تحلیل و پیش‌بینی بیماری کووید -۱۹ با استفاده از الگوریتم‌های رگرسیون یادگیری ماشین

مهسا ملک‌پور شهرکی[۱]، سجاد رحمانی*[۲]

[۱] دانشکده ریاضی و علوم کامپیوتر، دانشگاه دامغان، دامغان، ایران
[۲] دانشکده ریاضی و علوم کامپیوتر، دانشگاه دامغان، دامغان، ایران

s_rahmani@du.ac.ir

چکیده: بیماری کووید -۱۹ یکی از بزرگترین چالش های بهداشتی است که جهان تاکنون با آن روبرو بوده است. به همین علت پیش‌بینی نرخ شیوع و پیشرفت این بیماری دارای اهمیت است. در این مقاله، روش‌های آماری و یادگیری ماشین را برای پیش‌بینی نرخ شیوع این بیماری در ایران مورد بررسی و مطالعه قرار خواهیم داد. نتایج این مطالعه نشان می‌دهد الگوریتم رگرسیون پرسپترون چند لایه با میزان خطای قابل قبولی می‌تواند در تعیین نرخ شیوع مورد استفاده قرار گیرد.

کلید واژه‌ها: کووید -۱۹، نرخ شیوع، پیش‌بینی، یادگیری ماشین، رگرسیون.

۱. مقدمه

در دسامبر سال ۲۰۱۹ میلادی ویروس جدیدی از خانواده کرونا در شهر ووهان کشور چین شناسایی شد، که یکی از عوارض آن به وجود آمدن سندروم تنفسی حاد و کشنده در فرد بیمار است. در ۱۱ مارس سال ۲۰۲۰ میلادی سازمان بهداشت جهانی شیوع بیماری همه‌گیر کووید -۱۹ ناشی از ویروس کرونا را به طور رسمی اعلام کرد. شیوع این بیماری زندگی تمام مردم جهان را از جهات بسیاری مانند اقتصاد، روابط اجتماعی و... تحت تأثیر قرار داده است. به دلیل عدم وجود دارو و واکسن مؤثر برای این بیماری و همچنین جهش‌های فراوان آن، روش‌هایی مانند قرنطینه، ماسک، رعایت فاصله اجتماعی و دیگر پروتکل‌های بهداشتی اعلام شده توسط سازمان بهداشت جهانی تنها استراتژی‌های کاربردی برای کاهش سرعت انتقال این بیماری است. در بسیاری از کشورها برای پیش‌گیری از همه‌گیری این بیماری قرنطینه‌های سخت‌گیرانه و تعطیلی‌های موقت اجرا شده است.
یکی از عوامل مهم در بررسی بیماری‌های همه‌گیر، عدد سرایت پایه یا نرخ شیوع R_0[۱] است. این پارامتر شاخصی برای اندازه‌گیری توان شیوع عامل بیماری‌زا (ویروس) است و نشان‌دهنده متوسط تعداد افرادی است که فرد آلوده می‌تواند عامل عفونی را به آن‌ها سرایت دهد. گروه‌های تحقیقاتی این پارامتر را برای ویروس کرونا بین ۱.۸ تا ۳.۸ تخمین زده‌اند. در نتیجه از R_0 برای ارزیابی توانایی یک بیماری عفونی برای حمله به یک جامعه استفاده می‌شود، زمانی که $R_0>1$ باشد نشان از گسترش بیماری دارد و بدین معنا است که هر فرد آلوده می‌تواند بیشتر از یک نفر را آلوده کند و کنترل بیماری سخت می‌شود. از طرفی زمانی که $R_0<1$ باشد بیماری در حال از بین رفتن و تحت کنترل است. بنابراین مشخص کردن مقدار R_0

[1] Reproduction Rate

نقش کلیدی در تخمین و کنترل بیماری‌های همه‌گیر دارد [۲].

به همین علت از همان ابتدای شیوع بیماری کووید -۱۹ تحقیقاتی در زمینه تخمین نرخ شیوع این بیماری صورت گرفته است. این روش‌ها بیشتر بر مبنای مدل پایه مستعد-درمعرض-مبتلا-حذف شده (SEIR)[۲] بوده‌اند. به طوری که این مدل توسط ۴ معادله دیفرانسیل کنترل می‌شود. به عنوان نمونه از این روش روی داده‌های کشورهایی مثل چین، هند و ایران استفاده کرده‌اند[۱، ۴، ۵]. همچنین در [۳] یک ماشین حساب اپیدمی آنلاین کووید -۱۹ به عنوان ابزاری برای ارزیابی این پاندمی شرح داده شده است. مشکلی که در ایجاد مدل SEIR وجود دارد این است که به دلیل عدم دسترسی به داده‌های دقیق برای تخمین یکسری از پارامترها در تشکیل معادلات دیفرانسیل، از نرم‌افزار متلب و با سعی و خطا استفاده شده است که همین مسئله دقت پیش‌بینی مدل را پایین می‌آورد. به علاوه در این روش ناچار هستند داده‌ها را به چند دوره زمانی تقسیم کنند. مشکل دیگری که این روش دارد محدودیت این مدل برای داده‌های کشور خاص و همچنین بازه‌های کوتاه مدت است. علاوه بر استفاده از مدل SEIR جهت تحلیل و پیش‌بینی بیماری کووید -۱۹، مشاهده می‌شود مطالعات زیادی نیز با استفاده از روش‌های هوش مصنوعی ،یادگیری ماشین و یادگیری عمیق انجام شده است[۶، ۷]. در این مطالعه ما با توجه به اهمیت پارامتر نرخ شیوع $R_\circ$ و همچنین مفاهیم یادگیری ماشین، قصد داریم با استفاده از انواع روش‌های رگرسیون یادگیری ماشین به تخمین این پارامتر حیاتی بپردازیم. چرا که تا کنون مطالعه‌ای با استفاده از این الگوریتم‌ها روی داده‌های کشور ایران و برای پیش‌بینی نرخ شیوع صورت نگرفته است.

۲. روش تحقیق

در این مطالعه از انواع الگوریتم‌های رگرسیون یادگیری ماشین جهت پیش‌بینی نرخ شیوع بیماری کووید -۱۹ استفاده شده است و سپس این الگوریتم‌ها از نظر میزان خطای میانگین‌شان با یکدیگر مورد مقایسه قرار می‌گیرند.

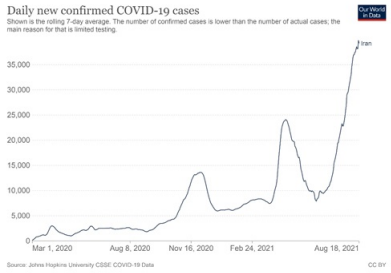
۲-۱. مجموعه داده‌ها

مجموعه داده‌های استفاده شده در این مطالعه به صورت گزارش روزانه از تاریخ ۱۹ فوریه ۲۰۲۰ تا تاریخ ۲۰ ژوئیه ۲۰۲۱ برای کشور ایران از سایت رسمی ourworldindata.org است. در این مطالعه داده‌های جمع‌آوری شده در یک جدول با فرمت csv ذخیره شده است. این مجموعه داده‌ها شامل ویژگی‌های موارد جدید ابتلا، موارد کل ابتلا، موارد جدید مرگ و میر، موارد کل مرگ و میر، میزان کل واکسناسیون، تعداد افراد بهبود یافته و تعداد کل افراد بهبود یافته است. شکل ۱ نشان می‌دهد موارد ابتلا جدید روزانه به صورت نمایی رشد کرده است و این امر نیاز به کنترل دارد.

۲-۲. الگوریتم‌ها

در این مطالعه برای پیش‌بینی نرخ شیوع بیماری کووید -۱۹ مدل‌های ریاضی پیشرفته‌ای بر اساس یادگیری ماشین به کار برده شده است. پس از پیاده‌سازی چندین الگوریتم، سه الگوریتم رگرسیون پرسپترون چند لایه (MLPR)[۳] و رگرسیون بردار پشتیبان (SVR)[۴] و رگرسیون خطی [۵] که

[2] Susceptible - Exposed - Infectious - Recovered
[3] Multi Layer Perceptron Regression
[4] Support Vector Machine Regression
[5] Linear Regression



شکل ۱: میزان موارد ابتلا جدید به صورت روزانه

کمترین میزان خطا را داشتند در اینجا می‌آوریم. این الگوریتم‌های رگرسیون یادگیری ماشین در نرم‌افزار وکا و با هدف پیش‌بینی مقدار نرخ شیوع R_0 روی داده‌های کشور ایران پیاده‌سازی شده است.

۳. یافته‌ها

در جدول ۱ سه الگوریتم رگرسیون پرسپترون چند لایه، رگرسیون بردار پشتیبان و رگرسیون خطی بر اساس چهار معیار ارزیابی با یکدیگر مقایسه شده‌اند. این چهار معیار عبارت‌اند از، خطای میانگین [6] یا MAE و خطای مطلق وابسته یا RAE و ریشه مربع خطای وابسته یا RRSE و ریشه خطای میانگین مربعات یا RMSE که از فرمول‌های نشان داده شده در شکل ۲ محاسبه می‌شوند، به طوری که p ها مقادیر پیش‌بینی شده و a ها مقادیر واقعی و n تعداد کل نمونه‌های به کار رفته هستند. طبق این جدول برای داده‌های جمع‌آوری شده در کشور ایران بهترین الگوریتم‌ها، رگرسیون پرسپترون چند لایه (MLPR) با کمترین میزان خطای ۰.۱۱۷۲ و بعد از آن رگرسیون بردار پشتیبان (SVR) با خطای ۰.۱۵۶۵ هستند. همانطور

Measure	Equation
Mean-absolute error (MAE):	$\mathrm{MAE} = \frac{\lvert p_1 - a_1\rvert + \ldots + \lvert p_n - a_n\rvert}{n}$
Relative absolute error (RAE):	$RAE = \frac{\sum_{i=1}^{n} \lvert p_i - a_i\rvert}{\sum_{i=1}^{n} \lvert a_i - \bar{a}\rvert}$
Root relative squared error (RRSE):	$RRSE = \sqrt{\frac{\sum_{i=1}^{n} (p_i - a_i)^2}{\sum_{i=1}^{n} (a_i - \bar{a})^2}}$
Root mean-squared error (RMSE):	$RMSE = \sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}}$

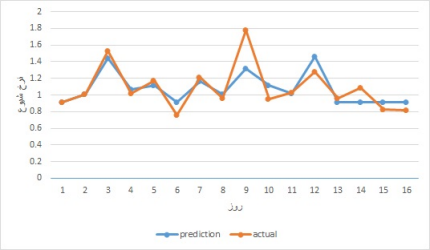
شکل ۲: معیارهای ارزیابی

که در نمودار شکل ۳ مشخص است میزان نرخ شیوع در ۱۵ روز متوالی در مقایسه با میزان پیش بینی شده بر اساس الگوریتم رگرسیون پرسپترون چند لایه اختلاف خیلی کمی دارد. همچنین نمودار شکل ۴ میزان نرخ شیوع در ۱۵ روز متوالی را در مقایسه با میزان پیش بینی شده بر اساس الگوریتم رگرسیون بردار پشتیبان نشان می‌دهد که اختلاف بیشتری نسبت به الگوریتم قبلی دارد. و نمودار شکل ۵ نیز میزان نرخ شیوع در ۱۵ روز متوالی را در مقایسه با میزان پیش‌بینی شده بر اساس الگوریتم رگرسیون خطی نشان می‌دهد که اختلاف بیشتری نسبت به الگوریتم‌های قبلی دارد. لازم به ذکر است تصویر خروجی نرم افزار وکا بعد از پیاده‌سازی الگوریتم رگرسیون پرسپترون چند لایه در شکل ۶ آورده شده است.

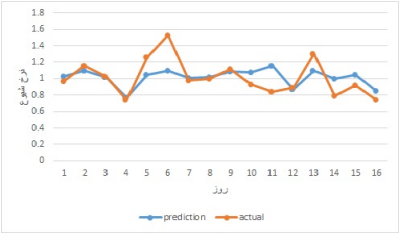
[6] Mean absolute error

جدول ۱: مقایسه الگوریتم‌ها بر اساس خطای میانگین

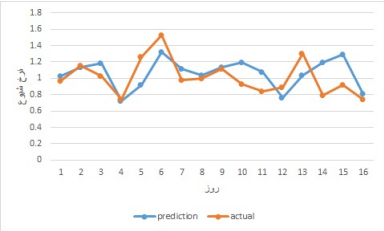
الگوریتم	MAE	RAE	RRSE	RMSE
رگرسیون پرسپترون چند لایه	۰/۱۱۷۲	۵۷/۰۶۳۲	۳۵/۷۱۷۵	۰/۱۵۶۹
رگرسیون بردار پشتیبان	۰/۱۵۶۵	۷۷/۹۲۸۶	۹۶/۸۹۸۴	۰/۴۲۵۷
رگرسیون خطی	۰/۱۹۱۷	۹۵/۹۱۴۹	۹۳/۱۹۴۴	۰/۴۰۹۴



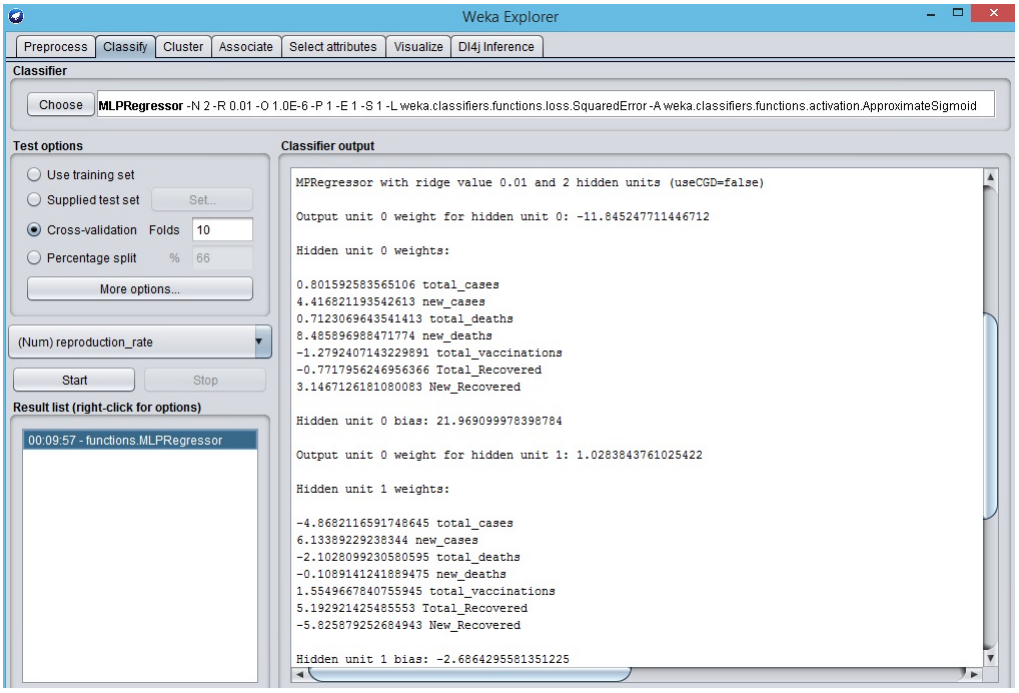
شکل ۳: میزان نرخ شیوع پیش‌بینی شده با الگوریتم رگرسیون پرسپترون چند لایه در مقایسه با نرخ شیوع واقعی



شکل ۴: میزان نرخ شیوع پیش‌بینی شده با الگوریتم رگرسیون بردار پشتیبان در مقایسه با نرخ شیوع واقعی



شکل ۵: میزان نرخ شیوع پیش‌بینی شده با الگوریتم رگرسیون خطی در مقایسه با نرخ شیوع واقعی



شکل ۶: خروجی نرم افزار وکا بعد از پیاده‌سازی الگوریتم رگرسیون پرسپترون چند لایه

۴. بحث و نتیجه‌گیری

در این مطالعه تجزیه و تحلیل داده‌های آماری بیماری کووید -۱۹ در ایران با استفاده از مدل‌های یادگیری ماشین انجام شده است. نتایج نشان می‌دهد که الگوریتم رگرسیون پرسپترون چند لایه دارای حداقل خطای میانگین نسبت به سایر الگوریتم‌ها در پیش‌بینی نرخ شیوع R_0 برای این بیماری است. پیشنهاد می‌شود برای مطالعات آینده از سایر مدل‌های یادگیری ماشین بهره برد.

مراجع

[۱] زارع، واثق، نسترن. مدلسازی و تحلیل گسترش کووید-۱۹ در ایران با استفاده از مدل کلاسیک SIR. مجله کنترل. ۱۰ Feb ۲۰۲۱؛۱۴(۵):۸۹-۹۶.
[2] Alimohamadi Y, Sepandi M. Basic reproduction number: An important indicator for the future of the COVID-19 epidemic in Iran. Journal of Military Medicine. 2020;22(1):96-7.
[3] Yap FF, Yong M. Implementation of An Online COVID-19 Epidemic Calculator for Tracking the Spread of the Coronavirus in Singapore and Other Countries. medRxiv. 2020 Jan 1.
[4] Batista M. Estimation of the final size of the COVID-19 epidemic. MedRxiv. 2020 Jan 1.
[5] Pandey G, Chaudhary P, Gupta R, Pal S. SEIR and Regression Model based COVID-19 outbreak predictions in India. arXiv preprint arXiv:2004.00958. 2020 Apr 1.
[6] Ahmad A, Garhwal S, Ray SK, Kumar G, Malebary SJ, Barukab OM. The number of confirmed cases of covid-19 by using machine learning: Methods and challenges. Archives of Computational Methods in Engineering. 2021 Jun;28(4):2645-53.
[7] Punn NS, Sonbhadra SK, Agarwal S. COVID-19 epidemic analysis using machine learning and deep learning algorithms. MedRxiv. 2020 Jan 1.

Analysis and prediction of Covid-19 disease using machine learning regression algorithms

1st Mahsa Malekpour Shahraki[1], 2nd Sajjad Rahmany[2]

[1] **Faculty of Mathematics and Computer Science, Damghan University, Damghan, Iran**
[2] **Faculty of Mathematics and Computer Science, Damghan University, Damghan, Iran**

s_rahmani@du.ac.ir

***Abstract*— The pandemic Covid-19 has been one of the biggest challenges that the world has ever faced. So it is important to predict the reproduction rate and rate of contagion of this disease. In this work, machine learning and statistical methods will be studied in order to predict the reproduction rate of the corona viruse in Iran. Results show proper performance of the multilayer perceptron regression algorithm for predicting reproduction rate due to its acceptable error.**

***Keywords*—Covid-19, Reproduction Rate, Prediction, Machine Learning, Regression**