نشریه آبیاری و زهکشی ایران
شماره۵، جلد ۱۳، آذر - دی ۱۳۹۸، ص. ۱۲۹۳-۱۲۸۳

Iranian Journal of Irrigation and Drainage
No. 5, Vol. 13, Dec.-Jan. 2019, p. 1283-1293

مقایسه‌ی مدل‌های درخت تصمیم M5 و الگوریتم نزدیک‌ترین همسایگی K در پیش‌بینی بارش ماهانه (مطالعه موردی: ایستگاه سینوپتیک بیرجند)

فاطمه پورصالحی[۱*]، علی شهیدی[۲]، عباس خاشعی سیوکی[۳]

تاریخ دریافت: ۱۳۹۷/۱۲/۱۷ تاریخ پذیرش: ۱۳۹۸/۴/۳

چکیده

باتوجه به واقع شدن ایران در اقلیم خشک و نیمه خشک، توزیع ناهمگن بارندگی و همچنین وقوع پدیده‌ی تغییراقلیم سبب ایجاد پدیده‌هایی مانند سیل، خشکسالی، بیان‌زایی و تولید ریزگردها و نیز ایجاد خسارت‌های مختلف اقتصادی، اجتماعی و زیست‌محیطی گردیده است. یکی از راهکارهای اولیه جهت کاهش این خسارات، پیش‌بینی رخداد بارندگی است. هدف از مطالعه‌ی حاضر پیش‌بینی بارش ماهانه با بکارگیری روش‌های داده‌کاوی الگوریتم‌های درخت تصمیم (M5) و نزدیک‌ترین همسایگی K (KNN) و مقایسه‌ی این دو روش در راستای تعیین روش کاراتر در زمینه‌ی پیش‌بینی بارندگی با استفاده از داده‌های هواشناسی ماهانه‌ی ایستگاه سینوپتیک بیرجند طی دوره‌ی آماری ۲۰۱۰-۱۹۶۱ میلادی در سه حالت داده خام، میانگین متحرک سه‌ساله و میانگین‌متحرک پنج‌ساله در نرم افزار Weka می‌باشد. نتایج نشان داد که در تمامی سناریوهای تعریف شده، مدل درختی M5 نسبت به مدل KNN توانایی بیشتری در پیش‌بینی بارش ماهانه‌ی این ایستگاه دارد. همچنین پس از بررسی معیارهای ارزیابی R، RMSE، MAE و NS، سناریو پانزدهم با پارامترهای ورودی اختلاف میانگین حداکثر و حداقل دما، متوسط رطوبت نسبی، میانگین سرعت باد و درجه روز سرمایش (بر پایه ۲۱ درجه سانتی‌گراد) در هر ماه به عنوان بهترین سناریو برای پیش‌بینی بارش همان ماه تعیین گردید. همچنین نتایج به دست آمده از مقایسه‌ی سناریوهای تعریف شده در هر مدل در سه حالت داده‌های خام، میانگین متحرک سه ساله و میانگین متحرک پنج ساله نشان می‌دهد که در اکثر سناریوها میانگین متحرک پنج ساله به طور میانگین با مقادیر R=۰/۹۰۴، RMSE=۶/۰۵۴ و MAE=۴/۷۸۰ در مدل M5 و به طور میانگین با مقادیر R=۰/۸۳۷، RMSE=۷/۶۹۸ و MAE=۵/۵۹۵ در مدل KNN پیش‌بینی دقیق‌تری از بارش ماهانه را ارائه می‌دهد.

واژه‌های کلیدی: خشکسالی، درخت تصمیم، روش‌های داده‌کاوی، نرم‌افزار Weka، نزدیک‌ترین همسایگی K

مقدمه

کشور ایران بر روی کمربند خشک جهان قرار دارد و با بارندگی معادل یک سوم متوسط جهانی، کشوری خشک است و به این دلیل، خشکی جزء صفات ذاتی آن محسوب می‌شود. روند بارندگی در ایران حاکی از آن است که این کشور به سوی خشکی پیش می‌رود و می‌بایست برنامه‌ریزی‌ها و تدابیر در مدیریت منابع آب بر این اساس پی‌ریزی شود. نتایج مطالعه و بررسی بلند مدت بارندگی و درجه حرارت نشان می‌دهد که در مجموع، ۶۵ درصد از اراضی کشور در گستره‌ی اقلیم‌های خشک و فراخشک قرار دارند (نعیمی و احقاقی، ۱۳۸۱). پیش‌بینی بارش و برآورد نزولات جوی، به عنوان یکی از مهم‌ترین پارامترهای اقلیمی در حوزه‌ی مدیریت منابع آب، از اهمیت ویژه‌ای جهت بهینه‌سازی صرف هزینه‌ها و استفاده از این منابع برخوردار است. بارش، پدیده‌ای بسیار پیچیده، غیرخطی و نسبت به زمان و مکان متغیر می‌باشد. عوامل مؤثر زیادی در تغییرات آن نقش دارند؛ که به طور کلی می‌توان آن‌ها را به دو دسته‌ی اقلیمی و جغرافیایی تقسیم کرد. از جمله عوامل اقلیمی می‌توان به رطوبت، فشار، دما، پوشش ابر، سرعت باد، عوامل فصلی و همچنین سیگنال‌های بزرگ مقیاس اقلیمی و به خصوص دوره‌های النینو و لانینا اشاره نمود. از عوامل جغرافیایی نیز می‌توان به دوری و نزدیکی از مراکز تولید جبهه (مثلاً دریاها و یا صحراهای بزرگ) و همچنین ارتفاع، اشاره کرد (مهدوی، ۱۳۷۴).

بنابراین با توجه به اهمیت بارش به عنوان یکی از پارامترهای مهم هواشناسی، شناخت کافی از مقدار و همچنین عوامل مؤثر بر این عنصر، بررسی تغییرات و پیش‌بینی آن، برای رسیدن به برنامه‌ریزی کاراتر در مدیریت بخش‌های کشاورزی، اقتصادی و اجتماعی و همچنین پیش‌بینی پدیده‌های هیدرولوژیکی مانند سیلاب و خشکسالی ضروری می‌باشد. در علم آمار روش‌های مختلفی برای

۱- دانشجوی دکتری مهندسی منابع آب دانشگاه بیرجند
۲- دانشیار گروه علوم و مهندسی آب دانشگاه بیرجند
۳- دانشیار گروه علوم و مهندسی آب دانشگاه بیرجند
(* - نویسنده مسئول: Email: Fatemehpursalehi@birjand.ac.ir)

دسته بندی، شناخت الگوها، پیش‌بینی و مدل‌سازی داده‌ها وجود دارد که در یک نگاه کلی می‌توان این روش‌ها را به دو دسته پارامتری و ناپارامتری تقسیم‌بندی نمود. در مدل‌های ناپارامتری، مرحله تخمین پارامترها وجود ندارد. یکی از شناخت شده‌ترین مدل های ناپارامتری روش K نزدیک‌ترین همسایه است. این روش به‌طور گسترده‌ای در علوم مختلف از جمله شناخت الگوهای داده‌ها و دسته‌بندی اطلاعات مورد استفاده قرار گرفته است (Lall and Yakowitz., 1985; Sharma., 1996). همچنین روش‌های داده کاوی برای مجموعه داده‌های بزرگ با متغیرهای زیاد ساخته شده‌اند، بنابراین متفاوت از روش‌های آماری قدیمی هستند که برای مجموعه داده‌های کوچک با متغیرهای اندک طراحی شده‌اند. روش‌های بر مبنای درخت یکی از تکنیک‌های داده کاوی است که در این روش‌ها خروجی به صورت یک مدل با سازه درختی با استفاده از داده‌های ورودی و خروجی می‌باشد. الگوریتم M5 رایج‌ترین طبقه‌بندی استفاده شده در خانواده مدل تصمیم‌گیری درختی است (ستاری و نهرین، ۱۳۹۲).

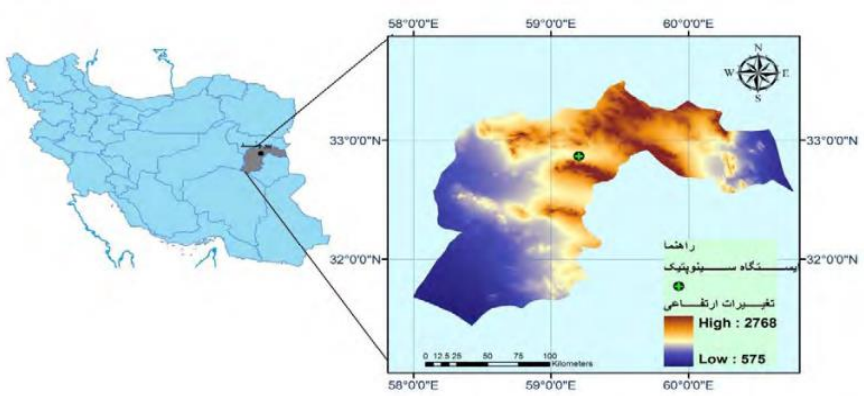
در سال‌های اخیر مطالعات بسیاری در زمینه‌ی پیش‌بینی بارش و تعیین پارامترهای مؤثر بر آن با استفاده از روش‌های مختلف انجام گردیده است که از این میان می‌توان به مطالعات Chuan (1997) ، Trafalis et al., (2002)، Maria et al., (2005)، (2007) Chattopadhyay و Hung et al., (2008) اشاره نمود که با استفاده از روش شبکه عصبی مصنوعی مقادیر بارش را برآورد نمودند.

در ایران نیز خلیلی و همکاران (۱۳۸۹)، با استفاده از روش شبکه عصبی مصنوعی مقدار بارش ماهانه‌ی ایستگاه سینوپتیک مشهد را پیش‌بینی نمودند که نتایج به دست آمده بر کارایی این روش دلالت داشت. فدایی‌کرمانی و همکاران (۱۳۹۳)، با استفاده از روش الگوریتم K نزدیکترین همسایگی خشکسالی بر مبنای شاخص بارش استاندارد را در شهرستان بم پیش‌بینی نمودند. نتایج به دست آمده بیان می‌کند که مدل KNN می‌تواند پیش‌بینی های قابل قبولی از وضعیت خشکسالی منطقه را ارائه نماید. سیدکابلی و همکاران (۱۳۹۱)، براساس روش ناپارامتریک نزدیکترین همسایگی مدلی جهت ریزمقیاس نمایی داده‌های اقلیمی ارائه نمودند. نتایج به دست آمده نشان داد که روش جمعی مبتنی بر الگوریتم نزدیکترین همسایگی عملکرد بهتری در مقایسه با روش جمعی غیرخطی احتمالاتی دارد و از عدم قطعیت کمتری در پیش‌بینی برخوردار است. قربانی (۱۳۹۳)، کارایی مدل‌های داده‌کاوی ماشین بردار پشتیبان، درخت تصمیم و نزدیک‌ترین همسایگی k را در ریزمقیاس نمایی بارش بر اساس داده‌های مدل گردش عمومی NCEP را بررسی نمود. نتایج این بررسی نشان داد که بارش شبیه‌سازی شده با هر یک از مدل‌های داده‌کاوی، دارای میانگین و انحراف معیار کمتری نسبت به داده‌های مشاهداتی هستند و مقادیر حدی را نمی‌توانند به خوبی پیش‌بینی کنند. با این وجود روش نزدیکترین همسایگی نسبت به دیگر روش‌ها نتایج بهتری را ارائه کرد. امیدوار و همکاران (۱۳۹۳)، توانایی مدل درخت تصمیم را در پیش بینی بارش ایستگاه سینوپتیک کرمانشاه مورد ارزیابی قرار دادند. نتایج نشان داد که در ایستگاه سینوپتیک کرمانشاه درخت تصمیم گیری رگرسیونی، مدلی نسبتاً کارا در پیش بینی بارش می‌باشد که استفاده از میانگین متحرّک نسبت به سایر حالات منجر به افزایش چشمگیر کارایی مدل درخت تصمیم می‌شود و در صورت تعدیل دامنه تغییرات داده‌های ورودی قادر است با ضریب اطمینان بالایی میزان بارش را ۳۰ ماه قبل از وقوع برآورد نماید که در شبیه سازی‌های صورت گرفته، زمانی که از میانگین متحرّک پنج ساله داده‌ها برای اجرای مدل استفاده گردیده، ترکیب بارش قبلی، دمای ماکزیمم به عنوان مناسب‌ترین حالت شناسایی شده است. دستورانی و همکاران (۱۳۹۱)، کارایی مدل درخت تصمیم را در پیش‌بینی بارش ایستگاه سینوپتیک یزد تعیین نمودند. نتایج نشان داد که در ایستگاه یزد، مدل درخت تصمیم‌گیری خصوصاً در شرایطی که از میانگین متحرک ۵ ساله داده‌ها استفاده گردد، دارای توانایی مناسبی در پیش‌بینی میزان بارش می‌باشد. ستاری و نهرین (۱۳۹۲)، با استفاده از سیستم‌های هوشمند (شبکه‌ی عصبی مصنوعی و برنامه‌ریزی ژنتیک) مقادیر حداکثر بارش روزانه‌ی ایستگاه‌های اهر و جلفا را پیش‌بینی و با مدل درختی M5 مقایسه نمودند. در حالت کلی می‌توان گفت که هر سه روش مذکور ضمن رقابت با یکدیگر نتایج نسبتا دقیقی را جهت پیش‌بینی حداکثر بارش روزانه در ماه مورد نظر در منطقه ارائه می‌کنند ولی به دلیل ارائه روابط خطی ساده و قابل فهم توسط مدل درختی M5، این روش می‌تواند به عنوان روشی کاربردی و جایگزین برای محاسبه حداکثر بارش روزانه در ماه مورد توجه قرار گیرد.

با توجه به مطالعات انجام شده در زمینه‌ی پیش‌بینی بارش، می‌توان به اهمیت برآورد این متغیر در میان سایر پارامترهای هواشناسی پی برد. در این مطالعه پارامترهای مؤثر بر پیش‌بینی بارش ماهانه تعیین و دو مدل ناپارامتریک درخت تصمیم M5 و الگوریتم KNN در راستای برآورد این پدیده‌ی هواشناسی با یکدیگر مقایسه و مدل برتر ارائه گردید، که تاکنون در ایران مقایسه‌ای بین این دو مدل داده کاوی جهت پیش‌بینی بارش ماهانه صورت نگرفته است.

مواد و روش‌ها

ایستگاه سینوپتیک بیرجند با ارتفاعی معادل ۱۴۹۱ متر از سطح دریا در موقعیت جغرافیایی ۱۲′ ۵۹° طول شرقی و ۵۲′ ۳۲° واقع شده است (شکل ۱).



شکل ۱- موقعیت جغرافیایی ایستگاه سینوپتیک بیرجند (هادی‌زاده و همکاران، ۱۳۹۰)

درخت تصمیم‌گیری M5[1]

روش درخت تصمیم یک روش سلسله مراتبی یا چند مرحله‌ای است که در آن به صورت بازگشتی مجموعه داده‌ها به روش دودویی به تقسیمات فرعی و کوچک‌تر تقسیم‌بندی می‌شود تا زمانی که تقسیمات فرعی نهایی نتوانند بیشتر از آن تجزیه شوند. درختان تصمیم استقرایی مجموعه‌ای از داده‌های معلوم را می‌گیرد و یک درخت تصمیم را از آن استنتاج می‌کند. سپس درخت می‌تواند به صورت مجموعه قوانینی برای پیش‌بینی نتیجه ویژگی‌های معلوم استفاده شود (طالبی و اکبری، ۱۳۹۲). اولین مرحله برای ایجاد یک مدل درختی، استفاده از یک معیار انشعاب است. معیار انشعاب برای الگوریتم M5 براساس عملکرد انحراف استاندارد مقادیر هر کلاس و یا طبقه است که در هر گره به دست آمده است. این روش اساس روش‌های طبقه‌بندی است که آنتروپی نامیده می‌شود. آنتروپی می‌تواند به عنوان معیار میزان آشفتگی و بی‌نظمی یک سیستم تفسیر شود. معیار انشعاب بیانگر میزان خطا در آن گره می‌باشد و مدل حداقل خطای مورد انتظار را به عنوان نتیجه‌ی آزمایش هر صفت در آن گره محاسبه می‌کند. خطای مدل عموما با اندازه‌گیری دقت پیش‌بینی مقادیر هدف موارد دیده نشده سنجش می‌شود. فرمول محاسبه‌ی کاهش انحراف استاندارد (SDR) به صورت روابط زیر می‌باشد:

$$SDR = Sd(T) - \sum_{i=1}^{N} \frac{|T_i|}{|T|} Sd(T_i) \quad (۱)$$

$$Sd(T) = \sqrt{\frac{1}{N}(\sum_{i=1}^{N} y_i^2 - \frac{1}{N}(\sum_{i=1}^{N} y_i)^2)} \quad (۲)$$

در این رابطه T مجموعه‌ای از نمونه‌هاست (موارد) که به هر گره وارد می‌شود، T_i نشان‌دهنده‌ی زیرمجموعه‌ای از نمونه‌هاست که i امین نتیجه‌ی تست پتانسیلی را دارند، Sd بیانگر انحراف معیار، y_i مقدار عددی ویژگی هدف نمونه‌ی i و N شماره‌ی داده‌ها را نشان می‌دهد (Alberg et al., 2012؛ ستاری و نهرین، ۱۳۹۲).

نزدیک‌ترین همسایگی k (KNN)[2]

در حالت کلی از این الگوریتم به دو منظور استفاده می‌شود: برای تخمین تابع چگالی توزیع داده‌های تعلیم و برای طبقه‌بندی داده‌های تست براساس الگوهای تعلیم. بر خلاف توابع انتقالی کلاسیک، مدل نزدیکترین همسایگی از هیچ تابع ریاضی از پیش تعریف شده‌ای برای تخمین متغیرهای مختلف استفاده نمی‌کند. به طور کلی می‌توان گفت که مدل نزدیکترین همسایگی، یکی از روش‌های داده‌کاوی است که هدف کلی آن طبقه‌بندی و تخمین ویژگی‌های یک سری داده‌های مجهول با توجه به بیش‌ترین شباهت این داده‌ها با داده‌های معلومی است که در همسایگی (نزدیکی) آن‌ها قرار دارند (Xindung and Kumar., 2009؛ فدایی کرمانی و همکاران، ۱۳۹۳). در این مدل داده‌های هدف مورد جستجو با توجه به نزدیک‌ترین فاصله نسبت به داده‌های آموزش (بانک داده مرجع)، طبقه‌بندی می‌شوند. نخستین گام در استفاده از این مدل، یافتن روش و رابطه‌ای برای محاسبه‌ی فاصله‌ی بین داده‌های مورد آزمایش و داده‌های تعلیم است. معمولا برای تعیین این فاصله از فاصله‌ی اقلیدسی زیر استفاده می‌شود (Jagtap et al., 2004):

$$d(X.Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (۳)$$

که در آن X نماینده‌ی داده‌های تعلیم با پارامترهای مشخص (x_1 تا x_n) و Y نماینده‌ی داده‌های آموزش با همان تعداد پارامترهای مشخص (y_1 تا y_n) است.

1- Decision Tree M5

2- K-Nearest Neighborhood

$$X = (x_1 . x_2 . \dots . x_n) \quad (۴)$$

$$Y = (y_1 . y_2 . \dots . y_n) \quad (۵)$$

پس از تعیین فاصله‌ی اقلیدسی بین داده‌ها، نمونه‌های بانک داده به ترتیب صعودی از کم‌ترین فاصله (حداکثر تشابه) تا بیشترین فاصله (حداقل تشابه) از نمونه‌ی مورد نظر، طبقه‌بندی و ارزش‌گذاری می‌شوند. قدم بعدی در این مدل یافتن تعداد نقاطی (k) از داده‌های آزمایش برای تخمین ویژگی‌های بانک داده‌ی موردنظر است. تعیین تعداد همسایه‌ها (k)، یکی از کلیدی‌ترین و مهم‌ترین مراحل در مدل یاد شده به شمار می‌آید و میزان کارایی این روش به طور قابل ملاحظه‌ای به کیفیت انتخاب نزدیک‌ترین (مشابه‌ترین) نمونه‌ها از داده‌های بانک مرجع بستگی دارد (Xindung and Kumar., 2009؛ فدایی کرمانی و همکاران، ۱۳۹۳).

نرم‌افزار Weka

این نرم افزار اولین بار در سال ۱۹۹۲ به منظور جمع آوری و یکپارچگی الگوریتم‌های یادگیری ماشین و ابزاری برای پردازش داده‌ای با استفاده از JAVA پیاده سازی و به صورت کد باز تحت مجوز عمومی GNU انتشار گردید و در سال ۱۹۹۳ با بکارگیری اهداف داده‌کاوی ارتقا داده شد. در حال حاضر این نرم افزار حاوی تعداد زیادی از تکنیک‌های یادگیری ماشین و داده‌کاوی است که امکان مقایسه‌ی تکنیک‌های مختلف یادگیری ماشین را می‌دهد. در این واسط گرافیکی کاربر اجازه‌ی دسترسی به قابلیت‌هایی مانند تجسم و تحلیل بسیاری از الگوریتم‌های داده کاوی را دارد (Witten and Frank., 2000).

در این مطالعه ابتدا پارامترهای اقلیمی مؤثر بر بارش مانند اختلاف میانگین حداکثر و حداقل دما ($\overline{T}max-\overline{T}min$)، دمای حداکثر (Tmax) و حداقل (Tmin)، میانگین رطوبت نسبی($R\overline{H}$)، میانگین حداکثر ($R\overline{H}max$) و حداقل رطوبت نسبی ($R\overline{H}min$)، میانگین فشار بخار آب ($W\overline{P}$)، میانگین فشار هوا ($A\overline{P}$)، میانگین سرعت باد ($W\overline{S}$)، درجه روز سرمایش (پایه ۲۱ درجه سانتی‌گراد) (CDD21)، تعداد روزهای ابری (NCD) و طوفانی (NSD) برای یک دوره‌ی آماری ۵۰ ساله طی سال‌های ۲۰۱۰-۱۹۶۱ میلادی به صورت ماهانه به عنوان داده‌های ورودی به نرم افزار Weka و داده‌های بارش ماهانه به عنوان خروجی نرم‌افزار، از سامانه‌ی اطلاعاتی ایستگاه سینوپتیک بیرجند استخراج گردید. سپس از تعداد ۶۰۰ داده‌ی موجود در هر پارامتر، ۷۰ درصد از داده‌ها معادل۴۲۰ داده به عنوان داده‌های تولید و ۳۰ درصد از داده‌ها برابر با۱۸۰ داده به عنوان داده‌های آزمون، در سه حالت داده‌های خام، میانگین متحرک سه ساله و میانگین متحرک پنج ساله داده‌ها در قالب بیست سناریوی تعریف شده در تحقیق (جدول ۱) جهت پیش‌بینی بارش در همان ماه به هر یک از مدل‌های درخت تصمیم M5 و نزدیک‌ترین همسایگی K وارد گردید.

پس از اجرای مدل‌ها، سناریوهای فوق در حالت‌های مختلف با استفاده از معیارهای ارزیابی مورد بررسی قرار گرفتند.

جدول ۱-سناریوهای تعریف شده در مدل‌های داده‌کاوی*

سناریو	پارامترهای ورودی تعریف شده در هر سناریو
1	$\overline{T}max-\overline{T}min$, Tmax , Tmin, $R\overline{H}$, $R\overline{H}max$, $R\overline{H}min$, $W\overline{P}$, $A\overline{P}$, $W\overline{S}$, CDD21, NCD, NSD
2	$R\overline{H}$, CDD21
3	$R\overline{H}max$, CDD21
4	Tmin, $R\overline{H}max$, NCD
5	$R\overline{H}$, $W\overline{P}$, $A\overline{P}$, $W\overline{S}$
6	Tmin, CDD21, NSD
7	$\overline{T}max-\overline{T}min$, $W\overline{S}$, NCD
8	Tmax , $R\overline{H}max$
9	Tmin, $R\overline{H}max$, $W\overline{P}$, CDD21, NCD
10	$W\overline{P}$, $A\overline{P}$, NCD
11	Tmax, $R\overline{H}min$, $A\overline{P}$, $W\overline{S}$
12	Tmin, $R\overline{H}max$, $W\overline{P}$, $W\overline{S}$, CDD21, NCD, NSD
13	$\overline{T}max-\overline{T}min$, $A\overline{P}$, $W\overline{S}$, NCD
14	$\overline{T}max-\overline{T}min$, $R\overline{H}$, $W\overline{P}$, $A\overline{P}$, $W\overline{S}$, CDD21
15	$\overline{T}max-\overline{T}min$, $R\overline{H}$, $W\overline{S}$, CDD21
16	$\overline{T}max-\overline{T}min$, $R\overline{H}$, $W\overline{P}$, $A\overline{P}$, $W\overline{S}$, CDD21, NCD, NSD
17	$R\overline{H}$, $W\overline{S}$, CDD21
18	Tmin, $R\overline{H}$, CDD21
19	Tmin, $R\overline{H}max$, $A\overline{P}$, CDD21
20	$R\overline{H}min$, $W\overline{P}$, $A\overline{P}$, $W\overline{S}$, NSD

*در کلیه‌ی سناریوهای مورد نظر بارش ماهانه به عنوان پارامتر خروجی به مدل معرفی شده است.

معیارهای ارزیابی

جهت ارزیابی مدل‌ها در سه حالت مختلف داده‌های خام، میانگین متحرک سه ساله و میانگین متحرک پنج ساله و همچنین تعیین بهترین سناریو از سه معیار ضریب همبستگی (r)، جذر میانگین مربعات خطا (RMSE) و میانگین قدر مطلق خطا (MAE) استفاده گردید. همچنین برای ارزیابی بهترین سناریو در حالت‌های مورد نظر علاوه بر سه مدل ارزیابی فوق معیار ضریب نش–ساتکلیف (NS) محاسبه گردید.

$$r = \frac{\sum_{i=1}^{n}(Oi - \overline{O})(Pi - \overline{P})}{\sqrt{\sum_{i=1}^{n}(Oi - \overline{O})^2}\sqrt{\sum_{i=1}^{n}(Pi - \overline{P})^2}} \quad (۶)$$

$$RMSE = \sqrt{\sum_{i=1}^{N}\frac{(Pi - Oi)^2}{N}} \quad (۷)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Pi - Oi| \quad (۸)$$

$$NS = 1 - \frac{\sum_{i=1}^{N}(Oi - Pi)^2}{\sum_{i=1}^{N}(Oi - \overline{O})^2} \quad (۹)$$

در روابط فوق Oi، $\overline{Oi}$، Pi، $\overline{Pi}$ و N به ترتیب مقادیر مشاهداتی، میانگین مقادیر مشاهداتی، مقادیر پیش‌بینی شده، میانگین مقادیر پیش‌بینی شده و تعداد مشاهدات را نشان می‌دهد. پایین بودن مقادیر خطا (RMSE و MAE) و بالا بودن مقادیر ضریب همبستگی و ضریب نش–ساتکلیف نشان دهنده‌ی دقت مدل و حالت در نظر گرفته شده است.

نتایج و بحث

نتایج ارزیابی سناریوها با استفاده از معیارهای ارزیابی (جدول ۲)، نشان می‌دهد که از میان بیست سناریوی تعریف شده در تحقیق حاضر، سناریوی پانزدهم با متغیرهای ورودی اختلاف میانگین حداکثر و حداقل دما، متوسط رطوبت نسبی، میانگین سرعت باد و درجه روز سرمایش (بر پایه ۲۱ درجه سانتی‌گراد) ماهانه با مقادیر R=۰/۷۲، RMSE=۱۳/۳۳ و MAE=۷/۰۰۸ در مدل KNN و R=۰/۸۶، RMSE=۹/۷۱ و MAE=۵/۵۷ در مدل M5 در داده‌های خام، R=۰/۸۷، RMSE=۷/۲۱۴ و MAE=۴/۶۵۵ در مدل KNN و R=۰/۹۳۳، RMSE=۵/۲۵۶ و MAE=۳/۵۴۵ در مدل M5 در داده‌های میانگین متحرک سه ساله و R=۰/۸۸۸، RMSE=۵/۵۸۲ و MAE=۴/۱۱۲ در مدل KNN و R=۰/۹۴۴، RMSE=۳/۹۳ و MAE=۳/۱۵۷ در مدل M5 در داده‌های میانگین متحرک پنج ساله پاسخ بهتری به پیش‌بینی مقادیر بارش در همان ماه می‌دهد. همچنین مقایسه‌ی حالت‌های بیان شده در هر یک از مدل‌ها حاکی از آن است که در تمامی سناریوها به جز دو سناریوی یازدهم و بیستم میانگین متحرک پنج ساله نسبت به میانگین متحرک سه ساله و داده‌های خام پیش‌بینی دقیق‌تری را ارائه می‌دهد. در دو سناریوی یازدهم و بیستم نیز پیش‌بینی بر اساس داده‌های میانگین متحرک سه ساله بهتر از میانگین متحرک پنج ساله و هر دوی این حالات از پیش‌بینی بر اساس داده‌های خام بهتر می‌باشد. بنابراین می‌توان دریافت که پیش‌بینی بر اساس داده‌های میانگین متحرک دقت بیشتری نسبت به داده‌های خام ایستگاه سینوپتیک بیرجند دارد. نتایج بیان شده در زمینه‌ی مقایسه‌ی حالات مختلف داده‌ها و کارایی مدل درختی M5 با نتایج دستورانی و همکاران (۱۳۹۱)، که حاکی از توانایی مدل درخت تصمیم‌گیری خصوصاً در شرایط استفاده از میانگین متحرک ۵ ساله داده‌ها می‌باشد، مطابقت دارد. از طرفی مقایسه‌ی بین دو مدل داده‌کاوی درخت تصمیم و الگوریتم k نزدیک‌ترین همسایه نشان می‌دهد که در تمامی سناریوها در پیش‌بینی بارش ماهانه مدل درختی M5 به طور میانگین با مقادیر R=۰/۸۱۳، RMSE=۱۱/۵۸۸ و MAE=۷/۴۴۶ در داده‌های خام، R=۰/۸۹۷، RMSE=۷/۴۴۶ و MAE=۵/۲۴۵ در داده‌های میانگین متحرک سه ساله و R=۰/۹۰۴، RMSE=۶/۰۵۴ و MAE=۴/۷۸۰ در داده‌های میانگین متحرک پنج ساله نسبت به الگوریتم KNN به طور میانگین با مقادیر R=۰/۷۱۹، RMSE=۱۵/۲۹۶ و MAE=۸/۴۹۳ در داده‌های خام، R=۰/۸۳۸، RMSE=۹/۰۷۱ و MAE=۵/۹۸۲ در داده‌های میانگین متحرک سه ساله و R=۰/۸۳۷، RMSE=۷/۶۹۸ و MAE=۵/۵۹۵ در داده‌های میانگین متحرک پنج ساله ارجحیت دارد.

جدول ۲- نتایج ارزیابی مدل‌ها

		داده های خام		میانگین متحرک ۳ ساله		میانگین متحرک ۵ ساله	
		KNN	M5	KNN	M5	KNN	M5
سناریو ۱	R	۰/۷۸	۰/۸۴	۰/۸۸۷	۰/۹۰۷	۰/۸۷۶	۰/۹۳۶
	RMSE	۱۳/۱۶	۱۰/۷۵	۷/۴۸۵	۸/۶۶۹	۷/۵۵۴	۵/۴۵۴
	MAE	۷/۲۷	۷/۵۷	۴/۷۹۶	۵/۸۳۴	۵/۲۱۳	۴/۴۲۲
سناریو ۲	R	۰/۷۲	۰/۷۵	۰/۷۷۵	۰/۸۳۵	۰/۸۱۷	۰/۸۶۱
	RMSE	۱۳/۶۶	۱۲/۷۱	۱۰/۰۹۴	۸/۴۶۵	۸/۲۰۲	۶/۷۲۷

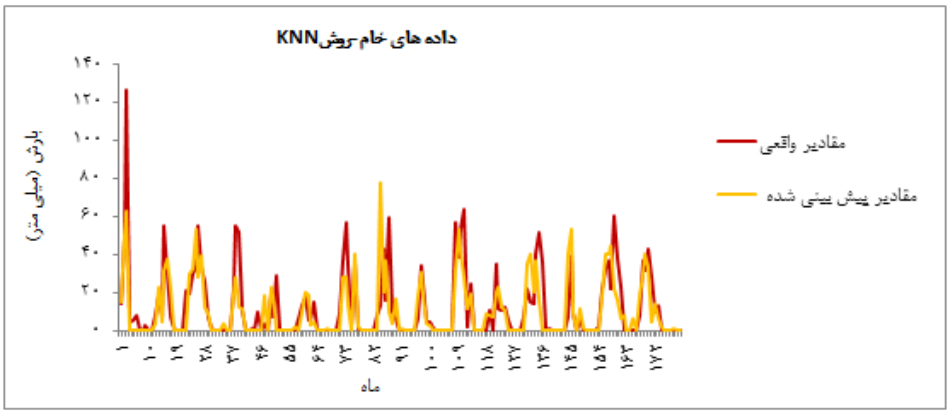
	MAE	۷/۳۱	۸/۵۶	۶/۹۰۴	۶/۵۴۸	۶/۳۴۳	۵/۵۱۶
	R	۰/۶۸	۰/۷۶	۰/۸۰۸	۰/۸۶۸	۰/۸۴۵	۰/۸۹
سناریو ۳	RMSE	۱۴/۵۴	۱۲/۵۲	۱۰/۰۶۳	۷/۸۷۶	۷/۷۸۳	۶/۴۴
	MAE	۸/۱۴	۷/۳۵	۶/۷۳۹	۵/۴۶۸	۵/۸۳۹	۵/۰۷۵
	R	۰/۶۸	۰/۸۳	۰/۸۳۹	۰/۹۲۳	۰/۸۳۳	۰/۹۱۸
سناریو ۴	RMSE	۱۵/۹۴	۱۰/۸	۹/۱۸۷	۶/۲۷۹	۷/۴۷۳	۵/۴۴۵
	MAE	۸/۹۳	۶/۶۷	۶/۱۰۳	۴/۳۷۳	۵/۳۵۵	۳/۸۷۹
	R	۰/۷۹	۰/۸۴	۰/۸۶۵	۰/۹۰۲	۰/۸۲۶	۰/۸۹۶
سناریو ۵	RMSE	۱۵/۵۷	۱۱/۵۷	۱۰/۷۸۸	۹/۶۳۶	۸/۷۸۲	۷/۷۱۵
	MAE	۸/۲۳	۷/۳۶	۷/۱۷۸	۶/۹۸	۶/۵۱۳	۶/۴۴۳
	R	۰/۵۸	۰/۷۳	۰/۷۶۹	۰/۸۲	۰/۷۵۱	۰/۸۳۹
سناریو ۶	RMSE	۱۹/۲۴	۱۳/۷۷	۱۰/۳۴	۹/۷۰۲	۹/۴۷۱	۷/۸۳
	MAE	۱۱/۵۴	۹/۳۸	۶/۷۱۸۸	۶/۷۹۹	۶/۸۹	۶/۱۳۴
	R	۰/۷۱	۰/۸۴	۰/۷۹۲	۰/۹۳	۰/۸۴۲	۰/۹۳۳
سناریو ۷	RMSE	۱۳/۶۳	۱۰/۳۶	۸/۹۶۵	۵/۴۸	۶/۶۷۶	۴/۵۰۴
	MAE	۷/۸۱	۶/۶۴	۶/۱۵۷	۳/۹۱۷	۴/۹۲۵	۳/۳۶۹
	R	۰/۶۶	۰/۸	۰/۸۲۷	۰/۸۹	۰/۸۰۲	۰/۸۸۵
سناریو ۸	RMSE	۱۸/۴۸	۱۲/۵۴	۱۰/۲۳	۷/۷۷۳	۸/۲۲۳	۶/۵۵۷
	MAE	۱۰/۰۹	۷/۶۵	۶/۶۷۶	۵/۳۴۳	۶/۱۷۵	۵/۰۸
	R	۰/۷۵	۰/۸۱	۰/۸۵۸	۰/۹۱۵	۰/۸۶۶	۰/۹۲۱
سناریو ۹	RMSE	۱۴/۱۸	۱۱/۲۷	۸/۴۶۵	۶/۵۳۸	۶/۵۲	۵/۱۷
	MAE	۷/۴۷	۷/۴۲	۵/۴۴	۴/۳۷	۴/۷۳	۳/۹۵۳
	R	۰/۶۷	۰/۷۲	۰/۷۵	۰/۸۳۹	۰/۸۲	۰/۸۵۳
سناریو ۱۰	RMSE	۱۵/۴۷	۱۳/۰۷	۱۰/۵۷۷	۸/۰۱۳	۸/۰۸۶	۶/۲۳
	MAE	۸/۹	۸/۱۷	۶/۸۰۷	۵/۴۴۹	۵/۶۸۷	۴/۵۴۹
	R	۰/۷۶	۰/۸۱	۰/۸۸۱	۰/۸۷۷	۰/۸۲۱	۰/۸۶۶
سناریو ۱۱	RMSE	۱۶/۹۹۶	۱۳/۱۷	۸/۹۰۳	۹/۵۱۷	۸/۶۱۱	۸/۷۳۲
	MAE	۹/۷۵	۹/۰۵	۵/۹۲۲	۷/۱۷۲	۶/۴۷۵	۷/۳۸۸
	R	۰/۷۲	۰/۸۵	۰/۸۷۱	۰/۹۱۷	۰/۸۷	۰/۹۲۲
سناریو ۱۲	RMSE	۱۴/۶۳	۱۰/۶۵	۷/۸۷۴	۶/۸۳۳	۶/۰۹	۵/۲۲۱
	MAE	۸/۱۱	۶/۶۷	۵/۰۹	۴/۴۱۸	۴/۲۷۱	۳/۷۹۲
	R	۰/۷۱	۰/۷۸	۰/۸۶۶	۰/۹۲۳	۰/۸۶۱۸	۰/۹۳۳
سناریو ۱۳	RMSE	۱۶/۶۹	۱۱/۷۸	۷/۸۹۱	۶/۰۰۷	۶/۴۷۸	۴/۴۲۵
	MAE	۹/۴۴	۸/۱۴	۵/۳۱۷	۴/۰۵۵	۴/۵۴۵	۳/۲۹۶
	R	۰/۸	۰/۸۶	۰/۹۰۸	۰/۹۳۴	۰/۸۶۹	۰/۹۴۴
سناریو ۱۴	RMSE	۱۳/۱۶	۱۱/۴۴	۷/۹۴۲	۶/۱۴	۷/۵۱	۴/۹۹۲
	MAE	۷/۳۷	۷/۱۷	۵/۲۱۹	۴/۳۳	۵/۴۵۴	۳/۹۶
	R	۰/۷۲	۰/۸۶	۰/۸۷	۰/۹۳۳	۰/۸۸۸	۰/۹۴۴
سناریو ۱۵	RMSE	۱۳/۳۳	۹/۷۱	۷/۲۱۴	۵/۲۵۶	۵/۵۸۲	۳/۹۳
	MAE	۷/۰۰۸	۵/۵۷	۴/۶۵۵	۳/۵۴۵	۴/۱۱۲	۳/۱۵۷
	R	۰/۷۶	۰/۸۷	۰/۸۸۶	۰/۹۳۸	۰/۸۷۴	۰/۹۳۶
سناریو ۱۶	RMSE	۱۴/۱۳	۱۰/۱۹	۷/۶۸۹	۶/۲۸۸	۷/۳۹۵	۵/۶۲۷
	MAE	۷/۶۳	۶/۵۲	۵/۰۲۷	۴/۲۶۸	۵/۱۰۸	۴/۱۸۷
	R	۰/۷۴	۰/۸	۰/۷۵۶	۰/۸۵۶	۰/۷۹۲	۰/۸۷۸
سناریو ۱۷	RMSE	۱۲/۹۲	۱۱/۲۷	۱۰/۱۹۷	۷/۶۰۸	۸/۱۱	۵/۷۵۴
	MAE	۶/۷۵	۶/۵۴	۶/۴۹۷	۵/۷۵	۵/۸۸۶	۴/۵۸۹

	R	۰/۶۷	۰/۸۳	۰/۸۲۹	۰/۹	۰/۷۹۵	۰/۸۹۳
سناریو ۱۸	RMSE	۱۶/۸۵	۱۱/۲۵	۹/۰۶	۷/۱۳	۸/۸۴۵	۶/۲۰۹
	MAE	۹/۱۲	۶/۸۱	۶/۱۷۲	۴/۹۷۱	۶/۵۴۲	۴/۹۶۷
	R	۰/۷۳	۰/۸۴۵	۰/۸۶۴	۰/۹۱۲	۰/۸۵۲	۰/۹۱۸
سناریو ۱۹	RMSE	۱۷/۴۵	۱۱/۵۴	۹/۱۹۷	۷/۳۹۸	۷/۹۸۱	۶/۴۷۷
	MAE	۱۰/۱۲	۷/۱۸	۶/۱	۵/۲۲۶	۵/۷۶۵	۵/۱۹۲
	R	۰/۷۶	۰/۸۴	۰/۸۶۷	۰/۹۲۴	۰/۸۳۷	۰/۹۲۳
سناریو ۲۰	RMSE	۱۵/۸۹	۱۱/۴	۹/۲۶۱	۸/۳۰۴	۸/۵۹۳	۷/۶۴۷
	MAE	۸/۸۸	۸/۵	۶/۱۲۱	۶/۰۷۷	۶/۰۷۲	۶/۶۵۹

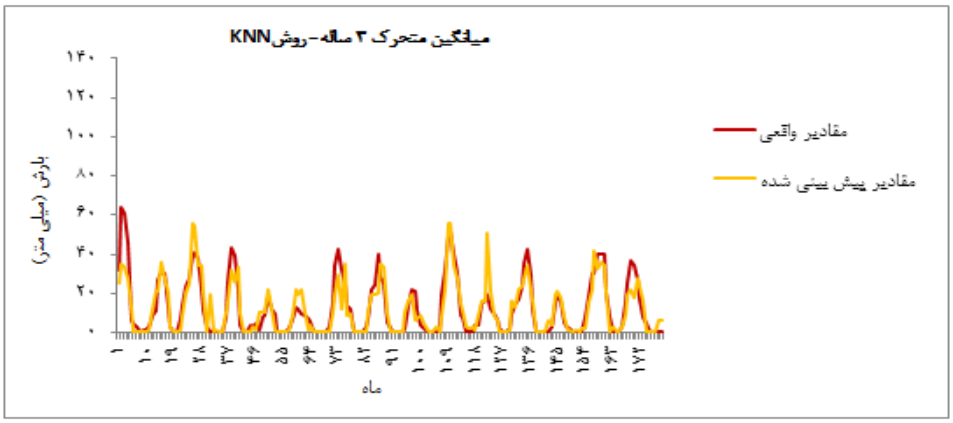
نتایج به دست آمده از برآورد ضریب نش–ساتکلیف برای مقایسه‌ی مدل‌ها و حالات مختلف داده‌ها در سناریوی برتر (جدول ۳) نیز بر دقت بیشتر مدل M5 نسبت به الگوریتم KNN دلالت دارد و همچنین نشان می‌دهد که در بین حالات مختلف داده‌ها پیش‌بینی بارش ماهانه با استفاده از داده‌های میانگین متحرک پنج ساله در مقایسه با دو حالت دیگر بیشتر به واقعیت نزدیک است. با مقایسه‌ی متناظر نمودار مدل‌های مختلف در هر یک از حالات داده‌های ورودی، می‌توان دریافت که در مدل درختی M5 روند مقادیر پیش‌بینی شده تطابق بیشتری بر روند داده‌های واقعی بارش ماهانه نسبت به روش KNN دارد. از طرفی مقایسه‌ی نمودارها در حالت‌های مختلف داده‌های خام، میانگین متحرک سه ساله و میانگین متحرک پنج ساله، نشان می‌دهد که میانگین متحرک پنج ساله به علت نوسان کمتری که در پیش‌بینی بارش دارد، نسبت به سایر حالات بهترین عملکرد را در پیش‌بینی این پارامتر اقلیمی داراست. (شکل‌های ۲–۷)

جدول ۳- نتایج ارزیابی سناریو ۱۵ با ضریب Nash-Sutcliffe

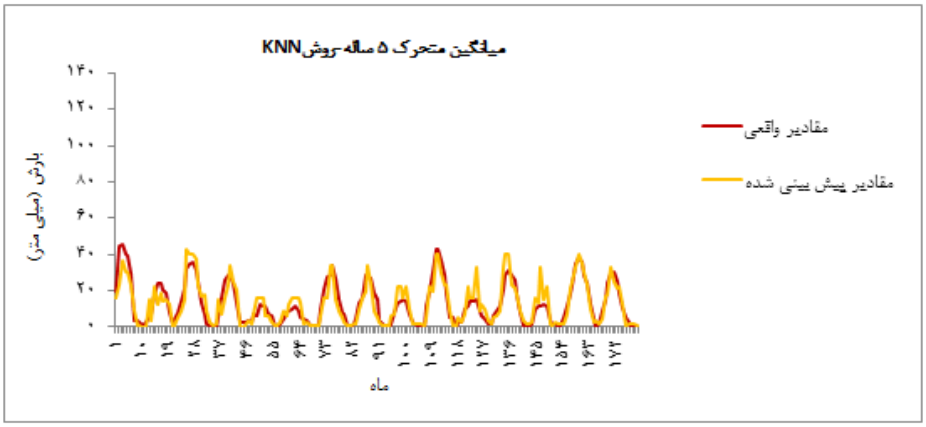
نوع داده‌های ورودی	خم		میانگین متحرک ۳ ساله		میانگین متحرک ۵ ساله	
مدل	KNN	M5	KNN	M5	KNN	M5
ضریب نش–ساتکلیف سناریو ۱۵	۰/۴۹۸	۰/۷۴۳	۰/۷۵۳	۰/۸۷۷	۰/۷۷۷	۰/۸۹۳



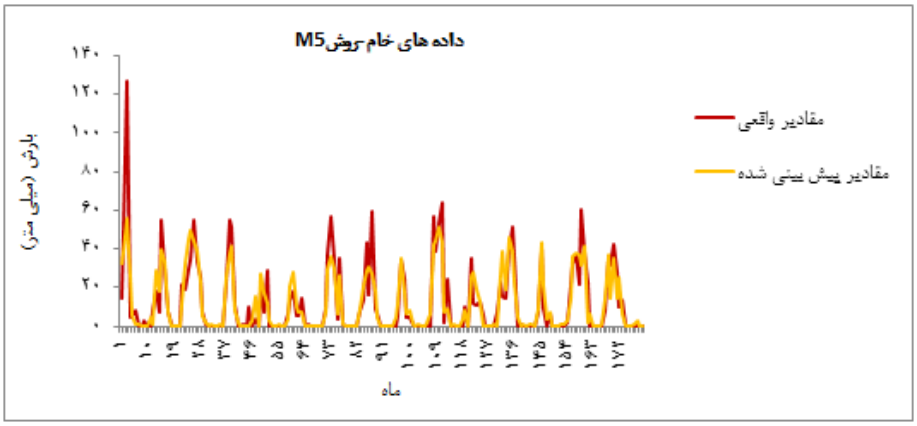
شکل ۲- مقایسه‌ی مقادیر واقعی (داده‌های خام) و پیش‌بینی شده‌ی بارش ماهانه با روش KNN- سناریو ۱۵



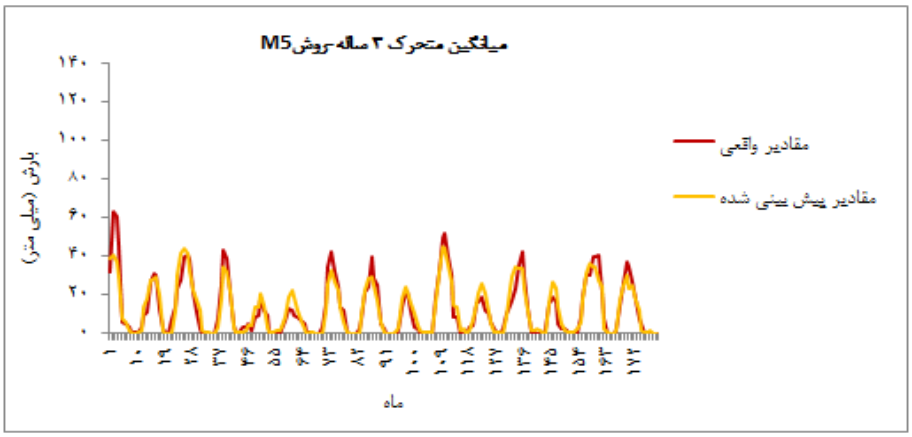
شکل ۳- مقایسه‌ی مقادیر واقعی (میانگین متحرک ۳ ساله) و پیش‌بینی شده‌ی بارش ماهانه با روش KNN- سناریو ۱۵



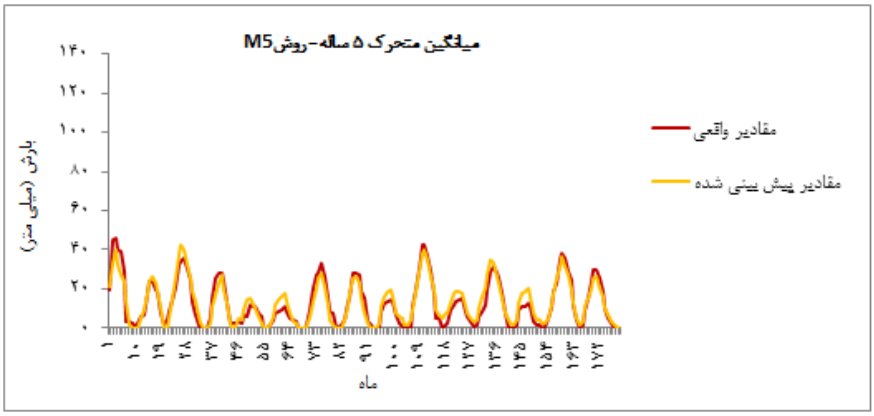
شکل ٤- مقایسه‌ی مقادیر واقعی (میانگین متحرک۵ ساله) و پیش‌بینی شده‌ی بارش ماهانه با روش KNN- سناریو ۱۵



شکل ۵- مقایسه‌ی مقادیر واقعی (داده‌های خام) و پیش‌بینی شده‌ی بارش ماهانه با روش M5- سناریو ۱۵



شکل ٦- مقایسه‌ی مقادیر واقعی (میانگین متحرک ۳ ساله) و پیش‌بینی شده‌ی بارش ماهانه با روش M5- سناریو ۱۵



شکل ۷- مقایسه‌ی مقادیر واقعی (میانگین متحرک۵ ساله) و پیش‌بینی شده‌ی بارش ماهانه با روش M5- سناریو ۱۵

نتیجه‌گیری

امروزه با توجه به خسارات قابل توجهی که در اثر وقوع پدیده‌های هیدرولوژیکی مانند خشکسالی و سیل به منابع طبیعی و انسانی وارد می‌گردد، سازماندهی و حفاظت از این منابع، نیازمند برنامه‌ریزی و تعیین موقعیت زمانی و مکانی وقوع این پدیده‌هاست، که پیش‌بینی بارندگی به عنوان یکی از مهم‌ترین عوامل مؤثر در برآورد این پدیده‌ها مطرح می‌گردد. این مطالعه با هدف پیش‌بینی بارش ماهانه از طریق ارائه‌ی یک مدل داده‌کاوی کارا و همچنین تعیین پارامترهای مؤثر در افزایش دقت در پیش‌بینی این پارامتر هواشناسی انجام گردیده است. به طور کلی نتایج تحقیق حاضر بیانگر این مطلب است که مدل درختی M5 به عنوان یک مدل داده‌کاوی می‌تواند بارش ماهانه را با دقت بیشتری نسبت به الگوریتم KNN برآورد نماید که این مهم در مواقع استفاده از میانگین متحرک داده‌ها با دقت بیشتری نسبت به داده‌های خام قابل دستیابی است. از طرفی با توجه به ارزیابی سناریوهای مختلف توصیه می‌گردد که از ترکیب متغیرهای ورودی اختلاف میانگین حداکثر و حداقل دما، متوسط رطوبت نسبی، میانگین سرعت باد و درجه روز سرمایش (بر پایه ۲۱ درجه سانتی‌گراد) برای پیش‌بینی بارش ماهانه، جهت تعیین زمان و مکان تقریبی وقوع پدیده‌هایی چون خشکسالی و سیلاب بر مبنای وقوع بارش، استفاده گردد.

منابع

امیدوار، ک.، شفیعی، ش.، تقی‌زاده، ز. و علی‌پور، م. ۱۳۹۳. ارزیابی کارایی مدل درخت تصمیم در پیش بینی بارش ایستگاه سینوپتیک کرمانشاه. نشریه تحقیقات کاربردی علوم جغرافیایی. سال چهاردهم، شماره ۳۴، صفحات ۸۹-۱۱۰.

خلیلی، ن.، خداشناس، س.ر.، داوری، ک. و موسوی‌بایگی، م. ۱۳۸۹. پیش‌بینی بارش روزانه با استفاده از شبکه‌های عصبی مصنوعی مطالعه موردی: ایستگاه سینوپتیک مشهد. مجله پژوهش‌های آبخیزداری (پژوهش و سازندگی). شماره ۸۹، صفحات ۱۵-۷.

دستورانی، م.ت.، حبیبی‌پور، ا.، اختصاصی، م.ر.، طالبی، ع. و محجوبی، ج. ۱۳۹۱. بررسی کارایی مدل درخت تصمیم در پیش بینی بارش (مطالعه موردی ایستگاه سینوپتیک یزد). مجله تحقیقات منابع آب

ایران. سال هشتم، شماره ۳، ۱۴-۲۷.

ستاری، م.ت. و نهرین، ف. ۱۳۹۲. پیش‌بینی مقادیر حداکثر بارش روزانه با استفاده از سیستم‌های هوشمند و مقایسه آن با مدل درختی M5؛ مطالعه موردی ایستگاه‌های اهر و جلفا. فصلنامه‌ی علمی-پژوهشی مهندسی آبیاری و آب. سال چهارم، شماره چهاردهم، صفحات ۸۳-۹۸.

سیدکابلی، ح.، آخوندعلی، ع.م.، مساح‌بوانی، ح. و رادمنش، ف. ۱۳۹۱. ارائه مدل ریزمقیاس نمایی داده‌های اقلیمی براساس روش ناپارامتریک نزدیکترین همسایگی (K-NN). نشریه آب و خاک (علوم و صنایع کشاورزی). جلد ۲۶، شماره ۴، صفحات ۸۰۸-۷۷۹.

طالبی، ع. و اکبری، ز. ۱۳۹۲. بررسی کارایی مدل درختان تصمیم-گیری در برآورد رسوبات معلق رودخانه‌ای (مطالعه موردی: حوضه سد ایلام). مجله علوم و فنون کشاورزی و منابع طبیعی، علوم آب و خاک. سال هفدهم، شماره ۶۳، صفحات ۱۲۱-۱۰۹.

فدایی‌کرمانی، ا.، خانجانی، م.ج. و بارانی، غ.ع. ۱۳۹۳. کاربرد الگوریتم K-نزدیک‌ترین همسایگی در پایش خشکسالی بر مبنای شاخص بارش استاندارد (SPI) (مطالعه موردی: شهرستان بم). فصلنامه بین المللی پژوهشی تحلیلی منابع آب و توسعه. شماره ۱، صفحات ۱۳۸-۱۳۱.

قربانی، خ. ۱۳۹۳. ارزیابی مدل‌های داده‌کاوی در ریزمقیاس نمایی بارش براساس داده‌های مدل گردش عمومی NCEP (مطالعه موردی: ایستگاه سینوپتیک کرمانشاه). مجله پژوهش آب ایران. سال هشتم، شماره ۱۵، صفحات ۱۷۷-۱۸۶.

مهدوی،م . ۱۳۷۴. هیدرولوژی کاربردی، چاپ دوم، انتشارات دانشگاه تهران، تهران.

نعیمی، م. و احقاقی، ا. ۱۳۸۱. بررسی و مدیریت خشکسالی در ایران. مرکز اطلاعات و مدارک علمی ایران. شماره ۴۸۴۸۲۴.

Alberg, D., M. Last and A. Kindle. 2012. Knowledge discovery in data streams with regression tree methods. WIREs Data Mining Knowl Discov (2): 69-78.

Chattopadhyay, S. 2007. Feed forward artificial neural network model to predict the average summer-monsoon rainfall in India, Acta Geophysical, No. 55(3), pp. 369-382.

Chuan, C.S. 1997. Weather prediction using artificial neural network, Journal of Hydrology, 230: 101-119.

Hung, NQ., Babel, MS., Weesakul, S., Tripathi, NK. 2008. An artificial neural network model for rainfall forecasting in Bankok, Thailand, Hydrology and Earth System Sciences Discussions, No. 5, pp. 183–218.

Jagtap SS. Lall U. Jones, JW. Gijsman AJ. Ritchie JT. 2004. Dynamic nearest-neighbor method for estimating soil water parameters.Trans. ASAE. 47:1437–1444.

Lall, U., and Sharma, A. 1996. A nearest neighbor bootstrap for resampling hydrologic time series, Water Resources Research, 32(3), 679-694.

Maria, C., Haroldo , F., Ferreira, N. 2005. Artificial neural network technique for rainfall forecasting

applied to the Sao Paulo region, Journal of Hydrology, No. 301, pp.146-162.

Nash, J. E. and Sutcliffe, J. V. 1970. River flow forecasting through conceptual models, Part I - A discussion of principles, J. Hydrol., 10, 282–290.

Trafalis, TB., White, A., Santosa, B., Richman, MB. 2002. Data mining techniques for improved WSR-88D rainfall estimation, Computers in Industrial Engineering, No. 43, pp. 775–786.

Witten, L. and Frank, E. 2000. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers.

Xindung, W. & Kumar, V. 2009. Top Ten Algorithm in Data Mining, First Edition, Taylor & Francis Group , USA.

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

Iranian Journal of Irrigation and Drainage
No. 5, Vol. 13, Dec.-Jan. 2019, p. 1283-1293

نشریه آبیاری و زهکشی ایران
شماره۵، جلد ۱۳، آذر - دی ۱۳۹۸، ص. ۱۲۹۳-۱۲۸۳

Comparison of Decision Tree M5 and K-Nearest Neighborhood Algorithm Models in the Prediction of Monthly Precipitation (Case Study: Birjand Synoptic Station)

F. Poursalehi[1*], A. Shahidi[2], A. Khasheisiuki[3]
Recived: Mar.08, 2019 Accepted: Jun.24, 2019

Abstract

Due to the location of Iran in dry and semi-arid climate, heterogeneous distribution of precipitation and also the occurrence of a climate change phenomenon has caused phenomena such as floods, drought, desertification and dust production and also creating the different economic, social and environmental damages. One of the primary strategies to reduce these losses, is prediction of the precipitation events. The goal of the present study is monthly precipitation prediction with using data mining methods of decision tree (*M5*) and K-Nearest Neighbor (*KNN*) algorithms and Comparing these methods in order to determining more efficient method in the field of predicting the precipitation using monthly meteorological data of Birjand synoptic station during the statistical period 1961-2010 in three cases the raw data, the three-year moving average and the five-year moving average in the Weka software. The results showed that in all defined scenarios, the tree model M5 has more ability than the KNN model to predict the monthly precipitation of the station. Also after investigation of the evaluation criteria R, RMSE, MAE and NS, the fifteenth scenario with input variables such as mean difference of maximum and minimum temperature, average relative humidity, average wind speed and cooling degree days (base 21 ° C) in every month was determined as the best scenario for predicting the same month precipitation. Also the obtained results from comparing the defined scenarios in each model in three states raw data, three-year moving average and five-year moving average show that in most scenarios The five-year moving average on average, with the values of R=0.904, RMSE=6.054 and MAE=4.780 in the M5 model and on average, with the values of R=0.837, RMSE=7.698 and MAE=5.595 in the KNN model offers more accurate prediction of monthly Precipitation.

Keywords: Data mining methods, Decision tree, Drouth, K-Nearest Neighborhood, Weka software

1- Ph.D. Student ,Water Resources Engineering at the University of Birjand
2- Associate Professor, Department of Water Science and Engineering University of Birjand
(* - Corresponding Author Email: Fatemehpursalehi@birjand.ac.ir)