Statistical Sciences Extension
Year 1 - Number 2 (P 7-14)

گسترهٔ علوم آماری
سال اول - شماره دوم (ص 7-14)

مقایسهٔ کارایی روش‌های رده‌بندی‌کننده رگرسیون لجستیک و رگرسیون درختی برای متغیر وابسته باینری

علیرضا پاک‌گهر *
مربی، آمار، دانشگاه پیام نور

تاریخ دریافت: 1395/02/04 **تاریخ پذیرش**: 1395/06/27

Performance Comparison of Logistic Regression and Classification Regression tree Models for Binary Dependent Variable

A. Pakgohar*
1. Lecturer, Statistics, Payame Noor University

Received: 2016/04/23 **Accepted:** 2016/09/17

چکیده

در این مقاله میزان کارایی مدل‌های رده‌بندی رگرسیون لجستیک باینری و رگرسیون درختی روی متغیر وابسته باینری بررسی می‌شود. شیوهٔ پردازش مدل، استفاده از تمام داده‌ها در مرحلهٔ آموزشی است. مجموعه داده‌های مورد مطالعه از یک گزارش مطالعاتی دربارهٔ سوابق بیماری زردی به دست آمده است که یک مجموعه داده شامل متغیرهای کمی و کیفی است. میزان کارایی دو روش طبقه‌بندی‌کننده رگرسیون لجستیک و رگرسیون رده‌بندی درخت تصمیم، بر اساس معیارهای کارایی آماری نظیر دقت، توجه به موارد خاص، و تحلیل حساسیت است. نتایج تجربی ما نشان می‌دهد که رگرسیون لجستیک، دقت بالای 83% و رگرسیون درختی میزان دقت حدود 73% را بر روی مجموعه نشان داده‌اند. به همین ترتیب میزان حساسیت رگرسیون لجستیک باینری برابر 77% و رگرسیون درختی برابر 66% است. همچنین اندازه توجه به موارد خاص مدل رگرسیون برابر 85% وبرای رگرسیون درختی برابر 76% است. نتایج کارایی مدل نشان می‌دهد رگرسیون لجستیک باینری بهتر از رگرسیون درختی عمل کرده است.

واژگان کلیدی

داده کاوی، رگرسیون لجستیک، رگرسیون درختی، دقت، حساسیت و مشخصه بودن.

Abstract

This paper describes the performance analysis of two classifier models common in statistics and data mining on binary dependent variable, binary Logistic Regression (B.LR) and Classification Regression Tree (CART). The evaluation method is using all data in training stage. The using data set is from ‘‘Evaluation of patients with Jaundice on children’’ report. Data set is collection of categorical and continues independent variables. The classification performance of two classifiers is presented by using statistical performance measures like accuracy, specificity and sensitivity. Experimental result showed that accuracy of LR is more than 83% and CLASSIFICATION AND REGRESSION TREE is nearly 73%. So the sensitivity measure for BINARY LOGISTIC REGRESSION is nearby 77% and 66% for CLASSIFICATION AND REGRESSION TREE as well the specificity scale is 85% for BINARY LOGISTIC REGRESSION and 76% for CLASSIFICATION AND REGRESSION TREE. The result shows the performance of BINARY LOGISTIC REGRESSION classifier is found to be better than CLASSIFICATION AND REGRESSION TREE.

Keywords

Data Mining, Binary Logistic Regression; Classification Regression tree; Accuracy; Sensitivity; Specificity.

* **نویسنده مسئول**: علیرضا پاک‌گهر

*Corresponding Author: a_pakgohar@pnu.ac.ir

مقدمه

معمولاً فرایند رده‌بندی یا طبقه‌بندی، یک فرایند دو مرحله‌ای است که تحت عنوان مرحلهٔ آموزشی و مرحلهٔ آزمایشی شناخته می‌شوند [1]؛ البته می‌توان تنها یک مرحله را برای رده‌بندی داده کاوی به کار برد. زمانی که تعداد داده‌ها محدود باشد، پیشنهاد محققان استفاده از یک مرحله‌ای بودن رده‌بندی داده‌کاوی است [2]. ما در این مطالعه این الگو را به کار برده‌ایم. در هر حال در مرحلهٔ نخست یا همان مرحله آموزش، ابتدا طبقه‌بندی کننده‌ای ساخته می‌شود که یک مجموعه معین از کلاس‌های داده یا مفاهیم را توضیح می‌دهد. به همین دلیل این مرحله را همان مرحلهٔ یادگیری یا فاز آموزش می‌نامیم که طی آن، طبقه‌بندی کننده‌ها به وسیلهٔ یک الگوریتم طبقه‌بندی و از طریق آنالیز یا "یادگیری از" یک مجموعه آموزشی متشکل از چندتایی‌های موجود در مجموعه داده‌ها و سطوح کلاس‌های اختصاص داده شده به آنها، ساخته می‌شوند. ما یک چندتایی X، را با یک بردار صفت n بعدی به صورت $X=(x_1 , x_2 \dots x_n)$ نشان می‌دهیم که n نمایندهٔ تعداد آزمودنی‌ها است بر این اساس فرض کنید هر چندتایی X به یک کلاس از پیش تعیین شده تعلق دارد که به وسیلهٔ صفت‌های موجود در پایگاه داده دیگر تعیین شده‌اند و "صفت‌های برچسب کلاس" نامیده می‌شوند. در این صورت، مقدارهای گرفته شده به وسیلهٔ صفت سطح کلاس به صورت گسسته و نامرتب هستند. این متغیر از نوع متغیر دسته‌ای است زیرا هر یک از مقدارهای آن به عنوان یک دسته، رده (category) یا کلاس عمل می‌کند. ما به چندتایی‌هایی که مجموعهٔ آموزشی را می‌سازند، چندتایی‌های آموزشی می‌گوییم و از آنجا که در این مرحله برچسب‌های کلاس مربوط به هر چندتایی آموزشی از قبل به الگوریتم داده می‌شوند، این مرحله را با نام "یادگیری سرپرستی شده" می‌شناسیم که تمایز بین مدل رده‌بندی و مدل خوشه‌بندی در همین تعریف است به طوری‌که اگر یادگیری غیر سرپرستی شده داشته باشیم با مدل خوشه‌بندی سر و کار داریم. گام نخست از فرایند طبقه‌بندی را می‌توانیم به صورت کشف و یادگیری نگاشت یا تابع $y=f(x)$ در نظر بگیریم که می‌تواند سطح کلاس y مربوط به ازای چندتایی X داده شده را پیش‌بینی کند. این نگاشت می‌تواند به صورت قواعد طبقه‌بندی، درخت‌های تصمیم، یا فرمول ریاضی ارائه شود. ما این قواعد را می‌توانیم برای دسته‌بندی چندتایی‌ها در داده‌های آینده نیز به کار ببریم. در مرحلهٔ دوم از این مدل به دست آمده برای طبقه‌بندی استفاده می‌شود. یک مجموعهٔ آزمایشی برای آزمایش چندتایی‌ها و برچسب‌های کلاس مربوط به آنها استفاده می‌شود. این چندتایی‌ها به صورت تصادفی از میان کل مجموعه داده‌ها انتخاب می‌شوند. که البته ما به دلیل تعداد داده‌های محدود آن را انجام نداده‌ایم.

دقت یک طبقه‌بندی کننده بر روی یک مجموعه داده‌های آزمایشی معین برابر با درصدی از چندتایی‌های مجموعهٔ آزمایشی است که طبقه‌بندی کننده توانسته آنها را به درستی طبقه‌بندی کند (یعنی آنها را در کلاس صحیح قرار داده است). سطح کلاس مربوط به هر یک از چندتایی‌های آزمایشی با کلاس پیش‌بینی شده به وسیلهٔ طبقه‌بندی کننده آموزش دیده برای آن چندتایی مقایسه می‌شود. اگر میزان دقت طبقه‌بندی کننده قابل قبول باشد، از آن طبقه‌بندی کننده می‌توانیم برای طبقه‌بندی چندتایی‌های داده‌های آینده استفاده کنیم که برچسب کلاس آنها را از قبل نمی‌دانیم.

در این مقاله روش‌های طبقه‌بندی رگرسیون لجستیک و رگرسیون درختی بر روی داده‌های گزارش مطالعاتی بررسی بیماری‌های اسهال، پنومونی و زردی اطفال در شهرستان‌های اردکان و میبد [7] مورد بررسی قرار گرفته‌اند. هدف از مطالعه، مقایسه‌ای بین الگوریتم‌های طبقه‌بندی برای پیش‌بینی وضعیت ابتلا به بیماری زردی در بین کودکان تازه متولد شده بوده است. کارایی هر یک از مدل‌ها با به کار بردن معیارهای مختلف آماری مانند دقت طبقه‌بندی، میزان خاص بودن آن و حساسیت طبقه‌بندی ارزیابی شده‌اند. هر یک از نمونه‌های مجموعه داده‌ها به یکی از دو دسته شامل بیمار یا سالم طبقه‌بندی شده‌اند.

مجموعه داده‌ها

داده‌های مورد استفاده در این مقاله از گزارش مطالعاتی بررسی بیماری‌های اسهال، پنومونی و زردی اطفال در شهرستان‌های اردکان و میبد [7] گرفته شده است. ما در این مقاله صرفاً اطلاعات مربوط به بیماری زردی را به عنوان متغیر وابسته و همچنین تعدادی متغیر مستقل و زمینه‌ای را به عنوان متغیرهای پیشگو در نظر گرفته‌ایم. این

مجموعه داده‌ها شامل اطلاعات مربوط به سن، جنسیت، وضعیت سلامت کودک هنگام تولد (بر اساس کارت سلامت) سابقه بیماری زردی در کودک، والدین و خواهر و برادران کودک، نوع زایمان، وضعیت اقتصادی خانواده، میزان تحصیلات والدین، زمان ابتلا به بیماری (ماه، و فصل سال) و سایر متغیرهای عمومی است. هر یک از نمونه‌های کارت سلامت) سابقهٔ بیماری زردی در کودک، والدین و خواهر و برادران کودک، نوع زایمان، وضعیت اقتصادی خانواده، میزان تحصیلات والدین، زمان ابتلا به بیماری (ماه، و فصل سال) و سایر متغیرهای عمومی در پرسش‌نامه‌ای که از قبل به همین منظور تهیه شده بود، ثبت گردید. و با استفاده از نرم‌افزار SPSS 18 (PASW) که یکی از

جدول 1. اطلاعات مشخصات متغیرهای مجموعه داده‌های مطالعه بیماری زردی کودکان

متغیر	کلاس	گونه	سطرهای گمشده	تعداد دسته‌ها
ابتلا به بیماری زردی	هدف	اسمی	0	2
جنسیت کودک	پیش‌بینی کننده	اسمی	0	2
فصل تولد کودک	پیش‌بینی کننده	اسمی	0	4
ماه تولد کودک	پیش‌بینی کننده	اسمی	0	12
رشد جسمانی کودک	پیش‌بینی کننده	ترتیبی	0	3
سن مادر	پیش‌بینی کننده	کمی	0	-
سن پدر	پیش‌بینی کننده	کمی	0	-
تعداد فرزندان	پیش‌بینی کننده	کمی	0	-
سابقه بیماری زردی در مادر	پیش‌بینی کننده	اسمی	0	2
سابقه بیماری زردی در پدر	پیش‌بینی کننده	اسمی	0	2
سابقه بیماری زردی در فرزندان دیگر	پیش‌بینی کننده	اسمی	0	2

مجموعه داده‌ها می‌تواند در یکی از دو دسته: مبتلا به بیماری زردی یا سالم طبقه‌بندی شود. در این مجموعه داده‌ها تعداد 11 صفت وجود دارد که اولین صفت به عنوان متغیر وابسته یا هدف و بقیه آنها به عنوان متغیرهای پیشگو یا ورودی در نظر گرفته می‌شوند. جدول (1) اطلاعات توصیفی مجموعه داده‌های مطالعهٔ بیماری زردی کودکان را نشان می‌دهد.

روش کار

این مطالعه از نوع توصیفی و جامعهٔ آماری آن شامل کلیه نوزادان کمتر از یک ماه است که به یکی از مراکز بهداشتی و درمانی شهر مراجعه کرده‌اند. در این مطالعه از بین 100 کودک مورد مطالعه اطلاعات 85 کودک، دربارهٔ وضعیت بیماری زردی، مورد تأیید قرار گرفته و تحلیل گردید [3]. نمونه‌گیری از نوع سهمیه‌ای بوده است. اطلاعات مربوط به سن، جنسیت، وضعیت سلامت کودک هنگام تولد (بر اساس

نرم‌افزارهای آماری- داده‌کاوی برای مدل‌سازی است [4] تحلیل آماری انجام یافته است.

رگرسیون لجستیک

رگرسیون لجستیک، که با نام رگرسیون اسمی[1] نیز نامیده می‌شود، یک روش آماری برای طبقه‌بندی ثبت‌ها بر پایهٔ مقادیر فیلدهای ورودی است. این روش مشابه رگرسیون خطی است، اما به جای یک متغیر هدف عددی، یک متغیر کیفی (مانند متغیر اسمی) را می‌پذیرد. این روش می‌تواند هم با مدل‌های دوجمله‌ای (برای هدف‌هایی که دارای دو دسته جدا از هم باشند) و هم با مدل‌های چند جمله‌ای (برای هدف‌هایی که دارای بیش از دو دسته باشند) کار کند. رگرسیون لجستیک با ساختن یک مجموعه از معادله‌ها کار می‌کند که مقادیر متغیر ورودی را به احتمالات مربوط به هر یک از دسته‌های ممکن برای بر آن متغیر (فیلد خروجی)

1. Nominal Regression

مرتبط می‌سازند. هنگامی‌که مدل ایجاد شد، می‌توانیم آن را برای تخمین زدن احتمالات برای داده‌های جدید به کار ببریم. برای هر ثبت، یک احتمال عضویت برای هر یک از دسته‌های ممکن خروجی محاسبه می‌شود. آن دسته هدف که دارای بیشترین میزان احتمال باشد، به عنوان مقدار پیش‌بینی شده خروجی برای آن ثبت درنظر گرفته خواهد شد. مدل‌های رگرسیون لجستیک معمولاً بسیار دقیق هستند. این مدل‌ها می‌توانند با متغیرهای ورودی نمادی یا عددی کار کنند؛ و احتمالات پیش‌بینی شده را برای تمام دسته‌های متغیر هدف به دست دهند. مدل‌های لجستیک هنگامی بیشترین تأثیر را دارند که یک متغیر واقعاً دسته‌ای داشته باشیم که عضویت در گروه را نشان دهد [2و4و5].

روش رگرسیون در ختی

مدل رگرسیون درختی[1] (CART) یک مدل ناپارامتری و بدون هرگونه پیش فرض به منظور سنجش رابطهٔ بین متغیرهای مستقل و متغیر وابسته یا هدف به کار می‌رود و از روش‌های مهم داده کاوی است، به طور گسترده در تجارت، صنعت، مهندسی و سایر علوم استفاده شده است. مدل CART، ابزاری قدرتمند در تعیین مهم‌ترین متغیرهای مستقل و حل مسائل دسته‌بندی و پیش‌بینی است. در این مدل، تعدادی رکورد وجود دارد که دستهٔ آنها از قبل معلوم است (متغیر وابسته در آنها معلوم است). هدف، تهیه درختی است که بتوان به وسیله آن متغیر وابسته یا همان کلاس را برای یک رکورد جدید پیش‌بینی و تعیین کرد. روش CART شاخه‌های خود را به صورت دوتایی و تنها بر اساس یک فیلد (متغیر مستقل) ایجاد می‌کند. یعنی هر گروه غیر برگ آن، به دو گروه دیگر تفکیک می‌گردد. منظور ما از گروه غیر برگ، گره مدل درختی است که خود به دو بخش دیگر جدا شده است. اولین قدم، پاسخ به این سوال است که کدام یک از فیلدها بهترین شاخه را تولید می‌کند. بهترین ایجاد شاخه هنگامی رخ می‌دهد که شاخه‌های حاصل طوری باشند که در هر شاخه یک کلاس بر سایر کلاس‌ها غلبه کند. معیاری که برای ارزیابی شاخه‌ها به کار می‌رود، عبارت است از گوناگونی. برای محاسبه گوناگونی در یک مجموعه از رکوردها روش‌های بسیاری وجود دارد که در تمامی آنها گوناگونی زیاد عبارت است از مجموعه‌هایی که از کلاس‌های گوناگون در خود داشته باشند و گوناگونی کم عبارت است از مجموعه‌هایی که اعضای یک کلاس در آن بر سایر کلاس‌ها غلبه کند و بهترین نحوه ایجاد شاخه آن است که گوناگونی در مجموعه‌ها را تا حد امکان کم کند. در مرحله بعد دو شاخه وجود دارد که هرکدام دارای یک سری رکورد هستند (هر یک از رکوردهای گره بالاتر در یکی از شاخه‌ها قرار گرفته است). حال برای هر شاخه مثل قبل عمل می‌گردد. یعنی برای هر یک از آنها دوباره فیلد طوری انتخاب می‌شود که بتوان بهترین شاخه‌های جدید را با حداقل گوناگونی ایجاد نمود. این مرحله آنقدر ادامه می‌یابد تا در هر زیر شاخه گره‌ای تولید شود که ایجاد شاخه جدید در آن گره مقدار گوناگونی را کاهش قابل توجه‌ای ندهد. به این گره نهایی برگ گفته می‌شود. برای جداسازی هر گره به دو زیر گره، شاخص‌های مختلفی وجود دارد که معروف‌ترین آن برای داده‌های اسمی، شاخص جینی است که به شکل رابطه (1) تعریف می‌شود:

$$P(j|m) = \frac{P(j,m)}{P(m)} \,,\; P(j,m) = \frac{\pi(j)N_j(m)}{N_j} \,,\; P(m) = \sum_{j=1}^{J} P(j,m) \qquad (1)$$

$$Gini(m) = 1 - \sum_{j=1}^{J} P^2(j|m) \qquad (2)$$

که در آن J تعداد دسته‌ها یا همان متغیرهای هدف، p(j)، احتمال اولیه مربوط به دسته j و تبه وسیلهٔ تصمیم گیرنده مشخص می‌شود. Nj(m) تعداد مشاهدات مربوط به دسته j در گره m، Nj تعداد کل مشاهدات مربوط به کلاس j در گره ریشه، P(j|m)، احتمال قرارگیری مشاهدات مربوط به دسته j در گره m و Gini(m)، که همان شاخص جینی است، معرف عدم خلوص یا ناهمگنی در گره m است. به این معنی که مثلاً اگر همه مشاهدات در یک گره از یک دسته باشند، Gini(m)، برابر صفر و مُبَین کمترین ناخالصی و به عبارت دیگر بیشترین خلوص در گره است و برعکس، بیشترین مقدار Gini(m)، زمانی حاصل می‌شود که از همه مشاهدات به یک نسبت در گره وجود داشته باشند. شاخص جینی در هر گره برای تمام متغیرها محاسبه شده و متغیری به عنوان متغیر جداکننده

1. Classification and Regression Tree

انتخاب می‌شود که کمترین مقدار برای جینی از آن به دست آید. احتمال اولیه، مبین سهم هر یک از دسته‌ها در جامع مرجع است. رشد درخت بر اساس شاخص جینی از همان گره ریشه، که اولین گره بوده و در برگیرندهٔ تمام مشاهدات است، آغاز شده و برای هر درختی که ایجاد می‌شود، هزینهٔ دسته‌بندی اشتباه آن - که می‌توان از آن به برای شاخص خوبی برازش استفاده کرد - طبق رابطه (3) محاسبه می‌شود:

$$\text{misclassificationcost} = \sum_{t=1}^{T} P(t)\left[1-\sum_{j=1}^{J} P^2(j|t)\right] \quad (3)$$

که در آن $P(t)$، سهم مشاهدات موجود در گره نهایی t از کل مشاهدات بوده و T، تعداد گره‌های نهایی است. رابطهٔ فوق مبین آن دسته از داده‌هایی است که به اشتباه در دسته‌های غیر مرتبط با خود، دسته‌بندی شده‌اند. برای ارزیابی درخت ایجاد شده به وسیلهٔ روش CART یا هر روش دیگری معیارهایی وجود دارند. از مهم‌ترین و اصلی ترین این معیارها نرخ خطا در درخت است. برای محاسبهٔ نرخ خطا در درخت ابتدا می‌باید نرخ خطا در هر شاخه به دست آید. نرخ خطا در هر برگ عبارت است از؛ نسبت تعداد رکوردهایی که کلاس یا دسته آنها درست پیش‌بینی نشده است. برای محاسبهٔ نرخ خطای کل درخت، مجموع وزنی نرخ خطاهای برگ‌ها به دست آورده می‌شود (وزن هر برگ در واقع نسبت جمعیت آن برگ به کل جمعیت رکوردها است). کیفیت درخت به دست آمده نیز مهم خواهد بود. جهت جلوگیری از تولید قانون‌های بی‌کیفیت در بعضی از شاخه‌ها، درخت تولیدی به اصطلاح (هرس) می‌شود. این کار با آنکه نرخ خطا را افزایش می‌دهد ولی از ایجاد بعضی قانون‌های ناکارا جلوگیری می‌کند. همچنین باید به این نکته توجه داشت که باید قطع کردن به نحوی صورت گیرد که خطا از مقدار معینی بیشتر نشود. درخت بهینه درختی است که کمترین هزینهٔ دسته‌بندی اشتباه را برای داده‌های آزمایشی داشته باشد [2و5].

همچنین مدل رگرسیون درختی وابستگی زیادی به شرایطی دارد که در الگوریتم آن تعریف شده است. یکی از آن شرایط تعداد داده‌ها در گره‌های والدین (بالاسری) و فرزند (پایین‌تری) است. ما در مدل برازش یافته خود حداقل تعداد داده در گره‌های بالاسری را 100 و حداقل تعداد داده در گره‌های پایین دستی را 50 در نظر گرفته‌ایم.

سنجش کارایی

کارایی هر یک از طبقه‌بندی‌کننده‌ها را می‌توانیم با استفاده از برخی معیارهای آماری بسیار شناخته شده یعنی دقت، حساسیت، و مشخصه بودن طبقه‌بندی ارزیابی کنیم [6]. این معیارها در چهار حالت مثبت درست (TP)، منفی درست (TN)، مثبت اشتباه (FP)، و منفی اشتباه (FN) تعریف می‌شوند. فرض کنید افرادی را مورد آزمایش قرار می‌دهیم تا وجود بیماری در بدن آنها را بسنجیم. برخی از این افراد واقعاً بیمار هستند و آزمون ما نیز می‌گوید مثبت هستند. این موارد مثبت درست نامیده می‌شوند. برخی از افراد بیمار هستند اما آزمون می‌گوید بیماری ندارد، به این موارد منفی اشتباه می‌گوییم. برخی از آنها بیماری ندارند و آزمون هم می‌گوید بیمار نیستند (منفی درست) در پایان، ممکن است برخی افراد واقعا سالم را داشته باشیم که نتیجه آزمون برای آنها مثبت است، اینها موارد مثبت اشتباه هستند. در جدول (2) ماتریسی آمده که تعداد موارد TP، TN، FP و FN را نشان می‌دهد. در اینجا N به معنای تعداد افرادی که دچار بیماری نشده‌اند و P به معنای تعداد افرادی است که دچار بیماری شده‌اند.

جدول 2. ماتریس مربوط به موارد واقعی و پیش‌بینی شده

	پیش‌بینی (سالم)	پیش‌بینی (بیمار)
مشاهده (بیمار)	تشخیص نادرست سالم (FN)	تشخیص درست بیمار (TP)
مشاهده (سالم)	تشخیص درست سالم (TN)	تشخیص نادرست بیمار (FP)

اگر تعداد کل موارد n باشد، پس بنا بر جدول بالا می‌توانیم معیارهای آماری زیر را که مربوط به کارایی هستند، ارزیابی کنیم.

دقت طبقه‌بندی (Classification Accuracy)

این معیار میزان کسری از پیش‌بینی‌های درست را با توجه به ورودی‌های مثبت و منفی می‌سنجد. این معیار تا حد

زیادی به توزیع مجموعه داده‌ها بستگی دارد و به همین دلیل می‌تواند به سادگی منجر به نتیجه‌گیری‌های نادرست درباره کارایی سیستم گردد. معیار دقت طبقه‌بندی به صورت زیر محاسبه می‌شود:

تعداد موارد در مجموعه داده‌ها / تعداد موارد درست پیش‌بینی شده=دقت طبقه‌بندی

$$= (TP + TN) / (P + N) \qquad (4)$$

حساسیت طبقه‌بندی (Classification Sensitivity)

این معیار میزان کسر مثبت‌های درست را می‌سنجد، یعنی، میزان قابلیت سیستم در پیش‌بینی مقادیر درست در کل مواردی که ارائه شده است را می‌دهد. این معیار با استفاده از فرمول زیر محاسبه می‌گردد:

تعداد کل مثبت‌ها / تعداد کل مثبت‌های درست=حساسیت

$$= TP / (TP + FN) \qquad (5)$$

مشخصه بودن طبقه‌بندی (Classification Specificity)

این معیار میزان کسر منفی‌های درست را می‌سنجد، یعنی قابلیت سیستم در پیش‌بینی مقادیر درست را برای مواردی که دقیقاً مخالف حالت‌های مورد علاقه ما هستند، ارزیابی می‌کند. محاسبهٔ آن به صورت زیر است:

تعداد کل موارد / تعداد منفی‌های درست = مشخصه بودن

$$= TN / (TN + FP) \qquad (6)$$

یافته‌ها

پیش از هر چیز کارایی هر یک از طبقه‌بندی‌کننده‌ها با استفاده از تمام متغیرهای ورودی بررسی شد. سپس روش انتخاب خصوصیات بر روی مجموعه داده‌ها اجرا شد و صفتهایی که مهم نبودند از مجموعه داده‌ها حذف شدند. سپس کارایی طبقه‌بندی کننده‌ها دوباره با استفاده از تعداد کاهش یافته صفت‌ها ارزیابی گردید. این مطالعهٔ تجربی با استفاده از نرم‌افزار SPSS انجام شده است. پس از به کار بردن مجموعه داده‌های آموزشی و مجموعه داده‌های آزمایشی بر روی هر طبقه‌بندی کننده، یک ماتریس رده‌بندی به دست می‌آید که مقدارهای مربوط به مثبت‌های درست، منفی‌های درست، مثبت‌های اشتباه و منفی‌های اشتباه را نشان می‌دهد و در زیر آورده شده است. جدول (3) ماتریس رده‌بندی را برای مجموعه داده‌ها نشان می‌دهند. همچنین جدول (4) معیارهای آماری مقایسه‌ای را برای دو مدل مورد بررسی برای مجموعه داده‌ها نشان می‌دهند. هر یک از خانه‌های جدول (3) در زیر حاوی تعداد خام نمونه‌های طبقه‌بندی شده و مربوط به ترکیب خروجی‌های دلخواه و واقعی هر مدل است. مقدارهای پیش‌بینی شده با مقدارهای واقعی کلاس‌ها مقایسه می‌شوند تا مثبت‌های درست، منفی‌های درست، مثبت‌های اشتباه و منفی‌های اشتباه را مشخص کنند. جدول (4) مقدارهای سه معیار آماری برای طبقه‌بندی، یعنی دقت، حساسیت و خاص بودن را برای دو مدل نشان می‌دهد.

جدول 3. ماتریس رده‌بندی مدل‌های مختلف برای مجموعه داده‌ها

مدل رده‌بندی	مشاهده شده		
	نتیجه مدل	سابقه بیماری ندارد	سابقه بیماری دارد
رگرسیون لجستیک	سابقه بیماری دارد	9	20
	سابقه بیماری ندارد	51	6
رگرسیون درختی	سابقه بیماری دارد	14	17
	سابقه بیماری ندارد	45	9

جدول 4. معیارهای آماری مقایسه‌ای مدل‌های مختلف برای مجموعه داده‌ها

مدل	اندازه معیار		
	معیار دقت	معیار حساسیت	معیار مشخصه بودن
رگرسیون لجستیک	%83/5	%76/9	%85
رگرسیون درختی	%72/9	%65/4	%76/3

همان‌طور که در جدول (4) مشاهده می‌شود، دقت به عنوان مهم‌ترین معیار خوبی برازش یک مدل، در مدل رگرسیون لجستیک بهتر از مدل رگرسیون درختی به دست آمده است. از لحاظ معیار آنالیز حساسیت که قابلیت مدل را در پیش‌بینی درست تشخیص طبقه، به ازای همه افرادی که می‌باید در آن طبقه قرار بگیرند؛ باز مدل رگرسیون لجستیک بهتر عمل کرده است. در خصوص شاخص مشخصه بودن که قابلیت مدل را در پیش‌بینی درست قرار ندادن موارد در طبقه مورد نظر نسبت به تمام مواردی که نباید در آن طبقه قرار بگیرند، همچنان مدل رگرسیون لجستیک نقش بهتری ایفا کرده است.

جمع‌بندی

در این مقاله کارایی دو مدل طبقه‌بندی کننده یعنی رگرسیون لجستیک و رگرسیون درختی روی داده‌های بیماری زردی بررسی شده است. کارایی طبقه‌بندی با استفاده از معیارهای آماری کارایی مانند دقت، حساسیت، و خاص بودن مورد تحقیق قرار گرفته است. میزان کارایی تمام طبقه‌بندی‌کننده‌ها در اینجا با استفاده از معیارهای کارایی آماری مانند دقت، توجه به موارد خاص و حساسیت طبقه‌بندی، نشان داده شده است. این مقاله یک مطالعهٔ مقایسه‌ای روی دو مدل رده‌بندی‌کننده اعم از رگرسیون لجستیک و رگرسیون درختی در آمار و داده‌کاوی است. بر این اساس، رگرسیون لجستیک، دقت بالای 83% و رگرسیون درختی میزان دقت حدود 73% را بر روی مجموعه داده‌های آزمایشی نشان داده‌اند. ضریب حساسیت مدل رگرسیون لجستیک در حدود 77% و برای رگرسیون درختی کمتر از 66% می‌باشد. ضریب مشخصه بودن مدل رگرسیون لجستیک برابر 85% و برای رگرسیون درختی بیشتر از 76% می‌باشد؛ بنابراین مدل رگرسیون لجستیک در هر سه مورد تشخیص کارایی بهتر از مدل رگرسیون درختی عمل کرده است.

به نظر می‌رسد مدل رگرسیون درختی به دلایل زیر ضعیف تر از مدل رگرسیون لجستیک عمل می‌کند:

- مدل رگرسیون درختی وابستگی زیادی به شرایطی دارد که در الگوریتم آن تعریف شده است. یکی از آن شرایط تعداد داده‌ها در گره‌های والدین (بالاسری) و فرزند (پایین‌تری) است.
- مدل رگرسیون درختی پایداری در ارائه الگوریتم ندارد به طوری‌که با افزایش تعداد داده‌ها (علی رغم برخورداری از حفظ خصوصیات قبلی مانند فراوانی نسبی هر یک از خصوصیات) می‌تواند خروجی‌های متنوعی ارائه دهد.
- بیش برآوردی و کم برآوردی‌های رگرسیون درختی هنگام به کاربردن آن برای یک متغیر پیوسته یا یک متغیر کمی بیشتر از زمانی است که متغیر ما کیفی است این به دلیل در نظر گرفتن مقدار میانگین مقادیر متغیر وابسته به عنوان نمایندهٔ طبقه است. این ضعف، زمانی بیشتر می‌شود که پراکندگی‌های داده‌ها در طبقات در نظر گرفته شده بیشتر باشد.

علی رغم این مدل‌های درختی انعطاف‌پذیری بسیار زیادی برای برازش مدل روی هر نوع داده‌ای دارند. همچنین برای مدل درختی می‌توان معیارهای کیفی دیگری نیز برشمرد که خصوصیات منحصر به فرد رگرسیون درختی باعث می‌شود این معیارها برای مدل‌های دیگر استفاده نشود. از این جمله بررسی عمق درخت در تحلیل‌های رگرسیونی است به طوری‌که هرچه عمق درخت افزایش یابد کیفیت درخت کاهش می‌یابد و می‌توان این را در مقایسه با شاخص دقت به عنوان یک شاخص کیفیت در تحلیل‌های آماری در نظر گرفت.

منابع

[1] Jiwaei Han, Kamber Micheline, Jian Pei Data mining: Concepts and Techniques, Morgam Kaufmann Publishers (Mar 2006).

[2] Pakgohar, Alireza. Statistical applications in data mining: special view in logistic regression. Islamic Azad University, branch of Mashad. department of Science. M.A degree thesis. 2006. [Persian language].

[3] Pakgohar, Alireza. Evaluation of patients with gastroenteritis, Pneumonia and Jaundice on children, Payame Noor

University, Report. 2012. [Persian Language].
[4] SPSS 18(PASW) help file. http//www-.spss.com
[5] Pakgohar, Alireza. Tabrizi, Reza Sigari. Khalili, Mohadeseh. Esmaeili, Alireza. The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach, Procedia Computer Science, Volume 3, 2011, Pages 764-769, ISSN 1877-0509, 0.1016/j.procs.2010.12.126.
[6] Alaa M. Elsayad "Predicting the severity of breast masses with ensemble of Bayesian classifiers" journal of computer science 6 (5): 576-584, 2010, ISSN 1549-3636.