Research Paper

Print ISSN: 2251-7480
Online ISSN: 2251-7400

Journal of
Water and Soil Resources Conservation (WSRCJ)

Web site:
https://wsrcj.srbiau.ac.ir

Email:
iauwsrcj@srbiau.ac.ir
iauwsrcj@gmail.com

Vol. 13
No. 4 (52)

Received:
2023-10-12

Accepted:
2024-01-17

Pages: 113-126

Comparison of Data Mining Models Performance in Rainfall Prediction Using Classification Approach
(Case Study: Hamedan Airport Synoptic Weather Station)

Morteza Salehi Sarbijan[1*] and Hamidreza Dezfoulian[2]

1) Assistance Professor, Department of Mechanical Engineering, Faculty of Engineering, Zabol University, Zabol, Iran.
2) Assistance Professor, Department of Industrial Engineering, Faculty of Engineering, Bu-Ali Sina University, Hamedan, Iran.
*Corresponding author email: m.salehisarbijan@uoz.ac.ir

Abstract:

Background and Aim: Rainfall is one of the complex natural phenomena and one of the most crucial component of the water cycle, playing a significant role in assessing the climatic characteristics of each region. Understanding the amount and trends of rainfall changes is essential for effective management and more precise planning in agricultural, economic, and social sectors, as well as for studies related to runoff, droughts, groundwater status, and floods. Additionally, rainfall prediction in urban areas has a significant impact on traffic control, sewage flow, and construction activities.

Method: The objective of this study is to compare the accuracy of classification models, including Chi-squared Automatic Interaction Detector (CHAID), C5 decision tree, Naive Bayes (NB), Quest tree, and Random Forest, k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Artificial Neural Network (ANN) in predicting rainfall occurrence using 50 years of data from the synoptic station at Hamedan Airport. In this study, 80% of the data is used for training the models, and 20% for model validation and the results obtained from the model executions are compared using metrics such as confusion matrix, Receiver Operating Characteristic (ROC) curve, and the Area Under the Curve (AUC) index. To create the classification variable for rainfall and non-rainfall data, based on rainfall data, the days of the year are categorized into two classes: days with rainfall (y) and days without rainfall (n). Data preprocessing is performed using Automatic Data Preprocessing (ADP). Then, Principal Component Analysis (PCA) is employed to reduce the dimensions of the variables.

Results: In this study, the PCA method reduces the dimensions of the variables to 5. Also, approximately 80% of the available data corresponds to rainless days, while 20% corresponds to rainy days. The research results indicated that the KNN model with an accuracy of 91.9% for training data and the SVM model with 89.13% for test data exhibit the best performance among the data mining models. The AUC index for the KNN model is 0.967 for training data and 0.935 for test data, while for the SVM algorithm, it is 0.967 for training data and 0.935 for test data. According to the ROC curve for Hamedan rainfall data, the KNN model outperforms other models. Considering the sensitivity index in the confusion matrix, the KNN and SVM models perform better in predicting non-rainfall occurrence for training data. In terms of the precipitation occurrence prediction, the RT and KNN models show better results according to the specificity index.

Conclusion: The results demonstrated that for the RT, C5, ANN, SVM, BN, KNN, CHAID, QUEST, accuracy metrics was obtained 86.82%, 89.78%, 89.55%, 89.96%, 88.06%, 91.9%, 88.29%, 87.46%, 91.9%, respectively for training data. Moreover, for test data, the accuracy metrics for this model was obtained 83.82%, 87.9%, 88.12%, 89.13%, 87.12%, 89.13%, 87.12%, 88.19%, 86.93%, 86.76%, respectively. The AUC index in the training data for RT, C5, ANN, SVM, BN, KNN, CHAID QUEST models was 0.94%, 0.99%, 0.94%, 0.94%, 0.93%, 0.97%, 0.93%, 0.89%, respectively. In addition, for the test data, this metric was evaluated 0.89%, 0.89%, 0.93%, 0.94%, 0.92%, 0.90%, 0.92%, 0.88% respectively. As observed, considering accuracy metric and AUC index for training data KNN model and for test data SVM model were more sufficient in rainfall prediction.

Keywords: Rainfall prediction, Decision tree models, K-nearest neighbors (KNN) model, Artificial neural network, Support vector machine (SVM)

10.30495/wsrcj.2024.75668.11411

مقاله پژوهشی

مقایسه عملکرد مدل‌های داده‌کاوی در پیش‌بینی بارش باران با استفاده از رویکرد دسته‌بندی

(مطالعه موردی: ایستگاه هواشناسی سینوپتیک فرودگاه همدان)

مرتضی صالحی سربیژن [1*] و حمید رضا دزفولیان [2]

۱) استادیار، گروه مهندسی مکانیک، دانشکده فنی و مهندسی، دانشگاه زابل، زابل، ایران.
۲) استادیار، گروه مهندسی صنایع، دانشکده فنی و مهندسی، دانشگاه بوعلی سینا، همدان، ایران.
* ایمیل نویسنده مسئول: m.salehisarbijan@uoz.ac.ir

شاپا چاپی: ۷۴۸۰-۲۲۵۱
شاپا الکترونیکی: ۷۴۰۰-۲۲۵۰

نشریه حفاظت منابع آب و خاک

آدرس تارنما:
https://wsrcj.srbiau.ac.ir

پست الکترونیک:
iauwsrcj@srbiau.ac.ir
iauwsrcj@gmail.com

سال سیزدهم
شماره ٤ (٥٢)

تاریخ دریافت:
۱۴۰۲/۰۷/۲۰

تاریخ پذیرش:
۱۴۰۲/۱۰/۲۷

صفحات: ۱۲۶-۱۱۳

چکیده:

زمینه و هدف: بارندگی یکی از پدیده‌های پیچیده طبیعی و از مهم‌ترین اجزای چرخه آب بوده و در سنجش خصوصیات اقلیمی هر منطقه نقش بسیار مهمی ایفا می‌کند. شناخت میزان و روند تغییرات بارش به‌عنوان یکی از عناصر مهم هواشناسی، از یک‌سو جهت داشتن مدیریت اثربخش و برنامه‌ریزی دقیق‌تر برای بخش‌های کشاورزی، اقتصادی و اجتماعی و از سوی دیگر برای مطالعاتی مانند رواناب‌ها، خشکسالی‌ها، وضعیت آب‌های زیرزمینی و سیلاب‌ها ضروری است. همچنین پیش‌بینی بارش در مناطق شهری تأثیر بسیار زیادی بر کنترل ترافیک، جریان فاضلاب‌ها و فعالیت‌های ساخت‌وساز دارد.

روش پژوهش: هدف این مطالعه مقایسه دقت مدل‌های کلاس‌بندی درخت تصمیم (چاید (CHAID)، درخت تصمیم C5، نیو بیزین (NB)، کوئست (Quest) و جنگل تصادفی)، k نزدیک‌ترین همسایگی (KNN)، ماشین بردار پشتیبان (SVM) و شبکه عصبی مصنوعی (ANN) جهت پیش‌بینی وقوع بارش باران با استفاده از داده‌های یک دوره ۵۰ ساله در ایستگاه سینوپتیک فرودگاه همدان است. در این مطالعه از ۸۰ درصد داده‌ها جهت آموزش و از ۲۰ درصد داده‌ها جهت صحت سنجی مدل‌ها استفاده‌شده و نتایج حاصل از اجرای مدل‌ها با استفاده از معیارهای ماتریس درهم‌ریختگی (اغتشاش)، منحنی ROC و شاخص AUC مقایسه شدند. برای ساخت متغیر کلاس‌بندی داده‌های بارش و عدم بارش، با توجه به داده‌های بارش، روزهای سال در دو کلاس روزهای وقوع بارش (y) و روزهای عدم وقوع بارش (n) دسته‌بندی شدند. در این تحقیق پیش‌پردازش داده‌ها با استفاده از پیش‌پردازش خودکار داده‌ها (ADP) انجام شده و آنگاه کاهش ابعاد متغی‌رها از روش PCA استفاده شد.

یافته‌ها: در این مطالعه با توجه به روش PCA ابعاد متغیرها به ۵ بعد کاهش یافت. همچنین از داده‌های موجود تقریباً ۸۰ درصد، روزها بدون بارش و ۲۰ درصد روزها با بارش هستند. نتایج تحقیق نشان داد که مدل KNN با معیار صحت ۹۱/۹ برای داده‌های آموزشی و مدل SVM، ۸۹/۱۳ درصد برای داده‌های آزمون بهترین عملکرد را بین مدل‌های داده‌کاوی داشتند. شاخص AUC مدل KNN برابر ۰/۹۷ در داده‌های آموزشی و در داده‌های آزمون مقدار ۰/۹۴ برای الگوریتم SVM به دست آمد. همچنین با توجه به منحنی عملکرد سیستم (ROC) برای داده‌های بارش همدان مدل KNN نسبت به سایر مدل‌ها عملکرد بهتری را دارا می‌باشد. توجه به شاخص حساسیت در ماتریس اغتشاش، مدل‌های KNN و SVM در پیش‌بینی عدم وقوع بارش برای داده‌های آموزش بهتر عمل کردند. با توجه به شاخص خاصیت در پیش‌بینی وقوع بارش مدل‌های RT و KNN نتایج بهتری داشتند.

نتایج: نتایج تحقیق نشان داد که در داده‌های آموزش مقدار معیار صحت برای مدل‌های RT، C5، ANN، SVM، BN، KNN، CHAID و QUEST به ترتیب ۸۶/۸۲، ۸۹/۷۸، ۸۹/۵۵، ۸۹/۹۶، ۸۸/۰۶، ۹۱/۹، ۸۸/۲۹ و ۸۷/۴۶ بدست آمده اند. همچنین این معیار در داده‌های آزمون برای این مدل‌ها به ترتیب ۸۳/۲، ۸۷/۹، ۸۸/۱۲، ۸۹/۱۳، ۸۷/۱۲، ۸۸/۱۹، ۸۶/۹۳ و ۸۶/۷۶ به دست آمد. مقدار شاخص AUC در داده‌های آموزش برای مدل‌های RT، C5، ANN، SVM، BN، KNN، CHAID و QUEST به ترتیب ۰/۹۴، ۰/۹۲، ۰/۹۴، ۰/۹۴، ۰/۹۳، ۰/۹۷، ۰/۹۳ و ۰/۸۹ به دست آمد. همچنین این معیار در داده‌های آزمون برای این مدل‌ها به ترتیب ۰/۸۹، ۰/۸۹، ۰/۹۳، ۰/۹۴، ۰/۹۲، ۰/۹۰، ۰/۹۲ و ۰/۸۸ برآورد شد. همان‌طور که مشاهده شد، با توجه به معیارهای صحت و شاخص AUC در داده‌های آموزش مدل KNN و با توجه به داده‌های آزمون مدل SVM کارا تر در پیش‌بینی بارش باران بودند.

کلیدواژه‌ها: پیش‌بینی بارش باران، مدل‌های درخت تصمیم، مدل K نزدیک‌ترین همسایگی، شبکه عصبی مصنوعی، ماشین بردار پشتیبان

10.30495/wsrcj.2024.75668.11411

مقدمه

بسیاری از پیش‌بینی‌های آب و هوایی مانند پیش‌بینی بارش، پیش‌بینی رعدوبرق، پیش‌بینی شرایط ابری، چالش‌های عمده‌ای برای تحقیقات اتمسفری است. پیش‌بینی‌های آب و هوایی برای پیش‌بینی شرایط مختلف جوی مانند باران، باد، گرما، فشار، رطوبت و غیره استفاده می‌شود. بارش باران و برف از اجزای اصلی منابع تأمین آب و از مهم‌ترین عناصر اقلیمی می‌باشند که در تعیین و پراکندگی دیگر عناصر اقلیمی نیز مؤثرند، پیش‌بینی آن‌ها می‌تواند در مدیریت تأمین آب کشاورزی و مدیریت منابع آب موجود در مخازن سدها مفید باشد. همچنین در مناطق شهری پیش‌بینی بارش در مناطق شهری تأثیر بسیار زیادی بر کنترل ترافیک، جریان فاضلاب‌ها و فعالیت‌های ساخت‌وساز دارد. در این میان کشور ایران به علت قرار گرفتن در کمربند خشک جغرافیای جهان، در زمره مناطق کم باران جهان محسوب می‌گردد علاوه بر آن نوسانات شدید بارش در مناطق مختلف کشور، وقوع خشک‌سالی‌های ضعیف تا شدید را به موضوعی گریزناپذیر تبدیل نموده است. بر اساس مطالعات مختلف، کمبود بارش نسبت به میانگین درازمدت یک منطقه به‌عنوان مؤلفه اصلی رخداد خشک‌سالی محسوب می‌گردد (Dastourani et al., 2013). فرایند بارش، فرایندی کاملاً غیرخطی و ازنظر زمان و مکان به شکل تصادفی است. عوامل مختلف اقلیمی مانند شدت بارش، مدت و دوره بارش، پراکنش شدت در مدت بارش و تبخیر در آن نقش دارند و تشریح آن در مدل‌های ساده و خطی به‌راحتی امکان‌پذیر نیست (Bahrami et al., 2017) اساساً پیش‌بینی بارش به سبب رفتار بسیار نامنظم و آشوب مندی که از خود نشان می‌دهد، دشوار، پیچیده و درعین‌حال حائز اهمیت است. با توجه به اهمیت این موضوع در سال‌های گذشته مدل‌های گوناگونی جهت ارزیابی، پیش‌بینی و تخمین بارندگی ارائه‌شده است، اما امروزه سامانه‌های هوشمند مبتنی بر کاوش داده‌ها در مدل‌سازی فرایندهای هیدرولوژیکی و مهندسی آب موردتوجه قرارگرفته‌اند. به‌عبارت‌دیگر، این سامانه‌ها به‌عنوان روش‌هایی معتبر و شناخته‌شده در مدل‌سازی فرآیندهای پیچیده غیرخطی قادرند با استفاده از داده‌های هواشناسی و هیدرولوژیکی، تخمین دقیق‌تری از شرایط منطقه ارائه دهند (Bhattacharya & Solomatine, 2005). اکثر مطالعات برای پیش‌بینی میزان بارش از مدل‌های دسته‌بندی شبکه عصبی، درخت تصمیم، نزدیک‌ترین همسایگی و رگرسیون استفاده کرده‌اند که در ادامه به برخی از آن‌ها اشاره می‌شود. بهرامی و همکاران (۱۳۹۶) از سه روش نرمال‌سازی شامل روش مینیمم-ماکزیمم، روش نرمال‌سازی رتبه‌ای و روش آماره نرمال استاندارد برای داده‌های میانگین بارش ماهانه، کمترین و بیشترین دما و رطوبت ایستگاه سینوپتیک شهرستان آباده در بازه‌ی زمانی ۱۳۵۵ تا ۱۳۹۲ استفاده شد. در این پژوهش از شبکه عصبی مصنوعی پرسپترون چندلایه[1] با تعداد لایه‌های پنهان و نورون‌های مختلف و الگوریتم یادگیری لونبرگ-مارکورات برای تخمین بارش استفاده شد. نتایج نشان داد که روش مینیمم- ماکزیمم با ساختار شبکه‌ی سه لایه و تعداد ۱۳ نورون در لایه‌ی پنهان با ضریب همبستگی برابر با ۰/۹۲ و میانگین مجذور خطا[2] (MSE) برابر با ۰/۱۲ در مقایسه با دیگر روش‌ها به‌عنوان بهترین روش انتخاب شد. مهتابی و همکاران (۱۳۹۷) به پیش‌بینی وقوع بارش روزانه با استفاده از داده‌های هواشناسی بین سال‌های ۲۰۰۰ تا ۲۰۰۹ شهر اصفهان پرداختند. آن‌ها برای پیش‌بینی از مدل‌های هوشمند ماشین بردار پشتیبان،K نزدیک‌ترین همسایگی، شبکه عصبی مصنوعی و درخت تصمیم استفاده نمودند. نتایج نشان داد که در هر چهار روش، دقت پیش‌بینی بهترین سناریوها با استفاده از داده‌های ۶ و ۷ روز قبل، کمتر از ۷۵ درصد بود، اما با استفاده از داده‌های روزهای ۱ تا ۵ روز قبل، بارش روزانه با دقت بیش از ۸۰ درصد پیش‌بینی شد. باگیروف[3] و همکاران (۲۰۱۷) ترکیب مدل رگرسیون خطی خوشه‌ای را برای پیش‌بینی ماهانه بارش طی سال‌های ۱۸۸۹ تا ۲۰۱۴ در ایالت ویکتوریای استرالیا ارائه دادند. نتایج نشان داد که عملکرد مدل رگرسیون خطی خوشه‌ای نسبت به مدل‌های شبکه عصبی، رگرسیون خطی چندگانه و ماشین‌های بردار پشتیبان بهتر بوده است. کرامر[4] و همکاران (۲۰۱۷) مدل زنجیره مارکف را با مدل‌های یادگیری ماشین شامل رگرسیون برداری پشتیبان، الگوریتم ژنتیک، قوانین M5، درخت تصمیم C5، شبکه‌های عصبی شعاعی پایه و K نزدیک‌ترین همسایگی برای پیش‌بینی بارش باران ماهیانه ۲۰ شهر از قاره اروپا و ۲۲ شهر از آمریکا مقایسه نمودند. نتایج نشان داد عملکرد مدل‌های رگرسیون برداری پشتیبان و الگوریتم ژنتیک در مقایسه با دیگر روش‌ها بهتر بوده است. سینگ[5] (۲۰۱۸) برای پیش‌بینی باران‌های موسمی هند از داده‌های ماهیانه بین سال‌های ۱۸۷۱ تا ۲۰۱۴ استفاده کرد. در این تحقیق از سه تکنیک مجموعه فازی، شبکه‌های عصبی و آنتروپی استفاده‌شده است. نتایج نشان داد که مدل پیشنهادی در برابر دیگر مدل‌ها نتایج مؤثر و کارآمدی داشت. آفتاب[6] و همکاران (۲۰۱۸) به پیش‌بینی بارندگی در شهر لاهور پاکستان با استفاده از پنج تکنیک داده‌کاوی شامل ماشین بردار پشتیبان، نیو بیزین،K نزدیک‌ترین همسایه، درخت تصمیم و پرسپترون چندلایه طی ۱۲ سال از داده‌های آب و هوایی از اول دسامبر ۲۰۰۵ تا ۳۱ نوامبر ۲۰۱۷ پرداخته‌اند. نتایج مدل‌های پیشنهادی نشان داد که تکنیک‌ها به‌خوبی توانسته‌اند کلاس بدون بارش را نسبت به کلاس بارش دسته‌بندی کنند. فقدان ویژگی آب و هوایی مناسب در مجموعه

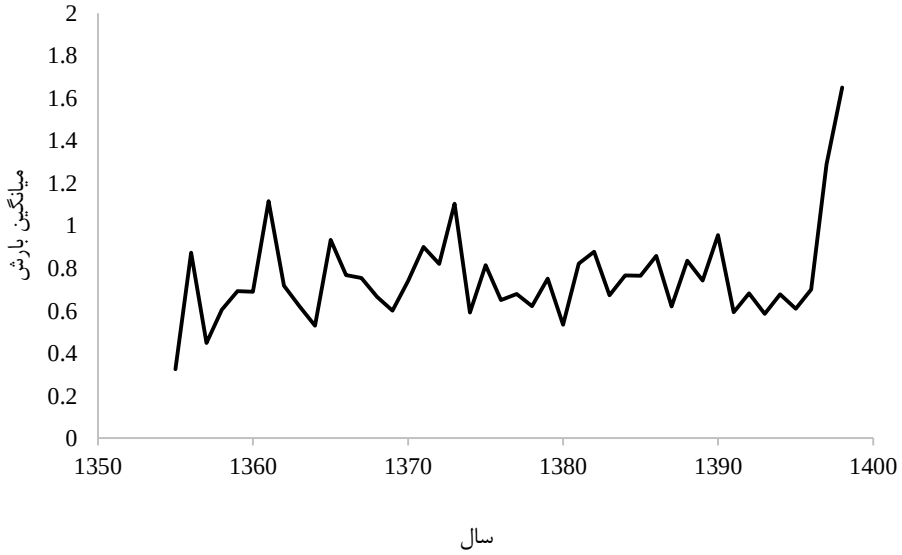
داده‌ها و نرخ پایین بارش از دلایلی بود که باعث شد دسته‌بندی کلاس با بارش با خطا همراه باشد. حیدر و ورما[۷] (۲۰۱۸) از مشخصه‌های بارندگی، میانگین، حداکثر و حداقل دما و شاخص نوسان جنوبی برای پیش‌بینی بارش شهر کوئینزلند استرالیا استفاده کردند. در این تحقیق الگوریتم ژنتیک مبتنی بر شناسایی بهترین ترکیب از ویژگی‌های ورودی و پارامترهای شبکه عصبی برای دستیابی به نتایج دقیق‌تر استفاده‌شده است. نتایج نشان داد که ترکیب الگوریتم ژنتیک و شبکه عصبی معیار RMSE کمتری نسبت به دیگر روش‌ها دارد. میشرا[۸] و همکاران (۲۰۱۸) تکنیک شبکه عصبی مصنوعی را برای پیش‌بینی بارش یک‌ماهه و دوماهه شمال هند طی سال‌های ۱۸۷۱ تا ۲۰۱۲ استفاده نمودند. در این مدل شبکه‌های عصبی پیش‌خور با استفاده از الگوریتم پس انتشار و الگوریتم لونبرگ-مارکورات استفاده‌شده است. از الگوریتم لونبرگ -مارکورات به‌عنوان یک روش استاندارد برای یافتن کمینه یک تابع غیر خطی چند متغیره استفاده می‌شود. نتایج نشان داد روش شبکه عصبی، پیش‌بینی یک‌ماهه را بهتر از الگوریتم لونبرگ پیش‌بینی کرده است. ملیکا و نیرمالا[۹] (۲۰۱۸) تکنیک‌های هموارسازی شامل میانگین متحرک (MA)، هموارسازی نمایی (ES)، مدل ARIMA و تکنیک داده‌کاوی k نزدیک‌ترین همسایگی برای پیش‌بینی بارندگی بین سال‌های ۱۹۰۱ تا ۲۰۱۳ در شهر چنای هند را مورداستفاده قراردادند. نتایج نشان داد که مدل ARIMA و مدل ترکیبی ES-KNN با استفاده از معیار آماری درصد میانگین مطلق خطا[۱۰] (MAPE) نتایج مطلوب‌تری ارائه دادند. داننده مهر و همکاران (۲۰۱۹) یک روش جدید برای توسعه‌ی مدل رگرسیون ترکیبی برای پیش‌بینی بارش یک‌ماهه در دو ایستگاه باران‌سنجی ایستگاه تبریز و ارومیه طی سال‌های ۱۹۹۰ تا ۲۰۱۴ ارائه دادند. این رویکرد مبتنی بر یکپارچه‌سازی رگرسیون بردار پشتیبان[۱۱](SVR) و الگوریتم کرم شب (FFA) است که منجر به پیش‌بینی‌های دقیق بارش می‌شود. برای هر دو مکان اندازه‌گیری نتایج نشان داد که مدل هیبرید به‌طور قابل‌توجهی عملکرد بهتری دارد. با توجه به میانگین بازده، بهبود ایجادشده در پیش‌بینی‌هایSVR با کاهش ۳۰ درصدی میانگین مربعات خطا (RMSE) و حدود ۱۰۰ درصدی افزایش کارایی NSE همراه بود. آدریانی و همکاران (۲۰۲۲) رویکرد یادگیری ماشین را در ایستگاه نیاوران تهران برای پیش‌بینی بارش باران استفاده کردند. نتایج نشان داد که عملکرد رویکردهای ترکیب بهینه‌سازی ازدحام ذرات (PSO) با ماشین بردار پشتیبان[۱۲] (PSO-SVR) و حافظه‌ی کوتاه‌مدت طولانی[۱۳] (LSTM) تقریباً یکسان و از روش شبکه عصبی کانولوشن[۱۴] (CNN) بهتر بوده‌اند. فهد و همکاران (۲۰۲۳) رویکرد خوشه‌بندی سلسله مراتبی را برای پیش‌بینی بارش باران برای داده‌های اقلیمی ۳۰ ساله از سال ۱۹۹۱ تا ۲۰۲۰ توسعه دادند. نتایج نشان داد که رویکرد پیشنهادی نسبت به شبکه‌های عصبی بازگشتی و شبکه عصبی متراکم از دقت بهتری برخوردار است. مارکنا[۱۵] و همکاران (۲۰۲۳) چهار تکنیک یادگیری ماشین، شامل رگرسیون خطی چندگانه، رگرسیون بردار پشتیبان، رگرسیون تطبیقی چند متغیر و جنگل تصادفی را برای پیش‌بینی روزانه و میانگین بارندگی هفتگی در ایستگاه رانی چائوری واقع در اوتاراکند هند را توسعه دادند. اعتبارسنجی مدل‌های توسعه‌یافته با استفاده از پارامترهای آماری ریشه میانگین مربعات خطا، شاخص تجمعی و ضریب همبستگی انجام شد. نتایج نشان داد که تکنیک جنگل تصادفی در مقایسه با دیگر روش‌ها در پیش‌بینی بارش بهتر عمل کرده است. هی[۱۶] و همکاران (۲۰۲۳) روش تقویت گرادیان بهبودیافته[۱۷] را برای پیش‌بینی کوتاه‌مدت بارش باران برای داده‌های آب‌وهوای استرالیا توسعه دادند. نتایج نشان داد که رویکرد تقویت گرادیان بهبودیافته نسبت به روش‌های رگرسیون لجستیک، جنگل تصادفی، آدابوست، شبکه عصبی مصنوعی و رگرسیون خطی بهتر عمل کرده است. همان‌طور که در ادبیات تحقیق مشاهده گردید مطالعه‌ای که از مدل‌های درخت تصمیم Quest، درخت C5 و درخت تصادفی برای پیش‌بینی وقوع بارش و عدم وقوع بارش استفاده کرده باشند، یافت نشد؛ بنابراین در این مطالعه برای اولین بار عملکرد و کارایی مدل‌های درخت تصمیم CHAID، درخت C5، نیو بیزین (NB)، Quest و درخت تصادفی (RT) با مدل‌های ماشین بردار پشتیبان، k نزدیک‌ترین همسایگی و شبکه عصبی مصنوعی در پیش‌بینی میزان بارش در ایستگاه سینوپتیک فرودگاه همدان باهم مقایسه شدند. همچنین استفاده از ۱۵ متغیر برای داده‌های روزانه‌ی هواشناسی و بکار بردن روش تجزیه مؤلفه‌های اصلی برای کاهش ابعاد از دیگر سهم‌های علمی مطالعه‌ی حاضر می‌باشند.

مواد و روش‌ها

ایستگاه سینوپتیک فرودگاه همدان در عرض شمالی ۳۴/۸۶ درجه و طول شرقی ۸۴/۵۳ واقع‌شده است. متوسط بارندگی سالیانه این ایستگاه ۴ میلی‌متر است. همچنین حداقل و حداکثر درجه حرارت در بلندمدت به ترتیب ۸/۳۲- و ۴۰ درجه سلسیوس گزارش‌شده است. در این مطالعه داده‌های هواشناسی از سامانه سازمان هواشناسی کشور، در بازه‌ی زمانی ۴۳ ساله از یازدهم خرداد سال ۱۳۵۵ تا بیست و نهم فروردین ۱۳۹۸ استخراج‌شده است. در جدول (۱) خصوصیات آماری مشخصه‌های استفاده‌شده در تحقیق خلاصه‌شده است. همان‌طور که در این جدول مشاهده می‌شود متوسط میزان بارندگی روزانه در بازه زمانی موردنظر برابر با ۰/۷۵ میلی‌متر بوده است. همچنین شکل (۱) نمودار متوسط بارش سالیانه برحسب میلی‌متر در سال‌های ۱۳۵۵ تا ۱۳۹۸ را نشان می‌دهد.

جدول ۱.خصوصیات آماری مشخصه‌های استفاده‌شده تحقیق

پارامتر	تعداد رکورد خالی	درصد داده‌های موجود	مینیمم	ماکزیمم	میانگین	انحراف معیار
میزان بارش روزانه	۱۲۰	۹۹/۲۳۳	۰	۱۳۳	۰/۷۵	۲/۸۶
میانگین ابرناکی روزانه	۱۹۷	۹۸/۷۴	۰	۸	۲/۵	۲/۳۴
ماکزیمم ابرناکی روزانه	۱۹۵	۹۸/۷۵	۰	۹	۴/۲۸	۳/۰۲
میانگین رطوبت نسبی	۱۲۸	۹۹/۱۸	۱۳/۵	۱۰۰	۵۱/۵۳	۱۹/۹۱
مینیمم رطوبت نسبی	۱۲۸	۹۹/۱۸	۱	۱۰۰	۳۲/۹۵	۱۹/۵۸
ماکزیمم رطوبت نسبی	۱۲۸	۹۹/۱۸	۱۸	۱۰۰	۷۳/۵۳	۱۸/۷۱
دمای ماکزیمم	۵۰۰	۹۶/۸	-۱۴/۳	۴۰	۱۹/۷۹	۱۱/۵۳
دمای میانگین	۶۱۴	۹۶/۰۷	-۱۸/۶	۳۲/۴	۱۲/۰۳	۱۰/۱۴
میانگین دمای تر	۳۹۱	۹۷/۵	-۱۹/۱۱	۱۹/۸	۶/۳۲	۶/۷۸
میانگین سرعت باد	۱۳۶	۹۹/۱۳	۰	۱۱/۱۲۵	۱/۷۷	۱/۵۱
ماکزیمم سرعت باد	۱۳۶	۹۹/۱۳	۰	۲۹	۵/۷	۳/۴۴
دمای مینیمم	۹۶۲	۹۳/۸۵	-۳۱/۱	۲۲/۲	۳/۹	۸/۱۸
میانگین دمای نقطه شبنم	۴۴۹	۹۷/۱۳	-۲۶/۷۳	۱۵/۳۳	۰/۳۳	۵/۷۲
میانگین فشار بخار	۱	۹۹/۹۹	۰/۶۳۷	۲۰/۶۴	۶/۸۲	۲/۷۱
فشار میانگین سطح ایستگاه	۲۰۸	۹۸/۶۷	۸۱۱/۱۲۵	۸۳۶/۵	۸۲۵/۴۶	۳/۵۶
فشار میانگین سطح دریا	۲	۹۹/۹۸	۹۸۹/۳۵	۱۰۵۳/۲	۱۰۱۲/۵۹	۹/۸۲



شکل۱. نمودار میانگین بارش روزانه ایستگاه سینوپتیک فرودگاه همدان به تفکیک سال

پیش‌پردازش داده‌ها

به دلیل خطاهای انسانی و ابزاری ممکن است مواردی مانند ثبت نشدن آمار، ثبت آمار غلط، خرابی یا از بین رفتن دستگاه‌های اندازه‌گیری یا تشخیص داده‌های پرت و حذف آن‌ها با عنوان داده‌های ناموجود پیش آید، بنابراین تخمین و برآورد این داده‌ها برای استفاده در مدل‌ها ضروری است. باید قبل از استفاده این داده‌ها در مطالعات نقایص آن‌ها برطرف و داده‌های ناموجود را بازسازی نمود. به همه‌ی مراحلی که قبل از مدل‌سازی و به‌منظور آماده‌سازی داده‌ها برای کم کردن خطا صورت می‌گیرد پیش‌پردازش گفته می‌شود. ازجمله این روش‌ها می‌توان به حذف داده‌های ناموجود، حذف داده‌های پرت و نرمال‌سازی داده‌ها اشاره نمود. ازآنجایی‌که از دست دادن داده باعث کاهش دقت مدل‌ها می‌شود، در این پژوهش جهت پر کردن داده‌های خالی از روش پیش‌پردازش خودکار داده‌ها[18] (ADP) استفاده‌شد. اساس کار روش ADP بر این اساس است که هرجایی که متغیر عددی باشد خانه خالی با مقدار میانگین متغیر تکمیل می‌شود و هر جا متغیر اسمی باشد به‌جای خانه خالی مقدار مد جایگزین می‌گردد. در این مطالعه به دلیل متفاوت بودن مقیاس داده‌ها و به دست آمدن نتایج دقیق، از

روش نرمال‌سازی استاندارد (Z-Score) استفاده شده است. همچنین در این مطالعه برای کاهش ابعاد از روش تجزیه مؤلفه‌های اصلی[۱۹] (PCA) استفاده شد. تجزیه مؤلفه‌های اصلی معمولاً برای کاهش ابعاد درداده‌های پیوسته استفاده می‌شود. درروش مذکور، محورهای مختصات جدیدی برای داده‌ها تعریف و داده‌ها بر اساس این محورها بیان می‌شوند. اولین محور بایستی در جهتی قرار گیرد که واریانس داده‌ها بیشینه شود. با استفاده از PCA می‌توان تعداد زیادی متغیر مستقل همبسته را با تعداد محدودی متغیر مستقل جدید که مؤلفه‌های اصلی نامیده می‌شوند و نا همبسته‌اند، جایگزین کرد.

روش‌های یادگیری ماشین با رویکرد دسته‌بندی

از روش‌های دسته‌بندی برای تخصیص یک برچسب به مجموعه‌ای از داده‌های اسمی که هنوز دسته‌بندی نشده‌اند استفاده می‌شود. در این مطالعه متغیر وابسته وقوع بارش از نوع اسمی است و برای کارایی و دقت پیش‌بینی بارش باران از مدل‌های دسته‌بندی استفاده شد. در مدل‌های دسته‌بندی کل مجموعه داده‌ها به دو قسمت مجموعه داده‌های آموزش و مجموعه داده‌های آزمون تقسیم‌بندی شدند. روش‌هایی که این مدل‌ها برای پیش‌بینی استفاده می‌کنند با یکدیگر تفاوت دارند به‌عنوان‌مثال بعضی از این مدل‌ها مانند درخت تصمیم در قالب یکسری قوانین عمل می‌کنند و برخی دیگر فقط به پیش‌بینی می‌پردازند.

درخت تصمیم

درختان تصمیم‌گیری از نسل جدید تکنیک‌های داده‌کاوی به شمار می‌آیند که در دو دهه اخیر توسعه زیادی یافته‌اند. درخت تصمیم یکی از روش‌های داده‌کاوی و متداول برای دسته‌بندی و پیش‌بینی است که برخلاف شبکه‌های عصبی به تولید قانون می‌پردازد یعنی درخت تصمیم پیش‌بینی خود را در قالب یکسری قوانین توضیح می‌دهد. فرایند ایجاد یک درخت تصمیم شامل چهار مرحله است:

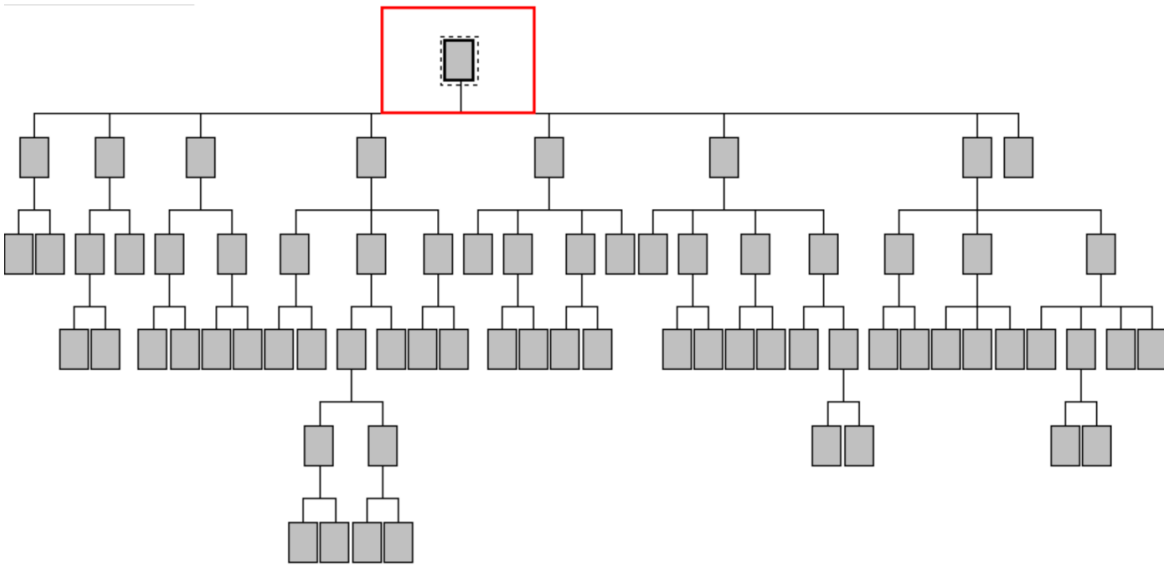
۱) **گره ریشه:** گره ریشه گره بالای درخت است که نقطه شروع فرآیند تصمیم‌گیری را نشان می‌دهد. این گره حاوی ویژگی است که آن را تبدیل به مهم‌ترین گره برای پیش‌بینی متغیر هدف می‌کند. در این تحقیق متغیر وقوع بارش و عدم وقوع بارش به‌عنوان متغیر هدف قرار داده‌شده است.

۲) **گره داخلی:** گره‌های داخلی گره‌هایی حاوی گره فرزند هستند. آن‌ها مراحل میانی در فرآیند تصمیم‌گیری را نشان می‌دهند. هر گره داخلی حاوی یک قانون تصمیم‌گیری است که داده‌ها را به دو یا چند شاخه تقسیم می‌کند.

۳) **شاخه‌ها**: شاخه‌ها خطوطی هستند که گره‌ها را به یکدیگر متصل می‌کنند. آن‌ها نتایج احتمالی یک تصمیم را نشان می‌دهند. هر شاخه به یک گره فرزند منتهی می‌شود.

۴) **گره برگ:** گره‌های برگ، گره‌هایی هستند که هیچ گره فرزندی ندارند. آن‌ها نشان‌دهنده نتیجه نهایی فرآیند تصمیم‌گیری هستند. هر گره برگ حاوی یک پیش‌بینی برای متغیر هدف است. شکل (۲) نمودار درخت تصمیم این تحقیق را برای متغیر هدف روزهای بارش و عدم بارش نشان می‌دهد.

در این مطالعه از مهم‌ترین مدل‌های درخت تصمیم، شامل مدل‌های CHAID، درخت C5، نیو بیزین (NB)، Quest و درخت تصادفی (RT) استفاده‌شده است.



شکل ۲. نمای کلی اجزای اصلی درخت تصمیم تحقیق

مدل CHAID روش آماری مؤثری برای دسته‌بندی است که با استفاده از مفهوم آزمون آماری تمام مقادیر صفت پیش‌بینی کننده بالقوه را ارزیابی می‌کند. در این مدل مقادیری را که به‌صورت آماری همگن در نظر گرفته‌شده‌اند را با توجه به متغیر هدف و حفظ تمام مقادیری که ناهمگن هستند، ادغام می‌کند. سپس این مدل بهترین پیش‌بینی کننده را از شاخه در درخت تصمیم انتخاب می‌کند به طوریکه هر فرزند گره از یک گروه از ارزش‌های همگون نسبت به صفت انتخاب‌شده می‌باشند. این فرایند به‌صورت بازگشتی تا آنجا که درخت به‌طور کامل رشد می‌کند، ادامه می‌یابد. اگر صفت هدف پیوسته باشد، آزمون آماری مورداستفاده F و اگر این صفت گسسته باشد آزمون مجذور مربع به‌کاربرده می‌شود(Kisi et al., 2016).

مدل درخت C5

مدل درخت C5 دسته‌بندی را با تقسیم داده‌ها به زیرمجموعه‌هایی که شامل رکوردهای همگن‌تر از والد خود می‌باشند، انجام می دهد. در مدل درخت تصمیم C5 تقسیم کردن نمونه‌ها بر اساس فیلدی که بیشترین بهره اطلاعات را شامل شود، صورت می‌گیرد. هر زیر نمونه به‌وسیله اولین انشعاب تعیین می‌گردد. سپس بر اساس فیلد جدید دوباره تقسیم‌بندی انجام می‌شود و این فرایند تا زمانی که زیر نمونه‌ها امکان تقسیم شدن را نداشته باشند تکرار می‌شود. درنتیجه انشعاب‌ها به پایین‌ترین سطح از نو آزموده می‌رسند و آن انشعاب‌هایی که ارزش چندانی ندارند از مدل حذف می‌گردند. در این مدل از دو رویکرد بهره اطلاعاتی[۲۰] و شاخص جینی[۲۱] را برای انتخاب مشخصه‌ها استفاده می‌شود. بهره اطلاعاتی به‌وسیله سنجش تغییرات آنتروپی پس از تقسیم‌بندی یک مجموعه داده بر اساس ویژگی‌ها انجام می‌شود. بهره اطلاعاتی بر روی مشخصه A از رابطه (۱) زیر به دست می‌آید.

$$\mathrm{GAIN(A)} = \mathrm{INFO(D)} - \mathrm{INFO_A(D)} \quad (۱)$$

در این رابطه INFO میزان آنتروپی را نشان می‌دهد که از رابطه(۲) زیر به دست می‌آید.

$$\mathrm{INFO(D)} = -\sum_{i=1}^{m} p_i \,\mathrm{LOG}_2\, p_i \quad (۲)$$

همچنین اطلاعات موردنیاز (بعد از استفاده از A برای تقسیم D به V پارتیشن) برای دسته‌بندی D از رابطه‌ی زیر به دست می‌آید. (Pang & Gong, 2009).

$$\mathrm{Info_A(D)} = \sum_{j=1}^{v} \frac{|\mathrm{D_j}|}{|\mathrm{D}|} \times \mathrm{Info}\left(\mathrm{D_j}\right) \quad (۳)$$

مدل نیو بیزین (NB)

قضیه بیزین یک روش از محاسبات احتمالی است و احتمال رخدادی که در آینده رخ خواهد داد وابسته به رخدادی است که قبلاً رخداده است. این تئوری توانایی خود یادگیری در سیستم هوشمند دارد که به‌صورت گسترده استفاده می‌شود. از تئوری بیزین می‌توان برای پیش‌بینی حوادث آینده بر اساس حوادث حاضر طبق نظریه‌ی آمار و احتمال استفاده کرد. دسته‌بندی نیو بیزین بر پایه قضیه Bayes و فرضیه‌های استقلال بین پیش‌بینی کننده‌ها است. یک مدل Bayesian برای ساختن، ساده و آسان است که بدون پارامتر پیچیده تکراری است که باعث می‌شود برای مجموعه‌های داده‌های بسیار بزرگ مفید باشد. به‌رغم سادگی، دسته‌بندی بیزین اغلب جالب است و به‌طور گسترده‌ای مورداستفاده قرار می‌گیرد(Anderson et al., 2023).

مدل Quest

مدل Quest تقریباً یک مدل رشد درخت باینری است. فرایند رشد درخت در مدل Quest شامل پیش‌بینی تقسیم، انتخاب یک نقطه انشعاب برای پیش‌بینی انتخاب‌شده و شرط توقف است. در این مدل فقط تک متغیر انشعابات در نظر گرفته می‌شود. ملاک تصمیم‌گیری برای انتخاب متغیرها با مقایسه‌ی P مقدار مربوط به آماره‌ی F در آزمون ANOVA برای متغیرهای کمی و P مقدار آماره‌ی خی دو مربوط به جدول تقاطعی برای متغیرهای کیفی انجام می‌شود. این الگوریتم با توجه به اینکه از p مقدار برای تصمیم‌گیری استفاده می‌نماید، موجب تشکیل درختی ناریب برای متغیرها می‌شود (Ritschard, 2010).

مدل جنگل تصادفی

مبنای روش جنگل تصادفی، یادگیری گروهی است، تکنیکی که بسیاری از طبقه‌بندی کننده‌ها را برای ارائه راه‌حل‌هایی برای حل مسائل پیچیده ترکیب می‌کند. مهم‌ترین ویژگی مدل جنگل تصادفی عملکرد بالای آن در اندازه‌گیری اهمیت متغیرها برای مشخص کردن اینکه هر متغیر چه نقشی در پیش‌بینی پاسخ دارد، است. این مدل درواقع یک الحاقی از درخت رگرسیون و طبقه‌بندی است. مدل پیش‌بینی کننده جنگل تصادفی بر اساس میانگین‌گیری از نتایج حاصل از تمامی درخت‌های تصمیم مربوطه استوار است و برای بسیاری از مجموعه داده‌ها، طبقه‌بندی با صحت بالایی را انجام می‌دهد. درخت تصادفی بردار ورودی را گرفته، آن را با هر درخت در جنگل طبقه‌بندی کرده و خروجی، برچسب‌های کلاسی هستند که از اکثریت آرا دریافت شده است (Josso et al., 2023).

مدل ماشین بردار پشتیبان تصمیم (SVM)

مدل ماشین بردار پشتیبان، یک روش یادگیری با نظارت است که برای طبقه‌بندی و پیش‌بینی به کار می‌رود. هدف روش طبقه‌بندی بردار پشتیبان، یافتن ابر صفحه‌ای با بهترین توانایی

تعمیم برای تفکیک داده‌های مربوط به دوطبقه با حاشیه (فاصله بین ابر صفحه و نزدیک‌ترین داده آموزشی) حداکثر است. در این روش برای به دست آوردن نتایج دقیق، انتخاب تابع کرنل بهینه که خصوصیات غیرخطی را به‌صورت خطی درمی‌آورد، حائز اهمیت است. به‌طورکلی تئوری یادگیری با استفاده از داده‌های آموزشی انجام می‌پذیرد در این صورت‌مسئله یادگیری شامل انتخاب تابعی از بین توابع $f(x,\alpha), \alpha \in R$ خواهد شد که بتواند پاسخ ناظر (هر بردار ورودی X و بر اساس تابع توزیع شرطی $P(y|x)$ که البته ثابت ولی نامعلوم است، مقدار y را بازمی‌گرداند) را به بهترین وجه پیش‌بینی نماید. این گزینش بر اساس مجموعه‌ای از داده‌های آموزشی $(x_1, y_1), \dots, (x_m, y_m)$ است که بر اساس احتمال شرطی $P(x,y) = P(x)P(y|x)$ انتخاب می‌شوند (Shin et al., 2005).

مدل k نزدیک‌ترین همسایگی (KNN)

مدل k نزدیک‌ترین همسایگی، یک مدل طبقه‌بندی بر اساس سنجش فاصله است. در این روش، ابتدا داده‌های آموزشی و طبقه‌های متناظر با آن‌ها در نظر گرفته می‌شود. سپس برای طبقه‌بندی یک نمونه جدید، فاصله آن با هریک از نمونه‌های آموزشی محاسبه‌شده و K همسایه از میان نزدیک‌ترین همسایه‌ها (نمونه‌های آموزشی) انتخاب می‌گردند. درنهایت، نمونه جدید در طبقه‌ای قرار می‌گیرد که بیشترین نمونه‌های موجود در K همسایه‌اش در آن طبقه هستند. برای محاسبه فاصله هر نمونه جدید از نمونه‌های مشاهداتی، از توابع فاصله مانند همینگ، اقلیدسی و چبیشف استفاده می‌شود که در این مطالعه، تابع اقلیدسی مورداستفاده قرار گرفت. در این مدل برای دستیابی به بهترین نتیجه، تعیین مقدار بهینه پارامتر K اهمیت زیادی دارد که برای این منظور، از روش سعی و خطا استفاده شد (N & K.Thanushkodi, 2010).

شبکه عصبی مصنوعی (ANN)

شبکه‌های عصبی مصنوعی، با تجزیه‌وتحلیل داده‌های ورودی و نتایج نظیر آن‌ها و بدون در نظر گرفتن فرآیند فیزیکی حاکم بر سیستم، قادر به ایجاد رابطه بین متغیر وابسته و مستقل هستند. به‌طورکلی، یک شبکه عصبی شامل سه لایه ورودی، پنهان و خروجی است. تعداد نورون‌های موجود در لایه‌های پنهان و تعداد این لایه‌ها با توجه به نوع مسئله و نیز با استفاده از روش سعی و خطا مشخص می‌شود. به‌طورکلی، شبکه عصبی یک ابزار پیش‌بینی برای ساخت یک مدل ریاضی از یک سیستم ناشناخته است. یکی از مهم‌ترین مدل‌های شبکه عصبی، مدل شبکه‌های عصبی پرسپترون چندلایه[۲۲] است(Alavi et al., 2010). در این تحقیق جهت اجرای مدل شبکه عصبی مصنوعی از شبکه پرسپترون چندلایه پیش‌خور با قانون یادگیری پس انتشار خطا و برای تابع آموزش از الگوریتم لونبرگ-مارکوات و برای تابع‌های انتقال نیز برای لایه اول تابع تانژانت هیپربولیک سیگوئید و برای لایه دوم تابع خطی استفاده شد. اطلاعات شبکه طراحی‌شده، مربوط به هر سه لایه شبکه (ورودی، مخفی و خروجی) است. این شبکه در لایه‌ی ورودی با ۹ متغیر، یک‌لایه‌ی خروجی تابع هدف و همچنین تعداد مناسب نورون‌های لایه پنهان برای دستیابی به بهترین پیش‌بینی‌ها و کمترین میزان خطا در شبکه بر مبنای روش آزمون‌وخطا و برابر پنج نورون با تابع انتقال سیگموییدی تعیین شد. در این مطالعه ۸۰ درصد داده‌ها جهت آموزش و ۲۰ درصد داده‌ها جهت صحت سنجی مدل‌ها استفاده شد.

معیارهای ارزیابی روش‌های دسته‌بندی

در این بخش معیارهایی جهت ارزشیابی دسته‌ها معرفی می‌شود و چگونگی عملکرد آن‌ها در پیشگویی برچسب کلاس‌ها بررسی می‌گردد. این شاخص‌ها هم در داده‌های آموزشی و هم در داده‌های آزمون مورداستفاده قرار گرفتند.

ماتریس درهم‌ریختگی (اغتشاش)

این ماتریس چگونگی عملکرد مدل دسته‌بندی را با توجه به مجموعه داده ورودی به تفکیک انواع دسته‌های مسئله دسته‌بندی نشان می‌دهد. جدول (۲) برای یک مسئله دو کلاس نشان می‌دهد. در این ماتریس مفاهیم TP، FP،FN و TN به شرح ذیل است. TP: بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و مدل دسته‌بندی نیز دسته آن‌ها را به‌درستی مثبت تشخیص داده است. FN: بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و مدل دسته‌بندی آن‌ها را به‌اشتباه منفی تشخیص داده است. FP: بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و مدل دسته‌بندی نیز دسته آن‌ها را به‌اشتباه مثبت تشخیص داده است. TN: بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و مدل دسته‌بندی نیز دسته آن‌ها را به‌درستی منفی تشخیص داده است. با توجه به ماتریس اغتشاش معیارهای دقت طبقه‌بندی[۲۳] (CA) و معیار خطای دسته‌بندی[۲۴] (EC) تعریف شد. معیار صحت دسته‌بندی (CA) این معیار نشان‌دهنده این است که طبقه طراحی‌شده چند درصد از کل مجموعه رکوردهای آزمایشی را به‌درستی دسته‌بندی کرده است و از رابطه (۴) به دست می‌آید. معیار EC در رابطه (۵) آمده است که دقیقاً برعکس معیار صحت دسته‌بندی است. کم‌ترین مقدار آن برابر صفر (بهترین کارایی) و بیشترین مقدار آن برابر یک (ضعیف‌ترین کارایی) است.

$$CA = \frac{TP + TN}{TP + TN + FN + FP} \quad (۴)$$

$$EC = \frac{FP + FN}{TP + TN + FN + FP} \quad (۵)$$

جدول ۲. ماتریس اغتشاش دو کلاسه جهت دسته‌بندی

		رکوردهای تخمینی	
		دسته مثبت	**دسته منفی**
رکوردهای واقعی	دسته مثبت	TP	**FN**
	دسته منفی	FP	**TN**

منحنی مشخصه عملکرد سیستم[۲۵] (ROC)

منحنی ROC ابزاری برای به تصویر کشیدن، سازمان‌دهی کردن و انتخاب دسته‌بندی کننده‌ها بر اساس کارایی آن‌هاست. منحنی‌های ROC منحنی‌های دوبعدی هستند که در آن‌ها نرخ تشخیص صحیح دسته مثبت (TPR) روی محور Y و نرخ تشخیص غلط دسته منفی (FPR) روی محور X رسم می‌شود. در نقاط بالای خط نیمساز منحنی ROC نقاطی قرارگرفته‌اند که مقدار حساسیت آن‌ها نسبت به FPR بیشتر است. به معنای آن‌که در این بخش، نرخ مثبت صحیح بیشتر از نرخ مثبت کاذب است. قرار گرفتن نقاط در این محیط مطلوب خواهد بود؛ بنابراین نقاطی که بالاترین TPR و پایین‌ترینFPR را دارند نقاط مطلوب این منحنی هستند. در این تحقیق مقادیر TPR و FPR از رابطه‌های (۶) و (۷) به دست آمدند.

$$TPR = \frac{TP}{TP + FN} \quad (۶)$$

$$FPR = \frac{FP}{FP + TN} \quad (۷)$$

شاخص AUC

معیار مهم دیگر که برای تعیین میزان کارایی یک طبقه بسیار مؤثر است معیار AUC است. این معیار نشان‌دهنده سطح زیر نمودار ROC است. هرچقدر عدد AUC مربوط به یک طبقه بزرگ‌تر باشد کارایی نهایی طبقه مطلوب‌تر ارزیابی می‌شود.

نتایج و بحث

در این مطالعه بعد از ایجاد مدل‌های دسته‌بندی در نرم‌افزارSPSS Modeler 18.0 انواع مدل‌ها با توجه به معیارهای ارزیابی ماتریس درهم‌ریختگی، منحنی مشخصه عملکرد سیستم و شاخص AUC مقایسه شدند.

مقایسه‌ی مدل‌ها و نتایج

در این تحقیق ابتدا عملیات پیش‌پردازش داده‌ها با استفاده از پیش‌پردازش خودکار داده‌ها[۲۶] (ADP) انجام و سپس برای کاهش ابعاد از روش PCA استفاده شد. جدول (۳) ارتباط همبستگی بین شاخص‌ها با استفاده از روش PCA را نشان می‌دهد. همان‌طور که مشاهده می‌گردد ابعاد به ۵ بعد کاهش‌یافته‌اند. با توجه به جدول (۳) میانگین فشار بخار (ewm)، میانگین دمای نقطه شبنم (tdm)، دمای میانگین (tm)، دمای ماکزیمم (tmax)، دمای مینیمم (tmin) و میانگین دمای تر (twetm) در بعد اول قرار گرفتند. در بعد دوم میانگین سرعت باد (ffm) و ماکزیمم سرعت باد (ffmax)، بعد سوم شامل میانگین ابرناکی روزانه (nm) و ماکزیمم ابرناکی روزانه (nmax)، بعد چهارم فشار میانگین سطح ایستگاه (p0m) و فشار میانگین سطح دریا (pm) و همچنین میانگین رطوبت نسبی (um)، مینیمم رطوبت نسبی (umin) و ماکزیمم رطوبت نسبی (umax) در بعد پنجم قرار گرفتند. جدول (۴) میزان ارزش و تأثیر هر یک از ۵ بعد را نشان می‌دهد. با توجه به‌ردیف سوم این جدول، مشاهده می‌گردد مقدار تجمعی درصد مقدار ویژه ابعاد یک تا پنج حدود ۹۴ درصد از تغییرات را پوشش داده‌اند؛ بنابراین در این تحقیق به‌جای استفاده از ۱۵ شاخص اولیه از ۵ شاخص به‌دست‌آمده از روش کاهش ابعاد استفاده شد. جدول ۵ نتایج محاسباتی معیارهای صحت (CA) و درصد خطا (ER) را برای روش‌های داده‌کاوی RT، C5، ANN، SVM، BN، KNN، CHAID و QUEST به تفکیک داده‌های آموزش و آزمون نشان می‌دهد. همان‌طور که از جدول (۵) مشاهده می‌شود روش k نزدیک‌ترین همسایگی با معیار صحت ۹۱/۹ برای داده‌های آموزشی و الگوریتم SVM مقدار ۸۹/۱۳ درصد برای داده‌های آزمون بهترین عملکرد را بین مدل‌های داده‌کاوی داشتند. با مقایسه‌ی AUC در جدول (۵) مشاهده می‌شود این مقدار برای مدل k نزدیک‌ترین همسایگی برابر ۰/۹۶۷ درداده نای آموزش و درداده‌های آزمون مقدار ۰/۹۳۵ برای الگوریتم SVM به دست آمد. منحنی عملکرد سیستم (ROC) برای مقایسه مدل‌ها در شکل (۳) آورده شده است. مطابق نمودار منحنی مشخصه عملکرد هر چه منحنی از خط قطری فاصله بیشتری داشته باشد نشان از دقت بیشتر مدل است. نتایج در شکل (۳) نشان می‌دهد که بهترین مدل ازلحاظ پیش‌بینی برای داده‌های آموزشی مدل KNN است. برای داده‌های آزمون منحنی ROC مدل‌ها تقریباً یکسان بوده است. مساحت زیر نمودار ROC که همان شاخص AUC است برای مدل KNN نسبت به دیگر مدل‌ها بهتر عمل کرده است. با توجه به شاخص‌های اندازه‌گیری شده برای داده‌های بارش همدان مدل KNN بهترین مدل بوده است.

جدول (۶) ماتریس اغتشاش (درهم‌ریختگی) تعداد روزهای وقوع و عدم وقوع بارش بین داده‌های واقعی و پیش‌بینی‌شده آموزش را برای مدل‌های مورد مقایسه نشان می‌دهد. ازآنجایی‌که توزیع کلاس‌های این مطالعه متعادل نیستند و

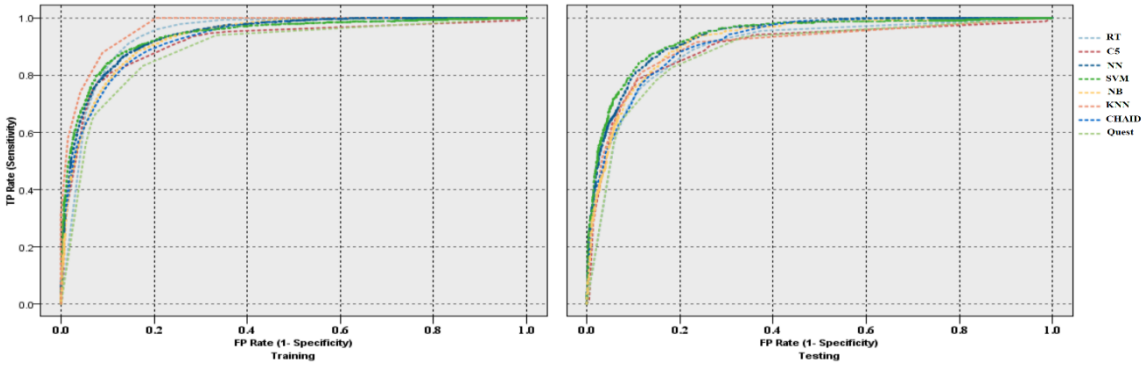
تعداد داده‌های برچسب کلاس‌های بدون بارش و با بارش باهم اختلاف زیادی دارند از دو سنجه میزان حساسیت و خاصیت نیز برای مقایسه‌ی مدل‌های داده‌کاوی استفاده‌شده است. شاخص حساسیت نرخ پاسخ‌های مثبت درست ($\frac{TP}{TP+FN}$) و شاخص خاصیت به‌عنوان نرخ پاسخ‌های منفی درست ($\frac{TN}{TN+FP}$) را نشان می‌دهند. همان‌طور که از جدول (۶) مشاهده می‌گردد با توجه به شاخص حساسیت مدل‌های KNN و SVM در پیش‌بینی عدم وقوع بارش برای داده‌های آموزش بهتر عمل کردند. همچنین با توجه به شاخص خاصیت در پیش‌بینی وقوع بارش مدل‌های RT و KNN نتایج بهتری داشتند. جدول (۷) ماتریس اغتشاش تعداد روزهای وقوع و عدم وقوع بارش بین داده‌های واقعی و پیش‌بینی‌شده آزمون را برای مدل‌های مورد مقایسه نشان می‌دهد. همان‌طور که از جدول (۷) مشاهده می‌گردد با توجه به داده‌های آزمون برای شاخص حساسیت کارایی مدل‌های SVM و CHAILD نسبت به دیگر مدل‌ها بهتر عمل کردند. همچنین با توجه به شاخص خاصیت برای پیش‌بینی وقوع بارش مدل‌های KNN و C5 کاراتر بودند. از آنجایکه مدل KNN یک مدل یادگیری ماشین نظارت‌شده است و از ویژگی‌های آن سادگی و آسانی این مدل برای پیاده‌سازی است. مدل KNN برای حل مسائل طبقه‌بندی و رگرسیون مورداستفاده قرار می‌گیرد. همچنین این مدل از دقت نسبتاً بالایی برخوردار است(Alcantara & Ahn, 2021). بنابراین نتایج حاصل از خروجی الگوریتم KNN برای پیش‌بینی بارش باران در ایستگاه سینوپتیک همدان قابل‌اطمینان هستند. نتایج این تحقیق با نتایج تحقیق لی و همکاران (۲۰۲۲) که پیش‌بینی بارش باران را در کشورهای ایتالیا و سوئیس با استفاده از مدل‌های یادگیری ماشین مطالعه کردند و نشان دادند که روش KNN کارایی بهتری داشته است در یک راستا است. همچنین نتایج این تحقیق با نتایج مطالعه‌ی آفتاب و همکاران (۲۰۱۸) که پیش‌بینی بارش باران را برای لاهور پاکستان با استفاده از تکنیک‌های داده‌کاوی بررسی کردند و نشان دادند که مدل‌های SVM و KNN توانسته‌اند به‌خوبی کلاس‌بندی بارش و عدم بارش را پیش‌بینی کنند، هم‌خوانی دارند. در مطالعه‌ی ملیکا و نیرمالا (۲۰۱۸) مدل ترکیبی ES-KNN با استفاده از معیار درصد میانگین مطلق خطا برای پیش‌بینی بارندگی در شهر چنای هند نتایج مطلوب‌تری به دست آمد. مشخص شد نتایج تحقیقات ذکرشده با این مطالعه هم سو و منطبق است. همچنین نتایج تحقیق حاضر با برخی از تحقیقات داخلی ازجمله مهتابی و همکاران (۱۳۹۷) که نشان دادند مدل‌های SVM و KNN به‌خوبی وقوع بارش روزانه در شهر اصفهان را پیش‌بینی کردند. همچنین مدرسی و عراقی نژاد (۲۰۱۴) و ستاری و همکاران (۲۰۱۶) که نشان دادند که مدل‌های SVM و KNN در داده‌های آزمون و آموزش کمترین خطا را برای کیفیت آب به دست آورده‌اند در یک راستا هستند.

نتیجه‌گیری

پیش‌بینی وضع هوا در محدوده وسیعی از زندگی روزمره دارای کاربرد است. برای نمونه‌هایی از این کاربرد می‌توان به پیش‌بینی وضع هوا در زندگی عادی، مسائل اقتصادی، ورزشی، پزشکی و نظایر آن اشاره نمود. بارش نیز به‌عنوان یکی از مهم‌ترین عناصر اقلیمی نقش مهمی را در زندگی تمام موجودات دارد. در کشور ایران همچنین به سبب وابستگی بخش عمده‌ای از فعالیت‌های اقتصادی به مقدار و توزیع زمانی بارش آگاهی از میزان و زمان وقوع بارش دارای اهمیت بسیار زیادی است؛ بنابراین در راستای اهمیت موضوع و باهدف ارزیابی و مقایسه‌ی مدل‌های موجود در داده‌کاوی از داده‌های روزانه ۴۳ سال از سال ۱۳۵۵ تا ۱۳۹۸ ایستگاه سینوپتیک فرودگاه همدان برای متغیر بارش یا عدم وقوع بارش باران با استفاده از ۱۵ متغیر تأثیرگذار بر میزان بارش مثل میانگین ابرناکی، میانگین رطوبت نسبی، دمای میانگین، میانگین دمای تر، میانگین سرعت باد، ماکزیمم سرعت باد و. پرداخته‌شده است. بعد از پیش‌پردازش و نرمال‌سازی داده‌ها به مقایسه مدل‌های مختلف داده‌کاوی پرداخته شد. نتایج مقایسه مدل‌های مختلف داده‌کاوی نشان داد که مدل‌های KNN و SVM با توجه به معیار صحت نسبت به دیگر مدل‌ها کارایی بهتری برای مشخص نمودن وقوع بارش و عدم وقوع بارش باران داشته‌اند. نتایج نشان داد که با توجه به شاخص AUC مدل KNN نسبت به دیگر مدل‌ها بهتر عمل کرده است. همچنین در این تحقیق از ماتریس اغتشاش برای مقایسه‌ی مدل‌های تحقیق بین داده‌های واقعی و پیش‌بینی‌شده استفاده شد. نتایج نشان داد در داده‌های آموزش شاخص حساسیت مدل‌های KNN و SVM بیشتر بوده است. همچنین با توجه به شاخص خاصیت مدل‌های RT و KNN نتایج بهتری داشته‌اند. در داده‌های آزمون شاخص حساسیت مدل‌های SVM و CHAID بیشتر و مدل‌های KNN و درخت C5 با توجه به شاخص خاصیت کاراتر بوده‌اند. با توجه به روند چشمگیر کاهش میزان بارش باران در استان همدان، کاهش منابع آب و میزان وابستگی فضاهای طبیعی به ریزش‌های جوی در این منطقه، پیش‌بینی دقیق باران با استفاده از روش‌های داده‌کاوی مانند KNN و SVM می‌تواند کمک شایانی در مدیریت و برنامه‌ریزی منابع آب، فعالیت‌های کشاورزی، گردشگری و غیره کند. محدودیت آب قابل‌دسترس، توزیع غیریکنواخت آن در سطح استان همدان، الگوی نامناسب شهرنشینی و مراکز سکونتگاهی و نوع و شیوه تولید محصولات زراعی، تأمین آب موردنیاز را در این منطقه از کشور مشکل ساخته است. همچنین این استان به لحاظ موقعیت ویژه

کشاورزی مستلزم برنامه‌ریزی‌های گسترده‌ای درزمینه کشت بعضی محصول‌های استراتژیک نظیر گندم و جو و ...، است، این برنامه‌ریزی بدون شناخت بارش و بدون در نظر گرفتن توزیع زمانی و مکانی بارش موفق به نظر نمی‌رسد ازاین‌رو مطالعات اقلیمی در استان و به‌روز کردن داده‌ها و اطلاعات اقلیم‌شناسی خصوصاً بارش و شرایط آن‌ها در آینده ضروری به نظر می‌رسد.

در نظر گرفتن برچسب کلاس‌های دیگر مثل روزهای آفتابی، روزهای برفی و روزهای بارانی و برفی به‌عنوان پیشنهاد ارائه می‌گردد. همچنین ازآنجایی‌که متغیر میزان بارش یک متغیر عددی است می‌توان برای پیش‌بینی بارش از مدل‌های رگرسیون، ARMA، شبکه‌های عصبی و مدل مارکف سوئیچینگ برای پیش‌بینی میزان بارش استفاده نمود.



شکل ۳: منحنی مشخصه عملکرد سیستم (ROC) برای مدل‌های تحقیق در داده‌های آموزش و آزمون

جدول ۳. ماتریس همبستگی بین مشخصه‌های داده‌های تحقیق

	ewm	ffm	ffmax	nm	nmax	p/m	pm	tdm	tm	tmax	tmin	twetm	um	umax	umin
ewm	۱	۰/۰۴۸	۰/۰۹۶	-۰/۰۷۱	-۰/۱۰۱	-۰/۲۸۷	-۰/۶۹۴	۰/۹۳۳	۰/۷۲۵	۰/۶۷۷	۰/۷۳۸	۰/۸۱۶	-۰/۲۵۲	-۰/۱۵۹	-۰/۲۲
ffm	۰/۰۴۸	۱	۰/۷۸۳	۰/۱۷۲	۰/۱۸۵	-۰/۲۸۳	-۰/۲۷	۰/۰۸۱	۰/۱۴۴	۰/۰۹۵	۰/۲۳۹	۰/۱۵	-۰/۱۸۱	-۰/۱۶۹	-۰/۱۴۹
ffmax	۰/۰۹۶	۰/۷۸۳	۱	۰/۱۲۸	۰/۱۶۲	-۰/۲۴۶	-۰/۳۰۴	۰/۱۳۷	۰/۱۹۳	۰/۱۶۷	۰/۲۸	۰/۲۰۴	-۰/۲۰۲	-۰/۱۶۵	-۰/۲۱۳
nm	-۰/۰۷۱	۰/۱۷۲	۰/۱۲۸	۱	۰/۸۹۱	-۰/۱۰۵۳	۰/۲۵۸	-۰/۱۰۵۲	-۰/۴۰۵	-۰/۴۸۳	-۰/۱۸۶	-۰/۳۰۷	۰/۶۰۳	۰/۴۸۷	۰/۶۱۷
nmax	-۰/۱۰۱	۰/۱۸۵	۰/۱۶۲	۰/۸۹۱	۱	-۰/۱۰۰۷	۰/۲۹۱	-۰/۱۰۸۹	-۰/۴۱۹	-۰/۴۸۶	-۰/۲۲۳	-۰/۳۳	۰/۵۶۹	۰/۴۹۵	۰/۵۴
p/m	-۰/۲۸۷	-۰/۲۸۳	-۰/۲۴۶	-۰/۱۰۵۳	-۰/۱۰۰۷	۱	۰/۶۷۲	-۰/۲۶۹	-۰/۲۸۱	-۰/۲۴۳	-۰/۳۶۴	-۰/۲۸	۰/۱۷۴	۰/۲۰۱	۰/۰۹۴
pm	-۰/۶۹۴	-۰/۲۷	-۰/۳۰۴	۰/۲۵۸	۰/۲۹۱	۰/۶۷۲	۱	-۰/۷۱۴	-۰/۸۵۸	-۰/۸۳۹	-۰/۸۶۹	-۰/۸۵۴	۰/۶۶۶	۰/۶۰۳	۰/۵۷۴
tdm	۰/۹۳۳	۰/۰۸۱	۰/۱۳۷	-۰/۱۰۵۲	-۰/۱۰۸۹	-۰/۲۶۹	-۰/۷۱۴	۱	۰/۷۵۸	۰/۷۰۲	۰/۷۸۶	۰/۸۶۴	-۰/۲۸۸	-۰/۱۸۹	-۰/۲۶۱
tm	۰/۷۲۵	۰/۱۴۴	۰/۱۹۳	-۰/۴۰۵	-۰/۴۱۹	-۰/۲۸۱	-۰/۸۵۸	۰/۷۵۸	۱	۰/۹۴۵	۰/۸۸۱	۰/۹۶۱	-۰/۷۸۳	-۰/۶۷۵	-۰/۷
tmax	۰/۶۷۷	۰/۰۹۵	۰/۱۶۷	-۰/۴۸۳	-۰/۴۸۶	-۰/۲۴۳	-۰/۸۳۹	۰/۷۰۲	۰/۹۴۵	۱	۰/۸۴۶	۰/۹۱۷	-۰/۸۱۱	-۰/۶۹۶	-۰/۷۴۷
tmin	۰/۷۳۸	۰/۲۳۹	۰/۲۸	-۰/۱۸۶	-۰/۲۲۳	-۰/۳۶۴	-۰/۸۶۹	۰/۷۸۶	۰/۸۸۱	۰/۸۴۶	۱	۰/۸۹۸	-۰/۶۲۴	-۰/۵۸۴	-۰/۵۲۸
twetm	۰/۸۱۶	۰/۱۵	۰/۲۰۴	-۰/۳۰۷	-۰/۳۳	-۰/۲۸	-۰/۸۵۴	۰/۸۶۴	۰/۹۶۱	۰/۹۱۷	۰/۸۹۸	۱	-۰/۶۷۳	-۰/۵۵۷	-۰/۶۰۸
um	-۰/۲۵۲	-۰/۱۸۱	-۰/۲۰۲	۰/۶۰۳	۰/۵۶۹	۰/۱۷۴	۰/۶۶۶	-۰/۲۸۸	-۰/۷۸۳	-۰/۸۱۱	-۰/۶۲۴	-۰/۶۷۳	۱	۰/۸۸۴	۰/۹۲۱
umax	-۰/۱۵۹	-۰/۱۶۹	-۰/۱۶۵	۰/۴۸۷	۰/۴۹۵	۰/۲۰۱	۰/۶۰۳	-۰/۱۸۹	-۰/۶۷۵	-۰/۶۹۶	-۰/۵۸۴	-۰/۵۵۷	۰/۸۸۴	۱	۰/۷۰۵
umin	-۰/۲۲	-۰/۱۴۹	-۰/۲۱۳	۰/۶۱۷	۰/۵۴	۰/۰۹۴	۰/۵۷۴	-۰/۲۶۱	-۰/۷	-۰/۷۴۷	-۰/۵۲۸	-۰/۶۰۸	۰/۹۲۱	۰/۷۰۵	۱

جدول ۴. مقادیر ویژه ابعاد به‌دست‌آمده به‌وسیله روش کاهش ابعاد

	بعد اول (ewm)	بعد دوم (ffm)	بعد سوم (nm)	بعد چهارم (p0m)	بعد پنجم (um)
مقدار ویژه	۷/۹۲	۲/۶۵	۱/۸۶	۰/۹۱	**۰/۶۷**
درصد مقدار ویژه	۵۲/۸۱	۱۷/۶۸	۱۲/۴۱	۶/۰۸	**۴/۴۶**
درصد تجمعی مقدار ویژه	۵۲/۸۱	۷۰/۴۹	۸۲/۹۰	۸۸/۹۹	**۹۳/۴۵**

جدول ۵. نتایج محاسباتی معیارهای درصد صحت (CA)، درصد خطا (ER) و شاخص AUC مدل‌های تحقیق به تفکیک داده‌های آموزش و آزمون

نوع داده‌ها	معیار (درصد)	RT	C5	ANN	SVM	BN	KNN	CHAID	QUEST
داده‌های آموزش	صحت (CA)	۸۶/۸۲(۱۰۹۰۹)	۸۹/۷۸ (۱۱۲۸۱)	۸۹/۵۵(۱۱۲۵۲)	۸۹/۹۶(۱۱۳۰۴)	۸۸/۰۶(۱۱۰۶۵)	۹۱/۹(۱۱۵۴۷)	۸۸/۲۹(۱۱۰۹۴)	۸۷/۴۶(۱۰۹۸۹)
	خطا (ER)	۱۳/۱۸(۱۶۵۶)	۱۰/۲۲(۱۲۸۴)	۱۰/۴۵(۱۳۱۳)	۱۰/۰۴(۱۲۶۱)	۱۱/۹۴(۱۵۰۰)	۸/۱(۱۰۱۸)	۱۱/۷۱(۱۴۷۱)	۱۲/۵۴(۱۵۷۶)
	شاخص AUC	۰/۹۴	۰/۹۲	۰/۹۴	۰/۹۴	۰/۹۳	۰/۹۷	۰/۹۳	۰/۸۹
داده‌های آزمون	صحت (CA)	۸۳/۲(۲۵۷۱)	۸۷/۹(۲۷۱۶)	۸۸/۱۲(۲۷۲۳)	۸۹/۱۳(۲۷۵۴)	۸۷/۱۲(۲۶۹۲)	۸۸/۱۹(۲۷۲۵)	۸۶/۹۳(۲۶۸۶)	۸۶/۷۶(۲۶۸۱)
	خطا (ER)	۱۶/۸(۵۱۹)	۱۲/۱(۳۷۴)	۱۱/۸۸(۳۶۷)	۱۰/۸۷(۳۳۶)	۱۲/۸۸(۳۹۸)	۱۱/۸۱(۳۶۵)	۱۳/۰۷(۴۰۴)	۱۳/۲۴(۴۰۹)
	شاخص AUC	۰/۸۹	۰/۸۹	۰/۹۳	۰/۹۴	۰/۹۲	۰/۹۰	۰/۹۲	۰/۸۸

جدول۶. ماتریس اغتشاش کلاس وقوع و عدم وقوع بارش برای داده‌های آموزش مدل‌های تحقیق

مدل	نوع داده	طبقه‌بندی کلاس‌ها	داده‌های پیش‌بینی		شاخص حساسیت	شاخص خاصیت
			عدم وقوع بارش (n)	وقوع بارش (y)	(درصد)	(درصد)
RT	داده‌های	عدم وقوع بارش (n)	۸۵۸۴	۱۴۴۰	۸۵/۶	
	واقعی	وقوع بارش (y)	۲۱۶	۲۳۲۵		۹۱/۵
C5	داده‌های	عدم وقوع بارش (n)	۹۳۷۸	۶۴۶	۹۳/۵	
	واقعی	وقوع بارش (y)	۶۳۸	۱۹۰۳		۷۴/۹
ANN	داده‌های	عدم وقوع بارش (n)	۹۴۵۸	۵۶۶	۹۴/۳	
	واقعی	وقوع بارش (y)	۷۴۷	۱۷۹۴		۷۰/۶
SVM	داده‌های	عدم وقوع بارش (n)	۹۶۰۷	۴۱۷	۹۵/۸	
	واقعی	وقوع بارش (y)	۸۴۴	۱۶۹۷		۶۶/۸
BN	داده‌های	عدم وقوع بارش (n)	۹۲۱۶	۸۰۸	۹۱/۹	
	واقعی	وقوع بارش (y)	۶۹۲	۱۸۴۹		۷۲/۷
KNN	داده‌های	عدم وقوع بارش (n)	۹۶۲۴	۴۰۰	۹۶	
	واقعی	وقوع بارش (y)	۶۱۸	۱۹۲۳		۷۵/۶
CHAID	داده‌های	عدم وقوع بارش (n)	۹۵۱۰	۵۱۴	۹۴/۸	
	واقعی	وقوع بارش (y)	۱۰۶۷	۱۴۷۴		۵۸
QUEST	داده‌های	عدم وقوع بارش (n)	۹۳۵۶	۶۶۸	۹۳/۳۳	
	واقعی	وقوع بارش (y)	۹۰۸	۱۶۳۳		۶۴/۲

جدول ۷. ماتریس اغتشاش کلاس وقوع و عدم وقوع بارش برای داده‌های آزمون مدل‌های تحقیق

مدل	نوع داده	طبقه بندی کلاس‌ها	داده‌های پیش‌بینی		شاخص حساسیت	شاخص خاصیت
			عدم وقوع بارش (n)	وقوع بارش (y)	(درصد)	(درصد)
RT	داده‌های واقعی	عدم وقوع بارش (n)	۲۰۵۱	۴۰۴	۸۳/۵	
		وقوع بارش (y)	۵۲۰	۱۱۵		۱۸
C5	داده‌های واقعی	عدم وقوع بارش (n)	۲۲۷۶	۱۷۹	۹۲/۷	
		وقوع بارش (y)	۱۹۵	۴۴۰		۶۹/۲۹
ANN	داده‌های واقعی	عدم وقوع بارش (n)	۲۲۹۶	۱۵۹	۹۳/۵	
		وقوع بارش (y)	۲۰۸	۴۲۷		۶۷/۲۴
SVM	داده‌های واقعی	عدم وقوع بارش (n)	۲۳۴۱	۱۱۴	۹۵/۳۵	
		وقوع بارش (y)	۲۲۲	۴۱۳		۶۵
BN	داده‌های واقعی	عدم وقوع بارش (n)	۲۲۳۶	۲۱۷	۹۱	
		وقوع بارش (y)	۲۲۹	۴۰۶		۶۳/۹
KNN	داده‌های واقعی	عدم وقوع بارش (n)	۲۳۰۰	۱۵۵	۹۳/۶۸	
		وقوع بارش (y)	۲۱۰	۴۲۵		۷۰
CHAID	داده‌های واقعی	عدم وقوع بارش (n)	۲۳۴۱	۱۱۴	۹۵/۳۵	
		وقوع بارش (y)	۲۹۰	۳۴۵		۵۴/۳
QUEST	داده‌های واقعی	عدم وقوع بارش (n)	۲۲۸۱	۱۷۴	۹۲/۹	
		وقوع بارش (y)	۲۳۵	۴۰۰		۶۳

Reference:

Adaryani, F. R., Jamshid Mousavi, S., & Jafari, F. (2022). Short-term rainfall forecasting using machine learning-based approaches of PSO-SVR, LSTM and CNN. *Journal of Hydrology, 614*, 128463. doi: https://doi.org/10.1016/j.jhydrol.2022.128463.

Aftab, S ،.Ahmad, M., Hameed, N., Bashir, S., Ali, I., & Nawaz, Z. (2018). Rainfall Prediction in Lahore City using Data Mining Techniques. *International Journal of Advanced Computer Science and Applications, 9*. doi: 10.14569/IJACSA.2018.090439.

Alavi, A., Gandomi ،A., Mollahassani, A., Heshmati R, A. A., & rashed, a. (2010). Modeling of maximum dry density and optimum moisture content of stabilized soil using artificial neural networks. *Journal of Plant Nutrition and Soil Science, 173*, 368-379. doi: 10.1002/jpln.2008002.

Alcantara, A. L., & Ahn, K.-H. (2021). Future flood riverine risk analysis considering the heterogeneous impacts from tropical cyclone and non-tropical cyclone rainfalls: Application to daily flows in the Nam River Basin, South Korea. *Advances in Water Resources, 154*, 103983. doi: https://doi.org/10.1016/j.advwatres.2021.103983.

Anderson, C. J., Cadeddu, R., Anderson, D. N., Huxford, J. A., VanLuik, E. R., Odeh, K., & Bortolato, M. (2023). A novel naïve Bayes approach to identifying grooming behaviors in the force-plate actometric platform. *Journal of Neuroscience Methods*, 110026. doi: https://doi.org/10.1016/j.jneumeth.2023.110026.

Bagirov, A. M., Mahmood, A., & Barton, A. (2017). Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach. *Atmospheric Research, 188*, 20-29. doi: https://doi.org/10.1016/j.atmosres.2017.01.003.

Bahrami, M., Amiri, M. J., Rezaei Maharluei, F., & Ghaffari, K. A. (2017). Data Pre-Processing Effects on the Artificial Neural Network Performance to Predict Monthly Rainfall (Case Study: Abadeh County). *Iranian journal of Ecohydrology, 4*(1), 29-37. doi: 10.22059/ije.2017.60880. [in Persian]

Bhattacharya, B., & Solomatine, D. P. (2005). Neural networks and M5 model trees in modelling water level–discharge relationship. *Neurocomputing, 63*, 381-396. doi: https://doi.org/10.1016/j.neucom.2004.04.016.

Cramer, S., Kampouridis, M., Freitas, A., & Alexandridis, A. (2017). An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. *Expert Systems with Applications, 85*. doi: 10.1016/j.eswa.2017.05.029.

Danandeh Mehr, A., Nourani, V., Khosrowshahi, V., & Ghorbani, M. A. (2018). A hybrid support vector regression–firefly model for monthly rainfall forecasting. *International Journal of Environmental Science and Technology, 16*, 1-12. doi: 10.1007/s13762-018-1674-2.

Dastourani, M. T., Habibipoor, A., Ekhtesasi, M. R., Talebi, A., & Mahjoobi, J. (2013). Evaluation of the Decision Tree Model in Precipitation Prediction (Case study: Yazd Synoptic Station). *Iran-Water Resources Research, 8*(3), 14-27 .[in Persian]

Fahad, S., Su, F., Khan, S. U., Naeem, M. R., & Wei, K. (2023). Implementing a novel deep learning technique for rainfall forecasting via climatic variables: An approach via hierarchical clustering analysis. *Science of The Total Environment, 854*, 158760. doi: https://doi.org/10.1016/j.scitotenv.2022.158760.

Haidar, A., & Verma, B. (2018). A novel approach for optimizing climate features and network parameters in rainfall forecasting. *Soft Computing, 22*. doi: 10.1007/s00500-017-2756-7.

He, S., Li, Z., & Liu, X. (2023). An improved GEV boosting method for imbalanced data classification with application to short-term rainfall prediction. *Journal of Hydrology, 617*.128882. doi: https://doi.org/10.1016/j.jhydrol.2022.128882.

Josso, P., Hall, A., Williams, C., Le Bas, T., Lusty, P., & Murton, B. (2023). Application of random-forest machine learning algorithm for mineral predictive mapping of Fe-Mn crusts in the World Ocean. *Ore Geology Reviews, 162*, 105671. doi: https://doi.org/10.1016/j.oregeorev.2023.105671.

Kisi, O., Genc, O., Dinc, S., & Zounemat-Kermani, M. (2016). Daily pan evaporation modeling using chi-squared automatic interaction detector, neural networks ،classification and regression tree. *Computers and Electronics in Agriculture, 122*, 112-117. doi: https://doi.org/10.1016/j.compag.2016.01.026.

Lee, S., Bae, J. H., Hong, J., Yang, D., Panagos, P., Borrelli, P., & Lim, K. J. (2022). Estimation of rainfall erosivity factor in Italy and Switzerland using Bayesian optimization based machine learning models. *CATENA, 211*, 105957. doi: https://doi.org/10.1016/j.catena.2021.105957.

Mahtabi, G., Taran, F., & Mozafari, S. (2018). Prediction of daily rainfall using meteorological data of previous days (case study: Isfahan city). *Journal of Physical Geography, 11*(39), 99-114 .[in Persian]

Mallika, M., & Nirmala, M. (2018). An environmental study on forecasting rainfall using data mining technique and ARIMA model: An integrated approach. *Ekoloji, 27*, 1133-1141.

Markuna, S., Kumar, P., Ali, R., Vishwakarma, D., Kushwaha, K., Kumar, R., & Kuriqi, A. (2023). Application of Innovative Machine Learning Techniques for Long-Term Rainfall Prediction. *Pure and Applied Geophysics, 180*. doi: 10.1007/s00024-022-03189-4.

Mishra, N., Soni, H., Sharma, S., & Upadhyay, A. (2018). Development and Analysis of Artificial Neural Network Models for Rainfall Prediction by Using Time-Series Data. *International Journal of Intelligent Systems and Applications, 10*, 16-23. doi: 10.5815/ijisa.2018.01.03.

Modaresi, F., & Araghinejad, S. (2014). A Comparative Assessment of Support Vector Machines, Probabilistic Neural Networks, and K-Nearest Neighbor Algorithms for Water Quality Classification. *Water Resources Management, 28*(12), 4095-4111. doi: 10.1007/s11269-014-0730-z.

Thanushkodi, N. K. (2010). An Improved k-Nearest Neighbor Classification Using Genetic Algorithm. *International Journal of Computer Science Issues, 7*.

Pang, S.-l,. &Gong, J. z. (2009). C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks. *Systems Engineering - Theory & Practice, 29*, 94-104. doi: 10.1016/S1874-8651(10)60092-0.

Ritschard, G. (2010). *CHAID and earlier supervised tree methods*.

Sattari, M. T., Joudi, A. R., & Kusiak, A. (2016). Estimation of Water Quality Parameters With Data-Driven Model. *Journal AWWA, 108*(4), E232-E239. doi: https://doi.org/10.5942/jawwa.2016.108.0012.

Shin, K.-s., Lee, T., & Kim, H.-j. (2005). Shin, K.S.: An Application of Support Vector Machines in Bankruptcy Prediction Model. Expert Systems and Applications 28, 127-135. *Expert Systems with Applications, 28*, 127-135. doi: 10.1016/j.eswa.2004.08.009.

Singh, P. (2018). Indian summer monsoon rainfall (ISMR) forecasting using time series data: A fuzzy-entropy-neuro based expert system. *Geoscience Frontiers, 9*(4), 1243-1257. doi: https://doi.org/10.1016/j.gsf.2017.07.011.

Zarei, M., Zandi, R., & Naemitabar, M. (2022). Assessment of Flood Occurrence Potential using Data Mining Models of Support Vector Machine, Chaid and Random Forest (Case study: Frizi watershed). *journal of watershed management research, 13*(25), 133-144. doi: 10.52547/jwmr.13.25.133.

یادداشت ها

1 Multi-layer perceptron
2 mean square error
3 bagirov
4 cramer
5 pritpal singh
6 shabib aftab
7 haidar and verma
8 mishra
9 mallika and nirmala
10 mean absolute percentage error
11 support vector regression
12 support vector regression
13 long-short term memory
14 convolutional neural network
15 markuna
16 he
17 an improved gradient boosting algorithm
18 automated data preparation
19 principal component analysis
20 information gain
21 gini index
22 multilayer perceptron neural network
23 classification accuracy
24 classification error
25 receiver operating characteristic