فصلنامهٔ پژوهش‌های پولی_بانکی
سال چهاردهم، شمارهٔ ۴۸، تابستان ۱۴۰۰
صفحات ۳۲۷_ ۳۶۰

# پیش‌بینی امتیاز اعتباری اشخاص با استفاده از الگوریتم‌های داده‌کاوی (مورد مطالعه: مشتریان یک بانک دولتی ایران)

**عالیه کاظمی** [*]
**کیمیا عظیم‌زاده طهرانی** [†]
**علی ابدالی** [‡]
**سارا آریایی** [§]

تاریخ دریافت: ۱۳۹۹/۱۲/۱۳ تاریخ پذیرش: ۱۴۰۰/۰۲/۲۷

## چکیده

امتیاز اعتباری مشتریان یکی از ابزارهای مهم برای مدیریت ریسک در سیستم‌های بانکی است. طراحی سیستمی که بتواند شناسایی مشتریان بانکی را به‌درستی انجام دهد، از چالش‌های اساسی در داده‌کاوی و یادگیری ماشین بوده که مطالعات بسیاری در مورد آن انجام شده است. در این مطالعه، عوامل مرتبط با امتیاز اعتباری معرفی، و پیش‌بینی امتیاز اعتباری برای مشتریان یک بانک دولتی ایران انجام شده است. بدین‌منظور، روش‌شناسی CRISP-DM به‌عنوان مدل مرجعی برای فرایند داده‌کاوی مورد استفاده قرار گرفته و مدل‌سازی داده‌ها با بهره‌گیری از الگوریتم‌های مختلف (نزدیک‌ترین همسایگی، درخت تصمیم، و جنگل تصادفی) انجام شده است. نتایج ارزیابی و مقایسهٔ حساسیت و دقت الگوریتم‌ها نشان داد الگوریتم نزدیک‌ترین همسایگی با دقت ۹۰/۳ درصد برای مجموعهٔ آموزش و ۷۶/۷ درصد برای داده‌های آزمون، عملکرد مناسبی برای پیش‌بینی امتیاز اعتباری دارد.

**واژه‌های کلیدی:** امتیاز اعتباری، داده‌کاوی، روش‌شناسی CRISP-DM، نزدیک‌ترین همسایگی، درخت تصمیم، جنگل تصادفی.
**طبقه‌بندی JEL: C02 , C52, G21**

[*] دانشیار گروه مدیریت صنعتی دانشکدهٔ مدیریت دانشگاه تهران(نویسندهٔ مسئول)؛ aliyehkazemi@ut.ac.ir
[†] کارشناس ارشد مدیریت صنعتی، دانشکدهٔ مدیریت دانشگاه تهران؛ kimia.azimzadeh@gmail.com
[‡] استادیار دانشگاه علوم انتظامی امین، تهران؛ abdali9192@gmail.com
[§] دانشجوی دکتری مدیریت صنعتی، دانشکدهٔ مدیریت دانشگاه تهران؛ sara.aryaee@ut.ac.ir


[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

## ۱ مقدمه

ریسک اعتباری یکی از مهم‌ترین ریسک‌هایی است که بانک‌ها با آن مواجه‌اند. بی‌توجهی به این ریسک منجر به ایجاد مطالبات معوق و سررسید گذشته می‌شود. برای تعیین و کنترل ریسک اعتباری، شناخت مشتریان خوش‌حساب و بدحساب از اهمیت ویژه‌ای برخوردار است. یکی از روش‌های مؤثر برای شناخت مشتریان، امتیاز اعتباری[1] است.

امتیاز اعتباری مجموعه‌ای از مدل‌های تصمیم‌گیری و روش‌های پایه‌ای است که به قرض‌دهنده برای اعطای اعتبار مصرفی کمک می‌کند. این روش‌ها به تصمیم‌گیرنده برای تصمیم‌گیری دربارهٔ اینکه چه‌کسی اعتبار دریافت‌کند و چقدر اعتبار می‌تواند دریافت کند، کمک می‌کند (توماس، کروک و ادلمن[2]، ۲۰۰۲).

تحقیقات بسیاری روی مدل امتیاز اعتباری در بانک‌ها و مؤسسات مالی صورت گرفته و روش‌های مختلفی استفاده شده است. در ابتدا، مدل‌های امتیاز اعتباری به‌صورت قضاوتی بودند. سپس، روش‌های تحلیلی مناسب‌تری برای محاسبهٔ این امتیاز ارائه شد (آسایش و جلالی، ۱۳۹۷).

هنگامی‌که یک تقاضای اعتباری (وام) از سوی یک مشتری حقیقی یا حقوقی به بانک ارائه می‌شود، بانک بایستی پیش از اعطای هرگونه تسهیلات، به بررسی ویژگی‌ها و عملکردهای گذشتهٔ دریافت‌کنندهٔ وام و احتمال عدم پرداخت اصل وام و سود حاصل از آن بپردازد (جلیلی، ۱۳۸۹). افزایش حتی یک درصدی دقت مدل‌هایی که بانک‌ها برای پی‌بردن به این مهم به‌کار می‌برند، می‌تواند از خسارت‌های مالی بسیاری جلوگیری کند (وانگ، ما و هوانگ و ژو[3]، ۲۰۱۲).

امتیاز اعتباری عامل بسیار مهمی برای درنظرگرفتن مقدار مناسب وام برای مؤسسات مالی و اعتباری است که علاوه بر تضمین بازگشت وجوه به این مؤسسات، باعث جلوگیری از دادن وام با استفاده از روش‌های سنتی و سلیقه‌ای به افراد می‌شود. دریافت وام‌های سنگین برای افراد با توانایی کمتر یا عدم دریافت وام سنگین برای افراد واجد شرایط هم به ضرر مؤسسات مالی و اعتباری مانند بانک‌ها و هم به ضرر افراد است.

موضوع امتیازدهی اعتباری از آنجا اهمیت می‌یابد که باوجود سرویس‌های مختلفی که به‌تازگی توسط بانک‌ها ارائه شده‌اند، واگذاری تسهیلات به مشتریان هنوز هم به‌عنوان یکی

---

[1] Credit Score

[2] Thomas, Crook and Edelman

[3] Wang, Ma, Huang, and Xu



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

از مهم‌ترین منابع درآمد برای بانک‌ها و مؤسسات مالی محسوب می‌شود. به استناد بخشنامهٔ بانک مرکزی در سال ۱۳۹۵، ماندهٔ مطالبات غیرجاری در سال ۱۳۹۵ حدود ۱۱ درصد است، درحالی‌که استاندارد این نرخ در دنیا ۴ درصد است که نشان‌دهندهٔ اهمیت این مسئله است. تاکنون تحقیقات مختلفی نیز در زمینهٔ اعتبارسنجی انجام شده است؛ ولی هنوز سالانه ماندهٔ معوق ۲۰ درصد رشد دارد. رقم قابل‌توجه مطالبات و رشد بالای آن نشان‌دهندهٔ اهمیت این موضوع و لزوم توجه به آن است (آسایش و جلالی، ۱۳۹۷).

در این پژوهش، تلاش بر آن است تا با رویکرد داده‌کاوی[1]، روشی مناسب برای پیش‌بینی[2] امتیاز اعتباری برای هر مشتری بانک به‌منظور اعتبارسنجی ارائه شود تا بانک موردنظر از این امتیاز برای نحوهٔ پرداخت وام و تسهیلات کمک بگیرد.

در ادامه به بررسی پیشینهٔ پژوهش پرداخته شده است؛ سپس روش‌شناسی پژوهش توضیح داده شده، و پس از آن، یافته‌های پژوهش ارائه شده و نهایتاً نتایج ذکر شده است.

## ۲ مبانی نظری و پیشینهٔ پژوهش

با توجه به اهمیت تعیین و پیش‌بینی امتیاز اعتباری مشتریان برای بانک‌ها و مؤسسات مالی و اعتباری، استفاده از روش‌های جدید در این حوزه موردتوجه قرار گرفته است. داده‌کاوی، به‌عنوان روشی برای شناسایی عوامل مؤثر در این امتیاز، تعیین امتیاز مشتریان، و همچنین کشف الگوهایی برای پیش‌بینی امتیاز اعتباری مشتریان، می‌تواند کمک بزرگی باشد.

محاسبهٔ امتیاز اعتباری بر پایهٔ روش تحقیق آماری یا عملیاتی است. روش‌های رگرسیون و تحلیل تفکیک‌کنندهٔ خطی[3] روش‌هایی‌اند که بیش از سایر روش‌ها برای ساخت امتیاز اعتباری مورد استفاده قرار گرفته‌اند. برخی از روش‌های دیگر عبارت‌اند از: رگرسیون

---

[1] Data mining

کشف دانش و داده‌کاوی یک حوزهٔ جدید میان‌رشته‌ای و درحال رشد است که حوزه‌های مختلفی چون پایگاه داده، آمار، یادگیری ماشین، و سایر زمینه‌های مرتبط را با هم تلفیق کرده تا اطلاعات و دانش ارزشمند نهفته در حجم بزرگی از داده‌ها را استخراج کند. هدف داده‌کاوی یافتن الگوها و یا مدل‌های موجود در پایگاه داده‌هاست که در میان حجم عظیمی از داده‌ها مخفی هستند (غضنفری, علیزاده و تیمورپور، ۱۳۸۷).

[2] پیش‌بینی عبارت است از تعیین مقدار یک متغیر پاسخ (متغیر وابسته) برحسب مقادیر متغیرهای مستقل (غضنفری و همکاران، ۱۳۸۷).

[3] Linear Discriminant Analysis



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

لجستیک[۱] (LogR) (لیائو و چین،۲۰۰۷; شواده و کرتی[۲]، ۲۰۰۳)، تحلیل پروبیت[۳] (سرور و همکاران[۴]، ۲۰۱۷)، نزدیک‌ترین همسایگی[۵] (KNN) (ساهیگارا و همکاران[۶]، ۲۰۱۳)، برنامه‌های ریاضی (کاستیلو و همکاران،۲۰۱۱) زنجیرهٔ مارکوف[۷] (کاپوزیلو و همکاران[۸]، ۲۰۱۱)، سیستم‌های خبره[۹] (لیائو[۱۰]، ۲۰۰۵)، الگوریتم ژنتیک[۱۱] (GA) (ریوز و رو[۱۲]، ۲۰۰۲) و شبکه‌های عصبی مصنوعی[۱۳] (ANN) (دونگار و همکاران[۱۴]، ۲۰۱۲)، درخت تصمیم CART[۱۵] (تیموفیو[۱۶]، ۲۰۰۴) برای طبقه‌بندی و رگرسیون، و ماشین بردار پشتیبان[۱۷] (SVM) (گویون و همکاران، ۲۰۰۲; فوری و همکاران[۱۸]، ۲۰۰۰) تکنیک‌های جدیدی‌اند که اخیراً برای ساخت امتیاز اعتباری از آن‌ها استفاده شده است (یپ، هنگ و هوساین[۱۹]، ۲۰۱۱).

سایجاده، تاگیو، و کاسمیران[۲۰] (۲۰۱۸) با استفاده از روش‌های رگرسیون لجستیک درخت تصمیم و جنگل‌های تصادفی، به پیش‌بینی امتیاز اعتباری افراد پرداختند. در این مطالعه، ۲۴ متغیر ورودی مورد بررسی قرار گرفت. برخی از این متغیرها جنسیت، تحصیلات،

---

[1] Logistic Regression
[2] Liao and Chin; Shevade and Keerthi
[3] Probit Analysis
[4] Sarwar et al.
[5] K-Nearest-Neighbours (KNN)
[6] Sahigara et al.
[7] Markov Chain
[8] Capozziello et al.
[9] Expert System
[10] Liao
[11] Genetic Algorithms (Ga)
[12] Reeves & Rowe
[13] Artificial Neural Network (ANN)
[14] Dongare et al.
[15] Classification and Regression Tree (CART)
[16] Timofeev
[17] Support Vector Machine (SVM)
[18] Guyon et al.; Furey et al
[19] Yap, Ong and Husain
[20] Sayjadah, Tagio and Kasmiran


[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

سن، اعتبار فعلی، و بدهی‌ها مربوط به شش ماه گذشتهٔ افراد بود. در این پژوهش، روش جنگل‌های تصادفی با دقت بالای ۸۰ درصد، بالاترین دقت را نسبت به سایر روش‌های ارائه‌شده برای پیش‌بینی امتیاز اعتباری افراد داشت .

اکتر (۲۰۱۸) با استفاده از روش‌های رگرسیون لجستیک، نزدیک‌ترین همسایگی، درخت تصمیم و جنگل‌های تصادفی به پیش‌بینی بهترین مقدار وام برای افراد، به‌طوری که سود وام‌دهنده حداکثر شود، پرداخت. در این پژوهش، از تعداد دادهٔ دریافت‌شده از بانک با ۲۰ متغیر استفاده شد که ازجملهٔ مهم‌ترین این متغیرها مقدار وام دریافت‌شدهٔ فرد، امتیاز اعتباری فعلی فرد، وضعیت محل سکونت، تعداد حساب‌های فرد، مقدار بدهی‌های پرداخت‌نشده و بیشترین مقدار بدهی بود. نتایج نشان داد روش رگرسیون لجستیک از سایر روش‌ها دقت بالاتری دارد، درحالی‌که روش درخت تصمیم دارای پایین‌ترین دقت است.

وانگ، ژیان، هووانگ و کایکوان[1] (۲۰۱۱) به بررسی روش‌های ترکیبی درخت تصمیم برای داده‌های پایگاه دادهٔ UCI برای کشورهای آلمان و استرالیا پرداختند. آن‌ها بیان کردند باوجود محبوبیت درخت تصمیم، این روش نسبت به روش‌های دیگر دارای نتیجهٔ ضعیف‌تری است، زیرا به‌راحتی تحت‌تأثیر داده‌های پرت قرار می‌گیرد. آن‌ها مشاهده کردند استفاده از درخت تصمیم به‌تنهایی نتیجهٔ ضعیف‌تری نسبت به روش‌های رگرسیون لجستیک، تحلیل تفکیک خطی، و نزدیک‌ترین همسایگی داشته است؛ اما با استفاده از روش‌های جنگل تصادفی، روش تقویتی، و روش بسته‌بندی، دقت مدل افزایش پیدا می‌کند.

ننی و لومینی[2] (۲۰۰۹) روش‌های ترکیبی را برای طبقه‌بندی داده‌های آلمان و استرالیا در پایگاه دادهٔ UCI برای پیش‌بینی امتیاز اعتباری و ورشکستگی مقایسه کردند. آن‌ها مشاهده کردند در روش‌های ترکیبی، کارایی نسبت به استفاده از روش‌هایی مانند شبکهٔ عصبی به‌تنهایی، افزایش می‌یابد. در این مطالعه، روش تقویتی بیشترین دقت را برای پیش‌بینی داشت.

لی[3] (۲۰۰۹) چهار روش را با روش طبقه‌بندی نزدیک‌ترین همسایگی ادغام کرد تا با استفاده از آن به انتخاب ویژگی‌های بااهمیت‌تر برای طبقه‌بندی بپردازد. وی از داده‌های پایگاه دادهٔ UCI برای کشور آلمان و استرالیا استفاده کرد. روش نزدیک‌ترین همسایگی با روش‌های تحلیل تفکیک خطی، درخت تصمیم، داده‌های سخت و روش‌های F ـ امتیاز[4]

---

[1] Wang, Jian, Huang and Kaiquan

[2] Nanni and Lumini

[3] Li

[4] F-Score



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

ترکیب شد. در نهایت، نتایج این چهار روش مقایسه شد و ترکیب روش F_امتیاز در مرحلهٔ پیش‌پردازش داده‌ها با روش نزیک‌ترین همسایگی، دارای بیشترین دقت برای داده‌های آلمان و استرالیا بود.

سای، خیودان، خیاوو، و ژانگ[1] (۲۰۰۹) مدلی از الگوریتم ژنتیک برای امتیاز اعتباری افراد ارائه دادند که از تابع ارزیابی مناسبی استفاده می‌کند و در آن از اطلاعات مشتریان خوب و بد در مجموعهٔ آموزش استفاده می‌شود. نتایج آزمایش‌ها روی پایگاه دادهٔ اعتباری آلمان، نشان‌دهندهٔ دقت قابل‌قبول ۹۱/۱۸ درصد برای مشتریان خوب بود؛ اما برای نمونه‌های متعلق به دستهٔ مشتریان بد کارایی ضعیف ۵۶/۲۵ درصد را داشت.

ابدو[2] (۲۰۰۹)، با استفاده از الگوریتم ژنتیک، به تحلیل مدل‌های امتیاز اعتباری با استفاده از داده‌های بخش عمومی بانک‌های یونان پرداخت و این روش را با تحلیل پروبیت که جایگزین خوبی برای رگرسیون لجستیک و معیار وزن شواهد[3] (مدلی از امتیاز اعتباری که تمرکز آن بر احتمال نسبت امتیازهای خوب به امتیازهای بد است) است، مقایسه کرد. نتایج تجربی نشان دادند که الگوریتم ژنتیک بالاترین دقت و پایین‌ترین خطای نوع اول و نوع دوم را در مقایسه با دو روش دیگر دارد. درعین‌حال، روش‌های الگوریتم ژنتیک، پایین‌ترین هزینهٔ طبقه‌بندی غلط را که مربوط به وزن شواهد است، ندارد.

هوانگ، چن، و وانگ[4] (۲۰۰۷) کارایی الگوریتم ژنتیک را با شبکه‌های عصبی، ماشین بردار پشتیبان، و مدل‌های درخت تصمیم C4.5، در کاربردهای امتیاز اعتباری با استفاده از داده‌های استرالیا و آلمان در پایگاه دادهٔ UCI[5] مقایسه کردند. داده‌های اعتباری استرالیا شامل دسته‌ای به نام «لایق اعتبار»[6] با ۳۰۷ نمونه و دسته‌ای به نام «نالایق برای دریافت اعتبار» با ۳۸۳ نمونه هرکدام با ۶ ویژگی اسمی، ۸ ویژگی عددی، و ستونی با برچسب «قبول» یا «رد» بود. داده‌های اعتباری آلمان شامل ۷۰۰ نمونه در دستهٔ لایق اعتبار و ۳۰۰ نمونه در دستهٔ «نالایق برای دریافت اعتبار» هرکدام با ۲۴ ویژگی به‌عنوان ورودی بود. مطالعهٔ آن‌ها

---

[1] Cai, Xiaodan, Xiaoyu and Zhong

[2] Abdou

[3] Weight Of Evidence (WOE)

[4] Huang, Chen & Wang

[5] UCI Repository of Machine learning Databases

[6] Creditworthy



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

دقت بهتر طبقه‌بندی الگوریتم ژنتیک را نسبت به روش‌های دیگر نشان داد، هرچند این تفاوت اندک بود.

ژانگ، هوانگ، چن، و ژیانگ[1] (۲۰۰۷) ابتدا به مقایسهٔ سه الگوریتم پس انتشارخطا[2] شبکهٔ عصبی، الگوریتم ژنتیک و ماشین بردار پشتیبان برای مدل امتیاز اعتباری پرداختند و سپس مدلی ترکیبی از این سه الگوریتم ایجاد کردند، به‌طوری‌که این مدل نتیجهٔ بهتری روی داده‌های استرالیا و آلمان نسبت به نتایج هرکدام از الگوریتم‌ها به‌شکل تنها ارائه داد.

هوانگ، تزنگ، و اونگ[3] (۲۰۰۶) برنامه‌نویسی دومرحله‌ای ژنتیک (به‌نام 2SGP[4]) را ارائه کردند. در این الگوریتم، با ترکیب مزایای قواعد «اگر–آن‌گاه[5]» و توابع تفکیک‌کننده، مشکلات امتیاز اعتباری برطرف شد. از نتایج تجربی به‌دست‌آمده روی پایگاه دادهٔ اعتباری استرالیا و آلمان، پژوهشگران به این نتیجه رسیدند که روش 2SGP دارای انعطاف بیشتری است و نسبت به مدل‌های دیگر موردبررسی (شبکه‌های عصبی چندلایه، درخت تصمیم CART، درخت تصمیم C4.5، قوانین نزدیک‌ترین همسایگی، داده‌های سخت و رگرسیون لجستیک) به اندازهٔ قابلِ‌توجهی بهتر کار می‌کند.

مرور ادبیات نشان می‌دهد بیشتر مطالعات انجام‌شده در زمینهٔ دسته‌بندی مشتریان بانکی به دو دستهٔ مشتریان خوب و بد انجام شده است و مطالعات برای داده‌های بدون امتیاز اعتباری بسیار محدود بوده است. همچنین، هریک از این تحقیقات عوامل مختلفی را در مدل‌سازی موردتوجه قرار داده‌اند و به نتایج و قوانین مختلفی دست یافته‌اند. در هریک از این تحقیقات از روش‌های متنوعی برای ارزیابی و تهیهٔ مدل استفاده شده و نتایج حاصل از روشی که بیشترین دقت را داشته، به‌عنوان نتیجهٔ تحقیق در نظر گرفته شده است. با توجه به کاربرد بالای روش‌های نزدیک‌ترین همسایگی، درخت تصمیم، و جنگل تصادفی در پژوهش‌های ارائه‌شده در قسمت مرور ادبیات و پیشینهٔ پژوهش، این سه روش به‌عنوان روش‌های منتخب برای تعیین و پیش‌بینی امتیاز اعتباری در این پژوهش انتخاب شده‌اند. پس از اجرای این سه روش و مقایسهٔ دقت آن‌ها با هم، روش با بالاترین دقت به‌عنوان بهترین روش برای پیش‌بینی امتیاز اعتباری اشخاص انتخاب می‌شود.

---

[1] Zhang, Huang, Chen and Jiang

[2] Backpropagation (BP)

[3] Huang, Tzeng and Ong

[4] Two-Stage Genetic Programming

[5] If-Then



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

## ۳ روش‌شناسی

این پژوهش براساس هدف پژوهش، از نوع پژوهش‌های کاربردی است. از نظر نحوهٔ گردآوری داده‌ها، از نوع پژوهش‌های توصیفی-پیمایشی است. هدف اصلی از انجام این پژوهش پیش‌بینی امتیاز اعتباری برای مشتریان بانک با استفاده از روش‌های داده‌کاوی است؛ به همین منظور، باید چگونگی تعیین امتیاز اعتباری برای افراد بررسی شود. همین‌طور باید متغیرهای ورودی برای تعیین امتیاز اعتباری و روش مناسب از میان روش‌های داده‌کاوی برای تعیین و پیش‌بینی امتیاز اعتباری افراد مشخص شود. به‌منظور نائل‌شدن به هدف اصلی و اهداف فرعی پژوهش سؤالات زیر پاسخ داده خواهد شد: امتیاز اعتباری مناسب برای هر فرد چقدر است؟ متغیرهای ورودی (اطلاعات مالی و غیرمالی موردنیاز) برای تعیین و پیش‌بینی امتیاز اعتباری فرد کدام‌اند؟ کدام‌یک از روش‌های داده‌کاوی (روش‌های نزدیک‌ترین همسایگی، درخت تصمیم، و جنگل تصادفی) برای تعیین امتیاز اعتباری افراد موردنظر نتایج بهتری می‌دهد؟

قلمرو موضوعی این پژوهش پیش‌بینی امتیاز اعتباری مشتریان یکی از بانک‌های کشور است. همچنین، از اطلاعات حساب مشتریانی که از فروردین ۱۳۹۶ تا اسفند ۱۳۹۷ از بانک مذکور وام دریافت کرده‌اند، استفاده شده است. جامعهٔ آماری این پژوهش ۷۵،۲۱۳ نفر از مشتریان یکی از بانک‌های کشور از سال ۱۳۹۶ تا سال ۱۳۹۷ است که به‌صورت تصادفی انتخاب شده‌اند.

فرایند اجرایی پژوهش بدین‌صورت است که ابتدا با بررسی مقالات و منابع موجود، متغیرهای ورودی (اطلاعات مالی و غیرمالی) تعیین؛ سپس داده‌ها از پایگاه دادهٔ موجود گردآوری شده است (داده‌ها با بررسی حساب‌های افرادی که در گذشته به بانک مراجعه کردند و وام‌های افراد که در گذشته ثبت شده است، به‌دست می‌آیند). پس از آماده‌سازی و پیش‌پردازش داده‌ها، الگوریتم‌های داده‌کاوی برای پیش‌بینی مشخص می‌شوند. در نهایت، مدل مناسب برای پیش‌بینی امتیاز اعتباری برای هر فرد انتخاب می‌گردد.

به‌منظور تعیین امتیاز اعتباری افراد، از رویکرد داده‌کاوی بر پایهٔ روش‌شناسی CRISP-DM استفاده می‌شود. روش‌شناسی CRISP-DM به‌عنوان روش مرجع فرایند داده‌کاوی به‌کار می‌رود. این روش‌شناسی از گام‌های شناخت سیستم، شناخت داده‌ها، آماده‌سازی داده‌ها، مدل‌سازی، ارزیابی، و توسعهٔ سیستم تشکیل شده است (غضنفری و همکاران، ۱۳۸۷).



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

در این پژوهش، ابتدا داده‌ها گردآوری، و پس از آماده‌سازی و پیش‌پردازش داده‌ها، عملیات خوشه‌بندی داده‌ها انجام می‌شود. سپس با به‌کارگیری نرم‌افزار پایتون[1]، به بررسی روش‌های مختلف پیش‌بینی پرداخته می‌شود.

پیش‌بینی نوعی عملیات برای تحلیل داده‌ها و استخراج مدل به‌منظور توصیف داده‌ها، فهم، و پیش‌بینی رفتار آیندهٔ آن‌هاست. مدل‌های دسته‌بندی در تحلیل داده‌های گسسته و طبقه‌ای به‌کار رفته و مدل‌های رگرسیونی بر روی داده‌های پیوسته عمل می‌کنند. بسیاری از روش‌های دسته‌بندی برای تخصیص یک برچسب به مجموعه‌ای از داده‌ها که هنوز دسته‌بندی نشده‌اند، استفاده می‌شود. پس از آن، داده‌ها براساس ویژگی‌هایشان به‌دسته‌هایی که برچسب آن‌ها از قبل مشخص‌اند، تخصیص داده می‌شوند. به بیان دیگر، دسته‌بندی برای یادگیری قواعد و یا ساختن مدلی به‌منظور پیش‌بینی دستهٔ داده‌های جدید به‌کار می‌رود. داده‌های مورداستفاده برای ساختن مدل، داده‌های آموزش[2] مدل نامیده می‌شوند. فرایند دسته‌بندی یا رگرسیون طی دو مرحلهٔ ساخت مدل و استفاده از مدل به شرح زیر انجام می‌شود:

- **ساخت مدل:** این مرحله شامل توصیف یکسری از داده‌های از پیش تعیین‌شده بر مبنای مجموعهٔ داده‌های آموزش مدل است و این فرایند «یادگیری» نامیده می‌شود. در این فرایند، سعی می‌شود با توجه به نمونه‌های موجود، مدلی ساخته شود که براساس آن بتوان برای داده‌های فاقد برچسب، برچسب مناسب مربوط به خودشان را پیش‌بینی کرد.
- **استفاده از مدل:** این مرحله دارای دو بخش است. در بخش نخست، مدل ساخته‌شده مورد آزمون واقع می‌شود تا دقت پیش‌بینی آن بررسی شود. در بخش دوم نیز مدلی که دارای دقت مناسبی است، برای دسته‌بندی داده‌ها به‌کار گرفته می‌شود. به‌منظور تخمین دقت پیش‌بینی مدل، مجموعه‌ای از داده‌های آموزش به‌طور اتفاقی از میان داده‌ها انتخاب و مدل روی آن‌ها اجرا می‌شود. هدف اصلی در اینجا بالاتربردن تخمین دقت مدل است تا به‌هنگام استفاده، به هر داده برچسب مناسب آن تخصیص داده شود.

برای این کار، داده‌های آموزشی را می‌توان به دو قسمت تقسیم کرد: اول آن دسته که مدل براساس آن‌ها ساخته می‌شود و دوم گروهی که برای ارزیابی مدل استفاده می‌شود. داده‌های گروه دوم را که برچسب آن‌ها مشخص است به مدل داده و خروجی مدل را با

---

[1] Python 3.4

[2] Training data



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

برچسب مشخص‌شده توسط مدل مقایسه کرده و دقت کل مدل استخراج می‌شود. در واقع، برچسب شناخته‌شده از نمونهٔ آزمون با نتایج پیش‌بینی مقایسه می‌شود. دقت مدل، درصد تعداد دفعاتی است که نمونه‌های آموزشی با موفقیت پیش‌بینی می‌شوند. اگر دقت مدل قابل‌قبول باشد، می‌توان مدل را برای پیش‌بینی داده‌هایی که برچسب آن‌ها مشخص نیستند به‌کار برد (غضنفری، علیزاده و تیمورپور، ۱۳۸۷).

روش‌های مختلفی برای پیش‌بینی داده‌ها وجود دارد. در این پژوهش در راستای پیش‌بینی امتیاز اعتباری مشتریان بانک، از روش تحلیل مؤلفه‌های اصلی[1] برای کاهش متغیرهای اصلی، از روش خوشه‌بندی k-means برای خوشه‌بندی نمونه‌ها، و از روش‌های نزدیک‌ترین همسایگی، درخت تصمیم، و جنگل تصادفی برای تعیین و پیش‌بینی امتیاز اعتباری استفاده می‌شود. همچنین، قوانین انجمنی استخراج می‌شود. در ادامه، این روش‌ها به‌اختصار توضیح داده شده است.

### ۳. ۱ تحلیل مؤلفه‌های اصلی

یکی از عمومی‌ترین روش‌های آماری به‌منظور کاهش ابعاد داده‌ها، روش تحلیل مؤلفه‌های اصلی است. روش تحلیل مؤلفه‌های اصلی در سال ۱۹۰۱ برای اولین بار توسط کارل پیرسن ارائه شد (پیرسون[2]، ۱۹۰۱). او توانست برای دو یا سه صفت خاصه (متغیر یا ویژگی) روش کاربردی ارائه دهد. تحلیل مؤلفه‌های اصلی در تعریف ریاضی، یک تبدیل خطی متعامد است که داده را به دستگاه مختصات جدید می‌برد، به‌طوری که بزرگ‌ترین واریانس داده بر روی اولین محور مختصات و دومین بزرگ‌ترین واریانس بر روی دومین محور مختصات قرار می‌گیرد و همین‌طور برای باقی مؤلفه‌های اصلی. تحلیل مؤلفه‌های اصلی می‌تواند برای کاهش ابعاد داده مورداستفاده قرار بگیرد؛ به‌این‌ترتیب، مؤلفه‌هایی از مجموعهٔ داده را که بیشترین تأثیر را در واریانس دارند، حفظ می‌کند. هدف تحلیل مؤلفه‌های اصلی کاهش تعداد زیاد متغیرهای اصلی به تعداد کمی از مؤلفه‌های اصلی است. تحلیل مؤلفه‌های اصلی وقتی نتیجهٔ بهتری خواهد داشت که متغیرهای اصلی با هم تا حد زیادی همبستگی داشته باشند (آذر و خدیور، ۱۳۹۳).

---

[1] Principal Component Analysis (PCA)

[2] Pearson



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

### ۳. ۲ روش خوشه‌بندی k-means

امروزه، الگوریتم k-means محبوب‌ترین ابزار خوشه‌بندی مورداستفاده در علوم و صنعت است. در روش خوشه‌بندی k-means، پارامتر k به‌عنوان ورودی گرفته می‌شود و مجموعهٔ n شیئی را به k خوشه افراز می‌کند. این روش به این دلیل k-means نام گرفته است که هرکدام از k خوشهٔ $C_j$، با میانگین (یا میانگین وزنی) $c_j$ آن نشان داده می‌شود و آن را سنترویید[1] می‌نامیم(برکین[2]، ۲۰۰۶).

### ۳. ۳ نزدیک‌ترین همسایگی

یکی از رایج‌ترین کاربردهای الگوریتم نزدیک‌ترین همسایگی، تشخیص الگوست. برای یک دادهٔ آزمایشی الگوریتم به‌دنبال k نمونه از نزدیک‌ترین نمونه‌ها می‌گردد (k نمونهٔ مشابه). نزدیکی دو نمونه با به‌دست‌آوردن تشابه و یا فاصله میان این دو نمونه محاسبه می‌شود. هر نمونه از انواع داده‌ها تشکیل شده است که باید تشابه میان آن‌ها بررسی شود. اگر مسئله از نوع طبقه‌بندی باشد، پس از یافتن k داده مشابه با نمونهٔ آزمایشی، با رأی اکثریت، برچسب طبقهٔ دادهٔ آزمایشی انتخاب می‌شود. اگر مسئله دارای برچسب از نوع پیوسته باشد، پس از یافتن k همسایه، میانگین مقادیر حاصل از برچسب این k نمونه به‌عنوان برچسب نمونهٔ آزمایشی برگزیده می‌شود (اسماعیلی، ۱۳۹۳).

### ۳. ۴ درخت تصمیم

درخت تصمیم در داده‌کاوی مدلی است که برای نمایش طبقه‌بندی‌ها و رگرسیون‌ها استفاده می‌شود. همان‌طور که از نام آن مشخص است، این درخت از تعدادی گره و شاخه تشکیل شده است. در درخت تصمیم، بالاترین گره درخت، گره‌های برگ، دسته‌ها یا توزیع دسته‌ها را نشان می‌دهند. در یکی از گره‌های دیگر (گره‌های غیر از برگ) با توجه به یک یا چند صفت خاصه، تصمیم‌گیری صورت می‌گیرد. درخت تصمیم به‌دلیل سادگی و قابلِ‌فهم‌بودن، تکنیک محبوبی در داده‌کاوی محسوب می‌شود. درخت تصمیم خود به‌تنهایی همهٔ مطالب را توصیف می‌کند و نیاز به فرد خبره‌ای نیست تا خروجی را تفسیر کند. در واقع، این یک روش گرافیکی است و بدین‌دلیل تفسیر آن شاید ساده‌تر از روش‌های دیگر باشد (غضنفری و همکاران، ۱۳۸۷).

---

[1] Centroid

[2] Berkhin



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

روش درخت تصمیم در تقسیم‌بندی داده‌ها به گروه‌های مختلف، به‌گونه‌ای است که هیچ داده‌ای حذف نمی‌شود. باوجود اینکه فهمیدن روش کار الگوریتم‌های سازندهٔ درخت چندان ساده نیست، فهمیدن نتایج به‌دست‌آمده از آن‌ها آسان است. دسته‌بندی‌هایی که توسط درخت تصمیم ایجاد می‌شود، از روی شباهت داده‌های ذخیره‌شده در پارامترهای پیش‌بینی‌کننده انجام‌پذیر است. درخت تصمیم، برخلاف شبکه‌های عصبی، به تولید قاعده می‌پردازند. در ساختار درخت تصمیم، پیش‌بینی به‌دست‌آمده از درخت، در قالب یک‌سری قواعد توضیح داده می‌شود؛ درحالی‌که در شبکه‌های عصبی تنها نتیجهٔ پیش‌بینی بیان می‌شود و چگونگی به‌دست‌آمدن آن‌ها در خود شبکه پنهان می‌ماند. همچنین در درخت تصمیم، برخلاف شبکه‌های عصبی، ضرورتی وجود ندارد که داده‌ها لزوماً به‌صورت عددی باشند (غضنفری، علیزاده و تیمورپور، ۱۳۸۷).

الگوریتم‌های متعددی برای ساخت درخت تصمیم وجود دارد که ازجملهٔ این موارد می‌توان به الگوریتم‌های C4.5 Quest (که یکی از تعمیم‌های الگورتیم ID3 است که بعدها با تغییراتی با نام الگوریتم C5.0 در بعضی از منابع مطرح شد)، CART، و CHAID اشاره کرد (غضنفری و همکاران، ۱۳۸۷).

در سال ۱۹۸۴، بریمن و همکارانش الگوریتم CART را ایجاد کردند. نتیجهٔ این الگوریتم یک درخت تصمیم دودویی است؛ بدین‌معنی که هر گره داخلی دارای دو انشعاب است. همچنین، این الگوریتم روشی برای هرس‌کردن دارد. یکی از ویژگی‌های مهم CART توانایی تولید درختان رگرسیون است. برگ‌ها در چنین درختی یک عدد واقعی را به‌جای برچسب کلاس تخمین می‌زنند (فریدمن، اولشن و استون[1]، ۱۹۸۴).

### ۳. ۵ الگوریتم جنگل تصادفی

جنگل تصادفی یکی از جدیدترین روش‌های یادگیری است. در این روش، مجموعه یا جنگلی از درختان تصمیم برای تخمین یا تصمیم‌گیری مورداستفاده قرار می‌گیرند. همچنین، استفاده از این روش برای کاربر بسیار ساده است، زیرا تنها دو پارامتر دارد (تعداد متغیرها در زیرمجموعهٔ تصادفی در هر گره و تعداد درختان جنگل). در جنگل‌های تصادفی، علاوه بر اینکه برای ساخت هر درخت از نمونهٔ تصادفی با جای‌گذاری متفاوت از داده‌ها استفاده می‌شود، نحوهٔ ساخت طبقه‌بندی یا رگرسیون برای هر درخت نیز متفاوت است. در درختان استاندارد، هر گره براساس بهترین انشعاب در میان متغیرها منشعب می‌شود. در جنگل‌های

[1] Friedman, Olshen and Stonr



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

تصادفی، هرگره براساس بهترین زیرمجموعهٔ پیشگو، که به‌طور تصادفی برای هرگره انتخاب شده‌اند، منشعب می‌شود. این راهبرد نسبت به کارایی بقیهٔ طبقه‌بندی‌ها و رگرسیون‌ها، ازجمله تحلیل تفکیک‌کننده و ماشین بردار پشتیبان، به‌خوبی کار می‌کند و نسبت به بیش‌برازش، قدرتمند است (Liaw and Wiener, 2002).

### ۳. ۶ قوانین انجمنی

استخراج قواعد انجمنی نوعی عملیات داده‌کاوی است که به جست‌وجو برای یافتن ارتباط بین ویژگی‌ها در مجموعهٔ داده‌ها می‌پردازد. این روش به‌دنبال استخراج قواعد به‌منظور کمّی‌کردن ارتباط میان دو یا چند خصوصیت است. قواعد انجمنی به‌شکل اگر و آنگاه به‌همراه دو معیار پشتیبان و اطمینان تعریف می‌شوند. قواعد انجمنی ماهیتاً قواعد احتمالی‌اند (غضنفری و همکاران، ۱۳۸۷).

## ۴ نتایج

مطابق با روش‌شناسی CRISP-DM، فازهای مختلف روش‌شناسی پیاده‌سازی می‌شود و پس از شرح مراحل مختلف مربوط به آماده‌سازی و پیش‌پردازش داده‌ها، از الگوریتم خوشه‌بندی استفاده می‌شود و برای داده‌ها تعداد مناسبی خوشه در نظرگرفته می‌شود؛ سپس از میان این خوشه‌ها تعداد ۱۲۰ داده به‌طور تصادفی انتخاب می‌شوند و برای این مجموعهٔ کوچک از داده‌ها با استفاده از نظر خبره، برچسب مناسب تعیین می‌شود؛ سپس با استفاده از الگوریتم‌های یادگیری با ناظر، مراحل آموزش و آزمون انجام می‌شود و با بررسی نتایج، از الگوریتمی که بالاترین دقت را دارد برای پیش‌بینی امتیاز اعتباری داده‌های باقیمانده استفاده می‌شود. الگوریتم‌های مورداستفاده عبارت‌اند از: الگوریتم k-means برای خوشه‌بندی داده‌ها به تعداد مناسب خوشه و الگوریتم‌های نزدیک‌ترین همسایگی، درخت تصمیم، و جنگل تصادفی برای پیش‌بینی.

### ۴. ۱ شناخت سیستم

مسئلهٔ اعتبارسنجی مشتریان بانک مسئله‌ای حیاتی برای این مؤسسات است، زیرا بحران‌های مالی می‌تواند هزینه‌های اجتماعی زیادی به‌بار آورد که می‌تواند بر سهام‌داران، مدیران، کارگران، وام‌دهندگان، تأمین‌کنندگان، مشتریان، جامعه، دولت، و غیره تأثیرگذار باشد. بنابراین، پیش‌بینی امتیاز اعتباری از اهمیت بالایی برخوردار است. با داشتن امتیاز اعتباری، بانک می‌تواند تصمیمات بهتری مانند قبول یا رد درخواست وام، اندازهٔ وام، مبلغ اقساط، و



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

نحوهٔ پرداخت وام برای مشتریانش بگیرد، و همچنین خطاهای ناشی از تصمیمات قضاوتی را کاهش دهد. در این پژوهش، داده‌های موجود از اطلاعات حساب‌های بانکی یکی از بانک‌های کشور برای گروهی از افراد گردآوری شده و با کمک نرم‌افزار پایتون و به‌کارگیری روش‌های داده‌کاوی، امتیاز اعتباری برای این افراد تعیین و پیش‌بینی می‌شود و مؤثرترین عوامل برای این امتیاز مشخص می‌شوند.

## ۴. ۲ شناخت داده‌ها

هدف از این گام شناخت مجموعهٔ داده‌ای است که باید مورد کاوش قرار گیرد و شامل به‌دست‌آوردن و شناخت داده‌هاست. در این پژوهش، بانکِ اطلاعاتی شامل اطلاعات حساب بانکی و وام بانکی بخشی از مشتریان یکی از بانک‌های کشور از سال ۱۳۹۵ تا ۱۳۹۶ است. این مجموعهٔ داده شامل ۳۳ متغیر برای ۷۵,۲۲۶ مشتری است. بررسی جامعی به‌منظور شناخت بیشتر داده‌های موجود در بانکِ اطلاعاتی و بازه مقادیر در خصوص هریک از متغیرها انجام شد که نتایج آن در جدول ۱ مشاهده می‌شود.

جدول ۱

*شناسایی متغیرهای پژوهش*

| نام متغیر | شرح | فراوانی | کمترین مقدار | بیشترین مقدار |
|---|---|---|---|---|
| اطلاعات مشتری | | | | |
| شناسهٔ مشتری | شناسهٔ منحصربه‌فرد برای هر شخص | ۱۶۵۸۳۸۰ نفر | -- | -- |
| سال بازکردن حساب | -- | -- | ۱۳۸۸ | ۱۳۹۴ |
| نوع شخص | حقیقی | ۱۶۵۷۳۷۱ | -- | -- |
| | حقوقی | ۱۰۰۹ | | |
| چک برگشتی | دارد | ۱۵۳۷۵ | -- | -- |
| | ندارد | ۱۶۴۳۰۰۵ | | |
| ملیت | ۳۳ کشور متفاوت | ۱۶۴۵۵۰۸ ایرانی | -- | -- |
| | | ۱۲۸۷۲ غیرایرانی | | |
| جنسیت | مرد | ۱۰۷۸۷۷۰ | -- | -- |
| | زن | ۳۵۲۰۹۲ | | |
| | نامعلوم | ۲۷۵۱۸ | | |
| وضعیت تأهل | مجرد | ۱۱۴۵۵۹۰ | -- | -- |



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

| | | | | |
|---|---|---|---|---|
| | متأهل | ۴۶۴۷۹۲ | | |
| | بیوه | ۱۹۸۷۳ | | |
| | مطلقه | ۲۳۳۲۷ | | |
| | نامعلوم | ۴۷۹۸ | | |
| | اطلاعات حساب | | | |
| سال | سال موجودی حساب | ـ ـ | ۱۳۹۵ | ۱۳۹۶ |
| ماه | ماه موجودی حساب | ـ ـ | ۱ | ۱۲ |
| کد شعبه | ـ ـ | ـ ـ | ـ ـ | -- |
| کد موجودی شماره | ـ ـ | ـ ـ | -- | -- |
| شماره حساب | شماره حساب(های) فرد | ـ ـ | -- | -- |
| رقم کنترلی | ـ ـ | ـ ـ | -- | -- |
| موجودی | موجودی حساب در ماه و سال مربوطه | ـ ـ | ـ۲۴۹۱۹۰۰۰۰ | ۱۹۳۰۹۵۲۴۰۰۹۷ |
| | اطلاعات وام | | | |
| مبلغ قسط | ـ ـ | ـ ـ | ۰ | ۱۱۰۵۳۸۷۰۰۰ |
| معوق ماه | تعداد ماه‌های معوق | ـ ـ | ۰ | ۷۶ |
| مبلغ وام | ـ ـ | ـ ـ | ۱۰۰۰ | ۱۲۶۴۰۰۰۰۰۰۰ |
| نوع وام | نوع وام | ـ ـ | ـ ـ | ـ ـ |
| مانده وام | مبلغ باقی‌مانده از وام | ـ ـ | ۰ | ۱۲۱۵۹۲۵۰۰۰۰ |
| مانده جاری | مبلغ باقی‌مانده وام که هنوز موعد پرداخت آن نرسیده است. | ـ ـ | ۰ | ۱۲۱۵۹۲۵۰۰۰۰ |
| مانده غیرجاری | مبلغ باقی‌مانده وام که از موعد پرداخت آن گذشته است. | ـ ـ | ۰ | ۲۷۲۲۰۲۳۰۰۰ |
| نرخ سود وام | نرخ سود تعیین‌شده توسط بانک | ـ ـ | ۰ | ۲۶/۵ |
| تعداد اقساط | ـ ـ | ـ ـ | ۰ | ۲۷۶ |
| تاریخ پرداخت | تاریخ آغاز وام | ـ ـ | ۱۳۹۰/۳ | ۱۳۹۶/۱۲ |
| تاریخ پایان | تاریخ پایان وام | ـ ـ | ۱۳۹۱/۳ | ۱۳۹۸/۱۲ |

## ۴. ۳ آماده‌سازی داده‌ها

هدف از این گام افزایش کیفیت داده‌های مورداستفاده است؛ به‌نحوی که بتوان مجموعهٔ داده‌ای مناسب برای فاز بعدی و مدل‌سازی داده‌ها تأمین کرد. این فاز مجموعه‌ای از یک سری فعالیت‌هاست که می‌بایست بر روی داده‌ها انجام شود و نتیجهٔ این فعالیت‌ها به‌عنوان ورودی به مرحلهٔ بعد انتقال یابد. این مرحله شامل گام‌های انتخاب داده، شناسایی داده‌های


[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

پرت، شناسایی داده‌های مفقوده، ادغام و ساخت شاخص‌ها، تلخیص توصیفی داده‌ها، و انتخاب مهم‌ترین متغیرها به‌عنوان ورودی مدل داده‌کاوی است.

**۴. ۳. ۱ حذف رکوردهای تکراری و نامربوط**

در این مرحله با بررسی مجموعهٔ داده‌ها در نرم‌افزار پایتون، رکوردهای تکراری موجود در مجموعهٔ داده‌ها شناسایی شده و رکوردهای اضافی حذف شدند. این رکوردها براثر اشتباه اپراتور و به‌علت نبودِ کنترل کافی در نرم‌افزار برای جلوگیری از ثبت رکوردهای تکراری به‌وجود آمده‌اند. برای شناسایی و حذف رکوردها با بررسی شناسهٔ مشتری، شناسه‌هایی که دوباره تکرار شده‌اند شناسایی شد. با بررسی در مجموعهٔ داده‌ها، رکوردهای اضافی حذف شدند که تعداد این رکوردها تنها ۲۱ رکورد بود.

با توجه به هدف پژوهش که تعیین امتیاز اعتباری افراد است، رکوردهایی که تحت عنوان «شخص حقوقی» ثبت شده‌اند ــ که تعداد آن ۱۳ رکورد است ــ حذف شدند، زیرا مربوط به حوزهٔ این پژوهش نیستند.

با بررسی موجودی حساب‌ها، تعداد ۱۶۳ نفر موجودی حسابشان مبلغی کمتر از صفر بود. به‌علت مشخص‌نبودن معنای موجودی منفی (وجود خطای انسانی در واردکردن اطلاعات یا بدهی به بانک)، رکوردهای مربوط به این ۱۶۳ مشتری از داده‌ها حذف شد.

وام انواع متفاوتی دارد. نحوهٔ پرداخت اقساط برخی از وام‌ها، به‌صورت خودکار در ابتدای ماه از حقوق فرد کاسته شده و به بانک پرداخت می‌شود. اطلاعات مربوط به این وام‌ها (۳۷ رکورد) از داده‌ها حذف شدند. براین‌اساس، از مجموعهٔ ورودی‌ها تعداد ۲۳۴ داده حذف شد.

**۴. ۳. ۲ حذف متغیرهای زائد**

در این مرحله پس از بررسی مقادیر موجود در خصوص هریک از متغیرها، برخی از متغیرها از مجموعهٔ داده‌های موردمطالعه حذف شدند. داده‌هایی که به‌عنوان فردی حقوقی در بانک حساب باز کرده بودند، حذف شدند. بنابراین، تمام حساب‌ها متعلق به افراد حقیقی است. پس، می‌توان ویژگی حقیقی/حقوقی را از میان ویژگی‌ها حذف کرد. اطلاعات مربوط به برخی از وام‌ها که مربوط به حوزهٔ این پژوهش نبودند در بخش قبل حذف گردید، بنابراین متغیر مربوط به نوع وام نیز کنار گذاشته شد. متغیرهای مربوط به کد شعبه، رقم کنترلی، کد موجودی شماره و شماره حساب در تصمیم‌گیری نحوهٔ ارائهٔ وام به مشتری بی‌تأثیر است. با توجه به این موضوع، ستون‌های مربوط به کد شعبه، رقم کنترلی، کد موجودی شماره، و شمارهٔ حساب حذف شد.



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

### ۴. ۳. ۳ شناسایی داده‌های مفقوده

با توجه به اینکه وجود داده‌های مفقوده از قدرت آماری آزمون‌های آماری می‌کاهد و خطای اندازه‌گیری را افزایش می‌دهد، باید این داده‌ها شناسایی و مدیریت شوند. از بررسی داده‌ها با دستور «fillna»، نرم‌افزار مقادیر مفقوده را شناسایی و با مقداری مشخص که توسط کاربر تعیین می‌شود، جایگزین می‌کند. مقدار جایگزین می‌تواند مقدار ثابت، میانگین ویژگی‌ها، مقدار رکورد بالایی، مقدار رکورد پایینی، یا مقادیر دیگری که برای آن تعیین می‌شود، باشد. در این پژوهش، مقادیر مفقوده با مقادیر متفاوتی ازجمله میانگین ویژگی‌ها، مقدار رکورد بالایی، و یا با تابعی از متغیر دیگر در همان رکورد جایگزین شدند.

در ستون مربوط به جنسیت، تعداد ۲۷،۵۱۸ داده با جنسیت نامعلوم وجود داشت، که برای آن‌ها جنسیت «مرد» در نظر گرفته شد. این انتخاب به‌علت فراوانی بیشتر رکوردها با جنسیت «مرد» نسبت به رکوردها با جنسیت «زن» بود. در ستون مربوط به وضعیت تأهل، ۴۷۹۸ داده موجود بود که وضعیت تأهل آن‌ها یا ذکر نشده، یا با علائم اشتباه ثبت شده بودند. در اینجا نیز به‌علت فراوانی بیشتر افراد «مزدوج»، رکوردها با مقادیر مفقوده در وضعیت تأهل، «مزدوج» فرض شدند. در ستون مربوط به موجودی آخر ماه حساب مشتری، تعداد ۱۱۳ مورد دارای مقادیر مفقوده بوده که این عناصر با مقدار عنصر بالایی جایگزین شدند. در ستون مربوط به تعداد اقساط، ۴۲ عنصر دارای مقدار مفقوده بوده که این مقادیر با میانگین این ستون جایگزین شدند.

### ۴. ۳. ۴ افزودن فیلدهای جدید

با بررسی داده‌ها به‌منظور انجام مدل‌سازی، متغیرهای پیشگو به مجموعهٔ داده‌ها اضافه شد. نحوهٔ به‌دست‌آمدن هریک از این متغیرها در ادامه شرح داده شده است.

تعداد وام‌های دریافتی هر فرد را در ستونی جدید ذخیره می‌کنیم. توجه شود که در داده‌های حاضر، وام‌هایی که درحال حاضر در جریان‌اند، موجود است؛ این به آن معناست که تعداد وام‌های دریافتی بیانگر تعداد وام‌هایی است که مشتری در زمان حال درحال پرداخت آن‌هاست و اطلاعاتی دربارهٔ وام‌های گذشتهٔ این مشتری نمی‌دهد. این متغیر طبق متغیرهای موجود در پژوهش چن و هونگ[1] اضافه شده است. با اضافه‌کردن ستون‌های کمترین موجودی و بیشترین موجودی، به‌ترتیب کمترین و بیشترین مقدار موجودی حساب مشتری در دو سال گذشته مشخص می‌شود. این متغیر طبق پژوهش فریتس و هوسمن[2] (۲۰۰۰) به

[1] Chen and Huang
[2] Fritz and Hosmann


[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

متغیرهای موجود اضافه شده است. از اختلاف بین موجودی حساب هر ماه یک مشتری، ستونی به نام نرخ تغییرات موجودی حساب به‌دست می‌آید. این ستون نشان‌دهندهٔ تغییر وضعیت موجودی حساب مشتری در هر ماه است و طبق آن می‌توان روند افزایش یا کاهش موجودی حساب را مشاهده کرد. در واقع، با این ستون تراکنش حساب مشتری در هر ماه مشخص می‌شود. این ستون نیز طبق متغیرهای استفاده‌شده در پژوهش یوباس و کروک[1] به متغیرهای موجود اضافه شده است. با میانگین ستون نرخ تغییرات موجودی حساب که به داده‌ها اضافه شد، نرخ رشد یا کاهش موجودی حساب برای هر مشتری به‌دست می‌آید. این ستون را میانگین نرخ تغییرات موجودی حساب می‌نامیم. تعداد حساب‌های هر فرد را تحت همین عنوان در ستونی جدید می‌توان ذخیره کرد. این ستون مطابق متغیرهای موجود در پژوهش فریتس و هوسمن (۲۰۰۰) اضافه شده است. ستون موجودی به ۳۰٪ اقساط، ستونی است که نسبت موجودی به ۳۰ درصد مبلغ اقساط را برای هر مشتری نشان می‌دهد (اگر فردی بیش از یک وام دریافت کرده باشد، در این ستون مجموع ۳۰ درصد همهٔ اقساط وی، ذخیره می‌شود). برای استفاده از روش vantagescore که جلوتر به تشریح آن پرداخته می‌شود، این ستون اضافه شده است(وستون، ۲۰۱۱[2]). ستون طول مدت وام نشان‌دهندهٔ بازه زمانی است که بانک تعیین کرده است تا وام‌گیرنده بدهی خود را به‌صورت کامل بپردازد. این ستون برحسب سال است. این ستون نیز در راستای استفاده از مدل vantagescore به متغیرهای اصلی اضافه شده است (وستون،۲۰۱۱) ستونی جدید به نام نرخ پرداخت به جدول اضافه شده است، که این ستون بیانگر نسبت مبلغی از وام که فرد باید پرداخت می‌کرده ولی پرداخت نکرده، به کل مبلغی است که تا آن لحظه باید پرداخت می‌کرده است. این ستون طبق مطالعهٔ هونگ و همکاران (۲۰۰۴) [3]به متغیرهای اصلی اضافه شده است. شکل ۱ برای بیان بهتر موضوع ارائه شده است.


*شکل ۱*. وضعیت پرداخت وام مشتری

[1] Yobas, Crook and Ross
[2] Weston
[3] Huang, et al.

### ۴. ۳. ۵ تغییر شکل داده‌ها

در این مرحله، به‌منظور تسهیل مدل‌سازی و استخراج نتایج قابل‌استناد، داده‌ها از حالت اسمی به متغیر دودویی تغییر شکل یافتند. با دودویی‌کردن (صفر و یک) متغیرهای وضعیت تأهل، جنسیت، ملیت، و چک برگشتی، این متغیرها به‌شکل عددی درآمد تا بتوان روی آن‌ها تحلیل لازم را انجام داد. متغیرهای مربوط به تاریخ به دو ستون ماه و سال تبدیل و سپس از آن‌ها استفاده شد، زیرا تاریخ‌های ارائه‌شده تاریخ‌های شمسی هستند و نرم‌افزار نمی‌تواند روی آن‌ها تحلیل انجام دهد.

### ۴. ۳. ۶ کاهش مقادیر یک متغیر

در برخی از متغیرها تعداد داده‌ها برای برخی از گروه‌ها به‌نسبت کم است که این موضوع سبب پراکندگی و عدم تعمیم‌پذیری قوانین به‌دست‌آمده از مدل‌ها خواهد شد. بنابراین، گروهی از داده‌های یک متغیر که امکان ترکیب با گروه دیگر را دارد، ادغام می‌شوند.

برای متغیر وضعیت تأهل، تعداد کمی از داده‌ها در کلاس بیوه و طلاق قرار دارند؛ بنابراین، به‌جهت ازبین‌بردن پراکندگی و نویز[1]، این دو کلاس با کلاس مجرد ادغام شدند. در متغیر ملیت، ۳۳ ملیت برای مشتریان بانک وجود دارد. با ترکیب تمام ملیت‌های غیر از ایران، تحت عنوان «غیر ایرانی»، دو کلاس برای این ستون تهیه شد.

### ۴. ۳. ۷ تشخیص رکوردهای پرت[2]

رکوردهایی که مقادیر پرت دارند، رکوردهای پرت در نظرگرفته می‌شوند و باید از ادامهٔ مراحل حذف شوند. رکوردهای پرت موارد نادری از اکثریت داده‌ها را متمایز کرده و داده‌هایی را که به‌وضوح از محدودهٔ گروه‌های داده خارج شده‌اند، انتخاب می‌کند. رکوردهای پرت ممکن است دلایلی زیادی داشته باشند، ازجمله زمانی‌که خطاهای انسانی یا خطاهای فنی در یک مجموعه داده رخ دهد. به‌منظور تشخیص نقاط پرت، با به‌کارگیری مدل Anomaly Detection با تنظیمات پیش‌فرض در نرم‌افزار پایتون هر رکورد با رکوردهای دیگر مقایسه می‌شود.

### ۴. ۳. ۸ توزیع فراوانی متغیرهای موردبررسی

توزیع فراوانی برای متغیرهایی که در جدول ۱ بیان شد، بر روی مشتریان بانک با فراوانی ۷۵,۰۱۲، در جدول ۲ آمده است.

---

[1] Noise

[2] Anomaly Detection



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

جدول ۲

*میانگین و انحراف معیار متغیرهای کمّی پژوهش*

| | نام متغیر | میانگین | انحراف معیار |
|---|---|---|---|
| حساب | موجودی | ۱۲۹۱۸۴۸۰۱/۳۰۷ | ۵۳۵۶۳۶۶۷۳/۰۰۳ |
| | نرخ تغییرات موجودی حساب | ۵۵۶۳۰۳۷/۱۸۰ | ۱۵۱۷۱۱۹۲۷/۲۶۲ |
| وام | مبلغ قسط | ۳۴۶۹۸۷۲/۳۸۶ | ۸۴۳۳۶۹۰/۹۹۹ |
| | معوق ماه | ۱/۳۲۵ | ۱/۹۴۷ |
| | مبلغ وام | ۱۳۳۶۰۰۰۷۷/۳۷۰ | ۱۷۸۷۶۱۶۰۵/۶۶۲ |
| | مانده | ۱۱۰۳۹۹۲۷۶/۳۸۴ | ۲۲۱۶۳۸۲۴۵/۱۱۲ |
| | مانده جاری | ۱۰۶۷۹۰۷۸۷/۵۸۵ | ۲۱۹۴۸۹۵۱۰/۴۲۰ |
| | مانده غیرجاری | ۳۶۰۸۴۸۸/۷۹۹ | ۱۹۰۶۷۳۴۸/۸۴۸ |
| | نرخ سود | ۹/۲۳۰ | ۷/۰۸۶ |
| | تعداد اقساط | ۵۴/۷۱۸ | ۳۱/۹۲۶ |
| | ۳۰٪ جمع اقساط | ۱۳۳۲۵۴۷ | ۳۰۰۱۳۰۰ |

## ۴. ۴ مدل‌سازی داده‌ها

در این گام، تلاش می‌شود با بهره‌گیری از الگوریتم‌های مختلف بر روی مجموعهٔ داده‌هایی که در مراحل قبل آماده شده است، مدل‌های مناسبی ارائه شود که ارائه‌دهندهٔ قوانینی با دقت بالا و توصیف‌کنندهٔ اطلاعات ارائه‌شده به مدل باشند. با توجه به اینکه نرم‌افزار پایتون انواع مدل‌های متداول طبقه‌بندی و رگرسیونی را پشتیبانی می‌کند، بسته به داده‌های مسئله و پس از آماده‌سازی آن‌ها، کافی است با فراخوانی کتابخانه‌های مناسب، مدل موردنظر را انتخاب کرده و آن را روی داده‌های آماده‌شده اجرا کرد. همچنین، هر مدل دارای تعدادی ورودی و خروجی است که به‌هنگام اجرای مدل باید مشخص باشند. بنابراین در هنگام اجرای هر مدل، باید مقدار پارامترهای موردنیاز مدل مشخص شوند. در مدل‌های این پژوهش، متغیر امتیاز اعتباری به‌عنوان خروجی و سایر متغیرها به‌عنوان متغیرهای پیشگو یا ورودی در نظر گرفته می‌شوند.

### ۴. ۴. ۱ ایجاد طرح آزمون

برای سنجش اعتبار و صحت مدل، داده‌های موجود به بخش آموزش و آزمون تقسیم می‌شود. داده‌های بخش آموزش، الگوریتم را تولید می‌کنند و داده‌های بخش آزمون با کمک تعدادی رکورد، الگوریتم تولیدشده را آزمون و برچسب مربوط به رکوردهای مذکور را پیش‌بینی می‌کنند. بنابراین پیش از آغاز مدل‌سازی، مجموعهٔ داده‌ها به دو بخش داده‌های آزمون و

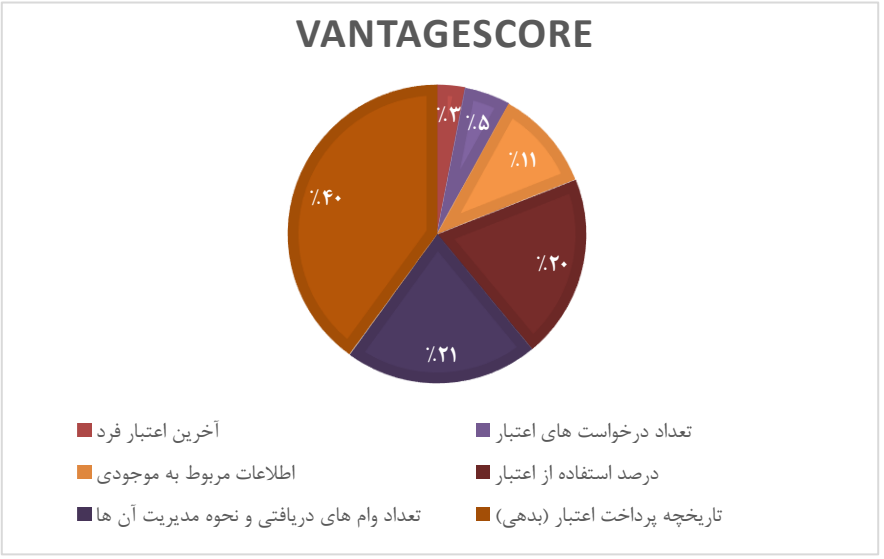


[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

آموزش تقسیم می‌شوند. بدین‌منظور با به‌کارگیری دستور train_test_split، ۷۵ درصد داده‌ها به‌صورت کاملاً تصادفی به‌عنوان داده‌های آموزش انتخاب می‌شوند و مدل‌سازی بر روی این داده‌ها انجام می‌گیرد و در نهایت پس از اجرای مدل، داده‌های باقی‌مانده به‌عنوان داده‌های آزمون برای ارزیابی مدل استفاده می‌شوند.

### ۴. ۴. ۲ نحوهٔ امتیازدهی امتیاز اعتباری

مدل‌هایی مانند vantagescore و Fico مدلی‌هایی‌اند که در بسیاری از مؤسسات مالی، از آن‌ها برای تعیین امتیاز اعتباری مشتریان استفاده می‌شود (وستون،۲۰۱۱). دراینجا، مدل vantagescore به‌دلیل اشتراک زیاد با داده‌های موجود در این پژوهش مورد بررسی قرار می‌گیرد و سپس با اعمال تغییراتی برای داده‌های موجود، قوانین امتیاز اعتباری معرفی می‌شود. در این مدل، اطلاعات مالی هر فرد مطابق با شکل ۲ دسته‌بندی می‌شود.




*شکل ۲*. درصد اهمیت متغیرها برای امتیاز اعتباری به روش vantagescore
منبع: وستون (۲۰۱۱)

عوامل تأثیرگذار در این امتیاز عبارت‌اند از:


[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

جدول ۳
*شرح عوامل تأثیرگذار در مدل امتیاز اعتباری vantagescore*

| شماره | عوامل | درصد اهمیت | توضیحات |
|---|---|---|---|
| ۱ | تاریخچهٔ پرداخت وام | ۴۰ | تاریخچهٔ وام‌های دریافتی هر مشتری حاوی اطلاعاتی چون دیرکرد اقساط مبلغ وام است. این اطلاعات کمک می‌کند تا بانک برای میزان و نحوهٔ وام جدید، تصمیم بهتری بتواند بگیرد. |
| ۲ | تعداد وام‌های دریافتی و نحوهٔ مدیریت آن‌ها | ۲۱ | در اینجا، مدت‌زمان وام‌ها مدنظر است. دریافت وام‌های مختلف (وام مسکن، وام خودرو، و...)، و مدیریت این وام‌ها، اعتبار را افزایش می‌دهد. |
| ۳ | درصد استفاده از اعتبار | ۲۰ | بالاترین امتیاز را از این قسمت افرادی دریافت می‌کنند که تا سقف ۳۰ درصد اعتبار خود را استفاده کنند. |
| ۴ | موجودی حساب مشتری | ۱۱ | اطلاعات مربوط به موجودی حساب‌های هر فرد |
| ۵ | تعداد درخواست‌های اعتبار | ۵ | اگر فردی از مؤسسات مختلف تقاضای اعتبار کند، اعتبار وی کاهش می‌یابد. |
| ۶ | اعتبار قبلی فرد | ۳ | اعتبار قبلی در نظرگرفته‌شده برای هر فرد |

منبع: وستون (۲۰۱۱)

به‌علت متفاوت‌بودن نظام بانکی ایران، با نظام‌های بانکی کشورهای پیشرفته‌تر که سال‌هاست از امتیاز اعتباری استفاده می‌کنند، تغییراتی در عوامل ذکرشده باید اعمال شود تا بتوان از آن‌ها در نظام بانکداری ایران استفاده کرد. در واقع برای محاسبهٔ امتیاز در این پژوهش، درصد اهمیت معیارهایی از امتیاز vantagescore را که در داده‌های موجود این پژوهش وجود ندارد، حذف می‌کنیم. برای تولید امتیاز اعتباری برای مشتریان بانک، چهار معیار برای داده‌ها تولید شد، که هریک از این معیارها مربوط به بخشی از امتیاز اعتباری است.


[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

جدول ۴

*شرح عوامل تأثیرگذار در مدل امتیاز اعتباری مورداستفاده در این پژوهش*

| معیار | امتیاز از ۱۰۰ | | متغیرها |
|---|---|---|---|
| ۱ | ۱۲ نمره | موجودی حساب مشتری | میانگین موجودی حساب<br>میانگین تغییرات موجودی هر ماه حساب<br>نسبت میانگین موجودی حساب‌های مشتری به جمع مبلغ وام<br>بیشترین موجودی حساب در ۲ سال گذشته |
| ۲ | ۲۱ نمره | نسبت میزان اقساط به موجودی | نسبت میانگین موجودی حساب در هر ماه به ۳۰٪ مبلغ اقساط (به‌ازای هر ماه)<br>نسبت میانگین تغییرات موجودی حساب در هر ماه به ۳۰٪ مبلغ اقساط (به‌ازای هر ماه) |
| ۳ | ۲۳ نمره | شامل ویژگی‌های مربوط به تعداد وام‌های دریافتی برای هر مشتری | تعداد وام‌های دریافتی<br>تغییرات مبلغ اقساط<br>تغییرات مبلغ وام<br>میانگین دورهٔ زمانی وام برای مشتری |
| ۴ | ۴۴ نمره | شامل ویژگی‌های مربوط به وام‌های دریافتی هر مشتری | جمع مبالغ اقساط<br>بیشترین مبلغ وام دریافتی<br>معوق وام<br>مبلغ غیرجاری (بدهی به بانک)<br>میانگین موجودی حساب |

### ۴. ۴. ۳ استفاده از روش نیمه‌نظارتی

در این پژوهش از روش نیمه‌نظارتی[۱] استفاده می‌شود؛ به این معنی که برای تعدادی از داده‌ها برچسب مناسب توسط خبره در نظر می‌گیریم و سپس از روش‌های با نظارت استفاده کرده و مدل‌های مناسب را روی داده‌های برچسب‌دار اجرا می‌کنیم.

[۱] یادگیری نیمه‌نظارتی به‌عنوان یک رویکرد مناسب در زمان کمبود داده‌های برچسب‌دار و استفاده از داده‌های بدون برچسب، در فاز آموزش مطرح شده است و با بهره‌گیری از مقادیر زیاد داده‌های بدون برچسب در کنار داده‌های برچسب‌دار، در جهت ساختن طبقه‌بندی‌های بهتر، روش‌هایی را برای غلبه بر این مشکل نشان می‌دهد. یادگیری نیمه‌نظارتی یک روش یادگیری بین یادگیری با ناظر و یادگیری بدون ناظر است. چهارچوب یادگیری نیمه‌نظارتی در هر دو رویکرد طبقه‌بندی و خوشه‌بندی قابلِ‌اعمال است. در مسائل طبقه‌بندی نیمه‌نظارتی، آموزش دادن یک طبقه‌بندی، با استفاده از هر دو مجموعه دادهٔ برچسب‌دار و بدون‌برچسب است، به‌طوری که کارایی آن از طبقه‌بندی با ناظری که فقط با داده‌های برچسب‌دار آموزش دیده باشد، بهتر باشد.



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

به‌علت هزینه‌بربودن و وقت‌گیربودن تعیین برچسب توسط خبره، تعداد داده‌هایی که برای برچسب‌گذاری انتخاب می‌شوند ۱۲۰ نمونه است. بنابراین پس از برچسب‌گذاری این داده‌ها، با ایجاد طرح آزمون، روش‌های مذکور برای این داده‌ها اجرا شده، و بهترین مدل با کمترین خطا برای برچسب‌گذاری داده‌های بدون برچسب انتخاب می‌شود. در استفاده از روش نیمه‌نظارتی، به‌علت محدودبودن داده‌هایی که برای آن‌ها برچسب تهیه می‌شود، روش انتخاب این نمونه‌ها حائز اهمیت است. این نمونه‌ها باید در بهترین حالت، تمام جامعهٔ داده‌ها را پوشش دهد.

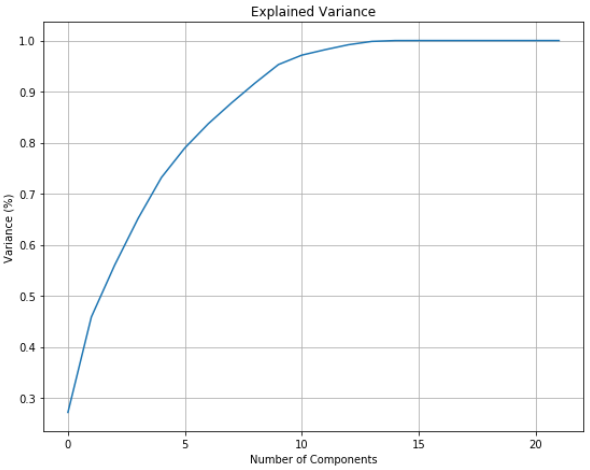
در این پژوهش، ابتدا از روش PCA برای تحلیل مؤلفه‌های اصلی داده‌ها استفاده می‌شود. پس از کاهش ابعاد داده‌ها، بهترین تعداد خوشه برای داده‌های موجود مشخص و سپس نمونه‌ها خوشه‌بندی می‌شوند. پس از آن از هر خوشه به‌طور تصادفی ۱۱ نمونه انتخاب و توسط خبره امتیازدهی می‌شوند. در مرحلهٔ آخر، سه روش با ناظر برای تعداد ۱۲۱ نمونه دارای برچسب اجرا شده و روشی که دارای بالاترین دقت باشد، برای پیش‌بینی نمونه‌های بدون برچسب انتخاب می‌شود.

### ۴. ۴. ۴ تحلیل مؤلفه‌های اصلی

از روش PCA برای کاهش ابعاد نمونه‌ها استفاده می‌شود. برای اجرای PCA، لازم است تمام متغیرهای موجود ابتدا به شکل استاندارد تبدیل شوند. در نرم‌افزار پایتون با دستور StandardScaler همهٔ متغیرها را استاندارد کرده و سپس به تحلیل مؤلفه‌ها پرداخته می‌شود. ابتدا، نموداری کشیده می‌شود تا با کمک آن، بهترین تعداد مؤلفهٔ اصلی برای توصیف داده‌ها انتخاب شود. همان‌طور که در شکل ۳ مشخص است، در محور افقی این نمودار، تعداد مؤلفه‌های اصلی و در محور عمودی مجموع واریانس به‌دست‌آمده توسط این مؤلفه‌ها قرار می‌گیرد. با توجه به شکل ۳، با انتخاب ۱۰ مؤلفهٔ اصلی، در حدود ۹۶ درصد از واریانس داده‌ها حفظ می‌شود. بنابراین، ۱۰ مؤلفهٔ اصلی اول برای کاهش داده‌ها انتخاب می‌کنیم. تأثیرگذارترین متغیرها برای ایجاد مؤلفه‌های اصلی عبارت‌اند از: موجودی به ۳۰٪ اقساط، جمع مبلغ اقساط، چک برگشتی، میانگین موجودی حساب، و جمع مبلغ وام.



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]




*شکل ۳*. نمودار واریانس حفظ‌شدهٔ داده‌ها برحسب تعداد مؤلفه‌های اصلی

### ۵ .۴ .۴ خوشه‌بندی

هدف از این گام یافتن بهترین تعداد خوشه برای متغیرهاست تا در هنگام نمونه‌گیری، از هر خوشه به‌طور تصادفی داده انتخاب شود. در واقع، این روش کمک می‌کند تا نمونه‌گیری بهتری برای برچسب‌گذاری داشته باشیم. با خوشه‌بندی، داده‌ها با ویژگی‌های نزدیک‌تر به هم، در یک خوشه قرار می‌گیرند و با نمونه‌گیری تصادفی ۱۲۰ داده از هر خوشه، مجموعهٔ داده‌ها به‌عنوان نمونه به‌دست می‌آید.

ابتدا بررسی تعداد مناسب خوشه برای خوشه‌بندی انجام می‌شود. در استفاده از الگوریتم k-means برای یافتن مناسب‌ترین k به‌عنوان تعداد خوشه‌ها، از روشی به نام روش آرنج[1] استفاده می‌شود. در این روش، نموداری کشیده می‌شود که در محور افقی آن تعداد خوشه‌های k و در محور عمودی آن مقدار هزینه براثر خوشه‌بندی داده‌ها به k خوشه قرار می‌گیرد.[2] این نمودار به‌شکل بازوست که بهترین مقدار k برای خوشه‌بندی قسمت آرنج نمودار بازو شکل است.

---

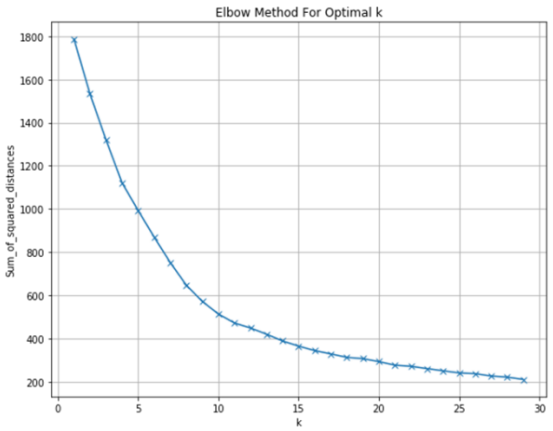
[1] Elbow

یکی از روش‌های پرکاربرد در زمینهٔ تعیین مقدار k، رسم نمودار آرنج است. در این نمودار، خطای به‌دست‌آمده به‌ازای مقدارهای متفاوت k در نموداری رسم می‌شود. این نمودار به‌شکل آرنج است که عدد k در قسمت بازوی نمودار، بهترین مقدار k می‌باشد (James, et al. , 2013).

[2] تابع هزینه برابر با $cost = \sum_{k=1}^{K}\sum_{x_n \in C_k}\|x_n - \mu_k\|^2$ است.



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]




*شکل ۴*. نمودار هزینه براساس تعداد خوشه‌های k

مطابق شکل ۴، k=11 بهترین تعداد خوشه برای داده‌های موردنظر است. بنابراین با استفاده از روش k-means و با پیش‌فرض k=11 در هنگام اجرای روش در نرم‌افزار، داده‌های حاضر را خوشه‌بندی می‌کنیم.

از هر خوشه با دستور sample در نرم‌افزار پایتون، ۱۱ نمونهٔ تصادفی به‌دست آورده و با قراردادن این نمونه‌ها در یک مجموعه، مجموعاً ۱۲۱ نمونه برای برچسب‌گذاری به فرد خبره ارائه می‌شود. همچنین، معیارهای به‌دست‌آمده در جدول ۴ را نیز در اختیار خبره قرار داده تا او با درنظرگرفتن این معیارها، به این نمونه‌ها امتیاز اعتباری دهد. امتیاز اعتباری به‌دست‌آمده توسط خبره عددی بین ۰ تا ۱۰۰ است. در هنگام امتیازدهی اعتباری، فرد خبره با توجه به محدودهٔ امتیاز هر معیار، و دانش خود، به ۱۲۱ نمونهٔ منتخب امتیاز می‌دهد.

### ۴. ۴. ۶ نزدیک‌ترین همسایگی

روش نزدیک‌ترین همسایگی بر روی داده‌های آموزش اجرا، و برای دستیابی به مدلی با بالاترین دقت، پارامترهای مدل تنظیم شد. پس از چندین‌بار اجرای مدل نزدیک‌ترین همسایگی، بهترین مدل برگزیده شد. در این مدل، تعداد نقاط همسایه برای هر داده $k = 4$ و فاصلهٔ اقلیدسی، برای پیش‌بینی امتیاز اعتباری مشتریان در نظر گرفته شد. در نهایت با تغییر عدد تصادفی تقسیم‌بندی داده‌ها، مدل بارها اجرا و مدل با بالاترین دقت[1] انتخاب شد. مدل نزدیک‌ترین همسایگی برای داده‌های آموزش، دقت ۰/۹۰۳ و برای داده‌های آزمون،

[1] $R^2 = (1 - (\sum(y_i - \hat{y}_i)^2)/(\sum(y_i - y_{mean})^2))$


[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

دقت ۰/۷۶۷ دارد. بنابراین، می‌توان نتیجه گرفت این مدل با دقت نسبتاً بالایی روی داده‌های بدون برچسب، پیش‌بینی امتیاز اعتباری را انجام خواهد داد.

### ۴. ۴. ۷ درخت تصمیم CART

روش درخت تصمیم CART برای داده‌های آموزش اجرا، و برای دستیابی به مدلی با بالاترین دقت، پارامترهای مدل تنظیم شد. با تغییر عدد تصادفی تقسیم‌بندی داده‌ها، مدل بارها اجرا و مدل با بالاترین حساسیت انتخاب شد. در این مدل، حداقل تعداد رکوردها برای هر زیرشاخه از درخت، ۴ رکورد تعیین شد.

مجموعهٔ قوانین حاصل از اجرای درخت CART برای داده‌های آموزش و آزمون بررسی شد. مدل درخت تصمیم CART برای داده‌های آموزش، دقت ۰/۹۸۷ و برای داده‌های آزمون، دقت ۰/۴۹۳ را دارد. بنابراین، می‌توان نتیجه گرفت این مدل دارای دقت نسبتاً بالایی برای داده‌های آموزش و دارای دقت ضعیفی برای پیش‌بینی داده‌های آزمون است. بنابراین، مدل درخت تصمیم CART مدل مناسبی برای پیش‌بینی داده‌های حاضر نیست.

### ۴. ۴. ۸ جنگل‌های تصادفی

جنگل تصادفی رگرسیونی برای داده‌های آموزش اجرا گردید و برای دستیابی به مدلی با بالاترین حساسیت و دقت، پارامترهای مدل تنظیم شد. مدل جنگل تصادفی با تشکیل ۱۰۰ درخت برای پیش‌بینی امتیاز اعتباری مشتریان در نظر گرفته شد. در نهایت با تغییر عدد تصادفی تقسیم‌بندی داده‌ها، مدل بارها اجرا و مدل با بالاترین دقت انتخاب شد. مدل جنگل‌های تصادفی برای داده‌های آموزش، دقت ۰/۷۵۹ و برای داده‌های آزمون، دقت ۰/۶۹۶ را داشت. بنابراین، می‌توان نتیجه گرفت این مدل دارای دقت نسبتاً بالایی برای داده‌های آموزش و دارای دقت ضعیفی برای پیش‌بینی داده‌های آزمون است. بنابراین، مدل جنگل‌های تصادفی مدل مناسبی برای پیش‌بینی داده‌های حاضر نیست.

## ۴. ۵ ارزیابی مدل‌های پژوهش

پس از مدل‌سازی، به ارزیابی نتایج حاصل از مدل‌سازی پرداخته شد. نتایج حاصل از ارزیابی باعث بهبود مدل می‌شود و مدل را قابل‌استفاده می‌کند. در جدول ۵، بهترین تابع توزیع برای متغیرهای مورداستفاده و P-value آن‌ها ارائه شده است.

در جدول ۶، نتایج ارزیابی الگوریتم‌ها، نزدیک‌ترین همسایگی، درخت تصمیم، و جنگل تصادفی ارائه شده است. همان‌طور که مشاهده می‌شود، روش نزدیک‌ترین همسایگی با دقت ۰/۷۶۷ برای مجموعهٔ آزمون، بالاترین دقت را دارد. بنابراین، جهت پیش‌بینی امتیاز برای داده‌های بدون برچسب از این روش استفاده می‌شود.



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

جدول ۵

*تعیین تابع توزیع متغیرهای مسئله*

| متغیر | تابع توزیع مناسب | P-value |
|---|---|---|
| میانگین موجودی حساب | lognorm | ۰/۳۹۲ |
| نسبت موجودی به وام | genextreme | ۰/۷۷۲ |
| نرخ تغییرات موجودی در ماه | گاما | ۰/۰۰۰ |
| نسبت موجودی به اقساط | lognorm | ۰/۷۳۸۴ |
| نسبت تغییرات موجودی به اقساط | exponweib | ۰/۰۰۰ |
| تغییرات اندازهٔ وام | نرمال | ۰/۰۰۰ |
| تغییرات اندازهٔ اقساط | نرمال | ۰/۰۰۰ |
| میانگین دورهٔ زمانی وام | genextreme | ۰/۰۰۰۹ |
| تعداد وام‌های دریافتی | نرمال | ۰/۰۰۰ |
| میانگین موجودی حساب | exponweib | ۰/۵۹۸ |
| جمع اقساط در هر سال | exponweib | ۰/۰۷۹ |
| بیش‌ترین اندازهٔ وام دریافتی | lognorm | ۰/۰۰۱ |
| جمع بدهی‌ها | نرمال | ۰/۱۳۸ |
| میانگین معوق ماه | exponweib | ۰/۰۰۰ |

جدول ۶

*مقایسهٔ نتایج حاصل از الگوریتم‌های داده‌کاوی برحسب درصد*

| روش مورداستفاده | دقت برای داده‌های آموزش | دقت برای داده‌های تست |
|---|---|---|
| نزدیک‌ترین همسایگی | ۰/۹۰۳ | ۰/۷۶۷ |
| درخت تصمیم CART | ۰/۹۸۷ | ۰/۴۹۳ |
| جنگل تصادفی | ۰/۷۵۹ | ۰/۶۹۶ |

### ۴. ۵. ۱ قوانین انجمنی

پس از امتیازدهی اعتباری به مشتریان بانک، با استفاده از قوانین انجمنی به بررسی ویژگی‌های مشتریان پرداخته شده است. با استفاده از دستور Apriori در نرم‌افزار پایتون و با تعیین پشتیبانی ۲۰ درصد و حداقل اطمینان ۵۰ درصد، مجموعهٔ داده‌ها مورد بررسی قرار گرفته و بهترین قوانین به‌دست‌آمده در جدول ۷ بیان شده است.



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

جدول ۷
*قوانین به‌دست‌آمده از الگوریتم Apriori*

| موارد | نتیجه (امتیاز اعتباری) | پشتیبانی | اطمینان |
|---|---|---|---|
| اگر تعداد ماه‌های معوق بین ۰ تا ۲ ماه و میانگین موجودی حساب بین ۵ میلیارد ریال تا ۱۰ میلیارد ریال و سال دریافت وام ۱۳۹۲ باشد، | آنگاه امتیاز اعتباری بین ۸۰ تا ۱۰۰ است. | ۰/۳۲۲ | ۰/۹۶۷ |
| اگر تعداد ماه‌های معوق بین ۵ تا ۱۰ ماه و مبلغ وام بین ۱ میلیارد ریال تا ۵ میلیارد ریال و نرخ سود وام بیش‌تر از ۲۰ درصد باشد، | آنگاه امتیاز اعتباری بین ۲۰ تا ۴۰ است. | ۰/۲۸۳ | ۰/۸۸۷ |
| اگر مانده وام بین ۱۰۰ میلیون تا ۵۰۰ میلیون ریال و مبلغ وام بین ۱ میلیارد ریال تا ۵ میلیارد ریال باشد، | آنگاه امتیاز اعتباری بین ۴۰ تا ۶۰ است. | ۰/۱۹۷ | ۰/۹۲۹ |

طبق نتایج به‌دست‌آمده، ازجملهٔ معیارهای مهم و تأثیرگذار می‌توان به نوع وام، تعداد ماه‌های معوق، میانگین موجودی حساب، مانده وام، مبلغ وام، و نرخ سود وام نام برد که با ویژگی‌های مهم ارائه‌شده برای تعیین امتیاز اعتباری در این پژوهش مطابقت دارد.

### ۶.۴ توسعه

در این فاز، نتایج حاصل از پژوهش در بانک یا مؤسسهٔ مالی دیگری به‌کار گرفته می‌شوند. بدین‌منظور باید یک طرح اجرایی آماده شود و سازوکارهای لازم برای پایش فرایندها در جهت اجرای نتایج مدل ایجاد شود. براساس نتایج به‌دست‌آمده، می‌توان یک طرح توسعهٔ اولیه برای مدل پیشنهاد کرد؛ به این‌صورت که تمام اطلاعات مشتریان از زمانی که در بانک حساب بازکرده‌اند تا اطلاعات مربوط به وام یا تغییرات موجودی حساب آن‌ها در پایگاه داده ثبت شود تا بتوان با مدل به‌دست‌آمده، امتیاز اعتباری مشتری را تهیه کرد. از قابلیت‌های بالای نرم‌افزار پایتون برای این امر می‌توان استفاده کرد. از میان الگوریتم‌های مورداستفاده در این مطالعه، الگوریتم نزدیک‌ترین همسایگی بالاترین دقت را داشت و به‌عنوان الگوریتم برتر انتخاب شد که می‌تواند به‌عنوان مدل مرجع اصلی تصمیم‌گیری استفاده شود.

## ۵ بحث و نتیجه‌گیری

یکی از موضوعات چالش‌برانگیز در بانک‌ها و مؤسسات مالی، تعیین امتیاز اعتباری صحیح برای مشتریان است، زیرا واگذاری تسهیلات به مشتریان یکی از مهم‌ترین منابع درآمد برای


[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

بانک‌ها و مؤسسات مالی محسوب می‌شود. امروزه، روش امتیاز اعتباری برای تصمیم‌گیری در مورد نحوهٔ تعیین وام به‌عنوان یک ابزار تصمیم‌گیری کمکی یا تصمیم‌گیری خودکار، برای طیف گسترده‌ای از مشتریان مورداستفاده قرار می‌گیرد. این روش برای بررسی میزان ریسک وام‌گیرنده، به هر مشتری یک امتیاز منحصربه‌فرد نسبت می‌دهد، که آن سطح ریسک وام‌گیرنده را با توجه به اطلاعات فرد مشخص می‌کند (عباس زاده و امیری ۱۳۹۵).

در این مقاله مطابق با روش CRISP-DM، ابتدا به شناخت سیستم و مجموعهٔ دادهٔ مورداستفاده در این پژوهش پرداخته شد؛ در گام بعدی پس از آماده‌سازی داده‌ها و به‌منظور فراهم‌آوردن امکان ارائهٔ قوانین و تفاسیر قابل‌درک، با به‌کارگیری الگوریتم‌های استفاده‌شده در این پژوهش، به شناسایی عواملی که بیشترین تأثیر را در پیش‌بینی امتیاز اعتباری دارند، پرداخته شد و سپس از روش‌های رگرسیونی استفاده کرده و با به‌کارگیری الگوریتم‌های نزدیک‌ترین همسایگی، درخت تصمیم، و جنگل تصادفی امتیاز اعتباری برای مشتریان بانک مشخص شد. در ادامه، مدل‌های استفاده‌شده توسط داده‌های آزمون ارزیابی شد و برترین مدل که مدل نزدیک‌ترین همسایگی بود، برای پیش‌بینی امتیاز اعتباری برگزیده شد.

در این مطالعه، روش نزدیک‌ترین همسایگی بالاترین دقت را برای داده‌های آموزش و آزمون داشت؛ لی نیز در پژوهش خود به همین نتیجه رسید؛ وی با ایجاد تغییراتی در مدل نزدیک‌ترین همسایگی، دقت بالایی برای پیش‌بینی امتیاز اعتباری در داده‌های خود به‌دست آورد (لی، ۲۰۰۹). روش درخت تصمیم CART دارای دقت بالایی در پیش‌بینی برای داده‌های آموزش بود، ولی برای داده‌های آزمون دقت آن بسیار پایین بوده ولی در جنگل‌های تصادفی، دقت برای هر دو این داده‌ها افزایش یافته است که این مسئله با نتایج ارائه‌شده در پژوهش ونگ و همکاران هم‌خوانی دارد. در بررسی ونگ و همکاران، مشاهده شد درخت تصمیم نسبت به روش‌های دیگر دارای دقت پایین‌تر بوده و استفاده از روش‌های ترکیبی مانند جنگل تصادفی اندازهٔ دقت را افزایش می‌دهد (ونگ و همکاران، ۲۰۱۲).

مطابق نتایج به‌دست‌آمده، بهترین نتایج از روش نزدیک‌ترین همسایگی با دقت ۹۰/۳ درصد برای مجموعهٔ آموزش و ۷۶/۷ درصد برای داده‌های آزمون جهت پیش‌بینی امتیاز اعتباری به‌دست آمد. بنابراین با توجه به این امتیاز، بانک می‌تواند برای دریافت تسهیلات به مشتریان تصمیم بهتری بگیرد تا از ضرر خود و مشتریان جلوگیری کند. در صورت استفادهٔ بانک‌ها از امتیاز اعتباری و بهبود این امتیاز با توجه به رفتارهای مشتری، این امتیاز می‌تواند به عاملی مهم برای تصمیم‌گیری نحوه و میزان وام به مشتریان تبدیل شود. به‌کارگیری سایر الگوریتم‌های پیش‌بینی‌کننده و مقایسهٔ دقت هریک از این روش‌ها با یکدیگر و مشخص‌ساختن عوامل مؤثر در این امتیاز می‌تواند زمینه‌های دیگری از پژوهش‌ها را جهت پیش‌گویی امتیاز



[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

اعتباری برای مشتریان بانک فراهم کند. در این پژوهش، از مدل امتیاز اعتباری VantageScore برای تعیین امتیاز اعتباری به مشتریان استفاده شد که در این مدل، از متغیرهای کیفی نامبرده‌شده استفاده نشده است. پیشنهاد می‌شود رابطهٔ (یا عدم رابطه) متغیرهای کیفی مانند جنسیت، سن، وضعیت تأهل و غیره در امتیاز اعتباری بررسی شود. همچنین، پیشنهاد می‌شود متغیر ارقام کنترلی در نظر گرفته شود و برای سنجش میزان اعتبار، موضوعی که برای آن درخواست وام ارائه شده است و فعالیتی که وام‌گیرنده می‌خواهد با استفاده از آن وام انجام دهد در نظر گرفته شود؛ به عبارت دیگر، بررسی شود که ریسک فعالیت در میزان وامی که بانک پرداخت می‌کند تا چه اندازه می‌تواند تأثیرگذار باشد.

## فهرست منابع


آذر، ع.، و خدیور، آ. (۱۳۹۳). *کاربرد تحلیل آماری چندمتغیره در مدیریت*. تهران: نگاه دانش.

آسایش، ح.، و جلالی، م. (۱۳۹۷). ارزیابی کارایی ماشین بردار پشتیبان در امتیازدهی اعتباری مشتریان حقیقی بانک تجارت. *سومین کنفرانس ملی سالانهٔ اقتصاد، مدیریت حسابداری*. اهواز.

هان، ژ.، پی، ژ.، کامبر، م. (۱۳۹۳). داده کاوی (مفاهیم و تکنیک ها). ترجمه اسماعیلی، م. تهران: نیاز دانش. (۱۳۹۰)

جلیلی، ح. (۱۳۸۹). سامانهٔ اعتبارسنجی مشتریان بانکی و بیمه‌ای – مطالعهٔ موردی: تجربهٔ شرکت مشاورهٔ رتبه‌بندی اعتباری ایران. *بیستمین کنفرانس سیاست‌های پولی و ارزی*. تهران: پژوهشکدهٔ پولی و بانکی.

عباس‌زاده، ح.، و امیری، ع. (۱۳۹۵). امتیاز اعتباری مشتریان با استفاده از الگوریتم RUBoost توسعه‌یافته و توصیف‌گر داده مبتنی بر بردار پشتیبان با گام برداری تصادفی. *سومین کنفرانس بین‌المللی مهندسی کامپیوتر و سیستم‌های خبره*. تربت حیدریه.

غضنفری، م.، علیزاده، س.، و تیمورپور، ب. (۱۳۸۷). *داده‌کاوی و کشف دانش*. تهران: دانشگاه علم و صنعت ایران.

Abdou, H. A. (2009). Genetic programming for credit scoring: the case of Egyptian public sector banks. *Expert systems with application*, 36(9), 11402-

Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping Multidimensional Data*, by J Kogan, C Nicholas and M Teboulle, 25-71. Berlin, Heidelberg: Springer.

Cai, W., Lin, X., Huang, X., & Zhong, H. (2009, December). A genetic algorithm model for personal credit scoring. In *2009 International Conference on Computational Intelligence and Software Engineering* (pp. 1-4). IEEE.




[ DOR: 20.1001.1.26453355.1400.14.48.3.9 ]

Capozziello, S., Lazkoz, R., & Salzano, V. (2011). Comprehensive cosmographic analysis by Markov chain method. *Physical Review D*, *84*(12), 124061.

Castillo, E., Conejo, A. J., Pedregal, P., Garcia, R., & Alguacil, N. (2011). *Building and solving mathematical programming models in engineering and science* (Vol. 62). John Wiley & Sons.

Chen, M. C., & Huang, S. H. (2003). Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications* 24 (4): 433-441.

Dongare, A. D., Kharde, R. R., & Kachare, A. D. (2012). Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology (IJEIT)*, *2*(1), 189-194.

Fritz, S., & Hosemann, D. (2000). Restructuring the credit process: Behaviour scoring for German corporates. *Intelligent Systems in Accounting, Finance & Management* 9.1: 9-21.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, *16*(10), 906-914.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1-3), 389-422.

Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, *33*(4), 847-856.

Huang, J. J., Tzeng, G. H., & Ong, C. S. (2006). Two-stage genetic programming (2SGP) for the credit scoring model. *Applied Mathematics and Computation*, *174*(2), 1039-1053.

Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision support systems*, *37*(4), 543-558.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Li, F. C. (2009, August). The hybrid credit scoring strategies based on knn classifier. In *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery* (Vol. 1, pp. 330-334). IEEE.

Liao, J. G., & Chin, K. V. (2007). Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *Bioinformatics*, *23*(15), 1945-1951.

Liao, S. H. (2005). Expert system methodologies and applications—a decade review from 1995 to 2004. *Expert systems with applications*, *28*(1), 93-103.

Liaw, A., & Wiener, M. (2002). Classificatio and regression by RandomForest. *R news* 2.3: 18-22.

Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert systems with applications*, *36*(2), 3028-3033.

Pearson, K. (1901). On line and planes of closet fit to systems of points in space. *Philosophical Magazine*, 2(11), 559-572.

Reeves, C., & Rowe, J. E. (2002). *Genetic algorithms: principles and perspectives: a guide to GA theory* (Vol. 20). Springer Science & Business Media.

Sahigara, F., Ballabio, D., Todeschini, R., & Consonni, V. (2013). Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. *Journal of cheminformatics*, *5*(1), 1-9.

Sarwar, M. T., Anastasopoulos, P. C., Golshani, N., & Hulme, K. F. (2017). Grouped random parameters bivariate probit analysis of perceived and observed aggressive driving behavior: a driving simulation study. *Analytic methods in accident research*, *13*, 52-64.

Sayjadah, Y., Hashem, I. A. T., Alotaibi, F., & Kasmiran, K. A. (2018, October). Credit Card Default Prediction using Machine Learning Techniques. In *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)* (pp. 1-4). IEEE.

Shevade, S. K., & Keerthi, S. S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, *19*(17), 2246-2253.

Thomas, L. C., Jonathan N. C., & Edelman, D. B. (2002). *Credit scoring and its applications.* Philadelphia, Pa: Society for Industrial and Applied Mathematics.

Timofeev, R. (2004). Classification and regression trees (CART) theory and applications. *Humboldt University, Berlin*, 1-40.

Wang, G, J Ma, L Huang, & K Xu. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems* 26: 61-68.

Wang, G., Ma, J., Huang, L., & Xu, K. (2011). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems*, *26*, 61-68.

Weston, L. P. (2011). *Your Credit Score: How to improve the 3-digit Number that Shapes Your Financial Future.* FT Press.

Yap, B. W., Ong, S. H., & Husain, N. H. M. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications*, *38*(10), 13274-13283.

Yobas, M. B., Crook, J. N., & Ross, P. (2000). Credit scoring using neural and evolutionary techniques. *IMA Journal of Management Mathematics*, *11*(2), 111-125.

Zhang, D., Huang, H., Chen, Q., & Jiang, Y. (2007, August). A comparison study of credit scoring models. In *Third International Conference on Natural Computation (ICNC 2007)* (Vol. 1, pp. 15-18). IEEE.