طبقه‌بندی داده‌های حسابهای جاری بانک با استفاده ازدرخت تصمیم

شبنم خلیلی، دانشگاه تبریز - گروه مهندسی کامپیوتر - sh-khalili91@ms.tabriz.ac.ir
سعید پاشازاده، دانشگاه تبریز - گروه مهندسی کامپیوتر - pashazadeh@tabriz.ac.ir

چکیده: داده‌کاوی مجموعه‌ای از شیوه‌های عمدتاً آماری و بدیع با هدف استخراج اطلاعات جدید از مجموعه داده‌های بزرگ و تبدیل آن به یک ساختار قابل فهم برای استفاده بیشتر است. طبقه‌بندی داده‌ها و داده‌کاوی پیش‌بینی فرایندها دو کاربرد عمده داده کاوی است. طبقه‌بندی شکلی از تجزیه وتحلیل داده‌ها برای مدل‌های توصیفی است که کلاس‌های مهم اطلاعات را استخراج می‌کند. طبقه‌بندی یکی از اشکال تجزیه و تحلیل داده‌ها است که می‌تواند برای استخراج مدل‌های توصیفی کلاس‌های مهم اطلاعات و یا برای پیش‌بینی برچسب‌های قطعی مورد استفاده قرارگیرد. یکی از حوزه‌های کاربردی طبقه‌بندی داده‌ها دسته‌بندی وام‌های بانکی و صاحبان حساب‌های بانکی است. تجزیه و تحلیل داده‌های بانکی می‌تواند در تولید دانش لازم در مورد صاحبان حساب‌های بانک به مدیران بانک کمک کند. لذا بانک‌ها می‌توانند با استفاده از مزایای دانش نهفته در داده‌های موجود حساب-های، آنها را برای توصیف مشتریان خود استفاده کنند و سپس براساس رفتار مشتریان می‌توانند اقدامات مختلفی، از جمله نگهداری مشتریان خوش‌حساب و جداکردن از مشتریان بدحساب را نیز انجام دهند. در این مقاله مرور کلی بر طبقه‌بندی داده‌ها و روش پایه درخت تصمیم‌گیری برای طبقه‌بندی داده‌های حسابهای جاری بانک با استفاده از نرم افزار Rapidminer ارائه شده است و نشان داده شده است که دقت آن در مقایسه با سایر روشهای دسته بندی مطلوب می‌باشد.

کلمات کلیدی: درخت تصمیم، داده‌های حساب‌های جاری بانک، طبقه‌بندی، Rapidminer

۱- مقدمه

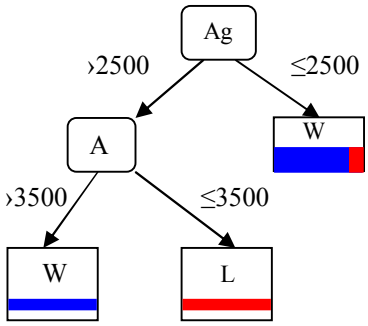
استخراج اطلاعات به طور گسترده در بسیاری از حوزه‌ها، از جمله خرده‌فروشی، مالی، ارتباط از راه دور و رسانه‌های اجتماعی استفاده می‌شود. تکنیک‌های اصلی برای داده‌کاوی شامل طبقه‌بندی، پیش‌بینی، خوشه‌بندی، تشخیص داده‌های زائد، استخراج قوانین انجمنی، تجزیه و تحلیل توالی، تجزیه و تحلیل سری‌های زمانی و متن کاوی و نیز شامل برخی از روش‌های جدید مانند تجزیه و تحلیل شبکه‌های اجتماعی و تجزیه و تحلیل احساسات می‌باشد. در دنیای واقعی، یک فرآیند داده‌کاوی را می‌توان به شش مرحله اصلی تقسیم کرد که عبارتند از:

۱- درک فرآیندکسب و کار ۲- درک داده‌ها ۳- آماده سازی داده‌ها ۴- مدل سازی ۵- ارزیابی مدل ۶- به کارگیری مدل که به مجموعه عملیات فوق عنوان CRISP-DM (استاندارد فرایند تعریف شده برای داده‌کاوی) اطلاق می‌گردد [۱۱].

۲- درخت تصمیم‌گیری

طبقه‌بندی داده‌ها یک فرآیند دو مرحله‌ای است. ابتدا، یک مدل برای توصیف مجموعه‌ای از پیش تعیین شده از کلاس داده‌ها و یا مفاهیم ساخته می‌شود. مدل با تجزیه و تحلیل رکوردهای پایگاه داده که توسط صفات توصیف شده هر رکورد فرض شده ساخته می‌شود شروع شده و این رکوردها متعلق به یک کلاس از پیش تعریف شده است، به صورتیکه بوسیله یکی از ویژگی‌ها که ویژگی برچسب نامیده می‌شود مشخص شده است. در زمینه طبقه‌بندی داده‌ها از رکوردها به عنوان نمونه‌ها، مثال‌ها و یا اشیاء اشاره می‌گردد. این ویژگی برچسب یا کلاس، همان ویژگی هدف در مدل نامیده می‌شود که در واقع پیش‌بینی خواهد شد و در یک درخت تصمیم در پایینترین سطح درخت یعنی برگ‌ها قرار می‌گیرد و در نتیجه انتخاب این فیلد از اهمیت خاصی برخوردار است. مقدار این ویژگی معمولاً مقادیری مانند{درست، نادرست} و یا {موفقیت، شکست} و یا {بله، خیر} و یا چیزی معادل آن خواهد بود[۱]. انواع الگوریتم‌های طبقه‌بندی درخت تصمیم‌گیری وجود دارد که می‌توان به الگوریتمهای ID3، C4.5، CART، CHAID اشاره نمود که هرکدام از این الگوریتم‌ها مزایایی دارند و برای ساخت درخت از اطلاعاتی استفاده می‌کنند. درخت‌های تصمیم‌گیری در طیف

گسترده‌ای از برنامه‌های کاربردی مختلف مورد استفاده قرار می‌گیرند. همچنین در رشته‌های مختلف بسیاری مورد استفاده قرار می‌گیرند از جمله تشخیص پزشکی، علوم شناختی، هوش مصنوعی، تئوری بازی‌ها، مهندسی و داده‌کاوی. درخت‌های تصمیم‌گیری از گره‌ها، شاخه‌ها و برگ‌ها تشکیل می‌شود. متغیرها، شرایط و نتایج را نشان می‌دهد که به ترتیب ساخته می‌شود. یکی از مهم‌ترین مزایای علمی درخت‌های تصمیم‌گیری این واقعیت است که دانش می‌تواند در شکل طبقه‌بندی قوانین (IF-THEN) استخراج شده و نشان داده شود. هر قانون می‌تواند نشان‌دهنده یک مسیر منحصر به فرد از ریشه به هر برگ باشد. درخت تصمیم‌گیری یک ساختار درختی فلوچارت مانند است که در آن هر شاخه نشان‌دهنده، یک خروجی از آزمون، و هرگره برگ نشان‌دهنده کلاس‌های ویژگی با بالاترین بهره اطلاعات (Information gain) است که به عنوان ویژگی آزمون برای گره فعلی انتخاب شده است. این ویژگی، اطلاعات مورد نیاز برای طبقه‌بندی نمونه را به حداقل می‌رساند. می‌توان گفت درخت تصمیم‌گیری یک ساختار درختی فلوچارت مانند است که در آن هر گره داخلی (گره‌های غیر برگ) نشان‌دهنده یک آزمون روی یک ویژگی است. هر شاخه نتیجه آزمون را نشان می‌دهد و هر گره برگ (یا گره پایانی) دارای یک برچسب کلاس است. بالاترین گره در درخت، گره ریشه است[۱۱][۱۲]. نمونه‌ای از یک درخت تصمیم‌گیری ساده در شکل ۱ نشان داده شده است.



شکل ۱ : نمونه‌ای از یک درخت تصمیم‌گیری ساده

یک درخت تصمیم‌گیری روشی است که برای کمک به انتخابهای خوب می‌توانیم از آن استفاده کنیم، به خصوص در مورد تصمیم‌گیری‌های که شامل هزینه‌ها و ریسک بالایی هستند. درخت‌های تصمیم‌گیری از روش گرافیکی برای مقایسه گزینه‌ها و مقادیر اختصاصی استفاده می‌کنند. فرض کنید ما می‌خواهیم خرید کامپیوتر انجام دهیم این نشان‌دهنده مفهوم اقدام به خرید کامپیوتر است، یعنی آن پیش‌بینی می‌کند که آیا یک مشتری به احتمال زیاد به خرید یک کامپیوتر راغب است یا نه. در هر درخت تصمیم گره‌های داخلی با مستطیل نشان داده می‌شود و برگها توسط بیضی نشان داده می‌شود. برخی از الگوریتم‌های درخت تصمیم‌گیری تنها درخت دودویی (هر گره داخلی دقیقا دو گره فرزند دارد) را تولید می‌کنند، در حالی که برخی دیگر از الگوریتم‌ها می‌توانند درخت‌های غیر دودویی تولید کنند[۱۲].

دو سوال در مورد ساخت یک درخت تصمیم مطرح می‌شود. اول اینکه کدام ویژگی برای گره ریشه در نظرگرفته شود؟ ودوم اینکه مدل چگونه ترتیب گره‌ها را از بالا به پایین تعیین می‌کند؟ برای پاسخگویی به این سوالات واضح است که ما نیاز به راهی برای تعیین کیفیت همگنی برای مجموعه داده‌ها داریم و در اینجا مفهومی به نام آنتروپی مطرح می‌شود. اگر زیرمجموعه داده‌ها ۱۰۰٪ همگن باشد یعنی همه بله و یا همه خیر باشد آنتروپی متغیر مورد بررسی صفر است و برای یک زیرمجموعه که ۵۰٪ آن بله و۵۰٪ آن خیر باشد آنتروپی هدف ۱ خواهد بود. در مجموعه داده‌هایی که درصد آن بین ۰ تا ۱۰۰ درصد باشد آنتروپی هدف بین ۰ تا ۱ خواهد بود[۱۵].

نرم افزار Rapidminer از Gain Ratio یا نسبت سود (که نسبت آنتروپی پس از تقسیم به آنتروپی قبل از تقسیم می‌باشد) و یا از معیار Information Gain یا بهره اطلاعات (که تفاوت بین آنتروپی‌ها قبل و بعد از تقسیم) و Gini Index یا شاخص جینی برای تعیین کیفیت همگنی استفاده می‌کند. این مقادیر برای تمام ویژگی‌ها محاسبه می‌شود.

آن ویژگی که این دو نسبت را حداکثر می‌کند به عنوان گره ریشه و یا گره بالایی درخت انتخاب می‌شود. مقدار این ویژگی که در این افزایش رخ می‌دهد به عنوان نقطه تقسیم در نظرگرفته خواهد شد. این تقسیم شدن در طول این ویژگی می‌تواند ادامه پیدا کرده و نتیجه این تقسیم، زیرمجموعه‌های کوچکتر وکوچکتر با خلوص بالاتر خواهد بود[۱۴].

۳- مزایای استفاده از درخت‌های تصمیم‌گیری

دراستفاده از درخت‌های تصمیم‌گیری چندین مزیت متمایز، در طبقه‌بندی و پیش‌بینی بسیاری از برنامه‌های کاربردی وجود دارد. این مزایا عبارتند از:

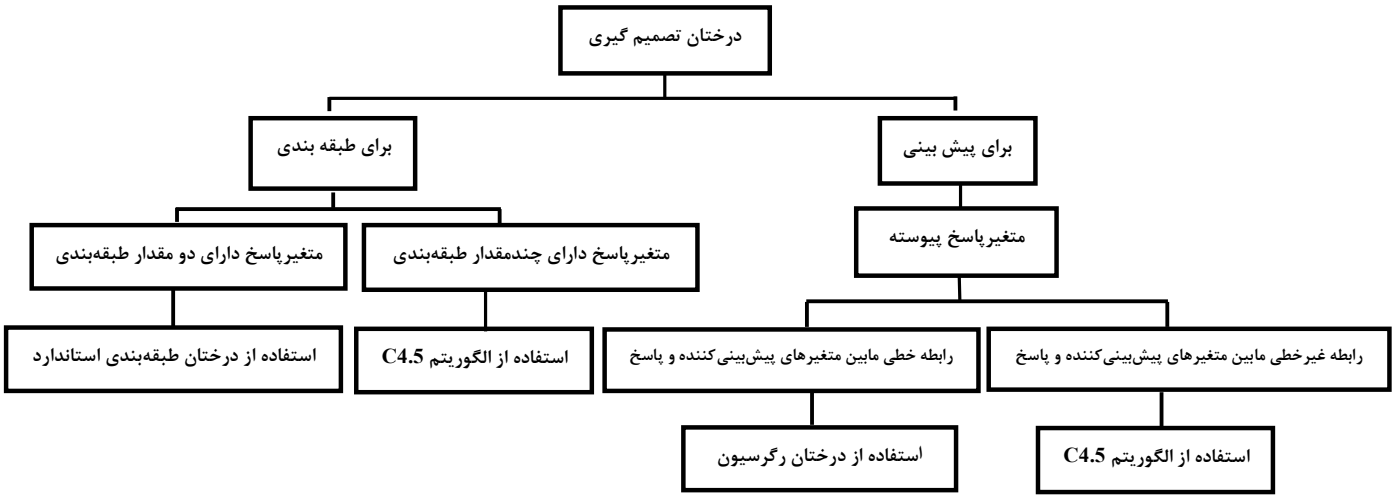
مزیت ۱: درخت‌های تصمیم‌گیری به طور ضمنی غربالگری متغیر و یا انتخاب ویژگی را انجام می‌دهند. انتخاب ویژگی در تجزیه و تحلیل داده‌کاوی مهم است. چند روش معمول برای انجام انتخاب ویژگی و یا غربالگری متغیر وجود دارد. هنگامی که ما یک درخت تصمیم‌گیری را با استفاده از یک مجموعه داده‌های آموزشی ایجاد کردیم، گره‌هایی در بالای چنین درختی و اساسا مهم‌ترین متغیرها درون مجموعه داده‌ها تقسیم می‌شود و انتخاب ویژگی به صورت خودکار کامل می‌شود[۱۳].

مزیت ۲: درخت‌های تصمیم‌گیری نیاز به تلاش نسبتا کمی برای آماده‌سازی داده‌ها از سوی کاربران دارد. یکی دیگر از ویژگی‌هایی که موجب صرفه‌جویی در زمان آماده‌سازی داده‌ها می‌شود، مقادیرگمشده از تقسیم داده‌ها برای ساختن درختان جلوگیری نخواهد کرد. همچنین درخت‌های تصمیم‌گیری به نقاط دورافتاده حساس نیستند[۱۳].

مزیت ۳: ارتباط غیرخطی بین پارامترها روی عملکرد درخت تاثیر نمی‌گذارد. درخت‌های تصمیم‌گیری هیچ فرضیات خطی در داده‌ها را نیاز ندارند. بنابراین ما می‌توانیم آنها را در حالاتی‌که پارامترهای غیرخطی مربوط به آن وجود دارد استفاده کنیم[۱۳].

مزیت ۴: بهترین ویژگی درختان تصمیم‌گیری این است که استفاده از این درختان امکان تجزیه و تحلیل و تفسیر آسانی از داده‌های انبوه را برای مدیران مهیا می‌کند. بنابراین درخت‌های تصمیم‌گیری بسیار شهودی و آسان برای توضیح هستند[۱۳].

با وجود تمام این مزایا، با یک نقطه ضعف کلیدی مزایای درخت‌های تصمیم‌گیری معتدل می‌شود، بدون هرس مناسب و یا محدود کردن رشد درخت، درختان تصمیم‌گیری تمایل به Overfit داده‌ها دارند، که آنها را تا حدودی ضعیف پیش بینی می‌کند و زمانی‌که درخت ساخته شده بزرگ باشد تفسیر آن بسیار پیچیده بوده و به سادگی امکان‌پذیر نخواهد بود[۱۳]. شکل ۲ دسته‌بندی کاربردهای درخت تصمیم‌گیری را نشان می‌دهد.



شکل ۲: دسته بندی کاربردهای درخت های تصمیم

۴- شناخت داده‌ها

داده‌های مورد استفاده در این تحقیق مربوط به اطلاعات صاحبان حسابهای جاری بانک ملت می‌باشد. تعداد رکوردها ۴۰۰ رکورد بوده و دراین مجموعه داده‌های ناقص وجود ندارد و در نتیجه در مرحله پیش پردازش داده‌ها حذف اطلاعات انجام نگردید. ویژگی‌های این جدول به صورت زیر می‌باشد:

۱- جنسیت: که شامل دو مقدار ۱ و۲ می باشد ۱ بیانگر زن و۲ بیانگر مرد است. این فیلد به صورت عددی است.

۲- شغل: شغل افراد نیز شامل مقادیر تاجر۱، تولید۲، خدمات ۳، کارگر۴، کارمند۵، عمده فروش ۶ و بازنشسته ۷ و به صورت Integer می‌باشد.

۳- تعداد تراکنش درسه ماه: این فیلد برای هر رکورد مجموع تعداد خریدهای یک فرد را درمدت زمان سه ماه را بدست می‌دهد که مقدار آن نیز Integer است.

۴- تحصیلات: این فیلد نیز در نمونه مورد مطالعه دارای هفت سطح می‌باشد و به صورت Integer است.
ابتدایی ۱ ، سیکل ۲، دیپلم ۳، فوق دیپلم ۴، لیسانس ۵ ، فوق لیسانس ۶ ، دکترا ۷ می‌باشد.

۵- میانگین مانده در ۳ ماه : این فیلد بصورت Numeric می‌باشد.

۶- سابقه فعالیت: این فیلد بصورت Integer بوده و سابقه فعالیت مشتریان را نشان می‌دهد که حداقل یکسال و حداکثر ۵۰ سال می‌باشد.

۷- سن: این فیلد بیانگر سن صاحبان حسابهای بانک می‌باشد که بین ۱۸ و۷۲ سال است.

۸- تعداد چکهای برگشتی درسه ماه: این فیلد نیز بصورت Integer بوده و مقدار آن حداقل صفر یعنی چک برگشتی نداشته و حداکثر ۷ عدد چک برگشتی می‌باشد[۸][۹].

تقسیم‌بندی مقادیرعددی به گروه‌های متفاوت Discretization یا گسسته‌سازی نامیده می‌شود. از آنجا که اپراتورهای زیادی زمانی که تمامی ویژگی‌ها Nominal باشند نتایج بهتری می‌دهند، درچنین مواردی ضروری است که خصوصیات عددی به نوع اسمی یا Nominal گسسته شود. عمل نرمال سازی یا Normalization یک تکنیک پیش پردازش است که برای ترسیم کوچکتر مقادیر ویژگی و مناسب کردن آنها در یک رنج خاص استفاده می‌شود. زمانی‌که با خصوصیات واحدها و مقیاس‌های مختلف سرکار داشته باشیم نرمال‌سازی داده خیلی مهم است. این دو تکنیک از تکنیک‌های پیش پردازش محسوب می‌شوند. بنابراین برای ساخت درخت تصمیم بایستی ابتدا داده‌ها را نرمال‌سازی نمود وسپس آنرا گسسته‌سازی کنیم همچنین تابع هدف را تعداد چکهای برگشتی در نظرگرفته و پس از گسسته‌سازی ویژگی میانگین مانده در ۳ ماه با حداقل مقدار موجود در داده‌ها و حداکثر مقدار این ویژگی (به روش Interval برای جداسازی این ویژگی و تبدیل آن به ویژگی اسمی) مدل ایجاد می‌شود. شکل ۳ ابر داده مورد تحقیق در

Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (400 examples, 1 special attribute, 7 regular attributes)

Role	Name	Type	Statistics	Range	...
label	تعداد چک های برگشتی در 3 ماه	nominal	mode = 0.0 (330), least = 6.0 (1)	3.0 (8), 2.0 (18), 0.0 (330), 1.0 (38), 5.0 (2), 6.0 (1), 4.0 (2), 7.0 (1)	0
regular	جنسیت	integer	avg = 1.345 +/- 0.476	[1.000 ; 2.000]	0
regular	شغل	nominal	mode = 1.0 (87), least = 7.0 (30)	1.0 (87), 5.0 (36), 2.0 (77), 3.0 (56), 4.0 (53), 6.0 (61), 7.0 (30)	0
regular	تعدا تراکنش در 3 ماه	integer	avg = 66.005 +/- 102.921	[0.000 ; 742.000]	0
regular	تحصیلات	nominal	mode = 4.0 (88), least = 7.0 (10)	2.0 (63), 7.0 (10), 5.0 (55), 4.0 (88), 1.0 (84), 6.0 (23), 3.0 (77)	0
regular	میانگین مانده در 3 ماه	numeric	avg = 1690933711.510 +/- 4896440323.494	[78757.000 ; 47594715474.000]	0
regular	سابقه فعالیت	integer	avg = 11.998 +/- 10.866	[1.000 ; 50.000]	0
regular	سن	nominal	mode = 20.0 (19), least = 63.0 (1)	63.0 (1), 47.0 (16), 25.0 (14), 29.0 (13), 30.0 (19), 23.0 (13), 44.0 (3), 43.0 (9), 61.0 (6), 42.0 (13), 38.0 (9), 36.	0

شکل ۳- نمایش ابرداده اطلاعات حساب های جاری بانک

نرم‌افزار Rapidminer را نشان می‌دهد.

مدل تحقیق به شدت تحت تاثیر متغیرهای تعداد چکهای برگشتی و ویژگی میانگین مانده در ۳ ماه هر مشتری است که می‌توان نتایج خوبی با استفاده از این متغیرها گرفت. در مدل ایجاد شده در تحقیق توسط نرم افزار Rapidminer، معیار تقسیم شاخه‌ها براساس (Gini Index) یا شاخص جینی تعریف شد و معیاراطمینان (Confidence) برای قوانین ایجاد شده ۰/۵ در نظرگرفته می‌شود.[۲][۳].

۵- ارزیابی مدل

پس از ایجاد مدل، ارزیابی آن انجام می‌شود. در ارزیابی مدل درخت تصمیم ساخته شده از Split Validation استفاده شده است. این اپراتور داده را به صورت تصادفی به دو دسته مجموعه آزمایشی (test) ومجموعه آموزشی (training) تقسیم می‌کند و سپس مدل را ارزیابی می‌کند. همچنین این اپراتور کارایی یک اپراتور یادگیرنده را تخمین می‌زند و عمدتا برای تخمین چگونگی صحت مدلی که در عمل اجرا خواهد شد استفاده می‌شود. مدل ابتدا با داده آموزشی یادگرفته می‌شود و سپس برروی داده تست بکار برده می‌شود. Split Validation راهی برای پیش‌بینی مناسب یک مدل به مجموعه تست فرضی است زمانیکه یک مجموعه تست واضحی در دسترس نباشد [۷][۶]. پس از اجرای مدل مقدار accuracy = ٪۷۳/۳۳ بدست می‌آید. میزان صحت در واقع درصد مشاهداتی است که بدرستی دسته‌بندی شده باشد و این دسته‌بندی بر روی مجموعه آموزشی انجام می‌شود[۹][۱][۱۶] .

۶-شرح درخت تصمیم

در درخت ایجاد شده تحقیق، ریشه درخت براساس ویژگی سطح تحصیلات بوده سپس میانگین مانده حساب وسایر ویژگیها دارای اولویت می‌باشد جدول یک حاوی اطلاعات بدست آمده از درخت تصمیم ساخته شده می‌باشد. درصد بیشتر مشتریان بانک دارای سطح تحصیلات پایین هستند و ۱۵/۲۵ درصد افراد دارای میانگین مانده بالایی هستند که بیشترین درصد مربوط به سطح تحصیلات لیسانس می‌باشد. از جمع ۷۰ مشتری که دارای چک برگشتی هستند ٪۳۳ مشتریان دارای سطح تحصیلات لیسانس به بالا و ٪۶۷ مابقی دارای سطح تحصیلات فوق دیپلم به پایین می‌باشند. به نظر می‌رسد که جنسیت مشتری در طبقه‌بندی تاثیری ندارد[۷]. ۱۹نفر یا ۴/۷۵ درصد مشتریان دارای میانگین مانده حساب بالایی هستند که ۳ نفر از این مشتریان دارای چک برگشتی می باشند و دارای سطح تحصیلات لیسانس و بالاتر بوده و ۹ نفر از این مشتریان دارای چک برگشتی دارای سطح تحصیلات فوق دیپلم به پایین می‌باشند. به نظرمی‌رسد علیرغم میانگین مانده بالای حساب سه ماه، مشتریان دارای سطح تحصیلات پایین احتمال برگشت چکهای بیشتری دارند.

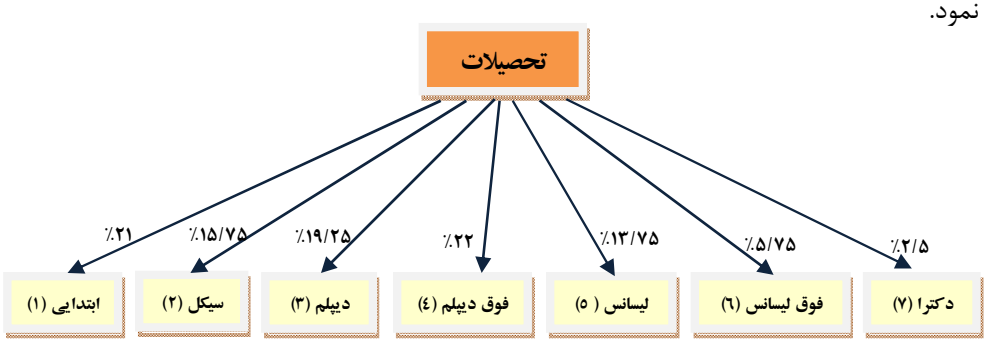
جدول ۱- اطلاعات بدست آمده از درخت تصمیم ایجاد شده

کد	عنوان	تعدادمشتری	درصدمشتری به کل مشتریان	چک برگشتی(تعداد- درصد نسبت به کل مشتریان)		میانگین مانده حساب (۷۰۰۰۰-۵۰۰۰۰۳۵۰۰)		میانگین مانده حساب (۵۰۰۰۰۳۵۰۰۰-۱۰۰۰۰۰۰۰۰۰)		میانگین مانده حساب ۱۰۰۰۰۰۰۰۰۰بیشتراز	
۱	ابتدایی	۸۴	٪۲۱	۱۱	٪۲/۷۵	۷۳	٪۸۶/۹	۳	٪۳/۵۷	۸	٪۹/۵۲
۲	سیکل	۶۳	٪۱۵/۷۵	۱۱	٪۲/۷۵	۴۷	٪۷۴/۶	۶	٪۹/۵۲	۱۰	٪۱۵/۸۷
۳	دیپلم	۷۷	٪۱۹/۲۵	۱۶	٪۴	۶۹	٪۸۹/۶	۲	٪۲/۵۹	۶	٪۷/۷۹
۴	فوق دیپلم	۸۸	٪۲۲	۹	٪۲/۲۵	۷۰	٪۷۹/۵۴	۳	٪۳/۴	۱۵	٪۱۷/۰۴
۵	لیسانس	۵۵	٪۱۳/۷۵	۱۶	٪۴	۴۷	٪۸۵/۴۵	۵	٪۹/۰۹	۳	٪۵/۴۵
۶	فوق لیسانس	۲۳	٪۵/۷۵	۴	٪۱	۲۱	٪۹۱/۳	۰	۰	۲	٪۸/۶۹
۷	دکترا	۱۰	٪۲/۵	۳	٪۰/۷۵	۱۰	٪۱۰۰	۰	۰	۰	۰
	جمع	۴۰۰	٪۱۰۰	۷۰	٪۱۷/۵	۳۳۷	٪۸۴/۲۵	۱۹	٪۴/۷۵	۴۴	٪۱۱

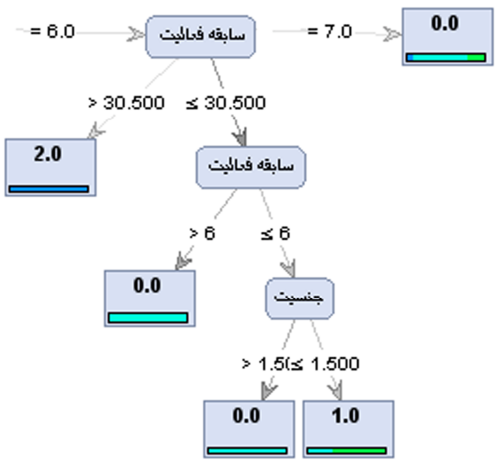
۲۷ نفر از مشتریان یا ۶/۷۵ درصد مشتریان دارای میانگین حساب پایینتری بوده که دارای چکهای برگشتی هستند و دارای سطح تحصیلات پایین می‌باشند. ٪۶۷ مشتریان دارای چک برگشتی دارای تحصیلات پایین می‌باشند. شکل ۴ درصد مشتریان هرسطح تحصیلات را نشان می‌دهد و درخت تصمیم ایجاد شده به تفکیک سطح تحصیلات بصورت شکل ۵ (الف- ب – ج – د- ه و و) نشان داده می‌شود[۴] [۵].

۷-نتیجه‌گیری

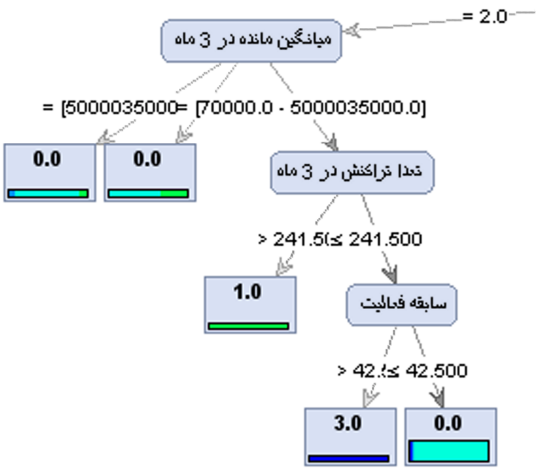
از آنجا که در بانکها، اطلاعات حجیم و متنوعی در خصوص مشتریان و دارندگان حسابهای جاری دربانکهای اطلاعاتی ذخیره می‌باشد وکشف دانش مورد نیاز مدیران از این حجم زیاد اطلاعات نیاز به ابزارها وتکنیکهای جدیدتری دارد لذا از تکنیکهای داده‌کاوی جهت این منظور میتوان سود برد. الگوریتمهای درخت تصمیم یکی از روش‌های مناسب در علم داده‌کاوی می‌باشد که علاوه بر دسته‌بندی داده‌ها و کاهش پیچیدگی محاسبات، روشهای ساده‌تری برای ساخت و فهم آسان و راحت آن ارائه می‌کند و دقت آن نیز در مقایسه با سایر روش‌های دسته بندی مطلوب می‌باشد. درصنعت بانکداری حفظ مشتری از موضوعات مهمی است بنابراین انتخاب بهترین زیرمجموعه از ویژگی‌ها و به کارگرفتن درختان تصمیم‌گیری برای طبقه‌بندی دارندگان حسابهای جاری بانک می‌تواند منجر به ارائه الگوهای مناسبی جهت پیش‌بینی‌های مورد نظر شود. نرم افزار Rapidminer ابزار مناسبی برای ایجاد درختان تصمیم‌گیری و پیش‌بینی‌های مختلف است که دارای محبوبیت زیادی دربین کاربران می‌باشد. با ایجاد درختان تصمیم مختلف می‌توان پیش‌بینی نمود که یک مشتری در چه دسته‌ای از مشتریان خوش‌حساب ویا بدحساب قرار دارد. همچنین با تحلیل درختان ایجاد شده می‌توان الگوهای جدیدی جهت مدیریت ریسک اعتباری بانکها (درصد ریسک‌پذیری وام برای مشتریان) و ارائه گزارشات برای مدیران بدست آورد. درصد ریسک‌پذیری یعنی اینکه چند درصد از مشتریان قادر به باز پرداخت وام خود هستند تا آنجا که می‌توان راههای مختلفی جهت بازاریابی و نگهداری دارندگان حسابها دربانک ارائه نمود.



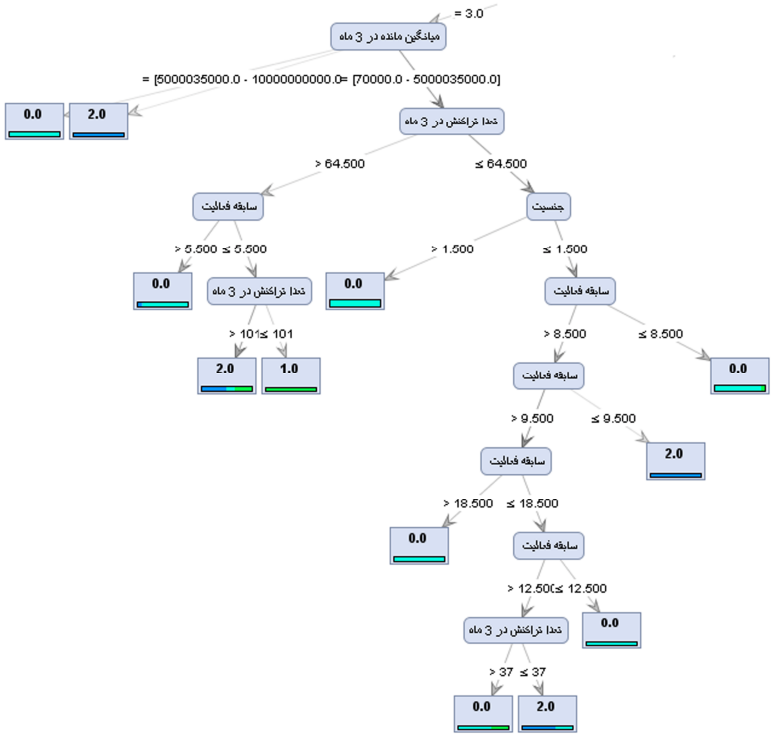
شکل ۴- طبقه‌بندی داده‌ها براساس سطح تحصیلات



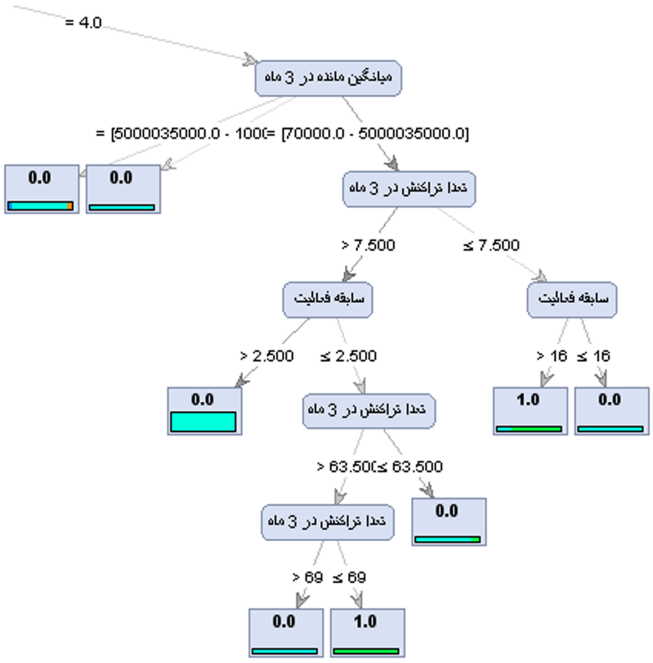
شکل ۵- ب) طبقه‌بندی سطح تحصیلات فوق‌لیسانس (۶) ودکترا (۷)



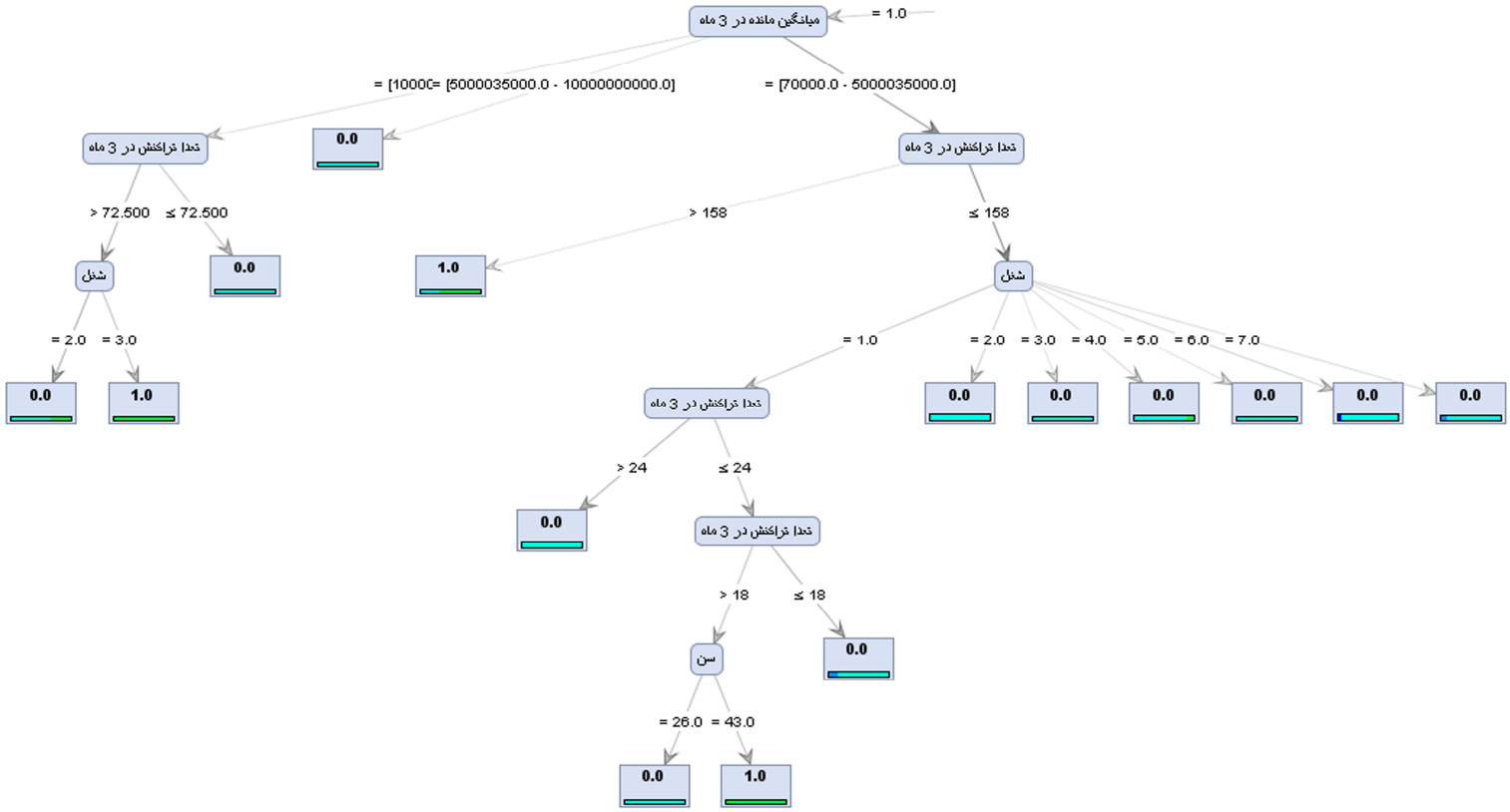
شکل ۵- الف) طبقه‌بندی سطح تحصیلات سیکل (۲)



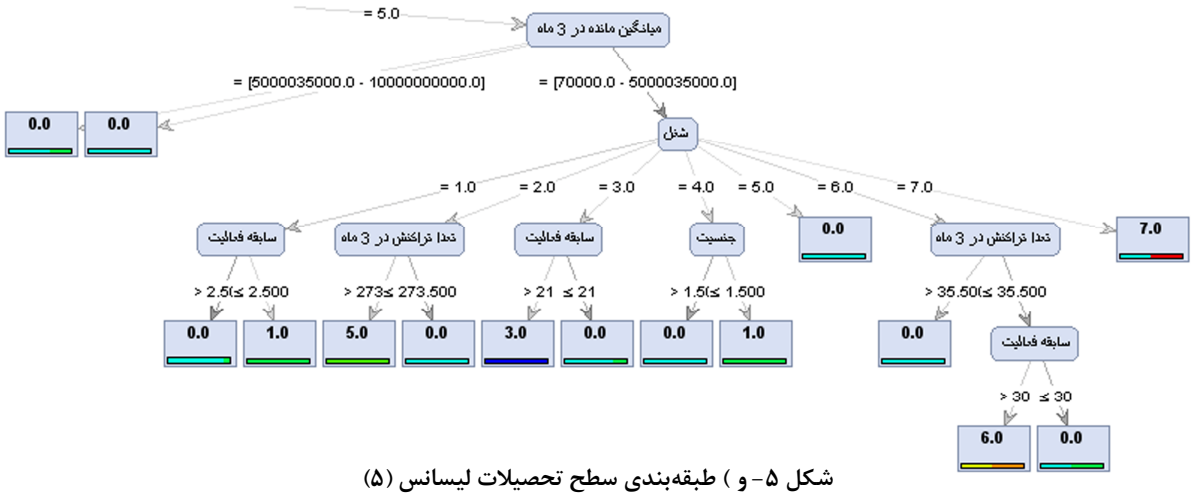
شکل ۵- ج) طبقه‌بندی سطح تحصیلات دیپلم (۳)



شکل ۵- د) طبقه‌بندی سطح تحصیلات فوق‌دیپلم (۴)



شکل ۵- ه) طبقه‌بندی سطح تحصیلات ابتدایی (۱)



شکل ۵- و) طبقه‌بندی سطح تحصیلات لیسانس (۵)

References

[۱] عامری، حکیمه، علیزاده، سمیه، برزگری، اکبر، مجله مدیریت سلامت ،"استخراج دانش از داده‌های بیماران دیابتی با استفاده از روش درخت تصمیم c5. 0 " ، سال۱۳۹۲ .

[۲] نامدار علی آبادی، عباس، آقاجانی، حسنعلی، غلامی، رمضان، نامدار علی آبادی، سوده، " شناسایی مشتریان هدف برای به کارگیری استراتژی بازاریابی مستقیم در بانک با استفاده ازداده‌کاوی" ، دومین کنفرانس بین المللی بازاریابی خدمات مالی .

[۳] البرزی، محمود، محمدپورزرندی، محمد ابراهیم ، خان بابایی، محمد،" به کارگیری الگوریتم ژنتیک در بهینه سازی درختان تصمیم گیری برای اعتبارسنجی مشتریان بانک‌ها"، نشریه مدیریت فناوری، دوره ۲، شماره۴، از صفحه ۲۳ تا ۳۸، بهار و تابستان ۱۳۸۹.

[۴] ناظمی، جمشید، جعفری، پژمان، هاشمی، حامد ،" کاوش خصوصیات مشتریان بانکداری خرد با استفاده از تکنیکهای داده‌کاوی"، مجله مدیریت بازاریابی ، شماره ۱۴، بهار۱۳۹۱.

[۵] خورشیدی، غلامحسین، کاردگر، محمدجواد، " شناسایی و رتبه‌بندی مهم‌ترین عوامل موثر بر وفاداری مشتریان با استفاده از روشهای تصمیم‌گیری چند معیاره"، چشم انداز مدیریت، شماره ۳۳، از صفحه ۱۷۷تا۱۹۱ ، زمستان ۱۳۸۸.

[۶] تقوا، محمدرضا، حسینی بامکان، سیدمجتبی،" ارائه خدمات مناسب به مشتریان بالقوه با استفاده از تکنیکهای داده‌کاوی درحوزه بانکداری الکترونیک"، فصلنامه پژوهشی مطالعات مدیریت صنعتی سال هشتم ، شماره ۲۳، از صفحه ۱۸۷تا ۲۰۷ ، زمستان ۱۳۹۰.

[۷] جماعت، علی، عسگری، فرید،" مدیریت ریسک اعتباری در سیستم بانکی با رویکرد داده‌کاوی"، فصلنامه مطالعات کمی در مدیریت.

[۸] جلائیان، زهرا ، " داده‌کاوی و کاربرد آن در بانکداری"، توسعه صادرات ، شماره ۷۳، مرداد ۱۳۸۷.

[۹] خان‌بابایی، محمد، زین العابدینی، سیده فاطمه، "مدل به کارگیری تکنیکهای داده‌کاوی در شناسایی و تحلیل رفتار مشتریان خدمات بانکداری الکترونیکی"، فصلنامه علمی پژوهشی تحقیقات بازاریابی نوین ، شماره دوم ، از صفحه ۱۷۵تا ۱۸۸ ، تابستان ۱۳۹۲.

[10] Moertini ,Veronica, “Towards The Use Of C4.5 Algorithm For Classifing Banking Dataset”, Pengetahuan Alam universitas, vol8. No 2, October 2003.

[11] Witten ,Ian ,Frank ,Eibe ,Hall ,Mark ,“Data Mining Practical Machine Learning Tools and Techniques Second Edition” , Amsterdam, pp.62–69,189.

[12] Olivas ,Rafael,“ Decision Trees A Primer for Decision-making Professionals”, pp. 1-20 ,2007.

[13] Deshpande ,Bala, “Decision Tree Digest-An eBook understand ,build and use decision trees for common business problems with Rapidminer”, pp. 3-19.

[14] Rapid-I , “Rapidminer-5.0-Manual-English v1.0 User Manual”,2010.

[15] Mitchell , “Machine Learning Decision Tree ”, Chapter3, pp. 1-29.

[16] Dass ,Rajanish, “Data Minig in Banking and Finance: A Note For Bankers”,Indian Institute of Management Ahmedabad.