DOR: https://dorl.net/dor/20.1001.1.28212606.1401.5.14.4.4

فصلنامه علمی آماد و فناوری دفاعی، سال پنجم، شماره چهاردهم، تابستان ۱۴۰۱

بررسی، تحلیل و پیش‌بینی دادگان فائواستات به منظور استخراج اطلاعات مفید برای مدیریت راهبردی در بخش کشاورزی

حسین هادیان[۱] سید ابوالفضل مطهری[۲] محمدرضا خسروی[۳]

تاریخ دریافت:۱۳۹۹/۰۵/۲۴ **تاریخ پذیرش:۱۴۰۰/۱۲/۰۵**

چکیده

کشاورزی، یک بخش مهم از تولید ناخالص داخلی کشورهاست که در کشور ایران، بخش کوچکی از آن را تشکیل می‌دهد. هدف از این تحقیق، بررسی، تحلیل و مدل‌سازی دادگان کشاورزی فائواستات به روش‌های هوش مصنوعی برای پیش‌بینی آینده میزان فروش و واردات/صادرات اقلام مختلف ایران و کشورهای دیگر به منظور مدیریت راهبردی بهتر در بخش کشاورزی است. در گذشته، پژوهش‌های مختلفی برای مدل‌سازی تولید اقلام مختلف در کشورهای دیگر و ایران انجام شده است ولی معمولاً محدود به یک کشور و محصول مشخص بوده است. در این تحقیق، ابتدا با بررسی همه‌جانبه دادگان فائواستات به خش‌هایی از آن‌که به سود و مدیریت راهبردی در کشاورزی مرتبط است شناسایی می‌شود. سپس تحلیل و بررسی انواع روش‌های یادگیری ماشین و استوکستیک بر روی محصولات کشاورزی در کشورهای مختلف انجام خواهد شد. با انجام آزمایش‌های مختلف سری‌های زمانی فائواستات با روش افزایش-گرادیان با خطای متوسط MAPE ۱۱,۳٪ برای یک سال آینده پیش‌بینی شدند که نسبت به خطای روش پایه ARIMA (که برابر با ۱۳,۷٪ است) ۱۸٪ بهبود نسبی نشان می‌دهد.

واژه‌های کلیدی: پیش‌بینی سری زمانی، تحلیل استوکستیک، دادگان فائواستات، مدیریت راهبردی، یادگیری ماشین

مقدمه

[۱] حسین هادیان، دکتری (مهندسی کامپیوتر)، نویسنده مسئول، hadianh@ce.sharif.edu
[۲] سید ابوالفضل مطهری، استادیار دانشگاه صنعتی شریف، دکتری، motahari@sharif.edu
[۳] محمدرضا خسروی، استادیار دانشگاه عالی دفاع ملی، دکتری، khosravi@sndu.ac.ir

با توجه به افزایش جمعیت در چند دههٔ گذشته، تغییر شرایط آب و هوایی و همچنین شرایط حساس اقتصادی موجود، بهبود تولید ناخالص داخلی هرچه بیشتر و بیشتر اهمیت پیدا می‌کند. یکی از به خش‌هایی که در تولید ناخالص داخلی اهمیت زیادی دارد، محصولات کشاورزی است.

در این تحقیق مقصود این است که با استفاده از روش‌های یادگیری ماشین، دادگان موجود فائواستات[1] که شامل داده‌های کشاورزی است، مدل‌سازی شود و دنباله‌های زمانی در این دادگان برای سال‌های آینده پیش‌بینی شوند تا بتوان از این مدل‌ها برای افزایش سود استفاده کرد. به عبارتی، مسئله مدنظر در این تحقیق، تحلیل دادگان کشاورزی فائواستات برای برنامه‌ریزی کشاورزی است. راه‌های زیادی برای بهبود این بخش از تولید کشور وجود دارد، ازجمله بهینه‌سازی فناوری‌های آب‌رسانی، حمایت از تولیدات کشاورزی، استفاده از علم مهندسی کشاورزی و غیره. ولی این موضوعات از حوزهٔ این تحقیق خارج هستند. آنچه در اینجا مدنظر است، استفاده از مدل‌های هوش مصنوعی برای برنامه‌ریزی سالانه در رابطه با کشت محصولات است. برای مثال، تصمیم در مورد افزایش کشت یک محصول و کاهش کشت محصول دیگر.

عوامل زیادی روی این تصمیم‌گیری و برنامه‌ریزی مؤثر هستند ولی در اینجا به‌طور خاص به عواملی پرداخته خواهد شد که دادگان آماری برای آن‌ها موجود است. این دادگان، از سازمان فائو[2] قابل‌دسترسی است.

به عبارت دیگر، در این تحقیق سعی است که با شناسایی سری‌های زمانی تولید و فروش در دادگان فائواستات و مدل‌سازی آن‌ها، گزارش‌هایی تولید شود که می‌توانند به مسئولین

[1] FAOSTAT
[2] FAO

کشوری در تصمیم‌گیری برای سرمایه‌گذاری و برنامه‌ریزی راهبردی کمک کنند. هستهٔ این تحقیق بررسی روش‌های ساخت مدل‌های پیش‌بینی سری زمانی برای داده‌های کشاورزی است.

در سال‌های اخیر روش‌های هوش مصنوعی و یادگیری ماشین رواج زیادی پیدا کرده است و خیلی از سازمان‌ها و شرکت‌ها و دولت‌ها از این روش‌ها برای تجزیه و تحلیل داده و تصمیم‌گیری بهتر استفاده می‌کنند. اگر بخش کشاورزی کشور ما از این فرصت‌ها و روش‌ها استفاده نکند، احتمال این وجود دارد که خیلی زود از دور رقابت در سطح جهانی خارج شود. همان‌طور که بقیه کشورها احتمالاً تحلیل‌های مشابهی در مورد تولید و فروش محصولات ما انجام می‌دهند، **ضرورت** دارد که در ایران نیز تولید و فروش آن‌ها مدل‌سازی و پیش‌بینی شود. در صورتی که چنین تحقیق‌هایی انجام نشوند، مسئولان ما اطلاعات کمتری برای اتخاذ تصمیمات مهم راهبردی خواهند داشت. هدف اصلی از انجام این تحقیق، بررسی دادگان فائواستات و پیدا کردن سری‌های زمانی مفید آن برای مدل‌سازی و پیش‌بینی آینده به منظور افزایش سود است. این هدف با پاسخ به سؤالات زیر ممکن خواهد بود:

سؤال اول: چه به خش‌هایی از دادگان فائواستات برای افزایش سود کشاورزی کشور مفید هستند؟

سؤال دوم: بهترین روش برای مدل‌سازی این به خش‌ها چیست و چگونه می‌توان مقادیر آینده آن‌ها را پیش‌بینی کنیم؟

ادبیات و مبانی نظری:

پیشینه تحقیق

با توجه به هدف اصلی این تحقیق، تمرکز بر تحلیل نمودار های تولید و فروش اقلام کشاورزی با استفاده از روش‌های یادگیری ماشین است؛ به عبارت دیگر، هسته این تحقیق، **مدل‌سازی سری‌های زمانی** است چراکه داده‌های تولید و فروش کشاورزی همگی از جنس سری زمانی هستند و قصد است که مقادیر این سری‌های زمانی برای آینده پیش‌بینی شوند. درنتیجه در اینجا پیشینه مدل‌سازی سری‌های زمانی به‌طورکلی و همین طور به‌طور خاص برای کشاورزی مرور خواهد شد.

به‌طورکلی، یکی از اولین روش‌های ارائه‌شده برای پیش‌بینی سری زمانی، روش‌های MA و ES بوده است که از سال ۱۹۷۱ مورداستفاده قرارگرفته است (کریستیانز[1]، ۱۹۷۱؛ ۹۰۰-۹۱۱). این روش‌ها البته ساده‌اند و مبتنی بر میانگین متحرک و نرم‌سازی نمایی[2] هستند. این روش ها در سال‌های بعد (از سال ۱۹۹۰) با روش بهتر ARIMA جایگزین شدند (میلز[3]، ۱۹۹۰).

روش ARIMA ترکیبی از میانگین متحرک و رگرسیون است و برای مدل‌سازی سری‌های زمانی نتایج خوبی داشته و دارد (میلز، ۱۹۹۰) و (دب[4]، ۲۰۱۷؛ ۹۲۴-۹۰۲). این روش همچنان بسیار رایج است و در موارد مختلف ازجمله کشاورزی مورداستفاده قرار می‌گیرد (دسموخ، ۲۰۱۶؛ ۲۲-۱۷).

[1] Christiaanse
[2] Exponential Smoothing
[3] Mills
[4] Deb

تقریباً از سال ۱۹۹۶ روش‌های مبتنی بر شبکه‌های عصبی ساده مثل پرسپترون برای این منظور مورد استفاده قرار گرفتند (گنزالس[1]، ۲۰۰۵؛ ۶۰۱-۵۹۵) (نیزامی[2]، ۱۹۹۵؛ ۱۰۴-۲۳). گرچه لازم به ذکر است که شبکه‌های عصبی خیلی قبل‌تر ارائه شدند ولی سال‌ها طول کشید تا به بلوغ لازم برسند و برای مدل‌سازی سری زمانی استفاده شوند (دب، ۲۰۱۷؛ ۹۲۴-۹۰۲).

تقریباً از سال ۲۰۰۴ روش‌های نسبتاً جدید ماشین بردار های پشتیبان (درا کر و همکاران، ۱۹۹۶؛ ۱۶۱-۱۵۵) برای مدل‌سازی سری زمانی به‌کار برده شدند (دب، ۲۰۱۷؛ ۹۲۴-۹۰۲) (چانگ[3]، ۲۰۰۴؛۳۰-۱۹) (لی کیو[4]، ۲۰۰۹؛ ۵۰-۶).

در ادامه، از سال ۲۰۱۰ به بعد روش‌های مبتنی بر شبکه‌های عصبی عمیق رایج شده و برای مدل‌سازی سری زمانی استفاده شدند و در جدیدترین تحقیق‌ها روش‌های مبتنی بر شبکه‌های عصبی عمیق بازگشتی مثل LSTM استفاده می‌شوند (چن[5]، ۲۰۱۵؛ ۲۸۲۸-۲۸۲۳)

اگر به‌طور خاص، استفاده از این تحلیل‌ها و روش‌ها برای کشاورزی در نظر گرفته شود، می‌توان گفت که روش ARIMA بیش از همه روش‌ها برای این منظور استفاده شده است (حسینی و همکاران، ۱۳۸۶؛ ۵۴-۴۱) (پرویز و همکاران، ۱۳۹۷؛ ۴۶-۳۵). برای مثال می توان به تحقیق (دسموخ، ۲۰۱۶؛ ۲۲-۱۷) اشاره کرد که میزان تولید شیر را در هند پیش‌بینی می‌کند. در ایران نیز در سال‌های اخیر با پیشرفت هوش مصنوعی، تقریباً از سال ۲۰۱۰ تحقیقات در این زمینه شروع شده است (حسینی و همکاران، ۱۳۸۶؛ ۵۴-۴۱) (رحمانی و همکاران، ۱۳۸۷؛ ۲۵-۳۷). البته لازم به ذکر است که بیشتر این تحقیقات با هدف بررسی تأثیر بارندگی و خشک‌سالی بر میزان تولید انجام شده‌اند ولی به هر حال روش های مورد استفاده همان

[1] Gonzales
[2] Nizami
[3] Chang M.W
[4] Li Q
[5] Chen

روش‌های تحلیل سری‌های زمانی و به‌طور خاص روش ARIMA و حالت‌های دیگر آن مثل SARIMA است.

جایگاه تحقیق در مدیریت راهبردی: برنامه‌ریزی راهبردی طبق (مینتزبرگ[1] و همکاران، ۱۹۹۶) عبارت است از مجموعه‌ای از روش‌ها و مدل‌ها برای تعیین راهبرد و جهت یک سازمان و تصمیم‌گیری در مورد چگونگی صرف کردن منابع برای رسیدن به اهداف و رقابت در محیطی که فعالیت می‌کند. برنامه‌ریزی راهبردی بخش مهمی از مدیریت راهبردی است (مینتزبرگ و همکاران، ۱۹۹۶). یکی از بخش‌های اساسی مدیریت راهبردی، تحلیل فرصت‌ها و تهدیدها است (پورتر[2]، ۱۹۸۰) و این دقیقاً بخشی است که نویسندگان در این تحقیق علاقه‌مندند از طریق روش‌های هوش مصنوعی روی آن کار کنند.

مروری بر ادبیات پایه‌ای سری‌های زمانی

مسئله پیش‌بینی داده‌های زمانی، عبارت است از مدل‌سازی مقادیر گذشته یک متغیر برای پیش‌بینی مقدار آن در آینده (ادهیکاری[3]، ۲۰۱۳)؛ بنابراین مدل‌سازی متغیر زمان (که معمولاً با t نشان داده می‌شود) یک بخش اصلی از این روش‌هاست (ادهیکاری، ۲۰۱۳). یک سری زمانی، عبارت است از یک متغیر تصادفی با تعریف X(t) که در محور زمان مقدار می‌گیرد. اگر یک سری زمانی متشکل از صرفاً یک متغیر باشد، مسئله تک-متغیره و اگر متشکل از بیش از یک متغیر باشد، مسئله چندمتغیره خواهد بود (ادهیکاری، ۲۰۱۳). یک مشخصه سری‌های زمانی، فاصله زمانی بین مقدارهای آن است. مثلاً اگر مقدارهای سری زمانی برای هرروز موجود باشند، نرخ نمونه آن روزانه است و اگر برای هر سال موجود باشند؛ سالانه است.

[1] Mintzberg
[2] Porter
[3] Adhikari

همچنین لازم به توضیح است که معمولاً داده‌های زمانی از نوع مقادیر حقیقی هستند (مثل قیمت در بورس یا داده‌های تولید و فروش کشاورزی که در این تحقیق به آن می‌پردازیم) و درنتیجه برای مدل‌سازی داده‌های زمانی به روش‌هایی از یادگیری ماشین که برای رگرسیون طراحی‌شده‌اند نیاز است (بیشاپ[1]، ۲۰۰۶). در مقابل این روش‌ها، روش‌های دسته‌بندی هستند (که در آن‌ها مقادیر دسته‌ای[2] هستند) که در این تحقیق نمی‌گنجند. از نگاهی دیگر سری‌های زمانی را می‌توان به گسسته و پیوسته نیز تقسیم کرد که از بحث این تحقیق خارج است.

به‌طورکلی منحنی‌های زمانی می‌توانند از چهار مؤلفهٔ اصلی تشکیل شوند:

- روند[3]: تمایل کلی یک سری زمانی برای افزایش یا کاهش در طولانی مدت، روند آن سری را نشان می‌دهد (ادهیکاری، ۲۰۱۳)؛ بنابراین روند، جریان و میانگین کلی و درازمدت یک سری زمانی را نشان می‌دهد.
- مشخصه فصلی: همان‌طور که از اسم این مؤلفه پیداست، تغییرات و بالا و پایین‌هایی که در طول سال در اثر فصل‌ها یا نرخ‌های کمتر رخ می‌دهد را نشان می‌دهد. برای مثال اگر نرخ نمونه یک سری زمانی بیشتر از سال باشد، این مؤلفه وجود نخواهد داشت. (ادهیکاری، ۲۰۱۳) (دب، ۲۰۱۷؛ ۹۲۴-۹۰۲)
- مشخصه دوره‌ای: مشابه مشخصهٔ فصلی است ولی طول دوره آن بیشتر است و مثلاً در طول ۲ یا ۳ سال رخ می‌دهد. خیلی از داده‌های زمانی مربوط به امور مالی و اقتصادی دارای این مشخصه هستند.

[1] Bishop
[2] Categorical
[3] Trend

- مؤلفه بی‌نظمی: نوسانات نامنظم و تصادفی در یک سری زمانی در اثر اتفاقات و اثرات پیش‌بینی‌نشده است که نظم و الگوی خاصی ندارند. مثال‌هایی از این اتفاقات، جنگ، زلزله، سیل و غیره است. (ادهیکاری، ۲۰۱۳)

به‌طور خلاصه، اگر یک منحنی در طول زمان رشد یا کاهش داشته باشد، اصطلاحاً گفته می‌شود روند دارد (ادهیکاری، ۲۰۱۳). در مقابل، در خیلی از مسائل، داده‌های زمانی روند خاصی ندارند (میانگین تقریباً ثابت است) ولی الگوی خاصی با نرخ مشخص تکرار می‌شود. در این حالت، اصطلاحاً گفته می‌شود که منحنی فصلی[1] است. اگر دوره تکرارها طولانی باشد، گفته می‌شود که منحنی مشخصه دوره‌ای[2] دارد (ادهیکاری، ۲۰۱۳). نهایتاً، ممکن است یک نمودار در بازه‌های مشخصی نه روند داشته باشد و نه فصلی/دوره‌ای باشد و اصطلاحاً نامنظم باشد. بسته به هر کدام از این خصیصه‌ها، روش‌هایی هستند که بهتر یا بدتر عمل می‌کنند. بعضی از روش‌ها (عموماً روش‌های یادگیری ماشین) نیز مستقل از نوع منحنی می‌توانند استفاده شوند.

دو رویکرد کلی برای ترکیب این مؤلفه‌ها وجود دارد: (ادهیکاری، ۲۰۱۳) رویکرد ضربی و رویکرد جمعی. در رویکرد ضربی، فرض می‌شود که یک سری زمانی از ضرب این چهار مؤلفه ساخته می‌شود. در این رویکرد، این مؤلفه‌ها مستقل از هم نیستند:

$$x(t) = T(t) * S(t) * C(t) * I(t)$$

[1] Seasonal
[2] Cyclical

که در آن، T مؤلفه روند، S مؤلفه فصلی، C مؤلفه دوره‌ای و I مؤلفه بی‌نظمی را نشان می‌دهند. در رویکرد جمعی، فرض می‌شود که سری زمانی از جمع این مؤلفه‌ها ساخته می‌شود و درنتیجه فرض می‌شود که این‌ها از هم مستقل هستند:

$$x(t) = T(t) + S(t) + C(t) + I(t)$$

معیارهای ارزیابی

هر مسئله یادگیری ماشین، نیاز به یک روش مشخص ارزیابی دارد تا بتوان عملکرد مدل‌ها و روش‌های مختلف را سنجید و مقایسه کرد (بیشاپ، ۲۰۰۶). برای مسئله سری‌های زمانی نیز همین طور است. روش معمول مدل‌سازی این است که یک بخش از سری زمانی به عنوان داده آموزشی انتخاب می‌شود (دب، ۲۰۱۷؛ ۹۲۴-۹۰۲). این داده آموزشی برای آموزش مدل یا اصطلاحاً برازش مدل استفاده می‌شود. مثلاً اگر سری زمانی از سال ۱۹۶۰ با نرخ سالانه تا سال ۲۰۰۰ مقدار دارد، بازه ۱۹۶۰ تا ۱۹۹۵ را می‌توان به عنوان داده آموزشی استفاده کرد. بقیه سری زمانی به عنوان (تست‌ست) انتخاب می‌شود. بعد از آموزش مدل، مقادیر سری زمانی برای ۵ سال آخر (تست‌ست) پیش‌بینی می‌شود. معیار ارزیابی مشخص می‌کند که مقادیر پیش‌بینی شده (تست‌ست) را چطور با مقادیر واقعی آن مقایسه کنیم (ادهیکاری، ۲۰۱۳). در ادامه مروری بر رایج‌ترین معیارهای ارزیابی انجام خواهد شد.

برای نشانه‌گذاری فرض کنید که y_t مقدار واقعی نمونه تستی را نشان می‌دهد، y_t مقدار پیش‌بینی‌شده را نشان می‌دهد، $e_t = y_t - f_t$ خطای پیش‌بینی را نشان می‌دهد و n تعداد نمونه‌های تستی را. همین طور $\bar{y} = \frac{1}{n}\sum_t y_t$ میانگین تست و $\sigma^2 = \frac{1}{n-1}\sum_t (y - \bar{y})^2$ واریانس تست را نشان می‌دهد.

- معیار MAE:

این معیار به صورت $\frac{1}{n}\sum_{t=1}^{n}|e_t|$ تعریف می‌شود (تای[1]، ۲۰۰۳؛ ۱۵۱۸-۱۵۰۶) (همزاچبی[2]، ۲۰۰۸؛ ٤٥٥٩-٤٥٥٠). این معیار مشابه MFE استبا این تفاوت که علامت‌دار نیست و میانگین قدر مطلق خطا را نشان می‌دهد. این معیار و معیار MFE هر دو تحت تأثیر مقیاس سری زمانی هستند و با تغییر مقیاس، تغییر می‌کنند که این مطلوب نیست. این معیار، خطاهای شدید مدل را در میانگین نشان نمی‌دهد (ادهیکاری، ۲۰۱۳).

- معیار MAPE

این معیار یکی از رایج‌ترین معیارهاست که نسبت خطا را نشان می‌دهد و درنتیجه تحت تأثیر مقیاس نیست (ادهیکاری، ۲۰۱۳) (تای، ۲۰۰۳؛ ۱۵۱۸-۱۵۰۶) (پرویز و همکاران، ۱۳۹۷؛ ٤٦-۳۵). این معیار به صورت $\frac{1}{n}\sum_{t=1}^{n}|\frac{e_t}{y_t}| * 100$ (تای، ۲۰۰۳؛ ۱۵۱۸-۱۵۰۶). مشابه MAE جهت خطا در این معیار نشان داده نمی‌شود.

- معیار MSE

معیار میانگین خطای مربعات یکی از معروف‌ترین معیارهای خطا در یادگیری ماشین و آمار است (ادهیکاری، ۲۰۱۳) (بیشاپ، ۲۰۰٦) و به صورت $\frac{1}{n}\sum_{t=1}^{n}e_t^2$ تعریف می‌شود. برخلاف معیارهای قبلی (که مربعات در آن‌ها نبود)، این معیار خطاهای شدید را به خوبی نشان می‌دهد (ادهیکاری، ۲۰۱۳). ولی به مقیاس حساس است و جهت خطا را نیز نشان نمی‌دهد. این معیار شهود خیلی خوبی در مورد عملکرد نمی‌دهد چراکه مربع است.

- معیار RMSE

[1] Tay
[2] Hamzacebi

این معیار همان MSE است ولی با این تفاوت که از آن جذر می‌گیرد (ادهیکاری، ۲۰۱۳). درنتیجه مشکل شهودی نبود MSE را تا حد خوبی حل می‌کند. به‌طور خلاصه، این معیار هم شهود خوبی در مورد خطا می‌دهد و هم خطاهای شدید مدل را به خوبی تأثیر می‌دهد.

روش‌شناسی

نوع این تحقیق توسعه‌ای-کاربردی است چراکه بخش زیادی از آن شامل تحقیق برای توسعهٔ روش‌هایی است که بتواند به خوبی مدل‌سازی پیش‌بینی را انجام دهد و خروجی نهایی کاربردی است. کاربرد آن کمک به مسئولان کشوری برای تصمیم‌گیری در مورد کشاورزی است. شیوه گردآوری اطلاعات غیر آزمایشگاهی است و همچنین مبنای رویکرد پژوهش کمی است چراکه این تحقیق با اعداد و ارقام و ارزیابی کمی مدل‌های یادگیری ماشین سروکار دارد.

روش تحقیق به صورت کتابخانه‌ای و تجربی است. به‌طور خلاصه ابتدا دادگان فائواستات که شامل انواع داده‌های کشاورزی است را آماده می‌کنیم. سپس این دادگان تمیزسازی و مرتب‌شده و بررسی می‌شود که چه به خش‌هایی از دادگان برای هدف این تحقیق (برنامه‌ریزی راهبردی برای سود بیشتر کشاورزی) مفید هستند. این داده‌ها تماماً به شکل سری‌های زمانی هستند و افزایش سود کشاورزی منوط به مدل‌سازی هرچه دقیق‌تر این سری‌های زمانی خواهد بود. بنابراین در ادامه، انواع روش‌های یادگیری ماشین و آماری/تصادفی برای مدل‌سازی سری‌های زمانی و پیشینه آن‌ها مرور می‌شوند و در پایان طبق این روش‌ها تحلیل و پژوهش دادگان انجام خواهد شد تا بهترین دقت به دست بیاید.

برای اندازه‌گیری دقت، از روش‌های شناخته‌شده ارزیابی عملکرد که در پیشینه بررسی و ارائه خواهند شد، استفاده می‌شود.

جامعه آماری و ابزار آمار/گردآوری اطلاعات

جامعه آماری اصلی و درواقع منبع اصلی اطلاعات در این پژوهش، دادگان سازمان جهانی فائو می‌باشد. کشورهای مختلف داده‌های کشاورزی خودشان را سالانه به این سازمان می‌فرستند و درنتیجه اطلاعات موجود در سایت این سازمان بسیار گسترده و کامل و همچنین به صورت رایگان قابل‌دسترسی می‌باشند. نام این دادگان فائواستات است که درواقع تمرکز اصلی این تحقیق روی این دادگان است. ازجمله به خش‌هایی از این دادگان که در راستای هدف این تحقیق است نمودارهای زمانی تولید و فروش اقلام کشاورزی برای کشورهای مختلف است که بر اساس آن‌ها می‌توان مدل‌های پیش‌بینی‌کننده ساخت و با پیش‌بینی آینده، برنامه‌ریزی راهبردی برای کشاورزی داشت.

روش تجزیه و تحلیل داده‌های تحقیق و روش اعتبارسنجی داده‌های گردآوری‌شده

هدف این است که برای تحلیل داده‌ها از روش‌های تصادفی مثل ARIMA و روش‌های کلاسیک یادگیری ماشین مثل درخت‌های تصمیم، رگرسیون خطی و SVM و همچنین روش‌های پیشرفته مثل شبکه‌های عصبی عمیق (شبکه‌های بازگشتی، چراکه داده‌ها زمانی هستند) استفاده شود.

ازآنجایی‌که از دادگان مشخص فائو استفاده می‌کنیم، نیاز به اعتبارسنجی نخواهد بود. با این حال برای مطمئن شدن از درستی داده‌ها، از روش چک کردن متقابل استفاده خواهد شد. همچنین محدوده‌ٔ عددها نیز چک می‌شوند تا داده‌های پرت پیدا شوند.

تجزیه و تحلیل

در این بخش از تحقیق به بررسی به خش‌های مختلف دادگان فائواستات و سپس تحلیل سری‌های زمانی آن پرداخته خواهد شد. این کار در سه زیر بخش انجام می‌شود. در زیر بخش اول، به مرور و بررسی به خش‌های مختلف دادگان فائواستات پرداخته خواهد شد. در زیر بخش دوم یک تحلیل استوکستیک از چند سری زمانی از این دادگان ارائه می‌شود و مدل ARIMA مناسب پیدا می‌شود. در ادامه، آزمایش‌های روش‌های یادگیری ماشین و مقایسه با روش ARIMA انجام خواهد شد.

بررسی دادگان فائواستات

دادگان فائو که با نام فائواستات شناخته می‌شود، یک مجمو عه داده از دنبا لـه‌های ز مانی از سال ۱۹۶۱ میلادی است که توسط سازمان فائو جمع‌آوری و ن گه‌داری می شود. ا ین داد گان برای ۲۴۵ کشور دیتا دارد، گرچه داده‌های کشورهای مخت لف لزو ماً کا مل نی ست. همچ نین میزان خطا در دادگان برای بعضی از کشورها بیش از دیگر کشورها است.

به‌طور خلاصه این دادگان شامل آماره‌های مختلف (تولید، فروش، برداشت، ...) در زمینه‌های مختلف برای کشورهای مختلف و اقلام مختلف است. این آماره‌ها به صورت دنباله‌های زمانی از سال ۱۹۶۱ جمع‌آوری‌شده‌اند. زمینه‌های مختلف این دادگان در ادامه توضیح داده می‌شوند:

- **تولید محصول**: این زمینه شامل میزان تولید و قیمت فروش اقلام مختلف برای کشورهای مختلف می‌شود که به صورت سالانه ارائه می‌شود و درواقع بخش اصلی دادگان لازم برای این تحقیق است.
- **تجارت**: این زمینه شامل دیتاهای صادرات و واردات اقلام مختلف برای کشورهای مختلف است.
- **امنیت غذا**: این زمینه شامل اطلاعات لازم برای جلوگیری از کمبود غذا و مشکلات غذایی می‌شود.
- **نشانگرهای زیست‌محیطی**: این زمینه شامل نشانگرهایی است که می‌تواند به دولت‌ها و دانشمندان کمک کند تا تصمیم‌های بهتری برای حفظ محیط‌زیست بگیرند. برای مثال تأثیر سیاست‌های مختلف روی محیط‌زیست و میزان تولید.
- **قیمت‌ها**: این زمینه شامل قیمت‌های مرجع و ایندکس برای تولیدکننده‌ها و خریداران است.
- **منابع**: این زمینه مربوط به توزیع زمین در کشورها و وضعیت زمین از نظر باروری و غیره می‌باشد.
- تشعشعات گازهای مضر: این زمینه شامل اطلاعات مربوط به تولید گازهای گلخانه‌ای می‌باشد. دو بخش اصلی در این زمینه وجود دارد که باعث تولید گاز گلخانه‌ای می‌شود: ۱. کشاورزی: تأثیر کودها و روش‌های مختلف کشاورزی در تولید گازهای گلخانه‌ای. ۲. زمین: تأثیر استفاده از زمین‌های مختلف (جنگلی، غیره) بر تولید گازهای گلخانه‌ای.
- **سرمایه‌گذاری**: این زمینه شامل دیتاهای مربوط به قیمت سهام مربوط به کشاورزی است.
- **جنگل**: این زمینه شامل اطلاعات تولید برای محصولات مختلف مربوط به درختان است، ازجمله چوب و کاغذ که به صورت سالانه جمع‌آوری و ارائه می‌شود.

دادگان فائواستات به صورت یک فایل زیپ از سایت فائو قابل دانلود است. حجم این فایل ۸۵۳ مگابایت است[۱]. داخل این فایل، ۷۸ فایل فشرده‌شده وجود دارد. هر کدام از این ۷۸ فایل، درواقع یک صفحه گسترده است با فرمت csv که اطلاعات مشخصی را ذخیره کرده است. با توجه به بررسی‌هایی که انجام شد، از این ۷۸ فایل، ۷ فایل دیتای لازم برای این تحقیق را تشکیل می‌دهد. این فایل‌ها عبارت‌اند از ٤ فایل برای میزان تولید، ۱ فایل که ارزش تولید را نشان می‌دهد، یک فایل برای میزان کل صادرات/واردات کشورها و نهایتاً یک فایل برای جزییات تجارت بین کشورها. میزان تولید به تن است و ارزش تولید به دلار. ازآنجایی‌که معمولاً حجم تولید زیاد اهمیت ندارد و بیشتر ارزش تولید مدنظر است، در به خش‌های بعد بیشتر تمرکز بر ارزش تولید خواهد بود.

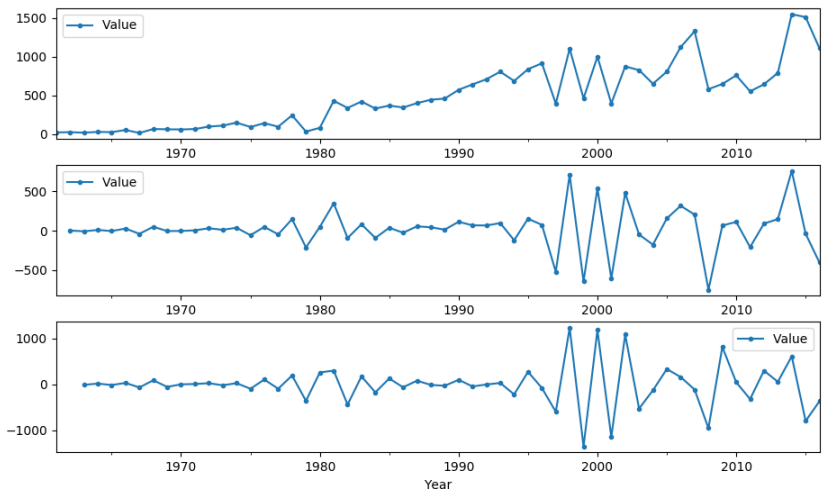
تحلیل استوکستیک سری‌های زمانی فائواستات

همان‌طور که در زیر بخش قبل گفته شد، دو بخش ارزش میزان تولید و صادرات/واردات از دادگان فائواستات، به سودآوری مربوط هستند و برای هدف این تحقیق مناسب‌اند. هر کدام از این به خش‌ها در دادگان فائواستات، شامل هزاران سری زمانی برای محصولات مختلف در کشورهای مختلف هستند. از این میان، دو سری زمانی را انتخاب کرده و در اینجا تحلیل استوکستیک انجام خواهد شد. سری اول مربوط به ارزش تولید پسته در ایران و سری دوم مربوط به ارزش تولید پسته در آمریکا است. این انتخاب از این جهت است که پسته یکی از اقلام پرسود برای کشاورزی ایران است و آمریکا یک رقیب جدید در این بازار جهانی است.

[۱] برای سال ۲۰۱۸.

برای هر کدام از این سری‌های زمانی، سه تحلیل مشتق زمانی، نرمال بودن و ACF و روش AIC برای پیدا کردن بهترین مدل ARIMA انجام می‌شود.

در نمودار شکل ۱ منحنی اصلی و مشتق درجه ۱ و ۲ سری زمانی تولید پسته در ایران مشاهده می‌شود.

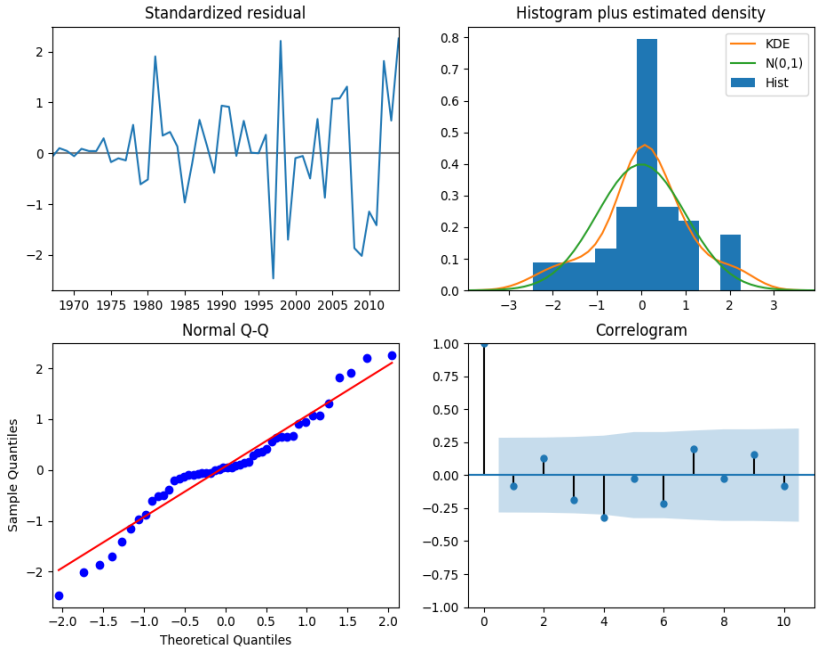


شکل ۱ مشتق اول و دوم سری زمانی ارزش تولید پسته در ایران طبق فائواستات.

می‌توان دید که میانگین مشتق اول و دوم تقریباً در طول زمان ثابت است که تا حدی ایستایی را نشان می‌دهد. البته واریانس در بعضی از نقاط (نزدیک انتها) تغییر داشته است با زمان که منطبق با ایستایی نیست؛ بنابراین مقادیر d=1 و d=2 هر دو در ARIMA مشابه عمل خواهند کرد.

در ادامه تحلیل AIC انجام می‌شود و بهترین مدل ARIMA پیدا می‌شود. این مدل از مرتبه (3,1,3) است و AIC به‌دست‌آمده برابر با ۶۹۰ است (واحد مقادیر نمودار میلیون دلار است).

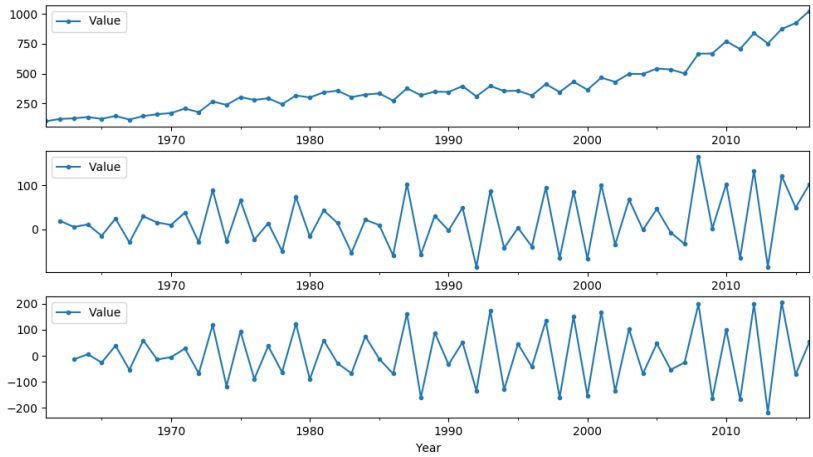
سپس تحلیل ACF و نرمال بودن را برای تولید پسته ایران انجام می گردد که در شکل ۲ نشان داده‌شده است.



شکل ۲ تحلیل نرمال بودن و ACF برای ارزش تولید پسته ایران.

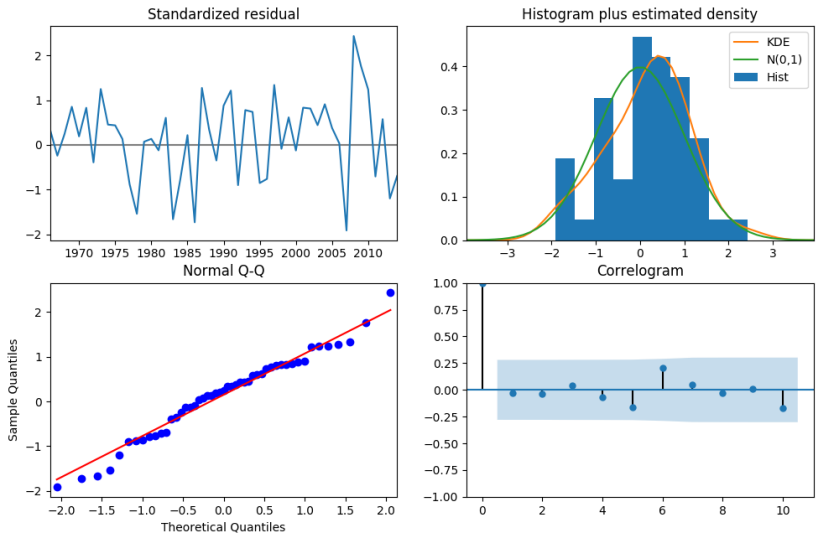
این شکل از چهار نمودار تشکیل‌شده است. نمودار بالا سمت چپ، باقیمانده‌های سری زمانی نسبت به نمودار ARIMA است. نمودار بالا سمت راست، یک هیستوگرام روی این مقادیر باقیمانده رسم کرده است که درواقع توزیع آن را نشان می‌دهد. طبق این نمودار می‌توان دید که با تخمین نسبتاً خوبی باقیمانده‌ها نرمال هستند. نمودار پایین سمت چپ نیز انطباق توزیع باقیمانده‌ها بر نمودار نرمال را نشان می‌دهند. می‌توان دید که انطباق کامل نیست ولی با تقریب خوبی به خط راست (انطباق کامل) نزدیک است. نهایتاً، نمودار پایین سمت راست، ACF را نشان می‌دهد. می‌توان دید که بیش‌ترین رابطه با اختلاف ٤ و ٦ و ٧ سال بین مشاهدات سری زمانی وجود دارد که نوعی بی‌نظمی به نظر می‌رسد و می‌توان گفت که بخش AR مدل احتمالاً کمک زیادی به مدل‌سازی نخواهد کرد.

به همین ترتیب، نمودار سری زمانی ارزش تولید پسته آمریکا به همراه مشتق اول و دوم در شکل ۳ دیده می‌شود. می‌توان دید که مشابه سری قبلی، مشتق اول و دوم هر دو میانگین تقریباً ثابت دارند با این تفاوت که میزان تغییرات واریانس در زمان کمتر است و درنتیجه با تخمین بهتری سری زمانی ایستا است.



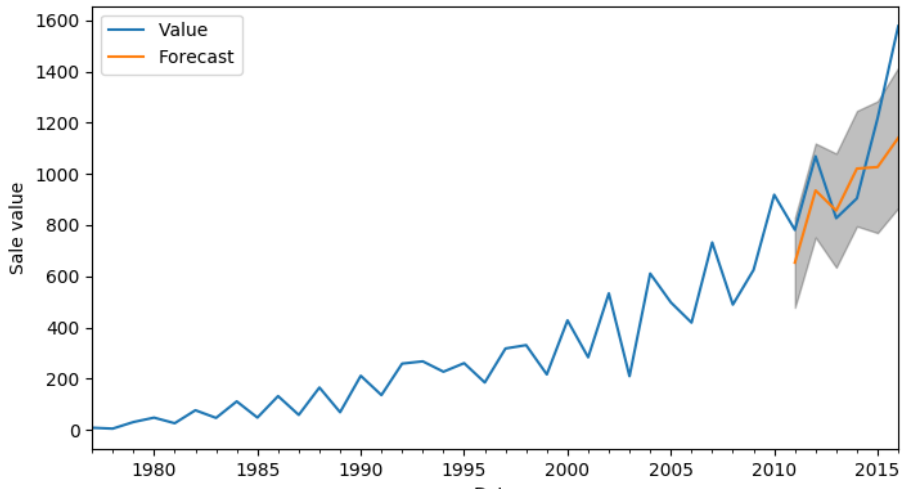
شکل ۳ مشتق اول و دوم سری زمانی تولید ارزش تولید پسته آمریکا.

مشابه سری قبل، تحلیل نرمال بودن برای این سری زمانی نیز در شکل ۴ نشان داده‌شده است. اولین چیزی که به چشم می‌آید این است که میزان نرمال بودن بیشتر است. این مسئله باعث تخمین بهتر و برازش بهتر مدل ARIMA می‌شود که در مقدار AIC (برابر با ٤٢٣) نیز پیداست.



شکل ٤ تحلیل نرمال بودن و ACF برای تولید پسته آمریکا.

همچنین می‌توان دید که رابطهٔ زیادی بین سال‌های متوالی وجود ندارد. همبستگی کمی بین فاصلهٔ سالی ۵ و ٦ وجود دارد که عملاً ناچیز است. اگر با این مدل، برای تولید پسته آمریکا پیش‌بینی انجام شود نمودار شکل ۵ حاصل می‌شود که دارای خطای MAPE برابر با ٦,۵ درصد است.



شکل ۵ پیش‌بینی ارزش تولید پسته در آمریکا از ۲۰۱۲ تا ۲۰۱۵ با روش ARIMA

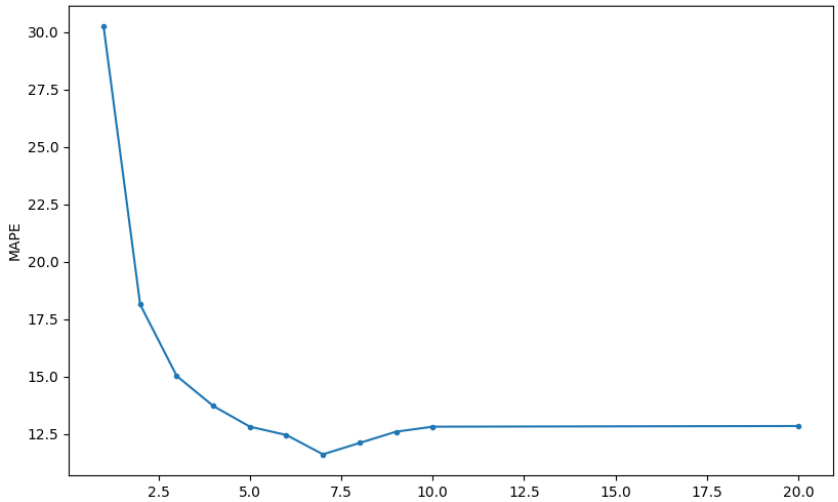
آزمایش روش‌های یادگیری ماشین

برای ارزیابی آزمایش‌ها از دو معیار رایج استفاده می‌شود: یکی RMSE است که درواقع ریشهٔ میانگین خطای مربعات است و به‌طور خلاصه برای یک سری زمانی نشان می‌دهد که به‌طور متوسط چقدر در پیش‌بینی مقدار بعدی دنباله خطای مطلق است. معیار دیگر، MAPE است که درواقع معیار خطای مطلق را به مقدار مرجع تقسیم می‌کند و با واحد درصد نشان داده می‌شود.

آزمایش‌های مختلف روی ترکیبی از سری‌های زمانی ارزش تولید اقلام کشور ایران، چین و آمریکا انجام می‌شود. کشور چین به این دلیل که بیش‌ترین واردات را از ایران دارد و کشور آمریکا به این دلیل که ممکن است رقیب ما در پسته و محصولات دیگر باشد و خوب است که بتوانی دیتای کشاورزی‌اش را پیش‌بینی و مدل‌سازی کرد. این دیتاست ترکیبی متشکل از ۳۸۹ سری زمانی با طول متوسط ٦٠ نمونه است. ۱۰٪ از این دادگان برای میزان سازی استفاده

می‌شود. روش آزمایش به این صورت است که از هر سری زمانی، به جز یک مقدار پایانی (که مربوط به سال اخیر است) در آموزش استفاده می‌شود. برای تست، این ۱ مقدار پایانی با دادن مقادیر سال‌های قبل به عنوان ورودی، پیش‌بینی می‌شود. هر نمونه آموزشی، متشکل از n مقدار متوالی به عنوان ویژگی‌های ورودی و مقدار بعدی به عنوان هدف (برچسب خروجی) است که در آن n طول بافت می‌باشد. لازم به ذکر است که برای مدل‌سازی روش‌های یادگیری ماشین، ابتدا مقیاس سری‌های زمانی نرمال‌سازی می‌شوند تا برازش بهتر انجام شود. نرمال‌سازی با استفاده از روش رایج MinMax انجام می‌شود.

روش‌هایی که در اینجا بررسی می‌شوند عبارت‌اند از روش افزایش گرادیان، رگرسیون خطی، روش ARIMA و نهایتاً شبکه عصبی چند لایه. روش پایه ARIMA در زیر بخش قبلی میزان سازی شد و بررسی‌های آن انجام شد. برای روش افزایش گرادیان، نیاز است که هایپرپارامتر بیشینه عمق میزان سازی شود. برای این کار، روی دادگان میزان سازی، چندین مقدار مختلف برای بیشینه عمق این روش ارزیابی شدند. شکل ٦ تأثیر این هایپرپارمتر بر معیار MAPE را نشان می‌دهد. می‌توان دید که بهترین مقادیر به ازای عمق‌های ۳ تا ۷ به‌دست‌آمده‌اند. برای شبکه عصبی، از تابع فعال‌سازی ReLU استفاده می‌شود. با توجه به محدود بودن دیتا، شبکه باید کوچک باشد تا از بیش برازش جلوگیری شود. چندین اندازه لایه و تعداد لایه بررسی شدند و نهایتاً، یک شبکه با ۳ لایه مخفی و تعداد ۳۰۰ نورون در هر لایه انتخاب شد. برای رگرسیون خطی از روش بهینه‌سازی کمینه–مربعات استفاده می‌شود.



شکل ٦ تأثیر بیشینه عمق بر MAPE درروش افزایش-گرادیان.

با توجه به میزان سازی اولیه‌ای که انجام شد، مدل‌سازی و ارزیابی طبق شرایط آزمایش که بالاتر توضیح داده شد انجام می‌گردد. نتایج آزمایش در جدول ۱ نشان داده‌شده‌اند. در این آزمایش‌ها طول بافت طبق یک سری آزمایش اولیه که در اینجا نشان داده نشده است، برابر با ۵ در نظر گرفته‌شده است.

جدول ۱ نتایج نهایی روش‌های بررسی‌شده.

روش	خطای RMSE	خطای MAPE ٪
ARIMA	818	13.7
رگرسیون خطی	948	18.9
افزایش-گرادیان	914	11.3
DNN	2105	24.2

بحث

شبکه عصبی چند لایه در بین روش‌های بررسی‌شده، ضعیف‌ترین عملکرد را داشته است. علی‌رغم اینکه شبکه‌های عصبی توانمندی زیادی برای مدل‌سازی دارند، نیاز آن‌ها برای دیتا نیز بسیار زیاد است. با توجه به اینکه سری‌های زمانی فائواستات سالانه هستند و فقط ۵۰ تا ۶۰ مقدار دارند، این دیتا کم است و بنابراین نتیجه‌ای که مشاهده می‌شود، دور از انتظار نیست. بهترین عملکرد MAPE (خطای ٪۱۱,۳) مربوط به روش افزایش-گرادیان است که درواقع گسترشی بر درخت‌های تصمیم است. بهترین عملکرد RMSE ولی مربوط به روش ARIMA است (مقدار ۸۱۸). روش رگرسیون خطی گرچه MAPE خوبی کسب نکرده است ولی مقدار RMSE نسبتاً خوبی در مقایسه با دو روش دیگر کسب کرده است. در مجموع می‌توان گفت که روش افزایش-گرادیان و ARIMA بهترین نتایج را داشته‌اند.

نتیجه‌گیری و پیشنهاد

نتیجه‌گیری

در این تحقیق، با رویکرد هوش مصنوعی و مدیریت راهبردی به دادگان فائواستات پرداخته شد. این دادگان، یک مجموعه داده بین‌المللی است که شامل اطلاعات مختلفی در حوزهٔ کشاورزی است. در این مقاله، اولاً به خش‌هایی از این دادگان که برای افزایش سود در حوزهٔ کشاورزی می‌توانند مفید باشند، شناسایی شدند. این به خش‌ها شامل میزان تولید اقلام مختلف کشاورزی و میزان صادرات/واردات آن‌ها است. دوما، تحلیل استوکستیک روی تعدادی از سری‌های زمانی انجام شد. به نظر می‌رسد باقیمانده‌های این سری‌های زمانی نرمال هستند و همچنین مشتق اول آن‌ها با تقریب خوبی ایستا است. همچنین مشاهده شد که سری‌های زمانی مربوط به ایران دارای نویز نسبتاً بالایی هستند. درنهایت، چندین روش یادگیری ماشین و روش ARIMA روی تعداد زیادی از سری‌های زمانی این دادگان برازش شده و ارزیابی شد و

نشان داده شد که با میزان سازی مناسب می‌توان به دقت متوسط MAPE برابر با ۸۸٫۷٪ رسید. سه روش ARIMA و رگرسیون خطی و افزایش گرادیان بهترین عملکرد را داشتند.

نتیجهٔ این تحقیق، می‌تواند برای پیش‌بینی تولید/صادرات/واردات اقلام مختلف موجود در دادگان فائواستات برای کشورهای رقیب ایران در حوزهٔ کشاورزی استفاده شود. درواقع، با پیش‌بینی مقادیر چند سال آینده محصولات کلیدی برای کشورهای مختلف، مدیران راهبردی در حوزهٔ کشاورزی می‌توانند تصمیمات مطلع‌تری اخذ کنند.

پیشنهاد

با توجه به عملکرد خوب سه روش رگرسیون خطی، ARIMA و افزایش گرادیان، یک پیشنهاد برای آینده این است که ترکیب این روش‌ها استفاده شود. پیشنهاد دیگر، ترکیب سری‌های زمانی مشابه برای ساخت مدل‌های یادگیری ماشین است. درواقع، ازآنجایی‌که یک سری زمانی متشکل از تعداد کمی نمونه است (برای فائواستات معمولاً ۵۰ تا ۶۰ نمونه)، می‌توان برای مدل‌سازی یک محصول مشخص، محصول‌های مشابه را نیز مورداستفاده قرار داد تا دیتای بیشتری در اختیار مدل قرار بگیرد.

منابع

Adhikari, Ratnadip (2013), "An Introductory Study on Time Series Modeling and Forecasting", Germany, Lambert Academic Publishing.

Bishop, Christopher M. (2006), Pattern Recognition and Machine Learning, USA, Springer.

Chang, Ming-Wei (Nov. 2004), Load forecasting using support vector machines: a study on EUNITE competition, Power Syst IEEE Trans, Vol. 19, No. 4.

Chen, Kai (Dec. 2015), A LSTM-based method for stock returns prediction: A case study of China stock market, IEEE international conference on big data.

Christiaanse, W. R (March 1971), Short-term load forecasting using general exponential smoothing, IEEE Transactions on Power Apparatus and Systems, Vol. 90, No. 2.

Deb, Chirag (July 2017), A review on time series forecasting techniques for building energy consumption, Renewable and Sustainable Energy Reviews, Vol 74, No.1.

Deshmukh, Sagar (March 2016), Forecasting of milk production in India with ARIMA and VAR time series models .Asian Journal of Dairy and Food Research, Vol. 35, No. 1.

Drucker, Harris, Burges, Christopher, Kaufman, Linda (July 1997), Support Vector Regression Machines .Advances in Neural Information Processing Systems, Vol. 88, No. 3.

González, Pedro (June 2005), Prediction of hourly energy consumption in buildings based on a feedback artificial neural network, Energy Build, Vol. 37, No. 6.

Hamzacebi, Coskun (Dec. 2008), Improving artificial neural networks' performance in seasonal time series forecasting .Information Sciences, Vol. 178, No. 23.

Li, Qiong, Meng, Qingling (Jan. 2009), Predicting hourly cooling load in the building: a comparison of support vector machine and different artificial neural networks, Energy Conversion and Management, Vol. 50, No. 1.

Mills, Terence C. (1990), Time Series Techniques for Economists, Cambridge, Cambridge University Press.

Mintzberg, Henry (1996), The Strategy Process: Concepts, Contexts, Cases, USA, Prentice Hall.

Nasr, G.E. (Jan. 2020), Neural networks in forecasting electrical energy consumption, International Journal of Energy Research, Vol. 26, No. 1.

Nizami, Javeed, Algerni, Ahmed (Dec. 1995), Forecasting electric energy consumption using neural networks, Energy Policy, Vol. 23, No. 12.

Porter, Michael (1998), Competitive Strategy, USA, Free Press.

Rumelhart, David, Hinton, Geoffrey (Oct. 1986), Learning representations by backpropagating errors, Nature, Vol. 323, No. 1.

Tay, Feh (Nov. 2003), Support Vector Machine with Adaptive Parameters in Financial Time Series Forecasting, IEEE Transaction on Neural Networks, Vol. 14, No. 6.

پرویز، لیلا، پیمایی، مهسا (زمستان ۱۳۹۷)، پیش‌بینی اقلیمی استوکستیکی عملکرد چهار گیاه گندم، جو، سیب‌زمینی و ذرت در استان‌های آذربایجان شرقی و غربی در راستای توسعه برنامه‌ریزی کشاورزی، نشریه تولید گیاهان زراعی, دوره یازدهم، شماره ۴.

حسینی، سید محمد طاهر، سی‌وسه مرده، عادل، فتحی، پرویز، سی‌وسه مرده، معروف (بهار ۱۳۸۶)، کاربرد شبکه‌های عصبی مصنوعی و رگرسیون چند متغیره در برآورد عملکرد گندم دیم منطقه قروه استان کردستان، پژوهش کشاورزی، دوره هفتم، شماره ۱.

رحمانی، الهام، خلیلی، علی، لیاقت، عبدالمجید (تابستان ۱۳۸۷)، بررسی کمی تأثیر خشک‌سالی بر عملکرد محصول جو در آذربایجان شرقی به روش رگرسیونی چند متغیره، علوم آب و خاک، دوره دوازدهم، شماره ۴۴.