



Missing Data Imputation Using Machine Learning: A Case Study on Imputing Maternal Age in Birth Records Registered in ۲۰۲۳

Sedigheh Omidvar Shalmani^{۱✉}, Alireza Sajedi^۲

۱. Corresponding Author, PhD is Statistics, National Population Monitoring Center, Civil Registration Organization, Tehran, Iran. E-mail: s.o۲۶۷@yahoo.com

۲. PhD is Statistics, National Population Monitoring Center, Civil Registration Organization, Tehran, Iran. E-mail: ali_sa۴۴@yahoo.com

ARTICLE INFO

Article type:

Research Article

Keywords:

Machin Learning, Missing Data Imputation, Missing Data Mechanism, Mother Age, Random Forest.

ABSTRACT

The accuracy of analyses and results can be substantially influenced by the absence of data in demographic studies. The objective of this investigation is to assess the efficacy of a variety of imputation techniques in ۲۰۲۳ to account for the absence of the maternal age variable in birth records. The efficacy of classical and machine learning-based imputation methods is evaluated in this study, taking into account varying missing data mechanisms and proportions. The results suggest that traditional methods are outperformed by machine learning-based methods, particularly the missForest algorithm, in terms of reducing imputation error. Nevertheless, traditional approaches, such as multiple imputation using chained equations, can also produce satisfactory results under specific circumstances. The significance of selecting an appropriate imputation method based on the dataset characteristics and missing data pattern is underscored by the results of this study.

Cite this article: Omidvar Shalmani, S., & Sajedi, A. R. (۲۰۲۳). Missing Data Imputation Using Machine Learning: A Case Study on Imputing Maternal Age in Birth Records Registered in ۲۰۲۳. *Population Journal*, 30(۱۲۳), ۱-۱۶.

© The Author(s).



پروہشگاہ علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

Extended Abstract

Introduction

A persistent challenge in the production of official statistics is the problem of missing data. Missingness occurs when values for certain variables are not recorded for a subset of, or all respondents.

Within vital registration systems, a notable example is the omission of the mother's date of birth on live birth records. Maternal age is a fundamental variable for demographic analysis and is indispensable for estimating key fertility indicators, most notably the Total Fertility Rate (TFR). The absence of precise maternal age information can introduce bias, distort fertility estimates, and lead to erroneous conclusions, ultimately undermining evidence-based decision-making in demographic and health policy.

Under varying assumptions regarding the proportion and mechanism of missingness, the objective of this study is to assess the implications of various imputation methods for maternal age.

Methods and Data

The dataset utilized in this study encompassed all registered births during the Persian calendar year ۱۴۰۲ (۲۰۲۳–۲۰۲۴ Gregorian year), obtained from the Population Information Database maintained by National Organization for Civil Registration of Iran. The analysis incorporated six variables: infant gender, birth order, province of birth occurrence, paternal age at birth, urban–rural classification of the birth location, and the primary variable of interest—maternal age at birth.

Six imputation methodologies were implemented to address absent values in maternal age data. These encompassed machine learning approaches, such as K-Nearest Neighbors (KNN), MiceForest, and MissForest, as well as statistical techniques, such as straightforward imputation, Expectation–Maximization (EM), and Multiple Imputation by Chained Equations (MICE). The investigation assessed the impact of different missingness mechanisms and proportions on the performance of imputation. Three distinct missing data mechanisms—Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR)—were systematically simulated across varying missingness proportions in relation to maternal age.

Findings

The results of this study demonstrate that, across varying mechanisms and proportions of missingness, machine learning–based imputation methods consistently outperform traditional approaches. Among the conventional techniques, the MICE method exhibited markedly superior performance compared to simpler strategies such as mode imputation or the EM algorithm. Within the category of machine learning approaches, the performance of the KNN method was found to be closely aligned with that of MissForest; nevertheless, MissForest consistently proved to be the most effective method for imputing missing values of maternal age. Moreover, while the performance of MICE (a traditional method) was relatively comparable to that of MiceForest (a machine learning–based method), the latter demonstrated a clear and consistent advantage. Finally, the findings suggest that the effectiveness of machine learning–based imputation techniques is sensitive to both the underlying mechanism and the proportion of missingness.

Conclusion and Discussion

In this study, the performance of various imputation methods for handling missing values in maternal age within the Persian calendar year ۱۴۰۲ (۲۰۲۳–۲۰۲۴ Gregorian year) birth registration data was evaluated under different mechanisms and proportions of missingness. The findings revealed that machine learning–based approaches, particularly the MissForest algorithm, consistently outperformed traditional methods in reducing imputation error, as measured by RMSE. Among the traditional methods, Multiple Imputation by Chained Equations (MICE) also demonstrated favorable performance in many cases, with results approaching those of machine learning techniques. The evidence suggests that MICE may serve as a suitable alternative when the implementation of advanced methods is not feasible.

Analysis of method performance in relation to missingness mechanisms and proportions further indicated that machine learning–based approaches performed particularly well under MNAR with lower levels of missingness (۰.۰۱–۰.۱۵). However, as the proportion of missingness increased, these methods exhibited superior performance under completely random missingness.

From a practical perspective, this study underscores the importance of selecting an imputation method based on the type of missingness mechanism, the rate of missingness, and the characteristics of the dataset. The use of advanced approaches such as MissForest can enhance the accuracy of analyses and ensure more reliable results.



کاربرد یادگیری ماشین در جانهای داده‌های گمشده در مطالعات جمعیتی: مطالعه موردی

جانهای سن مادر در ولادت‌های ثبت‌شده سال ۱۴۰۲

صدیقه امیدوار شلمانی^۱✉، علیرضا ساجدی^۲

۱. نویسنده مسئول، دکتری آمار، دفتر آمار و اطلاعات جمعیتی و مهاجرت، مرکز رصد جمعیت کشور، سازمان ثبت احوال کشور، تهران، ایران. رایانامه:

s.o267@yahoo.com

۲. دکتری آمار، دفتر آمار و اطلاعات جمعیتی و مهاجرت، مرکز رصد جمعیت کشور، سازمان ثبت احوال کشور، تهران، ایران. رایانامه: ali_sa94@yahoo.com

چکیده

اطلاعات مقاله

نوع مقاله:

علمی-پژوهشی

وجود داده‌های گمشده در مطالعات جمعیتی می‌تواند دقت تحلیل‌ها و نتایج را به‌طور قابل توجهی کاهش دهد. این پژوهش با هدف ارزیابی کارایی روش‌های مختلف جانهای داده‌های گمشده برای متغیر سن مادر در داده‌های ولادت سال ۱۴۰۲ انجام شده است. در این مطالعه، با در نظر گرفتن سازوکارها و نسبت‌های گمشدگی مختلف کارایی روش‌های کلاسیک و روش‌های مبتنی بر یادگیری ماشین بررسی شده است. نتایج مطالعه نشان می‌دهد روش‌های مبتنی بر یادگیری ماشین، به‌ویژه الگوریتم missForest، در کاهش خطای جانهای نسبت به روش‌های سنتی عملکرد بهتری دارند. با این حال، روش‌های سنتی مانند جانهای چندگانه با معادلات زنجیره‌ای نیز در شرایط خاص نتایج مطلوبی ارائه داده‌اند. کارایی هر روش به نوع داده‌ها، نسبت گمشدگی و همچنین سازوکار گمشدگی داده‌ها بستگی دارد. یافته‌های این پژوهش بر اهمیت انتخاب روش جانهای مناسب براساس الگوی گمشدگی و ویژگی‌های داده‌ها تأکید دارد.

کلیدواژه‌ها:

جانهای داده‌های گمشده، جنگل تصادفی، سن مادر، مکانیسم گمشدگی داده‌ها، یادگیری ماشین.

استناد: امیدوار شلمانی، صدیقه و ساجدی، علیرضا (۱۴۰۲). کاربرد یادگیری ماشین در جانهای داده‌های گمشده در مطالعات جمعیتی: مطالعه موردی جانهای سن مادر در ولادت‌های

ثبت‌شده سال ۱۴۰۲. فصلنامه جمعیت، ۳۰ (۱۲۳)، ۱-۱۶.

© نویسندگان.

مقدمه

یکی از چالش‌های رایج در فرایند گردآوری آمارهای رسمی، وجود داده‌های گمشده^۱ است. اصطلاح گمشدگی به حالتی اشاره دارد که در آن برخی یا تمامی متغیرها برای تعدادی از واحدهای پاسخ‌دهنده ثبت نمی‌شوند. عوامل متعددی می‌توانند منجر به گمشدگی داده‌ها شوند، از جمله اشتباهات در مراحل گردآوری، اندازه‌گیری و ورود اطلاعات، ورود دستی داده‌ها، عدم تمایل پاسخ‌دهندگان به همکاری یا اجتناب از پاسخ به برخی سؤالات، ترک مطالعه پیش از اتمام آن، مشکلات سخت‌افزاری یا نرم‌افزاری، نقص در سیستم‌های جمع‌آوری داده‌ها و اختلال در زمان انتقال اطلاعات (متی و همکاران، ۲۰۱۱؛ امانوئل و همکاران، ۲۰۲۱؛ آستین و همکاران، ۲۰۲۱؛ جگادسیواری و همکاران، ۲۰۲۲؛ ریزوی و همکاران، ۲۰۲۳؛ لی و همکاران، ۲۰۲۴).

شکاف‌های موجود در داده‌ها به دلیل گمشدگی می‌تواند پیامدهای متعددی داشته باشد، از جمله کاهش بازنمایی نمونه، افزایش خطا یا ابهام در تفسیر داده‌ها، ایجاد اریبی در تحلیل‌های آماری، کاهش توان آماری، دقت پایین در محاسبه شاخص‌های مربوطه، تأثیر منفی بر عملکرد مدل‌های پیش‌بینی، کاهش دقت پیش‌بینی‌ها، مدل‌سازی ناکارآمد که در نهایت منجر به تولید نتایج نامعتبر می‌شوند (هاول، ۲۰۰۷؛ امانوئل و همکاران، ۲۰۲۱؛ فؤاد و همکاران، ۲۰۲۱؛ ریزوی و همکاران، ۲۰۲۳؛ لی و همکاران، ۲۰۲۴). برای مدیریت داده‌های گمشده دو روش وجود دارد؛ روش اول حذف سطرها یا ستون‌های حاوی مقادیر گمشده است. این روش ساده به‌طور گسترده‌ای برای مدیریت مقادیر گمشده در طول پیش‌پردازش داده‌ها استفاده شده است (وانگ و همکاران، ۲۰۲۲b). اگرچه این روش در شرایطی که مجموعه داده‌ها بسیار بزرگ بوده و نرخ گمشدگی پایین است، قابل استفاده است (لین و همکاران، ۲۰۱۷)، اما در داده‌های با نرخ گمشدگی بالا کارایی چندانی ندارد. در چنین مواردی، ممکن است اطلاعات ارزشمندی که در داده‌های گمشده نهفته است، از دست برود (فؤاد و همکاران، ۲۰۲۱؛ جگادسیواری و همکاران، ۲۰۲۲). همچنین در مجموعه داده‌های کوچک، حذف موارد گمشده می‌تواند نتایج تحلیل را تحت تأثیر قرار دهد (ریزوی و همکاران، ۲۰۲۳). روش دوم برای مدیریت داده‌های گمشده، جایگزینی آن‌ها با یک مقدار مناسب است که به آن جانهی داده‌های گمشده^۲ گفته می‌شود (فؤاد و همکاران، ۲۰۲۱؛ لی و همکاران، ۲۰۲۴). برای جانهی داده‌های گمشده، رویکردهای متعددی مورد بررسی قرار گرفته است که از روش‌های آماری کلاسیک تا الگوریتم‌های یادگیری ماشین را شامل می‌شوند. روش‌های آماری مانند میانگین، مد، میانه، درونیابی خطی، رگرسیون و روش‌های یادگیری ماشین مانند K-نزدیک‌ترین همسایگی، جنگل تصادفی، شبکه عصبی و درخت تصمیم از جمله این رویکردها هستند. با این حال، هیچ نتیجه قطعی درخصوص بهترین مدل جانهی وجود ندارد؛ زیرا کارایی هر روش به نوع داده‌ها، نسبت و سازوکار گمشدگی داده‌ها بستگی دارد (لین و همکاران، ۲۰۱۷؛ آستین و همکاران، ۲۰۲۱؛ ریزوی و همکاران، ۲۰۲۳).

یکی از موارد گمشدگی در ولادت‌های ثبتی، عدم درج تاریخ تولد مادر در ثبت رویداد ولادت است که مرتبط با برخی فرایندهای قانونی از جمله تنظیم سند براساس رأی ابطال و صدور، موارد مشمول بند ۶ ماده ۹۷۶ قانون مدنی و مواردی از ماده ۴۵ قانون ثبت‌احوال است. در این موارد برای جلوگیری از طولانی‌شدن فرایند ثبت، رویداد ولادت با عدم درج برخی اقلام اطلاعاتی به ثبت می‌رسد. در این راستا، در این مقاله، شش روش جانهی مبتنی بر روش‌های آماری کلاسیک (روش‌های جانهی ساده^۳، امیدگیری - بیشینه‌سازی^۴ (EM) و جانهی چندگانه با معادلات زنجیره‌ای^۵ (MICE) و یادگیری ماشین (K-نزدیک‌ترین همسایگی^۶، MiceForest و MissForest) برای داده‌های مربوط به موارد نامشخص سن مادر از مجموعه داده‌های ولادت سال ۱۴۰۲ انتخاب شده است. همچنین برای ارزیابی روش‌های جانهی، از نرم‌افزار پایتون^۷ استفاده شده است.

سن مادر به‌عنوان یکی از متغیرهای کلیدی در تحلیل داده‌های ولادت، نقش مهمی در محاسبه شاخص‌های باروری به‌ویژه

۱. Missing data

۲. Missing data imputation

۳. Simple

۴. Expectation-Maximization

۵. Multiple Imputation by Chained Equations

۶. K-Nearest Neighbors

۷. Python

میزان باروری کل دارد. نبود اطلاعات دقیق در مورد سن مادر ممکن است منجر به محاسبات نادرست یا ناقص و ارائه نتایج گمراه کننده در مورد این شاخص ها شود که می تواند تصمیم گیری های نادرستی را در زمینه سیاست های جمعیتی و بهداشتی به همراه داشته باشد. این مقاله تلاش کرده است تا با در نظر گرفتن مقادیر مختلف برای نسبت و سازوکارهای گمشدگی، تأثیر روش های جانپی بر متغیر سن مادر را مورد ارزیابی قرار دهد.

پیشینه تحقیق

مطالعات متعددی به بررسی و توسعه روش های جانپی داده های گمشده پرداخته اند. روبین (۱۹۷۶) برای اولین بار سازوکارهای گمشدگی داده ها را به سه دسته گمشدگی کاملاً تصادفی^۱ (MCAR) گمشدگی تصادفی^۲ (MAR) و گمشدگی غیر تصادفی^۳ (MNAR) تقسیم بندی کرد و این چارچوب به عنوان پایه ای برای تحقیقات بعدی در این حوزه مورد استفاده قرار گرفت. او تأکید کرد که شناخت سازوکار گمشدگی برای انتخاب روش مناسب جانپی ضروری است؛ زیرا هر سازوکار گمشدگی تأثیر متفاوتی بر نتایج تحلیل ها دارد.

در سال های بعد، روش های آماری سنتی مانند جانپی ساده، جانپی EM و جانپی چندگانه^۴ توسعه یافتند. روش جانپی ساده شامل جایگزینی مقادیر گمشده با میانگین، میانه یا مد است که در مطالعات اولیه مورد استفاده قرار گرفته اند (لیتل و روبین، ۲۰۰۲). هر چند این روش ها ساده و سریع هستند، اما اغلب منجر به تخمین های نادرست و کاهش تغییرپذیری داده ها می شوند. الگوریتم EM به عنوان یک روش تکرارشونده برای برآورد پارامترها در حضور داده های گمشده توسط دمپستر و همکاران (۱۹۷۷) پیشنهاد و به دلیل توانایی در مدل سازی روابط پیچیده بین متغیرها، در بسیاری از تحقیقات به کار گرفته شد. به عنوان مثال، اسکفر (۱۹۹۷) از الگوریتم EM برای جانپی داده های چندمتغیره با توزیع نرمال استفاده کرد و نشان داد که این روش در مقایسه با روش های ساده تر مانند جانپی میانگین برآوردهای دقیق تری ارائه می دهد. همچنین، گراهام و همکاران (۲۰۰۷) نشان دادند الگوریتم EM در جانپی داده های گمشده با الگوی گمشدگی تصادفی نتایج قابل اطمینان تری تولید می کند. این مطالعات نشان می دهند که الگوریتم EM به دلیل رویکرد تکراری و مبتنی بر احتمال، یکی از روش های مؤثر برای جانپی داده های گمشده در تحقیقات مختلف بوده است.

جانپی چندگانه توسط روبین (۱۹۸۷) معرفی شد. این روش با تولید چندین مجموعه داده جانپی شده و ترکیب نتایج حاصل از تحلیل های انجام شده بر روی هر مجموعه داده، امکان محاسبه دقیق تر خطای استاندارد و فواصل اطمینان را فراهم می سازد. رویکرد اصلی این روش، در نظر گرفتن عدم قطعیت ذاتی ناشی از فرایند جانپی است که آن را از روش های تک مقداری متمایز می کند. مطالعات متعددی مانند آزور و همکاران (۲۰۱۱) و وایت و همکاران (۲۰۱۱) کارایی این روش را مورد بررسی قرار داده و نشان دادند که روش جانپی چندگانه تحت سازوکارهای مختلف گمشدگی داده ها بهینه عمل می کنند.

با گسترش الگوریتم های پیشرفته تر مانند MICE که توسط ون بورن و گروتویس-اودشوین (۲۰۱۱) توسعه یافت، امکان مدلسازی جداگانه برای هر متغیر با توجه به نوع آن فراهم شد که این امر تحولی در کاربردهای عملی این روش ایجاد کرده است. روش های جانپی چندگانه کاربردهای متعددی در حوزه های جمعیتی، تحقیقات اپیدمیولوژیک، پزشکی و پژوهش های اقتصادی-اجتماعی داشته که نشان دهنده آن است که جانپی چندگانه، به ویژه در ترکیب با روش های پیشرفته تر، می تواند تا حد زیادی از سوگیری های ناشی از داده های گمشده جلوگیری کند (آستین و همکاران، ۲۰۲۱؛ بیسلی و همکاران، ۲۰۲۱؛ وانگ و همکاران، ۲۰۲۲؛ لی و همکاران، ۲۰۲۴).

با پیشرفت فناوری و گسترش روش های یادگیری ماشین، رویکردهای نوینی برای جانپی داده های گمشده توسعه یافتند. در این میان، تروییاتسکایا و همکاران (۲۰۰۱) روش K را به عنوان نزدیک ترین همسایگی را به عنوان یک راهکار ساده اما مؤثر برای

۱. Missing Completely at Random

۲. Missing at Random

۳. Missing Not at Random

۴. Multiple imputation

جانهی معرفی کردند. این روش با بهره‌گیری از مفهوم شباهت بین نمونه‌های داده، مقادیر گمشده را پیش‌بینی می‌کند. مزیت اصلی این روش، انعطاف‌پذیری بالا و عدم نیاز به فرضیات پارامتری درباره توزیع داده‌ها است که باعث شده در حوزه‌های متنوعی از جمله علوم زیستی تا پژوهش‌های اجتماعی به‌طور گسترده مورد استفاده قرار گیرد (جرز و همکاران، ۲۰۱۰؛ دی سیلوا و پررا، ۲۰۱۶؛ پوجیانته و همکاران، ۲۰۱۹؛ ریزوی و همکاران، ۲۰۲۳؛ سانیا و همکاران، ۲۰۲۵). با این حال، مطالعات متعددی به محدودیت‌های این روش نظیر حساسیت به انتخاب تعداد همسایگی (K) و هزینه محاسباتی بالا در مجموعه داده‌های بزرگ اشاره کرده‌اند (باتیستا و مونارد، ۲۰۰۳؛ رحمان و اسلام، ۲۰۱۳؛ لیو و همکاران، ۲۰۱۵؛ عابدین و اسماعیل، ۲۰۲۱). در پاسخ به این چالش‌ها، پژوهشگران روش‌های بهبودیافته‌ای را ارائه کرده‌اند که کارایی الگوریتم‌های مبتنی بر روش KNN را افزایش داده‌اند (امانوئل و همکاران، ۲۰۲۱).

در دهه‌های اخیر، الگوریتم‌های مبتنی بر جنگل تصادفی^۱ به دلیل توانایی در مدیریت داده‌های پیچیده و غیرخطی مورد توجه پژوهشگران قرار گرفته‌اند (استخوون و بولمان، ۲۰۱۲؛ والچی و همکاران، ۲۰۱۳؛ داو و همکاران، ۲۰۱۴؛ تانگ و ایشواران، ۲۰۱۷؛ الصابر و همکاران، ۲۰۲۱؛ پترزینی و همکاران، ۲۰۲۱؛ وانگ و همکاران، ۲۰۲۲a و لی و همکاران، ۲۰۲۴). از جمله این روش‌ها الگوریتم missForest است که توسط استخوون و بوهلمن (۲۰۱۲) توسعه یافت. این الگوریتم با به‌کارگیری رویکرد تکراری جنگل تصادفی، مقادیر گمشده را با دقت بالایی پیش‌بینی می‌کند. مطالعات متعددی نشان داده‌اند که missForest در مقایسه با روش‌های سنتی جانهی، عملکرد بهتری در داده‌های چندبعدی دارد (والچی و همکاران، ۲۰۱۳؛ تانگ و ایشواران، ۲۰۱۷؛ الصابر و همکاران، ۲۰۲۱؛ پترزینی و همکاران، ۲۰۲۱؛ وانگ و همکاران، ۲۰۲۲a). با این حال در کاربرد این روش، برخی محدودیت‌ها نظیر وجود اربیی در برآوردها به دلیل داده‌های با گمشدگی غیرتصادفی یا همبستگی بالا بین متغیرها گزارش شده است (شاه و همکاران، ۲۰۱۴؛ هانگ و لین، ۲۰۲۰؛ آراکری و همکاران، ۲۰۲۳). همچنین روش MICEForest به‌عنوان ترکیبی از روش جانهی چندگانه (MICE) و جنگل تصادفی توسعه یافته که با ادغام قابلیت‌های این دو روش، انعطاف‌پذیری و دقت بیشتری را در مقایسه با روش‌های پیشین ارائه می‌دهد (داو و همکاران، ۲۰۱۴).

در حوزه مطالعات جمعیتی، داده‌های گمشده به‌ویژه در متغیرهای کلیدی مانند سن مادر، چالش‌های جدی برای محاسبه شاخص‌هایی نظیر نرخ باروری کل ایجاد می‌کند. امانوئل و همکاران (۲۰۲۱) نشان دادند که استفاده از روش‌های پیشرفته جانهی، به‌ویژه زمانی که نسبت گمشدگی بالا باشد، می‌تواند دقت تحلیل‌های جمعیتی را بهبود بخشد. همچنین لی و همکاران (۲۰۲۴) در مطالعه‌ای بر روی داده‌های ثبتی به کارایی بالای روش‌های یادگیری ماشین در مقایسه با روش‌های سنتی اشاره و بر اهمیت انتخاب روش مناسب با توجه به الگوی گمشدگی تأکید کرده‌اند.

روش‌شناسی

سازوکار گمشدگی داده‌ها: برای تعریف سازوکارهای گمشدگی داده‌ها، فرض کنید X ماتریس کل مجموعه داده‌ها باشد که شامل p ستون و n سطر است ($X = [X_{ij}; i = 1, \dots, n, j = 1, \dots, p]$). ماتریس نشانگر $R = [R_{ij}; i = 1, \dots, n, j = 1, \dots, p]$ به صورت زیر تعریف می‌شود:

$$R_{ij} = \begin{cases} X_{ij} \text{ گمشده است} & ۱ \\ \text{در غیر این صورت} & ۰ \end{cases}$$

در این صورت، R ماتریس الگوی گمشدگی X نامیده می‌شود. براساس این تعریف، ماتریس X را می‌توان به دو بخش داده‌های مشاهده شده و گمشده به صورت $X = (X_{obs}, X_{mis})$ تجزیه کرد. به منظور تعریف سازوکارهای گمشدگی مختلف، فرض کنید X و R متغیرهای تصادفی با توابع توزیع احتمال P_X و P_R باشند و پارامتر توزیع گمشدگی P_R را با ϕ نشان می‌دهیم.

^۱. Random forest

گمشدگی کاملاً تصادفی (MCAR): در این حالت احتمال گمشدگی به هیچ یک از مقادیر مشاهده شده و گمشده بستگی ندارد. به عبارت دیگر:

$$P_R(R|X; \phi) = P_R(R) \quad \forall \phi$$

به عنوان مثال، فرض کنید در یک نظرسنجی برخی از پاسخ دهندگان به طور تصادفی به یک یا چند سؤال پاسخ ندهند. در این حالت، عدم پاسخ دهی به هیچ ویژگی خاصی از پاسخ دهندگان یا سوالات بستگی ندارد و کاملاً تصادفی است.

گمشدگی تصادفی (MAR): در این حالت احتمال گمشدگی تنها به مقادیر مشاهده شده (یعنی X_{obs}) بستگی دارد. در این صورت:

$$P_R(R|X; \phi) = P_R(R|X_{obs}; \phi) \quad \forall \phi, X_{mis}$$

در این وضعیت، احتمال گمشدگی یک مقدار خاص به مقادیر سایر متغیرها بستگی دارد؛ بنابراین، یک عامل مشترک را می توان در تمام مشاهداتی که مقادیر گمشده دارند شناسایی کرد. به عنوان مثال فرض کنید در یک مطالعه پزشکی، برخی از بیماران به دلیل سن بالا یا وضعیت سلامتی ضعیف تر نتوانند در برخی از آزمایش ها شرکت کنند. در این حالت، گمشدگی داده ها به ویژگی های مشاهده شده (سن یا وضعیت سلامتی) بستگی دارد اما به مقادیر گمشده وابسته نیست.

گمشدگی غیر تصادفی (MNAR): گمشدگی غیر تصادفی زمانی رخ می دهد که احتمال گمشدگی به مقادیر گمشده و مشاهده شده بستگی داشته باشد. به عبارت دیگر:

$$P_R(R|X; \phi) = P_R(R|X_{obs}, X_{mis}; \phi) \quad \forall \phi, X_{obs}, X_{mis}$$

در این روش معمولاً مدیریت مقادیر گمشده غیر ممکن است؛ زیرا به داده های مشاهده نشده بستگی دارد. به عنوان مثال، در یک نظرسنجی در مورد درآمد، افرادی که درآمد بالاتری دارند ممکن است تمایلی به افشای درآمد خود نداشته باشند، و این احتمال گمشدگی در مردان بیشتر از زنان است. در این حالت گمشدگی داده ها هم به متغیر جنسیت و هم به مقادیر گمشده (درآمد بالا) بستگی دارد.

برای ارزیابی عملکرد نوع سازوکار گمشدگی داده ها و انواع روش های جانیی داده های گمشده، لازم است مقادیر گمشده شبیه سازی شوند. اگر داده ها از قبل شامل مقادیر گمشده باشند، به دلیل عدم دسترسی به مقادیر واقعی، ابتدا باید مقادیر گمشده نادیده گرفته شوند. سپس با استفاده از سازوکارهای گمشدگی، مقادیر معلوم در داده ها حذف و در نهایت از روش های مختلف برای جانیی داده ها استفاده شود. در این مقاله برای تولید داده های گمشده تحت سه سازوکار گمشدگی مختلف از کدهای ارائه شده توسط موزلک و همکاران (۲۰۲۰) استفاده شده است.

روش های جانیی داده های گمشده

۱. جانیی ساده (Simple)

در این روش، مقادیر گمشده با یک مقدار ثابت یا آماره های محاسبه شده از بخش غیر گمشده هر ستونی که داده های گمشده در آن قرار دارند، جایگزین می شوند. این آماره ها می توانند شامل میانگین یا میانه برای متغیرهای کمی و مد برای متغیرهای کیفی باشند.

روش جانیی ساده به دلیل سادگی در اجرا، در بسیاری از مطالعات به عنوان یک روش مرجع متداول شناخته می شود. با این حال، این روش به دلیل نادیده گرفتن همبستگی های بین ویژگی ها در مجموعه داده هایی با روابط پیچیده، ممکن است نتایج نادرست یا اریب ایجاد کند. همچنین در دنیای داده های با حجم بالا، این روش ممکن است عملکرد ضعیفی داشته و در نتیجه برای پیاده سازی روی چنین مجموعه های داده ای ناکافی باشد (خان و هوک، ۲۰۲۰؛ فؤاد و همکاران، ۲۰۲۱؛ لی و همکاران، ۲۰۲۴).

۲. جانیی امیدگیری-بیشینه سازی (EM)

الگوریتم امیدگیری-بیشینه‌سازی (EM) یک ابزار قدرتمند و روش تکراری برای جانی داده‌های گمشده است. این الگوریتم از رویکرد «جایگزینی، برآورد و تکرار تا همگرایی» استفاده می‌کند. هر تکرار شامل دو مرحله امیدگیری و بیشینه‌سازی است. در مرحله امیدگیری، امید ریاضی مقادیر گمشده براساس داده‌های مشاهده‌شده محاسبه می‌شود. در مرحله بیشینه‌سازی، از داده‌های مشاهده‌شده و مقادیر برآوردشده مقادیر گمشده که در مرحله امیدگیری به‌دست آمده‌اند، برای بیشینه‌سازی توزیع احتمال داده‌های کامل (شامل مقادیر گمشده و مشاهده‌شده) استفاده می‌شود (امانوئل و همکاران، ۲۰۲۱).

مطالعات متعدد نشان داده‌اند که بسته به نوع داده‌های مورد مطالعه، روش EM عملکرد بهتری نسبت به روش‌های جانی ساده نظیر حذف لیستی^۱ و رگرسیونی دارد (روبین و همکاران، ۲۰۰۷). با این حال، این روش به‌ویژه در مجموعه داده‌های با ابعاد بالا نیازمند محاسبات ماتریسی پیچیده و پرهزینه است (دلانو و همکاران، ۲۰۱۲). علاوه‌براین، به دلیل اینکه این روش از کل اطلاعات داده‌ها برای جانی مقادیر گمشده استفاده می‌کند فقط برای مجموعه داده‌هایی که همبستگی قوی بین ویژگی‌ها دارند، مناسب است (دب و لئو، ۲۰۱۶).

۳. جانی چندگانه با معادلات زنجیره‌ای (MICE)

یکی از تکنیک‌های آماری مدیریت داده‌های گمشده، روش جانی چندگانه است که در آن به جای جانی ساده، از مدل‌های آماری مختلف برای جانی یک مقدار گمشده استفاده می‌شود. نتایج به‌دست‌آمده از هریک از مجموعه داده‌های جانی‌شده به‌طور مجزا تحلیل و سپس با استفاده از روش‌های ترکیبی مناسب با یکدیگر ترکیب می‌شوند (روبین، ۱۹۸۷). این رویکرد عدم قطعیت مرتبط با داده‌های گمشده را در نظر گرفته و طیفی از مقادیر قابل قبول برای مشاهدات گمشده ارائه می‌دهد.

اگرچه روش‌های جانی چندگانه در عمل پیچیده به نظر می‌رسند، اما برخلاف روش‌های جانی ساده، تحت تأثیر اربیی قرار نمی‌گیرند و خطای ناشی از جانی کاهش می‌یابد (آزور و همکاران، ۲۰۱۱؛ خان و هوک، ۲۰۲۰). یکی از تکنیک‌های جانی چندگانه که به‌طور گسترده مورد استفاده قرار گرفته است روش MICE است که برای جانی مقادیر گمشده از چندین مدل رگرسیونی استفاده می‌کند. این روش به‌صورت تکراری و زنجیره‌ای عمل کرده و هر متغیر گمشده را به‌عنوان تابعی از سایر متغیرها مدل‌سازی می‌کند.

۴. جانی K- نزدیک‌ترین همسایگی (KNN)

روش KNN یکی از روش‌های یادگیری ماشین نظارت‌شده^۲ برای جانی مقادیر گمشده است. در این روش، ابتدا k نزدیک‌ترین همسایگی که بیشترین شباهت را به داده گمشده دارند شناسایی می‌شوند. سپس مقادیر گمشده با استفاده از مقادیر غیر گمشده (مشاهده‌شده) نقاط همسایگی برآورد می‌شوند. برای اندازه‌گیری شباهت از معیارهای فاصله مانند فاصله مینکوفسکی^۳، فاصله منهتن^۴، فاصله همینگ^۵ و فاصله اقلیدسی^۶ استفاده می‌شود که فاصله اقلیدسی رایج‌ترین آن است (داو و همکاران، ۲۰۱۴). از مزایای این روش می‌توان به سادگی در اجرا، انعطاف‌پذیری در برابر نوع داده (اعم از گسسته یا پیوسته)، امکان جانی هم‌زمان مقادیر گمشده در چندین ویژگی یا ستون و عدم نیاز به فرضیات قوی در مورد توزیع داده‌ها اشاره کرد (امانوئل و همکاران، ۲۰۲۱؛ اسماعیل و همکاران، ۲۰۲۲؛ گلدانی، ۲۰۲۴).

اگرچه این روش به دلیل کارایی بالا در جانی داده‌های گمشده، به‌طور گسترده در مطالعات مورد توجه قرار گرفته و نسخه‌های بهبودیافته آن توسط محققان متعددی معرفی شده است (زو و چنگ، ۲۰۱۵؛ پوجیانو و همکاران، ۲۰۱۹)، اما هنوز هیچ روش استاندارد و اثبات‌شده‌ای برای انتخاب بهترین پارامترهای آن نظیر تعداد همسایه‌ها و معیار فاصله وجود ندارد. این امر

۱. List_wise

۲. Supervised machine learning

۳. Minkowski distance

۴. Manhattan distance

۵. Hamming distance

۶. Euclidean distance

می‌تواند منجر به نتایج متفاوت و ناپایدار در مطالعات مختلف شود. علاوه بر این، در برخی مطالعات اثر سازوکارهای گمشده نادیده گرفته شده است که این موضوع می‌تواند منجر به نتایج غیرواقعی و اریب شود. استفاده از این روش در مجموعه داده‌های بزرگ نیز به دلیل نیاز به جستجو در کل مجموعه داده برای یافتن رکوردهای مشابه ممکن است هزینه‌بر باشد (رحمان و اسلام، ۲۰۱۳؛ فؤاد و همکاران، ۲۰۲۱).

۵. روش‌های مبتنی بر جنگل‌های تصادفی

الگوریتم‌های جانهی داده‌های گمشده مبتنی بر جنگل تصادفی^۱ (RF) یکی از روش‌های مؤثر برای جانهی داده‌های گمشده هستند. این الگوریتم بر مبنای مفهوم یادگیری جمعی^۲ و به‌ویژه تکنیک ادغام نمونه‌های بوت‌استرپ^۳ بنا شده است (بریمن، ۲۰۰۱). در این روش، مجموعه‌ای از درخت‌های تصمیم^۴ به صورت مستقل و موازی آموزش داده می‌شوند. برای آموزش هر درخت، یک نمونه بوت‌استرپ از داده‌ها استخراج و در هر گام از تقسیم گره، تنها یک زیرمجموعه تصادفی از ویژگی‌ها برای انتخاب بهترین تقسیم‌بندی مورد استفاده قرار می‌گیرد. این سازوکار باعث ایجاد تنوع بالا در بین درخت‌های تصمیم شده و از بیش‌برازش^۵ مدل پایه (درخت تصمیم منفرد) جلوگیری می‌کند. نتیجه نهایی از طریق تجمیع^۶ خروجی تمامی درخت‌ها به دست می‌آید، به این صورت برای متغیرهای طبقه‌ای از رأی‌گیری اکثریت^۷ و برای متغیرهای عددی از میانگین‌گیری استفاده می‌شود. الگوریتم جنگل تصادفی با برخورداری از ویژگی‌هایی مانند توانایی مدیریت انواع مختلف داده‌های گمشده، مدلسازی روابط غیرخطی و برهمکنش بین متغیرها، قابلیت مقیاس‌پذیری برای داده‌های بزرگ، به عنوان روشی انعطاف‌پذیر و با دقت بالا، به صورت گسترده در جانهی داده‌های گمشده مورد استفاده قرار می‌گیرد (تانگ و ایشواران، ۲۰۱۷). در حال حاضر، الگوریتم‌های جانهی RF مختلفی وجود دارند. در این مقاله به دو الگوریتم missForest و MICEForest می‌پردازیم.

۱-۵. روش جانهی missForest

الگوریتم missForest یک تکنیک جانهی مبتنی بر یادگیری ماشینی است که در آن از الگوریتم جنگل تصادفی برای جانهی داده‌های گمشده استفاده می‌شود. این الگوریتم توسط استخوون و بولمان در سال ۲۰۱۲ معرفی شد. روش جانهی missForest با یک حدس ابتدایی برای مقادیر گمشده آغاز می‌شود و تا رسیدن به معیار همگرایی، از فرایند برازش مدل جنگل تصادفی بر روی بخش مشاهده‌شده از مجموعه داده‌ها برای پیش‌بینی و به‌روزرسانی مقادیر داده‌های گمشده استفاده می‌کند. این الگوریتم علاوه بر آنکه برای داده‌های پیوسته و گسسته مناسب است و نیازی به تنظیم پارامترها یا فرضیات خاص ندارد، نتایج بسیار خوبی در داده‌های پیچیده با ابعاد بالا ارائه داده است. همچنین، در مقایسه با روش‌های دیگر مانند روش KNN و MICE، عملکرد بهتری نشان داده و می‌تواند کیفیت جانهی را بدون نیاز به داده‌های آزمون یا اعتبارسنجی پیچیده ارزیابی کند (والجی و همکاران، ۲۰۱۳). با این حال برخی مطالعات اشاره کرده‌اند که برآوردهای حاصل از روش missForest برای متغیرهای به‌طور تصادفی گمشده اریب بوده و فواصل اطمینان حاصل از آن پوشش ضعیفی دارند (شاه و همکاران، ۲۰۱۴).

۲-۵. جانهی MICEForest

روش جانهی MICEForest یک روش پیشرفته برای جانهی داده‌های گمشده است که با استفاده از جنگل‌های تصادفی به عنوان مدل جانهی، اصول MICE را گسترش داده و انعطاف‌پذیری MICE را با توان پیش‌بینی جنگل‌های تصادفی ترکیب می‌کند. این روش به دلیل دقت بالا و انعطاف‌پذیری در مدیریت داده‌های گمشده بسیار مورد توجه قرار گرفته است (داو و همکاران، ۲۰۱۴).

۱. Random Forest
۲. Ensemble Learning
۳. Bootstrap aggregation
۴. Decision trees
۵. Overfitting
۶. Aggregation
۷. Majority Voting

اگرچه هر دو الگوریتم مذکور از جنگل تصادفی برای جانهی داده‌های گمشده استفاده می‌کنند، یک تفاوت اصلی در اجرا وجود دارد. در الگوریتم missForest تمرکز بر جانهی واحد است، به این معنی که برای جانهی داده‌های گمشده از جنگل تصادفی استفاده شده و این فرآیند تا زمانی که مقادیر جانهی شده بهبود یابد تکرار می‌شود. اما در روش MICEForest جانهی‌های متعددی برای مقادیر گمشده در نظر گرفته می‌شود که می‌تواند رویکرد جامع‌تری برای مدیریت داده‌های گمشده فراهم نماید.

منبع داده‌ها

داده‌های مورد بررسی در این مطالعه ولادت‌های رخ داده در سال ۱۴۰۲ است که در پایگاه اطلاعات جمعیت کشور ثبت شده‌اند. برای اهداف این مطالعه، هشت متغیر جنسیت فرزند، مرتبه ولادت، استان محل رخداد ولادت، چندقلوایی، سن پدر در زمان ولادت فرزند، شهری یا روستایی بودن ولادت، فاصله ازدواج تا تولد فرزند و متغیر اصلی یعنی سن مادر در زمان ولادت فرزند در نظر گرفته شده‌اند. آماره‌های توصیفی برای متغیرهای پیوسته و گسسته و درصد رکوردهای دارای مقادیر گمشده برای هر متغیر در جدول ۱ ارائه شده است.

یافته‌های تحقیق

آماره‌های توصیفی و درصد رکوردهای دارای مقادیر گمشده برای هر متغیر در جدول ۱ ارائه شده است. با توجه به جدول ۱، درصد گمشدگی داده‌ها از صفر درصد در متغیرهای جنسیت فرزند، نوع ولادت (شهری یا روستایی) و چندقلوایی تا ۲۰۱ درصد در متغیر فاصله ازدواج تا تولد فرزند متغیر است. به منظور بررسی دقت روش‌های جانهی پیشنهادی و حذف اثر مقادیر گمشده بر نتایج به دست آمده، رکوردهایی که مقادیر هر یک از متغیرها در آن‌ها نامشخص است قبل از مطالعه از مجموعه داده‌ها حذف شده‌اند. با حذف این موارد، تعداد ۱۰۲۱۷۲۷ رکورد با اطلاعات کامل وارد مطالعه شده‌اند.

برای بررسی اثر سازوکار گمشدگی داده‌ها و نسبت گمشدگی بر روی روش‌های جانهی معرفی شده، سه سازوکار گمشدگی شامل MAR، MCAR و MNAR و نسبت‌های گمشدگی ۰،۰۱۵، ۰،۰۱، ۰،۰۰۵، ۰،۰۱، ۰،۰۰۱ بر روی متغیر سن مادر در نظر گرفته شده است. مقادیر واقعی سن مادر قبل از اعمال سازوکار گمشدگی و بعد از حذف و برازش آن با استفاده از روش‌های جانهی مورد بررسی و مقایسه قرار گرفته‌اند. از شاخص ریشه توان دوم خطا^۱ (RMSE) برای ارزیابی روش‌های جانهی مختلف استفاده شده است که به صورت زیر تعریف می‌شود:

$$RMSE_{Method} = \sqrt{\frac{1}{m} \sum_{i=1}^m (x_i - \hat{x}_i)^2}$$

که در آن m تعداد داده‌های گمشده در مجموعه داده، x_i مقدار واقعی و $\hat{x}_i$ مقدار داده جانهی شده توسط هر یک از روش‌های جانهی و اندیس «Method» اشاره به یکی از روش‌های جانهی داده‌های گمشده دارد. در جانهی ساده، مقادیر نامشخص سن مادر با مقداری که بیشترین فراوانی را دارد (مد^۲) جایگزین شده است. همچنین به منظور انتخاب مقدار K بهینه در روش KNN از روش جستجوی شبکه‌ای^۳ براساس معیار RMSE بر روی مقادیر ۲، ۳، ۴، ۵، ۶، ۷، ۸، ۹ استفاده شده است. کارایی روش‌های مختلف جانهی داده‌های نامشخص سن مادر به ازای مقادیر مختلف نسبت گمشدگی داده‌ها و سازوکارهای گمشدگی در جدول ۲ و نمودار ۱ ارائه شده است. براساس نتایج مشاهده می‌شود براساس معیار RMSE:

- به ازای سازوکارها و مقادیر مختلف نسبت گمشدگی، روش‌های جانهی مبتنی بر الگوریتم‌های یادگیری ماشین به طور قابل توجهی عملکرد بهتری نسبت به روش‌های سنتی جانهی داده‌های گمشده دارند. برای مثال برای سازوکار گمشدگی MCAR و نسبت گمشدگی ۰،۰۱۵، $p = ۰،۰۹۵۱$ ، $RMSE_{MICE} = ۰،۰۹۵۱$ در حالی که $RMSE_{missForest} = ۰،۰۷۸۱$.

۱. Root mean squared error

۲. Mode

۳. Grid search

- در بین روش‌های سنتی جانهای داده‌های گمشده، بالاترین مقادیر RMSE به روش EM و کمترین مقادیر RMSE به روش MICE اختصاص دارد. به عبارت دیگر، روش جانهای MICE به‌طور قابل توجهی عملکرد بهتری نسبت به روش‌های ساده‌تر مانند جانهای مقادیر گمشده با بیشترین فراوانی و جانهای EM دارد. برای مثال تحت سازوکار گمشدگی MAR و $p = 0.05$ ، $RMSE_{MICE} = 0.982$ درحالی‌که $RMSE_{EM} = 1.577$ و $RMSE_{Simple} = 1.047$
- در بین روش‌های مبتنی بر الگوریتم‌های یادگیری ماشین، مقدار RMSE روش جانهای KNN بسیار نزدیک به روش جانهای missForest است. باین حال الگوریتم missForest بهترین روش برای جانهای داده‌های نامشخص سن مادر است. برای مثال تحت سازوکار گمشدگی MAR و $p = 0.01$ ، $RMSE_{missForest} = 0.655$ درحالی‌که $RMSE_{KNN} = 0.691$
- مقادیر RMSE مربوط به روش جانهای MICE (یکی از روش‌های سنتی جانهای داده‌های گمشده) به روش جانهای MICEForest (از روش‌های مبتنی بر الگوریتم‌های یادگیری ماشین) نزدیک است، اما کارایی روش MICEForest به‌طور قابل توجهی بهتر از روش MICE است. به‌عنوان مثال تحت سازوکار گمشدگی MNAR و نسبت گمشدگی $RMSE_{MICE} = 0.967$ ، $p = 0.2$ درحالی‌که $RMSE_{MICEForest} = 0.916$
- در روش‌های جانهای داده‌های گمشده مبتنی بر الگوریتم‌های یادگیری ماشین، تحت سازوکار گمشدگی MCAR، با افزایش نسبت گمشدگی از 0.15 به 0.2 مقادیر RMSE به‌طور قابل ملاحظه‌ای کاهش می‌یابد. برای مثال به ازای $RMSE_{KNN} = 0.809$ ، $p = 0.15$ درحالی‌که به ازای $p = 0.2$ ، $RMSE_{KNN} = 0.650$. این روند در سازوکارهای گمشدگی MAR و MNAR چندان مشهود نیست.
- ارزیابی روش‌های جانهای مبتنی بر الگوریتم‌های یادگیری ماشین نشان می‌دهد که به ازای نسبت‌های گمشدگی بین 0.01 تا 0.15، مقادیر RMSE تحت سازوکار گمشدگی غیرتصادفی کمتر از سایر سازوکار گمشدگی داده‌ها است. اما به ازای نسبت گمشدگی $p = 0.2$ ، کارایی روش‌های جانهای تحت سازوکار گمشدگی کاملاً تصادفی (MCAR) بهتر از سایر سازوکارهای گمشدگی داده‌ها است.
- در بین روش‌های سنتی جانهای داده‌های گمشده، روش جانهای MICE عملکردی مشابه روش‌های مبتنی بر الگوریتم‌های یادگیری ماشین دارد به قسمی که به ازای نسبت‌های گمشدگی 0.01 تا 0.15، مقدار RMSE تحت سازوکار گمشدگی MNAR و به ازای $p = 0.2$ تحت سازوکار گمشدگی MCAR کمترین مقدار را دارد. در مقابل مقدار RMSE در الگوریتم EM و جانهای ساده تحت سازوکار گمشدگی کاملاً تصادفی به حداقل می‌رسد.

جدول ۱. آماره‌های توصیفی و درصد رکوردهای گمشده برای متغیرهای مرتبط با ولادت، ۱۴۰۲

ردیف	متغیر	میانگین و انحراف استاندارد، یا درصد	تعداد		درصد رکوردهای گمشده
			رکوردهای مشاهده شده	رکوردهای گمشده	
۱	استان محل ولادت فرزند	آذربایجان شرقی: ۴/۳؛ آذربایجان غربی: ۴/۸؛ اردبیل: ۱/۴؛ اصفهان: ۵/۰؛ البرز: ۲/۱؛ ایلام: ۰/۷؛ بوشهر: ۱/۳؛ تهران: ۱۲/۳؛ چهارمحال و بختیاری: ۱/۳؛ خراسان جنوبی: ۱/۵؛ خراسان رضوی: ۹/۵؛ خراسان شمالی: ۱/۳؛ خوزستان: ۸/۴؛ زنجان: ۱/۳؛ سمنان: ۰/۶؛ سیستان و بلوچستان: ۸/۴؛ فارس: ۵/۲؛ قزوین: ۱/۳؛ قم: ۱/۷؛ کردستان: ۲/۰؛ کرمان: ۴/۶؛ کرمانشاه: ۲/۲؛ کهگیلویه و بویراحمد: ۱/۰؛ گلستان: ۲/۹؛ گیلان: ۱/۸؛ لرستان: ۲/۳؛ مازندران: ۲/۵؛ مرکزی: ۱/۲؛ هرمزگان: ۳/۱؛ همدان: ۲/۱؛ یزد: ۱/۶	۱,۰۴۳,۳۴۱	۵۹۰	۰/۰۰۱
۲	جنسیت فرزند	دختر: ۵۱/۸؛ پسر: ۴۸/۲	۱,۰۴۳,۳۹۱	.	.
۳	نوع ولادت	شهری: ۷۵/۵؛ روستایی: ۲۴/۵	۱,۰۴۳,۳۹۱	.	.
۴	سن مادر	میانگین: ۳۰/۳؛ انحراف استاندارد: ۶/۵	۱,۰۳۵,۳۶۳	۸۵۶۸	۰/۰۰۸
۵	سن پدر	میانگین: ۳۵؛ انحراف استاندارد: ۶/۴	۱,۰۴۳,۲۱۰	۷۲۱	۰/۰۰۱
۶	فاصله ازدواج تا تولد فرزند (سال)	میانگین: ۸۰؛ انحراف استاندارد: ۵/۴	۱,۰۲۱,۷۲۷	۲۲۲۰۴	۰/۰۲۱
۷	چندقلو زایی	یک‌قلو: ۹۶/۳؛ دوقلو: ۳/۶؛ سه‌قلو و بیشتر: ۰/۲	۱,۰۴۳,۹۳۱	.	.
۸	مرتبه ولادت فرزند	اول: ۳۶/۴؛ دوم: ۳۸/۵؛ سوم: ۱۷/۸؛ چهارم: ۵/۲؛ پنجم و بیشتر: ۲/۱	۱,۰۳۵,۴۷۷	۸۴۵۴	۰/۰۰۸

نتیجه‌گیری

در این مطالعه کارایی روش‌های مختلف جانی داده‌های گمشده برای متغیر سن مادر در داده‌های ولادت سال ۱۴۰۲ تحت سازوکارها و نسبت‌های مختلف گمشدگی مورد بررسی قرار گرفت. یافته‌های تحقیق نشان داد که روش‌های جانی مبتنی بر الگوریتم‌های یادگیری ماشین به‌ویژه الگوریتم missForest در کاهش خطای جانی براساس معیار RMSE عملکرد بهتری نسبت به روش‌های سنتی دارند. این نتایج با یافته‌های پژوهش استخوون و بولمان (۲۰۱۲) و تانگ و ایشواران (۲۰۱۷) که بر دقت بالاتر روش missForest در داده‌های با ساختار پیچیده تأکید داشته‌اند، همسو است.

در بین روش‌های سنتی جانی داده‌های گمشده، روش جانی MICE نیز در بسیاری از موارد عملکرد مطلوبی داشته و نتایج آن به روش‌های یادگیری ماشین نزدیک بوده است. این یافته با نتایج پژوهش ون بورن و گروتویس-آودشورن (۲۰۱۱) که بیان کرده‌اند روش MICE می‌تواند به دقت روش‌های یادگیری ماشین نزدیک شود مطابقت دارد. این یافته نشان می‌دهد که روش MICE می‌تواند در مواقعی که پیاده‌سازی روش‌های پیشرفته امکان‌پذیر نیست به‌عنوان جایگزین مناسبی برای جانی مقادیر گمشده مورد استفاده قرار گیرد.

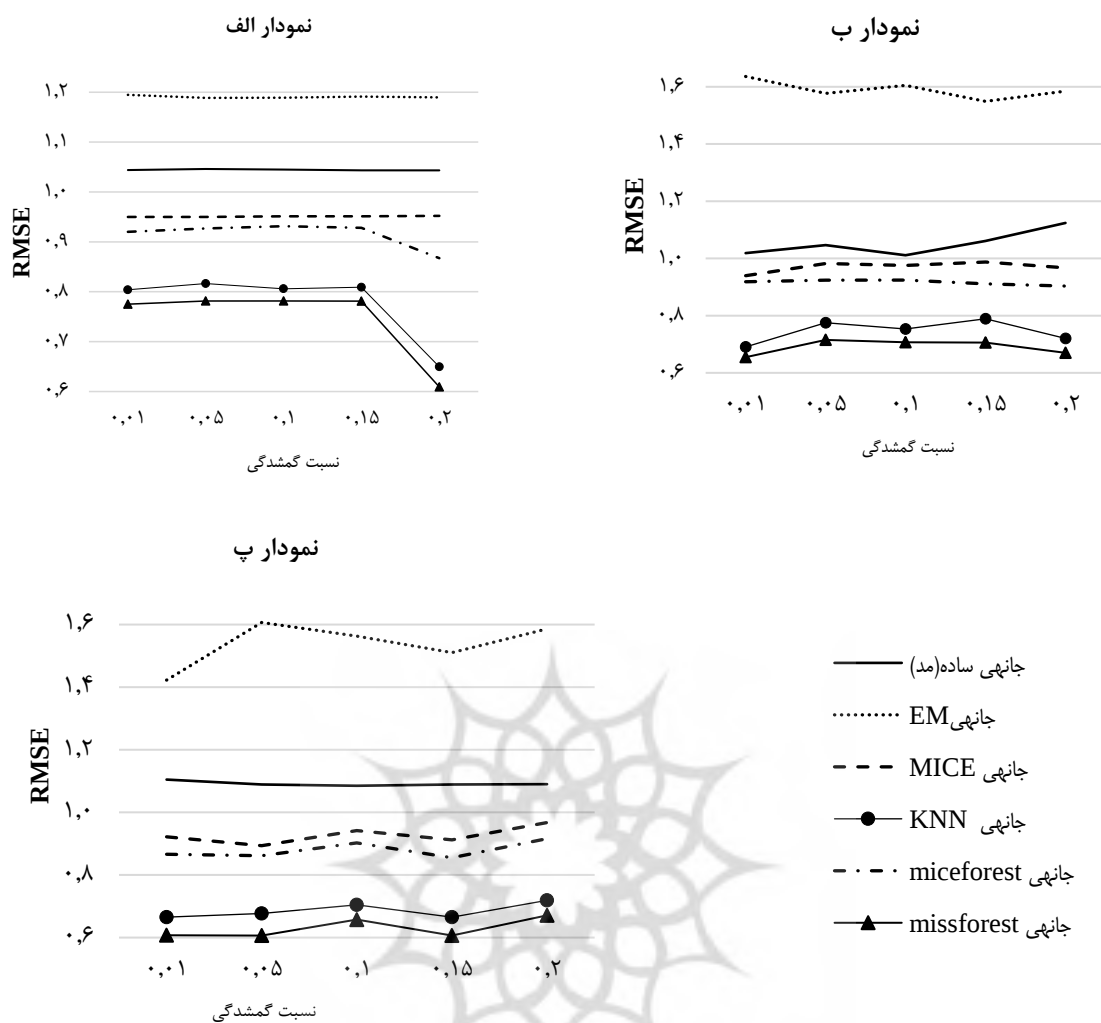
تحلیل وابستگی عملکرد روش‌های جانی برحسب سازوکارها و نسبت‌های گمشدگی نشان داد روش‌های یادگیری ماشین در شرایط گمشدگی غیرتصادفی و نسبت‌های پایین گمشدگی (۰/۰۱ تا ۰/۱۵) عملکرد بهتری داشتند، درحالی‌که با افزایش نسبت گمشدگی ($p = ۰/۲$)، این روش‌ها تحت شرایط گمشدگی کاملاً تصادفی عملکرد مطلوب‌تری ارائه دادند. این الگوها با یافته‌های وانگ و همکاران (۲۰۲۲b) که به تأثیر سازوکارها و نسبت‌های گمشدگی بر عملکرد روش‌های جانی داده‌های گمشده اشاره کرده‌اند همخوانی دارد.

از دیدگاه کاربردی این مطالعه نشان داد که انتخاب روش بهینه جانی باید براساس نوع سازوکار گمشدگی داده‌ها، نرخ گمشدگی و ساختار داده‌ها صورت گیرد. همچنین استفاده از روش‌های پیشرفته مانند missForest می‌تواند به افزایش دقت تحلیل‌ها و ارائه نتایج معتبرتر منجر شود. با توجه به برتری روش missForest در این مطالعه و مطالعات مشابه پیشنهاد می‌شود

از این روش برای جانهی داده‌های نامشخص سن مادر استفاده شود. همچنین برای مواردی که پیاده‌سازی روش‌های پیچیده یادگیری ماشین امکان‌پذیر نیست، روش جانهی MICE به‌عنوان جایگزین مناسب پیشنهاد می‌شود.

جدول ۲. مقدار RMSE روش‌های جانهی داده‌های نامشخص سن مادر تحت سازوکار گمشدگی MCAR، MAR و MNAR و نسبت‌های گمشدگی مختلف

نسبت گمشدگی					روش جانهی / سازوکار گمشدگی	
۰/۲	۰/۱۵	۰/۱	۰/۰۵	۰/۰۱		
۱/۰۴۴	۱/۰۴۳	۱/۰۴۵	۱/۰۴۶	۱/۰۴۴	MCAR	Simple (Mode)
۱/۱۳۴	۱/۰۶۱	۱/۰۱۱	۱/۰۴۷	۱/۰۱۹	MAR	
۱/۰۹۰	۱/۰۸۹	۱/۰۸۵	۱/۰۸۹	۱/۱۰۵	MNAR	
۱/۱۹۰	۱/۱۹۱	۱/۱۸۹	۱/۱۸۹	۱/۱۹۵	MCAR	EM
۱/۵۸۵	۱/۵۴۹	۱/۶۰۵	۱/۵۷۷	۱/۳۶۳	MAR	
۱/۵۸۶	۱/۵۱۰	۱/۵۶۳	۱/۶۰۷	۱/۴۲۳	MNAR	
۰/۹۵۲	۰/۹۵۱	۰/۹۵۱	۰/۹۵۰	۰/۹۵۰	MCAR	MICE
۰/۹۶۷	۰/۹۸۷	۰/۹۷۶	۰/۹۸۲	۰/۹۴۰	MAR	
۰/۹۶۷	۰/۹۱۲	۰/۹۴۲	۰/۸۹۴	۰/۹۲۲	MNAR	
۰/۶۵۰	۰/۸۰۹	۰/۸۰۶	۰/۸۱۶	۰/۸۰۴	MCAR	KNN
۰/۷۲۱	۰/۷۸۹	۰/۷۵۳	۰/۷۷۵	۰/۶۹۱	MAR	
۰/۷۱۹	۰/۶۶۵	۰/۷۰۴	۰/۶۷۷	۰/۶۶۵	MNAR	
۰/۸۶۷	۰/۹۲۸	۰/۹۳۱	۰/۹۲۷	۰/۹۲۰	MCAR	MICEForst
۰/۹۰۳	۰/۹۱۱	۰/۹۲۴	۰/۹۲۴	۰/۹۱۹	MAR	
۰/۹۱۶	۰/۸۵۵	۰/۹۰۲	۰/۸۶۲	۰/۸۶۶	MNAR	
۰/۶۰۹	۰/۷۸۱	۰/۷۸۱	۰/۷۸۱	۰/۷۷۵	MCAR	missForest
۰/۶۷۰	۰/۷۰۶	۰/۷۰۷	۰/۷۱۵	۰/۶۵۵	MAR	
۰/۶۷۱	۰/۶۰۷	۰/۶۵۷	۰/۶۰۷	۰/۶۰۷	MNAR	



نمودار ۱. مقادير RMSE روش‌هاي جانپي داده‌هاي گمشده سن مادر به ازاي سازگار گمشدگي MCAR (نمودار الف)، MAR (نمودار ب) و MNAR (نمودار پ) به ازاي نسبت‌هاي گمشدگي مختلف

پروژه نگاه علوم انساني و مطالعات فرهنگي
پرتال جامع علوم انساني

References

- Abidin, N. Z., & Ismail, A. R. (۲۰۲۱). An improved K-Nearest neighbour with grasshopper optimization algorithm for imputation of missing data. *International Journal of Advances in Intelligent Informatics*, ۷(۳), ۲۰۴.
- Alsaber, A., Al-Herz, A., Pan, J., Al-Sultan, A. T., Mishra, D., & KRRD Group. (۲۰۲۱). Handling missing data in a rheumatoid arthritis registry using random forest approach. *International Journal of Rheumatic Diseases*, ۲۴(۱۰), ۱۲۸۲-۱۲۹۳.
- Aracri, F., Bianco, M. G., Quattrone, A., & Sarica, A. (۲۰۲۳). Imputation of missing clinical, cognitive and neuroimaging data of Dementia using missForest, a Random Forest based algorithm. In *2023 IEEE 36th International Symposium on Computer-based Medical Systems*. Pp. 684-688. IEEE.
- Austin, P. C., White, I. R., Lee, D. S., & van Buuren, S. (۲۰۲۱). Missing data in clinical research: A tutorial on multiple imputation. *The Canadian Journal of Cardiology*, ۳۷(۹), ۱۳۲۲-۱۳۳۱.
- Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (۲۰۱۱). Multiple imputation by chained equations: what is it and how does it work?: Multiple imputation by chained equations. *International Journal of Methods in Psychiatric Research*, ۲۰(۱), ۴۰-۴۶.
- Batista, G. E. A. P. A., & Monard, M. C. (۲۰۰۳). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, ۱۷(۵-۶), ۵۱۹-۵۳۳.
- Beesley, L. J., Bondarenko, I., Elliot, M. R., Kurian, A. W., Katz, S. J., & Taylor, J. M. (۲۰۲۱). Multiple imputation with missing data indicators. *Statistical methods in medical research*, ۳۰(۱۲), ۲۶۸۵-۲۷۰۰.
- Breiman, L. (۲۰۰۱). Random Forests. *Machine Learning*, ۴۵(۱), ۵-۳۲.
- De Silva, H., & Perera, A. S. (۲۰۱۶). Missing data imputation using Evolutionary k-Nearest neighbor algorithm for gene expression data. In *۲۰۱۶ Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)* (pp. ۱۴۱-۱۴۶). IEEE.
- Deb, R., & Liew, A. W.-C. (۲۰۱۶). Missing value imputation for the analysis of incomplete traffic accident data. *Information Sciences*, ۳۳۹, ۲۷۴-۲۸۹.
- Delalleau, O., Courville, A., & Bengio, Y. (۲۰۱۲). Efficient EM training of Gaussian mixtures with missing data. *arXiv preprint arXiv:1209.0521*.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (۱۹۷۷). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, ۳۹(۱), ۱-۲۲.
- Doove, L. L., Van Buuren, S., & Dusseldorp, E. (۲۰۱۴). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, ۷۲, ۹۲-۱۰۴.
- Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (۲۰۲۱). A survey on missing data in machine learning. *Journal of Big Data*, ۸(۱), ۱۴۰.
- Fouad, K. M., Ismail, M. M., Azar, A. T., & Arafa, M. M. (۲۰۲۱). Advanced methods for missing values imputation based on similarity learning. *PeerJ. Computer Science*, ۷(e۶۱۹), e۶۱۹.
- Goldani, M. (۲۰۲۴). Comparative analysis of missing values imputation methods: A case study in financial series (S&P۵۰۰ and bitcoin value data sets). *Iranian Journal of Finance*, ۸(۱), ۴۷-۷۰.

- Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (۲۰۰۷). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science: The Official Journal of the Society for Prevention Research*, ۸(۳), ۲۰۶-۲۱۳.
- Hong, S., & Lynn, H. S. (۲۰۲۰). Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Medical Research Methodology*, ۲۰(۱), ۱۹۹.
- Howell, D. C. (۲۰۰۷). The treatment of missing data. *The SAGE Handbook of Social Science Methodology*, ۲۱۲-۲۲۶.
- Ismail, A. R., Abidin, N. Z., & Maen, M. K. (۲۰۲۲). Systematic review on missing data imputation techniques with machine learning algorithms for healthcare. *Journal of Robotics and Control*, ۳(۲), ۱۴۳-۱۵۲.
- Jegadeeswari, K., Ragunath, R., & Rathipriya, R. (۲۰۲۲). Missing data imputation using ensemble learning technique: a review. *Soft Computing for Security Applications: Proceedings of ICSCS 2022*, ۲۲۳-۲۳۶.
- Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (۲۰۱۰). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, ۵۰(۲), ۱۰۵-۱۱۵.
- Khan, S. I., & Hoque, A. S. M. L. (۲۰۲۰). SICE: an improved missing data imputation technique. *Journal of big Data*, ۷(۱), ۳۷.
- Li, J., Guo, S., Ma, R., He, J., Zhang, X., Rui, D., ... & Guo, H. (۲۰۲۴). Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. *BMC Medical Research Methodology*, ۲۴(۱), ۴۱.
- Lin, W. C., Ke, S. W., & Tsai, C. F. (۲۰۱۷). When should we ignore examples with missing values?. *International Journal of Data Warehousing and Mining*, ۱۳(۴), ۵۳-۶۳.
- Little R.J. A. & Rubin, D. B. (۲۰۰۲) *Statistical analysis with missing data*, John Wiley & Sons.
- Liu, Z. G., Liu, Y., Dezert, J., & Pan, Q. (۲۰۱۵). Classification of incomplete data based on belief functions and K-nearest neighbors. *Knowledge-Based Systems*, ۸۹, ۱۱۳-۱۲۵.
- Mattei, A., Mealli, F., & Rubin, D. B. (۲۰۱۱). Missing data and imputation methods. *Modern Analysis of Customer Surveys: with Applications using R*, ۱۲۹-۱۵۴.
- Muzellec, B., Josse, J., Boyer, C., & Cuturi, M. (۲۰۲۰). Missing data imputation using optimal transport. In *International Conference on Machine Learning* (pp. ۷۱۳۰-۷۱۴۰). PMLR.
- Petrazzini, B. O., Naya, H., Lopez-Bello, F., Vazquez, G., & Spangenberg, L. (۲۰۲۱). Evaluation of different approaches for missing data imputation on features associated to genomic data. *BioData mining*, ۱۴, ۱-۱۳.
- Pujianto, U., Wibawa, A. P., & Akbar, M. I. (۲۰۱۹). K-nearest neighbor (k-NN) based missing data imputation. In *2019 5th International Conference on Science in Information*. ۸۳-۸۸. IEEE.
- Rahman, M. G., & Islam, M. Z. (۲۰۱۳). Data quality improvement by imputation of missing values. In *International Conference on Computer Science and Information Technology* (pp. ۸۲-۸۸). Springer-Verlag London Ltd..
- Rizvi, S. T. H., Latif, M. Y., Amin, M. S., Telmoudi, A. J., & Shah, N. A. (۲۰۲۳). Analysis of Machine Learning Based Imputation of Missing Data. *Cybernetics and Systems*, ۱-۱۵.
- Rubin, D. B. (۱۹۷۶). Inference and missing data. *Biometrika*, ۶۳(۳), ۵۸۱-۵۹۲.

- Rubin, D. B. (۱۹۸۷). *Multiple imputation for nonresponse in surveys* (Vol. ۸۱). John Wiley & Sons.
- Rubin, L. H., Witkiewitz, K., Andre, J. S., & Reilly, S. (۲۰۰۷). Methods for handling missing data in the behavioral neurosciences: Don't throw the baby rat out with the bath water. *Journal of Undergraduate Neuroscience Education*, ۶(۲), A۷۱.
- Sania, A., Pini, N., Nelson, M. E., Myers, M. M., Shuffrey, L. C., Lucchini, M., ... & Fifer, W. P. (۲۰۲۵). K-nearest neighbor algorithm for imputing missing longitudinal prenatal alcohol data. *Advances in Drug and Alcohol Research*, ۴, ۱۳۴۴۹.
- Schafer, J. L. (۱۹۹۷). *Analysis of incomplete multivariate data*. CRC press.
- Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (۲۰۱۴). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American journal of epidemiology*, ۱۷۹(۶), ۷۶۴-۷۷۴.
- Stekhoven, D. J., & Bühlmann, P. (۲۰۱۲). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, ۲۸(۱), ۱۱۲-۱۱۸.
- Tang, F., & Ishwaran, H. (۲۰۱۷). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, ۱۰(۶), ۳۶۳-۳۷۷.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... & Altman, R. B. (۲۰۰۱). Missing value estimation methods for DNA microarrays. *Bioinformatics*, ۱۷(۶), ۵۲۰-۵۲۵.
- Van Buuren, S., & Groothuis-Oudshoorn, K. (۲۰۱۱). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, ۴۵, ۱-۶۷.
- Waljee, Akbar K., Ashin Mukherjee, Amit G. Singal, Yiwei Zhang, Jeffrey Warren, Ulysses Balis, Jorge Marrero, Ji Zhu, and Peter DR Higgins. (۲۰۱۳). Comparison of imputation methods for missing laboratory data in medicine. *BMJ open*, ۳(۸), e۰۰۲۸۴۷.
- Wang, K., Luo, M., Deng, M., & Chen, H. (۲۰۲۲a). Nested Random Forest: A Personalized Imputation Method for Missing Data. In *2022 8th International Conference on Big Data and Information Analytics (BigDIA)* (pp. ۱۱۹-۱۲۶). IEEE.
- Wang, H., Tang, J., Wu, M., Wang, X., & Zhang, T. (۲۰۲۲b). Application of machine learning missing data imputation techniques in clinical decision making: taking the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example. *BMC Medical Informatics and Decision Making*, ۲۲, ۱-۱۴.
- White, I. R., Royston, P., & Wood, A. M. (۲۰۱۱). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, ۳۰(۴), ۳۷۷-۳۹۹.
- Zhu, M., & Cheng, X. (۲۰۱۵). Iterative KNN imputation based on GRA for missing values in TPLMS. In *2015 4th international conference on computer science and network technology (ICCSNT)* (Vol. ۱, pp. ۹۴-۹۹). IEEE