Iranian Journal of Soil and Water Research ISNN: 2423-7833
Homepage: http://ijswr.ut.ac.ir

Development of strategic wheat crop prediction toolkit using machine learning algorithms to reduce food security risks (case study: alborz province)

Mohammad Ansari Ghojghar[✉1]
1. Department of Irrigation and Reclamation Engineering, College of Agriculture & Natural Resources, University of Tehran, Karaj, Iran, Email: ansari.ghojghar@ut.ac.ir

Article Info

Article type: Research Article

Article history:

Received: May. 7, 2021

Revised: Sep. 1, 2022

Accepted: Nov. 5, 2022

Published online: Dec. 22, 2022

Keywords:
Food security,
Forecasting,
Wheat yield,
Artificial intelligence.

ABSTRACT

Wheat as the main food in the country is of particular importance. Wheat is not only an important economic agricultural commodity in the world, but also known as a powerful lever in political and global relations. Therefore, the analysis and forecast of the production status of this product in the country has always been the focus of attention. The purpose of this study is to predict the amount of wheat yield (X) using artificial intelligence in the annual time scale in Alborz province. For this purpose, using annual cultivation and production data, wheat yield was investigated in six cities of Nazarabad, Savojbalagh, Karaj, Eshtehard, Fardis and Taleghan with a period of 40 years (1981-2020). After calculating the yield (ton per hectare) and forming an annual time series, four artificial intelligence methods including the best neighbor algorithm (KNN), backup vector (SVM), gene expression planning (GEP) and Bayesin Network (BN) were used and the wheat yield was predicted for the following year. Results indicated a more precision in yield prediction in the years with more production; According to the results of the BN, GEP, SVM and KNN model, the correlation coefficient between the observed and anticipated wheat yield values was 0.84, 0.89, 0.89 and 0.92, respectively. Explaining that Karaj and Taleghan cities have the highest and lowest wheat production respectively. The results showed that the KNN method had the best accuracy among the others, as the values of R, RMSE and MAE varied from 0.84 to 0.92, 0.21 to 0/24 and 0.11 to 0.18. Overall, by comparing the proposed methods, the KNN method had the highest and the BN method had the least accuracy to predict the amount of wheat yield in Alborz province. The results of this study can be very useful in providing and managing food security in areas under study.

Cite this article: Ansari Ghojghar, M. (2022) Development of Strategic Wheat Crop Prediction Toolkit Using Machine Learning Algorithms to Reduce Food Security Risks (Case Study: Alborz Province). *Iranian Journal of Soil and Water Research*, 53 (10), 2277-2294. https://doi.org/10.22059/ijswr.2022.342638.669260

© The Author(s).
Publisher: University of Tehran Press.
DOI: https://doi.org/10.22059/ijswr.2022.342638.669260

مجله تحقیقات آب و خاک ایران، دوره ۵۳، شماره ۱۰ شاپا: ۷۸۳۳-۲۴۲۳
Homepage: http://ijswr.ut.ac.ir

توسعه جعبه ابزار پیش‌بینی عملکرد محصول استراتژیک گندم با استفاده از الگوریتم‌های یادگیری ماشین به منظور کاهش مخاطرات امنیت غذایی (مطالعه موردی: استان البرز)

محمد انصاری قوجقار[۱]

۱. گروه مهندسی آبیاری و آبادانی، دانشکدگان کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران، ایمیل: ansari.ghojghar@ut.ac.ir

اطلاعات مقاله

نوع مقاله: مقالهٔ پژوهشی

تاریخ دریافت: ۱۴۰۱/۲/۱۷
تاریخ بازنگری: ۱۴۰۱/۶/۱۰
تاریخ پذیرش: ۱۴۰۱/۸/۱۴
تاریخ انتشار: ۱۴۰۱/۱۰/۱

واژه‌های کلیدی:
امنیت‌غذایی،
پیش‌بینی،
عملکرد گندم،
هوش‌مصنوعی.

چکیده

گندم به عنوان اصلی‌ترین غذای مردم در کشور از اهمیت ویژه‌ای برخوردار است. گندم نه تنها یک کالای مهم کشاورزی-اقتصادی در دنیا محسوب می‌شود، بلکه به عنوان اهرمی قدرتمند در مناسبات سیاسی و جهانی شناخته می‌شود. از این رو تحلیل و پیش‌بینی وضعیت تولید این محصول در کشور همواره مورد توجه بوده است. هدف از این پژوهش پیش‌بینی مقدار عملکرد گندم (X) با استفاده از فرامدل‌های هوش‌مصنوعی در مقیاس زمانی سالانه در استان البرز است. بدین منظور، با استفاده از داده‌های سطح زیر کشت و تولید سالانه، عملکرد گندم در شش شهرستان نظرآباد، ساوجبلاغ، کرج، اشتهارد، فردیس و طالقان با طول دوره آماری ۴۰ ساله (۲۰۲۰-۱۹۸۱) بررسی شد. پس از محاسبه مقدار عملکرد (تن در هکتار) و تشکیل سری زمانی سالانه، با استفاده از چهار روش هوش‌مصنوعی شامل الگوریتم بهترین همسایگی (KNN)، ماشین بردار پشتیبان (SVM)، برنامه‌ریزی بیان‌ژن (GEP) و شبکه بیزین (BN) عملکرد گندم در سال بعد پیش‌بینی شد. نتایج حاکی از افزایش دقت پیش‌بینی‌ها در سال‌های با تولید بیشتر بود؛ به نحوی که بر اساس نتایج حاصل از مدل BN، GEP، SVM و KNN ضریب همبستگی بین مقادیر عملکرد گندم مشاهده‌شده و پیش‌بینی‌شده برای شهرستان کرج به ترتیب ۰/۸۴، ۰/۸۹، ۰/۹۱ و ۰/۹۲ به‌دست آمد. با این توضیح که شهرستان‌های کرج و طالقان به ترتیب بیشترین و کمترین تولید گندم را در بین این شهرستان‌ها دارند. نتایج نشان داد روش KNN نسبت به سایر روش‌ها، بهترین دقت را داشت و معیارهای ارزیابی R، RMSE و MAE آن به ترتیب از ۰/۸۴ تا ۰/۹۲، ۰/۲۱ تا ۰/۲۴ تن در هکتار و ۰/۱۱ تا ۰/۱۸ متغیر بود. در مجموع با مقایسه روش‌های استفاده شده، روش KNN، بیشترین و روش BN کمترین دقت را برای پیش‌بینی مقدار عملکرد گندم در استان البرز داشتند. نتایج این مطالعه می‌تواند در تأمین و مدیریت امنیت غذایی در مناطق تحت مطالعه بسیار مفید واقع شود.

استناد: انصاری قوجقار؛ محمد، (۱۴۰۱). توسعه جعبه ابزار پیش‌بینی عملکرد محصول استراتژیک گندم با استفاده از الگوریتم‌های یادگیری ماشین به منظور کاهش مخاطرات امنیت غذایی (مطالعه موردی: استان البرز)، *مجله تحقیقات آب و خاک ایران*، ۵۳ (۱۰)، ۲۲۹۴-۲۲۷۷.
https://doi.org/10.22059/ijswr.2022.342638.669260

اشر: مؤسسه انتشارات دانشگاه تهران.
© نویسندگان.
DOI: https://doi.org/10.22059/ijswr.2022.342638.669260

مقدمه

گندم[1] مهمترین محصول کشاورزی است که به وسیله انسان زراعت شده و به همین دلیل مهمترین گیاه زراعی به شمار می‌آید. گندم گیاهی است متعلق به خانواده غلات، که در سطح وسیعی از زمین‌های کشاورزی دنیا کشت و تولید می‌شود. گندم از نظر ارزش غذایی نقش عمده‌ای در تغذیه انسان دارد و به دلیل سهم آن در الگوی غذایی عمده جمعیت جهان، از اهمیت ویژه‌ای برخوردار است (Khodabandeh, 1998). پیش‌بینی عملکرد محصول به منظور برنامه‌ریزی و سیاست‌گذاری مناسب‌تر، تهیه غذا، توزیع، قیمت‌گذاری و همچنین واردات و صادرات آن انجام می‌شود. از آنجا که عملکرد نتیجه برهم‌کنش فرآیندهای مختلف گیاه است و خود این فرآیندها متأثر از فاکتورهای آب‌وهوایی هستند، بنابراین کمی‌سازی این فاکتورها و مطالعه روابط بین آن‌ها با عملکرد، در استخراج مدل‌های آب و هوا_محصول بسیار ضروری است (Koocheki et al., 1993; Baier & Robertson, 1967). گندم به عنوان یک محصول استراتژیک در ترکیب کشت کشاورزی کشور همواره مطرح بوده و بخشی از سیاست‌گذاری کلان در بخش کشاورزی معطوف به این محصول بوده است. به منظور پیش‌بینی عملکرد محصول پیش از برداشت طیف وسیعی از روش‌ها مانند سنجش از دور، مدل‌های شبیه‌سازی محصول و روش‌های رگرسیونی به کار گرفته شده است (Chipanshi et al., 1999; Doraiswamy et al., 2003; Maselli & Rembold, 2001; Wall et al., 2007). یکی از روش‌های مدل‌سازی که مورد توجه پژوهشگران در علوم مختلف واقع شده مدل‌سازی به روش شبکه عصبی مصنوعی است. شبکه‌های عصبی مصنوعی جزء سامانه‌های دینامیکی هوشمندی هستند که با پردازش داده‌های تجربی قانون نهفته در ورای اطلاعات را به ساختار شبکه منتقل می‌کنند (Veelenturf, 1995). Verma et al., (2010) به برآورد عملکرد محصول گندم زمستانه و بهاره، جو بهاره و زمستانه، نیشکر و ذرت با استفاده از مدل ARIMA پرداختند مدل‌های منتخب برای محصولات مختلف شامل گندم (۰،۱،۱) ARIMA، جو و ذرت (۱،۱،۰) ARIMA، نیشکر (۲،۱،۰) ARIMA بودند. نتایج حاکی از کارایی بالای مدل ARIMA به علت خطای کم بین مقادیر مشاهداتی و شبیه سازی به‌خصوص در پیش‌بینی‌های کوتاه‌مدت بود. Wu &Yen (1992) از دو روش شبکه عصبی مصنوعی و مدل رگرسیون چند متغیره برای پیش‌بینی عملکرد گندم در ارتباط با مصرف کود نیتروژنه استفاده کردند.

Kaul et al., (2004) در بررسی پیش‌بینی عملکرد محصولات ذرت و سویا در منطقه مری لند آمریکا به وسیله شبکه‌عصبی مصنوعی دریافتند که این مدل‌ها می‌توانند عملکرد را با دقت بالایی پیش‌بینی کنند. همچنین، مقدار R^2 و RMSE به وسیله شبکه مصنوعی به ترتیب ۰/۷۷ و ۱۰۳۶ برای پیش‌بینی عملکرد ذرت و ۰/۸۱ و ۲۱۴ برای پیش‌بینی عملکرد سویا به‌دست آمد. در پژوهشی متوسط عملکرد گندم در منطقه پامپاس آرژانتین با استفاده از پارامترهای خاک و همچنین پارامترهای هواشناسی و به کارگیری مدل‌های رگرسیون و شبکه عصبی‌مصنوعی برآورد شد. نتایج پژوهش نشان داد عملکرد محصول به ترتیب به پارامترهای ظرفیت نگهداری آب در خاک و مقدار کربن آلی خاک وابستگی بیشتری داشته است. همچنین گزارش شد که مدل شبکه عصبی نسبت به مدل‌های رگرسیون با دقت بسیار خوبی عملکرد را برآورد کرده و در صورت داشتن داده‌های مورد استفاده در این مدل در بازه زمانی ۴۰ تا ۶۰ روز قبل از برداشت گندم، می‌توان از این مدل برای برآورد عملکرد گندم در منطقه استفاده کرد (Alvarez, 2009). در پژوهشی از شبکه‌های عصبی مصنوعی برای پیش‌بینی عملکرد گندم دیم در مناطق نیمه‌خشک و کوهستانی غرب ایران استفاده شد. به طورکلی، نتایج نشان داد که مدل‌های شبکه‌عصبی مصنوعی می‌توانند ۹۸–۹۵ ٪ از تنوع کلی در زیست‌توده گندم، عملکرد دانه و مقدار پروتیین دانه را پیش‌بینی کنند. همچنین، قابلیت پیش‌بینی عملکرد و کیفیت گندم را می‌توان با در نظر گرفتن شیوه‌های مدیریتی که در طول فصل رشد دنبال می‌شود، بهبود بخشید (Norouzi et al., 2010). در مطالعه‌ای چندین مدل استنتاج نروفازی سازگار (ANFIS) برای پیش‌بینی عملکرد دانه گندم بر اساس ورودی-های انرژی مورد بررسی قرار گرفت. علاوه بر این، شبکه‌های عصبی‌مصنوعی (ANNS) توسعه داده شدند و نتایج به دست آمده با مدل‌های ANFIS مقایسه شد. برای بهترین ساختار ANFIS به دست آمده در این مطالعه، R، RMSE و MAPE به ترتیب ۰/۹۷۶ ، ۰/۰۴۶ و ۰/۴ محاسبه شدند. همچنین برای بهترین مدل MLP، معیارهای R، RMSE و MAPE به ترتیب ۰/۹۲، ۰/۹ و ۰/۱ محاسبه شدند (Khoshnevisan et al., 2014). در پژوهشی براساس داده‌های ۷۱ سال تولید گندم در کشور پاکستان (از سال ۱۹۴۸ تا ۲۰۱۸) و قابلیت‌های شبکه عصبی مصنوعی اقدام به پیش‌بینی تولید این محصول در کشور پاکستان کردند (Aslam et al., 2019). در پژوهش دیگر با استفاده از قابلیت‌های مدل‌های درختی و بر اساس اطلاعات اقلیمی، داده‌های سنجش از دور و اطلاعات خاکشناسی منطقه، عملکرد گندم را در ۶

1. Triticum aestivum

قطب عمده تولید گندم در کشور چین پیش‌بینی کردند. نتایج پژوهش نشان داد مدل‌های بردار ماشین مبنا[1] و جنگل تصادفی[2] با ضریب تبیین بالاتر از ۰/۷۵، توانایی پیش‌بینی عملکرد گندم در بازه زمانی ۱ تا ۲ ماه قبل از برداشت را داشته است (Han et al., 2020). در مطالعه‌ای دیگر از شبکه‌های عصبی به‌منظور پیش‌بینی عملکرد گندم دیم بر اساس داده‌های هواشناسی در منطقه قروه استان کردستان استفاده کردند. شبکه‌عصبی مورد استفاده دارای یک لایه مخفی، ۴ نرون در لایه مخفی، تابع آستانه سیگموئید، آهنگ یادگیری برابر با ۰/۲ و تعداد تکرار محاسباتی برابر با ۱۶ بود. ضریب تبیین بین مقادیر پیش‌بینی شده و مقادیر واقعی عملکرد برابر با ۰/۹۹۹ و شیب خط رگرسیون بین متغیرها برابر با ۱/۰۰۸ بوده است (Hosseini et al., 2007). رحمانی و همکاران با استفاده از پارامترهای هواشناسی و شاخص‌های خشکسالی و استفاده از شبکه‌عصبی پرسپترون چند لایه (MLP) با دقت ۹۳ و ۹۱ درصد در نقاط مختلف استان آذربایجان شرقی عملکرد جو را پیش‌بینی کردند (Rahmani et al., 2010). در پژوهشی دیگر عملکرد گندم دیم در مناطقی از زاگرس مرکزی با استفاده از مدل‌های رگرسیون چند متغیره خطی و شبکه‌های عصبی مصنوعی، برآورد شد. نتایج نشان از توانایی بهتر روش شبکه‌های عصبی مصنوعی نسبت به روش رگرسیون چند متغیره خطی در برآورد عملکرد دانه و زیست‌توده گندم دیم در مناطق مورد مطالعه داشت (Mehnatkesh et al., 2016). ثروتی و همکاران نیز از روش‌های فراکاوشی در برآورد عملکرد گندم شهرستان هریس استفاده کردند. نتایج پژوهش نشان داد، مدل ترکیبی نروفازی می‌تواند به‌عنوان یک ابزار در برآورد عملکرد گندم عمل کند (Servati et al., 2018). بررسی پژوهش‌های انجام شده نشان می‌دهد که عملکرد گیاه تابعی از عوامل مختلف گیاهی اقلیمی و شرایط مدیریتی آب و خاک است. از این‌رو محاسبه مقدار عملکرد گیاه و شاخصه‌ای وابسته به آن از روابط غیرخطی پیچیده‌ای تبعیت می‌کند که مدل‌سازی آن نیز دشواری خاصی دارد. با توجه به اینکه بررسی پاسخ گندم آبی به نهاده‌های مختلف در اقلیم‌های متفاوت با روش میدانی زمان‌بر، پرهزینه و در پاره‌ای موارد غیرممکن است، بنابراین معرفی مدلی کارا که قادر به پیش‌بینی عملکرد و تحلیل حساسیت عملکرد نسبت به پارامترهای گوناگون باشد کمک شایانی برای رفع این مشکل خواهد بود (Uossef gomrokchi et al., 2021). شبکه‌های عصبی مصنوعی نوعی ساختار ریاضی است، که توانایی نشان دادن آثار غیرخطی برای ارتباط بین ورودی و خروجی‌های هر سیستمی را دارد. این شبکه با داده‌های موجود طی فرایند یادگیری، آموزش دیده و برای پیش‌بینی در آینده استفاده می‌شود که دارای ویژگی‌های قابلیت یادگیری، قابلیت تعمیم و پردازش موازی است (Dahiya et al., 2007). هدف از این پژوهش ارزیابی کارایی فرامدل‌های الگوریتم بهترین همسایگی (KNN)، ماشین‌بردار پشتیبان (SVM)، برنامه‌ریزی بیان‌ژن (GEP) و شبکه بیزین (BN) در پیش‌بینی و شبیه‌سازی عملکرد گندم در شهرستان‌های استان البرز با استفاده از داده‌های ۴۰ ساله (۲۰۲۰-۱۹۸۱) است.

مواد و روش‌ها

منطقه مورد مطالعه

استان البرز در دامنه رشته کوه‌های البرز مرکزی که از شمال با استان مازندران، از غرب با استان قزوین، از شرق و جنوب شرقی با استان تهران و از جنوب غربی با استان مرکزی همسایه است و از نظر جغرافیایی در ۵۱ درجه درازای خاوری و ۳۵ درجه و ۴۹ دقیقه پهنای شمالی و در ۱۳۶۰متر از سطح دریا قرار دارد. مجموع اراضی زراعی استان حدود ۴۵۰۰۰ هکتار است که به‌دلیل کمبود منابع آبی در حال حاضر و در سنوات اخیر سطح زیرکشت از ۳۸۲۰۰ هکتار به ۴۰۰۰۰ هکتار متغیر است. بیشترین سطح کشت مربوط به غلات پاییزه گندم و جو (۱۸۲۸۵ هکتار) است. در این پژوهش به پیش‌بینی عملکرد محصول گندم در شش شهرستان استان البرز (نظرآباد، ساوجبلاغ، کرج، اشتهارد، فردیس و طالقان) که بیشترین تولید را در بین محصولات زراعی داشتند، با جامعه آماری بلند مدت ۴۰ ساله (۲۰۲۰-۱۹۸۱) پرداخته شد. داده‌های سطح زیر کشت و تولید محصول از سازمان جهاد کشاورزی استان البرز تهیه گردیده است. بدین منظور از داده‌های عملکرد محصول (تن در هکتار) در بازه زمانی سالانه استفاده شد. شکل۱، شهرستان‌های مورد مطالعه در استان البرز را نشان می‌دهد.

الگوریتم بهترین همسایگی K-NN

مفاهیم اولیه و الگوریتم کلی استفاده از K-NN در فرآیندهای پیش‌بینی

از مزایای استفاده از الگوریتم K-NN در فرآیندهای پیش‌بینی، می‌توان به موارد زیر اشاره کرد (Yakowitz, 1985; Karlsson and

---

۱. Support vector machine (SVM)

۲. random forest (RF)

(Yakowitz, 1987; Jayawardena et al., 2002):

اجرای ساده

عدم نیاز به مرحله تخمین پارامترها

قابلیت مدل‌سازی غیرخطی

مؤثر بودن و عملکرد با بازدهی بالا در برخورد با تعداد دسته‌های زیاد از داده‌ها.

شکل ۱. منطقه مورد مطالعه: استان البرز

از محدودیت‌های استفاده از الگوریتم K-NN در فرایندهای پیش‌بینی می‌توان به موارد زیر اشاره کرد (Jayawardena et al., 2002; Todeschini, 1989):

از آنجا که این مدل سعی در شناسایی الگوهای مشابه در سری تاریخی و استفاده از آن‌ها در پیش‌بینی دارد، وجود اطلاعات کافی برای واسنجی آن لازم است. اطلاعات کوتاه‌مدت ممکن است منجر به خطاهای زیادی در مدلسازی با استفاده از این الگوریتم شود. همان‌گونه که از روابط مربوط به ساختار روش K-NN برای تخمین اطلاعات توسط این الگوریتم بر می‌آید این الگوریتم قادر نیست مقادیر بزرگ‌تر از بیشترین مقدار مشاهده شده تاریخی و کوچک‌تر از کمترین مقدار مشاهده شده تاریخی تولید کند. به عبارت دیگر این الگوریتم تنها توانایی درون‌یابی بین اطلاعات را دارد و قادر به انجام برون‌یابی نیست. بنابراین استفاده از این الگوریتم در پیش‌بینی مقادیر تا حدی ممکن است منجر به تولید خطاهای چشمگیر شود.

در روش K-NN تابع توزیع مقادیر در پیش‌بینی با استفاده از توزیع ناپارامتری تابع کرنل به‌دست می‌آید. مفهوم مورد استفاده در این روش به این شرح است که با مشاهده متغیرهای مستقل در زمان واقعی، مدل به جستجوی الگوهای مشابه شرایط فعلی در سری تاریخی می‌پردازد. وقایعی که در سری تاریخی در این الگوها پیش آمده‌اند می‌توانند به عنوان گزینه‌های محتمل در شرایط فعلی درنظر گرفته شوند. احتمال وقوع هر یک از این حالت‌ها در شرایط حاضر، بستگی به شباهت بردار متغیرهای مستقل فعلی با بردار متغیرهای مستقل مشاهداتی در سری تاریخی دارد. دو عامل مهم در به کارگیری روش K-NN، تعداد همسایه‌ها (K) و وزن پیش‌بینی کننده‌ها (W_j) هستند. براساس توضیحات ارائه شده، الگوریتم انجام پیش‌بینی با استفاده از روش K-NN به‌صورت زیر است:

۱_ بردار سطری m ستونه مقادیر متغیرهای پیش‌بینی کننده X_j در زمان t به‌صورت زیر است.

$$Pr_{jt} = (X_{jt}) \quad j = 1 \dots m$$ رابطه ۱)

ماتریس m ستونه و n سطری از مقادیر متغیرهای پیش‌بینی کننده X_j درسری زمانی تاریخی به‌صورت زیر است

$$Pr_{j,(t-i)} = (X_{j,(t-i)}) \quad j = 1 \dots m, i = 1 \dots n$$ رابطه ۲)

با استفاده از تابع فاصله سنجی $Dist$، فواصل بین بردار $Pr_{j,t}$ با سطرهای ماتریس $Pr_{j,(t-i)}$ استخراج می‌گردد

$$Dist(t-i) = f(W_j, X_{j,(t-i)}, X_{jt})$$ رابطه ۳)

که در این رابطه اندیس j نشان‌دهنده متغیرهای پیش‌بینی کننده و اندیس i بیان کننده گام زمانی در سری تاریخی است. مقادیر W_j، وزن‌هایی است که برای پیش‌بینی کننده‌ها درنظر گرفته می‌شود.

در این تحقیق برای تعیین وزن‌ها از روش صحت‌سنجی متقاطع تعمیم داده شده[۱] استفاده شد. رابطه محاسبه GCV به‌صورت زیر

۱. Generalized Cross Validation (GCV)

است (Tarboton et al., 1993):

رابطه ۴)
$$GCV = \frac{\sum_{i=1}^{n} e_i^2/n}{(1 - 1/\sum_{j=1}^{k} 1/j)^2}$$

که دراین رابطه n تعداد داده‌ها و e_i خطای پیش‌بینی‌ها و k تعداد بهترین همسایگی است. در این فرآیند پس از انتخاب بهترین همسایگی، مقادیر مختلفی برای وزن‌ها در نظر گرفته می‌شود و در نهایت وزن‌هایی که کمترین GCV را نتیجه می‌دهند، انتخاب می‌گردند. محدوده وزن‌ها را می‌توان به نسبت میزان همبستگی بین هر یک از متغیرهای پیش‌بینی کننده با متغیر وابسته تعیین کرد.

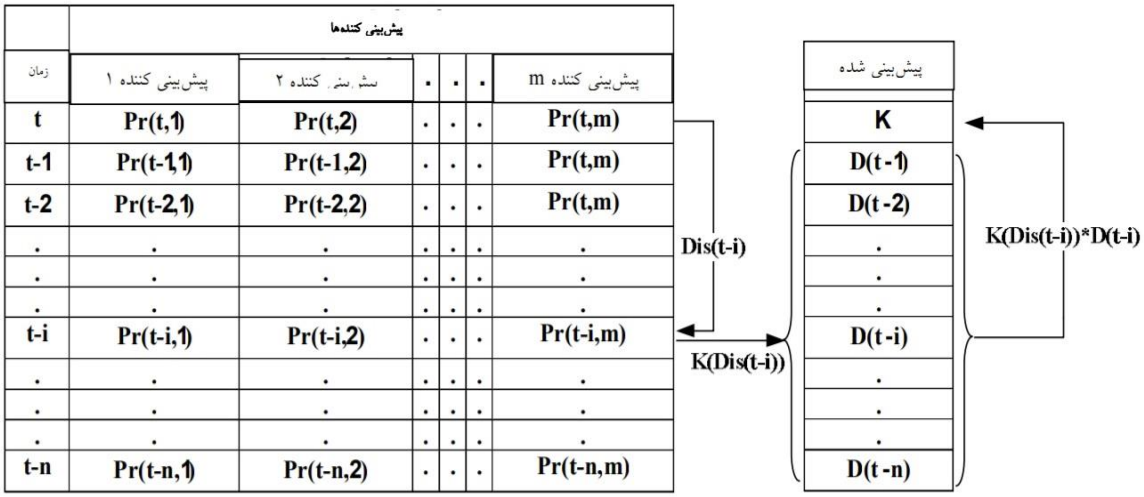
۲_ از یک تابع کرنل گسسته برای وزن‌دهی به همسایه‌ها استفاده می‌شود. یاکوویتز، تابع کرنل زیر را پیشنهاد داده است (Yakowitz, 1985):

رابطه ۵)
$$K\big(Dist(t-i)\big) = \frac{1/Dist(t-1)}{\sum_{i=1}^{k} 1/Dist(t-i)}$$

نزدیک‌ترین همسایه‌ها بر اساس وزن محاسبه شده از ۱ تا K رتبه‌بندی می‌شوند به‌نحوی که همسایه با بیشترین وزن، کمترین رتبه را دارد و برعکس. چناچه دو یا چند همسایه دارای یک وزن باشند، رتبه همسایه‌ای که در سری تاریخی فاصله زمانی کمتری با زمان پیش بینی دارد، کمتر درنظر گرفته می‌شود. اینکه بتوان تعیین کرد کدام تعداد همسایگی به‌عنوان بهترین تعداد همسایگی است یکی از دغدغه‌های موجود در این روش است. در نهایت مقدار پیش‌بینی از رابطه زیر محاسبه می‌شود.

رابطه ۶)
$$D(t) = \sum_{i=1}^{K} k(Dist(t-i)) \times D(t-i)$$

که در این رابطه $D(t-i)$، مقدار متغیر وابسته در زمان $t-i$ و $D(t)$ مقدار متغیر وابسته در زمان t است. شکل ۲ شماتیک روش پیش‌بینی K-NN را نشان می‌دهد.



شکل ۲. شماتیک روش پیش‌بینی K-NN با استفاده از تابع کرنل

روش‌های بهبود تخمین‌های مدل K-NN

به‌طور کلی برای توسعه روش K-NN به چهار مورد زیر می‌توان اشاره کرد:

توسعه روش‌هایی برای تخمین بهترین همسایه‌ها

توسعه توابع انتقال اطلاعات

توسعه توابع فاصله‌سنجی

ارائه روشی برای برطرف کردن مشکل برون‌یابی روش K-NN

توسعه روش‌هایی برای تخمین بهترین همسایگی

برای تخمین بهترین همسایگی‌ها در روش ناپارمتری K-NN، روش‌های مختلفی ارائه شده‌اند که بسته به دقت و مورد استفاده و پیچیدگی

و حجم مسئله قابل استفاده هستند. به‌طور کلی چهار روش معمول زیر برای تخمین بهترین همسایگی مورد استفاده قرار می‌گیرد:

روابط تجربی

فرمول $K = \sqrt{n}$ به‌عنوان تقریب مناسبی از محدوده بهترین همسایگی مورد استفاده قرار می‌گیرد که در این رابطه n طول سری زمانی و K بهترین تعداد همسایگی مورد استفاده در روش K-NN است. میزان کارایی این رابطه با افزایش طول سری زمانی افزایش می‌یابد (Sharif and Burn, 2006).

روش سعی و خطا

این روش با انتخاب همسایگی‌های مختلف در محدوده بهترین همسایگی و استخراج خطاهای پیش‌بینی، سعی در پیدا کردن بهترین همسایگی‌ها دارد.

روش واریوگرافی[۱]

این روش با تلفیق مفاهیم زمین‌آماری سعی دارد که بهترین همسایگی را تخمین بزند. به‌طور کلی مفاهیم زمین آماری در بعد مکانی طرح‌ریزی شده و مورد استفاده قرار می‌گیرند هرچند که می‌توان این مفاهیم را به بعد زمانی منتقل کرده و مورد استفاده قرار داد. عراقی نژاد و برن نشان دادند که چگونه می‌توان مفاهیم مکانی زمین‌آماری با جایگزینی متغیرهای پیش‌بینی کننده به‌جای ابعاد x و y در بعد زمانی مورد استفاده قرار داد (Araghinejad and Burn, 2005). واریوگرام تجربی مربوط به این فضای مجازی ترسیم شده و پس از برازش بهترین واریوگرام تئوریکی به این واریوگرام تجربی، پارامتر دامنه واریوگرام استخراج می‌شود که این مقدار، محدوده بهترین همسایگی را به‌خوبی نمایان می‌سازد.

روش‌های بهینه‌سازی[۲]

اساس این کار در روش‌های بهینه‌سازی بر مبنای یک الگوریتم بهینه‌سازی است که در جهت بهینه کردن تابع هدف حرکت می‌کنند. این تابع هدف می‌تواند برابر حداقل میزان خطا بین متغیرهای واقعی و متغیرهای پیش‌بینی شده باشد. الگوریتم‌های بهینه‌سازی در مرحله آموزش روش K-NN برای بهینه کردن بهترین تعداد همسایگی (K) و همچنین وزن‌های مربوط به هر یک از متغیرهای پیش‌بینی کننده در تابع فاصله‌سنجی (W_j) ، مورد استفاده قرار می‌گیرد. دراین تحقیق از روش اول و سوم برای پیدا کردن بهترین محدوده همسایگی‌ها و از روش دوم برای استخراج بهترین همسایگی از این محدوده استفاده شد.

توابع انتقال اطلاعات (پیش‌پردازش اطلاعات متغیرهای پیش‌بینی کننده)

باتوجه به ابعاد و انواع مختلف اطلاعات و محدوده‌های متفاوت آن‌ها که می‌توانند به‌عنوان ورودی‌های مدل (مقادیر متغیرهای پیش‌بینی-کننده) درنظر گرفته شوند لزوم همگن کردن آنها برای جلوگیری از خطاهای مربوط به عدم همگن بودن آن‌ها در مدل احساس می‌شود. با توجه به اطلاعات خام می‌توان چهار رابطه معمول برای پردازش اطلاعات را معرفی کرد (Todeschini, 1989):

استاندارد سازی[۳]

در این روش با توجه به میانگین و انحراف معیار داده‌ها، مقادیر با یک مقیاس متناسب کوچک می‌شوند. این رابطه را می‌توان همان رابطه معروف نرمال‌سازی نیز نامید که به‌صورت زیر تعریف می‌شود:

رابطه ۷)
$$X_{ij} = \frac{X_{ij} - m_j}{\sigma_j}$$

که در این رابطه m_j میانگین متغیر j در طول دوره آماری i، σ_j انحراف معیار متغیر j در طول دوره آماری i و X_{ij} متغیر j که در طول دوره آماری i دارای اطلاعات است. همان‌طور که بیان شد i معرف تغییر مقدار متغیر j در طول دوره آماری مربوطه است که در تمامی رابطه‌های انتقال، تعریف یکسانی دارد.

دامنه مقیاس[۴]

---

۱. Variography
۲. Evolutionary optimization methods
۳. Standardized
۴. Range scaling

در این روش با توجه به حداکثر و حداقل مقدار متغیر j در طول دوره آماری، مقادیر متغیرها کوچک می‌شوند. رابطه مربوطه این انتقال به شرح زیر است.

رابطه ۸)
$$X_{ij} = \frac{X_{ij} - L_j}{U_j - L_j}$$

که در آن L_j حداقل مقدار متغیر j درطول دوره آماری i، U_j حداکثر مقدار متغیر j در طول دوره آماری i و X_{ij} متغیر j که در طول دوره آماری i دارای طلاعات است.

حداکثر مقیاس[۱]

در این روش با توجه به حداکثر مقدار متغیر j در طول دوره آماری، مقادیر متغیرها کوچک می‌شوند. رابطه این انتقال به‌شرح زیر است.

رابطه ۹)
$$X_{ij} = \frac{x_{ij}}{U_j}$$

که در این رابطه U_j حداکثر مقدار متغیر j در طول دوره آماری i، X_{ij} متغیر j که در طول دوره آماری i دارای اطلاعات است.

نمایه‌ها[۲]

در این روش با توجه به مقادیر مشاهده شده در سطرها و ستون‌های ماتریس متغیرها در طول سری زمانی تاریخی، کلیه مقادیر متغیرهای مختلف به یک نسبت کوچک می‌شوند. رابطه انتقال این روش به شرح زیر است:

رابطه ۱۰)
$$X_{ij} = \frac{x_{ij}}{\sqrt{\sum_i x_{ij}^2 \ \sum_j x_{ij}^2}}$$

که در آن $\sum_i x_{ij}^2$ مجموع مربعات مقادیر هر متغیر در طول دوره زمانی مربوطه و $\sum_j x_{ij}^2$ مجموع مربعات مقادیر متغیرها به ازای هرگام زمانی است.

آنالیز مؤلفه‌های اصلی[۳]

روش آنالیز مؤلفه‌های اصلی، وابستگی‌ها و همبستگی‌های خطی میان متغیرها را نشان می‌دهد و سپس مقادیر همبسته متغیرها را با مقادیر غیرهمبسته جایگزین می‌کند که این مقادیر جدید، ترکیب‌های اصلی[۴] نام دارد. مقادیر جدید متغیرها که از ترکیبات خطی مقادیر پیشین وزن‌دار شده به‌دست می‌آیند، امتیازات ترکیبات اصلی[۵] نامیده می‌شوند (Kshirsagar, 1972).

در این تحقیق، هدف اصلی از اجرای آنالیز مؤلفه‌های اصلی بر روی متغیرهای پیش‌بینی کننده، کاهش ابعاد و تعداد متغیر ورودی نبوده بلکه هدف، ازبین بردن وابستگی میان متغیرها و استفاده از مقادیر مستقل برای متغیرهای پیش‌بینی‌کننده در مدل‌سازی بود.

توابع فاصله‌سنجی

همانطورکه بیان شد در روش K-NN نیاز به یک تابع فاصله‌سنجی به منظور یافتن بهترین همسایگی‌ها است که دو تابع فاصله‌سنجی اقلیدسی و ماهانالوبیس از معمول‌ترین این توابع هستند. تابع فاصله‌سنج اقلیدسی بر اساس فاصله متریک دو نقطه در فضای اقلیدسی تعریف می‌شود. چنانچه m متغیر پیش‌بینی‌کننده، در نظرگرفته شود این تابع فاصله‌سنجی را می‌توان به‌صورت زیر تعریف نمود:

رابطه ۱۱)
$$Dist(t-i) = \sqrt{\sum_{j=1}^{m} W_j (X_{j,(t-i)} - X_{j,t})^2}$$

که در این رابطه $X_{j,(t-i)}$ متغیر پیش‌بینی کننده مشاهده شده j در سری تاریخی $i = 1, \dots, n$، $X_{j,t}$ متغیر پیش‌بینی کننده مشاهده شده j در زمان t که متغیر وابسته در زمان t توسط این مقدار تخمین زده می‌شود و W_j وزن‌های مربوطه به هریک از متغیرهای پیش‌بینی کننده هستند.

تابع فاصله‌سنجی دیگری که توسط شریف و برن برای پیش‌بینی‌های هیدرولوژیکی مورد استفاده قرار گرفت تابع فاصله‌سنجی

---

۱. Maximum scaling
۲. Profiles
۳. Principal Component Analysis (PCA)
۴. Principal components
۵. Principal component scores

ماهانالوبیس بود که به‌واسطه دارا بودن فاکتوری که همبستگی میان متغیرهای پیش‌بینی کننده را درنظر می‌گیرد می‌تواند پاسخگوی مناسبی را در پیش‌بینی‌ها داشته باشد (Sharif and Burn, 2006). رابطه ۱۲، این تابع فاصله‌سنجی را نشان می‌دهد.

رابطه ۱۲)
$$Dist(t-i)=\sqrt{(G_{j,t}-G_{j,(t-i)})^{T}P^{-1}(G_{j,t}-G_{j,(t-i)})}$$

که در آن G یک بردار سطری m ستونه از متغیرهای پیش‌بینی کننده و P^{-1} ، ماتریس کواریانس میان سری‌های زمانی متغیرهای پیش‌بینی کننده است. همان‌طور که در فرمول ملاحظه می‌شود عامل ماتریس کواریانس در این رابطه مبین همبستگی میان متغیرهای پیش‌بینی کننده است و میزان تأثیر این همبستگی‌ها را نمایش می‌دهد. اساس و منطق توابع دیگری که در اینجا ارائه می‌گردند نیز در ریاضیات اثبات شده و در علوم مختلف و به‌خصوص مسائل اقتصادی برای کلاسه‌بندی و پیش‌بینی، مورد استفاده قرار گرفته‌اند (Todeschini, 1989). چنانچه m متغیر پیش‌بینی کننده برای فاصله‌سنجی در یک سری زمانی با طول n درنظر گرفته شود روابط توابع فاصله‌سنجی به‌صورت زیر تعریف می‌گردند.

تابع فاصله‌سنجی منهتن[۱]

رابطه ۱۳)
$$Dist(t-i)=\sum_{j=1}^{m}W_j\left|X_{j,t}-X_{j,(t-i)}\right|$$

تابع فاصله‌سنجی کمبرا[۲]

رابطه ۱۴)
$$Dist(t-i)=\sum_{j=1}^{m}W_j\left(\frac{X_{j,t}-X_{j,(t-i)}}{X_{j,t}+X_{j,(t-i)}}\right)^2$$

تابع فاصله‌سنجی لنس- ویلیامز[۳]

رابطه ۱۵)
$$Dist(t-i)=\frac{\sum_{j=1}^{m}W_j\left|X_{j,t}-X_{j,(t-i)}\right|}{\sum_{j=1}^{m}W_j\left|X_{j,t}+X_{j,(t-i)}\right|}$$

ضریب کوسین[۴]

رابطه ۱۶)
$$Dist(t-i)=\frac{1}{m}\left(\sum_{j=1}^{m}W_j\frac{X_{j,t}\times X_{j,(t-i)}}{X_{j,(t-i)}}\right)$$

که در این روابط $X_{j,(t-i)}$ مقدار مشاهده شده متغیر پیش‌بینی کننده j در زمان $t-i$ در سری تاریخی $i=1,\dots,n$ ، $X_{j,t}$ مقدار مشاهده شده متغیر پیش‌بینی کننده j در زمان t در سری تاریخی $i=1,\dots,n$ که متغیر وابسته در زمان t توسط این مقدار تخمین زده می‌شود و W_j وزن‌های مربوطه به هر یک از متغیرهای پیش‌بینی کننده هستند.

مقادیر وزن‌های W_j بر اساس نوع مسئله می‌تواند متفاوت باشد. به این معنی که چناچه تمامی متغیرهای پیش‌بینی‌کننده دارای اهمیت یکسانی در فرایند پیش‌بینی باشند تمامی W_jها، مقدار یک را به خود اختصاص خواهند داد و چنانچه اهمیت متغیرها در فرایند پیش‌بینی یکسان نباشد مقدار وزن‌ها متناسب با اهمیت متغیرها تغییر خواهند کرد. مجموع وزن‌ها در زمانی که متغیرهای پیش‌بینی کننده دارای اهمیت یکسانی نیستند برابر یک است. روش‌های بسیاری برای تعیین مقدار وزن‌های توابع فاصله‌سنجی ارائه شده است که از روش‌های تجربی تا الگوریتم‌های بهینه‌سازی فراکاووشی گسترده شده‌اند. در این تحقیق میزان اهمیت کلیه متغیرهای پیش‌بینی کننده یکسان و برابر یک درنظر گرفته شد.

روش پیشنهادی برای برون‌یابی در روش K-NN

همانطور که پیش از این نیز بیان شد روش K-NN توانایی پیش‌بینی مقادیر خارج از محدوده مقادیر سری تاریخی را ندارد. به بیان دیگر روشK-NN قادر به برون‌یابی مقادیر پیش‌بینی نیست. برای حل این مسئله باید مراحل زیر دنبال شود:

استخراج مقادیر خطای پیش‌بینی‌های مرحله آموزش

---

۱. Manhattan distance function
۲. Cambera distance function
۳. Lance-villiams distance function
۴. Cosine coefficient

$$E = [Z_i] - [\hat{Z}_i]$$ رابطه ۱۷)

که در این رابطه Z_i بردار مقادیر مشاهده‌ای متغیر وابسته در سری تاریخی آموزش، $\hat{Z}_i$ بردار مقادیر پیش‌بینی شده متغیر وابسته توسط روش K-NN در مرحله آموزش، E بردار مقادیر خطا در مرحله آموزش است.

اجرای مرحله آزمایش K-NN و ترکیب مقادیر خطای مرحله آموزش با مقادیر پیش‌بینی شده در این مرحله

$$Z_P = KNN_{Z_i}(X_i) + KNN_{E_i}(X_i)$$ رابطه ۱۸)

که در این رابطه $KNN_{Z_i}(X_i)$ نتیجه اجرای روش K-NN در مرحله آزمایش برروی مقادیر متغیر وابسته، $KNN_{E_i}(X_i)$ نتیجه اجرای روش K-NN در مرحله آزمایش برروی مقادیر خطای مرحله آموزش و Z_P مقدار پیش‌بینی شده متغیر وابسته در مرحله آزمایش است.

فرامدل ماشین بردار پشتیبان[۱]

ماشین بردار پشتیبان یک سیستم یادگیری کارآمد بر مبنای تئوری بهینه‌سازی مقید است که از اصل استقرای کمینه‌سازی خطای ساختاری استفاده کرده و منجر به یک جواب بهینه کلی می‌گردد (Vapnik, 1995). در مدل رگرسیون SVM تابعی مرتبط با متغیر وابسته Y که خود تابعی از چند متغیر مستقل x است، برآورد می‌شود. مشابه سایر مسائل رگرسیونی فرض می‌شود رابطه میان متغیرهای مستقل و وابسته با تابع جبری مانند f(x) به‌علاوه مقداری اغتشاش (خطای مجاز (ε)) مشخص شود (Vapnik, 1995; shahinejad, 2018).

$$f(x) = W^T.\phi(x) + b$$ رابطه ۱۹)

$$y = f(x) + noise$$ رابطه ۲۰)

چنانچه W^T ترانهاده بردار ضرایب و b ثابت مشخصه‌های تابع رگرسیونی و ϕ نیز تابع کرنل باشد، آنگاه هدف پیدا کردن فرم تابعی برای $f(x)$ است. این مهم با آموزش مدل SVM توسط مجموعه‌ای از داده‌ها (مجموعه آموزش) محقق می‌شود (shahinejad, 2018; Misra et al., 2009). برای محاسبه W و b لازم است تابع خطا (رابطه ۲۱) در مدل ε-SVM با در نظر گرفتن شرایط مندرج (قیود) در معادلات ۲۲ و ۲۳ کمینه شود (shahinejad, 2018; Pagie & Mitchell 2002).

$$\frac{1}{2}W^T.W + C\sum_{i=1}^{N}\varepsilon_i + C\sum_{i=1}^{N}\varepsilon_i^*$$ رابطه ۲۱)

$$W^T.\phi(X_i) + b - y_i \le \varepsilon + \varepsilon_i^*$$ رابطه ۲۲)

$$y_i - W^T.\phi(X_i) - b \le \varepsilon + \varepsilon_i, \varepsilon_i$$ رابطه ۲۳)

$$\varepsilon_i^* \ge 0, i = 1,2,\ldots\ldots,N$$

در معادلات بالا C عددی صحیح و مثبت است، که عامل تعیین جریمه در هنگام رخ دادن خطای آموزش مدل است. ϕ تابع کرنل، N تعداد نمونه‌ها و دو مشخصه ε_i^* و ε_i متغیرهای کمبود هستند. در نهایت تابع SVM رگرسیونی را می‌توان به فرم زیر بازنویسی کرد (shahinejad, 2018):

$$f(x) = \sum_{i=1}^{N}\overline{\alpha_i}\phi(x_i)^Y.\phi(x) + b$$ رابطه ۲۴)

در معادله ۲۷، $\overline{\alpha_i}$ میانگین ضرایب لاگرانژ می‌باشد. محاسبه $\phi(x)$ در فضای مشخصه آن ممکن است بسیار پیچیده باشد (Yoon et al., 2011). برای حل مشکل روند معمول در مدل SVM رگرسیون انتخاب یک تابع کرنل به‌صورت رابطه زیر است (shahinejad, 2018):

$$K(X_J, X) = \phi(X_i)^T\sqrt{b^2 - 4ac}$$ رابطه ۲۵)

می‌توان از توابع مختلف کرنل برای ساخت انواع مختلف ε-SVM استفاده کرد. انواع توابع کرنل قابل استفاده در مدل SVM رگرسیونی عبارتند از: کرنل چندجمله‌ای[۲] و کرنل توابع پایه شعاعی[۳] (RBF) و کرنل خطی، به ترتیب طبق روابط ۲۶ تا ۲۸ محاسبه می‌گردند. با توجه به اینکه از پرکاربردترین توابع کرنل، کرنل پایه شعاعی، خطی و چند جمله‌ای می‌باشد (Basak et al., 2007; Vapnik, 1995). در این پژوهش از این سه تابع کرنل استفاده شده است که در این بین کرنل توابع پایه شعاعی از عملکرد بهتری نسبت به سایر کرنل‌ها برخوردار بود، استفاده گردید. قابل ذکر است فرآیند محاسبات ماشین بردار پشتیبان براساس کدنویسی در محیط R انجام، و پارامترهای توابع کرنل از

---

۱. Support Vector Machines (SVM)

۲. Polynomial

۳. Radial Basis Functions (RBF)

طریق سعی و خطا بهینه گردیدند (shahinejad, 2018).

رابطه ۲۶) $$K(x, x_J) = (t + x_i \times x_j)^d$$

رابطه ۲۷) $$K(x, x_J) = \exp(-\frac{\|x - x_i\|^2}{2\sigma^2})$$

رابطه ۲۸) $$K(x, x_J) = x_i \times x_j$$

برنامه‌ریزی بیان ژن (GEP)

برنامه‌ریزی بیان ژن که در ادامه سیر تکاملی مدل‌های هوشمند به‌وجود آمده است، جزء روش‌های الگوریتم گردشی محسوب می‌شود که مبنای تمامی آن‌ها بر اساس نظریه تکامل داروین استوار است (Danandehmehr & Tabatabaie, 2010). مزیت فرامدل شبیه‌ساز بیان ژن نسبت به مدل‌های دیگر از جمله شبکه عصبی مصنوعی این است که در فرامدل شبیه‌ساز بیان ژن، ابتدا ساختار (متغیرهای ورودی، هدف و مجموع توابع) تعریف‌شده، سپس ساختار بهینه مدل و ضرایب طی فرآیند آموزش تعیین می‌شوند. درحالی که در شبکه‌های عصبی، ابتدا باید ساختار تعیین شده، فقط ضرایب مدل طی فرآیند آموزش حاصل می‌شوند. همچنین، این الگوریتم به‌طور خودکار می‌تواند متغیرهای ورودی که در این روش کروموزوم‌های خطی و ساده با طول ثابت، مشابه با آنچه که در الگوریتم ژنتیک استفاده می‌شود و ساختارهای شاخه‌ای با اندازه‌ها و اشکال متفاوت، مشابه با درختان تجزیه در برنامه‌ریزی ژنتیک ترکیب می‌شوند. اولین مرحله در الگوریتم مدل، تولید جمعیت اولیه از راه حل‌هاست. این امر می‌تواند به‌وسیله فرآیند تصادفی و یا در نظر گرفتن اطلاعات ورودی درباره مسأله انجام شود. سپس کروموزوم‌ها به‌صورت بیان درختی ارائه شده، به‌وسیله تابع برازش ارزیابی می‌شوند. در صورت دستیابی به راه‌حل مطلوب و یا رسیدن نسل‌ها به تعداد معین، تکامل متوقف شده و بهترین راه‌حل ارائه می‌شود. اگر شرایط توقف یافت نشود، نخبه‌گزینی انجام می‌شود و باقی راه‌حل‌ها به فرآیندی گزینشی واگذار می‌شوند. این فرآیند برای چندین نسل تکرار می‌شود و با پیش رفتن نسل به جلو کیفیت جمعیت نیز به‌طور نسبی بهبود می‌یابد (Ferreira, 2006).

فرامدل شبیه‌ساز بیان ژن نیز همانند الگوریتم ژنتیک و برنامه‌ریزی ژنتیک، یک الگوریتم ژنتیکی است به‌طوری که از جمعیتی از افراد استفاده می‌کند که آن‌ها را مطابق برازندگی انتخاب می‌کند و تغییرات ژنتیکی را با استفاده از یک یا چند عملگر ژنتیکی اعمال می‌کند. تفاوت اساسی بین این الگوریتم، مربوط به ماهیت افراد آن‌ها است به‌طوری که در الگوریتم ژنتیک، افراد رشته‌های خطی با طول ثابت (کروموزوم‌ها) و در برنامه‌ریزی ژنتیک، نهاده‌های غیر خطی با اندازه‌ها و اشکال متفاوت (درختان تجزیه) می‌باشند، در حالی که در فرامدل شبیه‌ساز بیان ژن، افراد به‌صورت رشته‌های خطی با طول ثابت (ژنوم یا کروموزوم‌ها) کدگذاری شده و سپس به شکل نهاده‌های غیر خطی با اندازه‌ها و اشکال متفاوت (یعنی نمایش دیاگرام ساده یا بیان درختی) بیان می‌شوند (Ferreira, 2001).

شبکه بیزین[۱] (BN)

نخستین بار در اوایل دهه ۱۹۹۰ به‌وسیله بنتین و بعدها Mackay (۱۹۹۲) برای شبکه‌های عصبی مصنوعی به‌کار برده شد و با سرعت به یک ابزار انتخاب برای کاربردهای هوش مصنوعی تبدیل شد (Mackay, 1992). در شبکه‌های بیزین برخلاف شبکه‌های عصبی متداول هدف یافتن یک بردار وزن بهینه نیست، بلکه ارائه عدم قطعیت صریح با توجه به مقادیر وزن به‌وسیله توزیع احتمال ثانویه است (Kingston et al, 2005). شبکه بیزین، نمایش با معنی روابط نامشخص مابین پارامترها در یک فرآیند می‌باشد و گرافی جهت‌دار غیر حلقوی از گره‌ها، برای نمایش متغیرهای تصادفی و کمان‌ها، برای نمایش روابط احتمالی میان متغیرها به‌شمار می‌رود. Mackay (۱۹۹۲) دو سطح از استنتاج‌های بیزین را که می‌توان در مدلسازی شبکه عصبی بیزین[۲] به‌کار رود، تشریح کرده است. اولین سطح شامل استخراج شبکه وزن‌ها می‌باشد، البته با این فرض که ساختار شبکه انتخاب شده صحیح است. سطح دوم شامل مقایسه مدل‌ها و انتخاب مدل است (Mackay, 1992). شبکه عصبی بیزین ابزاری مفید برای تجزیه و تحلیل عدم قطعیت در مدل‌سازی شبکه‌های عصبی مصنوعی و یکی از روش‌های سامانه پشتیبان تصمیم‌گیری است که از یک شکل هوش مصنوعی استفاده کرده، بر اساس قانون یادگیری شبکه بیزین آموزش دیده، با استفاده از اطلاعات مبهم، ناکافی، متضاد و غیر قطعی، هم در گذشته و هم در حال حاضر، پارامترها را به‌صورت عصب به یکدیگر ارتباط می‌دهد که فعالیت سلسله اعصاب یک شبکه و چگونگی پیوند بین آن‌ها مبنای هوشمندی شبکه را نشان می‌دهد. شبکه‌های عصبی بیزین در حل مسائلی که روابط بین متغیرها به درستی مشخص نباشد، بسیار مفید هستند.

---

۱. Bayesian Network

۲. Bayesian Neural Network

الگوریتم بیزین

شبکه بیزین یک مدل گرافیکی احتمالاتی است که مجموعه‌ای از متغیرها و احتمالات را نشان می‌دهد. این شبکه، یک گراف مستقیم و غیرچرخه‌ای است که در آن گره‌ها در حکم متغیرهای مسئله هستند. شبکه‌های بیزین محدودیتی در نمایاندن متغیرهای تصادفی ندارند. ساختار یک شبکه بیزین در واقع یک نمایش نموداری از اثرات متقابل متغیرهایی است که باید الگوبندی شوند، علاوه بر اینکه کیفیت، رابطه بین متغیرها را نیز به نمایش می‌گذارد. این روش بر مبنای محاسبات احتمالات شرطی (قانون بیز[۱]) است. معادله زیر رابطه بیز را نشان می‌دهد.

رابطه ۲۹)
$$P(a|b) = \frac{P(b|a) \times P(a)}{P(b)}$$

که در آن، $P(a)$ احتمال وقوع پیشامد a، $P(b)$ احتمال وقوع پیشامد b، $P(b|a)$ احتمال شرطی پیشامد b است به شرطی که پیشامد a اتفاق افتاده باشد و $P(a|b)$ احتمال شرطی پیشامد a است

مدل‌های پیش‌بینی عملکرد گندم

برای پیش‌بینی کمّی عملکرد محصول گندم (تن در هکتار) چهار مدل مختلف مورد استفاده قرار گرفت. سپس بر اساس معیارهای ارزیابی بهترین مدل در هر پیش‌بینی انتخاب شد. متغیر مقدار عملکرد گندم (X) هر شهرستان برای پیش‌بینی عملکرد گندم آتی می‌تواند شامل یک، دو، سه و چهار سال قبل از متغیر عملکرد گندم باشد. در مدل ۱، برای پیش‌بینی متغیر تناژ محصول در سال بعد، از سری زمانی متغیر مذکور با یک گام تأخیر استفاده شد. بدین صورت که برای پیش‌بینی در زمان t+1، از مقدار آن در زمان t استفاده شد. در مدل ۲، پیش‌بینی در سال بعد بر اساس مقادیر تناژ محصول تا دو سال قبل استفاده شد و به طور مشابه در مدل ۳ و ۴، پیش‌بینی‌ها بر اساس مقادیر متغیر مورد نظر تا سه و چهار سال قبل صورت پذیرفت. روابط زیر در تکمیل توضیحات بیان شده ارائه می‌شود

رابطه ۳۰) مدل ۱
$$X_{(t+1)} = f(X_{(t)})$$

رابطه ۳۱) مدل ۲
$$X_{(t+1)} = f(X_{(t)}, X_{(t-1)})$$

رابطه ۳۲) مدل ۳
$$X_{(t+1)} = f(X_{(t)}, X_{(t-1)}, X_{(t-2)})$$

رابطه ۳۳) مدل ۴
$$X_{(t+1)} = f(X_{(t)}, X_{(t-1)}, X_{(t-2)}, X_{(t-3)})$$

ارزیابی مدل‌های پیش‌بینی عملکرد گندم

برای ارزیابی مدل‌ها و مقایسه نتایج، از سه معیار ارزیابی یعنی ضریب همبستگی[۲] (R)، جذر میانگین مربعات خطا[۳] (RMSE) و میانگین قدرمطلق خطا[۴] (MAE) استفاده شد. رابطه ریاضی متغیرها به شرح زیر است:

رابطه ۳۴)
$$R = \frac{\frac{1}{n}\sum_{i=1}^{n}(O_i - \bar{O})(f_i - \bar{f})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - \bar{O})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - \bar{f})^2}}$$

رابطه ۳۵)
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - f_i)^2}$$

رابطه ۳۶)
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|O_i - f_i|$$

در این روابط، O_i مقادیر مشاهده‌شده درگام زمانی iام، f_i مقادیر پیش‌بینی شده در گام زمانی iام، $\bar{o}$ میانگین مقادیر مشاهده‌شده، $\bar{f}$ میانگین مقادیر پیش‌بینی‌شده و n تعداد داده‌ها است. هر مدلی که دارای R بیشتر، RMSE و MAE کمتر باشد، از مطلوبیت بیشتری

---

۱. Bayes theorem

28. *Correlation*
29. *Root Mean Square Error*
30. *Mean Absolute Error*

برخوردار است.

آماده سازی داده‌ها

پس از بررسی و کنترل کیفی آمار شهرستان‌ها و رفع نواقص آماری، همگنی داده‌ها با استفاده از آزمون Run Test ارزیابی و تصادفی بودن داده‌ها در سطح اطمینان ۹۵٪ پذیرفته شد. در ادامه با توجه به وجود سری زمانی متغیر فراوانی عملکرد گندم، اقدام به پیش‌بینی آن با استفاده از روش‌های BN، GEP، SVM و KNN شد. ۸۰ درصد از داده‌ها برای آموزش شبکه و ۲۰ درصد داده‌ها برای آزمون و خطا تعیین شد. پارامترهای مناسب روش‌های مختلف از طریق آزمون و خطا تعیین شد. همچنین برای جلوگیری از بیش برازش از صحت‌سنجی متقاطع[1] استفاده شد. در روش صحت‌سنجی متقاطع در هر بار اجرای مدل ها برای پیش‌بینی با یک پارامتر مشخص، یک یا تعدادی از داده‌های مجموعه آموزش کنار گذاشته می‌شوند (به عنوان داده های صحت‌سنجی) و مدل بر اساس داده‌های باقیمانده آموزش می‌بینند. سپس داده‌های کنار گذاشته شده به مجموعه آموزش برگردانده شده و داده یا مجموعه دیگری از داده‌های آموزش به عنوان داده صحت‌سنجی از این مسیر جدا می‌شوند و بار دیگر شبکه آموزش داده می‌شود. این فرایند تا زمانی که تمام داده‌های آموزش در قسمت صحت‌سنجی استفاده شوند، ادامه می‌یابد. پس از آن براساس میانگین خطای مدل در برآورد داده‌های صحت‌سنجی، پارامتر(های) مناسب شبکه‌عصبی تعیین می‌شوند. برای پیش‌بینی از چهار مدل معرفی شده فوق استفاده شد. مدل مناسب بر اساس معیارهای ارزیابی R، MAE و RMSE برای هر روش در مقیاس سالانه استخراج شد. پس از آن، تحلیل نتایج بر پایه بهترین مدل برای هر یک از روش‌ها صورت پذیرفت.

نتایج و بحث

شکل ۳، تغییرات عملکرد تولید گندم (X) را در شش شهرستان استان البرز نشان می‌دهد. همانطور که در شکل مشخص است، نوسانات سری زمانی شهرستان کرج (با بیشترین عملکرد تولید گندم) نسبت به سایر سری‌های زمانی در بین شهرستان‌های استان البرز، کمتر و همچنین نوسانات سری زمانی شهرستان‌های اشتهارد و طالقان با کمترین عملکرد تولید گندم بیشتر بود. بنابراین، فرامدل‌های هوش‌مصنوعی نسبت به سری‌های زمانی محصولاتی که از ناهمواری کمتری برخوردار هستند، رفتار بهتری به نمایش می‌گذارند.

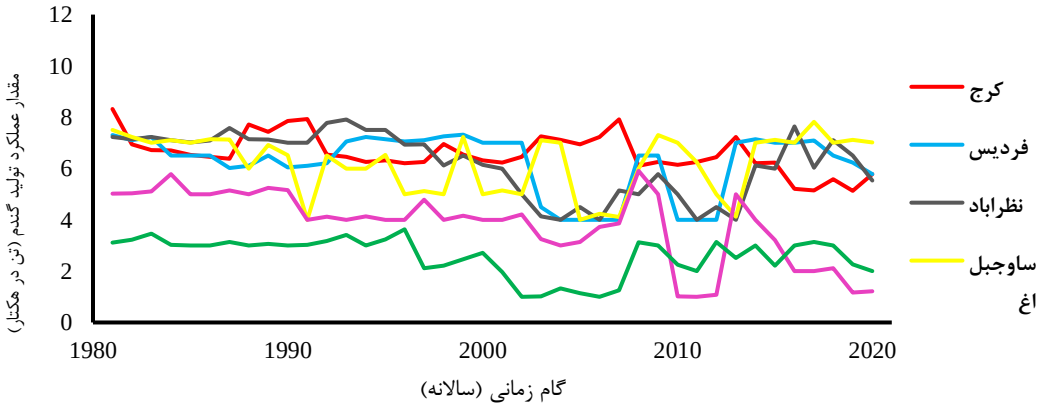
نتایج پیش‌بینی فراوانی تولید گندم حاکی از عملکرد مناسب تمام روش‌های استفاده شده در شهرستان‌های مورد مطالعه (نظرآباد، ساوجبلاغ، کرج، اشتهارد، فردیس و طالقان) بود (جدول ۱). با توجه به اینکه در جدول ۱ ترتیب شهرستان‌ها از طالقان تا کرج به صورت صعودی بوده و متوسط مقدار عملکرد گندم در مقیاس سالانه از شهرستان طالقان تا شهرستان کرج به ترتیب از ۲ تا ۶/۲ متغیر است، می‌توان دریافت که بهبود عملکرد تمامی روش‌ها با افزایش عملکرد گندم رابطه مستقیمی دارد. بدین معنی که با افزایش مقدار عملکرد گندم، مقادیر معیارهای ارزیابی مورد استفاده همگی بهبود می‌یابند. جدول ۱ مقادیر معیارهای ارزیابی مربوط به مرحله آزمایش روش‌های مختلف را نشان می‌دهد. براساس جدول ۱، تمامی روش‌ها در پیش‌بینی عملکرد گندم شهرستان طالقان که در بین شهرستان‌های مورد مطالعه کمترین فراوانی را داشته است، کمترین دقت؛ و در سمت مقابل، روش‌های هوش‌مصنوعی در شهرستان کرج که بیشترین فراوانی تولید گندم را در بین شهرستان‌های مورد بررسی دارند، بیشترین دقت را داشته‌اند. به نحوی که در روش BN ضریب همبستگی بین مقادیر مشاهداتی عملکرد گندم و مقادیر محاسباتی آن از ۰/۸۰ به ۰/۸۴ افزایش می‌یابد. همچنین برای روش‌های GEP، SVM و KNN نیز ضریب همبستگی به ترتیب از ۰/۸۳، ۰/۸۵ و ۰/۸۴ تا ۰/۸۹، ۰/۹۱ و ۰/۹۲ برای پیش‌بینی در شهرستان‌های مورد مطالعه متغیر است.

جدول ۱. معیارهای ارزیابی محاسبه‌شده مدل‌های مختلف هوش‌مصنوعی در پیش‌بینی فراوانی عملکرد گندم

روش	معیارهای ارزیابی	شهرستان‌ها					
		طالقان	اشتهارد	ساوجبلاق	نظرآباد	فردیس	کرج
	مدل بهینه	۴	۴	۴	۴	۴	۳
BN	R	۰/۸۰ (۴)	۰/۸۱ (۴)	۰/۸۲ (۴)	۰/۸۱ (۴)	۰/۸۲ (۴)	۰/۸۴ (۴)
	MAE	۰/۲۷ (۴)	۰/۲۵ (۴)	۰/۲۴ (۴)	۰/۲۲ (۴)	۰/۲۱ (۴)	۰/۱۹ (۴)

۱. Cross Validation

GEP	RMSE	۰/۳۳ (۴)	۰/۳۱ (۴)	۰/۲۹ (۴)	۰/۲۹ (۴)	۰/۲۸ (۴)	۰/۲۷ (۴)
	مدل بهینه	۴	۴	۴	۴	۳	۴
	R	۰/۸۳ (۳)	۰/۸۴ (۳)	۰/۸۵ (۳)	۰/۸۵ (۳)	۰/۸۷ (۳)	۰/۸۹ (۳)
	MAE	۰/۱۹ (۳)	۰/۱۸ (۳)	۰/۱۷ (۳)	۰/۱۶ (۳)	۰/۱۵ (۳)	۰/۱۴ (۳)
	RMSE	۰/۳۰ (۳)	۰/۲۹ (۳)	۰/۲۸ (۳)	۰/۲۷ (۳)	۰/۲۶ (۳)	۰/۲۵ (۳)
SVM	مدل بهینه	۳	۴	۴	۴	۴	۴
	R	۰/۸۵ (۱)	۰/۸۶ (۲)	۰/۸۷ (۲)	۰/۸۸ (۲)	۰/۸۹ (۲)	۰/۹۱ (۲)
	MAE	۰/۱۷ (۱)	۰/۱۷ (۲)	۰/۱۶ (۲)	۰/۱۵ (۲)	۰/۱۴ (۲)	۰/۱۳ (۲)
	RMSE	۰/۲۵ (۱)	۰/۲۷ (۲)	۰/۲۵ (۲)	۰/۲۴ (۲)	۰/۲۳ (۲)	۰/۲۲ (۲)
KNN	مدل بهینه	۴	۳	۴	۴	۴	۴
	R	۰/۸۴ (۲)	۰/۸۹ (۱)	۰/۹۰ (۱)	۰/۹۰ (۱)	۰/۹۱ (۱)	۰/۹۲ (۱)
	MAE	۰/۱۸ (۲)	۰/۱۶ (۱)	۰/۱۵ (۱)	۰/۱۴ (۱)	۰/۱۳ (۱)	۰/۱۱ (۱)
	RMSE	۰/۲۸ (۲)	۰/۲۴ (۱)	۰/۲۳ (۱)	۰/۲۳ (۱)	۰/۲۲ (۱)	۰/۲۱ (۱)
X (تن در هکتار)		۲	۳	۵/۵	۵/۵	۵/۷۸	۶/۲



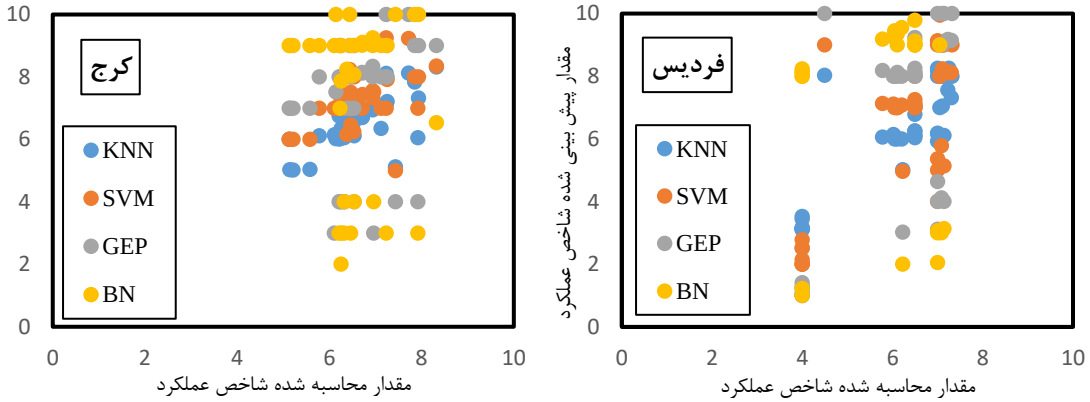
شکل ۳. سری زمانی شاخص عملکرد تولید گندم (x) در مقیاس زمانی سالانه برای شهرستان‌های مورد مطالعه در استان البرز

در جدول۱ اعداد داخل پرانتز روبروی مقادیر معیارهای ارزیابی، رتبه‌بندی هر روش بر اساس معیار ارزیابی مورد نظر در پیش‌بینی فراوانی عملکرد گندم را نشان می‌دهد. بر این مبنا، در شهرستان‌های اشتهارد، ساوجبلاغ، نظرآباد، فردیس و کرج روش KNN و در شهرستان طالقان روش SVM بیشترین تعداد رتبه اول (بهترین پیش‌بینی) را داشته اند. با ترکیب تعداد معیارهای ارزیابی و تعداد شهرستان-های مورد مطالعه برای پیش‌بینی، ۱۸ رتبه (۳*۶) برای هر روش وجود خواهد داشت. بدین ترتیب روش BN و GEP اصلاً رتبه نخستی به خود اختصاص نداده‌اند، در حالی‌که روش KNN و SVM به ترتیب ۱۵ و ۳ بار در رتبه نخست قرار گرفته‌اند. در مجموع و با این حساب، عملکرد روش KNN بهتر از سایر روش‌ها بوده و بالاترین دقت پیش‌بینی را داشته است. نکته قابل توجه، آموزش سریع این شبکه و کاهش زمان آزمون و خطا برای انتخاب پارامتر بهینه آن است. پس از روش KNN، روش SVM با ۳ بار قرارگیری در رتبه اول و ۱۵ بار قرارگیری در رتبه دوم بهترین عملکرد را داشته است. نکته مهم دیگری که باید به آن اشاره کرد، عملکرد مناسب روش SVM است. هرچند که این مدل در تمامی‌شهرستان‌های مورد مطالعه عملکرد تقریباً مشابهی داشته و تنها در شهرستان طالقان از تمامی روش‌های مورد استفاده، عملکرد بهتری داشت. نتیجه‌گیری دیگری که از جدول۱ قابل استنتاج است، انتخاب مدل‌هایی پیچیده‌تر به عنوان مدل بهینه پیش‌بینی‌کننده در شهرستان‌های استان البرز است. به نحوی که برای پیش‌بینی عملکرد گندم در شهرستان‌های ساوجبلاغ و نظرآباد از مدل ۴ (با چهار گام تأخیر) استفاده شده است. در حالی‌که برای شهرستان‌های طالقان، اشتهارد، فردیس و کرج استفاده از مدل بهینه شماره ۳ و ۴ (با سه و چهار گام تأخیر) نتیجه بهتری را گزارش می‌کند. شکل ۴، مقادیر مشاهده‌شده و پیش‌بینی‌شده عملکرد گندم را برای داده‌های آزمایش نشان می‌دهد. این شکل بهبود کارایی روش‌های مورد استفاده با افزایش تولید گندم را به خوبی بیان می‌کند. به‌طوری که شهرستان‌های

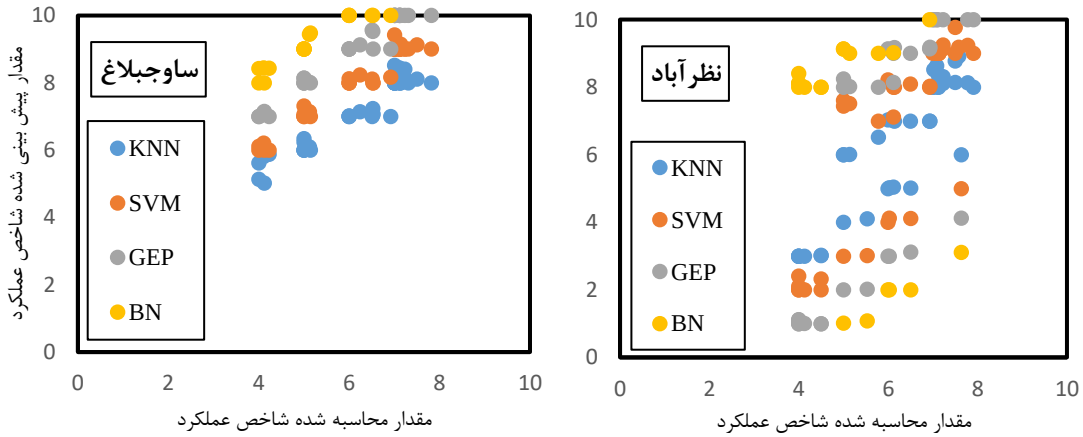
فردیس و کرج به عنوان فراوان‌ترین شهرستان‌ها از نظر عملکرد گندم، بهترین مطابقت بین مقادیر پیش‌بینی و محاسبه‌شده عملکرد گندم را داشتند. همان‌طور که در شکل ۴مشخص است؛ ارتباط بین مقادیر محاسباتی و پیش‌بینی‌شده از یک خط بدون شیب در شهرستان طالقان به خطی با شیب تقریباً ۴۵ درجه در شهرستان‌های فردیس و کرج رسیده است. مقایسه میانگین مقادیر مشاهده‌شده و پیش‌بینی‌شده بر اساس آزمون Z، نشان می‌دهد که فرض صفر مبنی بر برابر بودن میانگین سری مقادیر مشاهده‌شده و پیش‌بینی‌شده در تمامی روش‌های مورد استفاده، صحیح می‌باشد و میانگین سری زمانی مشاهداتی را حفظ کرده‌اند (جدول ۲). این ویژگی نیز می‌تواند تأکیدی بر کارایی روش‌های هوش‌مصنوعی در پیش‌بینی عملکرد گندم باشد.

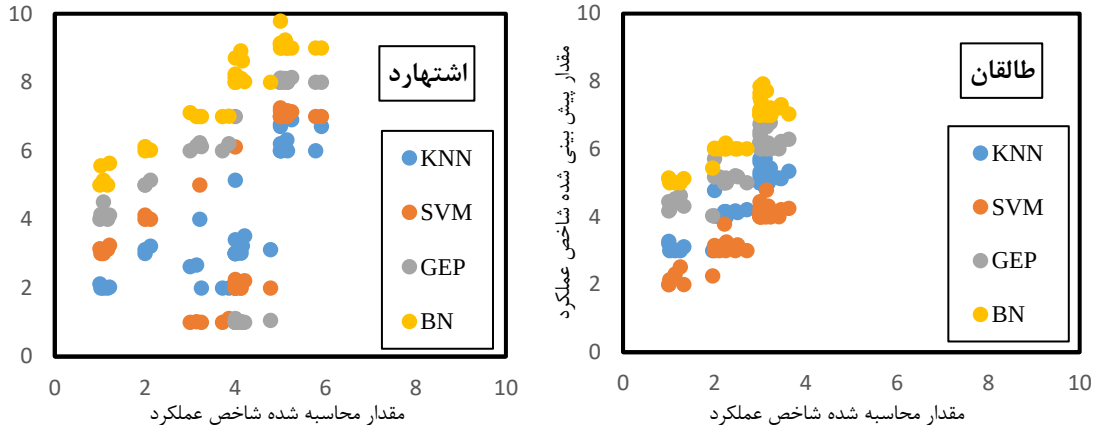
جدول ۲. آزمون مقایسه میانگین سری زمانی مشاهداتی و پیش‌بینی‌شده شاخص X

شهرستان	میانگین				آمارهZ			
	مشاهداتی	BN	SVM	GEP	KNN	BN	SVM	GEP
طالقان	۰/۸۳۱	۰/۸۳۳	۰/۸۳۵	۰/۸۲۹	۰/۸۴۸	۰/۸۳۷	۰/۸۴۰	۰/۸۴۲	۰/۸۵۲
اشتهارد	۰/۸۲۴	۰/۸۲۵	۰/۸۳۱	۰/۸۲۶	۰/۸۴۲	۰/۸۲۷	۰/۸۳۳	۰/۸۲۹	۰/۸۴۷
ساوجبلاق	۰/۸۲۷	۰/۸۲۹	۰/۸۳۳	۰/۸۲۸	۰/۸۴۷	۰/۸۳۰	۰/۸۳۵	۰/۸۳۱	۰/۸۵۱
نظرآباد	۰/۸۳۳	۰/۸۳۸	۰/۸۴۰	۰/۸۳۹	۰/۸۴۹	۰/۸۳۸	۰/۸۴۱	۰/۸۴۷	۰/۸۵۳
فردیس	۰/۸۳۴	۰/۸۴۰	۰/۸۴۲	۰/۸۴۱	۰/۸۵۱	۰/۸۳۹	۰/۸۴۲	۰/۸۵۴	۰/۸۵۹
کرج	۰/۸۳۷	۰/۸۴۱	۰/۸۴۳	۰/۸۴۲	۰/۸۵۷	۰/۸۴۰	۰/۸۴۸	۰/۸۵۸	۰/۸۶۷



شکل ٤. نتایج مقادیر محاسبه‌شده و پیش‌بینی‌شده شاخص عملکرد گندم (تن در هکتار) در شهرستان‌های مورد مطالعه





ادامه شکل ٤ نتایج مقادیر محاسبه‌شده و پیش‌بینی‌شده شاخص عملکرد گندم (تن در هکتار) در شهرستان‌های مورد مطالعه

نتیجه‌گیری

هدف از این پژوهش پیش‌بینی عملکرد گندم با استفاده از فرامدل‌های هوش‌مصنوعی BN، GEP، SVM و KNN در مقیاس زمانی سالانه در استان البرز بود. بدین منظور از داده‌های سالانه عملکرد در شش شهرستان استان البرز با طول دوره آماری ۴۰ ساله (۲۰۲۰-۱۹۸۱) استفاده شد. در ادامه با در دست بودن سری زمانی عملکرد گندم، با استفاده از چهار روش BN، SVM، GEP و KNN اقدام به پیش‌بینی گندم شد. نتایج نشان داد که روش‌های KNN و SVM به‌منظور پیش‌بینی عملکرد سالانه گندم بهترین و فرامدل‌های BN و GEP کمترین عملکرد را به همراه داشتند. با این توضیح که تمامی چهار روش مورد استفاده میانگین سری زمانی مشاهداتی را حفظ کرده بودند که این موضوع حکایت از توانمندی فرامدل‌های هوش‌مصنوعی برای پیش‌بینی گندم بود. پیش‌بینی KNN در شهرستان طالقان با دقت کمتری انجام شد؛ به نحوی که روش KNN با معیارهای ارزیابی R، MAE و RMSE برابر با ۰/۸۴، ۰/۱۸ و ۰/۲۸ عملکرد کمتری نشان داد. همچنین، با افزایش فراوانی عملکرد در شهرستان‌های مورد مطالعه، دقت پیش‌بینی تمام روش‌ها افزایش داشته است. با این توضیح که تمام روش‌ها در شهرستان طالقان که کمترین فراوانی را داشته، کمترین عملکرد و در شهرستان‌های فردیس و کرج که بیشترین فراوانی را به همراه داشتند، بیشترین عملکرد را از خود نشان داده‌اند.

"هیچ‌گونه تعارض منافع وجود ندارد"

REFERENCES

Alvarez R. (2009). Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach. *European Journal of Agronomy*. 30, 70-77. doi:10.1016/j.eja.2008.07.005.

Araghinejad, S. H. & Burn, D. (2005). Probabilistic forecasting of hydrological events using geostatistical analysis. *Hydrological Sciences Journal- des Sciences Hydrologiques*, 50(5), 57-66.

Aslam F., Salman A. & Jan I. (2019). Predicting wheat production in Pakistan by using an artificial neural network approach. *Sarhad Journal of Agriculture*, 35(4), 1054-1062.

Baier, W., and Robertson, G.W. (1967). Estimating yield components of wheat from calculated soil moisture. *Can. J. Plant*. Sci. 47: 617-30, doi:10.4141/cjps67-108.

Basak, D., Pal, S. & Patranabis, DC. (2007). Support vector regression. *Neural Information Processing*, 11, 203- 225.

Chipanshi A.C., Ripley E.A., & Lawford R.G. (1999). Large-scale simulation of wheat yields in a semi-arid environment using a crop-growth model. Agricultural Systems. 5: 57–66, doi:10.1016/s0308-521x (98)00082-1.

Dahiya, S., Singh, B., Gaur, S., Garg, V. K., & Kushwaha, H. S. (2007). Analysis of groundwater quality using fuzzy synthetic evaluation. Journal of Hazardous Materials, 147(3): 938-946, doi:10.1016/j.jhazmat.2007.01.119.

Danandehmehr, A. and M.R. Majdzadeh Tabatabai. 2010. Prediction of daily discharge trend of river flow based on genetic programming. Journal of Water and Soil, 24(2): 325-33 (in Persian).

Doraiswamy P.C., Moulin S., Cook P.W., and Stern, V. (2003). Crop yield assessment from remote sensing. Photogrammetric Engineering and Remote Sensing, 69: 665−674, doi:10.14358/pers.69.6.665.

Ferreira, C. (2001). Gene expression programming: a new adaptive algorithm for solving problems. *Complex System*, 13, 87-129.

Ferreira, C. (2006). Gene expression programming: mathematical modeling by an artificial intelligence (studies in computational intelligence). *Springer-Verlag New York*, Inc. Secaucus, NJ, USA.

Han J., Zhang Z., Cao J., Luo Y., Zhang L., Li Z., and Zhang J. (2020). Prediction of winter wheat yield based on multi-source data and machine learning in China. Remote Sensing 236(12): 1-22, doi:10.3390/rs12020236.

Hosseini, S., Sioseh Mardeh, A., Fathi, P., & Sioseh Mardeh, M. (2007). Application of artificial neural network (ANN) and multiple regression for estimating assessing the performance of day farming wheat yield in ghorveh region, Kurdistan province. *Agricultural reserch*, 7(1), 41-45.

Jayawardena, A. W., Li, W. K. & Xu, P. (2002). Neighbor selection for local modelling and prediction of hydrological time series. *J. of Hydrology*, 258, 40-57.

Karlsson, M. & Yakowitz, S. (1987). Nearest-neighbor methods for nonparametric rainfall-runoff forecasting. *Water Resources Research*, 23(7), 1300-1308.

Kaul, M., R. Hill, C. Walthall. (2005). Artificial neural networks for corn and soybean yield prediction. Agricultural Systems. 85: 1-18. doi:10.1016/j.agsy.2004.07.009.

Khodabandeh, n. (1998). Cereals. Fifth Edition. University of Tehran Press, Tehran, 538.

Khoshnevisan B., Rafiee S., Omid M. & Mousazadeh H. (2014). Development of an intelligent system based on ANFIS for predicting wheat grain yield on the basis of energy inputs. *Information Processing in Agriculture*, 1(1), 14-22. doi:10.1016/j.inpa.2014.04.001

Kingston, G.B., Lambert, M.F. & Maier, H.R. (2005). Bayesian training of artificial neural networks used for water resources modeling. *Water Resources Research*, 41(12), 11. https://doi.org/10.1029/2005WR004152.

Koocheki, A., Kamali, GH.A. & Banaian, M. (1993). Simulation of primary production. The center of agrobiological research and department of theoretical production ecology, Wageningen, Netherlands. Published by World Meteorological Organization. Geneva, July. 219p.

Kshirsagar, A.M. (1972). Multivariate analysis. Marcel Decker, Inc., New York.

MacKay, D.J.C. (1992). A practical Bayesian framework for backpropagation networks. Neural Computation, 4(3), 448-472.

Maselli و F., and Rembold, F. (2001). Analysis of GAC NDVI data for cropland identification and yield forecasting in Mediterranean African countries. *Photogrammetric Engineering and Remote Sensing*, 67: 593−602.

Mehnatkesh, A., Ayyubi, S., Jalalyan, A., & Dehgani, A. (2016). Comparison of multivariate linear regression and artificial neural networks models for estimating of rainfed wheat yield in some central Zagros areas. *Iranian Dryland Agronomy Journal*, 5(2), 119-133.

Meshkani, A. & Nazemi, A. (2009). Introduction to data mining. *Ferdowsi University of Mashhad*, 456 pages. (in Farsi)

Misra, D., Oommen, T., Agarwa, A., Mishra, SK. & Thompson, AM. (2009). Application and analysis of supportvector machine based simulation for runoff andsediment yield. *Biosystems Engineering*. 103(3), 527–535.

Norouzi M., Ayoubi S., Jalalian A., Khademi H., and Dehghani A.A. (2010). Predicting rainfed wheat quality and quantity by artificial neural network using terrain and soil characteristics. *Acta Agric Scandinavica, Section B-Plant Soil Sciences*, 60, 341-352. doi:10.1080/09064710903005682.

Pagie, L. & Mitchell, MA. (2002). Comparison of evolutionary and coevolutionary search. *International Journal of Computational Intelligence and Application*, 2, 53–69.

Rahmani, E., Liaghat, A. & Khalili, A. (2010). Estimating Barley Yield in Eastern Azerbaijan Using Drought Indices and Climatic Parameters by Artificial Neural Network (ANN). *Iranian Journal of Soil and Water Research*, 39(1), 47-56.

Servati, M., Barikloo, A., Alamdari, P. & Moravej, K. (2018). Application of Heuristic Methods in Prediction of Wheat Yield. *Applied Soil Research*, 6(3), 106-117.

shahinejad, B. (2018). Comparison of wavelet neural network models, support vector machine and gene expression programming in estimating the amount of oxygen dissolved in rivers. *Iran-Water Resources Research*, 14(3), 226-238.

Sharif, M. H. & Burn, D. (2006). Simulating climate change scenarios using an improved K-nearest neighbor

model. *J. of Hydrology*, 325, 179-196

Singh, V.P., Translation, M.R. Najafi. (2002). Hydrological systems for rainfall modeling. *Tehran University Press, First Edition*, 578 pagesM. (in Farsi)

Tarboton, D. G., Sharma, A., and Lall, U. (1993). The use of non-parametric probability distribution in streamflow modeling. *In Proceeding of the 6 South African National Hydrological Symposium*, Ed. S. A

Todeschini, R. (1989). K-nearest neighbour method: Influence of data transformations and metrics. *Chemometrics and Intelligent Laboratory Systems*, 6, 213-220.

Uossef gomrokchi, A., Baghani, J. & Abbasi, F. (2021). Evaluating the Capability of Data Mining Models in Predicting Irrigated Wheat Yield in Iran. *Water and Soil*, 35(2), 189-202. doi: 10.22067/jsw.2021.15029.0

27. Vapnik VN (1995) The nature of statistical learning theory. *Springer, New York.*

Veelenturf L.P.J. (1995). Analysis applications of artificial neural networks. *Simon and Schuster International Group*, United States of America.

Verma, U., Koehler, W., and Goyal, M. (2012). A study on yield trends of different crops using ARIMA analysis. Environ. Ecol., 30(4A), 1459-1463.

Wall L., Larocque D., and Leger P.M. (2007). The early explanatory power of NDVI in crop yield modeling. *International Journal of Remote Sensing*. 29: 2211−2225.

Wu F.Y., and Yen K.K. (1992). Application of neural network in regression analysis. *Computer and Industrial Engineering*. 23: 93-98.

Yakowitz, S. J. (1985). Nonparametric density estimation, prediction, and regression for markov sequences. *J. Am. Stat. Assoc*., 80, 215-221.

Yoon H, Jun SC, Hyun Y, Bae GO, Lee KK (2011) Acomparative study of artificial neural networks andsupport vector machines for predicting groundwaterlevels in a coastal aquifer. *Journal of Hydrology*, 396(4), 128–138.