



Research in Production and Operations Management
University of Isfahan E-ISSN: 2981-0329
Vol. 16, Issue 2, No. 41, summer 2025



DOI: [10.22108/pom.2025.144489.1610](https://doi.org/10.22108/pom.2025.144489.1610)

(Research paper)

Cost-Sensitive Machine Learning for Predicting Production Defects: a Novel Approach Based on MetaCost

Ahmad Jafarnejad *

Department of Industrial Management, Faculty of Management, University of Tehran, Tehran, Iran,
jafarnjd@ut.ac.ir

Arman Rezasoltani

Department of Industrial Management, Faculty of Management, University of Tehran, Tehran, Iran,
armanrezasoltani@ut.ac.ir

Amir Mohammad Khani

Department of Industrial Management, Faculty of Management, University of Tehran, Tehran, Iran,
amir.mo.khani@ut.ac.ir

Purpose: This paper aims to apply cost-sensitive machine learning (CSML) for predicting production defects in industrial processes to minimize the costs associated with false negatives. It investigates the MetaCost algorithm to assess its effectiveness in enhancing accuracy for defect detection when misclassification costs are incorporated into predictive models.

Design/methodology/approach: Real-world industrial production data is utilized to train multiple machine learning models: Random Forest, XGBoost, LightGBM, CatBoost, Gradient Boosting, SVM, and Logistic Regression to predict high-defect days. Subsequently, we implement various techniques to address class imbalance and enhance generalization, including Borderline-SMOTE, Recursive Feature Elimination with Cross Validation (RFECV), and Optuna hyperparameter optimization. After modeling, MetaCost is applied as a post-processing step to convert base learners into cost-sensitive classifiers with a 5:1 cost ratio focused on minimizing false negatives.

Findings: Findings indicated that among all tested classifiers, the most effective model for identifying high defect production days is the Random Forest model combined with MetaCost, as it boasts the highest recall (98.9%). In industrial settings, this high recall is particularly crucial, as failing to detect

* Corresponding author, 0000-0001-6763-5033

2981-0329 / © University of Isfahan



This is an open access article under the CC-BY-NC-ND 4.0 License (<https://creativecommons.org/licenses/by-nc-nd/4.0/>)

defects can lead to significant losses, waste, and quality degradation for the company. CatBoost and LightGBM also performed well in terms of precision and false negatives, matching Random Forest's performance, but they recorded slightly more false negatives. Additionally, the Random Forest model achieved a balanced F1 score, making it reliable and robust for imbalanced datasets. This highlights the importance of cost-sensitive models in quality control tasks, especially when the cost of false negatives far exceeds that of false positives. Random Forest consistently outperforms in this study, making it a promising candidate for practical application in real-world manufacturing, as it can predict problematic production days, prompting timely interventions and continuous process improvement. The findings thus confirm both the strategic and technical value of the model for industrial decision-making processes.

Research limitations/implications: The industrial variations in the real world may not be fully represented by the study's dataset. Results are confined to batch-based manufacturing data and require validation in continuous or real-time processes. Future research should concentrate on deep learning models (CNN or RNN) and real-time sensor data or IoT-based streams for dynamic settings.

Practical implications: This study offers a data-driven framework for manufacturers to predict and address production defects, thereby reducing waste and rework costs. The approach minimizes undetected high-defect days while enhancing production quality, decreasing operational losses, and enabling the use of simple ML-based solutions instead of costly diagnostic equipment.

Social implications: Enhanced defect prediction for sustainable manufacturing is achieved through reduced material waste, prolonged product lifespan, and improved quality control. Greater product reliability fosters stronger consumer trust and diminishes the environmental impact of industrial waste.

Originality/value: The MetaCost initiative has implemented this algorithm as one of the first of its kind in the real industrial defect prediction domain. Cost-sensitive learning formulations with modern ML tools are employed to bridge a gap in research that addresses production optimization while balancing quality and cost. This research is valuable both academically and for industry practitioners interested in economic considerations in predictive modeling. Furthermore, while previous works focused primarily on healthcare, finance, and software defect detection, this research introduces a novel application of MetaCost in a manufacturing environment characterized by highly imbalanced class distributions and complex processes. The study proposes a replicable and scalable pipeline by integrating techniques such as Borderline-SMOTE, RFECV, and Optuna, which can be utilized for future research as well as in real-time applications. It effectively closes the gap between theoretical ML design and the application of ML in high-impact quality improvement of industrial systems.

Keywords: Cost-sensitive learning, MetaCost, Defect prediction, Manufacturing, Machine learning, Production quality



پژوهش در مدیریت تولید و عملیات، دوره ۱۶، شماره ۲، پیاپی ۴۱، تابستان ۱۴۰۴
دریافت: ۱۴۰۳/۱۲/۱۲ پذیرش: ۱۴۰۴/۰۲/۰۸ ص ۷۳-۹۴



DOI: [10.22108/pom.2025.144489.1610](https://doi.org/10.22108/pom.2025.144489.1610)

(مقاله پژوهشی)

یادگیری ماشین با حساسیت هزینه برای پیش‌بینی نقص‌های تولید:

رویکردی نوین مبتنی بر MetaCost

احمد جعفرنژاد*^۱؛ آرمان رضاسلطانی^۲؛ امیرمحمد خانی^۳

۱- استاد گروه مدیریت صنعتی، دانشکده مدیریت، دانشگاه تهران، تهران، ایران، jafarnjd@ut.ac.ir

۲- دانشجوی دکتری گروه مدیریت صنعتی، دانشکده مدیریت، دانشگاه تهران، تهران، ایران، armanrezasoltani@ut.ac.ir

۳- دانشجوی دکتری گروه مدیریت صنعتی، دانشکده مدیریت، دانشگاه تهران، تهران، ایران، amir.mo.khani@ut.ac.ir

چکیده: کنترل کیفیت و کاهش هزینه‌های تولید، به پیش‌بینی دقیق عیوب در فرآیندهای صنعتی وابسته است. در این پژوهش، رویکرد یادگیری ماشین حساس به هزینه، با استفاده از الگوریتم MetaCost بررسی شده است. MetaCost یک تکنیک پس‌پردازش برای تبدیل مدل‌های یادگیری ماشین به مدل‌های حساس به هزینه است که با در نظر گرفتن ماتریس هزینه خطاها، تصمیم‌گیری مدل را بهینه می‌کند. هدف اصلی، کاهش خطاهای منفی کاذب در شناسایی روزهای پرنقص تولید است. برای این منظور، از چندین الگوریتم شامل Gradient Boosting, Random Forest, XGBoost, LightGBM, SVM و رگرسیون لجستیک استفاده شد. داده‌ها از دیتاست «Predicting Manufacturing Defects» بر گرفته از پلتفرم Kaggle، شامل اطلاعات مربوط به ۳۲۴۰ روز تولید صنعتی جمع‌آوری شدند. نتایج نشان داد که الگوریتم جنگل تصادفی با دستیابی به صحت برابر ۹۶٫۹٪ و بازخوانی برابر ۹۸٫۹٪، بهترین عملکرد را در میان مدل‌ها داشت. به‌ویژه توانایی بالای این مدل در شناسایی صحیح روزهای پرنقص، آن را به گزینه مناسبی برای کاربردهای واقعی در صنعت تبدیل کرد. دیگر مدل‌ها نیز عملکرد پذیرفتنی داشتند؛ اما در مقایسه با Random Forest، در کاهش نرخ منفی کاذب ضعیف‌تر ظاهر شدند. این نتایج، کارایی رویکردهای حساس به هزینه را در بهبود پیش‌بینی نقص تولید، تأیید می‌کند.

واژه‌های کلیدی: یادگیری ماشین، حساسیت به هزینه، MetaCost، پیش‌بینی نقص، تولید صنعتی



در دنیای پرشتاب تولید و عملیات امروزی، توانایی پیش‌بینی عیوب برای تضمین کیفیت و عملکرد بسیار مهم است. با اتکای بیشتر صنایع به تصمیم‌گیری مبتنی بر داده، یادگیری ماشین (ML) به ابزار قدرتمندی برای پیش‌بینی عیوب تولید تبدیل شده است (Barzizza et al., 2024., Kang et al., 2020., van Vuuren, 2024). با این حال در کاربردهای حیاتی، رویکردهای سنتی و ساده ML، هزینه‌های متفاوت مرتبط با انواع مختلف طبقه‌بندی‌های اشتباه را نادیده می‌گیرند که به‌طور بالقوه، به نتایج ضعیف منجر می‌شود. در این مطالعه، از چارچوب MetaCost برای پیش‌بینی نقص تولید استفاده و رویکرد جدیدی برای یادگیری ماشین حساس به هزینه (CSML) ایجاد می‌شود. این مقاله به این سؤال تحقیقاتی پاسخ می‌دهد که چگونه پیش‌بینی دقیق عیوب تولید با در نظر گرفتن هزینه‌های طبقه‌بندی اشتباه مرتبط است. هزینه منفی کاذب (شناسایی نکردن نقص در صورت لزوم) در بسیاری از زمینه‌های صنعتی بسیار بیشتر از هزینه مثبت کاذب (تشخیص نادرست نقص) است. برای بررسی این نبود تعادل، باید یک رویکرد حساس به هزینه برای مدل‌های پیش‌بینی‌کننده اتخاذ کرد که به دنبال به حداقل رساندن هزینه‌های طبقه‌بندی اشتباه به جای حداکثر کردن صحت‌اند. همان‌طور که بررسی پیشینه موجود نشان می‌دهد، یادگیری حساس به هزینه، زمینه بسیار مهمی است که پیشرفت خوبی را به خود دیده است؛ اما در اینجا نیز بیشتر کارها یا بر چارچوب نظری یا در برخی کاربردهای خاص، بدون در نظر گرفتن پیچیدگی‌های پیش‌بینی نقص تولید باقی مانده است؛ برای مثال، ژو و لیو^۳ (۲۰۱۰) اشاره می‌کنند که کمبود اکتشاف در یادگیری چند کلاس حساس به هزینه وجود داشته است و لیو و همکاران^۴ (۲۰۲۱) بر لزوم گنجانیدن اطلاعات هزینه در آموزش مدل تأکید می‌کنند. منولد و همکاران^۵ (۲۰۲۴) به‌جای استفاده از یادگیری مستقیم حساس به هزینه در زمینه تولید، بر کاربرد یادگیری ماشین در حوزه مراقبت‌های بهداشتی تمرکز کرده‌اند. این نتایج نشان می‌دهند که لازم است بررسی‌های بیشتری درباره استفاده از روش‌های پیش‌بینی نقص تولید با رویکرد حساس به هزینه انجام شود. در زمینه بهینه‌سازی مدل‌های پیش‌بینی و یادگیری توالی‌های زمانی، به استفاده از الگوریتم‌های تکاملی مانند Modified Asexual Reproduction Optimization و Asexual Reproduction Optimization، برای آموزش مدل‌های مارکوف پنهان نیز، توجه شده است (Mansouri et al., 2021).

در حال حاضر، هزینه‌های منفی کاذب در پیش‌بینی نقص‌های تولید، به‌ویژه در صنعت، یکی از مسائل اساسی به شمار می‌آید. این هزینه‌ها به دلیل شناسایی نکردن دقیق نقص‌ها و عواقب مالی ناشی از آنها، اهمیت زیادی دارند. در کشورهای مختلف، از جمله ایران، توجه به کاهش این هزینه‌ها و بهبود پیش‌بینی‌ها در سال‌های اخیر افزایش یافته است. به‌ویژه در صنایع تولیدی ایران، شناسایی نکردن صحیح نقص‌های تولید، به زیان‌های مالی درخور توجهی منجر می‌شود؛ برای مثال، مطالعات انجام‌شده در ایران نشان می‌دهند که در صنایع مختلف، مانند خودروسازی و الکترونیک، هزینه‌های منفی کاذب به دلیل شناسایی نکردن صحیح نقص‌ها، به زیان‌های مالی زیادی برای تولیدکنندگان منجر می‌شود (Motamedi et al., 2024; Ataei et al., 2025). در سطح جهانی نیز، تحقیقاتی نشان داده‌اند که هزینه‌های منفی کاذب در صنایع مختلف، از جمله صنایع دارویی، تولید خودرو و الکترونیک، فشار زیادی به تولیدکنندگان وارد می‌کند؛ برای نمونه، مطالعات اخیر در کشورهای صنعتی پیشرفته نشان می‌دهند که این

هزینه‌ها، آثار مالی چشمگیری در صنایع مختلف دارند (Barzizza et al., 2024). به همین دلیل، استفاده از رویکردهای حساس به هزینه مانند MetaCost برای کاهش این هزینه‌ها در صنایع تولیدی، اهمیت ویژه‌ای دارد. سهم اصلی این پژوهش، توسعه MetaCost برای پیش‌بینی عیوب ساخت و ارزیابی مدل‌های مختلف یادگیری ماشین از نظر پیش‌بینی عیوب ساخت بوده است تا بهترین مدل برای پیش‌بینی نقص‌های ساخت شناسایی شود. کاربرد عملی تحقیق درباره بهبود تصمیمات مدیریتی در صنعت تولید، کاهش هزینه ساخت عیوب، ساخت و بهبود کیفیت فرآیند حائز اهمیت بوده است. این تحقیق فرصتی را برای تولیدکنندگان فراهم کرده است تا از زیان‌های مالی و افت کیفیت در روزهای پرنقص جلوگیری کنند. نقشه راه این مقاله به شرح زیر بوده است: ابتدا مقدمه‌ای بر روش‌های یادگیری ماشین حساس به هزینه و MetaCost ارائه و در ادامه، مبانی نظری و پیشینه پژوهش بررسی شده است؛ سپس، روش تحقیق و مجموعه داده‌ها توضیح داده شده است. در ادامه، مدل‌های مختلف یادگیری ماشین برای پیش‌بینی نقص‌ها بررسی و ارزیابی شده‌اند. در نهایت، نتایج و تحلیل‌های مربوط به ارزیابی عملکرد مدل‌ها و انتخاب بهترین مدل برای پیش‌بینی نقص‌های تولید، ارائه شده است.

پیشینه پژوهش

مطالعات جدول یک نشان داده‌اند که MetaCost و دیگر تکنیک‌های یادگیری ماشین حساس به هزینه، عملکرد مدل‌ها را در داده‌های نامتعادل و پرهزینه بهبود می‌دهند. این روش‌ها در غربالگری پزشکی، پیش‌بینی نقص نرم‌افزار، کشف تقلب مالی و سرمایه‌گذاری کاربرد داشته‌اند. همچنین، ترکیب این الگوریتم‌ها با روش‌هایی مانند یادگیری فعال یا نمونه‌برداری بیش از حد، دقت پیش‌بینی را در مسائل پیچیده، مانند ورشکستگی و نقص نرم‌افزار افزایش داده است. این رویکردها هزینه اشتباهات طبقه‌بندی را کاهش می‌دهند و به تصمیم‌گیری دقیق‌تر کمک می‌کنند.

جدول ۱- پیشینه پژوهش

Table 1- Research background

نویسندگان	عنوان مقاله	هدف	دیتاست	نتیجه‌گیری
ستی و همکاران ^۶ (۲۰۲۴)	«یادگیری ماشینی حساس به هزینه برای حمایت از تصمیمات سرمایه‌گذاری راه‌اندازی»	بررسی استفاده از MetaCost برای کاهش ریسک در سرمایه‌گذاری استارت‌آپ‌ها	داده‌های ریسک را به طرز درخور توجهی برای استارت‌آپ‌ها	مدل‌های یادگیری ماشین حساس به هزینه، ریسک را به طرز درخور توجهی برای سرمایه‌گذاران کاهش می‌دهند.
منولد و همکاران (۲۰۲۴)	«یادگیری ماشین در غربالگری خودکار برای مرورهای سیستماتیک و متا آنالیز در اورولوژی»	برای ارزیابی پتانسیل یادگیری حساس به هزینه در غربالگری داده‌های پزشکی	مجموعه داده‌های مرور پزشکی	این مطالعه نتیجه می‌گیرد که رویکردهای حساس به هزینه، صحت فرآیندهای غربالگری خودکار را افزایش می‌دهند
مائلکوئیست و همکاران ^۷ (۲۰۲۳)	«پشتیبانی تصمیم‌گیری حساس به هزینه برای فرآیندهای دسته‌ای صنعتی»	توسعه یک سیستم پشتیبانی تصمیم که شامل یادگیری حساس به هزینه برای فرآیندهای دسته‌ای می‌شود.	داده‌های فرآیند دسته‌ای صنعتی	این مطالعه نشان می‌دهد که پشتیبانی تصمیم‌گیری حساس به هزینه، کارایی عملیاتی را به‌طور درخور توجهی در تنظیمات صنعتی افزایش می‌دهد.
چن ^۸ (۲۰۲۱)	«پژوهش در روش‌های طبقه‌بندی حساس به هزینه برای داده‌های نامتوازن»	معرفی الگوریتم ترکیبی جدید شامل یادگیری فعال و MetaCost برای بهبود دسته‌بندی نامتوازن	داده‌های عمومی و ترکیبی	ترکیب یادگیری فعال و MetaCost به‌طور مؤثری عملکرد پیش‌بینی را در داده‌های نامتوازن بهبود می‌بخشد.
نیو و همکاران ^۹ (۲۰۲۰)	«یادگیری دیکشنری حساس به هزینه برای پیش‌بینی نقص نرم‌افزار»	بهبود پیش‌بینی نقص‌های نرم‌افزاری با یادگیری دیکشنری	مجموعه داده‌های نرم‌افزاری	مدل‌های جدید دیکشنری یادگیری حساس به هزینه، عملکرد بهتری در

نویسندگان	عنوان مقاله	هدف	دیتاست	نتیجه گیری
		حساس به هزینه	NASA و AEEEM	پیش بینی نقص های نرم افزاری نسبت به روش های قبلی دارند.
قطعه‌هاشه و همکاران ^{۱۰} (۲۰۲۰)	«تجزیه و تحلیل کسب و کار در بهبود دقت و صحت پیش بینی در بازاریابی تلفنی: تحلیل حساس به هزینه کمپین های بازاریابی تلفنی با استفاده از شبکه های عصبی حساس به هزینه»	بررسی روش های oversampling و یادگیری حساس به هزینه برای پیش بینی های تلفنی	داده های کمپین بازاریابی تلفنی	نتایج نشان می دهد که مدل های حساس به هزینه به طور درخور توجهی بهتر از مدل های سنتی در مجموعه داده های نامتعادل عمل می کنند
وریکه و همکاران ^{۱۱} (۲۰۲۰)	«مبانی طبقه بندی علی حساس به هزینه»	پیشنهاد چارچوبی برای طبقه بندی علی حساس به هزینه برای افزایش تصمیم گیری	مجموعه داده های مختلف	این چارچوب راه های جدیدی را برای بهبود تصمیمات تجاری مبتنی بر داده از طریق روش های حساس به هزینه باز می کند
مالهوترا و کمال ^{۱۲} (۲۰۱۹)	«یک مطالعه تجربی برای بررسی روش های افزایشی نمونه به منظور بهبود پیش بینی نقص نرم افزار با استفاده از داده های نامتوازن»	بررسی روش های oversampling و یادگیری حساس به هزینه برای پیش بینی نقص های نرم افزاری	مجموعه داده های NASA	استفاده از oversampling و یادگیری حساس به هزینه با MetaCost پیش بینی های بهتری را برای داده های نامتوازن ارائه می دهد.
لی و همکاران ^{۱۳} (۲۰۱۹)	«یک رویکرد ترکیبی با استفاده از تکنیک نمونه برداری بیش از حد و یادگیری حساس به هزینه برای پیش بینی ورشکستگی»	برای ایجاد یک مدل ترکیبی برای پیش بینی ورشکستگی که نمونه برداری بیش از حد و یادگیری حساس به هزینه را ترکیب می کند.	مجموعه داده ورشکستگی کشور کره	رویکرد ترکیبی دقت پیش بینی بهبود یافته ای را در مجموعه داده های بسیار نامتعادل به همراه دارد.
کامالاروبان و ویلیامسون ^{۱۴} (۲۰۱۸)	«تعیین حداقل حد پایین برای طبقه بندی حساس به هزینه»	تجزیه و تحلیل طبقه بندی باینری حساس به هزینه و پیامدهای آن در کاربردهای مختلف	مجموعه داده های مختلف	نتایج نشان می دهد که نادیده گرفتن هزینه های طبقه بندی نادرست به عملکرد نامناسب در کاربردهای حیاتی منجر می شود
کیم و همکاران ^{۱۵} (۲۰۱۶)	«شناسایی تحریفات مالی با قصد کلاه برداری با استفاده از یادگیری حساس به هزینه چند طبقه»	توسعه مدل های یادگیری چند کلاسه برای تشخیص کلاه برداری در گزارش های مالی	داده های مالی شرکت ها	یادگیری حساس به هزینه با استفاده از MetaCost کمک می کند تا قصد کلاه برداری در بیانیه های مالی شناسایی شود.
سلطانی و همکاران ^{۱۶} (۲۰۲۳)	«رویکرد ترکیبی پیش بینی تقاضای کانال همه جانبه یکپارچه، با استفاده از یادگیری ماشین»	هدف این مقاله کاهش عدم قطعیت در پیش بینی تقاضا در خرده فروشی چندکاناله از طریق استفاده از الگوریتم های یادگیری ماشین و خوشه بندی سری های زمانی است.	داده های فروش ماهیانه از یک خرده فروشی لوازم آرایشی از فوریه ۲۰۲۰ تا ژوئن ۲۰۲۲.	استفاده از الگوریتم خوشه بندی با Dynamic Time Warping و به کارگیری شبکه های عصبی غیرخطی خودبازگشتی با ورودی های خارجی (NARX) به بهبود دقت پیش بینی تقاضا برای محصولات در خرده فروشی های چندکاناله کمک می کند
میرزایی و همکاران ^{۱۷} (۲۰۲۳)	«مقایسه کارایی مدل های آماری و مدل های یادگیری ماشین و انتخاب مدل بهینه برای پیش بینی سود خالص و جریان نقد عملیاتی»	این تحقیق، کارایی مدل های آماری و مدل های یادگیری ماشین را برای پیش بینی سود خالص و جریان نقد عملیاتی مقایسه می کند و هدف آن انتخاب مدل بهینه برای پیش بینی این دو متغیر مالی در شرکت هاست.	داده های مربوط به ۱۸۴ شرکت ثبت شده در بورس تهران از سال ۲۰۱۲ تا ۲۰۲۱.	نتایج نشان دادند که متغیرهای تعهدی قدرت تبیینی بیشتری نسبت به متغیرهای نقدی برای پیش بینی سود خالص و جریان نقد عملیاتی دارند.

پژوهش‌های مختلف در حوزه یادگیری ماشین حساس به هزینه^{۱۸} CSL و رویکرد MetaCost نشان داده‌اند که این روش در زمینه‌های گوناگونی از جمله سرمایه‌گذاری استارت‌آپ‌ها، پیش‌بینی نقص‌های نرم‌افزاری، تحلیل‌های پزشکی و پیش‌بینی ورشکستگی به کار رفته است. با این حال، هر مطالعه دارای مزایا، محدودیت‌ها و شکاف‌های پژوهشی خاص خود است. مطالعه ستی و همکاران (۲۰۲۴) نشان داد که MetaCost، ریسک سرمایه‌گذاری را در استارت‌آپ‌ها کاهش می‌دهد؛ اما نیازمند بررسی در صنایع سنتی است. این روشی که منولد و همکاران (۲۰۲۴) در غربالگری پزشکی به کار بردند، به بهبود دقت منجر شد، اما به ارزیابی‌های بالینی بیشتری نیاز دارد. چن (۲۰۲۱) ترکیب یادگیری فعال و MetaCost را برای داده‌های نامتعادل بررسی کرد؛ اما کاربرد آن در صنایع تولیدی کمتر بررسی شده است. درنهایت، نیو و همکاران (۲۰۲۰) از یادگیری دیکشنری حساس به هزینه برای پیش‌بینی نقص‌های نرم‌افزاری استفاده کردند؛ اما این مطالعه به مجموعه داده‌های بزرگ‌تر و متنوع‌تری نیاز دارد. علاوه بر این، مطالعه مانلکویست و همکاران (۲۰۲۳) یکی از معدود مطالعاتی است که CSL را در بستر فرآیندهای دسته‌ای صنعتی بررسی کرده است؛ اما دامنه آن محدود به شرایط خاص فرآیندهای دسته‌ای^{۱۹} بوده و کاربرد در محیط‌های تولید پیوسته یا مبتنی بر داده‌های بلادرنگ را بررسی نکرده است. مطالعات داخلی نظیر سلطانی و همکاران (۲۰۲۳) و میرزایی و همکاران (۲۰۲۳) نیز، از یادگیری ماشین برای حل مسائل پیش‌بینی در حوزه‌هایی مانند تقاضای خرده‌فروشی و تحلیل مالی بهره برده‌اند؛ ولی رویکرد حساس به هزینه در این مطالعات محوریت نداشته و مسئله طبقه‌بندی نابرابر و هزینه‌محور را به صورت مستقیم بررسی نکرده است.

در مجموع، این تحقیقات نشان‌دهنده تأثیرگذاری یادگیری حساس به هزینه در بهبود پیش‌بینی در داده‌های نامتعادل‌اند؛ اما شکاف پژوهشی در کاربردهای صنعتی و بررسی عملیاتی آنها همچنان وجود دارد. این تحقیق بر استفاده از یادگیری ماشین حساس به هزینه، برای پیش‌بینی نقص‌های تولیدی در فرآیندهای صنعتی تمرکز دارد. شکاف پژوهشی اصلی، محدودیت کاربرد MetaCost در این حوزه است؛ در حالی که تحقیقات پیشین بیشتر بر سرمایه‌گذاری استارت‌آپ‌ها، نقص‌های نرم‌افزاری و پزشکی متمرکز بوده‌اند. این پژوهش، با تمرکز بر شناسایی دقیق روزهای پرنقص و کاهش هزینه‌های پیش‌بینی اشتباه، این شکاف را پر می‌کند. مشارکت کلیدی پژوهش، استفاده از MetaCost برای تشخیص نقص‌های تولیدی است که همراه با Synthetic Minority Over-sampling Technique (SMOTE) برای متوازن‌سازی داده‌ها، RFECV^{۲۰} برای انتخاب ویژگی و Optuna برای تنظیم فرآیندها اجرا شده است. برخلاف مطالعات پیشین، این پژوهش از داده‌های پیچیده تولیدی استفاده و مدل‌های مختلف یادگیری ماشین را ارزیابی و مقایسه کرده است. انتخاب بهترین مدل براساس کاهش خطای منفی کاذب، که برای فرآیندهای صنعتی حیاتی است، انجام می‌شود. مشارکت اصلی این پژوهش در ارائه یک چارچوب نوآورانه برای پیش‌بینی نقص‌های تولیدی، با استفاده از یادگیری ماشین حساس به هزینه نهفته است. برخلاف مطالعات پیشین که عمدتاً از الگوریتم MetaCost در حوزه‌هایی مانند پزشکی، نرم‌افزار و مالی بهره برده‌اند، این تحقیق برای نخستین بار به طور متمرکز، از MetaCost در محیط‌های واقعی تولید صنعتی استفاده کرده است. این چارچوب با ترکیب تکنیک‌هایی نظیر SMOTE، برای متوازن‌سازی داده‌ها، RFECV^{۲۱} برای انتخاب ویژگی‌های مؤثر و Optuna برای تنظیم بهینه فرآیندها، یک مدل بهینه را برای داده‌های نامتوازن ارائه می‌دهد. همچنین پژوهش حاضر با انجام تحلیل مقایسه‌ای میان چندین الگوریتم یادگیری ماشین، از جمله CatBoost, LightGBM, XGBoost, Random Forest و

SVM در بستر حساس به هزینه، مدلی را معرفی می‌کند که با کاهش خطاهای منفی کاذب، امکان تصمیم‌گیری مؤثرتر را در فرآیندهای تولیدی فراهم می‌کند.

روش تحقیق

در این تحقیق، هدف اولیه بهبود پیش‌بینی عیوب تولید از طریق محاسبه هزینه‌های تصمیم‌گیری بود. به‌منظور انجام این کار، ما از تکنیک یادگیری ماشین حساس به هزینه (CSL)، با استفاده از MetaCost استفاده کردیم. این تکنیک الگوریتم‌های یادگیری ماشین را تغییر می‌دهد تا هزینه تصمیم‌گیری را در فرآیند پیش‌بینی بگنجانند و خطاهای پرهزینه را براساس احتمال کاهش دهد. نکته کلیدی برای مدیران تولید و مهندسان کیفیت، شناسایی روزهایی است که ایرادات فراوان دارند. اگر یک روز، یک مدل یادگیری ماشینی با نقص زیاد را به یک روز بدون نقص (منفی کاذب بالا) به‌اشتباه طبقه‌بندی کند، ممکن است اقدامات اصلاحی ضروری انجام نشود و به زیان مالی و کاهش کیفیت تولید در مقیاس منجر شود. در یک خط تولید صنعتی، هر خط از مجموعه داده حاوی اطلاعات مربوط به یک روز تولید خاص است؛ در نتیجه، یک (FN^{۲۲}) یک روز کامل تولید با عیوب احتمالی را نشان می‌دهد که تشخیص داده نشده است و به‌طور بالقوه، کیفیت محصولات تولیدشده را در آن روز به خطر می‌اندازد. داده‌های این مطالعه از مجموعه داده «Predicting Manufacturing Defects» استخراج شده‌اند که ربیع الخاروا^{۳۳} (۲۰۲۴) آن را در پلتفرم Kaggle منتشر کرده است و شامل ۱۶ ویژگی مستقل و یک متغیر هدف^{۲۴} است و وضعیت تولید را در یک روز نشان می‌دهد. تحلیل اولیه نشان داد که داده‌ها نامتعادل‌اند و روزهای با نقص زیاد به‌مراتب بیشتر از روزهای با نقص کم‌اند. برای رفع این مشکل، از Borderline-SMOTE برای متوازن‌سازی داده‌ها استفاده شد (Javad Hemmatian et al., 2025). برای کاهش پیچیدگی مدل و حذف ویژگی‌های کم‌اهمیت، روش RFECV با یک مدل تقویت‌گرادیان بهینه‌شده با Optuna به کار گرفته شد و ویژگی‌های کلیدی را انتخاب کرد (Shi et al., 2024). این فرآیند دقت مدل را افزایش و خطر بیش‌برازش را کاهش داد. چندین الگوریتم یادگیری ماشین حساس به هزینه از جمله CatBoost، LightGBM، XGBoost، Gradient Boosting، Random Forest، SVM^{۲۵} و رگرسیون لجستیک برای پیش‌بینی نقص تولید استفاده شد. این مدل‌ها، توانایی تخمین احتمال را دارند که برای اجرای MetaCost ضروری است. Optuna، برای بهینه‌سازی هایپرپارامترهای مدل‌ها به کار گرفته شد. در این مطالعه، MetaCost به‌گونه‌ای تنظیم شد که جریمه‌ای ۵ برابر برای خطای منفی کاذب در کلاس نقص زیاد در نظر گرفته شود؛ زیرا این خطاها آثار مهمی بر کیفیت و هزینه تولید دارند (Hassan, 2017., Barzizza et al., 2024). عملکرد مدل‌ها براساس صحت، دقت، بازخوانی و امتیاز F1 ارزیابی و تمام تحلیل‌ها با زبان برنامه‌نویسی پایتون انجام شد. شکل ۱، فلوچارت مراحل اجرای تحقیق را نشان می‌دهد.



شکل ۱- فلوچارت مراحل اجرای تحقیق

Fig. 1- Flowchart of research implementation steps

منبع داده و ویژگی‌های داده

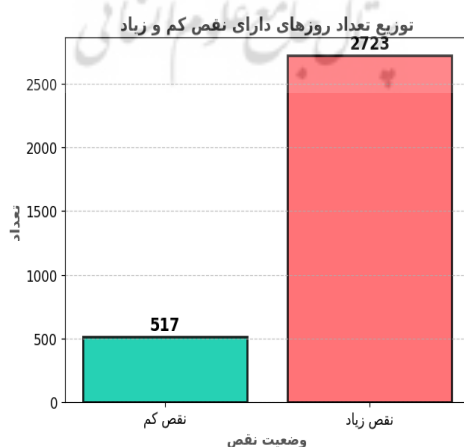
مجموعه داده این مطالعه از Kaggle بازیابی شده است و شامل ۱۶ ویژگی مستقل و یک متغیر هدف^{۲۶} است که روزهای با نقص زیاد (۱) و روزهای با نقص کم (۰) را در فرآیند تولید نشان می‌دهد. این مجموعه داده، ۳۲۴۰ روز تولیدی را پوشش می‌دهد (Rabie El Kharoua, 2024).

جدول ۲- متغیرها و ویژگی‌های تأثیرگذار بر کیفیت تولید و نقص‌ها

Table 2- Variables and characteristics affecting production quality and defects

متغیر	توضیحات	نوع داده	بازه مقادیر
ProductionVolume	تعداد واحدهای تولیدشده در روز	Integer	۱۰۰ تا ۱۰۰۰ واحد/روز
ProductionCost	هزینه تولید در روز	Float	۵۰۰۰ دلار تا ۲۰۰۰۰ دلار
SupplierQuality	رتبه کیفیت تأمین‌کنندگان	Float (%)	۸۰ درصد تا ۱۰۰ درصد
DeliveryDelay	میانگین تأخیر در تحویل	Integer(روز)	۰ تا ۵ روز
DefectRate	نرخ نقص در هر هزار واحد تولیدی	Float	۰/۵ تا ۵/۰ نقص
QualityScore	ارزیابی کلی کیفیت	Float (%)	۶۰ درصد تا ۱۰۰ درصد
MaintenanceHours	ساعات تعمیر و نگهداری در هفته	Integer	۰ تا ۲۴ ساعت
DowntimePercentage	درصد زمان توقف تولید	Float (%)	۰ درصد تا ۵ درصد
InventoryTurnover	نسبت گردش موجودی	Float	۲ تا ۱۰
StockoutRate	نرخ اتمام موجودی انبار	Float (%)	۰ درصد تا ۱۰ درصد
WorkerProductivity	سطح بهره‌وری نیروی کار	Float (%)	۸۰ درصد تا ۱۰۰ درصد
SafetyIncidents	تعداد حوادث ایمنی در ماه	Integer	۰ تا ۱۰ حادثه
EnergyConsumption	انرژی مصرف‌شده (کیلووات ساعت)	Float	۱۰۰۰ تا ۵۰۰۰ کیلووات ساعت
EnergyEfficiency	ضریب بهره‌وری انرژی	Float	۰/۱ تا ۰/۵
AdditiveProcessTime	زمان تولید افزوده	Float(ساعت)	۱ تا ۱۰ ساعت
AdditiveMaterialCost	هزینه مواد افزودنی در هر واحد	Float (\$)	۱۰۰ دلار تا ۵۰۰ دلار
DefectStatus	وضعیت پیش‌بینی نقص (۰: نقص کم، ۱: نقص زیاد)	Binary	۰ یا ۱

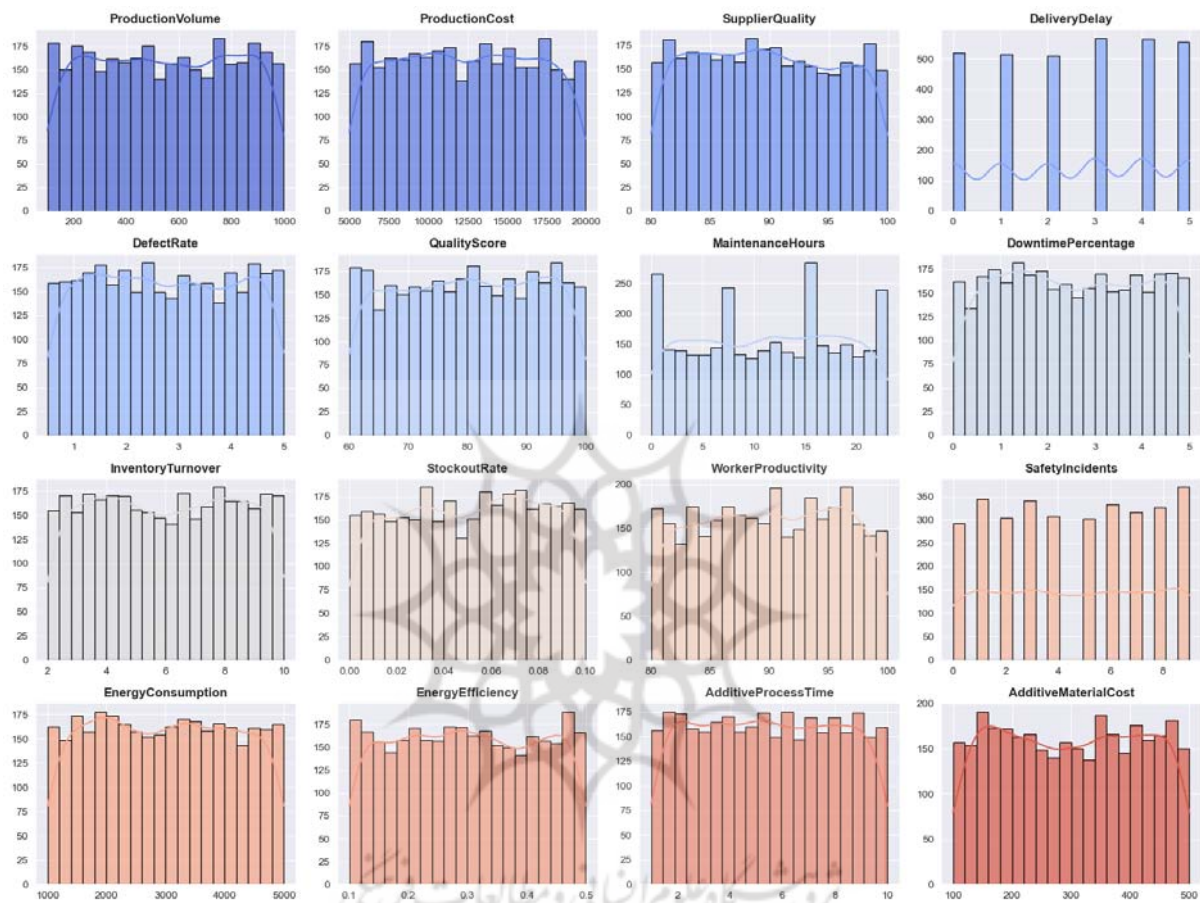
مجموعه داده این مطالعه بر روزهای با نقص زیاد تمرکز دارد؛ زیرا این روزها کمتر رخ می‌دهند؛ اما شناسایی آنها در فرآیند تولید بسیار مهم است. برای بهبود تعادل داده‌ها، روزهای کم نقص نیز به مجموعه اضافه شدند؛ اما همچنان مجموعه داده نامتوازن باقی ماند. این نبود توازن به پیش‌بینی بیش از حد کلاس اکثریت (روزهای با نقص زیاد) توسط مدل‌های یادگیری ماشین منجر می‌شود. براساس شکل ۲، کلاس ۱ (روزهای با نقص زیاد) ۸۴,۰۴٪ از داده‌ها را تشکیل می‌دهد؛ در حالی که کلاس ۰ (روزهای کم نقص) تنها ۱۵,۹۶٪ را شامل می‌شود که نشان‌دهنده توزیع نامتوازن نمونه‌ها در مجموعه داده است.



شکل ۲- توزیع تعداد روزها در متغیر هدف

Fig. 2- Distribution of the number of days in the target variable

در شکل ۳، توزیع ۱۶ ویژگی اصلی مجموعه داده از طریق هیستوگرام‌ها و خطوط چگالی نمایش داده شده است که به تحلیل الگوهای داده کمک می‌کند. برخی متغیرها مانند حجم تولید، هزینه تولید و کیفیت تأمین‌کننده، دارای توزیع یکنواخت‌اند؛ در حالی که متغیرهایی مانند نرخ نقص و امتیاز کیفیت در محدوده خاصی متمرکز شده‌اند. همچنین متغیرهایی مانند تأخیر در تحویل و تعداد حوادث ایمنی گسسته بوده است؛ اما بهره‌وری انرژی و هزینه مواد افزودنی مقادیر پیوسته‌ای دارند که در بازه وسیعی تغییر می‌کند.



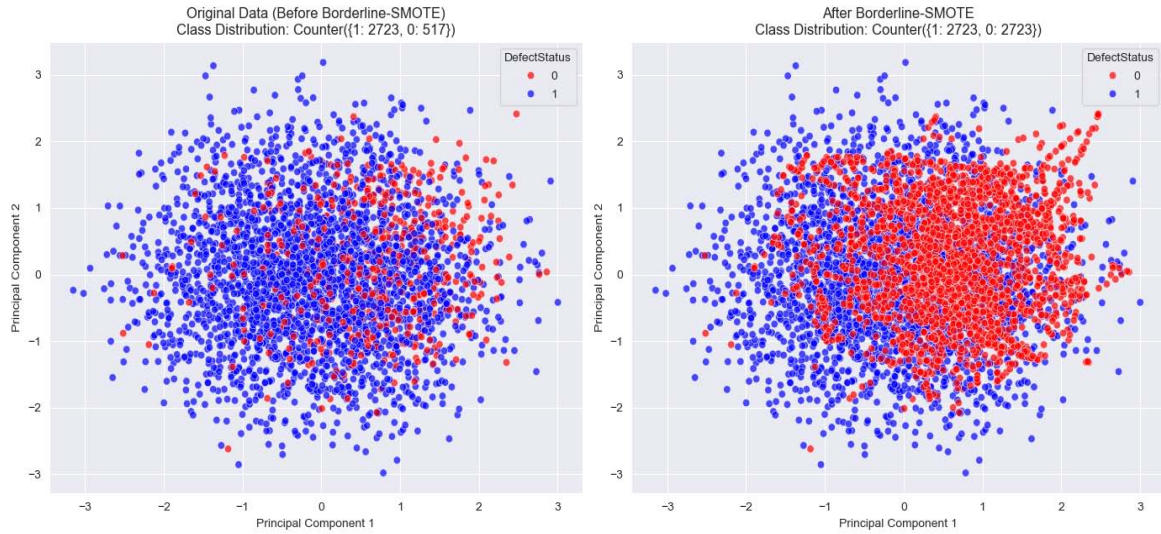
شکل ۳- توزیع داده‌ها بر هیستوگرام‌ها و نمودارهای چگالی

Fig. 3- Data distribution on histograms and density plots

پیش‌پردازش داده‌ها و تحلیل همبستگی

مجموعه داده این مطالعه کیفیت بالایی داشت و فاقد داده‌های گمشده یا نویزی بود؛ اما چالش اصلی، نامتعادل بودن کلاس‌ها بود. برای رفع این مشکل، از Borderline-SMOTE استفاده شد که برخلاف SMOTE معمولی، داده‌های اقلیت نزدیک به مرز، تصمیم‌گیری را نمونه‌برداری می‌کند (Ahsan et al., 2023)؛ (Matharaarachchi et al., 2024). این روش باعث شد مدل با تعادل بهتری میان کلاس‌ها یاد بگیرد و پیش‌بینی روزهای کم‌نقص را بهبود ببخشد. در شکل ۴، توزیع داده‌ها قبل و بعد از اعمال این روش، نشان داده شده است. در نمودار سمت چپ، کلاس کم‌نقص (قرمز) به مراتب کمتر از کلاس نقص زیاد (آبی) است؛ اما پس از اعمال

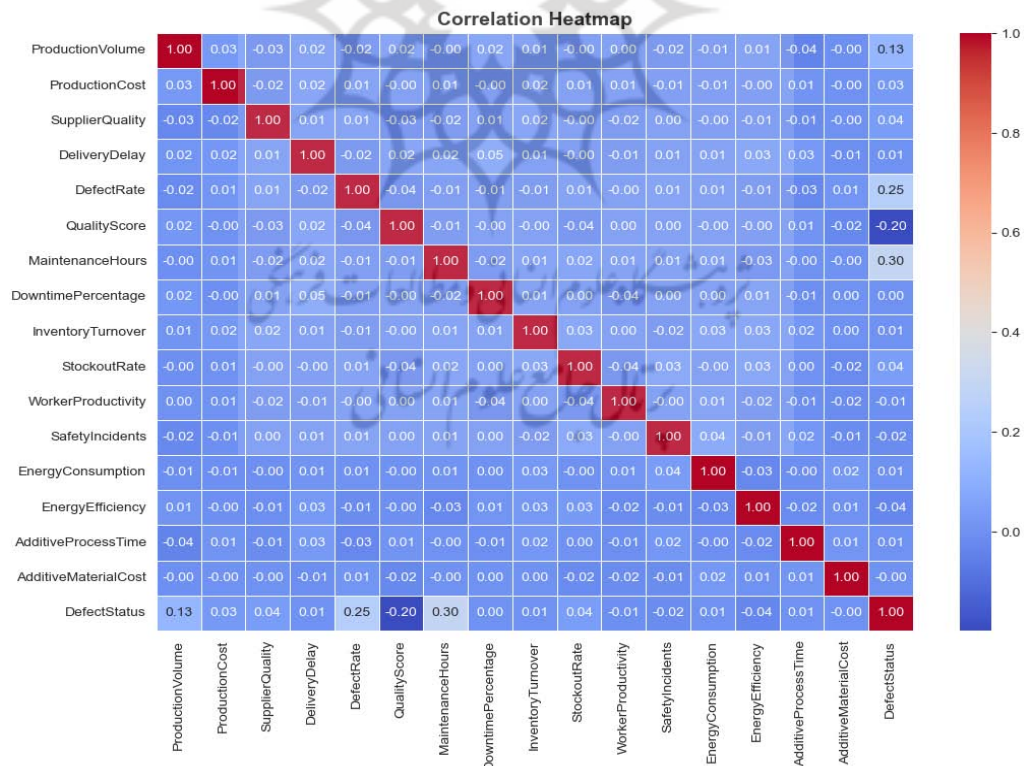
Borderline-SMOTE (نمودار سمت راست)، تعادل برقرار شده است. برای نمایش بهتر، PCA^{۳۷} برای کاهش ابعاد و تجسم داده‌ها در دو بعد استفاده شد.



شکل ۴- وضعیت داده‌ها قبل و بعد از متوازن‌سازی

Fig. 4- Data status before and after balancing

این مطالعه، از تحلیل همبستگی برای بررسی روابط خطی بین متغیرها در مجموعه داده استفاده کرد (Mehregan & Khani, 2024). هدف این تجزیه و تحلیل، تعیین این است که کدام متغیرها تأثیر معنی‌داری بر متغیر هدف دارند و کدام متغیرها ضعیف و زائدند.



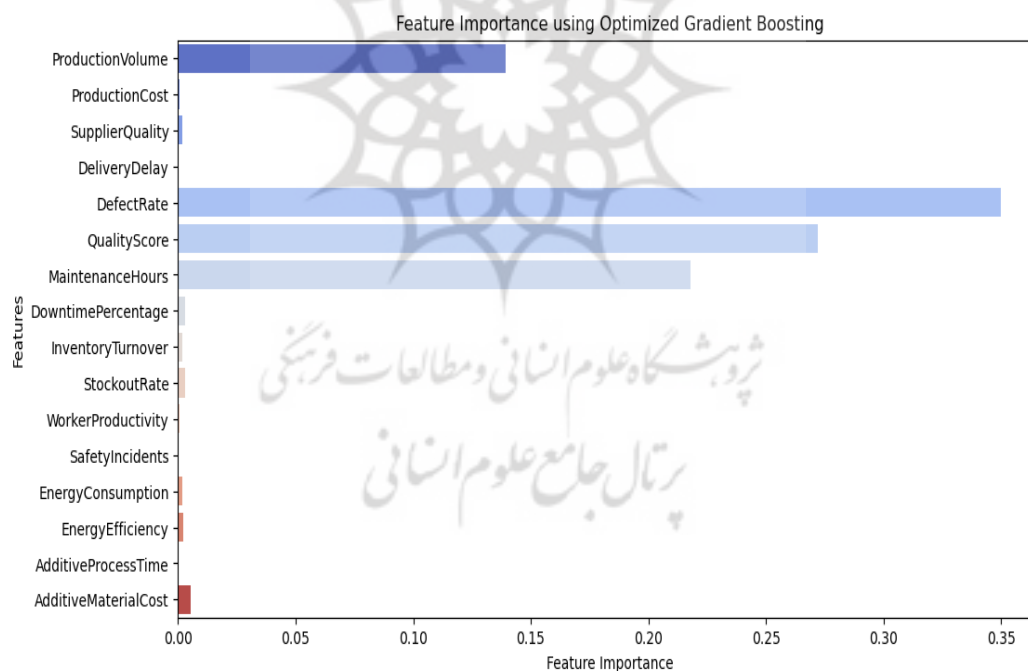
شکل ۵- نقشه حرارتی همبستگی

Fig. 5- Correlation heat map

همبستگی بین ویژگی‌های مجموعه داده در شکل ۵ نشان داده شده است. تمرکز بر این روابط، نشان می‌دهد که هیچ رابطه آماری یا معنی‌داری بین ویژگی‌های مستقل وجود ندارد؛ یعنی هیچ دو ویژگی، اطلاعات مشابه یا تکراری ارائه نمی‌دهند. به عبارت دیگر، این نتیجه به این معنی است که همه ویژگی‌ها در مدل‌سازی به کار می‌روند و لازم نیست هیچ ویژگی حذف شود؛ زیرا همبستگی بالایی دارند (Jafarnejad et al., 2024). برای این مطالعه، ما از StandardScaler برای مقیاس‌بندی داده‌ها استفاده کردیم. هر متغیر در بازه [۰، ۱] با میانگین صفر و انحراف استاندارد واحد نرمال شد. داده‌ها نیز به دو قسمت تقسیم شدند که ۸۰ درصد برای آموزش مدل‌ها و ۲۰ درصد برای ارزیابی عملکرد این مدل‌ها بود.

تحلیل اهمیت و انتخاب ویژگی‌ها

تحلیل اهمیت ویژگی‌ها برای تعیین تأثیر متغیرها بر پیش‌بینی نقص در تولید، انجام شد. در این مطالعه، الگوریتم تقویت گرادیان بهینه‌شده با Optuna برای ارزیابی میزان تأثیر ویژگی‌ها، بر عملکرد مدل به کار رفت (Lu et al., 2024). نتایج نشان داد DefectRate، QualityScore و MaintenanceHours از مهم‌ترین عوامل تأثیرگذار بر نقص‌های تولیدند و بیانگر ارتباط مستقیم عیوب با کیفیت تولید و فرآیند نگهداری‌اند. همچنین، ProductionVolume نیز نقش کلیدی دارد و نشان می‌دهد که تغییرات در مقیاس تولید، احتمال نقص را تحت تأثیر قرار می‌دهد. شکل ۶ اهمیت ویژگی‌ها را با الگوریتم تقویت گرادیان بهینه‌شده با Optuna نشان می‌دهد.



شکل ۶- اهمیت ویژگی‌ها

Fig. 6- Importance of features

در این مطالعه، فرایند انتخاب ویژگی با استفاده از حذف ویژگی بازگشتی، با روش اعتبارسنجی متقاطع (RFECV)، همراه با مدل پایه تقویت گرادیان بهینه‌شده از طریق Optuna اجرا شد. ابتدا با استفاده از الگوریتم Optuna، هایپرپارامترهای مدل پایه برای دستیابی به بهترین عملکرد مدل، تنظیم شدند؛ سپس این مدل پایه با تنظیمات بهینه‌شده درون فرایند RFECV قرار گرفت تا ویژگی‌هایی به صورت بازگشتی انتخاب شوند که بیشترین

تأثیر را بر معیار انتخاب (صحت) دارند. این روش با دقت آن دسته از ویژگی‌هایی را انتخاب می‌کند و از بین می‌برد که کمترین عملکرد را در عملکرد مدل دارند و در نهایت فهرستی از ویژگی‌های اصلی را نگه می‌دارد که تأثیر زیادی در دقت مدل دارند (Adler & Painsky, 2022). اجرای این روش چندین مرحله دارد. اهمیت هر ویژگی براساس روشی اندازه‌گیری می‌شود که مدل اولیه از کل مجموعه داده می‌آموزد. در نهایت در هر مرحله، کم‌اهمیت‌ترین ویژگی را حذف می‌کنیم و به‌صورت بازگشتی ادامه می‌دهیم تا زمانی که حذف بیشتر ویژگی‌ها به عملکرد مدل آسیب برساند. مجموعه‌ای از ویژگی‌های بهینه انتخاب می‌شود که عملکرد را دارد. پس از اجرای این فرآیند، هفت ویژگی کلیدی برای پیش‌بینی نقص‌های تولیدی انتخاب شدند که عبارت‌اند از: حجم تولید^{۲۸}، نرخ نقص^{۲۹}، امتیاز کیفیت^{۳۰}، ساعات تعمیر و نگهداری^{۳۱}، درصد زمان توقف تولید^{۳۲}، نرخ کمبود موجودی^{۳۳} و هزینه مواد افزودنی^{۳۴}. انتخاب این ویژگی‌ها با کاهش پیچیدگی مدل و جلوگیری از بیش‌برازش، دقت پیش‌بینی را افزایش داده است. نتایج نشان می‌دهد که تمرکز بر مدیریت کیفیت، تعمیر و نگهداری و کنترل حجم تولید، نقص‌ها را کاهش می‌دهد و بهره‌وری تولید را بهبود می‌بخشد.

مدل‌سازی

در این مطالعه از روش MetaCost، یک رویکرد یادگیری حساس به هزینه، برای بهینه‌سازی فرآیند پیش‌بینی نقص تولید استفاده می‌شود. یک تکنیک پس‌پردازش، MetaCost، یک مدل یادگیری ماشین معمولی را به یک مدل حساس به هزینه تبدیل می‌کند (Farmasso et al., 2020). ما سعی می‌کنیم پیش‌بینی‌ها را به گونه‌ای تطبیق دهیم که هزینه کل در سیستم به حداقل برسد؛ این کار با در نظر گرفتن یک ماتریس هزینه انجام می‌شود. در مشکل ما، کاهش منفی کاذب (FN) بسیار مهم است؛ زیرا شناسایی نادرست عیوب تولید، برای فرآیند تولید و عرضه یک محصول هزینه بر است؛ در نتیجه، هزینه پیش‌بینی نادرست در کلاس «عیب زیاد» به‌طور درخور توجهی، بالاتر از هزینه دیگر پیش‌بینی‌های نادرست است. هدف روش MetaCost این است که برچسب‌های پیش‌بینی‌شده مدل یادگیری ماشین را با پارامتری که هزینه کل خطاهای آن به حداقل رسیده است و با استفاده از یک ماتریس هزینه ثابت، اصلاح کند. MetaCost یک ماتریس هزینه را تعریف می‌کند:

$$C_{ij} = \begin{bmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{bmatrix} \quad (1)$$

که

C_{ij} نشان‌دهنده هزینه پیش‌بینی کلاس j در حالی که مقدار واقعی کلاس i است.

C_{01} هزینه پیش‌بینی نادرست کلاس ۱ (False Negative - FN)

C_{10} هزینه پیش‌بینی نادرست کلاس ۰ (False Positive - FP)

C_{00} و C_{11} معمولاً صفر در نظر گرفته می‌شوند (زیرا پیش‌بینی درست هزینه‌ای ندارد).

هر مدل یادگیری ماشینی که قابلیت خروجی احتمال کلاس‌ها را داشته باشد، احتمال قرارگیری هر نمونه را در کلاس j محاسبه می‌کند:

$$P(Y = j|X) = \hat{P}_j \quad (2)$$

که در آن:

$P(Y = j|X)$ احتمال تعلق نمونه X به کلاس j است. و \hat{P}_j خروجی مدل برای کلاس j .

پس از آن، هزینه موردانتظار برای هر کلاس محاسبه می‌شود:

$$E(c_0|x) = \hat{P}_0 c_{00} + \hat{P}_1 c_{10} \quad (۳)$$

$$E(c_1|x) = \hat{P}_0 c_{01} + \hat{P}_1 c_{11} \quad (۴)$$

که در آن:

$E(c_0|x)$ هزینه مورد انتظار برای پیش‌بینی کلاس ۰ و $E(c_1|x)$ هزینه مورد انتظار برای پیش‌بینی کلاس ۱.

برچسب نهایی به صورت زیر تعیین می‌شود:

$$\hat{Y} = \arg \min_{j \in \{0,1\}} E(c_j|x) \quad (۵)$$

یعنی کلاسی انتخاب می‌شود که هزینه موردانتظار کمتری داشته باشد.

این مطالعه، مجموعه‌ای از مدل‌های یادگیری ماشین را انتخاب کرد که احتمالات پیش‌بینی را برای اجرای MetaCost ایجاد می‌کنند. اتخاذ این رویکرد ضروری است؛ زیرا MetaCost از ماتریس هزینه برای تعیین بهترین کلاس برای هر نمونه استفاده می‌کند و براساس احتمال تعلق نمونه به هر کلاس، تصمیم می‌گیرد. در این مطالعه، MetaCost هزینه منفی کاذب را ۵ برابر بیشتر در نظر گرفت؛ زیرا از دست دادن روزهای با نقص زیاد، هزینه‌برتر از روزهای کم‌نقص است. علاوه بر این منطق مفهومی، نسبت ۵ به ۱ در جریان آزمایش‌های اولیه پژوهش نیز از لحاظ تجربی بررسی و مشخص شد که در مقایسه با نسبت‌های دیگر، این مقدار بالاترین میزان Recall را برای کلاس روزهای پرنقص ارائه می‌دهد و به افت محسوس در دیگر معیارهای ارزیابی منجر نمی‌شود؛ از این رو این مقدار، به‌عنوان ساختار نهایی ماتریس هزینه انتخاب و در الگوریتم MetaCost پیاده‌سازی شد. برای تنظیم بهینه‌های پارامترها، از Optuna به‌عنوان روش جست‌وجوی هوشمند استفاده و باعث کاهش تنظیمات دستی و افزایش دقت مدل‌ها شد. جدول زیر مدل‌های یادگیری ماشین و هایپرپارامترهای بهینه‌یافته از طریق Optuna را نمایش می‌دهد.

جدول ۳- مدل‌های یادگیری ماشین، توضیحات و هایپرپارامترهای بهینه‌شده از طریق Optuna

Table 3- Machine learning models, descriptions, and hyperparameters optimized by Optuna

مدل یادگیری ماشین	توضیحات	هایپرپارامترهای بهینه‌شده از طریق Optuna
جنگل تصادفی ^{۳۵}	مدلی مبتنی بر مجموعه‌ای از درخت‌های تصمیم که با ترکیب نتایج آنها، دقت و پایداری بالایی را در پیش‌بینی ارائه می‌دهد.	n_estimators=100, max_depth=19, min_samples_split=2, min_samples_leaf=1
گرادیان بوستینگ ^{۳۶}	الگوریتمی مبتنی بر درخت‌های تصمیم که با استفاده از یادگیری مرحله‌ای، دقت مدل را بهبود می‌بخشد.	n_estimators=100, learning_rate=0/2447, max_depth=8
XGBoost	نسخه پیشرفته گرادیان بوستینگ با بهینه‌سازی مصرف حافظه و سرعت پردازش بالا.	n_estimators=300, learning_rate=0/2834, max_depth=10
LightGBM	مدل بوستینگ سبک و سریع که برای پردازش مجموعه داده‌های حجیم، بهینه شده است.	n_estimators=300, learning_rate=0/2849, max_depth=8
CatBoost	مدل بوستینگ کارآمد که مخصوص داده‌های نامتوازن و متغیرهای رسته‌ای ^{۳۷} بهینه شده است.	iterations=300, learning_rate=0/1321, depth=10
ماشین بردار پشتیبان ^{۳۸}	مدلی مبتنی بر یافتن مرزهای جداساز بهینه، برای طبقه‌بندی داده‌های پیچیده و غیرخطی.	C=9/656, kernel=rbf
رگرسیون لجستیک ^{۳۹}	مدلی خطی که احتمال تعلق نمونه به یک کلاس را تخمین می‌زند و به‌عنوان یک مدل پایه استفاده می‌شود.	C=4/767, solver=liblinear

عملکرد مدل‌های یادگیری ماشین با استفاده از چهار معیار کلیدی، ارزیابی می‌شود. این ارزیابی، مفهوم مثبت واقعی (TP) را در نظر می‌گیرد و تعداد نمونه‌هایی است که مدل به درستی آن را معیوب شناسایی می‌کند و منفی واقعی (TN)، تعداد نمونه‌هایی است که به درستی، غیر معیوب طبقه‌بندی می‌شوند؛ در حالی که مثبت کاذب (FP) به تعداد نمونه‌هایی اشاره دارد که به اشتباه با مدل پیش‌بینی شده‌اند و معیوب‌اند؛ در حالی که این‌طور نیستند. هنگامی که یک نقص وجود دارد، اما مدل آن را به اشتباه غیر معیوب تشخیص می‌دهد - این علامت منفی کاذب (FN) است. معیارهای صحت، دقت، بازخوانی و امتیاز F1 براساس این معیارها، برای ارزیابی عملکرد کلی مدل محاسبه می‌شوند.

جدول ۴- شاخص‌های ارزیابی مدل‌های یادگیری ماشین

Table 4- Evaluation indicators for machine learning models

فرمول	تعریف	شاخص
$\frac{TP + TN}{TP + FP + FN + TN}$	نسبت پیش‌بینی‌های صحیح به کل نمونه‌ها.	صحت ^{۴۰}
$\frac{TP}{TP + FP}$	نسبت پیش‌بینی‌های درست برای یک کلاس به کل نمونه‌هایی که در آن کلاس پیش‌بینی شده‌اند.	دقت ^{۴۱}
$\frac{TP}{TP + FN}$	نسبت پیش‌بینی‌های درست برای یک کلاس به کل نمونه‌های واقعی آن کلاس.	بازخوانی یا یادآوری ^{۴۲}
$\frac{2 \times Precision \times Recall}{Precision + Recall}$	میانگین هارمونیک بین Precision و Recall برای ایجاد تعادل بین آنها.	امتیاز F1

این مدل با استفاده از روش اعتبارسنجی متقاطع، با استفاده از 5 بخش برای اعتبارسنجی تعمیم‌پذیری مدل و پایداری مدل از نظر داده‌های آزمایشی مختلف اعتبارسنجی می‌شود (Wong & Yeh, 2020).

یافته‌ها

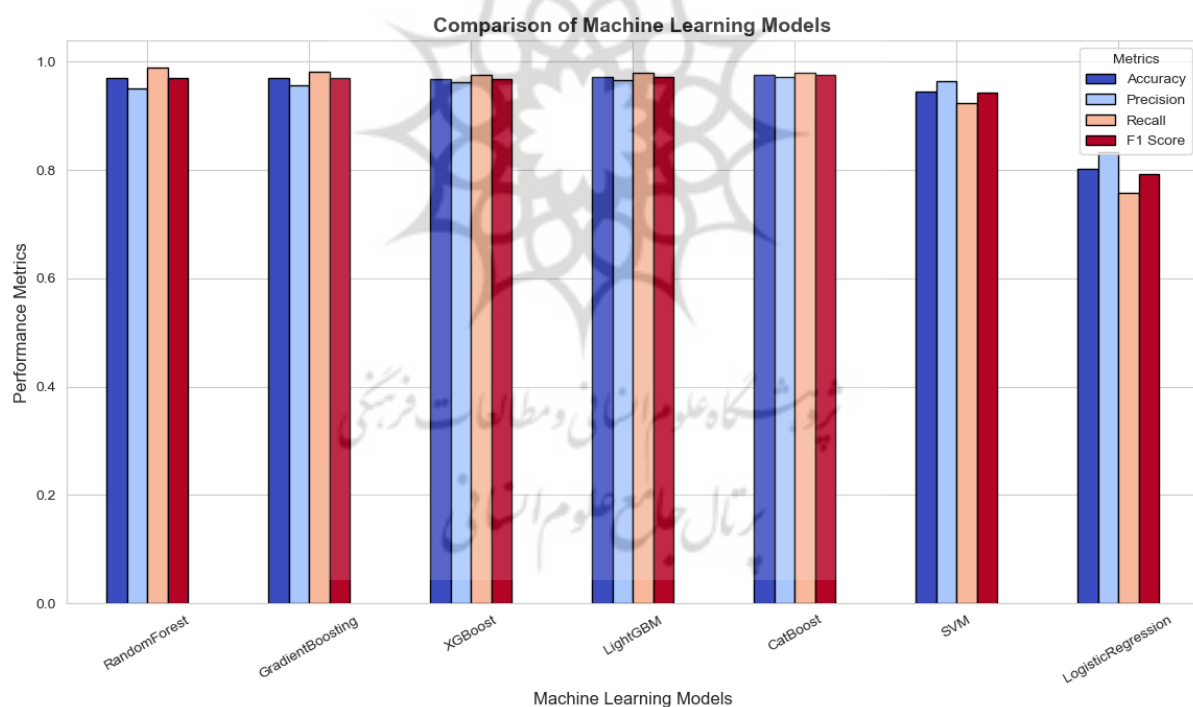
در این بخش، نتایج مدل‌های یادگیری ماشین را بررسی می‌کنیم. این مطالعه با استفاده از پردازنده Intel Core i7-13700H، ۱۶ گیگابایت رم و پایتون ۳،۱۲ اجرا شد. زبان برنامه‌نویسی پایتون، در زیر نتایج عملکرد مدل‌ها آورده شده است.

جدول ۵- مقایسه نتایج مدل‌های یادگیری ماشین

Table 5- Comparison of machine learning model results

مدل	صحت	دقت	بازخوانی	امتیاز F1
Random Forest	۰/۹۶۹۰۰۷	۰/۹۵۰۳۹۸	۰/۹۸۹۹۰۰	۰/۹۶۹۶۹۷
Gradient Boosting	۰/۹۶۹۰۰۸	۰/۹۵۷۵۶۵	۰/۹۸۱۶۳۱	۰/۹۶۹۴۱۴
XGBoost	۰/۹۶۸۰۸۸	۰/۹۶۲۰۸۱	۰/۹۷۴۷۴۲	۰/۹۶۸۳۲۲
LightGBM	۰/۹۷۱۹۹۱	۰/۹۶۴۸۱۰	۰/۹۷۹۷۹۲	۰/۹۷۲۲۰۵
CatBoost	۰/۹۷۴۹۷۸	۰/۹۷۰۹۱۹	۰/۹۷۹۳۴۰	۰/۹۷۵۰۹۴
SVM	۰/۹۴۴۴۴۴	۰/۹۶۴۴۹۱	۰/۹۲۲۸۵۵	۰/۹۴۳۱۹۹
Logistic Regression	۰/۸۰۲۳۴۰	۰/۸۳۲۶۳۱	۰/۷۵۷۰۹۹	۰/۷۹۲۸۹۰

در این مطالعه، مدل جنگل تصادفی به دلیل داشتن بالاترین مقدار بازخوانی (۹۸٫۹۹٪)، بهترین مدل انتخاب شد. بازخوانی بالا نشان می‌دهد که این مدل قادر است تقریباً تمامی روزهای با نقص زیاد را شناسایی کند که در محیط‌های تولیدی، حیاتی است. اگر یک روز پرنقص به اشتباه بدون نقص تشخیص داده شود (منفی کاذب بالا)، اقدامات اصلاحی انجام نمی‌شود و این زیان مالی و کاهش کیفیت تولید را به همراه دارد. با توجه به اینکه هر ردیف از مجموعه داده نمایانگر یک روز تولیدی است، یک پیش‌بینی نادرست به معنای شناسایی نکردن عیوب در یک روز کامل تولید خواهد بود. اگرچه دقت نیز مهم است، در این مطالعه بازخوانی اولویت بالاتری دارد؛ زیرا هزینه مثبت کاذب (خطای برچسب‌گذاری نادرست روزهای کم‌نقص) به مراتب کمتر از هزینه منفی کاذب (شناسایی نکردن روزهای پرنقص) است؛ بنابراین، مدلی که بازخوانی بالاتر دارد، بر مدل‌هایی با دقت بالاتر، ترجیح داده شده است. Random Forest به غیر از داشتن بالاترین مقدار بازخوانی، در دیگر شاخص‌ها نیز در مقایسه با باقی مدل‌ها تعادل خوبی دارد. در نهایت دقت مدل ۹۵/۰۳ درصد، امتیاز $F1$ ۹۶/۹۶ درصد و صحت کلی ۹۶/۹۰ درصد است که نشان می‌دهد این مدل به طور کلی، روزهایی را با ایرادهای زیاد پیش‌بینی می‌کند. مدل‌های دیگری مانند CatBoost و LightGBM نیز عملکرد خوبی داشتند؛ اما در مقایسه با جنگل تصادفی، بازخوانی کمتری داشتند. با این حال، مدل‌هایی مانند SVM و رگرسیون لجستیک عملکرد ضعیفی داشتند؛ زیرا نمی‌توانستند تمام روزهای دارای نقص زیاد را شناسایی کنند و بازخوانی آنها به اندازه کافی بالا نبود که مؤثر باشد.



شکل ۷- مقایسه نتایج مدل‌ها

Fig. 7- Comparison of model results

در شکل ۷، عملکرد ۴ مدل مختلف یادگیری ماشین با استفاده از ۴ معیار ارزیابی (صحت، دقت، بازخوانی و امتیاز $F1$) مقایسه شده است؛ از این رو، می‌بینیم که مدل جنگل تصادفی، بالاترین مقدار بازخوانی را دارد، به این معنی که روزهای با نقص‌های بالا را به خوبی پیش‌بینی می‌کند. در واقع، این معیار برای مشکل بررسی شده ضروری است؛ زیرا این مستلزم شناسایی روزهایی است که عیوب بالایی دارند؛ زیرا این موارد تأثیرات درخور توجهی هم

از نظر شدت و از نظر کنترل کیفیت و هم بر کاهش هزینه‌های تولید دارند. برعکس، مدل‌های CatBoost، LightGBM و XGBoost تقریباً به خوبی Random Forest بودند؛ اما ارزش بازخوانی آنها کمی کمتر از این مدل بود. با این حال، مدل CatBoost دقیق‌تر از Random Forest است؛ اما بازخوانی کمتری دارد، به این معنی که برخی از روزهای با نقص‌های بالاتر را نادیده می‌گیرد. علاوه بر این، مدل تقویت گرادین عملکرد متعادلی در تمام معیارها دارد؛ اما بازخوانی آن کمتر از مدل Random Forest است. به‌طور مشابه، مدل ماشین بردار پشتیبان^{۴۳} عملکرد معقولی در مقایسه با مدل‌های دیگر دارد؛ اما بازخوانی کمتری نسبت به مدل‌های مبتنی بر درخت دارد. نتایج نشان می‌دهد در حالی که SVM به‌طور کامل در شناسایی تمام موارد عیوب بالا شکست نمی‌خورد، عیوب بالا بحرانی می‌شود. برخلاف مدل‌های قبلی، مدل رگرسیون لجستیک، ضعیف‌ترین عملکرد را در مدل‌های مطالعه‌شده ما دارد؛ زیرا مقادیر صحت و بازخوانی آن بسیار کمتر از مدل‌های دیگر بود. این نشان می‌دهد که مدل‌های ساده مانند رگرسیون لجستیک، با مشکلات پیچیده‌ای مانند پیش‌بینی عیوب در ساخت مقابله نمی‌کنند. در نهایت مدل جنگل تصادفی، بهترین مدل برای این مطالعه انتخاب شد؛ زیرا دارای بهترین بازخوانی و تعادل در دیگر معیارهاست. علاوه بر این، این مدل توانسته است دقت، بازخوانی و امتیاز F1 را متعادل کند، روزهای دارای نقص بالا را به‌طور دقیق شناسایی کند و در فرآیند تولید صنعتی مؤثرتر باشد.

بحث

این مطالعه پیش‌بینی نقص‌های تولید را از طریق الگوریتم‌های یادگیری ماشین حساس به هزینه، با تمرکز بر الگوریتم MetaCost بررسی می‌کند. هدف اصلی این پژوهش، پیش‌بینی دقیق‌تر عیوب در فرآیندهای صنعتی، با در نظر گرفتن هزینه‌های مختلف خطاهای طبقه‌بندی (مانند منفی کاذب) بود. استفاده از MetaCost هزینه پیش‌بینی نادرست را کاهش داده و با شناسایی روزهای پرنقص، بازده درخور توجهی در کاهش هزینه و بهبود کیفیت تولید فراهم کرده است. در این مطالعه، از الگوریتم‌های XGBoost، Gradient Boosting، Random Forest، SVM، CatBoost، LightGBM و رگرسیون لجستیک برای پیش‌بینی نقص تولید استفاده شد. چهار معیار صحت^{۴۴}، دقت پیش‌بینی^{۴۵}، بازخوانی^{۴۶} و امتیاز F1 برای ارزیابی عملکرد مدل‌ها به کار گرفته شد. در نهایت، نتایج نشان داد که مدل جنگل تصادفی، بهترین عملکرد را از نظر بازخوانی ارائه می‌دهد و قادر است روزهای معیوب را به‌طور مؤثر، شناسایی کند. این موضوع برای فرآیند تولید بسیار مهم است؛ زیرا از ضررهای مالی و کاهش کیفیت محصول جلوگیری می‌کند.

مقایسه نتایج با مطالعات قبلی نشان می‌دهد که جنگل تصادفی، مشابه برخی تحقیقات مانند چن (۲۰۲۱) و مائلکویست و همکاران (۲۰۲۳)، تعادلی را میان دقت و بازخوانی ایجاد می‌کند. همچنین، مطالعات پیشین استفاده از تکنیک‌های حساس به هزینه، مانند MetaCost را در مدل‌هایی مانند XGBoost و LightGBM گزارش کرده‌اند که به بهبود عملکرد این مدل‌ها در مواجهه با داده‌های نامتعادل منجر شده است. در این مطالعه نیز، هر دو مدل CatBoost و LightGBM عملکرد مطلوبی داشتند؛ اما جنگل تصادفی از نظر کاهش منفی کاذب، عملکرد بهتری ارائه داد. این ویژگی برای مدیریت کیفیت و جلوگیری از خطاهای پرهزینه در فرآیند تولید، حیاتی است. علاوه بر

این، SVM و رگرسیون لجستیک در مقایسه با دیگر مدل‌ها ضعیف‌تر عمل کردند. نتایج نشان داد که این مدل‌ها، نقص‌های پیچیده تولید را پیش‌بینی نمی‌کنند. این یافته‌ها در تضاد با برخی مطالعات قبلی است که گزارش کرده‌اند SVM در برخی شرایط عملکرد بهتری دارد. با این حال، در این تحقیق مشخص شد که SVM برخی از روزهای بحرانی را به‌طور مؤثر متمایز نکرده و در شناسایی آنها ناکام مانده است.

نتیجه‌گیری

در این پژوهش مشخص شد که الگوریتم جنگل تصادفی از نظر دقت و یادآوری، بهترین الگوریتم برای پیش‌بینی عیوب ساخت در مقایسه با الگوریتم‌های دیگر است. با این الگوریتم، روزهای غنی به‌طور مؤثر پیش‌بینی و از مشکلات کیفی و زیان‌های مالی جلوگیری می‌شود. مطالعه حاضر همچنین نشان می‌دهد که MetaCost تأثیر زیادی بر به حداقل رساندن هزینه‌های خطا در فرآیند تولید دارد؛ به‌خصوص زمانی که این مدل‌ها در بهینه‌سازی استفاده می‌شوند. تحقیقات آینده نیز، کاربرد MetaCost را در دیگر حوزه‌های صنایع مختلف از جمله خودرو، الکترونیک، و داروسازی و غیره مطالعه می‌کنند. شبکه‌های عصبی کانولوشن (CNN^{۴۷}) یا شبکه‌های عصبی تکراری (RNN^{۴۸})، دقت پیش‌بینی را برای داده‌های پیچیده‌تر افزایش می‌دهند.

علاوه بر این، پیش‌بینی نقص یا تحلیل هزینه‌های پیش‌بینی تحت شرایط پویاتر و نامطمئن‌تر، با آزمایش مدل‌ها بهبود می‌یابد. یکی دیگر از حوزه‌های پیشنهادی برای تحقیقات آینده، بهره‌گیری از داده‌های حسگر و اینترنت اشیا^{۴۹} است؛ زیرا این فناوری‌ها با کاهش خطاهای پیش‌بینی و هزینه‌های تولید، به تصمیم‌گیری دقیق‌تر و کارا تر کمک می‌کنند. علاوه بر این، برخی از مدل‌های یادگیری ماشین حساس به هزینه برای پیش‌بینی عیوب، تحت داده‌های غیرثابت مانند متن، تصویر یا داده‌های صوتی به کار می‌روند. این مدل‌ها همچنین در سیستم‌های تولید واقعی آینده آزمایش می‌شوند و خروجی آنها با سیستم‌های سنتی مقایسه می‌شود. برای روشن کردن مزایای روش‌های نوین یادگیری ماشین در تصمیم‌گیری‌های عملیاتی، این روش‌ها با تکنیک‌های سنتی مقایسه شدند. در کل طیف، استفاده از یادگیری ماشین برای تمایز بین عیوب تولید، به‌ویژه در داده‌های پیچیده و نامتعادل چالش‌برانگیز، حوزه تشویق‌کننده و گسترده‌ای برای تحقیقات است که تأثیر مثبتی بر فرآیندهای تولید مرتبط و تصمیم‌گیری دارد. با توجه به اینکه مدل جنگل تصادفی با استفاده از الگوریتم MetaCost به‌گونه‌ای آموزش داده شده است که هزینه‌های ناشی از عدم شناسایی نقص‌ها (منفی کاذب) را کاهش دهد، پیاده‌سازی آن در خطوط تولید، به کاهش دوباره‌کاری، جلوگیری از توقف‌های غیرضروری و صرفه‌جویی در هزینه‌های ناشی از تولید معیوب منجر می‌شود. همچنین با بهینه‌سازی مدل و انتخاب ویژگی‌های کلیدی، امکان پیاده‌سازی آن بر داده‌های بلادرنگ^{۵۰} وجود دارد؛ به‌طوری که با دریافت ورودی‌های روزانه از حسگرها یا سامانه‌های تولید، هشدارهای پیشگیرانه ارائه دهد و تصمیمات مدیریتی به‌موقع اتخاذ شود.

References

- Adler, A. I., & Painsky, A. (2022). Feature Importance in Gradient Boosting Trees with Cross-Validation Feature Selection. *Entropy*, 24(5), 687. <https://doi.org/10.3390/e24050687>
- Ahsan, M. M., Raman, S., & Siddique, Z. (2023). *BSGAN: A Novel Oversampling Technique for Imbalanced Pattern Recognitions*. (Cornell University). PsyArXiv. <https://doi.org/10.48550/arxiv.2305.09777>

- Ataei, S., Adibnazari, S., & Ataei, S. T. (2025). *Data-driven Detection and Evaluation of Damages in Concrete Structures: Using Deep Learning and Computer Vision*. (Cornell University). PsyArXiv. <https://doi.org/10.48550/arxiv.2501.11836>
- Barzizza, E., Biasetton, N., Ceccato, R., & Molena, A. (2024). Machine learning-based decision-making approach for predicting defects detection: a case study. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 13(3), 3052. <https://doi.org/10.11591/ijai.v13.i3.pp3052-3060>
- Chen, Y. (2021). Research on Cost-sensitive Classification Methods for Imbalanced Data. *International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)*, Xi'an, China, 224-228. <https://doi.org/10.1109/caibda53561.2021.00054>
- Frumosu, F. D., Khan, A. R., Schiöler, H., Kulahci, M., Zaki, M., & Westermann-Rasmussen, P. (2020). Cost-sensitive learning classification strategy for predicting product failures. *Expert Systems with Applications*, 161, 113653. <https://doi.org/10.1016/j.eswa.2020.113653>
- Ghatasheh, N., Faris, H., AlTaharwa, I., Harb, Y., & Harb, A. (2020). Business Analytics in Telemarketing: Cost-Sensitive Analysis of Bank Campaigns Using Artificial Neural Networks. *Applied Sciences*, 10(7), 2581. <https://doi.org/10.3390/app10072581>
- Hassan, D. (2017). The Impact of False Negative Cost on the Performance of Cost Sensitive Learning Based on Bayes Minimum Risk: A Case Study in Detecting Fraudulent Transactions. *International Journal of Intelligent Systems and Applications*, 9(2), 18–24. <https://doi.org/10.5815/ijisa.2017.02.03>
- Jafarnejad Chaghoshi, A. , Rezasoltani, A. and Khani, A. M. (2024). Unleashing the Power of Ensemble Learning: Predicting National Ranks in Iran's University Entrance Examination. *Industrial Management Journal*, 16(3), 457-481. DOI: [10.22059/imj.2024.381521.1008178](https://doi.org/10.22059/imj.2024.381521.1008178)
- Kamalaruban, P., & Williamson, R. C. (2018). *Minimax Lower Bounds for Cost Sensitive Classification*. ArXiv (Cornell University). <https://doi.org/10.48550/arxiv.1805.07723>
- Kang, Z., Catal, C., & Tekinerdogan, B. (2020). Machine learning applications in production lines: A systematic literature review. *Computers & Industrial Engineering*, 149, 106773. <https://doi.org/10.1016/j.cie.2020.106773>
- Kim, Y. J., Baik, B., & Cho, S. (2016). Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Systems with Applications*, 62, 32–43. <https://doi.org/10.1016/j.eswa.2016.06.016>
- Le, T., Vo, M. T., Vo, B., Lee, M. Y., & Baik, S. W. (2019). A Hybrid Approach Using Oversampling Technique and Cost-Sensitive Learning for Bankruptcy Prediction. *Complexity*, 2019, 1–12. <https://doi.org/10.1155/2019/8460934>
- Liu, Y., Li, Q., Wang, K., Liu, J., He, R., Yuan, Y., & Zhang, H. (2021). Automatic Multi-Label ECG Classification with Category Imbalance and Cost-Sensitive Thresholding. *Biosensors*, 11(11), 453. <https://doi.org/10.3390/bios11110453>
- Mählkvist, S., Ejenstam, J., & Kyprianidis, K. (2023). Cost-Sensitive Decision Support for Industrial Batch Processes. *Sensors*, 23(23), 9464. <https://doi.org/10.3390/s23239464>
- Malhotra, R., & Kamal, S. (2019). An empirical study to investigate oversampling methods for improving software defect prediction using imbalanced data. *Neurocomputing*, 343, 120–140. <https://doi.org/10.1016/j.neucom.2018.04.090>
- Mansouri, T., Sadeghimoghadam, M., & Ghasemian Sahebi, I. (2021). *A New Algorithm for Hidden Markov Models Learning Problem*. PsyArXiv. <https://doi.org/10.48550/arXiv.2102.07112>
- Matharaarachchi, S., Domaratzki, M., & Muthukumarana, S. (2024). Enhancing SMOTE for imbalanced data with abnormal minority instances. *Machine Learning with Applications*, 18, 100597. <https://doi.org/10.1016/j.mlwa.2024.100597>
- Mehregan, M. R. and Khani, A. M. (2024). Improving organizational performance: the role of supply chain 4.0 and financing in reducing supply chain risk. *Journal of International Business Administration*, 7(3), 39-59. DOI: [10.22034/jiba.2024.60005.2164](https://doi.org/10.22034/jiba.2024.60005.2164)

- Menold, H. S., Wieland, V. L. S., Haney, C. M., D. Uysal, Wessels, F., G.C. Cacciamani, Michel, M. S., Seide, S., & Kowalewski, K. F. (2024). Machine learning enables automated screening for systematic reviews and meta-analysis in urology. *World Journal of Urology*, 42(1), 396. <https://doi.org/10.1007/s00345-024-05078-y>
- Mirzaei, S. , Ashtab, A. and Zavari Rezaei, A. (2023). Comparing the Efficiency of Statistical Models and Machine-Learning Models and Choosing the Optimal Model for Predicting Net Profit and Operating Cash Flows. *Journal of Asset Management and Financing*, 11(2), 53-74. DOI: [10.22108/amf.2023.136720.1784](https://doi.org/10.22108/amf.2023.136720.1784)
- Motamedi, M., Shidpour, R. & Ezoji, M. (2024). LSTM-based framework for predicting point defect percentage in semiconductor materials using simulated XRD patterns. *Sci Rep* 14, 24353. <https://doi.org/10.1038/s41598-024-75783-6>
- Niu, L., Wan, J., Wang, H., & Zhou, K. (2020). Cost-sensitive Dictionary Learning for Software Defect Prediction. *Neural Processing Letters*, 52(3), 2415–2449. <https://doi.org/10.1007/s11063-020-10355-z>
- Rabie El Kharoua. (2024). *Predicting Manufacturing Defects Dataset* [Data set]. Kaggle. <https://doi.org/10.34740/KAGGLE/DSV/8715500>
- Setty, R., Yuval Elovici, & Schwartz, D. (2024). Cost-sensitive machine learning to support startup investment decisions. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 31(1),1-17. <https://doi.org/10.1002/isaf.1548>
- Shi, K., Shi, R., Fu, T., Lu, Z., & Zhang, J. (2024). A Novel Identification Approach Using RFECV–Optuna–XGBoost for Assessing Surrounding Rock Grade of Tunnel Boring Machine Based on Tunneling Parameters. *Applied Sciences*, 14(6), 2347–2347. <https://doi.org/10.3390/app14062347>
- Soltani, M., Khatami Firouzabadi, S. M. A., Amiri, M. and Hajian Heidary, M. (2023). Proposing an integrated approach for omnichannel demand forecasting using machine learning-time series clustering with dynamic time warping algorithm and artificial neural networks. *Research in Production and Operations Management*, 14(1), 121-140. DOI: [10.22108/pom.2023.136202.1485](https://doi.org/10.22108/pom.2023.136202.1485)
- van Vuuren, J. H. (2024). A MACHINE LEARNING FRAMEWORK FOR DATA-DRIVEN DEFECT DETECTION IN MULTISTAGE MANUFACTURING SYSTEMS. *South African Journal of Industrial Engineering*, 35(2), 154-170. <https://doi.org/10.7166/35-2-3008>
- Verbeke, W., Olaya, D., Berrevoets, J., Verboven, S., & Maldonado, S. (2020). *The foundations of cost-sensitive causal classification*. ArXiv (Cornell University). <https://doi.org/10.48550/arxiv.2007.12582>
- Wong, T., & Yeh, P. (2020). Reliable Accuracy Estimates from k-Fold Cross Validation. *IEEE Transactions on Knowledge and Data Engineering*, 32(8), 1586–1594. <https://doi.org/10.1109/TKDE.2019.2912815>
- Zhou, Z.-H., & Liu, X.-Y. (2010). ON MULTI-CLASS COST-SENSITIVE LEARNING. *Computational Intelligence*, 26(3), 232–257. <https://doi.org/10.1111/j.1467-8640.2010.00358.x>

¹ machine learning

² cost sensitive machine learning

³ Zhou & Liu

⁴ Liu et al.

⁵ Menold et al.

⁶ Setty et al.

⁷ Mählkvist et al.

⁸ Chen

⁹ Niu et al.

¹⁰ Ghatasheh et al.

¹¹ Verbeke et al.

¹² Malhotra and Kamal

¹³ Le et al.

- 14 Kamalaruban and Williamson
- 15 Kim et al.
- 16 Soltani et al.
- 17 Mirzaei et al.
- 18 Cost-Sensitive Learning
batch
- 20 Recursive Feature Elimination with Cross-Validation
- 21 Recursive Feature Elimination with Cross-Validation
- 22 FalseNegative
- 23 Rabie El Kharoua
- 24 DefectStatus
- 25 Support Vector Machine
- 26 DefectStatus
- 27 Principal Component Analysis
- 28 ProductionVolume
- 29 DefectRate
- 30 QualityScore
- 31 MaintenanceHours
- 32 DowntimePercentage
- 33 StockoutRate
- 34 AdditiveMaterialCost
- 35 Random Forest
- 36 Gradient Boosting
- 37 Categorical
- 38 SVM
- 39 Logistic Regression
- 40 Accuracy
- 41 Precision
- 42 Recall
- 43 SVM
- 44 Accuracy
- 45 Precision
- 46 Recall
- 47 Convolutional Neural Network
- 48 Recurrent Neural Network
- 49 IoT
- 50 real-time

