

## انتخاب بهترین روش یادگیری ماشین برای پیش بینی بارندگی

مریم باقری نسب بهابادی

مقطع کارشناسی ارشد دانشگاه آزاد اسلامی واحد یزد.

نام نویسنده مسئول:

مریم باقری نسب بهابادی

### چکیده

پیش بینی بارندگی از تاریخچه طولانی برخوردار است. هدف از این پیش بینی کاهش خطر مالی در شرایط نامساعد آب و هوایی است با توجه به اهمیت روز افزون منابع آب و نیاز به هزینه های هنگفت برای ساخت و بهره برداری از سامانه های منابع آب، از یک طرف و نیز نوپا بودن علوم مرتبط با مهندسی و مدیریت منابع آب در جهان به طور عام و در ایران به طور خاص، از طرف دیگر اهمیت را در پیش بینی بارش و برآورد میزان آن در هر زمان و برای هر منطقه و حوضه آبریز، به عنوان یکی از مهم ترین پارامترهای جوی، از اهمیت ویژه ای در استفاده بهینه از منابع آبی برخوردار کرده است. بطوری که پیش بینی بارش نقش اساسی و موثری را در مدیریت و بهره برداری صحیح از حوضه، مدیریت سدها و مخازن و به حداقل رسانی خسارات ناشی از سیلاب، خشکسالی و مدیریت منابع آب ایفا می کند. پیش بینی هر واقعه ای، اساس مدیریت بحران آن را تشکیل می دهند و این امکان زمانی حاصل می شود که بتوان مدل های پیش بینی متناسب را در اختیار داشت. روش های مختلفی برای پیش بینی وقایع هیدرولوژیکی (از جمله بارندگی) به کار می روند. نتایج حاصل از بکارگیری هر یک از این روش ها همواره با مقداری خطا همراه می باشد. پیش بینی درست در سیگنالهای هیدرولوژیکی نظیر بارندگی می تواند اطلاعات مفیدی را به منظور پیش بینی مقدار بارندگی و مدیریت منابع آب و خاک در یک حوضه ارائه دهد علاوه بر این پیش بینی درست در سیگنالهای هیدرولوژیکی نقش مهمی را در کاهش اثرات خشکسالی بر سیستم های منابع آب ایفا می کند با توجه به ماهیت غیرخطی، عدم قطعیت و عدم صراحت زیاد و ویژگی های متغیر زمانی و مکانی در سیستم گردش آبی، هیچ یک از مدل های آماری و مفهومی پیشنهاد شده به منظور الگو سازی دقیق بارش نتوانسته اند به عنوان یک مدل برتر و توانا شناخته شوند و این مسئله همواره از پیچیدگی های ویژه های برخوردار است. پژوهش حاضر با هدف بررسی انتخاب روش مناسب برای پیش بینی بارندگی انجام شده به منظور پیاده سازی روش پیشنهادی از نرم افزار متلب نسخه 2016a استفاده شده است.

**واژگان کلیدی:** پیش بینی، بارندگی، یادگیری ماشین، شبکه عصبی.

## مقدمه

بارش<sup>۱</sup> یک پدیده مهمی در یک سیستم آب و هوایی است که ماهیت بی نظمی آن تأثیر مستقیم بر برنامه ریزی منابع آب، کشاورزی و سیستم های بیولوژیکی دارد. در بخش مالی، میزان بارندگی در طی یک دوره زمانی برای ارزیابی ارزش یک امنیت مالی حیاتی است. در طی سال های اخیر، توانایی دانشمندان در درک و پیش بینی بارش باران، به دلیل مدل های ناسازگار برای افزایش دقت پیش بینی بارش افزایش یافته است.

در نتیجه، چنین تلاش هایی در تکنیک های جدید می تواند منجر به پیش بینی صحیح میزان بارش برای مشتقات آب و هوا<sup>۲</sup> شود [۷].

مشتقات آب و هوایی یک روش جدید برای کاهش خطر مالی ناشی از شرایط نامساعد یا شرایط آب و هوایی است. این قراردادها مربوط به افرادی است که به طور مستقیم یا غیر مستقیم تحت تاثیر بارندگی قرار دارند. به عنوان مثال، کشاورزان که محصولات اصلی آنها درآمدشان هست، نیاز به طیف وسیعی از بارش باران در طی یک دوره زمانی برای به حداکثر رساندن درآمد دارند. مشتقات باران یک راه حل ساده برای حل مشکلات مربوط به حفاظت مالی در برابر بارندگی نامناسب ارائه می دهند [۶].

مشتقات باران شامل دو مشکل می شود. اولین مشکل پیش بینی بارندگی های متراکم در یک دوره زمانی خاص است. مشکل دوم این است که فرایند قیمت گذاری (مشتقات آب و هوایی، به عنوان قرارداد بین دو یا چند طرف تعریف می شوند، از این رو، در مورد مشتقات آب و هوا، دارایی اساسی نوع آب و هوا، مانند بارندگی است) را توسعه دهیم. دومی ویژگی های خاص خود را دارد، زیرا مشتقات باران در بازار ناقص هستند. این پایان نامه در بخش اول پیش بینی میزان بارندگی را بر عهده دارد. توجه داشته باشید که مهم است که یک مدل داشته باشید که می تواند سطح بارندگی را دقیق پیش بینی کند، زیرا قراردادهای مربوط به پیش بینی تراکم بارش باران در یک دوره زمانی است. همچنین اندازه گیری دقیق بارندگی می تواند بسیار دشوار باشد. بنابراین هدف وسیع ما توسعه یک روش برای پیش بینی دقیق بارش است که باید خطر سرمایه گذاری را کاهش دهد [۷].

## موجک

تبدیل موجک [۱۶-۱۷] تجزیه یک تابع بر مبنای توابع موجک میباشد. موجکها (که به عنوان موجکهای دختر شناخته میشوند) نمونه های انتقال یافته و مقیاس شده یک تابع (موجک مادر) با طول متناهی و نوسانی شدیداً میرا هستند. در مقایسه با تبدیل فوریه می توان گفت که تبدیل موجک دارای خصوصیت محلی سازی بسیار خوبی است. بطور مثال تبدیل فوریه یک پیک تیز دارای تعداد زیادی ضریب است، چرا که توابع پایه تبدیل فوریه توابع سینوسی و کسینوسی هستند که دامنه آنها در کل بازه ثابت است، در حالی که توابع موجک توابعی هستند که بیشتر انرژی آنها در بازه کوچکی متمرکز شده است و به سرعت میرا می شوند. بنابراین با انتخاب مناسب موجک های مادر می توان فشرده سازی بهتری در مقایسه با تبدیل فوریه انجام داد. دو نوع تبدیل موجکی وجود دارد:

تبدیل پیوسته موجکی

تبدیل گسسته موجکی

تبدیل گسسته موجکی جهت پیش پردازش و آنالیز سری های زمانی مناسب تر از تبدیل موجکی پیوسته می باشد، زیرا داده های تبدیل یافته با DWT اجزای اضافی نداشته و بنابراین تبدیل معکوس را می توان برای هردسته از داده های زمان-فرکانس بکار برد.

تبدیل موجک گسسته

برای آسانتر شدن کار با تبدیل موجک معمولاً گسسته سازی آن به صورت دودویی انجام می گیرد. بدین ترتیب که مقیاس و مکان به صورت توانهایی صحیح از ۲ می باشد.

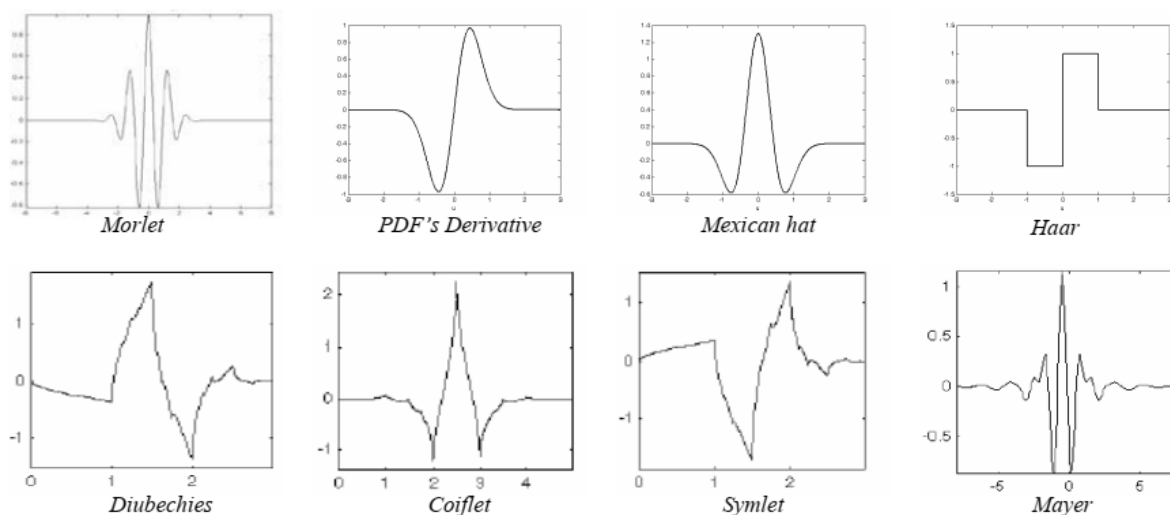
تبدیل موجک گسسته که بر روی سیگنال های گسسته عمل می کند، باعث افزایش سرعت تبدیل موجک شده است. این نوع تبدیل به راحتی قابل پیاده سازی است و همچنین دارای سرعت بالایی است و به منابع کمی نیاز دارد.

تبدیل موجک گسسته را می توان بر روی سیگنال های دو بعدی همچون تصویر اعمال کرد. تبدیل موجک گسسته از ساختار چند دقتی استفاده می کند که یک سیگنال تبدیل شده را با استفاده از یک مجموعه فیلتر برای چندین بار آنالیز می کند.

<sup>1</sup> Rainfall

<sup>2</sup> Weather derivatives

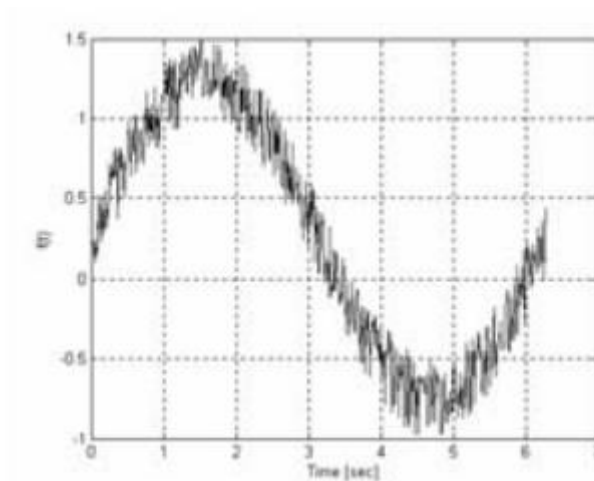
دو فیلتر یکی بالاگذر و دیگری پایین گذر مورد استفاده قرار می‌گیرد. اگر ورودی تصویر باشد در ابتدا عملیات فیلترینگ به صورت افقی انجام می‌شود و در مرحله بعد عملیات فیلترینگ با همان فیلترها به صورت عمودی روی سیگنال اعمال می‌شود. بر اساس کاربردی که از موجک انتظار می‌رود، انواع گوناگونی از موجکهای مادر قابل تولید هستند. تعدادی از این موجکهای مادر در شکل ۳-۳ معرفی شده‌اند. هر یک از این موجکها مادر خانواده‌ای هستند که بسته به کاربرد می‌توان از آنها استفاده کرد یا انواعی دیگر را تعریف کرد. ایده موجک از سادهترین شکل آن یعنی موجک Haar آغاز شد. برای روشنتر شدن عملکرد و قابلیت موجک در قسمت بعد مثالی آورده شده است در این پژوهش از موجک مادر Haar به منظور استخراج ویژگی از نمودار سری زمانی میزان بارش استفاده شده است.



شکل ۲-۱- برخی از انواع موجک مادر [۱۷]

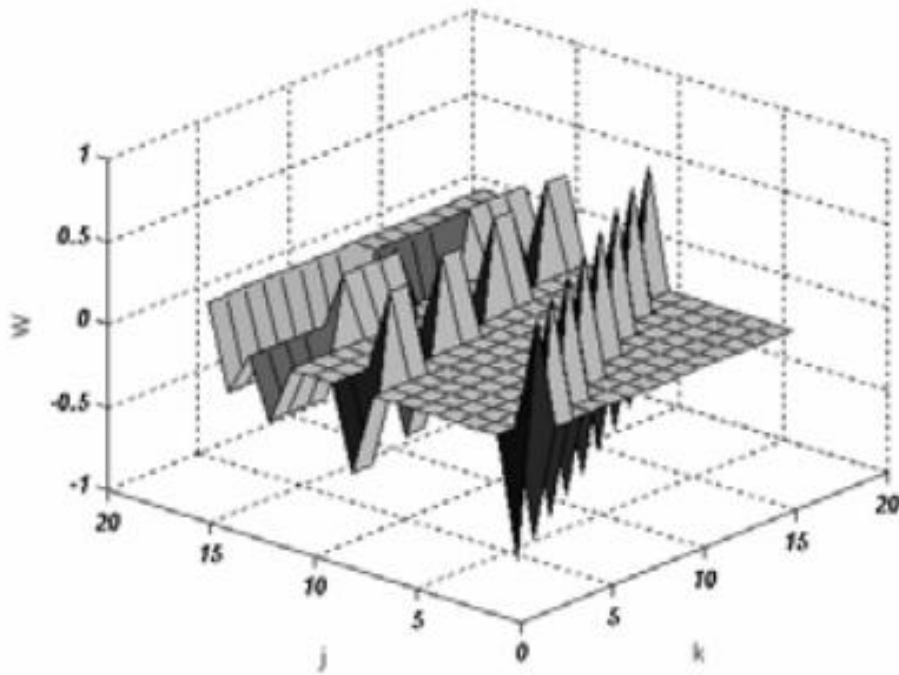
### مثالی از اعمال موجک Haar بر یک سیگنال

سیگنال تصادفی سوار بر موجی سینوسی به عنوان سیگنال آزمایش انتخاب شده است. این سیگنال در شکل ۳-۴ نشان داده شده است



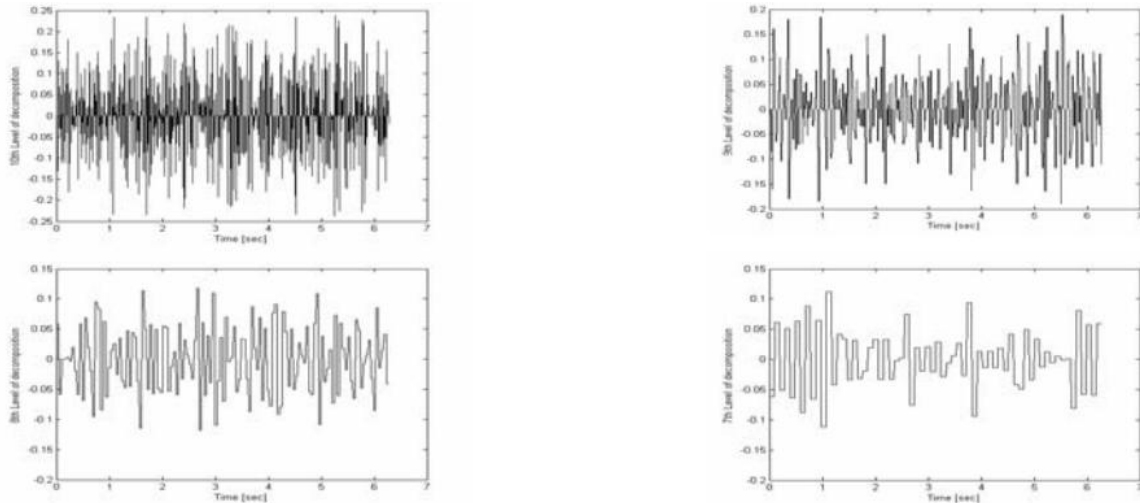
شکل ۲-۲- سیگنال تست موجک Haar [۱۷]

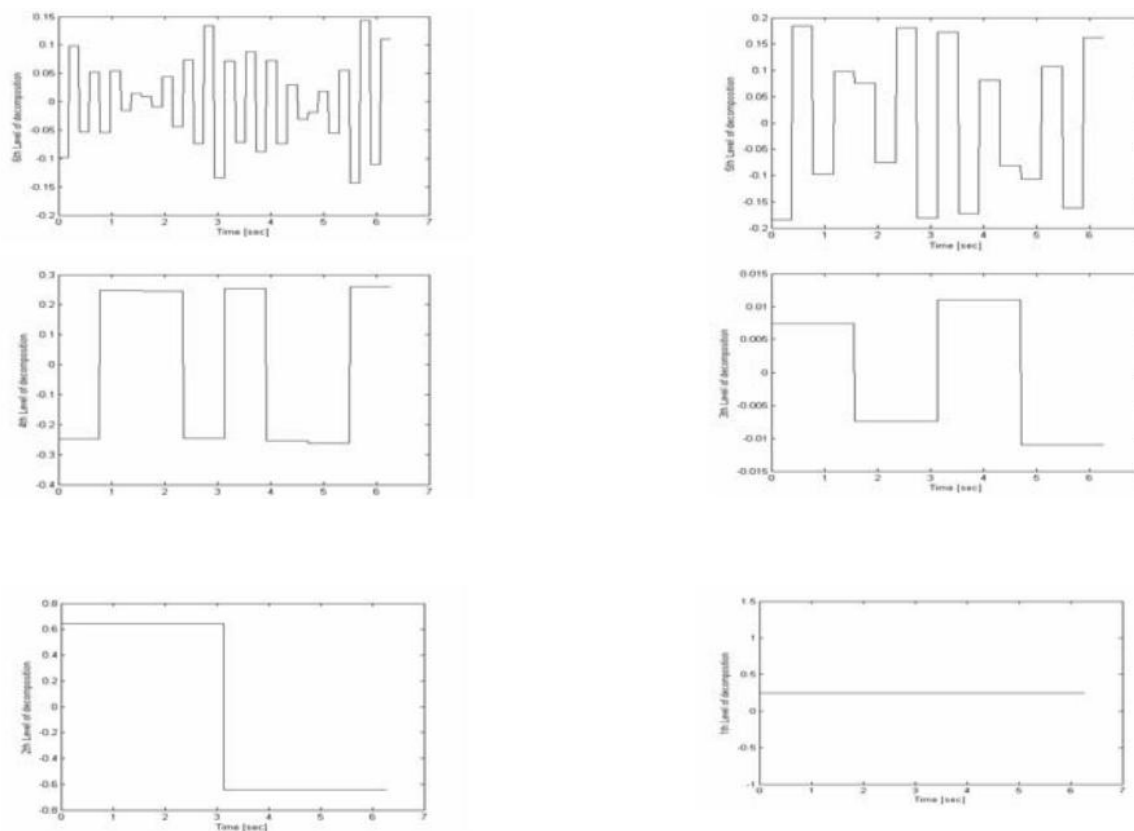
تبدیل موجک در حالت گسسته می‌تواند به فرم ماتریسی نوشته شود یعنی، تبدیل موجک با ضرب ماتریس تبدیل در نمایش برداری سیگنال حاصل آید. ماتریس تبدیل به صورت گرافیکی در شکل ۳-۵ مشاهده می‌شود



شکل ۲-۳- ماتریس تبدیل موجک Haar [۱۷]

سیگنال آزمایش پس از اعمال تبدیل موجک، در هر مقیاس بازسازی شده است که به ترتیب از مقیاس کوچک (فرکانس بالا) به مقیاس بزرگ (فرکانس پایین) در شکل ۳-۶ نشان داده شده است.





شکل ۲-۴- مؤلفه های سیگنال آزمایش بر حسب مقیاس [۱۷]

### روش برنامه ریزی ژنتیک<sup>۲</sup>

برنامه ریزی ژنتیک که جزو روشهای الگوریتم گردشی محسوب می شود. مبنای تمامی این روش ها بر اساس نظریه تکاملی داروین استوار است. الگوریتم های یاد شده اقدام به تعریف یک تابع هدف در قالب معیارهای کیفی نموده و سپس تابع هدف را برای اندازه گیری و مقایسه روش های مختلف حل کرده و در یک فرایند گام به گام تصحیح ساختار داده ها به کار گرفته و در نهایت، روش حل مناسب را ارائه می نمایند.

برنامه ریزی ژنتیک از روش های جدید بین روش های الگوریتم گردشی می باشد که به دلیل دارا بودن دقت کافی، به عنوان یک روش کاربردی مطرح می شود. از برنامه ریزی ژنتیک به طور وسیعی در کاربرد هوش مصنوعی در حل مسائل مهندسی استفاده شده است. از برنامه ریزی ژنتیک به طور موفقیت آمیزی می توان در شرایط زیر بهره جست:

- (۱) هرگاه تشخیص ارتباط داخلی میان متغیرهای وابسته بسیار پیچیده باشد
- (۲) یافتن اندازه و یا شکل متغیر نهایی بسیار پیچیده است
- (۳) حل تحلیلی مسئله به روش های ریاضی متداول غیر ممکن و یا بسیار پیچیده بوده ولی حصول جواب های تقریبی کفایت می کند

(۴) اصلاحات جزئی و متداول اجرایی بر روی نتایج به راحتی قابل لحاظ و اندازه گیری می باشد

(۵) حجم بالای داده های ورودی نیازمند به آزمون و طبقه بندی (مانند زنجیره DNA در مباحث زیست مولکولی، داده های ماهواره ای و یا داده های مربوط به امور مالی و بانکی).

**برنامه ریزی ژنتیک** شامل یک معادله ارتباط دهنده بین متغیرهای ورودی و خروجی بوده، لذا قادر به انتخاب خودکار متغیرهای مناسب الگو و حذف متغیرهای غیر مرتبط است که این امر سبب کاهش ابعاد متغیرهای ورودی خواهد شد. انتخاب ورودی های مناسب، یکی از مهمترین مواردی است که بایستی در این روش مورد توجه قرار گیرد.

این امر در شرایطی که از داده‌های ورودی مختلفی بهره برده می‌شود، از اهمیت مضاعفی برخوردار خواهد شد. چرا که ارائه داده‌های ورودی غیر مرتبط، سبب کاهش دقت الگو و ایجاد الگوهای پیچیده تری می‌شود که تفسیر آنها با دشواریهای بیشتری مواجه است. در کاربردهای مهندسی، از برنامه ریزی ژنتیک به طور وسیعی در الگوسازی مسائل مربوط به تعیین ساختار پدیده‌ها استفاده می‌شود. [۱۸]

## روش تحقیق

در این قسمت پیش بینی به کمک سایر روشهای یادگیری ماشین ارائه شده است.

### ۱-۲- کم کردن خطرات جغرافیایی مشتقات آب و هوا با استفاده از یک مدل بارندگی چندریخت<sup>۴</sup>

بازارهای خطرناک آب و هوای ناخوشایند دارای نرخ مشارکت پایین هستند، یعنی فاصله ای بین تمایل به پرداخت مشتقات آب و هوایی یا بیمه آب و هوای مبتنی بر شاخص و تمایل به پذیرش چنین بیمه ای وجود دارد [Dischel و همکاران ۲۰۰۲]. یکی از دلایلی که در میان دیگران به دلیل عدم تمایل خریداران بالقوه برای وارد شدن به بازار آب و هوا، وجود خطر پایه جغرافیایی است. Diaz و همکاران نشان دادند که خطر ترکیب جغرافیایی مشتقات آب و هوایی بارندگانی را می‌توان با ترکیب مشتقات ایستگاه‌های مرجع مجاور کاهش داد. با تقریبی بازده یک مفهوم فرض شده توسط ترکیب وزنی K نزدیکترین همسایگان، هنگام استفاده از بیش از یک همسایه، خطا حدود ۲۰ درصد کاهش می‌یابد. [Diaz -Caneja et al 2009].

برای محاسبه وزن مطلوب برای نمونه کارها، یک مدل بارندگی چند ریخت پیشنهاد شده است که به داده‌های تاریخی کالیبره شده است و سپس بارندگی‌های آینده را شبیه سازی می‌کند. عملکرد این مدل با دو رویکرد معیار مقایسه می‌شود: یک ایده بصری، که در آن وزن انتخاب شده است و تنها بسته به فاصله تا ایستگاه مرجع دارد. معلوم شد که مدل بارندگی چند ریختی منجر به کاهش ریسک جغرافیایی بین ۲۰ تا ۴۰ درصد می‌شود و به طور کلی از روش‌های دیگر بهتر است. نتایج نشان می‌دهد که اگر حدود ۵ ایستگاه استفاده شود، خطا حداقل است. این به این معنی است که همبستگی با افزایش فاصله کاهش می‌یابد، به طوری که ایستگاه‌هایی که دورتر قرار گرفته‌اند، در تقریب تفاوت ایجاد نمی‌کنند [۷].

### ۲-۲- الگوریتم تجزیه ژنتیک برای پیش بینی بارندگی در مشتقات آب و هوایی<sup>۵</sup>

در حال حاضر روند پیش بینی بارندگی در مشتقات باران تحت تاثیر مدل‌های آماری قرار گرفته است، یعنی زنجیره مارکوف با پیش بینی بارش (MCRP)<sup>۶</sup> افزایش یافته است. Cramer و همکاران از الگوریتم جدیدی برای تجزیه بارش، که هر دو الگوریتم برنامه ریزی ژنتیک / ژنتیک (GP / GA)<sup>۷</sup> است، پیشنهاد کردند. از این رو، مشکل کلی برای حل آن آسان تر می‌شود. عملکرد هر دو GP / GA را در برابر MCRP، عملکرد پایه شعاعی<sup>۸</sup> و GP بدون تجزیه<sup>۹</sup>، مقایسه می‌کند. هدف در اینجا نشان دادن اثر الگوریتم تجزیه بر حوزه مشکل است. نتایج نشان می‌دهد که به طور کلی تجزیه اثر بسیار مثبتی دارد که از لحاظ آماری بهتر از GP بدون تجزیه و MCRP است.

انگیزه برای اینکار این است که روند پیش بینی بارندگی را در یک فضای مشکل ساده تر بسازد. بنابراین، مسئله کلی پیش بینی بارش را به مجموعه ای از پارتیشن‌ها تجزیه می‌کند، جایی که بخش GP از ترکیب GP / GA می‌تواند معادلات رگرسیون چندگانه (درختان) را برای پیش بینی هر پارتیشن دقیق تر ایجاد کند. با پیاده سازی GP پیشنهاد شده، آن را با GA پیشنهادی برای طبقه بندی (پیش از انجام رگرسیون) ترکیب می‌کند. یک روش طبقه بندی لازم است، زیرا باید معیار ارزیابی رگرسیون را تعیین کند.

با استفاده از الگوریتم تجزیه پیشنهادی عملکرد شهرها در سراسر اروپا را در مقدار بارندگی با استفاده از بازه زمانی چند روز گذشته ارزیابی کرده و شواهد کافی مبنی بر این که الگوریتم پیشنهادی دارای قدرت پیش بینی کننده برتر نسبت به یک GP است. بنابراین، می‌تواند نشان دهد الگوریتم تجزیه، سهم قابل توجهی در قدرت پیش بینی مشاهده شده دارد اما این روش دقت بالایی برای تعیین میزان باران روزانه ندارد [۲].

<sup>4</sup> Minimizing Geographical Basis Risk of Weather Derivatives Using A Multi-Site Rainfall Model

<sup>5</sup> A Genetic Decomposition Algorithm for Predicting Rainfall within Financial Weather Derivatives

<sup>6</sup> Markov-chain extended with rainfall prediction

<sup>7</sup> Genetic Programming/Genetic Algorithm

<sup>8</sup> Radial Basis

<sup>9</sup> GP without decomposition

## ۳-۲- پیش بینی ماهانه بارش بر اساس شبکه عصبی مصنوعی: مطالعه موردی در ایستگاه تنگارگون، شرق کلیمانتان -

اندونزی<sup>۱۰</sup>

برای پیش بینی دقیق بارندگی، Mislan و همکاران ترکیب یک شبکه عصبی مصنوعی (ANN)<sup>۱۱</sup> با الگوریتم شبکه عصبی برگشتی (BPNN)<sup>۱۲</sup> را اعمال کرده اند. در این آزمایش، داده های بارندگی با استفاده از دو لایه پنهان از معماری BPNN با سه دوره مختلف [۱]-۱۰-۵۰-۲، دوره ۵۰۰ و [۱]-۲۰-۵۰-۲، دوره ۱۰۰۰ و ۱۵۰۰ مورد آزمایش قرار گرفتند. خطای میانگین مربع (MSE)<sup>۱۳</sup> برای اندازه گیری عملکرد طبقه بندی استفاده می شود. نتایج تجربی نشان داد که معماری [۲]-۵۰-۲۰-۱، دوره ۱۰۰۰ نتیجه خوبی با ارزش MSE برابر با ۰۰۰۹۶۳۴۱ ارائه کرده است. علاوه بر این، الگوریتم پیشنهادی یک مدل خوب برای پیش بینی بارندگی در ایستگاه تنگارگون، شرق کلیمانتان - اندونزی ارائه کرده است.

الگوریتم پیشنهادی برای مدل سازی و پیش بینی بارندگی در ایستگاه تنگارگون، شرق کلیمانتان - اندونزی استفاده شده است. پس از آزمایش سه معماری با دوره های مختلف؛ ۵۰۰، ۱۰۰۰ و ۱۵۰۰، سپس بهترین مقدار MSE ۰۰۰۹۶۳۴۱، با معماری ۱-۲۰-۵۰-۲ و دوره های ۱۰۰۰ به دست آمده است، نتایج این مطالعه نشان می دهد که این مدل ها را می توان به عنوان یک الگوریتم پیش بینی استفاده کرد زیرا پیش بینی دقیق خوبی را ارائه می دهد. نتایج پیش بینی شده نشان می دهد که مناسب منطقه Tenggarong است که دارای نوع استوایی دو فصلی بارانی در ماه های آوریل و نوامبر است این الگوریتم پیشنهادی نتایج دقیقی بر روی انواع مختلف آب و هوایی ندارد [۴]. انگیزه این مقاله این است که در غلبه بر برخی از مشکلات موجود در سری زمانی بارندگی روزانه، روند پیش بینی بارندگی ساده تر و موثرتر باشد. از این رو، این مقاله دارای دو بخش اصلی است: (۱) تجمع بارش در چندین تکنیک یادگیری ماشین، و (۲) ارزیابی اینکه آیا ترکیبی از تجمع داده ها و الگوریتم ها در سراسر اقلیم های مختلف تحت تاثیر قرار می گیرد. نتایج نشان می دهد که شواهد کافی وجود دارد که انباشتگی مقدار بارندگی منجر به قدرت پیش بینی برتر نسبت به پیش بینی استفاده از مقادیر روزانه می شود. علاوه بر این، هنگامی که داده های متراکم را به کار بردیم، توابع پایه شعاعی، رگرسیون برداری پشتیبان<sup>۱۴</sup> و برنامه نویسی ژنتیکی<sup>۱۵</sup> به طور کلی بهترین الگوریتم بودند. این نشان می دهد که یک مجموعه پارامتر قوی موثر است، و مهمتر از همه، هیچ تفاوت معناداری در خطای پیش بینی الگوریتمی در اروپا و ایالات متحده نداریم [۱]

## ۳- مقایسه روش های موجود

معایب	مزایا	الگوریتم		
چون مشتقات بارندگی قابل تبدیل به پول نیستند معاملات واقعی صورت نمی گیرد	انعطاف پذیری بالا و تخمین قیمت واقعی بازار	توزیع گاوس عادی معکوس و تبدیل Esscher	Patel و همکاران ۲۰۱۵	۱
توسعه DMT دارای دامنه مشکوک است چون می تواند در روش های مختلف علوم اجتماعی جایگزین شود	توسعه تکنیک های داده کاوی	استفاده از تکنیک های داده کاوی طبق مقالات ۲۰۰۰ تا ۲۰۱۱	Liao و همکاران ۲۰۱۲	۲
همبستگی با افزایش فاصله کاهش می یابد	هنگام استفاده از بیش از یک همسایه خطا ۲۰ درصد کاهش می یابد	KNN	Diaz و همکاران ۲۰۰۹	۳
در پیش بینی باران روزانه دقت چندانی ندارد	قدرت پیش بینی بهتر نسبت به GP در بارندگی چند روزه	الگوریتم پیشنهادی تلفیقی GA/GP	Cramer و همکاران ۲۰۱۶	۴
مناسب نبودن برای انواع آب و هوا	ازانه پیش بینی دقیق	ترکیب شبکه عصبی	Mislan و همکاران	۵

<sup>10</sup> Rainfall Monthly Prediction Based on Artificial Neural Network: A Case Study in Tenggarong Station, East Kalimantan - Indonesia

<sup>11</sup> Artificial Neural Network

<sup>12</sup> Backpropagation Neural Network

<sup>13</sup> mean square error

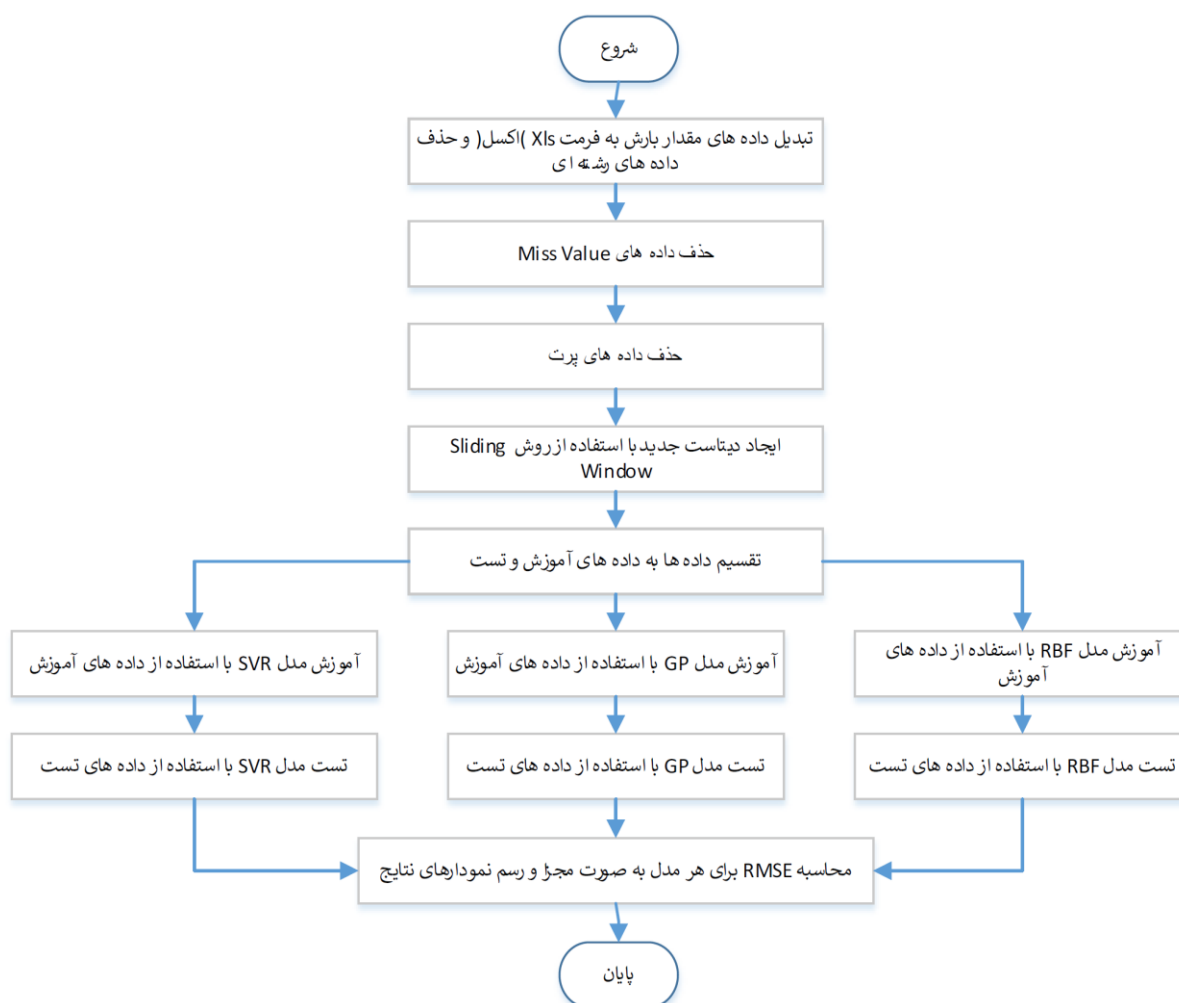
<sup>14</sup> Support vector regression

<sup>15</sup> Genetic Programming

		مصنوعی ANN و شبکه عصبی برگشتی BPNN	۲۰۱۵	
نیاز به پارامترهای زیادی در یک زمان دارد	جلوگیری از همگرایی زود هنگام در پیش بینی بارندگی	ترکیب بهینه سازی ازدحام ذرات Pso با الگوریتم ژنتیک (HPSOGA)	Wu و همکاران ۲۰۱۵	۶
برای پیش بینی بیشتر از ۱۰ روز مناسب نیست	افزایش دقت پیش بینی سهام	ANN، Svm، جنگل تصادفی، Naive Bayes	Patel و همکاران ۲۰۱۴	۷
نداشتن پارامتر بهینه و تعیین نکردن دقیق خطا	پیش بین دقت بارندگی در باران انباشته و روزانه	GP، RBF، SVR قوانین M5، درختان مدل M5، KNN	Cramer و همکاران ۲۰۱۷	۸

### ۳-۱-۳ مراحل مقایسه سه روش GP، SVR و RBF

در ابتدا و با توجه به هدف این پژوهش سه روش رگرسیون بردار پشتیبان، شبکه‌ی عصبی پایه شعاعی و برنامه ریزی ژنتیک با یکدیگر مطابق شکل ۳-۱ با یکدیگر مقایسه می‌شوند. مهمترین بخش الگوریتم مقایسه این سه روش استفاده از الگوریتم پنجره لغزان می‌باشد. مراحل مقایسه این سه روش و نیز الگوریتم پنجره لغزان در ادامه بیان خواهد شد.



شکل ۳-۱-۳- مراحل مقایسه سه روش رگرسیون بردار پشتیبان، برنامه ریزی ژنتیک و شبکه عصبی پایه شعاعی برای پیش بینی میزان بارش باران



**مرحله‌ی ۱) تبدیل داده‌ها):**

در این مرحله داده‌های موجود در دیتاست که به صورت متنی می‌باشند با استفاده از نرم افزار Excel و به صورت دستی به یک ماتریس عددی تبدیل می‌شود. هر سطر در فایل اکسل بیانگر میزان بارش باران در روزهای یک ماه می‌باشد.

مرحله‌ی ۲) حذف داده‌های بدون مقدار<sup>۱۶</sup>):

در مرحله‌ی دوم داده‌هایی که دارای مقدار Null می‌باشند در دیتاست شناسایی می‌شوند. چنانچه میزان بارش باران در یک روز به هر علتی ثبت نشده باشد در فیلد مربوطه مقدار Null قرار گرفته است. بعد از شناسایی داده‌های بدون مقدار نوبت به نحوه‌ی برخورد با این داده‌ها می‌باشد.

روش‌های مختلفی برای رفع این مشکل وجود دارد از جمله حذف این داده‌ها و یا محاسبه مقدار تقریبی این درایه‌ها با استفاده از درون‌یابی داده‌های قبلی و بعدی. در این پژوهش برای حل این مشکل مقدار داده‌های بدون مقدار با مقدار متوسط داده‌ی روز قبلی و روز بعدی طبق رابطه‌ی ۱-۳ جایگزین می‌شود.

$$x_t = \frac{x_{t-1} + x_{t+1}}{2} \quad (1-3)$$

که در آن  $x_t$  داده‌ی بدون مقدار در زمان  $t$  است.

مرحله‌ی ۳) حذف داده‌های پرت<sup>۱۷</sup>):

یک داده پرت، مشاهده‌ای است که به طور غیرعادی یا اتفاقی از وضعیت عمومی داده‌های تحت آزمایش و نسبت به قاعده‌ی آن که براساس آن آنالیز می‌شود، انحراف داشته باشد. برای حذف داده‌های پرت در این پژوهش داده‌های که اختلاف آنها از میانگین داده‌ها بزرگتر از ضریبی از واریانس داده‌ها باشد طبق رابطه‌ی ۲-۳ حذف می‌شوند.

$$x \text{ is outlier if } (x - \bar{x}) > a \times \delta \quad (2-3)$$

که  $\bar{x}$  میانگین داده‌ها،  $\delta$  واریانس داده‌ها و  $a$  یک ضریب می‌باشد که به صورت آزمایش و خطا بدست می‌آید.

مرحله‌ی ۴: ایجاد دیتاست جدید با استفاده از پنجره لغزان :

در این مرحله با استفاده از داده‌های دیتاست و بر اساس سری زمانی و روش پنجره‌ی لغزان<sup>۱۸</sup> یک دیتاست جدید ایجاد می‌شود. در روش پنجره لغزان ابتدا یک طول مشخص برای پنجره محاسبه می‌شود. سپس با توجه به طول پنجره آن را روی داده‌ها حرکت داده می‌شود. داده‌های درون پنجره برای پیش‌بینی داده‌ای که بعد از پایان پنجره قرار دارد، استفاده می‌شود. سپس پنجره یک واحد به جلو حرکت داده می‌شود و داده‌ی جدید بعدی تولید می‌شود. این فرآیند تا زمانی که پنجره به انتهای بردار داده‌ها برسد ادامه خواهد یافت. در این پژوهش از روش آزمایش و خطا برای تعیین طول پنجره‌ی لغزان استفاده می‌شود. برای مثال در جدول ۱-۳ میزان بارش باران در ۱۰ روز متوالی نشان داده شده است.

جدول ۱-۳- مثالی از میزان بارش باران در ۱۰ روز متوالی

t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
1	5	7	0	9	4	7	2	9	5

با فرض آنکه اندازه پنجره برابر ۳ باشد دیتاست جدید به صورت جدول ۲-۳ می‌باشد.

جدول ۲-۳- دیتاست جدید با اندازه پنجره ۳

ورودی ۱	ورودی ۲	ورودی ۳	خروجی
1	5	7	0
5	7	0	9

<sup>16</sup> Miss Vale

<sup>17</sup> Outlier

<sup>18</sup> Sliding Window

7	0	9	4
0	9	4	7
9	4	7	2
4	7	2	9
7	2	9	5

مرحله‌ی ۵) تقسیم داده‌ها به دو مجموعه آموزش و تست):

در این مرحله داده‌های موجود در دیتاست جدید به دو دسته داده‌های آموزش و تست تقسیم می‌شوند. داده‌های آموزش به منظور آموزش سه مدل SVR، RBF و GP استفاده می‌شود. داده‌های تست نیز برای ارزیابی عملکرد روش رگرسیون آموزش دیده شده استفاده می‌شود. روش تقسیم داده‌ها به صورت تصادفی و با توزیع یکنواخت می‌باشد. داده‌های آموزش و تست برای هر سه مدل یکسان می‌باشد.

مرحله‌ی ۶: آموزش مدل پیش‌بینی کننده:

در این مرحله مدل‌های SVR، RBF و GP، با استفاده از داده‌های آموزش جدا شده در مرحله‌ی ۵ آموزش داده می‌شوند.

مرحله‌ی ۷: تست مدل‌های SVR، RBF و GP:

مدل‌های SVR، RBF و GP که در مرحله‌ی ۶ آموزش داده شده‌اند، با استفاده از داده‌های تست مرحله‌ی ۵ تست می‌شوند و نتایج هر یک در قالب یک بردار مجزا ذخیره می‌شود. برای ارزیابی نتایج هر یک از مدل‌ها از معیار RMSE مطابق رابطه‌ی ۳-۳ استفاده می‌شود:

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (r_t - \bar{r}_t)^2} \quad (3-3)$$

که N تعداد داده‌های تست،  $r_t$  خروجی تولید شده برای داده‌ی تست  $t$ ام و  $\bar{r}_t$  مقدار مورد انتظار برای خروجی داده‌ی تست  $t$ ام است. در بخش ۳-۵ به بررسی ساختار الگوریتم‌های SVR، RBF و GP پرداخته می‌شود. و ساختار کروموزوم و تابع هدف در الگوریتم GP برای الگوریتم پیشنهادی بیان می‌شود.

#### دیتاست مورد استفاده

برای ارزیابی روش پیشنهادی از داده‌های میزان بارش باران در دو شهر portlan و postdam استفاده شده است. این داده‌ها از وب سایت <https://www.weather.gov/> استخراج گردیده و برای هر شهر در یک فایل اکسل جداگانه قرار داده شده‌اند. هر سطر در فایل اکسل داده‌های روزهای یک ماه را نمایش می‌دهد. تعداد روزهای ماه در این دیتاست ۳۱ روز فرض شده است. تعداد سطرها (ماه‌های) دیتاست، را نشان می‌دهد. داده‌های مربوط به سال‌های ۲۰۰۴ تا ۲۰۱۳ در دیتاست‌ها برای ارزیابی استفاده شده است.

#### ۳-۴- ابزارها و نرم افزارهای مورد استفاده در پیاده سازی روش پیشنهادی

به منظور پیاده‌سازی روش پیشنهادی از نرم‌افزار متلب نسخه 2016a استفاده شده است. الگوریتم پیشنهادی با استفاده از یک سیستم کامپیوتری با مشخصات میزان RAM، ۴ گیگابایت و پردازنده Core i5 مورد ارزیابی قرار گرفته است.

SVR و RBF در متلب دارای تابع می‌باشند و لذا برای آنها از پیاده‌سازی متلب استفاده شده است. اما از آنجایی که در متلب برای GP کتابخانه داخلی وجود ندارد لذا برای الگوریتم GP از کتابخانه GPTIPS 2 استفاده شده است. توضیحات، کدها و مثال‌هایی از این کتابخانه در آدرس زیر که منبع GTIPS موجود می‌باشد. (<https://sites.google.com/site/gptips4matlab/gptips-2-tutorial-1>)

قبل از اجرای کدها باید مسیر کتابخانه GPTIPS2 را به مسیریهای پیشفرض متلب اضافه کرد.

#### ۴-۴- معیار ارزیابی

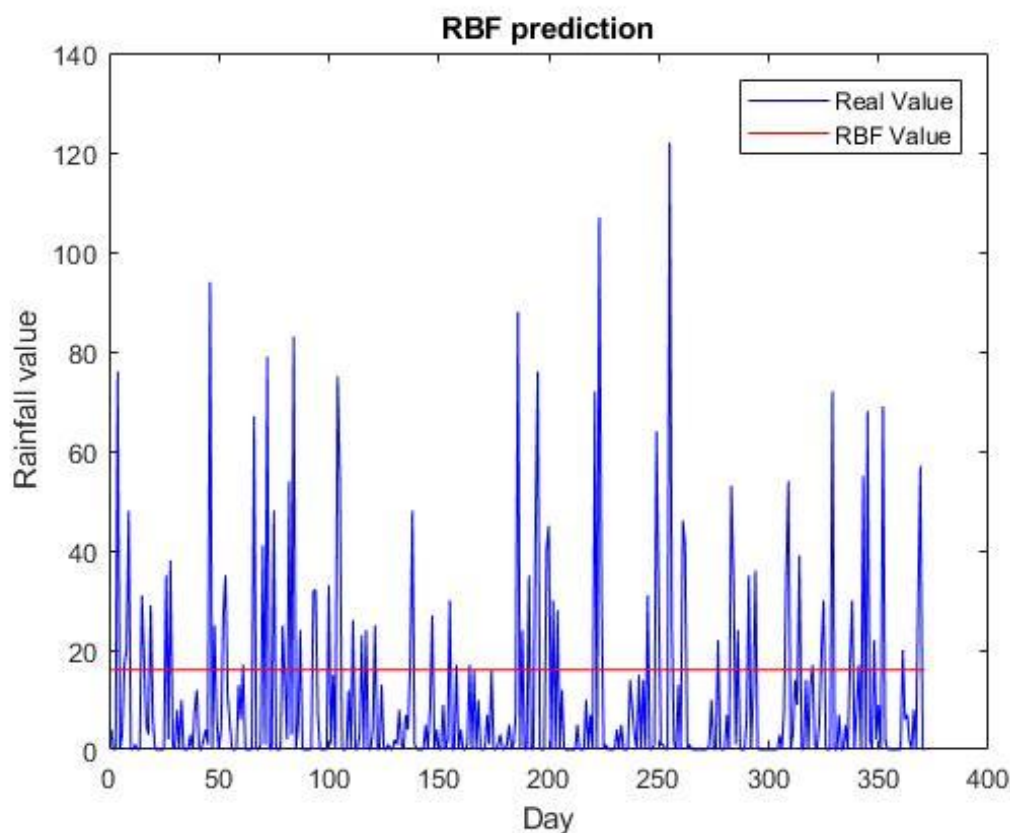
برای ارزیابی روش‌های ارائه شده از معیار RMSE استفاده شده است. مقدار RMSE برای هر مدل با توجه به خروجی‌های تولید شده توسط مدل‌ها برای داده‌های تست و خروجی‌های مورد انتظار آن‌ها محاسبه می‌شود. فرمول RMSE طبق رابطه ۴-۱ محاسبه می‌شود:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (r_i - \bar{r}_i)^2} \quad (1-4)$$

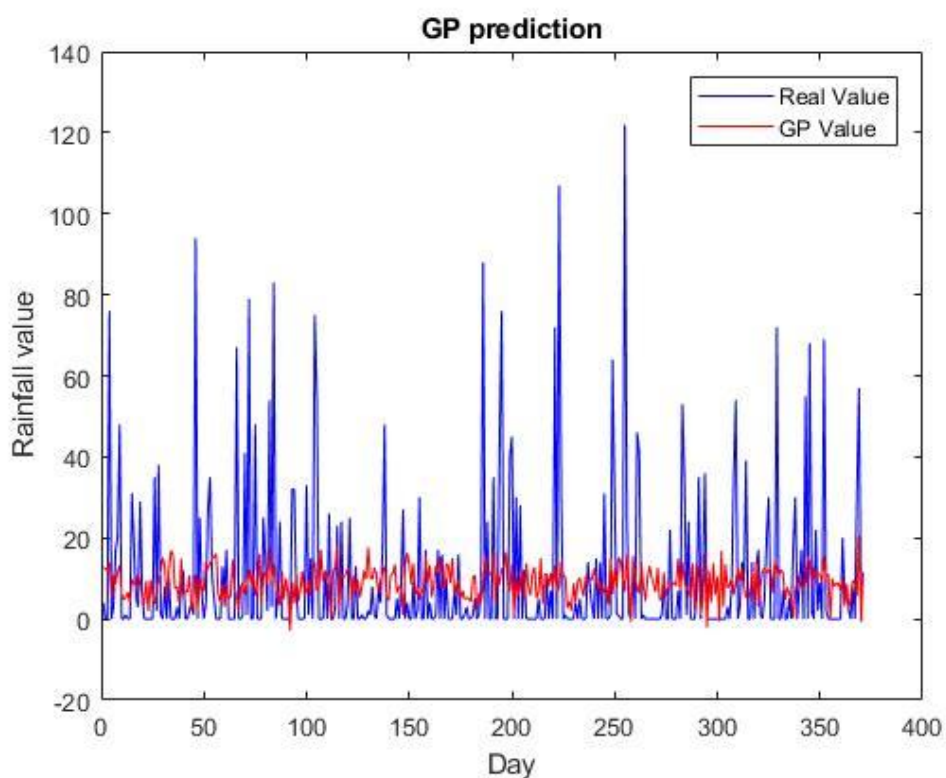
که N تعداد داده‌های تست،  $r_i$  خروجی تولید شده برای داده‌ی تست  $i$ ام و  $\bar{r}_i$  مقدار مورد انتظار برای خروجی داده‌ی تست  $i$ ام است.

#### ۴-۵- نتایج روش پیشنهادی و مقایسه

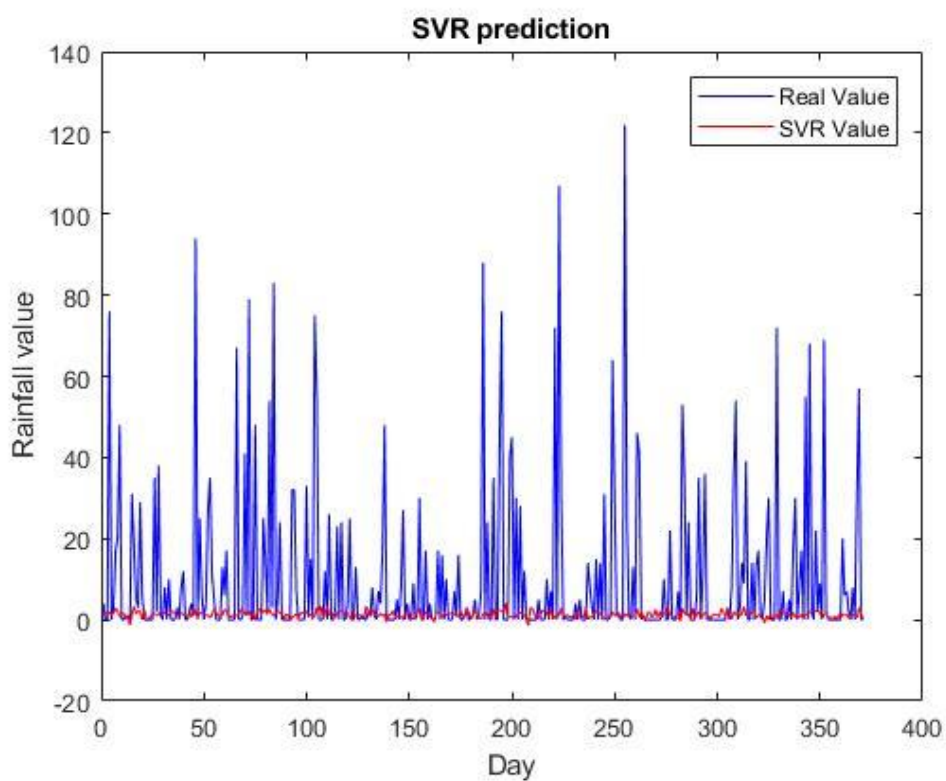
در نمودارهای ۱-۴ تا ۳-۴ خروجی حاصل از الگوریتم پیشنهادی برای مقایسه‌ی سه روش GP، SVR و RBF برای داده‌های شهر Portland نمایش داده شده است. همچنین تصاویر ۴-۴ تا ۶-۴ نتایج الگوریتم مقایسه پیشنهادی را برای شهر Postdam نمایش می‌دهند. نتایج RMSE حاصل از مقایسه سه روش GP، SVR و RBF برای هر دو شهر Portland و Postdam در جدول ۱-۴ آورده شده است.



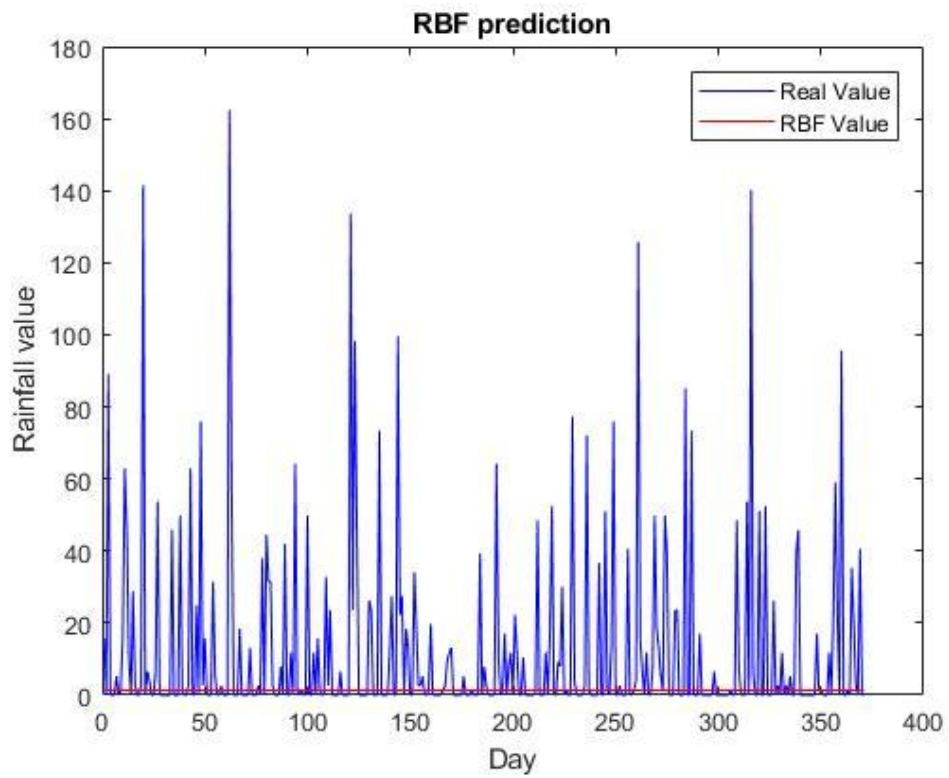
شکل ۴-۱- نمودار برآزش داده میزان بارش برای Portlan با استفاده از الگوریتم RBF



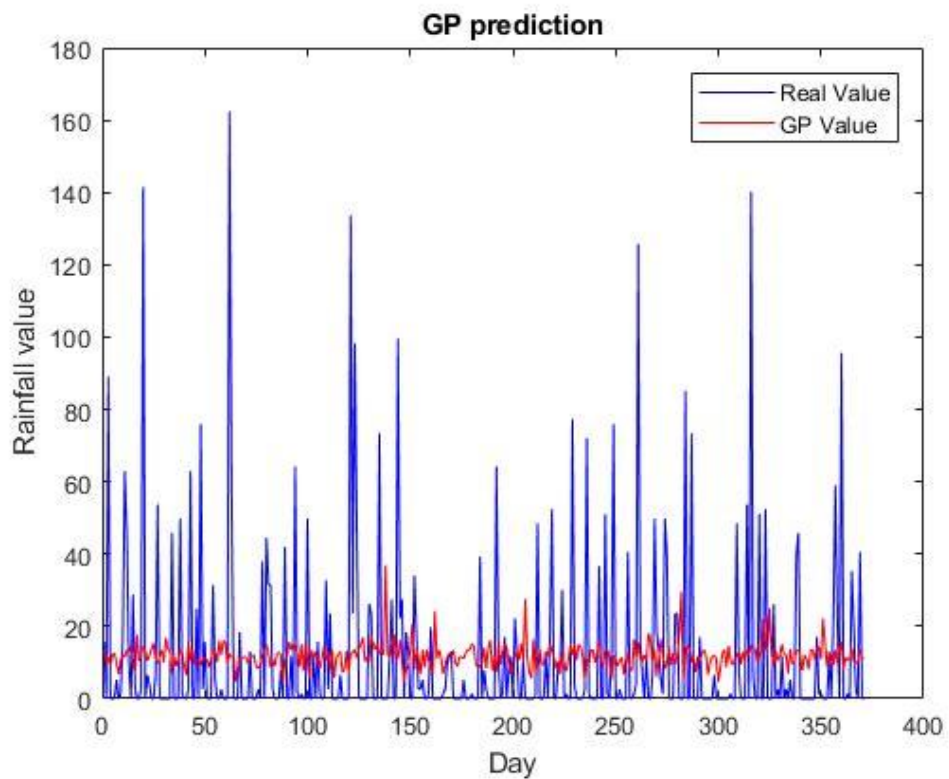
شکل ۲-۴- نمودار برازش داده میزان بارش برای Portlan با استفاده از الگوریتم GP



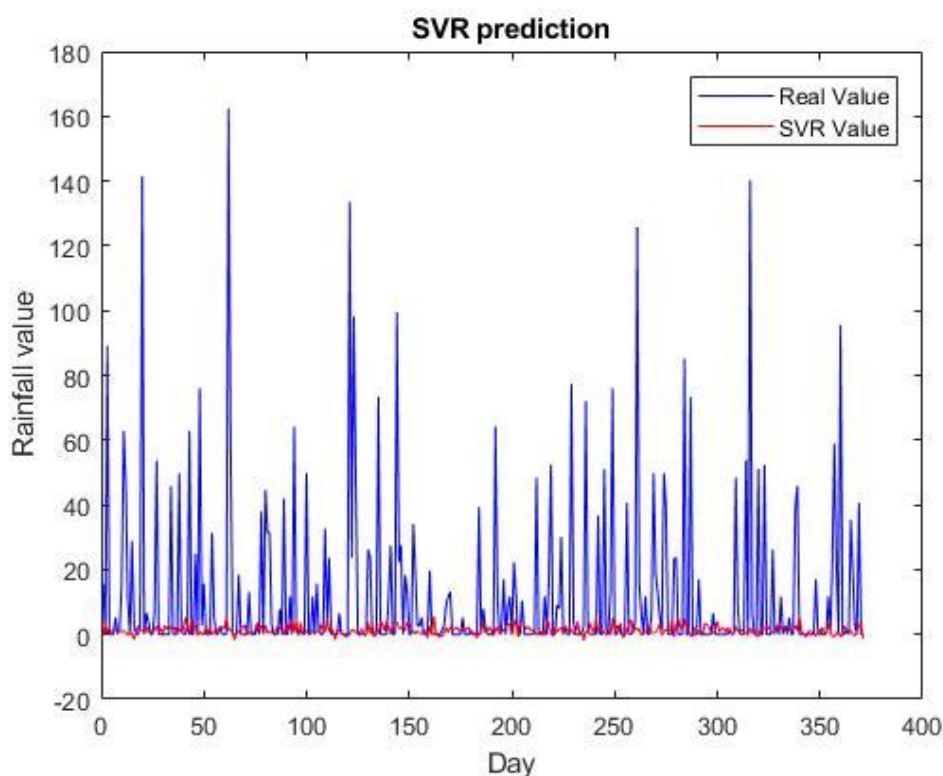
شکل ۳-۴- نمودار برازش داده میزان بارش برای Portlan با استفاده از الگوریتم SVR



شکل ۴-۴- نمودار برازش داده میزان بارش برای Postdam با استفاده از الگوریتم RBF



شکل ۴-۵- نمودار برازش داده میزان بارش برای Postdam با استفاده از الگوریتم GP



شکل ۴-۶- نمودار برازش داده میزان بارش برای Postdam با استفاده از الگوریتم SVR

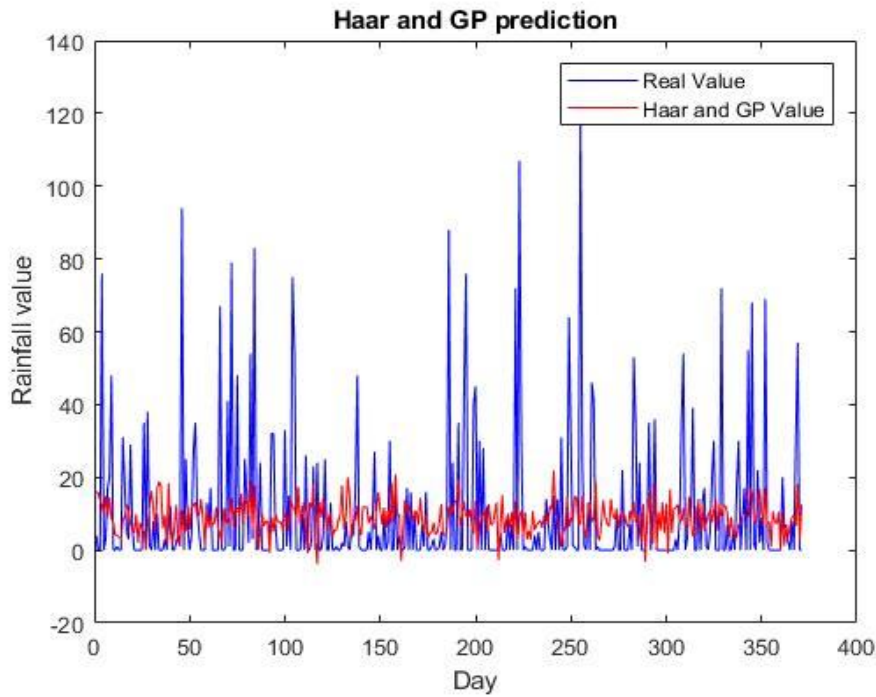
جدول ۴-۱- نتایج حاصل از مقایسه RMSE برای سه مدل GP، SVR و RBF

شهر / RMSE	RBF	GP	SVR
Portlan	158.5039	40.6101	60.98
Postdam	98.24701	30.532	40.8026

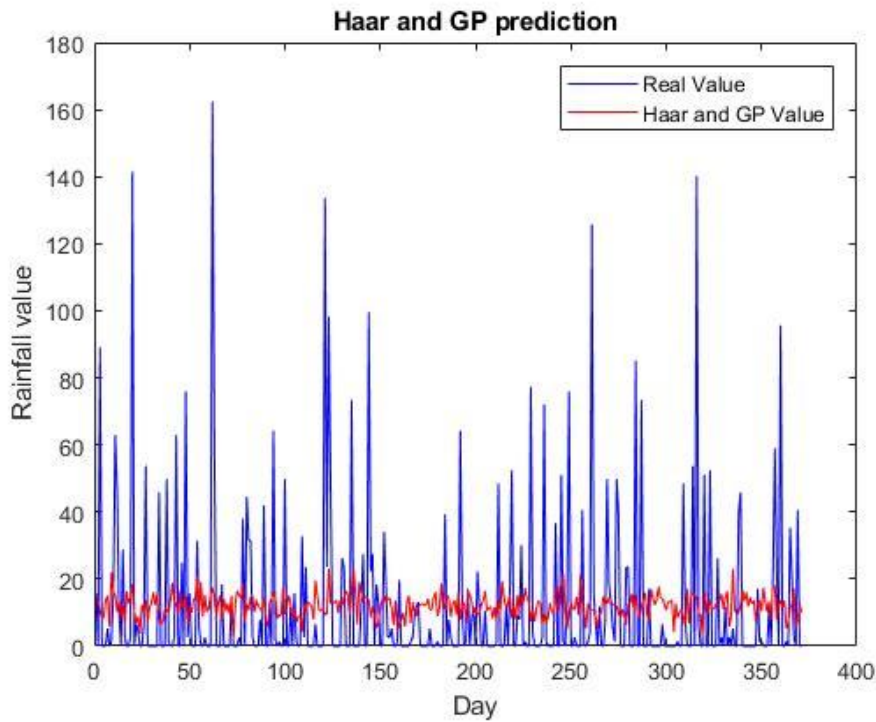
همانطور که از نتایج تصاویر ۴-۱ تا ۴-۶ و نیز جدول ۴-۱ مشاهده می‌شود در الگوریتم مقایسه اراپه شده الگوریتم GP دارای برتری بالایی نسبت به دو روش دیگر می‌باشد و روش شبکه عصبی RBF دارای کمترین کارایی می‌باشد. بنابراین با توجه به نتایج بدست آمده در این قسمت الگوریتم GP به عنوان روش پیش‌بینی کننده جهت بهبود انتخاب می‌شود. در ادامه نتایج حاصل از روش پیشنهادی که ترکیبی از الگوریتم پنجره‌ی لغزان، تبدیل موجک Haar و روش پیش‌بینی کننده‌ی GP است نمایش داده می‌شود.

#### ۴-۶- نتایج روش بهبود یافته

در قسمت ۴-۵ الگوریتم GP به عنوان کاراترین روش پیش‌بینی در مقایسه با دو روش دیگر SVR و RBF انتخاب گردید. در تصاویر ۴-۷ و ۴-۸ نتایج حاصل از روش پیشنهادی را بر روی همان داده‌های آموزش و تست مرحله ۴-۵ نمایش می‌دهد. همچنین جدول ۴-۲ نتایج RMSE روش پیشنهادی را برای دو شهر Portland و Postdom نمایش می‌دهد.



شکل ۴-۷- نمودار برازش داده میزان بارش برای Portland با استفاده از الگوریتم ترکیبی پنجره‌ی لغزان، تبدیل موجک Haar و الگوریتم GP



شکل ۴-۸- نمودار برازش داده میزان بارش برای Postland با استفاده از الگوریتم ترکیبی پنجره‌ی لغزان، تبدیل موجک Haar و الگوریتم GP

جدول ۲-۴- نتایج RMSE برای الگوریتم پیشنهادی

شهر	RBF	SVR	GP	روش پیشنهادی
Portlan	158.5039	60.98	18.6101	14.50198
Postdam	98.24701	40.8026	23.69871	19.532

همانطور که نتایج نمودارهای ۷-۴ و ۸-۴ و نیز جدول ۲-۴ نشان می‌دهد الگوریتم پیشنهادی دارای کارایی بالاتری نسبت به سه روش اولیه می‌باشد.

### نتیجه‌گیری

امروزه پیش‌بینی میزان بارش بارندگی از اهمیت بالایی برخوردار است. بارش یک پدیده مهم در یک سیستم آب و هوایی است که ماهیت بی‌نظمی آن تأثیر مستقیم بر برنامه‌ریزی منابع آب، کشاورزی و سیستم‌های بیولوژیکی دارد. در بخش مالی، میزان بارندگی در طی یک دوره زمانی برای ارزیابی ارزش یک امنیت مالی در مشتقات آب و هوا یا مشتقات بارش حیاتی است. در طی سال‌های اخیر، روش‌های متعددی برای پیش‌بینی میزان بارش باران ارائه گردیده است. در این بین روش‌های یادگیری ماشین از اهمیت بالایی برخوردار می‌باشند. هدف اصلی در این پژوهش ارائه یک الگوریتم جهت مقایسه کارآمدی سه روش پیش‌بینی کننده GP، RBF و SVR برای پیش‌بینی میزان بارش باران می‌باشد. الگوریتم پیشنهادی شامل یک سری مراحل پیش‌پردازش داده‌ها و سپس استفاده از روش پنجره‌ی لغزان می‌باشد. در ادامه پس از تعیین کاراترین روش پیش‌بینی کننده، یک روش بهبود یافته مبتنی بر استفاده از تبدیل موجک Haar برای بهبود دقت پیش‌بینی ارائه گردید.



## منابع و مراجع

- [1] Cramer, S., Kampouridis, M., Freitas, A. A., & Alexandridis, A. K. (2017). An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. *Expert Systems with Applications*, 85, 169-181.
- [2] Cramer, S. , Kampouridis, M. , & Freitas, A. (2016a). A genetic decomposition algorithm for predicting rainfall within financial weather derivatives. In *Gecco '16: Proceedings of the 2016 genetic and evolutionary computation conference* (pp. 885–892). Denver, USA: ACM .
- [3] Cramer, S. , Kampouridis, M. , & Freitas, A. A. (2016b). Feature engineering for improving financial derivatives-based rainfall prediction. In Y.-S. Ong (Ed.), *Proceedings of 2016 IEEE congress on evolutionary computation (cec 2016)* (pp. 3483–3490). Vancouver: IEEE Press .
- [4] Mislan , Haviluddin , Hardwinarto, S. , Sumaryono , & Aipassa, M. (2015). Rainfall monthly prediction based on artificial neural network: A case study in tenggarong station, east kalimantan - indonesia. *Procedia Computer Science*, 59 , 142–151 . International Conference on Computer Science and Computational Intelligence (ICCSCI 2015).
- [5] Wu, J. , Long, J. , & Liu, M. (2015). Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm. *Neurocomputing*, 148 , 136–142.
- [6] Patel, J. , Shah, S. , Thakkar, P. , & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, 42 (1), 259–268 .
- [7] Ritter, M. , Muhoff, O. , & Odening, M. (2014). Minimizing geographical basis risk of weather derivatives using a multi-site rainfall model. *Computational Economics*, 44 (1), 67–86 .
- [8] Cabrera, B. L. , Odening, M. , & Ritter, M. (2013). Pricing rainfall futures at the CME. *Journal of Banking & Finance*, 37 (11), 4286–4298 .
- [9] Liao, S.-H. , Chu, P.-H. , & Hsiao, P.-Y. (2012). Data mining techniques and applications a decade review from 2000 to 2011. *Expert Systems with Applications*, 39 (12), 11303–11311.
- [10] Marr, Marr. "A Short History of Machine Learning-Every Manager Should Read". *Forbes*. Retrieved 28 Sep 2016.
- [11] Michalski RS, Carbonell JG, Mitchell TM, editors. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media; 2013 Apr 17.
- [12] Freund, Y.; Schapire, R. E. (1999). "Large margin classification using the perceptron algorithm" (PDF). *Machine Learning*. 37 (3): 277–296.
- [13] David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, Volume 3, 3/1/2003
- [14] Breiman, L., 2001. Random forests. *Machine learning*, 45(1), pp.5-32.
- [15] Koza, J. R. (1992). *Genetic programming: on the programming of computers by means of natural selection* (Vol. 1). MIT press. Chicago
- [16] Rajae, T., Jafari, H., & Rahimi, R. (2017). Performance Improvement of Biological BOD in Rivers based on De-noising Comparison Wavelet-ANN Conjunction, GP, ANN and MLR Methods (Case Study: Karaj Dam Outlet Station).
- [17] Drucker, Harris; Burges, Christopher J. C.; Kaufman, Linda; Smola, Alexander J.; and Vapnik, Vladimir N. (1997); "Support Vector Regression Machines", in *Advances in Neural Information Processing Systems 9*, NIPS 1996, 155–161, MIT Press.
- [18] Suykens, Johan A. K.; Vandewalle, Joos P. L.; Least squares support vector machine classifiers, *Neural Processing Letters*, vol. 9, no. 3, Jun. 1999, pp. 293–300.
- [19] Smola, Alex J.; Schölkopf, Bernhard (2004). "A tutorial on support vector regression" (PDF). *Statistics and Computing*. 14 (3): 199–222. Archived (PDF) from the original on 2012-01-31.

- [20] Diaz-Caneja, M. B., Conte, C. G., Pinilla, F. J. G., Stroblmair, J., Catenaro, R., & Dittmann, C. (2009). Risk management and agricultural insurance schemes in Europe. JRC Reference Report, EU-23943, EN-2009, European Commission.