیشتر به شناسایی دقیق تر موارد شدید و غیر شدید کرونا کمک می‌کند، در شاخه‌های مختلف درخت می‌توان اطلاعات و نتایج دقیق تری به دست آورد. با توجه به درخت موجود در شکل ۱ مشاهده می‌شود که ۱۱ قانون برای تصمیم گیری در مورد شدید یا عدم شدید بودن کرونا در افراد مراجعه کننده استخراج شده است. این قوانین در گره‌های ۳، ۴، ۹، ۱۱، ۱۲، ۱۳، ۱۴، ۱۶، ۱۷، ۱۹ و ۲۰ قابل نمایش است. در این بخش به تفسیر این قوانین خواهیم پرداخت.
طبق شکل ۱، گره شماره سه نشان می‌دهد اگر فرد مراجعه کننده دارای بیماری قلبی باشد و سن کمتر از ۴۵ سال داشته باشد با احتمال ۸۳/۷۸ درصد با شدید بودن بیماری کرونا درگیر خواهد بود. از طرف دیگر، همین فرد اگر دارای سن بین ۴۵ تا ۶۰ سال باشد احتمال شدید بودن کرونا در آن ۰۳.۳۵ درصد خواهد بود (گره شماره چهار). سایر قوانین به دست آمده از درخت به صورت زیر تفسیر می‌شوند:
- گره شماره ۹: احتمال شدید بودن (شدید نبودن) کرونا برای مراجعین کمتر از ۴۰ سال که مبتلا به بیماری قلبی نیستند اما به بیماری دیابت مبتلا هستند تقریبا ۱۸/۹۰ درصد (۸۱/۱۰ درصد) است.
- گره شماره ۱۱: احتمال شدید بودن (شدید نبودن) کرونا برای مراجعین مرد که بین ۴۰ تا ۶۰ سال سن دارند و مبتلا به بیماری دیابت هستند تقریبا ۱۷/۵۵ درصد (۸۲/۴۵ درصد) است.
- گره شماره ۱۲: احتمال شدید بودن (شدید نبودن) کرونا برای مراجعین زن که سنی بین ۴۰ تا ۶۰ سال دارند و مبتلا به بیماری دیابت هستند تقریبا ۷/۹۶ درصد (۹۲/۰۴ درصد) است. در حالیکه شدید بودن بیماری کرونا در مردان نسبت به زنان به اندازه ۹/۵۹ درصد بیشتر است.
- گره شماره ۱۳: احتمال شدید بودن (شدید نبودن) کرونا برای مراجعین بین ۴۰ تا ۶۰ سال که مبتلا به دو بیماری قلبی و دیابت نیستند اما دارای چربی خون هستند تقریبا ۶/۳۲ درصد (۹۳/۶۸ درصد) است.
- گره شماره ۱۴: احتمال شدید بودن (شدید نبودن) کرونا برای مراجعین بین ۴۰ تا ۶۰ سال که مبتلا به دو بیماری قلبی و دیابت نیستند و دارای چربی خون هم نیستند تقریبا ۳۵.۳ درصد (۹۶/۶۵ درصد) است. بنابراین، نتیجه می‌شود تحت چنین شرایطی چربی خون داشتن باعث افزایش ۲/۹۷ درصدی در احتمال شدید بودن بیماری کرونا نسبت به افرادی که چربی خون ندارند می‌شود.
- گره شماره ۱۶: احتمال شدید بودن (شدید نبودن) کرونا برای مراجعین کمتر از ۱۹ سال که مبتلا به دو بیماری قلبی و دیابت نیستند تقریبا ۲۸.۰ درصد (۹۹/۷۲ درصد) است.
- گره شماره ۱۷: احتمال شدید بودن (شدید نبودن) کرونا برای مراجعین بین ۱۹ تا ۴۰ سال که بیماری قلبی، دیابت و بیماری تنفسی دارند تقریبا ۲۵/۹۰ درصد (۷۴/۱۰ درصد) است.
- گره شماره ۱۹: احتمال شدید بودن (شدید نبودن) کرونا برای مراجعین بین ۱۹ تا ۴۰ سال که بیماری قلبی، دیابت و بیماری تنفسی ندارند اما مبتلا به بیماری کلیه هستند، تقریبا ۵۷/۵۵ درصد (۴۲/۴۵ درصد) است.
- گره شماره ۲۰: ااحتمال شدید بودن (شدید نبودن) کرونا برای مراجعین بین ۱۹ تا ۴۰ سال که بیماری قلبی، دیابت و بیماری تنفسی ندارند و مبتلا به بیماری کلیه هم نیستند، تقریبا ۱/۷۹ درصد (۹۸/۲۱ درصد) است، در صورتی که افراد مبتلا به بیماری کلیه نسبت به افرادی که نیستند افزایش چشمگیری در احتمال ابتلا به کرونای شدید به اندازه ۵۵/۷۶ درصدی

است.

۲.۴. تفسیر مدل درختی برای گروه سنی مراجعین بالای ۶۰ سال

نتایج نشان می‌دهد زمانی که از سایر عوامل یا متغیرها اطلاعی موجود نباشد، افراد در سنین بالاتر از ۶۰ سال با احتمال ۸۲/۰۹ درصد، به شکل خفیف بیماری کرونا مبتلا شده و ۱۷/۹۱ درصد احتمال، با شدید بودن این بیماری مواجه می‌باشند.

همانند قبل، با توجه به اینکه درنظر گرفتن اطلاعات بیشتر به شناسایی دقیق تر موارد شدید و غیر شدید کرونا کمک می‌کند، به بررسی اطلاعات موجود در شاخه‌های مختلف درخت در شکل ۴ می‌پردازیم. این درخت ۸ قانون مهم در زمینه تشخیص شدید یا عدم شدید بودن کرونا را به کمک تعدادی از متغیرهای پیشگو نشان می‌دهد. قوانین مستخرج از این درخت در گره‌های ۳، ۵، ۷، ۸، ۹، ۱۱، ۱۳ و ۱۴ قابل مشاهده است. اولین گره یعنی گره شماره سه نشان می‌دهد اگر فرد مراجعه کننده دارای بیماری قلبی باشد و سن بیشتر از ۷۴ سال داشته باشد با احتمال ۲۷.۵۱ درصد با شدید بودن بیماری کرونا درگیر خواهند بود. از طرف دیگر، همین فرد اگر دارای سن بین ۶۰ تا ۷۴ سال باشد احتمال شدید بودن کرونا در آن ۹۷.۳۱ درصد خواهد بود (گره شماره پنج) به عبارت دیگر، زمانی که تنها وجود بیماری قلبی مشهود باشد، بودن افراد در سنین بالای ۷۴ سال با افزایش ۳.۱۹ درصدی در احتمال شدید بودن بیماری کرونا نسبت به افراد در سنین بین ۶۰ تا ۷۴ سال مواجهه هستند. سایر قوانین به دست آمده از درخت به صورت زیر تفسیر می‌شوند:

- گره شماره ۷: احتمال شدید بودن (شدید نبودن) کرونا برای مراجعین بیشتر از ۸۱ سال که مبتلا به بیماری قلبی نیستند تقریبا ۳۱.۳۶ درصد (۶۳/۶۹ درصد) است.
- گره شماره ۸: احتمال شدید بودن (شدید نبودن) کرونا برای مراجعینی که بین ۷۴ تا ۸۱ سال سن دارند و مبتلا به بیماری قلبی نیستند تقریبا ۲۰/۸۳ درصد (۷۹/۱۷ درصد) است.
- گره شماره ۹: احتمال شدید بودن (شدید نبودن) کرونا برای مراجعین که سنی بین ۶۰ تا ۷۴ سال دارند و مبتلا به بیماری قلبی نیستند اما به بیماری کلیه مبتلا هستند تقریبا ۷۱/۵۸ درصد (۲۸/۴۲ درصد) است. در حالیکه شدید بودن بیماری کرونا در مردان نسبت به زنان به اندازه ۵۹.۹ درصد بیشتر است.
- گره شماره ۱۳: احتمال شدید بودن (شدید نبودن) کرونا برای مراجعین که سنی بین ۶۰ تا ۷۴ سال دارند و مبتلا به بیماری قلبی، کلیه و تنفسی نیستند اما دیابت دارند تقریبا ۱۵/۶۸ درصد (۸۴/۳۲ درصد) است.
- گره شماره ۱۱: احتمال شدید بودن (شدید نبودن) کرونا برای مراجعین که سنی بین ۶۰ تا ۷۴ سال دارند و مبتلا به بیماری قلبی و کلیه نیستند اما به بیماری تنفسی مبتلا هستند تقریبا ۵۰/۵۸ درصد (۴۸/۴۲ درصد) است.
- گره شماره ۱۴: احتمال شدید بودن (شدید نبودن) کرونا برای مراجعین که سنی بین ۶۰ تا ۷۴ سال دارند و مبتلا به هیچ یک از بیماری‌های قلبی، کلیه، تنفسی و دیابت نیستند تقریبا ۸/۸۴ درصد (۹۱/۱۶ درصد) است.

این نشان می‌دهد در شرایطی که مراجعه کنندگان بین ۶۰ تا ۷۴ ساله از نظر عدم ابتلا به بیماری هایی چون قلبی، کلیه و تنفسی یکسان هستند و با این بیماری ها درگیر نیستند، وجود بیماری دیابت (نسبت به عدم وجود بیماری دیابت) به اندازه ۸۴.۶ درصد در افزایش احتمال ابتلا به شدید بودن بیماری کرونا نقش دارد.

۵. نتیجه گیری

با توجه به نوع مجموعه داده‌ها و حجم بالای داده، روش درخت رگرسیون و طبق بندی روشی مناسب و قابل تفسیر برای این تحقیق می‌باشد. با مدلی که از این روش به داده‌ها برازش شد می‌توان کنترل جامعه را با توجه به عوامل بررسی شده در این تحقیق به خوبی انجام داد. نتایج مدل رگرسیون و طبقه بندی درختی نشان می‌دهد که از بین متغیرهای برای سن زیر ۶۰ سال به ترتیب اهمیت بیماری قلبی، سن، دیابت، بیماری تنفسی و چربی خون و برای سن بالای ۶۰ سال به ترتیب اهمیت سن، بیماری قلبی، بیماری کلیوی، بیماری تنفسی و دیابت موثر می‌باشند.

۱.۵. معیارهای ورود و خروج از مطالعه

معیار ورودی بیماران در این پژوهش "افراد دارای تست PCR مثبت که در پیک دوم (اواخر خرداد ماه ۹۹ تا اواخر شهریور ماه ۹۹) و پیک چهارم(۱۴ فروردین ۱۴۰۰ تا اخر اردیبهشت ماه ۱۴۰۰) تشخیص داده شده و دارای پرونده الکترونیک سلامت فعال در سامانه سینا می‌باشند" و معیار خروج بیماران " فردی که پرونده الکترونیک فعال سلامت ندارد (عدم مراجعه به مرکز بهداشت یا عدم تکمیل فرم‌های مربوطه)" است.

۶. محدودیت هاي مطالعه

این مطالعه دارای چندین محدودیت است. با توجه به شیوه جمع آوری داده ها از سامانه سینا و بیماران منتخب با فعال بودن پرونده الکترونیک سلامت، بیماران بسیاری وارد مطالعه نمی شوند. همچنین در اطلاعات بیماران داده‌های تکمیل نشده و ناقص نیز زیاد مشاهده می‌شود. لذا پیشنهاد می‌شود مشابه این مطالعه برای داده‌های بیمارستانی نیز انجام پذیرد.

۷. پیشنهادات

در آینده تحقق پیشنهاد می‌شود روش‌های دیگر داده‌کاوی را برای تحلیل و بررسی بر روی این مجموعه داده و تاثیراتی که در این تحقیق بررسی نشده است کار شود. یکی از روش‌های مدرن در این زمینه شبکه‌های عصبی است که می‌توان روی این داده‌ها بررسی کرد و همچنین با روشی که در این تحقیق کار شد مقایسه شود. مدل شبکه‌های عصبی با دقت پیش‌بینی خیلی بالاتر از مدل‌های درختی رگرسیون می‌تواند نقص این مدل ها را رفع کند. این تحقیقات پیشنهاد می‌شود برای دیگر بخش‌های داده‌های استخرج شده از سامانه سینا در دانشگاه علوم پزشکی مشهد و همچنین در سطح وسیع‌تر از سامانه‌های نظام سلامت کشور انجام شود تا باعث برنامه‌ریزی دقیق‌تر مدیران، پزشکان و پیراپزشکان از حجم وسیع داده‌های ثبت شده شود.

۸. تقدیر و تشکر

از گروه آمار زیستی و گروه اپیدمیولوژی در دانشکده بهداشت دانشگاه علوم پزشکی مشهد بابت همکاری علمی و همچنین معاونت بهداشتی دانشگاه علوم پزشکی مشهد بابت همکاری در اختیار گذاشتن داده‌های

سامانه سینا تشکر و قدردانی به عمل می‌آید.

۸.۱. ملاحضات اخلاقی

این مطالعه مستخرج از طرح پژوهشي کد ۴۰۱۰۸۳۸ و مصوب کمیته اخلاق دانشگاه علوم پزشکی مشهد با کد IR.MUMS.REC.1401.236 ثبت شده است.

۹. سهم نویسندگان

تمامی نویسندگان معیارهای استاندارد نویسندگی بر اساس پیشنهادات کمیته بین المللی ناشران مجلات پزشکی را دارا بودند.

۱۰. تضاد منافع

نویسندگان تصریح می نمایند که هیچگونه تضاد منافعی در خصوص پژوهش حاضر وجود ندارد.

۱۱. منابع مالی

کلیه منابع مالی و بودجه این مطالعه توسط دانشگاه علوم پزشکی مشهد تامین شد.

مراجع

1. McIntosh K. Official reprint from UpToDate® 2020 UpToDate [Available from: www.uptodate.com.

2. Gude-Sampedro F, Fernández-Merino C, Ferreiro L, Lado-Baleato Ó, Espasandín-Domínguez J, Hervada X, et al. Development and validation of a prognostic model based on comorbidities to predict COVID-19 severity: a population-based study. International journal of epidemiology. 2021;50(1):64-74.

3. El Aidaoui K, Haoudar A, Khalis M, Kantri A, Ziati J, El Ghanmi A, et al. Predictors of Severity in Covid-19 Patients in Casablanca, Morocco. Cureus. 2020;12(9):e10716.

4. Rottoli M, Bernante P, Belvedere A, Balsamo F, Garelli S, Giannella M, et al. How important is obesity as a risk factor for respiratory failure, intensive care admission and death in hospitalised COVID-19 patients? Results from a single Italian centre. European journal of endocrinology. 2020;183(4):389-97.

5. Liang WH, Guan WJ, Li CC, Li YM, Liang HR, Zhao Y, et al. Clinical characteristics and outcomes of hospitalised patients with COVID-19 treated in Hubei (epicentre) and outside Hubei (non-epicentre): a nationwide analysis of China. The European respiratory journal. 2020;55(6).

6. Turcotte JJ, Meisenberg BR, MacDonald JH, Menon N, Fowler MB, West M, et al. Risk factors for severe illness in hospitalized Covid-19 patients at a regional hospital. PloS one. 2020;15(8):e0237558.

7. Saiphoklang N, Kanitsap A. Prevalence, clinical manifestations and mortality rate in patients with spontaneous pneumothorax in Thammasat University Hospital. Journal of the Medical Association of Thailand = Chotmaihet thangphaet. 2013;96(10):1290-7.

8. Bialek S, Boundy E, Bowen V, Chow N, Cohn A, Dowling N, et al. Severe outcomes among patients with coronavirus disease 2019 (COVID-19)—United States, February 12–March 16, 2020. Morbidity mortality weekly report. 2020;69(12):343.

9. Boskabadi M, Doostparast M, Sarmad M. Survival analyses with dependent covariates: A regression tree-base approach. Journal of Algorithms and Computation. 2020;52(1):105-29.

10. Choi SB, Kim WJ, Yoo TK, Park JS, Chung JW, Lee YH, et al. Screening for prediabetes using machine learning models. Computational and mathematical methods in medicine. 2014;2014:618976.

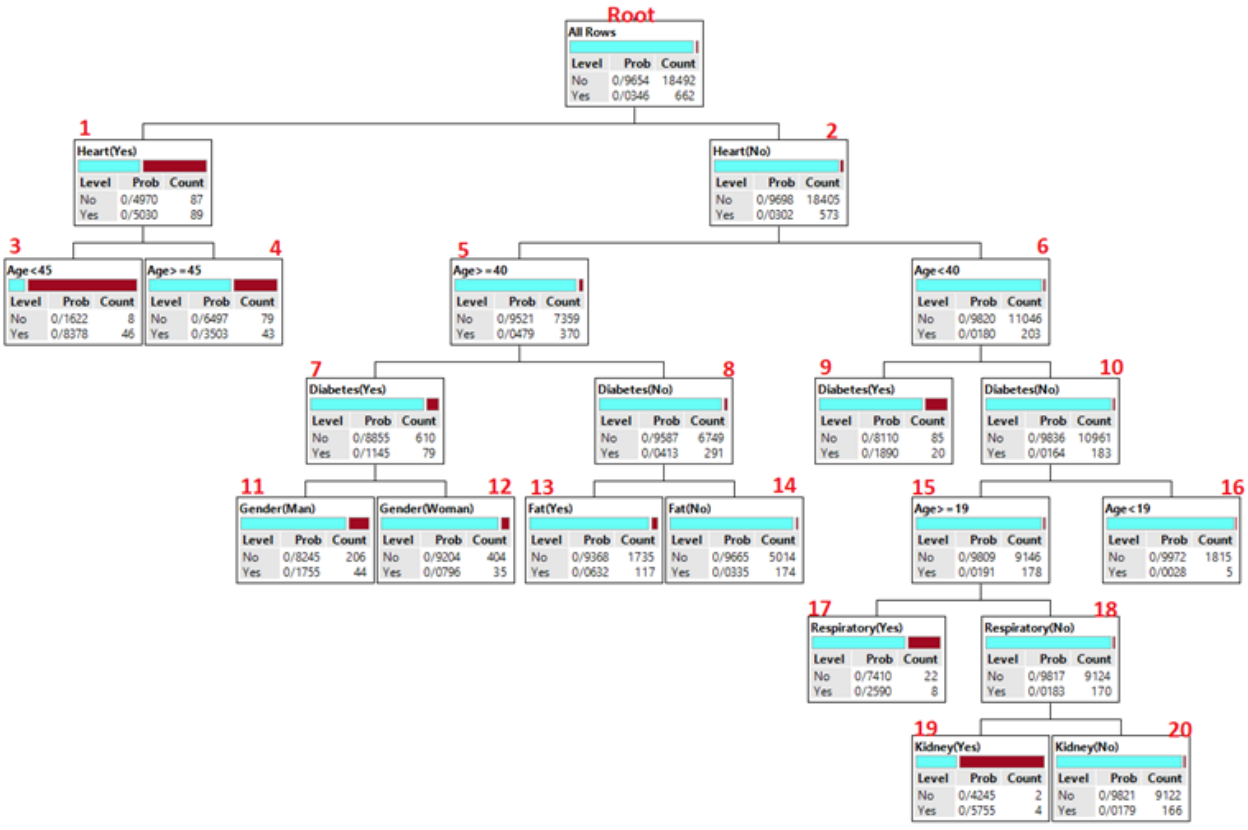
11. Muhammad LJ, Islam MM, Usman SS, Ayon SI. Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery. SN computer science. 2020;1(4):206.

12. Haque MR, Islam MM, Iqbal H, Reza MS, Hasan MK, editors. Performance Evaluation of Random Forests and Artificial Neural Networks for the Classification of Liver Disorder. 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2); 2018 8-9 Feb. 2018.

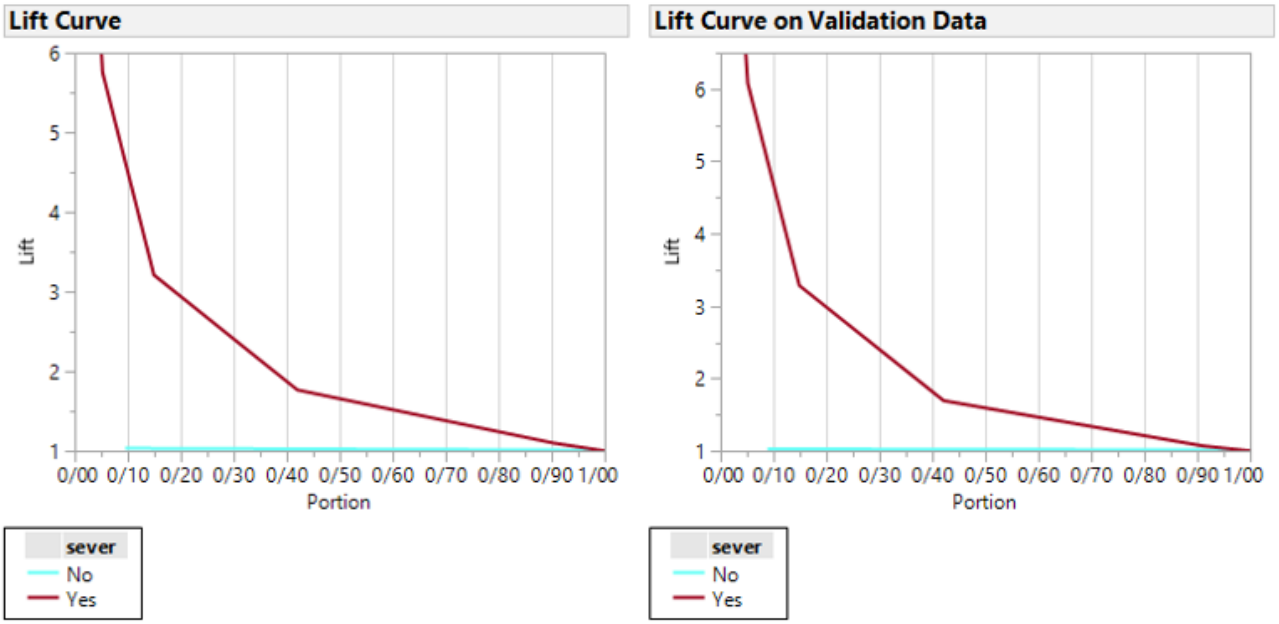
13. Boskabadi M, Doostparast M. Modeling and data mining of global data on patients with COVID-19. Iranian Journal of Emergency Medicine. 2020;7(1):1 [e40]-6.

بیماری زمینه‌ای	چاقی	دیابت	سرطان	کلیوی	قلبی عروقی	مزمن کبدی	مزمن ریوی	بارداری	افسردگی
تعداد	۶۰۵۶	۲۶۴۲	۳۰۸	۷۸	۷۳۴	۱۴۱	۱۹۳	۴۲۱۰	۱۴۸۹
درصد	۱۹/۹	۸/۷	۱	۰/۰۲	۲/۴	۰/۰۵	۰/۰۶	۱۳/۹	۴/۹

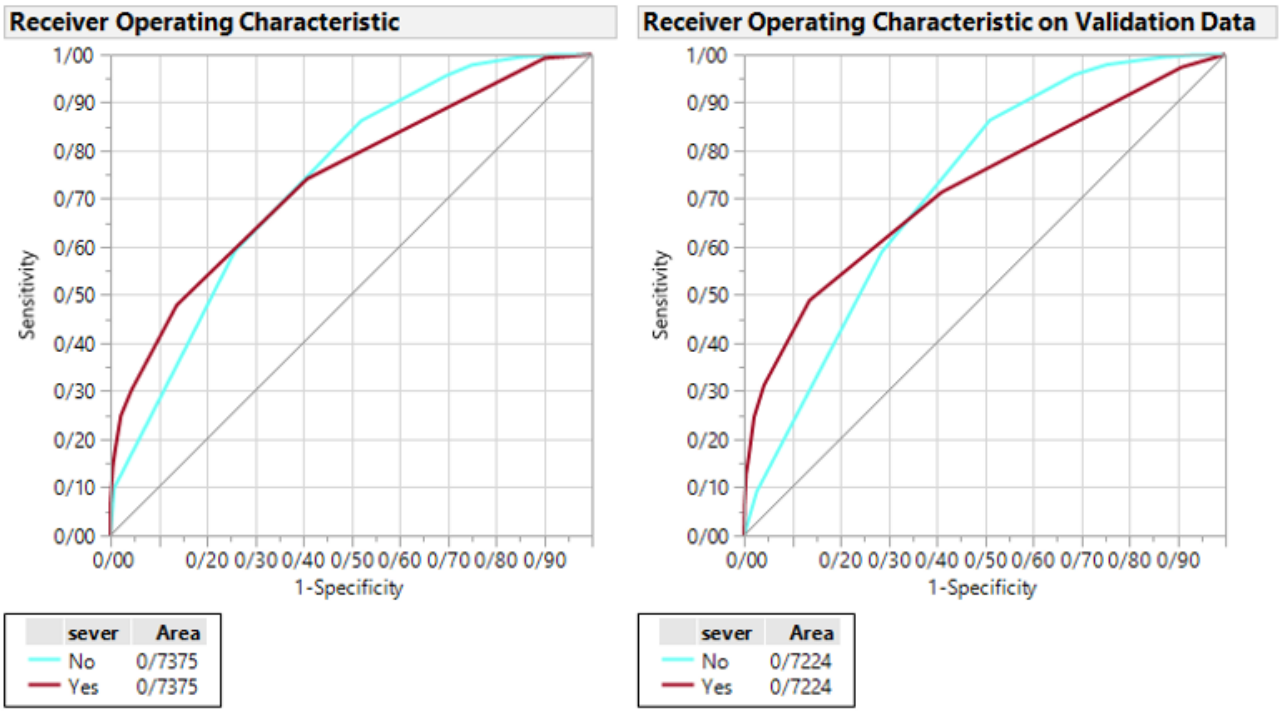
جدول ۱. تعداد و درصد بیماری زمینه ای در کل بیماران



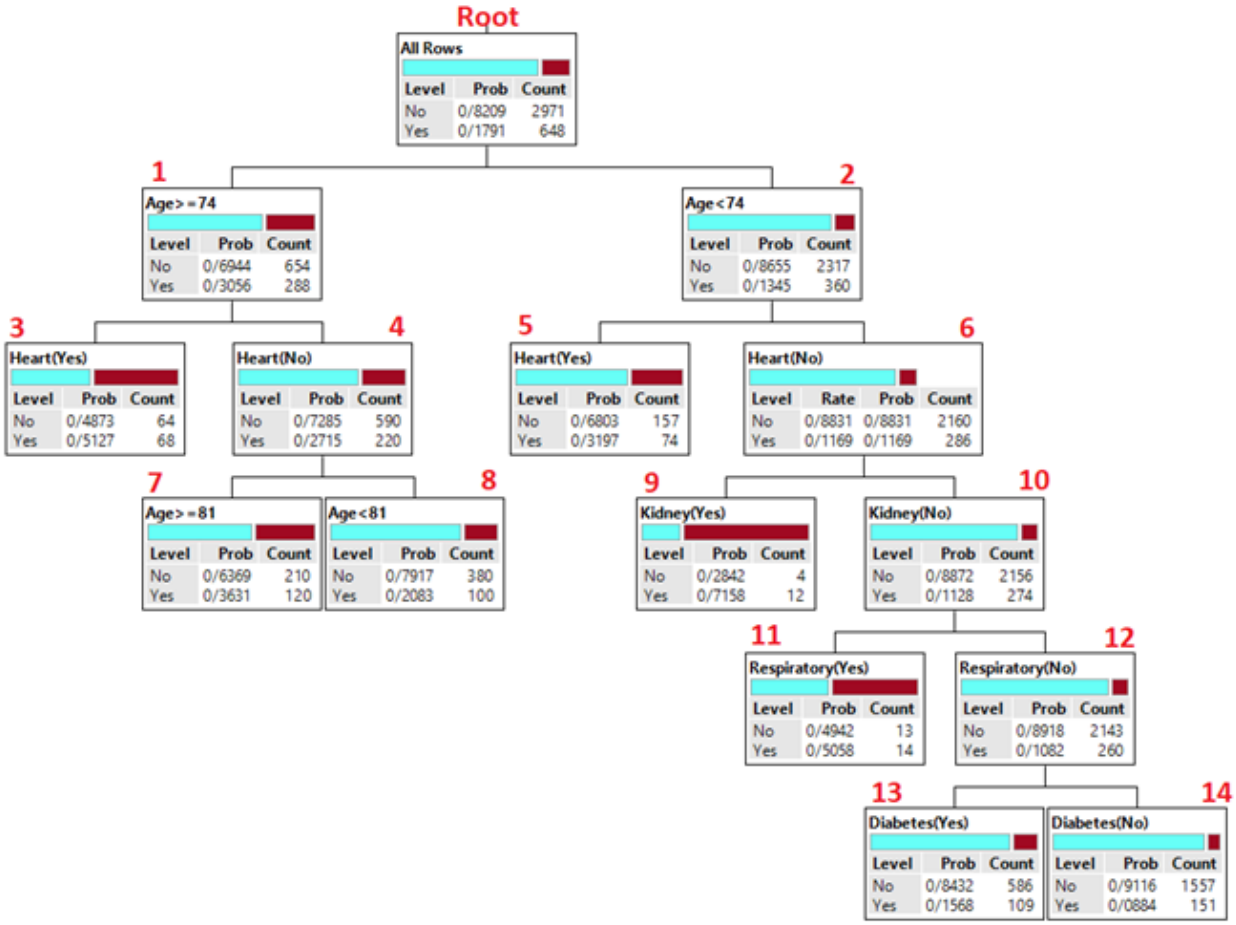
شکل ۱. مدل درختی برازش داده شده به بیماران مبتلا به کرونا زیر ۶۰ سال



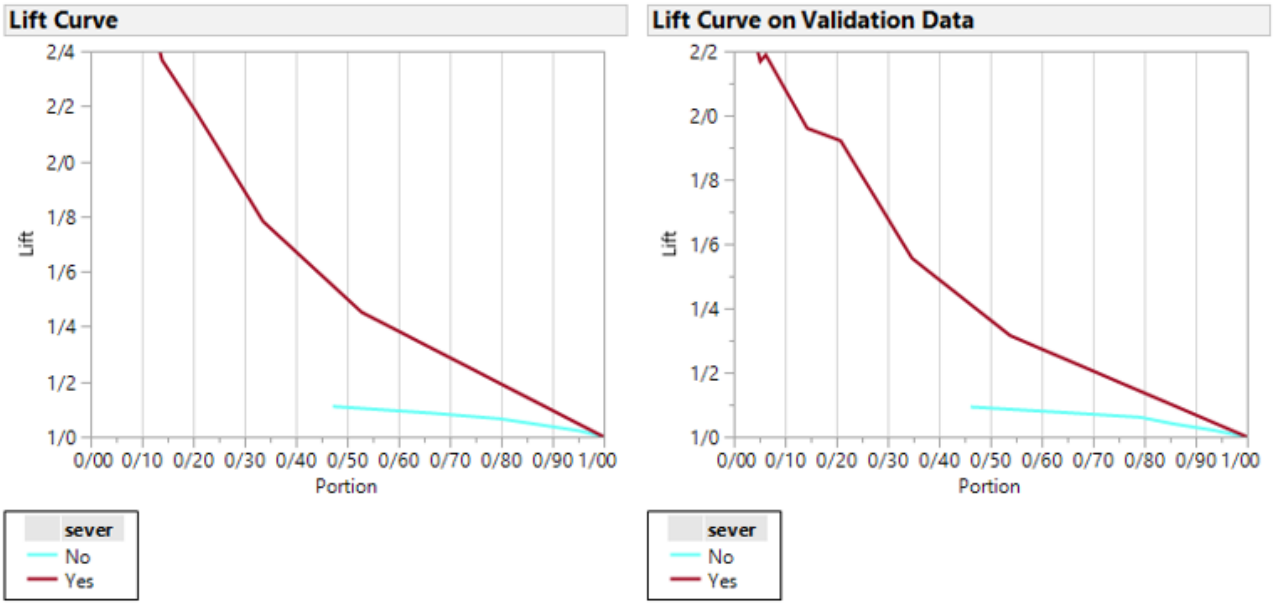
شکل ۲. نمودار منحنی قدرت تصمیم گیری مدل بیماران زیر ۶۰ سال



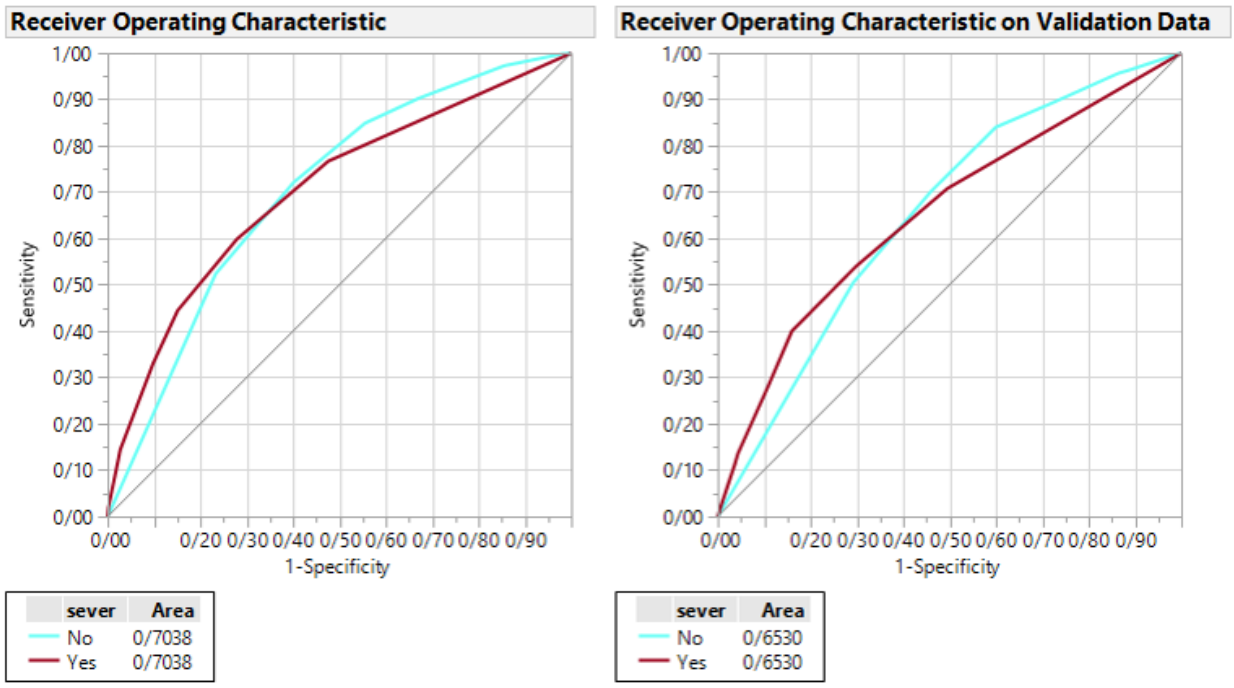
شکل ۳. نمودار منحنی ROC بیماران زیر ۶۰ سال



شکل ۴. مدل درختی برازش داده شده به بیماران مبتلا به کرونا بالای ۶۰ سال



شکل ۵. نمودار منحنی قدرت تصمیم گیری مدل بیماران بالای ۶۰ سال



شکل ۶. نمودار منحنی ROC بیماران بالای ۶۰ سال

ORIGINAL ARTICLE

Modeling the Impact of some Variables the COVID-19 Severe with CART Algorithm in Mashhad University of Medical Sciences

Mostafa Boskabadi[1*], Monavar Afzalaghaee[2], Nasrin Talkhi[1], Zahra Jamalian[2], Ehsan Musa Farkhani[2], Habibollah Esmaily[1]

[1]Department of Biostatistics, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran.
[2]Department of Epidemiology, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran.

[*]**Corresponding author:** Mostafa Boskabadi; School of Health, Mashhad University of Medical Sciences, Mashad, Iran. Email: Boskabadym1@mums.ac.ir.

Recieved Date: August 2022; **Accept Date:** November 2022

Abstract

Introduction: Considering that the new corona virus (COVID -19) is still prevalent, one of the important concerns is the variables affecting the severity of the corona disease in the health of the society. In this study, the CART algorithm was fitted to predict and determin the status of patients infected with the of COVID-19 in Mashhad University of Medical Sciences. **Methods:** This paper is a cross sectional-analytical study. Dataset were obtained from all of people referred for the disease of COVID -19 collected at the Sinai system during the second peak and the fourth peak of the disease in Mashhad University of Medical Sciences. Data analysis was performed using JMP statistical software version 13. Then for modeling, data mining methods and CART algorithm are used. **Results:** The descriptive findings of our study showed that 6% of patients with positive PCR suffer from severe disease of COVID-19. The age variable was very important in the severity of the disease. The age of 60 years old is the cut-off point for the severity of the disease, which increases the COVID-19 severe from about 3% under the age of 60 to about 18% over the age of 60. The diseases of heart, kidney, respiratory, blood fat, and diabetes were other important variables. **Conclusion:** The results of the CART model showed that for the age under 60 years the variables of heart disease, age, diabetes, respiratory disease, fat, gender and kidney, and for the age over 60 years the variables of age, heart disease, kidney, respiratory and diabetes were respectivly the most critical risk factors. According to the ROC curve, the fitted model has a good performance for COVID-19 severe disease, so that it increases up to 6 times the prediction of the COVID-19 severe disease.

Key words: COVID-19, Datamining, CART algorithm, ROC curve, prediction