تشخیص زودهنگام سندرم داون با استفاده از الگوریتم‌های یادگیری ماشین

عبدالرضا سیاره[1]، سعیده عبدالله‌زاده[2]

[1]گروه علوم کامپیوتر و آمار، دانشگاه صنعتی خواجه نصیرالدین طوسی

[2]گروه ریاضی کاربردی، دانشگاه صنعتی خواجه نصیرالدین طوسی

چکیده: با پیشرفت فناوری‌های توالی‌یابی، آزمایش غیرتهاجمی NIPT توسعه یافته است و در غربالگری تریزومی ۲۱ از طریق تشخیص DNA جنین موجود در خون مادر، استفاده می‌شود. برای تحلیل داده‌های NIPT معمولاً از آزمون Z استفاده می‌شود. در روش‌های مورد استفاده برای تشخیص سندرم داون احتمال تشخیص اشتباه وجود دارد. بنابراین ارائهٔ روشی که بتواند در کنار روش‌های تشخیصی به‌کار برده شود و کارایی این روش‌ها را بهبود بخشد؛ ضروری است. هدف اصلی این مقاله طراحی مدلی بر اساس الگوریتم‌های یادگیری ماشین برای تشخیص زودهنگام سندرم داون است؛ به‌طوری که بتوان از این روش‌ها برای افزایش دقت تشخیص استفاده کرد. در این مقاله به بهبود روش‌های تشخیصی به کمک الگوریتم‌های یادگیری ماشین مانند: ماشین بردار پشتیبان، بیز ساده، درخت تصمیم، جنگل تصادفی و نزدیکترین همسایه برای بررسی یک مجموعه دادهٔ مربوط به سندرم داون پرداخته شده است. عملکرد هر یک از مدل‌ها در مجموعه دادهٔ سندرم داون بررسی و در نهایت مناسب‌ترین مدل برای این هدف معرفی شده است. نتایج نشان می‌دهند که این الگوریتم‌ها دقت بسیار مناسبی در تشخیص این بیماری دارند.

واژه‌های کلیدی: یادگیری ماشین، سندرم داون، بیز ساده، ماشین بردار پشتیبان، درخت تصمیم.

آدرس الکترونیکی نویسنده مسئول مقاله: عبدالرضا سیاره، asayyareh@kntu.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): 62P10, 68W20.

۱ مقدمه

سندرم داون نوعی اختلال ژنتیکی ناشی از وجود نسخهٔ اضافی از کروموزم ۲۱ است. این اختلال تحت عنوان تریزومی ۲۱ نیز شناخته می‌شود. افراد مبتلا به این سندرم، دارای سه کروموزوم ۲۱ به‌جای دو کروموزوم هستند. بر اساس مطالعات موریس و البرمن (۲۰۰۹) تقریباً از هر ۸۰۰ جنین در سراسر جهان یک جنین مبتلا به سندرم داون است. تشخیص زودهنگام این بیماری به مدیریت مناسب و ارائهٔ مداخلات درمانی به‌موقع این جنین‌ها کمک می‌کند. در اغلب موارد، سندرم داون در نتیجهٔ خطاهایی در فرایند تقسیم سلولی رخ می‌دهد. در فرایند میوز که نوعی تقسیم سلولی است؛ کروموزوم‌های یک جفت کروموزوم مشابه به‌طور طبیعی در اولین تقسیم میوزی (میوز اول) از یکدیگر جدا می‌شوند. اما گاهی این جدایی رخ نمی‌دهد و هر دو کروموزوم متعلق به یک جفت کروموزوم به داخل یک سلول واحد می‌روند و در نتیجهٔ این عدم جدایی کروموزوم‌ها، یک سلول به‌جای دریافت ۲۳ کروموزوم تعداد ۲۴ کروموزوم و سلول دیگر ۲۲ کروموزوم دریافت می‌کند. اگر در هنگام لقاح، سلول دارای ۲۳ کروموزوم با سلول دارای ۲۴ یا ۲۲ کروموزوم یکی شود؛ فردی با ۴۷ کروموزوم (تریزومی) یا ۴۵ کروموزوم (مونوزومی) متولد می‌شود. اگر عدم جدایی مربوط به کروموزوم ۲۱ باشد؛ منجربه تریزومی ۲۱ یا سندرم داون خواهد شد. این افراد دارای ویژگی‌های فیزیکی مشخص مانند پوست بدن با نقص ارتجاعی، صورت بزرگ نسبت به بدن، چهره‌ای با بینی کوچکتر و سر بزرگتر از حد معمول و همچنین ضریب هوشی پایین هستند.

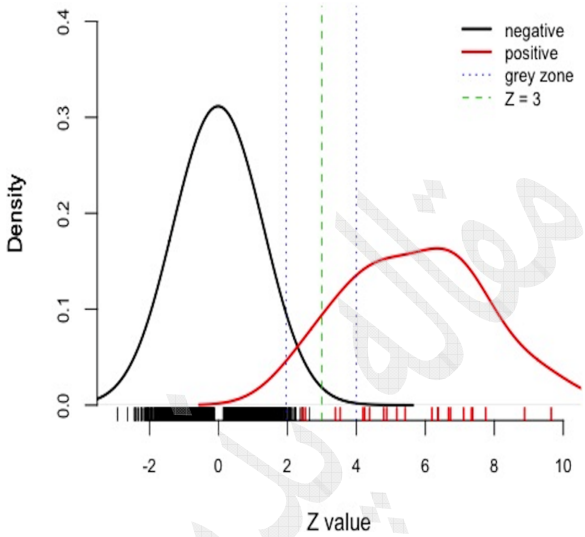
روش‌های متفاوتی برای تشخیص سندرم داون وجود دارد. آزمایش‌های آمنیوسنتز یا نمونه‌برداری از پرزهای کوریونی (CVS)، استانداردهای طلایی برای تشخیص ناهنجاری‌های کروموزومی هستند. اما این روش‌ها تهاجمی بوده و خطر عفونت و آسیب به جنین در طول معاینه وجود دارد. روش‌های غیرتهاجمی مختلفی نیز وجود دارد که برخی قدرت تشخیص مناسبی نداشته و به‌علاوه در تشخیص بیماری نتایج مثبت و منفی کاذب مشاهده می‌شود. بنابراین، علاقهٔ فزاینده‌ای به توسعهٔ روش‌های خودکار و مقرون به‌صرفه برای تشخیص زودهنگام سندرم داون با استفاده از الگوریتم‌های یادگیری ماشین وجود دارد. آشور و همکاران (۲۰۱۲) نوع دیگری از آزمایش‌های غیرتهاجمی پیش از تولد تحت عنوان آزمایش NIPT[1] را مطالعه کردند. تست NIPT یا تست غربالگری غیرتهاجمی که به آن cell free fetal DNA

[1] Non-Invasive Prenatal Testing

نیز می‌گویند به تجزیه و تحلیل DNA آزاد جنینی در گردش خون مادر می‌پردازد. این روش فقط با گرفتن مقداری از خون مادر امکان‌پذیر است و نیازی به روش‌های تهاجمی مانند آمنیوسنتز یا گرفتن پرزهای کوریونی نیست. در حالت طبیعی سلول‌ها به‌طور نرمال ۲۳ جفت کرموزوم دارند و هر کروموزوم از رشته‌های DNA حلقوی تشکیل شده است که حاوی هزاران ژن است. در هر کروموزوم سکانس‌ها و یا توالی خاصی از مولکول DNA وجود دارد که ویژهٔ هر کروموزوم است. در این آزمایش با شمارش تعداد این توالی‌ها، مقدار دقیق هر کروموزوم محاسبه می‌شود و در نتیجه می‌توان مجموع کل کروموزوم‌های ۲۱ موجود در خون مادر را نسبت به سایر کروموزوم‌ها مقایسه کرد. اگر جنین مبتلا به تریزومی ۲۱ باشد؛ به میزان خیلی کم توالی‌های مربوط به کروموزوم ۲۱، نسبت به سایر کروموزوم‌ها بیشتر است.

در مطالعات مختلف، گزارش‌های متفاوتی از قدرت تشخیص این تست بیان شده است. به‌طور مثال در مطالعهٔ لیم و همکاران (۲۰۱۲) قدرت تشخیص این آزمایش تقریباً ۸۳ درصد بیان شده است. چیو وهمکاران (۲۰۰۸) پیشنهاد کردند از آنجایی که جنین دارای نسخهٔ اضافی از یک کروموزوم ۲ تا ۵ درصد cffDNA بیشتری نسبت به جنین نرمال دارد؛ آزمون Z اغلب در اندازه‌گیری انحراف استفاده شود. در واقع در آزمایش NIPT، امتیاز Z برای نشان دادن و ارزیابی تفاوت میانگین اطلاعات ژنتیکی یک نمونه با میانگین جمعیت مرجع مورد استفاده قرار می‌گیرد که در بخش سوم مقاله نحوهٔ محاسبهٔ آن به‌طور کامل توضیح داده می‌شود. با توجه به استاندارهای تأیید شدهٔ CFDA (سازمان کنترل غذا و دارو در چین) امتیاز $Z = ۳$ در NIPT حد فاصل تشخیص نمونه‌های منفی و مثبت از نظر ابتلا به سندرم داون است (لیائو و همکاران، ۲۰۱۴). از این رو در NIPT معمولی، داده‌های دارای $Z > ۳$ نمونه‌های سندرم داون مثبت در نظر گرفته می‌شوند. به این معنی که به‌طور قابل‌توجهی از خط پایهٔ مجموعه دادهٔ مرجع انحراف دارند. در حالی که نمونه‌های با $Z < ۳$ جنین با سندرم داون منفی در نظر گرفته می‌شوند. اما این خط نمی‌تواند مرز بین داده‌های مثبت و منفی باشد. زیرا برخی داده‌ها مثبت پیش‌بینی می‌شوند درحالی که نتیجهٔ واقعی منفی است؛ این داده‌ها مثبت کاذب هستند و مواردی نیز به‌عنوان منفی پیش‌بینی می‌شوند ولی در واقع مثبت بوده‌اند. این نوع داده‌ها، منفی کاذب هستند. این تشخیص اشتباه می‌تواند منجربه از بین رفتن جنین سالم و یا تولد جنینی مبتلا به سندرم داون شود. علاوه بر این، یانگ و همکاران (۲۰۱۸) نشان دادند که کسر جنینی در تعیین تریزومی ۲۱ بسیار مهم است درحالی که در رویکرد مبتنی بر آزمون Z

در تحلیل داده‌های NIPT دخالت ندارد. این مشکلات می‌توانند منجر به پیش‌بینی نادرست، هزینهٔ زیاد آزمایش مجدد و تأخیر در درمان شوند.



شکل ۱: NIPT مبتنی بر آزمون Z

در شکل ۱ نمونه‌های منفی و مثبت به‌ترتیب با رنگ‌های مشکی و قرمز نشان داده شده‌اند. خط سبز نشان‌دهندهٔ برش $Z = ۳$ است که اغلب به‌عنوان معیاری در تمایز استفاده می‌شود. خط آبی فاصلهٔ منطقهٔ خاکستری بین $Z = ۱/۹۶$ و $Z = ۴$ را نشان می‌دهد که به معنی شکست در تشخیص با استفاده از مقدار $Z = ۳$ است و نیاز به آزمایش مجدد دارد. یانگ و همکاران (۲۰۱۸) نتیجه گرفتند که هر نمونه با نمرهٔ Z در بازهٔ (۴ ، ۱/۹۶)، به نام "منطقهٔ خاکستری"، نیاز به آزمایش مجدد دارد. زیرا استفاده از $Z = ۳$ به‌عنوان خطی برای جداسازی مقادیر مثبت از مقادیر منفی ممکن است منجربه نتایج نادرست شود. بنابراین، توسعهٔ روشی دقیق‌تر برای تحلیل داده‌های NIPT ضروری است. با توجه به اهمیت تشخیص زودهنگام سندرم داون و مزایای استفاده از الگوریتم‌های یادگیری ماشین، تحقیقاتی در این زمینه انجام شده است. این تحقیقات نشان داده‌اند؛ با استفاده از داده‌های بالینی و آزمایشگاهی، عملکرد الگوریتم‌های یادگیری ماشین در تشخیص زودهنگام سندرم داون، بهبود چشمگیری داشته است.

مطالعاتی در زمینهٔ تشخیص سندرم داون با استفاده از الگوریتم‌های یادگیری ماشین انجام شده است.

به‌طور مثال، راماناتان و همکاران (۲۰۱۸) در مقاله‌ای با عنوان تعیین احتمالی سندرم داون با استفاده از تکنیک‌های یادگیری ماشین، برخی از الگوریتم‌های یادگیری ماشین مانند الگوریتم بیزساده، درخت تصمیم، جنگل تصادفی و شبکهٔ عصبی مصنوعی را برای تشخیص زودهنگام سندرم داون بررسی کردند. هدف این مقاله، ساخت مدلی است که قادر باشد به‌عنوان ابزاری برای کمک به متخصصان زنان استفاده شود و به کمک آن بتوان احتمال ابتلا به سندرم داون را بررسی کرد. همچنین هی و همکاران (۲۰۲۱)، در مقاله‌ای تحت عنوان مدل یادگیری ماشینی برای پیش‌بینی سندرم داون در غربالگری سه ماههٔ دوم بارداری، مجموعه داده‌ای که شامل ۵۸۹۷۲ زن باردار ازجمله ۴۹ مورد سندرم داون که تحت غربالگری سندرم داون درسه ماههٔ دوم قرار گرفته بودند را مورد بررسی قرار دادند. چندین مدل یادگیری ماشین مانند درخت تصمیم، تقویت گرادیان، شبکه عصبی، ماشین بردار پشتیبان و جنگل تصادفی مورد بررسی قرار گرفته است و طبق نتایج، مدل جنگل تصادفی عملکرد خوبی برای این نوع مجموعه داده از خود نشان داده است.

مطالعه‌ای تحت عنوان کاربرد الگوریتم‌های هوشمند در غربالگری سندرم داون در سه ماههٔ دوم بارداری توسط ژانگ و همکاران (۲۰۲۱) انجام شد. در این مطالعه از ماشین بردار پشتیبان، الگوریتم‌های طبقه‌بندی و درخت رگرسیون (CART) برای ساخت مدل طبقه‌بندی برای غربالگری سندرم داون استفاده شده است. نتایج نشان می‌دهند که الگوریتم‌های ماشین بردار پشتیبان و CART به نتایج خوبی در مجموعه دادهٔ غربالگری سندرم داون دست یافته‌اند. همچنین یانگ و همکاران (۲۰۱۸) در مقاله‌ای، عملکرد الگوریتم‌های یادگیری ماشین روی داده‌های NIPT برای پیش‌بینی تریزومی‌های ۱۳، ۱۸ و ۲۱ را بررسی کردند. چندین الگوریتم یادگیری ماشین مانند ماشین بردار پشتیبان و درخت تصمیم بر روی همان مجموعه دادهٔ مورد نظر در این مطالعه آزمایش شدند. طبق نتایج، مدل ماشین بردار پشتیبان در مجموعهٔ داده‌های این مطالعه نسبتاً بهتر عمل کرد. همچنین موسوی و گلعلی‌زاده (۱۴۰۲) از الگوریتم ماشین بردار پشتیبان برای یافتن مجموعه‌ای از ژن‌های موثر در بروز سرطان پروستات، به‌منظور افزایش دقت رده‌بندی استفاده کردند. به‌طور کلی توسعهٔ روشی که بتواند؛ عملکرد مطلوبی در تشخیص سندرم داون داشته باشد؛ پراهمیت است. در این مقاله روشی مبتنی بر یادگیری ماشین مورد بررسی قرار می‌گیرد و نتایج آن در تشخیص سندرم داون بررسی می‌شود؛ به‌طوری‌که روش‌های یادگیری ماشین در راستای کاهش خطر تولد نوزادان بیمار و تصحیح نتایج اشتباه در سایر روش‌ها مورد استفاده قرار می‌گیرند تا بتوانند نرخ تشخیص

بهتری داشته باشند. این روش‌ها خواه ناخواه از روش‌های متداول آماری بهره می‌گیرند. مدل‌های یادگیری ماشین می‌توانند در کنار روش‌های‌های آماری و یا به‌طور مستقل مورد استفاده قرار گرفته و دقت تشخیص بهبود یابد و تا حد امکان خطا در تشخیص بیماری یا عدم تشخیص بیماری به حداقل برسد. بدیهی است که کاهش میزان خطا و مقایسهٔ روش‌های یادگیری ماشین و روش‌های متداول بستگی به نوع روش‌های به کار گرفته شدهٔ تشخیصی، آماری و یادگیری ماشین دارد.

در بخش دوم مقاله به بررسی الگوریتم‌های یادگیری ماشین و همچنین معیارهای ارزیابی پرداخته می‌شود. در بخش سوم نمونه‌های مورد استفاده در مطالعه و ویژگی‌های آن بررسی می‌شود و در نهایت نتایج پیاده‌سازی برخی الگوریتم‌های یادگیری ماشین و عملکرد هر یک گزارش می‌شود. در بخش چهارم نتایج نهایی مدل‌های یادگیری ماشین و مقایسه با روش آماری موجود، آورده شده است.

۲ برخی الگوریتم‌های یادگیری ماشین

برای مجموعه‌ای از داده‌ها همراه با برچسب آن‌ها، فرض می‌شود داده‌های آموزشی به‌صورت $X = \{x_1, ..., x_n\}$ است، که در آن $x_i \in R^d$ با برچسب $Y = \{y_1, ..., y_n\}$ و $y_i \in \{+1, -1\}$ و $D = \{(x_1, y_1), ..., (x_n, y_n)\}$ مجموعه دادهٔ آموزشی[۲] همراه با برچسب هر داده است.

۱.۲ الگوریتم ماشین بردار پشتیبان

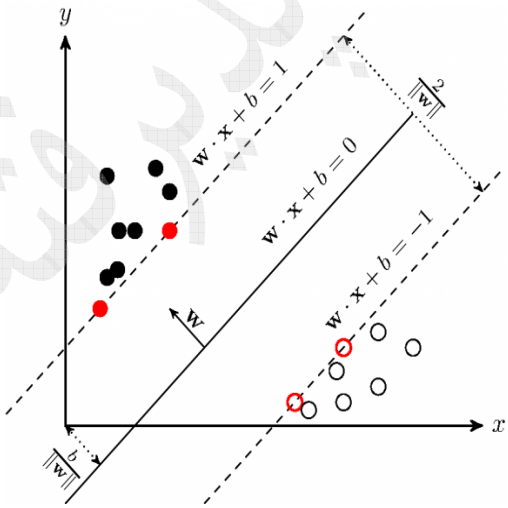
الگوریتم ماشین بردار پشتیبان[۳] یکی(SVM) از الگوریتم‌های شناخته شده تحت نظارت است. در ساده‌ترین حالت این الگوریتم یعنی SVM خطی، به‌دنبال جداسازی نمونه‌های مثبت و منفی با یک ابرصفحه است؛ به‌طوری که حداکثرحاشیه[۴] را داشته باشد. فرض می‌شود؛ یک ابرصفحه موجود است که قادر به جداسازی

[2] Train Data
[3] Support Vector Machine
[4] Marjin

نمونه‌های مثبت و منفی از یکدیگر است. در حالت کلی برای جدا کننده از نوع ابرصفحه

$$\sum_i w_i x_i + b = ۰ \tag{۱}$$

است، که در آن w بردار نرمال و b فاصلهٔ این ابرصفحه از مبدأ است که با تغییر آن موقعیت ابرصفحه در فضای ویژگی‌ها تغییر می‌کند (شالو-شوارتز و بن، ۲۰۱۴). نزدیکترین نقاط به ابرصفحهٔ (۱) بر روی صفحه‌های $wx + b = -۱$ و $wx + b = +۱$ قرار دارند. در واقع به فرض تفکیک‌پذیر بودن داده‌های مثبت و منفی، بردارهای مرزی به‌عبارتی نقاطی که در نزدیک‌ترین فاصله به ابرصفحه قرار گرفته‌اند؛ روی ابرصفحه‌های ذکر شده قرار خواهند گرفت. ادعا می‌شود؛ این نقاط روی دو صفحه در دو طرف ابرصفحهٔ هدف قرار دارند. به ناحیهٔ بین این دو ابرصفحه، حاشیه گفته می‌شود. بنابراین فاصلهٔ بین دو ابر صفحهٔ



شکل ۲: نمونه‌ای از یک مسئلهٔ قابل تفکیک در فضای دوبعدی.

مورد نظر از رابطهٔ $\frac{۲}{w^T w}$ به‌دست می‌آید؛ (داوم، ۲۰۱۷). برای مسئلهٔ SVM داریم

$$\max_{w,b} \frac{۲}{\|w\|^۲} = \min_{w,b} \frac{۱}{۲}\|w\|^۲ y_i(w^T x_i + b) \geq ۱; \quad (x_i, y_i) \in D \tag{۲}$$

مسئلهٔ (۲) یک مسئلهٔ بهینه‌سازی مقید است و برای حل مسئله، ضرایب لاگرانژ مثبت $\alpha_i \geq ۰$ در قیود ضرب و از تابع هدف کسر می‌شود،

$$L_p = \frac{۱}{۲}\|w\|^۲ - \sum_i \alpha_i(y_i(\vec{w}\vec{x_i} + b)) - ۱ \qquad (۳)$$

شرایط kkt[۵]، شرایط لازم و کافی برای داشتن پاسخ بهینه برای معادلات مقید را بیان می‌دارد و طبق این شرایط باید مشتق تابع نسبت به متغیرها برابر با صفر باشد. این شرایط روی مسئله (۳) اعمال شده، از معادله نسبت به w و b مشتق گرفته و مساوی صفر قرار داده می‌شود. پس از حل معادلهٔ (۳) و پیدا کردن ضرایب لاگرانژ خواهیم داشت،

$$\max_{w,b} -\frac{۱}{۲}\sum_i\sum_j \alpha_i\alpha_j y_i y_j x_i.x_j + \sum_i \alpha_i \qquad (۴)$$

که در آن $\sum_{i,j}\alpha_i\alpha_j = ۰$، $۰ \leq \alpha_i \leq C$ و $w = \sum_i \alpha_i y_i x_i$ است.

برای داده‌هایی که به‌صورت خطی جدایی‌پذیر نیستند؛ یک حاشیهٔ نرم[۶] تعریف می‌شود. در این حالت یک مقدار ϵ در نظر گرفته می‌شود که نشان‌دهندهٔ خطا است. $\sum_i \epsilon_i$ به‌عنوان کران بالا برای تعداد خطای داده‌های آموزشی در نظر گرفته شده و مسئله به صورت

$$L_p = \frac{۱}{۲}\|w\|^۲ + C(\sum_i \epsilon_i) - \sum_i (y_i(\vec{w}\vec{x} + b) - ۱ + \epsilon_i - \sum_i \mu_i\epsilon_i$$

بازنویسی می‌شود، که در آن μ_i ها ضرائب لاگرانژ هستند که سبب مثبت شدن ϵ_i می‌شوند و C یک پارامتر تنظیم است. در حالتی که داده‌ها به‌هیچ‌وجه، به‌صورت خطی از یکدیگر تفکیک نشوند؛ می‌توان داده‌ها را به یک فضای دیگر با بُعد بالاتر تصویر کرد. بنابراین تابع $\varphi : R^d \rightarrow H$ داده‌ها را به فضای H که بُعدی بالاتر از d دارد می‌نگارد. تصویر نقاط x_i و x_j در فضای جدید به‌صورت $\varphi(x_i)$

[5] Karush-Kuhn-Tucker
[6] Soft Margin

و $\varphi(x_j)$ خواهد بود و ضرب داخلی آن‌ها در فضای H به‌شکل $\varphi(x_i).\varphi(x_j)$ است که برای تشکیل ابرصفحه، به این ضرب داخلی نیازاست (واپنیک، ۱۹۹۵). پس از بازنویسی مسئله با تابع هسته رابطهٔ $\max -\frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j k(\vec{x_i}\vec{x_j}) + \sum_i \alpha_i$ را خواهیم داشت که در آن $\sum_{i,j}\alpha_i\alpha_j = 0$ و $0 \leq \alpha_i \leq C$. تابع هسته انواع متفاوتی مانند هسته خطی $k(x,y) = x.y$ و هسته گاوسی تابع پایهٔ شعاعیRBF[7] به صورت $k(x,y) = \exp\left(-\sigma\|x-y\|^2\right)$ است.

۲.۲ درخت تصمیم

الگوریتم درخت تصمیم[8]، یکی ازمتداول‌ترین الگوریتم‌های یادگیری ماشین برای مسائل طبقه‌بندی و پیش‌بینی است (شالو-شوارتز و بن، ۲۰۱۴). سوتاهاران (۲۰۱۶) بیان کرد؛ درخت تصمیم، به‌شکل درخت وارونه است که شامل گرهٔ ریشه[9]، شاخه‌ها و گره‌های برگ[10] است؛ به‌طوری‌که یک ریشه در بالای آن وجود دارد. به‌عبارتی بالاترین گرهٔ درخت ریشه است و برگ‌های آن در پایین هستند. هر گرهٔ برگ نشان‌دهندهٔ یک برچسب کلاس است. درختان تصمیم، نمونه‌ها را با مرتب کردن آن‌ها در درخت از گرهٔ ریشه به‌سمت گره‌های برگ، کلاس‌بندی می‌کنند. در ابتدا کل مجموعهٔ آموزشی به‌عنوان ریشه در نظر گرفته می‌شود. هر نمونه با شروع از گرهٔ ریشه و آزمایش ویژگی مشخص‌شده؛ توسط این گره و حرکت در شاخهٔ متناظر با مقدار ویژگی داده‌شده در نمونه، کلاس‌بندی می‌شود. این فرایند برای هر زیردرختی که گرهٔ جدید، ریشهٔ آن است؛ تکرار می‌شود. الگوریتم درخت تصمیم برای هر ویژگی در مجموعه داده، یک گره تشکیل می‌دهد. مهم‌ترین ویژگی در گرهٔ ریشه قرار می‌گیرد. معیارهای مختلفی برای انتخاب ویژگی‌هایی وجود دارد که می‌توانند به‌عنوان گرهٔ ریشه در نظر گرفته شوند. یکی از این معیارها کسب اطلاعات است، که با استفاده از آن سعی می‌شود اطلاعات موجود در هر ویژگی تخمین زده شود و از آنتروپی در این روش استفاده می‌شود. آنتروپی، ناخالصی در مجموعه داده را اندازه گیری می‌کند. از آن‌جایی که آنتروپی در واقع درجهٔ هرج و مرج (یا عدم قطعیت) در سیستم است؛ کاهش آنتروپی را کسب اطلاعات می‌نامند. آنتروپی

[7] Gaussian Radial Basis Function
[8] Decision tree
[9] Root Node
[10] Leaf

مجموعه دادهٔ S از رابطهٔ

$$\text{Entropy}(S) = \sum_{i=۱}^{C} -p_i \log ۲(p_i) \tag{۵}$$

به‌دست می‌آید، که در آن C تعداد کلاس‌ها و p_i احتمال مرتبط با کلاس iاُم است.

کسب اطلاعات در واقع تفاوت بین آنتروپی قبل از تقسیم و میانگین آنتروپی پس از تقسیم مجموعه داده را بر اساس مقادیر مشخصهٔ داده‌شده محاسبه می‌کند. با داشتن بی‌نظمی به‌عنوان ناخالصی در یک مجموعه مثال آموزشی، می‌توان معیار مؤثر بودن یک صفت در کلاس‌بندی داده‌های آموزشی را تعریف نمود. نفع اطلاعاتی یک صفت A مربوط به مجموعه داده‌های S با رابطهٔ

$$Gain(S, A) = \text{Entropy}(S) - \sum_{v \in Val(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \tag{۶}$$

تعریف می‌شود، که در آن $Val(A)$ مجموعهٔ مقادیر برای ویژگی A و $S_v = \{s \in S; A(s) = v\}$ است. جملهٔ اول در رابطهٔ (۶) بی‌نظمی(آنتروپی) درمجموعهٔ اصلی، یعنی کل داده‌ها و جملهٔ دوم مقدار موردانتظار بی‌نظمی بعد از تفکیک مجموعه داده با استفاده از صفت A است. رابطهٔ (۶) برای تمام ویژگی‌ها محاسبه می‌شود و متغیری که بیشترین مقدار $Gain$ را داشته باشد؛ به‌عنوان متغیر شاخه‌بند انتخاب می‌شود. الگوریتم درخت تصمیم ID3[۱۱] از آنتروپی برای محاسبهٔ بهرهٔ اطلاعات استفاده می‌کند.

۳.۲ الگوریتم جنگل تصادفی

الگوریتم جنگل تصادفی[۱۲] یکی از الگوریتم‌های یادگیری ماشین تحت نظارت مبتنی بر یادگیری گروهی است که بر اساس ترکیب تعدادی درخت تصمیم[۱۳] که از انتخاب تصادفی نمونه‌های آموزشی ساخته می‌شوند؛ عمل می‌کند. بریمن (۲۰۰۱) بیان می‌کند در این الگوریتم، چندین درخت تصمیم به‌صورت موازی و مستقل

[11] Iterative Dichotomiser
[12] Random Forest
[13] Decision Trees

از هم ساخته می‌شوند. سپس نتایج آن‌ها ترکیب می‌شود تا به یک پاسخ نهایی برسد. در نتیجه جنگلی از درختان ایجاد می‌شود و از این رو جنگل تصادفی نامیده می‌شود.

۴.۲ الگوریتم kنزدیکترین همسایه

الگوریتم kنزدیک‌ترین همسایه[۱۴] (kNN) یکی از پرکاربردترین الگوریتم‌های تحت نظارت یادگیری ماشین برای کلاس‌بندی داده‌ها است. ایدهٔ اصلی الگوریتم این است که مجموعهٔ آموزشی را به‌خاطر سپرده و سپس برچسب هر نمونهٔ جدید را بر اساس برچسب‌های نزدیک‌ترین همسایگان آن در مجموعهٔ آموزشی پیش‌بینی کند. روش کار این الگوریتم در ادامه آورده شده است (کرامر ، ۲۰۱۳). اگر x_q، نمونهٔ جدید باشد که باید کلاس‌بندی شده و برچسب آن مشخص گردد؛ الگوریتم kNN فاصلهٔ بین یک نقطه و سایر نقاط دادهٔ آموزشی را محاسبه می‌کند. در این الگوریتم، k تعداد نزدیک‌ترین همسایگان به نقطهٔ مورد نظر است. برای کلاس‌بندی نقطهٔ x_q، ابتدا باید نقاطی را که در نزدیکترین فاصله به x_q قرار دارند را یافت. برای این منظور، فاصلهٔ نقطهٔ مورد نظر از تمام نقاط مجموعهٔ دادهٔ آموزشی محاسبه می‌شود و به‌ترتیب از کوچک به بزرگ مرتب می‌گردد. پس از در نظر گرفتن k نقطه که کمترین فاصله از x_q را دارند و برچسب اکثریت داده‌ها، همان برچسب x_q خواهد بود؛ برای محاسبهٔ فاصله از نقاط مجموعهٔ آموزشی، از فاصلهٔ اقلیدسی، منهتن و یا هر نوع مقیاس دیگر استفاده می‌شود. فاصلهٔ اقلیدسی و همچنین فاصلهٔ منهتن حالت‌های خاصی از فاصلهٔ مینکوفسکی هستند که از رابطهٔ، $d(x,y) = (\sum_{i=1}^{n}(|x_i - y_i|)^p)^{\frac{1}{p}}$ محاسبه می‌شود.

۵.۲ بیز ساده

الگوریتم بیزساده[۱۵] با ترکیب نظریهٔ بیز و استقلال ویژگی‌ها به طبقه‌بندی می‌پردازد. فرض می‌شود که داده‌ها دارای تعدادی ویژگی (ستون‌های مجموعه داده‌ها) هستند. الگوریتم بیزساده بر این فرضیه استوار است که

[14] k-Nearest Neighbors Classifier
[15] Naive Bayes

این ویژگی‌ها مستقل از یکدیگر هستند و به‌صورت

$$P[Y|\ X] = \frac{P[X|\ Y]P[Y]}{P[X]} \tag{۷}$$

تعریف می‌شود (شالو-شوارتز و بن، ۲۰۱۴). در این الگوریتم دادهٔ $\tilde{x} \in X$ در نظر گرفته می‌شود و هدف پیش‌بینی برچسب آن است. طبق الگوریتم بیزساده، احتمال متعلق بودن $\tilde{x}$ به هر یک از کلاس‌ها محاسبه می‌شود. تخمین توزیع P با توجه به داده‌های مجموعهٔ آموزشی D. فرض کنید؛ $\tilde{x} \in X$ یکی از داده‌ها و مقادیر هدف به‌صورت $\tilde{y} \in \{+۱, -۱\}$ باشد؛ هدف محاسبهٔ احتمال شرطی $P[Y = \tilde{y}|X = \tilde{x}]$ است. تابع هدف به‌صورت، $\hat{y} = \arg\max_{\tilde{y}}[Y = \tilde{y}|X = \tilde{x}]; \quad \tilde{y} \in \{+۱, -۱\}$ تعریف می‌شود. با توجه به تعریف احتمال شرطی خواهیم داشت،

$$P[Y = \tilde{y}|X = \tilde{x}] = \frac{P[X = \tilde{x}, Y = \tilde{y}]}{P[X = \tilde{x}]} \tag{۸}$$

رابطهٔ (۸) بدین معنا است که دادهٔ مورد نظر $\tilde{x}$ و دقیقاً با همان ویژگی‌ها باید داخل مجموعه داده‌ها موجود باشد. طبق صورت کسر، دادهٔ $\tilde{x}$ با برچسب $\tilde{y}$ باید درون مجموعه داده‌ها موجود باشد و احتمال وجود دادهٔ مورد نظر، داخل مجموعه داده‌ها بسیار کم است؛ مگر این‌که حجم داده‌ها بسیار بزرگ باشد. بنابراین رابطهٔ فوق تقریباً همیشه صفر است. پس

$$\hat{y} = \arg\max_{\tilde{y}} \frac{P[X = \tilde{x}|Y = \tilde{y}].P[Y = \tilde{y}]}{P[X = \tilde{x}]} \tag{۹}$$

که در آن $\hat{y}$ برچسب پیش‌بینی شده است. برای سهولت در محاسبات، مخرج رابطهٔ (۹) یعنی توزیع حاشیه‌ای درنظر گرفته نمی‌شود. بنابراین رابطهٔ (۹) به صورت

$$\hat{y} = \arg\max_{\tilde{y}} P[X = \tilde{x}|Y = \tilde{y}].p[Y = \tilde{y}] \tag{۱۰}$$

در نظر گرفته می‌شود. می‌دانیم $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_d)$ و تمامی ویژگی‌ها مستقل از یکدیگر در نظر گرفته می‌شوند؛ بنابراین رابطهٔ (۱۱) را به‌صورت

$$\hat{y} = \arg\max_{\tilde{y}} \Pi_{i=1} P[x_i = \tilde{x}_i | Y = \tilde{y}] P[Y = \tilde{y}] \qquad (۱۱)$$

داریم. تمام مقادیری که در رابطهٔ (۱۱) در هم ضرب می‌شوند؛ کسری و بسیار کوچک هستند؛ بنابراین ممکن است حاصلضرب آن‌ها نیز عددی بسیار کوچک و نزدیک به صفر شود. برای حل این مشکل از آن لگاریتم گرفته می‌شود. در نهایت داریم

$$\hat{y} = \arg\max_{\tilde{y}} \quad \exp(\log P[Y = \tilde{y}] + \sum_{i=1}^{d} \log P[x_i = \tilde{x}_i | Y = \tilde{y}]). \qquad (۱۲)$$

۶.۲ ارزیابی مدل‌های یادگیری ماشین

ارزیابی الگوریتم‌های تحت نظارت یادگیری ماشین با استفاده از روش‌های مختلف صورت می‌گیرد. برخی از مهمترین روش‌ها شامل ماتریس در هم‌ریختگی، دقت، حساسیت و دقت پیش‌بینی است. ماتریس درهم‌ریختگی[۱۶] به‌صورت یک جدول نمایش داده می‌شود. اگر تعداد دسته‌ها برابر با m باشد؛ ماتریس درهم‌ریختگی یک ماتریس مربعی m*m است که سطرها(یا ستون‌ها) نشان‌دهندهٔ پیش‌بینی مدل و در مقابل ستون‌ها(سطرها) نشان‌دهندهٔ برچسب واقعی نمونه‌ها است. در این ماتریس FN به تعداد نمونه‌هایی که به اشتباه منفی پیش‌بینی شده‌اند؛ گفته می‌شود. همچنین تعداد نمونه‌هایی که به اشتباه مثبت پیش‌بینی شده‌اند با FP، تعداد نمونه‌هایی که به‌درستی منفی پیش‌بینی شده‌اند با TN و تعداد نمونه‌هایی که به‌درستی مثبت پیش‌بینی شده‌اند با TP نشان داده می‌شوند. همچنین دقت یک مدل برابر با نسبت تعداد مواردی که به‌درستی پیش‌بینی شده‌اند؛ به تعداد کل نمونه‌ها است و به‌عنوان معیاری کلی برای ارزیابی عملکرد مدل مورد استفاده قرار می‌گیرد. دقت از رابطهٔ $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ محاسبه می‌شود. معیار دیگر حساسیت است که نسبت مواردی که به‌درستی مثبت پیش‌بینی شده‌اند(TP) به مجموع تعداد

[16] Confusion Matrix

کل موارد مثبت واقعی را نشان می‌دهد و از رابطهٔ، $Recall = \frac{TP}{TP+FN}$ به دست می‌آید. به‌علاوه دقت پیش‌بینی، نسبت مواردی که به‌درستی مثبت پیش‌بینی شده‌اند (TP) به مجموع تمام نمونه‌هایی که مثبت پیش‌بینی شده‌اند را نشان می‌دهد و از رابطهٔ $Precision = \frac{TP}{TP+FP}$ به دست می‌آید. یکی دیگر از معیارهای پرکاربرد، $F۱Score = \frac{۲\times Precision\times Recall}{Precision+Recall}$ است که میانگین هارمونیک دقت و حساسیت را نشان می‌دهد و برای مواقعی که توازن بین دقت و حساسیت اهمیت دارد؛ مناسب است.

۳ مطالعهٔ داده‌های واقعی

در این بخش تحلیل داده‌های NIPT NGS به‌دست آمده از مارس تا ژوئیه ۲۰۱۶ در مرکز آزمایشگاهی بالینی گوانگژو داآن در نظر گرفته شده است. از این داده‌ها در مطالعهٔ یانگ و همکاران (۲۰۱۸) استفاده شده است. تعداد افراد نمونه در این مجموعه داده ۵۵۱۸ نفر است. بیست‌وهفت مورد از نمونه‌ها، دارای سندرم داون هستند. نمونه‌های تریزومی با روش آمنیوسنتز تأیید شده‌اند. آزمایش‌های NIPT بر اساس پلتفرم توالی‌یابی نیمه رسانا است (لیائو و همکاران، ۲۰۱۴). مجموعه داده دارای متغیرهایی (ویژگی‌هایی) است که برای نشان دادن اطلاعات این داده‌ها فهرست شده‌اند. در این مجموعه داده، ستون اول مربوط به متغیر Z-run است که نشان‌دهندهٔ خروجی امتیاز Z توسط برنامهٔ Seqboost است. ستون دوم، Real-state نتایج تأیید شده و قطعی توسط تشخیص قبل از تولد یا پس از زایمان و ستون‌های سوم تا هشتم مربوط به شش نمرهٔ Z است که نحوهٔ محاسبهٔ آن‌ها در ادامه توضیح داده می‌شود. ستون نهم مربوط به متغیر fetal نشان‌دهندهٔ کسر جنین، یعنی نسبت DNA آزاد متعلق به جفت موجود در گردش خون مادر است. ستون دهم در مجموعهٔ داده‌ها مقدار اوج طول خواندن [۱۷] است که به معنای تعداد نوکلئوتیدهایی است که در هر خواندن توالی‌یابی به‌دست می‌آید، ستون یازدهم با نام read تعداد خواندن‌های منحصر بفرد هر کروموزوم در فرایند توالی‌یابی و ستون دوازدهم سن مادر و ستون سیزدهم نشان‌دهندهٔ هفته بارداری است. از این ۵۵۱۸ داده، ۷۶۶ داده با خوانش منحصر بفرد کمتر از سه میلیون یا پیش‌بینی کسر جنینی کمتر از پنج درصد بر اساس کنترل کیفیت برچسب QC-filtered دارند. ۴۷۵۲ دادهٔ باقی‌مانده برچسب

[17] peak

QC-pass دارند که این برچسب گذاری در ستون چهاردهم نشان داده شده است. همچنین ستون پانزدهم group که نشان‌هندهٔ پیش‌بینی تریزومی توسط امتیاز Z_{run} است. داده‌ها به سه گروه برای کروموزوم ۲۱ طبقه‌بندی شدند: گروه N به‌عنوان کسانی که امتیاز Z_{run} کوچک‌تر از ۱/۹۶ داشتند، یعنی به‌طور قابل‌توجهی بالاتر از خط پایه مجموعه داده مرجع نیستند؛ گروه P به‌عنوان کسانی که امتیاز Z_{run} بزرگتر از ۴ دارند و گروه "طبقه بندی نشده" به‌عنوان افرادی که امتیاز Z_{run} بین ۱/۹۶ تا ۴ دارند به این معنی که برای بررسی دوباره در NIPT نیاز به آزمایش مجدد است که در ستون پانزدهم نشان داده شده است.

در ادامه به محاسبهٔ امتیاز Z_{run} پرداخته می‌شود. تشخیص تریزومی ۲۱ بر اساس محاسبهٔ بازنمایی ژنومی کسری (FGR) کروموزوم ۲۱ که به‌عنوان درصد کروموزوم ۲۱ یا نسبت کروموزوم ۲۱ شناخته می‌شود؛ انجام می‌گیرد. برای تشخیص آنوپلوئیدی برای هر نمونه، ابتدا درصد قرائت‌های منحصربه‌فرد یعنی FGR، تعیین می‌شود. این مقدار با تقسیم تعداد خواندن‌های منحصر به فرد از کروموزوم ۲۱ بر تعداد کل خواندن‌های تصحیح شدهٔ همهٔ اتوزوم‌ها(کروموزوم‌های شمارهٔ ۱ تا ۲۲) انجام می‌شود. برای مطالعهٔ بیشتر به لائو و همکاران (۲۰۱۲) مراجعه شود. انحراف FGR از مقدار مورد انتظار با استفاده از امتیاز Z_{run} که با تفریق میانگین گروه کنترل و تقسیم بر انحراف معیار ($s.d.$) گروه کنترل، محاسبه می‌شود. لازم به ذکر است؛ متغیر Z_{run} همان امتیاز Z بیان شده در مقدمه است که در مجموعهٔ داده با این نام آورده شده است. از طرفی $Z_{baseline-v-n_i} = \frac{\text{mean}_i - \text{mean}(ref_i)}{s.d(ref_{.i})}$ معیار دیگری است که در مطالعات جنینی به کار گرفته می‌شود و در آن mean_i میانگین عمق خواندن، تعداد خواندن‌های توالی‌شده از یک نمونه در موقعیت خاصی در ژنوم یا یک ناحیهٔ خاص، برای کروموزوم i است و $\text{mean}(ref_i)$ میانگین مرجع برای کروموزوم i است. $Z_{baseline-v-n_i}$ به‌معنای نمرهٔ Z نرمال‌شده در رابطه با میانگین نمونه‌های مرجع منفی در کروموزوم i است. به‌عبارتی به‌معنای مقادیر نرمال شدهٔ عمق خواندن نمونه‌های مرجع منفی (نرمال) است. معیاری دیگر به صورت

$$Z_{baseline-vs-p_i} = \frac{\text{mean}_i - \text{mean}(ref_{.i}) * (۱ + fetal\,٪/۲)}{s.d(ref_{.i})} \qquad (۱۳)$$

تعریف می‌شود، که به‌معنای نمرهٔ Z نرمال‌شده نسبت به میانگین داده‌های مرجع مثبت پیش‌بینی شده

است. $fetal\%$ به‌معنای کسر DNA جنین است. داده‌های مرجع مثبت پیش‌بینی‌شده در اینجا برابر با حاصلضرب مقدار میانگین داده‌های مرجع منفی (داده‌های مرجع که نتایج آن‌ها در کروموزوم i نرمال یا به‌عبارتی منفی است) در ضریب $(۱ + fetal\ \% /۲)$ است؛ زیرا فرض بر این است که نیمی از کسر جنینی در هنگام وقوع تریزومی افزایش می‌یابد.

معیار دیگر $Z_{chr-vs-n_i} = \frac{\text{mean}_i - \text{median}(\ \text{mean}(sample_{chr})}{s.d(ref_{.i})}$ است که در آن نمرهٔ Z نرمال‌شده نسبت به مقدار اتوزوم مرجع داخلی محاسبه می‌شود. مجموعه مرجع، میانهٔ تمام میانگین‌های عمق خواندن نرمال‌شده در هر اتوزوم از این نمونه است. به‌همین ترتیب داریم

$$Z_{chr-vs-p_i} = \frac{\text{mean}_i - \ \text{median}(\ \text{mean}(sample_{chr})) * (۱ + fetal\ \% /۲)}{s.d(ref_{.i})} \quad (۱۴)$$

که در آن نمرهٔ Z نرمال‌شده در رابطه با مقدار اتوزوم مرجع داخلی مثبت پیش‌بینی شده است که میانهٔ میانگین‌های داده‌های مرجع مثبت پیش‌بینی شده در هر اتوزوم از این نمونه است. به همین ترتیب

$$Z_{sample-vs-n_i} = \frac{-(\ \text{mean}(ref_{.i}) - \ \text{mean}_i}{s_m MAD_i/\sqrt{window_i}} \quad (۱۵)$$

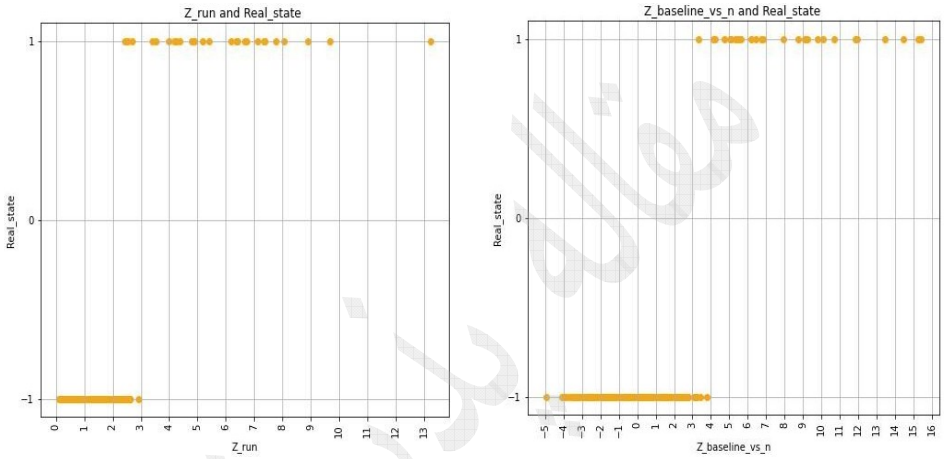
که در آن نمرهٔ Z نرمال‌شده نسبت به میانگین نمونه‌های مرجع نرمال محاسبه می‌گردد. به‌علاوه MAD به‌معنای انحراف مطلق متوسط عمق خواندن، window تعداد پنجره‌های کروموزوم i و s_m ضریبی برابر با ۱/۴۸۲۶است. پنجره کروموزوم به کروموزوم‌های مجاور هم اشاره دارد. درنهایت خواهیم داشت

$$Z_{sample-vs-n_i} = \frac{-(\ \text{mean}(ref_{.i})) * (۱ + fetal\ \% /۲) - \ \text{mean}_i}{s_m MAD_i/\sqrt{window_i}} \quad (۱۶)$$

که در آن نمرهٔ Z در رابطه با میانگین عمق خواندن نمونه‌های مرجع مثبت پیش‌بینی شده است.

۱.۳ پیش پردازش نمونه‌ها

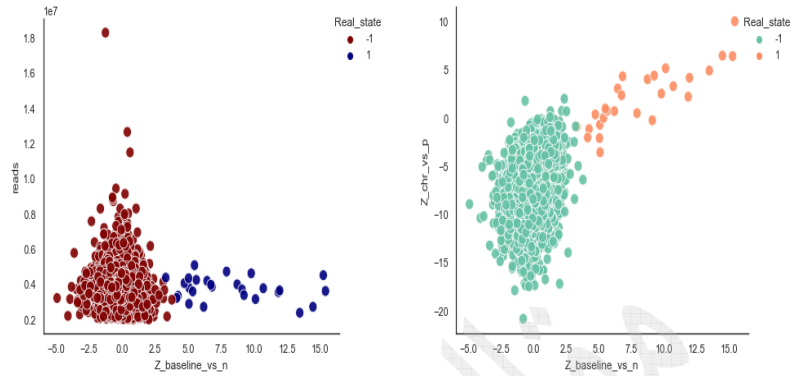
در ابتدا نمودار متغیرها با متغیر هدف یعنی ستون نتیجهٔ واقعی سندرم داون رسم شد تا رابطهٔ ویژگی‌های مختلف با نتیجهٔ نهایی بررسی شود.



(ب) (الف)

شکل ۳: الف–رابطهٔ بین $Z_{baseline-vs-n}$ و نتیجهٔ قطعی سندرم داون، ب– رابطهٔ بین نتیجهٔ قطعی سندرم داون و Z_{run}

شکل ۳ الف رابطهٔ بین نتیجهٔ قطعی سندرم داون و $Z_{baseline-vs-n}$ را نشان می‌دهد. می‌توان در نمودار مشاهده کرد که خط ۴=$Z_{baseline-vs-n}$ می‌تواند مرز بین نمونه‌های مثبت و منفی باشد؛ اما باز هم ممکن است منفی کاذب وجود داشته باشد؛ زیرا نقطه‌ای وجود دارد که سندرم داون مثبت است و امتیاز کمتر از ۴ دارد. به‌علاوه باتوجه به شکل ۳ ب، اگر مقدار Z_{run} کمتر از ۳ باشد؛ با احتمال زیاد نتیجهٔ سندرم داون منفی و اگر بیشتر از ۳ باشد؛ سندرم داون مثبت خواهد بود. اما با توجه به همین نمودار، ممکن است اگر از $Z =$ ۳ برای جداسازی و مرز بین داده‌های مثبت و منفی استفاده شود؛ بخش کوچکی از داده‌ها مثبت یا منفی کاذب تشخیص داده شوند. زیرا آن‌طور که از نمودار مشخص است؛ تعدادی از نمونه‌های مثبت، دارای امتیاز Z_{run} کمتر از ۳ هستند. شکل ۳ نشان می‌دهد برای تشخیص سندرم داون آزمون مبتنی بر Z به تنهایی نمی‌تواند؛ نمونه‌های مثبت و منفی را از یکدیگر تفکیک کند. شکل ۴ الف،



(ب) (الف)

شکل ۴: الف–نمودار مربوط به متغیرهای $Z_{baseline-vs-n}$ و $Z_{chr-vs-p}$ ، ب–رابطهٔ بین تعداد خواندن و $Z_{baseline-vs-p}$

نشان می‌دهد؛ زمانی که مقدار $Z_{baseline-vs-n}$ در بازهٔ ۵- تا ۳/۵ و $Z_{chr-vs-p}$ در بازه ۲۰- تا ۲/۵ است؛ سندرم داون منفی و درغیر این صورت نتیجه مثبت است. این نمودار نشان می‌دهد؛ داده‌ها را می‌توان با خطی از یکدیگر تفکیک کرد. همچنین در شکل ۴ ب، مشاهده می‌شود؛ نمونه‌های مثبت در بازه‌ای قرار گرفته‌اند که تعداد خوانش منحصربفرد کمتر از ۶۰۰۰۰۰۰ و همزمان $Z_{baseline-vs-n}$ در بازهٔ تقریبی ۳ تا ۱۵ است. از روی نمودار می‌توان؛ تشخیص داد که تجمع داده‌ها به چه صورت است و داده‌ها را می‌توان به صورت خطی تفکیک کرد. الگوریتم‌های یادگیری ماشین، روی مجموعه داده در چند حالت مختلف پیاده‌سازی شدند. مدل‌ها یکبار با نرمال‌سازی داده‌ها و سپس بدون نرمال‌سازی داده‌ها آموزش داده شدند. همچنین با جداسازی و بدون جداسازی داده‌های کلاس‌بندی‌نشده، مدل‌های یادگیری ماشین، ساخته شده و نتایج هر یک بررسی شد.

۲.۳ آموزش و ارزیابی مدل ماشین بردار پشتیبان

در ابتدا SVM با تمام داده‌های اصلی و بدون نرمال‌سازی آموزش داده می‌شود. باتوجه به نتایج به‌دست آمده؛ مدل SVM در این حالت، قادر به شناسایی هیچ یک از داده‌های مثبت نیست و تعداد ۲۱ دادهٔ مثبت، همگی منفی تشخیص داده شدند. همچنین زمانی که الگوریتم SVM با تمام داده‌های اصلی و با

بردار ویژگی‌های نرمال‌شده؛ آموزش داده می‌شود؛ نتایج به‌طور قابل توجهی بهبود می‌یابد. مدل ماشین بردار پشتیبان در این حالت تمام ۲۷ دادهٔ مثبت را در داده‌های آزمایشی و آموزش به‌خوبی شناسایی کرده است و تنها یک مثبت کاذب در پیش‌بینی‌ها وجود دارد. در مرحلهٔ بعدی برای غلبه بر داده‌های نامتوازن از کلاسی که دادهٔ کمتری دارد؛ نمونه تولید می‌شود. به‌عبارتی از روش‌های تولید نمونهٔ مصنوعی استفاده شده است. دراین مرحله، ابتدا نمونه‌هایی که برچسب کلاس‌بندی‌نشده[۱۸] دارند و در مجموع ۷۶ نمونه که شامل ۶ نمونهٔ مثبت و ۶۹ نمونهٔ منفی است؛ جداسازی می‌شوند و مدل SVM با باقی‌ماندهٔ داده‌ها، آموزش داده می‌شود. برای این منظور، پارامترهای مختلفی تنظیم شد؛ نتایج مدل‌های مختلف تفاوت چندانی باهم نداشت.

زمانی که از هستهٔ خطی استفاده می‌شود؛ عملکرد مدل مطلوب نیست و تعداد زیادی منفی و مثبت کاذب در پیش‌بینی وجود دارد. پس از آن در آموزش مدل از هستهٔ RBF استفاده می‌شود و پس از تنظیم پارامترهای بهینه و آموزش مدل، نتایج ارزیابی شد. دقت الگوریتم در داده‌های آزمایشی و آموزشی ۱۰۰ درصد است ولی در داده‌های کلاس‌بندی‌نشده که جدا شده بودند؛ هیچ یک از داده‌های مثبت به درستی تشخیص داده نشدند و ۶ منفی کاذب در پیش‌بینی وجود دارد. نتایج نشان می‌دهد؛ در این حالت زمانی که نرمال‌سازی انجام شود؛ نتایج برای داده‌های جداسازی‌شده بسیار بهتر است و فقط یک منفی کاذب در پیش‌بینی وجود دارد. نتایج پیاده‌سازی الگوریتم ماشین بردار پشتیبان در جدول ۱ گزارش شده است.

جدول ۱: نتایج پیاده‌سازی مدل ماشین بردار پشتیبان

نوع الگوریتم	FN	FP	accuracy	$f1$	recall	precision	عملکرد
با تمام داده های اصلی و بدون نرمال‌سازی	۲۷	۰	۰/۹۹	۰/۵	۰/۵	۰/۵	–
با داده‌های اصلی و پس از نرمال‌سازی	۰	۱	۰/۹۹	۰/۹۹	۱	۰/۹۸	–
پس از تولید داده با هستهٔ خطی	۹	۲۶۷۵	۰/۵	۰/۳۴	۰/۵۴	۰/۵	–
پس از تولید داده و بدون نرمال سازی با هسته RBF	۰	۰	۱	۱	۱	۱	FN= ۶
پس از تولید داده و بدون جداسازی داده‌ها	۰	۰	۱	۱	۱	۱	–
پس از تولید داده و نرمال‌سازی و جداسازی داده‌ها	۰	۰	۱	۱	۱	۱	FN= ۱
با داده های اصلی و نرمال سازی و جداسازی داده‌ها	۰		۱	۱	۱	۱	FN= ۱
وزن دهی کلاس ها و با نرمال سازی و بدون جداسازی داده‌ها	۰	۱	۰/۹۹	۰/۹۹	۱	۰/۹۸	–
وزن دهی کلاس ها و نرمال سازی بعد از جداسازی داده‌ها	۰	۰	۱	۱	۱	۱	FN= ۲

[18] Unclassified

۳.۳ آموزش و ارزیابی مدل بیز ساده

الگوریتم بیزساده نیز با تمام نمونه‌ها و بدون نرمال‌سازی، آموزش داده می‌شود. طبق نتایج، مدل بیز ساده نیز در این مرحله قادر به تشخیص هیچ یک از داده‌های مثبت نبوده و تمام آن‌ها را منفی پیش‌بینی می‌کند. همین کار با نرمال‌سازی داده‌ها تکرار شد. نتایج پس از نرمال‌سازی به‌طور قابل توجهی بهتر می‌شود. در مجموع تعداد مثبت کاذب در داده‌های آزمایشی ۷ عدد و در نمونه‌های آموزش ۸ عدد است. در این مدل منفی کاذب وجود ندارد. الگوریتم بیز ساده نیز پس از تولید داده و نرمال‌سازی با جداسازی داده‌های کلاس‌بندی‌نشده آموزش داده می‌شود. نتایج نهایی در جدول ۲ گزارش شده است. در الگوریتم بیز ساده نیز نتایج پس از نرمال‌سازی به‌طور قابل توجهی بهبود می‌یابد. دقت الگوریتم در داده‌های آزمایشی و آموزشی ۱۰۰ درصد و در داده‌های جداسازی‌شده، یک منفی کاذب وجود دارد. در این داده‌ها دقت الگوریتم ۰/۹۸ است.

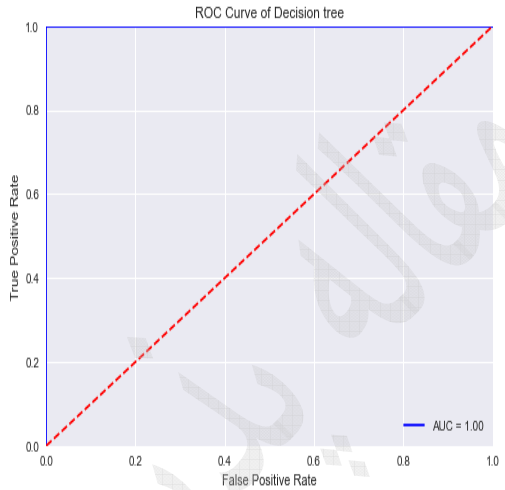
جدول ۲: نتایج پیاده‌سازی مدل بیز ساده

نوع الگوریتم	FN	FP	accuracy	f۱	recall	precision	عملکرد
با تمام داده های اصلی و بدون نرمال‌سازی	۲۷	۰	۰/۹۹	۰/۵	۰/۵	۰/۵	-
با داده‌های اصلی و پس از نرمال‌سازی	۰	۱۵	۰/۹۹	۰/۸۹	۱	۰/۸۲	-
پس از تولید داده و بدون نرمال‌سازی	۱۸	۰	۰/۹۹	۰/۷۵	۰/۶۷	۱	-
پس از تولید داده و با نرمال‌سازی	۰	۱۸	۰/۹۹	۰/۸۷	۱	۰/۸	-
پس از تولید داده و با جداسازی داده‌ها بدون نرمال‌سازی	۱۲	۰	۰/۹۹	۰/۸	۰/۷۱	۱	FN= ۲
پس از تولید داده و نرمال‌سازی و جداسازی داده‌ها	۰	۰	۱	۱	۱	۱	FN= ۱
با داده‌های اصلی و جداسازی داده‌ها و نرمال‌سازی	۰	۲	۰/۹۹	۰/۹۵	۱	۰/۹۶	FN=۱ FP= ۲

۴.۳ آموزش و ارزیابی مدل درخت تصمیم

زمانی که الگوریتم درخت تصمیم با تمام داده‌های اصلی و بدون نرمال‌سازی آموزش داده می‌شود؛ دقت مدل درخت تصمیم روی داده‌های آزمایشی و آموزش ۱۰۰ درصد است و به‌خوبی نمونه‌های مثبت و منفی را از هم تفکیک می‌کند. نتایج قبل و بعد از نرمال‌سازی بردار ویژگی‌ها تفاوتی ندارد و نتایج به‌شرح قبل است. مدل درخت تصمیم پس از جداسازی داده‌هایی که برچسب کلاس‌بندی‌نشده دارند؛ با داده‌های باقی‌مانده، آموزش داده شد. طبق نتایج، دقت مدل در داده‌های آموزش و آزمایشی ۱۰۰ درصد است. دقت مدل آموزش‌داده‌شده در پیش‌بینی داده‌های جدید ۹۷ درصد است و فقط تعداد ۲ منفی کاذب وجود دارد. می‌توان

گفت مدل درخت تصمیم روی مجموعه دادهٔ سندرم داون، عملکرد مطلوبی دارد و می‌تواند با دقت خوبی داده‌هارا از یکدیگر تفکیک کند. در جدول ۳ نتایج حاصل از مدل درخت تصمیم آورده شده است. شکل ۵ منحنی عملکرد این مدل را نشان می‌دهد.



شکل ۵: منحنی عملکرد مدل درخت تصمیم. مساحت سطح زیر نمودار یک است.

جدول ۳: نتایج مدل درخت تصمیم

نوع الگوریتم	FN	FP	accuracy	f۱	recall	precision	عملکرد
با تمام داده های اصلی	۰	۰	۱	۱	۱	۱	–
پس از جداسازی داده‌های کلاس‌بندی‌نشده	۰	۰	۱	۱	۱	۱	FN= ۲

۵.۳ آموزش و ارزیابی مدل جنگل تصادفی

زمانی که الگوریتم جنگل تصادفی با تمام داده‌های اصلی آموزش داده می‌شود؛ دقت مدل جنگل تصادفی برای مجموعهٔ دادهٔ سندرم داون، ۱۰۰ درصد است. بدین معنی‌که الگوریتم توانسته است داده‌های مثبت و منفی را به‌خوبی تفکیک کند و دقت در داده‌های آزمایشی نیز ۱۰۰ درصد است. مدل یادگیری ماشین در مرحلهٔ بعد و پس از جداسازی داده‌هایی که برچسب کلاس‌بندی‌نشده دارند؛ با داده‌های باقی‌مانده آموزش داده شد. طبق نتایج، دقت در داده‌های آموزش و آزمایشی ۱۰۰ درصد است. مدل آموزش داده شده روی

نمونه‌های تفکیک شده بررسی شد. در تشخیص نمونه‌ها تعداد دو منفی کاذب وجود داشت و دقت روی این نمونه‌ها ۰/۹۷ است. نتایج مدل جنگل تصادفی مطابق با جدول ۴ است،

جدول ۴: نتایج مدل جنگل تصادفی

نوع الگوریتم	FN	FP	accuracy	f۱	recall	precision	عملکرد
با تمام داده های اصلی	۰	۰	۱	۱	۱	۱	–
پس از جداسازی داده‌های کلاس‌بندی‌نشده	۰	۰	۱	۱	۱	۱	FN= ۲
پس از جداسازی داده‌های QC_filterd	۰	۰	۱	۱	۱	۱	FN= ۰

۶.۳ آموزش و ارزیابی مدل نزدیک‌ترین همسایه

الگوریتم k نزدیک‌ترین همسایه نیز در چند حالت مختلف پیاده‌سازی شد. زمانی که این الگوریتم پس از نرمال‌سازی آموزش داده می‌شود؛ دقت در داده‌های آزمایشی و آموزش ۱۰۰ درصد و در داده‌های دیده‌نشده ۴ منفی کاذب وجود دارد. همچنین دقت در این نمونه‌ها ۰/۹۴ است. نتایج ارزیابی مدل آموزش داده شده پس از نرمال‌سازی و با مجموعه دادهٔ اصلی و بدون جداسازی داده‌های کلاس‌بندی نشده نشان می‌دهد؛ دقت در داده‌های آموزش ۱۰۰ و در داده‌های آزمایشی ۰/۹۹ است و یک مثبت کاذب در داده‌های آزمایشی دیده می‌شود و میانگین امتیاز f۱ برابر ۰/۹۵ است. طبق نتایج، پس از تولید داده و نرمال‌سازی با جداسازی داده‌های کلاس‌بندی‌نشده در داده‌های آزمایشی ۲ مثبت کاذب و در داده‌های آموزشی ۳ مثبت کاذب وجود داشت. همچنین در داده‌های دیده‌نشده ۵ مثبت کاذب وجود دارد. نتایج کلی طبق جدول ۵ است،

جدول ۵: نتایج مدل k نزدیکترین همسایه

نوع الگوریتم	FN	FP	accuracy	f۱	recall	precision	عملکرد
با تمام داده های اصلی و پس از نرمال‌سازی	۰	۱	۰/۹۹	۰/۹۵	۱	۰/۹۲	–
پس از جداسازی داده‌های کلاس‌بندی‌نشده و بدون نرمال‌سازی	۲۷	۰	۰/۹۹	۰/۵	۰/۵	۰/۵	FN= ۴
پس از جداسازی داده‌های کلاس‌بندی‌نشده و بعد از نرمال‌سازی	۰	۰	۱	۱	۱	۱	FN= ۴
پس از تولید داده و نرمال‌سازی و جداسازی داده‌های کلاس‌بندی‌نشده	۰	۲	۰/۹۹	۰/۹۲	۱	۰/۸۶	FP= ۵

۴ بحث و نتیجه‌گیری

برای مجموعه دادهٔ سندرم داون، از آن‌جایی که تشخیص اشتباه عدم بیماری (منفی) عواقب جبران‌ناپذیری دارد؛ بنابراین سعی بر کاهش نتیجهٔ منفی کاذب است. از میان الگوریتم‌های آموزش داده شدهٔ ماشین بردار پشتیبان؛ مدلی که با داده‌های اصلی و پس از نرمال‌سازی ساخته می‌شود؛ تنها یک مثبت کاذب در پیش‌بینی دارد. در مدل بیزساده عملکرد مطلوب زمانی است که نمونه تولید شده و بردار ویژگی‌ها نرمال می‌شوند. زمانی که داده‌های کلاس‌بندی نشده را جداسازی کرده و مدل را آموزش دادیم و عملکرد مدل روی داده‌های جداسازی‌شده آزمایش شد؛ عملکرد مدل در پیشینی برچسب این داده‌ها ضعف دارد. اما مدل ماشین بردار پشتیبان پس از جداسازی این داده‌ها تنها یک تشخیص اشتباه و مدل‌های درخت تصمیم و جنگل تصادفی دو تشخیص اشتباه در این داده‌ها داشتند. در مدل بیزساده فقط یک مثبت کاذب در پیش‌بینی داده‌های جداسازی شده وجود داشت. اما بدون جداسازی این داده‌ها الگوریتم‌های ماشین بردار پشتیبان، درخت تصمیم، جنگل تصادفی به دقت صد درصدی در داده‌های آموزش و آزمایشی دست یافتند. مدل نزدیک‌ترین همسایه نیز با داده‌های اصلی و پس از نرمال‌سازی تنها یک مثبت کاذب داشت. اگرچه تعدادی از مدل‌ها به دقت صددرصدی دست یافتند؛ اما برای گسترش این مدل‌ها نیاز به تعداد بیشتری نمونه است تا عملکرد مدل در مواجهه با داده‌های جدید بهبود یابد. می‌توان این مدل‌ها را با یکدیگر ترکیب و از نتایج تمام آن‌ها استفاده کرد تا پیش‌بینی، اطمینان بیشتری داشته باشد. همچنین با توجه به این‌که اگر این الگوریتم‌ها در کنار یکدیگر استفاده شوند؛ نتایج نهایی در نتیجهٔ پیش‌بینی تمام مدل‌ها است؛ می‌توان زمانی که داده‌های کلاس‌بندی‌نشده جداسازی و مدل با داده‌های باقی‌مانده آموزش داده می‌شود؛ مدل‌های بیان شده را همزمان با یکدیگر به‌کار برده و مدلی ترکیبی ساخت و در نهایت عملکرد را روی داده‌های جدید بررسی کرد.

در نتیجه با توجه به این‌که الگوریتم‌ها به دقت کامل در پیش‌بینی دست یافتند؛ می‌توان الگوریتم‌های یادگیری ماشین استفاده شده در این مطالعه را در کنار سایر روش‌های تشخیصی برای بهبود نتایج به‌کار برد. به‌علاوه روش‌های مبتنی بر آزمون Z قادر به شناسایی تمام نمونه‌ها نیستند. هم‌چنین در این روش‌ها مثبت و منفی کاذب در تشخیص وجود دارد. اما مدل‌های یادگیری ماشین به دقت مطلوبی در تشخیص نمونه‌ها، دست یافتند. می‌توان از مدل‌های آموزش داده شده در کنار روش‌های تشخیصی حاضر برای تمام

نمونه‌ها بخصوص مواردی که روش‌های حاضر قادر به شناسایی آن‌ها نیستند؛ استفاده کرد. از محدویت‌های این پژوهش می‌توان به دسترسی دشوار به داده‌ها اشاره کرد. همچنین داده‌های مورد استفاده، محدود به نمونه‌های آزمایش NIPT است و داده‌های دیگر مانند آزمایشات سرم مادر و یا داده‌های اولتراسونوگرافی در دسترس نبود. پژوهش حاضر پژوهشی پایه‌ای است. می‌توان از مجموعه داده‌های دیگر مانند آزمایشات سرم مادر و یا داده‌های سونوگرافی استفاده نمود و الگوریتم‌های مورد استفاده در این پژوهش را روی این مجموعه داده پیاده‌سازی کرد. همچنین می‌توان به بررسی سایر الگوریتم‌های یادگیری ماشین و یادگیری عمیق در تشخیص زودهنگام سندرم داون پرداخت. علاوه بر این می‌توان تکنیک‌های یادگیری ماشین را در تشخیص سایر ناهنجاری‌ها به‌کار گرفت و عملکرد این روش‌ها را در تشخیص سایر نانجاری‌ها ارزیابی کرد.

تقدیر و تشکر

نویسندگان بر خود لازم می‌دانند؛ مراتب سپاس خود را از سردبیر محترم، داوران گرامی و ویراستار مجله که با نظرات ارزشمند خود ما را در انجام و ارتقای این مقاله یاری دادند، اعلام نمایند.

مراجع

موسوی، ن. و گلعلی‌زاده، م. (۱۴۰۲)، رویکردی نوین در بکارگیری روش دسته ماشین بردار پشتیبان تصادفی در تحلیل داده های بیان ژن سرطان پروستات، مجله علوم آماری ایران، ۱۷، ۴۷۶-۴۵۹.

Abu-Mostafa, Y. S., Magdon-Ismail, M. and Lin, H.T. (2012), *Learning from Data*, **4**, New York: AMLBook.

Ashoor, G., Syngelaki, A., Wagner, M., Birdir, C. and Nicolaides, K. H. (2012), Chromosome-Selective Sequencing of Maternal Plasma Cell–Free DNA for

First-Trimester Detection of Trisomy 21 and Trisomy 18, *American Journal of Obstetrics and Gynecology*, **206**, 322-e1.

Breiman, L. (2001), Random Forests, *Machine Learning*, **45**, 5-32.

Chiu, R. W., Chan, K. A., Gao, Y., Lau, V. Y., Zheng, W., Leung, T. Y. and Lo, Y. D. (2008). Noninvasive Prenatal Diagnosis of Fetal Chromosomal Aneuploidy by Massively Parallel Genomic Sequencing of DNA in Maternal Plasma, *Proceedings of the National Academy of Sciences*, **105**, 20458-20463.

Daumé, H. (2017), A Course in Machine Learning, 100-103.

Hand, D. J. (2007), Principles of Data Mining, *Drug Safety*, **30**, 621-622.

He, F., Lin, B., Mou, K., Jin, L., and Liu, J. (2021), A Machine Learning Model for the Prediction of Down Syndrome in Second Trimester Antenatal Screening, *Clinica Chimica Acta*, **521**, 206-211.

Kramer, O. (2013), Dimensionality Reduction with Unsupervised Nearest Neighbors, Berlin, 13-23.

Lau, T. K., Chen, F., Pan, X., Pooh, R. K., Jiang, F., Li, Y., Jiang, H., Li, X., Chen, S.H., Zhang, X. (2012), Noninvasive Prenatal Diagnosis of Common Fetal Chromosomal Aneuploidies by Maternal Plasma DNA Sequencing, *The Journal of Maternal-Fetal & Neonatal Medicine*, **25**, 1370-1374.

Liao, C., Yin, A. H., Peng, C. F., Fu, F., Yang, J, x., Li, R., and Zhang, K. (2014), Noninvasive Prenatal Diagnosis of Common Aneuploidies by Semiconductor Sequencing, *Proceedings of the National Academy of Sciences*, **111**, 7415-7420.

Lim, B. H., Adams, L. A. and Lilly, M. M. (2012), Self-worth as a Mediator Between Attachment and Posttraumatic Stress in Interpersonal Trauma, *Interpersonal Violencet*, **27**, 2039-2061.

Morris, J. K. and Alberman, E. (2009), Trends in Down's Syndrome Live Births and Antenatal Diagnoses in England and Wales from 1989 to 2008: Analysis of Data from the National Down Syndrome Cytogenetic Register, *Bmj*, **339**.

Ramanathan, S., Sangeetha, M., Talwai, S. and Natarajan, S. (2018), Probabilistic Determination of Down's Syndrome Using Machine Learning Techniques, IEEE, 126-132.

Shalev-Shwartz, S. and Ben-David, S. (2014), Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press.

Suthaharan, S. (2016), Decision Tree Learning. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, 237-269.

Vapnik, V. (1995), Support-Vector Networks, *Machine Learning*, **20**, 273-297.

Yang, J., Ding, X. and Zhu, W. (2018), Improving the Calling of Non-Invasive Prenatal Testing on 13-18-21-Trisomy by Support Vector Machine Discrimination, *PLoS One*, **13**, e0207840.

Zhang, H. G., Jiang, Y. T., Dai, S. D., Li, L., Hu, X. N. and Liu, R. Z. (2021), Application of Intelligent Algorithms in Down Syndrome Screening During Second Trimester Pregnancy, *World Journal of Clinical Cases*, **9**, 4573.

Early Detection of Down Syndrome Using Machine Learning Algorithms

Sayyareh[1], A., Abdollahzadeh, S.[2]

[1]Department of Computer Sciences and Statistics, K.N. Toosi University of Technology, Tehran, Iran.

[2]Department of Applied Mathematics, K. N. Toosi University of Technology, Tehran, Iran.

Abstract: Non-invasive NIPT test has been used in trisomy 21 screening. However, there is a possibility of misdiagnosis in the methods used to diagnose Down syndrome. Therefore, it is essential to provide a process that can be used alongside these methods to improve efficiency. The main goal of this article is to design a model based on machine learning algorithms for the early diagnosis of Down syndrome. Machine learning algorithms such as support vector machine, simple Bayes, decision tree, random forest, and nearest neighbor, which are frequently used to improve the diagnosis of disorders, have been implemented on the mentioned dataset. The performance of each model on the Down syndrome dataset was investigated, and the most suitable model for this purpose was introduced.

Keywords: Decision tree, Down syndrome, Machine learning, Naive Bayes, Support vector machine.

Mathematics Subject Classification (2010): 62P10, 68W20.