



## The effect of preprocessing and reducing the input dimensions of the flow prediction model on optimized Support Vector Regression by Genetic Algorithm

Seyed Ehsan Fatemi<sup>1</sup> , Sedigheh Darabi Cheghanaleki<sup>2</sup> , Maryam Hafezparast Mavadat<sup>3</sup>

<sup>1</sup> Corresponding Author, Assistant professor, Department of Water Science and Engineering, Faculty of Agriculture, Razi University, Kermanshah, Iran. E-mail: e\_fatemi78@yahoo.com

<sup>2</sup> M. Sc graduated of Water Resources, Department of Water Science and Engineering, Faculty of Agriculture, Razi University, Kermanshah, Iran. E-mail: seddigheh.darabi73@gmail.com

<sup>3</sup> Assistant professor, Department of Water Science and Engineering, Faculty of Agriculture, Razi University, Kermanshah, Iran. E-mail: m.hafezparast@razi.ac.ir

### ABSTRACT

#### Introduction

Accurate and correct prediction of surface water flow plays an important role in the principled planning and proper management of water resources. To achieve this, various prediction models using mathematical relationships based on hydrological information can be used. In recent years, machine learning models have received much attention due to their high accuracy in prediction through black box learning. Therefore, by combining or comparing these methods, more favorable results can be used to increase the accuracy of predictions. In this study, monthly discharge of Polechehr hydrometric station for a 48-year has been used.

#### Methodology

Support Vector Machine (SVM), Feature Selection and Random Forest Technique

The Support Vector Machine (SVM) is a non-statistical binary classifier that has received much attention in recent years and it is supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. The SVM is an algorithm that finds unique linear models and uses them to get the maximum hyper-plane margin. Maximizing the superstructure margin maximizes segregation between classes and improves accuracy in the modeling process. The nearest training points at the edge of the cloud are referred to as support vectors and are used to define boundaries between classes.

Feature Selection (FS) is the process of reducing the number of input variables when developing a predictive model. Feature Selection can be defined as the process of identifying related features and removing unrelated and duplicate features. They are desirable to reduce the number of input variables to both reduce the computational cost of modeling, and in some cases to improve the performance of the model.

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. It is a class of group classification methods designed specifically for decision trees.

#### Results and discussion

The first scenario results

The results of monthly flow prediction in the first scenario, with no-preprocessing on data, two-time series and non-time series approaches and feature selection cases are shown in the Table 1.

**Table 1. Model Evaluation results in scenario1 for different approach and cases**

Exp.Var	Train		Test							
	NSE	RMSE	Exp.Var	NSE	RMSE	C	Epsilon	SVR	Feature Selection	Time Series
0.029	-0.13	46.11	-6.70E-07	-0.046	22.81	1	0.01	Simple	No	Yes
0.063	-0.09	45.32	-1.47E-06	-0.051	22.86	2.181	0.0063	GA	No	Yes
0.115	-0.057	44.61	0.16	0.156	20.5	1	0.01	Simple	Yes	Yes
0.479	0.42	33.02	0.243	0.222	19.67	15.31	0.0063	GA	Yes	Yes
0.03	-0.148	43.67	-1.45E-06	-0.147	45.018	1	0.01	Simple	No	No
0.08	-0.098	42.7	-4.00E-06	-0.147	45.01	2.755	0.0097	GA	No	No
0.121	-0.049	41.73	0.096	-0.053	43.12	1	0.01	Simple	Yes	No
0.505	0.452	30.17	0.234	0.174	38.21	16.97	0.0014	GA	Yes	No

The second scenario results

The results of monthly flow prediction in the second scenario, by applying standardization on data, two-time series and non-time series approaches and feature selection cases are shown in table 2.

**Table 2. Model Evaluation results in scenario 2 for different approach and cases**

Train			Test							
Exp.Var	NSE	RMSE	Exp.Var	NSE	RMSE	C	Epsilon	SVR	Feature Selection	Time Series
0.668	0.66	0.583	1.55E-01	0.105	0.376	1	0.01	Simple	No	Yes
0.787	0.784	0.465	1.77E-01	0.126	0.372	3.965	0.0028	GA	No	Yes
0.55	0.539	0.678	0.019	-	0.404	1	0.01	Simple	Yes	Yes
0.552	0.541	0.677	0.019	0.032	0.404	1.04	0.0063	GA	Yes	Yes
0.704	0.698	0.533	5.56E-01	0.54	0.77	1	0.01	Simple	No	No
0.855	0.854	0.371	6.13E-01	0.604	0.714	8.89	0.009	GA	No	No
0.576	0.567	0.638	0.599	0.58	0.736	1	0.01	Simple	Yes	No
0.599	0.588	0.623	0.609	0.59	0.726	7.31	0.0013	GA	Yes	No

**Conclusions**

In the first scenario, without applying preprocessing on data, the results showed that, whether the feature selection is not selected and the whole 12 dimensions are considered, the results of simple and optimized SVR model would be approached to a fixed number for both time series and non-time series approaches. However, by using the model feature selection by the random forest algorithm, the dimensions are reduced from 12 to 1 and monthly flow fluctuations are initially predicted, but the accuracy of the results is still poor. The accuracy of the results in the time series approach has better results than the non-time series approach in similar cases as well. Furthermore, the optimizing effect of genetic algorithm in optimizing the SVR model parameters in the accuracy of results has been significant, but it still had low accuracy in peak flows. In the second scenario, results are significantly improved by applying standardization on data for the all cases. The optimizing effect of the genetic algorithm on the model accuracy was more tangible than the simple SVR when feature selection is not considered to predict peak flood flows. Finally, the best results is for the time series approach no feature selection case. In case, there is no significant difference in the results of the simple SVR model and optimized with the genetic algorithm. As a general conclusion, if there is no any preprocessing on data, time series approach and using feature selection with random forest algorithm will lead to better results in SVR prediction model and also use of genetic algorithm optimizer would be caused a significant effect on improving the results in case. In contrast, by doing the standardization on data, the best results would be achieved by the non-time-series approach and considering feature selection. In case, the genetic algorithm optimization model has not a significant effect on the prediction model accuracy compared to the simple SVR model.

**Keywords:** Discharge Forecasting, Genetic Algorithm, Support Vector Regression, SVM

**Article Type:** Research Article

**Article history:** Received: 09 July 2021 Revised: 26 July 2021 Accepted: 06 August 2021 ePublished: 20 February 2022

**Cite this article:** Fatemi, S.E., Darabi Cheghabaleki, S., & Hafezparast Mavadat, M. (2022). The effect of preprocessing and reducing the input dimensions of the flow prediction model on optimized Support Vector Regression by Genetic Algorithm., *Advanced Technologies in Water Efficiency*, 1(1), 24-47, DOI: 10.22126/ATWE.2021.6660.1002

Publisher: Razi University

© The Author(s).





## اثر پیش پردازش و کاهش ابعاد ورودی مدل پیش بینی دبی بر عملکرد مدل رگرسیون بردار پشتیبان بهینه شده الگوریتم ژنتیک

سید احسان فاطمی<sup>۱</sup>، صدیقه دارابی چقالیکی<sup>۲</sup>، مریم حافظ پرست مودت<sup>۳</sup>

<sup>۱</sup> نویسنده مسئول، استادیار گروه علوم و مهندسی آب، دانشکده کشاورزی، دانشگاه رازی، کرمانشاه، ایران. رایانامه: e\_fatemi78@yahoo.com  
<sup>۲</sup> دانش آموخته کارشناسی ارشد مهندسی منابع آب، دانشکده کشاورزی، دانشگاه رازی، کرمانشاه، ایران. رایانامه: seddigheh.darabi73@gmail.com  
<sup>۳</sup> استادیار گروه علوم و مهندسی آب، دانشکده کشاورزی، دانشگاه رازی، کرمانشاه، ایران. رایانامه: maryam.hafezparast@gmail.com

### چکیده

پیش بینی دقیق و صحیح جریان آب های سطحی در برنامه ریزی اصولی و مدیریت صحیح منابع آب نقش بسزایی دارد. برای دست یافتن به این مهم مدل های پیش بینی مختلف که با استفاده از روابط ریاضی بر پایه اطلاعات هیدرولوژی بنا شده اند، می توانند داده های مورد نیاز را با دقت کافی پیش بینی کنند. در این مطالعه از داده های دبی جریان ماهانه ایستگاه هیدرومتری پل چهر در یک دوره آماری ۴۸ ساله (۱۳۹۷-۱۳۹۸) شهر یور-۱۳۵۰ (مهر) استفاده شد. دو سناریوی اصلی با و بدون اعمال پیش پردازش (استانداردسازی) با دو رویکرد سری زمانی یا غیرسری زمانی بررسی شد. همچنین از الگوریتم جنگل تصادفی برای بررسی کاهش ابعاد ورودی مدل استفاده شد. در هر سناریو معیارهای ارزیابی مدل تغییرات واریانس، ضریب ناش-ساتکلیف و مجذور مربعات خطا محاسبه گردید. در همه حالت ها به ترتیب ۸۰ و ۲۰ درصد داده ها برای آموزش و آزمودن مدل در نظر گرفته شده است. مدل نوشته شده به زبان برنامه نویسی پایتون است. از الگوریتم ژنتیک برای بهینه سازی پارامترهای روش رگرسیون بردار پشتیبان استفاده شد. نتایج نشان داد که ابتدا استانداردسازی سپس در نظر نگرفتن توالی سری زمانی داده ها، کاهش ابعاد ورودی مدل نیز استفاده از الگوریتم ژنتیک برای بهینه سازی پارامترهای مدل رگرسیون بردار پشتیبان به ترتیب بیشترین اثر بر دقت پیش بینی را دارد به طوری که بالاترین ضریب NSE برای داده های آموزش برابر ۰/۸۵ و برای آزمودن معادل ۰/۶ است. چنانچه عمل استانداردسازی داده ها صورت نگیرد، منظور نمودن رویکرد سری زمانی و کاهش ابعاد ورودی مدل منجر به نتایج بهتری در پیش بینی مدل SVR خواهد شد و استفاده از بهینه ساز الگوریتم ژنتیک نسبت به مدل ساده آن تأثیر معنی داری بر بهبود نتایج خواهد داشت.

**واژه های کلیدی:** پیش بینی دبی، ماشین بردار پشتیبان، الگوریتم ژنتیک، SVM

نوع مقاله: مقاله پژوهشی

سابقه مقاله: دریافت: ۱۸ تیر ۱۴۰۰ اصلاح: ۰۴ مرداد ۱۴۰۰ پذیرش: ۱۵ مرداد ۱۴۰۰ چاپ الکترونیکی: ۰۱ اسفند ۱۴۰۰

**استناد:** فاطمی، س.ا.، دارابی چقالیکی، ص.، و حافظ پرست مودت، م. (۱۴۰۰). اثر پیش پردازش و کاهش ابعاد ورودی مدل پیش بینی دبی بر عملکرد مدل

رگرسیون بردار پشتیبان بهینه شده الگوریتم ژنتیک. *فناوری های پیشرفته در بهره وری آب*، (۱)، ۴۷-۲۴. شناسه

دیجیتال: 10.22126/ATWE.2021.6660.1002



## مقدمه

در مدل‌سازی دبی رودخانه برای مدیریت منابع آب، مدیریت ریسک مخصوصاً در مناطق کوهستانی به دلیل وابستگی مناطق پایین‌دست به آب موردنیاز برای مصارف و همچنین نیاز به برق تولیدشده اهمیت زیادی دارد؛ همچنین با توجه به ماهیت جریان که فراخور تغییرات مکانی و زمانی دستخوش تغییر می‌شود و ثبات ندارد و همچنین برداشت داده‌های جریان که با استفاده از سری زمانی داده‌های ثبت‌شده از ایستگاه‌های هیدرومتری است و احتمال بروز خطا در آن وجود دارد بنابراین نیاز به یک مدل قوی و جامع برای مدیریت منابع و پیش‌بینی داده‌ها امری ضروری است. طی سال‌های اخیر، مدل‌های یادگیری ماشینی به دلیل دقت بالا در پیش‌بینی از طریق یادگیری به‌صورت جعبه سیاه موردتوجه زیادی قرار گرفته‌اند. از این رو با مطالعه ترکیبی و یا مقایسه‌ای این روش‌ها می‌توان به نتایج مطلوب‌تری برای بالا بردن دقت پیش‌بینی‌ها استفاده کرد.

## مبانی نظری و پیشینه پژوهش

سیوپراگاسم و همکاران (۲۰۰۱) پیش‌بینی بارش باران و رواناب با روش SSA-SVM را بررسی کردند در این مطالعه یک فن پیش‌بینی ساده و کارآمد بر اساس تحلیل طیف مفرد (SSA) همراه با ماشین بردار پشتیبانی (SVM) پیشنهاد شده است. درحالی‌که SSA تجزیه سری زمان اصلی را به مجموعه‌ای از اجزای فرکانس بالا و پایین، SVM در برخورد کارآمد با عملکرد محاسباتی و عمومی‌سازی در یک فضای ورودی با ابعاد بالا کمک می‌کند. فن پیشنهادی برای پیش‌بینی داده‌های آبیگر تریگوارلد (دانمارک) و داده‌های بارش سنگاپور به‌عنوان مطالعات موردی اعمال می‌شود. نتایج با روش پیش‌بینی غیرخطی (NLP) مقایسه می‌شوند. مقایسه‌ها نشان می‌دهد که دقت فن پیشنهادی به‌طور قابل‌توجهی بالاتر از فن NLP است. زارع زاده مهریزی و بزرگ حداد (۱۳۸۹) شبیه‌سازی و پیش‌بینی آینده با استفاده از الگوریتم ترکیبی ANN-GA<sup>۱</sup> را بررسی کرده‌اند، در این مطالعات مؤثر بودن یا نبودن هر یک از عوامل دما، بارش، آبدی و تأخیرهای زمانی این پارامترها در شبیه‌سازی و پیش‌بینی آبدی رودخانه موردبررسی قرار گرفته است. به‌منظور دستیابی آسان‌تر به این پارامترها و نیز تعیین ترکیب بهینه پارامترهای مؤثر در واسنجی ضرایب ANN از الگوریتم ژنتیک (GA) به‌عنوان یک ابزار بهینه‌سازی استفاده گردیده است. در این تحقیق تابع هدف کاهش مقدار میانگین مربعات خطای کل در نظر گرفته شد. مقادیر ضریب همبستگی بین آبدی‌های محاسباتی و موجود برای دو مدل شبیه‌سازی و پیش‌بینی به ترتیب برابر با ۰/۸۶ و ۰/۷۹ به دست آمد که بیانگر کارایی الگوریتم ANN-GA در شبیه‌سازی و پیش‌بینی آبدی با دقت مطلوب می‌باشد. همچنین نتایج نشان‌دهنده برتری جواب‌های به‌دست‌آمده از مدل شبیه‌سازی نسبت به مدل پیش‌بینی می‌باشند. از دلایل این برتری می‌توان به وجود عوامل هواشناسی ماه موردنظر در تعیین میزان آبدی اشاره نمود. احمدی و همکاران (۱۳۹۳) روش‌های برنامه‌ریزی ژنتیک و ماشین بردار پشتیبان را در پیش‌بینی جریان روزانه رودخانه باراندوزچای بررسی کردند. داده‌های جریان روزانه این رودخانه در ایستگاه دیزج در خلال سال‌های ۱۳۸۵ تا ۱۳۸۹ برای ایجاد مدل استفاده شد که ۸۰ درصد داده‌ها برای آموزش و ۲۰ درصد برای آزمودن مدل بکار رفته است. نتایج نشان داد که در هر دو روش، مدل‌های شامل جریان یک، دو و سه روز قبل بالاترین دقت را در مرحله صحت‌سنجی داشتند. همچنین دقت هر دو مدل با افزایش مقادیر دبی کاهش می‌یابد. مقایسه نتایج دو مدل نشان داد که گرچه دقت روش برنامه‌ریزی ژنتیک با  $R = ۰/۹۷۸$  و  $RMSE^۲ = ۱/۶۶ \text{ m}^3/\text{s}$  نسبت به روش ماشین بردار پشتیبان  $R = ۰/۹۷۶$  و  $RMSE = ۱/۸۰ \text{ m}^3/\text{s}$  اندکی بیشتر بود، اما روش SVM به‌مراتب ساده‌تر از روش

<sup>1</sup> Sivapragasam et al

<sup>2</sup> Singular Spectrum Analysis- Support Vector Machine

<sup>3</sup> Singular Spectrum Analysis

<sup>4</sup>Support Vector Machine

<sup>5</sup> Trykgevælde

<sup>6</sup> Nonlinear programming

<sup>7</sup> Artificial Neural Networks- Genetic Algorithm

<sup>8</sup>Genetic Algorithm

<sup>9</sup> Root Mean Square Error

GP می‌باشد و این روش می‌تواند به‌عنوان یک روش کاربردی برای پیش‌بینی جریان روزانه بکار رود. سیدیان و همکاران (۱۳۹۳) پیش‌بینی دبی جریان رودخانه با استفاده از داده‌کاوی و سری زمانی را بررسی کردند که در آن دبی جریان در سه ایستگاه حاجی‌قوشان، قره‌شور و تمر در حوضه آبخیز گرگان رود برای سال‌های آبی ۹۰-۱۳۸۱ شبیه‌سازی شد. به‌منظور شبیه‌سازی از روش آماری سری زمانی در قالب الگوی اتو رگرسیون (AR) و داده‌کاوی در قالب ماشین بردار پشتیبان (SVM) به دو صورت ماهانه و هفتگی استفاده شد. نتایج در مقیاس ماهانه نشان داد هر دو روش در ایستگاه تمر، دقت کم و در ایستگاه حاجی قوشان، دقت خوبی دارند. در ایستگاه قره‌شور SVM توانست ضریب تعیین سری زمانی ماهانه را به مقدار ۰/۲۹ افزایش و خطای RMSE را ۳۵ درصد کاهش دهد و شبیه‌سازی دقیق‌تری انجام دهد. هر دو روش در ایستگاه‌های تمر و قره‌شور دبی هفتگی را با دقت کمی پیش‌بینی کردند. در ایستگاه حاجی‌قوشان ضریب تعیین روش سری زمانی هفتگی ۰/۹۱ و SVM برابر ۰/۸۶ است. آماره DDR نشان داد در ایستگاه حاجی‌قوشان در مقیاس ماهانه روش SVM نسبت به سری زمانی دارای دقت بیشتری است و در مقیاس هفتگی دقت این دو روش برابر است. نتایج این مطالعه نشان داد که روش SVM در هر دو مقیاس ماهانه و هفتگی دقت بیشتری نسبت به سری زمانی دارد؛ همچنین دقت هر دو روش در مقیاس ماهانه بیشتر از مقیاس هفتگی است. سادھیر و همکاران (۲۰۱۳) پیش‌بینی جریان توسط SVM با بهینه‌سازی ازدحام ذرات رفتار کوانتومی را مطالعه کردند در این مطالعه با بررسی صحت مدل هیبریدی SVM-QPSO (بهینه‌سازی ازدحام ذرات رفتار شده ماشین بردار-کوانتوم) در پیش‌بینی جریان‌های ماهانه پرداخت. مدل پیشنهادی SVM-QPSO در پیش‌بینی مقادیر جریان ایستگاه ویجاواوا و ایستگاه پولواوارام آندرا پرادش در هند به کار گرفته شده است. مدل SVM با ساختارهای ورودی مختلف ساخته شده و بهترین ساختار با استفاده از میانگین خطای مربع نرمال شده (NMSE) و ضریب همبستگی (R) تعیین می‌شود. عملکرد بهینه‌سازی ازدحام ذرات با رفتار کوانتومی بیشتر در این مطالعه برای تعیین مقادیر مطلوب پارامترهای SVM با به حداقل رساندن NMSE اقتباس شده است. بعدها عملکرد مدل SVM-QPSO به‌طور کامل با مدل‌های پیش‌بینی محبوب مقایسه شد. نتایج نشان داد که SVM-QPSO به‌مراتب روش بهتری برای پیش‌بینی جریان‌های ماهانه است، زیرا درجه بالایی از دقت و قابلیت اطمینان را فراهم می‌کند. معینی و همکاران (۱۳۹۴) مدل‌سازی دبی ماهانه ورودی به مخزن سد جامیشان با مدل‌های خود همبسته با میانگین متحرک تجمعی و سامانه استنتاج فازی-عصبی انطباقی را مطالعه نمودند. این مدل‌ها به ترتیب بر پایه روش‌های تصادفی<sup>۱</sup> و هوش مصنوعی (AI) هستند. برای مدل‌سازی تا بیشینه پنج پارامتر برای مدل ARIMA در نظر گرفته شده که ۱۲۹۶ مدل تولید و بر سری زمانی برازش داده شده است. برای مدل ANFIS نیز ۱۴ نوع ترکیب ورودی با استفاده از دبی‌های با تأخیرهای مختلف تعیین گردیده است. دبی‌های ماهانه مدل‌سازی توسط مدل‌های ARIMA و ANFIS به کمک شاخص‌های مختلف از جمله شاخص میانگین قدر مطلق خطای نسبی (MARE) با یکدیگر مقایسه شدند که مقدار آن به ترتیب برابر ۰/۳۹۸ و ۰/۸ برای هر مدل به‌دست آمده است. نتایج به‌دست آمده نشان داد مدل ARIMA در مدل‌سازی دبی‌های با مقادیر کم و مدل‌سازی‌های کوتاه‌مدت و درازمدت بسیار توانمندتر از ANFIS است. ناظری تهرودی و همکاران (۱۳۹۵) بررسی دقت مدل‌های ANFIS، SVM و GP در مدل‌سازی مقادیر دبی جریان رودخانه را مطالعه کردند. برای مدل‌سازی دبی ماهانه رودخانه نازلوچای در محل ایستگاه هیدرومتری تپیک واقع در غرب دریاچه ارومیه تحت تأثیر بارش حوضه رودخانه مطالعه شده بررسی و مقایسه شده است. در همه روش‌های یادشده الگوهای M1 تا M5 داده‌های دبی جریان با تأخیر یک تا پنج و الگوهای M6 تا M10 الگوی ترکیبی با داده‌های بارش و دبی و با تأخیرهای یک تا پنج ماه بررسی شدند. برای بررسی مقادیر خطای ناشی از مدل‌سازی از سه روش ضریب تبیین، مجذور میانگین مربعات خطا و معیار کارایی مدل استفاده شده است. نتایج

<sup>1</sup> Auto-Regressive

<sup>2</sup> Data Driven Regression

<sup>3</sup> Sudheer et al

<sup>4</sup> Stochastic

<sup>5</sup> Artificial Intelligence

<sup>6</sup> Auto Regressive Integrated Moving Average

<sup>7</sup> Adaptive Neuro-Fuzzy Inference Systems

بررسی دقت و میزان خطای مدها نشان داد الگوی ترکیبی فقط در مدل SVM بهترین نتیجه را داده است و در دو مدل GP و ANFIS الگوهای تک سری بهترین نتیجه را ارائه کردند. از بین سه مدل بررسی‌شده، مدل ANFIS با الگوی ورودی چهار و پنج تأخیر بهترین نتیجه را داد. به‌طور کلی، نتایج نشان داد با به‌کارگیری مدل ANFIS در مدل‌سازی دبی جریان ماهانه رودخانه نازلوچای، خطای مدل نسبت به دو مدل GP و SVM به ترتیب حدود ۲۳ و ۳ درصد (در واحد دبی جریان) کاهش و دقت مدل نیز نسبت به دو مدل GP و SVM به ترتیب حدود ۱۰ و ۴ درصد افزایش پیدا کرده است. نوذری و توکلی (۱۳۹۶) پیش‌بینی جریان با استفاده از مدل ماشین بردار پشتیبان بر مبنای سری‌های زمانی دبی و بارش در ایستگاه‌های بالادست (مطالعه موردی: ایستگاه هیدرومتری تله زنگ) را مطالعه کردند؛ به‌منظور پیش‌بینی جریان ماهانه ایستگاه هیدرومتری تله زنگ از مدل ماشین بردار پشتیبان و بر اساس آمار و اطلاعات ۱۰ ایستگاه هیدرومتری و ۸ ایستگاه باران‌سنجی بالادست حوضه در طول یک دوره آماری ۲۰ ساله (۱۳۷۱-۱۳۹۰) استفاده شده است. به همین منظور و در اولین گام، تأثیر استفاده از سری‌های زمانی میزان جریان، بارش و ترکیبی از این دو پارامتر به‌عنوان ورودی و در گام بعد تأثیر تعداد ایستگاه‌های هیدرومتری و باران‌سنجی بر نتایج پیش‌بینی، مورد بررسی قرار گرفت. مقایسه نتایج به کمک چهار شاخص آماری: ضریب تبیین (R)، جذر میانگین مربعات خطا (RMSE)، خطای استاندارد (SE) و درصد خطای مطلق (MAPE) صورت گرفت. نتایج نشان داد میزان ضریب تبیین، جذر میانگین مربعات خطا، خطای استاندارد و درصد خطای مطلق زمانی که از آمار بارندگی و دبی، به‌عنوان ورودی در مدل استفاده شود به ترتیب برابر است با ۰/۹۸۴، ۳۳/۱۹، ۰/۱۲ و ۰/۱۰۸۵. اگر تنها از سری زمانی بارندگی برای پیش‌بینی میزان جریان استفاده شود، آنگاه مقادیر ضریب تبیین، جذر میانگین مربعات خطا، خطای استاندارد و درصد خطای مطلق به ترتیب برابر با ۰/۵۱۱، ۱۸۴/۹، ۰/۶۸ و ۸۱/۹۵ می‌شود. همچنین با افزایش تعداد ایستگاه‌های هیدرومتری و باران‌سنجی در بالادست، مدل قادر خواهد بود دبی را با دقت بیشتری پیش‌بینی نماید. دهقانی و همکاران (۱۳۹۶) مقایسه عملکرد مدل‌های ماشین بردار پشتیبان، برنامه‌ریزی بیان ژن و شبکه بیزین در پیش‌بینی جریان روزانه رودخانه‌ها (مطالعه موردی: رودخانه کشکان) را مورد بررسی قرار دادند. نتایج نشان داد هر سه مدل، در ساختاری متشکل از ۱ تا ۵ تأخیر زمانی نتایج بهتری نسبت به سایر ساختارهای تعیین‌شده دارد. همچنین، با توجه به معیارهای ارزیابی مشخص شد که از بین مدل‌های به‌کار رفته، مدل ماشین بردار پشتیبان، بیشترین دقت  $R=0/910$ ، کمترین ریشه میانگین مربعات خطا و  $m^3/s$  داده است  $RMSE=0/002$  کمترین میانگین قدر مطلق خطا  $MAE=0/001 m^3/s$  در مرحله صحت‌سنجی را دارا بوده است. در مجموع، نتایج نشان داده است افزایش تعداد پارامتر مؤثر در مدل‌های مختلف جهت شبیه‌سازی موجب عملکرد بهتر در تخمین دبی جریان می‌شود. همچنین نتایج نشان داده است مدل ماشین بردار پشتیبان عملکرد بهتری نسبت به مدل‌های شبکه بیزین و برنامه‌ریزی بیان ژن دارد. صمدی و فتح‌آبادی (۱۳۹۷) استفاده از مدل‌های سری زمانی، شبکه عصبی و ماشین بردار پشتیبان جهت پیش‌بینی دبی ورودی به سد گرگان را بررسی کردند. بدین منظور از داده‌های هیدرومتری ایستگاه قزاقلی با دوره آماری ۴۷ سال و سه مدل سری زمانی، شبکه عصبی و ماشین بردار پشتیبان جهت پیش‌بینی استفاده شد و نتایج مدل‌های مختلف مورد مقایسه قرار گرفت. با توجه به نتایج به‌دست‌آمده در مدل سری زمانی بر اساس معیارهای آکاییک و شوارتز، مدل  $ARIMA(1,0,0)(1,0,1)$  به‌عنوان مدل بهینه انتخاب شده است. در مدل شبکه عصبی، شبکه با ورودی ۲ و ۱۰ نرون به‌عنوان شبکه برتر انتخاب شده و در مدل ماشین بردار پشتیبان شبکه با ورودی ۱ به‌عنوان شبکه برتر انتخاب شد. در نهایت با توجه به نتایج به‌دست‌آمده و با توجه به معیارهای ارزیابی مدل‌ها، مدل ماشین بردار پشتیبان بهترین عملکرد را نسبت به دو مدل دیگر داشت. مقادیر RMSE و MAPE برای مدل ماشین بردار پشتیبان به ترتیب ۵/۳۱ و ۱/۰۷، برای مدل شبکه عصبی به ترتیب ۹/۸۸ و ۲/۷۸ و برای مدل سری زمانی به ترتیب ۸/۸۴ و ۱/۲۰ به‌دست‌آمده آمد. بر اساس نتایج این پژوهش، بهترین مدل برای پیش‌بینی دبی ماهانه ورودی به سد گرگان مدل ماشین بردار پشتیبان می‌باشد. حشمتی و حافظ پرست مودت (۱۳۹۷) پیش‌بینی دبی با استفاده از تحلیل سری‌های زمانی با مدل SARIMA در شرایط خشکسالی (مطالعه موردی: ایستگاه دبی سنجی پیر سلمان در رودخانه جامیشان) را بررسی کرده و برای این منظور از داده‌های بارندگی ماهانه در یک دوره ۲۵ ساله

<sup>1</sup> Standard Error

<sup>2</sup> Seasonal Auto Regressive Integrated Moving Average

(۱۳۹۰-۱۳۶۵) استفاده شده است. با حذف عوامل نا ایستایی در نهایت پارامترهای مدل غیر فصلی SARIMA با استفاده از نرم افزار MINITAB جهت پیش بینی باران ماهانه محاسبه شده و در انتها مقادیر دبی رودخانه به کمک مدل  $(3,0,1)(3,0,1)_2$  SARIMA نسبت به مدل های دیگر با داشتن میانگین خطای کمتر در پیش بینی برآورد گردیده است.

صمدی و همکاران (۱۳۹۷) پیش بینی دبی ماهانه ورودی به سد بوستان در استان گلستان با استفاده از مدل های داده کاوی و ترکیبی را بررسی کردند؛ مقادیر دبی ماهانه ورودی به سد بوستان برای آینده با استفاده از داده های هیدرومتری ایستگاه تمر و به کارگیری سه مدل سری زمانی، شبکه عصبی مصنوعی و ماشین بردار پشتیبان و همچنین، سه مدل ترکیبی پیش بینی شده و سپس، با استفاده از معیارهای ارزیابی اقدام به مقایسه عملکرد هر کدام از مدل ها شده است. با توجه به نتایج به دست آمده در مدل، سری زمانی بر اساس کمینه بودن معیارهای آکاییک و شوارتز، مدل  $(2,0,0)(1,0,1)$  ARIMA به عنوان مدل برتر انتخاب شد. در مدل شبکه عصبی، شبکه با ورودی ۲ و ۴ نرون و در مدل SVM شبکه با ورودی ۹، به عنوان شبکه برتر انتخاب شدند. در نهایت، با توجه به نتایج به دست آمده از معیارهای ارزیابی، مدل سری زمانی بهترین عملکرد را داشته است که مقادیر معیارهای میانگین مربعات خطا، متوسط مقادیر مطلق خطای نسبی، میانگین مطلق خطا و نش-ساتکلیف برای این مدل به ترتیب برابر با  $0/88$ ،  $4/71$ ،  $0/24$  و  $0/36$  به دست آمده آمد. در نتیجه، مدل سری زمانی به عنوان بهترین مدل برای پیش بینی دبی ماهانه در این ایستگاه معرفی شد. زینعلی و همکاران (۱۳۹۷) ارزیابی مدل های هوش مصنوعی در مدل سازی جریان رودخانه گاماسیاب را مورد مطالعه قرار دادند که در آن به ارزیابی سه مدل به نام های، برنامه ریزی بیان ژن (GEP)، شبکه بیزین (BN) و ماشین بردار پشتیبان (SVM) پرداخته شده است؛ داده های مورد استفاده برای این پژوهش، داده های بارش و جریان روزانه رودخانه گاماسیاب نهادند در یک دوره ۱۰ ساله (۱۳۹۱-۱۳۸۱) می باشد. نتایج نشانگر برتری نسبی مدل برنامه ریزی بیان ژن نسبت به بقیه مدل ها بوده و مدل ماشین بردار پشتیبان (SVM) تقریباً عملکرد بهتری نسبت به شبکه بیزین در مدل سازی جریان روزانه رودخانه داشته است. علاوه بر این، سرعت اجرای مدل برنامه ریزی بیان ژن نسبت به بقیه مدل ها بیشتر بوده و در زمان کوتاهی قادر به ارائه نتایج بوده است. همچنین، مدل SVM در تخمین مقادیر کمینه پایانی نیز دچار بیش برآزش شده است. در نهایت مدل برنامه ریزی بیان ژن با ضریب تبیین  $0/923$  و جذر میانگین مربعات  $0/5867$  در مرحله آموزش و ضریب تبیین  $0/9025$  و جذر میانگین مربعات  $0/4936$  در مرحله آزمودن، به عنوان مدل برتر انتخاب شده است. جلالی و همکاران (۱۳۹۸) انتخاب مدل برتر سری زمانی جهت پیش بینی دبی ورودی رودخانه را مورد مطالعه قرار دادند که در آن آبدهی ماهانه حوضه آبریز جامیشان با استفاده از روش های تحلیل سری زمانی مورد پیش بینی قرار گرفت. در این راستا، از داده های ماهانه دبی ورودی به سد جامیشان در شهرستان سنقر استان کرمانشاه در بازه زمانی ۱۳۶۵-۱۳۹۴ استفاده گردید. آنالیز اولیه داده ها شامل بررسی ترم های قطعی سری (تناوب، روند و پرش) بر روی سری زمانی انجام و پس از اطمینان از حذف این ترم ها، داده ها نرمال و ایستایی داده ها مشخص شد، با توجه به معیار آکاییک، مدل  $(1,1,1)$  ARIMA به عنوان بهترین مدل انتخاب گردید. صمدی و همکاران (۱۳۹۸) بررسی کارایی مدل های ماشین بردار پشتیبان (SVM) و شبکه عصبی در پیش بینی دبی ورودی به سد گلستان را مطالعه کردند. بدین منظور از داده های هیدرومتری چهار ایستگاه قوچمز، پلکوسه، قره شور و اوغان و دو مدل شبکه عصبی و ماشین بردار پشتیبان جهت پیش بینی استفاده شد و نتایج مدل های مختلف مورد مقایسه قرار گرفت. با توجه به نتایج به دست آمده و با استفاده از معیارهای ارزیابی، در سه ایستگاه قره شور، پل کوسه و اوغان مدل ماشین بردار پشتیبان عملکرد بهتری را نسبت به مدل شبکه عصبی داشته است. و تنها در ایستگاه قوچمز عملکرد مدل شبکه عصبی بهتر بوده است. بر اساس نتایج این تحقیق، بهترین مدل برای پیش بینی دبی ماهانه ورودی به سد گلستان، مدل ماشین بردار پشتیبان می باشد. علی نژادی و همکاران (۱۳۹۹) روش های برنامه ریزی بیان ژن و رگرسیون های پارامتریک و ناپارامتریک در پیش بینی دبی میانگین روزانه رودخانه کارون (مطالعه موردی: ایستگاه هیدرومتری ملاثانی) را برای پیش بینی میانگین دبی روزانه رودخانه کارون در محل ایستگاه هیدرومتری ملاثانی طی دوره آماری (۱۳۹۶-۱۳۴۶) مورد مقایسه قرار دادند. ترکیب های مختلفی از داده های ثبت شده

<sup>1</sup> Gene Expression Programming

به‌عنوان الگوی ورودی برای پیش‌بینی دبی جریان استفاده شد. نتایج به‌دست‌آمده حاکی از عملکرد بهتر مدل برنامه‌ریزی بیان ژن با ضریب تبیین ( $R^2 = 0/827$ ) جذر میانگین مربعات خطا ( $RMSE = 59/45$ ) و میانگین خطای مطلق ( $MAE = 26/64$ ) در مرحله صحت‌سنجی برای پیش‌بینی دبی روزانه رودخانه کارون در ایستگاه ملاثانی با تأخیر ۵ روز، در مقایسه با روش‌های LR، NLR و NN-K بوده است. همچنین، ارزیابی عملکرد مدل‌ها در پیش‌بینی مقادیر حداکثر آبدهی جریان نشان داد که همه مدل‌ها میزان جریان را در بیشتر موارد کمتر از مقدار مشاهده‌ای تخمین زده بودند. رحیمی و حافظ‌پرست مودت (۱۳۹۹) مقایسه مدل‌های رگرسیون بردار پشتیبان، برنامه‌ریزی بیان ژن و آیهکرس را در پیش‌بینی تغییرات رواناب تحت تأثیر تغییر اقلیم (مطالعه موردی: سد جامیشان) مورد بررسی قرار دادند و مقادیر ماهانه بارش، دما و دبی سد جامیشان در سال‌های ۲۰۱۷-۱۹۸۸ را به‌عنوان دوره پایه در نظر گرفته‌اند. نتایج حاصل از پیش‌بینی دبی در هر سه مدل رگرسیون بردار پشتیبان، برنامه‌ریزی بیان ژن و آیهکرس حاکی از کاهش رواناب است که بیشترین کاهش رواناب مربوط به SVM در مدل  $ESM\_FLO$  با  $28/9$  درصد و کمترین کاهش رواناب مربوط به GEP در مدل  $CM5\_CNRM$  با  $14/1$  درصد می‌باشد و در انتها نتایج به‌دست‌آمده حاکی است که مدل‌های آیهکرس و بیان ژن نسبت به روش رگرسیون بردار پشتیبان از دقت مطلوب‌تری برخوردار هستند.

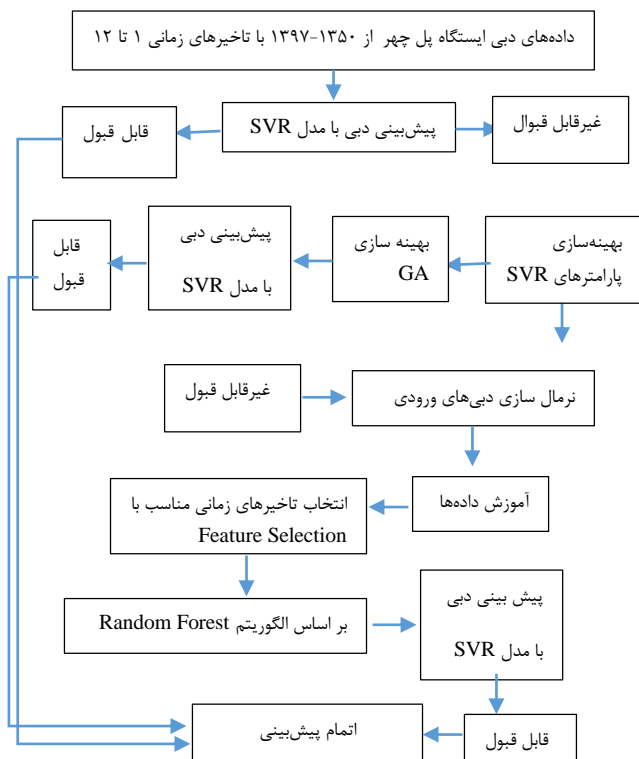
### روش پژوهش

در این پژوهش از داده‌های ماهانه دبی (۵۵۲ گام زمانی) در ایستگاه پل چهر برای پیش‌بینی استفاده شده است که مراحل انجام پژوهش حاضر در نمودار شکل ۱ مشخص شده است.

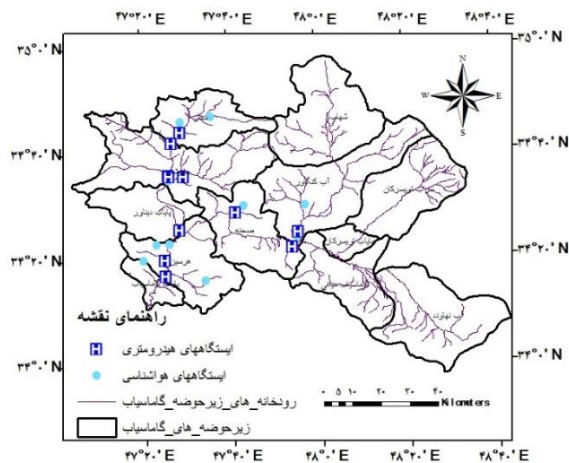
### منطقه مورد مطالعه

رودخانه گاماسیاب در بخش شمال شرقی حوزه آبریز کرخه جریان دارد. این رودخانه از سراب سنگ سوراخ و چشمه‌سارهای اطراف نهبوند سرچشمه می‌گیرد. در مسیر این رودخانه در رشته‌های فرعی مختلفی به آن ملحق می‌شود. در جنوب بیستون، شاخه فرعی دینور آب به آن پیوسته و به نام گاماسیاب ادامه مسیر می‌دهد و در نزدیکی فرامان به رودخانه قره‌سو متصل شده و پس از آن به نام سیمره وارد خاک لرستان می‌شود و سپس در استان خوزستان به نام کرخه جریان یافته و وارد هورالعظیم می‌شود. رودخانه گاماسیاب یکی از سرشاخه‌های مهم کرخه است و حوزه آبریز آن با مساحتی در حدود ۱۱۶۹۰ کیلومتر مربع بین طول جغرافیایی ۴۷ درجه و ۷ دقیقه تا ۴۹ درجه و ۱۰ دقیقه شرقی و عرض جغرافیایی ۳۳ درجه و ۴۸ دقیقه تا ۳۴ درجه و ۸۵ دقیقه شمالی واقع شده است. ارتفاع از سطح دریا در این حوزه بین ۱۲۴۲ تا ۳۵۸۰ متر می‌باشد (جبرانی و همکاران، ۲۰۱۱). شکل ۲ حوزه آبریز گاماسیاب را نشان می‌دهد.





شکل ۱. فلوچارت مراحل انجام پژوهش حاضر



شکل ۲. موقعیت حوضه آبریز گاماسیاب در جنوب غربی کشور، آبراهه‌ها و ایستگاه‌های هیدرومتری و هواشناسی منطقه

داده‌های دبی ایستگاه هیدرومتری پل چهر به مدت ۴۸ سال (شهریور ۱۳۹۷-۱۳۵۰ مهر) در دسترس است؛ برای مقایسه عملکرد مدل‌های ماشین بردار پشتیبان، برنامه‌ریزی بیان ژن در پیش‌بینی جریان رودخانه‌ها؛ ابتدا جریان روزانه مشاهده‌ای این حوضه در ایستگاه هیدرومتری پل چهر جهت واسنجی و اعتبارسنجی مدل‌ها به کار گرفته شد. برای این منظور، در ابتدا ۸۰ درصد از داده‌های جریان روزانه برای واسنجی مدل‌ها انتخاب شده و ۲۰ درصد داده‌ها جهت اعتبارسنجی مدل‌ها استفاده شد. معیارهای ضریب همبستگی، ریشه میانگین مربعات خطا، میانگین قدر مطلق خطا برای ارزیابی و نیز مقایسه عملکرد مدل‌ها در این پژوهش مورد استفاده قرار گرفته است.

### روش ماشین بردار پشتیبان (SVM)

- ماشین بردار پشتیبان یک طبقه‌بندی کننده دودویی غیر آماری است که در سال‌های اخیر بسیار مورد توجه قرار گرفته و به اختصار آن را بانام SVM معرفی می‌نمایند (ماترو و همکاران، ۲۰۰۵).
- SVM دارای خواص زیر است (احمدی، ۱۳۹۸):
۱. ایجاد حداکثر تعمیم در طراحی طبقه‌بندی کننده‌ها
  ۲. دارا بودن توانایی یافتن جواب بهینه کلی تابع
  ۳. تهیه خودکار ساختار و سازوکار بهینه در حل مسائل طبقه‌بندی
  ۴. مدل‌سازی توابع غیرخطی با استفاده از کرنل‌های غیرخطی و شیوه حاصل ضرب داخلی در فضاها ی هیلبرت

ماشین بردار پشتیبان الگوریتمی است که مدل‌های خطی منحصربه‌فرد را یافته و با استفاده از آن‌ها حداکثر حاشیه ابر صفحه را حاصل می‌کند. حداکثر شدن حاشیه ابر صفحه، حداکثر شدن تفکیک بین طبقات را حاصل نموده و موجب بهبود دقت در فرآیند مدل‌سازی می‌گردد. به نزدیک‌ترین نقاط آموزشی موجود در حاشیه ابر صفحه، بردارهای پشتیبان اطلاق گردیده و از آن‌ها برای مشخص کردن مرز بین طبقات استفاده می‌شود (شین و همکاران، ۲۰۰۵).

تفاوت اساسی این روش با طبقه‌بندی کننده‌های مرسوم آماری در قابلیت پردازش و طبقه‌بندی داده‌های ابر طیفی بوده و دیگر کاهش تعداد باندها در فرآیندهای مختلف مدل‌سازی نیاز نمی‌باشد. همچنین در این روش با استفاده از یک الگوریتم بهینه‌سازی تعداد نمونه‌های مورد نیاز برای تشکیل مرزهای کلاس‌ها به دست آمده و یک مرز تصمیم‌گیری خطی بهینه برای جداسازی کلاس‌ها محاسبه می‌شود. اگر داده‌ها به صورت خطی و جدا از هم باشند، SVM با استفاده از ماشین‌های خطی یک سطح بهینه را با کمترین خطا و حداکثر فاصله میان صفحه و نزدیک‌ترین نقاط آموزشی (بردارهای پشتیبان) تفکیک نموده و آموزش می‌دهد (شین و همکاران، ۲۰۰۵).

اگر نقاط آموزشی را به صورت  $[x_i, y_i]$  و بردار ورودی حائز شرط  $x_i \in \mathbb{R}^n$  باشد، در این صورت ارزش هر طبقه به صورت  $y_i \in \{-1, 1\}$   $i=1, \dots, i$  تعریف می‌گردد. آنگاه قواعد تصمیم‌گیری که تعریف می‌شود توسط یک صفحه بهینه که طبقات تصمیم‌گیری باینری را تفکیک می‌کند، به صورت رابطه زیر است:

$$Y = \text{sgn}(\sum_{i=1}^N y_i a_i (X \times X_j) + b) \quad (۱)$$

که در رابطه فوق  $Y$  خروجی رابطه،  $y_i$  ارزش طبقه نمونه  $X_i$  و  $a_i$  و  $b$  پارامترهای تعیین کننده ابر صفحه می باشند. اگر امکان تفکیک داده ها به صورت خطی فراهم نباشد، در این صورت رابطه (۲) به شرح زیر تغییر می یابد.

$$Y = \text{sgn}(\sum_{i=1}^N y_i a_i K(X \times X_j) + b) \quad (۲)$$

در این رابطه  $K(X \times X_j)$  ماشین های بردار پشتیبان باحالت های مختلف از سطوح تصمیم گیری غیرخطی در فضای داده ها، ضرب های داخلی تولید کرده و بدین منظور نیاز به تعریف معادله خط است. معادله خط در فضای دوبعدی با استفاده از رابطه (۳)، معادله صفحه از رابطه (۴) و معادله صفحه نمایش از رابطه (۵) محاسبه می گردد (چن و همکاران، ۲۰۰۲).

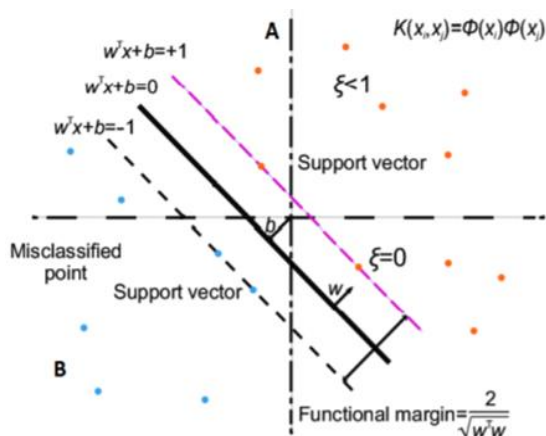
$$w_1 x_1 + w_2 x_2 + b = 0 \quad (۳)$$

$$w_1 x_1 + w_2 x_2 + w_2 x_2 + b = 0 \quad (۴)$$

$$\sum_{i=1}^n w_i x_i + b = 0 \rightarrow w^T x + b = 0, w = \begin{bmatrix} w_1 \\ \dots \\ w_i \end{bmatrix}, x = \begin{bmatrix} x_1 \\ \dots \\ x_i \end{bmatrix} \quad (۵)$$

در شکل (۳) خط پرنگ ممتد با معادله  $w^T x + b = 0$  به عنوان خط تفکیک کننده داده ها در صفحه بوده و آن ها را به دو دسته  $A$  و  $B$  تقسیم نموده است. این خط فضایی را تشکیل می دهد که در آن داده های متعلق به دسته  $A$  عدد مثبت و داده های متعلق به دسته  $B$  عدد منفی به خود بگیرند. اما در SVM علاوه بر استفاده از خط جداکننده از یک حاشیه اطمینان نیز برای دسته بندی استفاده می شود (شکل ۳). در این حالت هیچ یک از داده ها اجازه قرارگیری در ناحیه میانی را ندارند. با فرض اینکه خط با معادله  $w^T x + b = 0$  نقطه صفر مرزی است بنابراین برای داده ها بسته به موقعیت قرارگیری در کلاس های  $A$  و  $B$  به ترتیب معادلات  $w^T x + b > 1$  و  $w^T x + b < -1$  برقرار است. ضخامت تفکیک کننده در SVM یک ناحیه را شامل شده و سبب می گردد که فرآیند طبقه بندی در مقابل وجود ریسک برازش اشتباه مقاوم تر شود (فرودی خور و همکاران، ۲۰۱۷).

<sup>1</sup> Chen et al



شکل ۳. نمایش گرافیکی بردارهای پشتیبان

در مسائلی که داده‌ها به‌طور خطی تفکیک‌پذیر نباشند، می‌توان با نگاشت داده‌ها به یک فضای ویژگی، آن‌ها را به‌صورت خطی جدا پذیر نمود. در حالت کلی ابعاد این فضا بی‌نهایت بوده، بنابراین انجام محاسبات در فضای ویژگی می‌تواند پرهزینه باشد. برای غلبه بر این مشکل از توابع کرنل استفاده می‌شود، در نتیجه معادله تفکیک‌کننده برای حالت غیرخطی با دخالت تابع کرنل  $K(x)$  به‌صورت رابطه زیر است (چن و همکاران، ۲۰۰۲).

$$w^T x + b = 0 \rightarrow w^T k(x) + b = 0 \quad (۶)$$

که در رابطه فوق تابع  $K(x)$  با انجام نگاشت، داده‌ها را از فضای غیرخطی به یک فضای خطی منتقل می‌نماید. می‌توان این تابع را به‌صورت  $K(x_i, x_j)$  نیز تعریف نموده و به‌وسیله آن فضاها پیچیده‌تری را مورد بررسی قرارداد. به این فرآیند ترفند کرنل اطلاق می‌گردد. در مدل SVM انتخاب تابع کرنل بسیار مهم بوده و در مسائل مختلف با توجه به ماهیت مسئله می‌توان موارد متفاوتی را در نظر گرفت. بنابراین نمی‌توان تابعی را به‌طور قطعی به‌عنوان تابع مناسب برای SVM معرفی کرد. انواع مختلف تابع در جدول (۱) ارائه شده است. تابع کرنل خطی، حالت خاصی از تابع کرنل چندجمله‌ای بوده و در مسائل ساده‌تر کارایی خوبی از خود نشان می‌دهد ولی در مسائل پیچیده تابع کرنل چندجمله‌ای می‌تواند بسیار مفیدتر باشد. در تابع کرنل چندجمله‌ای پارامتر  $d$  باید از طرق آزمون و خطا یا از طریق الگوریتم‌های بهینه‌سازی محاسبه شود. تابع کرنل گوسی یا شعاعی (RBF) نوع دیگری از توابع کرنل بوده و معمولاً در مسائلی به‌کاربرده می‌شود که هیچ‌گونه شناختی از نوع و ماهیت داده‌ها در دسترس نباشد. به‌طور کلی در توابع کرنل می‌باید مقادیر بهینه ضرایب  $c$ ،  $\epsilon$ ،  $d$  و  $\gamma$  برآورد گردد (احمدی، ۱۳۹۸).

جدول ۱. توابع کرنل رایج در ماشین‌های بردار پشتیبان

نوع تابع	تابع کرنل
خطی	$k(X_i, X_j) = X_i^T \times X_j$
جند جمله‌ای	$k(X_i, X_j) = (\gamma X_i^T \times X_j)^d$
تانزانته هیپربولیک	$k(X_i, X_j) = \tanh(\gamma X_i^T \times X_j)$
RBF	$k(X_i, X_j) = -\exp(-\gamma  X_i - X_j ^2)$

## الگوریتم ژنتیک

علم ژنتیک، علمی است که در مورد چگونگی توارث و انتقال صفات بیولوژیکی از نسلی به نسل دیگر صحبت به عمل می‌آورد. عامل اصلی صفات بیولوژیکی در موجودات زنده کروموزوم‌ها و ژن‌ها است و نحوه عملکرد آن‌ها به گونه‌ای است که در نهایت ژن‌ها و کروموزوم‌های برتر و قوی‌تر باقی‌مانده و ژن‌های ضعیف‌تر از بین می‌روند. (کزا و همکاران، ۱۹۹۲).

الگوریتم ژنتیک (GA) برای اولین بار توسط هلند<sup>۱</sup> (۱۹۷۵) استفاده و به‌عنوان ابزار قوی بهینه‌سازی توسعه پیدا کرد. GA یک الگوریتم جستجو برگرفته از طبیعت بیولوژیکی و فرآیند انتخاب طبیعی می‌باشد. اساس این روش مبتنی بر نظریه داروین که در محیط متغیر همواره موجوداتی ادامه حیات دارند که از همه پایدارتر ترند، می‌باشد. این الگوریتم از مجموعه‌ای از جواب‌های تصادفی اولیه به نام جمعیت آغاز می‌گردد. هر جمعیت از مجموعه کروموزوم‌ها که هر کدام یک جواب مسئله هستند، تشکیل می‌شود و هر کروموزوم مجموعه‌ای از ژن‌ها و در واقع متغیرهای تصمیم مسئله می‌باشد.

به‌طور کلی در چرخه الگوریتم ژنتیک ابتدا یک جمعیت اولیه از افراد بدون در نظر گرفتن معیار خاصی و به‌طور تصادفی انتخاب می‌شود. برای تمامی کروموزوم‌های نسل صفر مقدار برازش با توجه به تابع هدف تعیین می‌شود. سپس با مکانیسم‌های مختلف تعریف‌شده برای عملگر انتخاب زیرمجموعه‌ای از جمعیت اولیه انتخاب خواهد شد. روی این افراد انتخاب‌شده عملیات برش و جهش در صورت لزوم با توجه به‌صورت مسئله اعمال خواهد شد. حال باید این افراد مکانیسم الگوریتم‌های ژنتیک در موردشان اعمال شده است، با افراد جمعیت اولیه (نسل صفر) از نظر مقدار برازش مقایسه شوند. قطعه انتظار می‌رود که افراد نسل اول با توجه به اعمال الگوریتم‌های ژنتیک روی آن‌ها از شایستگی بیشتری برخوردار باشند. اما لزوماً این چنین نخواهد بود. به‌رحال افرادی باقی خواهند ماند که بیشترین برازش را داشته باشند. چنین افرادی در مقام یک مجموعه به‌عنوان جمعیت اولیه برای مرحله بعدی الگوریتم عمل خواهد کرد. (گلدبرگ، ۲۰۰۲).

هر مرحله تکرار الگوریتم یک نسل جدید را ایجاد می‌کند که با توجه به اصلاحاتی که در آن صورت پذیرفته است، رو به‌سوی تکامل خواهد داشت. لازم به ذکر است که در استفاده از الگوریتم ژنتیک می‌توان برای انتخاب نسل، نوع ترکیب کروموزوم‌ها، چگونگی جهش و مواردی از این قبیل روش‌های متنوع و متفاوتی به کار برد.

الگوریتم ژنتیک مانند سایر روش‌های بهینه‌سازی با معرفی پارامترها، تابع هدف، قیودات، تابع هزینه و هزینه آغاز می‌شود و هم‌چنین خاتمه‌ی آن نیز مانند سایر روش‌های بهینه‌سازی با کنترل همگرایی موردنظر هر مسئله موردنظر است. در این میان اگرچه الگوریتم ژنتیک از نقطه‌نظر آغاز و خاتمه مانند سایر روش‌هاست اما از لحاظ عملکرد داخلی بهینه‌سازی نسبت به سایر روش‌ها بسیار متفاوت است. به‌طور کلی مسیر بهینه‌سازی به روش الگوریتم ژنتیک در شکل زیر نشان داده شده است.

<sup>1</sup> Koza et al

<sup>2</sup> John Holland

<sup>3</sup> Goldberg

## پیش‌پردازش داده‌ها

عملیاتی که روی داده‌ها قبل از ورود به مدل‌ها انجام می‌شود پیش‌پردازش نام دارد. روش‌های مختلفی برای از بین بردن پرش، تناوب و عدم ایستایی در داده‌ها وجود دارد. نرمال‌سازی داده‌ها به منظور هماهنگ‌تر شدن بهتر داده‌ها و افزایش سرعت اجرای مدل‌ها یکی از روش‌های رایج در مدل‌سازی است که در این روش داده‌ها با استفاده از فرمول زیر استاندارد شده و در ادامه محاسبات در نظر گرفته شده است.

$$Z_{i,T} = \left( \frac{X_{i,T} - \bar{X}_T}{S_T} \right) \quad (7)$$

در این رابطه  $Z_{i,T}$  و  $X_{i,T}$  مقادیر استاندارد شده و اولیه داده‌ها در سال  $T$  و در  $i$  و  $\bar{X}_T$  و  $S_T$  به ترتیب میانگین و انحراف معیار ماهانه در طول سال‌های مختلف می‌باشد.

## روش انتخاب ویژگی<sup>۱</sup>

انتخاب ویژگی را می‌توان به عنوان فرآیند شناسایی ویژگی‌های مرتبط و حذف ویژگی‌های غیر مرتبط و تکراری باهدف مشاهده زیرمجموعه‌ای از ویژگی‌ها که مسئله را به خوبی و با حداقل کاهش درجه کارایی تشریح می‌کنند تعریف کرد. روش‌های انتخاب ویژگی به منظور مواجهه با داده‌های ابعاد بالا، به موافقه‌های جدایی‌ناپذیر از فرآیند یادگیری مبدل شده‌اند. یک انتخاب ویژگی صحیح می‌تواند منجر به بهبود یادگیرنده استقرایی از جهت‌های گوناگون از جمله سرعت یادگیری، ظرفیت تعمیم و سادگی مدل استنتاج شده شود.

## الگوریتم جنگل تصادفی<sup>۲</sup>

جنگل تصادفی یک کلاس از روش‌های دسته‌بندی گروهی می‌باشد که مخصوص درخت تصمیم طراحی گردیده است. روش پوششی مبتنی بر درخت‌های تصادفی به طور گسترده‌ای برای مشخص کردن ویژگی‌های بارز استفاده می‌شود که نقشی اساسی در کیفیت پیشگویی دارد. اهمیت کل ویژگی‌ها بر اساس دقت دسته‌بندی محاسبه می‌شود که این کار در زمانی که تمامی ویژگی‌ها در درخت به صورت تصادفی پخش باشند، انجام می‌شود. یک جنگل جدید به صورت تکرار شونده به وسیله حذف ویژگی‌هایی که در عملکرد بهتر دسته‌بندی نقشی ندارند ساخته می‌شود. این حلقه تکرار تا زمانی که ویژگی برای حذف باقی نمانده باشد ادامه می‌یابد. در نهایت بهترین مجموعه از ویژگی‌ها که کمترین خطا را تولید می‌کند، از جنگل استخراج می‌شود. مدل جنگل‌های تصادفی در مسائل مختلف به صورت گسترده استفاده شده است و این به علت دقت بالای پیشگویی این روش است (بریمان، ۲۰۰۶).

## معیارهای ارزیابی مدل

جهت انتخاب مدل برتر برای انجام پیش‌بینی از معیارهای ارزیابی زیر استفاده شده است:

واریانس تبیین شده (Exp\_Var)، ناش-ساتکلیف (NSE) و ریشه میانگین مربعات خطا (RMSE) بین مقادیر مشاهده شده و مدل شده به صورت زیر محاسبه گردید:

<sup>1</sup> Feature Selection

<sup>2</sup> Random Forest

<sup>3</sup> Breiman

$$\text{Exp.Var}(Q, \hat{Q}) = 1 - \frac{\text{Var}(Q_i - \hat{Q}_i)}{\text{var}(Q_i)} \quad (۸)$$

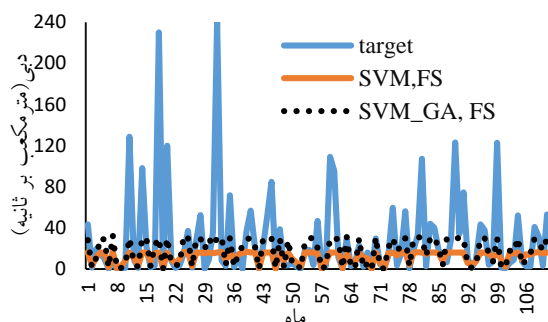
$$\text{NSE}(Q, \hat{Q}) = 1 - \frac{\sum_{i=1}^n (Q_i - \hat{Q}_i)^2}{\sum_{i=1}^n (Q_i - \bar{Q}_i)^2} \quad (۹)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (Q_i - \hat{Q}_i)^2}{n}} \quad (۱۰)$$

که در آن  $Q_i$  مقادیر مشاهده شده،  $\hat{Q}_i$  مقادیر محاسبه شده،  $\bar{Q}_i$  میانگین مقادیر مشاهده شده است.

### یافته‌ها

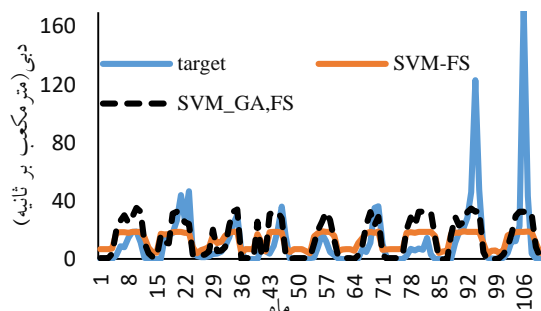
در این مطالعه ابتدا داده‌ها را از یک تا دوازده دوره تأخیر زمانی صورت گرفت و ۱۲ بعد برای پیش‌بینی ترسیم گردید. سپس دو سناریو برای پیش‌بینی سری زمانی دبی، با و بدون اعمال پیش‌پردازش (استانداردسازی) در نظر گرفته شد. در هر سناریو نیز در نظر گرفتن داده‌های آموزش و آزمودن با دو رویکرد سری زمانی یا غیر سری زمانی انجام شد. بدین معنی که توالی داده‌های ماهانه در رویکرد مدنظر قرار گرفته اما در رویکرد غیر سری زمانی این توالی در فرآیند پیش‌بینی مدنظر قرار نگرفته است. در هر رویکرد نیز دو روش با و بدون انتخاب ویژگی مدل بررسی شده است. نهایتاً برای حالت‌های گفته شده از مدل SVR ساده و بهینه شده با الگوریتم ژنتیک در پیش‌بینی استفاده شده است. در همه حالت‌ها به ترتیب ۸۰ و ۲۰ درصد داده‌ها برای آموزش و آزمودن مدل در نظر گرفته شده است. کلیه فرآیند کد نویسی به زبان برنامه‌نویسی پایتون صورت گرفته و از کتابخانه‌های ویژه اس کی لرن<sup>۱</sup> و اپتیمال<sup>۲</sup> نیز استفاده شده است. پیش‌بینی دبی‌های جریان ماهانه در سناریوی اول - بدون پیش‌پردازش (بدون اعمال استانداردسازی داده‌ها) در دو رویکرد سری زمانی و غیر سری زمانی و دو حالت انتخاب ویژگی و بدون انتخاب ویژگی انجام شد. نتایج نشان داد که در سناریوی اول در هر دو رویکرد و برای حالتی که انتخاب ویژگی مدل صورت نگیرد و کل ۱۲ بعد (تأخیرهای ۱ تا ۱۲) مدنظر باشد نتایج مدل SVR ساده و بهینه شده فقط به یک عدد ثابت معطوف شده و دقت لازم برای پیش‌بینی را ندارد. اما در سناریوی اول و با انتخاب ویژگی مدل که تأخیر زمانی ۱ را با حدود ۹۲ درصد خصوصیات از ۱۲ بعد در نظر گرفت، نتایج بهتری حاصل گردید. همان‌طور که در اشکال ۴ و ۵ نشان داده شده نتایج ظاهری و محاسبه معیارهای خطا نشان داد در این حالت (انتخاب ویژگی مدل) نتایج بهتری نسبت به حالت اول حاصل شده اما به نتایج پیش‌بینی‌ها اطمینانی وجود ندارد. و همچنان نتایج بسیار ضعیفی حاصل شده است.



شکل ۴. نتایج مدل پیش‌بینی در سناریو اول- رویکرد غیر سری زمانی، حالت انتخاب ویژگی مدل

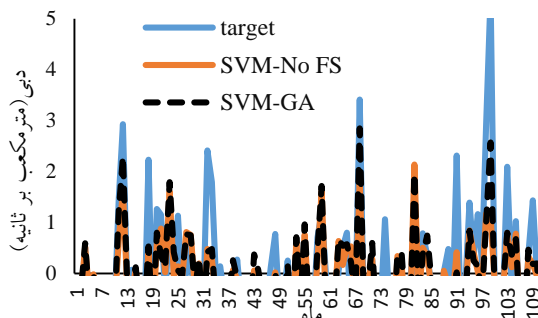
<sup>۱</sup> Sklearn

<sup>۲</sup> Optimal



شکل ۵. نتایج مدل پیش‌بینی در سناریو اول- رویکرد سری زمانی، حالت انتخاب ویژگی مدل

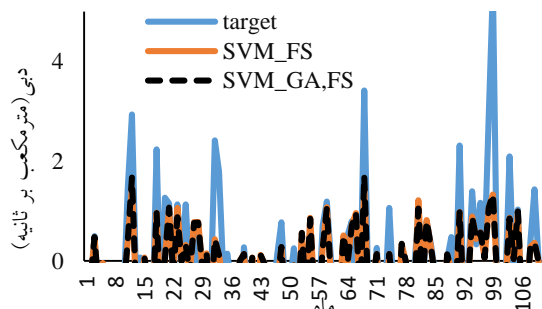
دقت نتایج در رویکرد سری زمانی نسبت به رویکرد غیر سری زمانی در حالت‌های مشابه نتایج بهتری داشته است. هرچند در شرایط استفاده از بهینه‌ساز الگوریتم ژنتیک در بهینه‌سازی پارامترهای مدل SVR نتایج بهتری حاصل شده است اما در دبی‌های بیک ماه‌های ۹۴ و ۱۰۶ در دوره آزمون مدل دقت پایینی داشته است. نتایج سناریوی دوم با اعمال پیش‌پردازش با عمل استانداردسازی در اشکال ۶ تا ۹ ارائه شده است. نتایج سناریو دو با رویکرد غیر سری زمانی و در حالت عدم اعمال انتخاب ویژگی مدل و منظور نمودن کلیه ۱۲ بعد به‌عنوان ورودی مدل در شکل ۶ نشان داده شده است.



شکل ۶. نتایج مدل پیش‌بینی در سناریو دوم- رویکرد غیر سری زمانی، حالت عدم انتخاب ویژگی مدل

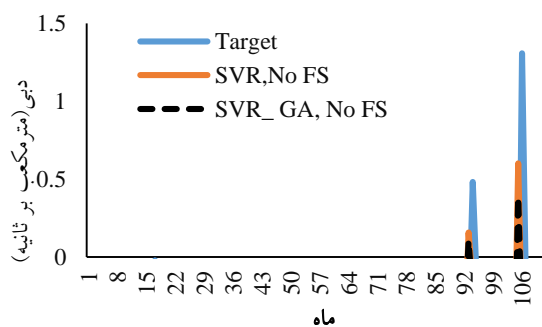
همچنین نتایج سناریو دو با رویکرد غیر سری زمانی در حالت اعمال انتخاب ویژگی مدل که منجر به تأخیرهای ۱ و ۱۲ با حدود ۹۲ درصد خصوصیات ۱۲ بعد گردید، در شکل ۷ ارائه شده است.



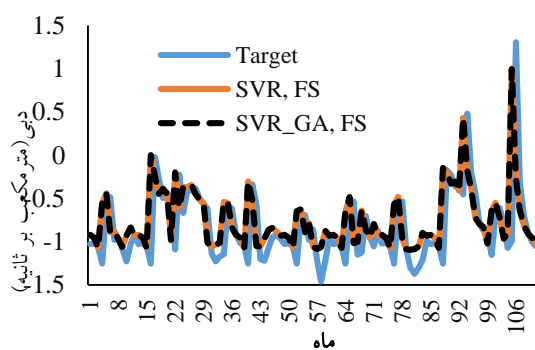


شکل ۷. نتایج مدل پیش‌بینی در سناریو دوم- رویکرد غیر سری زمانی، حالت انتخاب ویژگی مدل

همان‌طور که از اشکال ۶ و ۷ ملاحظه می‌گردد، نتایج پیش‌بینی با اعمال استانداردسازی بر داده‌ها در شرایط مشابه سناریو یک به‌صورت معنی‌داری بهبود یافته است. اثر بهینه‌ساز الگوریتم ژنتیک بر دقت مدل نسبت به SVR ساده در حالت عدم انتخاب ویژگی مدل جهت پیش‌بینی دبی سیلاب بیشتر بوده است. همچنین نتایج مدل در سناریوی دوم رویکرد سری زمانی در دو حالت بدون و با اعمال انتخاب ویژگی مدل در اشکال ۸ و ۹ ارائه شده است.



شکل ۸. نتایج مدل پیش‌بینی در سناریو دوم- رویکرد سری زمانی، حالت عدم انتخاب ویژگی مدل



شکل ۹. نتایج مدل پیش‌بینی در سناریو دوم- رویکرد سری زمانی، حالت انتخاب ویژگی مدل

همان‌طور که از اشکال ۸ و ۹ مشاهده می‌شود بهترین نتایج به‌دست‌آمده در بین همه موارد ذکرشده مربوط به سناریوی دوم و رویکرد سری زمانی است که انتخاب ویژگی مدل اعمال نشده و کلیه ۱۲ بعد در ورودی مدل SVR در نظر گرفته شده است. در این شرایط تفاوت محسوسی در نتایج مدل ساده SVR و بهینه‌شده با الگوریتم ژنتیک ندارد. برای تحلیل بیشتر در جداول ۲ و ۳ شاخص‌های ارزیابی دقت مدل در شرایط آموزش و آزمون مدل SVR در دو سناریوی مطرح‌شده، رویکردهای گوناگون و حالت‌های مختلف ارائه شده است. همان‌طور که از جدول ۲ مشاهده می‌شود، شاخص‌های ارزیابی مدل در دو حالت آموزش و آزمون مدل در حالت بدون انتخاب ویژگی مدل در هر دو رویکرد سری زمانی و غیر سری زمانی نتایج بسیار ضعیفی برای هر دو مدل SVR ساده و بهینه‌شده با الگوریتم ژنتیک حاصل شده است. از طرفی بهترین نتیجه در این سناریو در حالت انتخاب ویژگی مدل در رویکرد سری زمانی با مدل بهینه‌شده SVR توسط الگوریتم ژنتیک اتفاق افتاده و مقادیر پیش‌بینی مدل در محدوده آزمون برای دبی‌های حداقل و میانگین از دقت خوبی برخوردار بوده اما توانایی لازم در پیش‌بینی دبی‌های پیک را نداشته است.

جدول ۲. نتایج ضرایب ارزیابی مدل در سناریوی اول (بدون استانداردسازی) در رویکرد و حالت‌های گوناگون

آموزش			تست			حالت		مدل SVR	انتخاب	رویکرد
Exp.Var	NSE	RMSE	Exp.Var	NSE	RMSE	C	Epsilon			
۰/۰۲۹	-۰/۱۳	۴۶/۱۱	-۶/۷E-۰۷	-۰/۰۴۶	۲۲/۸۱	۱	۰/۰۱	ساده	بدون	با
۰/۰۶۳	-۰/۰۹	۴۵/۳۲	-۱/۴۷E-۰۶	-۰/۰۵۱	۲۲/۸۶	۲/۱۸۱	۰/۰۰۶۳	الگوریتم ژنتیک	بدون	با
۰/۱۱۵	-۰/۰۵۷	۴۴/۶۱	۰/۱۶	۰/۱۵۶	۲۰/۵	۱	۰/۰۱	ساده	با	با
۰/۴۷۹	۰/۴۲	۳۳/۰۲	۰/۲۴۳	۰/۲۲۲	۱۹/۶۷	۱۵/۳۱	۰/۰۰۶۳	الگوریتم ژنتیک	با	با
۰/۰۳	-۰/۱۴۸	۴۳/۶۷	-۱/۴۵E-۰۶	-۰/۱۴۷	۴۵/۰۱۸	۱	۰/۰۱	ساده	بدون	بدون
۰/۰۸	-۰/۰۹۸	۴۲/۷	-۴E-۰۶	-۰/۱۴۷	۴۵/۰۱	۲/۷۵۵	۰/۰۰۹۷	الگوریتم ژنتیک	بدون	بدون
۰/۱۲۱	-۰/۰۴۹	۴۱/۷۳	۰/۰۹۶	-۰/۰۵۳	۴۳/۱۲	۱	۰/۰۱	ساده	با	بدون
۰/۵۰۵	۰/۴۵۲	۳۰/۱۷	۰/۲۳۴	۰/۱۷۴	۳۸/۲۱	۱۶/۹۷	۰/۰۰۱۴	الگوریتم ژنتیک	با	بدون

جدول ۳. نتایج ضرایب ارزیابی مدل در سناریوی دوم (با استانداردسازی) در رویکرد و حالت های گوناگون

آموزش			تست							
Exp.Var	NSE	RMSE	Exp.Var	NSE	RMSE	C	Epsilon	مدل SVR	حالت	
سری										
انتخاب										
ویژگی										
۰/۶۶۸	۰/۶۶	۰/۵۸۳	۰/۱۵۵	۰/۱۰۵	۰/۳۷۶	۱	۰/۰۱	ساده	بدون	با
۰/۷۸۷	۰/۷۸۴	۰/۴۶۵	۰/۱۷۷	۰/۱۲۶	۰/۳۷۲	۳/۹۶۵	۰/۰۰۲۸	الگوریتم ژنتیک	بدون	با
۰/۵۵	۰/۵۳۹	۰/۶۷۸	۰/۰۱۹	-۰/۰۳۲	۰/۴۰۴	۱	۰/۰۱	ساده	با	با
۰/۵۵۲	۰/۵۴۱	۰/۶۷۷	۰/۰۱۹	-۰/۰۳۲	۰/۴۰۴	۱/۰۴	۰/۰۰۶۳	الگوریتم ژنتیک	با	با
۰/۷۰۴	۰/۶۹۸	۰/۵۳۳	۰/۵۵۶	۰/۵۴	۰/۷۷	۱	۰/۰۱	ساده	بدون	بدون
۰/۸۵۵	۰/۸۵۴	۰/۳۷۱	۰/۶۱۳	۰/۶۰۴	۰/۷۱۴	۸/۸۹	۰/۰۰۹	الگوریتم ژنتیک	بدون	بدون
۰/۵۷۶	۰/۵۶۷	۰/۶۳۸	۰/۵۹۹	۰/۵۸	۰/۷۳۶	۱	۰/۰۱	ساده	با	بدون
۰/۵۹۹	۰/۵۸۸	۰/۶۲۳	۰/۶۰۹	۰/۵۹	۰/۷۲۶	۷/۳۱	۰/۰۰۱۳	الگوریتم ژنتیک	با	بدون

همچنین با توجه به نتایج جدول ۳ مشاهده می گردد که در سناریو ۲ با عمل استانداردسازی به طرز معنی داری نتایج مدل پیش بینی SVR در هر شرایطی نسبت به سناریوی ۱ دقت بالاتری یافته است. در شرایطی که داده ها استاندارد شوند نتایج رویکرد غیر سری زمانی در حالت های مشابه نسبت به سری زمانی بهتر بوده بطوریکه مقادیر NSE و واریانس تبیین محاسبه شده در مرحله آزمون مدل بالای ۵۰ درصد بوده است. در مقایسه حالت انتخاب ویژگی مدل نتایج هر دو مدل SVR ساده و بهینه شده نسبت به عدم انتخاب ویژگی مدل تفاوت معنی داری نسبت به حالت بدون اعمال انتخاب ویژگی مدل نداشته است. این مسئله در خصوص نتایج مدل بهینه شده نسبت به مدل ساده SVR نیز صادق است. در نهایت بهترین نتیجه در این سناریو مربوط به رویکرد غیر سری زمانی و با انتخاب ویژگی مدل توسط مدل SVR بهینه شده با الگوریتم ژنتیک است.

### بحث

در این مطالعه به پیش بینی داده های دبی ماهانه مربوط به ۴۶ سال آمار ایستگاه پل چهر واقع بر رودخانه گاماسیاب با مدل SVR ساده و بهینه شده با الگوریتم ژنتیک پرداخته شد. برای این منظور ابتدا به داده ها از یک تا دوازده دوره تأخیر زمانی داده شد و بدین ترتیب ۱۲ بعد به عنوان ورودی مدل پیش بینی حاصل گردید. به منظور بررسی مقایسه اثر استانداردسازی و تأثیر مدل بهینه ساز الگوریتم ژنتیک بر دقت نتایج مدل SVR حالت های مختلفی مورد ارزیابی قرار گرفت. دو سناریوی اصلی با و بدون اعمال پیش پردازش (استانداردسازی) در نظر گرفته شد. جهت بررسی اثر توالی داده ها بر نتایج پیش بینی در مراحل آموزش و آزمون مدل SVR در هر سناریو نیز دو رویکرد سری زمانی یا غیر سری زمانی در نظر گرفته شد. همچنین به منظور بررسی اثر ابعاد ورودی مدل در دقت پیش بینی مدل نیز در هر رویکرد نیز دو حالت با و بدون انتخاب ویژگی مدل با استفاده از الگوریتم جنگل تصادفی بررسی شده است. در همه حالت ها به ترتیب ۸۰ و ۲۰ درصد داده ها برای آموزش و آزمون مدل در نظر گرفته شده

است. کلیه فرآیند کد نویسی به زبان برنامه‌نویسی پایتون صورت گرفته است. نتایج نشان داد که در سناریوی اول، بدون اعمال پیش‌پردازش (استانداردسازی داده‌ها) در دو رویکرد سری زمانی و غیر سری زمانی در حالتی که انتخاب ویژگی مدل صورت نگیرد و کل ۱۲ بعد (تأخیرهای ۱ تا ۱۲) مدنظر باشد نتایج مدل SVR ساده و بهینه‌شده فقط به یک عدد ثابت نزدیک شده و نوسانات دبی در ماه‌های مختلف را برآورد نکرده لذا دقت لازم برای پیش‌بینی را ندارد. اما با انتخاب ویژگی مدل توسط الگوریتم جنگل تصادفی ابعاد از ۱۲ به ۱ که مربوط به تأخیر زمانی ۱ است کاهش یافته و نوسانات دبی ماهانه پیش‌بینی می‌گردد اما همچنان دقت نتایج ضعیف است. اما در این سناریو دقت نتایج در رویکرد سری زمانی نسبت به رویکرد غیر سری زمانی در حالت‌های مشابه نتایج بهتری داشته است. همچنین اثر بهینه‌ساز الگوریتم ژنتیک در بهینه‌سازی پارامترهای مدل SVR در دقت نتایج معنی‌دار شده است ولی همچنان در دبی‌های پیک ماه‌های ۹۴ و ۱۰۶ در دوره آزمون مدل دقت پایینی داشته است.

در سناریوی ۲، با اعمال استانداردسازی بر داده‌ها در شرایط مشابه سناریو یک نتایج به صورت معنی‌داری بهبود یافته است. اثر بهینه‌ساز الگوریتم ژنتیک بر دقت مدل پیش‌بینی نسبت به SVR ساده در حالت عدم انتخاب ویژگی مدل که توسط الگوریتم جنگل تصادفی دو بعد با تأخیرهای ۱ و ۱۲ محاسبه شده‌اند، در پیش‌بینی دبی سیلاب بیشتر بوده است. نهایتاً بهترین نتایج به دست آمده در بین همه موارد ذکر شده مربوط به سناریوی دوم و رویکرد سری زمانی است که انتخاب ویژگی مدل اعمال نشده و کلیه ۱۲ بعد در ورودی مدل SVR در نظر گرفته شده است. در این شرایط تفاوت محسوس در نتایج مدل ساده SVR و بهینه‌شده با الگوریتم ژنتیک ندارد.

### نتیجه‌گیری

به‌عنوان نتیجه‌گیری کلی چنانچه عمل استانداردسازی بر داده‌ها صورت نگیرد منظور نمودن رویکرد سری زمانی (توالی داده‌ها مدنظر باشد) و کاهش ابعاد ورودی مدل با الگوریتم جنگل تصادفی منجر به نتایج بهتری در مدل پیش‌بینی SVR خواهد شد و استفاده از بهینه‌ساز الگوریتم ژنتیک نسبت به مدل ساده آن تأثیر معنی‌داری بر بهبود نتایج خواهد داشت. در مقابل با انجام عمل استانداردسازی بهترین جواب در رویکرد غیر سری زمانی که توالی داده‌ها در انتخاب دوره‌های آموزش و آزمون مدنظر نیست و حالت کاهش ابعاد ورودی مدل حاصل گردید. در این شرایط مدل بهینه‌ساز الگوریتم ژنتیک تأثیر معنی‌داری بر دقت نتایج پیش‌بینی نسبت به مدل ساده SVR نداشته است.

### تقدیر و تشکر

نویسندگان بر خود لازم می‌دانند مراتب تشکر صمیمانه خود را از شرکت سهامی آب منطقه‌ای کرمانشاه برای در اختیار گذاشتن اطلاعات موردنیاز این پژوهش اعلام کنند.

### منابع

- احمدی، فرشاد. (۱۳۹۹). ارزیابی عملکرد روش‌های ماشین‌بردار پشتیبان و سیستم استنتاج عصبی فازی تطبیقی در پیش‌بینی جریان ماهانه رودخانه‌ها (مطالعه موردی رودخانه‌های نازلو و سزار). تحقیقات آب و خاک ایران، (۳)۵۱، ۶۸۶-۶۷۳.  
<https://dx.doi.org/10.22059/ijswr.2019.290994.668356>
- احمدی، فرشاد، رادمنش، فریدون، و میرعباسی نجف‌آبادی، رسول. (۱۳۹۳). مقایسه روش‌های برنامه‌ریزی ژنتیک و ماشین‌بردار پشتیبان در پیش‌بینی جریان روزانه رودخانه (مطالعه موردی: رودخانه باراندوزچای). نشریه آب‌و خاک (علوم و صنایع کشاورزی)، (۶)۲۸، ۱۱۷۱-۱۱۶۲.  
<https://dx.doi.org/10.22067/jsw.v0i0.32406>
- جلالی، محمد نبی، قزلباش نادری، بابک، گویلی، افروز، و بهرامی، جمیل. (۱۳۹۸). انتخاب مدل برتر سری زمانی جهت پیش‌بینی دبی ورودی رودخانه. سومین کنفرانس ملی هیدرولوژی مناطق نیمه خشک با محوریت آب، انسان، طبیعت، سندج. <https://civilica.com/doc/903801>

- جیرانی، فرزین، مرید، سعید. و مریدی، علی. (۱۳۹۰). اثر دقت مکانی نقشه رقومی ارتفاع در واسنجی و برآورد رواناب و رسوب با استفاده از مدل CUP-SWAT. مجله پژوهش‌های حفاظت آب‌و خاک، ۱۸(۴)، ۱۰۲-۸۱.  
<https://dorl.net/dor/20.1001.1.23222069.1390.18.4.5.5>
- حشمتی، سارا، و حافظ پرست مودت، مریم. (۱۳۹۷). پیش‌بینی دبی با استفاده از تحلیل سری‌های زمانی با مدل SARIMA در شرایط خشک‌سالی مطالعه موردی: ایستگاه دبی سنجی پیر سلمان در رودخانه جامیشان. مجله علمی سامانه‌های سطوح آبیگیر باران، ۶(۱)، ۸۲-۷۳.  
<https://dorl.net/dor/20.1001.1.24235970.1397.6.1.7.8>
- دهقانی، رضا، یونسی، حجت ...، و ترابی پوده، حسن. (۱۳۹۶). مقایسه عملکرد مدل‌های ماشین بردار پشتیبان، برنامه‌ریزی بیان ژن و شبکه بیزین در پیش‌بینی جریان رودخانه‌ها (مطالعه موردی: رودخانه کشکان). نشریه پژوهش‌های حفاظت آب‌و خاک، ۲۴(۴).  
<https://dx.doi.org/10.22069/jwfst.2017.12398.2701>
- رحیمی، بنفشه، و حافظ پرست مودت، مریم. (۱۳۹۹). مقایسه مدل‌های رگرسیون بردار پشتیبان، برنامه‌ریزی بیان ژن و آیه‌کرس در پیش‌بینی تغییرات رواناب تحت تأثیر تغییر اقلیم (مطالعه موردی: سد جامیشان). مجله تحقیقات آب‌و خاک ایران، ۵۱(۱۰).  
<https://dx.doi.org/10.22059/ijswr.2020.303779.668640>
- زارع‌زاده مهریزی، محبوبه، و بزرگ حداد، امید. (۱۳۹۸). شبیه‌سازی و پیش‌بینی آبدی با استفاده از الگوریتم ترکیبی ANN-GA. مجله آب و خاک (علوم و صنایع کشاورزی)، ۲۴(۵)، ۹۵۴-۹۴۲.  
<https://dx.doi.org/10.22067/jsw.v0i0.5295>
- زینعلی، معصومه، گلایبی، محمدرضا، شریفی، محمدرضا، و حافظ پرست مودت، مریم. (۱۳۹۷). ارزیابی مدل‌های هوش مصنوعی در مدل‌سازی جریان رودخانه، مطالعه موردی: رودخانه گاماسیاب. نشریه علمی-پژوهشی مهندسی و مدیریت آبخیز، ۱۱(۴)، ۹۵۴-۹۴۱.  
<https://dx.doi.org/10.22092/ijwmse.2018.115870.1370>
- سیدیان، سید مرتضی، سلیمانی، مریم، و کاشانی، مجتبی. (۱۳۹۳). پیش‌بینی دبی جریان رودخانه با استفاده از داده‌کاوی و سری زمانی. نشریه آکو هیدرولوژی، ۱(۳)، ۱۶۷-۱۷۹.  
<https://dx.doi.org/10.22059/ije.2014.54219>
- صمدی، میثم، بهره‌مند، عبدالرضا، و فتح‌آبادی، ابوالحسن. (۱۳۹۷). پیش‌بینی دبی ماهانه ورودی به سد بوستان در استان گلستان با استفاده از مدل‌های داده‌کاوی و ترکیبی. نشریه علمی-پژوهشی مهندسی و مدیریت آبخیز، ۱۱(۴)، ۱۰۵۸-۱۰۴۴.  
<https://dx.doi.org/10.22092/ijwmse.2018.108926.1251>
- صمدی، میثم، مسلم، امیر پیام، بیت‌اله پور چهارمحالی، ایمان، کوشافر، لیلا، و فرمهمینی فراهانی، محمد ایمان. (۱۳۹۸). بررسی کارایی مدل‌های ماشین بردار پشتیبان (SVM) و شبکه عصبی در پیش‌بینی دبی ورودی به سد گلستان. نهمین کنفرانس علمی پژوهشی آبخیزداری و مدیریت منابع آب و خاک، کرمان.  
<https://civilica.com/doc/1007309>
- صمدی، میثم، و فتح‌آبادی، ابوالحسن. (۱۳۹۷). استفاده از مدل‌های سری زمانی، شبکه عصبی و ماشین بردار پشتیبان جهت پیش‌بینی دبی ورودی به سد گرگان. مجله محیط‌زیست و مهندسی آب، ۴(۴)، ۳۰۹-۲۹۹.  
<https://dx.doi.org/10.22034/jewe.2018.128256.1256>
- علی‌نژادی، مهدی، موسوی، سید فرهاد، و حسینی، خسرو. (۱۳۹۹). مقایسه روش‌های برنامه‌ریزی بیان ژن و رگرسیون‌های پارامتریک و ناپارامتریک در پیش‌بینی دبی میانگین روزانه رودخانه کارون (مطالعه موردی: ایستگاه هیدرومتری ملاثانی). نشریه علوم آب‌و خاک، ۲۵(۱)، ۴۳-۶۲.  
<http://jstnar.iut.ac.ir/article-1-3999-fa.html>
- فرودی خور، علی، صناعی، مجتبی، و اژدری مقدم، مهدی. (۱۳۹۶). مقایسه سامانه استنتاج فازی-عصبی تطبیقی (ANFIS) و سامانه ماشین بردار پشتیبان (SVM) در تخمین میزان ضریب دبی سرریزهای لبه‌تیز. نشریه آبیاری و زهکشی ایران، ۱۱(۵)، ۷۸۴-۷۷۲.  
[http://idj.iaid.ir/article\\_59839.html](http://idj.iaid.ir/article_59839.html)
- معینی، حمید، بندکداری، حسین، فاطمی، سید احسان، و ابتهاج، عیسی. (۱۳۹۴). مدل‌سازی دبی ماهانه ورودی به مخزن سد جامیشان با مدل‌های خود همبسته با میانگین متحرک تجمعی و سامانه استنتاج فازی-عصبی انطباقی. نشریه دانش آب‌و خاک، ۲۶(۱/۲)، ۲۸۵-۲۷۳.  
[https://water-soil.tabrizu.ac.ir/article\\_4894.html](https://water-soil.tabrizu.ac.ir/article_4894.html)

ناظری تهرودی، محمد، هاشمی، سید رضا، احمدی، فرشاد، و ناظری تهرودی، زهرا. (۱۳۹۵). بررسی دقت مدل‌های ANFIS، SVM و GP در مدل‌سازی مقادیر دبی جریان رودخانه. نشریه آکو هیدرولوژی، ۳(۳)، ۳۶۱-۳۴۷. <https://dx.doi.org/10.22059/ije.2016.60024>

نوذری، حامد، و توکلی، فاطمه. (۱۳۹۶). پیش‌بینی جریان با استفاده از مدل ماشین بردار پشتیبان بر مبنای سری‌های زمانی دبی و بارش در ایستگاه‌های بالادست (مطالعه موردی: ایستگاه هیدرومتری تله زنگ). مجله مدل‌سازی در مهندسی، ۱۶(۵۴)، ۹۵-۱۰۴. <https://dx.doi.org/10.22075/jme.2017.11363.1112>

## References

- Ahmadi, F., Radmanesh, F., & Mirabbasi, R. (2016). Comparison between Genetic Programming and Support Vector Machine Methods for Daily River Flow Forecasting (Case Study: Barandoozchay River). *Water and Soil*, 28(6), 1162-1171. <https://dx.doi.org/10.22067/jsw.v0i0.32406> [In Persian]
- Ahmadi, F. (2020). Evaluation of Support Vector Machine and Adaptive Neuro-Fuzzy Inference System Performance in Prediction of Monthly River Flow (Case Study: Nazlu chai and Sezar Rivers). *Iranian Journal of Soil and Water Research*, 51(3), 673-686. <https://dx.doi.org/10.22059/ijswr.2019.290994.668356> [In Persian]
- Alinezhadi, M., Mousavi, S.F., & Hosseini, K. (2021) Comparison of Gene Expression Programming (GEP) and Parametric and Non-Parametric Regression Methods in the Prediction of the Mean Daily Discharge of Karun River (A case Study: Mollasani Hydrometric Station). *JWSS*, 25(1). 43-62. <http://jstnar.iut.ac.ir/article-1-3999-fa.html> [In Persian]
- Banzhaf, W., Nordin, P., Keller, R.E., & Francone, F.D. (1998). *Genetic Programming*. Morgan Kaufmann, San Francisco, CA. [https://dx.doi.org/10.1007/978-3-642-14344-1\\_10](https://dx.doi.org/10.1007/978-3-642-14344-1_10)
- Breiman, L., (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://link.springer.com/article/10.1023/A:1010933404324>
- Chen, Q., Dai, G., & Liu, H. (2002). Volume of fluid model for turbulence numerical simulation of stepped spillway overflow. *Journal of Hydraulic Engineering*, 128(7), 683-688. [https://doi.org/10.1061/\(ASCE\)0733-9429\(2002\)128:7\(683\)](https://doi.org/10.1061/(ASCE)0733-9429(2002)128:7(683))
- Dehghani, R., Yonesi, H., & Torabi Poudeh, H. (2017). Comparing the performance of Support Vector Machines, Gene Expression Programming and Bayesian networks in predicting river flow (Case study: Kashkan River). *Journal of Water and Soil Conservation*, 24(4), 161-177. <https://dx.doi.org/10.22069/jwfst.2017.12398.2701> [In Persian]
- Goldberg, D.E. (2002). *The Design of Innovation: Lessons from and for Competent Genetic Algorithms*. Addison-Wesley. Reading, MA, 2002 edition.
- Foroudi Khowr, A., Saneie, M., & Azhdari Moghaddam, M. (2017). Comparison of Adaptive Neuro Fuzzy Inference System (ANFIS) and Support Vector Machines(SVM) for discharge capacity prediction of a sharp-crested weirs. *Iranian Journal of Irrigation & Drainage*, 11(5), 772-784. [http://idj.iaid.ir/article\\_59839.html](http://idj.iaid.ir/article_59839.html) [In Persian]
- Jalali, M.N., Ghezelbash Naderi, B., Guili, A., & Bahrami, J. (2017). Select the best time series model to predict the river inlet flow. *Third National Conference on Semi-Arid Hydrology with a focus on water, humans, nature, Sanandaj*. <https://civilica.com/doc/903801> [In Persian]
- Jeirani, F., Morid, S., & Moridi, A. (2011). Effect of spatial accuracy of digital elevation map on calibration and estimation of runoff and sediment using CUP-SWAT model. *Journal of Soil Conservation Research*, 18(4), 81-102. <https://dorl.net/dor/20.1001.1.23222069.1390.18.4.5.5> [In Persian]
- Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press. Ann Arbor.

- Heshmati, S., & Hafez Parast Movadat, M. (2018). Flow prediction using time series analysis with SARIMA model in drought conditions Case study: Pir Salman flowmeter station in Jamishan river. *Rainwater catchment systems*, 6(1), 73-82. <http://jircsa.ir/article-1-309-fa.html> [In Persian]
- Koza, J., (1992). *Genetic programming on the programming of computers by natural selection*. MIT Press, Cambridge, MA.
- Mantero, P., Moser, G., & Serpico, S.B. (2005). Partially supervised classification of remote sensing images through SVM-based probability density estimation. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3), 559-570. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.467&rep=rep1&type=pdf>
- Moeeni, H., Bonakdari, H., Fatemi, S.A., & Ebtehaj, I. (2016). Modeling the Monthly Inflow to Jamishan Dam Reservoir Using Autoregressive Integrated Moving Average and Adaptive Neuro- Fuzzy Inference System Models. *Water and Soil Science*, 26(1-2), 273-285. [https://water-soil.tabrizu.ac.ir/article\\_4894.html](https://water-soil.tabrizu.ac.ir/article_4894.html) [In Persian]
- Nazeri Tahroudi, M., Hashemi, S., Ahmadi, F., & Nazeri Tahroudi, Z. (2016). Evaluation the Accuracy of ANFIS, SVM and GP Models to Modeling the River Flow Discharge. *Iranian journal of Ecohydrology*, 3(3), 347-361. <https://dx.doi.org/10.22059/ije.2016.60024> [In Persian]
- Nozari, H., & tavakoli, F. (2018). Stream Flow Prediction Using Support Vector Machine Based on Discharge and Precipitation Time series on Upstream Stations (Case Study: Taleh Zang Hydrometric Station). *Journal of Modeling in Engineering*, 16(54), 95-104. <https://dx.doi.org/10.22075/jme.2017.11363.1112> [In Persian]
- Samadi, M., & Fathabadi, A. (2019). Application of Time Series, ANN, and SVM Models in Forecasting the Gorgan Dam Inflow Rate. *Environment and Water Engineering*, 4(4), 299-309. <https://dx.doi.org/10.22034/jewe.2018.128256.1256> [In Persian]
- Samadi, M., Bahremand, A., & Fathabadi, A. (2019). The Boustan Dam monthly inflow forecasting using data-driven and ensemble models in the Golestan Province. *Watershed Engineering and Management*, 11(4), 1044-1058. <https://dx.doi.org/10.22092/ijwmse.2018.108926.1251> [In Persian]
- Samadi, M., Moslem, A.P, Beitollahpour Chaharmahali, A., Kushafar, L., & Farmahini Farahani, M.A. (2019). Investigating the efficiency of support vector machine (SVM) and neural network models in predicting the inlet flow to Golestan Dam. 9th Scientific Conference on Watershed Management and Management of Water and Soil Resources, Kerman. <https://civilica.com/doc/10073> [In Persian]
- Seyedian, S., Soleimani, M., & Kashani, M. (2014). Predicting streamflow using data-driven model and time series. *Iranian journal of Ecohydrology*, 1(3), 167-179. <https://dx.doi.org/10.22059/ije.2014.54219> [In Persian]
- Shin, K.S., Lee, T.S., & Kim, H.J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert systems with applications*, 28(1), 127-135. <https://doi.org/10.1016/j.eswa.2004.08.009>
- Sivapragasam, C, Shie-Yui Liang, S, Y., & Pasha, M.F.K., (2001). Rainfall and runoff forecasting with SSA-SVM approach. *Journal of Hydro informatics*. 3(3), 141-152. <https://doi.org/10.2166/hydro.2001.0014>
- Sudheer, Ch., Anand, N., Panigrahi B.K., & Mathur, S., (2013). Stream flow forecasting by SVM with quantum behaved particle swarm optimization. *Neuro computing*, 101, 18-23. <https://doi.org/10.1016/j.neucom.2012.07.017>
- Rahimi, B., & Hafez Parast Movadat, M. (2020). Comparison of support vector regression models, gene expression programming and IHACRES in predicting runoff changes under the influence of climate

- change (Case study: Jamishan Dam). Iranian Soil and Water Research, 51(10), 2483-2499. <https://dx.doi.org/10.22059/ijswr.2020.303779.668640> [In Persian]
- Zarezadeh, M., & Bozorg Haddad, O. (2010). Inflow Simulation and Forecasting Optimization Using Hybrid ANN-GA Algorithm. Water and Soil, 24(5), 942-954. <https://dx.doi.org/10.22067/jsw.v0i0.5295> [In Persian]
- Zeinalie, M., Golabi, M.R., Sharifi, M.R., & Hafez Parast Movadat, M. (2018). Evaluation of Artificial Intelligence Models in River Flow Modeling, Case Study: Gamasiab River. Journal of Watershed Engineering and Management, 11(4), 941-954. <https://dx.doi.org/10.22092/ijwmse.2018.115870.1370>. [In Persian]