تشخیص بیماری دیابت با استفاده از روش رده‌بند تقویتی کت‌بوست و بیزی

زهرا احمدیان[۱]، فرزاد اسکندری[۲]

تاریخ دریافت: ۱۴۰۲/۰۲/۰۷
تاریخ پذیرش: ۱۴۰۲/۱۱/۱۷

چکیده:
امروزه تشخیص بیماری‌ها با استفاده از هوش مصنوعی و الگوریتم‌های یادگیری ماشین از اهمیت بسیار بالایی برخوردار است، چراکه با استفاده از داده‌های موجود در زمینه مطالعاتی بیماری موردنظر می‌توان به اطلاعات و نتایج سودمندی دست‌یافت که از رخداد بسیاری از مرگ‌ومیرها می‌کاهد. از جمله این بیماری‌ها می‌توان به تشخیص بیماری دیابت که امروزه با توجه به رشد زندگی شهرنشینی و کاهش فعالیت افراد گسترش‌یافته است، اشاره کرد. پس تشخیص این موضوع که فرد به بیماری دیابت مبتلا می‌گردد یا خیر از اهمیت بسیار بالایی برخوردار است. در این مقاله از مجموعه داده مربوط به اطلاعات افرادی که آزمایش تشخیص دیابت را انجام داده‌اند استفاده‌شده است. این اطلاعات مربوط به ۵۲۰ نفر است، عمل رده‌بندی افراد به دو دسته که آیا نتیجه آزمایش دیابتشان مثبت است یا خیر صورت می‌گیرد و از روش‌های رده‌بند بیزی مانند ماشین بردار پشتیبان بیزی، بیز ساده، CNK و روش رده‌بند ترکیبی کت‌بوست استفاده‌شده است تا بتوان نتیجه گرفت که کدام‌یک از این روش‌ها می‌توانند توانمندی بهتری برای تحلیل داده‌ها داشته باشند و همچنین برای مقایسه این روش‌ها از معیارهای دقت، صحت، وضوح، حساسیت و نمودار راک استفاده‌شده است.

واژه‌های کلیدی: رده‌بندی، رده‌بندی ترکیبی، رده‌بندی بیزی.

۱ مقدمه

در حوزه علم داده‌ها سعی بر آن است که مبانی و مفاهیم نظری طوری گسترش یابند که بتوانند پاسخگوی حل بسیاری از مسائل واقعی باشند که امروزه بررسی توانایی و مقایسه مدل‌ها بسیار مورد توجه واقع گردیده است. لذا در این مقاله نیز هدف پیش بردن تئوری و مفاهیم نظری در راستای حل مسائل واقعی است به‌عنوان‌مثال یکی از مسائلی که امروزه مورد توجه قرار می‌گیرد مربوط به تشخیص بیماری دیابت است. در واقع به این صورت که افرادی که آزمایش دیابت از آن‌ها گرفته‌شده است و نتیجه آزمایش‌هایشان در دو رده مثبت و منفی قرار دارند را می‌توان با استفاده از الگوریتم‌های یادگیری ماشین، رده‌بندی کرد تا بتوان در نهایت یک تعمیم‌پذیری داشت به این صورت که یک فرد جدید با ویژگی‌های جدید آیا مبتلا به بیماری دیابت خواهد شد یا خیر.

اساساً دیابت قندی یا به عبارتی مرض قند که به‌اختصار به آن دیابت گویند بیماری است که در اثر کاهش ترشح انسولین توسط غده لوزالمعده ایجاد می‌شود. انسولین ماده‌ای است که پس از تولید به داخل خون ریخته می‌شود و نقش آن سوزاندن مواد نشاسته‌ای و قندهای ساده در بدن است تا در نهایت قند خون بدن انسان تنظیم گردد. پس کاهش ترشح آن سبب افزایش قند خون و آسیب دیدن نواحی مختلف بدن می‌گردد. منظور از کاهش ترشح انسولین نیز دو معنای متفاوت می‌تواند داشته باشد یا ترشح این ماده از لوزالمعده کاهش می‌یابد یا اثربخشی آن در خون به دلیل حذف برخی فاکتورها کاهش می‌یابد. این بیماری جزو بیماری‌های متابولیسمی قرار دارد. [۱] در اهمیت تشخیص این بیماری می‌توان به دو جنبه مهم آن اشاره کرد که عبارت‌اند از: الف) این بیماری هرساله افراد بسیار زیادی را در سرتاسر جهان درگیر خود می‌کند. فدراسیون بین‌المللی دیابت تخمین زده است که در سال ۲۰۱۷، ۴۵۱ میلیون بزرگ‌سال مبتلا به دیابت در سراسر جهان زندگی می‌کنند که پیش‌بینی می‌شود در صورت عدم اتخاذ روش‌های پیشگیری مؤثر این تعداد به ۶۹۳ میلیون تا ۲۰۴۵ برسد. [۹] ب) نوع رایج این بیماری، بیماری دیابت نوع دو است. دیابت نوع دو اکثراً بدون هیچ علائم

[۱] دانشجوی کارشناسی ارشد علم‌داده‌ها، دانشگاه علامه طباطبائی
[۲] عضو هیئت‌علمی گروه آمار، دانشگاه علامه طباطبائی

خاصی است و ممکن است فرد تا چندین مدت از بیماری خودآگاه نباشد و در این مدت وجود قند خون اضافی در جریان خون سبب آسیب‌دیدگی در نواحی مختلفی چون چشم، کلیه‌ها، قلب و مغز می‌شود. [۱۰]

دیابت نوع دو: این نوع از دیابت اکثراً در افراد میان‌سال گزارش می‌شود. این بیماران اکثراً دارای اضافه‌وزن یا چاق هستند. علائم آن عبارت‌اند از: مصرف بسیار زیاد آب، ادرار مکرر، دفع بیش‌ازحد ادرار در شب، تاری دید، خستگی زودرس و کاهش وزن است. یکی از عوارض شایع این بیماری بی‌حسی در ناحیه پا است. [۱۳]

تحقیقات متعددی در راستای پیش‌بینی و تشخیص بیماری دیابت صورت گرفته است که به ذکر برخی از آن‌ها می‌پردازیم.

[۴] از مجموعه داده دیابت هندی $PIMA$ در تکنیک‌های یادگیری ماشین مانند شبکه‌های عصبی مصنوعی، درخت تصمیم، جنگل تصادفی، بیز ساده، $k-$نزدیک‌ترین همسایگی، ماشین‌های بردار پشتیبان و رگرسیون لجستیک استفاده‌شده و نتایج آن‌ها موردبحث قرارگرفته است.

[۵] از الگوریتم آدابوست و مقایسه آن با روش‌های مختلف رده‌بندی اعم از ماشین بردار پشتیبان، بیزساده و درخت تصمیم برای رده‌بندی داده‌های دیابت استفاده‌شده است.

[۶] از یک مدل جمعی برای تشخیص دیابت نوع دو که شامل بیزساده، ماشین بردار پشتیبان، درخت تصمیم و شبکه عصبی است بهره گرفته و به تشخیص اینکه آیا یک فرد از بیماری رنج خواهد برد یا خیر پرداخته‌شده است.

[۸] برای رده‌بندی بیماران، پنج مدل مختلف پیش‌بینی اعم از ماشین بردار پشتیبان با تابع هسته خطی، $k-$نزدیک‌ترین همسایگی، شبکه عصبی مصنوعی (ANN) استفاده‌شده و به مقایسه پرداخته‌شده است.

[۷] برای تشخیص دیابت از شبکه عصبی عمیق (DNN) بهره گرفته و به پیش‌بینی پرداخته‌شده است.

لازم به ذکر است که در بین تحقیقات متفاوتی که در راستای تشخیص بیماری دیابت صورت گرفته است از روش‌های رده‌بندی بیزی و مقایسه آن با الگوریتم کت‌بوست استفاده‌نشده است به همین دلیل در این مقاله به ذکر و مقایسه این روش‌ها با یکدیگر پرداخته‌شده است. در بخش ۲ به بیان الگوریتم‌های استفاده‌شده برای تحلیل داده‌های دیابت پرداخته خواهد شد.

۲ روش‌های رده‌بند بیزی

۱.۲ ماشین بردار پشتیبان بیزی

در ماشین بردار پشتیبان هدف کمینه کردن تابع هزینه‌ای به‌صورت زیر است

$$d_\alpha(\beta, v) = \sum_{i=1}^{n} \max(1 - y_i x_i^T \beta) + v^{-\alpha} \sum_{j=1}^{k} |\frac{\beta_j}{\sigma_j}|^\alpha \quad (1)$$

[۱۲] در این مقاله هدف برآورد پارامترهای ماشین بردار پشتیبان به‌صورت بیزی است یعنی به شیوه بیزی به برآورد پارامتر β پرداخته می‌شود. پس به‌عنوان تابع هزینه اولیه از ماشین بردار پشتیبان رابطه‌ی ۱ در نظر گرفته می‌شود. کمینه کردن تابع هزینه ۱ جهت برآورد پارامترها معادل با برآورد مد پسین $p(\beta|v, \alpha, y)$ است.

$$p(\beta|v, \alpha, y) \propto exp(-d_\alpha(\beta, v)) \quad (2)$$
$$\propto C_\alpha(v) L(y|\beta) p(\beta|v, \alpha)$$

مقدار $C_\alpha(v)$ یک ضریب ثابت نرمال‌سازی پسین است که در تحلیل نادیده گرفته می‌شود و $L(y|\beta)$ نیز تابع درستنمایی نوع اول نام دارد که به‌صورت رابطه ۳ است.

$$L(y|\beta) = \prod_i L_i(y_i|\beta) = \exp(-2\sum_{i=1}^{k} \max(1 - y_i x_i^T \beta, 0)) \quad (3)$$

(۴)

برقرار شدن این تساوی به این دلیل است که نشان داده می‌شود تابع درستنمایی نوع اول داری توزیع نرمال آمیخته است. پس:

$$L_i(y_i|\beta) = \quad (5)$$
$$= \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_i}} exp(-\frac{1}{2} \frac{(1 + \lambda_i - y_i x_i^T \beta)^2}{\lambda_i})$$
$$= \int_0^\infty \frac{1}{\sqrt{2\pi\lambda}} exp(-u - \frac{\lambda}{2} - \frac{u^2 \lambda^{-1}}{2}) d\lambda$$
$$= \int_0^\infty \frac{1}{\sqrt{2\pi\lambda}} exp(\frac{-(u + \lambda)^2}{2\lambda}) d\lambda$$
$$= exp(-2\max(u, 0)) \quad (6)$$

که در آن

$$\int_0^\infty \phi(u| - \lambda, \lambda) d\lambda = exp(-2\max(u, 0)) \quad (7)$$

و $\phi(u|.,.)$ نشان‌دهنده توزیع نرمال است. طبق نتایج به‌دست‌آمده از رابطه ۵ می‌توان توزیع پیشین را به‌صورت ۸ نوشت.

$$p(\beta|v,\alpha)=\prod_{j=1}^{k}p(\beta_j|v,\alpha)= \\ (\frac{\alpha}{v\tau(1+\alpha^{-1})})^k exp(-\sum_{i=1}^{k}|\frac{\beta_j}{v\sigma_j}|^{\alpha}) \qquad (8)$$

و می‌توان توزیع پسین را در نهایت به‌صورت ۹ نوشت.

$$p(\beta,\lambda,w|y,v,\alpha)\propto \\ \prod_{i=1}^{n}\lambda_i^{\frac{-1}{2}}\exp(\frac{-1}{2}\sum_{i=1}^{n}(\frac{1+\lambda_i-y_ix^T\beta}{\lambda_i})^2)\times \\ \prod_{j=1}^{k}w_j^{\frac{-1}{2}}exp(\frac{-1}{2v^2}\sum_{j=1}^{k}\frac{\beta_j^2}{\sigma_j^2 w_j})p(w_j|\alpha) \qquad (9)$$

که در آن $p(w_j|\alpha)\propto w_j^{\frac{-2}{2}}St^{+}_{\frac{\alpha}{2}}(w_j^{-1})$ است. که در نتیجه با استفاده الگوریتم امید ریاضی-بیشینه‌سازی (EM)، β برآورد می‌شود. [۱۲]

۲.۲ بیز ساده

رده‌بندی بیز ساده خانواده‌ای از الگوریتم‌های احتمالی یادگیری ماشین بر اساس قضیه بیز است و این فرض را ایجاد می‌کند که ویژگی‌های مورداستفاده برای رده‌بندی، با توجه به برچسب کلاس، به‌طور مشروط مستقل هستند؛ به‌عبارت‌دیگر، فرض بر این است که وجود یا عدم وجود یک ویژگی تأثیری بر وجود یا عدم وجود هیچ ویژگی دیگری ندارد.
ایده اصلی این الگوریتم به این صورت است که احتمال تعلق نمونه‌ای که قبلاً دیده نشده است را برای هر کلاس محاسبه و سپس محتمل‌ترین کلاس را انتخاب می‌کند. فرض کنید $X=(x_1,x_2,\ldots,x_n)$ برداری از n ویژگی را بیان می‌کند که به‌صورت متغیرهای مستقل می‌باشند؛ و احتمال رخداد C_K به‌طوری K بیان‌گر یک رده از $(1,\ldots,k)$ رده‌ی متفاوت است که طبق وجود فرض استقلال شرطی در بیز ساده به‌صورت رابطه‌ی ۱۰ به دست می‌آید.

$$P(C_K|X)=\frac{P(X|C_K)P(C_K)}{P(X)} \\ P(C_K|X)=\frac{P(x_1,\ldots,x_n|C_K)P(C_K)}{P(X)} \\ P(C_K|X)=\frac{P(x_1|C_K)\ldots P(x_n|C_K)P(C_K)}{P(X)} \qquad (10)$$

در رابطه‌ی ۱۰ مخرج کسر یک مقدار ثابت است که می‌توان از آن صرف‌نظر نمود. در نهایت رابطه‌ی فوق، به‌صورت زیر به دست می‌آید.

$$P(C_K|X)\propto P(x_1|C_K)\ldots P(x_n|C_K)P(C_K) \\ P(C_K|x_1,\ldots,x_n)\propto P(C_K)\prod_{i=1}^{n}P(x_i|C_K) \qquad (11)$$

بنابراین در بیز ساده نیز هدف یافتن C_Kای است که منجر به بیشینه نمودن رابطه‌ی ۱۱ و حصول رابطه‌ی ۱۲ شود.

$$C^*=arg\max_K P(C_K|X) \\ \propto arg\max_K P(C_K)\prod_{i=1}^{n}P(x_i|C_K) \qquad (12)$$

۳.۲ ترکیب بیزساده و $k-$ نزدیک‌ترین همسایه (CNK)

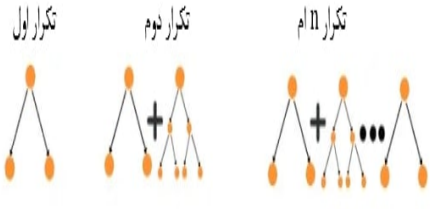
الگوریتم CNK یک روش جدید رده‌بندی است که الگوریتم $k-$ نزدیک‌ترین همسایگی را با رده‌بندی بیز ساده‌ترکیب می‌کند و دارای قدرت پیش‌بینی بالایی می‌باشد.
در الگوریتم بیز ساده یکی از فاکتورهای مهم سروکار داشتن با ویژگی‌های عددی می‌باشد چون درون الگوریتم نیازمند محاسبه احتمال‌های شرطی است که برای این موضوع نیازمند به گسسته سازی کردن متغیرهای عددی هستیم تا ویژگی‌های عددی را به چندین رده تقسیم کنیم؛ بنابراین تکنیک گسسته سازی نقش بسیار مهمی دارد. چندین راهکار و روش‌هایی به وجود آمده‌اند که دقت الگوریتم بیز ساده را با استفاده از روش‌های گسسته سازی بهبود می‌بخشند. [۱۴]. الگوریتم $k-$ نزدیک‌ترین همسایگی وضعیت کاملاً برعکس است و با ویژگی‌های رسته‌ای مشکل دارد. ازآنجایی‌که الگوریتم به محاسبه فاصله می‌پردازد لذا باید یک طرح اندازه‌گیری فاصله برای داده‌ها معرفی شود که به‌طورکلی از انواع روش‌های اندازه‌گیری شباهت استفاده می‌شود که مطالعات بسیاری برای یافتن یک طرح اندازه‌گیری فاصله انجام‌شده است. در این روش یک الگوریتم جدیدی ارائه‌شده است که دو روش بیز ساده و $k-$ نزدیک‌ترین همسایگی را باهم ترکیب می‌کند طوری که مشکلاتی که وجود داشت نیز حل می‌گردد یعنی نیازی به گسسته سازی متغیرهای عددی و محاسبه فاصله بین ویژگی‌های رسته‌ای نیست.
حال به توضیح ایده ترکیب دو رده‌بند بیز ساده و $k-$ نزدیک‌ترین همسایگی می‌پردازیم: برای رده‌بندی یک داده جدید ابتدا از الگوریتم $k-$نزدیک‌ترین همسایگی برای یافتن k نزدیک‌ترین همسایه از مجموعه داده آموزشی استفاده می‌کنیم. در حین پیاده‌سازی الگوریتم

$k-$نزدیک‌ترین همسایگی ویژگی‌های رسته‌ای درج نمی‌شوند و فاصله فقط با ویژگی‌های عددی محاسبه خواهد شد. بعد از یافتن k تا نزدیک‌ترین همسایه برای داده، با استفاده از الگوریتم بیز ساده یک مدل ساخته می‌شود اما به این صورت که فقط شامل ویژگی‌های رسته‌ای خواهد بود. پس فرایند، دومرحله‌ای است که در مرحله اول تنها از ویژگی‌های عددی برای انتخاب نزدیک‌ترین داده‌ها به داده جدید استفاده می‌شود و امری منطقی است زیرا داده‌های نزدیک به داده جدید باید خاصیت و ویژگی‌های یکسانی داشته باشند و در مرحله دوم به رابطه داده‌های رسته‌ای با کلاس نگاه می‌شود که از الگوریتم بیز ساده استفاده می‌گردد به‌این‌ترتیب از هر دو ویژگی عددی و رسته‌ای برای رده‌بندی یک داده جدید بدون هیچ تغییری در داده‌ها استفاده می‌شود. [۳]

۳ روش رده‌بند تقویتی

۱.۳ کت‌بوست

پیش از تعریف و بررسی الگوریتم کت‌بوست به توضیح کلی روش‌های رده‌بندی ترکیبی خواهیم پرداخت، همان‌طور که از نام آن پیدا است این روش‌ها از مجموعه‌ای از یادگیرنده‌ها مثل روش‌ها و الگوریتم‌های مختلف رده‌بندی به‌طور تکراری استفاده می‌کند که منجر به ایجاد مدل‌ها و رده‌بندی‌هایی می‌شود که به‌مراتب دارای دقت بالایی هستند و نقاط ضعف کمتری دارند. به‌طورکلی به سه بخش الگوریتم‌های انبوهش تصادفی، الگوریتم بگینگ و الگوریتم بوستینگ (تقویتی) تقسیم می‌شوند که در مقاله از الگوریتم کت بوست که زیرمجموعه الگوریتم بوستینگ (تقویتی) است استفاده‌شده است. الگوریتم کت‌بوست یکی از الگوریتم‌های روش تقویتی است که در سال ۲۰۱۷ توسط یاندکس ارائه گردیده است. این الگوریتم توانسته است با موفقیت ورودی‌های رسته‌ای شده را بدون انجام اعمال کدگذاری مستقیم در مرحله پیش‌پردازش داده‌ها مدیریت کند و همچنین سایر ویژگی‌های عددی و متنی را نیز مدیریت می‌کند. وجود این کدگذاری مناسب در داخل الگوریتم سبب شده است تا این الگوریتم معرفی شود و دارای مزایای بیش‌تری نسبت به سایر الگوریتم‌های رده‌بندی شود و مورد توجه در این مقاله قرار گیرد. همچنین در طول آموزش این الگوریتم، مجموعه‌ای از درختان تصمیم متقارن (درخت متقارن: درختی که هر گره والد آن شامل هیچ یا دو فرزند باشد) به‌صورت دنباله‌ای ساخته می‌شود که هر درختی با کاهش دادن هزینه نسبت به قبلی ایجاد می‌شود. این الگوریتم از یک طرح‌واره جدید برای محاسبه مقادیر برگ‌ها هنگام انتخاب ساختار درخت استفاده می‌کند که به کاهش بیش برازش کمک می‌کند. پس وجود ویژگی‌هایی مانند کدگذاری متغیرهای رسته‌ای به شیوه جدید (با استفاده از آماره هدف مرتب‌شده) و کاهش بیش برازش سبب برتری و قدرت این الگوریتم نسبت به سایر الگوریتم‌ها شده است. شکل ۱.۳ نمونه‌ای از ساختار رده‌بند کت‌بوست را نشان می‌دهد.



شکل ۱.۳. نمونه‌ای از ساختار رده‌بند کت‌بوست

الگوریتم کت‌بوست در واقع به‌عنوان راه‌حلی برای حل مشکل انتقال پیش‌بینی الگوریتم تقویت گرادیان ارائه شد. در واقع می‌توان نشان داد که توزیع $F(x)$ به‌طوری‌که x عضو مجموعه داده آموزشی باشد با توزیع $F(x)$ به‌طوری‌که x عضو داده آزمایشی باشد متفاوت است که در آن F یک مدل پیش‌بینی است. در نتیجه با ارائه اصل ترتیب این مشکل برطرف شد. از سویی روش استفاده‌شده در گرادیان بوستینگ برای بهره بردن از متغیرهای گسسته در محاسبات مربوط به نزول گرادیان بهینه نیست و یکی از بهترین روش‌ها تبدیل متغیرهای گسسته به آماره‌های هدف مربوطه به آن‌ها است. در نتیجه با ترکیب اصل ترتیب و آماره‌های هدف، الگوریتم کت بوست متولد می‌شود. اگر یک مدل گرادیان بوستینگ را به‌صورت زیر در نظر بگیریم.

$$F^t = F^{t-1} + \alpha h^t \qquad (۱۳)$$

که در رابطه (۱۳) h^t یک پیش‌بینی کننده پایه و α اندازه گام است، هدف کمینه کردن مقدار زیان مورد انتظار $EL(y, F(x))$ است.

به زبان ساده‌تر داریم:

$$h^t = arg\min_h EL(y, F^{t-1}(x) + h(x)) \qquad (۱۴)$$

که با استفاده از روش نیوتن به کمینه‌سازی پرداخته می‌شود. الگوریتم کت‌بوست در واقع نسخه‌ای از گرادیان تقویتی است که در آن از درخت‌های دودویی به‌عنوان برآوردگر پایه استفاده‌شده است.

حال به بیان نحوه کدگذاری توسط الگوریتم کت‌بوست می‌پردازیم همان‌طور که می‌دانیم یک روش محبوب برای مواجه با ویژگی‌های رسته‌ای استفاده از روش کدگذاری داغ است اما زمانی که سطوح متغیر رسته‌ای (کاردینالتی) بسیار زیاد می‌شود، روش کدگذاری داغ موجب

ایجاد ویژگی‌های جدید بسیاری می‌شود. یک راه‌حل مرسوم استفاده از آماره هدف به جهت خوشه‌بندی مقادیر متغیر رسته‌ای است. در واقع آماره هدف امید متغیر پاسخ در هر خوشه را برآورد می‌کند. در واقع می‌توان نشان داد در میان تمام خوشه‌بندی‌های دو دسته‌ای با توجه به یک آستانه برای آماره هدف می‌توان با استفاده از ضرایب جینی، روش حداقل مربعات جزئی و دیگر روش‌ها، یک خوشه‌بندی دو رده بهینه برای متغیر رسته‌ای در مسئله رده‌بندی یا رگرسیون یافت.

اگر مقدار متغیر رسته‌ایiام در بررسی نمونه k ام برابر با$\hat{x}_k^i$، باشد در این صورت مقدار آماره هدف این متغیر عبارت است از $\hat{x}_k^i \approx E(y|x^i = x_k^i)$ به جهت آنکه با مشکل تغییر توزیع متغیر پاسخ در گرادیان بوستینگ مواجه نشویم نیاز به دو ویژگی برای این آماره هدف داریم

۱- $E(\hat{x}^i|y = v) = E(\hat{x}_k^i|y_k = v)$که در آن$(x_k, y_k)$ k امین نمونه هستند.

۲- استفاده بهینه از تمام نمونه‌های آموزشی برای محاسبه آماره هدف و آموزش مدل.

این دو ویژگی از بررسی روش‌های متفاوت محاسبه آماره هدف ارائه‌شده‌اند؛ و هرکدام سعی در حل بخشی از مشکلات روش‌های پیشین را دارند. روش‌های متفاوتی برای محاسبه آماره هدف وجود دارد. ازاین‌دست می‌توان به آماره هدف حریصانه $GreedyTS$، هولد اوت $HoldoutTS$، لیو وان اوت $Leave-one-outTS$ اشاره کرد.

۱- آماره هدف حریصانه: یک راه ساده برای برآورد= $E(y|x^i$ x_k^i)استفاده از میانگین متغیر پاسخ برای نمونه‌هایی است که در آن‌ها $x^i = \hat{x}_k^i$ برای هموارسازی میانگین از مقادیر زیر استفاده می‌شود.

$$\hat{x}_k^i = \frac{\sum_{j=1}^{n} \mathbb{1}_{x_j^i = x_k^i} . y_j + a.p}{\sum_{j=1}^{n} \mathbb{1}_{x_j^i = x_k^i} + a}$$

(۱۵)

که در رابطه (۱۵) ، $a > 0$ یک پارامتر است. یک مقدار متداول برای p عبارت است از میانگین متغیر پاسخ برای تمام نمونه‌ها. مشکل رویکرد حریصانه ریزش هدف است ویژگی $\hat{x}_k^i$ با استفاده ازy_k محاسبه‌شده است؛ که در واقع متغیر پاسخ است. این موضوع منجر می‌شود تا توزیع شرطی $\hat{x}^i|y$ در مجموعه داده آموزشی و آزمایشی متفاوت شود. مثال زیر نشان می‌دهد که این موضوع تا به‌اندازه‌ای می‌تواند بر روی خطای تعمیم مدل آموزش‌دیده شده اثر بگذارد. فرض کنید که ویژگی i ام رسته‌ای باشد و همه مقادیرش هم یکتا هستند برای دسته‌ای مانند A، $p(y = 1|x^i = A) = 0.5$ را داریم. همچنین در مجموعه داده آموزشی رابطه $x_k^i = \frac{y_k + ap}{1 + a}$ نیز برقرار است. پس فقط کافی است که یک حد آستانه‌ای مانند $t = \frac{0.5 + ap}{1 + a}$ را برای رده‌بندی تمام نمونه‌های آموزشی در نظر بگیریم. گرچه برای تمام نمونه‌های داده آزمایشی مقدار آماره هدف حریصانه برابر با p است و مدل مشاهده‌شده برای $p < t$ مقدار صفر را می‌دهد و در غیر این صورت مقدار ۱ را می‌دهد. بنابراین دقت ۵۰ درصد را برای هر دو حالت خواهیم داشت. به همین دلیل یکی از ویژگی‌های آماره هدف حریصانه به‌صورت رابطه $E(\hat{x}^i|y = v) = E(\hat{x}_k^i|y_k = v)$ است که پیش‌تر نیز بیان نمودیم. که در آن $E(\hat{x}_k^i|y_k) = \frac{y_k + ap}{1 + a}$ و همچنین $E(\hat{x}^i|y) = p$ باهم متفاوت هستند. همچنین چندین راه برای جلوگیری از تغییر یافتن توزیع شرطی در مجموعه داده آموزشی و آزمایشی وجود دارد که یکی از عمده‌ترین آن‌ها محاسبه آماره هدف برایx_k بر روی زیرمجموعه‌ی $D_K \subset D \setminus x_k$ است که در نهایت به‌صورت زیر تعریف می‌شود.

$$\hat{x}_k^i = \frac{\sum_{x_k \in D_k} \mathbb{1}_{x_j^i = x_k^i} . y_j + a.p}{\sum_{x_k \in D_k} \mathbb{1}_{x_j^i = x_k^i} + a}$$

(۱۶)

۲-آماره هدف هولد اوت: یک روش بخش‌بندی مجموعه داده آموزشی به دو بخش $D = \hat{D}_0 \cup \hat{D}_1$ و استفاده از $D_k = \hat{D}_0$ به‌عنوان بخش مربوط به آماره هدف رابطه (۱۶) و $\hat{D}_1$ برای آموزش است این موضوع بخشی از مشکلات روش حریصانه را حل می‌کند اما موجب کاهش دیتاست استفاده‌شده برای آموزش می‌شود.

۳- آماره هدف لیو وان اوت:

در این روش در قسمت آموزش $D_k = D \setminus x_k$ و در قسمت تست $D_k = D$ را قرار می‌دهیم. متأسفانه این روش نیز نمی‌تواند مانع از تغییر یافتن توزیع شرطی مجموعه داده آموزشی و آزمایشی شود.

اما روش استفاده‌شده در کت‌بوست آماره هدف مرتب‌شده نام دارد که بهینه‌تر از روش‌های پیشین می‌باشد. آماره هدف مرتب‌شده بر مبنای اصل ترتیب عمل می‌کند و ایده آن از الگوریتم‌های یادگیری آنلاین به‌دست‌آمده است. واضح است که محاسبه آماره هدف نیازمند داده‌های تاریخی و قبلی است به همین جهت و به دلیل تبدیل ایده روش‌های آنلاین به آفلاین یک‌زمان مصنوعی در فضای داده‌ای در دسترس تعریف می‌شود. در واقع یک جایگشت تصادفی بر روی داده‌های آموزشی انجام می‌دهیم و سپس از تمام داده‌هایی که ازنظر این جایگشت تصادفی برای یک نمونه خاص قبلی هستند برای محاسبه آماره هدف استفاده

می‌کنیم. اگر σ را جایگشت انجام‌شده در نظر بگیریم در این صورت مجموعه دیتای استفاده‌شده برای محاسبه آماره هدف برای نمونه k ام به شیوه زیرنشان داده می‌شود. $D_k = x_j : \sigma(j) < \sigma(k)$ این روش محاسبه آماره هدف هر دو ویژگی لازم را دارا است. حال با ترکیب این روش برای محاسبه آماره‌های هدف و استفاده از درخت‌های باینری و جایگزینی تمام این موارد در روش تقویت گرادیان الگوریتم کت بوست حاصل می‌شود [۱۱].

مثال ۱.۳. *فرض نمایید می‌خواهیم در جدول (۱.۳) به کدگذاری مجموعه داده با استفاده از آماره هدف مرتب‌شده بپردازیم.*

جدول ۱.۳. کدگذاری

flower	*size*	*color*
۱	۵	*red*
۲	۵	*red*
۱	۴	*red*
۰	۵	*green*
۰	۷	*green*
۱	۲	*blue*

به کدگذاری سطر سوم از جدول می‌پردازیم.
رابطه آماره هدف سفارش شده به شرح زیر است:

$$\frac{current\ count + (a * p)}{maximum\ count + a} \qquad (۱۷)$$

در رابطه (۱۷) مقادیر a و p پارامترهای ثابتی هستند و به‌طور پیش‌فرض برابر ۱ و ۰.۵ هستند.
current count: مجموع همه‌ی مقادیر کلاس‌هایی است که دسته یکسان و مشابهی دارند که داده به آن دسته تعلق دارد.
maximum count: مجموع آیتم‌هایی که دسته مشابه دارند و بالای سطر مدنظر قرار دارند.
همان‌طور که مشاهده می‌کنیم تعداد maximum count برابر با ۲ است زیرا تنها دو مقدار red در بالای این سطر قرار دارد. همچنین تعداد current count نیز برابر است با ۳ = ۱ + ۲ پس داریم:

$$\hat{x}_۳ = \frac{۳ + (۱ \times ۵/۰)}{۲ + ۱} = ۱/۶۶$$

برای سطرهای دیگر نیز به‌این‌ترتیب محاسبه می‌شود.

۴ تحلیل و بررسی

همان‌طور که می‌دانیم الگوریتم‌های متعددی در حوزه رده‌بندی داده‌ها وجود دارد وجود این الگوریتم‌های متفاوت سبب گردیده است که محققان ابزارها و روش‌های متنوعی را جهت ارزیابی دقت الگوریتم‌ها ابداع کنند که در این بخش به توضیح آن‌ها خواهیم پرداخت.

۱.۴ ماتریس اغتشاش

ماتریس اغتشاش (ماتریس درهم‌ریختگی) به بررسی چگونگی عملکرد الگوریتم رده‌بندی با توجه به مجموعه داده ورودی به تفکیک انواع رده‌های مسئله رده‌بندی، می‌پردازد. در این ماتریس مقدار مثبت‌های درست (TP) و منفی‌های درست (TN) عملکرد درست رده‌بند و مقدار مثبت‌های نادرست (FP) و منفی‌های نادرست (FN) عملکرد نادرست رده‌بند را بازخورد می‌کنند.

- TP تعداد موارد مثبتی است که رده‌بند به‌درستی آن‌ها را به‌عنوان رده‌ی مثبت شناسایی کرده است.
- FP تعداد موارد مثبتی است که رده‌بند به‌اشتباه آن‌ها را به‌عنوان رده‌ی مثبت شناسایی کرده است و در واقعیت به‌عنوان رده‌ی منفی می‌باشند.
- TN تعداد موارد منفی است که رده‌بند به‌درستی آن‌ها را به‌عنوان رده‌ی منفی شناسایی کرده است.
- FN تعداد موارد منفی است که رده‌بند به‌اشتباه آن‌ها را به‌عنوان رده‌ی منفی تشخیص داده است و در واقعیت به‌عنوان رده‌ی مثبت می‌باشند.

با توجه به توضیحات و تعاریف ذکرشده، می‌توان عملکرد یک رده‌بندی که دارای دو رده هستند را به کمک ۱.۴ بررسی کرد.

جدول ۱.۴. ماتریس اغتشاش

شرح		رده‌ی پیش‌بینی شده: مثبت	رده‌ی پیش‌بینی شده: منفی	
رده‌ی واقعی	مثبت	مثبت‌های درست (TP)	منفی‌های نادرست (FN)	P
	منفی	مثبت‌های نادرست (FP)	منفی‌های درست (TN)	N
مقدار کل		P'	N'	$P+N$

با توجه به ماتریس اغتشاش می‌توان معیارهای مختلف ارزیابی رده‌بندی را بیان کرد. این معیارها شامل صحت، وضوح، حساسیت، دقت، نرخ خطا، خطای نوع اول و خطای نوع دوم می‌باشند که آن‌ها را توضیح خواهیم داد.

تعریف ۱.۴. صحت: *یکی از رایج‌ترین شاخص‌های اندازه‌گیری کیفیت یک رده‌بند، شاخص صحت می‌باشد که کسری از موارد مثبت راستین*

را در بین نمونه‌هایی که به‌عنوان مثبت پیش‌بینی‌شده‌اند، نشان می‌دهد شاخص صحت بر اساس مقادیر حاصل از ماتریس اغتشاش به‌صورت رابطه‌ی (۱۸) حاصل می‌شود.

$$Accuracy = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN} \quad (۱۸)$$

تعریف ۲.۴. دقت: *از شاخص دقت به‌عنوان یک شاخص درستی استفاده می‌شود. در واقع برای دست‌یابی به میزان مشاهداتی که توسط یک مدل رده‌بند با دو رده، به‌درستی در رده‌ی مثبت تشخیص داده‌شده‌اند، از شاخص دقت بهره می‌گیریم.*

$$Precision = \frac{TP}{P} = \frac{TP}{TP+FN} \quad (۱۹)$$

تعریف ۳.۴. وضوح: *شاخص وضوح که تحت عنوان شاخص نرخ منفی‌های درست شناخته می‌شود، بیان‌گر آن است که مدل رده‌بند چه درصدی از موارد منفی را به‌درستی شناسایی کرده است. این شاخص طبق رابطه‌ی (۲۰) حاصل می‌شود.*

$$TNR = \frac{TN}{N} = \frac{TN}{TN+FP} \quad (۲۰)$$

تعریف ۴.۴. حساسیت: *شاخص حساسیت یا نرخ مثبت‌های درست، بیان‌گر میزان موارد مثبتی است که مدل رده‌بند به‌درستی شناسایی کرده است. حساسیت به‌عنوان نرخ تمامیت نیز شناخته می‌شود.*

$$TPR = \frac{TP}{P} = \frac{TP}{TP+FN} \quad (۲۱)$$

مدلی که به‌خوبی برازش شده باشد باید از مقدار TPR بالا (در حالت ایدئال برابر یک) برخوردار باشد.

ارزیابی

مجموعه داده استفاده‌شده در این بخش مربوط به اطلاعات ۵۲۰ فرد مراجعه‌کننده به بیمارستان است که از آن‌ها آزمایش دیابت گرفته‌شده است. [۲] این مجموعه داده دارای ویژگی‌هایی به شرح سن، جنسیت، ادرار مکرر، عطش بیش‌ازحد، کاهش وزن ناگهانی، ضعف، پرخوری زیاد، برفک تناسلی، تاری دید، خارش، کج‌خلقی، تأخیر در بهبودی، فلج نسبی، اسپاسم عضلانی، ریزش مو، چاقی و نتیجه آزمایش به‌عنوان متغیر پاسخ است که به دو رده مثبت و منفی تقسیم‌شده است و باقی متغیرها نیز به‌عنوان متغیرهای ورودی یا ویژگی‌ها تعریف می‌شوند. هدف پیش‌بینی نتیجه آزمایش یک شخص با استفاده از ویژگی‌های مذکور با به کار بردن الگوریتم‌های بیان‌شده در بخش‌های ۲ و ۳ و انتخاب مدلی با بیش‌ترین دقت است. گام اول از شروع به بررسی و تحلیل داده‌ها فرایند پیش‌پردازش داده‌ها که جزئی از آماده‌سازی داده‌ها است می‌باشد که یک مرحله مقدماتی خیلی مهم برای فرایند داده‌کاوی محسوب می‌شود. اخیراً، تکنیک‌های پیش‌پردازش داده‌ها برای آموزش مدل‌های یادگیری ماشین و مدل‌های هوش مصنوعی بسیار پیشرفت کرده است. پیش‌پردازش داده‌ها، داده‌ها را به قالبی تبدیل می‌کند که در داده‌کاوی و یادگیری ماشین، پردازش آسان‌تر و مؤثرتر باشد. چندین روش برای پیش‌پردازش داده‌ها وجود دارد که به شرح آن‌ها خواهیم پرداخت:

۱- تبدیل داده، تبدیل داده‌های خام به ساختار قابل‌قبول.

۲-حذف نویز یا خطا که به هنگام جمع‌آوری داده‌ها پدید آمده است.

۳-جانهی کردن مقادیر گمشده یا حذف آن‌ها درصورتی‌که اطلاعات از دست نرود.

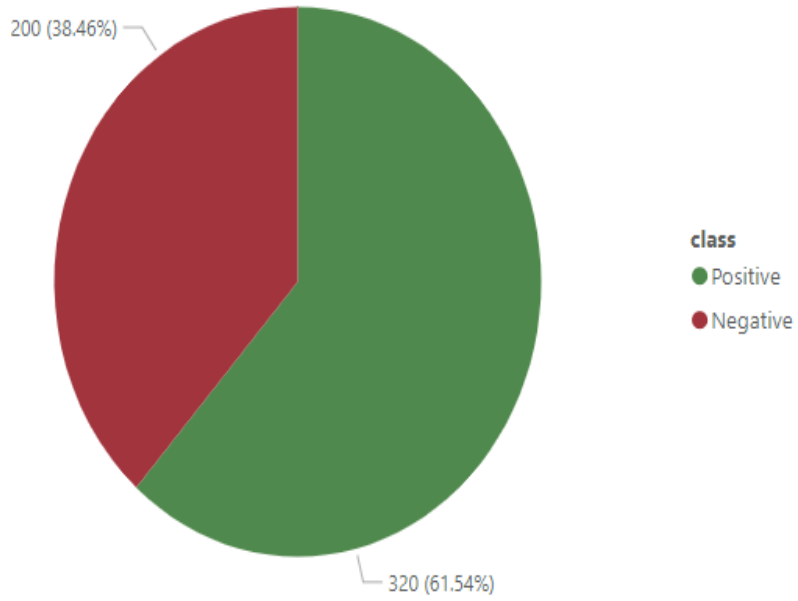
۴-استاندارد کردن داده‌ها

۵-استخراج ویژگی که یک زیرمجموعه ویژگی مرتبط را که در یک زمینه خاص مهم است، استخراج می‌کند.

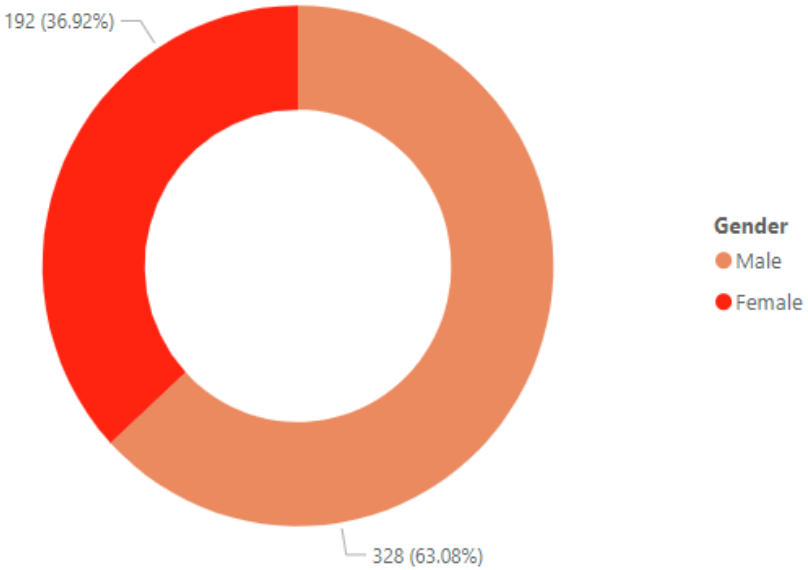
برای پیش‌پردازش مجموعه داده دیابت به بررسی داده‌های گمشده، استخراج ویژگی‌های مهم و استانداردسازی پرداخته‌ایم؛ و ازآنجایی‌که اکثر متغیرهای ورودی گسسته بودند نیازی به استانداردسازی داده‌ها نبود و تنها متغیر سن استانداردسازی شده است. از روش‌های متداول کدگذاری نیز برای کدگذاری متغیرها استفاده‌شده است اما لازم به ذکر است که به هنگام آموزش الگوریتم کت بوست از داده‌های قبل از کدگذاری استفاده کردیم چراکه این الگوریتم خود از کدگذاری خاصی بهره می‌برد.

داده موردنظر شامل هیچ مقدار گمشده‌ای نبود. با پیاده‌سازی روش انتخاب ویژگی خی دو نیز به این نتیجه رسیدیم که لازم است متغیرهای ادرار مکرر، کاهش وزن ناگهانی، تاری دید، تأخیر در بهبودی از مجموعه ویژگی‌های موردنظر حذف شوند و از باقی متغیرها به‌عنوان ویژگی استفاده شود. حال پیش از آغاز رده‌بندی داده‌ها به مصورسازی داده‌ها خواهیم پرداخت چراکه مصورسازی داده‌ها باعث می‌شود که داده‌ها با استفاده از اشکال مختلف قابل‌نمایش شوند و ما را به درک هر چه‌بهتر اطلاعات موجود در داده‌ها می‌رساند. ابتدا به فراوانی داده‌ها در متغیر پاسخ می‌پردازیم که در شکل ۱.۴ نشان داده می‌شود. از نتیجه به‌دست‌آمده متوجه می‌شویم که بین رده‌های متغیر پاسخ تعادل وجود ندارد یعنی فراوانی یک کلاس از کلاس دیگر به‌طور قابل‌توجهی بیشتر است پس لازم است از روش‌های متعادل‌سازی استفاده کنیم تا فراوانی

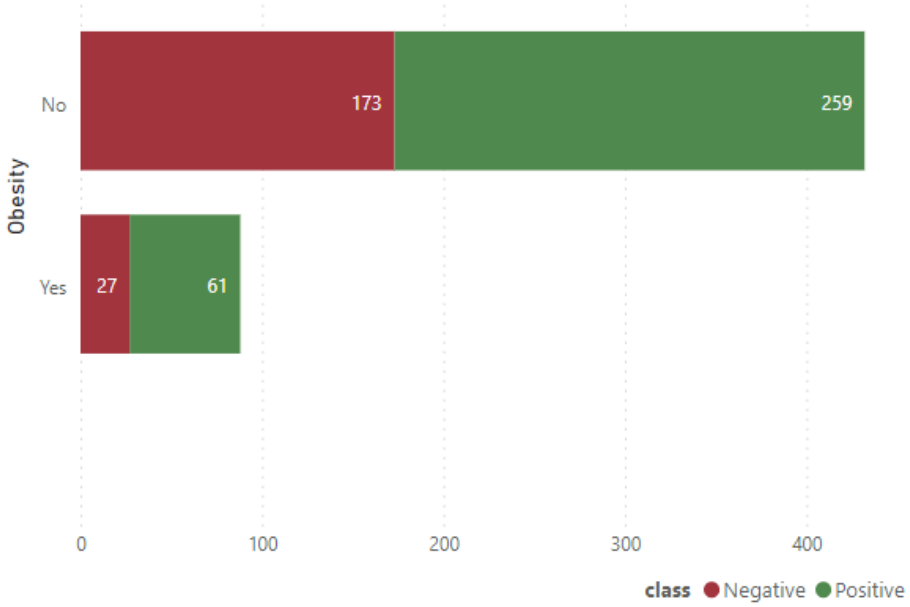
مشاهدات به یک اندازه برسد، عدم تعادل بین مشاهدات رده‌ها باعث می‌شود که به هنگام آموزش الگوریتم‌ها رده‌ای که بیش‌ترین فراوانی را دارد بیش‌تر آموزش ببیند و در نتیجه به هنگام تعمیم‌پذیری و استفاده از نمونه‌های آزمایشی بیشتر داده‌ها برچسب کلاسی را اختیار کنند که فراوانی بیش‌تری داشته باشد.



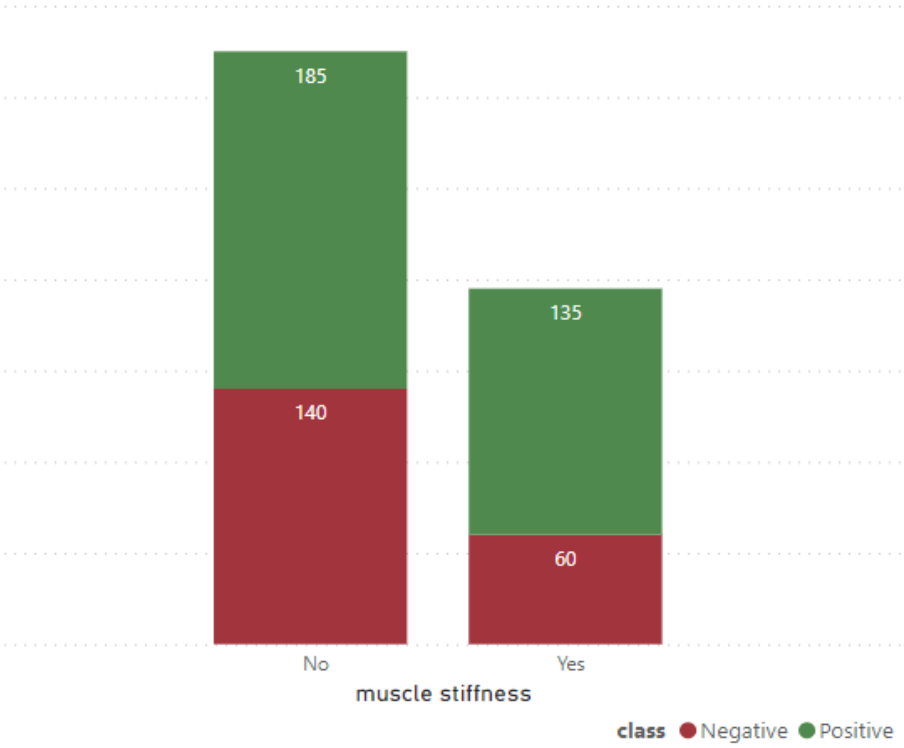
شکل ۰.۴ نمودار دایره‌ای نتایج آزمایش



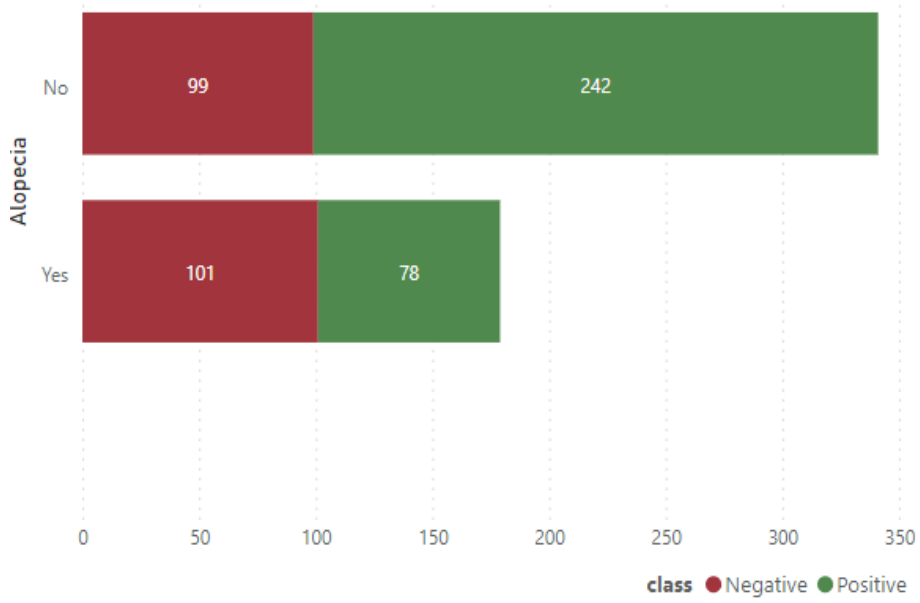
شکل ۰.۳ نمودار دونات جنسیت



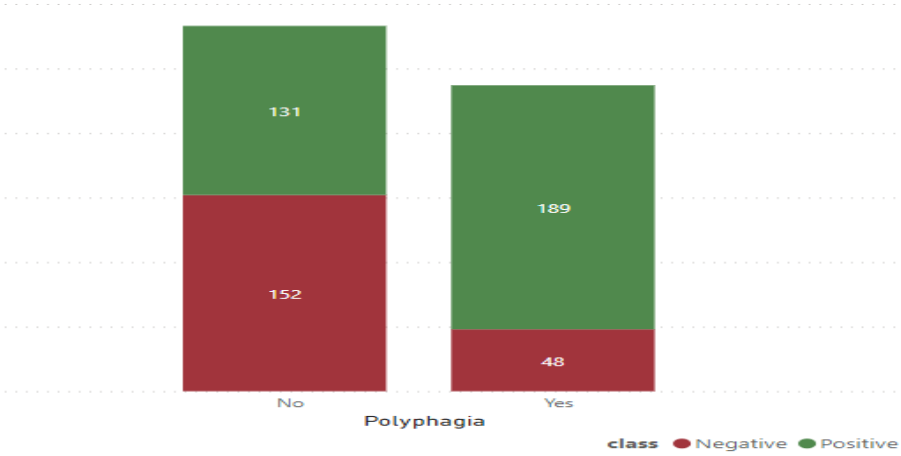
شکل ۴.۴. نمودار چاقی در برابر نتیجه آزمایش



شکل ۵.۴. نمودار اسپاسم عضلانی در برابر نتیجه آزمایش



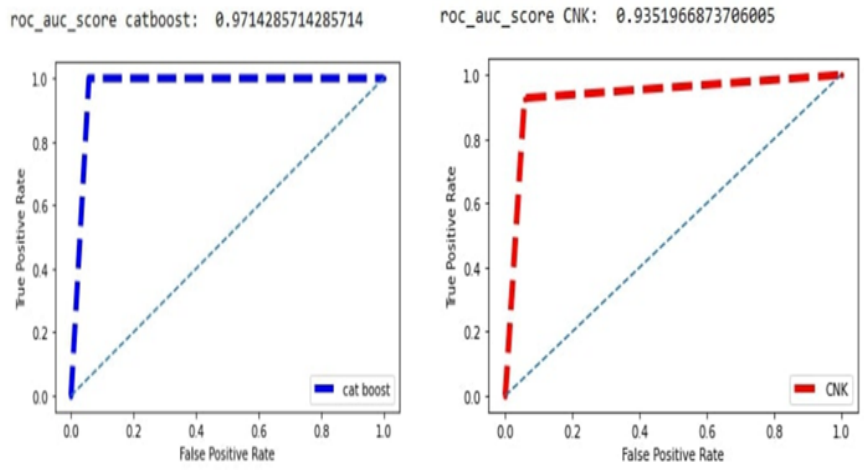
شکل ۶.۴ نمودار ریزش مو در برابر نتیجه آزمایش



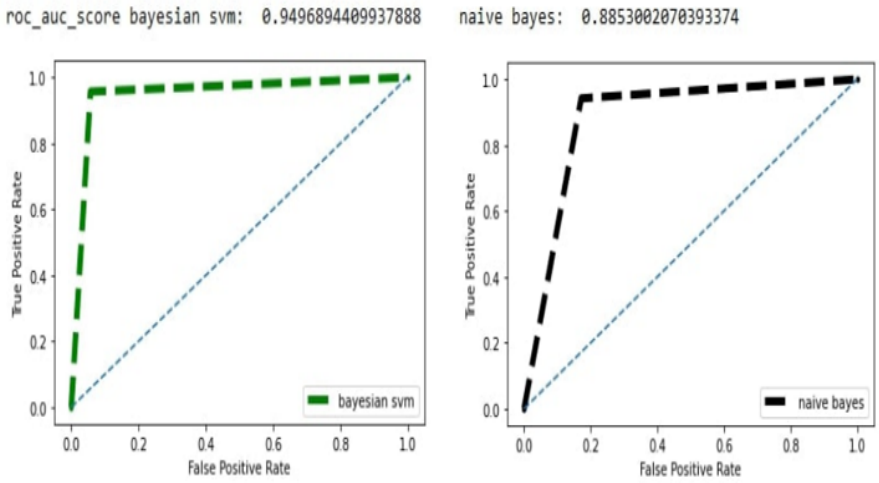
شکل ۷.۴ نمودار پرخوری در برابر نتیجه آزمایش

۵ بحث و نتیجه‌گیری

پس از انجام پیش‌پردازش داده‌ها با استفاده از متغیرهای ورودی و متغیر هدف به رده‌بندی داده‌ها پرداخته شد که دارای نتایجی بر روی مجموعه داده آزمایشی به شرح زیر شدند. لازم به ذکر است در ارزیابی الگوریتم‌ها داده‌های آموزشی و آزمایشی از یکدیگر جدا شده‌اند و ۷۰ درصد آن به‌عنوان مجموعه داده آموزشی و ۳۰ درصد آن به‌عنوان مجموعه داده آزمایشی در نظر گرفته‌شده‌اند. نتایج نهایی رده‌بندی در جدول ۲.۴ قابل‌مشاهده‌اند و می‌توان نتیجه گرفت با توجه به نتایج به‌دست‌آمده الگوریتم کت‌بوست برای رده‌بندی افراد جدید مناسب است و بهتر می‌توان نتیجه گرفت که فرد جدید به کدام‌یک از رده‌ها تعلق خواهد داشت. در زیر نمودار راک الگوریتم‌ها قابل‌مشاهده هستند.



شکل ۸.۴. نمودار راک CNK و کت‌بوست



شکل ۹.۴. نمودار راک بیز ساده و ماشین بردار پشتیبان بیزی

نتایج نهایی انواع روش‌های ارزیابی بر روی الگوریتم‌ها به‌صورت زیر هستند.

جدول ۲.۴. نتایج نهایی ارزیابی الگوریتم‌ها

precision	recall	F1-score	accuracy	انواع الگوریتم ها
99%	97%	98%	98%	کت‌بوست
94%	95%	95%	95%	ماشین بردار پشتیبان بیزی
89%	90%	89%	90%	بیز ساده
92%	94%	93%	93%	CNK

مراجع

[1] Berringer, R., Shibley, M. C., Cary, C. C., Pugh, C. B., Powers, P. A., Rafi, J. A. (1999). Outcomes of a community pharmacy-based diabetes monitoring program.*Journal of the American Pharmaceutical Association*, **39(6)**, 791-797.

[2] Ergün, Ö. N., İLHAN, H. O. (2021). Early stage diabetes prediction using machine learning methods. *Avrupa Bilim ve Teknoloji Dergisi*, (29), 52-57.

[3] Ferdousy, E. Z., Islam, M. M., Matin, M. A. (2013). Combination of naive bayes classifier and K-Nearest Neighbor (cNK) in the classification based predictive models.*Computer and information science*, **6(3)**, 48

[4] Choudhury, A.,, Gupta, D. (2019). A survey on medical diagnosis of diabetes using machine learning techniques. *In Recent Developments in Machine Learning and Data Analytics*: IC3 2018 (pp. 67-78). Springer Singapore.

[5] Vijayan, V. V.,, Anjali, C. (2015, December). Prediction and diagnosis of diabetes mellitus—A machine learning approach.*In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, (pp. 122-127). IEEE.

[6] Sarwar, A., Ali, M., Manhas, J.,, Sharma, V. (2020). Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. *International Journal of Information Technology*, 12, 419-428.

[7] Ayon, S. I.,, Islam, M. M. (2019). Diabetes prediction: a deep learning approach. *International Journal of Information Engineering and Electronic Business*, 12(2), 21.

[8] Kaur, H.,, Kumari, V. (2022). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied computing and informatics*, 18(1/2), 90-100.

[9] Lin, X., Xu, Y., Pan, X., Xu, J., Ding, Y., Sun, X., ..., Shan, P. F. (2020). Global, regional, and national burden and trend of diabetes in 195 countries and territories: an analysis from 1990 to 2025. *Scientific reports*, 10(1), 14790.

[10] Makrilakis, K., Liatis, S., Grammatikou, S., Perrea, D., Stathi, C., Tsiligros, P., Katsilambros, N. (2011). Validation of the Finnish diabetes risk score (FINDRISC) questionnaire for screening for undiagnosed type 2 diabetes, dysglycaemia and the metabolic syndrome in Greece. *Diabetes metabolism*, **37(2)**, 144-151.

[11] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

[12] Polson, N. G., Scott, S. L. (2011). *Data augmentation for support vector machines*., *Bayesian Analysis*, **6(1)**, 1-24.

[13] Reach, G., Pechtner, V., Gentilella, R., Corcos, A., Ceriello, A. (2017). Clinical inertia and its impact on treatment intensification in people with type 2 diabetes mellitus. *Diabetes metabolism*, **43(6)**, 501-511.

[14] Yang, Y., Webb, G. I. (2002, August). A comparative study of discretization methods for naive-bayes classifiers. In Proceedings of PKAW (Vol. 2002).

Diagnosing diabetes using Catboost and bayesian methods

Zahra Ahmadian[1] and Dr Farzad Eskandari[2]

Abstract:

Today, the diagnosis of diseases using artificial intelligence and machine learning algorithms is of paramount importance. Leveraging the available data in the field of study of a particular disease can yield valuable insights and results, ultimately reducing the occurrence of many fatalities. Among these diseases, diabetes diagnosis stands out, given its prevalence in modern urban life and the sedentary lifestyles of many individuals. Therefore, accurate identification of diabetes is crucial. In this article, a dataset containing information from 520 individuals who have undergone diabetes diagnosis tests is utilized. These individuals are categorized into two groups based on whether their diabetes test results are positive or negative. Bayesian classification methods, including Bayesian Support Vector Machine, Naive Bayes, CNK, and the CatBoost ensemble classification method, are employed to determine which of these methods exhibits superior data analysis capabilities. To compare these methods, metrics such as accuracy, precision, F1-score, recall, and ROC diagrams are utilized.

Keywords: classification, bayesian classification, ensemble classification.

[1] **Masters student**

[2] **Faculty of statistics department**