Journal of Statistical Sciences, Spring and Summer, 2025
Vol. 19, No. 1, pp 161-182
DOI: 10.61186/jss.19.1.9

Advanced Missing Value Imputation Techniques: Machine Learning Methods with an Emphasis on an Ensemble Method for Multiple Imputation by Chained Equations

Ghaderi, M. [1], Rezaei Ghahroodi, Z. [1], Gandomi M.[2]
[1]School of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, Iran.
[2]Statistical Center of Iran, Tehran, Iran.

Corresponding author: Z. Rezaei Ghahroodi, z.rezaeigh@ut.ac.ir

Received: 13/10/2024 **Revised:** 14/3/2025 **Accepted and Published Online:** 17/3/2025.

Introduction
Missing data is a well-known problem common in all organizations that gather data on persons or enterprises. It causes issues like performance degradation, analysis challenges, issues with data processing, and biased results due to the disparities between missing and complete values.

Material and Methods
Traditional imputation techniques, including mean, regression, K-nearest neighbor, and ensemble methods, have been proposed to handle missing values. Multiple imputation by chained equations (MICE) is one of the most common methods for imputation. In theory, any imputation model can be used to predict the missing values. However, if the predictive models are incorrect, it can lead to biased estimates and invalid inferences. One of the latest solutions for dealing with missing data is machine learning methods and the SuperMICE method, which is a combination of the two methods of SuperLearner and MICE.

Results and Discussion
In the random mechanism, when missingness is 10%, 20%, and 30%, the performance of PMM, SuperMICE, and BLR is similar. However, at higher missingness levels (40% and 50%), SuperMICE performs better, with a significant difference compared to the other two methods. In real data with a

random missingness mechanism, SuperMICE outperforms the other methods significantly. In cases of non-random missingness, SuperMICE exceeds PMM, BLR, hotdeck, and irmi methods, with performance similar to missRanger and missForest. However, KNN and CART methods occasionally outperform SuperMICE, while in other cases, they perform worse.

Conclusion
In conclusion, the SuperMICE method shows strong performance in the MAR mechanism but struggles in the MCAR mechanism, as evidenced by bias and coverage percentages. Using R software, this paper presents simulations demonstrating that SuperMICE yields parameter estimates with lower bias and better coverage compared to other common imputation methods. Additionally, the paper discusses the application of SuperMICE, machine learning methods, and an ensemble algorithm on industrial establishment survey data, which faces unit non-response. The simultaneous imputation of various variables in this dataset is explored, and the evaluation of different methods reveals that SuperMICE outperforms previous approaches.

Keywords: Missing data, Super learner ensemble algorithm, Multiple imputation by chained equations, Machine learning methods, Industrial establishment survey.

Mathematics Subject Classification (2010): 97K80, 47N30

مجله علوم آماری، بهار و تابستان ۱۴۰۴
جلد ۱۹، شماره ۱، ص ۱۶۱ -- ۱۸۲
DOI: 10.61186/jss.19.1.9
مقاله پژوهشی

روش‌های پیشرفته جانهی مقادیر گم‌شده: روش‌های یادگیری ماشین با تاکید بر یک روش ترکیبی برای جانهی چندگانه با استفاده از معادله‌های زنجیره‌ای

مهرداد قادری[۱]، زهرا رضائی قهرودی[۱]، مینا گندمی[۲]

[۱]دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه تهران
[۲]مرکز آمار ایران

نویسنده مسئول : زهرا رضائی قهرودی، z.rezaeigh@ut.ac.ir

تاریخ دریافت: ۱۴۰۳/۷/۲۲ تاریخ بازنگری: ۱۴۰۳/۱۲/۲۴ تاریخ پذیرش و انتشار: ۱۴۰۳/۱۲/۲۷

چکیده: نحوه برخورد با داده‌های گم‌شده یکی از مسائلی است که اغلب محققان با آن روبرو هستند. جانهی چندگانه با استفاده از معادله‌های زنجیره‌ای یکی از رایج‌ترین و انعطاف‌پذیرترین روش‌ها برای جانهی است. از دیدگاه تئوری، هر مدل جانهی می‌تواند برای پیش‌بینی مقادیر داده‌های گم‌شده استفاده شود اما اگر مدل‌های پیشگویی نادرست باشند می‌تواند منجر به برآوردهای اریب و استنباط‌های نامعتبر شود. یکی از جدیدترین راه‌حل‌ها برای برخورد با داده‌های گم‌شده، روش ترکیبی یادگیری ماشین و ابریادگیرنده است. در این مقاله، چند شبیه‌سازی برای نشان دادن رویکرد بهتر این روش از نظر اریبی کمتر و همگرایی بهتر برآورد پارامتر نهایی نسبت به روش‌های جانهی رایج ارائه شده است. همچنین، به پیاده‌سازی برخی روش‌های یادگیری ماشین و یک الگوریتم ترکیبی از ابریادگیرنده، روی داده‌های کارگاه‌های صنعتی پرداخته شده است که در آن جانهی متغیرهای مختلف در داده‌ها به‌طور همزمان صورت می‌گیرد. همچنین به ارزیابی روش‌های مختلف و معرفی روش دارای عملکرد برتر، پرداخته شده است.

واژه‌های کلیدی: داده‌های گم‌شده، الگوریتم ترکیبی ابریادگیرنده، جانهی چندگانه با استفاده از معادلات زنجیره‌ای، روش‌های یادگیری ماشین، آمارگیری از کارگاه‌های صنعتی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 97K80 ، 47N30.

۱ مقدمه

مسئله داده‌های گم‌شده یکی از مشکلات فراگیر در تحقیقات رشته‌های مختلف است. توجه به مسئله داده‌های گم‌شده از زمان‌های گذشته مد نظر دانشمندان حوزه علم آمار و سایر علوم که با آمار و داده‌ها سروکار دارند، بوده است. با توجه به اهمیت موضوع گم‌شدگی و رخداد این پدیده در اکثر داده‌های آمارهای رسمی کشور، برخورد مناسب با داده‌های گم‌شده بسیار مهم است؛ زیرا برخورد نامناسب با داده‌های گم‌شده باعث ایجاد خطاهای بزرگ یا نتایج نادرست می‌شود. در خصوص مبحث مواجهه با داده‌های گم‌شده، در ادبیات آماری رویکردهای متفاوتی وجود دارد. یکی از رویکردهای رایج که در این مقاله به آن پرداخته می‌شود، رویکرد جانهی است. جانهی، پیش‌بینی مقادیر گم‌شده بر اساس داده‌های مشاهده شده است. برای جانهی داده‌های گم‌شده، در ابتدا از روش‌های جانهی تکی استفاده می‌شد و برای هر مقدار گم‌شده یک مقدار جانهی برآورد می‌گردید. اما با گذشت زمان، با توجه به این‌که در روش‌های جانهی تکی، عدم قطعیت در نظر گرفته نمی‌شد؛ روش‌های جانهی چندگانه، معرفی گردید که خطای استاندارد دقیق‌تری نسبت به جانهی تکی دارد.

از طرف دیگر، اکثر روش‌های جانهی موجود تنها به یک نوع متغیر پیوسته یا رسته‌ای محدود می‌شوند و یا به طور مجزا پردازش و جانهی می‌شود و روابط ممکن بین انواع متغیرها نادیده گرفته می‌شوند. بنابراین، استفاده از روش‌های پیشرفته و هوشمند جانهی که بتواند روابط ممکن بین انواع متغیرها را در نظر بگیرد و محدودیتی در نوع متغیرها نداشته باشد، می‌تواند منجر به عملکرد بهتری شود. با گسترش روش‌های یادگیری ماشین و با توجه به حجم بالای اطلاعات موجود در تحقیقات آماری، در سال‌های اخر، بحث استفاده از روش‌های یادگیری ماشین در مباحث جانهی داده‌های گم‌شده مطرح گردید. برخی از روش‌های نوین، روش‌های یادگیری ماشین و روش‌های ترکیبی براساس جانهی داده‌های آمیخته (انواع مختلف متغیرهای پیوسته و رسته‌ای) با در نظر گرفتن ساختار همبستگی بین متغیرها به طور همزمان است که عملکرد بهتری نسبت به روش‌های قبلی دارند. برخی از مهم‌ترین روش‌های جانهی مورد استفاده در این مقاله، روش جانهی چندگانه با معادله‌های زنجیره‌ای[۱] $(MICE)$، روش $SuperMICE$، الگوریتم K - نزدیک‌ترین همسایه[۲] (KNN)، درخت تصمیم، درخت تصمیم رگرسیون و رده‌بندی[۳] $(CART)$، $missForest$، $missRanger$ است (راقاناتان و همکاران، ۲۰۰۱؛ اِستِخاون و بالمَن، ۲۰۱۲). الگوریتم‌های منعطف و پیچیده مانند جنگل تصادفی، تمایل به تولید واریانس بالاتر و اریبی کم‌تر دارند ولی مستعد بیش‌برازشی داده‌ها است؛ این در حالی است که الگوریتم‌های ساده‌تر با ساختارهای غیر منعطف، تمایل به تولید واریانس کمتر و مستعد کم‌برازشی هستند. از آنجا که که فرم تابعی مدل‌های جانهی کمتر شناخته شده‌اند، محقق می‌بایست یک تصمیم بهینه بگیرد تا تعادل بین واریانس و اریبی در فرایند برآورد برقرار شود. برای افزایش دقت و کارایی الگوریتم‌ها روش‌های ترکیبی معرفی شدند. وَندِرلین و همکاران (۲۰۰۷) روش ابریادگیرنده که یک روش یادگیری ترکیبی مبتنی بر تابع زیان است را معرفی کردند. این روش امکان ترکیب کلاسی از الگوریتم‌ها را برای پیش‌بینی فراهم می‌کند.

[1] Multivariate Imputation by Chained Equations
[2] K nearest neighbors
[3] Classfication and regression tree

نتایج نشان داد مجموع وزنی الگوریتم‌ها به صورت مجانبی به خوبی و یا حتی بهتر از الگوریتم‌های تکی عمل می‌کنند. یکی از جدیدترین راه‌حل‌های معرفی شده در زمینه برخورد با داده‌های گم‌شده، الگوریتم $SuperMICE$ است که توسط لاکویر و همکاران (۲۰۲۲) معرفی شده است. این روش در واقع تلفیقی از دو روش ابریادگیرنده[1] و $MICE$ است. آلواتر و همکاران (۲۰۲۴) به بررسی جامع روش‌های مختلف جانهی پرداخته‌اند و آن‌ها را به سه رویکرد اصلی روش‌های قطعیتی، مدل‌های احتمالی، و الگوریتم‌های یادگیری ماشین دسته‌بندی کرده‌اند. همچنین با توجه به اهمیت برخورد مناسب با داده‌های گم‌شده در داده‌های پزشکی برای دست‌یابی به پیش‌بینی‌های پزشکی دقیق، تواسکار و همکاران (۲۰۲۴) به بررسی جامع اثرات ناشی از تکنیک‌های جانهی مبتنی بر یادگیری ماشین به‌ویژه KNN، $MICE$ و $missForest$ پرداخته است. در ادامه به معرفی برخی الگوریتم‌های نوین یادگیری ماشین برای جانهی داده‌های گم‌شده پرداخته می‌شود و سپس با شبیه‌سازی و مثال واقعی، الگوریتم‌های مختلف مقایسه و بهترین روش پیشنهاد می‌شود.

۲ روش‌های جانهی مبتنی بر یادگیری ماشین

مراکز آماری، داده‌های مورد نیاز خود را از منابع مختلف مانند آمارگیری‌های نمونه‌ای، سرشماری‌ها، آمارهای ثبتی و داده‌های استخراج شده از وب‌سایت‌ها گردآوری می‌کنند. از آن‌جا که داده‌های گردآوری شده با بی‌پاسخی روبرو است، اعمال روش‌های وزن‌دهی و جانهی روی این داده‌ها، ضروری است. این موضوع یک مرحله بسیار مهم از فرایندها در مراکز آماری به‌منظور اطمینان از کیفیت استنباط‌های نهایی و جداول استخراجی در مراکز آماری است. این فرآیند حیاتی در سازمان‌های آماری با روش‌های مختلفی انجام می‌شود. بهبود روش‌های جانهی و دانش زمینه‌ای متخصصان موضوع، یکی از مولفه‌های مهم و ضروری در فرآیند جانهی است. با توجه به پیشرفت‌های صورت‌گرفته در زمینه یادگیری ماشین و خودکارسازی فرایندهای مراکز آماری، استفاده از روش‌های یادگیری ماشین برای جانهی داده‌های بی‌پاسخ و ناکامل ضروری است (کوپینلن، ۱۹۸۷). از مزایای استفاده از روش‌های یادگیری ماشین، امکان بررسی حجم زیادی از اطلاعات از طریق خودکارسازی فرایندها و همچنین انجام محاسبات دقیق‌تر و سریع‌تر است. در این بخش برخی از روش‌های جانهی که مبتنی بر روش‌های یادگیری ماشین هستند، معرفی می‌گردند. الگوریتم‌های یادگیری ماشین برای جانهی مقادیر گم‌شده در ابتدا بر روی جانهی‌های تکی تمرکز داشته است (وَن‌بردن، ۲۰۱۸). سپس تلاش‌هایی برای استفاده از این روش‌ها در جانهی چندگانه صورت گرفت. به‌عنوان مثال، الگوریتم درخت تصمیم در جانهی چندگانه استفاده شد. هدف این الگوریتم به عنوان الگوریتم پرطرف‌دار یادگیری ماشین، مدل‌بندی بهتر روابط غیرخطی است. به‌عنوان مثال دیگر از کاربرد روش‌های یادگیری ماشین در جانهی چندگانه، می‌توان به الگوریتم $CART$ و جنگل تصادفی اشاره کرد (وَن‌بردن و گروتوز–کودشورن، ۲۰۱۱)، که در بسته $MICE$ در نرم‌افزار R قابل دسترس است.

[1] Super learner

۲.۱ جانهی چندگانه با معادله‌های زنجیره‌ای

جانهی چندگانه با استفاده از $MICE$ یکی از رایج‌ترین روش‌های جانهی است. در $MICE$ یک مدل پیشگو برای هر متغیر دارای مقادیر گم‌شده به شرط سایر متغیرهای موجود در مجموعه داده به صورت تکراری برازش داده می‌شود. الگوریتم $MICE$ به مشخص‌سازی مدل برای هر متغیر دارای مقادیر گم‌شده و همچنین به انتخاب پیشگوها که باید در جانهی متغیرها درنظر گرفته شوند، نیاز دارد (وَندرلین و همکاران، ۲۰۰۷). فرض کنید $Y = (Y_۱, \ldots, Y_p)$ و هر یک از p متغیر می‌تواند دارای مقادیر گم‌شده باشد. فرض کنید، Y_j برای $j = ۱, \ldots, p$ یکی از p متغیر دارای مقادیر گم‌شده است. رکوردهای مشاهده شده و گم‌شده jامین متغیر یعنی Y_j به ترتیب با Y_j^{obs} و Y_j^{mis} نمایش داده می‌شود. به‌گونه‌ای که $Y^{obs} = (Y_۱^{obs}, \ldots, Y_p^{obs})$ و $Y^{mis} = (Y_۱^{mis}, \ldots, Y_p^{mis})$ است. تعداد جانهی‌ها $m > ۱$ در نظر گرفته می‌شود. hامین مجموعه داده جانهی شده با Y^h نمایش داده می‌شود که در آن $h = ۱, \ldots, m$ است. با قرار دادن $Y_{-j} = (Y_۱, \ldots, Y_{j-۱}, Y_{j+۱}, \ldots, Y_p)$ و Q کمیت مورد نظر مانند ضرایب رگرسیون است که لازم است برآورد شود. یک مجموعه داده کامل فرضی، که یک نمونه تصادفی مشاهده شده جزئی از توزیع p متغیره $P(Y|\theta)$ است را در نظر بگیرید. توزیع چند متغیره Y از طریق بردار پارامترهای نامعلوم θ، مشخص شده است. مساله این است که چگونه می‌توان توزیع چند متغیره θ را به طور صریح یا ضمنی بدست آورد. در الگوریتم $MICE$ توزیع پسین θ از طریق نمونه‌گیری تکراری از توزیع شرطی به صورت $P(Y_۱|Y_{-۱}, \theta_۱), \ldots, P(Y_p|Y_{-p}, \theta_p)$ به دست می‌آید. پارامترهای $\theta_۱, \ldots, \theta_p$ مربوط به چگالی‌های شرطی مربوطه هستند و لزوماً حاصل‌ضرب تجزیه به عامل‌های توزیع توام $P(Y|\theta)$ نیستند. با استخراج یک نمونهٔ ساده از توزیع‌های حاشیه‌ای مشاهده شده، تکرار tام معادلات زنجیره‌ای، یک نمونه‌گیر گیبز است که به طور متوالی از توزیع

$$
\begin{aligned}
&\theta_۱^{*(t)} \sim P(\theta_۱|Y_۱^{obs}, Y_۲^{(t-۱)}, \ldots, Y_p^{(t-۱)})\\
&Y_۱^{*(t)} \sim P(Y_۱|Y_۱^{obs}, Y_۲^{(t-۱)}, \ldots, Y_p^{(t-۱)}, \theta_۱^{*(t)})\\
&\vdots\\
&\theta_p^{*(t)} \sim P(\theta_p|Y_p^{obs}, Y_۱^{(t)}, \ldots, Y_{p-۱}^{(t)})\\
&Y_p^{*(t)} \sim P(Y_p|Y_p^{obs}, Y_۱^{(t)}, \ldots, Y_p^{(t)}, \theta_p^{*(t)})
\end{aligned}
$$

استخراج می‌شود، که در آن $Y_j^{(t)} = (Y_j^{obs}, Y_j^{*(t)})$، j امین متغیر جانهی‌شده در t امین تکرار است. توجه داشته باشید که جانهی‌های قبلی $Y_j^{*(t-۱)}$ از طریق ارتباط آن با سایر متغیرها و نه مستقیماً وارد $Y_j^{*(t)}$ می‌شود. در این روش، همگرایی بر خلاف بسیاری از روش‌های دیگر $MCMC$، می‌تواند بسیار سریع باشد. نام معادلات زنجیره‌ای به این واقعیت اشاره دارد که الگوریتم $MICE$ را می‌توان به راحتی به عنوان ترکیبی از رویه‌های تک‌متغیره برای پر کردن داده‌های گم‌شده پیاده سازی کرد. جانهی چندگانه با استفاده از معادلات زنجیره‌ای یک رویکرد بسیار انعطاف‌پذیر است و برای انواع متغیرها قابل استفاده است. در این روش، داده‌های گم‌شده بر اساس متغیرهای مشاهده‌شده برای

یک مورد خاص و ارتباط میان آن‌ها در کل داده جانهی می‌شوند. از آنجا که جانهی چندگانه شامل چندین پیش‌بینی برای هر مقدار گم‌شده است، این رویکرد جانهی عدم قطعیت را درنظر می‌گیرد و می‌توان خطاهای استاندارد را محاسبه نمود. در روش $MICE$ فرض بر این است که متغیرهای مورداستفاده در فرایند جانهی از مکانیسم MAR پیروی می‌کنند. بنابراین اگر این رویکرد بر روی داده‌ای که از مکانیسم MAR پیروی نمی‌کند، پیاده‌سازی شود، ممکن است باعث اریبی برآوردها شود. فرایند جانهی براساس معادلات زنجیره‌ای را می‌توان به چهار گام اصلی تقسیم کرد (لاکویر و همکاران، ۲۰۲۲):

گام ۱: یک جانهی ساده برای تمامی مقادیر گم‌شده اعمال می‌شود (جانهی میانگین برای متغیرهای پیوسته و جانهی نما برای متغیرهای رسته‌ای). در پایان این مرحله همه متغیرها دارای مقدار هستند و اصطلاحاً داده‌ها کامل هستند.

گام ۲: جانهی گم‌شدگی‌های متغیر Y_j با استفاده از واحدهای مشاهده‌شده متغیر Y_j و مقادیر جانهی شدهٔ Y_i برای $i \neq j$ انجام می‌شود. به ترتیب هر بار یک متغیر که در شروع کار دارای مقدار گم‌شده بود انتخاب می‌شود و داده‌های جانهی شده آن که در مرحله یک کامل شده بود، به حالت اول یعنی گم‌شدگی بازگردانده خواهد شد. مقادیر گم‌شده این متغیر با استفاده از داده‌های مشاهده شده و مقادیر جانهی شده سایر متغیرها و یک مدل جانهی که توسط کاربر تعیین می‌شود، پیش‌بینی شده و مجدداً جانهی خواهند شد.

گام ۳: مقادیر گم‌شدهٔ Y_j با استفاده از پیش‌بینی‌های حاصل از مدل رگرسیونی با استفاده از سایر متغیرهای موجود و جانهی‌شده، جانهی می‌شود.

گام ۴: گام‌های ۲ و ۳ برای هر متغیر که گم‌شدگی دارد تکرار می‌گردد.

گام ۵: گام‌های ۲ تا ۴ به تعداد دفعات موردنیاز، n بار، تکرار می‌شود (راقاناتان و همکاران، ۲۰۰۱).

فرآیند گفته‌شده تنها یک مجموعه دادهٔ جانهی‌شده تولید می‌کند. برای دستیابی به چندین جانهی مختلف، این فرایند به تعداد m بار اجرا می‌شود (گراهام و همکاران ، ۲۰۰۷).

یک روش خوب برای $MICE$ این است که یک توزیع نمونه‌گیری برای هر مقدار گم‌شده تولید شود که به صورت منطقی منعکس‌کننده اطلاعات موجود در مورد مقادیر مورد انتظار شرطی و همچنین عدم قطعیت ذاتی برآورد باشد. به عنوان مثال، BLR از یک توزیع نرمال با میانگین و واریانسی که بوسیله مدل رگرسیونی تولید شده است، نمونه‌گیری می‌کند و PPM مقادیری را از تعدادی همسایگی که امیدریاضی شرطی مشابه دارند تولید می‌کند. در ادامه، موارد زیر با یکدیگر مقایسه می‌شوند.

- جانهی $MICE$ با استفاده از ابر یادگیرنده
- جانهی $MICE$ با استفاده از PPM
- جانهی $MICE$ با استفاده از BLR

۲.۱.۱ جانهی $MICE$ با استفاده از PPM

PPM یک روش جانهی نیمه‌پارامتری است که از داده‌های مشاهده شده برای جانهی مقادیر گم‌شده نمونه می‌گیرد (لیتل، ۱۹۸۸). برای متغیر Y با مقادیر گم‌شده و مجموعه‌ای از متغیرهای پیشگوی X، در این روش، یک رگرسیون خطی ساده از Y روی مجموعه متغیرهای پیشگوی X برازش داده می‌شود تا از این طریق مجموعه‌ای از ضرایب β برآورد شود. سپس به تصادف از توزیع پیشگوی پسین β برای تولید یک مجموعه جدید β^* استفاده می‌شود تا تمام مقادیر Y شامل مقادیر مشاهده شده و گم‌شده پیش‌بینی شود. این روش برای هر مقدار گم‌شده، k مقدار مشاهده شدهٔ Y را که مقادیر پیش‌بینی شدهٔ آن به مقدار پیش‌بینی شدهٔ داده‌های گم‌شده نزدیکتر است پیدا می‌کند و به تصادف این مقادیر را به مقادیر گم‌شده اختصاص می‌دهد. این روش برای هر مجموعه داده کامل تکرار می‌شود. یکی از مزایای PMM نسبت به روش‌های استاندارد مبتنی بر رگرسیون خطی این است که توزیع داده‌های جانهی، بهتر توزیع داده‌های مشاهده شده را منعکس می‌کنند و عموماً در شبیه‌سازی‌ها خوب عمل می‌کند (مارشال و همکاران، ۲۰۱۰). روش PMM استاندارد در وضعیت گم‌شدگی تک‌متغیره به شرح زیر عمل می‌کند. فرض کنید که متغیر y با n واحد یا آزمودنی وجود دارد و تنها این متغیر دارای مقادیر گم‌شده است. همچنین، این متغیر متناظر با ماتریس $n \times p$ بُعدی متغیرهای پیش‌بینی‌کننده مشاهده‌شده X است. فرض کنید (y_{obs}, y_{mis}) به ترتیب نشان‌دهنده مقادیر مشاهده‌شده و گم‌شده متغیر y باشد. گام‌های این الگوریتم عبارتند از:

گام ۱: یک رگرسیون خطی به صورت $y_{y_{obs}} = X_{y_{obs}}\beta + \epsilon_{y_{obs}}$ برای برآورد ضرایب رگرسیونی β برازش دهید.

گام ۲: یک نمونه تصادفی از توزیع پیشگوی پسین $\hat{\beta}$ با استفاده از رگرسیون بیزی گرفته می‌شود تا ضرایب رگرسیونی جدیدی به‌نام $\hat{\beta}^*$ به‌دست آید. معمولاً، این نمونه، نمونهٔ تصادفی از یک توزیع نرمال چندمتغیره با میانگین $\hat{\beta}$ و ماتریس کوواریانس برآورد شدهٔ β است.

گام ۳: با استفاده از $\hat{\beta}^*$، مقادیر پیش‌بینی‌شده برای y، به‌نام $\hat{y}^*$، را برای تمام n واحد یا آزمودنی، که شامل مقادیر گم‌شده نیز است، تولید می‌شود.

گام ۴: برای $y_{y_{miss}}$، که مربوط به مقدار گم‌شده است، مجموعه‌ای از k مشاهدهٔ $y_{y_{obs}}$ شناسایی می‌شود که مقادیر پیش‌بینی‌شده آن‌ها به مقدار پیش‌بینی‌شده برای مشاهدات با داده‌های گم‌شده نزدیک است. به‌عبارت دیگر، $\hat{y}^*_j = X_{yobs}\hat{\beta}^*$ با $\hat{y}^* = X_{y_{miss}}\hat{\beta}^*$ جور و تطبیق پیدا می‌کند.

گام ۵: به‌تصادف یکی از نامزدهایی که مشخص شده‌اند، انتخاب می‌شود و مقدار مشاهده‌شده آن به‌عنوان جایگزینی برای مقدار گم‌شده اختصاص داده می‌شود.

گام ۶: برای جانهی چندگانه، مراحل ۲ تا ۵، m بار تکرار می‌شود تا مجموعه دادهٔ جانهی‌شده به‌دست آید.

روش PMM مشابه بسیاری از روش‌های دیگر، زمانی که مقادیر گم‌شده در بیش از یک متغیر وجود دارد از روش EM استفاده می‌کند. این کار از طریق فرآیند دنباله‌ای به‌گونه‌ای انجام می‌شود که در آن جانهی به صورت متغیر به متغیر و با تکرار انجام می‌شود تا مقادیر جانهی شده به همگرایی برسند.

۲.۱.۲ جانهی $MICE$ با استفاده از BLR

روش BLR یک رویکرد رایج و آسان برای جانهی چندگانهٔ متغیرهای پیوسته با توزیع نرمال است که برای انجام آن ابتدا یک رگرسیون Y روی X برای برآورد ضریب β انجام می‌شود، سپس از توزیع پیشگوی پسین β، به تصادف β^* تولید می‌شود. این مجموعه از ضرایب برای پیش‌بینی مشاهدات گم‌شده Y استفاده می‌شوند. با این وجود برای متغیرهای غیرنرمال و روابط غیرخطی، رگرسیون خطی می‌تواند برای پیش‌بینی‌های دقیق مناسب نباشد. براساس وَن‌بردن (۲۰۱۸) فرض کنید که متغیر y با n واحد یا آزمودنی وجود دارد و تنها این متغیر دارای مقادیر گم‌شده است. همچنین، این متغیر متناظر با ماتریس $n \times p$ بُعدی متغیرهای پیش‌بینی‌کننده مشاهده‌شده X است. فرض کنید (y_{obs}, y_{mis}) به ترتیب نشان‌دهنده مقادیر مشاهده‌شده و گم‌شده متغیر y باشد. نماد X_{obs} زیرمجموعه n_1 ردیف از X را نشان می‌دهد که y آن مشاهده شده است، و X_{mis} زیرمجموعه مکمل n_0 ردیف از X است که y آن گم‌شده است. بردار حاوی n_1 داده مشاهده شده در y با y_{obs} و بردار حاوی n_0 دادهٔ گم‌شده y با y_{mis} نشان داده می‌شود. در ادامه چهار روش مختلف جانهی تحت مدل خطی عادی بر اساس وَن‌بردن (۲۰۱۸) معرفی می‌گردد.

۱. روش پیش‌بینی: پیش‌بینی مقادیر جانهی با استفاده از رابطهٔ $\dot{y} = \hat{\beta}_0 + X_{mis}\hat{\beta}_1$ که در آن $\hat{\beta}_1$ و $\hat{\beta}_0$ برآورد کمترین توان‌های دوم هستند که از مقادیر مشاهده شده به‌دست آمده‌اند. این روش، جانهی رگرسیون نامیده می‌شود و در روش $MICE$ این روش با دستور $norm.predict$ در دسترس است.

۲. روش پیش‌بینی به‌علاوهٔ نویز: پیش‌بینی مقادیر جانهی با استفاده از رابطهٔ قبلی به علاوه عبارت خطا قابل دستیابی است. $\dot{y} = \hat{\beta}_0 + X_{mis}\hat{\beta}_1 + \dot{\epsilon}$ که در آن $\dot{\epsilon}$ به‌صورت تصادفی از توزیع $\dot{\epsilon} \sim N(0, \hat{\sigma}^2)$ انتخاب می‌شود. این روش، جانهی رگرسیون تصادفی نامیده می‌شود و در روش $MICE$ این روش با دستور $norm.nob$ در دسترس است.

۳. روش جانهی چندگانه بیزی: فرض کنید $\dot{y} = \dot{\beta}_0 + X_{mis}\dot{\beta}_1 + \dot{\epsilon}$ که در آن $\dot{\epsilon} \sim N(0, \dot{\sigma}^2)$ است. همچنین پارامترهای $\dot{\beta}_0, \dot{\beta}_1, \dot{\sigma}$ از توزیع پسین به شرط داده‌های مشاهده شده به تصادف انتخاب می‌شوند. در روش $MICE$ این روش با دستور $norm$ در دسترس است.

۴. روش جانهی چندگانه بوت‌استرپ: فرض کنید $\dot{y} = \dot{\beta}_0 + X_{mis}\dot{\beta}_1 + \dot{\epsilon}$ که در آن $\dot{\epsilon} \sim N(0, \dot{\sigma}^2)$ و پارامترهای $\dot{\beta}_0, \dot{\beta}_1, \dot{\sigma}$ برآورد کمترین توان‌های دوم حاصل از نمونهٔ بوت‌استرپ از داده‌های مشاهده‌شده است. در روش $MICE$ این روش با دستور $norm.boot$ در دسترس است.

۲.۲ MissForest

این الگوریتم که توسط اِستِخاون و بالمَن (۲۰۱۲) معرفی شده است، همان الگوریتم جانهی به روش جنگل تصادفی است. این روش جانهی، از الگوریتم جنگل تصادفی برای پیش‌بینی مقادیر گم‌شده بر اساس مقادیر مشاهده شده در مجموعه داده استفاده می‌کند. این روش می‌تواند با انواع مختلف متغیرها به‌طور همزمان جانهی را انجام دهد. اکثر روش‌های رایج و متداول جانهی به‌طور مثال جانهی‌های چندمتغیره با استفاده از معادلات زنجیره‌ای $MICE$ برای

داده‌های آمیخته، بستگی به پارامترهای تنظیم یا تعیین یک مدل پارامتری دارد. انتخاب چنین پارامترهای تنظیم یا مدل‌هایی، بدون دانش قبلی مشکل است و ممکن است تأثیر چشمگیری بر عملکرد یک روش داشته باشد. روش $MissForest$، با رویکردی متفاوت بدون نیاز به هیچگونه پارامتر و بدون نیاز به دانش پیشین دربارهٔ مجموعه داده‌ها، امکان جانهی داده‌های آمیخته را فراهم می‌کند.

این الگوریتم ابتدا مجموعه داده را به دو زیرمجموعه تقسیم می‌کند، یکی زیرمجموعه داده‌های کامل و دیگری، زیرمجموعه مقادیر گم‌شده است. سپس یک مدل جنگل تصادفی بر روی زیرمجموعه داده کامل آموزش داده می‌شود تا مقادیر گم‌شده در زیرمجموعه دیگر را پیش‌بینی کند. این فرآیند چندین بار تکرار می‌شود تا دقت جانهی بهبود یابد. این روش به ویژه برای مجموعه داده‌هایی با روابط پیچیده بین متغیرها و تعداد زیادی مقادیر گم‌شده روبرواست، مناسب است.

برای اجرای الگوریتم بر اساس اِستِخاون و بالمَن (۲۰۱۲)، فرض کنید $Y = (Y_1, \ldots, Y_p)$ یک ماتریس $n \times p$ است که سطر iام این ماتریس به صورت $Y_i = (Y_{i1}, \ldots, Y_{ip})$ نمایش داده می‌شود. مجموعه مقادیر گم‌شده با y^{mis} و مجموعه مقادیر مشاهده شده با y^{obs} نمایش داده می‌شود. جانهی روش جنگل تصادفی RF با استفاده از داده‌های جانهی شده با روش میانگین و آموزش روش جنگل تصادفی RF روی این داده‌ها به پیش‌بینی بخش گم‌شدهٔ متغیر هدف می‌پردازد. در این روش، مقادیر گم‌شده با استفاده از روش RF آموزش داده شده روی بخش مشاهده شدهٔ مجموعه داده، پیش‌بینی می‌شود. برای یک متغیر دلخواه Y_d که شامل مقادیر گم‌شده است، می‌توان مجموعه داده را به چهار بخش تقسیم کرد:

- مقادیر مشاهده dامین متغیر که آن‌ها را با نماد y_d^{obs} نشان می‌دهیم.
- مقادیر گم‌شده dامین متغیر که آن‌ها را با نماد y_d^{mis} نشان می‌دهیم.
- سایر متغیرها (به غیر از متغیر Y_d) که واحدهای متغیر dام آن گم‌شده است با نماد x_d^{mis} نشان می‌دهیم.
- سایر متغیرها (به غیر از متغیر Y_d) که واحدهای متغیر dام آن مشاهده شده است با نماد x_d^{obs} نشان می‌دهیم.

برای نقطه آغاز، یک برآورد اولیه برای مقادیر گم‌شده با استفاده از روش جانهی میانگین یا سایر روش‌ها برای Y انجام می‌دهیم. سپس تمام متغیرهای دارای مقادیر گم‌شده، براساس میزان گم‌شدگی به صورت $y_1, \ldots, y_p$ از کمترین به بیشترین میزان گم‌شدگی مرتب می‌شوند. برای هر متغیر گم‌شدهٔ d، با استفاده از متغیر هدف y_d^{obs} و متغیرهای پیشگوی x_d^{obs}، مدل RF برازش داده می‌شود. سپس مقادیر گم‌شدهٔ y_d^{mis} با استفاده از مدل آموزش داده شدهٔ RF روی x_d^{mis} جانهی می‌گردند. این فرایند جانهی تا رسیدن به شرط توقف ادامه می‌یابد. شرط توقف زمانی است که تفاوت میان داده‌های جدید و داده‌های قبلی برای نخستین بار با درنظرگرفتن هر دو نوع متغیر پیوسته و رسته‌ای افزایش یابد. برای بررسی موضوع همگرایی، یک نقطه بحرانی به نام δ تعریف می‌شود که تفاضل مقدار جانهی شده قبلی و جدید است. وقتی اختلاف مقدار جانهی قدیم و جدید برابر این نقطه بحرانی یا کمتر از آن شود، تکرار الگوریتم متوقف می‌گردد.

۲.۳ MissRanger

این الگوریتم که تحت عنوان جانهی سریع داده‌های گم‌شده [۱] از آن یاد می‌شود، یک الگوریتم جانهی است که برای داده‌هایی که نوع آن‌ها آمیخته‌ای از انواع داده‌ها است، استفاده می‌شود و بر اساس روش جنگل تصادفی عمل می‌کند. این الگوریتم از زنجیره‌ای کردن الگوریتم‌های جنگل‌های تصادفی برای پیش‌بینی مقادیر گم‌شده استفاده می‌کند. این الگوریتم ابتدا مدل جنگل تصادفی را بر داده‌های موجود برازش می‌دهد و سپس از این مدل برای پیش‌بینی مقادیر گم‌شده استفاده می‌کند. الگوریتم MissRanger برای پیاده‌سازی جنگل تصادفی، امکان تنظیم پارامترهای مختلف مانند تعداد درختان، عمق درختان و...را فراهم می‌کند. این الگوریتم، جانهی سریع داده‌های گم‌شده را با استفاده از جنگل تصادفی زنجیره‌ای انجام می‌دهد. همچنین جایگزینی برای الگوریتم $MissForest$ است. اساساً، هر متغیر با پیش‌بینی‌های یک جنگل تصادفی با استفاده از همه متغیرهای دیگر به عنوان متغیرهای کمکی جانهی می‌شود. الگوریتم به گونه‌ای است که جانهی چندین بار روی همه متغیرها تکرار می‌شود تا زمانی که میانگین خطای پیش‌بینی خارج از کیسهٔ (OOB) [۲] مدل‌ها، بهبود یابد. میانگین خطای پیش‌بینی خارج از کیسه (OOB) به روشی اشاره دارد که در آن از نمونه‌های داده‌ای که در حین آموزش مدل استفاده نشده‌اند، برای ارزیابی عملکرد مدل استفاده می‌شود. در واقع، در روش‌های یادگیری مانند جنگل تصادفی، هر درخت تصمیم با یک زیرمجموعه تصادفی از داده‌ها آموزش می‌بیند و نمونه‌های باقی‌مانده می‌توانند برای ارزیابی دقت مدل استفاده شوند. این الگوریتم از ترکیب روش جانهی جنگل تصادفی و جورسازی میانگین پیشگو استفاده می‌کند که باعث می‌شود از جانهی مقادیری که در داده‌های اولیه وجود ندارد، خودداری کند. این الگوریتم به صورت خودکار بازه‌های اطمینان برای مقادیر پیش‌بینی شده ارائه می‌دهد و از جورسازی میانگین پیشگو یا pmm برای افزایش واقع‌بینی در تکمیل داده‌ها استفاده می‌کند. درکل، این الگوریتم یک روش قدرتمند و موثر برای تکمیل داده‌های گم شده است که می‌تواند مورد استفاده قرار گیرد.

۲.۴ روش جانهی SuperMICE

در روش لاکویر و همکاران (۲۰۲۲) برای جانهی مقادیر گم‌شده، از ترکیب یک ابرباد گیرنده و $MICE$ که یک رویکرد نیمه‌پارامتری است، استفاده شده است. این روش، یکی از روش‌های جدید در خصوص استفاده از رویکرد یادگیری ماشین ترکیبی برای جانهی چندگانه است. الگوریتم ابریادگیرنده که در الگوریتم $SuperMICE$ از آن استفاده می‌شود، یک روش یادگیری ترکیبی (ترکیب چندین مدل یادگیری ماشین) است تا یک مدل نهایی با دقت بالاتر ایجاد کند. این کار از طریق یافتن بهترین ترکیب از چندین مدل یادگیری ماشین اتفاق می‌افتد. بنابراین محقق می‌تواند طیف وسیعی از یادگیرنده‌های متنوع شامل مدل‌های پارامتری ساده تا پیچیده را برای پیش‌بینی ترکیب کند. در روش $SuperMICE$، برای برآورد مقادیر داده‌های گم‌شده از یک توزیع نرمال استفاده می‌شود. میانگین این توزیع از یک مدل ابر یادگیرنده به دست می‌آید و واریانس آن نیز از برآورد موضعی واریانس (آئرت و همکاران،

[1] Fast imputation of missing values
[2] Out of bag

۲۰۰۲) محاسبه می‌شود. فرض کنید مجموعه دادهٔ X با p متغیر $X = (x_1, \ldots, x_p)$ و هر متغیر دارای n واحد نمونه است. مقادیر هر یک از متغیرها می‌تواند گم‌شده یا ناکامل باشد. ماتریس دودویی $n \times p$ متناظر با مجموعه دادهٔ معرفی شده، برای $i = 1, \ldots, n$ و $j = 1, \ldots, p$ به‌صورت

$$\delta_{ij} = \begin{cases} 0 & x_{ij} \text{ها گم‌شده هستند} \\ 1 & x_{ij} \text{ها مشاهده شده هستند} \end{cases} \qquad (1)$$

تعریف می‌شود. علاوه بر این فرض کنید $\mathcal{L}$ یک کتابخانه (یک مجموعه) از الگوریتم‌های پیشگو به تعداد q است. تکرار الگوریتم با نماد t نمایش داده می‌شود به طوری که $0 \leq t \leq T$ است. فرض کنید m تعداد جانهی در روش جانهی چندگانه است. در نهایت، kامین مجموعه داده در t امین تکرار الگوریتم با بالانویس (k,t) نشان داده شده است. در ادامه مراحل الگوریتم $MICE$ با استفاده از ابریادگیرنده شرح داده می‌شود.

گام ۱. مقداردهی اولیه: در تکرار صفر $(t = 0)$، m مجموعه داده $X^{(1,0)}, \ldots, X^{(m,0)}$ به شرح زیر ساخته می‌شود. برای هر مقدار گم‌شده که $\delta_{ij} = 0$ است، مقدار x_{ij} صفر قرار داده می‌شود. سپس مقادیر گم‌شده با میانگین مقادیر مشاهده‌شده همان ستون و متناظر با متغیر x_j با استفاده از رابطه

$$\text{برای j ثابت} \quad x_{ij}^{(k,0)} = \frac{\sum_{i=1}^{n} x_{ij}\delta_{ij}}{\sum_{i=1}^{n} \delta_{ij}},$$

جانهی می‌شود اگر متغیرها دودویی باشند، میانگین محاسبه شده به صفر یا یک گرد می‌شود.

گام ۲. پیش‌بینی: برای مجموعه دادهٔ k $1 \leq k \leq m$ و متغیر j که $1 \leq j \leq p$ و حداقل یک مقدار گم‌شده دارد، هر کدام از الگوریتم‌های $\ell \varepsilon \mathcal{L}$ برازش داده می‌شود که منجر به پیشگویی $X_j^{(k,t)}$ از طریق $X_{(j)}^{(k,t)}$ می‌شود. نماد $X_{(j)}^{(k,t)}$ به‌معنای حذف متغیر j است. به‌عبارت دیگر، متغیر j که می‌خواهیم مقادیر گم‌شده آن را برآورد کنیم از مجموعه متغیرها حذف می‌کنیم که این مجموعه جدید را با نماد $X_{(j)}^{(k,t)}$ نشان می‌دهیم. بر اساس این مجموعه جدید مقادیر متغیر $x_j^{(k,t)}$ توسط هر کدام از الگوریتم‌های موجود در کتابخانه الگوریتم‌ها پیش‌بینی می‌شود. در نهایت به هر کدام از الگوریتم‌های موجود در کتابخانه یک وزن داده می‌شود که این وزن‌ها $\hat{\alpha}_j = (\hat{\alpha}_{j1}, \ldots, \hat{\alpha}_{jq})$ به‌گونه‌ای تعیین می‌شوند که مجموع آن‌ها برابر یک شود یعنی $\sum_{\ell=1}^{q} \hat{\alpha}_{\ell,j} = 1$. این وزن‌ها طوری به‌دست می‌آیند که مخاطرهٔ اعتبارسنجی متقابل [1] به‌دست آمده، کمترین مقدار ممکن گردد (آئرت و همکاران، ۲۰۰۲). مقدار پیش‌بینی متغیر مورد نظر که دارای مقدار گم‌شده است از رابطه

$$\hat{x}_j^{(k,t)} = \hat{\Psi}_{SL,j}^{(k,t)}(X_{(j)}^{(k,t)}) = \sum_{\ell=1}^{q} \hat{\alpha}_{j\ell}^{(k,t)} \hat{\Psi}_{j,\ell}^{(k,t)}(X_{(j)}^{(k,t)})$$

[1] Cross-validation risk

به دست می‌آید، که در آن، $\Psi(\cdot)$ تابعی است که پیشگوها را از یک الگوریتم و مجموعه‌ای از وزن‌های برآورد شدهٔ $\hat{\alpha}_j = (\hat{\alpha}_{j1}, \ldots, \hat{\alpha}_{jq})$ برمی‌گرداند.

گام ۳. جانهی: در این مرحله، از رویکردی نیمه‌پارامتری برای جانهی داده‌های گم‌شده استفاده شده است. اگر نوع داده‌ها، دودویی باشد، از توزیع برنولی با احتمال‌های برابر با پیشگوهای SL نمونه گرفته می‌شود و به جای مقادیر گم‌شده جایگزین می‌شود. برای متغیرهای پیوسته، مقادیر داده‌های گم‌شده، به تصادف از توزیع نرمال با میانگین پیشگوهای SL و برآورد واریانس موضعی، استخراج و جایگزین مقادیر گم‌شده می‌شود. واریانس موضعی به‌عنوان یک واریانس وزن‌دهی شده است که به مشاهداتی که مقادیر پیشگوی آن‌ها مشابه مقادیر گم‌شده است و مشاهداتی که در نواحی که گم‌شدگی بیشتری وجود دارد، وزن بیشتری می‌دهد. به طور خاص، برای یک تابع هسته $K_h(.)$ با پهنای باند h، وزن‌های $w_{ij}(x) = \frac{\delta_{ij}\hat{w}_{ij}(x_{ij})}{\hat{\pi}(x)}$ محاسبه می‌شوند، که در آن $\hat{w}_{ij}(x) = \frac{K_h(x-\hat{x}_{ij}^{(k,t)})}{\sum_{i=1}^{n} K_h(x-\hat{x}_{ij}^{(k,t)})}$ وزن‌های معمول ناداریا (۱۹۶۴) هستند و $\hat{\pi}(x) = \frac{K_h(x-\hat{x}_{ij}^{(k,t)})\delta_{ij}}{\sum_{i=1}^{n} K_h(x-\hat{x}_{ij}^{(k,t)})}$ برآورد موضعی از گم‌شدگی است. واریانس وزنی می‌تواند به صورت

$$\hat{\sigma}_{ij}^{2(k,t)} = \frac{\sum_{i=1}^{n} w_{ij}(\hat{x}_{ij}^{(k,t)})(\hat{x}_{ij}^{(k,t)} - \hat{x}_{ij}^{*(k,t)})}{\sum_{i=1}^{n} w_{ij}(\hat{x}_{ij})}$$

محاسبه شود که در آن $\hat{x}_{ij}^{*}$ میانگین وزنی است. در نهایت، هر مقدار گم‌شده از توزیع متناظر

$$x_{ij}^{(k,t+1)} \sim \mathcal{N}(\hat{x}_{ij}^{(k,t)}, \hat{\sigma}_{ij}^{2(k,t)})$$

نمونه‌گیری می‌شود. برای T تکرار (معمولاً ۱۰ تا ۲۰ تکرار کافیست) گام‌های ۲ و ۳ تکرار می‌شود.

گام ۴. برآورد: برآورد و ادغام برآوردها و خطای استاندارد برای هر یک از m مجموعه دادهٔ جانهی شده، طبق رویّه استاندارد $MICE$ و طبق قواعد روبین ترکیب می‌شوند (روبین، ۲۰۰۴).

۳ مطالعهٔ شبیه‌سازی

در این بخش، الگوریتم $MICE$ با سه روش $SuperLearner$، PMM و BLR با درصدهای مختلف گم‌شدگی، روی داده‌های شبیه‌سازی شده، اجرا و عملکرد این سه روش براساس معیار اریبی و درصد پوشش با یکدیگر مقایسه شده است. برای این منظور ابتدا دو متغیر کمکی X_1 و X_2 و یک متغیر پاسخ Y را تعریف نموده و بین دو متغیر کمکی، چهار رابطه مختلف ایجاد می‌شود. سپس گم‌شدگی مصنوعی در یکی از متغیرهای کمکی (X_2) ایجاد می‌شود. مکانیسم گم‌شدگی در متغیر مورد نظر به دو صورت گم‌شدگی تصادفی و گم‌شدگی کاملاً تصادفی است. در این بخش، چهار دادهٔ مختلف شبیه‌سازی شده است. اندازهٔ نمونه در هر شبیه‌سازی، سه حالت اندازهٔ نمونه ۱۰۰، ۴۰۰ و ۷۰۰ در نظر گرفته شده است. در فرآیند شبیه‌سازی، گم‌شدگی مصنوعی ایجاد شده در متغیر X_2 با پنج حالت ۱۰٪، ۲۰٪، ۳۰٪، ۴۰٪، ۵۰٪ می‌باشد. سایر متغیرها (X_1, Y) فاقد گم‌شدگی هستند. بنابراین جانهی فقط بر

روی متغیر $X_۲$ صورت می‌گیرد. در شبیه‌سازی‌ها از دو مکانیسم برای ایجاد گم‌شدگی استفاده شده است. در روش اول احتمال گم‌شدگی در متغیر $X_۲$ به متغیر $X_۱$ وابسته است. این احتمال با افزایش مقدار $X_۱$ افزایش می‌یابد. در روش دوم، گم‌شدگی در متغیر $X_۲$ به صورت مستقل از متغیر $X_۱$ و به صورت تصادفی ایجاد شده است. بنابراین، داده‌های تولیدشده در حالت دوم دارای گم‌شدگی کاملاً تصادفی هستند. به دلیل حجم محاسباتی سنگین و زمان‌بر، هر کدام از شبیه‌سازی‌ها به تعداد ۱۰۰ بار تکرار شده است.
در روش اول، رابطه بین دو متغیر کمکی از نوع درجه دوم و از روابط زیر برای شبیه‌سازی استفاده شده است.

$$
\begin{aligned}
&X_۱ \sim Uniform(-۳, ۳)\\
&X_۲ = ۱ + X_۱^۲ + \epsilon_۱\\
&Y = X_۱ + X_۲ + \epsilon_۲\\
&\epsilon_i \sim \mathcal{N}(۰, ۱), \qquad i = ۱, ۲
\end{aligned}
$$

در روش دوم (روش نمایی)، داده‌های شبیه‌سازی به‌صورت زیر ساخته شده‌اند.

$$
\begin{aligned}
&X_۱ \sim Uniform(۰, ۱)\\
&X_۲ = X_۱ + e^{X_۱} + \epsilon_۱\\
&Y = X_۱ + X_۲ + \epsilon_۲\\
&\epsilon_i \sim \mathcal{N}(۰, ۱), \qquad i = ۱, ۲
\end{aligned}
$$

در روش سوم، رابطه بین دو متغیر کمکی از نوع توزیع پواسون متورم در نقطه صفر[1] است و داده‌ها از روابط زیر تولید شده‌اند.

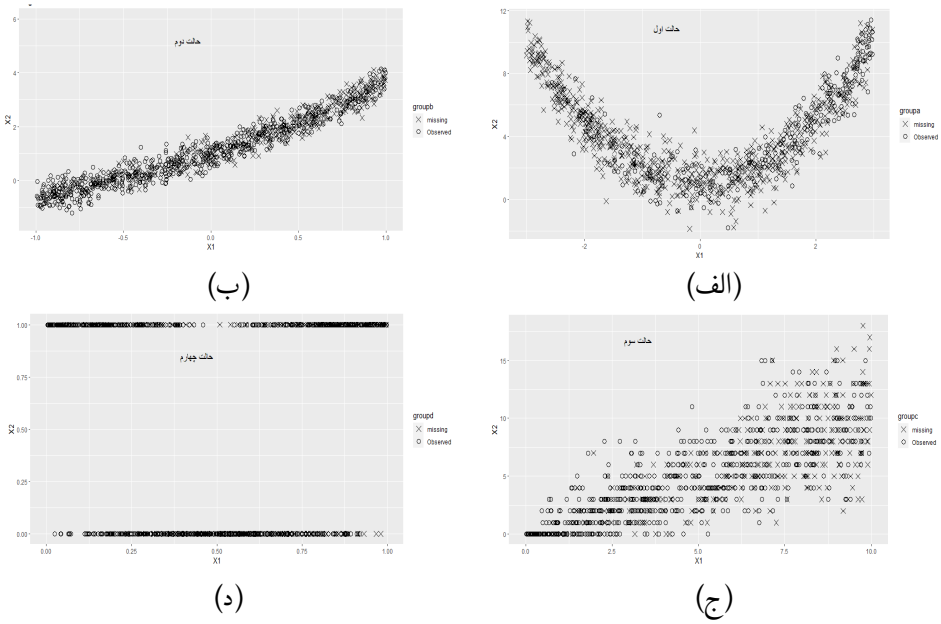
$$
\begin{aligned}
&X_۱ \sim Uniform(۰, ۱۰)\\
&X_۲|X_۱ \sim \begin{cases} ۰ & \text{با احتمال } e^{-X_۱} \\ Poisson(\lambda = X_۱) & \text{با احتمال } ۱ - e^{-X_۱} \end{cases}\\
&Y = X_۱ + X_۲ + \epsilon\\
&\epsilon \sim \mathcal{N}(۰, ۱)
\end{aligned}
$$

[1] Zero-inflated poisson distributation

در روش چهارم، که رابطه بین دو متغیر کمکی، دودویی است، متغیر اول از توزیع یکنواخت $Uniform(۰, ۱)$، تولید شده است. متغیر دوم به‌صورت یک توزیع برنولی که احتمال موفقیت این توزیع، وابسته به متغیر اول است، تولید شده است. برای این روش از روابط

$$
\begin{aligned}
X_۱ &\sim Uniform(۰, ۱) \\
X_۲ &\sim Bernouli(|۲X_۱ - ۱|) \\
Y &= X_۱ + X_۲ + \epsilon \\
\epsilon &\sim \mathcal{N}(۰, ۱)
\end{aligned}
$$

استفاده شده است. رابطه بین متغیرهای کمکی $X_۱$ و $X_۲$ در شکل ۱ نشان داده شده است. بعد از ایجاد چهار رابطهٔ



شکل ۱. رابطه بین متغیرهای کمکی $X_۲, X_۱$

مختلف بین متغیرهای کمکی، داده‌های مورد نیاز تولید شده است. سپس گم‌شدگی در متغیر کمکی $X_۲$ ایجاد شده است. اکنون با استفاده از الگوریتم $MICE$ با روش‌های $SuperLearner$، PMM و BLR، داده‌ها جانهی می‌شود. سپس بر اساس داده‌های جانهی شده، یک مدل رگرسیونی خطی ساده به صورت $Y = \beta_۰ + \beta_۱X_۱ + \beta_۲X_۲ + \epsilon$ برازش داده شده است. بر اساس مدل رگرسیونی تعریف شده، ضرایب $\beta_۱$ و $\beta_۲$ به‌ترتیب برای متغیرهای $X_۱$ و $X_۲$

برآورد شده است. سپس میزان اریبی و درصد پوشش برای ضریب متغیر X_2 محاسبه و این دو معیار مبنای مقایسه عملکرد سه روش $SuperLearner$، PMM و BLR، قرار گرفته شده است. نتایج جداول ۱ و ۲ به‌ترتیب میزان اریبی و درصد پوشش در حالتی که مکانیسم گم‌شدگی تصادفی هستند، برای درصدهای گم‌شدگی، سه اندازهٔ نمونه ۱۰۰، ۴۰۰ و ۷۰۰ و چهار نوع داده مختلف تولید شده (دودویی، درجه دوم، پواسون و نمایی)، ارایه شده است.

جدول ۱. نتایج شبیه‌سازی اریبی با مکانیسم تصادفی (MAR)

درصد گم‌شدگی	روش تولید شبیه‌سازی	اندازهٔ نمونه ۷۰۰ PMM	۷۰۰ BLR	۷۰۰ S.L	۴۰۰ PMM	۴۰۰ BLR	۴۰۰ S.L	۱۰۰ PMM	۱۰۰ BLR	۱۰۰ S.L
۱۰	دودویی	۰٫۰۶۰	۰٫۰۶۰	۰٫۰۶۰	۰٫۰۷۴	۰٫۰۷۵	۰٫۰۷۳	۰٫۱۴۷	۰٫۱۵۱	۰٫۱۵۳
	درجه دو	۰٫۰۰۷	۰٫۰۰۷	۰٫۰۰۶	۰٫۰۱۰	۰٫۰۰۹	۰٫۰۰۹	۰٫۰۱۹	۰٫۰۲۰	۰٫۰۱۹
	پواسون	۰٫۰۱۹	۰٫۰۱۶	۰٫۰۱۶	۰٫۰۲۳	۰٫۰۱۹	۰٫۰۱۸	۰٫۰۴۶	۰٫۰۳۸	۰٫۰۴۲
	روش نمایی	۰٫۰۵۴	۰٫۰۵۳	۰٫۰۵۳	۰٫۰۷۲	۰٫۰۷۸	۰٫۰۷۳	۰٫۱۳۶	۰٫۱۳۳	۰٫۱۳۶
۲۰	دودویی	۰٫۰۶۰	۰٫۰۶۰	۰٫۰۵۷	۰٫۰۸۶	۰٫۰۸۲	۰٫۰۷۴	۰٫۱۷۵	۰٫۱۶۹	۰٫۱۷۲
	درجه دوم	۰٫۰۰۷	۰٫۰۰۶	۰٫۰۰۸	۰٫۰۱۰	۰٫۰۱۱	۰٫۰۱۱	۰٫۰۲۰	۰٫۰۲۱	۰٫۰۲۰
	پواسون	۰٫۰۲۲	۰٫۰۲۲	۰٫۰۱۸	۰٫۰۳۶	۰٫۰۲۱	۰٫۰۱۸	۰٫۰۴۸	۰٫۰۵۳	۰٫۰۴۹
	نمایی	۰٫۰۵۲	۰٫۰۵۳	۰٫۰۵۰	۰٫۰۸۳	۰٫۰۷۷	۰٫۰۷۸	۰٫۱۵۹	۰٫۱۳۸	۰٫۱۴۲
۳۰	دودویی	۰٫۰۷۰	۰٫۰۷۲	۰٫۰۶۷	۰٫۱۱۰	۰٫۱۰۲	۰٫۰۹۱	۰٫۱۹۷	۰٫۲۰۷	۰٫۱۸۷
	درجه دوم	۰٫۰۰۸	۰٫۰۰۹	۰٫۰۰۷	۰٫۰۱۳	۰٫۰۱۳	۰٫۰۱۱	۰٫۰۲۵	۰٫۰۲۱	۰٫۰۲۱
	پواسون	۰٫۰۳۵	۰٫۰۳۶	۰٫۰۲۱	۰٫۰۴۳	۰٫۰۳۶	۰٫۰۳۱	۰٫۰۶۳	۰٫۰۶۰	۰٫۰۶۴
	نمایی	۰٫۰۷۵	۰٫۰۵۵	۰٫۰۵۰	۰٫۱۰۳	۰٫۰۷۸	۰٫۰۸۱	۰٫۱۸۳	۰٫۱۸۴	۰٫۱۸۲
۴۰	دودویی	۰٫۰۹۵	۰٫۰۸۰	۰٫۰۷۵	۰٫۱۱۹	۰٫۱۱۰	۰٫۰۹۱	۰٫۲۰۳	۰٫۲۰۰	۰٫۱۸۷
	درجه دوم	۰٫۰۱۸	۰٫۰۱۵	۰٫۰۱۱	۰٫۰۱۷	۰٫۰۱۶	۰٫۰۱۲	۰٫۳۷	۰٫۰۳۴	۰٫۰۲۷
	پواسون	۰٫۰۴۰	۰٫۰۳۹	۰٫۰۲۸	۰٫۰۳۵	۰٫۰۴۲	۰٫۰۳۵	۰٫۰۶۸	۰٫۰۷۶	۰٫۰۷۰
	نمایی	۰٫۰۷۵	۰٫۰۷۲	۰٫۰۶۵	۰٫۱۰۳	۰٫۱۰۲	۰٫۰۹۸	۰٫۱۸۳	۰٫۱۹۳	۰٫۱۷۶
۵۰	دودویی	۰٫۱۳۶	۰٫۰۹۹	۰٫۰۷۷	۰٫۱۴۵	۰٫۱۲۵	۰٫۰۹۱	۰٫۲۰۰	۰٫۱۹۴	۰٫۱۸۱
	درجه دوم	۰٫۰۴۸	۰٫۰۴۷	۰٫۰۲۰	۰٫۰۵۱	۰٫۰۴۷	۰٫۰۲۷	۰٫۰۶۵	۰٫۰۶۶	۰٫۰۵۲
	پواسون	۰٫۰۵۰	۰٫۰۶۱	۰٫۰۴۰	۰٫۰۵۶	۰٫۰۶۵	۰٫۰۵۰	۰٫۱۰۷	۰٫۰۸۹	۰٫۰۸۹
	نمایی	۰٫۱۰۹	۰٫۱۴۶	۰٫۱۰۸	۰٫۱۱۸	۰٫۱۴۷	۰٫۱۰۴	۰٫۲۴۳	۰٫۲۵۵	۰٫۲۱۲

با توجه به این‌که تئوری ارایه شده برای الگوریتم $SuperLearner$ براساس مکانیسم گم‌شدگی تصادفی عملکرد بالایی دارد، نتایج شبیه‌سازی در تمام حالت نشان از عملکرد بالاتر الگوریتم جانهی $SuperLearner$ دارد که در ادامه جزییات نتایج ارایه می‌شود. نتایج شبیه‌سازی جدول ۱ نشان می‌دهد که الگوریتم $SuperMICE$ نسبت به الگوریتم PMM و BLR، اریبی کمتری برای برآورد β_2 (در تمام حالت‌هایی که بین متغیرهای کمکی ایجاد شده است)، دارد. در تمام حالت‌ها، میزان اریبی با افزایش درصد گم‌شدگی زیادتر می‌شود. در سطوح گم‌شدگی ۱۰، ۲۰ و ۳۰ درصد میزان اریبی هر سه روش به یکدیگر نزدیک‌تر است و با افزایش درصد گم‌شدگی اختلاف بیشتر می‌شود.

در حالتی که رابطه بین متغیرها درجه دو است در گم‌شدگی ۵۰ درصد، بیشترین اختلاف در اریبی در الگوریتم $SuperMICE$ نسبت به الگوریتم PMM و BLR ، وجود دارد. به بیان دیگر بهترین عملکرد $SuperMICE$ در این حالت مشاهده شده است. در این حالت اختلاف اریبی روش $SuperMICE$ با دو روش دیگر بیش از دو برابر است. روش $SuperMICE$ در روش درجه دو و در گم‌شدگی ٪۱۰ کمترین میزان اریبی را داشته است.

جدول ۲. نتایج شبیه‌سازی درصد پوشش با مکانیسم تصادفی (MAR)

درصد گم‌شدگی	روش تولید شبیه‌سازی	اندازهٔ نمونه ۷۰۰			۴۰۰			۱۰۰		
		PMM	BLR	S.L	PMM	BLR	S.L	PMM	BLR	S.L
۱۰	دودویی	۱۰۰	۱۰۰	۹۹	۹۹	۹۷	۹۹	۹۶	۹۶	۹۶
	درجه دو	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰
	پواسون	۱۰۰	۱۰۰	۱۰۰	۹۹	۱۰۰	۱۰۰	۹۸	۱۰۰	۱۰۰
	نمایی	۹۸	۱۰۰	۹۹	۹۷	۹۸	۹۹	۹۵	۹۶	۹۹
۲۰	دودویی	۹۸	۱۰۰	۹۷	۹۸	۹۶	۹۷	۹۴	۹۴	۹۷
	درجه دوم	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰
	پواسون	۹۹	۱۰۰	۱۰۰	۹۶	۱۰۰	۱۰۰	۹۹	۹۸	۹۹
	نمایی	۹۸	۹۸	۹۹	۹۴	۹۶	۹۷	۹۴	۹۵	۹۸
۳۰	دودویی	۹۵	۹۶	۹۷	۹۱	۹۵	۹۹	۹۱	۹۲	۹۲
	درجه دوم	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۹۹	۱۰۰	۱۰۰
	پواسون	۹۵	۱۰۰	۱۰۰	۹۲	۹۹	۱۰۰	۹۳	۹۵	۹۳
	نمایی	۹۸	۹۹	۹۸	۹۷	۹۶	۹۶	۹۱	۹۱	۹۷
۴۰	دودویی	۸۶	۹۱	۹۴	۸۵	۹۱	۹۴	۸۷	۹۱	۹۰
	درجه دوم	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۹۶	۹۷	۹۹
	پواسون	۹۴	۹۷	۹۷	۹۵	۹۹	۹۷	۸۹	۸۷	۹۵
	نمایی	۹۲	۹۷	۹۷	۸۶	۹۴	۹۵	۹۲	۹۲	۹۵
۵۰	دودویی	۶۹	۸۸	۹۳	۷۷	۸۷	۹۸	۹۱	۸۸	۹۴
	درجه دوم	۸۰	۸۰	۹۹	۷۹	۸۰	۹۹	۷۹	۷۴	۸۵
	پواسون	۸۳	۷۶	۹۵	۸۳	۸۴	۹۵	۷۵	۸۴	۹۴
	نمایی	۷۶	۷۴	۹۷	۸۳	۸۲	۹۳	۸۲	۸۴	۹۲

بیشترین میزان اریبی در الگوریتم $SuperMICE$ در درصد گم‌شدگی v درصد و در روش دوم با حجم نمونه ۱۰۰ اتفاق افتاده است. در روش نمایی میزان اریبی هر سه روش بسیار به یکدیگر نزدیک بوده که نسبت به سایر روش‌ها، کمترین اختلاف را داشته است. روش PMM در مقایسه با روش BLR در حالت درجه دوم و دودویی عملکرد کاملاً ضعیف‌تری داشته است. در حالت‌های دیگر در درصدهای گم‌شدگی ۵۰ و ۴۰ عملکرد بهتری داشته و در سایر درصدهای گم‌شدگی عملکرد آن ضعیف بوده است. درصد پوشش در تمام روش‌ها شیب نزولی آرامی دارد، اما در درصد گم‌شدگی ۵۰ درصد، شیب نزولی بالایی پیدا می‌کند. در تمام حالت‌ها، درصد پوشش $SuperMICE$ نسبت به دو روش دیگر بهتر بوده است. در سطوح گم‌شدگی ۱۰، ۲۰ و ۳۰ درصد در تمامی حالت‌ها درصد پوشش برای بازهٔ اطمینان ٪۹۵ برای β_2 بیش از ٪۹۰ است. در گم‌شدگی ۴۰ و ۵۰ ٪ کاهش میزان درصد پوشش در برخی موارد به حدود ٪۷۵ نیز رسیده است. در روش درجه دوم برای درصدهای گم‌شدگی ۴۰ و کمتر از آن عملکرد هر سه روش بسیار به یکدیگر نزدیک بوده است، اما در گم‌شدگی ٪۵۰، درصد پوشش برای روش‌های دیگر تا حدود ٪۸۰ نیز افت داشته است اما روش $SuperMICE$ درصد پوشش ٪۹۹ در حجم نمونه‌های ۴۰۰ و ۷۰۰ را نشان داده است. نتایج مربوط به گم‌شدگی براساس مکانیسم گم‌شدگی کاملاً تصادفی نشان داد که روش $SuperLearnear$ ضعیف‌ترین عملکرد را داشته است که نتایج آن ارایه نشده است.

۴ تحلیل داده‌های طرح آمارگیری کارگاه‌های صنعتی

مجموعه داده مربوط به طرح آمارگیری از کارگاه‌های صنعتی کشور است. کد ۲۸۲۱ $= ISIC$ مربوط به تولید ماشین آلات کشاورزی و جنگلداری است که شامل ۸۰ کارگاه در سال ۱۴۰۰ است. چند متغیر اصلی و مهم در این طرح، ارزش نهاده، ارزش ستانده، ارزش افزوده و I/O (نسبت ارزش نهاده به ارزش ستانده)، کارگاه‌ها است. فرایند انجام جانهی و ارزیابی الگوریتم‌های مختلف یادگیری ماشین به‌گونه‌ای طراحی شده است که در داده‌های پردازش و تمیزشده که بدون گم‌شدگی است، بی‌پاسخی واحد با درصدهای مختلف ۱۰، ۲۰ و ۳۵ درصد در تمام متغیرها غیر از متغیرهای مربوط به اطلاعات چارچوب مانند تعداد کارکن، وضع مالکیت کارگاه، وضعیت حقوقی کارگاه و ...ایجاد شده است. در داده‌های پردازش و تمیزشده که بدون گم‌شدگی است، مجموع ارزش نهاده، ارزش ستانده و ارزش افزودهٔ تمام ۸۰ کارگاه مورد بررسی در این کد فعالیت، در جدول ۳ آورده شده است. این مقادیر برای ارزیابی عملکرد جانهی و

جدول ۳. مجموع مقادیر مولفه‌های نهاده، ستانده و ارزش افزوده در ۸۰ کارگاه (میلیون ریال)

عنوان	مجموع
ارزش نهاده	۴۰۹۷۴/۹۴
ارزش ستانده	۵۵۵۸۶/۲۹
ارزش افزوده	۱۴۶۱۱/۳۶

میزان خطای تولید شده براساس جانهی، کمک‌کننده خواهد بود. از آنجا که بسیاری از مولفه‌های تشکیل دهنده ارزش ستانده، نهاده و نهایتاً ارزش افزوده شامل درصدهای بالایی از مقادیر صفر است (بیش از ۹۰ درصد)، این ویژگی می‌تواند در برخی روش‌های جانهی تاثیرگذار و کیفیت عملکرد روش را کاهش دهد. در خصوص اجرای الگوریتم‌های $SuperMICE$ و PMM و BLR بر روی داده‌های واقعی باید به چند نکته که در لاکویر و همکاران (۲۰۲۲) اشاره شده است، توجه کرد.

- داده‌های تُنُک [1] یا داده‌هایی که تعداد واحدهای مشاهده‌شده در آن کم است، جزء محدودیت‌هایی برای پایین آوردن عملکرد الگوریتم‌هایی مانند PMM به حساب می‌آیند.
- الگوریتم $SuperMICE$ بر روی داده‌هایی با مکانیسم تصادفی عملکرد خوبی دارد، که نتایج آن در بخش شبیه‌سازی ارایه شد. بنابراین نوع مکانیسم گم‌شدگی در عملکرد الگوریتم‌های جانهی بسیار حایز اهمیت است و باید مد نظر باشد.

در خصوص نکته اول، در داده‌های واقعی، متغیرهایی که حداقل ۹۰٪ مقادیر آن‌ها صفر است از مجموعه داده کنار گذاشته شده است. به این ترتیب مشکل تُنُک بودن داده‌ها برطرف می‌شود. همچنین متغیرهایی که نقشی در محاسبه متغیرهای کلیدی مانند ارزش افزوده نداشتند کنار گذاشته شده‌اند. در خصوص ایجاد گم‌شدگی در داده‌ها، بر اساس

[1] Sparse

تابع چگالی متغیر ارزش افزوده، به هر ردیف یک مقدار احتمال به‌گونه‌ای تخصیص داده شده است که با افزایش مقدار ارزش افزوده، احتمال گم‌شدگی افزایش می‌یابد. بنابراین، کارگاه‌های با سود بالا، با احتمال بیشتری به سوالات پاسخ نخواهند داد که معمولاً مکانیسم گم‌شدگی و بی‌پاسخی در کارگاه‌ها در مراکز آماری از این الگو پیروی می‌کند. لذا از این طریق این ویژگی در داده‌های واقعی شبیه‌سازی شده است.

جدول ۴. مقایسه RMSE با مکانیسم گم‌شدگی غیر تصادفی بر اساس چگالی ارزش افزوده (میلیون ریال)

درصد	نام متغیر	$S.L$	PMM	BLR	KNN	$irmi$	$hotdeck$	$CART$	$MissForest$	$MissRanger$
	ارزش مصرف	۱۵۳٫۵	۳۰۷٫۸	۷۲۵۹٫۵	۲۴۶٫۵	۶۳٫۷	۱۰۲٫۵	۷۸٫۸	۱۴۵٫۸	۲۵۴٫۶
	ارزش قطعات مصرفی برای ساخت	۰٫۱	۰٫۲	۱۷٫۱	۰٫۳	۰٫۱	۰٫۱	۰٫۱	۰٫۱	۰٫۱
	ارزش تولید	۱۴۶٫۶	۳۹۵٫۷	۳۷۹۰٫۴	۲۵۹٫۷	۶۳٫۳	۱۳۷٫۳	۴۳٫۸	۲۰۳٫۷	۲۱۴٫۳
	ارزش غذای طبخ شده توسط شاغلین	۰٫۳	۰٫۵	۳۵٫۷	۰٫۲	۰٫۳	۰٫۶	۰٫۲	۰٫۲	۰٫۲
	کالاهای موجود در انبار در اول فروردین	۳۵	۲۲٫۳	۳۳۲٫۹	۱٫۸	۳٫۹	۳٫۷	۳٫۹	۲٫۷	۳
	کالاهای موجود در انبار در اول اسفند	۶٫۴	۴٫۵	۵۷۶٫۳	۳٫۸	۴٫۶	۳٫۲	۲۴٫۱	۳٫۵	۴٫۱
	مواد خام و اولیه موجود در انبار فروردین	۱۱٫۱	۲۴٫۵	۲۲۶۴٫۲	۱۲٫۹	۲۳٫۸	۱۸٫۳	۲۱٫۲	۱۱	۸٫۱
۱۰	ارزش مصرف گازوئیل	۰	۰	۲٫۴	۰	۰٫۱	۰	۰	۰	۰
	ارزش مصرف گاز مایع	۰٫۱	۰٫۲	۵٫۰	۰٫۲	۰٫۱	۰	۰	۰٫۱	۰٫۱
	ارزش مصرف گاز طبیعی	۰٫۲	۰٫۲	۸٫۹	۰٫۱	۰٫۲	۰٫۲	۰٫۱	۰٫۱	۰٫۱
	ارزش مصرف بنزین	۰	۰٫۱	۱٫۴	۰	۰	۰	۰	۰	۰
	ساخت و تعمیر اموال سرمایه‌ای	۰٫۲	۰٫۴	۱۶٫۲	۰٫۳	۰٫۲	۰٫۲	۰٫۲	۰٫۱	۰٫۱
	ارزش نهاده	۱۵۳٫۱	۳۰۷٫۳	۷۲۴۰٫۳	۲۴۵٫۹	۲۹۴۰٫۱	۲۹۴۰٫۱	۷۹٫۴	۱۴۵٫۳	۲۵۴٫۲
	ارزش ستانده	۳۴۷٫۱	۴۹۸٫۶	۳۵۷۵٫۹	۴۰۷٫۱	۳۷۰۸٫۷	۳۷۰۸٫۸	۳۱۹	۳۷۴٫۳	۳۸۰٫۴
	ارزش افزوده	۳۱۳٫۷	۴۳۱٫۷	۳۸۳۶٫۲	۳۱۱٫۳	۷۸۳	۷۸۳	۳۳۰	۳۱۸٫۲	۳۲۲٫۴
	I/O	۰٫۴	۰٫۸	۰٫۶	۰٫۴	۰٫۶	۰٫۶	۱٫۸	۰٫۴	۰٫۴
	ارزش مصرف	۲۰۵٫۹	۴۱۸٫۲	۱۸۳۸٫۷	۸۳٫۵	۱۹۴٫۷	۲۸۹۸٫۸	۹۸٫۲	۱۲۶٫۹	۲۴۹٫۸
	ارزش قطعات مصرفی برای ساخت	۰٫۲	۰٫۳	۱٫۵	۰٫۲	۰٫۱	۰٫۱	۰٫۳	۰٫۱	۰٫۱
	ارزش تولید	۲۰۹	۴۶۷٫۶	۲۵۲۸٫۵	۱۱۴٫۴	۲۰۰٫۷	۳۳۲۸٫۷	۱۳۵٫۵	۳۰۷٫۱	۲۴۷٫۴
	ارزش غذای طبخ شده توسط شاغلین	۰٫۳	۰٫۵	۸٫۲	۰٫۵	۰٫۳	۰٫۳	۰	۰٫۱	۰٫۱
	کالاهای موجود در انبار در اول فروردین	۳٫۷	۱۸٫۷	۲۰٫۳	۱٫۸	۱٫۳	۲۳٫۱	۱۱٫۷	۲٫۵	۲٫۹
	کالاهای موجود در انبار در اول اسفند	۵٫۷	۴٫۳	۹۳٫۷	۱٫۴	۲٫۶	۲۷٫۷	۵٫۱	۴٫۱	۳٫۹
	مواد خام و اولیه موجود در انبار فروردین	۴۷	۵۳٫۷	۵۸۵	۵۰٫۹	۴۸٫۲	۵۰٫۷	۵۳٫۴	۴۷٫۵	۴۷٫۹
۲۰	ارزش مصرف گازوئیل	۰٫۱	۰٫۱	۰٫۲	۰	۰٫۱	۰٫۱	۰٫۱	۰	۰
	ارزش مصرف گاز مایع	۰٫۱	۰	۰٫۸	۰	۰٫۱	۰	۰٫۲	۰٫۱	۰٫۱
	ارزش مصرف گاز طبیعی	۰٫۲	۰٫۶	۱۲٫۲	۰٫۱	۰٫۴	۰٫۴	۰٫۲	۰٫۱	۰٫۱
	ارزش مصرف بنزین	۰	۰٫۱	۰٫۹	۰	۰	۰	۰	۰	۰
	ساخت و تعمیر اموال سرمایه‌ای	۰٫۲	۰٫۵	۲٫۱	۰٫۳	۰٫۱	۰٫۲	۰٫۵	۰٫۱	۰٫۱
	ارزش نهاده	۲۰۵٫۵	۴۱۸٫۴	۱۸۵۱٫۶	۸۳٫۶	۲۹۴۰٫۱	۲۹۴۰٫۱	۹۸٫۱	۱۲۶٫۴	۲۴۹٫۴
	ارزش ستانده	۳۷۶٫۱	۵۵۲٫۸	۲۶۲۲٫۷	۳۳۲٫۴	۳۷۰۸٫۷	۳۷۰۸٫۸	۳۴۳٫۲	۴۳۸٫۵	۳۹۸٫۱
	ارزش افزوده	۳۱۸٫۹	۵۷۲٫۱	۸۵۳٫۴	۳۱۱٫۸	۷۸۳	۷۸۳	۳۳۰٫۴	۳۶۹٫۱	۳۱۱٫۵
	I/O	۰٫۶	۵٫۱	۰٫۷	۰٫۶	۰٫۶	۰٫۶	۳٫۱	۰٫۶	۰٫۶
	ارزش مصرف	۳۵۲٫۵	۲۶۴۰	۲۵۲۷۲۳۷	۳۵۳٫۷	۲۸۸٫۳	۴۱۲٫۵	۴۰۸۴٫۲	۳۰۹	۳۱۹٫۲
	ارزش قطعات مصرفی برای ساخت	۰٫۳	۰٫۴	۱۰۷۸٫۹	۰٫۴	۰٫۴	۰٫۴	۰٫۴	۰٫۴	۰٫۴
	ارزش تولید	۳۶۸٫۷	۴۱۵٫۶	۵۱۵۸۶۱٫۷	۳۹۱٫۶	۲۶۱٫۷	۶۵۰٫۸	۴۷۰۲٫۵	۳۵۴٫۶	۳۴۴٫۳
	ارزش غذای طبخ شده توسط شاغلین	۰٫۵	۰٫۶	۱۴۵	۰٫۶	۰٫۷	۰٫۵	۰٫۶	۰٫۵	۰٫۵
	کالاهای موجود در انبار در اول فروردین	۸٫۶	۲۷٫۱	۲۵۰۸۹۱۳٫۳	۴٫۹	۳٫۳	۱۶۸	۳۷٫۷	۷٫۶	۸٫۱
	کالاهای موجود در انبار در اول اسفند	۹٫۹	۱۷٫۷	۲۶۶۸۳۱٫۶	۴٫۸	۴٫۱	۲۵۵	۴۰٫۵	۹	۹٫۱
	مواد خام و اولیه موجود در انبار فروردین	۲۲٫۵	۷۹٫۲	۵۸۷۰۹۰٫۴	۳۸٫۹	۵۲٫۲	۸۹٫۲	۷۷٫۲	۲۰	۲۳٫۹
۳۵	ارزش مصرف گازوئیل	۰٫۱	۰٫۱	۲۱۱۵٫۱	۰	۰٫۱	۰	۰	۰	۰
	ارزش مصرف گاز مایع	۰٫۲	۰٫۲	۱۳۱٫۶	۰٫۲	۰٫۲	۰٫۲	۰٫۲	۰٫۲	۰٫۲
	ارزش مصرف گاز طبیعی	۰٫۴	۱٫۷	۳۰۲۵٫۶	۰٫۴	۰٫۷	۱٫۸	۱٫۹	۰٫۴	۰٫۳
	ارزش مصرف بنزین	۰٫۱	۰٫۱	۷۵۴٫۱	۰٫۱	۰٫۱	۰٫۱	۰٫۱	۰	۰
	ساخت و تعمیر اموال سرمایه‌ای	۰٫۶	۰٫۶	۲۶۸۹٫۴	۰٫۶	۰٫۵	۰٫۵	۰٫۶	۰٫۵	۰٫۵
	ارزش نهاده	۳۵۲٫۳	۲۶۳٫۳	۲۵۲۸۰۳۸	۳۵۴۰	۲۹۴۰٫۱	۲۹۴۰٫۱	۴۰۸۳۸	۳۰۸٫۷	۳۱۹
	ارزش ستانده	۴۸۱٫۹	۵۰۹٫۳	۵۱۶۳۸۴۲٫۷	۴۹۹٫۶	۳۷۰۸٫۷	۳۷۰۸٫۸	۴۷۰۶٫۹	۴۷۲٫۴	۴۶۴٫۳
	ارزش افزوده	۳۱۸٫۹	۴۸۳٫۶	۲۶۳۵۸۰۵٫۶	۳۱۳٫۹	۷۸۳	۷۸۳	۴۳۵۷٫۹	۳۲۳	۳۲۵٫۴
	I/O	۰٫۳	۴٫۷	۰٫۱	۰٫۲	۰٫۶	۰٫۶	۷٫۳	۰٫۱	۰٫۱

جدول ۴ مقادیر $RMSE$ متغیرهای اصلی را با استفاده از روش‌های مختلف جانهی شامل الگوریتم‌های $S.L$،

PMM، BLR، $MissRanger$، $Missforest$، $K-NN$، $hotdeck$، $CART$ و $irmi$[1] و درصدهای مختلف ایجاد بی‌پاسخی واحد را نشان می‌دهد. همانطور که ملاحظه می‌شود عملکرد روش BLR مناسب نبوده و برتری روش $S.L$ بر روش PMM مشخص است. برای سه متغیر مهم و متغیری که بیشترین مقدار $RMSE$ را دارند، اگر مقدار $RMSE$ به مجموع مقادیر متغیرهای موردبررسی قبل از ایجاد گم‌شدگی (جدول ۳) تقسیم شود، سهم این خطا از کل در آن متغیر مشخص می‌گردد. نتایج جدول ۵ نشان می‌دهد که:

جدول ۵. سهم $RMSE$ ارزش نهاده، ستانده و افزوده از مجموع متغیرها قبل از ایجاد گم‌شدگی (درصد)

درصد گم‌شدگی	متغیر	S.L	PMM	BLR	KNN	irmi	Hotdeck	CART	MissForest	MissRanger
	ارزش نهاده	۰/۴	۰/۷	۱۷/۷	۰/۶	۷/۲	۷/۲	۰/۲	۰/۴	۰/۶
۱۰	ارزش ستانده	۲/۱	۳/۰	۲۶/۳	۰/۷	۶/۷	۶/۷	۰/۶	۰/۷	۰/۷
	ارزش افزوده	۲/۱	۳/۰	۲۶/۳	۲/۱	۵/۴	۵/۴	۲/۳	۲/۲	۲/۲
	ارزش نهاده	۰/۵	۱	۴/۵	۰/۲	۷/۲	۷/۲	۰/۲	۳/۱	۰/۶
۲۰	ارزش ستانده	۰/۷	۱	۴/۷	۰/۶	۶/۷	۶/۷	۰/۶	۰/۸	۰/۷
	ارزش افزوده	۲/۲	۳/۹	۵/۸	۲/۱	۵/۴	۵/۴	۲/۳	۲/۵	۲/۱
	ارزش نهاده	۰/۹	۰/۶	۶/۵	۸/۶	۷/۲	۷/۲	۱۰	۰/۸	۰/۸
۳۵	ارزش ستانده	۰/۹	۰/۹	۹۲۸۶	۰/۹	۶/۷	۶/۷	۸/۵	۰/۸	۰/۸
	ارزش افزوده	۲/۲	۳/۳	۱۸۰۳۹	۲/۱	۵/۴	۵/۴	۲۹/۸	۲/۲	۲/۲

- روش SL برتری کامل نسبت به روش BLR دارد.
- روش SL نسبت به روش PMM در اکثریت موارد برتری دارد.
- روش SL نسبت به روش KNN در برخی موارد برتری خوبی داشته است و در برخی موارد عملکرد مشابه این روش داشته است.
- روش SL نسبت به روش‌های $irmi$ و $Hotdeck$ همیشه عملکرد بهتری داشته است.
- روش SL نسبت به روش‌های $missForest$ و $missRanger$ عملکرد مشابهی داشته است.
- روش SL نسبت به روش $CART$ در درصدهای گم‌شدگی بالا (گم‌شدگی ۳۵ درصد) عملکرد بهتری داشته است ولی در درصدهای گم‌شدگی پایین، عملکرد ضعیف‌تری داشته است.

۵ بحث و نتیجه‌گیری

روش ابریادگیرنده در حالتی که داده‌ها تُنُک و دارای مقادیر صفر زیادی در متغیرها باشد، قادر به جانهی داده‌های گم‌شده نیست. عملکرد الگوریتم ابریادگیرنده به مکانیسم ایجاد گم‌شدگی داده‌ها وابسته است و نتایج نشان داده است که عملکرد این روش در مکانیسم گم‌شدگی تصادفی نسبت به سایر روش‌ها برتری دارد. به نظر می‌رسد در متغیرهایی که میانگین و انحراف معیار داده‌ها بسیار بالا است، عملکرد روش‌هایی مانند $CART$ بهبود می‌یابد و عملکرد

[1] Iterative robust model-based imputation

$SuperMICE$ ضعیف‌تر می‌گردد. با توجه به محاسبات بالا و زمان‌بر بودن اجرای الگوریتم $SuperMICE$، بهبود الگوریتم برای کاهش زمان محاسبات الگوریتم‌های ترکیبی جدید پیشنهاد می‌شود.

تقدیر و تشکر

نویسندگان مقاله مراتب قدردانی و سپاس خود را از پیشنهادات ارزنده داوران، سردبیر و ویراستار محترم مجله که باعث ارائه بهتر و افزایش سطح کیفی مقاله شده است، کمال قدردانی و تشکر را دارند.

مراجع

Aerts, M., Claeskens, G., Hens, N., and Molenberghs, G. (2002), Local Multiple Imputation, Biometrika, **89**(2), 375-388.

Alwateer, M., Atlam, E. S., Abd El-Raouf, M. M., Ghoneim, O. A., and Gad, I. (2024), Missing Data Imputation: A Comprehensive Review, Journal of Computer and Communications, **12**(11), 53-75.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., and Tabona, O. (2021), A Survey on Missing Data in Machine Learning, Journal of Big Data, **8**(1), 1-37.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007), How Many Imputations Are Really Needed? Some Practical Clarifications of Multiple Imputation Theory, Prevention Science, 8, 206-213.

Laqueur, H. S., Shev, A. B., and Kagawa, R. M. (2022), SuperMICE: An Ensemble Machine Learning Approach to Multiple Imputation by Chained Equations, American Journal of Epidemiology, **191**(3), 516-525.

Little, R. J. (1988), Missing-data Adjustments in Large Surveys, Journal of Business and Economic Statistics, **6**(3), 287-296.

Marshall, A., Altman, D. G., Royston, P., and Holder, R. L. (2010), Comparison of Techniques for Handling Missing Covariate Data Within Prognostic Modelling Studies: A Simulation Study, BMC Medical Research Methodology, **10**(1), 1-16.

Nadaraya, E. A. (1964), On Estimating Regression. Theory of Probability and Its Applications, **9**(1), 141-142.

Quinlan, J. R. (1987), Simplifying Decision Trees, International Journal of Man-Machine Studies, **27**(3), 221-234.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J.,and Solenberger, P. (2001), A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models, Survey Methodology, **27**(1), 85-96.

Rubin, D. B. Multiple Imputation for Nonresponse in Surveys. Toronto, ON, Canada: John Wiley and Sons, Inc.; 2004.

Stekhoven, D. J., & Bühlmann, P. (2012), MissForest—non-Parametric Missing Value Imputation for Mixed-type Data, Bioinformatics, **28**(1), 112-118.

Van Buuren, S. (2018), Flexible Imputation of Missing Data, CRC press.

Tiwaskar, S., Rashid, M., and Gokhale, P. (2024), Impact of Machine Learning-based Imputation Techniques on Medical Datasets-a Comparative Analysis, *Multimedia Tools and Applications*, DOI:10.1007/s11042-024-19103-0.

Van Buuren, S., and Groothuis-Oudshoorn, K. (2011), mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, **45**, 1-67.

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007), Super Learner, Statistical Applications in Genetics and Molecular Biology, **6**(1), DOI:10.2202/1544-6115.1309.