International Conference on
Management By Intelligence
Business and organizational Intelligence / Tehran university /13 july 2017

دومین کنفرانس سالانه
مدیریت بر مبنای هوشمندی
پارادایم‌های نوین مدیریت در حوزه هوشمندی / دانشگاه تهران / ۲۲ تیر ۱۳۹۶

به‌کارگیری امکانات تعبیه شده در نرم افزار هوش تجاری Tableau به منظور داده‌کاوی

هدا مرادیان ۱

چکیده

داده‌کاوی تلاش برای استخراج دانش از انبوه داده‌های موجود است. فرآیند داده‌کاوی به کمک مجموعه‌ای از روش‌های آماری و مدل‌سازی، می‌تواند الگوها و روابط پنهان موجود در پایگاه‌های داده را تشخیص دهد. تاکنون ابزارها و روش‌های مختلفی برای پردازش داده‌های ساخت‌یافته توسعه داده شده است که در نتیجه آن‌ها، ساخت پایگاه‌های داده و ایجاد انباره‌های داده به سادگی صورت می‌گیرد. در این مقاله قصد داریم از کتابخانه‌های نرم‌افزار داده‌کاوی R در محیط نرم افزار هوش تجاری Tableau استفاده نماییم. در نتیجه، ابتدا مقدمه‌ای راجع به داده‌کاوی ارائه نموده و سپس به ارائه شرحی مختصر در رابطه با دو نرم افزار Tableau و R می‌پردازیم. پس از آن، درباره به‌کارگیری برخی از روش‌های مرسوم داده‌کاوی مانند رده‌بندهای درخت تصمیم CART و Random Forest توضیحاتی ارائه می‌نماییم و بصورت عملی با کمک نرم‌افزارهای اشاره شده با چگونگی استفاده از این الگوریتم‌ها در Tableau آشنا خواهیم شد. در نهایت، جمع‌بندی این مقاله در بخش نتیجه ارائه خواهد شد.

کلمات کلیدی

داده‌کاوی، رده بندی، Tableau، R، درخت تصمیم CART، Random Forest

Utilizing Embedded Tools in Tableau Business Intelligence Software for Data mining

Hoda Moradian

MSc in Software Engineering, ICT Department, Water and Wastewater Company, Esfahan, Iran

ABSTRACT

Data Mining is an attempt to extract knowledge from tremendous amount of available data. This process can detect hidden patterns and relationships in databases using a set of statistical methods and modeling. So far, different methods have been developed to process structured data, thus, facilitating the creation of databases and data warehouses. In this paper, we intend to utilize R data mining software libraries in Tableau Business Intelligence environment. As a result, at first, an introduction of data mining is offered and then, a brief description of both Tableau and R is provided. After that, we offer an explanation about utilizing some conventional data mining methods such as CART and Random Forest classifiers, and then we will become practically familiar with how to use these algorithms in Tableau. Finally, the conclusion of this paper will be presented in the summary section.

KEYWORDS

Data mining, Classification, Tableau, R, CART Decision Tree, Random Forest.

۱. مقدمه

به علت استفاده همگانی از وب و اینترنت، دنیای اطلاعات با حجم زیادی از داده‌ها و اطلاعات مواجه شده است. این رشد انفجاری در داده‌های ذخیره شده، نیاز مبرم به وجود فناوری‌های جدید و ابزارهای خودکاری را ایجاد کرده است که تا از طریق تبدیل این حجم زیاد داده به اطلاعات و دانش، به صورت هوشمند به انسان یاری رسانند. در نتیجه، داده‌کاوی به عنوان مهمترین فناوری جهت بهره‌برداری موثر از این داده‌های بزرگ[۱] مورد توجه قرار گرفته و اهمیت آن رو به فزونی است. داده‌های بزرگ علیرغم حجیم بودن، اغلب بدون ارزش می باشند چرا که به تنهایی قابل استفاده نبوده و این دانش نهفته در این نوع داده‌هاست که قابل استفاده می باشد. در حال حاضر داده‌کاوی به عنوان یک راه‌حل اساسی برای چالش‌های مربوط به داده‌های بزرگ مطرح شده است. در یک تعریف غیر رسمی داده‌کاوی فرآیندی است خودکار برای استخراج الگوهایی که دانش را بازنمایی می کنند، که این دانش به صورت ضمنی در پایگاه داده‌های عظیم، انباره داده و دیگر مخازن بزرگ اطلاعات، ذخیره شده است.

داده‌کاوی بطور همزمان از چندین رشته علمی نظیر فناوری پایگاه داده‌ها، هوش مصنوعی، یادگیری ماشین، شبکه‌های عصبی، آمار، شناسایی الگو، سیستم‌های مبتنی بر دانش[۲]، اکتساب دانش[۳]، بازیابی اطلاعات[۴] و مصورسازی داده[۵] بهره می‌برد. داده‌کاوی در اواخر دهه ۱۹۸۰ پدیدار گشته، در دهه ۱۹۹۰ گام‌های بلندی در این شاخه از علم برداشته شد و انتظار می رود در این قرن با سرعتی فزاینده به رشد و پیشرفت خود ادامه دهد[۲].

همانگونه که پیش از این نیز بیان شد، پیشرفت و تکامل فناوری‌های پایگاه داده و استفاده فراوان آن‌ها در کاربردهای مختلف سبب جمع‌آوری حجم انبوهی از داده‌ها شده است. شکاف موجود بین داده‌ها و اطلاعات، ضرورت وجود ابزارهایی خاص جهت انجام داده‌کاوی را باعث شده است تا بدین بوسیله، داده‌های بی ارزش به دانشی ارزشمند تبدیل گردند. ابزارهای داده‌کاوی داده‌ها را تحلیل کرده و الگوهای موجود در داده‌ها را کشف می‌کنند. در نتیجه می‌توان از این الگوها در کاربردهایی نظیر تعیین راهبردهای کسب‌وکار، پایگاه دانش، تحقیقات علمی و پزشکی و دیگر موارد استفاده نمود. در بخش زیر به مراحل کشف دانش از طریق داده‌کاوی می‌پردازیم.

۱،۱. مراحل کشف دانش

کشف دانش دارای مراحل تکراری زیر است:

1. پاکسازی داده‌ها (از بین بردن نویز و ناسازگاری داده‌ها)
2. یکپارچه سازی داده‌ها (چندین منبع داده ترکیب می‌شوند)
3. انتخاب داده‌ها (داده های مرتبط با آنالیز از پایگاه داده بازیابی می‌شوند)
4. تبدیل‌کردن داده‌ها (تبدیل داده‌ها به فرمی که مناسب برای داده‌کاوی باشد مثل خلاصه سازی و همسان سازی)

---

۱. کارشناس ارشد نرم افزار، دفتر فناوری اطلاعات و ارتباطات، شرکت آب وفاضلاب استان اصفهان، اصفهان، ایران hoda.moradian@gmail.com

5. داده‌کاوی (فرایند اصلی که روال‌های هوشمند برای استخراج الگوها از داده‌ها به کار گرفته می شوند)
6. ارزیابی الگو (برای مشخص کردن الگوهای صحیح و مورد نظر به‌وسیله معیارهای اندازه‌گیری)
7. ارائه دانش (نمایش بصری، فنون بازنمایی دانش برای ارائه دانش کشف شده به کاربر استفاده می شود)[۳].

شکل (۱) مراحل داده‌کاوی را به اختصار نشان می‌دهد.

ادامه مقاله به شرح زیر سازماندهی شده است:

بخش ۲ به معرفی و نحوه استفاده از ابزارهای داده‌کاوی و مجموعه داده‌های مورد استفاده پرداخته، بخش ۳ رده‌بندها را شرح می‌دهد، بخش ۴ نحوه انجام آزمایش‌ها و بدست آوردن نتایج را تشریح کرده و نهایتاً بخش ۵ مقاله را جمع‌بندی می‌کند.

۲. مبانی و چارچوب نظری تحقیق

۱،۲. نرم افزار هوش تجاری Tableau

Tableau، یک شرکت نرم‌افزاری آمریکایی واقع در سیاتل آمریکاست. این شرکت خانواده‌ای از محصولات مصورسازی تعاملی داده‌ها را که متمرکز بر روی هوشمندی کسب‌وکار است تولید میکند. ابزار Tableau از محبوب‌ترین محصولات درزمینه تجزیه وتحلیل داده‌ها و گزارش‌گیری می‌باشد که مبتنی بر وب بوده و با در اختیار داشتن قابلیتهای امنیتی بالا، اطلاعات جمع‌آوری‌شده از حوزه‌های مختلف را در داشبوردهای اطلاعاتی عرضه میکند. Tableau از فنون استاندارد داده‌کاوی بویژه پردازش اطلاعات، خوشه‌بندی، طبقه‌بندی، رگرسیون، مصورسازی و گزینش ویژگی پشتیبانی می‌کند.

مصورسازی داده‌ها در یک تعریف ساده، نمایش اعداد به‌صورت تصویر، نمودار یا نقشه با هدف درک روند آنها در یک نگاه است که با تغییر یک متغیر، تصویر نیز تغییر می‌کند. Tableau ازجمله شرکت‌هایی است که در زمینه مصورسازی داده‌ها فعالیت گسترده‌ای دارد [۱].

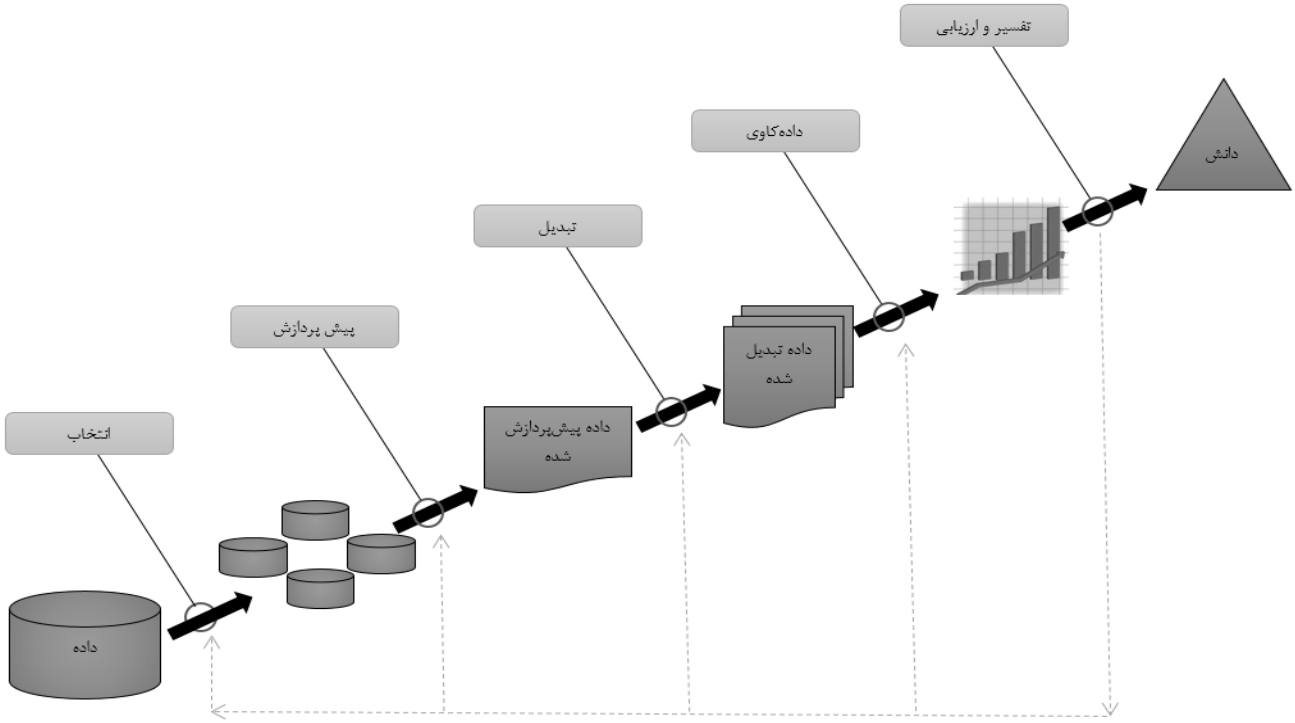
۲،۲. نرم‌افزار داده‌کاوی R

نرم‌افزار R ابزاری قدرتمند برای تجزیه و تحلیل داده‌های آماری است. اولین عاملی که آن را از سایر نرم‌افزارهای آماری متمایز می‌نماید، رایگان و منبع باز[۶] بودن آن است. R دارای یک محیط کدنویسی است و عملیات مختلف در آن بوسیله تایپ دستورات صورت می‌گیرد. برای هر کاری، یا هر روش آماری جدید و قدیمی، به احتمال زیاد یک بسته در نرم افزار R وجود دارد. این بسته‌ها توسط کاربران حرفه‌ای طراحی و در دسترس عموم قرار می‌گیرند[۱].

R حاوی طیف گسترده‌ای از فنون آماری از جمله مدل‌سازی خطی و غیرخطی، آزمون‌های کلاسیک آماری، تحلیل سری‌های زمانی، رده‌بندی، خوشه‌بندی و غیره می باشد. R همچنین نرم‌افزاری قدرتمند برای ایجاد اشکال گرافیکی و نمودارهاست.

در این پژوهش نیز ما از این نرم افزار بویژه دو بسته بسیار پر کاربرد آن یعنی Rserve وrpart [۱] به منظور پیاده سازی الگوریتم های یادگیری ماشین و ترسیم نمودارهای مربوطه استفاده خواهیم کرد.

در بخش بعدی پس از ارائه مقدمه ای اجمالی در رابطه با دو الگوریتم رایج داده‌کاوی، به منظور آشنایی با قابلیت های



شکل (۱): مراحل داده‌کاوی

داده‌کاوی نرم افزار R در محیط Tableau، دو نمونه الگوریتم داده‌کاوی را بر روی مجموعه داده Iris که در مخزن یادگیری ماشینی UCI[۶] موجود است بررسی می‌کنیم.

۳. روش شناسی تحقیق

۱،۳. رده بندها

در این مقاله به منظور انجام داده‌کاوی از دو الگوریتم CART و Random Forest استفاده‌شده است. این دو الگوریتم جهت داده‌کاوی در پژوهش های زیادی مورد استفاده قرارگرفته‌اند[۴, ۵].

بلوک‌بندی بازگشتی یکی از ابزارهای اساسی در داده‌کاوی می‌باشد. این ابزارها برای کاوش ساختار مجموعه داده، به‌منظور توسعه راحت قواعد تصمیم‌گیری برای پیش‌بینی نتایج گسسته(درخت طبقه بند) یا پیوسته(درخت رگرسیون) به ما کمک می‌کنند.

هنگامی‌که داده‌ها ویژگی‌های زیادی دارند و تعامل آن‌هابه‌صورت غیرخطی است(یعنی به‌صورت خطی جدایی‌پذیر نیستند)، آنگاه پیدا کردن یک فرمول رگرسیون کلی مانند فرمول پیش‌بینی که روی همه مجموعه داده به‌کار رود بسیار مشکل است.

یک روش جایگزین، بخش کردن فضا به نواحی کوچک‌تر و سپس تبدیل آن‌ها به قطعات کوچک‌تر(قطعه‌بندی بازگشتی) است تا زمانی که همه قطعات بزرگ، قابل شرح بوسیله یک مدل ساده باشند.

در این مقاله از مدل CART، جهت پیاده‌سازی‌های درخت رده‌بند و درخت رگرسیون استفاده گردید. همچنین علاوه بر الگوریتم CART، از روش جمعی[۷] Random Forest نیز استفاده شد و برای جلوگیری از بروز پدیده بیش برازش[۸] [۱] در این الگوریتم‌ها از هرس کردن استفاده شد. در ادامه به شرح این الگوریتم‌ها می‌پردازیم.

۱،۱،۳. درخت CART

الگوریتم CART ابتدا یک درخت پیچیده را تولید کرده و سپس ساختار درخت با توجه به اعتباردهی متقاطع و اعتباردهی مجموعه داده هرس می‌شود[۱].

CART به دلیل سادگی مطالعه درخت‌های تصمیم‌گیری، ابتکارات فنی ارائه شده بوسیله آن و مباحث پیچیده و پیشرفته داده‌های درختی، حائز اهمیت است.

شبه کد نحوه عملکرد الگوریتم CART در Tableau به شرح زیر می‌باشد:

```
[۱] SCRIPT_STR
{
[۲] LIBRARY (RPART);
[۳] DETERMINE TARGET VARIABLE & PREDICTOR VARIABLES;
[۴] METHOD OF ALGORITHM IS CLASSIFICATION;
[۵] READ DATA AND TRAIN ALGORITHM USING DATASET;
[۶] PRUNE TREE & PREVENT FROM OVERFITING;
[۷] TEST DATA USING CREATED TREE IN ABOVE;
}
```

شکل(۲): شبه کد الگوریتم CART

۲،۱،۳. Random Forest

Random Forest، به‌عنوان یکی از روش‌های یادگیری جمعی شناخته شده است. یک روش یادگیری جمعی، یادگیرهای منحصربه‌فرد زیادی را تولید کرده و نتایج را جهت رای‌گیری حداکثر جمع‌آوری می‌کند. در یک رده‌بند درخت تصمیم معمولی، تصمیم‌گیری در یک نود قابل انشعاب بر اساس همه ویژگی‌ها ساخته می‌شود. اما در Random Forest، بهترین پارامتر در هر گره موجود در درخت تصمیم به‌صورت تصادفی از یک تعداد ویژگی انتخاب‌شده ساخته‌می‌شود. این انتخاب تصادفی ویژگی‌ها نه تنها به داشتن یک معیار خوب هنگامی‌که تعداد زیادی بردار ویژگی وجود دارد کمک می‌کند بلکه به کاهش همبستگی میان ویژگی‌ها نیز کمک می‌کند و بنابراین در برابر نویزهای موجود در داده‌ها کمتر آسیب‌پذیرمی‌شود[۷].

شبه کد بکار رفته برای پیاده‌سازی Random Forest به‌صورت زیر است:

```
[۱]SCRIPT_STR
{
[۲]LIBRARY(RPART);
[۳]LIBRARY(RANDOM FOREST)
[۴]DETERMINE AIM VARIABLE & PREDICTOR VARIABLE
[۵]READ DATA AND TRAIN ALGORITHM USING DATASET;
[۶]PREDICT TEST DATA USING CREATED TREES;
}
```

شکل (۳): شبه کد الگوریتم Random Forest

در بخش بعدی ارزیابی این الگوریتم‌ها را بر روی مجموعه داده Iris بررسی می‌کنیم.

۴. تحلیل داده‌ها و یافته‌ها

۱,۴. درخت مربوط به الگوریتم CART

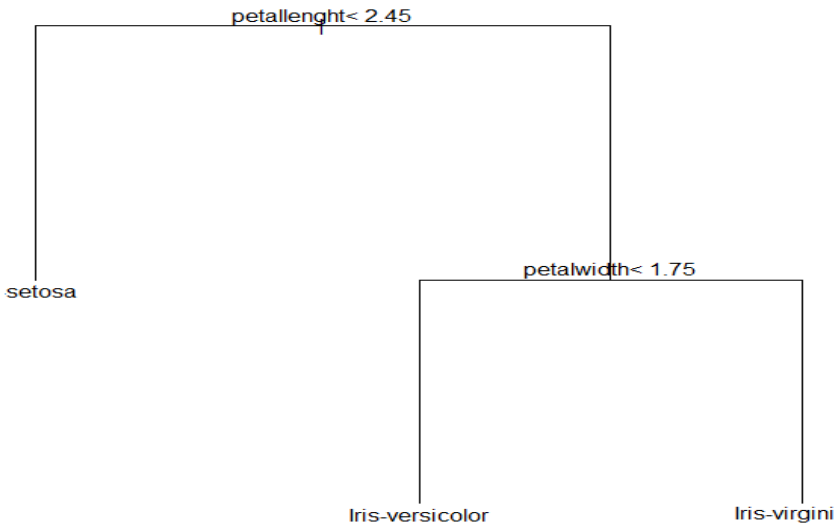
برای ایجاد درخت CART میتوان از بسته rpart استفاده نمود.

وارد محیط Tableau شده، اتصال Tableau را به R برقرار نموده و در قسمت انتخاب داده‌ها برای رده‌بندی از مجموعه داده Iris که برای تعیین رده گل‌های زنبق است استفاده می‌کنیم. این مجموعه داده دارای ۶ فیلد اطلاعاتی به نام‌های petallength (طول گلبرگ)، petalwidth (عرض گلبرگ)، sepallength (طول کاسبرگ)، sepalwidth (عرض کاسبرگ)، فیلد مهمی به نام class که مشخص کننده رده قرارگیری گل است که در این مجموعه داده گل‌ها در ۳ دسته (setosa, versicolor, virginca) قرار می‌گیرند و نهایتا یک فیلد به نام ID است.

رده بندی انجام شده در Tableau در شکل (۴) نمایش داده شده است. همچنین، همانطور که قبلا نیز بیان شد، از بسته rpart برای ایجاد درخت CART استفاده می شود که در شکل (۵) به تصویر کشیده شده است.

ID	petalleng..	sepallen..
1	1.400	5.100
2	1.400	4.900
3	1.300	4.700
4	1.500	4.600
5	1.400	5.000
6	1.700	5.400
7	1.400	4.600
8	1.500	5.000
9	1.400	4.400
10	1.500	4.900
11	1.500	5.400
12	1.600	4.800
13	1.400	4.800
14	1.100	4.300
15	1.200	5.800
16	1.500	5.700
17	1.300	5.400
18	1.400	5.100
19	1.700	5.700
20	1.500	5.100
21	1.700	5.400
22	1.500	5.100

شکل (۴): رده بندی انجام شده در Tableau توسط الگوریتم CART



شکل (۵): درخت ایجاد شده توسط الگوریتم CART در R

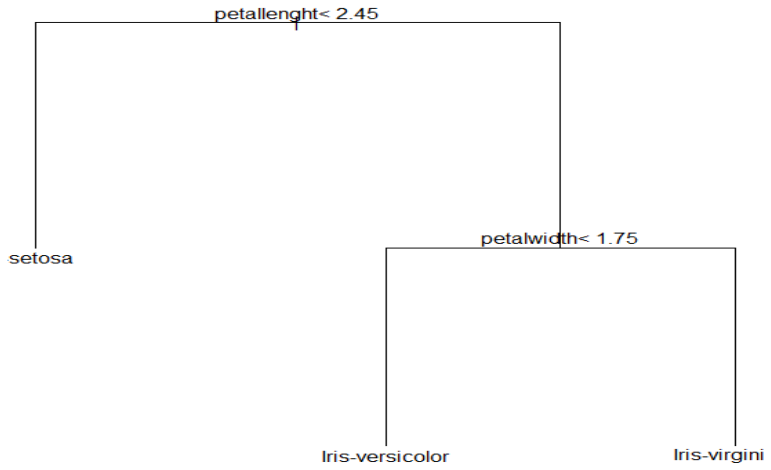
از همین مجموعه داده برای بررسی الگوریتم Random Forest نیز استفاده نمودیم که در بخش زیر بررسی می نماییم.

۲,۴.درخت مربوط به الگوریتم Random Forest

در این مقاله به‌منظور استفاده از الگوریتم Random Forest در محیط Tableau از کتابخانه randomForest استفاده‌شده است. همان‌طور که می‌دانیم، روش Random Forest برای پیش‌بینی، چندین درخت به‌صورت کاملاً تصادفی ساخته سپس بر اساس رأی‌گیری اکثریت بهترین پیش‌بینی را انجام می‌دهد. نتایج این الگوریتم در شکل‌های (۶) و (۷) نمایش داده شده که بسیار مشابه با نتایج الگوریتم CART است.

ID	Classification	class
42	Iris-setosa	Iris-setosa
43	Iris-setosa	Iris-setosa
44	Iris-setosa	Iris-setosa
45	Iris-setosa	Iris-setosa
46	Iris-setosa	Iris-setosa
47	Iris-setosa	Iris-setosa
48	Iris-setosa	Iris-setosa
49	Iris-setosa	Iris-setosa
50	Iris-setosa	Iris-setosa
51	Iris-versicolor	Iris-versicolor
52	Iris-versicolor	Iris-versicolor
53	Iris-versicolor	Iris-versicolor
54	Iris-versicolor	Iris-versicolor
55	Iris-versicolor	Iris-versicolor
56	Iris-versicolor	Iris-versicolor
57	Iris-versicolor	Iris-versicolor
58	Iris-versicolor	Iris-versicolor
59	Iris-versicolor	Iris-versicolor
60	Iris-versicolor	Iris-versicolor
61	Iris-versicolor	Iris-versicolor
62	Iris-versicolor	Iris-versicolor

شکل (۶): رده‌بندی انجام شده در Tableau توسط الگوریتم Random Forest



شکل (۷): رده بندی انجام شده در Tableau توسط الگوریتم Random Forest

۵. نتیجه‌گیری

این مقاله تلاشی بود در جهت آشنایی با فرآیند داده‌کاوی از طریق نرم افزار هوش تجاری Tableau، که این مهم با استفاده از امکان برقراری اتصال به نرم‌افزار داده‌کاوی R که به صورت تعبیه شده در Tableau موجود می‌باشد فراهم گردید. بدین منظور، پس از بیان مفاهیم کلی و معرفی ابزارها، پیاده سازی عملی الگوریتم‌های بیان شده را مشاهده کردیم. در این پژوهش از مجموعه داده Iris، برای پیاده سازی‌ها استفاده گردید. با توجه به گستردگی الگوریتم‌های داده‌کاوی که در این مقاله نمی‌گنجید، فقط از روش‌های رده‌بندی با استفاده از درخت تصمیم CART و Random Forest استفاده نمودیم و نتایج حاصله نیز به تصویر کشیده شد.

مراجع

[۱] مرادیان، هدا؛ محمدی، مسعود؛ فرهودی نژاد، اکبر؛ **" مقایسه تکنیک‌های داده‌کاوی برای پیش بینی ورشکستگی به وسیله فاکتورهای کیفی"** ، مجله بین‌المللی پژوهش‌های رایانه و فناوری اطلاعات و ارتباطات، شماره دوره ۲، شماره مجله ۴، شماره صفحه۲۷-۱۶، ۱۳۹۴.

[۲] Hand, D.J., Mannila, H. & Smyth, P., *Principles of data mining*. MIT press. ۲۰۰۱.

[۳] Fayyad, U. M., Wierse, A., & Grinstein, G. G. *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann. ۲۰۰۲.

[۴] Pecchia, L., Melillo, P., & Bracale, M., *Remote health monitoring of heart failure with data mining via CART method on HRV features*. Biomedical Engineering, IEEE Transactions on, ۵۸(۳), pp. ۸۰۰-۸۰۴. ۲۰۱۱.

[۵] Farooq, F., & Kidmose, P., *Random forest classification for p۳۰۰ based brain computer interface applications*. In Signal Processing Conference (EUSIPCO), ۲۰۱۳ Proceedings of the ۲۱st European, pp. ۱-۵. IEEE. ۲۰۱۳.

[۶] http://archive.ics.uci.edu/ml/datasets/

[۷] Criminisi, A., Shotton, J., & Konukoglu, E., *Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning*. Foundations and Trends® in Computer Graphics and Vision, ۷(۲–۳), pp. ۸۱-۲۲۷. ۲۰۱۲.

زیر نویس‌ها

[1] Big Data
[2] Knowledge-based Systems
[3] Knowledge Acquisition
[4] Information Retrieval
[5] Data Visualization
[6] Open Source
[7] Ensemble
[8] Over Fitting