نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد ۳۵، شماره ۲، خرداد – تیر ۱۴۰۰، ص. ۲۰۲-۱۸۹

Journal of Water and Soil
Vol. 35, No. 2, May-June 2021, p. 189-202

مقاله پژوهشی

ارزیابی قابلیت مدل‌های مبتنی بر داده‌کاوی در پیش‌بینی عملکرد گندم آبی در کشور

افشین یوسف گمرکچی[۱*]- جواد باغانی[۲]- فریبرز عباسی[۳]

تاریخ دریافت: ۱۳۹۹/۰۶/۲۵
تاریخ پذیرش: ۱۳۹۹/۱۱/۲۷

چکیده

گندم و نان به‌عنوان اصلی‌ترین غذای مردم در کشور از اهمیت ویژه‌ای برخوردارند. گندم نه‌تنها یک کالای مهم تجاری در دنیا محسوب می‌شود، بلکه به‌عنوان سلاحی برتر در مناسبات سیاسی و جهانی روز به‌روز بر اهمیت استراتژیک آن افزوده می‌شود. از این رو تحلیل و پیش‌بینی وضعیت تولید این محصول همواره مورد توجه بوده است. در این تحقیق کارایی سه مدل شبکه عصبی مصنوعی، رگرسیون خطی چند متغیره و مدل درختی به‌منظور پیش‌بینی عملکرد گندم آبی در مناطق عمده تولید در سطح کشور، بر اساس اطلاعات میدانی ثبت شده ۲۴۱ مزرعه، ارزیابی شد. نتایج تحقیق نشان داد ضریب تبیین مدل شبکه عصبی مصنوعی و مدل رگرسیون خطی چند متغیره به ترتیب برابر ۰/۶۷۲ و ۰/۵۷۷ بود که با اعمال گروه‌بندی داده‌ها به روش درختی ضریب تبیین مدل پیش‌بینی به ۰/۷۶۲ افزایش یافت. نتایج خروجی مدل درختی نشان داد مناطق عمده تولید گندم در سطح کشور از نظر حجم آب مصرفی، به ۴ گروه مستقل قابل تفکیک است. نهایتاً می‌توان نتیجه گرفت مدل درختی با اعمال گروه‌بندی هدفمند در داده‌های ورودی، می‌تواند به‌عنوان یک ابزار قدرتمند در تخمین عملکرد گندم آبی در قطب‌های عمده تولید گندم در سطح کشور مورد استفاده قرار گیرد.

واژه‌های کلیدی: داده‌کاوی، حجم آب مصرفی، گروه‌بندی، مدل‌سازی

مقدمه

با افزایش قیمت مواد غذایی در سال‌های اخیر اهمیت پیش‌بینی دقیق‌تر و به‌موقع تولید محصول در مقیاس ملی و جهانی افزایش یافته است. چنین اطلاعاتی برای ایجاد برنامه‌ریزی آگاهانه در بخش کشاورزی در سطح ملی و بین‌المللی، تثبیت بازارها، افزایش دسترسی به بازار و جلوگیری از کمبود مواد غذایی ضروری به نظر می‌رسد (۸). گندم به‌عنوان 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. به‌منظور پیش‌بینی عملکرد محصول قبل از برداشت، طیف وسیعی از تکنیک‌ها مانند استفاده از ارزیابی کمی تناسب اراضی، روش‌های مبتنی بر سنجش از دور، مدل‌های شبیه‌سازی محصول و روش‌های رگرسیونی به کار گرفته شده است (۶، ۷، ۱۴ و ۲۶). یکی از روش‌های مدل‌سازی که در سال‌های اخیر مورد توجه محققین 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 (۲۵). نوروزی و همکاران (۱۷) از شبکه‌های عصبی مصنوعی به‌منظور پیش‌بینی عملکرد گندم دیم در مناطق نیمه‌خشک و کوهستانی غرب ایران استفاده نمودند. محنت کش و همکاران (۱۵) با استفاده از مدل‌های رگرسیون چند متغیره خطی و شبکه‌های عصبی مصنوعی، عملکرد گندم دیم را در مناطقی از زاگرس مرکزی برآورد نمودند. نتایج نشان از توانایی بهتر شبکه‌های عصبی مصنوعی نسبت به رگرسیون چند متغیره خطی در برآورد عملکرد دانه و زیست‌توده گندم دیم در مناطق مورد مطالعه داشت. خوشنویسان و همکاران (۱۲) از روش نروفازی و شبکه‌های عصبی برای پیش‌بینی عملکرد گندم در منطقه فریدون‌شهر واقع در استان اصفهان استفاده کردند. داده‌های

۱- استادیار بخش تحقیقات فنی و مهندسی کشاورزی، مرکز تحقیقات و آموزش کشاورزی و منابع طبیعی استان قزوین، سازمان تحقیقات، آموزش و ترویج کشاورزی، قزوین، ایران
(*- نویسنده مسئول: Email: a.gomrokchi@areeo.ac.ir)
۲ و ۳- به‌ترتیب استادیار و استاد مؤسسه تحقیقات فنی و مهندسی کشاورزی، سازمان تحقیقات، آموزش و ترویج کشاورزی، کرج، ایران
DOI: 10.22067/jsw.2021.15029.0

مورد استفاده در این تحقیق از ۲۶۰ مزرعه گندم جمع‌آوری شد. نتایج نشان داد که مدل نروفازی می‌تواند عملکرد گندم را دقیق‌تر از مدل شبکه عصبی مصنوعی پیش‌بینی کند. ثروتی و همکاران (۲۲) از روش‌های فراکاوشی در تخمین عملکرد گندم شهرستان هریس استفاده نمودند. نتایج تحقیق نشان داد، مدل ترکیبی نروفازی می‌تواند به‌عنوان یک ابزار در تخمین عملکرد گندم عمل کند. وو و ین (۲۷) از دو روش شبکه عصبی مصنوعی و مدل رگرسیون چند متغیره برای پیش‌بینی عملکرد گندم در ارتباط با مصرف کود نیتروژنه استفاده نمودند. آلوارز (۲) با به‌کارگیری شبکه‌های عصبی مصنوعی، متوسط عملکرد گندم را در منطقه پامپاس آرژانتین برآورد کرد. وی با استفاده از پارامترهای خاک و همچنین پارامترهای هواشناسی و به‌کارگیری مدل‌های رگرسیون و شبکه عصبی مصنوعی، میزان عملکرد گندم را برآورد نمود. نتایج تحقیق نشان داد عملکرد محصول به ترتیب به پارامترهای ظرفیت نگهداری آب در خاک و محتوای کربن آلی خاک وابستگی بیشتری داشته است. وی گزارش نمود که مدل شبکه عصبی با دقت بسیار خوبی نسبت به مدل‌های رگرسیون عملکرد را برآورد کرده و چنانچه داده‌های مورد استفاده در این مدل در بازه زمانی ۴۰ تا ۶۰ روز قبل از برداشت گندم موجود باشند، می‌توان از این مدل به‌منظور برآورد عملکرد گندم در منطقه استفاده کرد. اسلام و همکاران (۳) بر اساس 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 نمودند.

علاوه بر مدل‌های شبکه عصبی مصنوعی و رگرسیونی، امروزه از قابلیت‌های روش‌های داده‌کاوی به‌منظور بهبود نتایج خروجی مدل‌های پیش‌بینی و تحلیل اطلاعات میدانی استفاده شده است. مدل‌های درختی (درختان تصمیم[1]) به همراه قوانین تصمیم یکی از روش‌های داده‌کاوی به شمار می‌آیند. مدل‌های درختی روشی برای نمایش یک سری از قوانین هستند که منتهی به یک رده یا مقدار می‌شود. این مدل‌ها، از طریق جداسازی متوالی داده‌ها به گروه‌های مجزا ساخته شده و هدف در این فرآیند افزایش فاصله بین گروه‌ها در هر جداسازی است (۲۸). رامش و واردان (۱۸) عملکرد محصولات کشاورزی را با استفاده از تکنیک‌های مختلف داده‌کاوی پیش‌بینی نمودند. نتایج تحقیق آنها نشان داد روش‌های داده‌کاوی برای پیش‌بینی عملکرد محصولات کشاورزی از دقت بالا و توانایی زیادی برخوردار است. ذکی دیزاجی و همکاران (۲۸) از قابلیت‌های الگوریتم درخت تصمیم‌گیری به‌منظور مدل‌سازی متغیرهای مؤثر بر عملکرد نیشکر استفاده کردند. همچنین ایوانسکا و همکاران (۱۰) از قابلیت های مدل درختی به‌منظور تجزیه و تحلیل عوامل مؤثر بر عملکرد گندم پائیزه در مزارع لهستان استفاده نمودند. هان و همکاران (۹) با استفاده از قابلیت‌های مدل‌های درختی و بر اساس اطلاعات اقلیمی، داده‌های سنجش از دور و اطلاعات خاکشناسی منطقه، عملکرد گندم را در ۶ قطب عمده تولید گندم در کشور چین پیش‌بینی کردند. نتایج تحقیق نشان داد مدل‌های بردار ماشین مبنا و جنگل تصادفی با ضریب تبیین بالاتر از ۰/۷۵، توانایی پیش‌بینی عملکرد گندم در بازده زمانی ۱ تا ۲ ماه قبل از برداشت را داشته است.

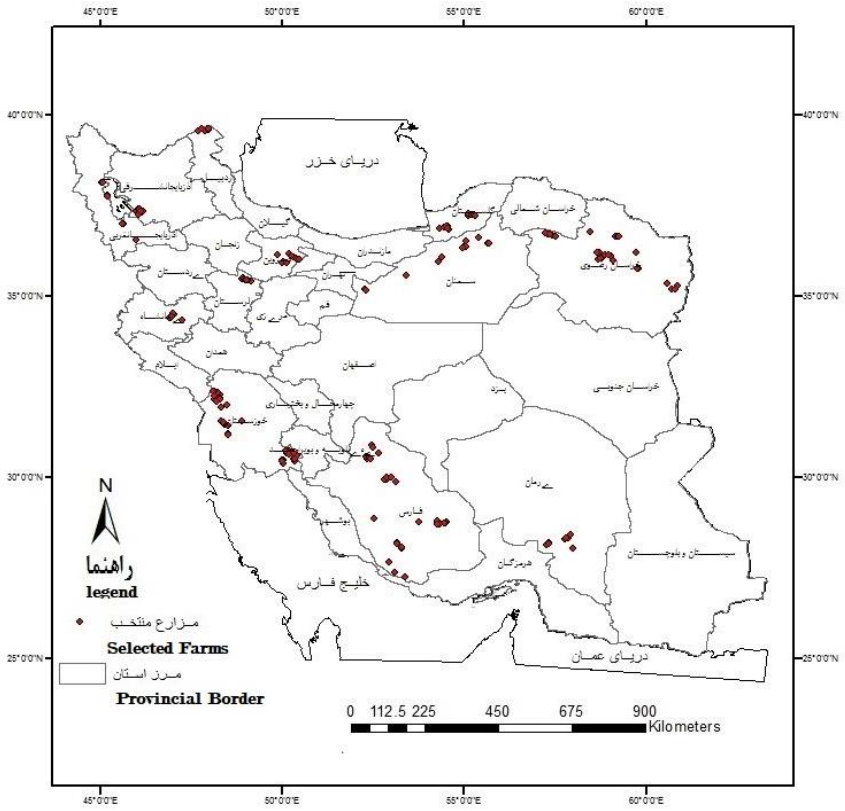
بررسی پژوهش‌های انجام شده، نشان می‌دهد که عملکرد گیاه تابعی از عوامل مختلف گیاهی، اقلیمی و شرایط مدیریتی آب و خاک است. از این‌رو محاسبه مقدار عملکرد گیاه و شاخص‌های وابسته به آن از روابط غیرخطی پیچیده‌ای تبعیت می‌کند که مدل‌سازی آن نیز دشواری خاصی دارد. با توجه به اینکه بررسی پاسخ گندم آبی به نهاده‌های 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. تحقیق حاضر با هدف توسعه و ارزیابی قابلیت سه مدل شبکه عصبی، درختی و رگرسیون خطی چند متغیره در پیش‌بینی عملکرد گندم بر اساس پارامترهای مؤثر بر عملکرد آن، در قطب‌های عمده تولید گندم در سطح کشور انجام شده است.

مواد و روش‌ها

پایگاه داده

اطلاعات مورد استفاده در این تحقیق شامل مقادیر حجم آب مصرفی و عملکرد گندم آبی و کمیت‌های مربوط به این دو شاخص در مزارع گندم آبی تحت مدیریت زارعین (تعداد ۲۴۱ مزرعه) در استان‌های خوزستان، فارس، گلستان، همدان، کرمانشاه، خراسان رضوی، اردبیل، آذربایجان شرقی، آذربایجان غربی، سمنان، جنوب کرمان و قزوین است که در یک تحقیق میدانی در سال زراعی ۹۶-۱۳۹۵ برداشت شده‌اند (۴). شاخص‌های مورد نظر از جمله حجم آب مصرفی، بدون دخالت در برنامهٔ آبیاری بهره‌دار و تحت مدیریت زارعین اندازه‌گیری شدند. برای انجام آزمایش‌ها، ابتدا مقدار دبی خروجی از منبع آبی انتخاب شده (کانال، چاه، قنات و یا چشمه) با وسیله مناسب (فلوم، کنتور حجمی، سرریز، میکرومولینه و یا دستگاه دبی سنج اولتراسونیک) اندازه‌گیری و سپس حجم آب مصرفی قطعه انتخابی گرفته شد. در برخی موارد با توجه به تغییرات احتمالی دبی در منابع آبی، کل دبی چند نوبت در طول فصل زراعی اندازه‌گیری شد.

1- Decision Trees



شکل ۱- پراکنش مکانی مزارع گندم مورد مطالعه در سطح کشور

Figure 1- Spatial distribution of studied wheat fields in Iran

در دشت‌هایی که منبع تأمین آب آبیاری، شبکه آبیاری بود، یک دریچه از شبکه مورد نظر انتخاب و دبی آن با استفاده از میکرومولینه، سرریز و یا فلوم چند نوبت در طول فصل زراعی اندازه‌گیری شد. علاوه بر آن زمان هر نوبت آبیاری و تعداد نوبت‌های آبیاری در طول فصل زراعی نیز ثبت شد. با استفاده از داده‌های اندازه‌گیری شده، عمق آب داده شده و حجم آب مصرفی در هر یک از نوبت‌های آبیاری در هر روش آبیاری (سطحی و یا تحت فشار) در فصل زراعی قابل محاسبه بود و در نهایت حجم آب مصرفی محصول گندم در طول یک فصل زراعی با اندازه‌گیری دبی منبع آبی و زمان کارکرد آن تعیین شد (۴). استان‌های مذکور بر اساس آمارنامه وزارت جهاد کشاورزی در سال زراعی ۹۶-۱۳۹۵ دارای بیشترین سطح زیر کشت گندم آبی در کشور بوده و حدود ۷۰ درصد سطح زیر کشت و تولید این محصول در کشور را پوشش داده است (۱). در شکل ۱ پراکنش مکانی مزارع گندم منتخب در سطح کشور نشان داده شده است.

مدل‌سازی عملکرد

مدل شبکه عصبی مصنوعی

شبکه‌های عصبی مصنوعی مدل‌های ریاضی انعطاف‌پذیری هستند که بر اساس ساختار و رفتار شبکه‌های عصبی طبیعی ساخته شده‌اند. شبکه‌های عصبی از مجموع چندین نرون ریاضی تشکیل شده‌اند. هر نرون ریاضی دارای شمای یک نرون واقعی می‌باشد. یکی از شبکه‌های عصبی نظارت شده و پرکاربرد، شبکهٔ چند لایه پرسپترون[1] با الگوریتم پس انتشار خطا[2] است که برای محدودهٔ گسترده‌ای از کاربردها از قبیل شناسایی الگو، درون‌یابی، پیش‌بینی و مدل‌سازی فرایند مناسب است (۱۶). در تحقیق حاضر به‌منظور توسعه شبکه عصبی از قابلیت‌های نرم‌افزار R با بسته Neuralnet استفاده شده است و برای آموزش و همگرایی سریع‌تر و دقت بیشتر شبکه، ابتدا ورودی‌های آن با استفاده از تکنیک حداقل-حداکثر (رابطه ۱) استاندارد شده و به داده‌های نرمال (بی‌بعد) در بازه ۱- تا ۱ تبدیل شدند (۲۱).

$$X_n = 2\left(\frac{X_i - X_{min}}{X_{max} - X_{min}}\right) - 1 \qquad (۱)$$

1- Multi Layers Perceptron
2- Back Propagation

که در آن، X_r، X_n، X_{min} و X_{max} به ترتیب نشان دهنده مقادیر واقعی، نرمال شده، حداقل و حداکثر داده‌های تحت بررسی است.

پس از مرحله نرمال‌سازی، تصادفی نمودن داده‌ها انجام شد. نتیجه این مرحله، داشتن مجموعه‌ای از ورودی و خروجی‌ها است که در آن دسته‌های ورودی- خروجی دارای نظام خاصی نیستند. پس از پایان تصادفی نمودن داده‌ها، میزان اطلاعاتی که باید در فرآیند آموزش شبکه استفاده شود، مشخص شده است. بر این اساس بخشی از داده‌ها برای آموزش (۷۰ درصد) و بخشی دیگر برای آزمون شبکه (۳۰ درصد) در نظر گرفته شد. توابع فعال‌ساز شبکه عصبی پرسپترون در مراحل پیاده‌سازی آموزش و آزمون شبکه، از تابع فعالیت تانژانت هذلولوی[1] برای محدودسازی دامنه داده‌های خروجی از هر نرون و روند آموزش الگو به الگو استفاده شده است (رابطه ۲).

$$F(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad (۲)$$

که در آن، x مقدار داده است.

مدل درختی (درخت تصمیم)

در تحقیق حاضر علاوه بر استفاده از قابلیت مدل‌سازی شبکه عصبی، از روش مدل درختی نیز به‌منظور پیش‌بینی عملکرد گندم استفاده شده است. مدل درختی یکی از ابزارهای قوی و متداول برای دسته‌بندی و پیش‌بینی می‌باشد. مدل درختی بر خلاف مدل شبکه عصبی به تولید قانون می‌پردازد. از مزایای درخت تصمیم نسبت به شبکه عصبی آن است که نسبت به نویز داده‌های ورودی مقاوم است (۲۴). مدل درختی بر اساس تقسیم‌های دودویی، داده‌ها را به بخش‌های مختلف تقسیم می‌کند. هر یک از افرازهای داده‌ها می‌توانند دوباره تحت یک تقسیم دودویی دیگر قرار گرفته و به هر زیر افراز، یک مدل برازش داده شود. تعیین نقطه مناسب افراز معمولاً قبل از برازش مدل انجام می‌شود. در این روش برای هر متغیر ورودی، مرتب‌سازی بر اساس مقدار ورودی انجام شده و نقاط شکست مختلف مورد آزمون قرار می‌گیرد. نقطه‌ای که بیشترین انحراف معیار بین مقادیر خروجی در دو گروه را ایجاد کند، نقطه شکست بهینه است. برای انتخاب نقطه شکست بهینه، یک افراز مشخص در نظر گرفته می‌شود و میزان کاهش انحراف استاندارد در دو بخش ایجاد شده (σ_1 و σ_2) نسبت به انحراف استاندارد کلیه داده‌ها (σ) به‌صورت میانگین وزن‌دار محاسبه می‌شود. افرازی که بیشترین کاهش را در σ ایجاد کند به‌عنوان نقطه شکست بهینه انتخاب می‌شود. در این تحقیق از قابلیت‌های نرم‌افزار WEKA برای اجرای مدل درختی استفاده شده است. شایان ذکر است پس از گروه‌بندی، مدل پیش‌بینی بر روی داده‌های گروه‌بندی شده اعمال شده است.

مدل رگرسیون خطی

هدف کلی از مدل‌سازی رگرسیون خطی چندگانه[2] پیدا کردن رابطه بین چند متغیر مستقل و یک متغیر وابسته است. داده‌های ورودی و خروجی در هر دو مدل شبکه عصبی مصنوعی و مدل رگرسیون خطی چند متغیره یکسان بود. در انجام فرآیند مدل‌سازی با استفاده از شبکه عصبی مصنوعی و مدل رگرسیون مجموعه داده‌های یکسان به کار گرفته شد.

در این راستا به‌منظور پیش‌بینی عملکرد گندم، یک متغیر وابسته یعنی عملکرد محصول و ۲۱ متغیر مستقل شامل اقلیم (X_1)، سطح سواد زارع (X_2)، ارتفاع (X_3)، نوع منبع آبی (X_4)، نوع شبکه آبیاری (X_5)، دبی منبع آبی (X_6)، شوری آب آبیاری (X_7)، سطح کل مزرعه (X_8)، سطح زیر کشت محصول گندم (X_9)، بافت خاک (X_{10})، شوری عصاره اشباع خاک (X_{11})، رقم (X_{12})، تاریخ کاشت (X_{13})، تاریخ برداشت (X_{14})، درجه روز رشد (X_{15})، روش آبیاری (X_{16})، متوسط عمق آب آبیاری در هر دور آبیاری (X_{17})، تعداد کل نوبت‌های آبیاری (X_{18})، حجم کل آب مصرفی (X_{19})، میزان بارندگی در طول دوره رشد (X_{20}) و نیاز آبشویی (X_{21}) مورد استفاده قرار گرفت. در نهایت با استفاده از روابط جذر میانگین مربعات خطا (رابطه ۳) و ضریب همبستگی پیرسون (رابطه ۴)، توانایی مدل‌ها در تخمین عملکرد گندم، مورد بررسی قرار گرفت (۵).

$$RMSE=\sqrt{\frac{1}{n}\times\sum_{i=1}^{n}\left(y_{i,predicted}-y_{i,observed}\right)^{2}} \quad (۳)$$

$$\rho=\frac{COV(y_{i,predicted},y_{i,observed})}{\sigma_{y_{i,observed}}\sigma_{y_{i,predicted}}} \quad (۴)$$

که در آن $y_{i,predicted}$ و $y_{i,observed}$: به ترتیب میزان عملکرد پیش‌بینی شده و اندازه‌گیری شده، N: تعداد مزارع، COV: تابع کوواریانس و σ: انحراف معیار متغیرهای مورد بررسی است.

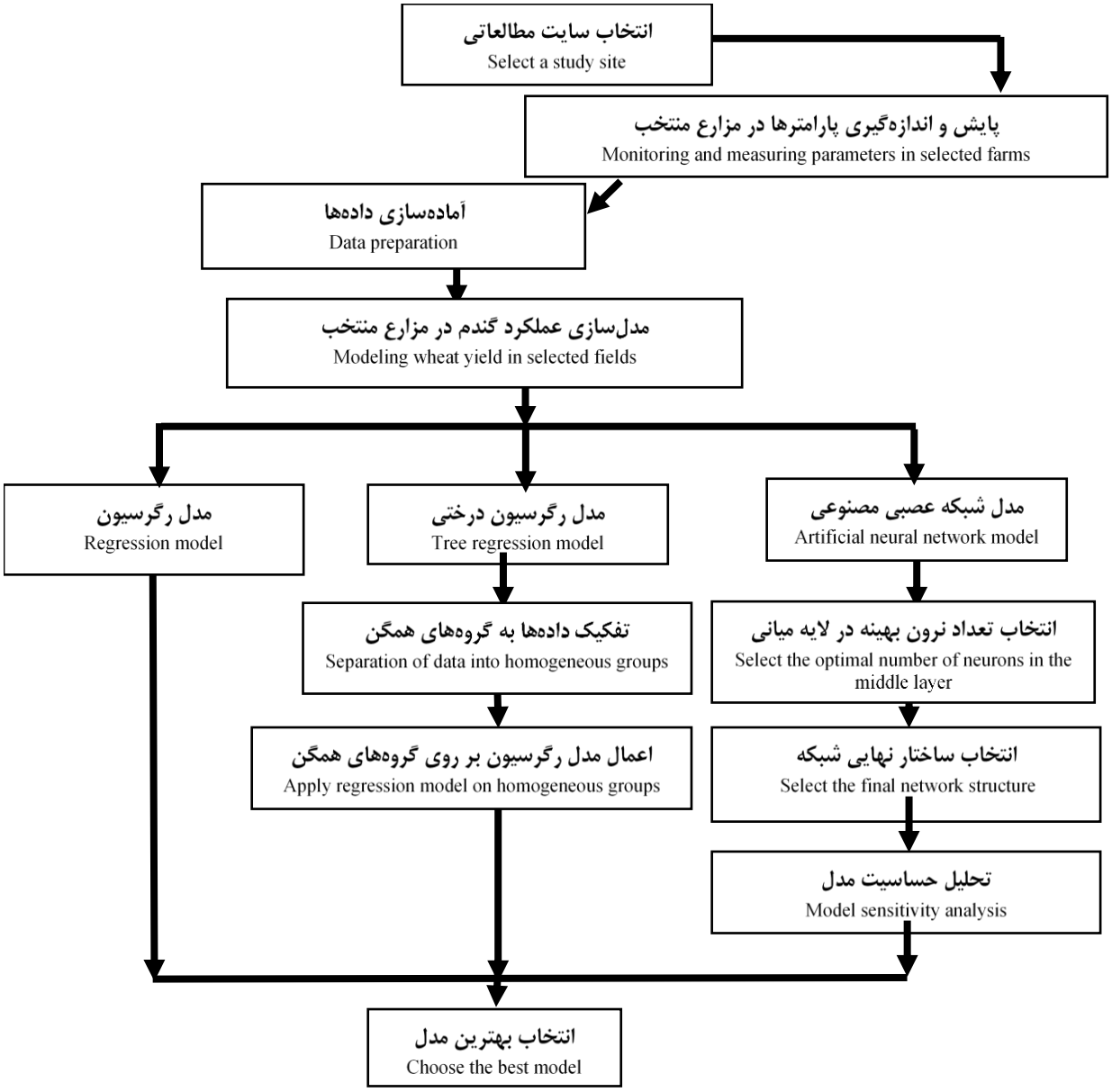
در شکل ۲ روند نمای انجام تحقیق نشان داده شده است.

نتایج و بحث

همان‌طور که اشاره گردید پارامترهای کمی شامل شوری آب آبیاری، شوری خاک، طول دوره رشد، تعداد دفعات آبیاری، میزان آب مصرفی و پارامترهای کیفی همانند اقلیم، سطح سواد زارع، بافت خاک که در ۲۴۱ مزرعه اندازه‌گیری شده‌اند، به‌عنوان ورودی برای تعیین مدل مناسب پیش‌بینی عملکرد مورد استفاده قرار گرفتند. در جدول ۱ تغییرات برخی پارامترهای آماری ثبت و اندازه‌گیری شده در ۲۴۱ مزرعه منتخب آورده شده است.

1- Hyperbulic Tangent

2- Multiple Linear Regression



شکل ۲- روند نمای مدل پیش‌بینی عملکرد گندم در مزارع منتخب

Figure 2- Flowchart of wheat yield prediction model in selected fields

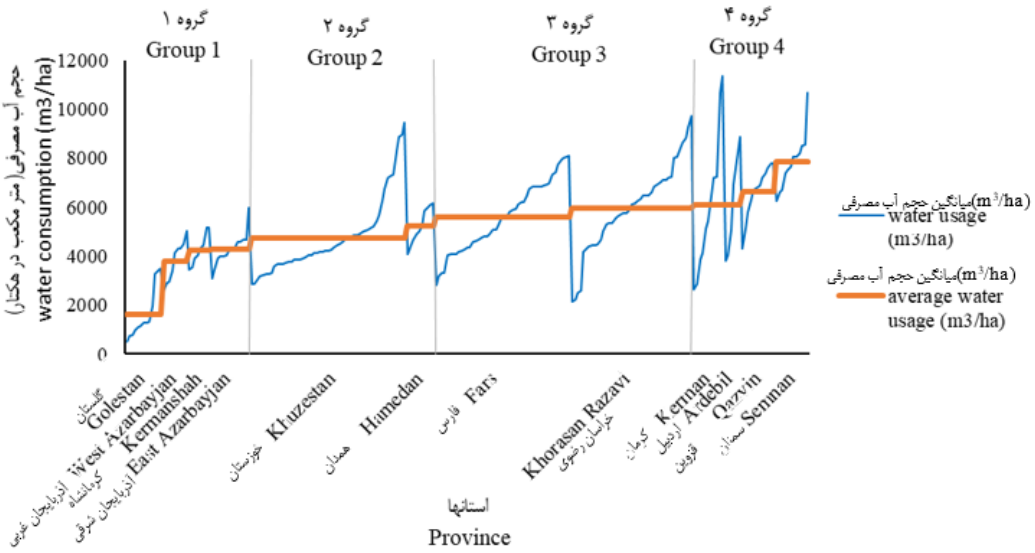
با توجه به آنکه حجم آب مصرفی، بیشترین همبستگی با عملکرد گندم را داشت، مزارع مطالعاتی بر اساس نتایج خروجی مدل درختی، از نظر میانگین حجم آب مصرفی در کشت محصول گندم به ۴ بخش تقسیم شده و مدل رگرسیونی برای هر بخش به‌صورت جداگانه اجرا شد. نکته حائز اهمیت آن است که به علت تعداد کم داده در گروه‌بندی‌های مدل درختی، امکان به‌کارگیری مدل شبکه عصبی مصنوعی در فرآیند پیش‌بینی میسر نگردید. در شکل ۳ تقسیم‌بندی استان‌ها بر مبنای میانگین حجم آب مصرفی با استفاده از قابلیت‌های مدل WEKA نشان داده شده است.

بر اساس نتایج خروجی مدل درختی، مناطق عمده تولید گندم در سطح کشور از نظر حجم آب مصرفی، به ۴ گروه تفکیک شد (شکل ۳). در این شکل استان‌هایی که میانگین حجم آب مصرفی کمتری نسبت به استان‌های دیگر داشتند در گروه ۱ و استان‌هایی که میانگین حجم آب مصرفی بیشتری جهت تولید گندم داشتند در گروه ۴ قرار گرفتند.

جدول ۱- محدوده تغییرات برخی پارامترهای ثبت و اندازه‌گیری شده در مزارع گندم منتخب

Table 1- Range of changes in some parameters recorded and measured in selected wheat fields

شاخص آماری Statistical index	EC (dS/m)	EC iw (dS/m)	تعداد دفعات آبیاری Number of irrigations	عملکرد Yield (kg/ha)	آب مصرفی Water consumption (m^3/ha)	بارش مؤثر در دوره کشت Effective rainfall during the growing season (mm)	بهره‌وری آب W.P (kg/m^3)
میانگین Mean	2.48	1.75	7.00	5123.94	5248.75	172.09	0.79
انحراف معیار Standard deviation	1.80	1.70	3.03	1431.08	1936.88	77.50	0.30



شکل ۳- گروه‌بندی استان‌ها بر اساس میانگین حجم آب مصرفی در تولید گندم

Figure 3- Grouping of provinces based on the average volume of water used in wheat production

نتایج خروجی مدل درختی نشان داد سه قطب تولید گندم در سطح کشور که شامل استان‌های فارس، خوزستان و خراسان رضوی بوده در گروه‌های میانه از دیدگاه حجم آب مصرفی قرار گرفتند. پس از گروه‌بندی داده‌ها، از قابلیت مدل رگرسیون خطی به‌منظور پیش‌بینی عملکرد گندم استفاده شد. در گام دوم برای انتخاب ساختار شبکه عصبی، پیچیدگی شبکه به‌تدریج افزایش یافته و خطای آموزش و آزمون مدل مورد بررسی قرار گرفت. بررسی داده‌ها نشان دهنده آن است که استفاده از ۲ گره در لایه پنهان منجر به افزایش دقت تخمین نخواهد شد و استفاده از یک گره در لایه پنهان کفایت داشته است. در نهایت مدل شبکه عصبی با ساختار ۱-۱-۹ (۹ گره در لایه ورودی، ۱ گره در لایه پنهان و ۱ گره در لایه خروجی) انتخاب شد. در جدول ۲ نتایج بررسی ساختار مختلف انتخاب شبکه عصبی در مدل‌سازی عملکرد گندم نشان داده شده است.

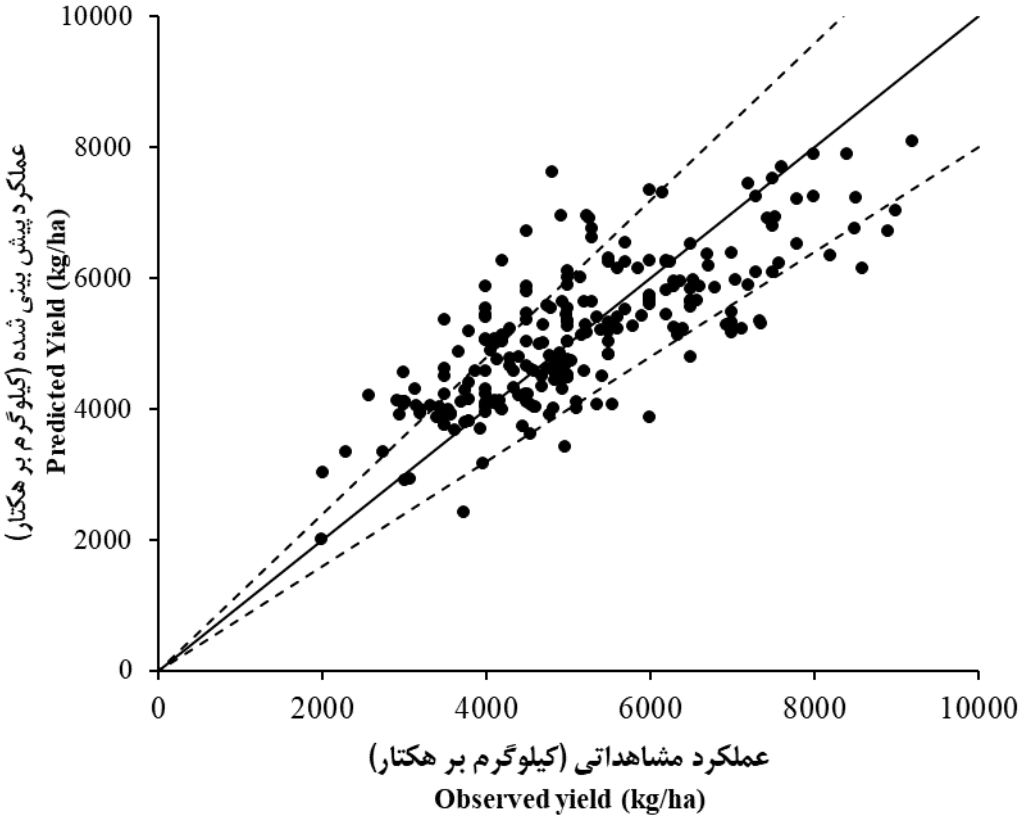
بر این اساس نتایج پیش‌بینی عملکرد گندم در سه روش مدل سازی (مدل درختی، شبکه عصبی و رگرسیون) در شکل‌های ۴ تا ۶ نشان داده شده است. در شکل‌های مذکور نتایج خروجی مدل، خط تطابق عالی و خطوط خطای ۲۰± درصد نیز نشان داده شده است.

نتایج خروجی نشان داد ضریب تبیین مدل در پیش‌بینی عملکرد محصول گندم در مدل شبکه عصبی مصنوعی و مدل رگرسیون خطی چند متغیره به ترتیب برابر ۰/۶۷۲ و ۰/۵۷۷ بود که با اعمال گروه‌بندی داده‌ها به روش درختی، ضریب تبیین مدل پیش‌بینی به ۰/۷۶۲ افزایش یافت. به عبارتی با گروه‌بندی هدفمند داده‌های ورودی، دقت پیش‌بینی عملکرد گندم افزایش یافته و مدل درختی دقت بالاتری نسبت به مدل شبکه عصبی و مدل رگرسیون در پیش‌بینی عملکرد گندم در سطح مزارع منتخب داشته است. با توجه به آنکه بر اساس نتایج مدل درختی مناطق تولید گندم (از منظر حجم آب مصرفی) به ۴ گروه تفکیک شده است در شکل‌های ۷ تا ۹ نمودار پراکنش خروجی مدل پیش‌بینی عملکرد گندم به‌صورت نمونه در چند قطب تولید گندم در سطح کشور نشان داده شده است.

جدول ۲- تحلیل خطای تخمین ساختار مختلف شبکه عصبی در مدل‌سازی عملکرد گندم

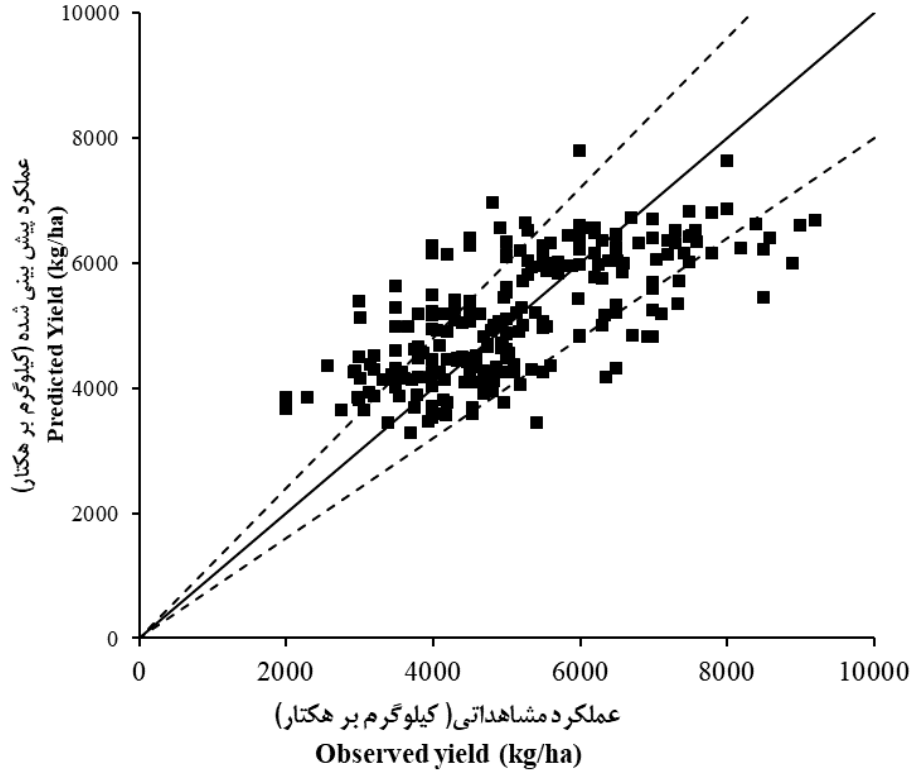
Table 2- Error analysis of estimating different neural network structure in wheat yield modeling

ساختار شبکه Network structure	جذر میانگین مربعات خطا Root Mean Squared Error	
	آموزش Train	آزمون Test
1-1-1	0.222	0.237
2-1-1	0.222	0.237
3-1-1	0.211	0.236
4-1-1	0.206	0.224
5-1-1	0.203	0.232
6-1-1	0.203	0.229
7-1-1	0.203	0.230
8-1-1	0.197	0.226
9-1-1	0.196	0.226
5-2-1	0.200	0.223
9-2-1	0.191	0.227



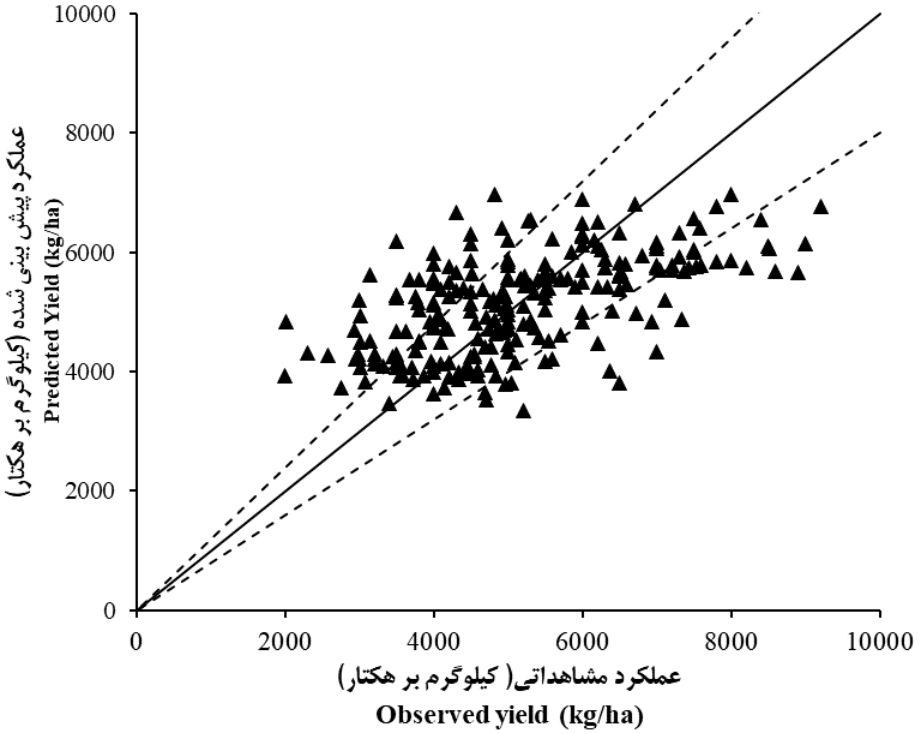
شکل ٤- نمودار پراکنش خروجی مدل پیش‌بینی عملکرد گندم به روش درختی

Figure 4- Output distribution diagram of wheat yield prediction model by tree method



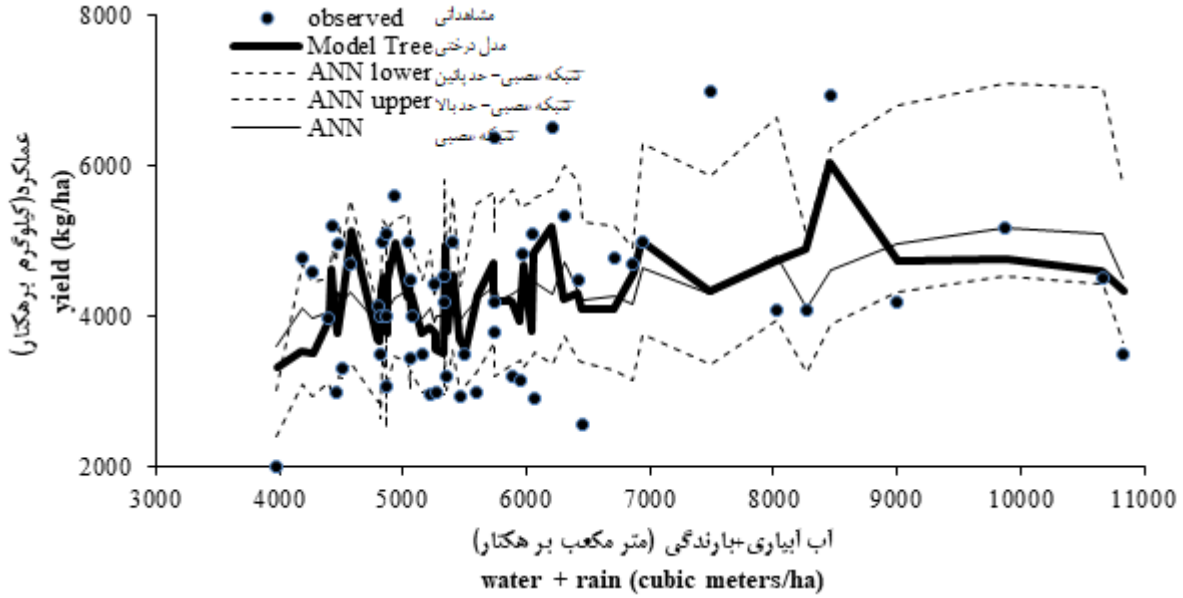
شکل ۵- نمودار پراکنش خروجی مدل پیش‌بینی عملکرد گندم به روش شبکه عصبی مصنوعی

Figure 5- Output distribution diagram of wheat yield prediction model by artificial neural network method



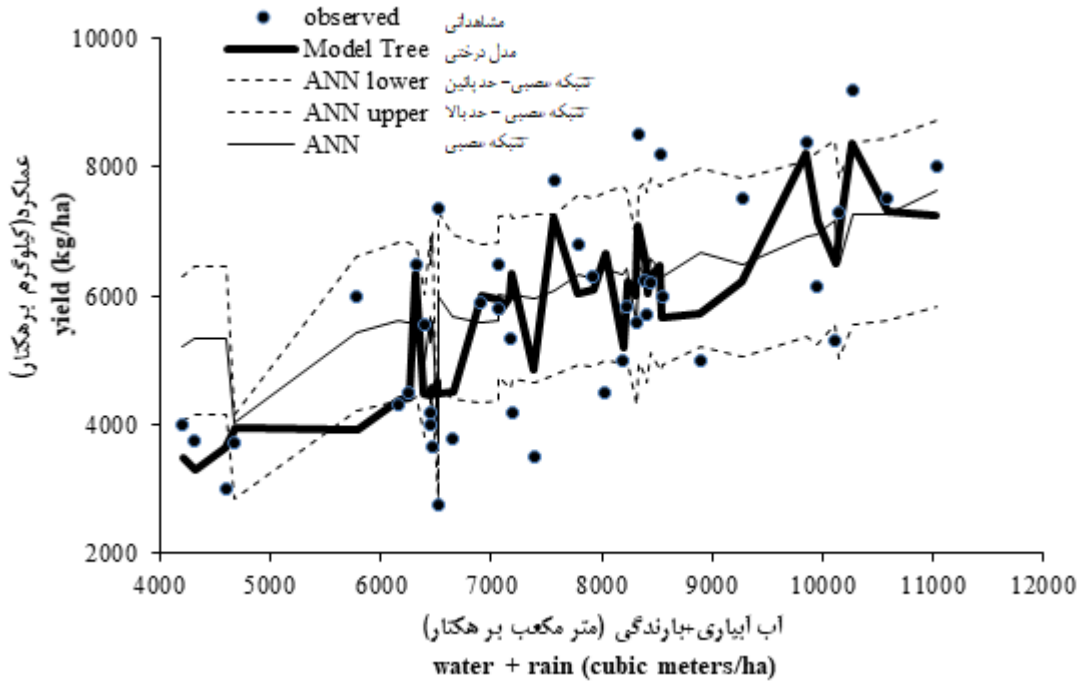
شکل ٦- نمودار پراکنش خروجی مدل پیش‌بینی عملکرد گندم به روش رگرسیون خطی

Figure 6- Output distribution diagram of wheat yield prediction model by linear regression method



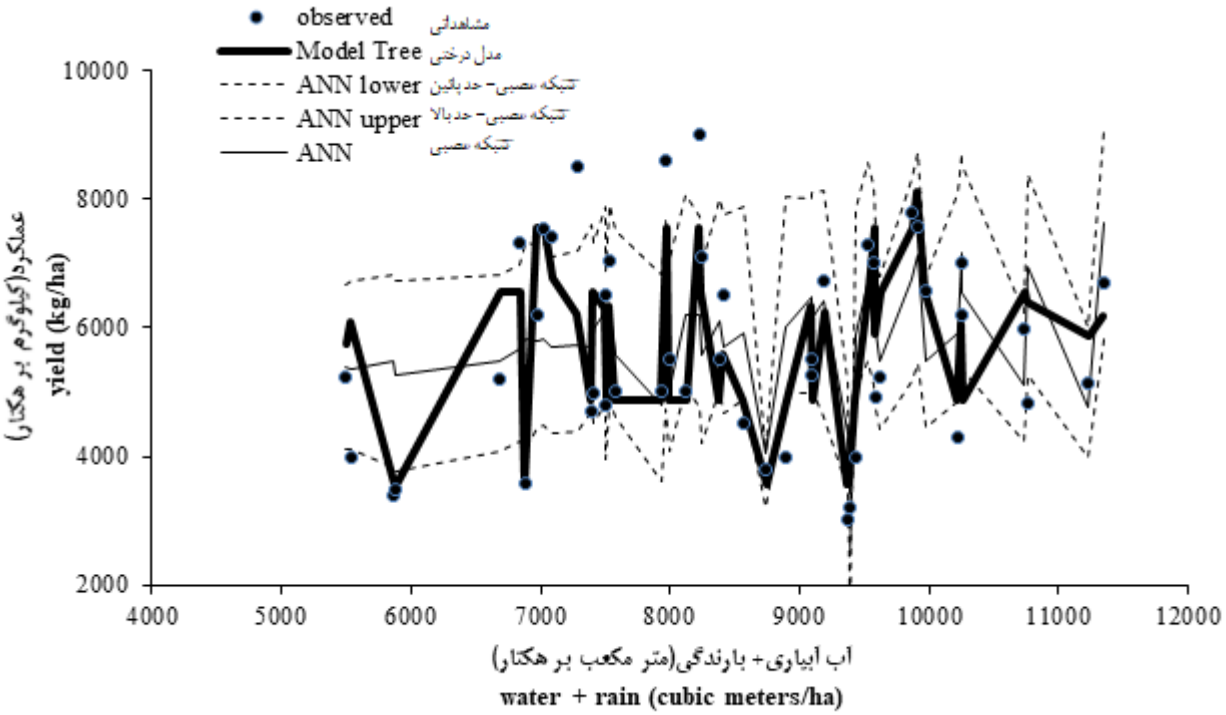
شکل ۷- نمودار پراکنش خروجی مدل پیش‌بینی عملکرد گندم به روش مدل درختی و شبکه عصبی در استان خوزستان

Figure 7- Output distribution diagram of wheat yield prediction model by tree model and neural network method in Khuzestan province



شکل ۸- مقایسه عملکرد گندم و نتایج پیش‌بینی مدل شبکه عصبی و مدل درختی در استان خراسان رضوی

Figure 8- Output distribution diagram of wheat yield prediction model by tree model and neural network method in Khorasan Razavi province



شکل ۹- مقایسه عملکرد گندم و نتایج پیش‌بینی مدل شبکه عصبی و مدل درختی در استان فارس

Figure 9- Output distribution diagram of wheat yield prediction model by tree model and neural network method in Fars province

در جدول ۳ دقت پیش‌بینی عملکرد گندم در استان‌های مختلف بر اساس نتایج خروجی دو مدل شبکه عصبی و درختی نشان داده شده است. نتایج حاصل از مدل‌سازی عملکرد در استان‌های مورد بررسی نشان داد مدل درختی در اغلب استان‌های کشور، دارای خطای تخمین کمتری در مقایسه با شبکه عصبی مصنوعی بوده است. از سوی دیگر ضمن تایید نتایج مطلوب تکنیک‌های شبکه عصبی، نتایج تحقیق نشان داد مدل رگرسیون چند متغیره دقت چندانی در پیش‌بینی عملکرد نداشت. این امر شاهدی بر وجود روابط غیر خطی و پیچیده موثر بر عملکرد گندم بوده و تاییدی بر استفاده از ابزارهای داده‌کاوی در پیش‌بینی عملکرد گندم است.

در این تحقیق گروه‌بندی هدفمند داده‌های ورودی بر اساس میانگین حجم آب مصرفی، منجر به افزایش دقت تخمین مدل پیش بینی شده است. کائول و همکاران (۱۱) اعلام کردند که عامل آب قابل دسترس یکی از عوامل اصلی در تخمین عملکرد محصولات کشاورزی می باشد. منتظر و همکاران (۱۶) نیز با ارزیابی کارایی مدل‌های شبکه عصبی مصنوعی در محاسبه عملکرد گندم به این نتیجه رسیدند که مدل شبکه عصبی دقت بالاتری نسبت به رگرسیون غیرخطی دارد و تحلیل حساسیت مدل ها نشان داد که عملکرد گندم بیشترین حساسیت را به عامل مقدار آب مصرفی داشته است. لیو و همکاران (۱۳) و ساداث و همکاران (۲۳) در تحقیقات مشابهی تغییرپذیری مکانی ویژگی‌های مورد بررسی را عامل ایجاد خطا در تخمین عنوان نمودند و پیشنهاد کردند در مدلسازی مقیاس وسیع، گروه‌بندی هدفمند داده‌های ورودی، منجر به کاهش خطای تخمین خواهد شد. به عبارت دیگر، همگن‌تر کردن داده‌ها و کاهش منابع تغییرات از پیچیدگی روابط بین متغیرها کاسته و در نتیجه کارایی شبکه عصبی برای مدل‌سازی این روابط را افزایش خواهد داد. بر این اساس رامش و واردان (۱۸) روش‌های مبتنی داده‌کاوی را برای پیش‌بینی عملکرد محصولات کشاورزی توصیه نمودند. هان و همکاران (۹) با استفاده از قابلیت‌های مدل‌های مبتنی بر داده‌کاوی، به ضریب تبیین بالاتر از ۰/۷۵، در پیش‌بینی عملکرد گندم دست یافتند.

جدول ۳- نتایج تحلیل خطای تخمین مدل شبکه عصبی و مدل درختی در استان‌های مختلف

Table 3- Results of error analysis of neural network model and tree model in different provinces

استان Province	جذر میانگین مربعات خطا Root Mean Squared Error	
	مدل درختی Tree model	مدل شبکه عصبی Neural network model
اردبیل (Ardebil)	0.041	0.072
آذربایجان غربی (West azarbayjan)	0.150	0.142
آذربایجان شرقی (East azarbayjan)	0.216	0.384
فارس (Fars)	0.164	0.259
گلستان (Golestan)	0.071	0.151
همدان (Hamedan)	0.092	0.213
کرمان (Kerman)	0.118	0.248
کرمانشاه (Kermanshah)	0.044	0.179
خراسان رضوی (Khorasan razavi)	0.211	0.285
خوزستان (Khuzestan)	0.240	0.257
قزوین (Qazvin)	0.159	0.186
سمنان (Semnan)	0.116	0.264
کل استان‌ها (All Provinces)	0.182	0.254

نتیجه‌گیری

در این تحقیق نیاز به بکارگیری روش‌های داده‌کاوی در تحلیل اطلاعات میدانی و ساماندهی پایگاه‌های بزرگ اطلاعات و نیز سودمندی روش‌های داده‌کاوی بخصوص درخت تصمیم در تخمین عملکرد محصول گندم، با سایر روش‌های پیش‌بینی مورد بررسی و آزمون قرار گرفت. نتایج کلی تحقیق نشان داد که تفکیک هدفمند داده‌های ورودی به مدل‌های پیش‌بینی می‌تواند منجر به افزایش دقت خروجی مدل‌های پیش‌بینی شود. هرچند نمی‌توان یک رویکرد عام در انتخاب و یا عدم انتخاب یک مدل پیش‌بینی در مناطق مختلف ارائه نمود. به نحوی در برخی تحقیقات انجام شده شبکه‌های عصبی توانایی بالائی در پیش‌بینی عملکرد محصولات مختلف از خود نشان داده‌اند (۲، ۳، ۱۵، ۱۷، ۲۱ و ۲۷) ولی نکته حائز اهمیت آن است که در صورت وجود داده‌های کافی و شناخت صحیح عوامل تأثیرگذار در متغیر وابسته، می‌توان با اعمال روش‌های داده‌کاوی دقت مدل‌های شبکه عصبی را نیز بهبود بخشید (۵). بنابراین تکنیک‌های داده‌کاوی در کنار قابلیت‌های هوش مصنوعی می‌تواند به عنوان ابزاری کارآمد برای تدوین برنامه‌های صحیح مدیریتی معرفی شود.

منابع

1- Ahmadi K., H.R. EbadZadeh F., Hatami R., HoseinPour and AbdShah H. 2019. Agricultural Statistics of 2017-2018. Ministry of Jihad for Agriculture, Deputy for Planning and Economy, Information and Communication Technology Office. Volume 3, Garden Products. 166 pp. (In Persian)

2- 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.

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

4- Baghani J. 2018. Determination of wheat water consumption in Iran. Final Research Report, Agricultural Engineering Research Institute. (In Persian)

5- Barikloo A., Alamdari P., Moravej K., and Servati M. 2017. Prediction of irrigated wheat yield by using hybrid algorithm methods of artificial neural networks and genetic algorithm. Journal of Water and Soil 30(3): 715-726. (In Persian with English abstract)

6- Chipanshi A.C., Ripley E.A., and Lawford R.G. 1999. Large-scale simulation of wheat yields in a semi-arid environment using a crop-growth model. Agricultural Systems 59: 57−66.

7- 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.

8- Franch B., Vermote E.F., Becker-Reshef I., Claverie M., Huang J., Zhang J., Justice C., and Sobrino J.A. 2015. Improving the timeliness of winter wheat production forecast in the United States of America, Ukraine and China using MODIS data and NCAR Growing Degree Day information. Remote Sensing of Environment 161: 131–148.

9- 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.

10- Iwańska M., Oleksy A., Dacko M., Skowera B., Oleksiak T., and Wójcik-Gront E. 2018. Use of classification and regression trees (CART) for analyzing determinants of winter wheat yield variation among fields in Poland. Biometrical Letters 55(2): 197-214.

11- Kaul M., Hill R.L., and Walthall C. 2005. Artificial neural networks for corn and soybean yield prediction. Agricultural Systems 85: 1-18.

12- Khoshnevisan B., Rafiee S., Omid M., and 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.
13- Liu J., and Goering C.E. 1999. Neural network for setting target corn yields. ASAE paper 99-3040, Toronto, Ontario, Canada, 18-21.
14- Maselli 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.
15- Mehnatkesh A., Ayyubi S., Jalalyan A., and Dehgani A.A. 2017. Comparison of multivariate linear regression and artificial neural networks models for estimating of rainfed wheat yield in some central Zagros areas. Iranian Journal of Dryland Agriculture 5(2): 119-133. (In Persian with English abstract)
16- Montazar A., Azadegan B., and Shahkary M. 2009. Assesing the Efficiency Of atifical neural network model to predict wheat yield and water prodoctivity based on climatic data and seasonal water-nitrogen variabls. Iranian Water Research Journal 3(5): 17-29.
17- 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.
18- Ramesh D., and Vishnu Vardhan B. 2013. Data mining techniques and applications to agricultural yield data. International Journal of Advanced Research in Computer and Communication Engineering 2(9): 3477-3480.
19- Raorane A.A., and Kulkarni R.V. 2013. Review role of data mining in agriculture. International Journal of Computer Science and Information Technologies 4(2): 270-275.
20- Rumelhart D.E., Hinton G.E., and Williams R.J. 1986. Learning internal representation by back-propagation errors. In: Rumelhart DE, McClelland JL, the PDP Research Group (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, MA.
21- Sepehri S., Abbasi F., and Nakhjavanimoghaddam M.M. 2019. Prediction of silage maize yield and sensitivity analysis of management parameters using artificial neural network models. Iranian Journal of Irrigation and Drainage 13(5): 1460-1470. (In Persian with English abstract)
22- Servati M., Barikloo A, Alamdari P., and Moravej K. 2017. Application of heuristic methods in prediction of wheat yield. Applied Soil Research 6(3): 106-117. (In Persian with English abstract)
23- Sudduth K.A., Drummond S.T., Birrell S.J., and Kitchen N.R. 1996. Analysis of spatial factors influencing crop yield, in Proc. 3rd Int. Conf. On Precision Agriculture, P.C. Robert et al. (ed.), pp. 129-140.
24- Toloei Ashlaghi A., Poorebrahimi A., Ebrahimi M., and Ghasemahmad L. 2013. Using data mining techniques for prediction breast cancer recurrence. Iranian Journal of Breast Diseases 5(4): 23-34. (In Persian with English abstract)
25- Veelenturf L.P.J. 1995. Analysis applications of artificial neural networks. Simon and Schuster International Group, United States of America.
26- 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.
27- Wu F.Y., and Yen K.K. 1992. Application of neural network in regression analysis. Computer and Industrial Engineering 23: 93-98.
28- Zakidizaji H., Bahrami H., Monjezi N., and Sheikhdavoodi M.J. 2019. Modeling of the variables that influence sugarcane yield using C5.0 and QUEST decision tree algorithms. Journal of Agricultural Machinery 9(2): 469-484. (In Persian with English abstract)

Journal of Water and Soil
Vol. 35, No. 2, May-June 2021, p. 189-202

نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد ۳۵، شماره ۲، خرداد – تیر ۱۴۰۰، ص. ۲۰۲-۱۸۹

Evaluating the Capability of Data Mining Models in Predicting Irrigated Wheat Yield in Iran

A.U. Gomrokchi[1]*- J. Baghani[2]- F. Abbasi[3]

Received: 15-09-2020
Accepted: 15-02-2021

Introduction: One of the modeling methods researchers have considered in various sciences in recent years is artificial neural network modeling. In addition to the artificial neural network and regression models, today, the capabilities of data mining methods have been used to improve the output results of prediction models and field information analysis. Tree models (decision trees) along with decision rules are one of the data mining methods. Tree models are a way of representing a set of rules that lead to a category or value. These models are made by sequentially separating data into separate groups, and the goal in this process is to increase the distance between groups in each separation. Research shows that plant yield is a function of various plant, climatic, and water, and soil management conditions. Therefore, calculating the amount of plant yield and related indices follows complex nonlinear relationships that also have special difficulty in modeling. Considering that the response of irrigated wheat to different inputs in dissimilar climates by field method is time-consuming, costly, and in some cases impossible, so the introduction of an efficient model that can predict yield and analyze yield sensitivity to various parameters is a great help. It will be to solve this problem. This study aimed to develop and evaluate the capability of three models of the neural network, tree, and multivariate linear regression in predicting wheat yield based on parameters affecting its yield in major wheat production hubs in the country.

Materials and Methods: The information used in this study includes the volume of water consumption and yield of irrigated wheat and the committees related to these two indicators in irrigated wheat fields under the management of farmers (241 farms) in the provinces of Khuzestan, Fars, Golestan, Hamadan, Kermanshah, Khorasan Razavi, Ardabil, East Azerbaijan, West Azerbaijan, Semnan, south of Kerman and Qazvin, which were harvested in a field study in the 2016-17 growing season. According to the Ministry of Jihad for Agriculture statistics, these provinces have the highest area under irrigated wheat cultivation in the country and cover about 70% of the area under cultivation and production of this crop in the country.

One of the most widely used monitored neural networks is the Perceptron multilayer network with error replication algorithm, which is suitable for a wide range of applications such as pattern recognition, interpolation, prediction, and process modeling. In the present study, in order to develop the neural network, the capabilities of R software with Neuralnet package have been used. After the normalization step, the data were randomized. This step aims to have a set of inputs and outputs in which the input-output categories do not have a special system. After the randomization of the data, the amount of information that should be used in the network training process is determined. This part of the data was considered for training (70%) and another part for network test (30%). Perceptron neural network activator functions in the implementation of network training and testing. The hyperbolic tangent activity function has been used to limit the range of output data from each neuron and the pattern-to-pattern training process. In the present study and the neural network modeling capability, the tree model method has been used to predict wheat yield. Tree modeling is one of the most powerful and common tools for classification and forecasting. The tree model, unlike the neural network model, produces the law. One of the advantages of the decision tree over the neural network is that it is resistant to input data noise. The tree model divides the data into different sections based on binary divisions. Each data partition can be re-subdivided into another binary, and a model fitted to each subdivision. In this research, the capabilities of WEKA software have been used to run a tree model. It is worth noting that after grouping, the prediction model is applied to the grouped data.

1- Assistant Professor, Agricultural Engineering Research Department, Qazvin Agricultural and Natural Resources Research and Education Center, AREEO, Qazvin, Iran
(*- Corresponding Author Email: a.gomrokchi@areeo.ac.ir)
2 and 3- Assistant Professor and Professor, Agricultural Engineering Research Institute, Agricultural Research, Education and Extension Organization, Karaj, Iran, respectively.
DOI: 10.22067/jsw.2021.15029.0

Results and Discussion: In this study, the efficiency of three models of the artificial neural network, multivariate linear regression, and tree model to predict the performance of irrigated wheat in major production areas in the country was evaluated based on field information recorded in 241 farms. The results showed that the coefficient of explanation of the model in predicting the yield of wheat production in the model of artificial neural network and a multivariate linear regression model was 0.672 and 0.577, respectively, which was applied by grouping the data by tree method. The coefficient of explanation has been increased to 0.762. The output results of the tree model showed that the major wheat production areas in Iran in terms of water consumption could be divided into four independent groups. Finally, it can be concluded that the tree model, considering the purposeful grouping in the input data, can be used as a powerful tool in estimating irrigated wheat yield in major wheat production areas in Iran.

Conclusion: In this study, the need to use data mining methods in analyzing field information and organizing large databases and the usefulness of data mining methods, especially the decision tree in estimating wheat crop yield, were investigated and compared with other forecasting methods. The general results of the research show that purposeful separation of input data into forecasting models can increase the output accuracy of forecasting models. However, it is not possible to provide a general approach to selecting or not selecting a forecasting model in different regions. In some studies, neural networks have shown a high ability to predict the performance of different products, but it is important to note that if there is sufficient data and correct understanding of the factors affecting the dependent variable, the accuracy of the models can be applied by data mining methods. It also improved the neural network. In a general approach, considering the accuracy of estimating the predicted models under study, these techniques can be used to estimate other late-finding characteristics of plants and soil.

Keywords: Data mining, Grouping, Modeling, Water consumption volume