DOR: 20.1001.1.2538161.1404.10.2.1.9

DOI: 10.22034/csj.2025.227836

تابستان ۱۴۰۴، دوره ۱۰، شماره ۲، پیاپی ۳۷

دریافت مقاله: ۱۴۰۳/۱۰/۱۱

پذیرش مقاله: ۱۴۰۳/۱۲/۱۱

نوع مقاله: پژوهشی

پیش‌بینی انتشار دی‌اکسیدکربن در زنجیره تامین فولاد با استفاده از جدول داده–ستانده و یادگیری ماشین

محمد پارسایی‌فر

دانش‌آموخته کارشناسی ارشد گروه مهندسی صنایع، دانشگاه شهید باهنر کرمان، کرمان، ایران

پست الکترونیک: mohammad.parsaeifar90@gmail.com

سید حامد موسوی‌راد*

دانشیار گروه مهندسی صنایع، دانشکده فنی و مهندسی، دانشگاه شهید باهنر کرمان، کرمان، ایران

پست الکترونیک: S.h.moosavirad@uk.ac.ir

منصوره نادری‌پور

استادیار گروه مهندسی صنایع، دانشکده فنی و مهندسی، دانشگاه شهید باهنر کرمان، کرمان، ایران

پست الکترونیک: m.naderipour@uk.ac.ir

چکیده

زنجیره تأمین فولاد یک سیستم پیچیده شامل فرآیندها و واحدهای تامین، تولید و توزیع است که منجر به انواع محصولات فولادی می‌شود. با گسترش صنعت فولاد، میزان انتشار دی اکسید کربن (CO_2) و سایر گازهای گلخانه‌ای افزایش می‌یابد که این امر پیامدهای زیست‌محیطی همچون گرمایش جهانی، تغییرات اقلیمی و آسیب به سلامت انسان‌ها را به همراه دارد. جداول داده–ستانده نمایانگر جریان کالاها و خدمات بین فعالیت‌های مختلف یک اقتصاد در یک دوره زمانی مشخص هستند و در زنجیره تأمین فولاد، شامل ورودی‌های صنعت فولاد از سایر صنایع و خروجی‌های صنعت فولاد به صنایع دیگر می‌شوند. این تحقیق به پیش‌بینی میزان انتشار CO_2 صنعت فولاد کشورها با استفاده از روش‌های یادگیری ماشین پرداخته است. این فناوری با تحلیل دقیق داده‌های جداول داده–ستانده، فرآیندهای تولید فولاد را مورد بررسی قرار داده و میزان انتشار CO_2 را پیش‌بینی می‌کند. مدل‌های یادگیری ماشین، از جمله تقویت گرادیان، XGBoost، و جنگل تصادفی با تکنیک Bootstrap، در این تحقیق به کار گرفته شده‌اند. نتایج نشان می‌دهد که مدل تقویت گرادیان با دقت ۸۵ درصد بهترین عملکرد را داشته است، در حالی که مدل‌های جنگل تصادفی و XGBoost به ترتیب با دقت‌های ۸۲ درصد و ۷۹ درصد کمترین دقت را از خود نشان داده‌اند. این تحقیق نشان می‌دهد که الگوریتم‌های یادگیری ماشین قادر به پیش‌بینی مؤثر میزان انتشار CO_2 در زنجیره تأمین فولاد بوده و می‌توانند به تصمیم‌گیری‌های

* نویسندهٔ مسئول

بهتر و کاهش انتشار CO_2 در این صنعت کمک کنند.

واژه‌های کلیدی: یادگیری ماشین، مدیریت زنجیره تامین، صنعت فولاد، جدول داده-ستانده، دی اکسید کربن

۱. مقدمه

فولاد یکی از مؤثرترین عناصر در توسعه و پیشرفت انقلاب صنعتی بوده است و به دلیل ویژگی‌های منحصربه‌فرد و کاربردهای گسترده، جایگاهی برتر نسبت به سایر فلزات به دست آورده است. فولاد، چه به‌عنوان ماده اولیه و چه به‌عنوان محصول واسطه، در بسیاری از صنایع نقش کلیدی ایفا می‌کند و می‌توان آن را پایه و اساس توسعه صنعتی در نظر گرفت [۱]. زنجیره تأمین فولاد مجموعه‌ای از فعالیت‌ها و فرآیندهای پیوسته است که از استخراج مواد خام تا تولید و توزیع محصول نهایی را شامل می‌شود. در مراحل ابتدایی، مواد معدنی همچون سنگ‌آهن و زغال‌سنگ از معادن استخراج می‌شوند. سپس این مواد به کارخانه‌های فولادسازی منتقل می‌گردند تا طی فرآیندهای ذوب و تصفیه به فولاد تبدیل شوند. در نهایت، محصولات فولادی تولیدشده به صنایع مختلفی از جمله ساخت‌وساز و خودروسازی عرضه می‌شوند. این صنایع به تنهایی نیمی از کل تولیدات فولاد در جهان را مصرف می‌کنند [۱].

با وجود نقش حیاتی و گسترده فولاد در صنعت، زنجیره تأمین آن با چالش‌های متعددی مواجه است. از جمله این چالش‌ها می‌توان به نوسانات قیمت مواد اولیه، تغییرات غیرمنتظره در تقاضا، مشکلات حمل‌ونقل، نیاز به فناوری‌های نوین، تأثیرات تنش‌های بین‌المللی و مسائل زیست‌محیطی اشاره کرد. این مسائل تأثیر قابل توجهی بر فرآیند تولید و توزیع فولاد دارند و برای حفظ پایداری این زنجیره، نیاز به راه‌حل‌های نوآورانه و مدیریتی احساس می‌شود [۲]. تولید فولاد با چالش‌های زیست‌محیطی همچون انتشار گازهای گلخانه‌ای مانند دی اکسید کربن (CO_2)، NO_x و SO2 همراه است که تأثیرات مخربی بر تغییرات اقلیمی و محیط زیست دارد. CO_2 به‌عنوان یکی از اصلی‌ترین گازهای گلخانه‌ای، از طریق فرآیندهای احتراق و تولید فولاد وارد جو می‌شود و تأثیرات منفی مانند گرمایش جهانی و آلودگی هوا را تشدید می‌کند. علاوه بر این، این گازها باعث اسیدی شدن باران‌ها، تغییرات الگوهای آب‌وهوایی، و کاهش کیفیت هوا می‌شوند که نتیجه آن آسیب به سلامت انسان‌ها و اکوسیستم‌ها است. مدیریت انتشار این آلاینده‌ها در زنجیره تأمین فولاد، به چالش اصلی صنعت فولاد تبدیل شده است، به‌ویژه با توجه به نیاز به کاهش اثرات زیست‌محیطی و دستیابی به تولید پایدار [۳]. پیش‌بینی میزان CO_2 در صنعت فولاد بسیار مهم است، زیرا این صنعت یکی از بزرگترین منابع انتشار این گاز به جو است. پیش‌بینی دقیق انتشار CO_2 به بهینه‌سازی فرآیندها و مدیریت بهتر منابع انرژی کمک کرده و همچنین نقش مهمی در حفظ محیط زیست ایفا می‌کند. آگاهی از میزان این انتشار می‌تواند تصمیم‌گیری‌های مؤثری را برای کاهش اثرات زیست‌محیطی از طریق استفاده از فناوری‌های پاک و بهینه‌سازی فرآیندها تسهیل کند [۳].

یادگیری ماشین به عنوان شاخه‌ای گسترده از هوش مصنوعی این امکان را به کامپیوترها می‌دهد که از داده‌ها بیاموزند و بدون نیاز به برنامه‌ریزی دقیق به تصمیم‌گیری بپردازند. در پیش بینی میزان CO_2 و نیز دیگر پدیده‌ها می‌توان از الگوریتم‌های یادگیری ماشین استفاده کرد. الگوریتم‌هایی نظیر درخت تصمیم، جنگل تصادفی و ماشین بردار پشتیبان با تحلیل حجم زیادی از داده‌های پیچیده، دقت پیش‌بینی انتشار CO_2 را بهبود می‌بخشند. این الگوریتم‌ها به‌طور خودکار روابط و الگوهای موجود در داده‌ها را کشف کرده و در تحلیل داده‌های پیچیده عملکرد مؤثری دارند. بهره‌گیری از این روش‌ها در پیش‌بینی، امکان سازگاری با تغییرات داده‌ها و شناسایی الگوهای جدید را فراهم می‌آورد. این رویکرد، علاوه بر افزایش دقت پیش‌بینی‌ها، به ارائه راهکارهای بهینه و مؤثر در مدیریت انتشار CO_2 نیز کمک می‌کند. بنابراین، یادگیری ماشین به

عنـوان ابزاری توانمند برای ارتقـای پایداری و بهره‌وری فرآیندهای صنعتی که تأثیر مستقیم بر محیط زیست دارند، شناخته می‌شود [٤].

هدف این پژوهـش، پیش‌بینی ضرایب CO_2 در زنجیره تأمین فولاد با اسـتفاده از الگوریتم‌هـای جنگل تصادفی، تقویت گرادیان و XGBoost اسـت. این الگوریتم‌ها به دلیل دقت بالا در پیش‌بینی، توانایی شناسـایی ویژگی‌های مهم، مدیریـت داده‌های پیچیده و مقـاوم بودن در برابر نویز و داده‌های ناقص انتخاب شده‌اند. علاوه بر این، انعطاف‌پذیری در تنظیـم پارامترها، قابلیـت پردازش مـوازی، و قدرت پیش‌بیـنی بالای آن‌ها، اسـتفاده از ایـن روش‌ها را برای تحلیل داده‌های بزرگ و متنوع مناسب می‌سازد [٥]. تحلیل مقایسه‌ای عملکرد این مدل‌ها نیز به شناسایی بهترین روش برای بهینه‌سـازی مصرف انرژی و کاهش انتشار CO_2 در صنعـت فولاد کمک می‌کند. در ادامه برای رسیدن به هدف این تحقیق، ابتدا مروری بر تحقیقات پیشنین در این زمینه ارائه خواهد شد. سـپس روش تحقیق مورد استفاده بیان خواهد شـد و در ادامه نتایج تحقیق تشـریح خواهد گشت و در انتها نتیجه گیری و پیشـنهادات آتی این تحقیق ذکر می‌گردد.

۲. پیشینه تحقیق

بخش مرور ادبیـات این تحقیق به بـررسی مطالعات پیشیـن در زمینه تحلیـل و پیش‌بیـنی آلاینده‌های هوا با اسـتفاده از روش‌های یادگیری ماشین پرداخته و با هدف شناسایی شکاف‌های موجود و استفاده از یافته‌های آن‌ها برای ارتقای پژوهش حاضر طراحی شـده اسـت. در این مطالعات، الگوریتم‌های مختلفی مانند شبکه عصبی، جنگل تصـادفی، درخت تصمیـم SVM[1]، CNN[2] و LSTM[3] به‌کار گرفتـه شـده‌اند. در بسیـاری از پژوهش‌ها، اسـتفاده از یادگیری ماشین بر روش‌های سنتی ارجحیت داشته است؛ از جمله ایونسـکو[4] که شـبکه عصبی را بـرای پیش‌بینی انتشـار CO_2 و NO_2 در صنعت فولاد اسـتفاده کرده [٦] و یو[5] کـه از جنگل تصادفی برای پیش‌بینی AQI (شـاخص کیفیت هوا) شهر پکن بهره برده است [٧]. همچنین، ریبیرو[6] و همکاران در پرتغال نشـان دادند یادگیری ماشین نسبت به روش‌های سنتی مؤثرتر است [٨]. این پژوهش‌ها نتایج مختلفی را برای آلاینده‌ها در کشـورهای مختلف از جمله هند، چین، اتریش و آرژانتین به‌دسـت آورده‌اند و نشـان داده‌اند که روش‌های یادگیری ماشین از جمله شـبکه‌های عصبی ترکیبی و گرادیان تقویت‌شده می‌توانند بهبود قابل توجـهی در دقت پیش‌بینی ایجاد کننـد و به تحلیل بهتر و جامع‌تری از آلاینده‌ها در شـهرها و مناطق صنعتی منجر شوند.

جـدول ۱ به بررسی و ارائه مقالات مرتبط با پیش‌بینی آلاینده‌های هوا با اسـتفاده از یادگیری ماشین می‌پردازد و خلاصه‌ای از پژوهش‌ها، نویسـندگان، سـال انتشـار، روش‌هـای به کار رفته و نتایج کلیدی هر مطالعه را در بر دارد. هـدف این جدول، ارائه نمایی کلی از وضعیت کنونی تحقیقات و شناسایی روندها و تحولات در این زمینه است. این اطلاعات به محققان کمک می‌کند تا با روش‌های مختلف و نتایج آن‌ها آشـنا شـوند و زمینه‌های تحقیقاتی جدید را شناسایی کنند.

با توجه به مقالات بررسی شـده، روش‌های یادگیری ماشین به‌ویـژه الگوریتم‌هایی مانند درخت تصمیم، جنگل تصادفی، ماشین بردار پشـتیبان و شـبکه‌های عصبی در پیش‌بینی انتشار آلاینده‌ها، نتایج موثرتری داشته‌اند. لذا این پژوهش در نظر دارد به بررسی و پیش‌بینی ضرایب تولید CO_2 در زنجیـره تأمین فولاد با اسـتفاده از الگوریتم‌های یادگیری ماشین بپردازد. به عبارتی هدف پیش‌بینی میزان انتشـار CO_2 صنعت فولاد کشـورها بر اسـاس ضرایب فناوری آنها می‌باشـد. نوآوری پژوهش حاضر تمرکز آن بر داده‌های اقتصادی و انتشـار CO_2 صنایع ٤٣ کشور در قالب جداول داده–ستانده می‌باشد که امکان تحلیل دقیق‌تری را برای هدف تحقیق فراهـم می‌آورد. لذا این تحقیق گامی

1-Support Vector Machines
2-Convolutional Neural Networks
3-Long Short-Term Memory
4-Ionescu
5-Yu
6-Ribeiro

جدول (۱): خلاصه مقالات مرتبط با موضوع پژوهش حاضر

ردیف	دسته‌بندی	نویسندگان (سال)	هدف	منطقه	روش‌های کلیدی
۱	پیش‌بینی دقیق‌تر مدل‌ها	شکیا و همکاران (۲۰۲۳) [۹]، کریمی و همکاران (۲۰۲۳) [۱۰]، بکار و همکاران (۲۰۲۱) [۱۱]، یو و همکاران (۲۰۱۶) [۷]	پیش‌بینی آلاینده‌ها	هند، ایران، چین	RNN,RF, LSTM
۲	تأثیر ترافیک بر آلودگی	موریرا و همکاران (۲۰۲۳) [۱۲]، جئونگ و همکاران (۲۰۲۳) [۱۳]، جلال‌الدین و همکاران (۲۰۲۳) [۱۴]، سیکان و همکاران (۲۰۲۳) [۱۵]	تحلیل آلودگی ناشی از ترافیک	برزیل، کره، مالزی، رومانی	RNN, LR, ANN
۳	مقایسه الگوریتم‌ها	لین و همکاران (۲۰۲۳) [۱۶]، گسیم حیدر و همکاران (۲۰۲۳) [۱۷]، ریبرو و همکاران (۲۰۲۱) [۸]	مقایسه مدل‌ها	چین، هند، پرتغال	LSTM, CNN, XGB,LightGBM
۴	پیش‌بینی AQI	دی‌هی و همکاران (۲۰۲۳) [۱۸]، فنگ و همکاران (۲۰۲۳) [۱۹]	مدل‌های ترکیبی AQI	چین، هند	RF, RNN, CNN
۵	پیش‌بینی CO_2	پائول و همکاران (۲۰۲۳) [۲۰]، ایونسکو و همکاران (۲۰۰۷) [۶]، نبی و همکاران (۲۰۲۱) [۲۱]	پیش‌بینی CO_2	کانادا، هلند، هند	ANN, GBT
۶	آلاینده‌های صنعتی و دریایی	استفان و همکاران (۲۰۲۳) [۲۲]، تنگ و همکاران (۲۰۲۳) [۲۳]	تخمین آلاینده‌ها	کلمبیا، چین	DRF, MLR

مهم برای پر کردن شکاف‌های موجود و ارتقای درک از انتشار آلاینده‌ها با استفاده از داده‌های جامع داده–ستانده و روش‌های نوین یادگیری ماشین است.

۳. روش تحقیق

برای این هدف، از مجموعه‌ای از داده‌های تاریخی که شامل ویژگی‌های مرتبط با ورودی‌های صنعت فولاد در کشورهای مختلف است، استفاده می‌شود. به منظور تحلیل و پیش‌بینی دقیق‌تر، روش‌های یادگیری ماشین متنوعی به کار گرفته می‌شوند. در ادامه جزئیات این روش تحقیق بیان خواهد شد.

۳-۱. داده‌های مورد استفاده در این پژوهش

داده‌های مورد استفاده در این پژوهش از پایگاه (WIOD (https://www.rug.nl/ggdc/valuechain/wiod/?lang=en)، حاوی ٦٤٥ نمونه و ٥٦ ویژگی که در بر گیرنده ورودی‌های (تامین) صنعت فولاد و انتشار CO_2 بخش‌های مختلف اقتصادی از جمله تولیدات کشاورزی، صنایع معدنی، صنایع تولیدی (مانند فولاد و تجهیزات الکتریکی)، حمل‌ونقل، خدمات مالی، خدمات عمومی و مدیریت پسماند ٤٣ کشور از سال ۲۰۰۰ تا ۲۰۱٤ است. لازم به ذکر است که مقادیر تمام این ویژگی‌ها، از نوع عددی پیوسته هستند. لازم به ذکر است از جداول داده–ستانده در بسیاری از مقالات با اهداف و روش‌های مختلف استفاده شده است از جمله اولزا[7] و همکاران، تراز تجاری اندونزی را با استفاده از جداول داده–ستانده برای شناسایی بخش‌ها و شرکای تجاری کلیدی از این جداول استفاده کردند [٢٤]. لو[8] و همکاران رقابت‌پذیری صنعت تجهیزات الکتریکی چین را در زنجیره ارزش جهانی با رویکرد درآمد زنجیره ارزش جهانی در پایگاه داده WIOD بررسی کردند [٢٥]. گونزالز[9] و همکاران، الگوهای اثرات زیست‌محیطی کشورهای ثروتمند را با تحلیل چندمنطقه‌ای جداول داده–ستانده مطالعه کردند [٢٦]. ال هیوتی[10] و همکاران نیز با به کارگیری جداول داده ستانده پایداری سیستم‌های با تأخیر متغیر در زمان را تحلیل کردند [٢٧]. سو[11] و همکاران اثرات تجمیع‌بخشی بر انتشار CO_2 مرتبط با انرژی در تجارت بین‌الملل را با داده‌های جداول داده–ستانده چین و سنگاپور را بررسی کردند [٢٨]. در هیچ کدام از این مقالات بررسی شده، از

7-Ulza
8-Lu
9-Gonzalez
10-El Haouti
11-Su

روش‌های یادگیری ماشین برای تحلیل جداول داده ستانده استفاده نشده است. در مقابل، مطالعه حاضر با استفاده از الگوریتم‌های یادگیری ماشین و داده‌های جدول داده-ستانده، مدل‌های پیش‌بینی ضریب CO_2 در زنجیره تأمین فولاد را توسعه می‌دهد و دقت آن‌ها را ارزیابی می‌کند.

۳-۲. جدول داده-ستانده و ضرایب فناوری

جدول داده-ستانده شامل داده های مربوط به ویژگی‌های ورودی های زنجیره تامین فولاد و تولید CO_2 در زنجیره تأمین فولاد است. در این جدول، ورودی‌های مربوط به صنعت فولاد شامل مقادیر مختلف مانند انرژی، مواد اولیه و غیره در نظر گرفته می‌شود و مقدار تولید CO_2 به عنوان خروجی مشخص می‌شود. این اطلاعات به‌صورت عددی پیوسته و در ابعاد مختلف ثبت می‌شوند. ضرایب فناوری نیز به میزان CO_2 انتشار یافته به ازای هر نوع ورودی در فرآیند تولید فولاد اشاره دارد. در شکل ۱، نمایی از یک جدول داده-ستانده با بخش‌های R1 و R2 ارائه شده است. در این جدول، سلولی که با حرف "a" مشخص شده است، نشان‌دهنده میزان ورودی (بر حسب دلار) از صنعت دوم کشور استرالیا به خود این صنعت در این کشور است. همچنین، سلول حاوی مقدار "c" نمایانگر کل خروجی (کل تولید بر حسب دلار) صنعت دوم کشور استرالیا است. به علاوه، سلول "b" نشان می‌دهد که هر صنعت نه تنها می‌تواند ورودی‌های لازم را از خود دریافت کند، بلکه قابلیت دریافت ورودی از صنایع دیگر کشورها نیز دارد.

ورودی / خروجی		استرالیا		اتریش		خروجی
		R1	R2	R1	R2	
استرالیا	R1					
	R2		a			c
اتریش	R1					
	R2	b				

شکل (۱): نمایشی از یک جدول داده-ستانده

۳-۳. یادگیری ماشین

یادگیری ماشین به مجموعه‌ای از الگوریتم‌ها و تکنیک‌ها اشاره دارد که به سیستم‌ها این امکان را می‌دهد تا بدون برنامه‌نویسی صریح، از داده‌ها یاد بگیرند و پیش‌بینی کنند [۲۹]. در این پژوهش، از دو روش پیش‌پردازش و سه روش یادگیری ماشین استفاده می‌شود.

۳-۳-۱. پیش‌پردازش داده‌ها

در این پژوهش، از دو روش توکی و KNNImputer برای پیش‌پردازش داده‌ها استفاده شده است.

۳-۳-۱-۱. روش توکی

این روش از چارک‌ها و دامنه میان‌چارکی برای شناسایی داده‌های پرت استفاده می‌کند. این روش یک حد بالا و یک حد پایین برای محدوده داده‌ها تعیین می‌کند و داده‌هایی که خارج از این حدود باشند به عنوان نقاط پرت در نظر گرفته می‌شوند. این فرآیند به‌طور موثری به شناسایی و حذف داده‌های غیرمعمول کمک می‌کند. مقدار ضریب k در روش توکی ۱/۵ در نظر گرفته شده است. این مقدار توازنی مناسب میان حساسیت به نقاط پرت و حفظ داده‌های واقعی ایجاد می‌کند. همچنین به دلیل عملکرد مؤثر در بسیاری از توزیع‌های داده و جلوگیری از حذف اطلاعات مفید، به‌عنوان یک معیار استاندارد پذیرفته شده و در اغلب تحلیل‌ها به کار می‌رود [۳۰].

۳-۳-۱-۲. روش KNNImputer

این الگوریتم تصحیح داده است که بر مبنای k-نزدیک‌ترین همسایه عمل می‌کند. پس از شناسایی داده‌های پرت با روش توکی، KNNImputer برای تصحیح مقادیر مفقود شده به کار می‌رود. این الگوریتم با اندازه‌گیری فاصله بین نمونه‌ها (معمولاً با استفاده از فاصله اقلیدسی) و انتخاب k نمونه نزدیک به هر نمونه پرت، مقادیر پرت را با میانگین وزن‌دار مقادیر مشابه در همسایه‌ها پیش‌بینی می‌کند و مقدار k در روش KNNImputer برابر با ۵ فرض شده است. در روش KNNImputer، مقدار برابر با

۵ به‌طور رایج انتخاب می‌شود، زیرا توازنی مناسب میان دقت و پایداری ایجاد می‌کند. این مقدار با استفاده از پنج نزدیک‌ترین همسایه، مقادیرگمشده را به‌طور دقیق تخمین می‌زند و در عین حال از تأثیر نقاط غیرمعمول یا نویز در داده‌ها جلوگیری می‌کند [۳۱].

۳-۴. روش‌های یادگیری ماشین

در این پژوهش از سه روش یادگیری ماشین شامل تقویت گرادیان، XGBoost و جنگل تصادفی استفاده شده است، زیرا این الگوریتم‌ها به دلیل قدرت بالای خود در مدل‌سازی روابط پیچیده، انتخاب‌های مناسبی برای پیش‌بینی ضرایب CO_2 در زنجیره تأمین فولاد هستند. تقویت گرادیان و XGBoost به‌واسطه توانایی در تنظیم دقیق پارامترها، استفاده از منظم‌سازی (Regularization) و یادگیری افزایشی، برای بهبود دقت و جلوگیری از بیش‌برازش بسیار مؤثر هستند. XGBoost همچنین به دلیل سرعت بالا و قابلیت پردازش موازی، برای تحلیل داده‌های حجیم مناسب است. جنگل تصادفی با استفاده از ترکیب چندین درخت تصمیم و به‌کارگیری بوت‌استرپ، به‌طور طبیعی در برابر نوفه مقاوم است و توانایی انتخاب خودکار ویژگی‌های مهم را دارد. این سه روش در کنار یکدیگر، دقت بالا، انعطاف‌پذیری و تفسیرپذیری قابل قبولی برای تحلیل این مسئله ارائه می‌دهند [۵].

۳-۴-۱. تقویت گرادیان

این روش برای حل مسائل طبقه‌بندی و رگرسیون به کار می‌رود. در این روش، یک مدل قوی و کارآمد از ترکیب چندین مدل ضعیف ایجاد می‌شود. در ابتدا، یک مدل اولیه مانند درخت تصمیم ایجاد می‌شود. سپس با محاسبه میزان خطا بین پیش‌بینی‌ها و مقادیر واقعی، یک مدل جدید برای تصحیح خطاها ایجاد می‌شود. این فرآیند تا زمانی ادامه می‌یابد که تعداد تعیین‌های مشخصی به دست آید. از آنجا که این روش با تعداد تعیین‌ها مواجه است و از خطای باقی‌مانده یادگیری می‌کند، نیازی به کاهش بعد یا نرمال‌سازی داده‌ها نیست. اَبَرپارامترهای این روش شامل تعداد درخت‌ها در مدل (n_estimators)، میزان تغییر در پیش‌بینی‌ها در هر مرحله (learning_rate)، حداکثر عمق درخت‌ها برای کنترل پیچیدگی مدل (max_depth) و حداقل تعداد نمونه‌ها برای انجام تقسیم در یک گره درخت (min_samples_split) است [۳۲].

۳-۴-۲. روش XGBoost

این روش نیز بر پایه تقویت گرادیان بنا شده و درآن از درخت‌های تصمیم به عنوان مدل‌های ضعیف استفاده می‌شود. در XGBoost، تابع هدف به‌صورت گرادیان تابع هزینه نسبت به پارامترهای مدل محاسبه می‌شود. با افزودن مدل‌های جدید، سعی می‌شود خطاهای موجود کاهش یابد. این روش از دو نوع جریمه (LASSO) L1 و (Ridge) L2 برای کنترل بیش‌برازش استفاده می‌کند. اَبَرپارامترهای XGBoost شامل تعداد درخت‌ها در مدل (n_estimators)، حداکثر عمق درخت‌ها برای کنترل پیچیدگی مدل (max_depth)، میزان تغییر در پیش‌بینی‌ها در هر مرحله (learning_rate)، پارامتر جریمه L1 برای کنترل اهمیت ویژگی‌ها (reg_alpha) و پارامتر جریمه L2 برای کنترل وزن ویژگی‌ها (reg_lambda) هستند [۳۳].

۳-۴-۳. روش جنگل تصادفی

این روش به‌عنوان یک الگوریتم ترکیبی از چندین درخت تصمیم عمل می‌کند. در جنگل تصادفی، از تکنیک بوت‌استرپ برای نمونه‌برداری تصادفی استفاده می‌شود و برای هر نمونه، یک درخت تصمیم آموزش داده می‌شود. این الگوریتم نیازی به کاهش بعد یا نرمال‌سازی داده‌ها ندارد و هر درخت به‌طور مستقل پیش‌بینی می‌کند. اَبَرپارامترهای جنگل تصادفی تعداد درخت‌ها در مدل (n_estimators)، حداکثر عمق درخت‌ها برای کنترل پیچیدگی مدل (max_depth)، میزان تغییر در پیش‌بینی‌ها در هر مرحله (learning_rate)، حداقل تعداد نمونه‌ها برای انجام تقسیم در یک گره (min_samples_split) و استفاده از تکنیک Bootstrap (bootstrap) هستند [۳۴].

۳-۵. ارزیابی مدل

به منظور ارزیابی عملکرد مدل‌های یادگیری ماشین از معیارهای مختلفی استفاده می‌شود که شامل R2 Score، میانگین مطلق خطا (MAE) و میانگین مربعات خطا (MSE) است. این معیارها به‌طورکلی نشان‌دهنده دقت و کارایی مدل‌ها در پیش‌بینی مقادیر انتشار CO_2 می‌باشند.

۳-۵-۱. روش R2 Score

این معیار نشان‌دهنده توانایی مدل در توضیح تغییرات داده‌های خروجی است و مقدار آن بین ۰ تا ۱ متغیر است. مقدار نزدیک به ۱ به معنی پیش‌بینی دقیق‌تر است . فرمول ۱ نحوه محاسبه این معیار را نشان می‌دهد که در آن n تعداد نمونه، yi مقدار واقعی نمونه iام، $\hat{y}_i$ مقدار پیش‌بینی برای مدل iام، $\bar{y}$ میانگین مقدار واقعی داده‌ها و R2 مقدار دقت به دست آمده را نشان می‌دهد [۳۵].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

۳-۵-۲. میانگین قدر مطلق خطا (MAE)

این معیار نشان‌دهنده میانگین خطاهای مطلق بین مقادیر پیش‌بینی شده و مقادیر واقعی است. هر چه مقدار MAE کمتر باشد، نشان‌دهنده دقت بیشتر مدل است که در آن n تعداد نمونه، yi مقدار واقعی نمونه iام، $\hat{y}_i$ مقدار پیش‌بینی برای مدل iام است [۳۶].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

۳-۵-۳. میانگین مربعات خطا (MSE)

این معیار به‌صورت میانگین مربعات اختلاف بین مقادیر پیش‌بینی‌شده و واقعی محاسبه می‌شود و برای کاهش تاثیر نقاط پرت نسبت به MAE حساس‌تر است که در آن n تعداد نمونه، yi مقدار واقعی نمونه iام، $\hat{y}_i$ مقدار پیش‌بینی برای مدل iام است [۳۶].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3)$$

۴. نتایج

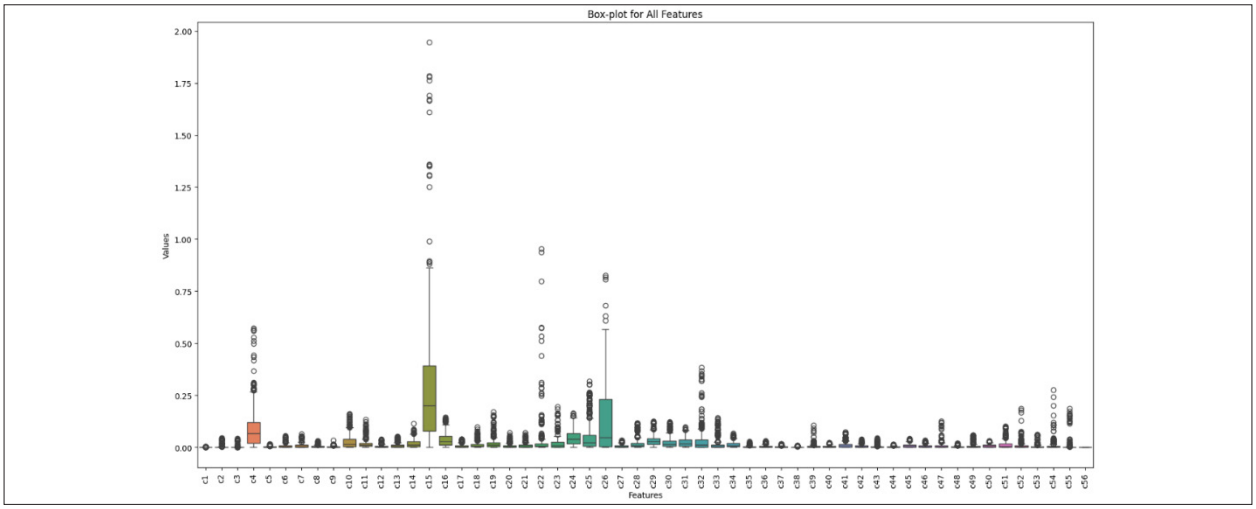
در این بخش، نتایج حاصل از پیاده‌سازی مدل‌های شرح داده شده در بخش قبل ارائه خواهد شد. این نتایج شامل نمودارهایی است که به درک بهتر عملکرد مدل‌ها کمک می‌کند. ابتدا نتایج حاصل از تحلیل و مصورسازی داده‌ها ارائه می‌شود. سپس، نتایج پیش‌پردازش داده‌ها و در نهایت، عملکرد مدل‌ها پس از پیاده‌سازی به تفصیل بیان خواهد شد.

۴-۱. مصورسازی داده‌ها

در این بخش، به تحلیل بصری داده‌ها از طریق نمودارها پرداخته می‌شود تا دیدگاهی بهتر نسبت به ساختار و توزیع داده‌های پژوهش ایجاد گردد. برای این منظور، از زبان برنامه‌نویسی پایتون استفاده شده و جداول و نمودارها مورد بررسی قرار گرفته‌اند.

۴-۱-۱. نمودار جعبه‌ای

نمودار جعبه‌ای که توسط جان توکی معرفی شده، ابزاری برای توصیف و تحلیل داده‌ها است. این نمودار شامل یک جعبه است که نشان‌دهنده دامنه میان‌چارکی است. نقاطی که خارج از این جعبه قرار دارند، داده‌های پرت را نمایش می‌دهند. جعبه شامل چارک اول، دوم (میانه) و سوم است و خطوط متصل به جعبه نشان‌دهنده کمترین و بیشترین مقادیر غیرپرت هستند [۳۷]. در این پژوهش، سه ویژگی اصلی یعنی «تولید فولاد»، «بخش فاضلاب» و «استخراج معادن» دارای بیشترین دامنه میان‌چارکی و چولگی به راست هستند که نشان از توزیع گسترده مقادیر دارد. همچنین داده‌های پرت در این ویژگی‌ها به وضوح مشاهده می‌شوند. در شکل ۲، نمودار جعبه‌ای برای ویژگی‌های داده‌ها نمایش داده شده است. ویژگی‌های «تولید فولاد»، «بخش فاضلاب» و «استخراج معادن» که دارای بیشترین دامنه میان‌چارکی هستند، نشان‌دهنده پراکندگی بالای داده‌ها در این بخش‌ها هستند. این ویژگی‌ها به دلیل تغییرات زیاد، نیاز به توجه و تحلیل

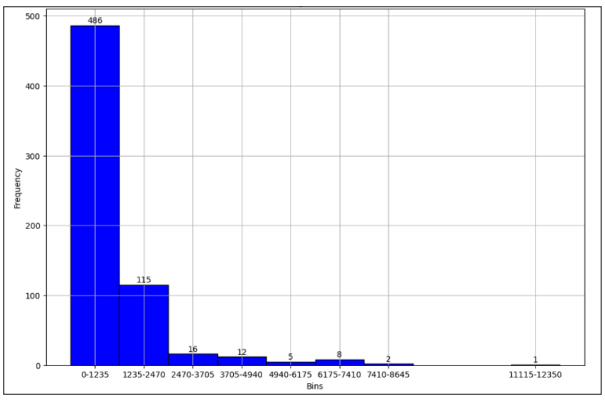


شکل (۲): نمایی از نموار جعبه‌ای داده‌ها

دقیق‌تری دارند. شناسایی چنین ویژگی‌هایی به کمک در اولویت‌بندی تحلیل‌ها، بهبود فرآیندها و تخصیص منابع به بخش‌هایی که بیشترین نوسان را دارند، می‌آید و می‌تواند به تصمیم‌گیری‌های راهبردی و بهبود دقت پیش‌بینی‌ها کمک کند.

۴-۱-۲. هیستوگرام

هیستوگرام، یک نمودار گرافیکی برای نمایش توزیع داده‌های کمّی است. محور افقی بیان‌کننده بازه‌های داده و محور عمودی نشان‌دهنده فراوانی هر بازه است با استفاده از فرمول استرجس، تعداد بازه‌ها در این پژوهش برابر با ۱۰ و طول هر بازه برابر ۱۲۳۵ محاسبه شده است [۳۸]. در این نمودار، توزیع داده‌ها به‌طور واضح نشان‌دهنده چوله به راست بودن مقادیر هدف است، به‌طوری که بیشترین داده‌ها در بازه‌های ابتدایی قرار دارند. این ویژگی نشان‌دهنده تمرکز بالای داده‌ها در مقادیر کمتر و توزیع نامتقارن آن‌ها است. شکل ۳، نمودار مربوطه را نشان می‌دهد که در آن محور افقی نمایانگر بازه‌های مختلف داده‌ها و محور عمودی نمایش‌دهنده فراوانی هر بازه است. این نمودار به تحلیل توزیع داده‌ها و شناسایی الگوهای خاص در آن‌ها کمک می‌کند.

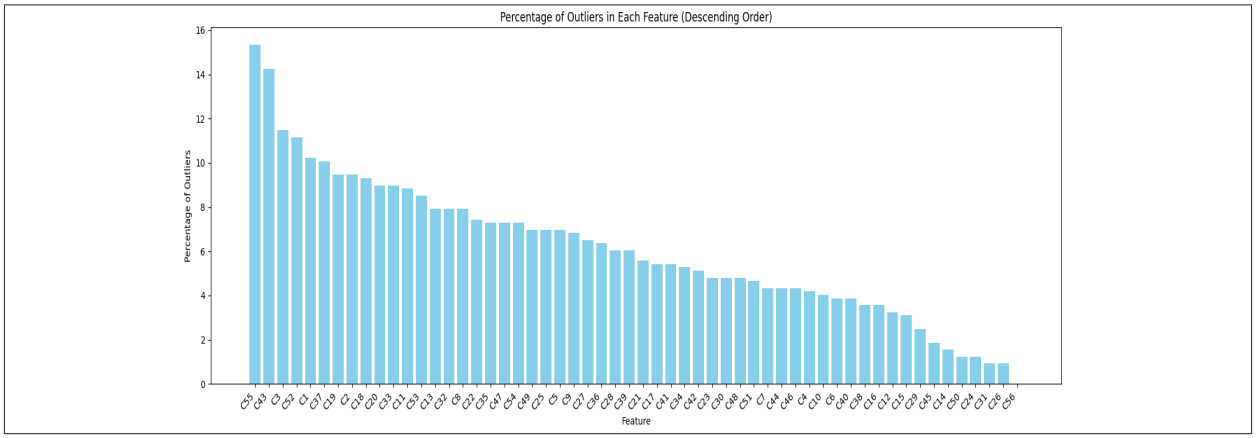


شکل (۳): نمایشی از نمودار هیستوگرام برچسب‌ها یا مقادیر هدف داده‌ها

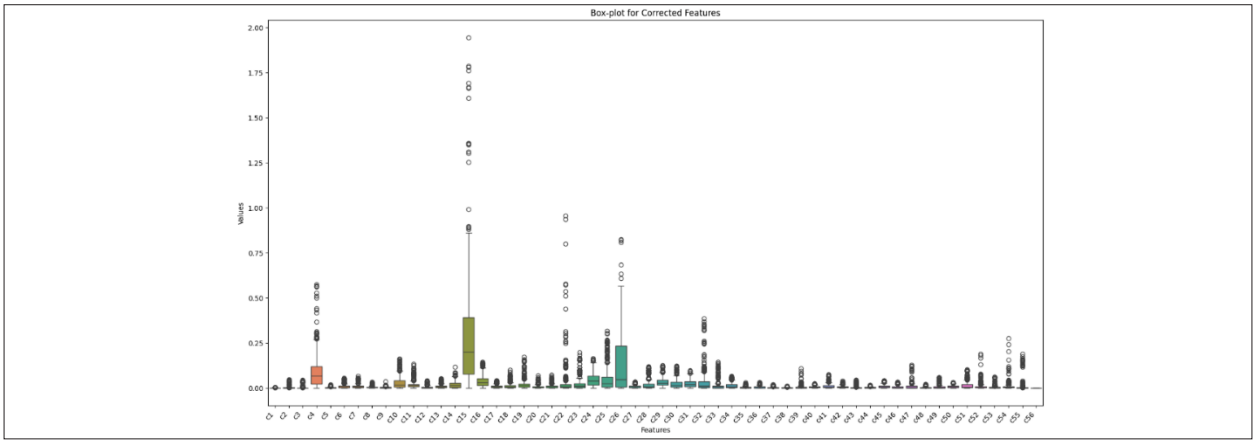
۴-۲. مدیریت داده‌های پرت

در این بخش، نتایج حاصل از پیاده‌سازی پیش‌پردازش داده‌ها، شناسایی داده‌های پرت با استفاده از روش توکی-فنس انجام شده و بر اساس ضریب k برابر با ۱/۵، تعداد ۲۱۹۶ مقدار پرت شناسایی شده است که حدود ٦ درصد کل داده‌ها را تشکیل می‌دهد. در نمودار میله‌ای شکل ٤، درصد داده‌های پرت برای هر ویژگی نشان داده شده است. برای این روش، کد به صورت دستی با استفاده از زبان برنامه‌نویسی پایتون پیاده‌سازی شده است.

همچنین، نمودار جعبه‌ای داده‌ها، قبل از تصحیح داده‌های پرت در نمودار شکل ٥ و بعد از تصحیح آنها با استفاده از روش KNNImputer در نمودار شکل ٦ ارائه



شکل (۴): نمودار درصد داده‌های پرت هر ویژگی با روش توکی-فنس



شکل (۵): نمودار جعبه‌ای قبل از تصحیح داده‌های پرت

شـــده است. در این روش، مقدار k برابر با ۵ در نظر گرفته شده اســـت. این نمودارها تغییرات مقادیر ویژگی‌ها را در فرآیند تصحیح داده‌های پرت نشان می‌دهند. نمودار شکل ۵ نشـــان‌دهنده وضعیت داده‌ها قبل از تصحیح اســـت که حاوی مقادیر پرت زیادی در برخی ویژگی‌ها اســـت. پس از اعمال KNNImputer، نمودار شـــکل ٦ نشان می‌دهد که مقادیر پرت کاهش یافته و پراکندگی داده‌ها در بسیاری از ویژگی‌ها بهبود یافته اســـت. این کار به کاهش اثرات منفی داده‌های پرت و بهبود دقت مدل کمک می‌کند.

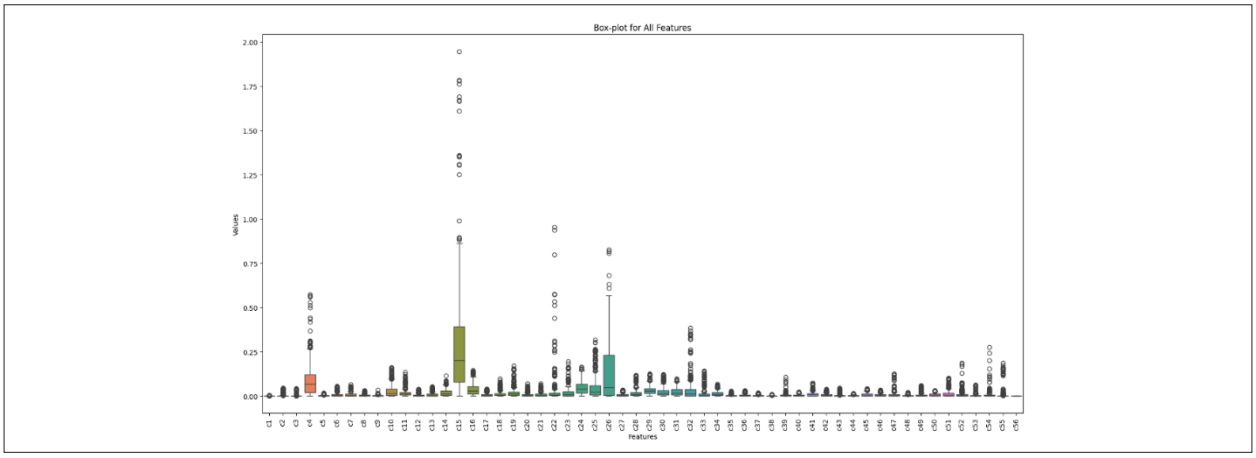
۴-۳. نتایج حاصل از پیاده‌سازی مدل

در این بخش، نتایج حاصل از پیاده‌سازی سه الگوریتم تقویـــت گرادیـــان، XGBoost و جنـــگل تصـــادفی بررسی شده‌اند. داده‌ها به نسبت ۷۰ به ۳۰ به دو بخش آموزش[۱۲] و آزمون[۱۳] تقسیم شـــده و برای ارزیابی مدل‌ها از معیارهای MSE، R2 Score و MAE استفاده شده است. تمامی تنظیمات بهینه پارامترهای این مدل‌ها با اســـتفاده از گریدسرچ[۱۴] به دست آمده‌اند. گریدسرچ یک روش جستجوی فراگیر است که تمامی ترکیب‌های ممکن از پارامترها را بررسی می‌کند تا بهترین مجموعه پارامترها را برای مدل انتخاب کند. این روش بـــه یافتن پارامترهای بهینه و افزایش دقت و کارایی مدل‌ها کمک می‌کند [۳۹].

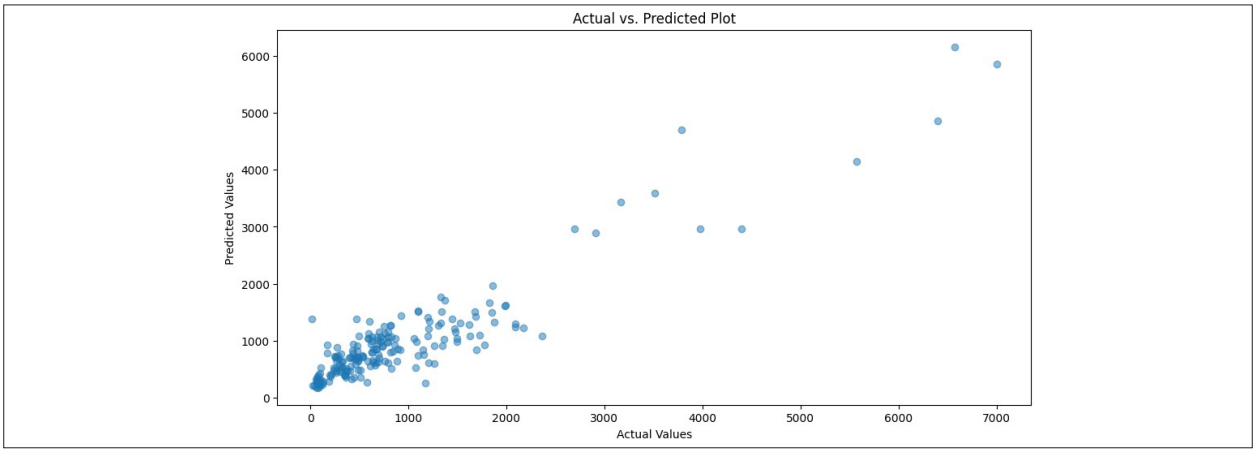
۴-۳-۱. روش تقویت گرادیان

برای روش تقویت گرادیان، بهترین مقادیر اَبَرپارامترها شـــامل ۵۰ درخت، نـــرخ یادگیری ۰/۰۷، عمـــق درخت ٤ و حداقل نمونه‌ها برای تقسیم ۳ بوده اســـت. نتایج نشـــان می‌دهند که دقت داده‌های آمـــوزش ۹۵ درصد و داده‌های آزمون ۸۵ درصد اســـت. خطای MSE در داده‌های آزمون

12-Train
13-Test
14-Grid Search



شکل (۶): نمودار جعبه‌ای بعد از تصحیح داده‌های پرت با روش KNNImputer



شکل (۷): نمودار مقادیر واقعی داده‌ها در مقابل مقادیر پیش‌بینی شده

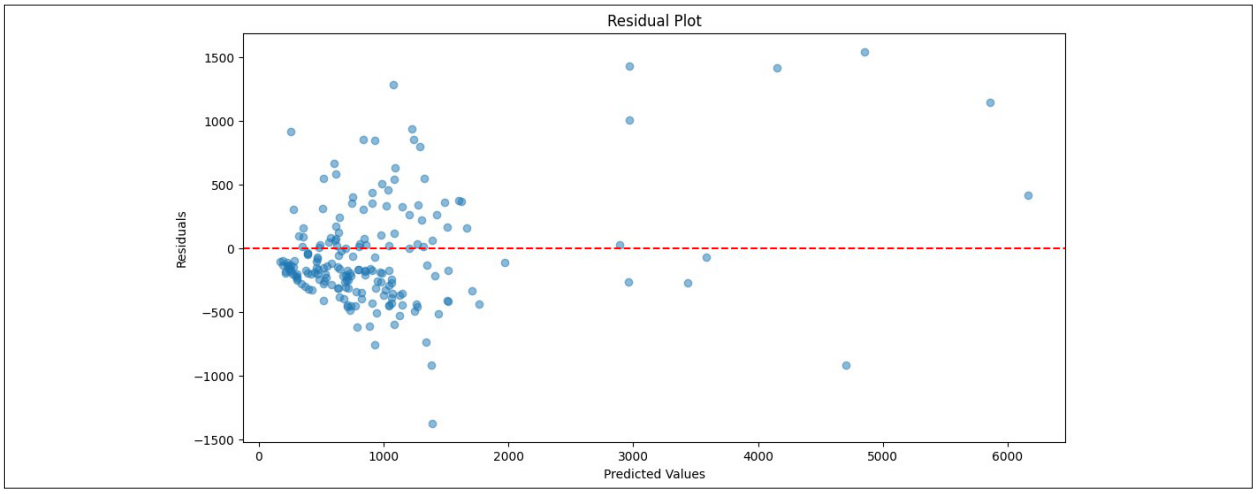
۲۰۸۸۵۵ و خطای MAE برابر با ۳۱۳ است. شکل ۷ و ۸ به ترتیب نمودار مقایسه مقادیر واقعی و پیش‌بینی شده و نمودار باقی‌مانده‌ها را نشان می‌دهند. در شکل ۷، توزیع مقادیر نشان‌دهنده همبستگی نسبی مناسبی بین مقادیر واقعی و پیش‌بینی‌شده است، هرچند در مقادیر بالا انحراف مشاهده می‌شود. در شکل ۸، نمودار باقی‌مانده‌ها حاکی از توزیع تصادفی است، اما پراکندگی باقی‌مانده‌ها در مقادیر پیش‌بینی‌شده بزرگ‌تر بیشتر است که ممکن است به مدل‌سازی نیازمند بهبود اشاره داشته باشد.

۴-۳-۲. روش XGBoost

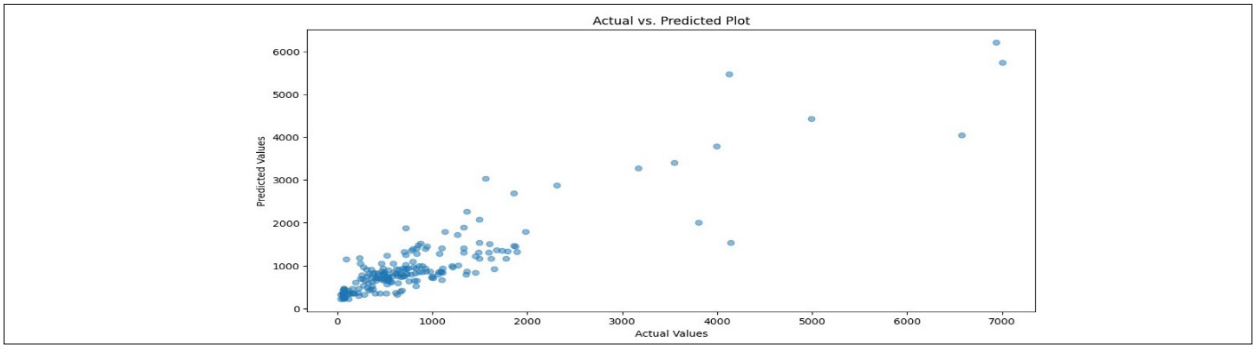
برای روش XGBoost، بهترین اَبَرپارامترها شامل ۱۰۰ درخت، نرخ یادگیری ۰٫۱، عمق ۲، و مقادیر جریمه و L2 به ترتیب ۰/٤ و ۰/۷ بوده است. نتایج نشان می‌دهند که دقت داده‌های آموزش ۸۷ درصد و داده‌های آزمون ۷۹ درصد است. خطای MSE در داده‌های آزمون ۳۳۳۱۱٤ و خطای MAE برابر با ٤۰۰ است. نمودارهای ۹ و ۱۰ به ترتیب مقایسه مقادیر واقعی و پیش‌بینی‌شده و باقی‌مانده‌های مدل را نشان می‌دهند. نمودار ۹ نشان می‌دهد که مدل در پیش‌بینی مقادیر هدف عملکرد نسبتاً خوبی داشته است، اما نمودار ۱۰ باقی‌مانده‌هایی با پراکندگی تصادفی و تقریباً نرمال را نشان می‌دهد که ممکن است ناشی از تأثیر عوامل ناشناخته یا تصادفی باشد.

۴-۴-۲. روش جنگل تصادفی

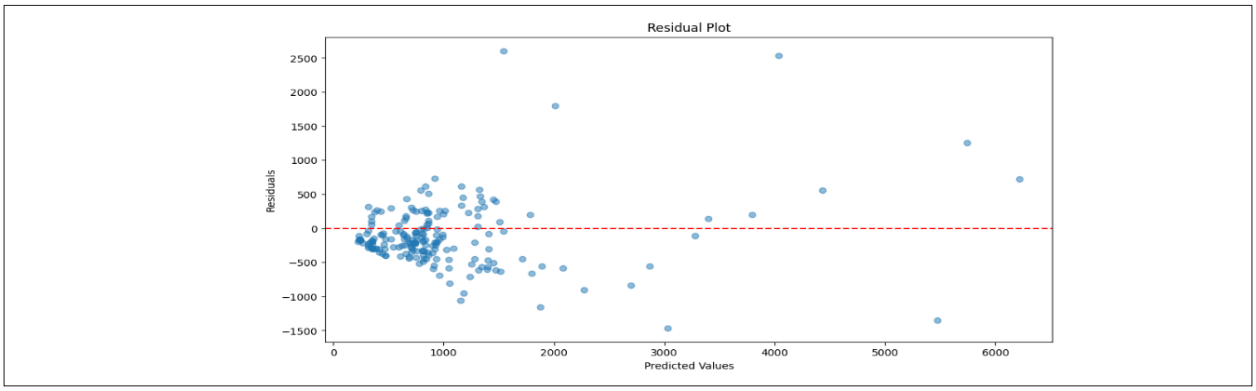
برای روش جنگل تصادفی، بهترین اَبَرپارامترها شامل ٦۰ درخت، عمق ۱۳، و استفاده از روش bootstrap بوده است. نتایج نشان می‌دهند که دقت داده‌های آموزش ۹٤ درصد و داده‌های آزمون ۸۲ درصد است. خطای MSE در داده‌های آزمون ۲۰٤۲۹۵ و خطای MAE برابر با ۲۸۸



شکل (۸): نمودار باقیمانده‌ها یا خطاهای مدل



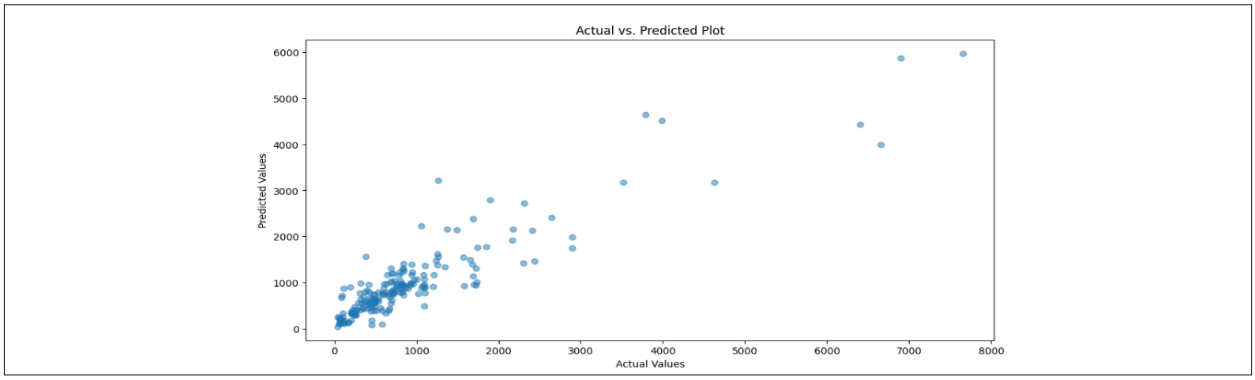
شکل (۹): نمودار مقادیر واقعی داده‌ها در مقابل مقادیر پیش‌بینی شده



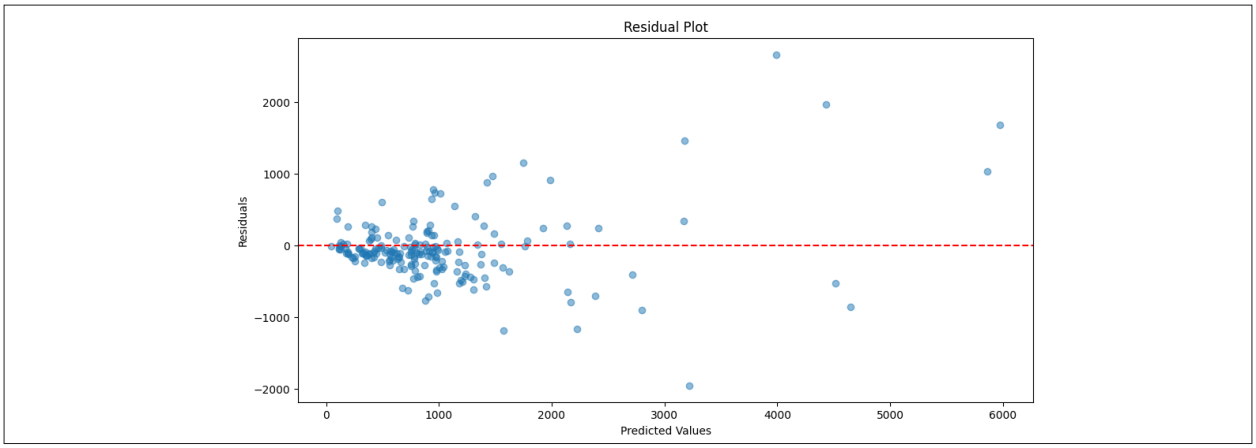
شکل (۱۰): نمودار باقیمانده‌ها یا خطاهای مدل

است. نمودارهای ۱۱و ۱۲ به ترتیب مقایسه مقادیر واقعی و پیش‌بینی‌شده و باقی‌مانده‌های مدل را نشان می‌دهند. نمودار ۱۱ نشان می‌دهد که مدل توانسته است مقادیر هدف را با دقت خوبی پیش‌بینی کند. نمودار ۱۲ نشان‌دهنده پراکندگی تصادفی باقی‌مانده‌ها حول خط قرمز است که احتمالاً ناشی از تأثیر عوامل تصادفی یا ناشناخته در داده‌ها است.

نتایج مقالات مختلف در زمینه تحلیل داده‌های زیست‌محیطی و اقتصادی نشان‌دهنده تفاوت‌های قابل توجه در دقت پیش‌بینی مدل‌های مورد استفاده است. در مقاله حاضر، از الگوریتم‌های یادگیری ماشین XGBoost، تقویت گرادیان و جنگل تصادفی برای پیش‌بینی شاخص انتشار CO_2 استفاده شده و بالاترین دقت مدل در داده‌های آزمون ۸۵ درصد گزارش شده است. مقایسه دقت نتایج این تحقیق



شکل (۱۱): نمودار مقادیر واقعی داده‌ها در مقابل مقادیر پیش‌بینی شده



شکل (۱۲): نمودار باقیمانده‌ها یا خطاهای مدل

با تحقیقات مشابه نشان می‌دهد که به‌طورکلی، استفاده از یادگیری ماشین عملکرد بهتری نسبت به روش‌های سنتی داشته است. در مطالعه حداد[15] و همکاران، انتشار CO_2 در صنعت ساخت‌وساز داخلی چین با استفاده از مدل داده–ستانده غیررقابتی تحلیل شده که دقت پیش‌بینی مدل ۷۸ درصد بوده است. این مطالعه نشان داده که تغییرات در زنجیره تأمین و وابستگی به صنایع مرتبط تأثیر زیادی بر انتشار CO_2 دارد، اما مدل مورد استفاده در تشخیص اثرات متقابل صنایع با محدودیت‌هایی مواجه بوده است [۴۰]. در تحقیق پوپوویسی[16] و همکاران، با بهره‌گیری از مدل داده–ستانده سنتی، تأثیر سرمایه‌گذاری‌های ساختاری اتحادیه اروپا (ESIF) بر اقتصاد سبز رومانی بررسی شده که دقت مدل ۷۵ درصد گزارش شده است. در این مطالعه، اثرات این سرمایه‌گذاری‌ها بر تولید و اشتغال مورد بررسی قرار گرفته، اما پیچیدگی‌های رفتاری اقتصاد و تأثیرات غیرخطی در نظر گرفته نشده است [۴۱]. در مطالعه پیچلر[17] و همکاران، مدل داده–ستانده پویا برای تحلیل اثرات اقتصادی شوک‌های ناشی از همه‌گیری استفاده شده که دقت پیش‌بینی ۸۰ درصد بوده است. نتایج این تحقیق نشان داده که اثرات همه‌گیری بر زنجیره تأمین و تولید در سطوح مختلف اقتصادی متفاوت بوده و شوک‌های اولیه تأثیرات ماندگاری بر برخی بخش‌ها گذاشته‌اند [۴۲]. در بررسی گونزالز و همکاران، که بر تأثیر سرویس‌های حمل‌ونقل برخط در برزیل متمرکز بوده، از مدل داده–ستانده بین‌منطقه‌ای استفاده شده که دقت مدل ۷۰ درصد گزارش شده است. این مطالعه تأکید کرده که تغییر در الگوهای حمل‌ونقل شهری، به‌ویژه در کلان‌شهرها، تأثیر قابل‌توجهی بر توزیع انتشار کربن و مصرف سوخت دارد، اما مدل قادر به در نظر گرفتن

15-Haddad
16-Popovici
17-Pichler

پویایی رفتـاری مصرف‌کننـدگان نبوده اسـت [۲۶]. در مجموع، مقاله حاضر که از روش‌های یادگیری ماشین بهره گرفته، موفق به ارائه دقت بالاتری در پیش‌بینی شده است، در حالی که سایر مقالات که عمدتاً مبتنی بر مدل‌های داده–ستانده بوده‌اند، به دلیل ماهیت ایستا و خطی این مدل‌ها، دقت پیش‌بینی پایین‌تری داشـته‌اند. این نشـان می‌دهد که در پیش‌بینی‌های پیچیده، روش‌های یادگیری ماشین با در نظر گرفتن ویژگی‌های غیرخطی داده‌ها و تعاملات پیچیده میان متغیرها، می‌توانند عملکرد بهتری نسبت به مدل‌های اقتصادسنجی سنتی داشته باشند.

۵. نتیجه‌گیری

این تحقیـق با بررسی و مقایسـه سـه روش تقویت گرادیـان، XGBoost، و جنگل تصادفی بـا Bootstrap به ارزیـابی کارایی این مدل‌هـا در پیش‌بیـنی ضرایب CO_2 پرداخت. نتایج نشـان دادند که مدل تقویت گرادیان با دقت ۸۵ درصد عملکرد بهتری در پیش‌بینی ضرایب CO_2 نسبت بـه دیگر مدل‌ها داشـت. در مقابـل، XGBoost با دقت ۷۹ درصد، ضعیف‌ترین عملکرد را نشان داد، و جنگل تصادفی نیـز با دقـت ۸۲ درصد، نتایـج مطلوبی به دسـت آورد. نمودارهای پراکندگی نتایج نشـان‌دهنده انطباق مناسـب بیـن مقادیر پیش‌بینی و واقـعی در اکثر مدل‌ها بود، که به تصادفی بودن خطاها اشـاره می‌کنـد و حاکی از صحت نسبی پیش‌بینی‌ها است.

این پژوهش محدودیت‌هایی نیز داشت، از جمله کمبود داده‌هـای کافی، کـه می‌تواند دقـت پیش‌بیـنی مدل‌ها را کاهش دهـد و منجر به بیش‌برازش شـود. علاوه بر این، نیاز به تجهیزات پیشـرفته برای اجـرای مدل‌ها یکی دیگر از چالش‌هـای تحقیق بـود. این محدودیت‌هـا به‌ویژه در محیط‌هایی با منابع محدود، تاثیر منفی بر انجام آزمایش‌ها و تحلیل‌ها می‌گذارد. پیشنهاد می‌شود که در تحقیقات آتی داده‌ها از صنایع مختلف و همچنیـن آلاینده‌های متنوع‌تر استفاده شود تا به درک جامع‌تری از تاثیر آلاینده‌ها دست یافـت. همچنین، ایجـاد مدل‌های خاص بـرای هر صنعت می‌تواند به شناسایی و پیشگیری از مشکلات زیست‌محیطی خاص آن کمک نماید.

منابع

1. Bjorhovde, R., 2004. Development and use of high performance steel. Journal of Constructional Steel Research, 60(3): p. 393-400.
2. Conejo, A.N., Birat, J. P., & Dutta, A. 2020. A review of the current environmental challenges of the steel industry and its value chain. Journal of Environmental Management, 259: p. 109782.
3. Burchart, D., Pichlak, M., & Kruczek, M. 2016. Innovative technologies for greenhouse gas emission reduction in steel production. Metalurgija, 55: p. 119-122.
4. Chahbi, I., N. Ben Rabah, & Tekaya, I. 2022. Towards an efficient and interpretable Machine Learning approach for Energy Prediction in Industrial Buildings: A case study in the Steel Industry. 1-8.
5. Mienye, I.D. & Sun, Y. 2022. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. IEEE Access, 10: p. 99129-99149.
6. Ionescu, A., & Candau, Y. 2007. Air pollutant emissions prediction by process modelling – Application in the iron and steel industry in the case of a re-heating furnace. Environmental Modelling & Software, 22: p. 1362-1371.
7. Yu, R., et al., 2016. RAQ–A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems. Sensors, 16: p. 86.
8. Ribeiro, V.M., 2021. Sulfur dioxide emissions in Portugal: Prediction, estimation and air quality regulation using machine learning. Journal of Cleaner Production, 317: p. 128358.
9. Shakya, D., et al., 2023. PM2.5 air pollution prediction through deep learning using meteorological, vehicular, and emission data: A case study of New Delhi, India. Journal of Cleaner Production, 427: p. 139278.
10. Karimi, S., et al., 2023. Machine learning-based white-box prediction and correlation analysis of air pollutants in proximity to industrial zones. Process Safety and Environmental Protectio. 178: p. 1009-1025.
11. Bekkar, A., et al., 2021. Air-pollution prediction in smart city, deep learning approach. Journal of Big Data, 8(1): p. 161.
12. Moreira, G.d.A., et al., 2023. Analyzing the Influence of Vehicular Traffic on the Concentration of Pollutants in the City of São Paulo: An Approach Based on Pandemic SARS-CoV-2 Data and Deep Learning. Atmosphere, 14(10): p. 1578.
13. Park, C., et al., 2023. Machine learning based estimation of urban on-road CO2 concentration in Seoul. Environmental Research, 231: p. 116256.

14. Jalaludin, J., et al., 2023. The Impact of Air Quality and Meteorology on COVID-19 Cases at Kuala Lumpur and Selangor, Malaysia and Prediction Using Machine Learning. Atmosphere, 14(6): p. 973.
15. Cican, G., Buturache, A. N., & Mirea, R. 2023. Applying Machine Learning Techniques in Air Quality Prediction-A Bucharest City Case Study. Sustainability, 15(11): p. 8445.
16. Wen, C., et al., 2023. Dioxin emission prediction from a full-scale municipal solid waste incinerator: Deep learning model in time-series input. Waste Management, 2023. 170: p. 93-102.
17. Ravindiran, G., et al., 2023. Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam. Chemosphere, 338: p. 139518.
18. Wu, C.L., et al., 2023. A hybrid deep learning model for regional O(3) and NO(2) concentrations prediction based on spatiotemporal dependencies in air quality monitoring network. Environ Pollut, 320: p. 121075.
19. Feng, X. 2023. Air pollution prediction in context of supervised machine learning and time series model. in Second International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2022). SPIE.
20. Natarajan, Y., et al., 2023. Forecasting Carbon Dioxide Emissions of Light-Duty Vehicles with Different Machine Learning Algorithms. Electronics, 12(10): p. 2288.
21. Yarragunta, S., et al., 2021. Prediction of Air Pollutants Using Supervised Machine Learning. 1633-1640.
22. Paternina- Arboleda, C.D., et al., 2023. Towards Cleaner Ports: Predictive Modeling of Sulfur Dioxide Shipping Emissions in Maritime Facilities Using Machine Learning. Sustainability, 15(16): p. 12171.
23. Xia, H., et al., 2023. Dioxin emission modeling using feature selection and simplified DFR with residual error fitting for the grate-based MSWI process. Waste Manag, 168: p. 256-271.
24. Muchdie, M., Ulza, E. & Setiawan, E. 2018. Sector and Country Balance of Trade Analysis Based on World Input-Output Database: Indonesian Economy. p. 109-129.
25. Lu, Y., 2017. China's electrical equipment manufacturing in the global value chain: A GVC income analysis based on World Input-Output Database (WIOD). International Review of Economics & Finance, 52: p. 289-301.
26. Pascual-González, J., et al., 2015. Statistical analysis of global environmental impact patterns using a world multi-regional input–output database. Journal of Cleaner Production, 90: p. 360-369.
27. El Haouti, K., El Aiss, H., & Barra, A. 2025. Input output stability analysis of delayed descriptor systems according to free weighting matrices. International Journal of Dynamics and Control, 13(2): p. 67.
28. Su, B. & Ang, B.W. 2014. Input–output analysis of CO2 emissions embodied in trade: A multi-region model for China. Applied Energy, 114: p.384- 377.
29. El Naqa, I. & Murphy, M.J. 2015. What is machine learning? 2015: Springer.
30. Gong, M., et al., 2023. Threshold of 25(OH)D and consequently adjusted parathyroid hormone reference intervals: data mining for relationship between vitamin D and parathyroid hormone. J Endocrinol Invest, 46(10): p. 2067-2077.
31. Yang, P., & Huang, B. 2008. KNN Based Outlier Detection Algorithm in Large Dataset. Education Technology and Training & Geoscience and Remote Sensing, 1: p. 611-613.
32. Fafalios, S., Charonyktakis, P., & Tsamardinos, I. 2020. Gradient Boosting Trees.
33. Chen, T., & Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. 785-794.
34. Lee, T.-H., Ullah, A., & Wang, R. 2020. Bootstrap Aggregating and Random Forest. p. 389-429.
35. Sapra, R.L., 2014. Using R2 with caution. Current Medicine Research and Practice, 4(3): p. 130-134.
36. Chicco, D., Warrens, M.J. & Jurman, G. 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7: p. e623.
37. Liu, Y. 2008. Box plots: use and interpretation. Transfusion, 48(11): p. 2279-2280.
38. Scott, D.W., 2010. Histogram. Wiley Interdisciplinary Reviews: Computational Statistics, 2(1): p. 4 48-4.
39. Tripathi, A. & Rani, P. 2024. An Improved MSER using Grid Search based PCA and Ensemble Voting Technique. Multimedia Tools and Applications, 83(34): p. 80497-80522.
40. Haddad, E.A., et al., Off We Go! An interregional input-output analysis of e-hailing in Brazil. Journal of Urban Management, 2022. 11(2): p. 188-197.
41. Davidescu, A.A., O.C. Popovici, and V.A. Strat, Estimating the impact of green ESIF in Romania using input-output model. International Review of Financial Analysis, 2022. 84: p. 102336.
42. Pichler, A., et al., Forecasting the propagation of pandemic shocks with a dynamic input-output model. Journal of Economic Dynamics and Control, 2022. 144: p. 104527.