# ارائۀ یک روش جدید برای تخمین مقادیر گمشده در مجموعه داده

سلیمه ضیاءالدینی[۱،*] و مینا ابارقی[۲]

**اطلاعات مقاله**

دریافت مقاله: ۱۳۹۵/۰۹/۲۴
پذیرش مقاله: ۱۳۹۶/۱۱/۱۷

واژگان کلیدی:
داده‌کاوی،
داده گمشده،
الگوریتم C4.5،
مجموعه داده،
درخت تصمیم.

**چکیده**

بیشتر مجموعه داده‌های مربوط به داده‌کاوی و ماشین یادگیری، دارای داده‌هایی با مقادیر Missing Values یا داده گمشده هستند. برخورد با داده گمشده و نیز ارائه راهکارهایی مبتنی بر تخمین مقدار مربوط به داده گمشده، منجر به بروز یک مسئله بسیار مهم در زمینه داده‌کاوی و ماشین یادگیری شده است. در بین الگوریتم‌های داده‌کاوی، الگوریتم C4.5، به‌دلیل کارایی، استفاده در کاربردهای مختلف داده‌کاوی و نیز توانایی در کار کردن و تخمین مقدار داده گمشده در مجموعه داده‌ها، به‌طور مکرر مورد استفاده قرار گرفته است. پژوهشگران، روش‌ها و الگوهای متعددی برای برخورد با مقادیر داده گمشده و تخمین مقدار آن در مجموعه داده‌های الگوریتم C4.5 ارائه کرده‌اند که هریک از روش‌ها به‌نحوی موجب افزایش دقت درخت تصمیم و در نتیجه، تولید یک درخت تصمیم مؤثر و کاراتر شده است. بنابراین در مقاله حاضر ابتدا به بررسی و مرور روش‌ها و راهکارهای ارائه‌شدۀ پیشین و سپس به ارائه روش پیشنهادی با عنوان روش جابه‌جایی خصوصیت‌ها جهت تخمین مقادیر گمشده در مجموعه داده، پرداخته خواهد شد. در پایان به مقایسه و ارزیابی دقت حاصل‌شده روش پیشنهادی با روش‌های حذف و میانگین خواهیم پرداخت.

## ۱- مقدّمه

داده‌کاوی، فرایندی است خودکار برای استخراج الگوهایی که دانش را بازنمایی می کنند. این دانش‌ها به‌صورت ضمنی در پایگاه داده‌های عظیم و انباره داده و دیگر مخازن بزرگ اطلاعات ذخیره شده‌اند. داده‌کاوی، روشی برای تجزیه و تحلیل داده‌های موجود در مجموعه داده، به‌منظور کشف روابط پنهانی بین داده‌های موجود در مجموعه داده است.
به بیان ساده‌تر، داده‌کاوی، به فرایند استخراج دانش ناشناخته، درست و مفید از داده اطلاق می‌شود و تکنیکی برای شناسایی اطلاعات یا دانش تصمیم‌گیری از داده است، به‌نحوی که با استخراج آن‌ها، در حوزه‌های تصمیم‌گیری، پیش‌بینی، پیش‌گویی و تخمین، مورد استفاده قرار گیرند. داده‌ها اغلب حجیم، اما بدون ارزش هستند. داده‌ها به‌تنهایی قابل استفاده نیستند، بلکه دانش نهفته در داده‌ها قابل استفاده است.

سنگ بنای عملیات داده‌کاوی خوب، دسترسی به داده‌های اولیه خوب و مناسب است. داده‌های موجود در مجموعه داده‌ها از طریق پرس‌وجو جمع‌آوری می‌شوند، اما در جمع‌آوری داده‌ها و ایجاد مجموعه داده ممکن است برخی از داده‌ها فاقد مقدار باشند که به این داده‌ها، داده‌های با مقادیر نامشخص یا داده‌های گمشده گفته می‌شود. درحال حاضر، بسیاری از مجموعه داده‌های مربوط به ماشین یادگیری و داده‌کاوی دارای داده‌هایی با مقادیر نامشخص هستند و طبقه‌بندی با وجود این گونه داده‌ها، به‌عنوان یک چالش و نقص محسوب می‌شود.

داده‌های موجود در یک مجموعه داده، زمانی که وارد

* پست الکترونیک نویسنده مسئول: sz220_zia@yahoo.com
۱. کارشناس ارشد مهندسی نرم‌افزار کامپیوتر، کرمان، ایران
۲. دانشجوی کارشناسی ارشد مهندسی نرم‌افزار کامپیوتر، کرمان، ایران

الگوریتم می‌شوند، باید کامل و بدون مقادیر گمشده باشند. این مشکل بیشتر ناشی از عدم یکپارچگی قسمت‌های جمع‌آوری‌کننده داده‌ها، منابع داده و شکل‌های جمع‌آوری داده‌ها است. وجود یک مجموعه داده خالی از مقادیر داده گمشده در کشف روابط پنهانی بین داده‌های موجود در مجموعه داده، نقشی بسیار مهم و حیاتی دارد. از معروف‌ترین تکنیک‌های طبقه‌بندی در داده‌کاوی، درخت تصمیم است. در حال حاضر، تعداد بسیار زیادی از الگوریتم‌های درخت تصمیم از جمله ID3[۱]،[۱] CART[۲]،[۲] C4.5[۳]،[۳] وجود دارد که هریک عملاً در کاربردهای متفاوت استفاده می‌شوند.

در بین الگوریتم‌های موجود، الگوریتم C4.5، به‌دلیل کارایی، استفاده در کاربردهای مختلف داده‌کاوی و نیز توانایی در کار کردن و پیش‌بینی داده‌های دارای مقادیر داده گمشده، به‌طور مکرر مورد استفاده قرار گرفته است. ادامه مقاله بدین صورت است: در بخش۲ خصوصیات داده گمشده، در بخش ۳ روش‌های مختلف گمشدگی، در بخش ۴ مروری بر روش‌های تخمین داده گمشده، در بخش ۵ روش پیشنهادی جابه‌جایی خصوصیت‌ها، در بخش ۶ مقایسه دقت روش پیشنهادی با روش‌های حذف و میانگین، در بخش ۷ نتیجه‌گیری و در نهایت، در انتهای مقاله به ذکر مراجع مرتبط با مقاله خواهیم پرداخت.

## ۲- خصوصیات داده گمشده

داده گمشده، یک مشکل عمومی در مجموعه داده است و نهایتاً دستیابی به یک نتیجه‌گیری و تصمیم‌گیری مفید از داده‌های جمع‌آوری‌شده را با مشکل مواجه می‌سازد. با وجود تمام این مشکلات، گمشدگی بهتر از جواب اشتباه در مجموعه داده است. اصطلاحات مختلف و غالباً مترادفی برای این مفهوم وجود دارد. این اصطلاحات عبارت‌اند از: داده گمشده، داده‌های ناقص و بی‌پاسخ. معمولاً داده‌های گمشده در مجموعه داده‌ها، با خالی گذاشتن خانه‌هایی از این مجموعه داده مشخص می‌شوند. به عبارت دیگر، وقتی داده‌ای از تعدادی از پاسخگویان در برخی از متغیرها وجود ندارد، هنگام ورود داده‌ها نیز در خانه‌های مربوط به این متغیرها که به این پاسخگویان تعلق دارند، مقداری وارد نمی‌شود. در مجموعه داده‌ها، این خانه‌های خالی، با علامت ? نشان داده می‌شوند.

در یک مجموعه داده، دو نوع گمشدگی وجود دارد. گمشدگی آزمودنی که تمامی اطلاعات برای یک نمونه گمشده باشد. به عنوان مثال، برخی از افراد از پاسخ دادن به سؤالات اجتناب می‌کنند. محققان، تکنیسین‌ها و جمع‌آوری‌کننده داده‌ها ممکن است اشتباهاتی را انجام دهند.

گمشدگی پژوهش که ممکن است به علت نوع طرح پژوهشی رخ داده باشد. به عنوان مثال، در مطالعه‌ای برای جمع‌آوری داده‌های موردنظر، نمونه‌گیری در دو مرحله انجام می‌شود. در مرحله اول، مقادیر داده‌هایی که اندازه‌گیری آن‌ها آسان و ارزان است، جمع‌آوری می‌شود و سپس در مرحله دوم، مقادیر داده‌هایی که اندازه‌گیری آن‌ها پرهزینه و پیچیده است،. در مطالعه جمع‌آوری می‌شود. بنابراین جزئیات در مطالعه موجود نیست. در برخی از مطالعات ممکن است برخی از اطلاعات در دسترس نباشد. همچنین این احتمال هم وجود دارد که به علت نقص یا ضعف دستگاه و تجهیزات، امکان مشاهده و اندازه‌گیری وجود نداشته باشد یا در برخی از مطالعات نظرسنجی، افراد قادر به اظهار دقیق جواب نباشند یا به هر دلیلی داده جمع‌آوری‌شده برای افراد شرکت‌کننده در مطالعه گم شود. وجود مقادیر گمشده، دقت محاسبه‌شده را کاهش می‌دهد. مقادیر گمشده از مشکلات معمولی در تجزیه و تحلیل داده‌های مجموعه داده محسوب می‌شوند. نتایج حاصل از تحلیل داده‌های ناقص می‌تواند به نتایج نادرست منجر شود؛ بنابراین اهمیت دارد که تحلیل این نوع داده‌ها در مسیری مناسب و صحیح قرار داده شود تا بتوان به‌راحتی در بسیاری از زمینه‌ها، تصمیمات مهم و کاربردی را اتخاذ کرد.

### ۲-۱- روش‌های برخورد با گمشدگی

وجود مقادیر گمشده، دقت محاسبه‌شده را کاهش می‌دهد. مقادیر گمشده، از مشکلات معمولی در پردازش داده‌های مجموعه داده محسوب می‌شود. با توجه به اینکه نتایج حاصل از پردازش و تحلیل داده‌های ناقص می‌تواند به نتایج نادرست منجر شود، اهمیت دارد که پردازش این نوع داده‌ها در مسیری مناسب و صحیح قرار داده شود. با توجه به اینکه گمشدگی، تهدیدی جدی برای درستی نتایج حاصل از تجزیه و تحلیل داده‌ها به شمار می‌آید، بررسی روش‌های برخورد با آن، اهمیت زیادی دارد. یکی از مشکلات عمومی

1.Iterative Dichotomiser 3
2.Classification And Regression Tree
3.Classifire

در جمع‌آوری اطلاعات، جمع‌آوری داده‌های مجموعه داده است. اگر مجموعه داده‌های دارای مقادیر گمشده به‌صورت منظم رخ دهند، نتایج، گمراه‌کننده خواهند بود و دقت محاسبات را به‌دلیل از دست‌دادن بخش عمده‌ای از اطلاعات کاهش می‌دهند. امروزه بسیاری از پژوهشگران، روش‌هایی را مبنی بر تحلیل گمشدگی ارائه داده‌اند که در قسمت‌های بعد به مرور و بررسی هریک از روش‌ها خواهیم پرداخت.

## ۳- روش‌های مختلف گمشدگی

در تحلیل داده‌های تولیدشده از مجموعه داده‌های دارای چند خصوصیت، با داده‌های گمشده، توجه به ساختار گمشدگی داده‌ها، از اهمیتی ویژه برخوردار است. داده‌های گمشده به سه دسته تقسیم می‌شوند. هریک از دسته‌ها ویژگی‌ها و شرایط خاصی از گمشدگی مربوط به داده‌های موجود در مجموعه داده را نشان می‌دهند.

### ۳-۱- گمشدگی کاملاً تصادفی

در چنین حالتی گفته می‌شود که گمشدگی داده‌ها به‌صورت کاملاً تصادفی است. برای داده‌هایی که گمشدگی آن‌ها به‌صورت کاملاً تصادفی است، در این حالت، احتمال آنکه مقدار خصوصیت $X_i$ گمشده باشد، مستقل از مقدار $X_i$یا سایر خصوصیت‌های مربوط است. در واقع این نوع گمشدگی، زمانی اتفاق می‌افتد که احتمال مشاهده نشدن مقدار در یک زمان به هیچ‌یک از مقدارهای مشاهده‌شده و مشاهده‌نشده (قبلی و بعدی) بستگی نداشته باشد. محققان، برخی از روش‌ها را برای سنجش گمشدگی کاملاً تصادفی ارائه داده‌اند [۴و۵]. لیتل [۶]، روشی جهت سنجش گمشدگی کاملاً تصادفی برای تمام خصوصیت‌های مجموعه داده‌ها ارائه کرده است.

### ۳-۲- گمشدگی تصادفی

زمانی که توزیع گمشدگی به داده‌ها وابسته باشد، گمشدگی از نوع گمشدگی تصادفی خواهد بود. احتمال گمشدگی خصوصیت $X_i$ مستقل از $X_i$ یا سایر خصوصیت‌های موجود در مجموعه داده است. فلایس و همکاران در [۷]، روشی برای تشخیص گمشدگی تصادفی ارائه داده‌اند.

### ۳-۳- گمشدگی غیرتصادفی

مکانیسم گمشدگی مقادیر خصوصیت Y را گمشدگی غیرتصادفی گویند، هرگاه شرط خصوصیت X و Y وابسته به Y باشد. تشخیص گمشدگی غیرتصادفی، بسیار مشکل است؛ زیرا نیازمند دانستن خود مقادیر گمشده در تمام مجموعه داده خواهد بود. گمشدگی غیرتصادفی مربوط به داده‌های گسسته است [۸].

## ۴- مروری بر روش‌های تخمین داده گمشده

روش‌ها و راهکارهای گوناگونی در زمینهٔ پیش‌بینی مقدار داده گمشده ارائه شده است که در این قسمت، به ذکر و توصیف روش‌های ارائه‌شده پرداخته می‌شود.

در [۹]، برخی از مقادیر در مجموعه داده، دارای مقایر داده گمشده هستند. وجود این مقادیر در محاسبهٔ بهره اطلاعات و نرخ بهره موجب ناپایداری می‌شود. در این مقاله با عنوان Similarity Method، برای تعیین مقدار داده گمشده از روش شباهت به سایر داده‌های موجود در مجموعه داده استفاده کرده، آن را تخمین می‌زند.

در [۱۰]، از روش نزدیکی و تکرار برای به دست آوردن مقادیر داده گمشده استفاده کرده است. بدین‌صورت مقدار داده گمشده برابر با مقدار داده‌ای است که آن مقدار داده بیش از سایر مقادیر موجود در مجموعه داده تکرار شده باشد. در نهایت، به مقایسه بهره اطلاعات و نرخ بهره پرداخته، در نتیجه دقت درخت تصمیم تولیدشده توسط الگوریتم C4.5 را افزایش می‌دهد.

در [۱۱]، مقدار خصوصیتی را که در یک مجموعه داده بیش از سایر مقادیر تکرار شده است، به‌عنوان مقدار داده گمشده انتخاب می‌کند. روش مذکور با عنوان Missing Value For Decision Tree، ساده‌ترین روش در بین روش‌های ارائه شدهٔ پیشین است و موجب افزایش دقت درخت تصمیم تولیدشده توسط الگوریتم C4.5 می‌شود. شایان ذکر است که الگوریتم CN2 نیز، با استفاده از روش مذکور، قادر به پیش‌بینی مقدار داده گمشده است.

در [۱۲]، پیش‌بینی مقدار داده گمشده بر اساس ترکیب دو الگوریتم C4.5 و الگوریتم LEM2 [۱۳][1] است. در این روش با عنوان LC4.5 در تولید درخت تصمیم با استفاده از قوانین الگوریتم LEM2، آن دسته از داده‌هایی که فاقد مقدار هستند نیز حذف می‌شوند. در این روش، در صورتی که تعداد داده‌های موجود در مجموعه داده کم باشد، کارایی چندانی ندارد، در حالی که برای مجموعه داده‌هایی با تعداد داده‌های زیاد، بسیار مناسب است و دقت را افزایش می‌دهد.

[1].Learnable Evolution Model

در [۱۴]، پیش‌بینی مقدار داده گمشده بر اساس مجموعه‌ای از مقادیر موجود با عنوان Existing Value است. در این روش، هر داده فاقد مقدار، توسط سایر مقادیر موجود در مجموعه داده، پیش‌بینی می‌شود. بدین صورت با توجه به ترتیب مقادیر مشخص‌شده در مجموعه داده، هریک از داده‌های فاقد مقدار، به‌ترتیب، مقداردهی می‌شوند. دقت به‌دست‌آمده در این روش، نسبت به سایر روش‌های ارائه‌شدهٔ پیشین کمتر است.

در [۱۵]، پیش‌بینی مقدار داده گمشده بر اساس نادیده گرفتن خصوصیت‌هایی شامل داده‌های فاقد مقدار در مجموعه داده با عنوان Ignore Attributes است. در این روش، آن دسته از خصوصیاتی را که حدقل دارای یک داده فاقد مقدار هستند، نادیده می‌گیرد و از باقی‌مانده خصوصیت‌های موجود در مجموعه داده، برای طبقه‌بندی درخت تصمیم الگوریتم C4.5 استفاده می‌کند. دقت به‌دست‌آمده در این روش، همانند روشExisting Value، نسبت به سایر روش‌های ارائه‌شده پیشین، کمتر و نسبت به روش Existing Value، بیشتر است.

در [۱۶]، برخی از مقادیر در مجموعه داده، دارای مقایر داده گمشده هستند. با توجه به مجموعه داده موردنظر، از روش حذف برای تخمین مقدار داده گمشده استفاده می‌کند و ردیف‌های شامل داده گمشده را از مجموعه داده حذف می‌کند. حذف ردیف‌های شامل داده گمشده، موجب حذف اطلاعات از مجموعه داده می‌شود.

در [۱۷]، از روش میانگین برای تخمین مقدار داده گمشده استفاده می‌کند. بدین صورت با توجه به مقادیر موجود در مجموعه داده، تمام مقادیر موجود در هر خصوصیت با یکدیگر جمع می‌شوند و سپس از طریق میانگین، حاصل جمع آن‌ها بر تعداد کل داده‌های موجود در آن خصوصیت، تقسیم شده، در نهایت، مقدار حاصل‌شده توسط میانگین، به‌عنوان مقدار داده تخمینی برای داده گمشده، در خصوصیت موردنظر جایگزین می‌شود.

## ۵- روش پیشنهادی جابه‌جایی خصوصیت‌ها

در یک مجموعه داده، هریک از خصوصیات می‌توانند مقادیر داده گمشده داشته باشند. در روش پیشنهادی با عنوان روش جایگزینی خصوصیت‌ها ابتدا با توجه به مجموعه داده موردنظر، آن دسته از خصوصیت‌هایی را که دارای مقادیر داده‌گمشده هستند، در نظر گرفته، سپس تمام سطرهای تشکیل‌دهنده مجموعه داده، شامل داده گمشده، از مجموعه داده حذف می‌شوند. در مرحله بعد، آن دسته از خصوصیت‌هایی را که در ابتدا دارای مقادیر داده گمشده بودند، به‌عنوان خصوصیت کلاس مجموعه داده قرار داده می‌شوند و به‌ازای هریک از آن خصوصیات در نظر گرفته شده به‌عنوان کلاس مجموعه داده، یک درخت تصمیم ساخته می‌شود.

شکل (۱) شبه الگوریتم روش تخمین داده گمشده را با توجه به کلاس‌های درخت تصمیم نشان می‌دهد.

***Algorithm1: Estimating Missing Values According to Multiple Decision Tree***

*1. Input:*
*1.1 Dataset (D), Attributes = {1,2,..,N}*
*1.2 Decision Attribute = Last column of dataset(D)*
*2. For each Attribute $A_i$*
*2.1 If $A_i$ has "Missing Values" then*
*Decision Attribute = $A_i$*
*2.2 Call C4.5 (D) to create a Decision tree $DT_i$ (C4.5) based on $A_i$ as the DecisionAttribute and other attributes as usual attributes*
*2.3 End if*
*2.4 End for // after performing this loop we may have several C4.5 trees ($DT_i$, …,$DT_j$)//*
*3.Estimating the empty fields (Missing Values) of the dataset based on the createddecision trees ($DT_i$, …, $DT_j$) to achieve fulldataset*

شکل ۱- شبه الگوریتم ارائه یک روش جدید برای تخمین مقادیر گمشده در مجموعه داده

روش کار شبه الگوریتم فوق به این صورت است که در ابتدا بعد از ورود یک مجموعه داده اگر مجموعه داده موردنظر دارای مقادیر گمشده باشد، تمام ردیف‌های شامل مقدار گمشده از مجموعه داده حذف شده، یک مجموعه داده فاقد مقدار داده گمشده ایجاد می‌شود. با توجه به مجموعه داده اولیه که دارای مقدار داده گمشده بوده، خصوصیت‌های شامل مقدار داده گمشده با خصوصیت تصمیم جایگزین می‌شوند و در خط 2.2، با فراخوانی الگوریتم C4.5، یک درخت تصمیم ساخته می‌شود. این روند مرتباً به‌ازای هریک از خصوصیت‌های شامل مقدار داده گمشده تکرار می‌شود و در نهایت، هریک از مقدارهای به‌دست‌آمده از درخت‌های تصمیم، در مجموعه داده اولیه که شامل مقدار داده گمشده

است، جایگزین می‌شوند و نهایتاً یک مجموعه داده کامل و فاقد مقدار داده گمشده به دست خواهد آمد که طبق این مجموعه داده کامل، یک درخت تصمیم واحد و یکتا ایجاد می‌شود.

به‌عنوان مثال، جدول ۱، قسمتی از مجموعه داده بزرگ و واقعی برگرفته‌شده از پایگاه داده مرکزی UCI به نام مجموعه داده اعداد با چهار خصوصیت A, B, C و خصوصیت تصمیم CLASS را نشان می‌دهد. در این مجموعه داده برخی از مقادیر موجود در خصوصیت‌های B و C دارای مقادیر داده گمشده هستند که این مقادیر با علامت ؟ در مجموعه داده مشخص شده‌اند.

جدول۱- مجموعه داده دارای مقادیر داده گمشده

| *Index* | *A* | *B* | *C* | *Class* |
|---|---|---|---|---|
| ۱ | ۱ | ۲ | ۲ | ۱ |
| ۲ | ۱ | ۲ | ۲ | ۱ |
| ۳ | ۲ | ? | ۲ | ۱ |
| ۴ | ۳ | ۲ | ? | ۲ |
| ۵ | ۳ | ۱ | ۱ | ۲ |
| ۶ | ۳ | ۱ | ? | ۱ |
| ۷ | ۲ | ۱ | ۲ | ۲ |
| ۸ | ۳ | ۲ | ? | ۲ |
| ۹ | ۲ | ? | ۱ | ۱ |
| ۱۰ | ۳ | ۱ | ۱ | ۲ |

طبق روش پیشنهادی، ابتدا با توجه به مجموعه داده موردنظر، تمام ردیف‌های دارای مقادیر داده گمشده از مجموعه داده حذف می‌شوند. از آن جایی که در مجموعه داده اولیه دو خصوصیت B و C دارای مقادیر گمشده هستند، در این قسمت طبق روش پیشنهادی در هر زمان یکی از خصوصیت‌های B و C به‌عنوان خصوصیت تصمیم یا CLASS مجموعه داده اعداد در نظر گرفته شده، یک مجموعه داده جدید با خصوصیت تصمیم B و C ایجاد می‌شود و در نهایت، طبق الگوریتم C4.5، درخت تصمیم متناظر با آن مجموعه داده را می‌سازند. جدول ۲ مجموعه داده کامل و بدون مقادیر داده گمشده اعداد را نشان می‌دهد که در این مجموعه داده مقادیر مربوط به خصوصیت B و خصوصیت تصمیم، جایگزین یکدیگر می‌شوند، به گونه‌ای که خصوصیت B، به‌عنوان خصوصیت تصمیم در نظر گرفته و با استفاده از الگوریتم C4.5 درخت تصمیم ساخته می‌شود.

جدول۲- مجموعه داده اعداد با جایگزینی مقادیر خصوصیت B به‌عنوان خصوصیت تصمیم

| Index | A | B | C | CLASS |
|---|---|---|---|---|
| ۱ | ۱ | ۱ | ۲ | ۲ |
| ۲ | ۱ | ۱ | ۲ | ۲ |
| ۳ | ۳ | ۲ | ۱ | ۱ |
| ۴ | ۲ | ۲ | ۲ | ۱ |
| ۵ | ۳ | ۲ | ۱ | ۱ |

علاوه بر جدول فوق، جدول ۳ نیز، مجموعه داده کامل و بدون مقادیر داده گمشده اعداد را نشان می‌دهد که در این مجموعه داده، تمامی مقادیر مربوط به خصوصیت C و خصوصیت تصمیم، جایگزین یکدیگر شده‌اند و خصوصیت C به‌عنوان خصوصیت تصمیم، در نظر گرفته می‌شود. طبق مجموعه داده جدید به‌دست‌آمده، با استفاده از الگوریتم C4.5 درخت تصمیم طبق مجموعه داده جدید ساخته می‌شود.

جدول۳- مجموعه داده اعداد با جایگزینی مقادیر خصوصیت C به‌عنوان خصوصیت تصمیم

| Index | A | B | C | CLASS |
|---|---|---|---|---|
| ۱ | ۱ | ۱ | ۲ | ۲ |
| ۲ | ۱ | ۱ | ۲ | ۲ |
| ۳ | ۳ | ۲ | ۱ | ۱ |
| ۴ | ۲ | ۲ | ۱ | ۲ |
| ۵ | ۳ | ۲ | ۱ | ۱ |

پس از تولید درختان تصمیم توسط الگوریتم C4.5، مقادیر حاصل‌شده از هریک از درختان تصمیم در مجموعه داده اولیه که دارای مقدار داده گمشده بوده است نیز جایگزین می‌شود. در نهایت، یک مجموعه داده کامل و بدون داده گمشده به دست می‌آید که طبق آن، با استفاده از الگوریتم C4.5، یک درخت تصمیم نهایی برای پیش‌بینی و تقسیم‌بندی حاصل می‌شود.

## ۶- مقایسه دقت روش پیشنهادی با روش‌های حذف و میانگین

دقت [۱۸]، یک معیار کلی اندازه‌گیری برای تعیین کمیت یک سیستم یادگیری است. دقت به‌عنوان یکی از مهم‌ترین و مشهورترین معیارها برای تعیین کارایی یک الگوریتم دسته‌بندی است. درواقع، این معیار مشهور، بیانگر این است که الگوریتم دسته‌بندی مربوط، چند درصد از کل داده‌های مجموعه داده را به‌درستی دسته‌بندی کرده است. دقت با استفاده از فرمول زیر محاسبه می‌شود.

$$\text{Precision} = \frac{\text{True Samples}}{\text{Total Samples}} \qquad (۱)$$

True Samples: تعداد داده‌های درست طبقه‌بندی‌شده در مجموعه داده.

Total Samples: تعداد کل داده‌های یک مجموعه داده.

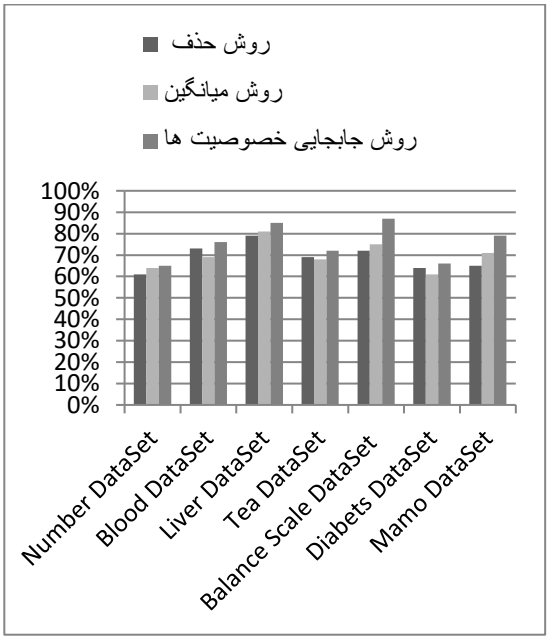
دقت یک درخت تصمیم با استفاده از یک مجموعه با عنوان مجموعه داده تست یا با استفاده از تکنیک‌های تخمین از جمله اعتبارسنجی، محاسبه و سنجیده می‌شود. پس از کامل کردن مجموعه داده و پیش‌بینی مقدار داده گمشده، برای محاسبه دقت درخت تصمیم، داده‌های موجود در مجموعه داده با نسبت ۷۰٪ به نسبت ۳۰٪، به دو قسمت با عنوان مجموعه داده یادگیری و مجموعه داده تست تقسیم می‌شود.

از مجموعه داده یادگیری برای یادگیری الگوریتم، ساخت مدل و تولید درخت تصمیم و از مجموعه داده تست برای برآورد درستی تولید درخت تصمیم و نیز تست و ارزیابی مدل ساخته‌شده استفاده می‌شود. این روند همچنان تکرار می‌شود تا زمانی که کل مجموعه داده، برای داده‌های یادگیری و داده‌های تست به کار گرفته شود. جدول ۴ به توصیف مشخصات مربوط به مجموعه داده‌های موردآزمایش از جمله تعداد داده، تعداد خصوصیت، نوع خصوصیت و نیز داده گمشده برای سنجش معیار دقت می‌پردازد. از آن جایی که الگوریتم C4.5 قابلیت دسته‌بندی داده‌های پیوسته و گسسته را دارد، ویژگی تعداد خصوصیت دارای دو مقدار پیوسته و گسسته است.

شکل (۳)، نمودار دقت به‌دست‌آمده از روش جابه‌جایی خصوصیت‌ها و نیز روش‌های میانگین و حذف را نشان می‌دهد. برای ارزیابی دقت، از مجموعه داده‌های واقعی UCI استفاده شده است.

جدول۴- مجموعه داده‌های موردآزمایش

| مجموعه داده | مشخصات مجموعه داده | | | |
|---|---|---|---|---|
| | تعداد داده | تعداد خصوصیت | نوع خصوصیت | داده گمشده |
| *Number DataSet* | ۳۰۰ | ۴ | پیوسته | بله |
| *Blood DataSet* | ۷۴۸ | ۴ | گسسته | بله |
| *Liver DataSet* | ۳۴۵ | ٦ | گسسته | بله |
| *Tea DataSet* | ۱۲۵ | ۵ | پیوسته | بله |
| *Balance Scale DataSet* | ۶۲۵ | ۴ | پیوسته | بله |
| *Diabets DataSet* | ۷۶۸ | ۸ | گسسته | بله |
| *Mamo DataSet* | ۹۶۱ | ۶ | پیوسته گسسته | بله |




شکل ۳- نمودار دقت روش جابه‌جایی خصوصیت‌ها، روش حذف و روش میانگین

طبق نمودار فوق، دقت حاصل‌شده در روش جایگزینی خصوصیت‌ها نسبت به روش‌های میانگین و حذف، بیشتر است. بنابراین، تخمین مقادیر گمشده در مجموعه داده‌ها از طریق روش پیشنهادشده در این مقاله، در حل مسائل و اتخاذ تصمیمات، کاربرد و مهم‌تر از آن، کارایی بیشتری دارد.

## ۷- نتیجه‌گیری

در حال حاضر، راهکارهای گوناگونی توسط پژوهشگران در زمینهٔ پیش‌بینی مقدار داده گمشده ارائه شده است. یکی از راهکارهای موجود، استفاده از الگوریتم C4.5 است که به‌دلیل کارایی، استفاده در کاربردهای مختلف داده‌کاوی و نیز توانایی در کارکردن و پیش‌بینی مقدار داده گمشده، به طور مکرر مورداستفاده قرار گرفته است. با به‌کارگیری الگوریتم C4.5 و نمونه داده‌های موجود، می‌توان مقدار داده گمشده موجود در مجموعه داده را پیش‌بینی کرد. بنابراین، در این مقاله، به پیشنهاد یک روش جدید برای تخمین مقادیر گمشده در مجموعه داده با عنوان جابه‌جایی خصوصیت‌ها پرداخته شد. پس از اجرا و پیاده‌سازی آن، دقت حاصل‌شده نشان داد که همواره روش پیشنهادی در این مقاله نسبت به روش تخمین داده گمشده طبق روش حذف و روش میانگین، دقت بالاتری دارد.

## مراجع


[1] R. Bhardwaj and S. Vatta, "Implemention of ID3 Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 82, 2013, pp. 317–329.

[2] S. Gey, E. Nedelec, "Model Selection for CART regr ession trees", Information Theory IEEE Transactions, Vol.51, Issue 2, 2005.

[3] H, Zhu and S. Chen, “Rang Tree:A Feauter Selection Algoritm for C4.5 Decission Tree”, 5th International Conference on Intelligent Networking and Collaborative Systems (INCoS), 2013.

[4] M.H. ,Chen and S.R. Lipsitz, "Bayesian methods for generalized linear models with covariates missing at random", Canadian Journal of Statistics,2008.

[5] T. Marwala, “Computational Intelligence for Missing Data Imputation, Estimation and Management: Knowledge Optimization Techniques”, South Africa University of Witwatersrand IGI Global, 2009.

[6] R.J.A. Little, "A test of missing completely at random for multivariate data with missing values", Journal of the American Statistical Association, 83, 1988, PP. 1198-1202.

[7] J.L. Fleiss, B. Levin and M.C Paik, “Statistical Methods for Rates and Proportions”, 3rd International New York, 2002.

[8] C. Priyadharsini, P.Selvadoss, "Prediction of Missing Values in Blood Cancer and Ocuurrence of Cancer Using Improved ID3 Algorithm", International Journal of Innovative Research in Computer and Communication Engineering, Vol.2, Issue 8, 2014.

[9] J. Kaiser, "Dealing with Missing Values in Data", Journal of Systems Integration, Vol.6, Issue 10, 2014.

[10] M. Augustin and S. Sakena, “Machine Learning with Missing Attributes Value Methods Implementation”, Procedding of the Worid Congress on Engineering and Computer, 2015.

[11] T. Chen, “A comparison of approaches for dealing with Missing Values”, proceedings of the international conference on machine learning and cybernetics, 2015.

[12] Li. Huaxiong, "Missing Values Imputation Based on Iterative Learning", International Journal of Intelligence Science, 3, 2013, pp. 50-55

[13] P. Clark and T. Niblett, "The LEM2 and C4.5 Inductiion Algorithm Machine", International Journal software Computer, 2012.

[14] W. Jerzy, "A Comparison of Rule Induction Using Feature Selection and the LEM2 Algorithm", Springer Verlag Berlin Heidelberg, 2015.

[15] J. Grzymala, “On the unknown attribute values in learning from examples”, 6th International Symposium on Methodologies for Intelligent systems, Charlotte, 2014.

[16] J. Quinlan, C4.5 Programs for Machine Learning, MorganKaufman Publishers, San Matteo CA, 2014.

[17] A. Sharma and N. Mehta, "Reasoning with Missing Values in Multi Attribute Datasets", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013, pp. 1035-1043.

[18] A. Jeerz, Coordinate Metology, Acuuracy of Systema and Measurements Series Springer Tracts in Mechanial Enginnering, 2016.